REPUBLIQUE ALGERIENNE DEMOCRATIQUEET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU
FACULTE DE GENIE ELECTRIQUE ET D'INFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

Mémoire

En vue de l'obtention de Master en Informatique option

Ingénierie des Système Informatique

Thème

Mise en place d'une BD NOSQL

Promotrice :

Mme. TAOURI Dalila

Réaliser par :

- M_r : CHERTOUH Jugurtha
- M_r : KERKACHE Ghiles

Mémoire soutenu publiquement le 09/07/2018 devant le jury composé de :

Mme. BOUSSENINA Lila **« Présidente »**

M. RADJA Hakim **« Examinateur »**

Mme. ADOUANE Farida **« Examinatrice »**

2017 / 2018

REMERCIMENT

Nous tenons à exprimer notre profonde gratitude à notre promotrice Mme. TAOURI Dalila De nous avoir suivi durant toutes les phases de notre projet ou il nous a orienté et éclairé par ces précieux conseils.

Nous la prions de bien vouloir agrée le témoignage notre plus vive reconnaissance.

Nous remercions chaleureusement les membres du jury pour l'honneur qu'ils nous font en acceptant de juger notre travail.

Enfin, nous remercions toutes les personnes ayant contribué de près ou de loin au bon accomplissement de notre travail.

Dédicaces

Ce travail, achevé avec l'aide de dieu le tout Puissant,

je le dédie à toutes les personnes que

J'aime :

Aux êtres les plus chers au monde qui n'ont jamais

Cessé de témoigner leurs affections et m'apporter

Leurs soutiens et encouragements depuis mon

existence, mes chers parents (mon père Rabah et ma mère Aldjia) et à ma grand-mere maternnelle.

A la mémoire de mon grand père maternnelle et paternnelle ,et ma grand mere paternnelle ,

A mon cher frère Juba.

A mon adorable sœur, Kamelia.

A toute ma famille,

A tous mes oncles paternels et mes oncles maternels Said, Chabane, mes tantes,

Mes cousins et cousines

A tous mes adorables ami (es) : said, amine, abdesslam, idir, rabah……

A mon cher ami : ghiles,.

A tous mes enseignants en particulier notre promotrice Mme TAOURI .

A toute la promotion d''informatique UMMTO

Année 2017/2018

Jugurtha

Dédicaces

Ce travail, achevé avec l'aide de dieu le tout

Puissant, je le dédie :

A mes chers parents en témoignage de ma reconnaissance pour leur Patience, leur sacrifice et leur soutien tout au long de mes études que dieu leurs prête santé.

A la memeoire de mes grands parents

A mon frère mohamed amine et à ma sœur

A mes cousins et cousines,

A mes meilleurs amis : Jugurtha , amine, abdesslam, idir, rabah……

A tous mes enseignants en particulier notre promotrice Mme TAOURI et mes amis de la promotion INFORMATIQUE

Et à toutes les personnes qui m'aiment.

Année 2017/2018

GHILES

Contents

I État de l'art **10**

1 L'Internet des objets **11**
1.1 Évolution de l'Internet . . . 11
1.2 Concepts de base et définitions de l'internet de objets (IoT) . . . 12
1.2.1 l'Internet des objet (IoT) . . . 12
1.2.2 Les pilliers de l'IoT . . . 12
1.2.2.1 Objets Connectés . . . 13
1.2.2.2 Les Données . . . 16
1.2.2.3 Les personnes . . . 17
1.2.2.4 Les processus . . . 17
1.3 Architecture de l'IoT . . . 19
1.3.1 Construction d'un block IoT: . . . 20
1.3.2 Les capteurs ou Actionneurs: . . . 20
1.3.3 La passerelle (Gateway) IoT: . . . 20
1.3.4 La plateforme cloud et Big Data Analytics: . . . 20
1.4 La sécurité dans l'IoT . . . 21
1.4.1 Les vulnérabilités les plus fréquemment rencontrées sur les objets connectés . . . 21
1.4.2 Les solutions pour se prémunir des attaques ciblant les IoT . . . 22

2 Big Data **24**
2.1 Les données en mouvement . . . 24
2.2 La donnée matière première de la transformation numérique . . . 24
2.3 Les évolutions technologiques vers le Big Data . . . 25
2.4 Le stockage des données . . . 26
2.5 Les acteurs du Big Data: . . . 26
2.5.1 Analyste de données (Data Analyst): . . . 27
2.5.2 Scientifique de données (data scientist) : . . . 27
2.5.3 Data_Engineer: . . . 28
2.6 Les quatres sources du big data : . . . 29
2.6.1 Les « logs » des sites web: . . . 29
2.6.2 Les « insights » des médias sociaux: . . . 30
2.6.3 Les données tierces « third party data » : . . . 31
2.6.4 L'Open data (donnée ouverte): . . . 31
2.7 La Gestion du Big data: (Les 5 V de Big Data) . . . 31
2.7.1 Le volume . . . 32
2.7.2 La Variabilité/Variété . . . 32
2.7.3 la Vélocité . . . 32
2.7.4 La Valeur . . . 33
2.7.5 La Véracité . . . 33
2.8 Les technologies Big Data: . . . 33
2.8.1 Big Data opérationnel: . . . 34
2.8.2 Big Data analytique: . . . 34
2.9 Impacte du big data dans le développement des entrprises: . . . 35
2.9.1 Vers un marketing en temps réel: . . . 36
2.9.2 Optimiser la prospection: . . . 36

2.9.3 Les Big data au secours de la logistique: . . . 36
2.9.4 Fiabilité et contrôle de la production: . . . 37
2.9.5 Améliorer le service informatique: . . . 37
2.9.6 De nouveaux outils pour la prise de décisions: . . . 38
2.10 Cycle de vie du Big Data: . . . 38
2.11 Problématique du stockage du Big Data . . . 39
2.11.1 Stockage HDFS pour le Big Data et Hadoop . . . 39
2.11.2 Stockage en mode objet . . . 40

3 Cloud computing **42**
3.1 L'informatique en nuage (Cloud Computing): . . . 42
3.2 Identification d'un cloud . . . 43
3.3 Caractéristiques du Cloud Computing: . . . 45
3.4 Data Centers . . . 46
3.5 Les Modèles de déploiement . . . 47
3.5.1 Cloud Public . . . 47
3.5.2 Cloud Privé . . . 48
3.5.3 Cloud Communautaire . . . 48
3.5.4 Cloud Hybride . . . 49
3.6 Les Services du Cloud computing: . . . 50
3.6.1 Software as a Service (SaaS): . . . 50
3.6.2 La Platform as a Service (PaaS): . . . 52
3.6.3 Infrastructure as a Service (IaaS) . . . 53
3.7 Architecture cloud computing . . . 54
3.8 Les Technologies du Cloud Computing: . . . 55
3.8.1 La Virtualisation . . . 55
3.8.2 Architecture orientée service (SOA): . . . 56
3.8.3 Calcul en grille (Grid Computing): . . . 57
3.9 Cloud computing et sécurité: . . . 58
3.9.1 Les composants sécurité d'un système de Cloud computing: . . . 58
3.10 Les avantages et inconvénients du cloud computing : . . . 58
3.10.1 Avantages du cloud: . . . 58
3.10.2 Inconvénients du cloud: . . . 59
3.11 Base de données Cloud Computing: . . . 59
3.11.1 NoSQL dans le Cloud: une alternative évolutive aux bases de données relationnelles . . . 60
3.11.2 Critères de migration vers le NOSQL: . . . 61
3.11.3 Critères de choix pour mieux choisir le type de BDD: . . . 61
3.11.4 Défis majeurs des BDD NOSQL: . . . 63

II Conception et Réalisation **65**

4 Analyse et conception **66**
4.1 Étude du contexte . . . 67
4.1.1 Problématique du système étudié . . . 67
4.1.2 Concepts de base liés au contexte à modéliser . . . 68
4.1.3 Les Systèmes de transport intelligents . . . 69
4.1.3.1 Le rôle du systèmes intelligent de gestion du trafic : . . . 69
4.1.3.2 La gestion des feux de circulation . . . 70
4.2 Cycle de vie de la donnée: . . . 71

5 Outils de développements et simulation(Préparation) 72
5.1 Les Différentes Technologies Utilisées . . . 72
5.1.1 hadoop . . . 72
5.1.1.1 Hadoop - Big Data Solutions . . . 72
5.1.1.2 Architecture Hadoop . . . 72
5.2 Installation d'hadoop: . . . 74
5.3 Java : . . . 77
5.3.1 définition . . . 77
5.3.2 Son mode de fonctionnement . . . 78
5.3.3 Entrées et sorties (Java Perspective) . . . 78
5.4 Matériel physique . . . 78
5.4.1 Conception générale du système . . . 79
5.4.2 Description détaillée du système . . . 79
5.5 simulation avec anylogic . . . 81
5.5.1 Environnement de modélisation multi-méthode . . . 81
5.5.2 Initiation sur AnyLogic: . . . 82
5.5.3 Les constructions de langage de simulation fournies par AnyLogic . . . 83
5.5.4 Installation . . . 83
5.5.5 Implémentation: . . . 84
5.5.6 Avantages du simulateur Anylogic: . . . 86
5.5.7 AnyLogic et le langage Java: . . . 86

6 Execussion 87
6.1 Préparation de environnement denone simulation . . . 87
6.2 Simulation de la gestion du trafic routier . . . 92
6.2.1 Simulation de la gestion du trafic routier (cas classique) . . . 93
6.2.2 Simulation de la gestion du trafic routier (après optimisation) . . . 94

List of Figures

1.1.1 Les phases de l'évolution de l'internet. . . . 12
1.2.1 Objets Connectés . . . 13
1.2.2 RFID . . . 16
1.2.3 M2M . . . 18
1.2.4 M2P . . . 18
1.2.5 P2P . . . 19
1.3.1 Architecture IoT . . . 20

2.2.1 les données nouvel OR NOIR. . . . 25
2.5.1 Acteurs du big data . . . 27
2.5.2 Architecture d'une BDD centrée . . . 29
2.6.1 Source du big data. . . . 29
2.6.2 Geopiq pour Instagram . . . 30
2.6.3 carte géographique des données ouvertes dans le monde . . . 31
2.7.1 Les 5 V Du big Data . . . 32
2.8.1 Hadoop MapReduce . . . 35
2.11.1Avantages du stockage en mode objet. . . . 40

3.1.1 **Cloud Computing** . . . 43
3.2.1 Les 4 points permettant d'identifier un Cloud . . . 44
3.5.1 Modèle de déploiement d'un cloud public. . . . 47
3.5.2 Modèle de déploiement d'un cloud privé. . . . 48
3.5.3 Modèle de déploiement d'un cloud communautaire . . . 49
3.5.4 Modèle de déploiement d'un cloud hybride . . . 50
3.6.1 service cloud computing . . . 50
3.6.2 Les services offerts par un Saas. . . . 51
3.6.3 PaaS vs IaaS . . . 52
3.6.4 Les services qu'offre chaque modéle. . . . 54
3.7.1 Architecture du cloud computing. . . . 55
3.8.1 Virtualisation. . . . 55
3.8.2 architecture orientée service cloud computing. . . . 57
3.8.3 Grid Computing . . . 57
3.11.1théorème CAP . . . 60
3.11.2BBD orientée documents . . . 62
3.11.3BDD orientée colomne . . . 62
3.11.4BDD orientée graphe . . . 63
3.11.5BDD orientée clé-valeur . . . 63

4.0.1 Démarches de la conduite de projet . . . 66
4.1.1 Modèle de carrefour . . . 69
4.1.2 Plan du feu . . . 70
4.2.1 Cycle de vie de la donnée . . . 71

5.1.1 Architecture Hadoop . . . 73
5.2.1 Mise a jour des paquets . . . 74
5.2.2 Installation java . . . 74
5.2.3 version du java . . . 74

5.2.4 output java version . . . 74
5.2.5 téléchargement hadoop-2.7.3 . . . 75
5.2.6 vérification de la version . . . 75
5.2.7 vérifiez la valeur SHA-256 . . . 76
5.2.8 Correspondance . . . 76
5.2.9 extraire le fichier . . . 76
5.2.10déplacer le fichier extrait vers l'emplacement / usr / local . . . 76
5.2.11obtenir le chemin java par défaut . . . 77
5.2.12ouvrez le hadoop-env.sh . . . 77
5.2.13Configuration d'une valeur statique . . . 77
5.2.14Utiliser le lien de lecture directement . . . 77
5.4.1 synoptique représentatif de la conception du feu intelligent . . . 79
5.4.2 Raspberry PI 2 . . . 80
5.4.3 Raspberry PI 2 avec module caméra embarquée . . . 80
5.5.1 simulation multi-agents . . . 81
5.5.2 dynamique système . . . 82
5.5.3 simulation par évènement discret . . . 82
5.5.4 Téléchargement . . . 83
5.5.5 chois de version . . . 84
5.5.6 Remplissage de formulaire . . . 84
5.5.7 création du model . . . 85
5.5.8 objet à simulé . . . 85
5.5.9 rend point . . . 86

6.2.1 Simulation avant l'optimisation(cas classique) . . . 94
6.2.2 Graphe montrant l'évolution du temps moyen . . . 95
6.2.3 Résultat de l'optimisation . . . 96
6.2.4 Les données nouvel Or Noir . . . 98

List of Tables

1.1 Exemples des objets connectés 14
1.2 capteurs 15
1.3 Contrôleurs 16

2.1 la mesure de la croissance des données produites 26
2.2 Big data et analytique classique 36

3.1 les data centers 46

5.1 Téléchargement 75
5.2 L'infrastructure de MapReduce 78

6.1 Etape1: Dessin des routes nord et sud 88
6.2 Etape2: Ajouter une animation 3D 89
6.3 Etape3: Dessiner la route Est pour former une intersection 90
6.4 Etape4: noneAjout du parking 91
6.5 Etape 5: Ajout des bus et de l'arret de bus 91
6.6 Etape6 Ajout des feux de signalisation 92
6.7 Etape 7 statistiques et optimisations. 92
6.8 Paramètres de base 93
6.9 Collecte des données et statistiques 93
6.10 Phase d'optimisation 94
6.11 Tableau montrant l'évolution du temps moyen 95

Motivations et objectifs

Les technologies les plus profondément enracinées sont les technologies invisibles. Elles s'intègrent dans la trame de la vie quotidienne jusqu'à ne plus pouvoir en être distinguées.» ***Mark Weiser***

*"**La création d'un "Internet des Objets", le développement et la diffusion ubiquitaire des technologies basées sur les capteurs, vont à terme brouiller les frontières entre monde virtuel et monde physique et pourraient modifier la nature même de la vie privée. Les enjeux de sécurité sur le long terme restent encore à analyser et à résoudre…** ".[3]*

Le développement de l'internet et la multiplication des objets connectés à travers le monde s'accompagnent d'une croissance exponentielle des données sur la toile. La multiplication des moyens de communication et d'échange n'y est pas étrangère ; en effet, les différents écrans nous suivent partout, tout au long de la journée. En 2011, il y avait près de 9 milliards de terminaux connectés dans le monde et ce chiffre devrait s'élever à 24 milliards en 2020, si l'on en croit cette étude de Valtech[1].

Outre les smartphones, tablettes et télévision connectées, les nouveaux objets connectés, tels que les voitures, les appareils électroménagers ou encore les montres connectées qui déferlent sur le marché devraient remonter une quantité phénoménale d'informations dans les années à venir. Si l'on en croit les résultats d'une étude récente, le monde a manipulé en 2012 plus de 2,8 zetaoctets d'informations soit 2,8 milliards de gigaoctets, ce chiffre est colossal, mais ***le plus intéressant dans cette étude est de savoir que seul 0,5% de ces 2,8 zetaoctets ont été analysés d'une manière ou d'une autre alors que l'étude estime que 25% d'entre elles représentent une valeur potentielle pour des entreprises.*** Les informations disponibles sur Internet ne sont plus seulement volumineuses, elles sont également très diverses, non structurées au sens où ***elles ne se présentent pas sous la forme de lignes et de colonnes comme aime à être structuré le web.***

Ces données doivent donc être structurées avant d'être analysées et exploitées par les technologies actuelles. Toutes leurs interactions avec les nouvelles technologies génèrent des données : téléchargement d'un fichier, consultation d'une vidéo, coup de téléphone, envoi de SMS, utilisation de GPS... ce n'est pas tant les interactions en tant que telles qui génèrent autant de données, mais c'est surtout l'ensemble des informations annexes (les métadonnées) et des communications cachées entre différents serveurs (publicitaires par exemple) qui ont lieu au même moment qui génèrent un flux impressionnant de données. ***L'ensemble de ces milliards de données, représente ce que l'on appelle communément les Big data***. Les premières entreprises à avoir compris leur intérêt sont les géants du web actuel tels que Google, Yahoo, Microsoft, Facebook ou bien Amazon. Du fait de leur succès ou de leur volonté de vouloir gérer une quantité très élevée d'informations, ***elles ont dû apprendre à maitriser ces Big data, car les outils et méthodes traditionnelles ne leur suffisaient plus***, et ont pour cela développé des technologies en interne qui sont désormais, pour la plupart, disponibles dans des versions libres et gratuites (open source).

Ces technologies sont assez récentes et sont pour la plupart issues des grandes sociétés du web américain telles que Google, Yahoo ou encore Linkedin. Ces dernières ont dû créer pour leurs propres besoins un ensemble d'outils afin de traiter les masses de données qu'ils devaient analyser chaque jour. La plupart d'entre elles ont été rendues publiques via une licence Open Source ou données à la fondation

[1] Valtech est une agence de marketing numérique et technologique créée en 1993. Le groupe est présent dans 14 pays (France, Royaume-Uni, Allemagne, Suède, Danemark, États-Unis, Inde, Singapour, Australie, Canada, Argentine, Pays Bas) et compte environ 1 800 collaborateurs. Valtech signifie littéralement (val)orisation par les (tech)nologies.

Apache pour pouvoir être réutilisées et améliorées par la communauté, tout le monde bénéficie donc de ce cercle vertueux.

Des milliards de données d'activités, de ressenties, d'intentions, ou juste de comportement sont stockées via différentes applications opérationnelles, ou outils et autres médias de tous les jours. Et les trois quarts de ces données ont été créés par les utilisateurs–consommateurs–patients que nous sommes. L'exploitation de ces données peut s'avérer complexe, et ne peut se résumer à un simple projet technologique de migration de données ou d'architecture informatique.

Traiter l'ensemble de ces données requiert un travail colossal. L'atteinte de cet objectif passe par la maîtrise et l'apprentissage des données accumulées, de leurs analyses, et des retours métiers qui vont améliorer le modèle, le rendre plus apte à anticiper.

Problématique générale

Etant donné l'aspect particulier des big data, les méthodes de traitement, de stockage et d'analyse classique ne sont pas assez efficaces, ce qui engendre des problématiques quant aux méthodes à utiliser pour parvenir à des résultats satisfaisant :

- **Comment valoriser et traiter un amas de données ?**

  Travailler les données pour les valoriser et obtenir des résultats, nécessite de l'intelligence, l'outil et de la réflexion. Le Big data n'a rien à voir chez une entreprise A ou avec celle de l'entreprise B et pourtant ils peuvent être clients de l'un et de l'autre. Le traitement doit être effectué de façon stratégique en fonction de la nature des données et de leur utilisation.

- **Comment stocker les données de manière à faciliter l'accès ?**

  Les SGBD classiques ne sont pas adaptés pour ce genre de stockage, il est préférable d'utiliser les BDD **non relationnelle.**

- **Comment exploiter ces données afin d'en tirer un maximum de profit ?**

  En effet, les entreprises peuvent exploiter le big data pour comprendre ce que leur clientèle attend vraiment d'eux et ainsi améliorer leurs produits afin de faire plus de bénéfices.Et dans notre cas elles peuvent utiliser les données récoltées afin de bien se situer par rapport à leur clientelle.

La première chose qui vient à l'esprit quand on parle de **Big Data** et d'**IoT** est l'augmentation du volume de données **qui va frapper le cadre de stockage de données des entreprises**. Les centres de données devront être configurés pour gérer toute cette charge de données supplémentaire.

Compte tenu de l'impact considérable de l'IoT sur l'infrastructure de stockage de données, les entreprises ont commencé à se tourner vers le modèle **P**latform-**a**s-**a-S**ervice, une solution basée sur le **cloud**, plutôt que de maintenir leur propre infrastructure de stockage. Contrairement aux systèmes de données internes qui doivent être constamment mis à jour à mesure que la charge de données augmente, PaaS offre flexibilité, évolutivité, conformité et architecture sophistiquée pour stocker toutes les données IoT de grande valeur.

Les options de stockage dans le cloud incluent les modèles publics, privés et hybrides. Si une entreprise possède des données sensibles soumises à des exigences de conformité réglementaire qui nécessitent une sécurité accrue, l'utilisation d'un cloud privé constitue la meilleure solution. Pour les autres entreprises, un cloud public ou hybride peut être utilisé pour le stockage des données IoT.

Pour notre part on essaye d'apporter une modeste participation en essayant d'apporter des solutions à cette problématique est ce à travers un exemple d'application qui est: ***l'optimisation du trafic routier*** qui sera confronté au problème suivant:

- ***Le calcul du nombre variable de véhicules dans chacune des voies de l'intersection.***
- ***La réduction du temps nécessaire à un véhicule pour quitter l'intersection.***

Objectifs

Dans le cadre de notre travail, nous allons nous intéresser au traitement et à l'analyse des données massives plus précisément à la manière utiliser et valoriser une donnée brute.

On se fixera l'objectif d'améliorer le rendu du trafic routier en exploitant les données récoltées via des capteurs positionnées dans une intersection.

Pour atteindre cet objectif nous devrons tout d'abord: étudier le contexte de ce travail à savoir l'IoT, le Big Data et le cloud computing puisque les technologies existante et récentes qui nous permettrons de mettre en évidence une solution qui répond à la problématique suivante:

- Chercher et proposer un modèle de stockage qu'on évaluera grâce à une simulation.

Pour ce faire ce mémoire se présente comme suit:

1. **Première partie:** Un état de l'art dans lequel nous traiterons les chapitres décrivant le contexte de ce mémoire.
    - Analyse et problématique.
    - L'internet des objets
    - Big Data
    - Cloud Computing
2. **Deuxième partie:** Conception et réalisation de la solution dans laquelle nous présenterons notre système pilote, la problématique et la simulation de la solution, nous terminerons par les testes et une évaluation.
    - Les outils technologiques choisis pour réaliser la solution.
    - Mise en œuvre (simulation) de la solution.

Part I

État de l’art

Chapter 1

L'Internet des objets

Introduction

Les objets connectés s'implantent petit à petit dans notre quotidien. Le marché de ces objets représentait déjà 655.8 milliards de dollars en 2014 et de nombreuses études démontrent que ce chiffre va connaître un essor fulgurant dans les prochaines années. Ainsi, d'après le cabinet IDC(International Data Conseil)[1], le marché des objets connectés devrait représenter plus de 1700 milliards de dollars d'ici à 2020 soit près de 50 milliards d'objets connectés dans le monde. En France, GfK(Gesellschaft für Konsumforschung)[2], annonce que 2 milliards d'objets connectés devraient être vendus d'ici 2020. Chaque foyer disposerait alors d'une trentaine d'objets connectés[3]. Mais qu'est-ce qu'un objet connecté et comment expliquer cet engouement ?

ce chapitre est consacré essentiellement à présenter les concepts et définitions de ce **triptyque *(Cloud – Big Data – IoT***)qui vont contourner le monde d'ici **2020.**

D'une manière générale, lorsque des personnes parlent d'Internet, ils ne se réfèrent pas aux connexions physiques présentes dans le monde réel. En revanche, ils ont tendance à considérer Internet comme un ensemble de connexions dénuées de forme. C'est en effet l'endroit où les gens se rendent pour chercher ou partager des informations. Il s'agit par exemple à la fois de la bibliothèque du 21è siècle, d'un magasin de vidéos et d'un album photo personnel.

1.1 Évolution de l'Internet

L'évolution d'Internet a connu quatre phases distinctes : La connectivité, l'économie en réseau, l'expérience en collaboration et l'internet des objets. Ces phases se sont suc-cédées chronologiquement sur une cinquantaine d'années. (voir FIG) :

[1] IDC:le premier fournisseur mondial de renseignements, de conseils et d'événements sur les marchés des nouvelles technologies

[2] GfK: société pour la recherche sur la consommation »), créé en 1934

[3] statistique concernant l'Europe ou la France

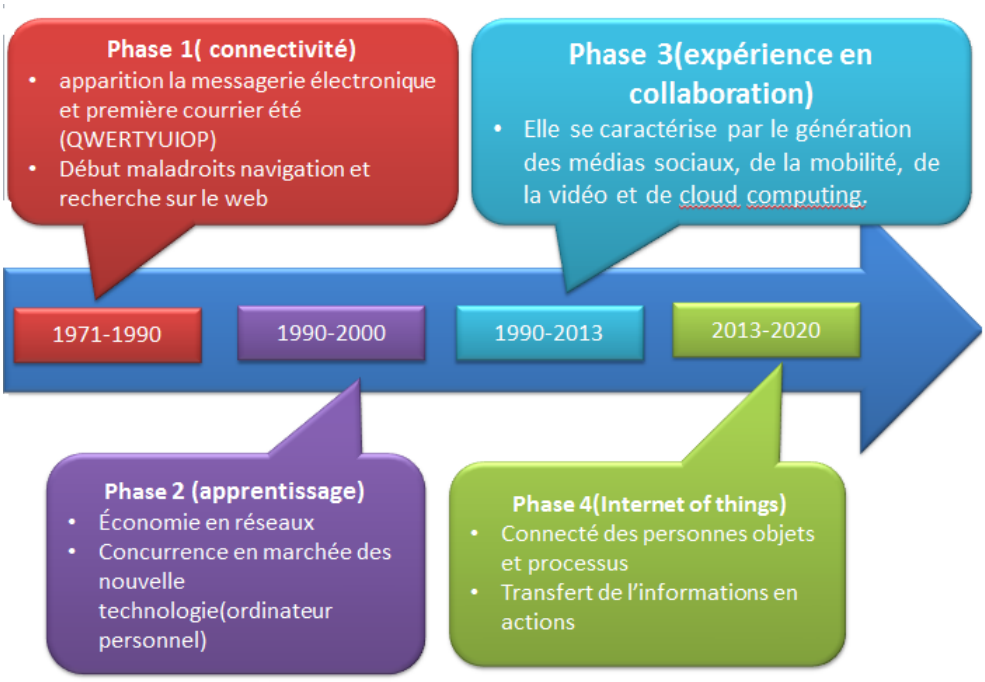


Figure 1.1.1: Les phases de l'évolution de l'internet.

1.2 Concepts de base et définitions de l'internet de objets (IoT)

Dans cette section nous allons donner quelques conceptes se rapportant à l'IoT à savoir : l'internet des objet et ses pilliers.

1.2.1 l'Internet des objet (IoT)

Définition 1. IoT (Internet of Things) est un système d'automatisation et d'analyse avancé qui exploite la technologie de réseautage, de détection, de **big data** et d'intelligence artificielle pour fournir des systèmes complets pour un produit ou un service. Ces systèmes permettent une plus grande transparence, contrôle et performance lorsqu'ils sont appliqués à n'importe quelle industrie ou système.Les systèmes IoT ont des applications dans toutes les industries grâce à leur flexibilité unique et leur capacité à s'adapter à n'importe quel environnement. Ils améliorent la collecte de données, l'automatisation, les opérations et bien plus encore grâce à des dispositifs intelligents et à une puissante technologie habilitante.[4]

Définition 2. Dans un premier temps, voyons plus précisément ce à quoi fait référence le terme d'objet connecté aussi appelé internet des objets ou encore IdO. La définition la plus courante est: L'internet des objets est un réseau de réseaux qui permet, via des systèmes d'identification électronique normalisés et unifiés, et des dispositifs mobiles sans fil, d'identifier directement et sans ambiguïté des entités numériques et des objets physiques et ainsi de pouvoir récupérer, stocker, transférer et traiter, sans discontinuité entre les mondes physiques et virtuels, les données s'y rattachant. Ces petits objets font appel à différentes technologies. Le langage n'étant pas normalisé, chaque constructeur utilise la technologie qu'il souhaite : La radio identification, la communication en champ proche, les capteurs, le Bluetooth ou encore la wifi. Le processus repose sur un système sophistiqué qui repose sur le lien entre le stockage et le traitement de la data. [3]

1.2.2 Les pilliers de l'IoT

L'IoT repose sur quatre piliers visant à rendre les connexions réseau plus efficaces et plus utiles qu'auparavant, à savoir les **personnes**, les **processus**, les **données** et les **objets**. Les informations issues de ces connexions conduisent à des décisions et des actions qui créent de nouvelles possibilités, des expériences plus riches et des opportunités économiques sans précédent, et ce, pour les utilisateurs, les entreprises et les pays.[3]

1.2.2.1 Objets Connectés

Définition. Les objets connectés sont dotés d'une technologie intégrée qui leur permet d'interagir avec des serveurs internes et leur environnement externe.Ces Objets sont capable de communiquer avec un autre objet (souvent un smartphone, une tablette ou un ordinateur) par l'intermédiaire d'une plate-forme réseau sécurisée, fiable et disponible. Cette communication permet à l'objet d'envoyer ou de recevoir des informations via une connexion Internet(d'ou l'internet des objets ou l'Internet of Things **IoT**). L'intérêt de cette interactivité est de pouvoir récupérer des informations, d'en tirer des statistiques, de créer des règles ...etc.[7]

Exemple: *la montre connectée. Elle relève des informations (nombre de pas, rythme cardiaque, ...etc.) pour les envoyer sur un smartphone. Ce dernier nous montre ensuite les résultats sous forme de statistiques et nous donne des conseils personnalisés en fonction de nos performances.*

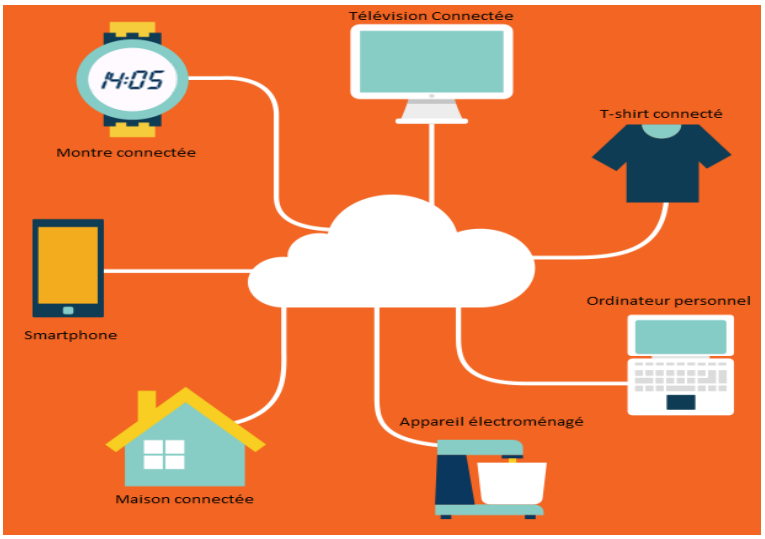


Figure 1.2.1: Objets Connectés

1. **Exemples d'objets connectés (les objets connectés du marché)**

   Les gens utilisent chaque jour de plus en plus des périphériques mobiles pour communiquer entre eux et effectuer des tâches quotidiennes, comme se **renseigner** sur la météo ou effectuer des opérations bancaires en ligne. À l'avenir, de nombreux objets présents dans la maison seront également connectés à Internet et il sera par conséquent possible de les contrôler et de les configurer à distance.Il existe également de nombreux périphériques connectés dans le monde extérieur et qui peuvent fournir des informations pratiques, utiles, voire même essentielles. **Nous présentons dans ce qui suit quelques exemples d'objet connectés :[7]**

Les smartphones peuvent se connecter à Internet presque partout. Ils réunissent les fonctions de nombreux autres appareils différents : téléphone, appareil photo, récepteur GPS, lecteur multimédia, ordinateur à écran tactile.

Une smartwatch peut se connecter à un smartphone pour transmettre à l'utilisateur les alertes et les messages qu'il reçoit. D'autres fonctions, telles que la surveillance du pouls et le comptage des pas, comme un podomètre, peuvent aider à contrôler l'état de santé.

Bon nombre **de voitures** modernes peuvent se connecter à Internet pour accéder aux cartes, à du contenu audio et vidéo ou aux informations sur une destination. Elles peuvent même envoyer un SMS ou un e-mail en cas de tentative de vol ou appeler les secours en cas d'accident. Ces voitures peuvent également se connecter aux smartphones et aux tablettes pour afficher des informations sur les différents systèmes du moteur, fournir des alertes relatives à l'entretien ou indiquer l'état du système de sécurité.

Les lunettes Google, véritable ordinateur miniature, sont dotées d'un tout petit écran qui renseigne la personne qui les porte, un peu comme le dispositif de visualisation tête haute du pilote d'un avion de combat. Un petit pavé tactile latéral permet de parcourir les menus sans cesser de regarder à travers les lunettes Google

Les appareils électroménagers tels que les réfrigérateurs, les fours et les chauffe-eau peuvent être connectés à Internet. Ainsi, le propriétaire de la maison peut les allumer ou les éteindre, contrôler l'état de l'appareil et également recevoir des alertes en fonction des conditions prédéfinies.

Les dispositifs médicaux tels que les pacemakers, les pompes à insuline et les systèmes de monitoring des hôpitaux renseignent ou alertent les patients ou le personnel médical lorsque les signes vitaux atteignent des niveaux spécifiques.

Table 1.1: Exemples des objets connectés

2. **Les Capteurs**

   Le matériel le plus important dans l'IoT pourrait être ces capteurs. Ces appareils comprennent des modules d'énergie, des modules de gestion de l'alimentation, des modules RF(Radio Frequency) et des modules de détection. Les modules RF gèrent les communications grâce à leur traitement du signal, WiFi, ZigBee, Bluetooth, émetteur-récepteur radio et duplexeur.Le module de détection gère la détection par l'intermédiaire de divers dispositifs de mesure actifs et passifs. Voici une liste de certains des appareils de mesure utilisés dans l'IoT :

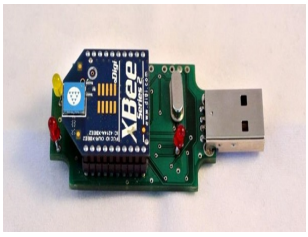
	
Capteur XBee	B+B SmartWorks capteurs sans fil[4]
	
accéléromètre	capteurs de pression
	
magnétomètres:Capteurs de proximité	capteurs d'humidité

Table 1.2: capteurs

3. **Les contrôleurs**

   Les capteurs peuvent être programmés pour prendre des mesures, convertir les données obtenues en signaux, puis envoyer ces données à un périphérique principal appelé contrôleur. Le rôle du contrôleur est de collecter les données en provenance des capteurs et de fournir une connexion Internet. Les contrôleurs peuvent parfois prendre des décisions immédiates ou envoyer les données à un ordinateur plus puissant à des fins d'analyse. Cet ordinateur peut se trouver dans le même réseau local que le contrôleur ou n'être accessible que par l'intermédiaire d'une connexion Internet.[3]

   Pour atteindre Internet, puis les ordinateurs plus puissants situés dans le data center de la figure, le contrôleur commence par envoyer les données à un routeur local. Ce routeur joue le rôle d'interface entre le réseau local et Internet, et peut transmettre des données entre les deux.

[4] `http://www.distrimedia.fr/wzzard-iIoT-solution.html`

contrôleur IoT TJA560

Contrôleur D'alarme S150[5]

Table 1.3: Contrôleurs

4. **Les RFID**

   Un type de capteur populaire utilise l'identification par radiofréquence (RFID). La technologie RFID utilise des champs électromagnétiques de radiofréquence pour échanger des informations entre de petites étiquettes codées (étiquettes RFID) et un lecteur RFID. En général, ces étiquettes servent à identifier et à suivre ce sur quoi elles ont été implantées, par exemple un animal domestique. Ces étiquettes étant de petite taille, elles peuvent être attachées quasiment n'importe où, y compris sur des vêtements ou de l'argent. Certaines d'entre elles ne contiennent pas de pile. L'énergie nécessaire à l'étiquette pour la transmission des informations provient de signaux électromagnétiques envoyés par le lecteur. L'étiquette reçoit ces signaux et utilise une partie de son énergie pour envoyer la réponse.

   Grâce à leur flexibilité et à leur faible consommation d'énergie, les étiquettes RFID constituent un moyen idéal pour connecter un périphérique autre qu'un ordinateur à une solution IoT en fournissant des informations à un lecteur RFID. Il est par exemple aujourd'hui très courant de trouver des usines automobiles dans lesquelles on installe des étiquettes RFID sur les carrosseries des voitures. Et ce, afin de permettre un meilleur suivi des véhicules tout le long de la chaîne de montage.

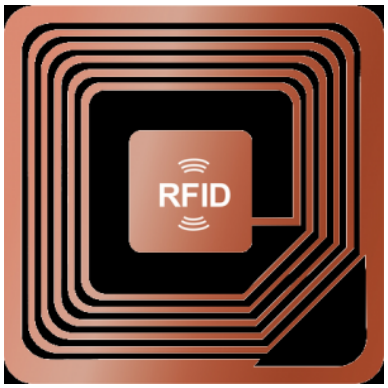


Figure 1.2.2: RFID

1.2.2.2 Les Données

Les données sont partout, les données sont une valeur affectée à tout ce qui nous entoure.Toutefois, les données seules peuvent être sans signification. Lorsqu'on interprète des données, par exemple en les corrélant ou en les comparant, elles deviennent alors plus utiles. Ces données utiles constituent désormais des informations. Une fois ces informations appliquées ou comprises, elles deviennent de la connaissance.

[5] https://m.fr.aliexpress.com/item/32817922359.html

Structuration des données

Les données structurées et non structurées sont des ressources précieuses pour les personnes, les organisations, les industries et les gouvernements. Comme d'autres ressources, les informations collectées à la fois à partir des données structurées et non structurées possèdent une valeur mesurable. Toutefois, la valeur de ces données peut augmenter ou diminuer selon la manière dont ces données sont gérées. Même les données les plus intéressantes perdent de la valeur au fil du temps.

Il est important pour les organisations de collecter toutes les formes de données (structurées, non structurées) et de déterminer des moyens de formater ces données afin de pouvoir les gérer et les analyser.

1. **Données structurées**

   Les données structurées sont des données qui sont entrées et mises à jour dans des champs fixes au sein d'un fichier ou d'un enregistrement. Les données structurées sont facilement entrées, classées, interrogées et analysées par un ordinateur. Par exemple, lorsqu' on entre un nom, une adresse et des données de facturation sur un site Web, on crée des données structurées.

   Afin de minimiser les erreurs et de faciliter l'interprétation des données par l'ordinateur, un format spécifique est requis lors de l'acquisition de ces données. **Exemple:** *Les données personnels d'un client d'une banque.*

2. **Données non structurées**

   Les données non structurées ne possèdent pas l'organisation que l'on retrouve dans les données structurées. Les données non structurées sont des données brutes. Il n'est par conséquent pas possible d'identifier leur valeur. Les données non structurées ne présentent pas de moyen défini permettant de les saisir, de les regrouper et de les analyser. *Comme exemples de données non structurées, on peut citer des photos, des fichiers audio et des vidéos.*

1.2.2.3 Les personnes

Les personnes doivent être connectées Les données seules ne servent à rien. Un grand nombre de données auxquelles personne ne peut accéder ne sert à rien. L'organisation de ces données et leur transformation en informations utilisables permettent aux personnes de prendre des décisions en meilleure connaissance de cause et d'adopter les mesures appropriées. Cela crée de la valeur économique dans une économie activée par l'IoT.

C'est pourquoi les personnes en constituent l'un des quatre piliers. Elles sont au cœur de tout système économique. Elles interagissent en tant que producteurs et consommateurs, l'objectif étant d'améliorer le bien-être par la satisfaction des besoins des êtres humains. Qu'il s'agisse de connexions de personne à personne (P2P), de machine à personne (M2P) ou de machine à machine (M2M), toutes les connexions, ainsi que les données générées à partir de celles-ci, sont utilisées pour créer de la valeur ajoutée pour les personnes.

Remarque: Les informations transforment le comportement d'après John Chambers« L'IoT n'est pas du tout une question de technologie. Il concerne la manière avec laquelle nous modifions les vies des gens. *», John Chambers, PDG de Cisco Systems.[3]*

La valeur est une mesure des avantages offerts par un système économique. Ce sont les personnes qui déterminent la valeur des offres par le biais d'un système d'échange. Il est important de souligner que si les données et les analyses sont importantes, c'est le jugement des personnes qui transforme les données en opinions, et les opinions en valeur de l'IoT.

L'IoT permet l'obtention d'informations précises et opportunes, susceptibles de provoquer une modification du comportement humain, et ce, au bénéfice de l'ensemble des personnes. Il facilite les commentaires qui permettent aux individus de prendre des décisions en toute connaissance de cause, diminuant ainsi les différences entre les résultats souhaités et réels. Cela porte le nom de boucle de rétroaction. Une boucle de rétroaction fournit des informations en temps réel, basées sur le comportement actuel, puis délivre des informations exploitables afin de modifier ce comportement.[9]

1.2.2.4 Les processus

Le quatrième pilier concerne les processus. Les processus jouent un rôle important dans la manière dont les autres piliers, à savoir les objets, les données et les personnes, interagissent afin de fournir de la

valeur dans le monde connecté de l'IoT.Internet a révolutionné la manière avec laquelle les entreprises gèrent leurs chaînes d'approvisionnement ainsi que celle avec laquelle les consommateurs effectuent leurs achats. Bientôt, nous disposerons d'une visibilité sur les processus jamais atteinte auparavant. Cela fournira des opportunités permettant de rendre ces interactions plus rapides et plus simples. Avec le processus adéquat, les connexions deviennent pertinentes et ajoutent de la valeur, car la bonne information est livrée à la bonne personne au bon moment, et de la manière la plus appropriée.Les processus facilitent les interactions entre les personnes, les objets et les données. Aujourd'hui, l'IoT rassemble tous ces éléments en combinant des connexions de machine à machine (M2M), de machine à personne (M2P) et de personne à personne (P2P).

1. **Connexions M2M**

   Les connexions de machine à machine (M2M) se produisent lorsque des données sont transférées d'une machine (ou « objet ») à une autre sur un réseau. Les machines incluent des capteurs, des robots, des ordinateurs et des périphériques mobiles. Ces connexions M2M sont souvent appelées « Internet of Things » (ou Internet des objets.

   Exemple de connexion M2M : une automobile connectée signale que le conducteur est presque arrivé chez lui, invitant ainsi le réseau domestique de cette personne à ajuster la température et l'éclairage dans la maison.

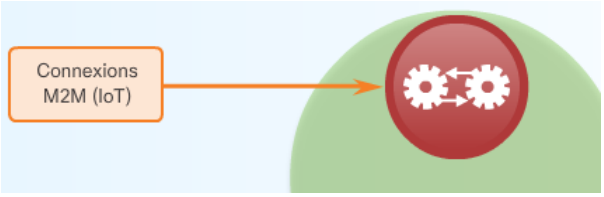


Figure 1.2.3: M2M

2. **Connexions M2P**

   Les connexions de machine à personne (M2P) se produisent lorsque des informations sont transférées entre une machine (par exemple un ordinateur, un périphérique mobile ou une signalisation numérique) et une personne, comme l'illustre dans la FIG . Lorsqu'une personne accède à des informations situées dans une base de données ou qu'elle effectue une analyse complexe, il s'agit d'une connexion M2P. Ces connexions M2P facilitent le déplacement et la manipulation des données, ainsi que la création de rapports, à partir de machines, afin d'aider les utilisateurs à se forger des opinions en toute connaissance de cause. Les actions entreprises par ces personnes sur la base de ces jugements éclairés complètent la boucle de rétroaction de l'IoT.

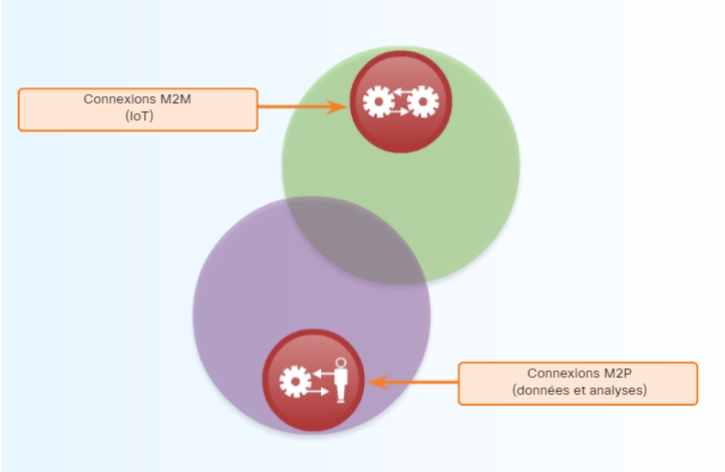


Figure 1.2.4: M2P

3. **Connexions P2P**

Des connexions de personne à personne (P2P) se produisent lorsque des informations sont transférées d'une personne à une autre. De plus en plus, les connexions P2P ont lieu par l'intermédiaire de la vidéo, des périphériques mobiles et des réseaux sociaux. Les connexions P2P sont souvent appelées « collaboration ».

Comme l'illustre la figure, la valeur de l'IoT est maximale lorsque le processus facilite l'intégration de connexions M2M, M2P et P2P.

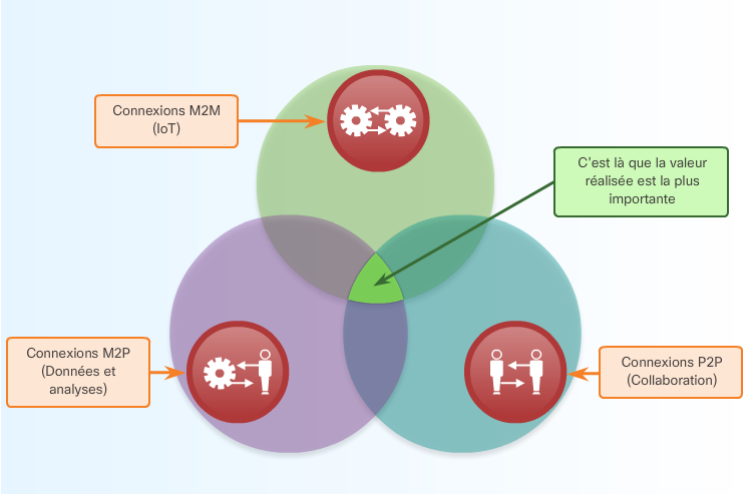


Figure 1.2.5: P2P

1.3 Architecture de l'IoT

IoT est un concept technologique et une architecture qui regroupe des technologies déjà disponibles. L'architecture IoT peut également être appelée en tant que modèle piloté par les événements. Les visionnaires ont également réalisé que cet écosystème IoT a des applications commerciales dans les domaines de l'automatisation des lignes d'usine et d'assemblage, du commerce de détail, des soins médicaux / préventifs, de l'automobile et plus encore.[6]

À partir du niveau inférieur, le flux de données est généré à partir de n'importe quelle objets grâce à des capteurs qui sont envoyés vers le ***Cloud***[7] via la passerelle de communication pour l'analyse, ce qui s'avère être une information utile, comme le montre la figure suivante:

[6] `https://www.cognixia.com/overview-architecture-IoT-works/`

[7] ***Cloud :*** *est est une infrastructure dans laquelle la puissance de calcul et le stockage sont gérés par des serveurs distants.auxquels les usagers se connectent via une liaison Internet sécurisée. L'ordinateur de bureau ou portable, le téléphone mobile, la tablette*

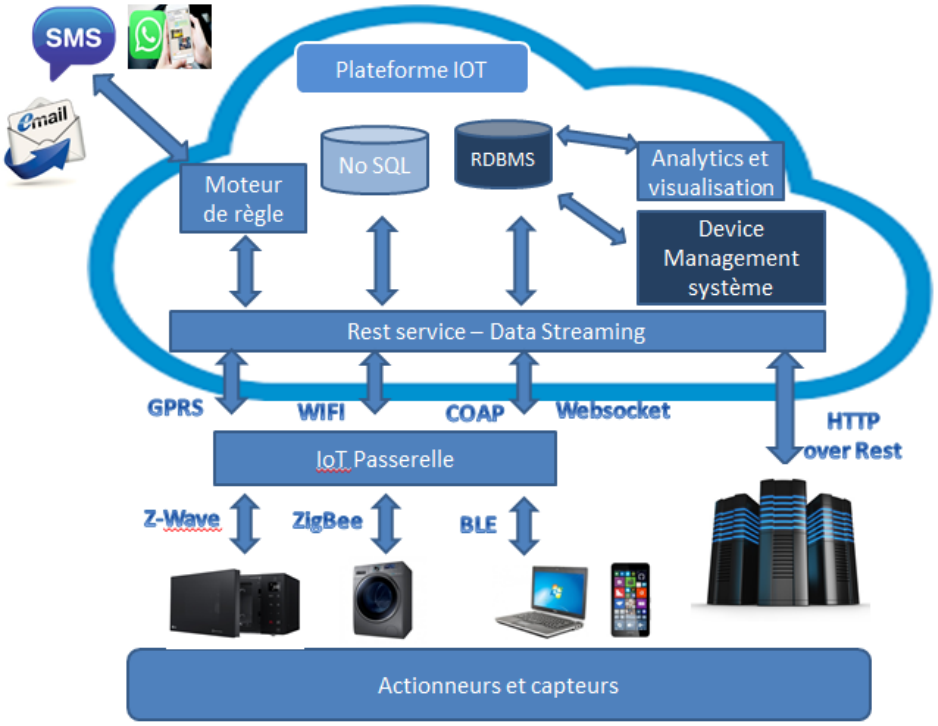


Figure 1.3.1: Architecture IoT

1.3.1 Construction d'un block IoT:

Un objet, dans l'Internet-des-objets peut être une personne ou un animal ou tout objet qui peut générer des données via un capteur intégré, ou tout autre objet naturel ou artificiel qui peut être assigné une adresse IP et fourni avec le capacité à transférer des données sur un réseau.

1.3.2 Les capteurs ou Actionneurs:

Un capteur est un transducteur, dont le but est de renifler une grande variété d'informations allant de l'emplacement, conditions météorologiques / environnement, paramètres de grille, mouvement sur les lignes d'assemblage, données de maintenance du moteur à réaction signal optique. Les capteurs de l'IoT sont appelés comme un nœud qui va collecter des informations et les envoyer au monde extérieur, via des protocoles de communication - Bluetooth, BLE, ZigBee, Z-Wave, Wi-Fi ou par communication filaire. Ces nœuds transmettront les données à un périphérique appelé Gateway .

1.3.3 La passerelle (Gateway) IoT:

La passerelle agit comme un pont entre ces objets IoT et Internet. Les passerelles peuvent se connecter aux périphériques IoT qui communiquent via des protocoles spécifiques, stocker et analyser les informations, puis les envoyer aux serveurs cloud pour traitement et analyse. Les passerelles IoT non seulement permettent d'abstraire le moyen de communication mais fournissent également le canal sécurisé nécessaire à la transmission de ces données. Les passerelles exécutent généralement des systèmes d'exploitation en temps réel (RTOS) ou une forme de Linux pour piloter leurs systèmes. Le cryptage au niveau matériel et logiciel est intégré directement dans la passerelle pour fournir un canal sécurisé pour la communication.

1.3.4 La plateforme cloud et Big Data Analytics:

Les protocoles supportant la plate-forme cloud sont: GPRS, Wi-Fi, CoAP, Websocket, RESTful, etc. Le cloud permettra à IoT de fournir une puissance de calcul, un stockage et une mise en réseau élastiques. Les données massives générées par IoT peuvent être analysées dans le cloud avec des solutions Big Data pour obtenir des informations et des schémas d'utilisation et de comportement des machines et des humains.

Cette intelligence d'entreprise nous permet à son tour de prédire la croissance prochaine de la demande de données et de déployer des ressources supplémentaires en conséquence. Ces modèles sont ensuite analysés, et s'ils ne sont pas pertinents, les informations sont envoyées à l'utilisateur, pour contrôler et surveiller leurs appareils (allant du thermostat d'ambiance aux moteurs à réaction et lignes d'assemblage) à distance. Ces applications transmettent les informations importantes sur des appareils portables et aident à envoyer des commandes à des appareils intelligents.

Cependant, le flux de communication peut également être inversé lorsque l'utilisateur ou bien fabricant souhaite actionner un objet:

- Utilisateur ou fabricant donnera une entrée sous la forme de SMS, Push, Email, Call, ...etc.Cette information est transmise à Internet, c-à-d le Cloud
- le Cloud traite ensuite l'information identifie l'objet particulier à travers l'adresse IP et pousse l'information à travers les protocoles de communication à la passerelle.
- La passerelle déclenchera l'actionneur qui sera responsable du contrôle et du déplacement du système ou de l'objet.[8]

1.4 La sécurité dans l'IoT

Une grande partie de l'internet des objets s'appuie à ce jour sur les technologies du machine-to-machine (M2M). En d'autres termes, les capteurs de l'IoT parlent les uns aux autres, au lieu de parler à un serveur centralisé. Si un thermostat intelligent dit à un lave-vaisselle quand démarrer, que la communication passe sur le réseau Wi-Fi ou Bluetooth, même sans passer par internet, on prend de grands risques. Il va sans dire que les protocoles Wi-Fi et Bluetooth sont facilement piratables, mais comment les deux nœuds de communication savent-ils que les informations venant de l'autre sont autorisées? N'importe quel type d'interaction M2M nécessite un certain niveau de confiance, seulement nous n'avons aucun moyen de prévoir cette confiance à priori, ou de pouvoir la révoquer si un incident survenait. Comment le lave-vaisselle peut-il savoir que quelqu'un a piraté le thermostat?[8]

L'augmentation du nombre d'appareils connectés et de la quantité de données qu'ils génèrent accroît la demande de sécurité de ces données.Les attaques des pirates informatiques sont quotidiennes et il semble qu'aucune organisation ne soit à l'abri. Étant donné la facilité avec laquelle il est aujourd'hui possible de voler et d'utiliser de façon malveillante des informations dans le monde connecté, il est naturel de se préoccuper de ce problème, car les personnes, les processus, les données et les objets seront à l'avenir tous connectés au sein de l'IoT. Mais trop souvent, les objets de l'IoT sont la cible de cyber attaques, et il est impératif d'adopter une approche globale de la sécurité de ces objets. On nous aide à :

- *Renforcer la sécurité du réseau de capteurs*
- *Mettre en œuvre une authentification renforcée*
- *Rationaliser la gestion des identités*

1.4.1 Les vulnérabilités les plus fréquemment rencontrées sur les objets connectés

Dans le monde de l'informatique et dans celui des objets connectés plus particulièrement les pirates informatiques tentent d'accéder aux systèmes en utilisant les failles et les vulnérabilités de ce derniers et parmi les vulnérabilités les plus répandues on a:

- **Mises à jour non sécurisées :** absence de chiffrement et de signature pour les mises à jour des micrologiciels,
- **Utilisation de secrets par défaut :** définition de clés et de mots de passe connus en environnement de production,
- **Communications non-sécurisées :** absence ou faiblesse du chiffrement et du contrôle d'intégrité par signature numérique sur les communications,

[8] `https://www.cognixia.com/overview-architecture-IoT-works/`

- **Stockage de données en clair :** absence de chiffrement sur le stockage local des données,
- **Présence des interfaces de débogage :** possibilité de prendre le contrôle des composants matériels de l'objet.

1.4.2 Les solutions pour se prémunir des attaques ciblant les IoT

Sécuriser une solution liée à l'Internet des Objets (IoT) exige de nouveaux réflexes dans la façon d'envisager le risque, d'évaluer le coût associé, de constituer une équipe pour les anticiper et y répondre.

Identifier les risques véritables

La cybersécurité pour les entreprises a longtemps consisté à protéger ses infrastructures. Mais avec l'IoT, elle ne porte plus seulement sur l'information mais sur le contrôle des actions pilotées par les objets connectés. Le détournement de ces fonctions peut avoir des conséquences graves pour l'entreprise ou la collectivité qui les a mises en place.

Exemple: En juillet 2015, le groupe Fiat Chrysler Automobiles a été contraint de rappeler 1,4 million de véhicules aux Etats-Unis pour effectuer une mise à jour de leurs systèmes informatiques embarqués, deux chercheurs en sécurité avaient piraté un modèle de véhicule connecté. ce qui est inquiétant!

Quantifier les risques

Tout risque n'est pas forcément à anticiper : il faut définir le rapport bénéfice-risque pour juger de ce qui sera acceptable. Lorsque de nombreux capteurs sont déployés à des fins de collecte d'information, pour établir des statistiques par exemple, on peut décider que l'altération des données transmises par un seul capteur n'a pas d'impact majeur. En revanche, si ces mêmes capteurs servent à piloter des actions concrètes, il faudra évaluer le coût d'une attaque, la perte potentielle associée pour l'entreprise et l'investissement nécessaire pour l'éviter afin de décider ou non de sécuriser la solution. Mais attention à ne pas sous-estimer la sécurisation. « Avec l'IoT, la sécurité industrielle devient un enjeu clé : elle permet de garantir la qualité et le délai de production. Si bien que revendiquer un haut niveau de sécurisation devient une valeur ajoutée, un critère de différenciation pour l'entreprise qui peut vendre la sécurité à ses clients » explique Benoît Lemaire, Responsable sécurité IoT chez Orange Cyberdefense.

Combiner les compétences et les expertises

Chaque solution IoT est spécifique et unique. Pour identifier les risques, il est nécessaire de recourir à des experts, des hackers éthiques et des consultants qui envisageront d'abord tous les risques théoriques. En adoptant une démarche similaire à celle des pirates, un hacker éthique tentera toutes les cyberattaques possibles : capter l'information transmise, s'attaquer physiquement à un capteur pour altérer ses composants ou comprendre comment en prendre le contrôle... Le consultant quant à lui pourra évaluer les conséquences de chaque type d'attaque sur les données et leur confidentialité ou toutes les conséquences potentielles (dégradation de la production, arrêt de l'activité, risque financier et risque d'image...). L'entreprise aura ainsi en sa possession tous les éléments pour prendre une décision et définir sa politique.

Lorsqu'un risque est identifié, l'entreprise peut avoir deux postures différentes:

1. ***apporter une réponse technique, en protégeant la solution IoT pour supprimer le risque.***
2. ***apporter une réponse légale, en se déchargeant de toute responsabilité.***

Dans les conditions générales d'utilisation ou les mentions légales de certains services, il est courant de lire des mentions précisant que l'entreprise ne saurait être tenue responsable en cas de... préjudice, dommage... Cela permet à l'entreprise d'anticiper le risque de piratage en se dédouanant des conséquences éventuelles sur les clients.

Imaginons par exemple une société de livraison qui, en traçant en temps réel ses véhicules grâce à des GPS connectés peut ainsi communiquer à ses clients le temps restant avant une livraison via un intranet. Elle pourra choisir de sécuriser l'ensemble de la solution et s'engager sur une information communiquée à tout moment en temps réel. Un engagement dont le coût de sécurisation sera reporté sur les tarifs. À l'inverse, elle pourra présenter ce service d'information.

Conclusion

Dans ce chapitre nous avons mis en évidences la quantité d'informations générées par l'IoT. Selon une étude de Gartner[9], les revenus générés par les produits et services activés par l'IoT dépasseront 300 milliards de dollars d'ici 2020. Cependant, ce n'est que la pointe de l'iceberg.

L'IoT générera de très nombreuses données et, dans le monde d'aujourd'hui, des données bien analysées sont extrêmement précieuses. L'impact de ce phénomène se fera sentir dans tout l'univers du Big Data, ce qui obligera les entreprises à mettre à jour rapidement leurs processus, outils et technologies actuels pour prendre en charge des volumes de données massifs et tirer parti des informations générées par le Big Data.

les données d'entreprise (courriels, documents, bases de données, historiques de processeurs métiers...) aussi bien que Les données issues de capteurs, les contenus publiés sur le web (images, vidéos, sons, textes), les transactions de commerce électronique, les échanges sur les réseaux sociaux, les données transmises par les objets connectés (étiquettes électroniques, compteurs intelligents, smartphones...), les données géolocalisées ...etc, forment ce que l'on appel communément le Big Data. Nous allons le voir plus en détails dans le prochain chapitre.

[9]Gartner Inc. est une entreprise américaine de conseil et de recherche dans le domaine des techniques avancées dont le siège social est situé à Stamford au Connecticut.

Chapter 2

Big Data

Introduction

Il y a dix ans, le volume des données qui était généré en un an l'est aujourd'hui en une semaine. Cela représente plus de 20 exaoctets de données produites chaque semaine. Au fur et à mesure de la connexion de nouveaux périphériques à Internet, le volume de données continue à croître exponentiellement.

L'ensemble de ces milliards de données, c'est ce que l'on appelle communément les Big data.

Le Big Data désigne le courant technologique que nous voyons émerger ces dernières années autour des données, des mégadonnées que nous permettent de stocker aujourd'hui les serveurs.

Le Big Data vient du fait que les données de certaines entreprises ou institutions ***sont devenues tellement volumineuses que les outils techniques classiques de gestion, de requête sur les bases dites structurées et de traitement des données sont devenus obsolètes, avec des difficultés dans l'instanciation de celles-ci, les temps d'extraction, de traitement devenant trop long.***

Les impacts du Big Data, vont bien au-delà des applications web ou informatiques. Les impacts sur la société se font de plus en plus sentir. ***La capacité de stockage et d'analyse des données dûe à l'émergence de nouvelles technologies de traitement et d'analyse de l'information nous permettent de toucher du doigt et d'accélérer les recherches dans tous les domaines :*** médical, pharmaceutique, informatique, bancaire ou électromécanique (Internet des Objets, automobile etc...).

2.1 Les données en mouvement

D'une manière générale, les données sont considérées comme des informations collectées avec le temps. Ces données peuvent par exemple avoir été collectées à l'occasion de diverses transactions représentant le processus de commande d'une organisation. Ces données ont de la valeur pour l'organisation et elles sont historiques par nature. Il s'agit de données statiques que l'on appelle **Les données au repos**.

Toutefois, avec la croissance accélérée du volume des données, la majeure partie de la valeur de ces données est perdue pratiquement dès la création de celles-ci. Les périphériques, les capteurs et la vidéo sont à l'origine de cette croissance permanente du volume de nouvelles données. Ces données présentent une valeur ajoutée maximale, étant donné qu'elles interagissent en temps réel. Ces données portent le nom de « données en mouvement ».Cet afflux de nouvelles opportunités de données offre de nouveaux moyens d'améliorer notre monde, par exemple en résolvant les problèmes globaux de santé ou en améliorant l'éducation. Il existe un potentiel incroyable pour des solutions intelligentes capables de collecter, de gérer et d'évaluer des données à la vitesse des communications humaines. Par conséquent, l'Internet of Everything concernera de plus en plus **Les données en mouvement**.

2.2 La donnée matière première de la transformation numérique

La donnée, qu'on appelle également pétrole numérique du XXIe siècle, constitue une matière première, renouvelable et pour partie inépuisable, comme au début de l'ère pétrolière, on commence à peine à discerner les multiples usages et bénéfices, et on tarde à prendre en compte la valeur stratégique.

Chaque jour, nous générons 2,5 trillions d'octets de données. A tel point que 90% des données dans le monde ont été créées au cours des quatre dernières années seulement. Ces données proviennent de

partout : de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS de téléphones mobiles, pour ne citer que quelques sources.Fig: (2.2.1)[1]

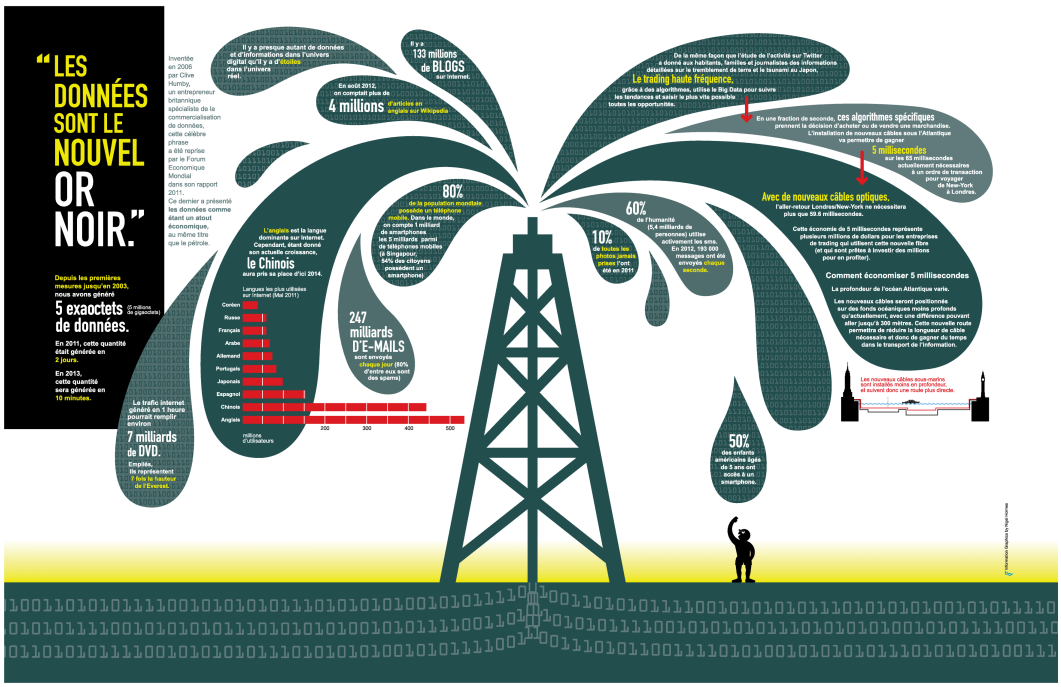


Figure 2.2.1: les données nouvel OR NOIR.

2.3 Les évolutions technologiques vers le Big Data

Trois grandes évolutions techniques ont permis la création et la croissance du Big Data:

il y a premièrement l'évolution du hardware de stockage capable de stocker de plus en plus de données dans des espaces de plus en plus petit et l'évolution des modèles de stockage passant de serveurs physiques internes à l'entreprise à des serveurs dit cloud, qui ont souvent une capacité de stockage bien supérieure aux serveurs des entreprises.

Le second point étant une remise en cause du modèle matériel existant, celui où il fallait acheter le plus gros serveur possible. Aujourd'hui, la nouvelle évolution consiste à mettre en série des petits serveurs remplaçables et de créer un système distribué résistant aux pannes. Ce paradigme a été popularisé par Google au début des années 2000 et est à l'origine de la première version open source du premier framework Big Data sortie il y a 10 ans : Hadoop.

La troisième révolution s'amorçant depuis 2009 est l'explosion des outils d'analyse, d'extraction et de traitement des données de manière non structurée que cela soit du NoSQL ou de nouveaux framework lié à l'écosystème de Hadoop. Les bases de données classiques ne permettant plus de gérer de tels volumes, les grands acteurs du web (Facebook, Google, Yahoo, Linkedin, Twitter) ont créé des Framework Big Data, permettant de gérer et traiter des grandes quantités de données à travers, par exemple, des lacs de données, où toutes données provenant de sources diverses sont stockées. Ces données sont ensuite "splittées" ou séparées pour être traitées parallèlement afin d'alléger les processus de calcul (dans l'ancien modèle, les traitements étaient fait les uns après les autres, dans un stack), puis réassemblées pour donner le résultat final. C'est cette technologie qui permet des vitesses de traitement aussi rapides sur de gros volumes de données. Au départ développée par Google, elle est maintenant sous le drapeau Apache et s'appelle Map Reduce dont voici le découpage du processus de traitement:

L'objectif de l'algorithme ci-dessus est de calculer le nombre de répétitions d'un mot clé dans le texte. L'algorithme Map Reduce répartit les données (ici des chaînes de caractères ou mots) dans plusieurs

[1] http://www.cil.cnrs.fr

noeuds (splitting), chaque noeud va effectuer ses calculs séparément (calcul du nombre de mot - Mapping et Shuffling) et enfin l'étape de Reduce va consolider les données de chaque calcul pour afficher le résultat final.[2]

2.4 Le stockage des données

Lorsque nous parlons d'espace de stockage, nous utilisons le terme d'octets (o). Un octet est une combinaison de 8 bits. Les unités utilisées sont les suivantes :

Le Bit est l'unité de mesure de base utilisée en informatique pour quantifier la taille de la mémoire d'un ordinateur, l'espace utilisable sur un disque dur, la taille d'un fichier ou d'un répertoire. Le bit (anglais binary digit, « chiffre binaire ») ne peut prendre que deux valeurs, la plupart du temps interprétées comme 0 ou 1. Les autres unités de mesure ne correspondent qu'à des regroupements de bits. Ainsi, un octet est composé de huit bits. Son symbole normalisé est « o ». Il peut prendre 28, soit 256, valeurs possibles. Les abréviations suivantes permettent de prendre la mesure de la croissance des données produites :

mesure	mesure transformer en Octets
kilo-octets(ko)	environ mille octets (10^3) .
pétaoctet(Po)	environ un quadrillion d'octets(10^5)
mégaoctet(Mo)	environ un million d'octets (10^6).
gigaoctet(Go)	environ un milliard d'octets (10^9).
téraoctet(To)	environ un trillion d'octets (10^12).
exaoctet(Eo)	environ un quintillion d'octets (10^18).

Table 2.1: la mesure de la croissance des données produites

Au fil des années, l'espace de stockage disponible a augmenté de manière exponentielle. Par exemple, il n'y a pas si longtemps, l'espace de stockage des disques durs s'exprimait généralement en mégaoctets. À l'heure actuelle, on parle plutôt de téraoctets. Il existe trois types principaux de modes de stockage des données :

- **Données locales:**

  Concerne des données accessibles directement à partir de périphériques locaux. Les disques durs, les lecteurs flash USB et les CD/DVD sont des exemples de stockage local de données.

- **Données centralisées:**

  Données stockées et partagées à partir d'un serveur centralisé unique. Ces informations sont accessibles à distance à l'aide de plusieurs périphériques sur le réseau ou sur Internet. L'utilisation d'un serveur de données centralisées peut générer des goulots d'étranglement et des dysfonctionnements, pouvant donner lieu à l'apparition d'un point de défaillance unique.

- **Données distribuées:**

  Données gérées par un système de gestion de bases de données centralisé (DBMS). Les données distribuées sont répliquées et stockées dans divers emplacements. Cela permet un partage à la fois aisé et efficace des données. Les données distribuées sont accessibles par le biais de l'utilisation d'applications locales et globales. Avec un système distribué, il n'y a pas de point de défaillance unique. Si un site n'est plus alimenté, les utilisateurs peuvent toujours accéder aux données à partir des autres sites.

2.5 Les acteurs du Big Data:

Le Big Data a besoin de nouvelles compétences, il est donc normal de voir apparaître de nouveaux rôles: (Fig: 2.5.1)[3]

[2] `https://www.memoireonline.com`

[3]:`http://architecture-nice.com/masters-in-data-science/masters-in-data-science-luxury-close-look-at-data-sci`

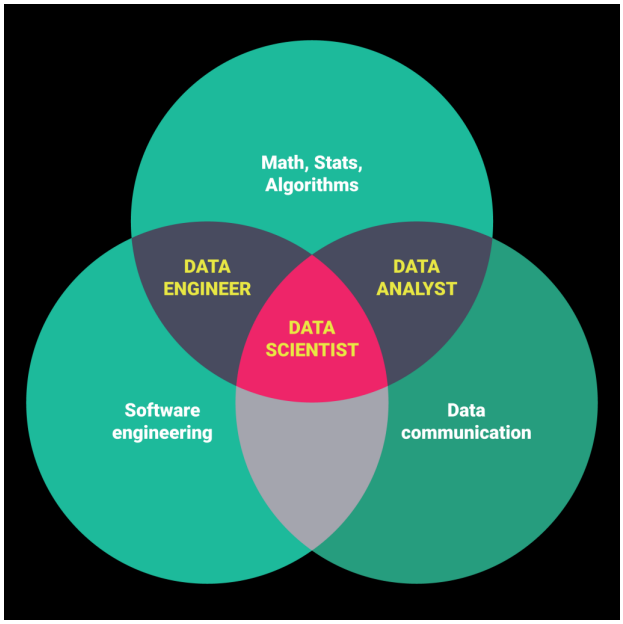


Figure 2.5.1: Acteurs du big data

2.5.1 Analyste de données (Data Analyst):

Un anlyste de données a une expérience dans l'extraction et l'analyse de données provenant d'entrepôts de données traditionnels utilisant SQL. Leurs taches sont soit du coté du stockage des données, soit de la présentantion des résultats d'affaires généraux.

De nombreuses organisations ont du mal à trouver des scientifiques de données compétents sur le marché. Il est toutefois judicieux de sélectionner des analystes de données potentiels et de leur enseigner les compétences pertinentes pour devenir des spécialistes des données.Ce n'est en aucun cas une tâche triviale et impliquerait normalement la personne qui fait un master dans un domaine quantitatif, mais c'est certainement une option viable. Les compétences de base qu'un analyste de données compétent doit avoir sont énumérées ci-dessous:

- *compéhension commerciale.*
- *programation SQL.*
- *conception et mise en oeuvre du rapport.*
- *Développement de tableau de bord.*

2.5.2 Scientifique de données (data scientist) :

le rôle d'un data scientist est assicié à des taches telles que la modélisation prédictive, le développement d'algorithmes de segmentation, les systèmes de recommandation, les frameworks de test A/B et souvent le travail avec des données brutes **non structurées**.La nature de leur travail exige une **compréhension approfondie des mathématiques**, des statistiques appliquées et de la programmation. Il existe quelques compétences communes entre un analyste de données et un chercheur de données, par exemple, la possibilité d'interroger des bases de données. Les deux analysent les données, mais la décision d'un data scientist peut avoir un plus grand impact dans une organisation.Voici un ensemble de compétences qu'un data scientist doit normalement avoir:

- Programmation dans un package statistique tel que: R, Python, SAS[4], SPSS[5] .
- Capable de nettoyer, d'extraire et d'explorer des données provenant de différentes sources.
- Recherche, conception et mise en œuvre de modèles statistiques.

[4]Statistical Analysis System: logiciel généraliste permettant la création et l'édition de base de données
[5]Statistical Package for the Social Sciences

- Connaissance approfondie des statistiques, des mathématiques et de l'informatique.

Dans l'analyse de Big Data, les gens confondent normalement le rôle d'un data scientist avec celui d'un architecte de données. En réalité, la différence est assez simple. Un architecte de données définit les outils et l'architecture de stockage des données, tandis qu'un data scientist utilise cette architecture. Bien sûr, un data scientist devrait être capable de mettre en place de nouveaux outils si nécessaire pour des projets ad-hoc[6], mais la définition et la conception de l'infrastructure ne devraient pas faire partie de sa tâche.

2.5.3 Data_Engineer:

Les ingénieurs de données **(Data_Engineer)** construisent des réservoirs massifs pour le Big Data. Ils développent, construisent, testent et entretiennent des architectures telles que des bases de données et des systèmes de traitement de données à grande échelle. s'occupe du côté applicatif permettant le travail des data scientist. Il développe et entretient les systèmes de collecte, stockage et mise à disposition des données.Il doit s'assurer que l'infrastructure reste fluide et opérationnelle.

Bien que les ingénieurs de données doivent posséder les compétences énumérées ci-dessus, le travail quotidien d'un ingénieur de données variera en fonction du type d'entreprise pour lequel il travaille. De manière générale, on peut classer les ingénieurs de données dans plusieurs catégories:

- Généraliste (Generalist)
- Axé sur le pipeline (Pipeline-centric)
- Base de données centrée (Database-centric)

Passons en revue chacune de ces catégories.

1. **Généraliste :**

   un ingénieur de données généraliste peut devoir tout faire, de l'ingestion des données à leur traitement en passant par l'analyse finale. Cela nécessite plus de compétences en science des données que la plupart des ingénieurs de données.travaille généralement sur une petite équipe.

2. **Axé sur le pipeline:**

   Les ingénieurs de données axés sur les pipelines ont tendance à être nécessaires dans les entreprises de taille moyenne qui ont des besoins complexes en matière de données scientifiques. Un ingénieur de données centré sur les pipelines travaillera avec des équipes de spécialistes des données pour transformer les données en un format utile pour l'analyse. Cela implique une connaissance approfondie des systèmes distribués et de l'informatique.

3. **Base de données centrée:**

   Un ingénieur de données centré sur la base de données se concentre sur la configuration et le remplissage des bases de données analytiques. Cela implique un certain travail avec les pipelines, mais plus travailler avec l'optimisation des bases de données pour l'analyse rapide et la création de schémas de table. Cela implique un travail ETL[7] pour obtenir des données dans les entrepôts . Ce type d'ingénieur de données se trouve généralement dans les grandes entreprises avec de nombreux analystes de données dont les données sont réparties entre les bases de données;

Un ingénieur de données centré sur la base de données peut concevoir une base de données analytique, puis créer des scripts pour extraire des informations de la base de données principale de l'application dans la base de données analytique.

[6]Exemple:`http://www.adhoc-project.com/`

[7]L'ETL (Extract, Transform, Load) est un processus d'intégration des données qui permet de transférer des données brutes d'un système source, de les préparer pour une utilisation en aval et de les envoyer vers une base de données, un entrepôt de données ou un serveur cible. Dans ce processus la transformation des données intervient sur un serveur intermédiaire avant le chargement sur la cible.

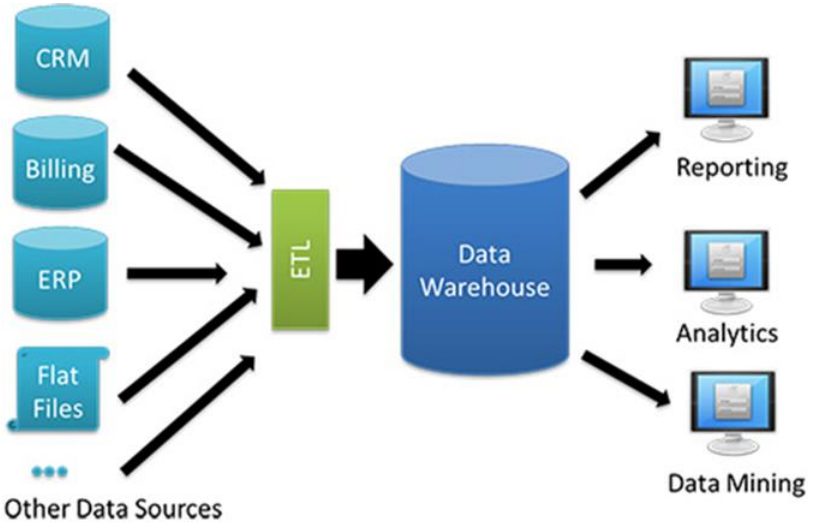


Figure 2.5.2: Architecture d'une BDD centrée

2.6 Les quatres sources du big data :

Pour piloter son activité, l'entreprise doit enrichir ses données avec celles du Big Data (volumes de données sans limite, réponses en temps réel, personnalisation accrue...). Le Big Data permet ainsi de faire passer l'entreprise de l'analyse reporting à l'analyse prescriptive.Tour d'horizon des quatre sources d'information sur lesquelles s'appuie le Big Data. L'information produite par l'entreprise (journal des ventes, états des stocks, liste des clients et prospects...) s'organise dans des bases de données (dites de production), elles-mêmes agrégées dans des entrepôts de données (datawarehouse ou datamarts). Ces données sont ensuite traitées sous forme de cubes décisionnels pour permettre de visualiser des indicateurs de performance sous différentes dimensions (temporelle, géographique, catégories de produits, segmentation client,...).Le Big Data s'appuie sur quatre sources de données :

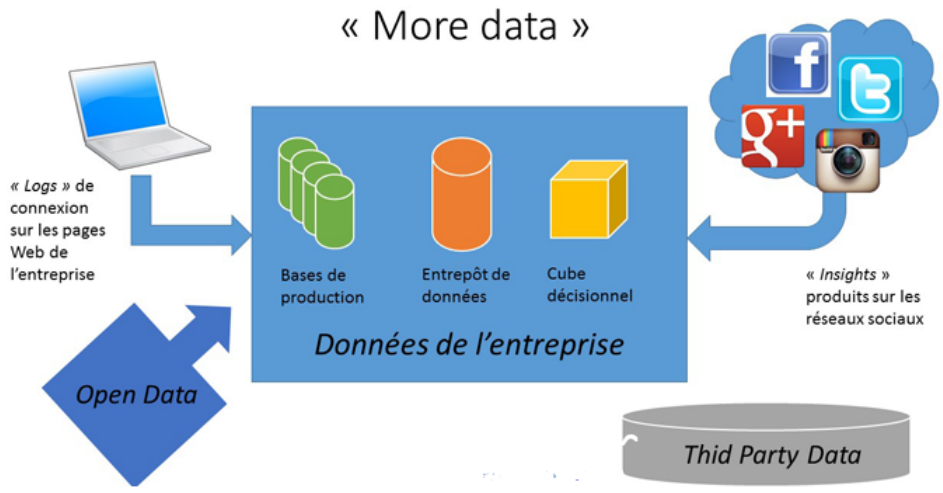


Figure 2.6.1: Source du big data.

2.6.1 Les « logs » des sites web:

Un log, événement ou journal, est simplement la notification d'un événement d'une importance plus ou moins élevée envoyée par un service, un système, un élément réseau ou une application. Les journaux d'événement permettent en administration systèmes, réseaux et en sécurité de retracer les actions et la

vie d'un système. Les logs ont un intérêt et une importance cruciale en informatique, car il s'agit là de savoir ce qu'il s'est passé sur un ensemble d'applications ou systèmes pour pouvoir, par exemple :

1. Expliquer une erreur, un comportement anormal, un crash sur un service comme un service web.
2. Retracer la vie d'un utilisateur, d'une application, d'un paquet sur un réseau sur les logs d'un proxy et des éléments réseau par exemple.
3. Comprendre le fonctionnement d'une application, d'un protocole, d'un système comme les étapes de démarrage d'un service SSH sous Linux.
4. Être notifié d'un comportement, d'une action, d'une modification tel qu'une extinction ou un démarrage système.

2.6.2 Les « insights » des médias sociaux:

C'est un insight consommateur issu des réseaux sociaux. Les insights sociaux sont établis à partir de l'observation et l'analyse des conversations sociales tenues à l'égard d'une marque, d'un besoin ou d'une problématique. Ils peuvent être identifiés à partir d'un dispositif de social listening utilisant des techniques de text-mining ou même de reconnaissance d'images, dans ce cas on prend l'exemple de ***Geopiq pour Instagram*** qui a comme principales caractéristiques de:[8]

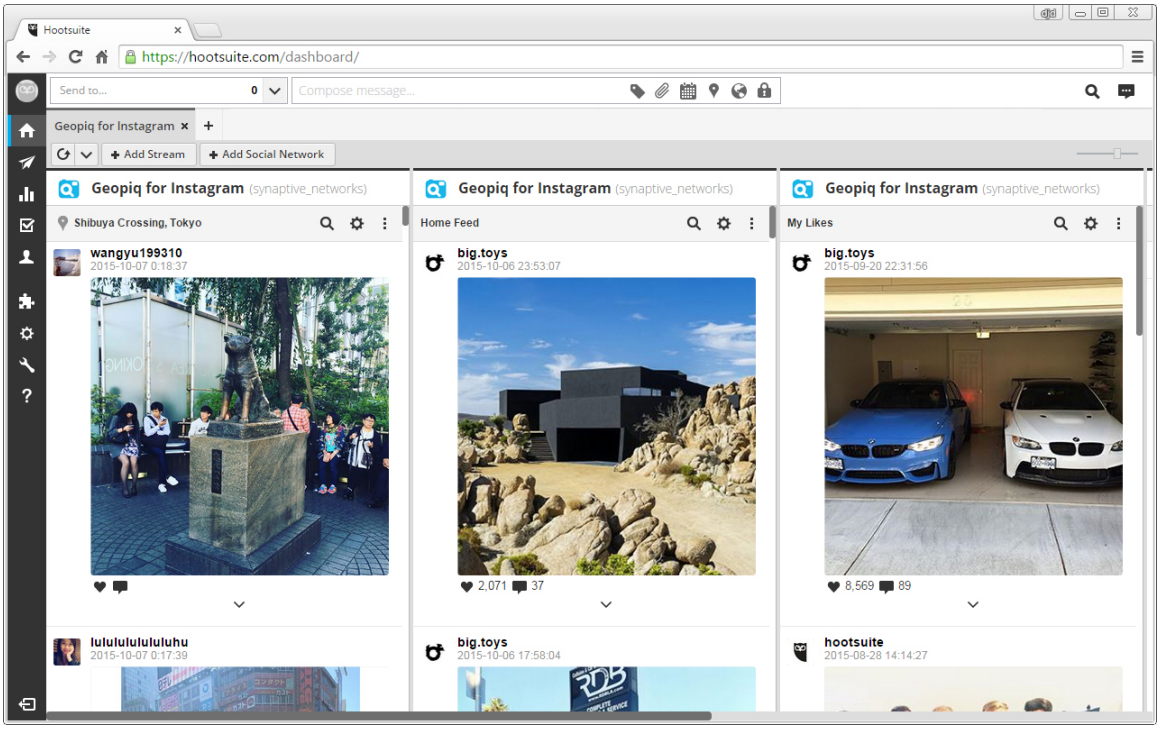


Figure 2.6.2: Geopiq pour Instagram

1. Passer en revue les photos publiées sur Instagram par lieu.
2. Surveiller les posts sur Instagram par lieu, hashtag ou pseudo.
3. Interagir avec l'utilisateur en publiant des commentaires.
4. Partager les photos Instagram sur les autres réseaux sociaux.
5. Consulter les photos et un flux des photos des utilisateurs, des photos populaires et des mentions j'aime, le tout dans un tableau de bord.
6. Consulter et publier des commentaires et des mentions j'aime Consultez les détails des utilisateurs: nombre de photos, nombre d'abonnés et d'abonnements.

[8] `http://apps.hootsuite.com/162/vidpiq`

Cette application permet de surveiller et d'interagir avec les utilisateurs qui publient dans notre ville, ou bien région qu'ils suivent. nous pouvons par exemple utiliser cet outil lors des événements pour voir ce qui a été publié et répondre aux commentaires de tous ceux présents à notre événement.

2.6.3 Les données tierces « third party data » :

Ce sont des informations collectées par une entité qui n'a pas de relation directe avec l'utilisateur sur lequel les données sont collectées. Souvent, les données de tiers sont générées sur une variété de sites Web et de plates-formes et sont ensuite regroupées par un fournisseur de données tiers

Ce sont généralement des données de ciblage publicitaire ou marketing Internet qui sont fournies à l'annonceur par une société tierce autre que l'éditeur utilisé comme site support pour une campagne.

Les données tierces peuvent être comportementales (tirées du comportement du consommateur) ou déclaratives (lorsque le consommateur remplit un formulaire, par exemple), et elles sont collectées et associées aux visiteurs à l'aide de cookies.

2.6.4 L'Open data (donnée ouverte):

Il s'agit de données auxquelles tout le monde peut accéder et que tout le monde peut utiliser et partager. Les gouvernements, les entreprises et les individus peuvent utiliser l'open data afin de créer des avantages sociaux, économiques et environnementaux.

L'Open Data, c'est, avant tout, une philosophie, une volonté citoyenne, celle de considérer l'information publique comme un bien commun.[9]

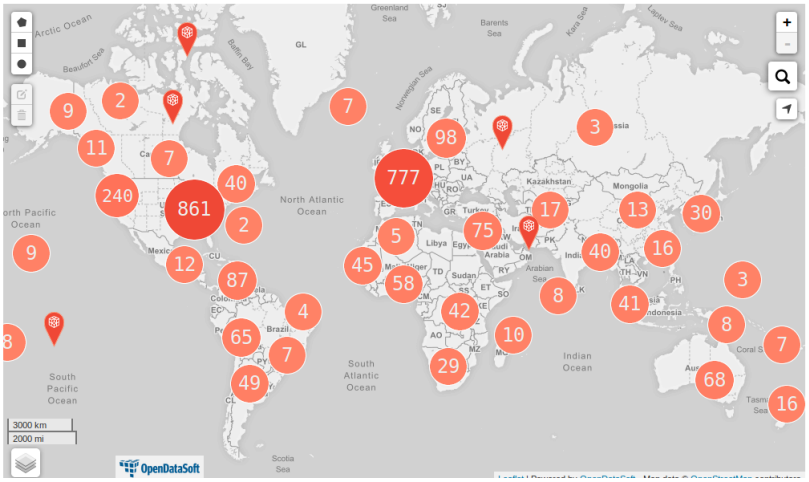


Figure 2.6.3: carte géographique des données ouvertes dans le monde

2.7 La Gestion du Big data: (Les 5 V de Big Data)

Parmi les facteurs contribuant à cette augmentation de la quantité d'informations, on peut citer le nombre de périphériques connectés à Internet ainsi que le nombre de connexions entre ces périphériques. Mais nous n'en sommes qu'au début. Chaque jour, de nouveaux périphériques sont connectés à Internet, créant ainsi une abondance de nouveaux contenus. Avec cette quantité d'informations, les organisations doivent apprendre à gérer les données et également à gérer le « Big Data ».

Pour les résumer en une phrase, les Big data sont un ensemble de solutions alternatives aux solutions traditionnelles de bases de données et d'analyse afin de traiter un volume très important de données, en temps réel et avec une très grande diversité de sources et de formats.

Les outils permettant de collecter des données, de les analyser et de leur donner du sens pour proposer des solutions business aux professionnels.Depuis les débuts du Big Data.Les 5V(volume, vélocité, variété, véracité et visibilité). Elles permettent de définir les attentes des utilisateurs et les besoins de l'industrie.

[9] https://www.opendatasoft.fr/ressource-liste-portails-open-data-dans-le-monde/

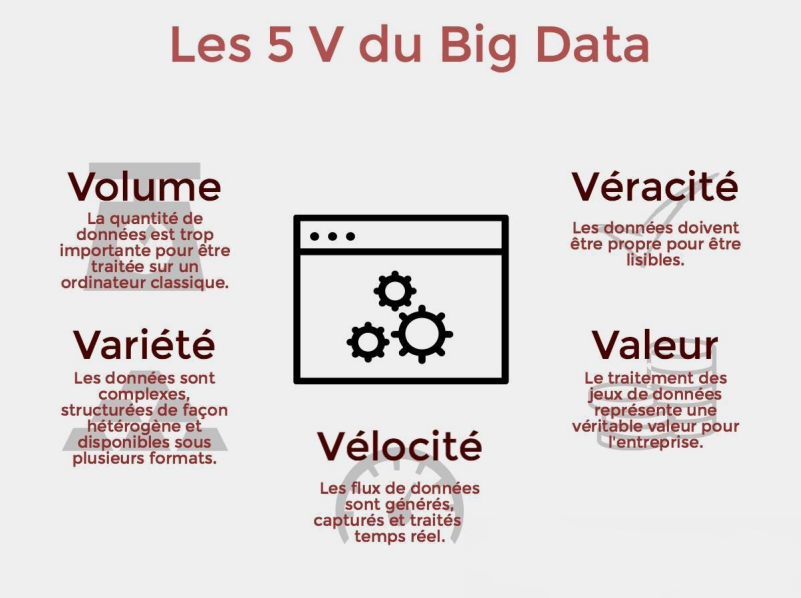


Figure 2.7.1: Les 5 V Du big Data

2.7.1 Le volume

Le volume correspond à la masse d'informations produite chaque seconde. Selon des études, pour avoir une idée de l'accroissement exponentiel de la masse de données, on considère que 90% des données ont été engendrées durant les années où l'usage d'internet et des réseaux sociaux a connu une forte croissance. L'ensemble de toutes les données produites depuis le début des temps jusqu'à la fin de l'année 2008, conviendrait maintenant à la masse de celles qui sont générées chaque minute. Dans le monde des affaires, le volume de données collecté chaque jour est d'une importance vitale. Les pratiques d'analytiques et de géolocalisation sont par exemple des domaines contribuant à l'explosion du volume des données et cette dernière devrait être renforcée par les données en provenance des objets connectés. Les entreprises font face à une augmentation exponentielle des données (jusqu'à plusieurs milliers de téraoctets) qui seront extraites soit:

des journaux de connexion des sites web(logs), des réseaux sociaux (insights) ou bien de l'analyse des données etc..., Les technologies traditionnelles (Business Intelligence, bases de données) n'ont pas été pensées pour de telles volumétries.

2.7.2 La Variabilité/Variété

Seulement 20% des données sont structurées puis stockées dans des tables de bases de données relationnelles similaire à celles utilisées en gestion comptabilisée. Les 80% qui restent sont non-structurées. Cela peut être des images, des vidéos, des textes, des voix, et bien d'autres encore... La technologie Big Data, permet de faire l'analyse, la comparaison, la reconnaissance, le classement des données de différents types comme des conversations ou messages sur les réseaux sociaux, des photos sur différents sites etc. Ce sont les différents éléments qui constituent la variété offerte par le Big Data.[5]

Les données à traiter dans une entreprise sont de natures multiples (**structurées, non structurées**).Exemple de données ***structurées*** :flux, RSS, XML, JSON,bases de données.Ce à quoi peuvent s'ajouter des données ***non structurées*** :mails, pages web, multimédia (son, image, vidéo, etc.).Ces données ***non structurées*** peuvent faire l'objet d'une analyse sémantique permettant de mieux les structurer et les classer, entraînant une augmentation du volume de données à stocker. La solution doit être évolutive car les formats de données ne sont pas tous actuellement connus (voir par exemple comment le format JSON a supplanté XML très rapidement).

2.7.3 la Vélocité

La velocité équivaut à **la rapidité** de l'élaboration et du déploiement des nouvelles données. Par exemple, si on diffuse des messages sur les réseaux sociaux, ils peuvent devenir viraux et se répandre en un rien

de temps. Il s'agit d'analyser les données au décours de leur lignée (appelé parfois analyse en mémoire) sans qu'il soit indispensable que ces informations soient entreposées dans une base de données.[2]

Dans certains cas l'accès et le partage des données doivent se faire en temps réel La vitesse de traitement élevée permet d'offrir des capacités temps réel d'analyse et de traitements des données, la vélocité, ou vitesse, fait référence à l'énorme rapidité avec laquelle les données sont générées et traitées. Jusqu'il y a quelques années, traiter les bonnes données et faire remonter à la surface les bonnes informations prenait beaucoup de temps. Aujourd'hui, les données sont disponibles en temps réel. Cela ne s'explique pas uniquement par la vitesse de l'Internet, mais aussi par la présence du Big Data. Plus nous créons des données, plus il nous faut de méthodes pour les analyser, ce qui augmente parallèlement le nombre de données à observer. C'est un cercle vicieux.

2.7.4 La Valeur

La notion de valeur correspond au profit qu'on puisse tirer de l'usage du Big Data. Ce sont généralement les entreprises qui commencent à obtenir des avantages incroyables de leurs Big Data. Selon les gestionnaires et les économistes, les entreprises qui ne s'intéressent pas sérieusement au Big Data risquent d'être pénalisées et écartées. Puisque l'outil existe, ne pas s'en servir conduirait à perdre un privilège concurrentiel.C'est un point essentiel du Big Data car il va permettre de monétiser les données d'une entreprise.Ce point n'est pas une notion technique mais économique.[5, 2]

On va mesurer le retour sur investissements de la mise en œuvre du Big Data et sa capacité à s'autofinancer par les gains attendus pour l'entreprise.

Plus on souhaite apporter de la valeur aux données, plus le coût et la complexité de la chaîne augmente :

2.7.5 La Véracité

C'est la capacité à disposer de **données fiables** pour le traitement.On va s'intéresser à la provenance des données afin de déterminer s'il s'agit de données de confiance.En fonction du critère de confiance, on accordera plus ou moins d'importance à la donnée dans les chaînes de traitement.Parmi les données dont il faut éventuellement se méfier on trouve les données des réseaux sociaux dont la provenance et l'objectivité est difficile à évaluer.De plus même pour des données dont on connaît la provenance, la pondération n'est pas constante.Par exemple il peut s'agir de données incomplètes, dont l'anonymisation a enlevé une partie de la valeur statistique ou encore de données trop anciennes.[5, 2]

Comme le Big Data permet de collecter un nombre indéfini et plusieurs formes de données, il est difficile de justifier l'authenticité des contenus, si l'on considère les post Twitter avec les abréviations, le langage familier, les hashTag, les coquilles etc. Toutefois, les génies de l'informatique sont en train de développer de nouvelles techniques qui devront permettre de faciliter la gestion de ce type de données notamment par le W3C.

Au final, ces 5 V permettent de définir le Big Data. Un outil qui permet de gérer des données qualitatives et volumineuses, très diverses, et traitées en temps réel. Surtout, un outil qui assure à l'utilisateur une visibilité des données, via des outils de reporting, qui lui permette de prendre de bonnes décisions.

2.8 Les technologies Big Data:

Les technologies Big Data sont importantes pour fournir une analyse plus précise, ce qui peut conduire à une prise de décision plus concrète entraînant une plus grande efficacité opérationnelle, des réductions de coûts et une réduction des risques pour l'entreprise.

Pour exploiter la puissance du Big Data, on a besoin d'une infrastructure capable de gérer et de traiter en temps réel d'énormes volumes de données structurées et non structurées et de protéger la confidentialité et la sécurité des données.

Il existe différentes technologies sur le marché de différents fournisseurs, y compris Amazon, IBM, Microsoft, etc., pour gérer les données volumineuses. En examinant les technologies qui traitent les mégadonnées, nous examinons les deux classes de technologies suivantes:

2.8.1 Big Data opérationnel:

Cela inclut des systèmes comme MongoDB qui fournissent des capacités opérationnelles pour les charges de travail interactives en temps réel où les données sont principalement capturées et stockées.

Les systèmes NoSQL[10]Big Data sont conçus pour tirer parti des nouvelles architectures de cloud computing apparues au cours de la dernière décennie pour permettre des calculs massifs à moindre coût et efficacement. Cela rend les charges de travail Big Data opérationnelles beaucoup plus faciles à gérer, moins chères et plus rapides à mettre en œuvre.

Certains systèmes NoSQL peuvent fournir des informations sur les tendances et les tendances basées sur des données en temps réel avec un codage minimal et sans avoir besoin de scientifiques de données et d'infrastructures supplémentaires.

Bases NoSQL (Not Only SQL):

Les bases de données relationnelles ont une philosophie d'organisation des données bien spécifiques, avec notamment le langage d'interrogation SQL, le principe d'intégrité des transactions (ACID), et les lois de normalisation. Bien utiles pour gérer les données qualifiées de l'entreprise, elles ne sont pas du tout adaptées au stockage de très grandes dimension et au traitement ultra rapide. Les bases NoSQL autorisent la redondance pour mieux servir les besoins en matière de flexibilité, de tolérance aux pannes et d'évolutivité.

2.8.2 Big Data analytique:

Cela inclut des systèmes tels que les systèmes de base de données Massively Parallel Processing (MPP) et MapReduce qui fournissent des capacités analytiques pour l'analyse rétrospective et complexe pouvant toucher la plupart ou la totalité des données.

MapReduce fournit une nouvelle méthode d'analyse des données qui est complémentaire aux capacités fournies par SQL, et un système basé sur MapReduce qui peut être étendu de simples serveurs à des milliers de machines haut et bas de gamme. Ces deux classes de technologies sont complémentaires et fréquemment déployées ensemble.

MapReduce:

Au départ, il y'a eu "Map Reduce", une méthode et une technologie de traitement massivement parallèle issues des laboratoires Google Corp avec gestion de la tolérance aux pannes et système de gestion de fichiers spécifiques (Google File System), c'est un Framework de traitement de données en clusters. Il est composé des fonctions Map et Reduce, il permet de répartir les tâches de traitement de données entre différents ordinateurs, pour ensuite réduire les résultats en une seule synthèse. On parle là de traitement sur des milliers de machines réparties en grappes (clusters). (Fig 2.8.1)[11]

[10]Not Only SQL : désigne les bases de données qui ne sont pas fondées sur l'architecture classique des bases de données on le verra plus en détail dans le chapitres qui suivent.

[11]`http://architecture-nice.com/big-data-hadoop/`

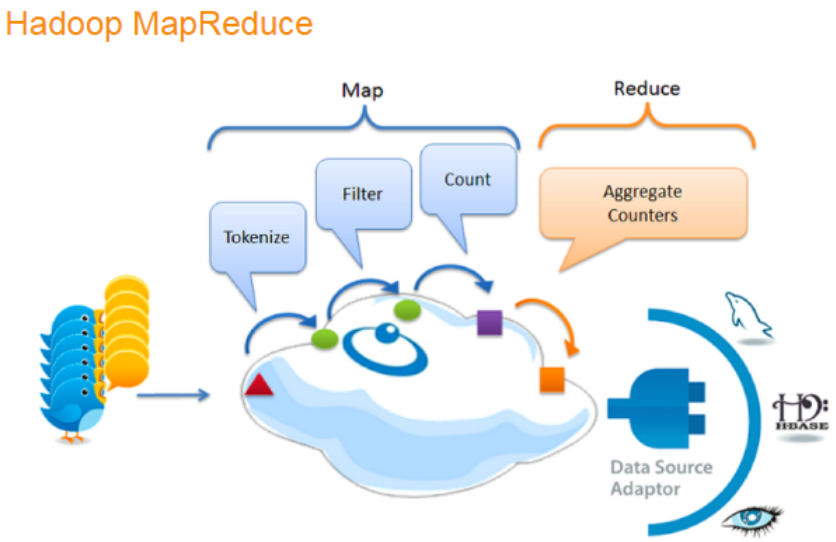


Figure 2.8.1: Hadoop MapReduce

Apache Hadoop:

Hadoop[12] a vu le jour, un framework largement utilisé aujourd'hui pour traiter de très gros volumes de données. Hadoop est composé de plusieurs éléments : un système de stockage (HDFS)[13], un système de planification des traitements (YARN)[14] et le framework de traitement (MapReduce). Un des cas d'utilisation les plus connus de Hadoop est le data lake[15]. Aujourd'hui Hadoop est Devenu opérationnel dans le monde du Big Data.

2.9 Impacte du big data dans le développement des entrprises:

Un récent sondage, mené en 2013 auprès des lecteurs de la Harvard Business Review, a permis de donner des chiffres intéressants sur la pénétration des Big data dans les entreprises. 28% des sondés ont indiqué que leur entreprise « ***utilisait des solutions Big data pour améliorer les décisions commerciales ou pour créer de nouvelles opportunités d'affaires*** ». 23% ont répondu que leur entreprise avait une stratégie Big data et seulement 3,5% ont déclaré que leur entreprise « savait comment appliquer le Big data à leur secteur ». Si les directions informatiques des entreprises commencent toutes à se pencher sur la question, seules 34% d'entre elles dans le monde se sont lancées dans un tel projet, selon Capgemini, la plupart ne sachant pas par où commencer et/ou ne voyant pas l'intérêt de mettre en place une telle structure d'analyse.[5]Les Big data ne sont pas qu'une simple amélioration de l'analytique tel que nous le connaissons actuellement ; il s'agit de repenser la façon dont les données sont analysées et restituées. Ce tableau résume assez bien les différences qui existent entre les deux technologies qui sont le Big data et l'analytique classique:

[12] `https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm`

[13] **H**adoop **D**istributed **F**ile **S**ystem

[14] **Y**et **A**nother **R**esource **N**egotiatorest une technologie de gestion de clusters, mieux adapté aux applications opérationnelles qui ne peuvent pas attendre la fin des traitements par lots

[15] Lacs de données, est un référentiel de données permettant de stocker une très large quantité de données brutes dans le format natif pour une durée indéterminée.

	Big data	Analytique classique
Type de données	Formats non structurés	En lignes et en colonnes
Volume de données	100 téraoctets à plusieurs pétaoctets	Dizaines de téraoctets ou moins
Disponibilité des données	Flux constant	Pool statique
Méthode d'analyse	Apprentissage automatique	À base d'hypothèses
Type d'analyse	Vision prospective, recommandations	Vision rétrospective, « dans le rétroviseur »
Objectif premier	Optimisation de l'activité et anticipation	Support aux décisions internes et supervision de l'activité

Table 2.2: Big data et analytique classique

Tout d'abord, il convient de savoir que toutes les industries et secteurs peuvent être concernés par la mise en place de Big data ; bien entendu, il existe des secteurs plus propices que d'autres tels que les assurances, les sociétés en ligne, les banques, les voyagistes, les transporteurs... Ces secteurs génèrent une quantité très importante de données et peuvent en extraire directement de la valeur afin d'améliorer leurs services ou proposer de nouvelles offres. A contrario de ces secteurs, certains sont historiquement moins fournis en données, mais cela tend à changer avec la e-transformation de leur activité et l'augmentation du recours à des outils informatiques.

2.9.1 Vers un marketing en temps réel:

Le marketing connait depuis bien longtemps l'utilisation de l'analytique traditionnel. Il utilise ces données pour prévoir et planifier ses différentes actions, mais l'arrivée du marketing prédictif, avec l'utilisation des Big data, lui permettrait d'obtenir une véritable connaissance du client et de pouvoir suivre son comportement sur différentes plateformes et canaux de distribution. Il s'agit pour eux d'améliorer encore plus l'efficacité de leurs campagnes en prenant en compte des éléments extérieurs aux données disponibles en interne, via la surveillance des commentaires sur les réseaux sociaux par exemple. Ils seront ainsi à même de proposer des offres beaucoup plus ciblées ou d'analyser plus précisément le phénomène d'attrition de leurs clients afin de leur proposer des offres personnalisées ou corriger les éléments identifiés comme bloquants dans leur relation avec la marque. Certes, il est déjà possible de le faire avec l'analytique traditionnel, mais l'apport du Big data leur permettra de prendre en compte beaucoup plus de paramètres, de calculer davantage de probabilités en un minimum de temps et du même coup de gagner en réactivité. Il s'agit pour eux de passer du marketing prédictif au marketing temps réel.

2.9.2 Optimiser la prospection:

Le service commercial des entreprises s'est considérablement amélioré ces dernières années avec l'arrivée de système de gestion de la relation client et de rapports clients beaucoup plus performants. Mais l'arrivée des Big data dans leur domaine permettrait d'atteindre une autre étape qui est celle de pouvoir prévoir plus efficacement l'aspect « chaud » de leur prospect. Via l'analyse des courriels ou bien des échanges téléphoniques, la mise en place d'un algorithme leur permettrait, par exemple d'attribuer une « note de chaleur » à chaque échange afin de produire une moyenne et surveiller l'évolution de cette note afin de proposer la bonne proposition commerciale au bon moment. Cet outil pourrait également vérifier des conditions externes à l'entreprise ou au particulier prospect, tel que la météo ou des événements de l'actualité afin d'évaluer les chances de pouvoir conclure la vente et du même coup améliorer l'efficacité du service commercial. Le Big data permettrait également de mieux connaître l'activité des commerciaux, via les données de géolocalisation de leurs voitures ou de leurs portables qui pourraient être comparées avec leurs ventes. Mais cette méthode risque de connaître une certaine réticence pour des raisons évidentes de vie privée et de son aspect un peu trop « Big Brother».

2.9.3 Les Big data au secours de la logistique:

Le service logistique est sans aucun doute celui qui peut le plus être amélioré par l'utilisation des Big data. Le tracking a ici sa place sur l'ensemble de la chaîne, de la conception du produit jusqu'à son

arrivée chez le client. Avec la démocratisation des technologies de suivi et d'identification, la masse de données disponibles dans ce secteur va exploser.

Récemment UPS, en utilisant les données collectées via les GPS de l'ensemble de ses véhicules, a revu l'ensemble de ses trajets pour la troisième fois de son histoire afin d'économiser de l'essence, et du temps11. Celui-ci calcule en temps réel le meilleur trajet pour le véhicule en fonction du contenu de son chargement. Cela évite au chauffeur de saisir une adresse et lui permet d'éviter l'ensemble des embouteillages. Cela a également permis à UPS de développer un nouveau service pour ses clients, My-Choice, qui leur permet de modifier l'heure et le lieu de livraison jusqu'à la dernière minute. L'entreprise gagne donc en flexibilité tout en offrant un service unique à ses clients. Une autre technologie, ILC, permet de surveiller les conditions environnementales auxquelles sont soumis les produits tout au long de la chaîne logistique. Cette technologie qui produit énormément de données ne peut être analysée que via le Big data afin de les traiter en temps réel et signaler tout dysfonctionnement ou anomalie aux opérateurs afin qu'ils puissent déclencher des opérations de maintenance plus rapidement ou détecter le vol de marchandises. Ce ne sont que quelques exemples de ce que peut apporter l'utilisation des Big data dans ce domaine. Il s'agit réellement d'un outil clé qui va permettre d'améliorer l'ensemble de la chaîne logistique de toutes les entreprises soucieuses de leur qualité de service, dans les années à venir.

2.9.4 Fiabilité et contrôle de la production:

La production se rapproche du service de la logistique avec qui elle travaille en étroite collaboration. La robotisation des usines de productions qui ne cesse de progresser déporte de plus en plus les hommes à un travail de supervision de ces robots bardés de capteurs. Ces derniers sont désormais capables d'indiquer leurs performances et de signaler leurs besoins de maintenance en détectant l'usure de leurs propres composants. Toutes ces données doivent donc être traitées en temps réel afin de remonter les alertes rapidement et ne pas interrompre la chaîne de production. Par exemple, General Electric a installé sur des turbines à gaz des technologies permettant le suivi en temps réel des conditions de fonctionnement des appareils(12). Si ses algorithmes détectent une anomalie, la commande de pièce et la mobilisation d'un technicien sont effectuées automatiquement afin de minimiser l'impact d'un arrêt de la production. Cette amélioration représente, pour les seules turbines à gaz, une économie de plus de 66 milliards de dollars en consommation de carburant sur les quinze prochaines années. Ces données génèrent 588Go de données par jour, pour avoir un ordre d'idée, cela représente sept fois le volume d'information généré par Twitter chaque jour. En y associant les données de la logistique, il est donc possible pour ce service de gagner en efficacité et d'assurer une délivrabilité constante grâce à l'anticipation que lui permet le recours au Big data.

2.9.5 Améliorer le service informatique:

Le service informatique, qui est la clé de voûte de toute solution Big data, peut également en bénéficier afin d'améliorer ses propres prises de décisions. Ainsi les deux domaines principalement impactés sont la fiabilité et la sécurité de l'infrastructure informatique. Ce n'est pas une nouveauté, l'ensemble des outils informatiques produit une quantité de données phénoménale. Toutes ces données sont autant d'informations précieuses pour évaluer le bon déroulement d'une tâche ou le bon fonctionnement d'un matériel. Malheureusement, dans beaucoup de services informatiques, elles restent souvent enfouies dans les fichiers « logs » que génèrent ces outils et ne sont analysées que lorsqu'un problème survient afin d'en comprendre l'origine. Les Big data permettraient d'améliorer la remontée d'information de par leur capacité à analyser en permanence des données structurées ou non ; elles sont capables de s'adapter à tous les types de reporting et peuvent ainsi signaler des anomalies avant que celles-ci n'impactent de façon plus grave l'ensemble du système. De même pour la sécurité informatique, désormais les entreprises ne peuvent plus se contenter de réagir ; elles se doivent d'anticiper toute tentative d'intrusion. Pour cela, il faut mettre en place une surveillance accrue des zones sensibles du système informatique et monitorer toute activité qui pourrait paraitre suspecte grâce à un algorithme utilisant le Big data. La vitesse d'analyse et la capacité de compréhension de l'algorithme utilisant plusieurs centaines de paramètres est ici un élément crucial de l'avantage d'une solution Big data pour effectuer ce genre de monitoring. Il semble évident, avec l'explosion de l'utilisation d'internet et des outils informatiques, que les sociétés qui sauront anticiper les problématiques de leur infrastructure informatique auront un avantage certain dans les années à venir sur celles qui ne font que constater les dégâts et réagir après

coup.

2.9.6 De nouveaux outils pour la prise de décisions:

Le cœur de la survie d'une entreprise repose sur les décisions que prennent les dirigeants. Avant la démocratisation d'internet dans les entreprises, les décisions étaient prises en fonction de paramètres très génériques sur le secteur dans lequel évoluait l'entreprise. Il était très dur de connaître les stratégies des concurrents et voir ce qui se faisait dans le monde. Aujourd'hui cela a bien changé et il serait inconsidéré de prendre une décision sans avoir récupéré au préalable des éléments très précis sur l'état du marché dans le monde et avoir analysé finement la concurrence. Le genre d'applications qui offrent déjà ce genre de service sont appelées Informatique décisionnel ou Business Intelligence. Mais désormais, la moindre information se doit d'entrer en ligne de compte dans les solutions de Business Intelligence. Celle-ci peut être sous forme structurée ou non, mais elle se doit d'être analysée et remontée afin de faciliter la prise de décisions. Elle doit également être fraîche ce qui signifie qu'il s'agit d'analyser en quasi temps réel l'état du marché et de l'entreprise. Les décideurs doivent disposer de l'ensemble des informations disponibles dans l'entreprise et à l'extérieur de celle-ci afin de prendre des décisions en corrélation avec l'état réel du marché. Les éditeurs actuels de solutions de Business Intelligence ont de plus en plus de mal à suivre la cadence et à répondre aux attentes des entreprises dans ce domaine, d'autant plus que le Big data revient généralement moins cher sans toutes les contraintes que présentent ces anciens systèmes. Les géants du web ont déjà positionné leurs produits, BigQuery chez Google ou Redshift chez Amazon, et proposent des solutions de Business Intelligence à la sauce Big data à des tarifs compétitifs en comparaison aux solutions traditionnelles (500€ par To/mois).

2.10 Cycle de vie du Big Data:

Dans un projet Big Data ces étapes constituent normalement la majeure partie du travail dans un projet réussi. Dans cette section, nous allons jeter un peu de lumière sur chacune de ces étapes du cycle de vie des données volumineuses.

- ***Définition du problème métier:*** C'est un point commun dans le cycle de vie traditionnel BI et Big Data Analytics. Normalement, c'est une étape non triviale d'un grand projet de données de définir le problème et d'évaluer correctement combien de gain potentiel il peut avoir pour une organisation. Il semble évident de le mentionner, mais il faut évaluer quels sont les gains et les coûts attendus du projet.

- ***Recherche:*** Analyser ce que d'autres entreprises ont fait dans la même situation. Cela implique de rechercher des solutions raisonnables pour une entreprise, même si cela implique d'adapter d'autres solutions aux ressources et aux exigences de l'entreprise. À ce stade, une méthodologie pour les étapes futures devrait être définie.

- ***Évaluation des ressources humaines:*** Une fois le problème défini, il est raisonnable de poursuivre l'analyse si le personnel actuel est capable de mener à bien le projet. Les équipes BI traditionnelles ne sont peut-être pas en mesure de fournir une solution optimale à toutes les étapes, il faut donc envisager de démarrer le projet s'il est nécessaire d'externaliser une partie du projet ou d'embaucher plus de personnes.

- ***L'acquisition des données***: Cette section est la clé d'un grand cycle de vie des données; Il définit quel type de profils serait nécessaire pour fournir le produit de données résultant. La collecte de données est une étape non-triviale du processus; cela implique normalement la collecte de données non structurées provenant de différentes sources. Pour donner un exemple, il pourrait s'agir d'écrire un robot d'exploration pour récupérer des avis sur un site Web. Cela implique de traiter du texte, peut-être dans différentes langues nécessitant normalement beaucoup de temps à compléter.

- ***Data Minging***: Une fois les données récupérées, par exemple, sur le Web, elles doivent être stockées dans un format facile à utiliser. Pour continuer avec les exemples de commentaires, supposons que les données sont extraites de différents sites où chacun a un affichage différent des données.

  - Supposons qu'une source de données donne des avis en termes d'évaluation dans les étoiles, il est donc possible de lire ceci comme un mappage pour la variable de réponse ***Y*** $\in$ ***{1, 2, 3, 4, 5}*** . Une autre source de données donne des avis en utilisant deux systèmes de flèches, l'un pour le vote en hausse et l'autre pour le vote en baisse. Cela impliquerait une variable de réponse de la forme ***Y*** $\in$ ***{positif, négatif}*** .
  - Afin de combiner les deux sources de données, une décision doit être prise afin de rendre ces deux représentations de réponse équivalentes. Cela peut impliquer de convertir la première représentation de la réponse de la source de données à la seconde forme, en considérant une étoile comme négative et cinq étoiles comme positive. Ce processus nécessite souvent une grande allocation de temps pour être livré avec une bonne qualité.

- ***Stockage de données:*** Cette étape du cycle est liée à la connaissance des ressources humaines en termes de leurs capacités à mettre en œuvre différentes architectures. Les versions modifiées des entrepôts de données traditionnels sont toujours utilisées dans des applications à grande échelle. Par exemple, teradata et IBM offrent des bases de données SQL capables de gérer des téraoctets de données; Des solutions Open Source telles que PostgreSQL et MySQL sont toujours utilisées pour des applications à grande échelle.

- ***L'analyse exploratoire des données:*** Une fois les données nettoyées et stockées de manière à pouvoir en extraire les informations, la phase d'exploration des données est obligatoire. L'objectif de cette étape est de comprendre les données, ce qui est normalement fait avec des techniques statistiques et de tracer les données. C'est une bonne étape pour évaluer si la définition du problème est logique ou faisable.

- ***Préparation des données pour la modélisation et l'évaluation:*** Cette étape consiste à remodeler les données nettoyées récupérées précédemment et à utiliser un prétraitement statistique pour l'imputation des valeurs manquantes, la détection des valeurs aberrantes, la normalisation, l'extraction des caractéristiques et la sélection des caractéristiques.

- ***La modélisation:*** L'étape précédente aurait dû produire plusieurs ensembles de données pour la formation et les tests, par exemple, un modèle prédictif. Cette étape implique d'essayer différents modèles et d'anticiper la résolution du problème commercial. En pratique, il est normalement souhaitable que le modèle donne un aperçu de l'activité. Enfin, le meilleur modèle ou la meilleure combinaison de modèles est sélectionné pour évaluer ses performances sur un ensemble de données abandonné.

- ***la mise en œuvre:*** A ce stade, le produit de données développé est implémenté dans le pipeline de données de l'entreprise. Cela implique la mise en place d'un schéma de validation pendant que le produit de données fonctionne, afin de suivre ses performances. Par exemple, dans le cas de la mise en œuvre d'un modèle prédictif, cette étape impliquerait l'application du modèle à de nouvelles données et, une fois la réponse disponible, évaluer le modèle.

2.11 Problématique du stockage du Big Data

Si les entreprises veulent tirer profit du Big Data, elles doivent tout d'abord trouver un moyen de stocker ces données, de préférence dans un environnement qui permet d'effectuer des analyses. Cet environnement de stockage doit être capable d'ingérer des données provenant de nombreuses sources. Il doit également pouvoir faire face à une croissance massive et rapide des données.

2.11.1 Stockage HDFS pour le Big Data et Hadoop

Au cours de la dernière décennie, le framework populaire Apache Hadoop a permis aux entreprises de stocker et d'analyser le Big Data. La technologie HDFS[16] elle a été conçue à une époque où le stockage était moins fiable et les réseaux plus lents, et n'est pas optimisée pour les infrastructures hautement performantes d'aujourd'hui.

[16]Hadoop Distributed File System:est le système de fichier distribué de Apache Hadoop, Il s'agit d'un composant central du Framework de Apache, et plus précisément de son système de stockage

Le stockage HDFS traditionnel n'est pas adapté aux besoins des entreprises. Son système de mise en miroir des données nous oblige à stocker trois copies complètes de chaque donnée, augmentant par le même coup la charge supplémentaire liée au stockage et les frais de gestion. En outre, la technologie HDFS standard ne prend pas en charge le multitenancy ni la géodistribution, ses fonctionnalités de reprise après sinistre sont limitées.

La taille maximale de chaque système de stockage est déterminée par l'évolutivité de son « espace de nommage ». Dans Hadoop, l'espace de nommage standard du système de fichiers HDFS est géré par un seul serveur et est maintenu en mémoire. La taille maximale de l'espace de nommage pouvant être géré par le « NameNode » central Hadoop est limitée par la quantité de mémoire disponible sur ce serveur. Les performances globales du système de fichiers sont par conséquent limitées par celles d'un seul serveur. La technologie HDFS est inefficace pour la gestion d'un gros volume de petits fichiers (tels qu'en utilisent généralement les applications IdO), car les métadonnées de chaque fichier doivent être stockées dans la mémoire du serveur NameNode. Si le serveur NameNode tombe en panne, tous les processus seront mis en attente pendant la durée des réparations.

2.11.2 Stockage en mode objet

Le stockage en mode objet, basé dans le Cloud, est une nouvelle forme de stockage que les entreprises peuvent utiliser pour exploiter les nouvelles opportunités offertes par l'Internet des objets et l'analytique Hadoop. Dans cette architecture de stockage, les données ne sont pas stockées dans des blocs, des fichiers ni des dossiers, mais dans des conteneurs de taille flexible appelés « objets ».

Le stockage en mode objet est idéal pour stocker de gros volumes de données Big Data non structurées, dans la mesure où il peut évoluer facilement, sans limites physiques. Ce modèle de stockage permet de surmonter les restrictions du stockage HDFS traditionnel, en offrant un espace de nommage quasiment illimité, pour une évolutivité totale et une gestion simplifiée.

Le stockage en mode objet a gagné en popularité en tant que technologie derrière le stockage Cloud, avec la forte croissance des services de Cloud public comme Amazon S3 et Microsoft Azure Storage, ainsi que plusieurs options sur site. Mais comment identifier l'option la plus adaptée pour la stratégie Big Data, et pour quelles situations ?

Exemple: Le stockage en mode objet pour le Big Data avec EMC ECS:

Figure 2.11.1: Avantages du stockage en mode objet.

Avec le stockage en mode objet sur site, on bénéficie de l'évolutivité et de la simplicité de gestion du Cloud public, mais on garde un contrôle total sur l'emplacement et la protection de nos données. Cette solution nécessite une infrastructure physique, mais on peut pour cela utiliser du matériel standard à moindre coût. Ce matériel sera intégré de façon intelligente dans la plate-forme de stockage software-defined. On bénéficie également des performances d'un stockage sur site pour nos applications Big Data et nos utilisateurs, en réalisant des économies sur le TCO par rapport au Cloud public.

Conclusion

Les technologies du Big Data s'inscrivent dans une évolution continue compte tenu du fait qu'elles sont jeunes et pas encore stables, ce qui leur vaut la réticence de certaines entreprises. Actuellement, le virage technologique est d'ores et déjà annoncé. Le Big Data s'impose tout doucement, mais certains aspects ne

sont pas encore à la hauteur des attentes, certaines pistes sont à explorer profondément avant l'intégration dans les systèmes d'information :

La sécurité : elle est encore balbutiante malgré quelques initiatives comme Apache Knox (système qui fournit un point unique d'authentification et d'accès pour les services Apache Hadoop dans un cluster. Le but est de simplifier la sécurité Hadoop pour les utilisateurs (qui ont accès aux données du cluster et exécutent des jobs) et les opérateurs (qui contrôlent les accès et de gèrent le cluster).

L'intégration avec le système d'information (SI), une plate forme Hadoop isolée et non intégrée au système d'information ne sera plus possible dans le futur (en tout cas certains besoins exigeront une interaction plus grande). Cette intégration entraînera une modification des processus et par conséquent des besoins de formation des ressources humaines.

Les ressources compétentes : actuellement les compétences ne sont pas encore assez poussées dans le domaine

Protection de la vie privée : la manipulation à grande échelle de des données pose aussi le problème de la vie privée. Trouver l'équilibre entre le respect de son intimité et les bénéfices tirés du big data n'est pas simple. Les utilisateurs des réseaux sociaux ignorent souvent que leurs données privées sont accessibles par un grand public, beaucoup reste à faire afin de garantir la protection des utilisateurs.

L**'IoT** génère une grande quantité de données, ce qui pèse lourdement sur l'infrastructure Internet. En conséquence, cela oblige les entreprises à trouver des solutions pour minimiser la pression et résoudre leur problème de transfert de grandes quantités de données. Le **cloud computing** est entré dans le courant dominant de la technologie de l'information, de nombreux fournisseurs de cloud peuvent autoriser le transfert de nos données via une connexion Internet traditionnelle ou via un lien direct dédié.

Chapter 3

Cloud computing

Les volumes de données augmentent constamment et représentent une complexité majeure en termes de conception. Les solutions de Big Data peuvent être un outil utile à leur gestion. Elles doivent être réalisées sur des architectures de Cloud évolutives, efficaces et basées sur un modèle de coûts fiable et pratique.

Les technologies de Cloud offrent un soutien efficace et économique pour obtenir des architectures évolutives et pour fournir à l'administration de l'entreprise les outils nécessaires à l'analyse du volume sans cesse croissant des informations afin de prendre des décisions stratégiques.

Pour ce faire, des fournisseurs comme Amazon et Google ont conçu des technologies d'analyse très rapides et utiles pour la collecte et la gestion des Big Data.

Pour profiter de tous les avantages du Cloud Computing, il est important de s'appuyer sur un partenaire qui est en mesure de superviser l'ensemble de l'activité, alliant de fortes compétences techniques à une parfaite compréhension des flux de l'entreprise impactés.À travers son réseau d'entreprises certifiées partenaires d'excellence par les grands leaders du marché du Cloud Computing, le groupe Reply est en mesure de soutenir les entreprises dans le transfert de leurs activités vers un centre de données extérieur ou encore dans la migration vers le Cloud de leurs applications depuis une architecture traditionnelle.

C'est un fait bien connu que le **cloud computing** et le **big data** deviennent rapidement deux des technologies les plus importantes actuellement en cours dans le secteur des technologies de l'information. Le cloud permet aux entreprises de remplacer les anciens systèmes obsolètes, de réduire les dépenses informatiques et d'améliorer la collaboration entre les employés.Les données massives consistent en de grands ensembles d'informations qui peuvent aider les entreprises à identifier les tendances pour améliorer l'efficacité ou d'autres domaines de leurs opérations. Les organisations prises au dépourvu quand il s'agit de cette tendance ne seront pas en mesure de gérer autant d'informations et peuvent rencontrer des problèmes qui peuvent avoir un impact sur des choses comme les performances de la base de données.

3.1 L'informatique en nuage (Cloud Computing):

Ceux d'entre nous qui ont été dans les tranchées de l'informatique pendant une ou deux décennies se souviendront que le premier type d'application client-serveur qui était populaire est l'application Mainframe et le terminal. À cette époque, le stockage et le processeur étaient très chers, et le Mainframe met en commun les deux types de ressources et a servi à des terminaux (clients légers). Avec l'avènement de la révolution du PC, qui a apporté stockage de masse et les processeurs pas chers sur les bureaux des entreprises moyennes, le serveur de fichiers a gagné en popularité en tant que moyen pour permettre le partage de documents et d'archivage. Fidèle à son nom, le serveur de fichiers a servi des ressources de stockage pour les clients de l'entreprise, tandis que les cycles de CPU nécessaires pour faire un travail productif ont tous été produits et consommés dans les limites du PC client.

Définition. Le cloud computing est un autre moyen de gérer et de stocker des données, ainsi que d'y accéder. Il implique un grand nombre d'ordinateurs reliés entre eux par l'intermédiaire d'un **réseau**. Les fournisseurs de cloud utilisent énormément la technologie de la virtualisation pour assurer leurs services. La virtualisation permet également de diminuer les coûts d'exploitation grâce à une utilisation plus efficace des ressources. Ces entreprises proposent quatre catégories distinctes de services.(SaaS, PaaS, IaaS, TaaS).

Le cloud computing permet aux utilisateurs d'accéder à leurs données en tout lieu et à tout moment. On a probablement déjà utilisé une certaine forme de cloud si on utilise des services de messagerie basés sur le Web. Il permet également aux entreprises de rationaliser l'exploitation de leur infrastructure IT en ne s'abonnant qu'aux services nécessaires. Son utilisation peut aussi permettre à des entreprises de se passer d'équipement informatique sur site, de maintenance et de gestion. Le cloud computing permet de réduire les coûts des organisations. Il permet de diminuer les coûts en matière d'équipement et d'énergie, les exigences physiques des usines ainsi que les besoins de formation du personnel d'assistance.

Figure 3.1.1: **Cloud Computing**

3.2 Identification d'un cloud

Un Cloud se concentre sur la donnée indépendamment du support, et est capable de la restituer indépendamment de sa localisation. En cherchant des définitions, on a identifié quatre points qui permettent de caractériser un Cloud :

- **Point 1 :** Un Cloud est toujours un **espace virtuel.**
- **Point 2 :** Un Cloud contient des **données** qui sont **fragmentées.**
- **Point 3 :** Les fragments de données d'un Cloud sont toujours **dupliqués** et **répartis** (ou distribués) dans cet espace virtuel, lequel peut être sur un ou plusieurs supports physiques.
- **Point 4 :** Un Cloud possède une fonction de restitution permettant de reconstituer les données. Cette fonction peut être intégrée à la gestion du Cloud ou déportée sur l'application qui fournit le service.

Si l'une de ces quatre conditions n'est pas établie, nous ne sommes pas en présence d'un Cloud.

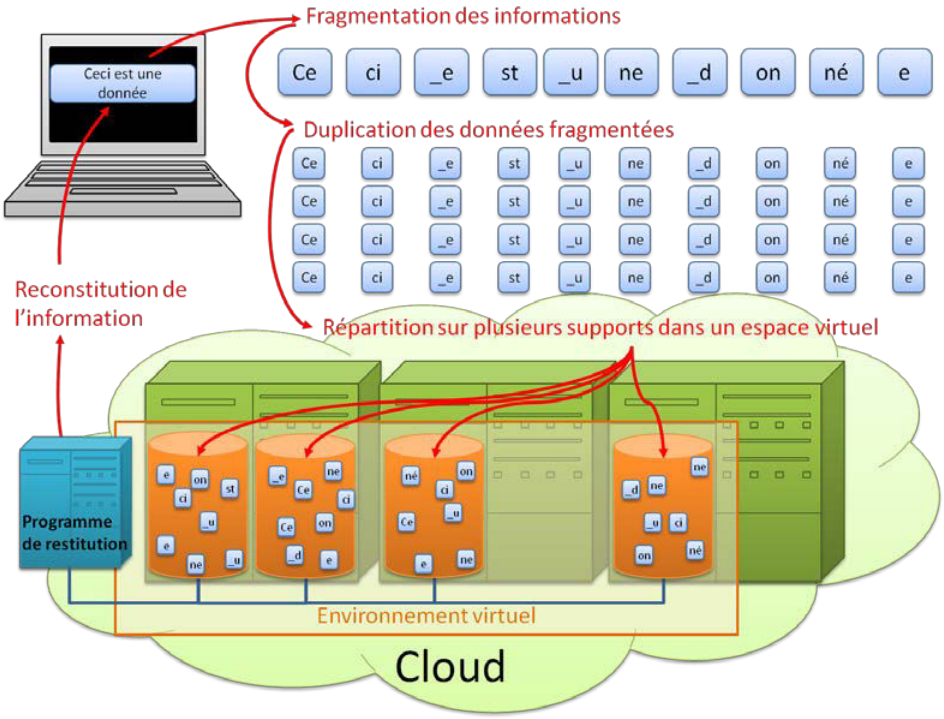


Figure 3.2.1: Les 4 points permettant d'identifier un Cloud

En plus de ces points indispensables, on a également remarqué:

- **Qu'il n'est pas possible de savoir où se trouve une information particulière** (d'où la notion de Cloud/nuage). En effet, les fragments des données la constituant sont répartis sur l'ensemble des supports/devices composant le Cloud, et seule une application de restitution peut les localiser pour reconstituer les données et fournir une information complète. Cette faculté de distribuer des données permet d'étendre un Cloud à plusieurs **datacenters** disposés dans des lieux géographiquement éloignés et reliés par des réseaux hauts débits. Mais de nombreuses offres de Cloud se limitent à une répartition sur un ensemble de serveurs dans un unique **datacenter**.

- Que la granularité[1] de la fragmentation est importante.

  - En effet, si les fragments sont trop signifiants, il sera possible d'en lire le contenu, mais aussi les traitements - pour une plus grande fragmentation - seront donc plus nombreux.
  - Si les fragments sont trop petits (au-delà du nécessaire sécuritaire), ce sera le nombre d'accès nécessaires pour reconstituer les données d'une information qui pourra être pénalisant en termes de performances.

  En tout état de cause, la fragmentation des données augmente la fiabilité et la sécurité mais peut ralentir leur agrégation. Dans ce cas, la qualité du réseau est importante, notamment dans le cas d'un Cloud qui s'appuie sur plusieurs datacenters reliés entre eux.

- Que la reconstitution des données pour la délivrance d'une information, l'exécution d'une application ou l'accès à une fonction constitue le « service » fourni.

- Que cette définition peut s'appliquer de manière identique quel que soit le modèle de service mis en œuvre : IaaS, PaaS ou SaaS.

En termes de mise en œuvre, d'infrastructure et de sécurité, il est alors de possible de déduire plusieurs éléments importants issus des quatre points cités précédemment :

• La perte d'une partie d'un Cloud (une machine par exemple) n'a pas d'effet sur les informations puisqu'elles sont dupliquées et réparties sur plusieurs machines ou sur plusieurs sites. Une donnée est donc stockée en de multiples endroits. Dans ce cas, la sauvegarde des serveurs d'un Cloud est-elle toujours nécessaire ?

[1] Quand on arrive au niveau de granularité d'un système, on ne peut plus découper l'information.

• De même, si un Cloud venait à manquer de ressources dans les serveurs existants, il est alors possible d'ajouter des devices12 supplémentaires. Un Cloud s'adapte alors naturellement à la volumétrie nécessaire, il est dimensionnable. Cette particularité permet notamment d'utiliser de manière optimum l'espace disponible : dans un Cloud les données « remplissent » la place disponible plutôt que d'être « assignées » à des espaces déterminés. Certains membres du CIGREF, en passant de datacenters classiques à un Cloud, ont diminué considérablement le nombre de serveurs mis en œuvre.

• De la même façon, le vol (ou la délivrance) des données d'un serveur ou d'un ensemble de serveurs ne permet pas de lire les informations s'y trouvant stockées puisque chacune d'elles ne contient que des fragments de données. Seul le programme de reconstitution des informations (« la console de restitution ») est capable de faire le lien entre les fragments. Si en plus les données ont été cryptées (avec une clé RSA par exemple) avant d'être fragmentées, la lecture des fragments en direct devient quasi impossible. Ce point particulier dépend néanmoins du niveau de fragmentation des données, des gros fragments seront plus signifiants et permettront de lire plus d'information.

• En termes de protection et de sécurité, l'élément critique n'est donc pas le Cloud en lui-même avec ses serveurs de données, mais « l'application (ou console) de restitution » qui permet de reconstituer les données nécessaires à la délivrance du service. C'est elle qu'il faut protéger. Il faut notamment se soucier de son emplacement géographique et de quelle législation elle dépend (par exemple dans le cadre du Cloud d'un prestataire).[1]

3.3 Caractéristiques du Cloud Computing:

Le Cloud computing tire partie d'un certain nombre de caractéristiques pour fournir des services dans des conditions techniques et économiques très avantageuses. C'est un peu comme la production d'électricité. La plupart des entreprises et des particuliers ont intérêt à utiliser des fournisseurs dont c'est le métier pour garantir la fiabilité et les meilleures conditions économiques. Parmi ces caractéristiques communes, on trouve généralement:

- **Agilité:** Le cloud fonctionne dans l'environnement informatique distribué . Il partage les ressources entre les utilisateurs et fonctionne très rapidement.
- **Haute disponibilité et fiabilité:** La disponibilité des serveurs est élevée et plus fiable, car les risques de défaillance de l'infrastructure sont minimes .
- **Haute évolutivité** : Signifie un approvisionnement "à la demande" de ressources à grande échelle , sans avoir d'ingénieurs pour les charges de pointe.
- **Partage multiple:** Grâce à l'informatique en nuage, de multiples utilisateurs et applications peuvent travailler plus efficacement avec des réductions de coûts en partageant une infrastructure commune.
- **Indépendance de l'appareil et de l'emplacement:** Le cloud computing permet aux utilisateurs d'accéder aux systèmes à l'aide d'un navigateur Web, quel que soit leur emplacement ou l'appareil qu'ils utilisent, par exemple PC, téléphone mobile, etc. L' infrastructure étant hors site (généralement fournie par un tiers) et accessible via Internet. peut se connecter de n'importe où .
- **Maintenance:** La maintenance des applications de cloud computing est plus simple, car elles n'ont pas besoin d'être installées sur l'ordinateur de chaque utilisateur et sont accessibles depuis différents endroits . Donc, cela réduit également le coût.
- **Virtualisation:** La virtualisation est une caractéristique indispensable qui présente de très nombreux avantages. Le matériel est remplacé par du logiciel avec tous les avantages du logiciel : créer une nouvelle machine ou sauvegarder son état consiste à copier un fichier d'où un énorme gain de temps et d'argent. La machine virtuelle ne tombe pratiquement jamais en panne ce qui accroit sérieusement la fiabilité des systèmes. On peut continuer à utiliser des machines qui ne sont plus fabriquées. Le pourcentage d'utilisation réelle d'un serveur physique dépasse rarement 15%. Sur la même puissance de calcul on peut faire fonctionner plusieurs serveurs. Lorsqu'une configuration est utilisée pour des développements, des opérations de recette ou des tests de charge, il est possible de libérer des ressources en archivant la configuration. La remise en ligne du système lorsque c'est nécessaire se fait en quelques minutes

- **faible coût :** En utilisant le cloud computing, le coût sera réduit parce que pour prendre les services de l'informatique en nuage, l' entreprise de TI n'a pas besoin de définir sa propre infrastructure et de payer l'utilisation des ressources.
- **Services en mode paiement à l'utilisation :** Des interfaces de programmation d'application (API) sont fournies aux utilisateurs afin qu'ils puissent accéder aux services sur le cloud en utilisant ces API et payer les frais conformément à l'utilisation des services .
- **Distribution géographique :** Les grandes plates-formes publiques disposent de centres répartis sur la planète pour réduire les risques et placer les données au plus près des utilisateurs. Par exemple, Amazon Web Services propose en 2011 des centres(**data center**s) en Europe (2) aux USA (4) et au Japon (2)

3.4 Data Centers

Les data centers constituent un élément critique du cloud computing. Un data center est une installation qui fournit les services nécessaires à l'hébergement des plus grands environnements informatiques qui existent à l'heure actuelle. Sa fonction principale est de permettre la continuité des activités en assurant la disponibilité des services informatiques, la plupart des organisations étant en effet dépendantes de leurs opérations IT.


Table 3.1: les data centers

Plusieurs facteurs doivent être pris en compte lors du déploiement d'un data center en vue de fournir le niveau de service nécessaire :

- **Emplacement**

  les data centers doivent être situés dans des endroits présentant un faible risque de catastrophes naturelles et suffisamment éloignés des zones à trafic intense de personnes (aéroports, centres commerciaux, 36729309_294595307751110_713955759393603584_netc.) ainsi que des zones revêtant une importance stratégique pour les gouvernements et les services publics (raffineries, barrages, centrales nucléaires, etc.).

- **Sécurité**

  un data center doit présenter des contrôles stricts en matière d'accès physique et de personnel sur site.

- **Électricité**

  un accès suffisant doit être prévu en matière d'alimentation électrique, avec une alimentation de secours composée de systèmes d'alimentation sans coupure, de groupes de batteries et de générateurs électriques.

- **Environnement**

  il faut prévoir un environnement physique étroitement contrôlé capable de maintenir une température et une humidité appropriées. Des systèmes sophistiqués d'extinction d'incendie doivent également être présents.

- **Réseau**

  l'infrastructure réseau doit être évolutive et fiable, avec une connectivité redondante.

  Il existe à l'heure actuelle plus de 3.000 data centers dans le monde, offrant des services d'hébergement généraux (IaaS) aux particuliers et aux organisations. Il existe toutefois encore bien plus de data centers détenus et exploités par des entreprises privées, et ce, pour leur usage personnel.

3.5 Les Modèles de déploiement

Le cloud computing utilise un pool partagé de ressources informatiques (par exemple des réseaux, des serveurs, des espaces de stockage, des applications et des services) afin de fournir un accès réseau à la demande. L'utilisation de la virtualisation dans les environnements de data center permet une évolutivité rapide du cloud computing, avec un minimum de gestion et d'efforts.Comme le montre la figure, le NIST[2](National Institute of Standards and Technology)[3] a défini quatre types de modèles de déploiement de cloud :

- Cloud Public,
- Cloud Privé,
- Cloud Hybride,
- Cloud Communautaire.

3.5.1 Cloud Public

Un cloud public est destiné à être utilisé par le grand public. Son infrastructure est physiquement située sur le site du fournisseur, mais elle peut être détenue par une ou plusieurs organisations, comme des entreprises, des institutions universitaires ou des gouvernements.

Dans un cloud public, l'environnement est entièrement détenu par la société qui met à disposition ses services cloud. Les utilisateurs et clients d'un cloud public n'ont de droit ni sur l'infrastructure, ni sur le matériel, ni sur les logiciels, ni sur quoi que ce soit d'autre. Le fournisseur de cloud met à disposition des utilisateurs des ressources. Il conçoit, gère, maintient et fait évoluer ces ressources en fonction des besoins des clients au fil du temps. Dans le domaine du cloud, et du cloud public en particulier, la sécurité est la clé de voute. L'infrastructure étant partagée avec de nombreux clients la sécurité est de la responsabilité du fournisseur qui en assure la gestion. Ce point est d'autant plus crucial que le client n'au qu'un degré de contrôle et de surveillance très faible des aspects physiques et logiques de sécurité sur les ressources qui sont mises à sa disposition. Le fournisseur doit donc tout mettre en œuvre pour garder la confiance de ses utilisateurs.

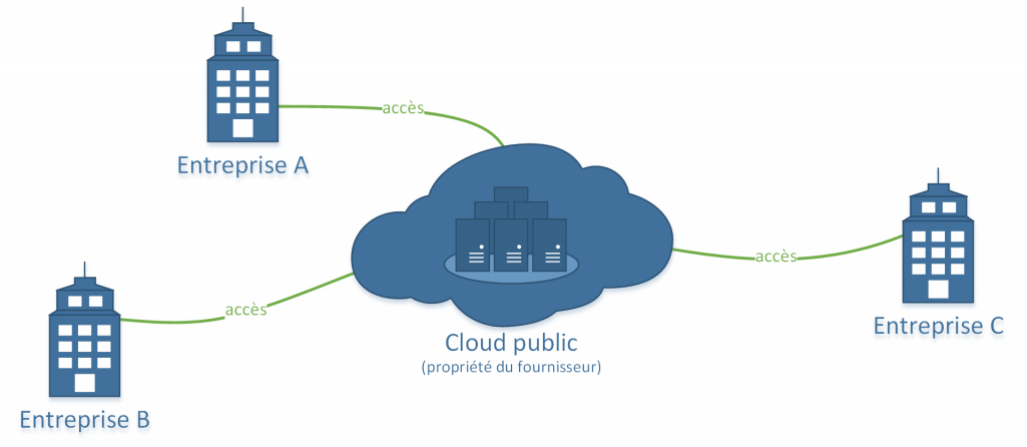


Figure 3.5.1: Modèle de déploiement d'un cloud public.

[2]NIST est une agence du département du Commerce des États-Unis. Son but est de promouvoir l'économie en développant des technologies

[3]`https://www.nist.gov/`

On distingue plusieurs types d'offres en matière de cloud:

- **les webmails:** peuvent offrir d'importanrs espaces de stockage .le plus connue tel plus utilisé étant Gmail (service de Google) . il offre 15 GO (à répartir surtous les autres services google) .Il offre 15Go (à répartir sur tous les autres services Google) pour stocker pièces jointes et mails.
- **Les fournisseurs d'accèsàinternet (FAI)** proposent très souvent un espace de stockage gratuit: 250 Go pour Free,20 Gochez Bouygues Telecom et de 50 à 100 Gopour Orange. Certains FAI proposent même, pour 5euros par mois (en plus de l'abonnement) de disposer d'un espace de stockage illimité.

3.5.2 Cloud Privé

Il est créé exclusivement pour une **organisation unique**. Son infrastructure peut être physiquement située sur le site ou en dehors de celui-ci, et elle peut appartenir à un fournisseur distinct. Un cloud privé n'offre de services qu'aux membres de cette seule organisation.

Le terme cloud privé est utilisé pour décrire l'infrastructure qui émule le cloud computing sur un réseau privé. Le cloud privé a pour ambition d'offrir certains avantages du cloud computing tout en en limitant ses inconvénients. Un cloud privé est détenu par l'entreprise utilisatrice, cela nécessite d'acheter, de construire et de maintenir l'ensemble de ses constituants, ce qui implique de supporter un investissement initial très important.

Les clouds privés diffèrent des clouds publics en ce que les réseaux, serveurs, et infrastructures de stockage qui lui sont associés sont dédiés à une seule entreprise et ne sont pas partagés avec d'autres. Puisque le cloud est entièrement contrôlé par l'entreprise elle-même, les risques de sécurité associés à un cloud privé sont minimisés. Ce haut degré de contrôle et de transparence permet au propriétaire d'un cloud privé de se conformer plus facilement à des normes, politiques de sécurités ou conformités réglementaires qui peuvent être requises dans certains domaines.

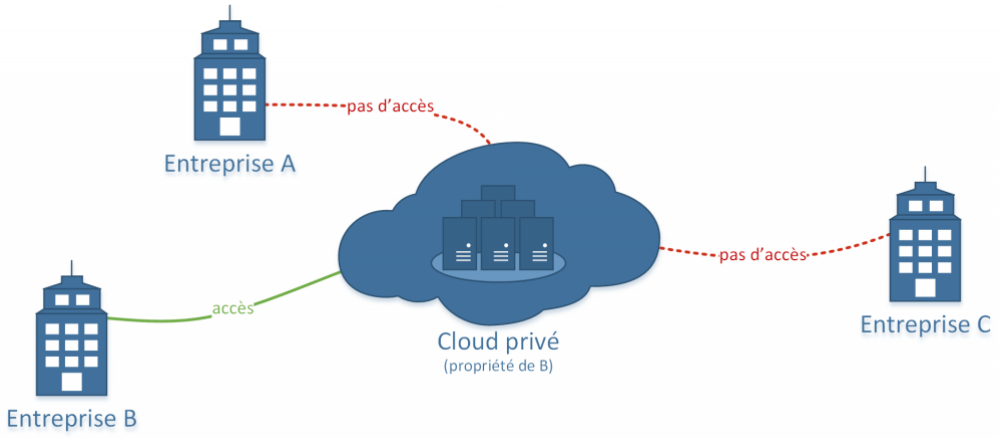


Figure 3.5.2: Modèle de déploiement d'un cloud privé.

3.5.3 Cloud Communautaire

Un cloud communautaire est créé pour une utilisation exclusive par une communauté spécifique. La communauté se compose de plusieurs organisations partageant les mêmes préoccupations (par exemple en matière de mission, d'exigences de sécurité, de stratégie ou de critères de conformité). L'infrastructure peut être physiquement située sur le site ou en dehors de celui-ci, et elle peut appartenir à un fournisseur distinct ou à une ou plusieurs des organisations de la communauté. Les différences entre clouds publics et clouds communautaires se réfèrent aux besoins fonctionnels qui ont été personnalisés pour la communauté. Par exemple, les organisations de soins de santé doivent se conformer à certaines stratégies et réglementations (par exemple, HIPAA) qui nécessitent une authentification et une confidentialité particulières. Les organisations peuvent partager les efforts d'implémentation liés à ces exigences au sein du déploiement d'un cloud commun.

Le cloud de type communautaire est un modèle de déploiement multitenant partagé entre plusieurs entreprises ou organisations et qui est régi, géré, et sécurisé par l'ensemble des participants ou par un fournisseur de service.

Un cloud communautaire est une forme hybride de cloud privé construit et exploité spécifiquement pour un groupe restreint et ciblé. Ces communautés ont des exigences semblables et réunissent leurs moyens humains et financiers pour atteindre leurs objectifs communs.

L'infrastructure commune est spécifiquement conçue pour répondre aux exigences d'une communauté ; à titre d'exemple, des organismes gouvernementaux, des hôpitaux ou des entreprises de télécommunication qui auraient des contraintes de réseau, de sécurité, de stockage, de calcul ou d'automatisation similaires pourraient trouver des intérêts communs à déployer collectivement un cloud communautaire.

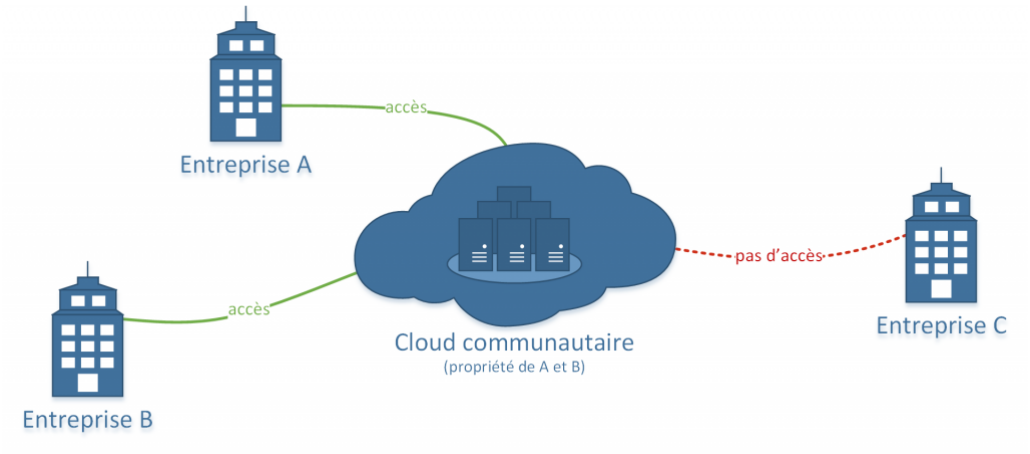


Figure 3.5.3: Modèle de déploiement d'un cloud communautaire

3.5.4 Cloud Hybride

Une infrastructure de cloud hybride est la combinaison d'au minimum deux infrastructures de cloud distinctes (cloud privé, communautaire ou public), représentant des entités uniques. Ces entités sont reliées par le biais d'une technologie permettant la portabilité des données et des applications. Cette portabilité permet à une organisation de conserver un point de vue unique en matière de solution de cloud, tout en profitant des avantages offerts par différents fournisseurs de cloud. Par exemple, la zone géographique (emplacement des utilisateurs finaux), la bande passante, les exigences en matière de stratégie ou de législation, la sécurité et le coût sont des caractéristiques susceptibles de différer d'un fournisseur à l'autre. Un cloud hybride offre une flexibilité permettant de s'adapter et de réagir aux services offerts par ces fournisseurs, et ce, à la demande.

Comme son nom l'indique, un cloud hybride est la combinaison de plusieurs modèles de déploiement de clouds. Avec un cloud hybride, une entreprise peut tirer parti de la simplicité et du faible coût d'un cloud public pour héberger des services classiques ne requérants pas de précautions particulières tout en créant son propre cloud privé pour des applications étroitement intégrées aux systèmes existants ou pour le stockage de données sensibles. Elle a également la possibilité de privilégier l'utilisation de son cloud privé tout en gardant la possibilité de déborder sur une offre de cloud public en cas de besoin temporaire.

Dans un cloud hybride, les clouds public, privé ou communautaire restent des entités uniques, mais sont reliés entre eux par une technologie normalisée ou propriétaire qui permet la portabilité des données et des applications. En raison de la complexité que la combinaison de plusieurs types de clouds engendre, la conception, la gestion et le maintien d'un cloud hybride peuvent être un véritable défi.

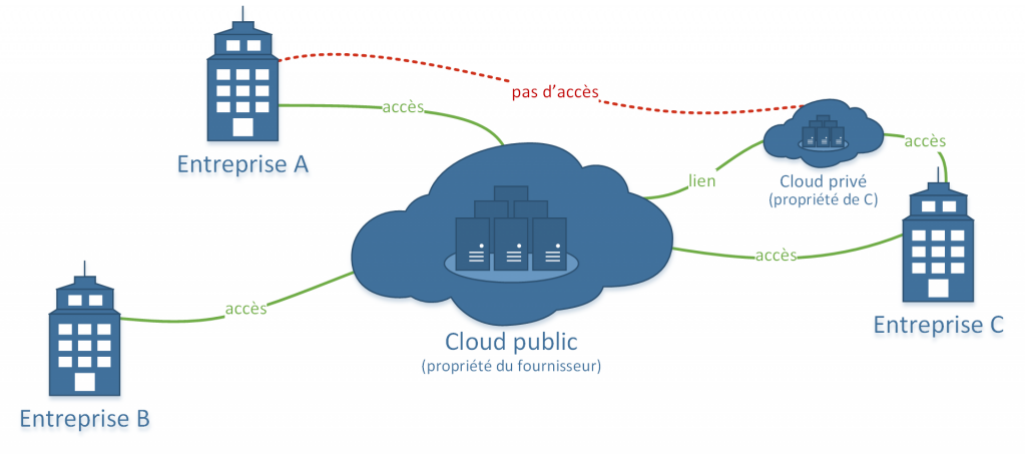


Figure 3.5.4: Modèle de déploiement d'un cloud hybride

3.6 Les Services du Cloud computing:

Le modèle du Cloud Computing est capable de traiter indifféremment les trois couches communément utilisées du modèle de service : 3.6.1[4]

- Le IaaS : Infratructure as a Service.
- Le PaaS : Plateform as a Service.
- Le SaaS : Software as a Service .

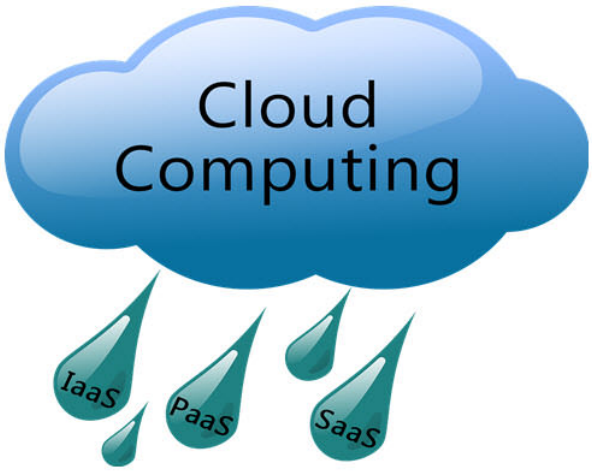


Figure 3.6.1: service cloud computing

La force du Cloud Computing réside dans le fait qu'il puisse soit traverser l'ensemble de ces couches, soit avoir un comportement spécifique à chacune d'entre elles.

3.6.1 Software as a Service (SaaS):

C'est un service qui est accessible à toutes les entreprises et il est facturé au nombre d'utilisateurs. L'entreprise loue les applications du fournisseur de services. Plus besoin d'acheter un logiciel. Ces applications sont accessibles via différentes interfaces, navigateurs Web...etc

Les logiciels disponibles dans le Cloud, en mode Saas :

[4] `https://www.guru99.com/cloud-computing-for-beginners.html`

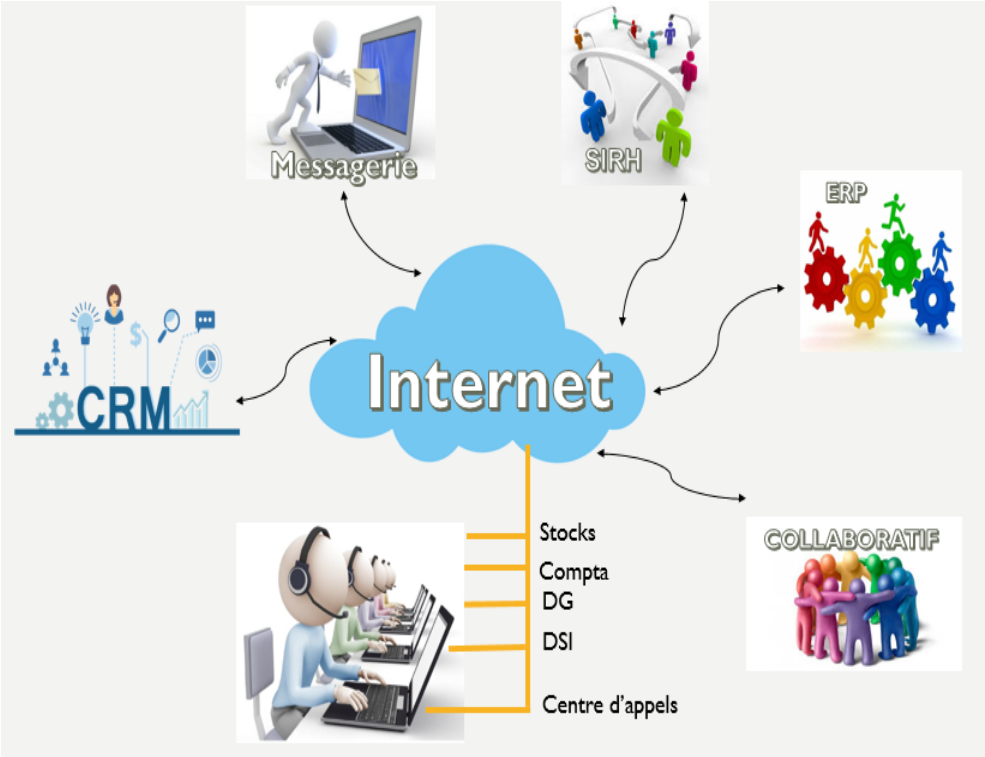


Figure 3.6.2: Les services offerts par un Saas.

- **Messagerie:** Outil incontournable d'informations et de communication, elle consiste a nous bénificier d'une solution de messagerie accessible (anywhere,anytime et anydevice). Cette solution regroupe toutes les fonctionnalités attendues par une société dans le cadre de son travail collaboratif ainsi que la sécurité de ses données.ainsi il doivent répondre a des nombreuses exigences: [5]

  - sécurité et confidentialité.
  - disponibilité.
  - gestion simple et facile.

- **CRM :** Un logiciel de Gestion Relation Client (GRC / CRM) – Customer Relationship Management en anglais ou CRM – est une base de données clients qui permet à une entreprise d'avoir une vision claire et constante de ses clients et futurs clients (prospects).Il représente donc l'ensemble des outils qui permettent d'identifier les prospects, puis traiter, analyser et fidéliser les clients

- **ERP :**Enterprise Ressource Planning a été traduit en français par l'acronyme PGI (Progiciel de Gestion Intégré) et se définit comme un groupe de modules relié à une base de données unique.un progiciel qui permet de gérer l'ensemble des processus opérationnels d'une entreprise en intégrant plusieurs fonctions de gestion. Pour être qualifiée de « Progiciel de Gestion Intégré », une solution logicielle ERP doit couvrir au moins deux principes fondamentaux qui sont les suivants :

  - Construire des applications informatiques sous forme de modules indépendants mais parfaitement compatibles sur une base de données unique et commune.
  - L'usage d'un moteur de Workflow permet de définir l'ensemble des tâches d'un processus et de gérer leur réalisation dans tous les modules du système qui en ont besoin.

- **Collaboration:** Les outils de collaboration(partage de documents, réseaux sociaux… se prêtent bien au mode Saas.

[5] `http://www.serians.fr/solutions-informatiques/offre-cloud/395-cloud-messagerie.html`

3.6.2 La Platform as a Service (PaaS):

Tout comme ITunes ou encore YouTube permettent d'accéder et de naviguer à travers du contenu via une interface web (ou logiciel / application web), l'offre PaaS est un moyen rapide et moins onéreux de développer et lancer sa propre application ou son logiciel.

Le PaaS (Platform as a Service) nous fournira l'infrastructure dont on a besoin pour développer et lancer une application rapidement. Les utilisateurs finaux pourront accéder à des applications personnalisées construites dans le Cloud via le SaaS. Le PaaS permettra au département IT de se concentrer sur l'ajout de valeur et l'innovation plutôt que le maintien des infrastructures techniques. La technologie PaaS permet aux entreprises d'investir les budgets sur l'innovation et la création d'applications à valeur ajoutées plutôt que sur le maintien d'une architecture lourde dans nos Data Centers.

Le PaaS permet la libération de l'innovation et des énergies en nous donnant de l'agilité pour porter la croissance de l'entreprise. Les développeurs peuvent se concentrer sur le développement sans se préoccuper de la tuyauterie technique, du hardware et de l'infrastructure logicielle (Framework).(Fig3.6.3)[6]

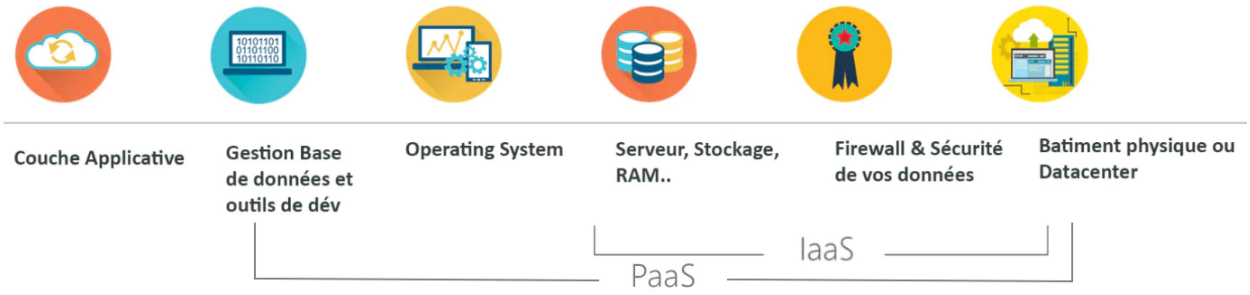


Figure 3.6.3: PaaS vs IaaS

Le PaaS nous permet donc de nous délester de la plupart des services informatiques pour nous concentrer sur le business et sur la valeur ajoutée. En général les services Platform as a Service couvrent les fonctionnalités suivantes :

- Système d'exploitation (Operating system);
- Système de gestion de bases de données (SGBD, NoSQL, etc. . .) ;
- Accessibilité au réseau ;
- Stockage / hébergement .

Avantages d'une offre PaaS (Platform as a Service) :

- **Suppression des infrastructures physiques**, coûteuses en ressources humaines, plus besoin de personnels pour gérer le matériel et des économies énergétiques importantes. Louer une infrastructure virtuelle permettra à nos équipes de développement de se concentrer sur un produit.
- **Pay as you use**, on ne paye que ce qu'on consomme.
- **Agilité**, on garde le contrôle sur notre outillage, on n'installe que les outils qui nous seront vraiment utiles.
- **Collaboration**, le PaaS permettra aux équipes de développeurs de construire des applications, site internet ou produit depuis n'importe où dans le monde avec la haute disponibilité Oracle.
- **Les développeurs ne perdront plus jamais leur code**, quand ils développeront sur le PaaS, avec la réplication et le back-up automatisés.

[6] http://stackify.com/wp-content/uploads/2017/11/img_59fc9c14be80f.png

3.6.3 Infrastructure as a Service (IaaS)

L'IaaS pour Infrastructure As A Service, désigne la couche basse du Cloud Computing et peut être considéré comme un point d'entrée dans les technologies du Cloud. L'IAAS est la mise à disposition de ressources matérielles comme des unités de puissance de calcul, de traitement, des capacités de stockage. Le fournisseur met donc à disposition de ses clients des serveurs « virtualisés » évolutifs suivant la demande. donc est une externalisation de l'infrastructure informatique matérielle. En souscrivant à un abonnement auprès d'un fournisseur, on lui délégue :

- L'installation des serveurs fichiers.
- Les réseaux.
- Le storage (stockage) de nos données.

Cette location évite de nous dôter de ces équipements et de devoir les payer par ailleurs. Le prestataire nous soulage de la gestion du cloud, du matériel, du stockage et des réseaux. Par contre, c'est à nous de gérer :

- Les serveurs.
- Les logiciels et leur paramétrage.
- L'intégration SOA (architecture orientée services).

Comme exemple d'IaaS, on peut citer l'IaaS d'Amazon, celui de Numergy ou encore les logiciels open source.

Les avantages de l'IaaS:

- **Gain d'efficacité :** les ressources sont virtualisées et regroupées, de sorte que l'infrastructure est utilisée au maximum de sa capacité
- **Meilleure réactivité :** les ressources IT peuvent être provisionnées à la demande et replacées dans le pool de ressources avec la même facilité
- **Évolutivité rapide :** allocation instantanée de ressources de calcul supplémentaires pour répondre aux besoins métiers en période de pointe, de croissance ou de décroissance de l'entreprise
- **Baisse des coûts :** dépenses liées à l'infrastructure, à l'énergie et aux locaux, modèle de « paiement à l'utilisation »
- **Gain de productivité du personnel IT :** provisionnement automatisé à travers un portail en libre-service
- **Moins de ressources gaspillées :** la transparence de la tarification et du suivi d'utilisation ainsi que les outils de facturation interne permettent aux administrateurs IT de savoir où ils peuvent réaliser des économies
- Taux d'utilisation supérieur des investissements IT
- Sécurité renforcée et protection accrue des informations.

En résumé

Dans le modèle classique, l'utilisation de logiciels d'entreprise classiques nécessite des investissements matériels (coût des locaux, des serveurs, du matériel de sauvegardes, des équipements réseau), logiciels (coût d'achat des licences, coûts annuels de l'assistance, des mises à jour, des changements de versions) et humains conséquents. En contrepartie, l'entreprise garde son indépendance et une totale maitrise de son infrastructure.

Avec le modèle IaaS, le prestataire héberge l'infrastructure informatique de l'utilisateur ou plus généralement de l'entreprise. Cette dernière peut gérer à distance son infrastructure comme si celle-ci se trouvait dans ses propres locaux, les contraintes matérielles en moins. Le prestataire s'occupe ici de la virtualisation, du stockage, des réseaux, du matériel, et de la continuité de service.

Le PaaS désigne de manière générale l'environnement fourni par le prestataire au client. C'est un socle fonctionnel et prêt à l'emploi qui lui est mis à disposition. L'utilisateur aura donc la possibilité de gérer ses applications au sein de celui-ci.

Enfin, le SaaS consiste à fournir des applications hébergées et accessibles en ligne par le biais d'un navigateur web. Les logiciels se présentent sous la forme de services. Une inscription et le paiement d'un droit d'accès permettent généralement à l'utilisateur d'exploiter ces derniers.

La figure suivante illustre les principaux composants de ces quatre modèles, ainsi que les niveaux de responsabilités qui incombent à l'entreprise et au fournisseur de cloud sur chacun des composants.

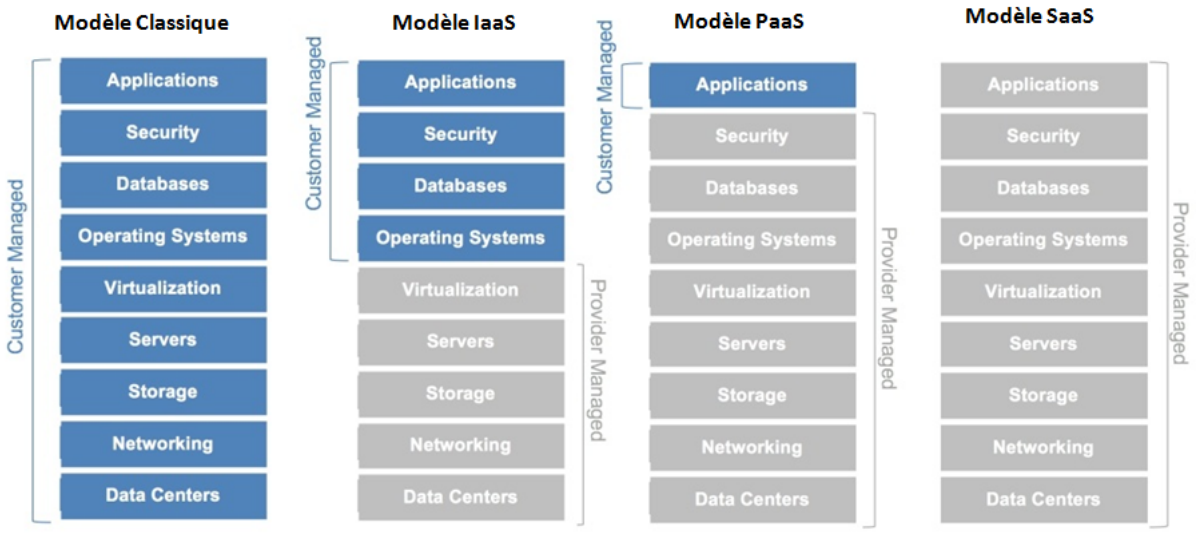


Figure 3.6.4: Les services qu'offre chaque modéle.

3.7 Architecture cloud computing

L'architecture Cloud Computing comprend de nombreux composants de cloud, qui sont faiblement couplés.

Chacune des extrémités est connectée via un réseau, généralement Internet. la figure suivant montre la vue graphique de l'architecture de cloud computing Nous pouvons diviser l'architecture du cloud en deux parties:

Frontend (L'extrémité avant):

Le Frontend fait référence à la partie client du système de cloud computing. Il se compose d'interfaces et d'applications nécessaires pour accéder aux plates-formes de cloud computing, exemple: Web Browser.

Back End:

L'extrémité arrière fait référence au nuage lui-même. Il comprend toutes les ressources nécessaires pour fournir des services de cloud computing. Il comprend un énorme stockage de données, des machines virtuelles, un mécanisme de sécurité, des services, des modèles de déploiement, des serveurs, etc.

Remarque. ***Le cloud computing distribue le système de fichiers qui s'étend sur plusieurs disques durs et machines. Les données ne sont jamais stockées dans un seul endroit et dans le cas où une unité échoue, l'autre prendra automatiquement le relais. L'espace disque de l'utilisateur est alloué sur le système de fichiers distribué, tandis qu'un autre composant important est l'algorithme pour l'allocation des ressources. Le cloud computing est un environnement fortement distribué et il dépend fortement d'un algorithme fort.***

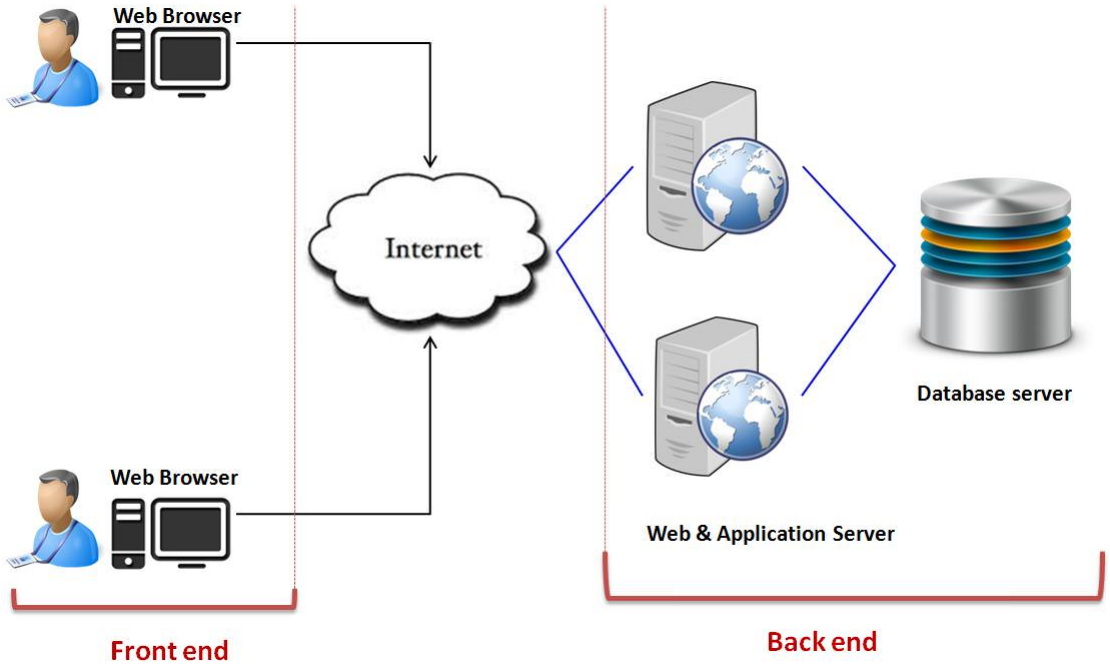


Figure 3.7.1: Architecture du cloud computing.

3.8 Les Technologies du Cloud Computing:

3.8.1 La Virtualisation

Historiquement, chaque ordinateur disposait de son propre système d'exploitation, de ses propres applications et de ses propres composants matériels dédiés. À l'heure actuelle, à l'aide de l'émulation logicielle, plusieurs ordinateurs virtuels peuvent s'exécuter sur un même ordinateur physique. Cela signifie que chaque ordinateur virtuel possède son propre système d'exploitation, ses propres applications et ses propres composants matériels dédiés. En informatique, cette technologie porte le nom de virtualisation. Chaque machine virtuelle illustrée dans la figure fonctionne de manière indépendante:

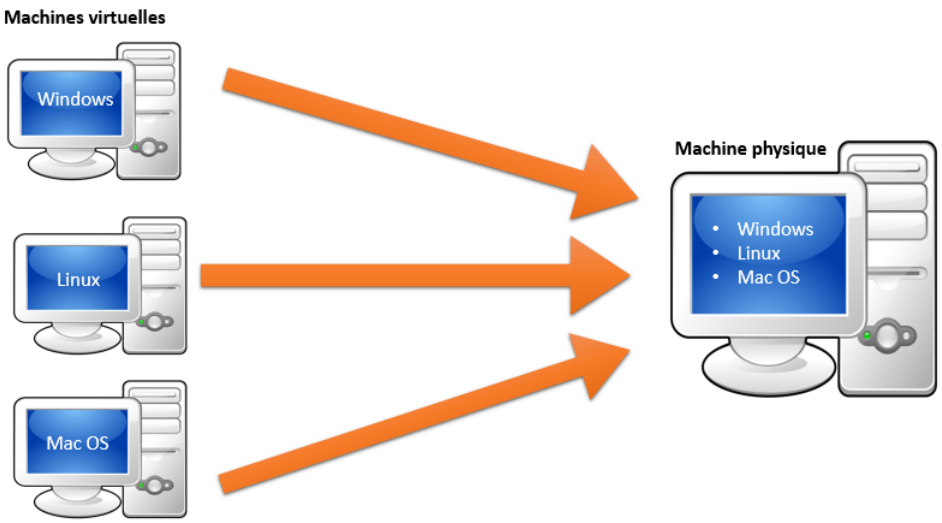


Figure 3.8.1: Virtualisation.

Dans le monde de l'entreprise, une infrastructure physique unique peut exécuter plusieurs infrastructures virtuelles. La virtualisation des serveurs et des réseaux permet aux entreprises de diminuer leurs frais de fonctionnement et leurs coûts administratifs. Des économies opérationnelles peuvent ainsi être réalisées grâce à la réduction des besoins en alimentation et en refroidissement, en raison de la diminution du nombre de machines physiques. On peut ajouter un serveur virtuel afin de prendre en charge

des applications supplémentaires.De nombreuses entreprises populaires comme VmWare et Microsoft fournissent des services de virtualisation, où au lieu d'utiliser un PC personnel pour le stockage et le calcul, on utilise leur serveur virtuel. Ils sont rapides, rentables et prennent moins de temps.

On peut également utiliser la virtualisation pour nos besoins personnels en informatique. On peut essayer un nouveau système d'exploitation sur un ordinateur sans endommager un système actuel. On peut naviguer sur Internet en toute sécurité avec une machine virtuelle. La machine virtuelle peut être supprimée en cas de problème.

Pour les développeurs de logiciels et les testeurs, la virtualisation est très pratique, car elle permet au développeur d'écrire du code qui s'exécute dans de nombreux environnements différents et, plus important encore, de tester ce code.La virtualisation est principalement utilisée à trois fins principales:

- **Virtualisation de réseau :** Il s'agit d'une méthode de combinaison des ressources disponibles dans un réseau en divisant la bande passante disponible en canaux, chacun étant indépendant des autres et chaque canal étant indépendant des autres et pouvant être affecté à un serveur ou à un périphérique spécifique. temps réél.

- **Virtualisation du stockage:** il s'agit de la mise en commun du stockage physique de plusieurs périphériques de stockage réseau dans ce qui semble être un périphérique de stockage unique géré à partir d'une console centrale. La virtualisation du stockage est couramment utilisée dans les réseaux de stockage (SAN).

- **Virtualisation du serveur:** la virtualisation du serveur est le masquage des ressources du serveur telles que les processeurs, la RAM, le système d'exploitation, etc., à partir des utilisateurs du serveur. L'intention de la virtualisation des serveurs est d'augmenter le partage des ressources et de réduire la charge et la complexité du calcul des utilisateurs.

La virtualisation est la clé pour débloquer le système Cloud, ce qui rend la virtualisation si importante pour le cloud, c'est qu'il déconnecte le logiciel du matériel. Par exemple, les PC peuvent utiliser la mémoire virtuelle pour emprunter de la mémoire supplémentaire sur le disque dur. Généralement, le disque dur a beaucoup plus d'espace que la mémoire. Bien que les disques virtuels soient plus lents que la mémoire réelle, s'ils sont gérés correctement, la substitution fonctionne parfaitement. De même, il existe un logiciel qui peut imiter un ordinateur entier, ce qui signifie qu'un ordinateur peut effectuer les fonctions égales à 20 ordinateurs.[7]

3.8.2 Architecture orientée service (SOA):

L'architecture orientée services permet d'utiliser des applications en tant que service pour d'autres applications, quel que soit le type de fournisseur, de produit ou de technologie. Par conséquent, il est possible d'échanger les données entre les applications de différents fournisseurs sans programmation supplémentaire ou apporter des modifications aux services.L'architecture orientée service cloud computing est représentée dans le diagramme ci-dessous.

[7] `https://www.guru99.com/cloud-computing-for-beginners.html`

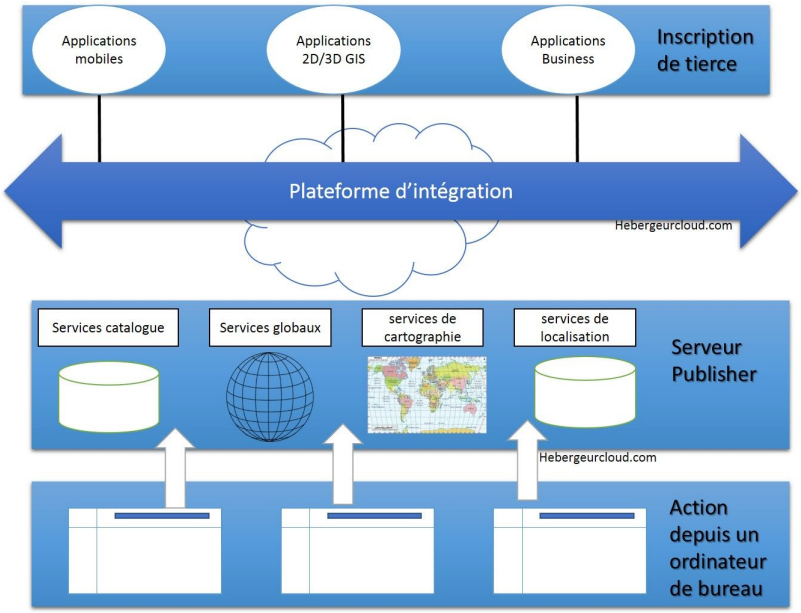


Figure 3.8.2: architecture orientée service cloud computing.

3.8.3 Calcul en grille (Grid Computing):

Grid Computing fait référence à l'informatique distribuée, dans laquelle un groupe d'ordinateurs provenant de plusieurs endroits sont connectés les uns aux autres pour atteindre un objectif commun. Ces ressources informatiques sont hétérogènes et géographiquement dispersées.

Grid Computing casse la tâche complexe en plus petits morceaux, qui sont distribués aux processeurs qui résident dans la grille.

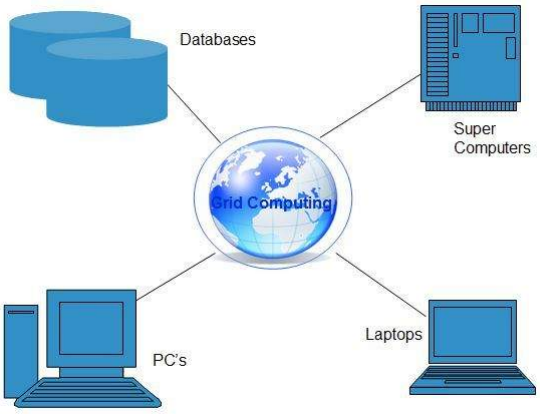


Figure 3.8.3: Grid Computing

Grid Computing est un moyen de coordonner des ressources informatiques disparates à travers un réseau, leur permettant de fonctionner dans son ensemble. Il est plus souvent utilisé dans la recherche scientifique et dans les universités à des fins éducatives. Par exemple, un groupe d'étudiants architectes travaillant sur un projet différent nécessite un outil de conception spécifique et un logiciel de conception, mais seulement deux d'entre eux ont accès à cet outil de conception, le problème est de savoir comment ils peuvent rendre cet outil disponible. élèves. Pour rendre disponible pour les autres étudiants, ils mettront cet outil de conception sur le réseau du campus, maintenant la grille connectera tous ces ordinateurs dans le réseau du campus et permettra aux étudiants d'utiliser l'outil de conception nécessaire pour leur projet de n'importe où.

Le cloud computing et l'informatique en grille sont souvent confus, bien que leurs fonctions soient presque similaires car leurs fonctionnalités sont différentes.

3.9 Cloud computing et sécurité:

Toutes les enquêtes montrent que la sécurité est la préoccupation majeure des organisations dans le processus d'adoption des technologies Cloud. Les questions sont nombreuses comme par exemple :

- Quelle confiance peut-on avoir dans le stockage des données à l'extérieur de l'entreprise ?
- Quels sont les risques associés à l'utilisation de services partagés ?
- Comment démontrer la conformité des systèmes à des normes d'exploitation ?

Les infrastructures Cloud sont de gigantesques systèmes complexes. Ils peuvent cependant être réduits à un petit nombre de primitives simples qui sont instanciées des milliers de fois et à quelques fonctions communes. La sécurité du Cloud est donc un problème gouvernable moins complexe qu'il n'y parait.

3.9.1 Les composants sécurité d'un système de Cloud computing:

Les différents composants qui participent à la sécurité d'un système de Cloud computing présentent les caractéristiques suivantes :

- **Service de console de gestion (Provisioning):** La mise en route et la reconfiguration des composants des systèmes sont très rapides. Il est possible de mettre en service plusieurs instances dans plusieurs centres de traitement répartis dans le monde en quelques minutes. Les reconfigurations réseau sont facilitées. En revanche, la sécurité d'utilisation de la console de gestion devient impérative (authentification multi-facteurs, connexion chiffrée, etc..).
- **Infrastructures de calcu:** Un des gros avantages du Cloud pour le développement et l'exploitation des applications réside dans la virtualisation. Elle permet de préparer des configurations maîtres sûres qu'il suffit de dupliquer pour déployer. Les défis restent la sécurisation des données dans les applications partagées et la sécurité entre les instances garantie par les hyperviseurs.
- **Services de support:** La principale caractéristique du Cloud est la mise en place a priori d'une sécurité renforcée et auditable (authentification, logs, pare-feux, etc..). Il reste à traiter les risques liés à l'intégration avec les applications des utilisateurs ainsi que les processus toujours délicats de mises à jour.
- **Sécurité périmétrique du réseau Cloud:** Ces grandes infrastructures partagées fournissent des moyens de protection au delà des capacités d'une entreprise normale comme par exemple la protection contre les attaques DDOS (Distributed Denial Of Service). Les mécanismes de sécurité périmétriques sont généralement bien conçus (fournisseur d'identité, authentification, pare-feux , etc..). En revanche, il reste à traiter les sujets liés à la mobilité.
- **Service de stockage des données:** Les avantages du stockage des données dans le Cloud dépendent des fournisseurs mais en général, ceux-ci fragmentent et répartissent les données. Celles ci sont aussi souvent recopiées dans des centres de traitement différents. Ces opérations améliorent considérablement la sécurité des données. Si leur contenu doit rester confidentiel, il convient de les chiffrer avant de les stocker.

3.10 Les avantages et inconvénients du cloud computing :

le cloud computing est une technologie qui à marqué sa présence aujourd'hui, elle a eu un énorme impacte sur le développement de l'informatique en général néanmoins comme toute technologie elle possède également ses avantages et ses inconvénients :

3.10.1 Avantages du cloud:

- **Externalisé :** sauvegarde , données et documents sont stockés sur un serveur distant. le disque dur de l'ordinateur est alégé et nos documents importants conservés. de plus , on a plus à nous occuper de la mises à jour, le fournisseur de services est là pour prévenir panne et mises à jour éventuelles.

- **Mobile et accessible:** l'accès au cloud est relativement aisé: une adresse et un mot de passe suffisant. les données conservées sont accessibles depuis n'importe quel appareil, depuis n'importe où et à n'importe quel moment. Nul besoin de puissants ordinateurs pour utiliser certains logiciels, il suffit seulement d'un accès à internet.

- **Économique:** fini les dépenses en matériel de stockage(clés USB, disque durs...)En cas de problème et n'être pas responsable.En principe, on a une totale liberté vis-à-vis de notre fournisseur de service, l'abonnement est résiliable à tous moment.

- **Collaboration et partage:** la plupart des services de stockage en ligne offrent la possibilité de partager des contenus, favorisant ainsi le travail de plusieurs personnes sur un même document. Le partage sur les réseaux sociaux est aussi tés utilisé, notamment pour les photos.

3.10.2 Inconvénients du cloud:

- **Pérennité:** Qu'arrive-t-il si le fournisseur décide de mettre fin au service? amazon, hubic Dropbox précisent dans leurs conditions d'utilisation qu'ils peuvent fermer leur service n'importe quand et sans préavis. on est donc prévenus... n'envisagez pas le cloud comme unique moyen de sauvegarde ! Qu'en est-il du risque lié à l'effacement de données? En supprimant une donnée du cloud, celle-ci devrait l'être sous toutes ses formes.Or , elles peuvent être conservées dans plusieurs data centers. Comment pouvons-nous avoir la preuve qu'une donnée a bien été supprimée définitivement?

- **Connexion:** Le cloud c'est tés pratique: accéder à ses documents n'importe quand et de n'importe où... Mais si l'utilisateur n'a pas de connexion internet ou une connexion insuffisante , il ne poura pas y accéder.

- **Sécurité:** où va l'information stockée dans les nuages ? qui y a accès?A priori, les données envoyées dans le cloud sont en sécurité. Quel que soit le prestataire, les serveurs qui conservent nos données sont concentrés dans d'immenses parcs informatiques (data centers) suveillés en permanence afin d'éviter les attaques matérielles. Mlfré tout, et bien qu'il soit faible, le risque d'intrusion de pirates informatiques dans ces serveurs est une réalité. En confiant nos données à un tiers,d'une certaine manière, on renonce à leur contrôle.

- **Confidentialité:** plus que de la sécurité des serveurs,c'est de leur géolocalisation dont il faut se méfier par exemple en France, grâce à la CNIL, ils sont bien protégés en matière de confidentialité des données personnelles.Mais rare sont les services qui hébergent ces données sur le territoire français. Le prestataire peut également utiliser nos données à des fins marketing. Google, par exemple, qui affiche des publicité en lien avec les contenus de nos mails ou Facebook qui vend les informations des profils des sociétés

3.11 Base de données Cloud Computing:

Le cloud computing fait référence à un large éventail de produits logiciels vendus en tant que service, gérés par un fournisseur tiers et distribués via un réseau basé sur le cloud. Infrastructure-as-a-Service (IaaS) est une solution de cloud computing qui offre des ressources de traitement, de stockage ou de réseau à la demande. IaaS a du sens pour beaucoup d'entreprises parce que c'est:

- **Moins cher :** Ne payer que pour ce dont on a besoin pas besoin d'investir dans les ressources nécessaires pour garantir la disponibilité.

- **Élastique :** Ajouter et supprimer facilement des ressources pour gérer les évènements inattendus tels que les pics de trafic sur une application.

- **Adapté:** Ajouter de la bande passante, des capacités de traitement et de stockage à n'importe quel taux ou incréments nécessaires.

- **Fiable:** Les serveurs distribués à travers les zones géographiques permettent une meilleure reprise après sinistre et la continuité des opérations, et les données peuvent être diffusées localement aux utilisateurs.

La nouvelle génération de bases de données NoSQL est particulièrement bien adaptée aux environnements de cloud computing car elles gèrent généralement la charge en répartissant les données entre de nombreux serveurs.

MongoDB, la principale base de données NoSQL, est un outil naturel pour le cloud. Grâce à son architecture native évolutive, MongoDB permet la mise à l'échelle horizontale par «sharding». Sharding nous permet de répartir automatiquement les données de manière homogène sur les clusters multi-nœuds et d'équilibrer les requêtes entre eux.

De nombreux utilisateurs de MongoDB exécutent leur déploiement de base de données dans le cloud pour tirer parti de ces avantages. Amazon Web Services (AWS), un partenaire de cloud computing MongoDB, est un choix populaire pour ces utilisateurs. C'est une excellente solution pour ceux qui ont besoin d'opérations hautes performances sur de grands ensembles de données.

3.11.1 NoSQL dans le Cloud: une alternative évolutive aux bases de données relationnelles

Si on utilise une base de données relationnelle traditionnelle, on devra peut-être travailler sur une stratégie complexe pour répartir la charge de la base de données entre plusieurs instances de base de données. Cette solution présentera souvent beaucoup de problèmes et ne sera probablement pas idéale pour la mise à l'échelle élastique. Pourquoi ne pas envisager d'utiliser une base de données **NoSQL** basée sur le cloud comme alternative?

Avec le passage actuel au cloud computing, la nécessité de dimensionner les applications se présente comme un défi pour le stockage des données. Si on utilise une base de données relationnelle traditionnelle, on doit peut-être travailler sur une stratégie complexe pour répartir la charge de la base de données entre plusieurs instances de base de données. Cette solution présentera souvent beaucoup de problèmes et ne sera probablement pas idéale pour la mise à l'échelle élastique. Comme alternative, on peut envisager une base de données NoSQL basée sur le cloud.et il est nécessaire de comprendre que les bases de données NoSQL distribuées atteignent une grande évolutivité par rapport à un SGBDR traditionnel en faisant des compromis importants.Un bon point de départ pour y penser est le théorème CAP , qui indique qu'une base de données distribuée peut, au mieux, fournir deux des éléments suivants: Cohérence, disponibilité et tolérance de partition. Nous définissons chacun de ceux-ci comme suit:

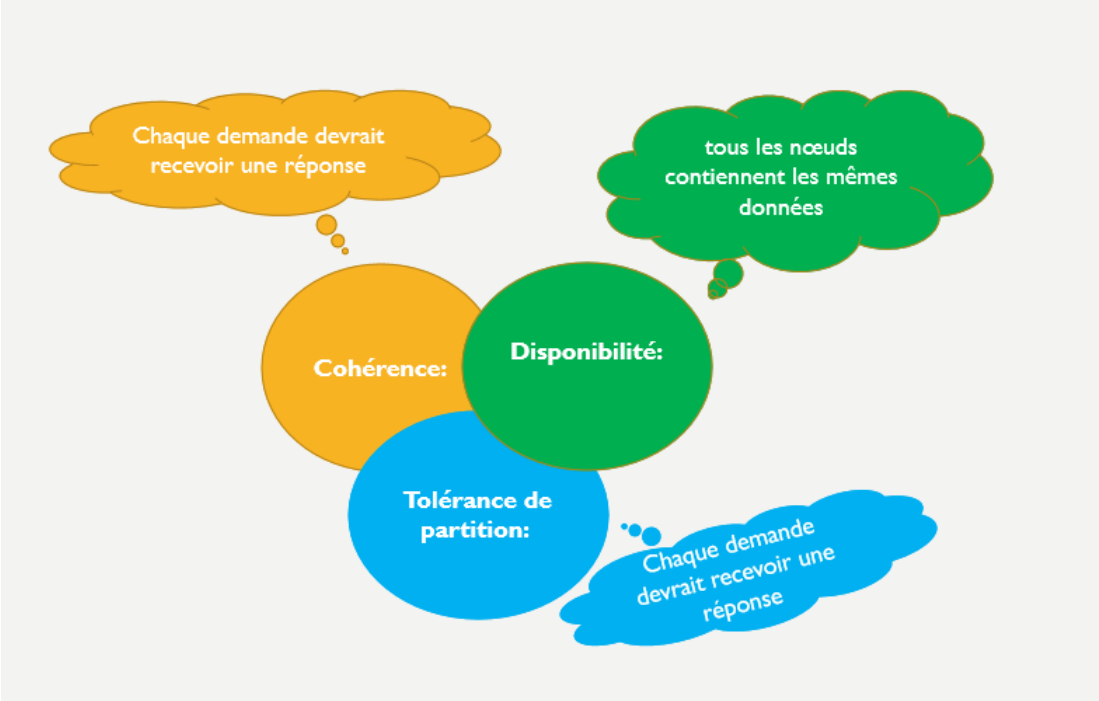


Figure 3.11.1: théorème CAP

3.11.2 Critères de migration vers le NOSQL:

C'est une évidence de dire qu'il convient de choisir la bonne technologie en fonction du besoin. Il existe cependant certains critères déterminants pour basculer sur du NoSQL.

- **Taille :** Nous sommes dans un monde où il y a des données ayant une masse considérable (qu'on appelle infobésité). Il sied d'avoir alors un système pouvant supporter un nombre important des opérations, d'utilisateurs, des données, etc. de manière optimale. Bien que tous les systèmes NoSQL ne soient pas conçus pour cette problématique, il est possible d'en trouver sans problème.
- **Performance en écriture :** Intérêt principal du géant Google, Facebook (135 milliards de messages par mois), Twitter (7TB de données par jour). Des données qui augmentent chaque année. A 80MB/s cela prend une journée pour stocker 7TB, donc l'écriture doit être distribuée sur un cluster, ce qui implique du MapReduce, réplication, tolérance aux pannes, consistance,... Pour des performances en écriture encore plus puissante, il convient de se tourner vers les systèmes in-memory.
- **Performance en lecture clé-valeur :** Certaines solutions NoSQL ne possèdent pas cet avantage mais comme il s'agit d'un point clé, la plupart d'entre elles en sont dotées.
- **Type de données flexibles :** Les solutions NoSQL supportent de nouveaux types de données et c'est une innovation majeure. Divers types de données, souvent des données complexes ne peuvent pas être mis sous forme de données relationnelles, d'où l'adaptation à un nouveau type des données.
- **ACID :** Bien que ce ne soit pas le but premier du NoSQL, il existe des solutions permettant de conserver certains (voire tous) aspects des propriétés ACID. Se référer au théorème CAP plus haut et aux propriétés BASE.
- **Simplicité de développement :** L'accès aux données est simple. Bien que le modèle relationnel soit simple pour les utilisateurs finaux (les données sont restituées selon la structure de la base), il n'est pas très intuitif pour les toujours être celle d'embauche d'un administrateur de base de données, le développeur devrait pouvoir être en mesure de le résoudre.
- **Parallel Computing :** Les solutions NoSQL améliorent les calculs parallèles.

3.11.3 Critères de choix pour mieux choisir le type de BDD:

Certaines comme HBase, MongoDB ou Neo4j apparaissent régulièrement dans l'actualité et sont toutes libellées en tant que NoSQL. Il existe cependant des différences significatives entre elles, liées notamment au théorème de CAP. Ainsi, nous proposons le critère de choix suivant lorsque nous sommes buttés au choix d'un système de gestion de base de données du type NoSQL :

- Choisir une base de données orientées-document car elle s'adapte aux données non planes (type profil utilisateur).

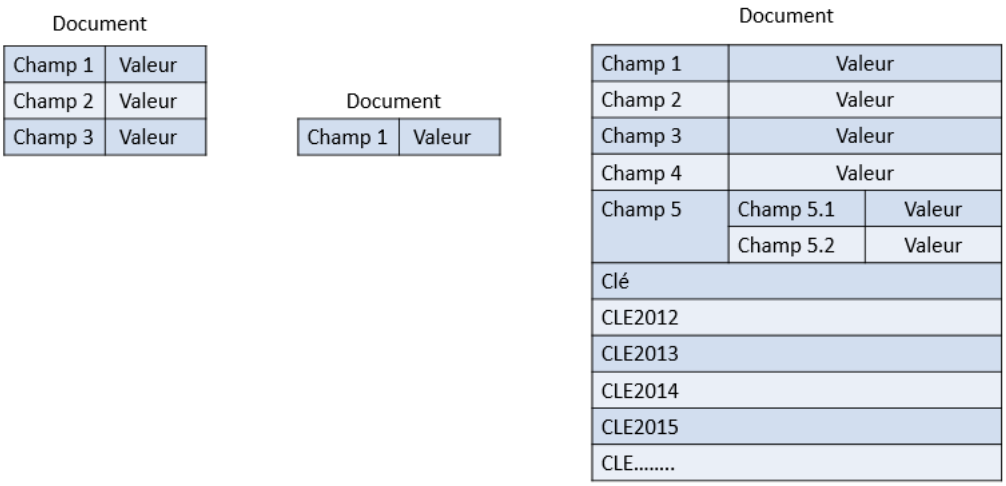


Figure 3.11.2: BBD orientée documents

Quelques SGBD orientées-document :

1. ***MongoDB :*** Développé en C++. Les API officielles pour beaucoup de langages.Protocole personnalisé BSON. Réplication master/slave. Licence AGPL (commercial et libre) ;
2. ***CouchDB :*** Développé en Erlang. Protocol http. Réplication master/master. Licence Apache.

- Choisir une base de données orientées-colonne car elle s'adapte mieux au stockage des listes (messages, postes, commentaires, ...).

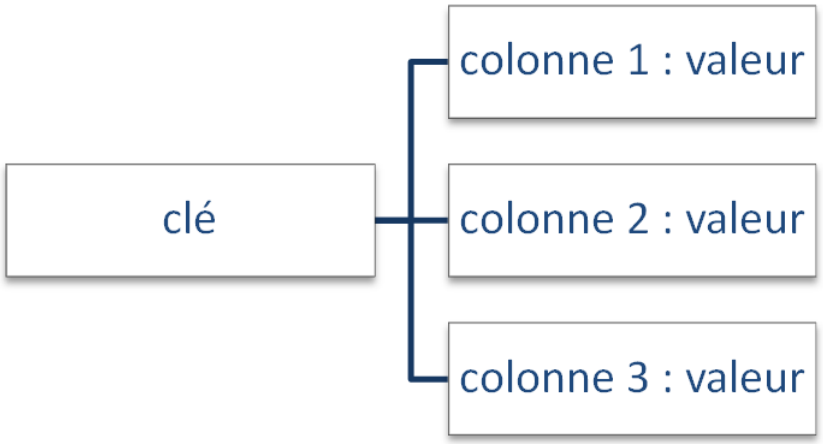


Figure 3.11.3: BDD orientée colomne

Quelques SGBD orientées-colonnes :

1. ***HBase:*** Utilise un API Java. Adopte un design CA. Présence de quelques SPOF.
2. ***Cassandra:*** Beaucoup d'API disponibles. Adopte un design AP avec consistance éventuelle. Moins performant que HBase sur les insertions de données.

- Choisir une base de données orientées-graphe pour mieux gérer les relations multiples entre objets (comme des relations dans un réseau social).

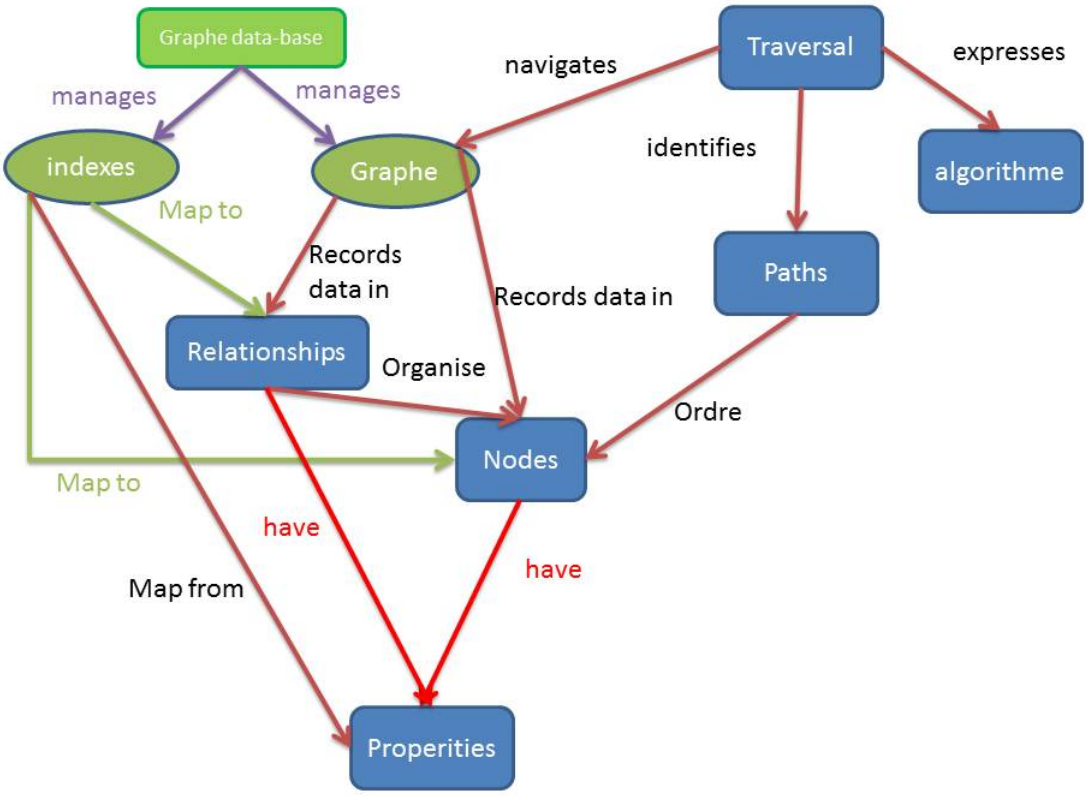


Figure 3.11.4: BDD orientée graphe

Quelques SGBD orientées-graphe :

1. ***Neo4J:*** Développé en Java. Supporte beaucoup de langages. Réplication master/slave. Propriétés ACID possibles. Langage de requêtes personnalisé «Cypher».
2. ***Titan:*** Haute disponibilité avec réplication master/master. Prise en compte d'ACID avec consistance éventuelle. Intégration native avec le framework TinkerPop.

- Choisir une base de données orientées-clé-valeur car elle permet d'accéder rapidement aux informations pour la gestion des caches.

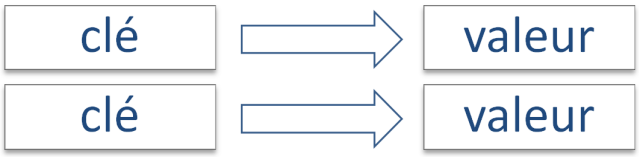


Figure 3.11.5: BDD orientée clé-valeur

Quelques SGBD orientées-clé-valeur :

1. ***DynamoDB:*** Solution d'Amazon à l'origine de ce type de base. Design de type AP selon le théorème de CAP mais peut aussi fournir une consistance éventuelle.
2. ***Voldemort:*** Implémentation open-source de Dynamo. Il y a possibilité d'en faire une base embarquée.

3.11.4 Défis majeurs des BDD NOSQL:

La plupart des organisations cherchant à migrer vers les NoSQL se révèlent des grandes organisations traitant d'énormes masses de données, ainsi ayant d'énormes besoins de stockage. Il faut aussi le signaler que, les petites organisations ne regardent presque pas à la même direction ce concept qu'est le NoSQL.

Dans une enquête menée par une communauté d'enquêté appelée Information Week, 44% des professionnels en activité de l'informatique n'ont pas entendu parler de NoSQL. En outre, seulement 1% des répondants ont indiqué que NoSQL est une partie de leur orientation stratégique. De toute évidence, NoSQL a sa place dans notre monde connecté, mais devra continuer à évoluer pour obtenir l'appel de masse que beaucoup pensent qu'elle pourrait avoir. La technologie NoSQL ne cesse de faire parler d'elle et semble avoir le vent en poupe. Attrayante, la barrière d'entrée pour un nouveau développement est d'ailleurs assez peu élevée pour tout développeur ayant bien compris les sous-jacents de la solution retenue. Néanmoins, il est essentiel de garder à l'esprit que NoSQL apporte une réponse à des besoins bien spécifiques. Dit autrement, il est nécessaire d'avoir identifié au préalable la nécessité d'utiliser cette technologie dans nos services.

Cependant, il faut se mettre d'accord que bien que robuste, le NoSQL reste encore une technologie en gestation et doit par conséquent ne pas cesser d'évoluer d'une manière quantitative que qualitative (en se dotant par exemple de solutions ORM éprouvées, en gommant l'absence d'un langage de requêtage commun et capitaliser sur l'utilisation, comme nos chers SGBDR l'ont fait sur les 20 dernières années).

Les bases de données NoSQL ont suscité beaucoup d'enthousiasme, mais il y'a de nombreux obstacles à surmonter avant de pouvoir faire appel aux principaux acteurs de l'industrie des bases de données. Voici quelques-uns des principaux défis.

Conclusion

Le Cloud Computing est aujourd'hui un élément clé de la transformation numérique des entreprises. Il permet à ces dernières de se dégager de la contrainte technique au profit de l'agilité et de l'adaptation du service aux besoins des Métiers. Il permet aussi de répondre efficacement à la problématique de la mobilité en donnant accès aux informations et services en tous lieux.[8]

[8] [BR15] R. BRUCHEZ: Les bases de données NoSQL et le Big Data. Comprendre et mettre oeuvre, 2ème Ed. EYROLLES, Paris, 2015.

Les bases de données NoSQL et le Big Data - Rudi Bruchez.

Part II

Conception et Réalisation

Chapter 4

Analyse et conception

Comme tout projet informatique, ce travail doit se référer à une démarche de conduite de projet celle-ci doit suivre les phases suivantes:

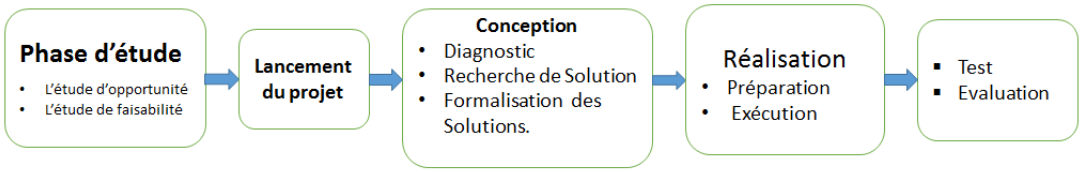


Figure 4.0.1: Démarches de la conduite de projet

- **L'étude d'opportunité** qui permet au commanditaire ou au sponsor et à la direction générale d'apprécier la pertinence économique de lancer ou non le projet. Cette pertinence s'analyse au regard du retour sur investissement ou de l'impact sur le business de l'entreprise ou des impacts sociaux ou réglementaires.
- **L'étude de faisabilité** technique qui permet au commanditaire ou au sponsor et à la direction générale de lancer ou non le projet en toute connaissance de ses différents impacts techniques.
- **Le lancement d'un projet** se fait à partir d'une note de lancement. Celle-ci officialise le lancement du projet auprès de l'ensemble des responsables et des personnes concernées par le projet dans l'entreprise. La note de lancement est rédigée et diffusée au démarrage du projet.

  Le document de lancement qui sera présenté à l'équipe projet doit reprendre les éléments suivants :

  1. Le contexte du projet
  2. Le rappel des enjeux et de la problématique
  3. Les objectifs fixés au projet
  4. L'organisation du projet
- **Le diagnostic de la situation** a pour objectif d'analyser la situation existante afin d'identifier, dans le cadre d'un projet de réorganisation, les différents axes d'amélioration. Dans cette perspective il est important de bien comprendre le contexte et les enjeux pour lesquels la réorganisation est souhaitée par les dirigeants de l'entreprise.
- **La recherche de solutions** est le résultat du diagnostic. Une bonne connaissance de l'existant facilite la recherche de solutions. Les solutions retenues doivent donner par ailleurs une réponse aux dysfonctionnements constatés.
- **La formalisation des solutions** doit être faite dans un dossier de choix permettant de faire le choix le plus adapté entre elles.

- **La réalisation:** son succès passe par le déroulement de deux étapes successives qui sont: la préparation et l'exécution.
  - **La préparation** comprend la planification des tâches, la définition du programme et la mobilisation des ressources.
  - **L'exécution** consiste à construire le produit fini qui répondra aux objectifs assignés dans le cahier des charges.
- **La phase de testes et évaluation** permet de s'assurer de la conformité du produit du projet par rapport au cahier des charges.

4.1 Étude du contexte

Ces dernières années ont vu une multiplication des médias de diffusion de l'information sur le trac routier. Le grand public peut être ainsi renseigné sur l'état actuel du trac grâce à un ordinateur ou un téléphone portable, mais surtout par l'intermédiaire d'appareils de navigation embarqués. L'information trac y est diffusée sous la forme d'événements, Big data et objets connectés Faire de la France un champion de la révolution numérique avril 2015artes de conditions de route ou de temps de parcours pré-calculés. Cette explosion des moyens de diffusion va de pair avec une amélioration du recueil de l'information brute. Les mesures des stations de comptage par boucles électromagnétiques sont complétées par des relevés issus de méthodes novatrices, nécessitant des infrastructures matérielles plus légères. L'analyse de traces de véhicules par relevés de géo-positionnement par satellite ou par signalisation liée à l'utilisation de réseaux de téléphonie portable est un exemple de ces nouvelles technologies. Parallèlement à cette amélioration, les bases de données géographiques décrivant le réseau routier se raffinent et permettent de constituer des historiques homogènes de mesures du trac sur plusieurs années. Cette révolution préfigure de nouveaux services intelligents de guidage et d'estimation des temps de trajets tenant compte de l'évolution du trac. L'utilisateur aura à sa disposition, en plus de l'information en temps réel, des contenus enrichis l'aidant à préparer et à planifier ses déplacements. Ces nouveaux services s'appuieront sur des méthodes de prévision et des algorithmes de routage capables d'intégrer une information évoluant au cours du temps.[6]

4.1.1 Problématique du système étudié

Entre 1958 et 1971, la part de la route en France passe de 27 % à 45,7 % du trafic exprimé en tonnes/kilomètres, tandis que celle du rail tombe de 62,1 % à 45 %.

Pratiqués en France depuis 1844, les recensements de la circulation ont pour but de connaître et de prévoir l'intensité du trafic sur les routes principales afin de mieux en assurer l'entretien et l'aménagement[6]

Lancement du projet de système de régulation du trafic routier en Algérie

Le projet du nouveau système de régulation du trafic routier dans la capitale confié à une entreprise algéro-espagnole a été lancé officiellement le 20 mars. Des feux tricolores intelligents ont été installé au niveau de la commune de Belouizded, au titre d'un projet pilote de gestion centralisée du trafic routier assurée de la joint-venture algéro-espagnole. Ce projet d'une valeur de 15 milliards de dinars algériens sera réalisé par la société algéro-espagnole "Mobilité et éclairage d'Alger" en vertu d'une convention signé en juillet 2016 entre l'Entreprise de gestion de la circulation et du transport urbain (EGCTU) et l'Etablissement de réalisation et de maintenance de l'éclairage public d'Alger (ERMA) du côté algérien et deux sociétés espagnoles spécialisées dans les systèmes de gestion du trafic routier.

Le "lancement des travaux du projet de système de régulation du trafic routier intelligent et l'installation de feux tricolores au niveau de la commune de Belouizded concerneront, dans une première étape, 200 intersections sur un ensemble de 500 au niveau de la wilaya d'Alger en vue de résoudre les problèmes liés aux embouteillages qui asphyxient la capitale. La réalisation de ce projet durera 25 mois, selon les responsables de la wilaya.

Ce système, selon ses concepteurs, **procédera en premier lieu à la collecte de données concernant la fluidité du trafic routier au niveau d'un centre spécialisé à Kouba à travers les caméras**

de surveillance et des puces magnétiques pour que ces données soient ensuite analysés en vue de trouver les solutions idoines[1].

Objectifs

Dans le cadre de ce travail, on va s'intéresser à la récolte de données statistiques concernant le flux routier (dans une route, auto-route ...etc). La pertinence de ces données statistiques peut intéresser plus d'un utilisateur, ces données peuvent servir à divers domaines d'activité à savoir :

- ***Le domaine commercial:*** En effet la décision d'installation d'un commerce donné sur une voie (dans une direction particulière) dépendra du flux de véhicule passant par cette voie.
- ***Les pouvoirs public:*** La décision de réaliser de grandes constructions tels que les routes et les passerelles dans un sens ou un autre dépend aussi des données liées au trafic routier.
- ***Amélioration du confort des usagers:*** En effet, les données collectées permettent d'identifier les bouchons vu qu'on récupère le nombre de véhicule dans une route donnée en temps réel, de ce fait les usagers pourront facilement les éviter ou les contourner.
- ***Amélioration du trafic routier: Bien évidement, les données collectées pourront aider à la gestion du trafic routier en réduisant le temps d'attente dans les intersection, et ce en réglant les durées des feux tricolores selon le flux de véhicule dans les routes.***

Notre objectif principal sera donc l'amélioration du trafic routier pour se faire nous procéderons comme suit:

1. Simulation d'un cas d'une intersection fréquenté par un nombre variant de véhicules.
2. Récolter des données statistiques.
3. Exploiter les données récoltées afin d'optimiser le trafic routier.

4.1.2 Concepts de base liés au contexte à modéliser

notre centre d'intérêt est de présenter le cas particulier de la gestion des feux de circulation en milieu urbain, ceci en s'appuyant sur la littérature utilisant les réseaux fixes de capteurs sans fil et en présentant des outils permettant de simuler le contrôle du trafic routier urbain nous introduisons quelques définitions autour des types de réseaux existant ainsi que leurs métriques et principaux paramètres

Définition 1. Les systèmes de transport intelligents (STI) apparaissent comme étant "l'application des technologies de l'information et de la communication au domaine des transports". Le terme système est vague et se décline en un ensemble de moyens mis en place pour gérer au mieux les contraintes liées au trafic routier, telles que les embouteillages, la sécurité ou même la pollution.[6]

Définition 2. Les réseaux de capteurs sans fil appartiennent à la famille des réseaux mobiles ad hoc (MANET) et se composent d'un large ensemble de capteurs à capacité et énergie généralement limitées. Dans de nombreux cas, les capteurs sont constitués des unités suivantes :

- Unité d'acquisition, permet le recueil de données environnementales et la conversion analogique vers numérique.
- Unité de calcul, permettant le lancement de procédures, protocoles et autres.
- Unité de communication, permettant la connexion au réseau (émission et réception).

Ceci, en sans-fil (ex : radio), souvent en multi-sauts et permettant de s'affranchir des inconvénients filaires (temps d'installation et facilité d'accès)[6]

Le rôle des capteurs: c'est de récolter un ensemble de données dans son environnement, et le transmettre de proche en proche jusqu'à atteindre généralement une station de base, qui peut jouer le rôle de coordinatrice du réseau et communiquer vers l'extérieur les données importantes recueillies.

[1] **Les solutions idoines:** Qui convient exactement à la situation

Un capteur électromagnétique est un type de capteur sans fil fixe utilisé notamment pour la gestion du trafic routier. L'unité d'acquisition utilisée est un magnétomètre, permettant d'enregistrer les variations du champ magnétique terrestre. Dans le cadre des STI, le rôle de tels dispositifs va être de relever des informations sur le trafic routier

Définition 3. Un réseau fixe regroupe un ensemble de technologies (dont détecteurs) étant fixes sur le terrain : leur position n'est pas amenée à être changée, mis à part pendant des périodes d'entraînement où l'objectif va être de déterminer leur emplacement définitif.[6]

Définition 4. Un système coopératif est un réseau hybride où des acteurs mobiles vont pouvoir interagir avec des acteurs fixes[6]

4.1.3 Les Systèmes de transport intelligents

Le trafic routier urbain s'est amplifié en l'espace de quelques années, augmentant ainsi les problèmes engendrés qui sont nombreux et qui coûtent quotidiennement temps, argent, santé et qualité environnementale : embouteillages, accidents, pollution ou encore infractions

La gestion du trafic routier s'inscrit dans le domaine des STI, qui visent à proposer des outils et modèles afin de gérer les aléas de ce dernier, ceci par le biais ou non d'équipements réactifs dits dynamiques. L'application de tels systèmes va avoir de multiples objectifs, parmi lesquels la fluidification du trafic, la détection d'incidents, la surveillance temps-réel du trafic, la diffusion d'informations ou de consignes variables aux automobilistes ou encore la réduction en conséquence de la pollution et des bruits

4.1.3.1 Le rôle du systèmes intelligent de gestion du trafic :

ces derniers sont majoritairement conçus pour fluidifier et gérer le trafic routier, notamment au niveau des intersections où ces derniers peuvent directement agir sur les feux de circulation, également au niveau de la politique de stationnement, de l'information de l'utilisateur à tout niveau, et de l'utilisation de stratégies particulières afin de gérer les situations de danger

le modèle de carrefour qui est typiquement utilisé dans la littérature afin de valider un modèle : une intersection composée de quatre directions avec un nombre fixé de voies pour chacune. Ici, les voies pour tourner à gauche sont séparées des voies allant tout droit ou à droite, ces deux derniers mouvements étant confondus. Ce modèle possède l'avantage de pouvoir être adaptable à de nombreuses situations, mais est instinctivement limité de part la distinction des mouvements et voies.

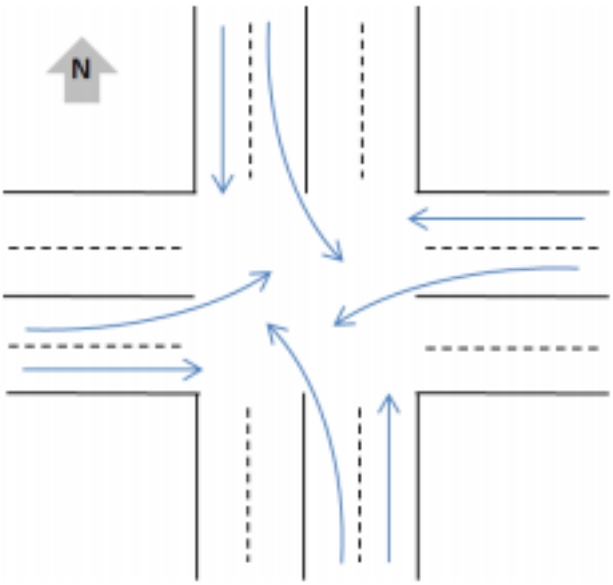


Figure 4.1.1: Modèle de carrefour

Remarque: Les feux quant à eux améliorent grandement la fluidité du trafic, mais uniquement en cas de débit suffisamment élevé.

Afin d'être efficaces, les auteurs proposent de baser le timing des feux en fonction d'une base de données historique, afin d'identifier les heures de pointes, et désactiver toute signalisation le reste du temps. Enfin, citons, où les auteurs proposent d'analyser trois approches afin de fluidifier le trafic dans les ronds-points : avec des signaux de ralentissement à l'arrivée, avec des feux de circulation à l'arrivée, et avec des feux de circulation à la fois à l'arrivée mais également à l'intérieur du rond-point, lorsqu'un usager prend la voie de gauche

on rappelle que dans ce travaille on s'intéresse à une approche qui est le calcule du temps nécessaire a une voiture pour quitter l'intersection.[6]

4.1.3.2 La gestion des feux de circulation

Il sont généralement gérés par une boite de contrôle, qui va posséder plus ou moins de propriétés en fonction des constructeurs. Typiquement, une boite est rattachée à une seule intersection et possèdent les éléments principaux suivants :

- Une unité d'énergie.
- Une unité de détection, connectée à des éléments de contrôle (détecteurs).
- Une unité de contrôle, donnant l'ordre d'enclenchement des feux.
- Une unité d'avertissement rapide, réagissant en cas d'erreur critique (par exemple : orange clignotant sur l'ensemble des feux.).
- Une unité de gestion des conflits, qui est programmée avec les combinaisons de feux verts autorisés (matrice de conflits) et qui vérifie les données envoyées par l'unité de contrôle : elle fait appel à l'unité précédente en cas d'erreur ou de faute constatée sur l'un des feux.
- Une unité d'administration, pour prendre le contrôle du carrefour (par la police par exemple)

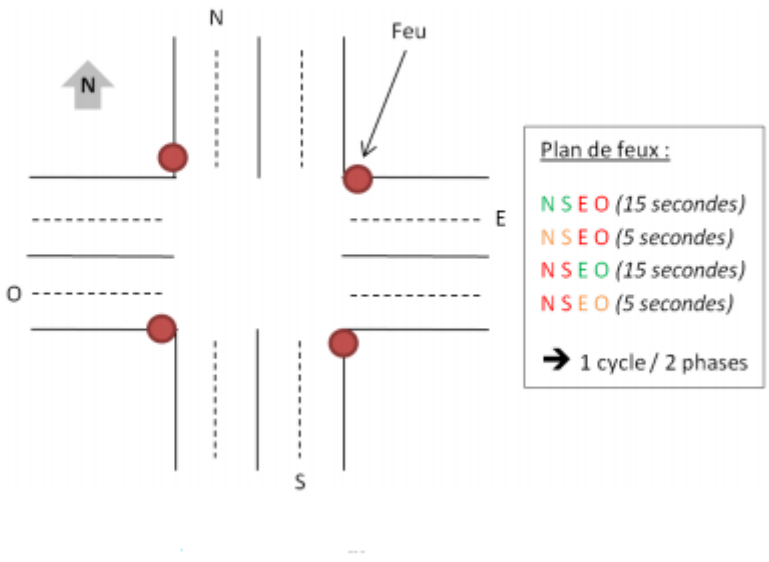


Figure 4.1.2: Plan du feu

4.2 Cycle de vie de la donnée:

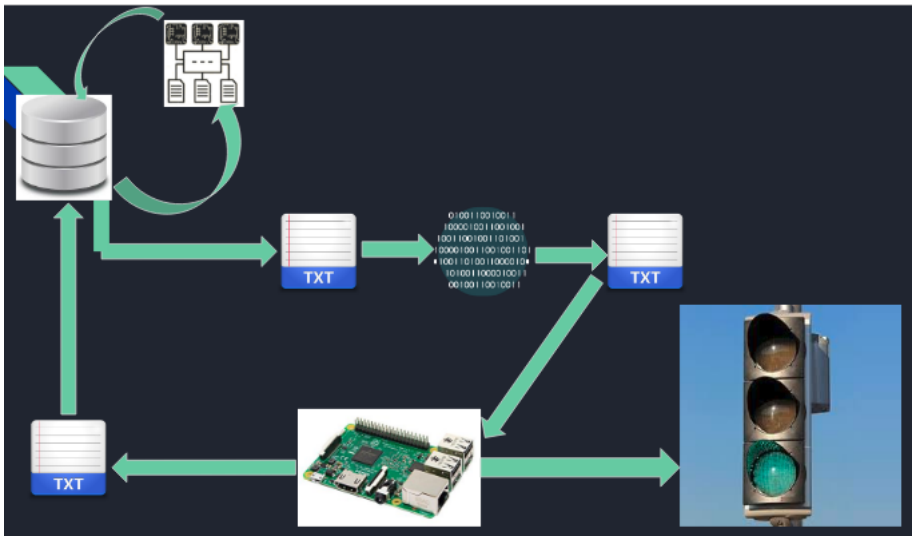


Figure 4.2.1: Cycle de vie de la donnée

Notre travail consiste à manipuler les données récoltées grâce au capteur vidéo relié à la rasberry pi, cette dernière s'est occupée de récupérer des données tels que le nombre de voitures et les temps nécessaires aux voiture pour sortir de l'intersection à l'aide d'un algorithme de traitement d'image qui utilise la bibliothèque d'apprentissage déjà configurée dans tensorflow (détection de voitures), ensuite ces données sont sauvegardées dans un fichier .Txt et enregistrées dans notre BDD NoSQL, l'algorithme de MapReduce récupère ces dernière et les traite pour en extraire uniquement les données souhaitées, dans notre cas c'est les temps et les enregistre dans la BDD, maintenant c'est au tour de l'algorithme d'optimisation d'intervenir en récupérant les données traités, le résultat du déroulement de notre algorithme génère les valeurs optimales des durées des feux qui vont être envoyées vers la rasberry et qui s'occupera de les envoyer à son tour vers les feux.

Chapter 5

Outils de développements et simulation(Préparation)

5.1 Les Différentes Technologies Utilisées

Pour la réalisation de notre travail on a eu recours à plusieurs technologies comme hadoop, rasberry pi, anylogic.

5.1.1 hadoop

Hadoop est un framwork libre et open source écrit en Java destiné à la création d'applications distribuées et scalables. Hadoop a été inspiré par les publications de MapReduce et GoogleFS de Google et fait partie des projets de la fondation Apache depuis 2009. Hadoop est le framework Big Data le plus largement utilisé dans le monde. Il regroupe un ensemble d'outils répondant à un très grand nombre de cas d'usages tels que le stockage et la gestion des données, la sécurité, la sérialisation, l'analyse, etc. Hadoop est utilisé chez un très grand nombre d'entreprises à l'instar des géants de l'IT Facebook, Yahoo, LinKedIn, Amazon, eBay, etc.[1]

5.1.1.1 Hadoop - Big Data Solutions

Hadoop exécute des applications à l'aide de l'algorithme MapReduce, dans lequel les données sont traitées en parallèle sur différents nœuds de processeur. En bref, le framework Hadoop est capable de développer des applications capables de fonctionner sur des clusters d'ordinateurs et d'effectuer une analyse statistique complète pour une grande quantité de données.

5.1.1.2 Architecture Hadoop

Le framework Hadoop comprend quatre modules:

[1] http://hadoop.apache.org/releases.html

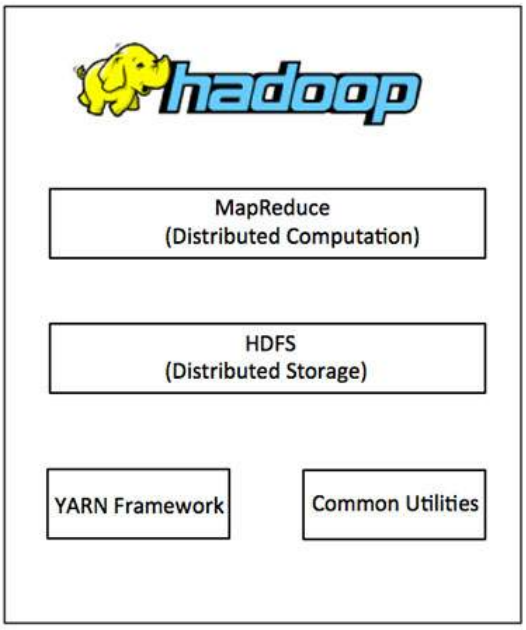


Figure 5.1.1: Architecture Hadoop

Hadoop Common: Il s'agit des biblIoThèques Java et des utilitaires requis par d'autres modules Hadoop. Ces biblIoThèques fournissent des abstractions au niveau du système de fichiers et du système d'exploitation et contiennent les fichiers Java et les scripts nécessaires au démarrage de Hadoop.

Hadoop YARN: Ceci est un cadre pour la planification des tâches et la gestion des ressources de cluster.

Système de fichiers distribués Hadoop (HDFS): système de fichiers distribué offrant un accès à haut débit aux données d'application.

Hadoop MapReduce: Ceci est un système basé sur YARN pour le traitement parallèle de grands ensembles de données.

Nous pouvons utiliser le diagramme suivant pour décrire ces quatre composants disponibles dans le framework Hadoop.

MapReduce Hadoop MapReduce est un framework logiciel permettant d'écrire facilement des applications qui traitent de grandes quantités de données en parallèle sur de grands clusters (des milliers de nœuds) de matériel de base de manière fiable et tolérant aux pannes.

Le terme MapReduce fait référence aux deux différentes tâches suivantes exécutées par les programmes Hadoop:

- **La tâche de mappe:** C'est la première tâche, qui prend les données d'entrée et les convertit en un ensemble de données, où les éléments individuels sont décomposés en tuples (paires clé / valeur).
- **La tâche Réduire:** cette tâche prend la sortie d'une tâche de carte en entrée et combine ces tuples de données dans un ensemble plus petit de tuples. La tâche de réduction est toujours effectuée après la tâche de carte.

Généralement, l'entrée et la sortie sont stockées dans un système de fichiers. Le framework prend en charge la planification des tâches, les surveille et ré-exécute les tâches échouées.

Système de fichiers distribué Hadoop Hadoop peut fonctionner directement avec n'importe quel système de fichiers distribué, tel que FS local, FS HFTP, S3 FS, etc., mais le système de fichiers le plus utilisé par Hadoop est Hadoop Distributed File System (HDFS).

Le système de fichiers distribués Hadoop (HDFS) est basé sur le système de fichiers Google (GFS) et fournit un système de fichiers distribué conçu pour fonctionner sur de grandes grappes (milliers d'ordinateurs) de petites machines informatiques de manière fiable et tolérante aux pannes.

5.2 Installation d'hadoop:

Pour des questions de simplicité et pour ne pas perturber notre travail courant, on choisis de faire tourner Hadoop sur une ubunto ,se système d'exploitation, j'ai retenu Centos (version 6.5 64 bits) qui correspond à la version libre de Red Hat. on à choisi la version « minimale » car on n'a pas besoin d'interface graphique : l'administration et la gestion de Hadoop se fera en mode « remote » à partir de notre machine hôte .

pour cette installation on suit ses étape:

- installation java:

  1. Penser à faire une mise à jour du cache des paquets du système.

```
$ sudo apt-get update
```

Figure 5.2.1: Mise a jour des paquets

  2. Après cela, installez Java sur Ubuntu

```
$ sudo apt-get install java
```

Figure 5.2.2: Installation java

  3. Maintenant, vérifiez la version java.

```
$ java -version
```

Figure 5.2.3: version du java

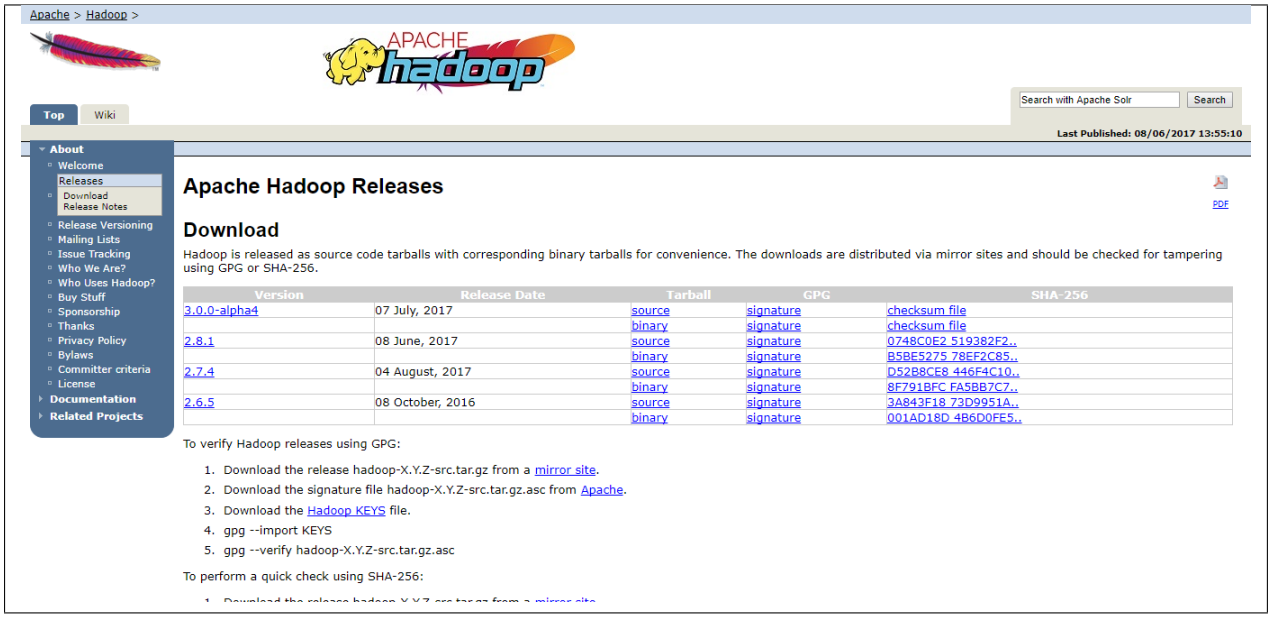
  4. On obtiendra la sortie suivante.

```
openjdk version "1.8.0_91"<font></font>
OpenJDK Runtime Environment (build 1.8.0_91-8u91-b14-<font></font>
3ubuntu1~16.04.1-b14)<font></font>
OpenJDK 64-Bit Server VM (build 25.91-b14, mixed mode)<font></font>
```

Figure 5.2.4: output java version

- installation hadoop:

  Nous avons installé Java et ensuite nous devons installer le Hadoop. Accédez à la page de version d'Apache Hadoop pour trouver la dernière version d'Apache `http://hadoop.apache.org/releases.html`



Télachargement la version

Téléchargement 2.7.3.tar.gz

Table 5.1: Téléchargement

On doit trouver la dernière version stable pour installer Hadoop 2.7 sur Ubuntu. Une fois que avoir trouvé la dernière version stable, copiez le lien en cliquant avec le bouton droit de la souris.

1. Utilisez la commande ci-dessous pour télécharger le fichier.

```
$ wget http://apache.mirrors.tds.net/hadoop/common/hadoop-2.7.3/hadoop
```

Figure 5.2.5: téléchargement hadoop-2.7.3

2. Après cela, faites la vérification en utilisant la commande ci-dessous.

```
$ shasum -a 256 hadoop-2.7.3.tar.gz
```

Figure 5.2.6: vérification de la version

3. On obtiendra la sortie suivante. Maintenant, vérifiez la valeur SHA-256 .

Figure 5.2.7: vérifiez la valeur SHA-256

4. Les deux sorties doivent correspondre.

```
~/hadoop-2.7.3.tar.gz.mds<font></font>
...<font></font>
hadoop-2.7.3.tar.gz: SHA256 = D489DF38 08244B90 6EB38F4D 081BA49E 50C
...<font></font>
```

Figure 5.2.8: Correspondance

5. On peut simplement ignorer les espaces. De cette façon, on peut vérifier si le fichier est corrompu lors du téléchargement.

   Aprés avoir vérifié que le fichier est original, nous devons utiliser la commande tar pour extraire le fichier.

```
$ tar -xzvf hadoop-2.7.3.tar.gz
```

Figure 5.2.9: extraire le fichier

6. Ici:

   -x est pour extraire le drapeau

   -z est pour décompresser le fichier.

   -v pour une sortie détaillée

   -f spécifie l'extraction du fichier.

   Maintenant, nous allons déplacer le fichier extrait vers l'emplacement / usr / local.

```
$ sudo mv hadoop-2.7.3 /usr/local/hadoop
```

Figure 5.2.10: déplacer le fichier extrait vers l'emplacement / usr / local

Maintenant, la prochaine étape consiste à démarrer l'environnement.

- configuration de hadoop pour l'utilisation de java

1. Nous devons configurer Hadoop pour qu'il utilise le fichier Java dans le fichier de configuration de Hadoop ou en utilisant la variable d'environnement.

   Ici / usr / bin / java et / etc / alternatives / java sont tous deux des liens symboliques entre eux.

   Ici, nous devons utiliser l'option -f pour suivre le lien symbolique dans chaque partie du chemin mentionné.

   Le sed sera utilisé ici pour couper le chemin pour obtenir le bin / java. Nous devons le faire pour obtenir la valeur correcte de java Home à partir de la sortie.

   Si on veut obtenir le chemin java par défaut.

```
$ readlink -f /usr/bin/java | sed "s:bin/java::"<font></font>
<font></font>
Output<font></font>
<font></font>
/usr/lib/jvm/java-8-openjdk-amd64/jre/<font></font>
```

Figure 5.2.11: obtenir le chemin java par défaut

2. Nous allons définir cette version de Java sur le chemin d'accueil de Hadoop.

   Il existe une autre façon d'utiliser la commande readlink pour définir dynamiquement le chemin si on utilise une version mise à jour.

   D'abord, ouvrez le hadoop-env.sh:

```
$ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

Figure 5.2.12: ouvrez le hadoop-env.sh

Il y a deux options disponibles. Voici eux.

(a) Configuration d'une valeur statique

```
/usr/local/hadoop/etc/hadoop/hadoop-env.sh<font></font>
 . . .<font></font>
#export JAVA_HOME=${JAVA_HOME}<font></font>
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre/<font></font>
. . . <font></font>
```

Figure 5.2.13: Configuration d'une valeur statique

(b) Utiliser le lien de lecture directement

```
/usr/local/hadoop/etc/hadoop/hadoop-env.sh<font></font>
. . .<font></font>
#export JAVA_HOME=${JAVA_HOME}<font></font>
export JAVA_HOME=$(readlink -f /usr/bin/java | sed "s:bin/java::")<fo
. . . <font></font>
```

Figure 5.2.14: Utiliser le lien de lecture directement

5.3 Java :

5.3.1 définition

Java est un langage de programmation moderne développé par Sun Microsystems (aujourd'hui racheté par Oracle), C'est un langage très utilisé, notamment par un grand nombre de programmeurs profession-

nels, ce qui en fait un langage incontournable actuellement et les programmes crée en java fonctionnent sous Windows, Mac, Linux, etc.... ce qui fait son excellente portabilité, rapidité et sécurité.

5.3.2 Son mode de fonctionnement

Syntaxe:

Syntaxe semblable à celle du C. Attention, comme dans C, les ' ;' sont des terminateurs d'instructions et non des séparateurs d'instructions comme en Pascal (en particulier il y a toujours un ' ;' à la fin d'une affectation qu'il y ait un else ou non derrière).

Concepts de base des langages objets:

Ils sont au nombre de trois:

- le concept de structuration qui lie entre eux les notions d'objet, de classe et d'instance.
- le concept de communication au travers de l'envoi de message et des méthodes.
- le concept d'héritage.

5.3.3 Entrées et sorties (Java Perspective)

L'infrastructure MapReduce fonctionne sur les paires <key, value>, c'est-à-dire que l'infrastructure considère l'entrée du travail comme un ensemble de paires <key, value> et produit un ensemble de paires <key, value> comme sortie du travail , peut-être de différents types.

La clé et les classes de valeur doivent être sérialisées par le framework et par conséquent, doivent implémenter l'interface Writable. De plus, les classes clés doivent implémenter l'interface Writable-Comparable pour faciliter le tri par le framework. Types d'entrée et de sortie d'un travail MapReduce: (Entrée) <k1, v1> -> carte -> <k2, v2> -> réduire -> <k3, v3> (Sortie).

	Entrée	Sortie
Map	<k1, v1>	liste (<k2, v2>)
Reduce	<k2, liste (v2)>	iste (<k3, v3>)

Table 5.2: L'infrastructure de MapReduce

L'interface Map<K, V>:

Ce type d'objets gère ses données avec un système de clé-valeur. Elle diverge donc franchement des autres collections vues précédemment, et c'est pourquoi ses interfaces et ses objets n'héritent pas de l'objet Collection<E>.

Reduce <k2, list(v2)>

Cette étape est la combinaison de l' étape de lecture aléatoire et de l' étape de réduction . Le travail du réducteur consiste à traiter les données provenant du mappeur. Après le traitement, il produit un nouveau jeu de sortie, qui sera stocké dans le HDFS.

5.4 Matériel physique

Dans ce chapitre on va présenter le système conçu, passant par l'explication de son fonctionnement de manière détaillée, la représentation synoptique du système, et la présentation des principaux éléments le constituant (carte Arduino Uno, carte Raspberry Pi) ainsi que le module Bluetooth qui assure la communication sans fils entre ces deux cartes. Une caméra embarquée reliée à la Raspberry Pi sera décrite.[9]

5.4.1 Conception générale du système

Le système que nous avons conçu et réalisé est constitué de deux modules principaux et essentiels, la carte Raspberry PI avec caméra embarquée, dont la fonction principale est d'effectuer un traitement d'image.[9]

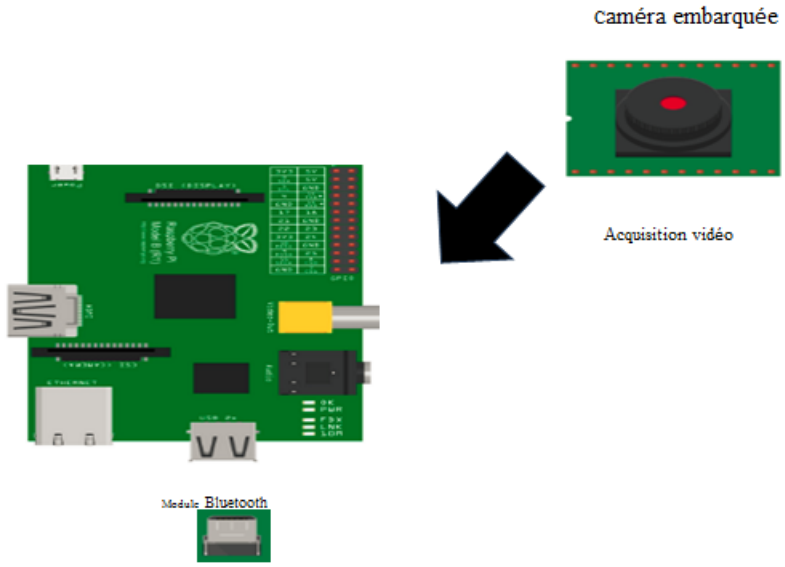


Figure 5.4.1: synoptique représentatif de la conception du feu intelligent

5.4.2 Description détaillée du système

Le système est composé essentiellement comme le montre la figure5.4.1

Partie de contrôle intelligente :

qui est structurée autour d'une carte Raspberry Pi qui s'occupe de l'exécution des taches du traitement des images requises par une caméra embarquée qui se charge de la visualisation du flux de circulation. Cette partie s'occupe aussi de la gestion des priorités. Carte de commande : c'est la carte Arduino Uno, qui se charge de la réception d'instructions, et de la gestion de l'allumage des feux. Module Bluetooth : qui se charge de la liaison sans fils entre ces deux parties.

Partie de contrôle intelligente :

La carte Raspberry Pi : Le Raspberry Pi est un nano-ordinateur mono-carte à processeur ARM conçu par le créateur de jeux vidéo David Braben, dans le cadre de sa fondation Raspberry Pi. Cet ordinateur, qui a la taille d'une carte de crédit, est destiné à encourager l'apprentissage de la programmation informatique ; il permet l'exécution de plusieurs variantes du système d'exploitation libre GNU/Linux et des logiciels compatibles. Il est fourni nu (carte mère seule, sans boîtier, alimentation, clavier, souris ni écran) dans le but de diminuer les coûts et de permettre l'utilisation de matériel de récupération.[**?**]

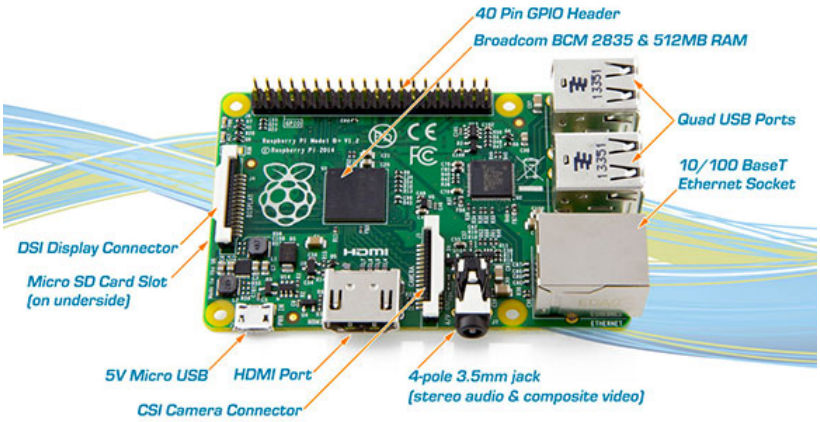


Figure 5.4.2: Raspberry PI 2

La Raspberry Pi au sein du système : C'est la partie qui va s'occuper principalement de l'acquisition des données du trafic routier à l'aide de la caméra au centre d'une intersection, faisant un balayage sur les 4 voies, avec un moteur pas à pas piloté par la carte Raspberry via le biais des sorties GPIO, puis de les traitées grâce au microprogramme que nous avons développé. Ce dernier fera le comptage des véhicules passants dans chaque voie, et la comparaison, ainsi que l'identification des véhicules prioritaires, prenant compte de tout cela, le passage sera accordé à la voie priorisée, les instructions sont envoyées vers l'ARDUINO via Bluetooth, fonctionnelles sur la Raspberry PI grâce a un module USB.

Le modèle qu'on a utilisé est équipé d'un processeur broadcom BCM2836, quatre cœurs ARMv7 à 900 MHz, accompagné de 1 Go de RAM, ainsi qu'une caméra omnivision 5647 capteurs dans un module de mise au point fixe. Le module s'attache à Raspberry Pi, au moyen d'un câble ruban Pin 15, à l'interface série dédiée MIPI caméra à 15 broches (CSI). Le bus CSI est capable de débits de données très élevés, et il comporte exclusivement des données de pixels au processeur BCM2836 Le capteur lui-même à une résolution native de 5 mégapixels, et dispose d'un objectif à focale fixe à bord.

La liaison avec la camera : La caméra prend en charge 1080 p @ 30 fps, 720 p à 60 images par seconde et 640 x 480 p enregistrement vidéo 60/90 aussi est pris en charge par la dernière version de Raspbian, système d'exploitation recommandé pour la Raspberry Pi.[**?**]

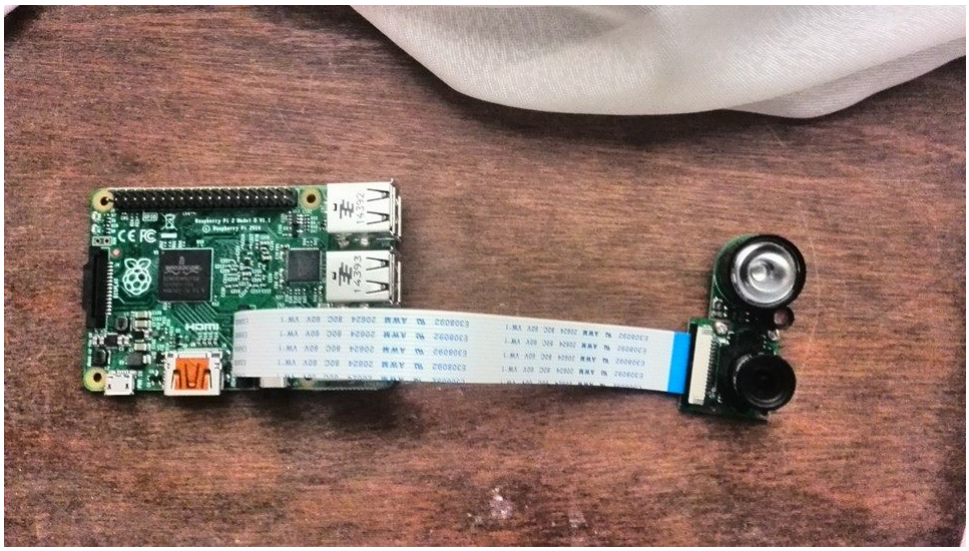


Figure 5.4.3: Raspberry PI 2 avec module caméra embarquée

Connectiques et ports disponibles sur la Raspberry Pi 2 :

- Double abaisseur de tension d'alimentation (buck) pour le 3,3 V et le 1,8 V

- Le 5V dispose d'une protection contre l'inversion de polarité, d'un fusible de 2A et d'une protection pour les branchements à chaud
- Nouvelle puce contrôleur USB / Ethernet
- 4 ports USB au lieu de 2 ports 40 broches GPIO dont
- 2 broches d'identification d'EEPROM La sortie vidéo composite (NTSC/PAL) intégrée au jack audio de 3,5mm
- Support de carte MicroSD.
- Quatre trous de montage en disposition rectangulaire
- Connecteur d'alimentation micro USB.
- Le port HDMI pour l'affichage sur un moniteur de haute définition.
- Connecteur CSI pour module caméra et d'affichage DSI.[9]

5.5 simulation avec anylogic

5.5.1 Environnement de modélisation multi-méthode

Développez des modèles à l'aide de trois méthodes de simulation modernes :

La modélisation multi-agents

se focalise sur les éléments individuels actifs d'un système. Cela contraste avec l'approche de dynamique système, plus abstraite, ou avec la méthode par événement discret, qui est axée sur les processus.

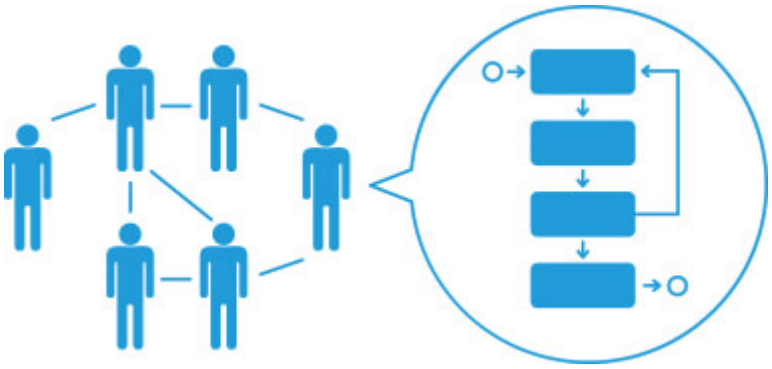


Figure 5.5.1: simulation multi-agents

Dans la simulation multi-agents, les entités actives, que l'on appelle agents, doivent être identifiées et leur comportement doit être défini. Il peut s'agir de personnes, de foyers, de véhicules ou d'équipements, même de produits ou d'entreprise, de tout élément pertinent pour le système. Les connexions entre ces éléments sont ensuite créées, les variables environnent-ales définies et les simulations exécutées. La dynamique globale du système émerge alors des interactions entre les nombreux comportements individuels.

AnyLogic associe la modélisation professionnelle multi-agents, par événement discret et par dynamique système au sein d'une plateforme unique, pour des résultats efficaces et sans compromis.

Dynamique de système

La dynamique de système est une méthode de modélisation à haut niveau d'abstraction. Elle ignore les détails les plus spécifiques d'un système, comme par exemple les propriétés individuelles des personnes, des produits et des événements et fournit une représentation générale d'un système complexe. Ces modèles abstraits peuvent être utilisés dans le cadre d'une modélisation et d'une simulation stratégique

à long terme. Par exemple, un opérateur téléphonique qui prévoit de réaliser une campagne marketing peut simuler et analyser le succès de nouveaux projets de forfaits, sans devoir modéliser les interactions de chaque client.

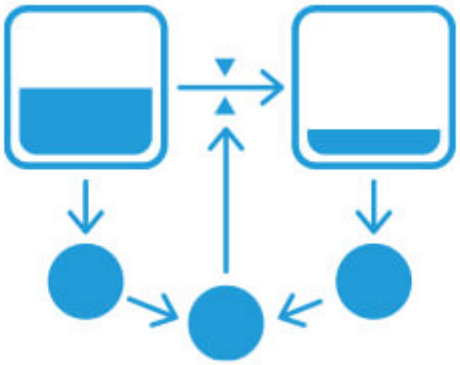


Figure 5.5.2: dynamique système

Modélisation de simulation par événement discret

La plupart des processus d'entreprise peuvent être décrits comme des événements discrets et séparés. Par exemple, dans notre cas un véhicule arrive au Rond point et les feus se passe du vert au rouge alors se dernier vas marquer sont arrêt et il ne départ pas jusque a se que les passes a la couleur vert

Dans une modélisation par événement discret, le mouvement d'un train d'un point A à un point B est modélisé avec deux événements, c'est à dire un départ et une arrivée. Le mouvement du train serait modélisé sous la forme du délai entre les événements d'arrivée et de départ. Ces événements et le mouvement entre eux peuvent être parfaitement animés.

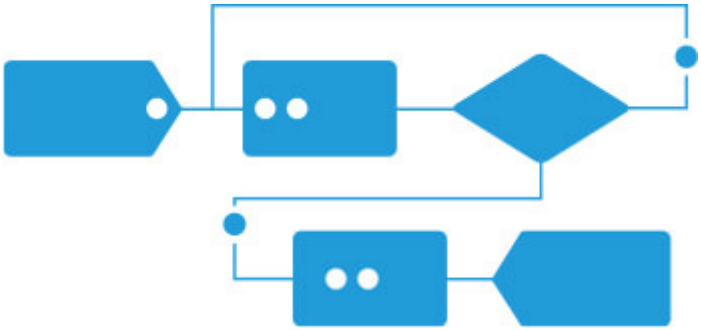


Figure 5.5.3: simulation par évènement discret

La modélisation par événement discret se focalise sur les processus dans un système à un niveau intermédiaire d'abstraction. Habituellement, les informations physiques spécifiques, comme la géométrie d'une voiture ou l'accélération d'un train, ne sont pas représentées. La modélisation par événement discret est massivement utilisée dans les secteurs de la santé, de la fabrication et de la logistique.

5.5.2 Initiation sur AnyLogic:

La méthode de la dynamique du système suppose un niveau d'abstraction élevé et est principalement utilisée pour les problèmes de niveau stratégique. La modélisation des événements discrets (centrés sur les processus) est principalement utilisée aux niveaux opérationnel et tactique. Les modèles basés sur des agents sont utilisés à tous les niveaux: les agents peuvent être des entreprises concurrentes, des consommateurs, des projets, des idées, des véhicules, des piétons ou des objets et tous se que on peut imaginé.

Tous les secteurs d'activité utilisent le logiciel de simulation AnyLogic pour une meilleure prise de décision tout au long du cycle de vie de l'entreprise. L'utilisation d'un outil pour tous les défis métier permet aux organisations d'économiser du temps et de l'argent, d'augmenter le partage des connaissances et de relier les modèles entre plusieurs services

En utilisant des technologies de simulation, on à facilement examiner et vérifier n'importe quel scénario, concevoir la meilleure configuration pour notre projet de construction, réorganiser nos processus de production ou construire un entrepôt . Dans tous les secteurs, de la chaîne d'approvisionnement et de la logistique à la recherche de marché et aux soins de santé, AnyLogic peut être appliqué à plusieurs niveaux des entreprises, y compris opérationnels, tactiques et stratégiques, devenant rapidement une partie intégrante de notre organisation.

AnyLogic fournit une large gamme d'outils de visualisation et d'animation pour nous aider à créer des modèles qui peuvent avoir un aspect personnalisé avec un impact visuel amélioré. La flexibilité unique du langage de modélisation permet à l'utilisateur de saisir la complexité et l'hétérogénéité des systèmes commerciaux, économiques et sociaux à n'importe quel niveau de détail souhaité.

5.5.3 Les constructions de langage de simulation fournies par AnyLogic

Le langage de simulation AnyLogic est composé des éléments suivants:[2]

- Les Diagrammes des Stocks et des Flux sont utilisés pour la modélisation de Dynamique de Système.
- Statecharts (diagrammes d'état) sont utilisés surtout dans les Systèmes Multi-Agents pour définir le comportement d'agents. Ils sont aussi souvent utilisés dans la modélisation par Événements Discrets : par exemple, simuler la panne de machine.
- Les diagrammes d'Action sont utilisés pour définir des algorithmes. Ils peuvent être utilisés dans la modélisation par Événements Discrets (par exemple, pour l'acheminement d'appels) ou dans les Systèmes Multi-Agents (par exemple pour la logique de décision d'agent).
- Les Diagrammes de Flux sont la base de la construction des processus dans la modélisation par Événements Discrets. Quand on regarde ces diagrammes, on comprend pourquoi l'approche par Événements Discrets est souvent appelée Approche Centrée Processus.

Le langage inclut aussi le niveau bas de constructions de la modélisation (variables, équations, paramètres, événements etc.), les formes de présentation (lignes, polylignes, ovales etc.), moyens d'analyse (ensembles de données, histogrammes, graphiques), outils de connectivité, images standard et les outils d'expérimentations.

5.5.4 Installation

Le logiciel de simulation est disponible à l'adresse suivante : http://www.anylogic.com/Le logiciel se trouve dans Download et il faut choisir Free PLE (Free Personal Learning Edition).Il faut choisir la version correspondante au système utilisé (ici Windows 64 bits) .

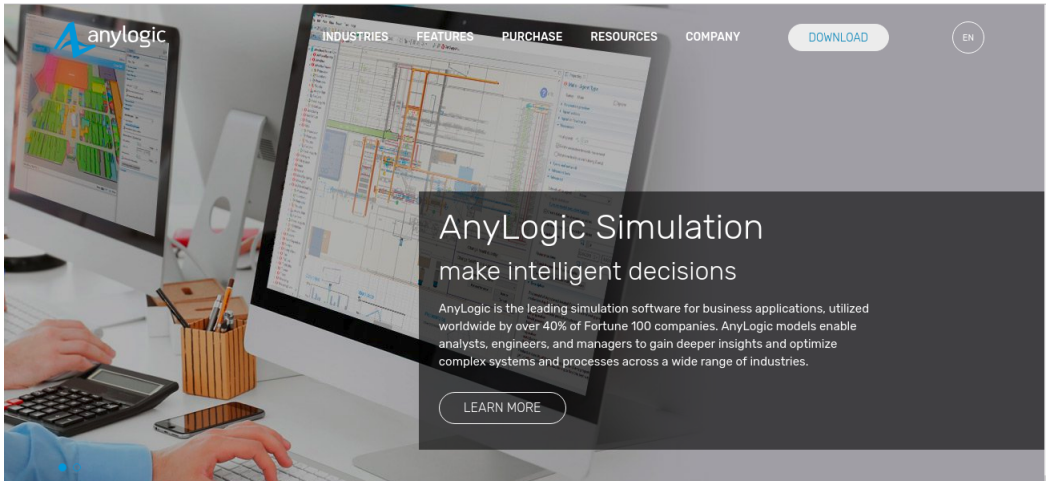


Figure 5.5.4: Téléchargement

[2] a et b AnyLogic on-line help on official vendor web-site

Il faut choisir la version correspondante au système utilisé (ici Windows 64 bits)[3]

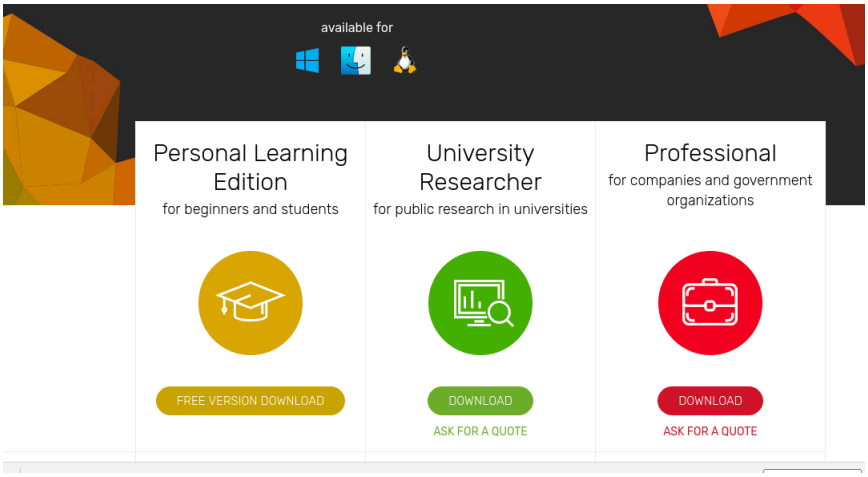


Figure 5.5.5: chois de version

Il ne reste plus qu'à remplir le formulaire

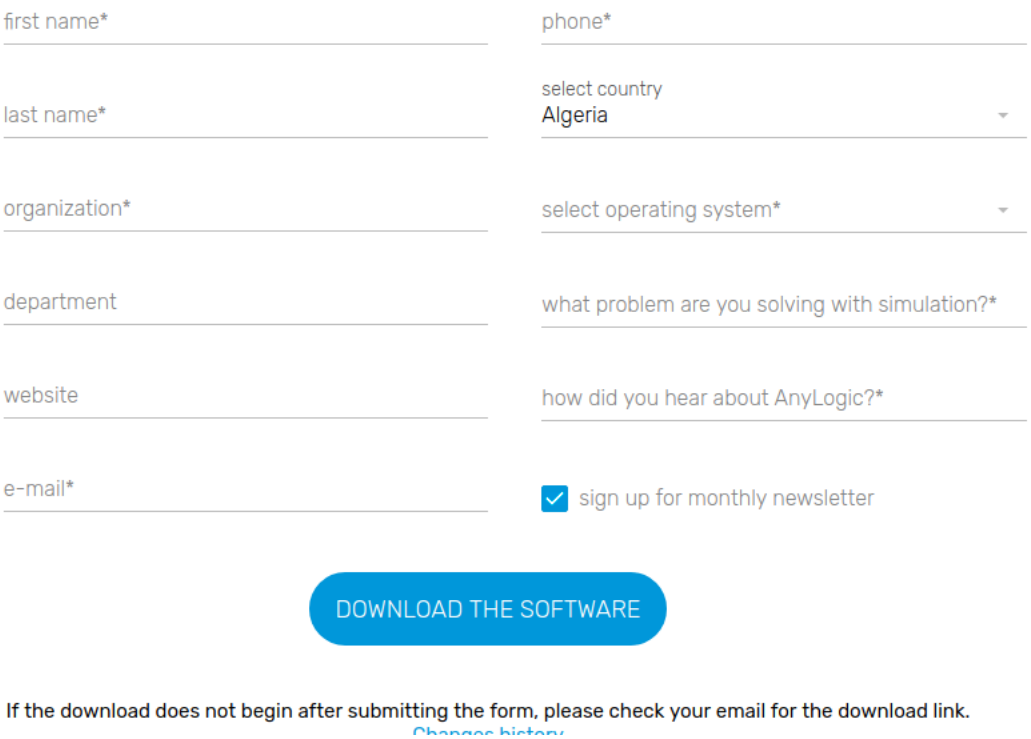


Figure 5.5.6: Remplissage de formulaire

Remarque:*Contrairement à beaucoup d'autres environnements, Anylogic permet de réaliser plusieurs types de modèle de simulation. Il s'agit là du point fort de l'environnement.*

5.5.5 Implémentation:

Modèle de simulation à événements discrets:

Dans notre projet on prend l'exemple d'un rend-point pour la simulation et à cause le manque de moyen on a pas pue implémenté notre travaille sur la réalités alors on à utilisé cette simulation pour illustré un peut notre aidé et pour y ' arrivée on vas se précédé comme suit :

[3] `https://www.anylogic.com`

first name*

last name*

organization*

department

website

e-mail*

phone*

select country
Algeria

select operating system*

what problem are you solving with simulation?*

how did you hear about AnyLogic?*

sign up for monthly newsletter

DOWNLOAD THE SOFTWARE

If the download does not begin after submitting the form, please check your email for the download link.

Figure 5.5.7: création du model

Objets à crée:

Le système à simuler se compose des feus tricolores ,des routes ,parking véhicule, des voitures ,...etc. on illustre ses objets cité en haut par cette figure:

(a) feus tricolore

Figure 5.5.8: objet à simulé

Produit final:

on final on implémente tous ses objets avec notre simulateur on vas aboutir a se résultat:

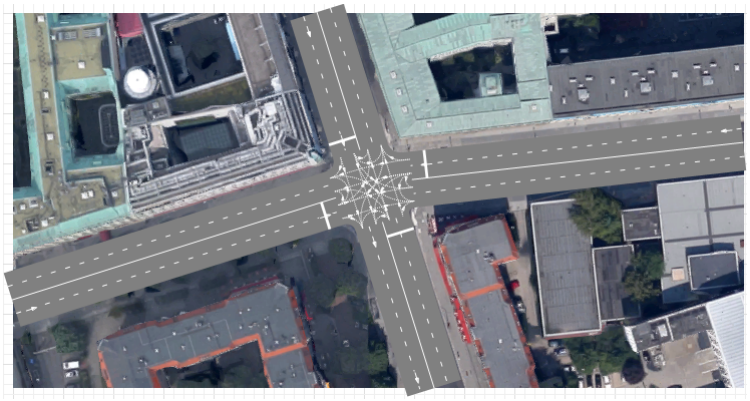


Figure 5.5.9: rend point

Intersection est un élément de balisage d'espace graphique qui est utilisé pour connecter deux routes ou plus. À l'aide de routes, d'intersections et d'autres éléments de balisage d'espace, nous traçons des réseaux routiers pour les modèles de biblIoThèque de circulation routière.

L'intersection contrôle les directions de la circulation à l'aide de connecteurs de voies qui indiquent les itinéraires pour les voitures sur chaque voie lorsque les véhicules traversent l'intersection.

5.5.6 Avantages du simulateur Anylogic:

Le logiciel de simulation AnyLogic offre de nombreux avantages, notamment:

- L'environnement Java natif prend en charge l'extensibilité illimitée, y compris le code Java personnalisé, les biblIoThèques externes et les sources de données externes.
- Un vaste ensemble de fonctions de distribution statistique fournit une excellente plate-forme pour simuler l'incertitude inhérente à tous les systèmes.
- Un cadre expérimental puissant, un support intégré pour les simulations de Monte Carlo et des formes avancées d'optimisation supportent une grande variété d'approches de simulation.
- Le paradigme de conception de modèle orienté objet pris en charge par AnyLogic fournit une construction modulaire, hiérarchique et incrémentielle de grands modèles.
- L'environnement de développement visuel accélère considérablement le processus de développement.
- Les biblIoThèques d'objets permettent d'intégrer rapidement des éléments de simulation prédéfinis.
- Réutilisation grâce à une structure entièrement orientée objet.
- Élaborer des modèles de systèmes basés sur des agents, de la dynamique du système, des événements discrets, continus et dynamiques, dans n'importe quelle combinaison, avec un seul outil.
- Prend en charge l'intégration transparente de simulations discrètes et continues.

5.5.7 AnyLogic et le langage Java:

AnyLogic comprend le langage de modélisation graphique et il permet aussi à l'utilisateur d'effectuer des modèles de simulation avec le code Java. La nature de l'utilisation de Java dans AnyLogic est liée à l'extension de modèles personnalisés via le codage en Java, aussi bien qu'à la création d'applettes Java, qui peuvent être ouvertes avec n'importe quel navigateur standard. Ces applettes rendent les modèles AnyLogic très faciles à partager ou à placer sur des sites Web. En plus des applettes, la version Professionnelle permet la création d'applications indépendantes Java qui peuvent être distribuées aux utilisateurs. Ces applications Java peuvent servir de base comme outil d'aide à la décision[9]

Chapter 6

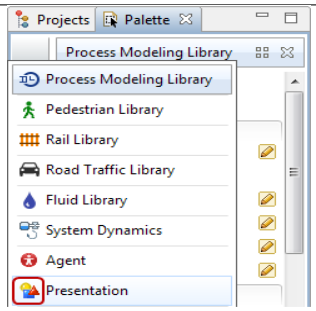
Execussion

6.1 Préparation de environnement denone simulation

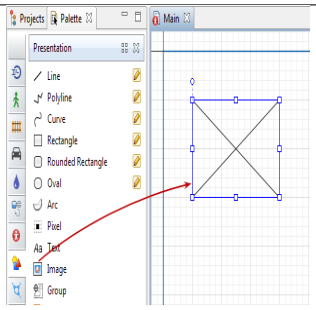
Nous allons créer un modèle basé sur la capture d'écran satellite disponible ci-dessous. On peut facilement voir toutes les particularités de la zone du réseau routier actuel. Les deux routes sont bidirectionnelles et contiennent une voie pour chaque direction de mouvement. Il y a un arrêt de bus à la route située au nord et un petit parking avec sept places de parking au bord de la route située à l'est.

Nous allons pas à pas considérer toutes ces particularités dans notre modèle, en démontrant la majorité des éléments de la biblIoThèque au fur et à mesure.

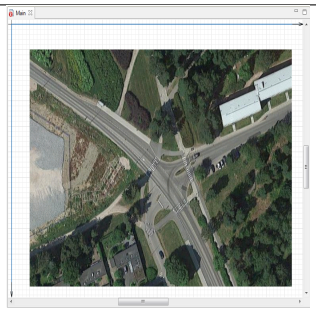
Etape 1: Dessin des routes nord et sud. à l'aide des blocs de la bibliothèque de la circulation routière



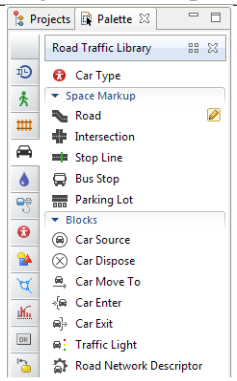
1. ouvrir la palette Présentation de la vue Palette. Pour ouvrir une palette, il suffit de cliquer sur l'icône correspondante dans le ruban vertical ancré dans la partie gauche de la palette.



2. Sélectionner la palette Présentation pour l'ouvrir. Faire glisser l'élément Image de la palette Présentation sur le diagramme graphique.



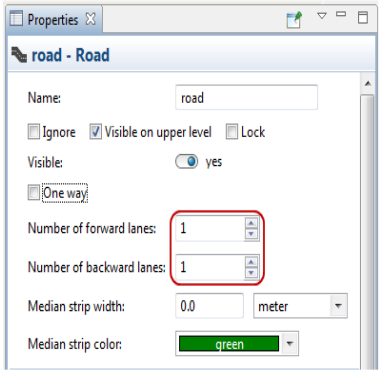
3. Choisir le fichier image à afficher. Accéder au dossier dans lequel on a enregistré le fichier, le sélectionner, puis cliquer sur Ouvrir dans la boîte de dialogue.



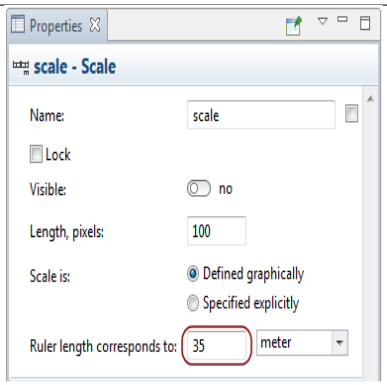
4. Dans la vue Palette, passer à la palette BiblIoThèque routière:

5. Double-cliquer sur l'élément Route dans la section Marquage d'espace.puis,Cliquez dans l'éditeur graphique pour dessiner le premier point de la route.

6. Mettre le dernier point de la route avec un double-clic.



7. Personnaliser les attributs de route dans la vue Propriétés.



8. Régler la règle d'échelle pour correspondre à 35 mètres.

Table 6.1: Etape1: Dessin des routes nord et sud

Etape 2: Ajouter une animation 3D

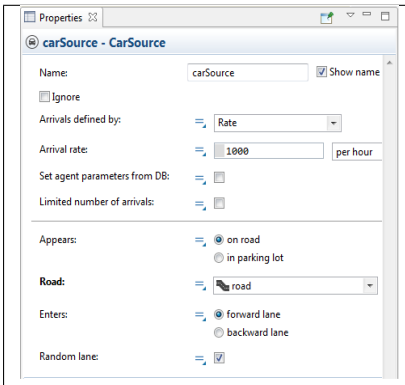
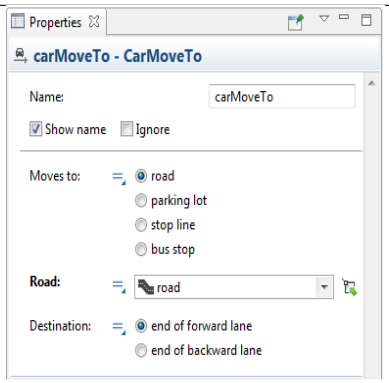
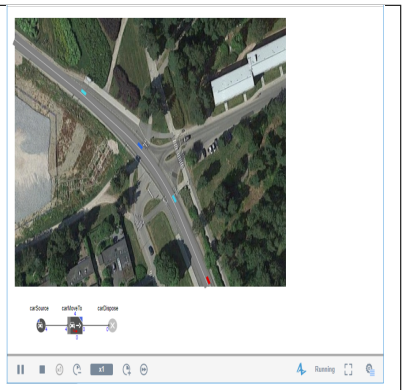
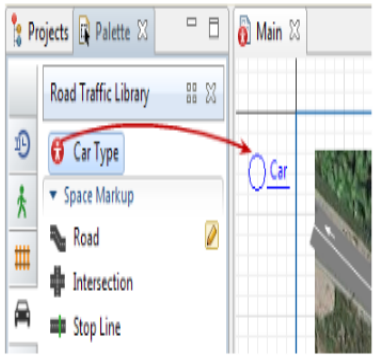
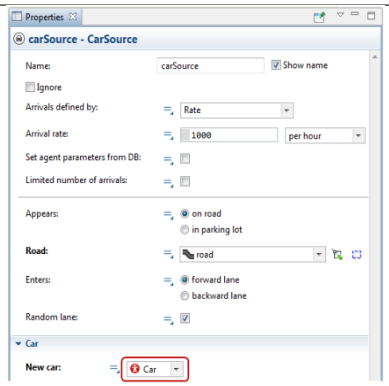
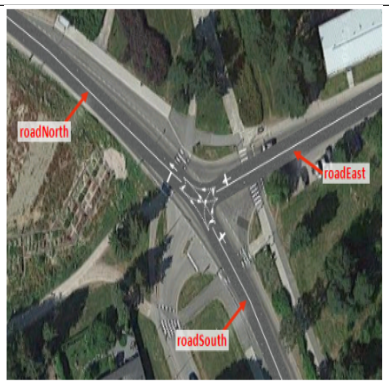
		
1. Sélectionner le bloc carSource, Dans la vue Propriétés, spécifier la fréquence d'arrivée des voitures.	2. Sélectionner road dans la liste déroulante.	3. Cliquez sur le bouton Exécuter de la barre d'outils.
		
14. Faites glisser l'élément Car Type dans l'éditeur graphique.	5. Dans la vue Propriétés, développez la section Car et choisissez Car dans la liste déroulante Nouvelle voiture.	6. Exécuter le modèle et passer en vue 3D pour voir les voitures se déplacer le long de la route en animation 3D.

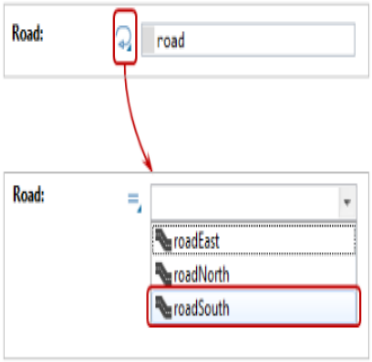
Table 6.2: Etape2: Ajouter une animation 3D

Etape 3: Dessiner la route Est pour former une intersection

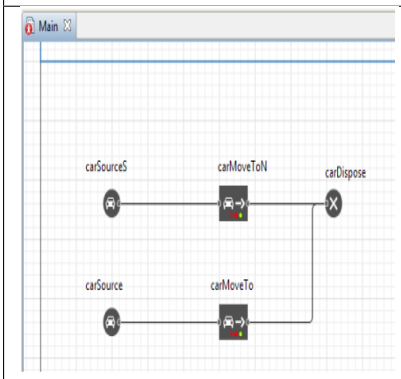
1. Commencez à dessiner la route en cliquant sur la bande médiane à la droite de la bordure de la carte.



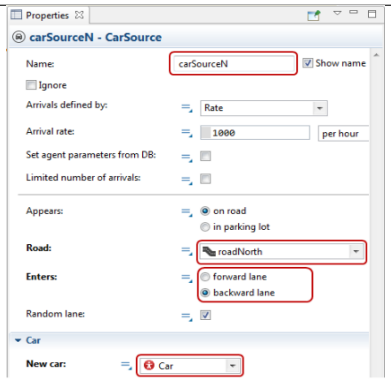
2. A la fin on obtiendra un résultat semblable à celui ci.



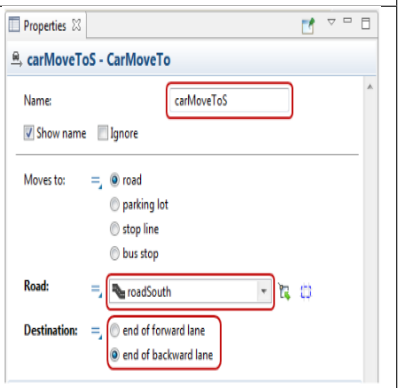
3. Séléctionnerinstitut montaigne :1771-6756, I. (Ed.) Big data et objets connectés Faire de la France un champion de la révolution numérique avril 2015 la route.



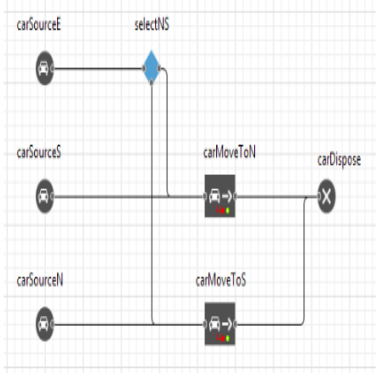
4. Nous allons maintenant ajouter des blocs qui modéliseront le mouvement de la voiture du nord au sud sur la même route



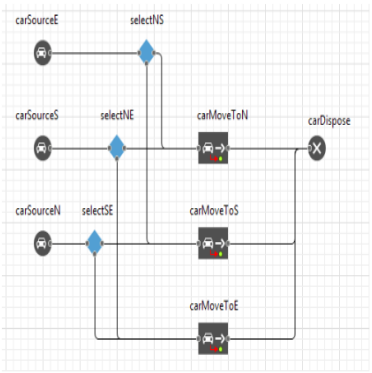
5.Renommer le bloc carSource en carSourceN. Dans le paramètre Road, choisir roadNorth. Choisir la voie arrière dans le paramètre Entrées. Ouvrir la section Propriétés de la voiture et choisir Car dans le paramètre Nouvelle voiture.



6. Apporter les mêmes modifications dans le bloc carMoveTo.



7. Ajouter un bloc SelectOutput à partir de la biblIoThèque de modélisation de processus.



8. Aprés avoir relié toutes les routes on est parvenu a ce résultat.

Table 6.3: Etape3: Dessiner la route Est pour former une intersection

Etape 4: Ajout du parking

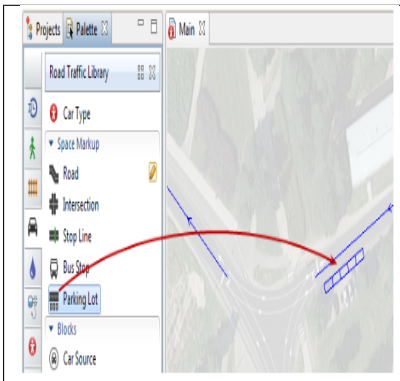
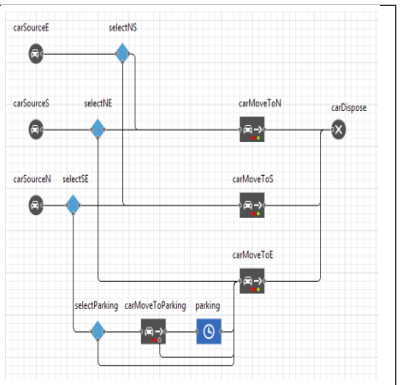
		
1. Faire glisser l'élément Parc de stationnement de la section Marquage d'espace de la palette BiblIoThèque de la circulation routière vers le diagramme de l'agent.	2. Passer à la vue Propriétés et définir le Type du parking: Parallèle ou Perpendiculaire. Ensuite définir le nombre de place de parking.	3. Pour exécuter la division de flux de voiture, ajouter un autre bloc SelectOutput à partir de la biblIoThèque de modélisation de processus.

Table 6.4: Etape4: noneAjout du parking

Etape 5: Ajout des bus et de l'arret de bus

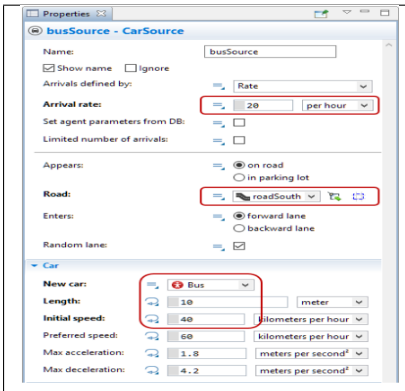
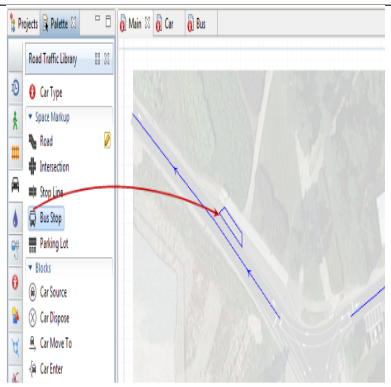
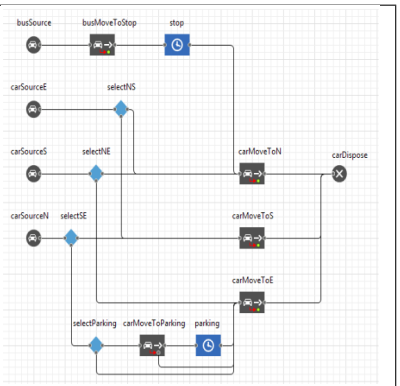
		
1. Pour créer le bus, il suffit de suivre les mêmes étapes que pour la création de la voiture et modifier les parametres comme le montre l'image ci-dessus.	2. Faire glisser l'élément Arrêt de bus de la section Marquage d'espace de la palette BiblIoThèque de la circulation routière vers l'éditeur graphique.	3. Ajoutez un autre bloc CarMoveTo à la branche de diagramme du bloc busSource, puis passez à la vue Propriétés du bloc busMoveToStop et choisir l'arrêt de bus dans le paramètre Déplacements vers. On doit maintenant modéliser la période de temps que les bus passent à l'arrêt de bus en allant dans les propriétés du bloc et en spécifiant le délai.

Table 6.5: Etape 5: Ajout des bus et de l'arret de bus

Etape 6: Ajouter les feux de signalisation

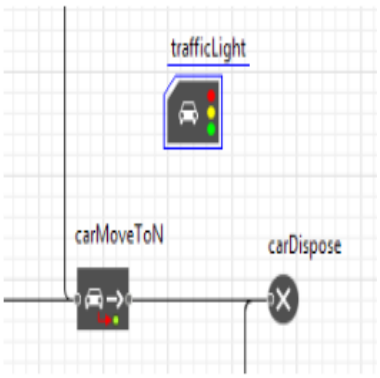
		
1. Ajouter le bloc de feu de signalisation à l'organigramme à partir de la section Bloc de la biblIoThèque de la circulation routière.	2. Cliquer sur une ligne d'arrêt dans l'éditeur graphique comme indiqué sur la capture d'écran ci-dessus et la couleur de la ligne d'arrêt passera du rouge au vert.	3. La capture ci-dessus montre le résultat final de notre simulation.

Table 6.6: Etape6 Ajout des feux de signalisation

Etape 7: Ajouter un graphe pour afficher des statistiques sur la fréquentation des routes en temps réel

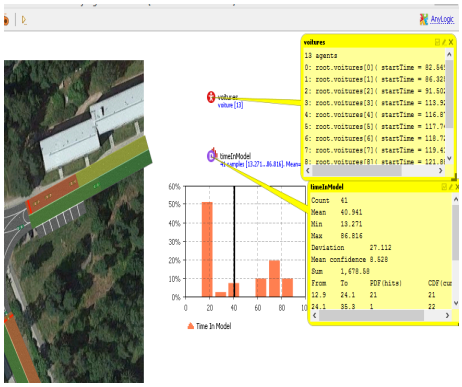
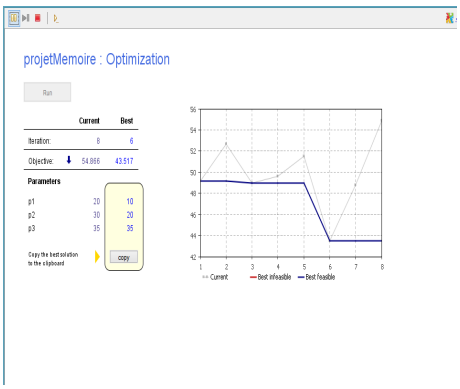
	
Ajouter le bloc diagramme à partir du bloc statistique. et créer un objets voitures sur lequel on implémentera une méthode qui récoltera les données de chaque voitures qui rentre dans notre intersection.	Utiliser un algorithme d'optimisation afin d'obtenir les valeurs optimales pour configurer les durées des feux de signalisation.

Table 6.7: Etape 7 statistiques et optimisations.

6.2 Simulation de la gestion du trafic routier

On rappel que le système de gestion de trafic routier classique, se base sur la logique de fixer une même durée pour tous les feux de signalisation d'une intersection données, ce qui engendre un problème de circulation dans la voie la plus fréquentée et également des délais d'attente inutile puisqu' il se peut que dans un voie il n'y est aucun véhicule.

6.2.1 Simulation de la gestion du trafic routier (cas classique)

Dans notre cas on doit passer par trois étapes qui sont le paramétrage, la collecte de données et l'affichage du résultat de la simulation.

Paramétrage

Avant de pouvoir lancer notre simulation on doit fixer les paramètres de base afin d'assurer son bon déroulement.

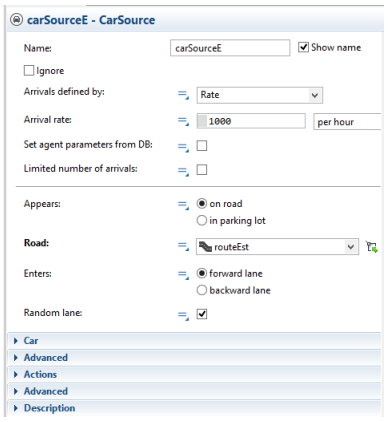
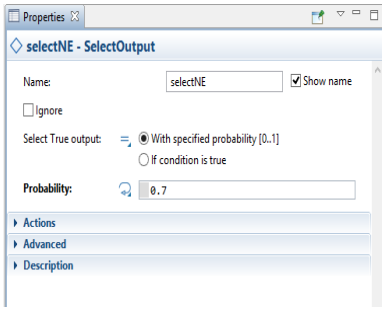
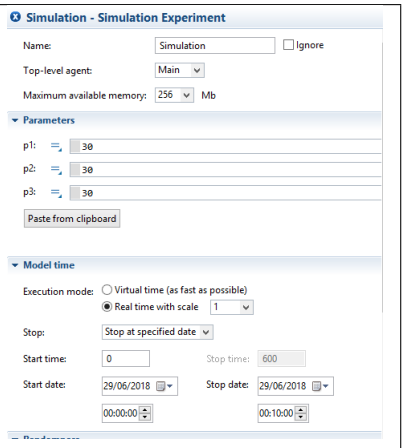
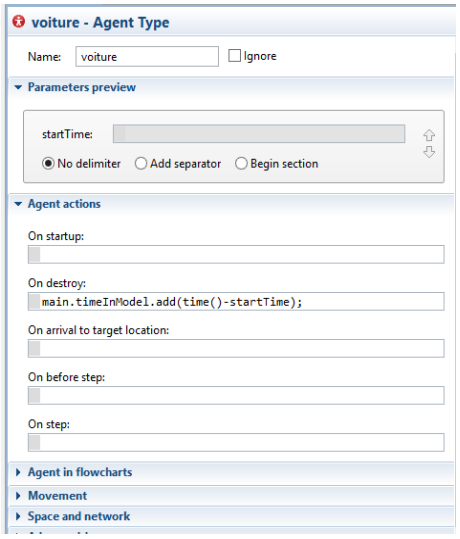
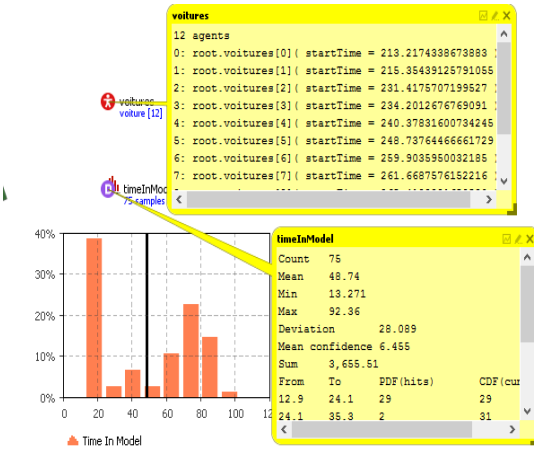
		
Fixer le nombre de véhicules générés sur chaque route.	fixer la probabilité qu'un véhicule prenne une route donnée.	Fixer la durée des feux et également la durée de la simulation.

Table 6.8: Paramètres de base

Collecte des données



Implémenter la méthode qui va récupérer les données sur chaque voiture lors de sa sortie de l'intersection.



Affichage des statistiques en temps réel de chaque voiture,les détails sur la circulation et le nombre de véhicule.

Table 6.9: Collecte des données et statistiques

On remarque que la méthode de collecte de données qu'on a implémenté récupère liées au véhicule et au trafic tel que :

- Le nombre de véhicule présent actuellement dans l'intersection.
- Le nombre de véhicule qui ont été générés depuis le début de la simulation.

- Le temps que met chaque véhicule depuis son entrée à sa sortie de l'intersection
- Le temps nécessaire au véhicule le plus lent pour sortir de l'intersection.
- Le temps nécessaire au véhicule le plus rapide pour sortir de l'intersection et diverses autres données qu'on a pas exploité dans notre travail.
- Dans notre cas nous nous sommes concentré sur le temps moyen nécessaire pour sortir de l'intersection.

Résultat de la simulation

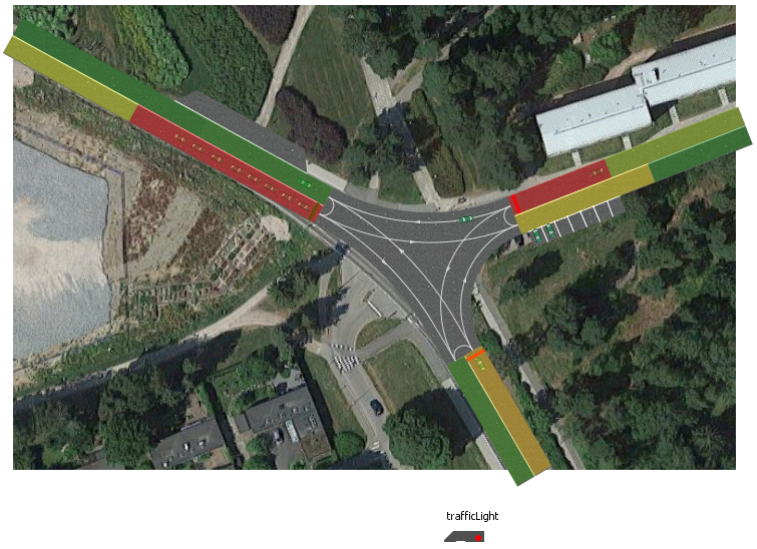
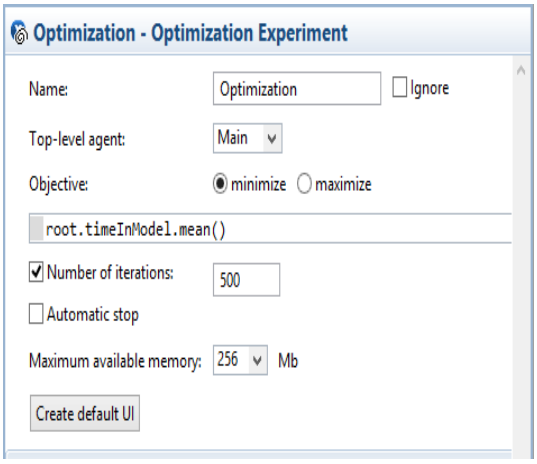


Figure 6.2.1: Simulation avant l'optimisation(cas classique)

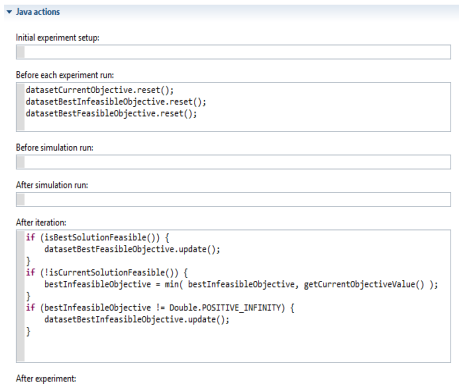
On remarque dans la voie nord un embouteillage est entrain de se former par contre dans les voies Est et sud il n'y a quasiment aucun véhicule. C'est ce qui nous a poussé a penser à une méthode pour rééquilibrer le trafic.

6.2.2 Simulation de la gestion du trafic routier (après optimisation)

Paramétrage



Fixer les paramètres de base de l'optimisation et appel de la méthode de récupération des données.



Implémentation du programme d'optimisation.

Table 6.10: Phase d'optimisation

Le programme d'optimisation est un ensemble d'appel de méthode qui sont implémenté par défaut dans les bibliothèque du simulateur, ces méthodes permettent de collecter les meilleur temps (temps moyen, temps minimal et temps maximal) et retourne également les durée des feux nécessaire pour les réaliser.

Tests et évaluations

Le tableau suivant représente le temps moyen retourné en fonction des valeurs définis dans les trois feux de notre intersection.

Feu Sud	Feu Nord	Feu Est	Temps moyen
30	35	15	51.48
25	35	25	49.46
10	25	20	48.62
25	30	25	49.97
20	15	35	43.51
15	25	35	40.78
10	35	35	49.46
15	30	10	51.68
35	25	35	49.03
10	25	10	40.59

Table 6.11: Tableau montrant l'évolution du temps moyen

Le déroulement de l'algorithme d'optimisation nous retourne le graphe suivant:

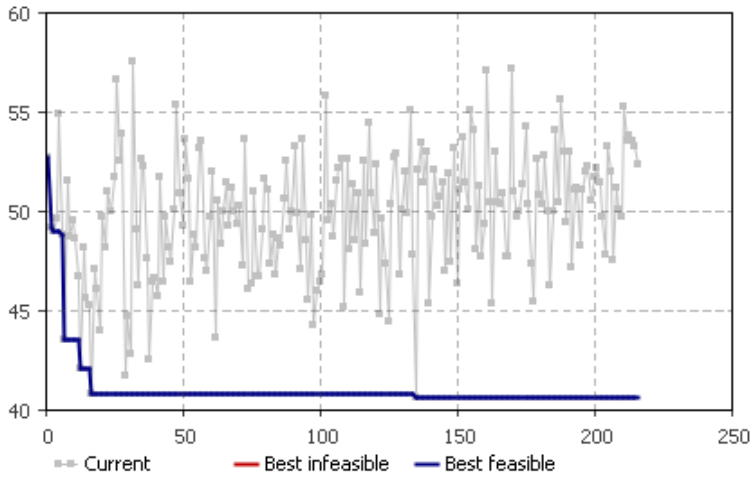


Figure 6.2.2: Graphe montrant l'évolution du temps moyen

Le graphe nous retourne en bleue l'évolution du temps moyen optimal nécessaire au véhicule pour sortir de l'intersection, on remarque que la valeur de ce temps se met à jour à chaque fois qu'il trouve une combinaison qui retourne un meilleur temps.

En gris ces les résultats de toutes les autres combinaisons.

Résultat de la simulation

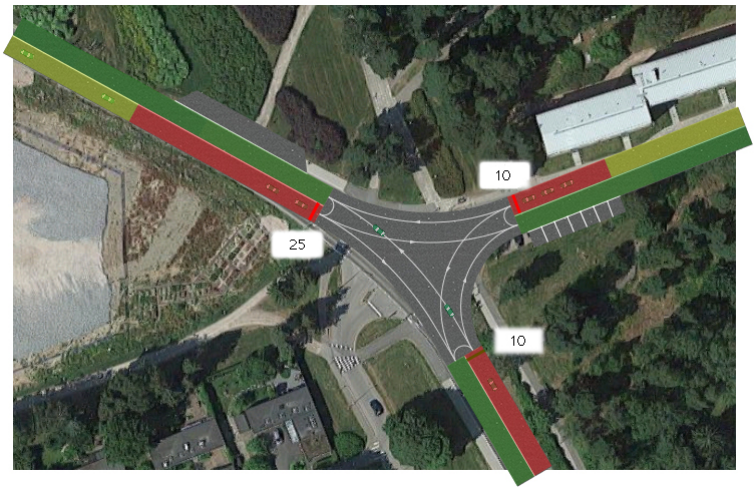


Figure 6.2.3: Résultat de l'optimisation

Après l'implémentation des temps récupérés de la phase d'optimisation on remarque que le trafic routier s'est rééquilibré sur les trois voies et le temps moyen a été réduit.

Conclusion

Dans cette simulation on a pas eu besoin d'avoir recours a une bdd NoSQL étant donné la faible quantité de données récoltée, et le format qui est le même pour toutes ces informations mais dans la réalité on aura besoin de récolter des données de divers formats tel que le MP4 pour la vidéo, le texte, le JSON, les différents formats d'images(jpg, png....).

Conclusion

Notre travail a consisté à récolter des données sur le trafic routier et de les exploiter afin d'améliorer la circulation des véhicules dans nos villes, pour ce faire on a dû avoir recours à diverses méthodes et technologies récentes comme le cloud, les capteurs et les BDD NoSQL.

Perspective d'avenir

- Réaliser le programme d'optimisation de la gestion du trafic.
- Réaliser le programme qui contrôlera les feux de signalisations.
- Implémenter ce système sur plusieurs intersections qui seront successive pour voir son réel impacte sur le trafic.

Annexe

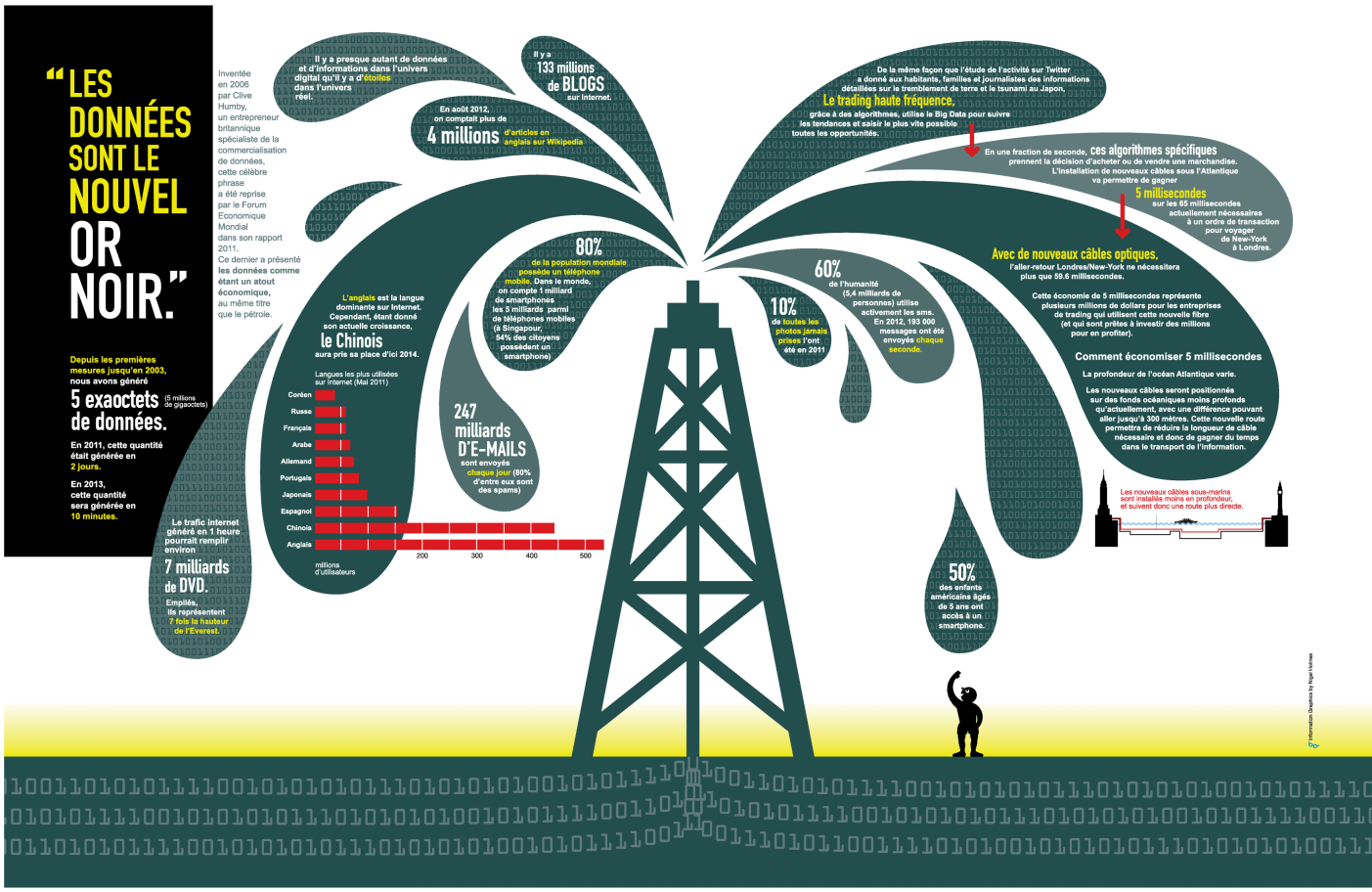


Figure 6.2.4: Les données nouvel Or Noir

Webographie

IoT

- http://www.squirrel.fr/tutoriel-IoT-intel-arduino-web-3/
- https://blog.xebia.fr/2016/02/26/linternet-des-objets-2-connecter-vos-capteurs-aux-resea
- https://www.cisco.com/c/en/us/solutions/internet-of-things/overview.html#~stickynav=3
- https://www.supinfo.com/articles/single/5256-evolution-internethttp://www.idnext.net/linternet-des-objets-phenomene-de-mode-ou-revolution/

Big data

- https://www.cetic.be/Comment-deployer-avec-succes-un-projet-Big-Data
- https://www.lebigdata.fr/definition-quest-data-analytics
- https://journals.openedition.org/sociologie/2613
- https://www.lebigdata.fr/definition-big-data
- https://www.memoireonline.com/05/14/8890/m_Big-data-rapport-de-stage0.html
- https://www.saagie.com/fr/blog/qu-est-ce-que-le-big-data-definition
- https://www.tutorialspoint.com/big_data_analytics/index.htm

Cloud

- http://www.tutorialspoint.com/articles/cloud-computing-and-big-data
- https://blog.blaisethirard.com/qu-est-ce-que-le-cloud-computing/
- https://www.guru99.com/cloud-computing-for-beginners.html
- https://www.figer.com/Publications/nuage.htm

NoSQL

- https://www.grafikart.fr/blog/sql-nosql
- https://dzone.com/articles/nosql-in-the-cloud-a-scalable-alternative-to-relat

Hadoop

- https://data-flair.training/blogs/install-hadoop-on-ubuntu/

Anylogic

- https://www.anylogic.com.

Bibliography

[1] les big data et lâInternet des objets AprÃšs lâInternet : le Cloud. *Les Enjeux de l'information et de la communication*. 2001 Ã 2017.

[2] Andrei Bors Christian Wartha, Momtchil Peev. Winter simulation confere. 2002.

[3] Cisco. Internet des objets (iot). 16 dÃ©c. 2015.

[4] StratÃ©gie Big Data. Thomas davenport. Edition Pearson.

[5] Thomas Davenport. *StratÃ©gie Big Data*.

[6] Minestry of labor Ganesh chandra Deka and Employement gouvenement of india. Asurvey of cloud database systems. avril 2014.

[7] Ismail MERZOUK. Article les piliers de l'ioe. PubliÃ© le 21/07/2017.

[8] Sylvain Bureau et FranÃ§oise Massit-FollÃ©a Pierre-J.Benghozi. *L'Internet des objets*. 10 septembre 2009.

[9] ThÃšse. Conceptiond'unfeudecirculation routiÃšre intelligent. Technical report, EncadrÃ© par Pr.Mr Larouche .-RÃ©alisÃ©:Yazdad.Abdesselam, 2014/2015.