Les fondamentaux du Big Data

Joyce MBIGUIDI • Data Scientist

Objectifs pédagogiques

Objectifs pédagogiques

- Comprendre les enjeux et paramètres à intégrer pour la mise en place du Big Data au sein de l'entreprise.
- Appréhender et choisir les outils appropriés pour l'installation du Big Data.
- Intégrer les principes et la terminologie de la Big Data.
- Manipuler des données à travers des cas pratiques.

Partie 1
Introduction

Plan du module

1. Présentation et définition du Big Data
2. Apports
3. Le marché du Big Data
4. Démystifier cette ressource
5. Les technologies concernées
6. Hadoop et son écosystème
7. Les métiers
8. Les distributions
9. TP : présentation des acteurs du marché

Présentation et définition du Big Data

Les systèmes d'informations

- Entre 1960 et 1980, on passe des grands systèmes centralisés aux systèmes « client-serveur ».

- Une pléthore d'entreprises adoptent des progiciels tels que SAP, reposant sur les bases de données relationnelles. C'est l'essor des ERP.

- Révolution d'internet en 2000 : partage et diffusion de l'information, commerce électronique à travers le monde.

Présentation et définition du Big Data

Les systèmes d'informations

- À partir des années 2000 :
  - Démocratisation des technologies de type : ordinateurs, smartphones, tablettes, livres électroniques…
  - Réseaux sociaux : Facebook, Twitter, LinkedIn…

➔ ***Les entreprises identifient le potentiel BtoC.***

- **Cloud computing :** utilisation de la mémoire, des capacités de calculs et passage à l'échelle.

Présentation et définition du Big Data

Les données massives

- Né du fait de l'accumulation de quantités de données de plus en plus importantes.

- Les données sont trop volumineuses et représentées sous différentes formes qu'elles ne peuvent plus être traités par les moyens classiques (base de données relationnelles, requêtes SQL).

- Les entreprises de moteurs de recherches ont été les premières à faire face à ce type de difficultés, notamment en raison de la volumétrie et des données non structurées.

Présentation et définition du Big Data

Les données massives

- En 2001 création d'une base de données compressées appelée **Big Table**.
- Création de l'algorithme **MapReduce**.
- Publication de ces travaux de recherche en 2004.

Présentation et définition du Big Data

Les données massives

- Doug Cutting lit les travaux de recherche de Google et lui vient l'idée de lancer Hadoop.

- La force de Hadoop repose sur :
  - Un modèle de programmation MapReduce
  - Un système de fichiers hautement distribués
  - La gestion de grands volumes de données non structurées

Présentation et définition du Big Data

Les données massives

- Doug Cutting finit par rejoindre Yahoo pour stabiliser le prototype Open Source Hadoop.
- Puis, il intègre Cloudera (une des sociétés qui va enrichir la solution Hadoop).
- Yahoo a son tour va créer la société Hortonworks qui apportera sa part de contribution au projet Hadoop.

Présentation et définition du Big Data

Les données massives

- Le terme Big Data fait son apparition vers 2010, bien que les contraintes étaient déjà identifiées au début des années 2000.

- Il s'agit donc :
  - De de grandes quantités de données structurées et non structurées.
  - De données impossibles à stocker et à traiter avec les moyens traditionnels.
  - Des données diverses produites en temps réel.

Présentation et définition du Big Data

Big Data, les 3 V : **volume**, vitesse et variété

- Le **volume** traite du stockage et du traitement de grandes volumétries de données.
- On est historiquement passé de quelques mégaoctets (10^6) à plusieurs exa (10^{18}) voire zettaoctets (10^{21}) en une dizaine d'années.

Présentation et définition du Big Data

Big Data, les 3 V : volume, **vitesse** et variété

- La notion de **vitesse** fait appel au gain d'efficacité engendré par le progrès technologique inhérent au Big Data.
- Auparavant on était capable de traiter quelques mégaoctets en plusieurs jours. Aujourd'hui, en quelques minutes.

Présentation et définition du Big Data

Big Data, les 3 V : volume, vitesse et **variété**

- La notion de **variété** intègre l'idée d'un stockage et d'un traitement de données de différents types et de différentes sources :
  - Texte, images, sons, vidéos, logs…
  - Ordinateurs, tablettes, smartphones, capteurs, puces RFID, GPS, caméras, réseaux sociaux, caméras, blogs

- Un quatrième **V** pour **valeur** peut être envisageable. Car l'objectif des entreprises est d'en tirer des insights.

Travaux pratiques

Associez chaque bloc à sa thématique et justifiez votre choix.

10 *minutes*

VOLUME

VARIETE

VITESSE

Séquences sonores

Transactions financières

Trading automatique

Blogs

Tickets d'opérateurs téléphoniques

Réseaux sociaux

Publicités comportementales sur internet

Vidéos

Photos

Emails

Surveillance d'un site de production

Géolocalisation des voyageurs

Relevés de consommation de gaz

Prévisions météorologique

Mesures par capteurs

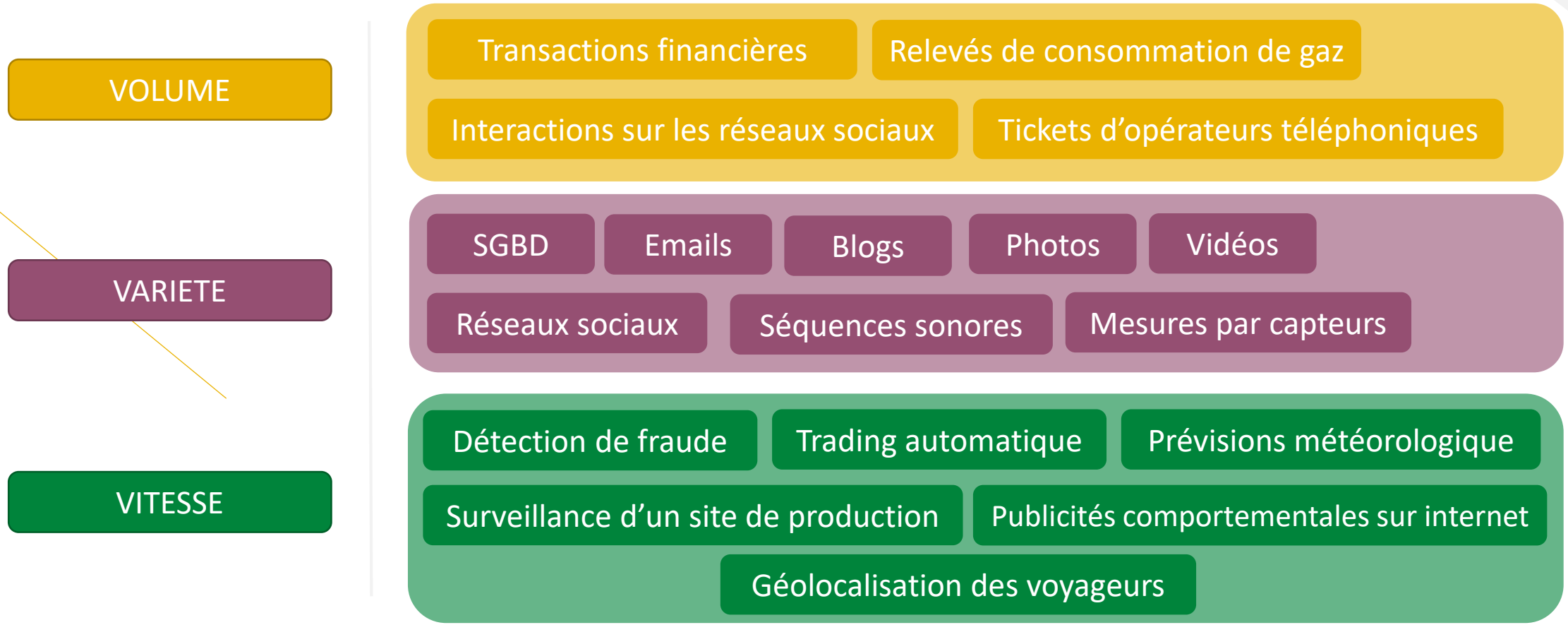
SGBD

Détection de fraude

Interactions sur les réseaux sociaux

Travaux pratiques

Solution possible



Apports du Big Data

Créer de la valeur à partir des données

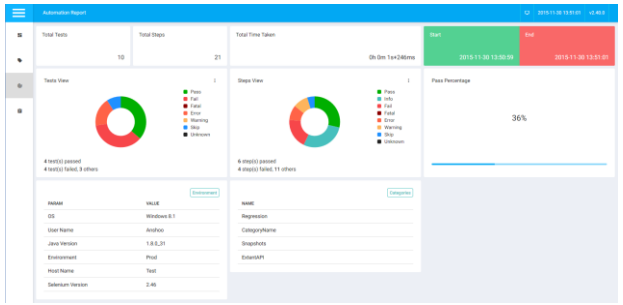
- Les données sont généralement sous exploitées en entreprise. Y accorder une attention particulière permet d'être **compétitif** vis-à-vis des concurrents.
  - Analyse comportementale des clients.
  - Moteurs de recommandations, règles d'associations de produits…

- Il est possible de **croiser** des données de toutes sortes à forte volumétrie, plutôt que de restreindre les analyses sur des échantillons plus petits.
  - On dispose ainsi d'une vision 360° du business.
  - On comprend mieux le présent, pour anticiper le futur.

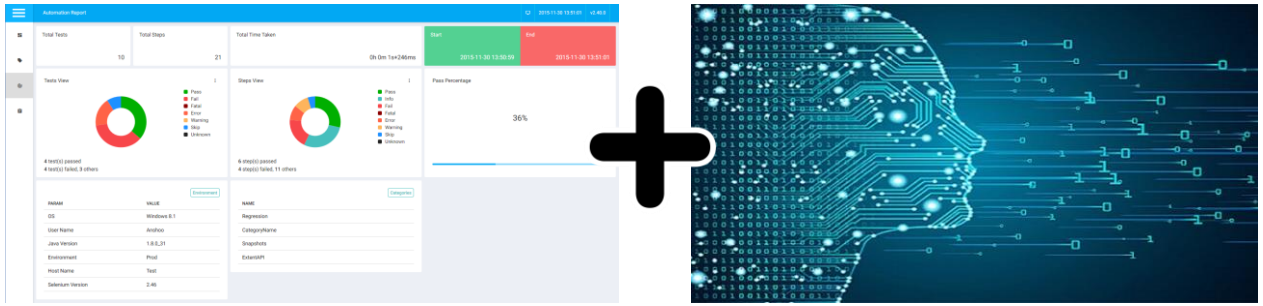
Apports du Big Data

Créer de la valeur à partir des données

- Un changement dans l'usage des données : avant vs aujourd'hui



- Business intelligence classique.
- Reporting des données de la veille.



- Aujourd'hui on dispose d'un modèle hybride, qui associe connaissance antérieure, actuelle et anticipation des événements futurs, grâce à l'association de la Business Intelligence et le Machine Learning.

Apports du Big Data

Défi technologique majeur

- L'accroissement exponentielle des données requiert de nouveaux procédés de stockage, de traitement et de visualisation des données.

- On est passé des systèmes classiques de type données structurées (SGBDR) aux données de type NoSQL, reposant sur des architectures plus sophistiquées et hautement distribuées pouvant gérer des transactions en temps réel.

Apports du Big Data

Les apports du Big Data dans le Marketing

- Dans les entreprises matures dans la data, le département Marketing a énormément recourt au Big Data.

- Les entreprises scrutent les réseaux sociaux : mesure d'audience, e-réputation, profils d'abonnés, centres d'intérêts, fake news…
  - Objectif : mieux comprendre les usages et les attentes de clients.

- Segmentation par profil de consommateur, recommandation et associations de produits en magasin physique ou en ligne.

Apports du Big Data

Les apports du Big Data dans l'énergie

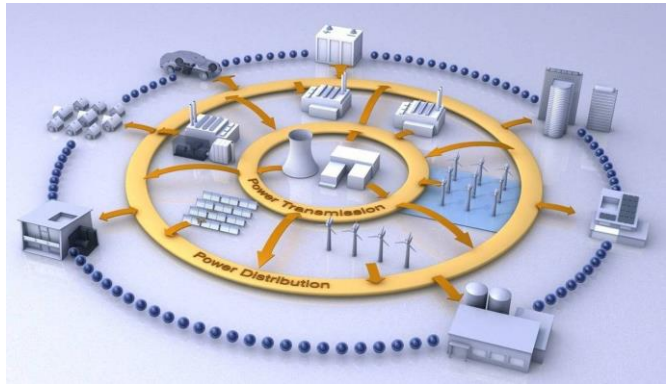
- L'émergence des **smart grids** : ou réseaux intelligents. Les opérateurs disposent de leviers pour augmenter ou diminuer la quantité d'énergie en fonction de la consommation.

- Ces réseaux permettent :
  1. D'optimiser la production et la distribution d'électricité.
  2. De trouver un équilibre entre l'offre et la demande d'électricité.

- Les relevés de compteurs à distance, parfois par pas de 15 minutes génèrent de grandes quantités de données à stocker et à traiter.

Apports du Big Data

Les apports du Big Data dans l'énergie



Issygrid : le smart grid à la française

- Premier smart grid français déployé en 2012.
- Issy-les-Moulineaux devient le premier quartier intelligent.
- Objectif : réduire les émissions de CO2
- +1600 logements et 4 immeubles rattachés au réseau intelligent.
- Efficacité énergétique : +20%
- Emprunte carbone : -20%

Apports du Big Data

Les apports du Big Data dans la santé

- Recherche médicale : association de médicaments, méthodes de diagnostics, protocoles de soins, imagerie médicale…
- Réduction du temps de calcul : super calculateur, parallélisation des calculs…
- Santé publique : prévenir les épidémies grâces aux données de requêtes effectuées sur les moteurs de recherches, données médicales des patients des hôpitaux…

Apports du Big Data

Les apports du Big Data dans les services publics

- Données gouvernementales sur les citoyens et les entreprises : les impôts pré remplissent les avis d'impositions, la CAF croise les données des usagers avec celles des finances publiques et de l'URSSAF…

- Des cas de fraudes sont détectés grâce aux données massives récoltées (cohérence du train de vie, déclarations suspectes...).

Apports du Big Data

Résumé des apports du Big Data dans l'entreprise

- Amélioration de la connaissance dans les données (clients, chaîne logistique, santé, fraude…).
- Anticipation et prise de décisions (modèles statistiques avancés, simulations).
- Amélioration de la compétitivité.
- Diminution des coûts liées au stockage de la données, aux études statistiques…

Le marché du Big Data

Les principaux acteurs

cloudera

- Créée en 2008, Cloudera est une Startup californienne spécialisée dans le développement d'Apache Hadoop.
- Son fondateur est Jeff Hammerbach.
- Hadoop est un projet Open Source, mais Cloudera a été la première société à l'avoir commercialisé.
- Client vers lesquels elle a déployé Hadoop : Groupon, Navteq, Samsung etc.

Le marché du Big Data

Les principaux acteurs

- Société californienne créée en 2011 par Yahoo, avant d'être rachetée par Cloudera en 2018.

- Spécialisée dans le développement d'Hadoop, ses partenaires sont Microsoft, Informatica et Teradata.

- Société californienne qui vend des solutions basées sur Hadoop.

Le marché du Big Data

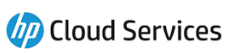
Quelques acteurs du stockage physique

- Principales sociétés spécialisées dans le stockage physique de données Big Data.

Le marché du Big Data

Quelques acteurs du stockage dans le Cloud

- Principales sociétés spécialisées dans le stockage IaaS (Infrastructure as a Service).



Le marché du Big Data

Quelques fournisseurs d'entrepôts de données

- Leader historique, propose ces services depuis plus de 30 ans.
- Leader mondial des SGBD, +300K clients dans le monde.
- Poids lourd du stockage de données. Plusieurs milliers de clients.
- Fournit plusieurs solutions dont Microsoft SQL Server.

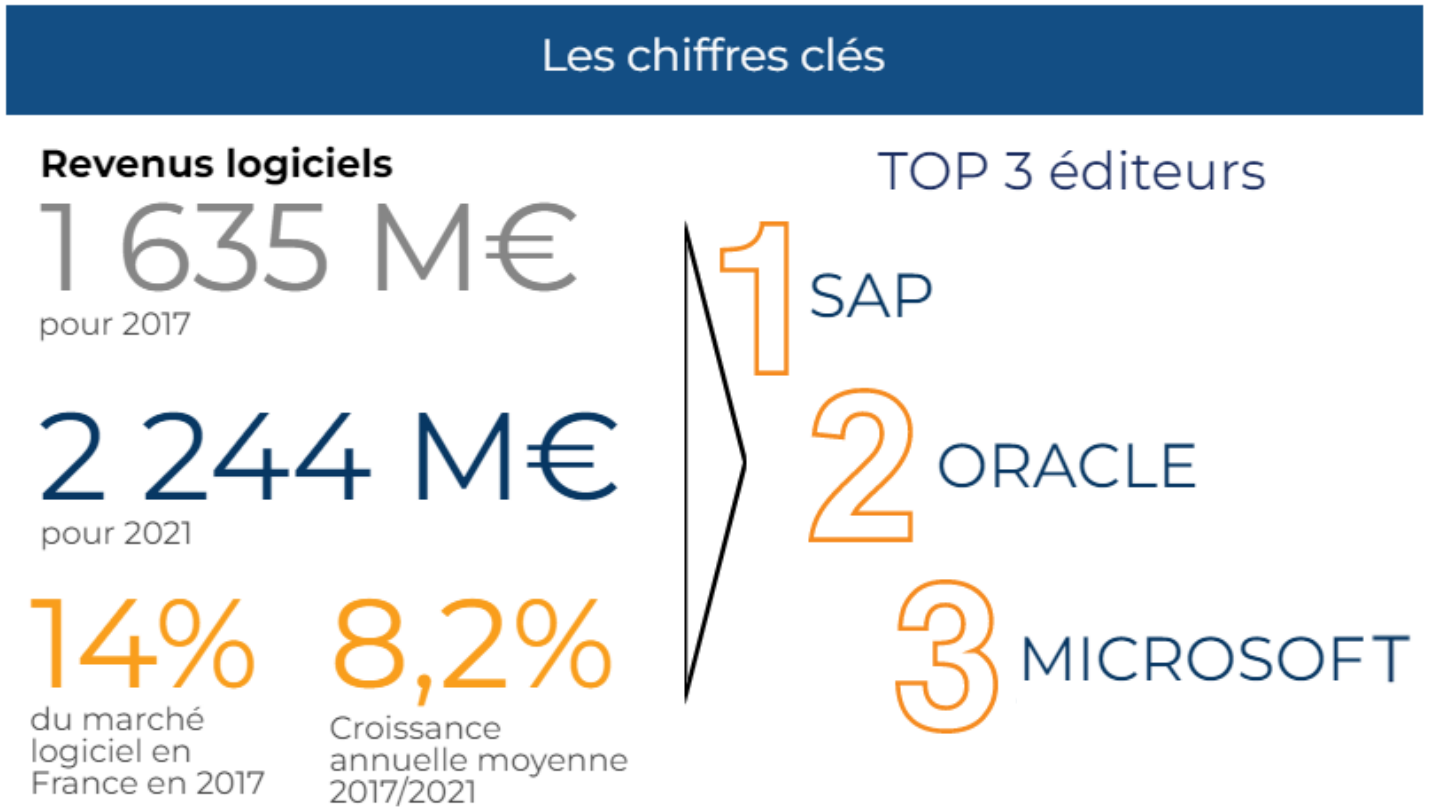
Le marché du Big Data

Quelques fournisseurs de solutions d'analyses

- Créé en 1976, c'est l'éditeur le plus répandu sur le marché de l'analytics.
- Business Intelligence et data visualisation. Plusieurs clients dans le monde.
- Éditeur spécialisé en Business Intelligence. Créé en 1989, propose sa solution à de nombreux clients dans le monde.
- Américain, spécialiste de la visualisation de données. Créé en 2003 et racheté en 2019 par Salesforce.

Le marché du Big Data

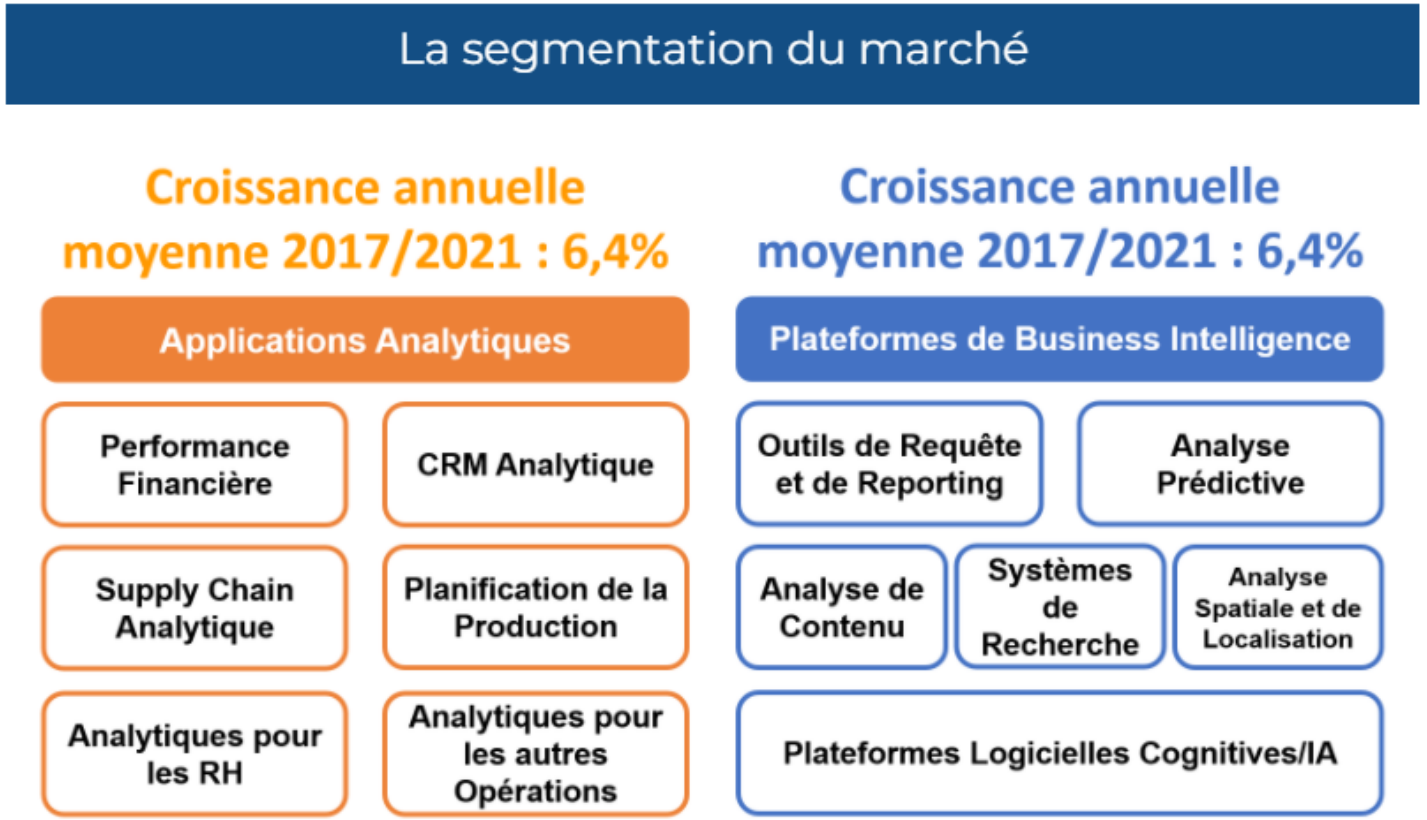
En France



Extraits de l'étude IDC France « Big Data Analytics », Juin 2017

Le marché du Big Data

En France



Extraits de l'étude IDC France « Big Data Analytics », Juin 2017

Démystifier cette ressource

Les composantes d'une architecture Big Data

1. **La source de données :** datawarehouse, cloud…
2. **Stockage :** magasin de données, data lake
3. **Batch processing :** traitement des données par lots
4. **Stream processing :** traitement de flux de données (temps réel)
5. **Préparation des données :** qualité des données
6. Catalogue des données (métadonnées accessoirement)
7. **Modélisation des données :** Data Mining, Machine Learning
8. **Technologie d'orchestration :** kubernetes, Docker…

Démystifier cette ressource

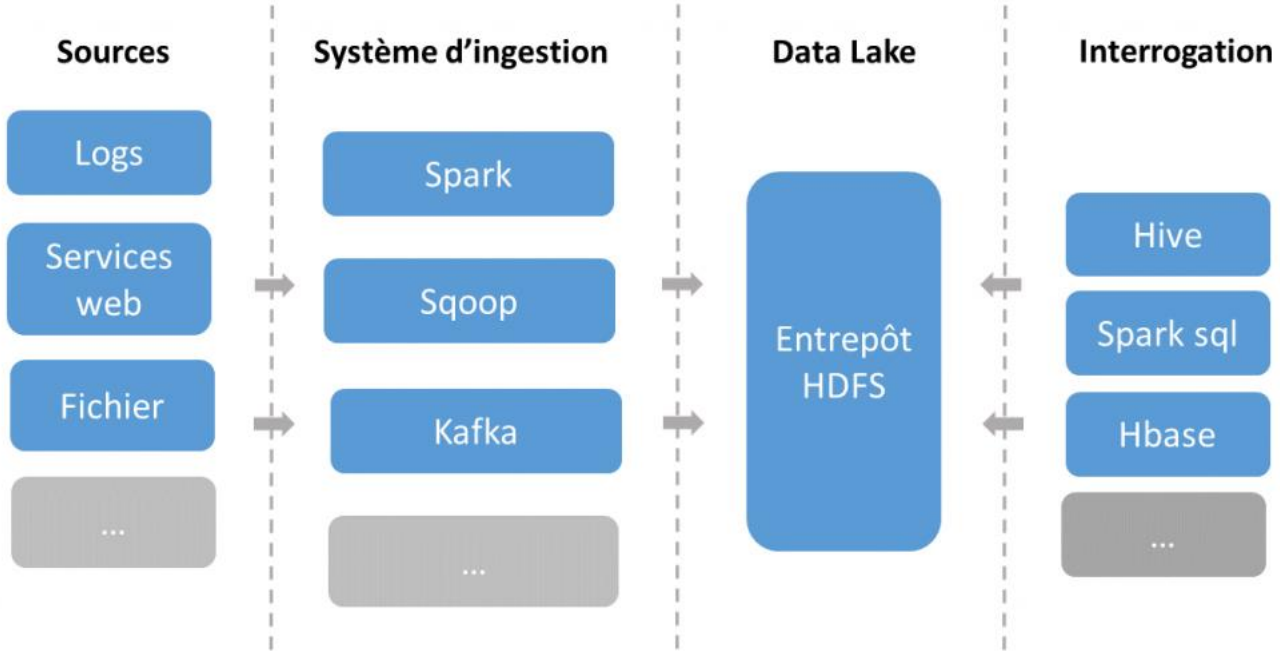
Les 3 types d'architectures Big Data : Datalake, Lambda et Kappa**.**

- Le Datalake offre la possibilité de stocker tous types de données au même endroit.
  - Données brutes
  - Données traitées
  - Données structurées
  - Données non structurées
  - Données semi-structurées

- Les Datalake reposent généralement sur le framework Hadoop

Démystifier cette ressource

Les 3 types d'architectures Big Data : Datalake, Lambda et Kappa.



https://www.cyres.fr/blog/architecture-lambda-big-data/

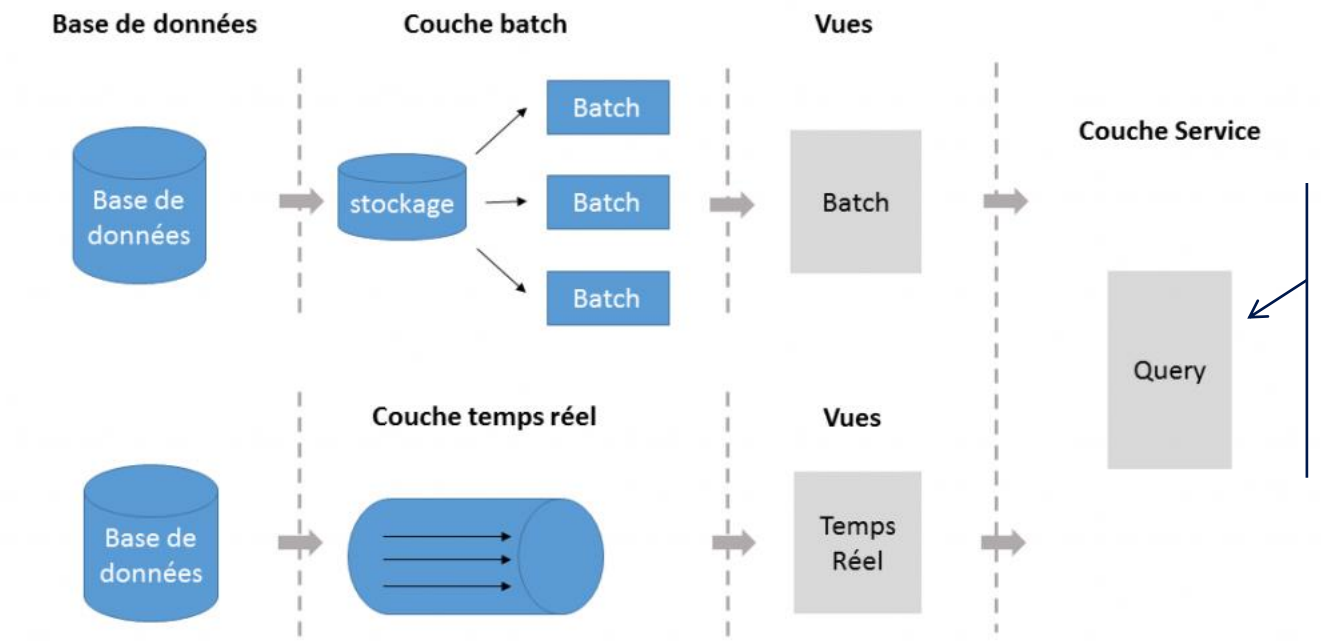
Démystifier cette ressource

Les 3 types d'architectures Big Data : Datalake, Lambda et Kappa.

- Architecture la plus utilisée pour le traitement des données en **temps réel** et en **batch** de façon simultanée.
- La couche de traitement par lots (batch) sert à récupérer les données et à les stocker au format brut dans des datalake. Des vues logiques sont créées dans la couche service.
- La couche en temps réel traite les **nouveaux flux de données** pour générer des vues avec des données récentes en temps réel.

Démystifier cette ressource

Les 3 types d'architectures Big Data : Datalake, Lambda et Kappa.



- Permet de stocker et d'exposer aux clients les vues créées par les couches batch et temps réel.
- Adapté à tous types de bases NoSQL.

https://www.cyres.fr/blog/architecture-lambda-big-data/

Démystifier cette ressource

Les 3 types d'architectures Big Data : Datalake, Lambda et Kappa.

- Modèle scalable et résistant aux incidents.

- Ce modèle impose la cohabitation de deux systèmes : batch et traitement en temps réel.

Démystifier cette ressource

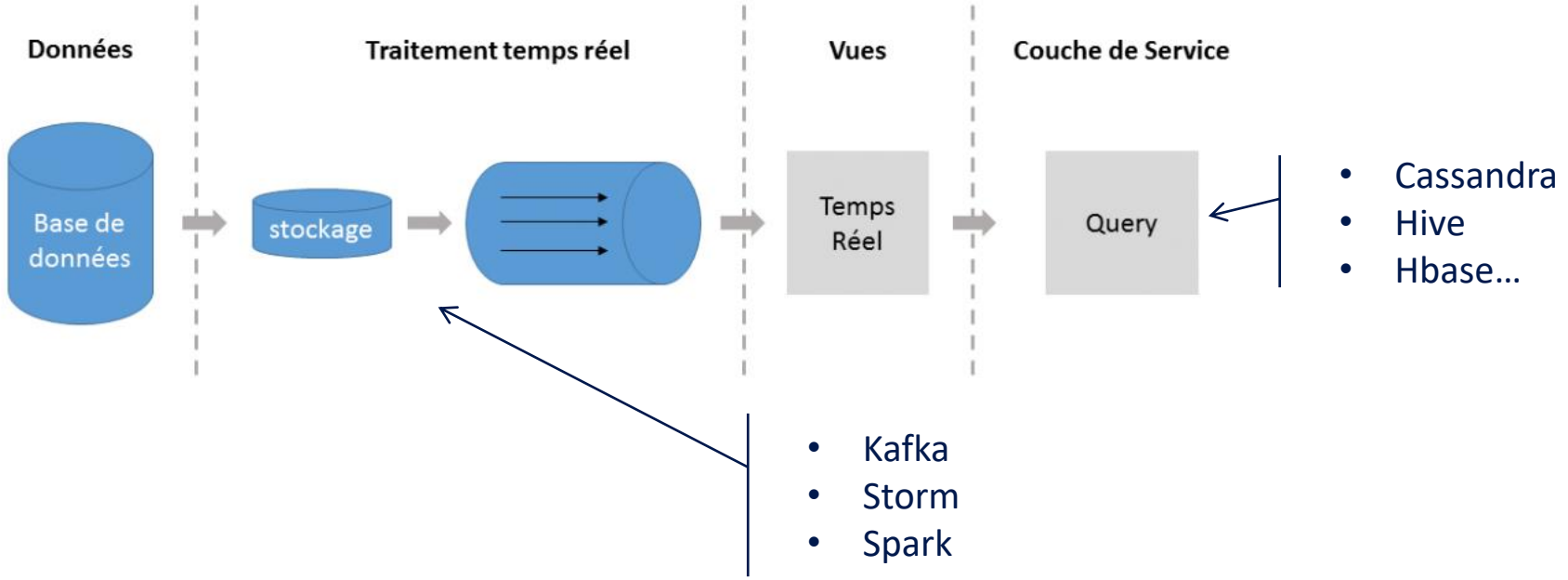
Les 3 types d'architectures Big Data : Datalake, Lambda et Kappa.

- Contrairement à l'architecture Lambda, Kappa fusionne la couche batch et la couche temps réel.

- Kappa ne stocke pas les données Big Data, mais y effectue des traitements.

Démystifier cette ressource

Les 3 types d'architectures Big Data : Datalake, Lambda et Kappa.



https://www.cyres.fr/blog/architecture-lambda-big-data/

Les technologies concernées

Apache Hadoop

- De loin, la solution la plus répandue pour traiter de très gros volumes de données.

- Hadoop est composé de plusieurs éléments :
  - Un système de stockage (HDFS)
  - Un système de planification des traitements (YARN)
  - Un framework de traitement (MapReduce).

Les technologies concernées

Traitements de type batch

- Les traitements opérés sur les données sont continus et incrémentaux c'est-à-dire que l'architecture va à chaque fois prendre en compte les nouvelles données sans avoir à traiter à nouveau les anciennes.

- Pour avoir une cohérence dans le traitement de ces données, les résultats ne sont visibles et accessibles qu'à la fin du traitement (une fois qu'il n'y a plus de données à l'entrée).

- Il existe comme traitement Big Data de type batch Map Reduce dans sa version Hadoop ou encore Apache Spark.

Les technologies concernées

Traitements en temps réel

- C’est l’inverse des traitements de type batch. Grâce à cette méthode, il n’est pas nécessaire d’attendre la fin de traitement des données pour accéder aux résultats.

- Cas d’usages :
  - Analyse du parcours de navigation d’un site de e-commerce : analyser les actions, l’engagement des internautes, anticiper les abandons de paniers…
  - Internet des objets : analyse des données en provenance de capteurs d’une chaîne d’assemblage ou d’une flotte d’avions afin de prévenir les pannes…

Les technologies concernées

Les bases de données NoSQL

- Les bases de données relationnelles classiques servent à gérer les données qualifiées de l'entreprise mais ne sont pas habilitées à stocker de la donnée à grande échelle avec un traitement rapide.

- Les bases NoSQL sont plus **flexibles, plus adaptables aux évolutions et moins sensibles aux pannes de systèmes**.

Les technologies concernées

Les bases de données orientées colonnes

- Parmi eux : Cassandra et Hbase
- Ces bases de données sont très performantes dans la lecture et l'écriture de gros volumes de données.
- Ce type de base sera capable d'assumer des montées en charge progressives sans sacrifier les fonctionnalités existantes.

Les technologies concernées

Le cloud computing

- Le cloud computing n'est pas une technologie Big Data stricto sensu, mais c'est une **méthode de déploiement pour les technologies Big Data**.

- Face aux capacités énormes de stockage et de traitement, le cloud est aujourd'hui le moyen le plus capable de supporter ces volumétries et à moindre coût comparé à une solution classique on-premise.

Les technologies concernées

Les sociétés et leurs technologies

Google

Google Cloud Bigtable

- Système de base de données distribuée reposant sur GFS (Google File System).
- N’est pas Open Source.

- Framework de développement pour traitements distribués.

Les technologies concernées

Les sociétés et leurs technologies

YAHOO!

- Plate-forme Java destinée au Big Data.
- Inspiré des projets Google : Big Table, MapReduce, Google File System.

S4

- Plate-forme dédiée au traitement des flux de données.

Les technologies concernées

Les sociétés et leurs technologies

facebook

- Base de donnée de type NoSQL et distribuée.
- Infrastructure de requêtes et d’analyses de données intégré dans Hadoop.

Les technologies concernées

Les sociétés et leurs technologies

twitter

APACHE STORM™

- Plate-forme de traitement de données massives.

DB DB.io Database of Databases FlockDB

- Base de données distribuée de type graphe.

Les technologies concernées

Les sociétés et leurs technologies

SenseiDB

- Système distribué de gestion des messages.
- Base de données temps réel distribuée et semi-structurée.
- Voldemort est une base de données distribuée orientée Big Data.

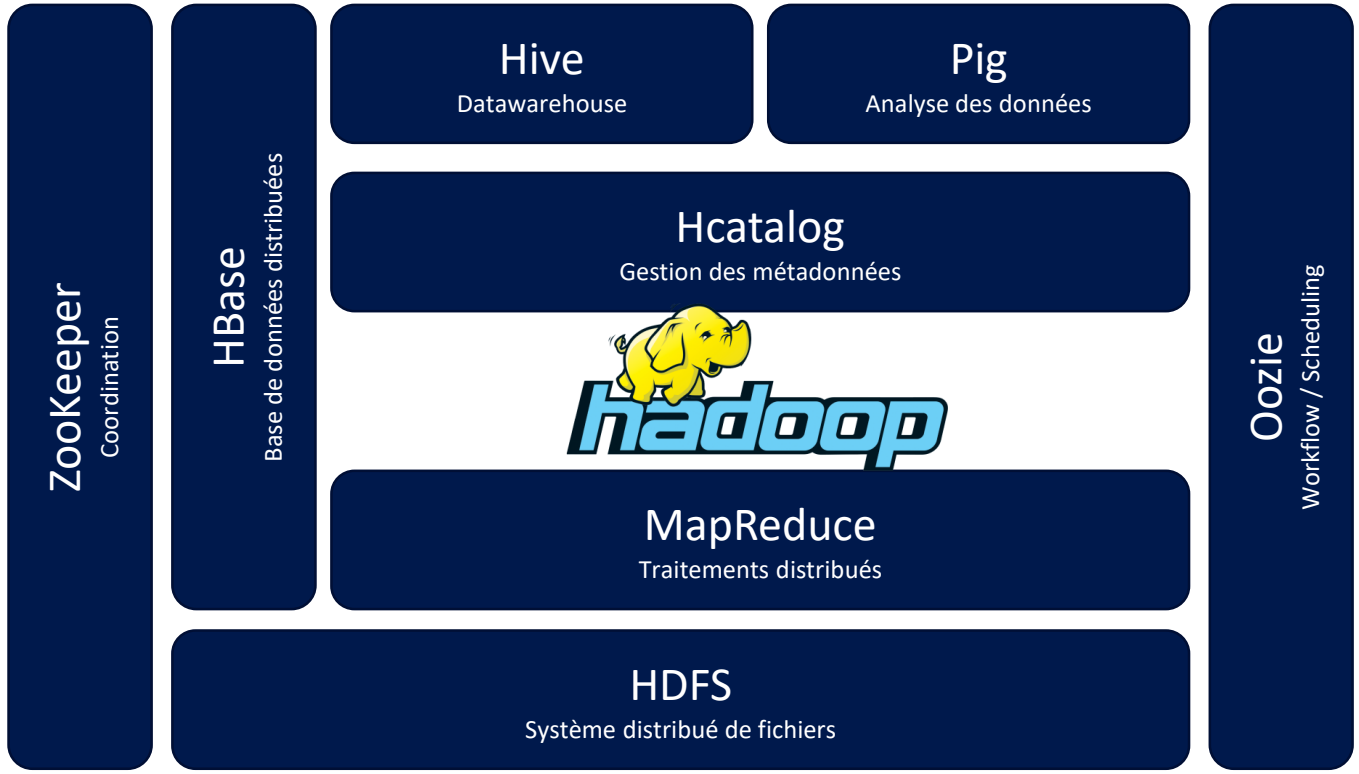
Hadoop et son écosystème

L’écosystème Hadoop

- Ce qu’on retrouve dans Hadoop :
  - Des composants de stockage
  - Des composants de répartition des données
  - Des composants de traitement distribués
  - L’entrepôt de données
  - Le workflow
  - La programmation

Hadoop et son écosystème

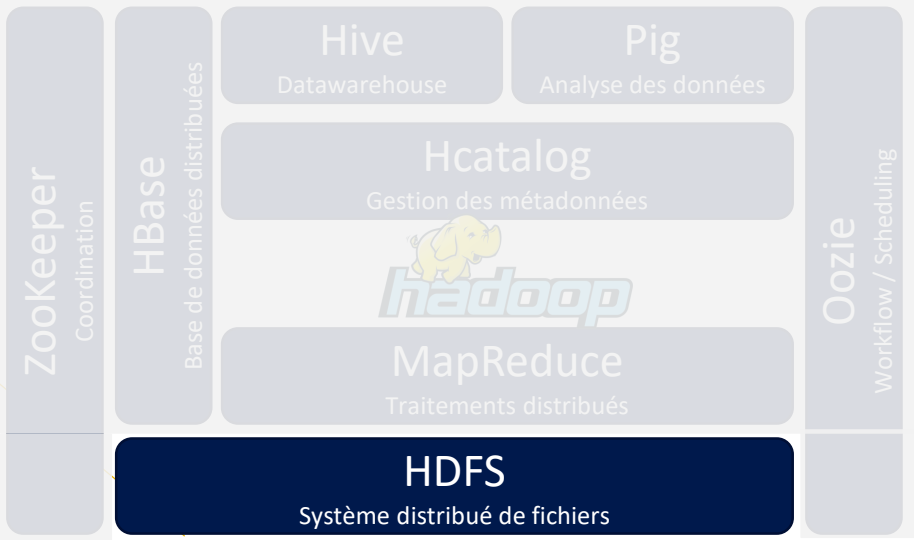
L'écosystème Hadoop



Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop



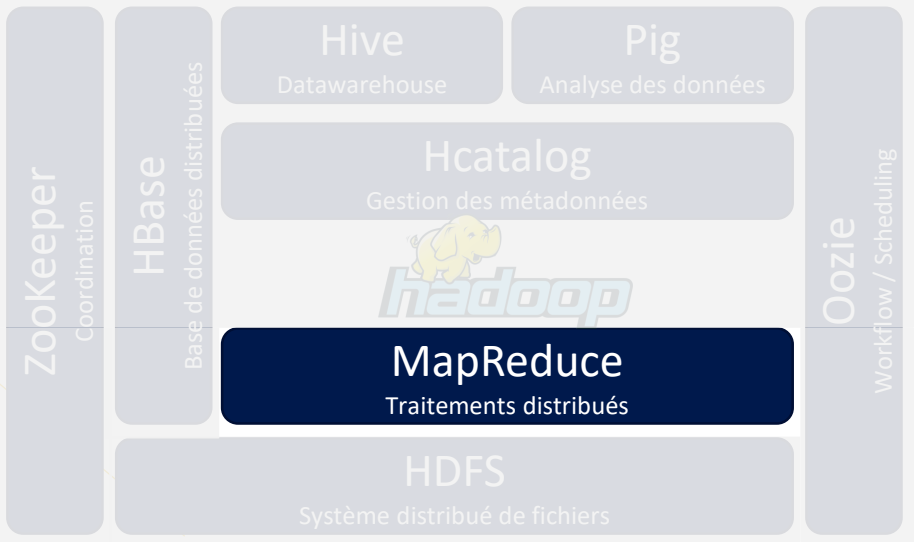
Hadoop Distributed File System :

- Elément central d'Hadoop.
- C'est le système de fichiers distribué qui permet de stocker et répliquer les données sur plusieurs serveurs.

Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop

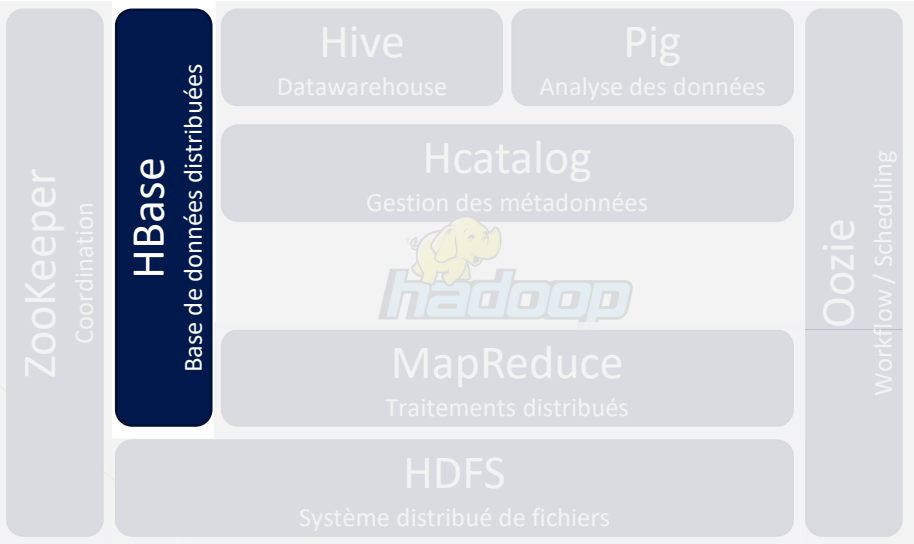


- MapReduce est une plate-forme de programmation qui exécute des algorithmes pour décomposer des données en datasets plus petits.
- MapReduce s'appuie sur deux fonctions : Map() et Reduce(), qui analysent les données rapidement et efficacement.
- La fonction Map() regroupe, filtre et trie plusieurs datasets en parallèle et génère des tuples (paires clés valeurs).
- La fonction Reduce() agrège ensuite les données de ces tuples pour produire le résultat souhaité.

Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop



- Système de gestion de bases de données non relationnelle, distribuée et orientée colonnes.
- Exemple de base de données classique :

EmpId	Nom	Prénom	Salaire
1	Durant	Jacques	40000
2	Dupont	Marie	50000
3	Martin	Jeanne	44000

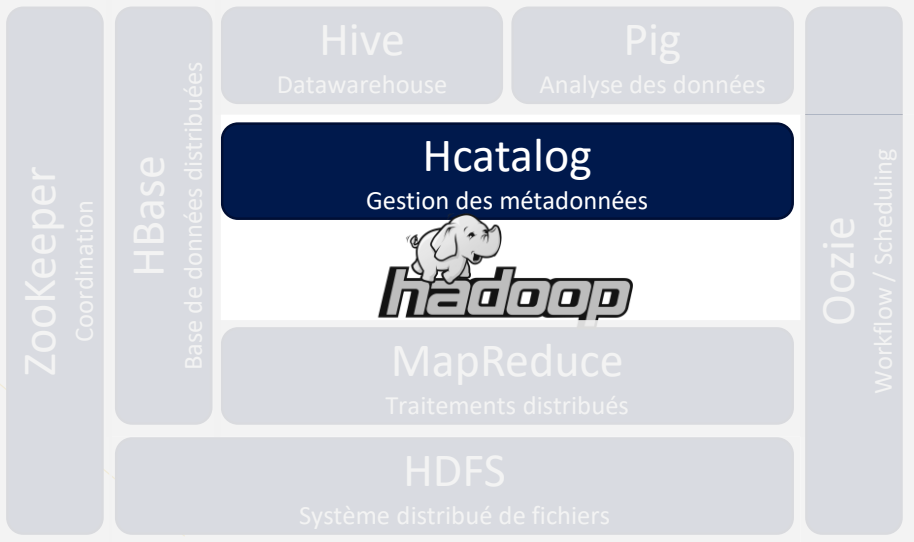
- Une base de données orientée colonne sérialise les valeurs d'une colonne ensemble, puis les valeurs de la colonne suivante :

```
1,2,3;Durant,Dupont,Martin;Jacques,Marie,Jeanne;40000,50000,44000;
```


Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop

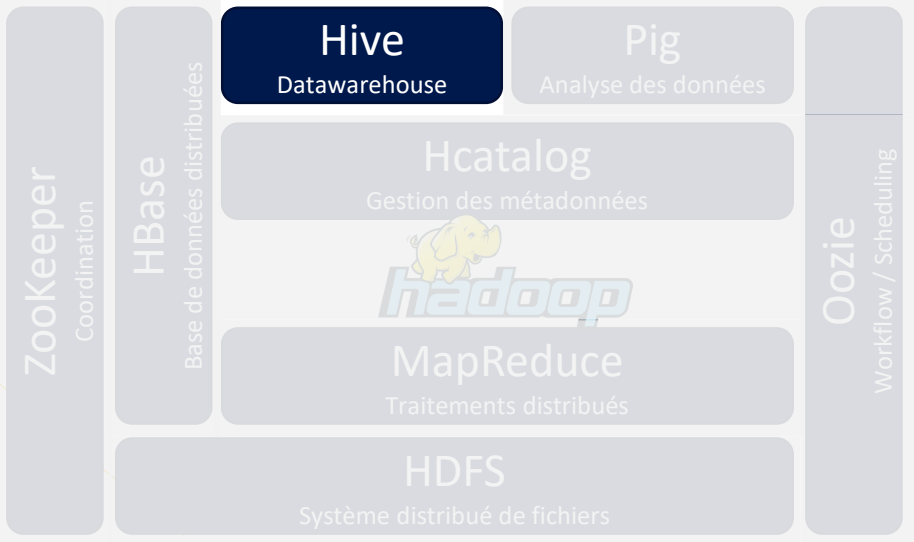


- Outil qui permet d'accéder aux données HDFS via des schémas de type tables de données.
- HCatalog a un client de ligne de commande et une interface REST qui permettent de créer des tables ou d'effectuer d'autres opérations

Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop

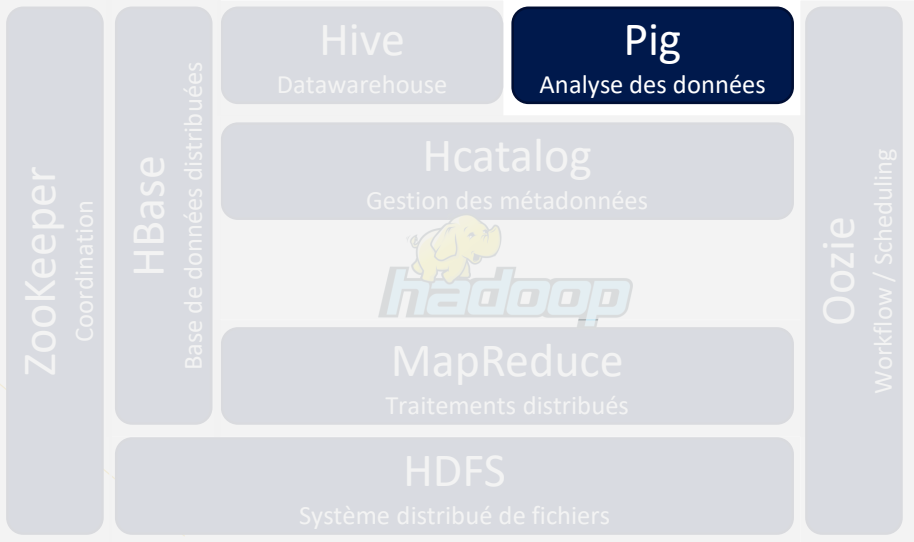


- Système d'entrepôt des données qui facilite l'agrégation et l'analyse des données.
- Il dispose d'un langage similaire au SQL appelé HiveQL.

Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop

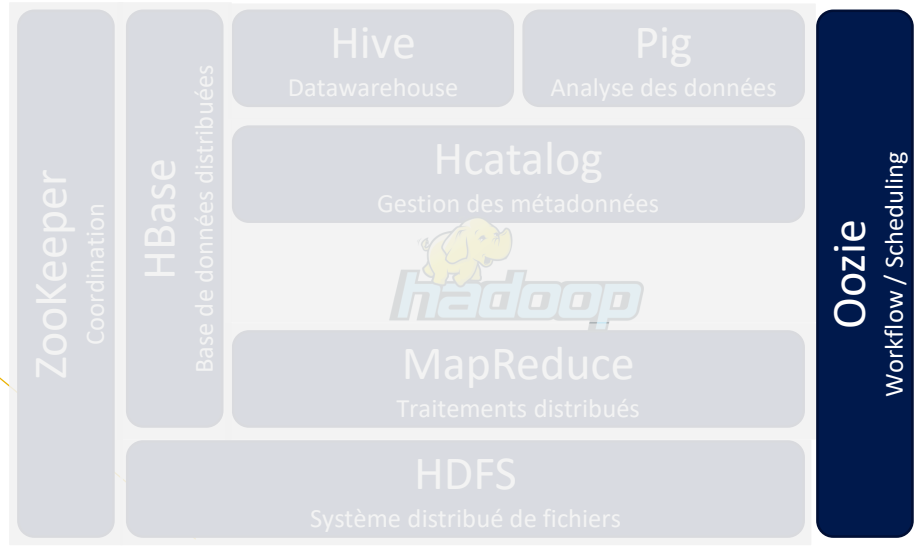


- Plate-forme d'analyse de données à forte volumétrie.
- Pig gère le calcul parallèle des analyses.
- Son langage est Pig Latin.

Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop

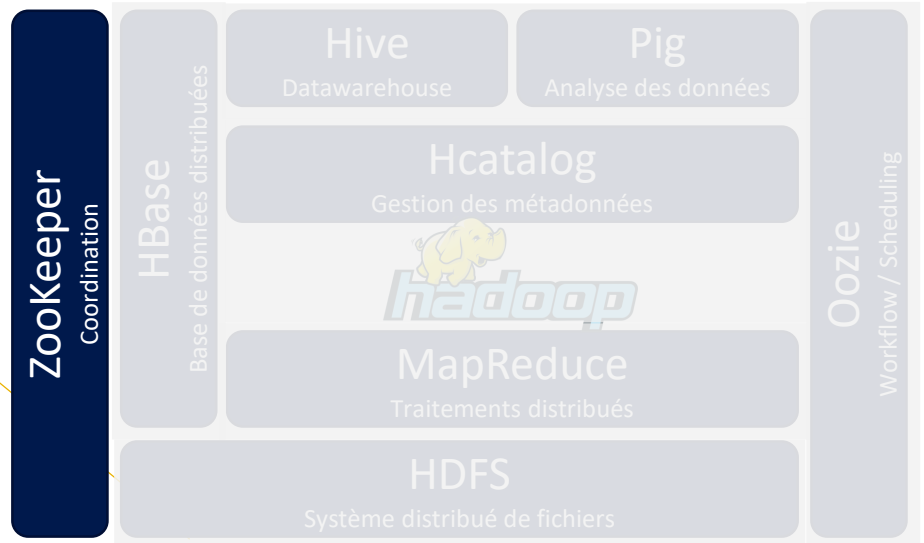


- Outil de workflow servant à coordonner les différents traitements.
- L'utilisateur peut définir la suite d'exécution de plusieurs jobs, voire créer des dépendances entre les jobs.

Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Hadoop et son écosystème

L'écosystème Hadoop



- ZooKeeper un service qui coordonne les applications distribuées. C'est un véritable outil d'administration.
- Les applications distribuées utilisent ZooKeeper pour stocker et arbitrer les mises à jour importantes concernant les informations de configuration.
- Apache ZooKeeper est écrit en Java.

Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

Les métiers

Les fonctions concernées par le Big Data : les Dirigeants

- Premiers concernés. Ils doivent être à l'initiative des projets Big Data.
- Doivent avoir une sensibilité sur ces sujets qui ont un impact sur la stratégie, les prises de décisions et le pilotage du business.
- Disposer de données volumineuses, fiables et exploitées permet d'être plus réactif et plus compétitif sur le marché.

Les métiers

Les fonctions concernées par le Big Data : les Managers

- Grâce aux données exploitées, ils remontent des informations fiables à la Direction Générale (connaissance client, concurrents, fournisseurs…).

- Généralement, ils motivent la poursuite ou l'arrêt des projets Big Data selon que les retours des POC soient pertinents ou non.

Les métiers

Les fonctions concernées par le Big Data : les opérationnels

- Les collaborateurs opérationnels utilisent la donnée au quotidien. Ils sont demandeurs des solutions Big Data.
- Cette donnée se doit d'être fraîche, pertinente, de bonne de qualité et documentée pour lui donner un sens métier.
- Un gestionnaire de stocks doit absolument éviter toute rupture de stock ou de commandes superflues.
- Un technicien de maintenance doit avoir les bonnes informations sur les opérations de maintenances qui ont été réalisées sur l'ensemble de ses machines.
- Une assistance commerciale doit être capable de confirmer une commande dans le bon timing et d'indiquer une date de livraison fiable à son client…

Les métiers

Les fonctions concernées par le Big Data : les responsables informatiques

- Basés au sein de la DSI, ils sont responsables de :
  - La qualité des données stockées dans le SI
  - La prise en compte effective d'une forte volumétrie de données
  - Déployer des technologies nouvelles en phase avec les besoins et les objectifs de l'entreprise

Les métiers

Les métiers du Big Data : Business Analyst

- Moins technique qu'un Data Analyst.
- Le Business Analyst est un spécialiste dans son domaine, il maitrise le secteur d'activité de son entreprise.
- Il fait très souvent le pont entre les équipes informatiques et les dirigeants de l'entreprise.

Les métiers

Les métiers du Big Data : Data Analyst

- Analyse les données, trouve des « insights » à partir des informations brutes stockées dans les bases de données, dans le but de résoudre les problèmes de son organisation.
- Sa mission consiste aussi à créer des rapports afin de partager ses découvertes avec les dirigeants et les autres équipes de l'entreprise.
- L'analyste de données doit posséder des compétences en analyse, en statistiques, en algorithmes et en visualisation de données.
- Communique efficacement les résultats d'analyse.

Les métiers

Les métiers du Big Data : Data Scientist

- Comprend les problématiques business.
- Collecte les données structurées ou non.
- Analyse les données de son organisation.
- Il crée des modèles statistiques avancés de type Machine Learning.
- Dans certaines entreprises, il met en production ses modèles.
- Communique efficacement les résultats d’analyse.

Les métiers

Les métiers du Big Data : Architecte Big Data

- Maîtrise toute l'infrastructure des bases de données.
- Conçoit de nouveaux prototypes de bases de données pour répondre aux besoins de l'entreprise.
- Grâce à sa connaissance des langages informatiques du Big Data, il organise et maintient les données dans des bases de données relationnelles et des répertoires d'entreprise. Il développe des stratégies d'architectures de données pour chaque zone d'activité de l'organisation.

Les métiers

Les métiers du Big Data : Administrateur de bases de données

- Assure la stabilité et la disponibilité des bases de données dans l'entreprise.
- Responsable de l'alimentation des bases et des éventuelles migrations de données.
- Veille aux mises à jour, aux sauvegardes des bases et à leurs restauration.
- Garant de la sécurité des données dans le SI.

Les distributions

Hortonworks

- HDP pour Hortonworks Data Platform, est une plateforme de gestion et de traitement analytique de données massives.

- La solution couvre l'ensemble des besoins d'entreprise, de l'acquisition, au stockage de masse, jusqu'au traitement de données, sans contraintes de formats ou de structures de ces données.

Les distributions

Hortonworks : composition de HDP

1. Cœur Hadoop (HDFS/MapReduce).
2. NoSQL (Apache HBase).
3. Méta-données (Apache HCatalog).
4. Plateforme de script (Apache Pig).
5. Requêtage (Apache Hive).
6. Planification(Apache Oozie).
7. Coordination (Apache Zookeeper).
8. Gestion et supervision (Apache Ambari).
9. Services d'intégration (HCatalog APIs, WebHDFS, Talend Open Studio for Big Data, Apache Sqoop).
10. Gestion distribuée des logs (Apache Flume).
11. Apprentissage (Apache Mahout).

Les distributions

Cloudera : composition de CDH (Cloudera Distribution for Hadoop)

cloudera

- CDH est une distribution Open Source de Hadoop et de composants complémentaires préparé par Cloudera.

1. Flume : Exploitation de fichiers (log) dans Hadoop.
2. HBase : Base de données NoSQL (accès read/write aléatoires).
3. Hive : Requêtage de type SQL.
4. Hue : SDK permettant de développer des interfaces utilisateur pour les applications Hadoop.
5. Impala : entrepôt de données natif d'Hadoop pour les requêtes.
6. Kafka : traitement distribué en temps réel des données Hadoop…

Les distributions

MapR

- MapR est une distribution complète pour Apache Hadoop qui regroupe plus d'une douzaine de projets de l'écosystème Hadoop afin de fournir un large éventail de capacités pour les mégadonnées.

1. HBase,
2. Pig,
3. Hive,
4. Mahout : algorithmes d'apprentissage automatique distribués
5. Cascading : crée et exécute des workflows,
6. Sqoop : transfère des données entre des bases de données relationnelles et Hadoop
7. Flume : collecte et à analyse de fichiers de log

Partie 2

Enjeux et perspectives d’avenir du Big Data

Plan du module

1. La qualité des données
2. Open Data
3. Traiter les données
4. La sécurité des données
5. L’image de la donnée

La qualité des données

1. La complétude

- La complétude sert à mesurer l’exhaustivité des données.
- L'analyse de complétude permet d'identifier les enregistrements dont les valeurs de données de la colonne n'ont pas de signification métier.
- Il est important de connaître le pourcentage de "données manquantes" dans une colonne.
- L’exhaustivité peut faire référence aux valeurs disponibles par enregistrement (par exemple, 80% des champs obligatoires sont remplis sur un enregistrement).

La qualité des données

2. L’unicité

- Chaque enregistrement doit être unique dans la base de données.

- Une vérification de la présence de doublons est primordiale :
  - Dans mon fichier client, deux entreprises ne peuvent pas posséder le même code SIRET.
  - Dans mon fichier client, mon client ne doit pas apparaître plus d’une fois.

La qualité des données

3. L’actualité

- Les données doivent être fraîches et refléter le monde réel.

- Les valeurs des capteurs de températures enregistrées 2 heures plus tôt ne valent peut-être plus rien à l’instant t.

La qualité des données

4. La validité

- Les données sont valides si elles respectent la syntaxe de leur définition.

- Une adresse e-mail valide a une syntaxe de type :
  - $partielocale@domaine.fr$ ➔ $jean.marc@yahoo.fr$

La qualité des données

5. La précision

- La précision concerne la question de savoir si les données décrivent correctement le monde réel.

- Si vos données de prospect indiquent que Mike Wazowski est un « Scare Executive » chez **Monsters Inc.**, mais qu'il est en fait « Chief Scare Officer » à **Monsters University**, vos données pour les champs « $role$ » et « $company$ » ne sont pas exactes.

- Les données sont valides si elles respectent la syntaxe de leur définition.

La qualité des données

6. La cohérence

- La cohérence décrit à quel point d'autres représentations du monde réel sont identiques dans des bases de données distinctes.

- Si le CRM indique que le client Mike Wazowski est né en 1981 mais que le système de commerce électronique indique 1980, alors les données ne sont pas cohérentes.

La qualité des données

Quelques exemples courants : les doublons

- Un même enregistrement apparaît plusieurs fois en base de données.
- En cas de mise à jour, un enregistrement peut être modifié et pas les autres.
- L'usage d'un enregistrement incorrect peut engendrer divers problèmes :
  - Mauvaise adresse de livraison, mauvaise facturation…
  - Un poids supplémentaire en base de données, données incohérentes…
  - Des analyses de données incohérentes, non pertinentes, contradictoires…

La qualité des données

Quelques exemples courants : les coordonnées des clients

- Les coordonnées doivent être mises à jour régulièrement :
  - Adresse postale, adresse email, téléphone…
- S’adresser au mauvais interlocuteur peut retarder les process de négociation ou de livraison / commande.
- Un message publicitaire adressé au mauvais prospect / client :
  - On manque sa cible, on communique sur le mauvais produit / service
  - On créé un désintérêt (spaming)

La qualité des données

Quelques exemples courants : les montants facturés

- Cas de sous-facturation :
  - Perte de chiffre d'affaires
- Cas de surfacturation :
  - Client mécontent, plaintes, résiliation
- Un compteur électrique communiquant défaillant envoie des relevés erronés. La facture du client peut être sous-estimée ou surestimée.

La qualité des données

Quelques exemples courants : la gestion des stocks

- Il doit y avoir une cohérence entre les flux sortants et entrants.

- Un système d'information défaillant peut causer :
  - Des ruptures de stocks.
  - Commander / fabriquer à tort un produit.
  - Des coûts inutiles de fabrications ou de commandes pour l'entreprise.

La qualité des données

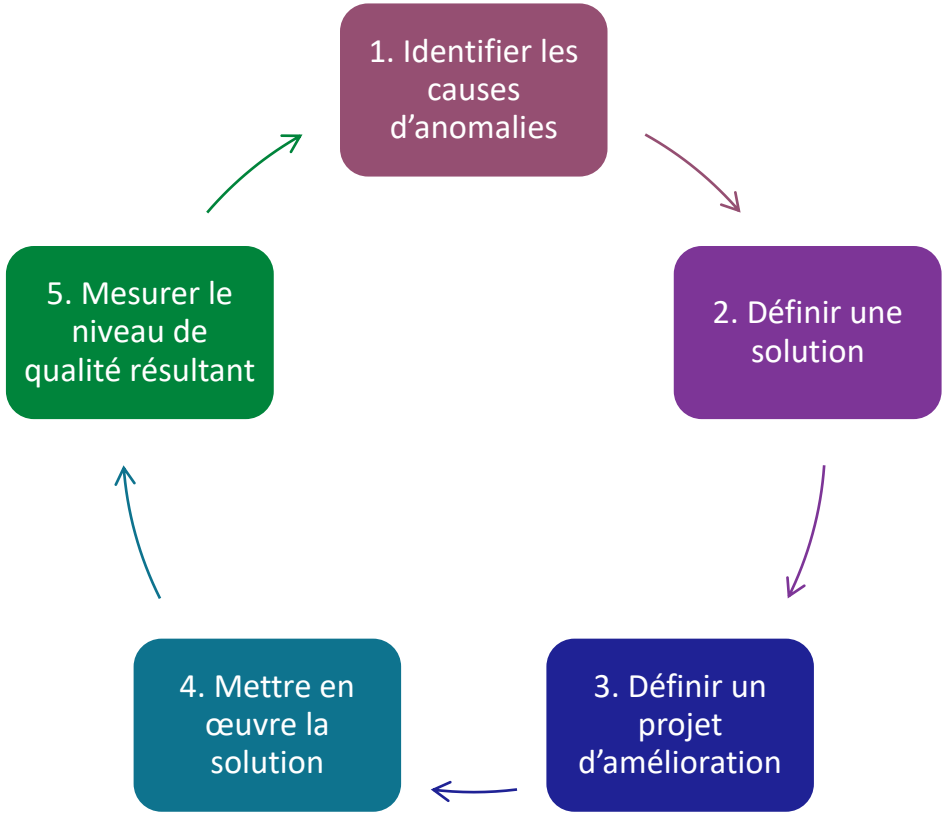
Des pistes pour améliorer la qualité de ses données

- Réaliser un audit complet de la qualité en contrôlant l'existant :
  - Identifier les sources d'erreurs
  - Mettre à jour et nettoyer les données
  - S'assurer de l'exactitude des données futures.

- Réaliser des rapports réguliers sur la qualité des données

La qualité des données

Des pistes pour améliorer la qualité de ses données



Open Data

Définition

- Selon la CNIL* :
  - « L'*open data* désigne un mouvement, né en Grande-Bretagne et aux États-Unis, d'ouverture et de mise à disposition des données produites et collectées par les services publics (administrations, collectivités locales...) ».
  - « Les administrations peuvent permettre au public de consulter en ligne **une partie de leur base de données** : il s'agit de l'open data (base de données ouverte) ».

* Commission Nationale de l'Informatique et des Libertés

Open Data

Cas d’usage : parking intelligent

- Le stationnement est au cœur de la « smart city » :
  - Les villes ont de plus en plus recours à l’usage de l’Open Data pour optimiser l’offre de stationnement.
  - En France, on estime que 15 à 30% de la circulation en ville est due à la recherche de places, selon opendatafrance.net
  - Les données sont fournies par des capteurs. On y trouves les informations sur la localisation des parkings, les zonages tarifaires, l’emplacement des places PMR et des horodateurs.

Open Data

Cas d’usage : guidage géolocalisé

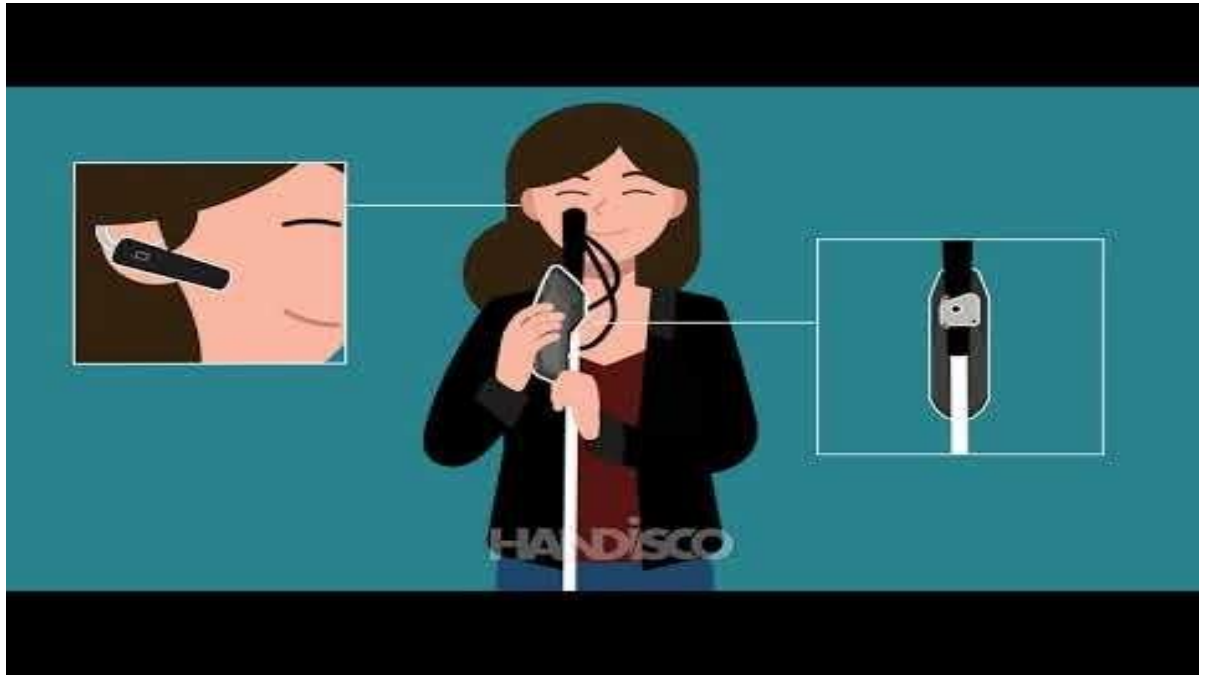
- Solution de guidage géolocalisé qui permet à partir d'un boîtier équipé d'un GPS qui s'installe sur une canne blanche de faciliter la mobilité quotidienne des malvoyants en environnement urbain.

- Développé par Handisco, **Sherpa** est un assistant intelligent pour personnes déficientes visuelles qui améliore les déplacements quotidiens.

- Elles peuvent se laisser guider jusqu’à une destination inconnue, prendre le bus en toute autonomie ou encore dialoguer avec les carrefours de leur ville. Pour améliorer l’accès aux transports en commun, **Sherpa** est en mesure de renseigner l’utilisateur sur les horaires et itinéraires des bus, tramways et métros. Pour ce faire, il exploite les données ouvertes fournies par les opérateurs locaux de transport.

Open Data

Cas d’usage : guidage géolocalisé



Traiter les données

Les différents types de données : le texte

- Premier format utilisé au début de l'ère informatique.
- En 1963, le code ASCII* fut créé pour associer chaque caractère à une suite de 0 et de 1. Ce code ne prenait pas en compte le caractère accentué.
- La norme Unicode (très récente) permet de gérer les caractères de toutes les langues.
- Les données des fichiers clients (nom, adresse…) sont donc en réalité encodés par une suite de 0 et de 1.

* American Standard Code for Information Interchange

Traiter les données

Les différents types de données : le dessin vectoriel

- Objet géométrique dont la forme et la position sont déterminés par des formules mathématiques.

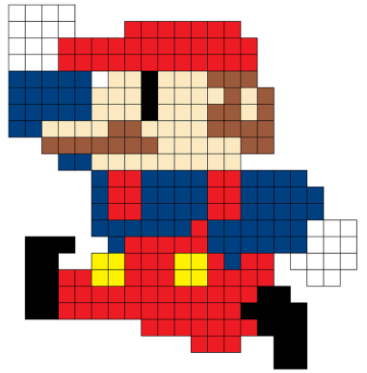
- On retrouve ce type de données dans l'industrie automobile, la construction (bâtiments, ponts) et dans le domaine des SIG (cartes vectorielles)…

Traiter les données

Les différents types de données : l'image matricielle

- C'est la représentation de l'image numérique telle que nous la connaissons de nos jours.

- L'image matricielle est une succession de pixels dont les nuances de couleurs impactent la qualité de l'image.



Traiter les données

Les différents types de données : la vidéo

- Succession d’images matricielles qui défilent de façon à donner une impression de continuité (25 images par secondes).
- Les formats de vidéos varient. Ils vont de la plus faible qualité à la plus haute.
- La vidéo est une donnée non structurée.

Traiter les données

Les différents types de données : le son

- Succession d'ondes générés par la vibration de l'air.
- Le son a deux caractéristiques :
  - La fréquence : nombre d'oscillations par seconde (Hertz)
  - L'intensité : force du son (décibels)
- L'ordinateur a donc besoin de convertir le son en valeurs numériques : on parle de numérisation.
- MP3 est un des formats de compression du son.

Traiter les données

Définition

- Selon la CNIL :
  - « Un traitement de données personnelles est une opération, ou ensemble d'opérations, portant sur des données personnelles, quel que soit le procédé utilisé ».
  - « Un traitement de données personnelles n'est pas nécessairement informatisé : les fichiers papier sont également concernés et doivent être protégés dans les mêmes conditions ».

Traiter les données

Définition

- Le traitement des données fait référence aux procédés de : collecte, enregistrement, organisation, conservation, modification, extraction, consultation, utilisation, communication...

  - Consultation de données de contacts comprenant des données personnelles
  - Campagne d'e-mails promotionnels
  - Conservation d'adresse IP
  - Conservation d'enregistrements de vidéosurveillance
  - Publication d'une photo d'une personne sur internet

- La collecte des données doit donc être limitée aux seuls besoins de l'entreprise de façon à remplir certains objectifs précis et licites : création d'un fichier clients, gestion des stocks…

La sécurité des données

Sécuriser son S.I.

- Face aux risques de fuites et de failles de sécurité, les organismes concernés par le RGPD ont pour obligation d'adapter leur système de protection des données.

- En effet, la collecte et la conservation des données peuvent représenter un danger pour la confidentialité et le respect de la vie privée.

- Afin de garantir la sécurité des utilisateurs, il convient donc de renforcer son système de sécurité.

La sécurité des données

Principes à respecter

- Pour remplir cette obligation de sécurité, les organismes concernés se doivent de respecter certains principes :
  - Transparence de traitement
  - Information claire et précise
  - Réduction de la quantité de données collectées
  - Limite de la durée de conservation
  - Respect des droits des utilisateurs (consultation, portabilité, droit à l'oubli)

La sécurité des données

Sanctions en cas d'infraction

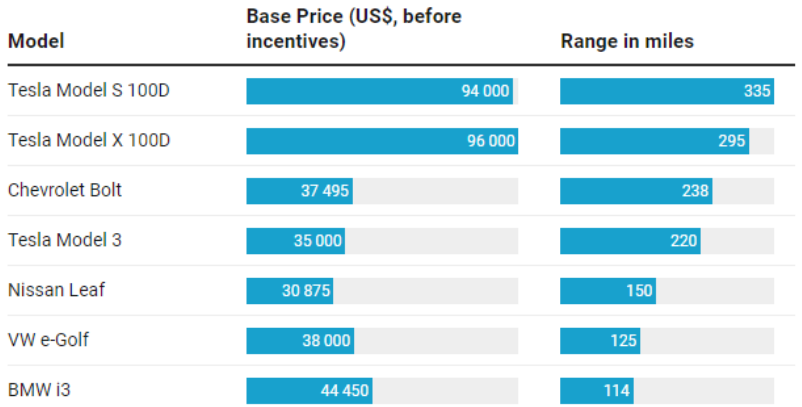
- En France, la CNIL est en charge de vérifier la bonne application du règlement européen et d'émettre des sanctions en cas d'infraction.
- Au maximum les sanctions peuvent aller jusqu'à 20 millions d'euros ou bien une somme égale à 4% du chiffre d'affaires annuel mondial de l'entreprise.
- En plus de l'amende, la sanction émise va avoir des effets sur l'image de l'organisme en question.

L’image de la donnée

Les techniques de visualisation

- **Objectif :**
  - L’objectif est de présenter l’information de façon claire, visuelle et pédagogique.
  - Un graphique est plus explicite qu’un tableau présentant la même information.

Model	Base Price	Range in miles
Tesla Model S 100D	94000	335
Tesla Model X 100D	96000	295
Chevrolet Bolt	37495	238
Tesla Model 3	35000	220
Nissan Leaf	30875	150
VW e-Golf	38000	125
BMW i3	44450	114



L’image de la donnée

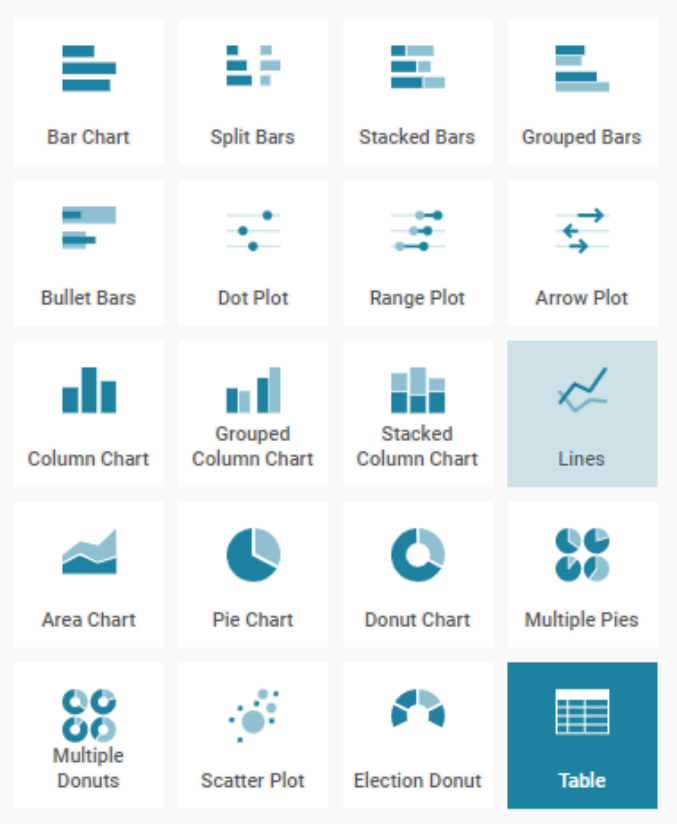
Les techniques de visualisation

- **Principes :**
  - *« L’efficacité d’une visualisation se mesure par la durée nécessaire à sa compréhension »,* (Christophe Brasseur).

  - Eléments à prendre en compte dans la présentation de sa visualisation :
    - La taille (prise en compte des variations de quantités)
    - La forme (ajout de pictogrammes)
    - La couleur (nuances de couleurs pour évoquer l’intensité d’un phénomène)
    - L’orientation (le sens de lecture, les axes du graphique…)

L’image de la donnée

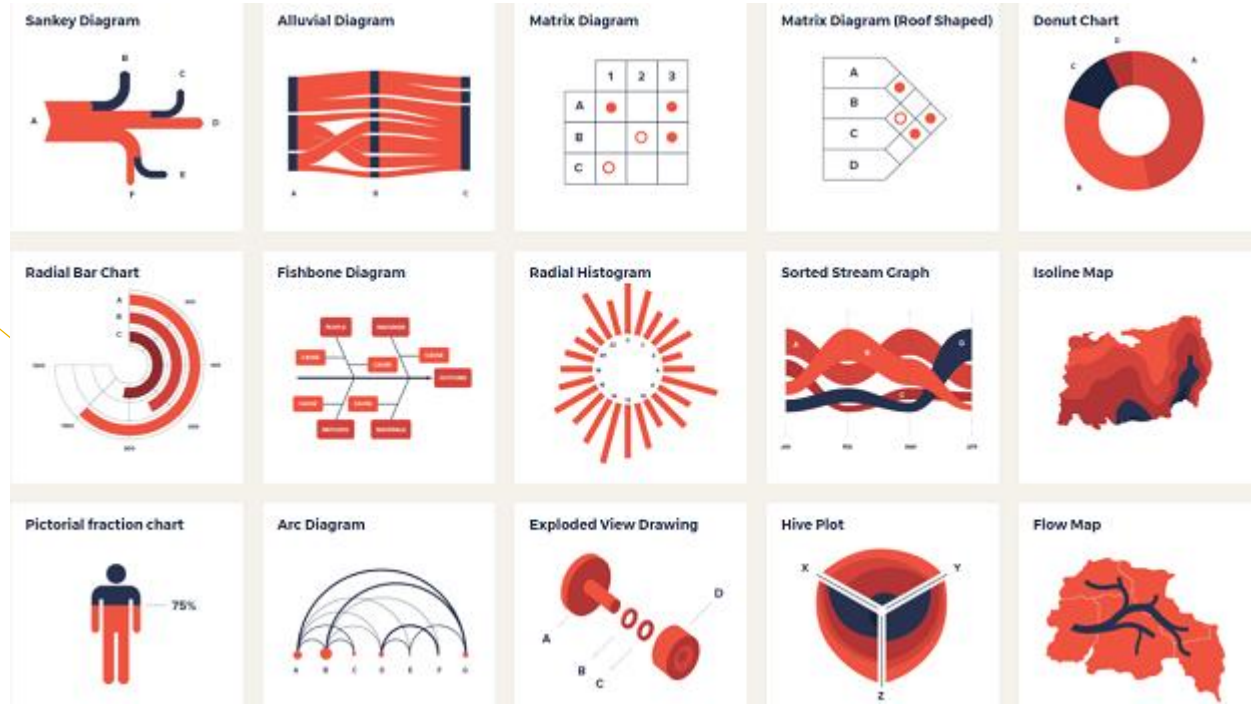
Visualisations classiques



- Histogrammes
- Courbes
- Camembert
- Aires
- Bulles
- Tableaux…

L’image de la donnée

Visualisations avancées

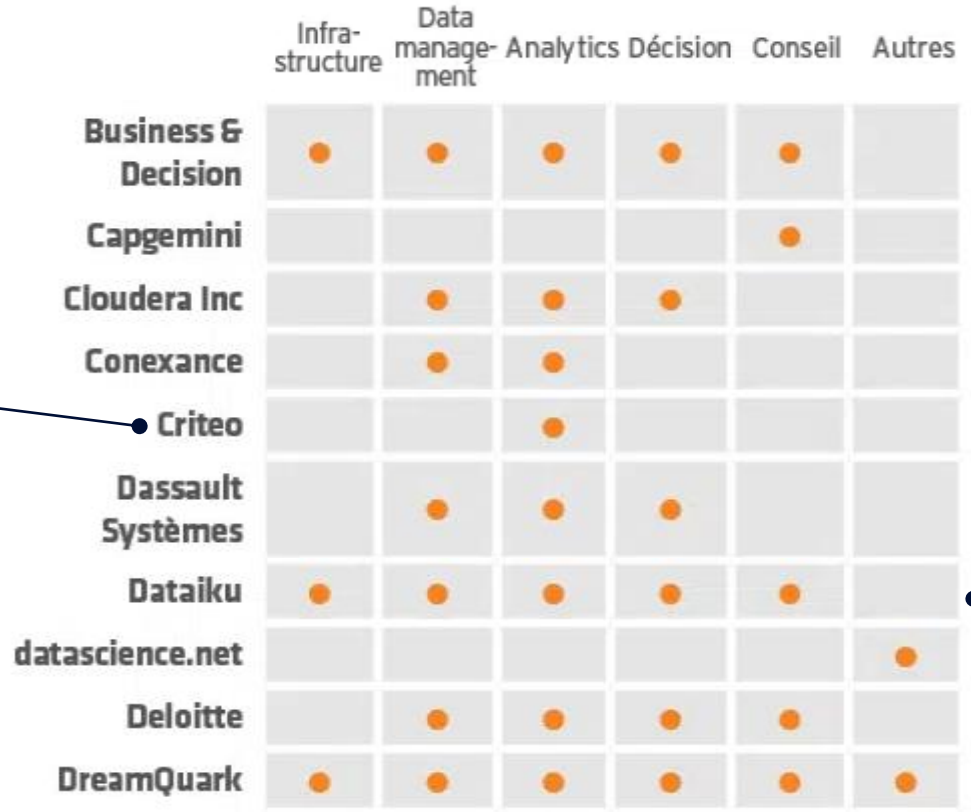


- Sankey
- Matrix diagram
- Radial Bar Chart
- Arc Diagram
- …

TP : Présentation des acteurs du marché

Entreprise française de re-ciblage publicitaire sur internet.

Criteo manipule des volumes de données considérables dans des temps de réponse de quelques dixièmes de seconde afin d'offrir le meilleur service aux annonceurs.



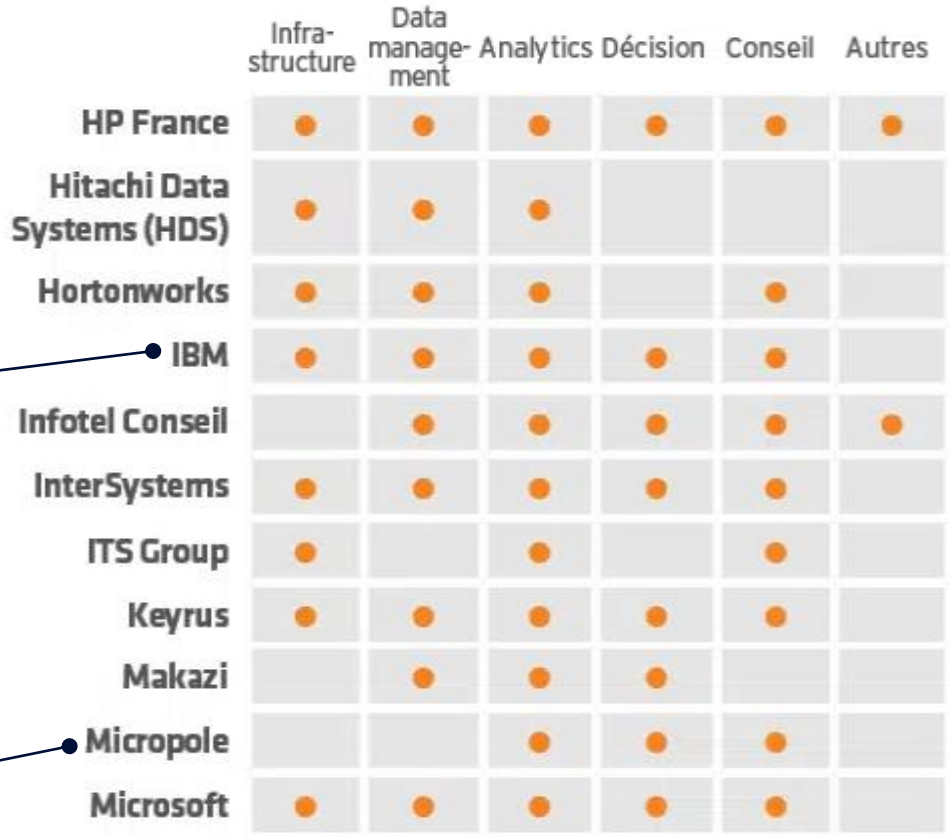
Source : https://www.alliancy.fr/infographie-les-50-acteurs-qui-font-le-big-data

Dataiku développe une plateforme pour analyser la donnée et développer des méthodes prédictives en environnement Big Data.

TP : Présentation des acteurs du marché

Fondé en 1911, pionnier du traitement des données.

Société de Consulting spécialisée en Data et Digital.



	Infra-structure	Data manage-ment	Analytics	Décision	Conseil	Autres
HP France	●	●	●	●	●	●
Hitachi Data Systems (HDS)	●	●	●			
Hortonworks	●	●	●		●	
IBM	●	●	●	●	●	
Infotel Conseil		●	●	●	●	●
InterSystems	●	●	●	●	●	
ITS Group	●		●		●	
Keyrus	●	●	●	●	●	
Makazi		●	●	●		
Micropole			●	●	●	
Microsoft	●	●	●	●	●	

Source : https://www.alliancy.fr/infographie-les-50-acteurs-qui-font-le-big-data

Partie 3

Aspects stratégiques et organisationnels de la donnée

Plan du module

1. Le défi technique de la donnée
2. Les investissements : capacité de stockage, analyses des données
3. Le web sémantique
4. Les défis économiques
5. L'impact sur l'organisation
6. La conduite du changement
7. Les nouveaux métiers
8. TP : présentation des usages du Big Data

3.1 Le défi technique de la donnée

Le défi sur l'intégration

- Le Big Data rassemble des données provenant de nombreuses sources et applications disparates.
- Les mécanismes traditionnels d'intégration de données, tels que ETL (extract, transform, and load) ne sont généralement pas à la hauteur de la tâche.
- Pour analyser des données massives, il est nécessaire d'adopter de nouvelles stratégies et technologies (Datawarehouse, Data Lake…).
- Lors de l'intégration, il faut **importer** les données, les **traiter** et s'assurer qu'elles sont **formatées** et **disponibles** sous une forme **accessible** à vos analystes.

3.1 Le défi technique de la donnée

Le défi sur la gestion

- La solution de stockage peut se trouver dans le cloud, sur site (On premise), ou les deux à la fois.
- Vous pouvez stocker vos données sous la forme de votre choix (tableur, XML…) et imposer à ces jeux de données vos exigences de traitement.
- Nombreux sont ceux qui choisissent leur solution de stockage en fonction de l'endroit où sont hébergées leurs données (Europe, Amérique...).
- Le cloud est de plus en plus adopté, car il prend en charge vos besoins informatiques actuels et laisse la possibilité d'augmenter les ressources en fonction des besoins (scalabilité).

3.1 Le défi technique de la donnée

Le défi sur l'analyse

- **Data Driven Company :** L'investissement dans le Big Data porte ses fruits dès lors que vous êtes en mesure d'analyser vos données et d'agir à partir de l'analyse.

- **La puissance de la Dataviz :** adoptez un nouveau point de vue grâce à une analyse visuelle de vos jeux de données. L'exploration des données offre de nouvelles opportunités de découvertes.

- **La puissance de l'IA :** Les modèles de données créés avec l'intelligence artificielle ajoutent une valeur supplémentaire à la Business Intelligence.

3.2 Les investissements

Capacité de stockage

- Exemple de tarification avec paiement à l'utilisation du stockage de données chez Microsoft Azure :

	Premium	À chaud	À froid	Archive
50 premiers To/mois	**$0,15** par Go	**$0,018** par Go	**$0,01** par Go	**$0,00099** par Go
450 To/mois suivant(s)	**$0,15** par Go	**$0,0173** par Go	**$0,01** par Go	**$0,00099** par Go
Plus de 500 To/mois	**$0,15** par Go	**$0,0166** par Go	**$0,01** par Go	**$0,00099** par Go

$$500\ To = 500\ \times 1000\ Go = 500\ 000\ Go\ \times \$0{,}15 = \$75\ 000\ par\ mois.$$

3.2 Les investissements

Analyse des données

- **Plusieurs solutions possibles :**
  - **Infogérance :** déléguer la gestion des ressources ou l'analyse de données à un cabinet de conseil de type ESN par exemple.
    - Facile à déployer des experts.
    - Très coûteux à terme.
  - **Embaucher :** recruter un analyste de données.
    - Poste permanent, expert dans son domaine, salaire négociable.
    - Recrutement difficile, chronophage, risque de turn-over.
  - **Recrutement en interne :** faire monter en compétences vos collaborateurs (un Business Analyst peut devenir Data Analyst, voire Data Scientist).
    - L'interne a une bonne maîtrise du métier, de l'environnement et des enjeux de l'entreprise.
    - Processus d'apprentissage très long, formations à financer.

3.3 Le web sémantique

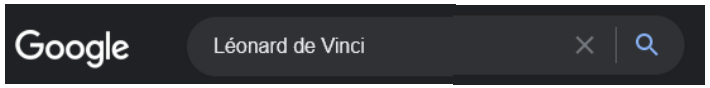
Définition

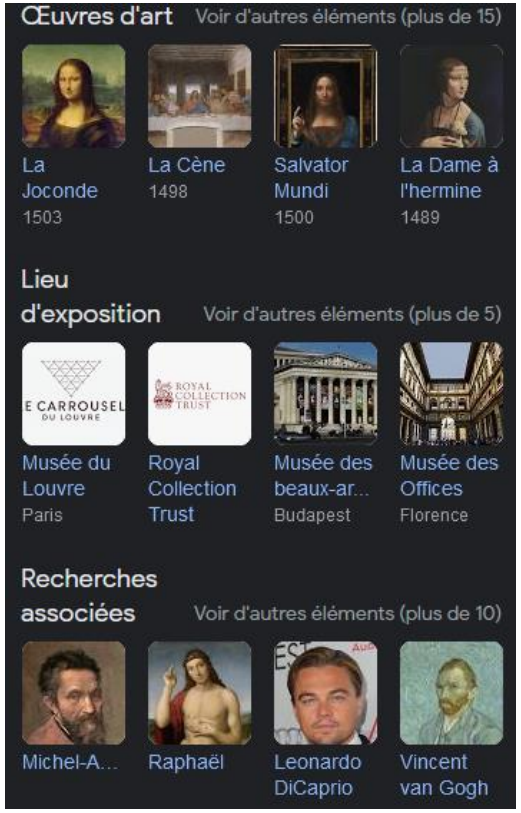
- Le web sémantique est une extension du web standardisé.
- Le web standardisé, tel que nous le connaissons, est composé de 3 technologies :
  - Le langage HTML pour représenter un document et formaliser des liens
  - Le protocole HTTP pour échanger un document
  - Le mécanisme URI pour identifier une ressource (page web, image, vidéo, son…).
- Le web sémantique a vocation à donner du sens au contenu des pages web, par l'interprétation des machines.
- Chaque ressource (page web, image, vidéo…) est associée à une métadonnée, qui va donner du sens à la ressource.

3.3 Le web sémantique

Quelques cas d’usage du web sémantique

- **Google :** qualifie les données en rendant plus efficace et plus pertinente ses réponses (indexation des pages web).
  - La recherche de Léonard de Vinci dans Google, fait apparaître les résultats les plus pertinents sur : le personnage lui-même, les lieux d’expositions, les personnages contemporains ou liés à son nom.

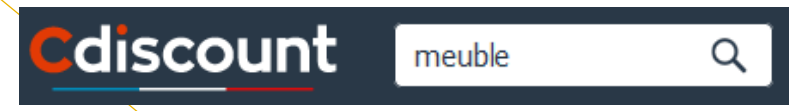




3.3 Le web sémantique

Quelques cas d’usage du web sémantique

- **E-commerce :** je recherche un meuble. Si le site est bien conçu, il doit être capable de me proposer tout type de meubles, des tables, des chaises, des bureaux, et tout autre mobilier en rapport avec la recherche. Même si ces termes n'ont pas été explicitement mentionnés.



3.4 Les défis économiques

Maîtriser les coûts

- Les bases de données traditionnelles ne peuvent supporter la montée en charge liée aux données massives.
- Le choix de la bonne technologie, i.e. celle qui est le plus adaptée à l'entreprise est essentiel.
- Quels sont dons les types de déploiements à envisager ?

3.4 Les défis économiques

Modèle de déploiement On Premise vs Cloud

- Les services et équipements du fournisseur (Microsoft par exemple) sont installés au sein de l'entreprise avec pour contre partie, le paiement d'une licence généralement renouvelée à l'année.
- L'exploitation et les coûts liés à l'usage du service sont à la charge de l'entreprise : fonctionnement du service, mise en place des processus, ressources humaines…
- On a généralement une licence par utilisateur unique. Si l'entreprise dispose donc de 500 collaborateurs, il faudra acheter 500 licences par an.
- Certaines entreprises se retrouvent avec un parc informatique difficilement gérable.

3.4 Les défis économiques

Modèle de déploiement On Premise vs Cloud

- Hadoop peut être déployée en distribution libre ou commerciale.
- Dans le cas d'une distribution libre On Premise, il suffit d'installer une distribution open source d'Hadoop gratuitement sur les serveurs de l'entreprise. Tous les collaborateurs pourront l'utiliser sans coûts financiers supplémentaires (pas de licence).
- Problème, il est difficile de suivre rythme des évolutions des versions de la communauté et de gérer les problèmes d'incompatibilités.
- En entreprise, la politique de gouvernance IT exige une pleine maîtrise de la technologie et une feuille de route claire.

3.4 Les défis économiques

Modèle de déploiement On Premise vs Cloud

- Dans le cas d'une distribution commerciale On Premise, l'éditeur va céder les droits d'exploitation de sa solution Big Data (HortonWorks par exemple).

- **Avantages :** Le logiciel livré par l'éditeur est fiable et l'éditeur vous apporte du support.

- **Point faible :** le déploiement est On Premise. Si les serveurs tombent ou sont hackés, il faudra réagir très rapidement.

3.4 Les défis économiques

Modèle de déploiement On Premise vs Cloud

- La ressource informatique est hébergé par l'éditeur et offert à l'entreprise sous forme de service Cloud.
- Ce service cloud est délivré aux clients sous forme de service Web, leur accès est contrôlé par une API et leur exploitation fait l'objet d'une facturation en fonction du volume de ressources utilisées (nombre de requêtes par minute, nombre de connexions par seconde, le nombre de commandes enregistrées par jour…).
- L'éditeur reste propriétaire de son produit, l'héberge et l'administre.
- En On Premise, les licences sont nominatives et tarifiées sur une durée fixe. Le modèle Cloud transforme les coûts IT d'une charge fixe à une charge variable.

3.5 L’impact sur l’organisation

Quelques impacts sur l’organisation

- **Les nouveaux métiers :** pour tirer parti du Big Data, de nouveaux profils sont recrutés pour collecter, analyser et visualiser les données. La stratégie de recrutement des entreprises changent.

- **Organisation repensée :** collaboration et partage de l’information entre équipes. Les équipes marketing travaillent conjointement avec les équipes Data Science.

- **Data / Customer centrics :** le Big Data a permis de repenser la relation avec les clients et les partenaires. La prise de décisions se fait de plus en plus avec l’aide des données.

3.6 La conduite du changement

Enjeux

- Le déploiement une solution Big Data n'est (totalement) réussi que lorsque tous les collaborateurs l'ont adopté.

- Il est conseillé d'anticiper le changement avant la mise en place de la solution Big Data, et non pas après l'avoir délivrée :
  - Intégrer les collaborateurs : recueil des besoins métiers, ateliers
  - Prendre en compte la dimension psychologique : tous les collaborateurs n'ont pas la même sensibilité au changement (peur, acceptation…)

3.6 La conduite du changement

Les 3 étapes clefs : la préparation, le changement, la pérennisation

- La préparation implique l'anticipation de l'acceptabilité du projet de la part des collaborateurs (préparer les esprits, expliquer la nécessité du changement…).Il faut donc :
  - Les équipes doivent adhérer au projet.
  - Rassurer les plus craintifs.

- Montrez les limites de la solution actuelle, et vantez la nouvelle solution (entendre l'avis des collaborateurs également).

3.6 La conduite du changement

Les 3 étapes clefs : la préparation, le changement, la pérennisation

- Le changement n'est possible qu'après avoir embarqué les collaborateurs et avoir levé la résistance.

- Invitez les collaborateurs à participer (être acteurs et non spectateurs) au changement :
  - Définition des objectifs
  - Qualification des besoins métiers
  - Choix de la solution Big Data
  - Déploiement au sein de l'organisation

3.6 La conduite du changement

Les 3 étapes clefs : la préparation, le changement, la pérennisation

- Une fois la solution déployée :
  - Mettre en place des ambassadeurs qui assureront la promotion et la bonne prise en main de l'outil.
  - Mettre en place des cas d'usages pertinents parlant aux équipes.
  - Mettre à disposition de la documentation technique.
  - Mettre en place un dispositif de formation (ateliers, webinaires, e-learning…).
  - Mesurer le changement (niveau d'adoption du nouvel outil versus l'ancien outil).

3.7 Les nouveaux métiers

Chief Data Officer (Directeur des données)

- **Missions :** Mettre en place l'environnement Big Data, Choisir les données à analyser, assurer la qualité et la cohérence des données, développer une stratégie Data Driven…

- **Compétences :** Outils analytiques, bases de données, mathématiques, communication, leadership, connaissance de l'entreprise et de son secteur d'activité.

- **Formations et salaire :** Informatique, statistiques, Big Data, école d'ingénieur spécialisée, salaire entre 3500 et 4900 euros par mois.

3.7 Les nouveaux métiers

Architecte Big Data

- **Missions :** collecter les données brutes, créer une infrastructure de stockage, manipulation des données, reporting, data management…
- **Compétences :** maitrise des technologies Big Data, maitrise des infrastructures serveur, travail en équipe, communication...
- **Formations et salaire :** Informatique, statistiques, Big Data, école d'ingénieur spécialisée, salaire moyen 3000 euros par mois.

3.7 Les nouveaux métiers

Data Analyst

- **Missions :** analyser les données pour les transformer en informations exploitables, définir la stratégie Data Driven, créer et maintenir des bases de données, élaborer les critères de segmentation…

- **Compétences :** maitrise des bases de données, mathématiques, statistiques, informatique, organisation, anglais...

- **Formations et salaire :** Informatique, statistiques, Big Data, salaire entre 2200 et 2500 euros par mois.

3.7 Les nouveaux métiers

Data Scientist

- **Missions :** collecter et convertir de larges quantités de données, détecter des tendances dans les ensembles de données, rédiger des rapports pour la direction…

- **Compétences :** maitrise des bases de données, Big Data, mathématiques, statistiques, informatique, R, Python, anglais...

- **Formations et salaire :** Informatique, statistiques, Big Data, école d'ingénieur spécialisée, salaire entre 50000 et 60000 euros par an.

TP : Présentation des usages du Big Data

Cas d'usage 1 : Développement de produits

- Des sociétés comme Netflix et Procter & Gamble utilisent le Big Data pour anticiper la demande des clients.
- Elles créent des modèles prédictifs pour de nouveaux produits et services, en classant les principaux attributs de produits ou services passés et présents et en modélisant la relation entre ces attributs et le succès commercial de leurs offres.
- De plus, P&G utilise les données et analyses émanant de groupes cibles, réseaux sociaux, marchés test et présentations en avant-première pour prévoir, produire et lancer de nouveaux produits.

TP : Présentation des usages du Big Data

Cas d'usage 2 : Maintenance prédictive

- Les facteurs permettant de prédire les défaillances mécaniques peuvent être profondément enfouis dans des données structurées, telles que l'année, la marque et le modèle de l'équipement, ainsi que dans des données non structurées couvrant des millions d'entrées de journal, de données de capteur, de messages d'erreur et de température du moteur.
- En analysant ces indications de problèmes potentiels avant que ceux-ci surgissent, les entreprises sont à même de déployer leur maintenance de manière plus rentable et d'optimiser le temps de fonctionnement de leurs pièces et équipements.

TP : Présentation des usages du Big Data

*Cas d'usage 3 : E*xpérience client

- Il est désormais possible d'avoir une meilleure vue d'ensemble de l'expérience client qu'auparavant.
- Le Big Data vous permet de rassembler des données provenant de médias sociaux, de visites Web, de journaux d'appels et d'autres sources pour améliorer l'expérience d'interaction et maximiser la valeur fournie.
- Commencez à proposer des offres personnalisées, à réduire la perte de clients et à traiter les problèmes de manière proactive.

TP : Présentation des usages du Big Data

Cas d'usage 4 : Fraude et conformité

- En matière de sécurité, il ne s'agit pas que de quelques pirates informatiques malhonnêtes : vous faites face à des équipes entières.
- Les paysages de la sécurité et les exigences de conformité sont en évolution constante. Le Big Data vous aide à identifier des modèles dans les données qui indiquent une fraude et à agréger de grands volumes d'informations permettant d'accélérer le reporting réglementaire.

TP : Présentation des usages du Big Data

Cas d'usage 5 : Machine Learning

Nous sommes désormais capables d'enseigner aux machines, plutôt que de simplement les programmer. La disponibilité du Big Data pour former des modèles de machine learning rend cela possible.

TP : Présentation des usages du Big Data

Cas d'usage 6 : Efficacité opérationnelle

- Grâce au Big Data, vous pouvez analyser et évaluer la production, les commentaires et retours des clients, ainsi que d'autres facteurs, afin de réduire les pannes et d'anticiper les demandes à venir.
- Le Big Data permet également d'améliorer la prise de décision, en adéquation avec la demande du marché.

TP : Présentation des usages du Big Data

Cas d'usage 7 : Dynamiser l'innovation

- Le Big Data peut vous aider à innover en étudiant les interdépendances entre les êtres humains, les institutions, les entités et les processus, puis en déterminant de nouvelles façons d'utiliser ces informations.
- Exploiter les informations pour améliorer les décisions dans les domaines financiers et de planification.
- Examiner les tendances et les souhaits des clients pour offrir de nouveaux produits et services.
- Mettre en place une tarification dynamique…

TP : Présentation des usages du Big Data

Prévention des risques sur les équipements

10 *minutes*

- Vous venez d'être recruté(e) en tant qu'Ingénieur(e) Big Data au sein de l'opérateur téléphonique **Antena**.
- Chaque année, des dizaines de milliers de Box internet sont mises hors service par la foudre, ce qui représente un coût élevé et un impact négatif sur la satisfaction client.
- **Votre mission consiste à mettre en place des procédures qui permettront de réduire ces incidents.**
- Décrivez étape par étape, le processus d'amélioration que vous proposerez.

TP : Présentation des usages du Big Data

Solution possible

1. Recueil des données météos : accès aux données via le SI de l'entreprise ou une API disponible sur le marché de la data.
2. Accès la base de données clients (plaignants et non plaignants).
3. Accès aux données de supervision des connexions de Box internet.

- En croisant toutes ces données, il est possible d'envoyer aux clients situés dans une zone à risque un mail et un SMS pour les inciter à débrancher temporairement leurs équipements.

- Après l'orage, un diagnostic à distance est réalisé pour détecter les Box foudroyées afin d'inviter le client à l'échanger en magasin.

Partie 4
Présentation des technologies du Big Data

Plan du module

1. Le stockage des données
2. Les prérequis du stockage
3. Le NoSQL
4. Hadoop et ses outils
5. Les distributions Big Data
6. La visualisation du Big Data

4.1 Le stockage des données

Les enjeux du stockage

- Stocker consiste à sauvegarder les données à l'aide d'une technologie de stockage de type cloud par exemple.

- On y retrouve des données structurées, non structurées et/ou semi structurées; collectées à des fins d'apprentissage automatique, de modélisation prédictive ou de développement d'applications.

- Disposer du meilleur stockage (celui qui convient à sa société) permet d'améliorer ses process, d'augmenter sa rentabilité et de prendre des décisions Data Driven éclairées.

4.1 Le stockage des données

Le stockage de type Datalake

- Lieu de stockage centralisé contenant des données massives sous un format brut, provenant d'un grand nombre de sources.

- Hadoop est très souvent associé aux Data lakes.

https://news.sap.com/france/2021/04/qu-est-ce-qu-un-data-warehouse/

4.1 Le stockage des données

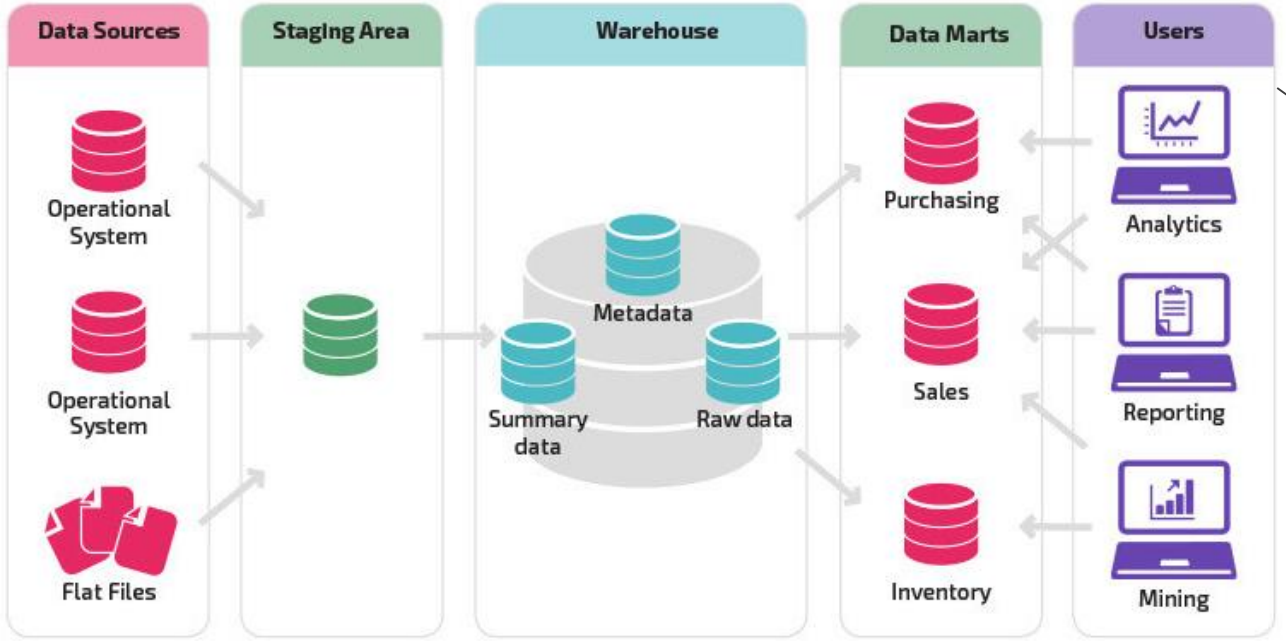
Le stockage de type Data warehouse

- Apparition dans les 80’s.
- Le Data warehouse regroupe, stocke et organise les données.
- Conçu pour supporter les données massives.
- Les caractéristiques d’un data warehouses sont :
  1. **Orienté sujet :** données classées par thèmes (commercial, marketing, production, logistique). Les informations doivent répondre directement aux demandes du métier.
  2. **Intégré :** la collecte des sources diverses de données doivent former un ensemble uniforme et cohérent.
  3. **Time-variant :** on doit pouvoir conserver les données de façon chronologique afin d’en ressortir les grandes tendance.
  4. **Non volatile :** les données sauvegardées ne doivent pas être supprimées

4.1 Le stockage des données

Le stockage de type Data warehouse

Sources de données diverses



Utilisateurs finaux avec usages divers.

Zone de transformation des données en données structurées.

Lieu de stockage des données structurées.

Sous-section du data warehouse, partitionné spécifiquement pour un service ou un secteur d'activité.

https://www.cartelis.com/blog/architecture-data-warehouse/

4.1 Le stockage des données

Data lake vs Data warehouse

	Data lake	**Data warehousse**
Structure des données	Brutes	Traitées/transformées
Objet des données	À déterminer	Données actives
Utilisateurs	Data scientists	Spécialistes
Accessibilité	Accès facile, mises à jour rapides	Modifications plus complexes et plus coûteuses

4.1 Le stockage des données

Quelques exemples d'utilisation

- **Santé :** ce secteur utilise généralement les Data lake pour stocker les notes de médecins et les données cliniques par exemple. Les data lakes permettent de combiner données structurées et données non structurées, ce qui convient généralement mieux aux prestataires de santé.

- **Enseignement :** ce secteur utilise les Data lake pour stocker les données de notes, d'assiduité, de bulletins et d'autres données brutes.

4.1 Le stockage des données

Quelques exemples d'utilisation

- **Finance :** ce secteur utilise généralement les Data warehouse pour y stocker des données structurées facilement accessibles aux métiers.

- **Transport :** ce secteur utilise les Data lake pour stocker les données de logistiques, chaînes d'approvisionnement…

4.2 Les prérequis du stockage

1. Identifier l'objectif

- Quel service de l'entreprise doit être optimisé ou rentabilité ?
- De quelle manière allons-nous l'optimiser ? (gain de temps, automatisation des tâches, baisse des coûts…)
- Quelles données sont à récolter et à croiser entre elles pour obtenir le résultat espéré ?
- …

4.2 Les prérequis du stockage

2. Identifier le système adéquat

- Data warehouse ou Data lake ?
- Cloud ou On-Premise ?
- L’avis d’un expert est requis.

4.2 Les prérequis du stockage

3. Créer un service Data

- Recruter un responsable Big Data qui mettra en place la gouvernance des données (procédures à mettre en place pour la bonne gestion des données de l'entreprise).

4.3 Le NoSQL

~~*Not SQL*~~ *Not Only SQL*

- Le terme « NoSQL » fait référence à des types de bases de données non relationnelles, et ces bases de données stockent les données dans un format différent des tables relationnelles.
- Les données d'un SGBDR sont stockées dans des objets de base de données appelés tables.
- Dans les bases de données NoSQL, les données peuvent être stockées sans définir le schéma à l'avance.
- Solutions actuelles permettant de faire du NoSQL :
  - Cassandra
  - MongoDB
  - SimpleDB

4.3 Le NoSQL

***Types de BDD NoSQL : valeur clé**, document, Graphique, En mémoire*

- **Valeur clé :** Il s'agit du type de base de données NoSQL le plus flexible, car l'application a un contrôle total sur ce qui est stocké dans le champ de valeur, sans aucune restriction.

- Une base de données clé-valeur est un type de base de données non relationnelle qui utilise une méthode clé-valeur simple pour stocker des données.

- Une base de données clé-valeur stocke les données sous forme de paires clé-valeur dans lesquelles une clé sert d'identifiant unique.

4.3 Le NoSQL

***Types de BDD NoSQL : valeur clé**, document, Graphique, En mémoire*

Cas d'utilisation:

- **Magasin de sessions :** Une application orientée session telle qu'une application Web ouvre une session lorsqu'un utilisateur se connecte, puis ferme la session lorsque l'utilisateur se déconnecte ou lorsque la session expire. Pendant cette période, l'application stocke toutes les données liées à la session dans la mémoire principale ou dans une base de données. Les données de session peuvent inclure des informations sur le profil d'utilisateur, des messages, des données et des thèmes personnalisés, des recommandations, des promotions ciblées et des remises. Chaque session d'utilisateur possède un identifiant unique. Les données de session sont uniquement interrogées par une clé primaire.

4.3 Le NoSQL

***Types de BDD NoSQL : valeur clé**, document, Graphique, En mémoire*

Cas d'utilisation:

- **Panier d'achat :** Pendant la période des achats de Noël, un site Web d'e-commerce peut recevoir des milliards de commandes en quelques secondes. Les bases de données clé-valeur peuvent gérer la mise à l'échelle de grandes quantités de données et de grands volumes de changements d'état tout en répondant aux besoins de millions d'utilisateurs simultanés grâce à un traitement et à un stockage distribués.

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé,* ***document****, Graphique, En mémoire*

- Une base de données de documents est un type de base de données non relationnelle conçu pour stocker et interroger des données sous forme de documents de type JSON (semi-structurées).

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé,* ***document****, Graphique, En mémoire*

Document JSON décrivant un livre

```JSON
[
    {
        "year" : 2013,
        "title" : "Turn It Down, Or Else!",
        "info" : {
            "directors" : [ "Alice Smith", "Bob Jones"],
            "release_date" : "2013-01-18T00:00:00Z",
            "rating" : 6.2,
            "genres" : ["Comedy", "Drama"],
            "image_url" : "http://ia.media-imdb.com/images/N/O9ERWAU7FS797AJ7LU8HN09AMUP908RLlo5JF90EWR7LJKQ7@@._V1_SX400_.jpg",
            "plot" : "A rock band plays their music at high volumes, annoying the neighbors.",
            "actors" : ["David Matthewman", "Jonathan G. Neff"]
        }
    },
    {
        "year": 2015,
        "title": "The Big New Movie",
        "info": {
            "plot": "Nothing happens at all.",
            "rating": 0
        }
    }
]
```

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé,* ***document****, Graphique, En mémoire*

Cas d'utilisation:

- **Gestion de contenu :** Une base de données de documents est un bon choix pour les applications de gestion de contenu comme les blogs et les plateformes vidéo. Avec une base de données de documents, chaque entité que suit l'application peut être stockée comme un document unique. La base de données de documents est plus intuitive pour qu'un développeur puisse mettre à jour une application à mesure que les exigences évoluent. De plus, si le modèle de données doit être modifié, seuls les documents concernés doivent être mis à jour. Aucun schéma de mise à jour ou interruption de la base de données n'est nécessaire pour effectuer les modifications.

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé,* ***document****, Graphique, En mémoire*

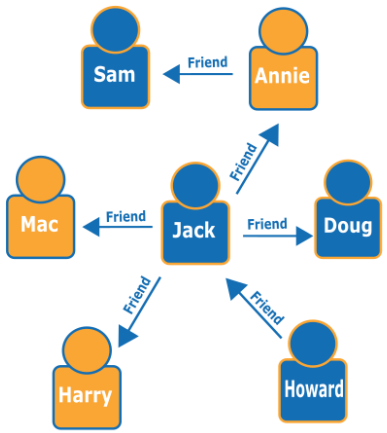
Cas d'utilisation:

- **Catalogues :** Les bases de données de documents sont un bon moyen de stocker des informations sur les catalogues. Par exemple, dans une application d'e-commerce, des produits différents ont souvent des attributs différents. Gérer des milliers d'attributs dans des bases de données relationnelles n'est pas efficace, et cela nuit aux performances de lecture. En utilisant une base de données de documents, les attributs de chaque produit peuvent être décrits dans un seul document pour une gestion aisée et une vitesse de lecture supérieure. Vous pouvez modifier les attributs d'un produit sans craindre de changer ceux d'un autre produit.

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé, document,* ***Graphique****, En mémoire*

- Cette base de données organise les données sous forme de nœuds et de relations, qui montrent les connexions entre les nœuds. Cela permet une représentation plus riche et plus complète des données. Les bases de données graphiques sont appliquées dans les réseaux sociaux, les systèmes de réservation et la détection des fraudes.



On peut déterminer qui sont les « amis des amis » d'une certaine personne : par exemple, les amis des amis d'Howard.

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé, document,* ***Graphique****, En mémoire*

Cas d'utilisation:

- **Détection de fraudes :** Les bases de données orientées graphe peuvent détecter les fraudes de manière sophistiquée. Avec des bases de données orientées graphe, vous pouvez utiliser les relations pour traiter les transactions financières et les transactions d'achat en temps presque réel. En formulant des requêtes de graphe rapides, vous pouvez par exemple détecter qu'un acheteur potentiel utilise la même adresse e-mail et la même carte de crédit enregistrées lors d'un précédent cas de fraude. Les bases de données orientées graphe peuvent également vous aider dans la conception de requêtes de graphe afin de facilement détecter les modèles de relations, tels que les cas d'utilisation d'une adresse e-mail personnelle par plusieurs personnes ou de partage d'adresse IP entre plusieurs personnes se trouvant à des adresses physiques différentes.

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé, document,* ***Graphique****, En mémoire*

Cas d'utilisation:

- **Moteurs de recommandation :** Les bases de données orientées graphe sont un bon choix pour les applications de recommandation. Avec des bases de données orientées graphe, vous pouvez stocker dans un graphe des relations entre des catégories d'informations telles que les intérêts d'un client, ses amis et son historique d'achat. Vous pouvez utiliser une base de données orientée graphe à disponibilité élevée afin de recommander des produits à un utilisateur en fonction des produits achetés par les autres utilisateurs qui sont abonnés à la même page sportive et dont l'historique d'achat est similaire. Vous pouvez aussi trouver les personnes qui ont un ami en commun, mais qui ne se connaissent pas encore, et ensuite émettre une recommandation de mise en relation.

4.3 Le NoSQL

Types de BDD NoSQL : valeur clé, document, Graphique, En mémoire

- Les bases de données en mémoire sont des bases de données spécialement conçues qui reposent principalement sur la mémoire pour le stockage des données, contrairement aux bases de données qui stockent les données sur disque ou SSD.

- Les magasins de données en mémoire sont conçus pour permettre des temps de réponse minimaux en éliminant le besoin d'accéder aux disques. Étant donné que toutes les données sont stockées et gérées exclusivement dans la mémoire principale, les bases de données en mémoire risquent de perdre des données en cas de défaillance d'un processus ou d'un serveur. Les bases de données en mémoire peuvent conserver des données sur des disques en stockant chaque opération dans un journal ou en prenant des instantanés.

- Les bases de données en mémoire sont idéales pour les applications qui nécessitent des temps de réponse de l'ordre de la microseconde ou qui connaissent des pics de trafic importants, telles que les classements de jeux, les magasins de sessions et les analyses en temps réel.

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé, document, Graphique,* ***En mémoire***

Cas d'utilisation:

- **Enchères en temps réel :** Les enchères en temps réel font référence à l'achat et à la vente d'impressions d'annonces en ligne. Habituellement, l'enchère doit être faite pendant que l'utilisateur charge une page Web, en 100 à 120 millisecondes et parfois aussi peu que 50 millisecondes. Pendant cette période, les applications d'enchères en temps réel demandent des offres à tous les acheteurs pour le spot publicitaire, sélectionnent une enchère gagnante en fonction de plusieurs critères, affichent l'enchère et collectent les informations post-affichage de l'annonce. Les bases de données en mémoire sont des choix idéaux pour ingérer, traiter et analyser des données en temps réel avec une latence inférieure à la milliseconde.

4.3 Le NoSQL

Types de BDD NoSQL : *valeur clé, document, Graphique,* ***En mémoire***

Cas d'utilisation:

- **Classements de jeu :** Un classement de jeu relatif montre la position d'un joueur par rapport à d'autres joueurs d'un rang similaire. Un classement de jeu relatif peut aider à renforcer l'engagement des joueurs et, en attendant, empêcher les joueurs de se démotiver lorsqu'ils ne sont comparés qu'aux meilleurs joueurs. Pour un jeu avec des millions de joueurs, les bases de données en mémoire peuvent fournir des résultats de tri rapidement et maintenir le classement à jour en temps réel.

4.3 Le NoSQL

Types de BDD NoSQL :** valeur clé, document, Graphique, **En mémoire

Cas d'utilisation:

- **Mise en cache:** Un cache est une couche de stockage de données à haut débit qui stocke un sous-ensemble de données, généralement de nature transitoire, de sorte que les demandes futures pour ces données soient traitées plus rapidement qu'il n'est possible en accédant à l'emplacement de stockage principal des données. La mise en cache vous permet de réutiliser efficacement les données précédemment récupérées ou calculées. Les données d'un cache sont généralement stockées dans un matériel à accès rapide tel que la RAM (mémoire à accès aléatoire) et peuvent également être utilisées en corrélation avec un composant logiciel. L'objectif principal d'un cache est d'augmenter les performances de récupération des données en réduisant le besoin d'accéder à la couche de stockage sous-jacente plus lente.

4.4 Hadoop et ses outils

Les 3 éléments de base d'Hadoop : HDFS, YARN et MapReduce

- HDFS (Hadoop Distributed File System) est le système de fichiers distribué et l'élément central de Hadoop permettant de stocker et répliquer des données sur plusieurs serveurs.
- HDFS utilise un NameNode et un DataNode.
- Le DataNode est un serveur standard sur lequel les données sont stockées.
- Le NameNode contient des métadonnées (informations sur les données stockées dans les différents nœuds).
- L'application interagit uniquement avec le NameNode, et celui-ci communique avec les nœuds de données selon besoin.

4.4 Hadoop et ses outils

Les 3 éléments de base d'Hadoop : HDFS, YARN et MapReduce

- YARN est l'abréviation de « Yet Another Resource Negotiator » (plus simplement, un négociateur de ressources).
- Cet élément assure la gestion et planification des ressources (clusters) Hadoop et décide de ce qui doit se passer dans chaque nœud de données.
- Le nœud maître central qui gère toutes les demandes de traitement est le « Resource Manager ».
- Le Resource Manager interagit avec les différents Node Managers : chaque DataNode esclave possède son propre Node Manager pour l'exécution des tâches.

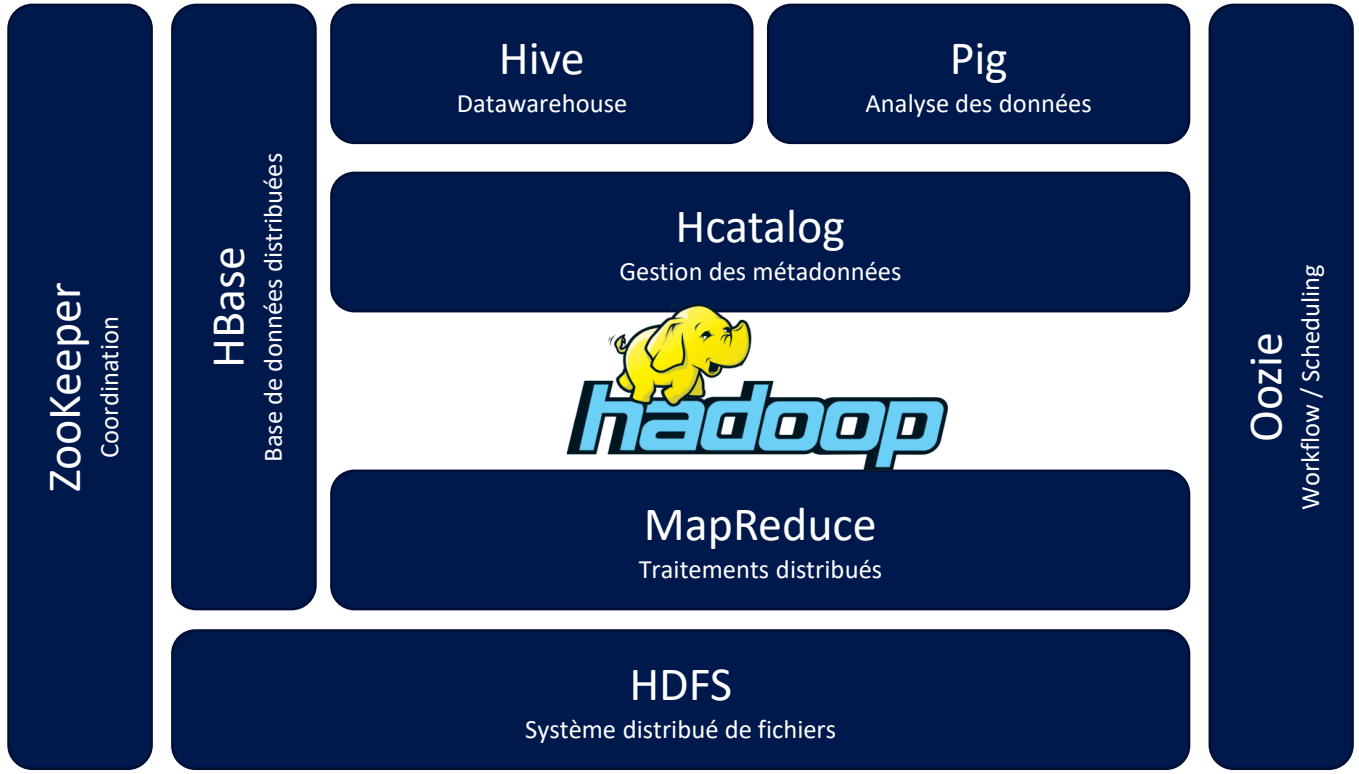
4.4 Hadoop et ses outils

Les 3 éléments de base d'Hadoop : HDFS, YARN et MapReduce

- MapReduce est un modèle de programmation qui a d'abord été utilisé par Google pour indexer ses opérations de recherche.
- Suivant cette logique, cet élément exécute des algorithmes pour décomposer des données en datasets plus petits. MapReduce s'appuie sur deux fonctions : Map() et Reduce(), qui analysent les données rapidement et efficacement.
- La fonction Map regroupe, filtre et trie plusieurs datasets en parallèle et génère des tuples (paires key value). La fonction Reduce agrège ensuite les données de ces tuples pour produire le résultat souhaité.

4.4 Hadoop et ses outils

Les composants d'Hadoop



Enjeux et usages du Big Data, Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361

4.4 Hadoop et ses outils

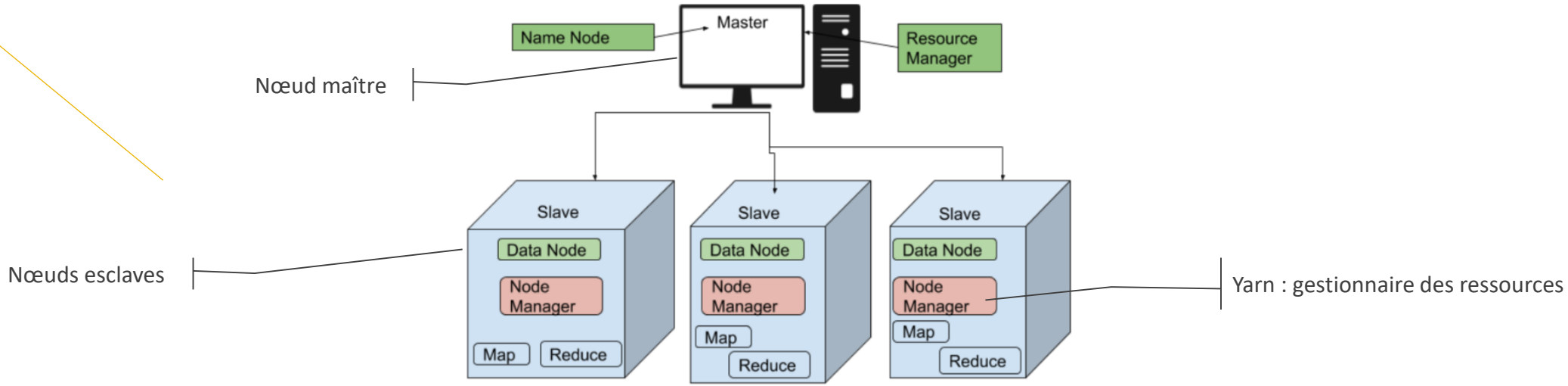
Le cœur d'Hadoop : HDFS

- HDFS est un système de fichiers Java utilisé pour stocker des données structurées ou non sur un ensemble de serveurs distribués.
- HDFS s'appuie sur le système de fichier natif de l'OS pour présenter un système de stockage unifié reposant sur un ensemble de disques et de systèmes de fichiers hétérogènes.
- La consistance des données est basée sur la redondance. Une donnée est stockée sur au moins n volumes différents.

4.4 Hadoop et ses outils

Le cœur d'Hadoop : HDFS

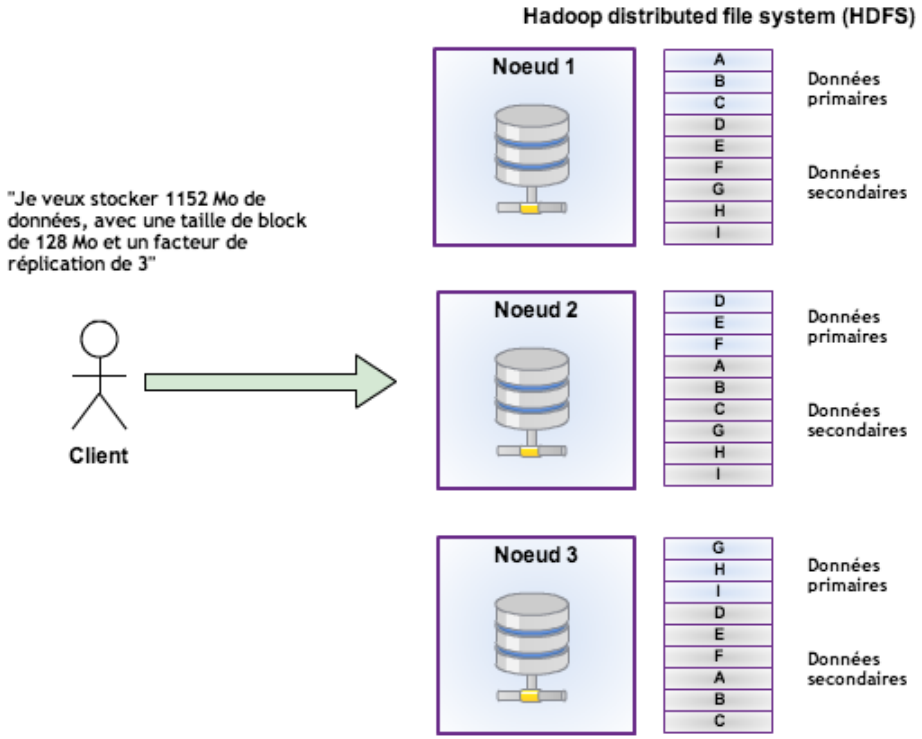
- Dans une architecture Hadoop chaque membre pouvant traiter des données est appelé node (nœud). Un seul d'entre eux peut être master même s'il peut changer au cours de la vie du cluster.



4.4 Hadoop et ses outils

Le cœur d'Hadoop : HDFS

- Au sein du cluster, les données sont découpées et distribuées en blocks selon les deux paramètres suivants :
  - **Blocksize :** Taille unitaire de stockage (généralement 64 Mo ou 128 Mo). C'est à dire qu'un fichier de 1 Go (et une taille de block de 128 Mo) sera divisé en 8 blocks.
  - **Replication factor :** C'est le nombre de copies d'une données devant être réparties sur les différents nœuds du cluster (souvent 3, c'est à dire une primaire et deux secondaires).



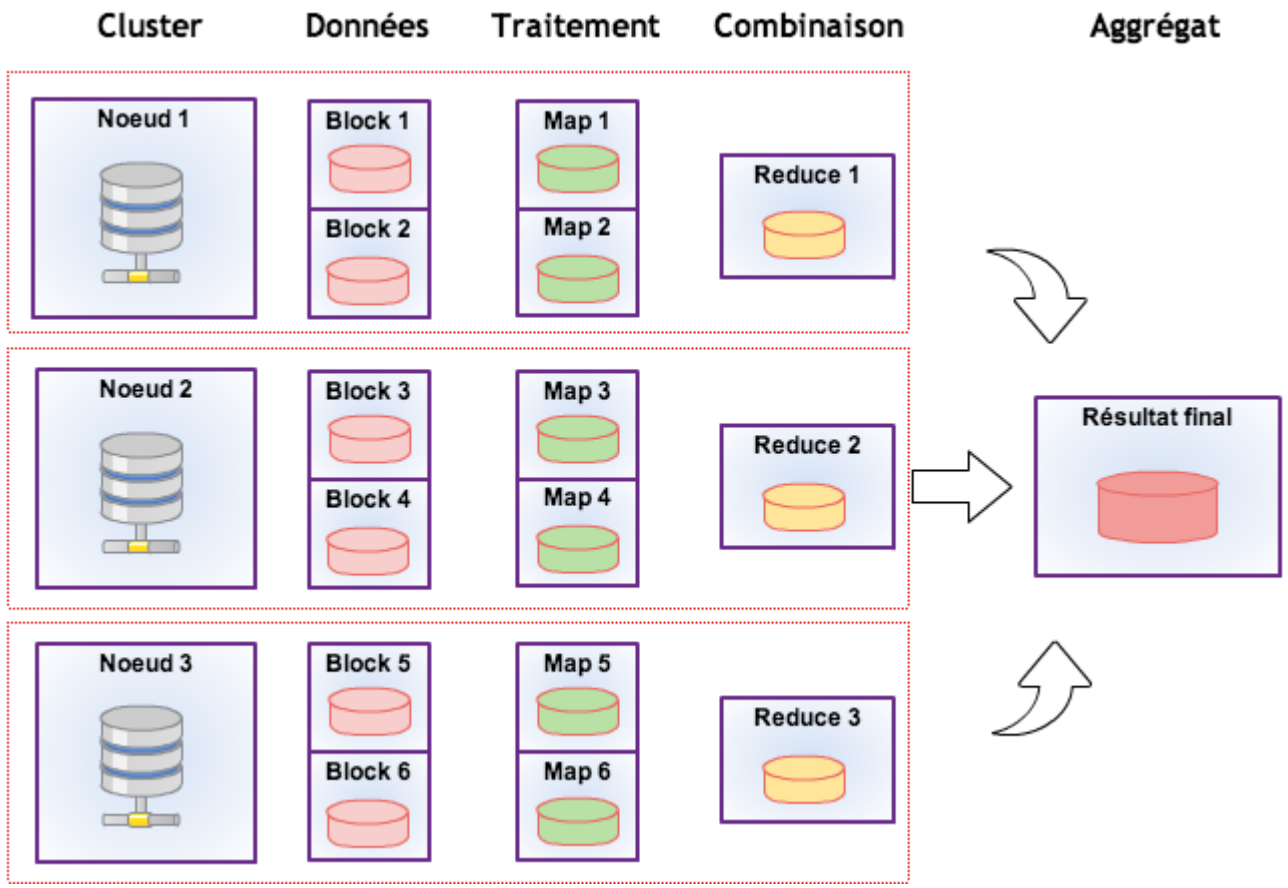
4.4 Hadoop et ses outils

Le cœur d'Hadoop : MapReduce

- C'est Google qui a inventé le MapReduce pour la recherche web.
- MapReduce est un framework qui va décomposer une requête importante en plusieurs autres petites requêtes qui vont elles aussi produire un sous ensemble du résultat final (fonction Map).
- L'ensemble des résultats est traité (agrégation, filtre) : c'est la fonction Reduce.

4.4 Hadoop et ses outils

Le cœur d'Hadoop : MapReduce



4.4 Hadoop et ses outils

Extensions d'Hadoop : Hive pour le requêtage

- Hive est à l'origine un projet Facebook qui permet de faire le lien entre le monde SQL et Hadoop.
- Il permet l'exécution de requêtes SQL sur un cluster Hadoop en vue d'analyser et d'agréger les données.
- Le langage SQL est nommé HiveQL. C'est un langage de visualisation uniquement, c'est pourquoi seules les instructions de type "Select" sont supportées pour la manipulation des données.
- Hive utilise un connecteur JDBC/ODBC.

4.4 Hadoop et ses outils

Extensions d'Hadoop : Pig pour le scripting

- Pig est à l'origine un projet Yahoo qui permet le requêtage des données Hadoop à partir d'un langage de script.
- Contrairement à Hive, Pig est basé sur un langage de haut niveau PigLatin qui permet de créer des programmes de type MapReduce.
- Contrairement à Hive, Pig ne dispose pas d'interface web.

4.4 Hadoop et ses outils

Intégration SGBD-R : Sqoop

- Sqoop permet le transfert des données entre un cluster Hadoop et des bases de données relationnelles.
- C’est un produit développé par Cloudera.
- Il permet d’importer/exporter des données depuis/vers Hadoop et Hive.
- Pour la manipulation des données Sqoop utilise MapReduce et des drivers JDBC.

4.4 Hadoop et ses outils

Ordonnanceur : Apache Oozie

- Oozie est une solution de workflow (au sens scheduler d'exploitation) utilisée pour gérer et coordonner les tâches de traitement de données à destination de Hadoop.

- Oozie s'intègre parfaitement avec l'écosystème Hadoop puisqu'il supporte les types de jobs suivant :
  - MapReduce (Java et Streaming).
  - Pig.
  - Hive.
  - Sqoop.

4.4 Hadoop et ses outils

Gestion des clusters Hadoop : Apache Zookeeper

- ZooKeeper est un service de coordination des services d'un cluster Hadoop.
- En particulier, le rôle de ZooKeeper est de fournir aux composants Hadoop les fonctionnalités de distribution.
- Pour cela il centralise les éléments de configuration du cluster Hadoop, propose des services de clusterisation et gère la synchronisation des différents éléments (événements).
- ZooKeeper est un élément indispensable au bon fonctionnement de HBase.

4.4 Hadoop et ses outils

Supervision : Apache Ambari

- Ambari est un projet d'incubation Apache initié par HortonWorks et destiné à la supervision et à l'administration de clusters Hadoop.
- C'est un outil web qui propose un tableau de bord. Cela permet de visualiser rapidement l'état d'un cluster.
- Ambari dispose d'un tableau de bord dont le rôle est de fournir une représentation :
  - De l'état des services.
  - De la configuration du cluster et des services.
  - Des informations issues de Ganglia et de Nagios.
  - De l'exécution des jobs.
  - Des métriques de chaque machine et du cluster.

4.4 Hadoop et ses outils

Autres outils : Apache Flume

- Flume est une solution de collecte et d'agrégation de fichiers logs, destinés à être stockés et traités par Hadoop.
- Il a été conçu pour s'interfacer directement avec HDFS au travers d'une API native.
- Flume est à l'origine un projet Cloudera, reversé depuis à la fondation Apache.
- Alternatives : Apache Chukwa.

4.4 Hadoop et ses outils

Autres outils : Apache Mahout

- Apache Mahout est un projet de la fondation Apache visant à créer des implémentations d'algorithmes d'apprentissage automatique et de datamining.

- Même si les principaux algorithmes d'apprentissage se basent sur MapReduce, il n'y a pas d'obligation à utiliser Hadoop. Apache Mahout ayant été conçu pour pouvoir fonctionner sans cette dépendance.

4.4 Hadoop et ses outils

Autres outils : Apache Drill

- Initié par MapR, Drill est un système distribué permettant d'effectuer des requêtes sur de larges données. Il implémente les concepts exposés par le projet Google Dremel.
- Drill permet d'adresser le besoin temps réel d'un projet Hadoop. MapReduce étant plutôt conçu pour traiter de larges volumes de données en batch sans objectif de rapidité et sans possibilité de redéfinir la requête à la volée.
- Drill est donc un système distribué qui permet l'analyse interactive des données, ce n'est pas un remplacement de MapReduce mais un complément qui est plus adapté pour certains besoins.

4.4 Hadoop et ses outils

Autres outils : Apache HCatalog

- HCatalog permet l'interopérabilité d'un cluster de données Hadoop avec des systèmes externes.

- HCatalog est un service de management de tables et de schéma des données Hadoop :
  - Permet d'attaquer les données HDFS via des schémas de type tables de données en lecture/écriture.
  - Permet d'opérer sur des données issues de MapReduce, Pig ou Hive.

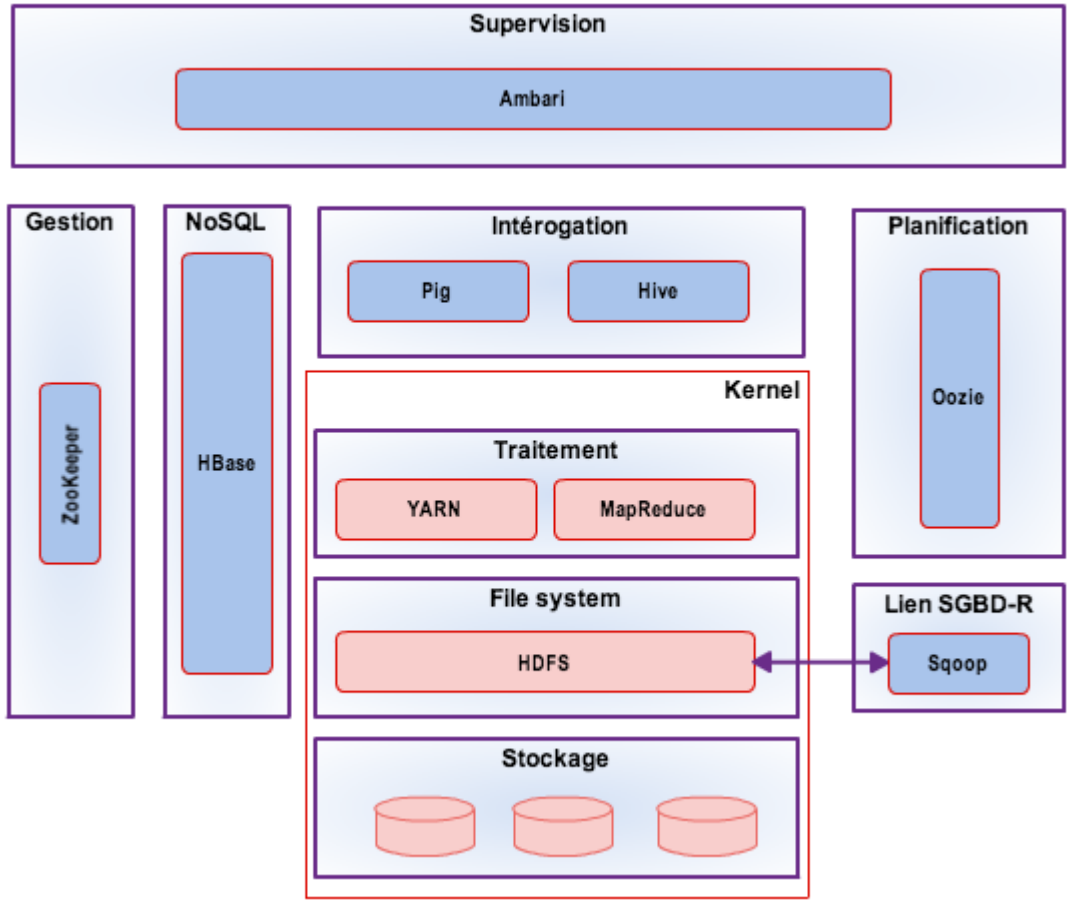
4.4 Hadoop et ses outils

Autres outils : Apache Tez

- Tez est un nouveau framework en incubation chez Apache.
- Utilisant YARN il remplace MapReduce afin de fournir des requêtes dites “temps réel”. La faible latence est en effet un pré requis à l’exploration interactive des données stockées sur un cluster Hadoop.
- C’est un concurrent d’Apache Drill (MapR) ou de Cloudera Impala.

4.4 Hadoop et ses outils

Autres outils : vue d'ensemble



4.5 Les distributions Big Data

HortonWorks

- Fondé en 2011 par les équipes de Yahoo!, Hortonworks Data Platform, est une plateforme de gestion et de traitement analytique de données massives.
- Leur but est de faciliter l'adoption de la plate forme Hadoop d'Apache, c'est pourquoi tous les composants sont open source et sous licence Apache.
- Le modèle économique d'HortonWorks est de ne pas vendre de licence mais uniquement du support et des formations.

4.5 Les distributions Big Data

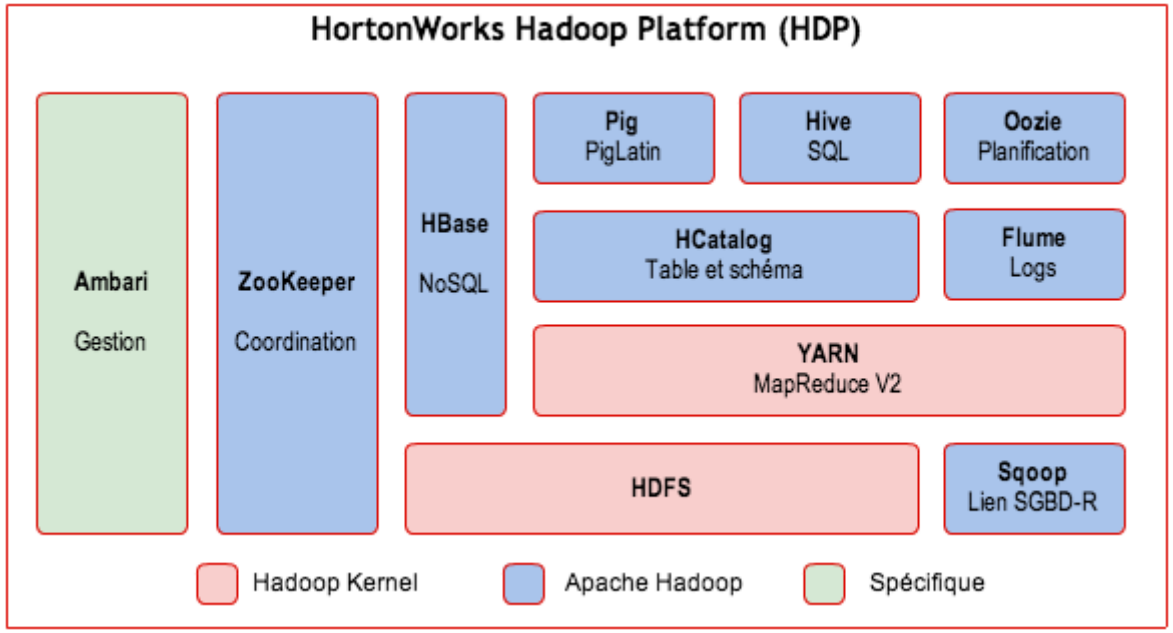
HortonWorks : les composants

1. Cœur Hadoop (HDFS/MapReduce).
2. NoSQL (Apache HBase).
3. Méta-données (Apache HCatalog).
4. Plate forme de script (Apache Pig).
5. Requêtage (Apache Hive).
6. Planification(Apache Oozie).
7. Coordination (Apache Zookeeper).
8. Gestion et supervision (Apache Ambari).
9. Services d'intégration (HCatalog APIs, WebHDFS, Talend Open Studio for Big Data, Apache Sqoop).
10. Gestion distribuée des logs (Apache Flume).
11. Apprentissage (Apache Mahout).

4.5 Les distributions Big Data

HortonWorks

Vue d'ensemble de la distribution



4.5 Les distributions Big Data

HortonWorks : déploiement de la plateforme

- **Machine virtuelle (VM) :**
  - HortonWorks met à disposition une machine virtuelle dans laquelle sont pré installés les composants de la plate forme Hadoop.
  - C'est l'idéal pour l'apprentissage de la plateforme mais **incompatible** avec les exigences de production ou même celles d'un POC.

4.5 Les distributions Big Data

HortonWorks : déploiement de la plateforme

- **Installation automatique avec Ambari :**
  - En plus de la gestion du cluster, Ambari permet le déploiement de l'ensemble des composants Hadoop de manière centralisée.

- **Installation manuelle avec Linux RPM :**
  - HortonWorks met à disposition des packages RPM.
  - En utilisant le principe des RPM Linux il est possible d'installer les composants HDP manuellement.

4.5 Les distributions Big Data

Cloudera

- Entreprise qui commercialise Hadoop.
- Si leur plate forme est en grande partie basée sur Hadoop d'Apache, elle est complétée avec des composants maison essentiellement pour la gestion du cluster.
- Le modèle économique de Cloudera est la vente de licences mais aussi du support et des formations.

4.5 Les distributions Big Data

Cloudera : les composants qui viennent d'Apache

1. HDFS : File System distribué.
2. MapReduce : Framework de traitement parallélisé.
3. HBase : Base de données NoSQL (accès read/write aléatoires).
4. Hive : Requêtage de type SQL.
5. Pig : Scripting et requêtage Hadoop.
6. Oozie : Workflow et planification de jobs Hadoop.
7. Sqoop : Intégration de bases SQL.
8. Flume : Exploitation de fichiers (log) dans Hadoop.
9. ZooKeeper : Service de coordination pour les applications distribuées.
10. Mahout : Framework d'apprentissage et de datamining pour Hadoop.

4.5 Les distributions Big Data

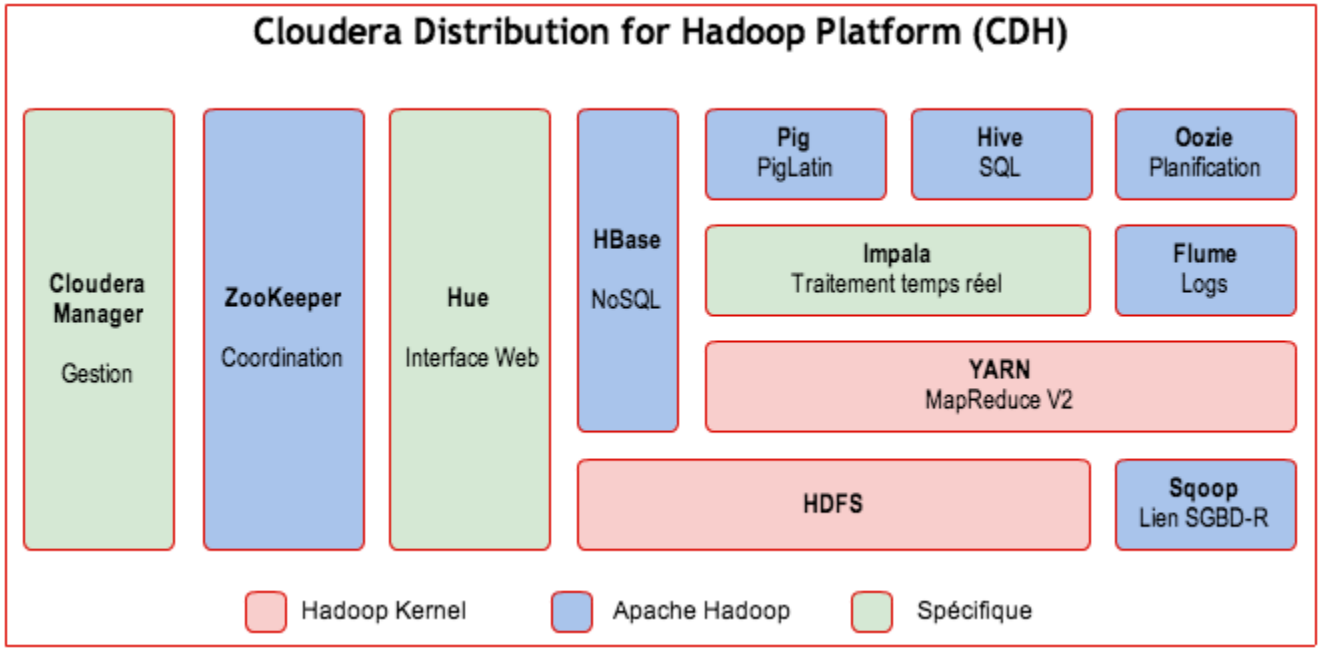
Cloudera : les composants d'origine

- Hadoop Common: Un ensemble d'utilitaires.
- Hue : SDK permettant de développer des interfaces utilisateur pour les applications Hadoop.
- Whirr : Librairies et scripts pour l'exécution d'Hadoop et de services liés dans le cloud.

4.5 Les distributions Big Data

Cloudera

Vue d'ensemble de la distribution



4.5 Les distributions Big Data

Cloudera : déploiement de la plateforme

- **Automatique avec Cloudera Manager :**
  - Cloudera Manager permet l'installation des composants de la plate forme sur une machine (y compris distante).
  - Cloudera Manager permet la configuration centralisée des composants du cluster.
  - Enfin Cloudera Manager permet de finaliser l'installation en vérifiant le bon fonctionnement de chacun des composants.

- Manuel avec les packages
  - Récupération des archives tarball (tgz) contenant la distribution.
  - Configuration et installation à l'aide des scripts fournis.

4.5 Les distributions Big Data

MapR

- MapR a été fondée en 2009 par d'anciens membres de Google.
- Bien que son approche soit commerciale, MapR contribue à des projets Apache Hadoop comme HBase, Pig, Hive, ZooKeeper et surtout Drill.
- MapR se distingue surtout de la version d'Apache Hadoop par sa prise de distance avec le cœur de la plate-forme. Ils proposent ainsi leur propre système de fichier distribué ainsi que leur propre version de MapReduce : MapR FS et MapR MR.

4.5 Les distributions Big Data

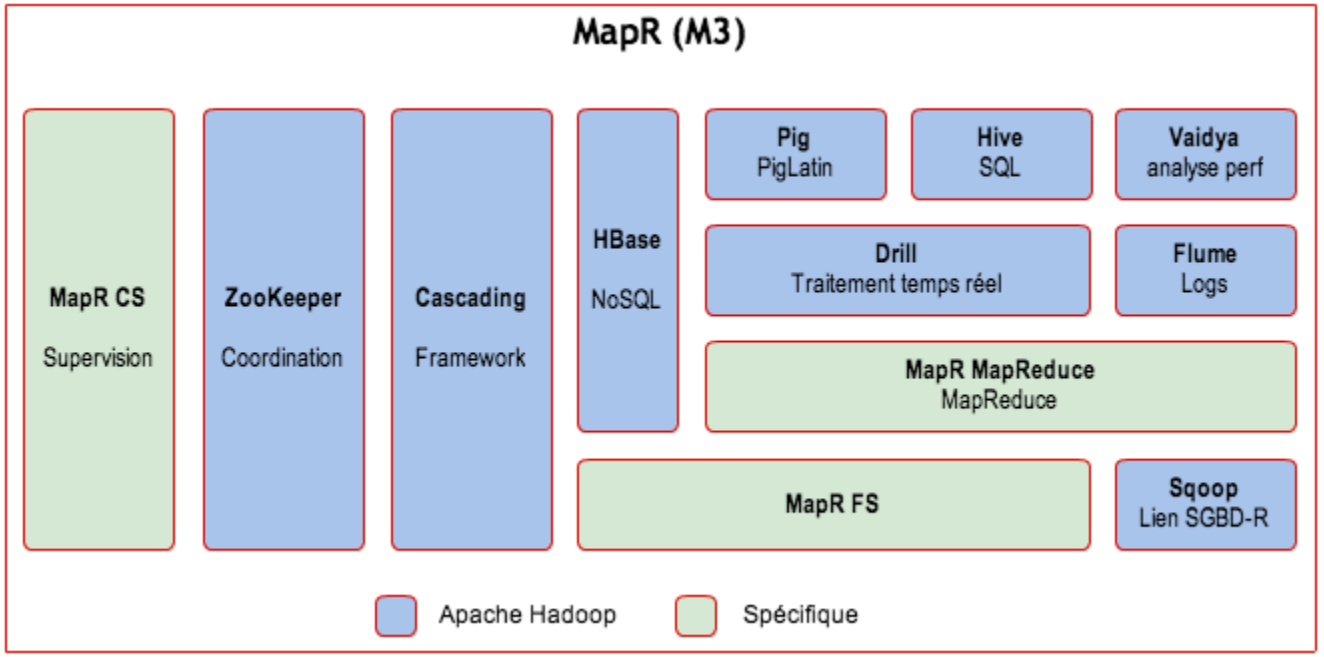
MapR : les composants

- HBase
- Pig
- Hive
- Mahout
- Cascading
- Sqoop
- Flume

4.5 Les distributions Big Data

MapR

Vue d'ensemble de la distribution



4.5 Les distributions Big Data

MapR : déploiement de la plateforme

- **Machine virtuelle :**
  - MapR fourni une machine virtuelle avec un seul nœud et l'ensemble des composants installés.
  - C'est l'idéal pour une prise en main de la plate-forme mais incompatible avec les exigences de production.

- **Manuelle avec les packages :**
  - Composants à installer :
    - Depuis le repository internet
    - Depuis un repository local
    - Avec des packages Debian/Linux

4.6 La visualisation du Big Data

Qu’est-ce que la visualisation Big Data ?

- La visualisation Big Data est un ensemble de techniques permettant de représenter les données massives sous une forme visuelle, dans le but de les interpréter plus facilement.

- Il va bien au-delà des graphiques de bases : nuages de points, bulles, histogrammes, camemberts, anneaux.

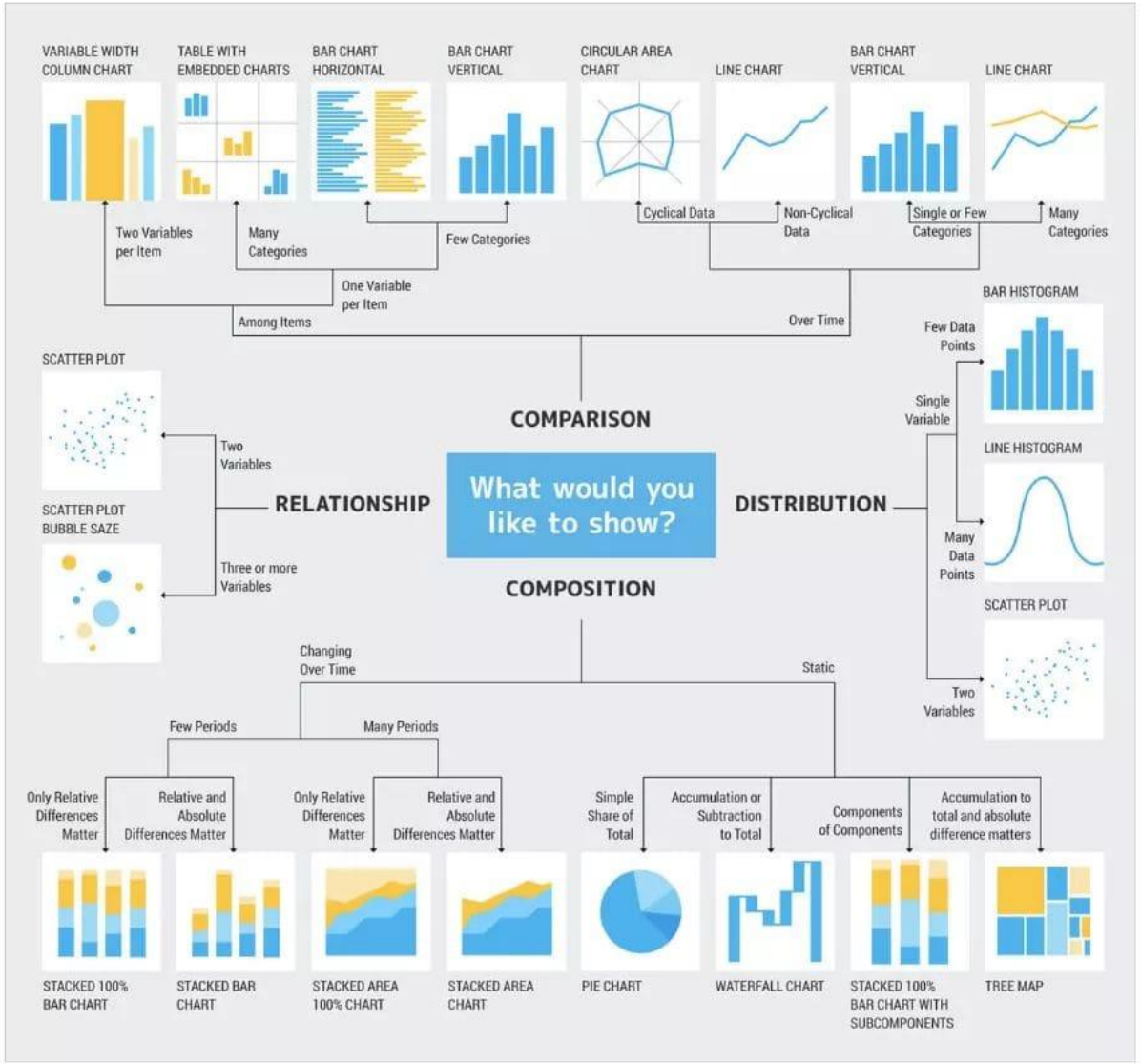
- Représentations complexes : cartes thermiques, boîtes à moustaches, Sankey, Sunburst…; permettant aux décideurs d'explorer des ensembles de données pour identifier des corrélations ou des modèles inattendus.

4.6 La visualisation du Big Data

L’intérêt de visualiser ses données Big Data

- Permettre aux décideurs de comprendre très rapidement ce que signifient les données (insights).
- Dégager des tendances : séries longues, saisonnalités, hausse, baisse…
- Révéler des modèles : identifier les corrélations et les connexions inattendues qui n'ont pas pu être trouvées avec la business intelligence classique.
- Fournissez un moyen très efficace de communiquer toutes les informations qui apparaissent aux autres.

4.6 La visualisation du Big Data



4.6 La visualisation du Big Data

Les principaux outils de visualisation de données Big Data

- Microsoft Power BI
- Tableau Desktop
- Google Data Studio
- Qlik solution : QlikSense et QlikView
- Oracle Visual Analyzer

Partie 5

Les enjeux juridiques, sécuritaires et éthiques

Plan du module

1. La protection des données : anonymisation d'une donnée, contrôle d'intégrité, chiffrement d'une donnée
2. Le RGPD
3. TP : étude d'une mise en place du Big Data dans le secteur de la mutuelle

5.1 La protection des données

Anonymisation d'une donnée, contrôle d'intégrité, chiffrement d'une donnée

- Pour se mettre en conformité avec le RGPD.
- L'anonymisation rend impossible l'identification d'une personne à partir d'un jeu de données et permet, ainsi, de respecter sa vie privée.
- L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, toute identification de la personne par quelque moyen que ce soit et de manière irréversible (une suppression par exemple).
- L'anonymisation ne doit pas être confondue avec la pseudonymisation.
- La pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.) ➔ Obfusquer.

5.1 La protection des données

Anonymisation d'une donnée, **contrôle d'intégrité**, chiffrement d'une donnée

- L'intégrité des données correspond à la capacité de garantir l'exhaustivité, la précision, l'exactitude et la validité des données durant tout leur cycle de vie ➔ Gouvernance des données.

- Un défaut d'intégrité correspond à une altération ou une destruction des données.

5.1 La protection des données

Anonymisation d'une donnée, **contrôle d'intégrité**, chiffrement d'une donnée

- **Les causes possibles d'altération des données :**
  - Les fraudes internes et/ou externes : malveillance, cyber crime, les virus informatiques…
  - Les défaillances techniques du Système d'Information (ex : un bug dans une application qui supprime la mauvaise donnée, une validation inadéquate…)
  - Les erreurs lors des transferts d'informations ou des réplications
  - Les erreurs humaines de saisie, d'exploitation ou de manipulation
  - Les risques matériels (incendie, panne…)

5.1 La protection des données

Anonymisation d'une donnée, **contrôle d'intégrité**, chiffrement d'une donnée

- **Conseils pour assurer l'intégrité des données :**
  - Fiabiliser la collecte des données. Toute entrée doit être vérifiée et validée et être cohérente avec le dictionnaire de données.
  - Contrôler les permissions et les droits d'accès et de modification.
  - Centraliser et garantir l'unicité de vos bases de données.
  - Surveiller toutes les modifications effectuées sur les données (ajout, suppression, modification…) et disposer d'un historique complet et non-falsifiable.
  - Assurer la sauvegarde les données et la capacité à restaurer des informations.
  - Documenter les procédures, définir les règles et les obligations du personnel.

5.1 La protection des données

Anonymisation d'une donnée, contrôle d'intégrité, **chiffrement d'une donnée**

- Le chiffrement des données consiste à convertir les données afin que seules les personnes pourvues d'une clé secrète ou d'un mot de passe soient en mesure de les lire.

- On utilise une clé de chiffrement pour chiffrer les données, et une clé de déchiffrement pour les déchiffrer.

- On distingue deux principaux types de chiffrement de données : le **chiffrement asymétrique** (deux clés sont utilisées : une clé publique, et une clé privée. La clé publique peut être partagée avec n'importe qui, mais la clé privée doit impérativement être protégée), et le **chiffrement symétrique** (la même clé est utilisée pour le chiffrement et le déchiffrement du fichier).

5.2 Le RGPD

La CNIL : le gendarme européen

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

- Veille à l'application du règlement européen et émet des sanctions en cas d'infraction.
- Au maximum les sanctions peuvent aller jusqu'à 20 millions d'euros ou une somme égale à 4% du chiffre d'affaires annuel mondial de l'entreprise.
- En plus de l'amende, la sanction émise va entacher l'image de l'organisme en question.

5.2 Le RGPD

Le RGPD en quelques lignes

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

- Transparence de traitement
- Information claire et précise
- Réduction de la quantité de données collectées
- Limite de la durée de conservation
- Respect des droits des utilisateurs (consultation, portabilité, droit à l'oubli)

5.2 Le RGPD

Quelques cas de vols / fuites de données

- En 2021, fuite de données médicales en France : 500.000 noms mis en vente sur Telegram.
- En 2020, Bouygues Construction a été victime d'un piratage de 700 To de données (rançon exigée : 10 M $).
- La société Transavia, filiale d'AirFrance a été victime d'un vol massif de données de ses salariés : CV, lettres de motivations, documents médicaux…
- L'AP-HP (Assistance Publique Hôpitaux de Paris) victime d'une fuite de données de dépistages Covid-19 de 1,4 M de personnes.

5.2 Le RGPD

Quelques cas de vols / fuites de données

- L’hébergeur OVH a été piraté en 2013. Les données de ses clients européens ont été volées : nom, le prénom, l'adresse, la ville, le pays, le numéro de téléphone, le numéro de fax, l'identifiant et le mot de passe du compte.

- LinkedIn victime de piratage de données en 2021. 700 M de comptes mises en vente sur le Dark web.

- Facebook assiste à une fuite de données de 533 M d’utilisateur sur le web…

5.2 Le RGPD

Quelques sanctions pour manquements

- La CNIL a sanctionné :
  - Google pour avoir rendu difficile aux utilisateurs le refus de cookies : 150 M €.
  - Un commerce de détail d'optique pour défaut de sécurité des données : 250 K€.
  - Une société d'assurance pour défaut de durée de conservation des données et d'information des personnes : 1,750 M €.
  - Un organisme de presse pour défaut de consentement des personnes (cookies) : 50 K €.
  - Une société de vente de mobilier sur internet et en magasin, pour non respect des demandes de suppression de données personnelles : 120 K €.
  - Un opérateur téléphonique pour non respect du droit, de modification de données personnelles et défaut de sécurité des données : 300 K €…

TP : étude d'une mise en place du Big Data dans le secteur de la mutuelle

- La mutuelle vitalys a été créée en 2011.

- Suite à un surcroît d'activité, vous avez été recruté pour mettre en place une infrastructure Big Data pour répondre aux nouveaux enjeux de l'entreprise :
  - Répondre aux défis du stockage
  - Prendre en compte les critères de performance
  - Croisement des données multi sources
  - Reporting, Business Intelligence et Machine Learning…

TP : étude d'une mise en place du Big Data dans le secteur de la mutuelle

- Ancienne infrastructure :
  - Base de données SQL Server
  - Fichiers plats

- Nouveauté à prendre en compte dans la nouvelle infra :
  - Fichiers texte, audio, vidéo, images, flux de données des réseaux sociaux…

Reproduisez le schémas permettant de collecter, traiter et analyser les données en y associant les technologies vues en séance.

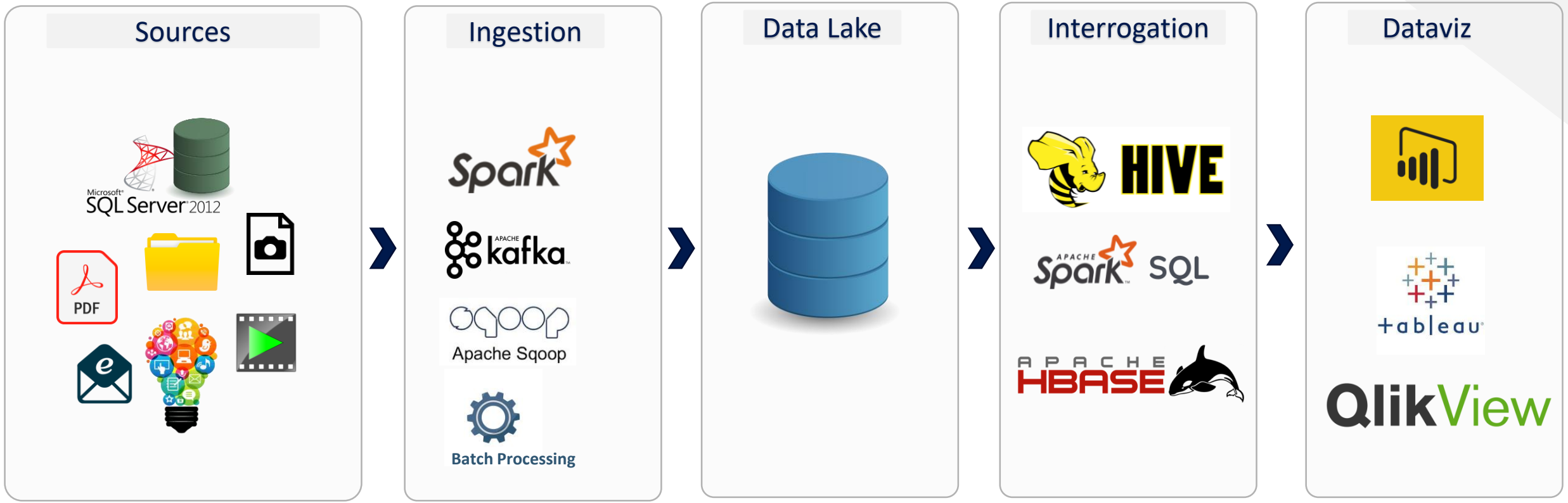
TP : étude d'une mise en place du Big Data dans le secteur de la mutuelle

Pistes de réflexion

1. *Choix de l'infrastructure Big Data : Hadoop Datawarehouse ou Datalake ? Discuter…*
2. *Intégration des données selon l'infra choisie: identifier les sources des données et les types de données à importer, intégrer ou non la data prep.*
3. *La stack technique d'Hadoop :*
    1. *Type d'architecture Big Data : Datalake, Lambda, Kappa*
    2. *Ingestion des données (Spark, Sqoop, Kafka, Batch)*
    3. *Interrogation des données (Hive, Spark SQL, Hbase)*
    4. *Outils Hadoop : Pig, Sqoop, Oozie, Zookeeper, Ambari, Flume, Mahout*
    5. *Outils de DataViz : Power BI ou Tableau*

TP : étude d'une mise en place du Big Data dans le secteur de la mutuelle

Solution possible



Partie 6

Découverte et manipulation de l’outil de Data Visualisation Power BI

6.1 Découverte des données

Sample Superstore : les indicateurs

- KPI n°1 : Nombre de clients ➔ Tiles
- KPI n°2 : Nombre de commandes ➔
- KPI n°3 : Total des quantités ➔ Tiles
- KPI n°4 : Total des ventes ➔ Tiles
- KPI n°5 : Total bénéfice ➔ Tiles

- Chart n°1 : Ventes par mois ➔ Graphique barres
- Chart n°2 : Bénéfice par mois ➔ Graphique barres
- Chart n°3 : Segments ➔ Anneau
- Chart n°4 : Catégorie de produits ➔ Anneau
- Chart n°5 : US map ➔ Cartographie

6.2 Mis en place d’un tableau de bord

Le storyboard

Sources

Documents	Auteur(s)
Enjeux et usages du Big Data	Christophe Brasseur, Ed. Lavoisier, ISSN 1635-7361
Guide du Big Data 2014/2015	Corp-agency
IssyGrid, un succès qui ouvre la voie au modèle français du smart grid	https://www.issy.com/issygrid
Big Data Analytics	IDC France
Architecture big data	https://www.talend.com/fr/resources/architecture-big-data/
Qu'est-ce que le Big Data et quelles sont ses Applications ?	https://www.saagie.com/fr/blog/qu-est-ce-que-le-big-data-definition/
Publication en ligne et réutilisation des données publiques (« open data »)	https://www.cnil.fr/fr/publication-en-ligne-et-reutilisation-des-donnees-publiques-open-data
Open Data : cas de réutilisation	https://www.opendatafrance.net/reutilisations/
Ippon Positive Technology	https://blog.ippon.fr

Merci pour votre attention