Ministère de l’enseignement supérieur et de recherches scientifiques
Institut Supérieur des Etudes Technologiques de Radès
Département Technologies de l’ information

Cours Gestion des Données Massives

Enseignante : Mme MSAKNI Imen

AU :2022/2023

Ministère de l’enseignement supérieur et de recherches scientifiques
Institut Supérieur des Etudes Technologiques de Radès
Département Technologies de l’ information

Cours Gestion des Données Massives

Public ciblé

- L3 DSI

Déroulement

- Cours : 1,5 heure/semaine

Evaluations

- Devoir surveillé
- Examen

Description

Objectif Général

Acquérir les connaissances basiques en gestion et traitement des données massives (Big Data) ainsi que des compétences pour l'utilisation d'outils adaptés.

Chapitres

Chapitre 1 : Introduction au Big Data
Chapitre 2 : Les bases de données NoSQL
Chapitre 3 : Hadoop

Description

Mots Clés

Données ouvertes (Open Data), Objets Connectés, Bases de Données, Masse de Données (Big Data), Business Intelligence, Intelligence Artificielle, Data Science, MapReduce, NoSQL,Hadoop,MangoDB,etc.

Chapitre 1 :
Introduction au Big Data

Enseignante : Mme MSAKNI Imen

Introduction

Avec l'essor des réseaux Internet et Wi-Fi, des smartphones, des objets connectés et des réseaux sociaux, de plus en plus de données de formes variées sont générées. En parallèle, le développement d'outils de stockage et d'analyse ainsi que de nouveaux outils de visualisation permettent la valorisation de ces données structurées ou non , variées et en très grande quantité : c'est le phénomène nommé big data.

SGBDR

1955-1960

SGF

1960-1980

SGBD hiérarchiques

SGBD réseaux

1970-1990

SGBD relationnel

Langage SQL

1985-1990

SGBD objet

Depuis 1995

BD et internet

BD et applications décisionnelles

SGBDR

Avantages

- Jointures entre les tables
- Construction de requêtes complexes
- Contraintes d'intégrité solides

Limites

- Surcharge sémantique
- Représentation complexe du réel avec le relationnel.
- Jointure très lourdes
- etc

Limites dans le contexte distribué: comment distribuer/partitionner les données

- Liens entre entités -> Même serveur
- Mais plus on a de liens, plus le placement des données est complexe

SGBDR

Propriétés ACID pour les transactions

- ***Atomicité:*** une transaction s'effectue entièrement ou pas du tout
- ***Cohérence:*** le contenu d'une base doit être cohérent au début et à la fin d'une transaction (mais pas forcément durant son exécution)
- ***Isolation:*** les modifications d'une transaction ne sont visibles/modifiables que quand celle-ci a été validée.
- ***Durabilité:*** une fois la transaction validée, l'état de la base est permanent (non affecté par les pannes ou autre)

Limites dans le contexte distribué: comment distribuer/partitionner les données

- Contraintes ACID très complexes à assurer dans un contexte distribué.
- Incompatible avec les performances (disponibilité des données).

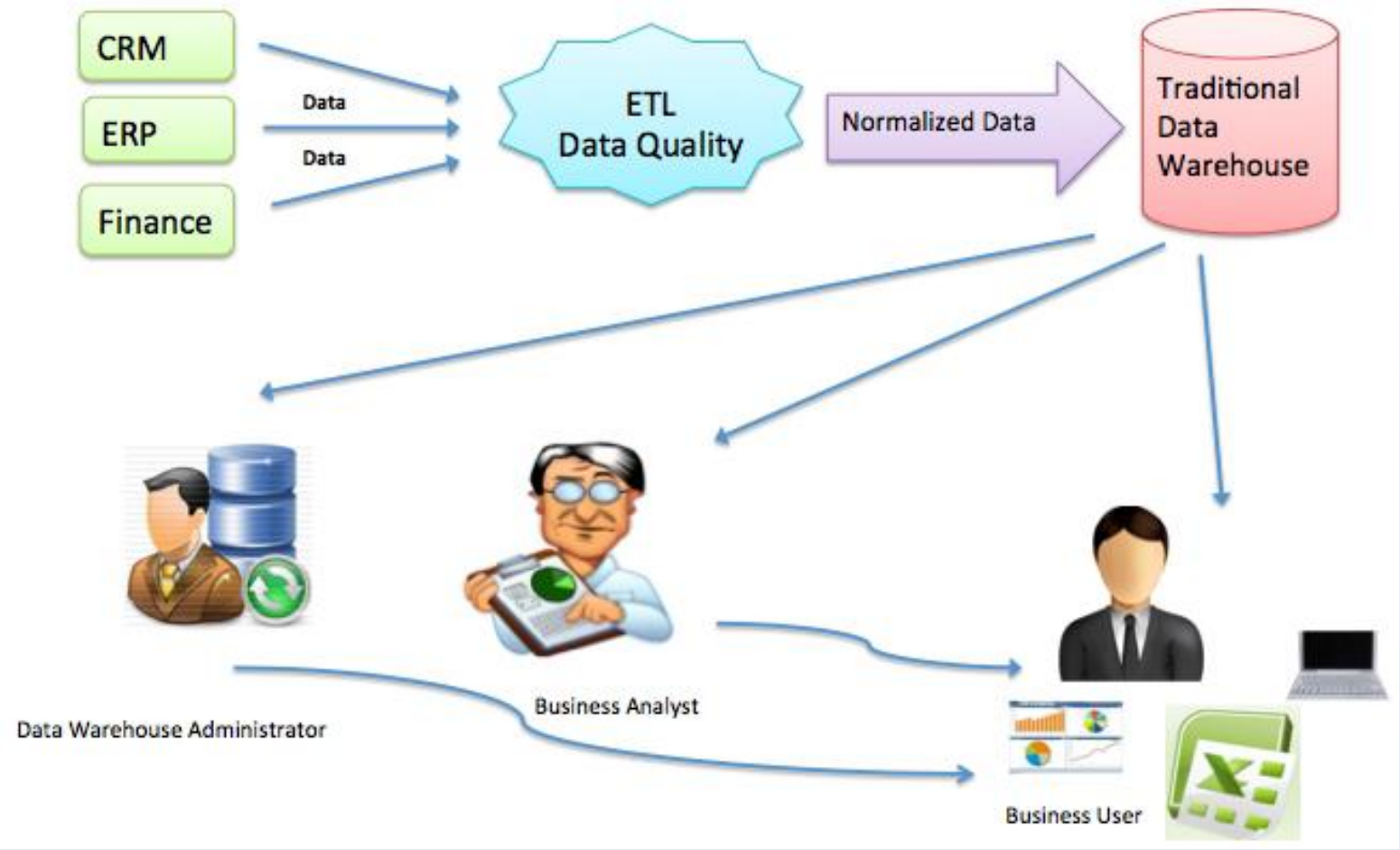
Entrepôts de Données(Décisionnel)

Le terme **Entrepôt de données** (ou base de données décisionnelle, ou encore data warehouse) désigne une **base de données** utilisée pour **collecter**, **ordonner**, **journaliser** et **stocker** des informations provenant de base de données opérationnelles et fournir une **aide à la décision** en entreprise.

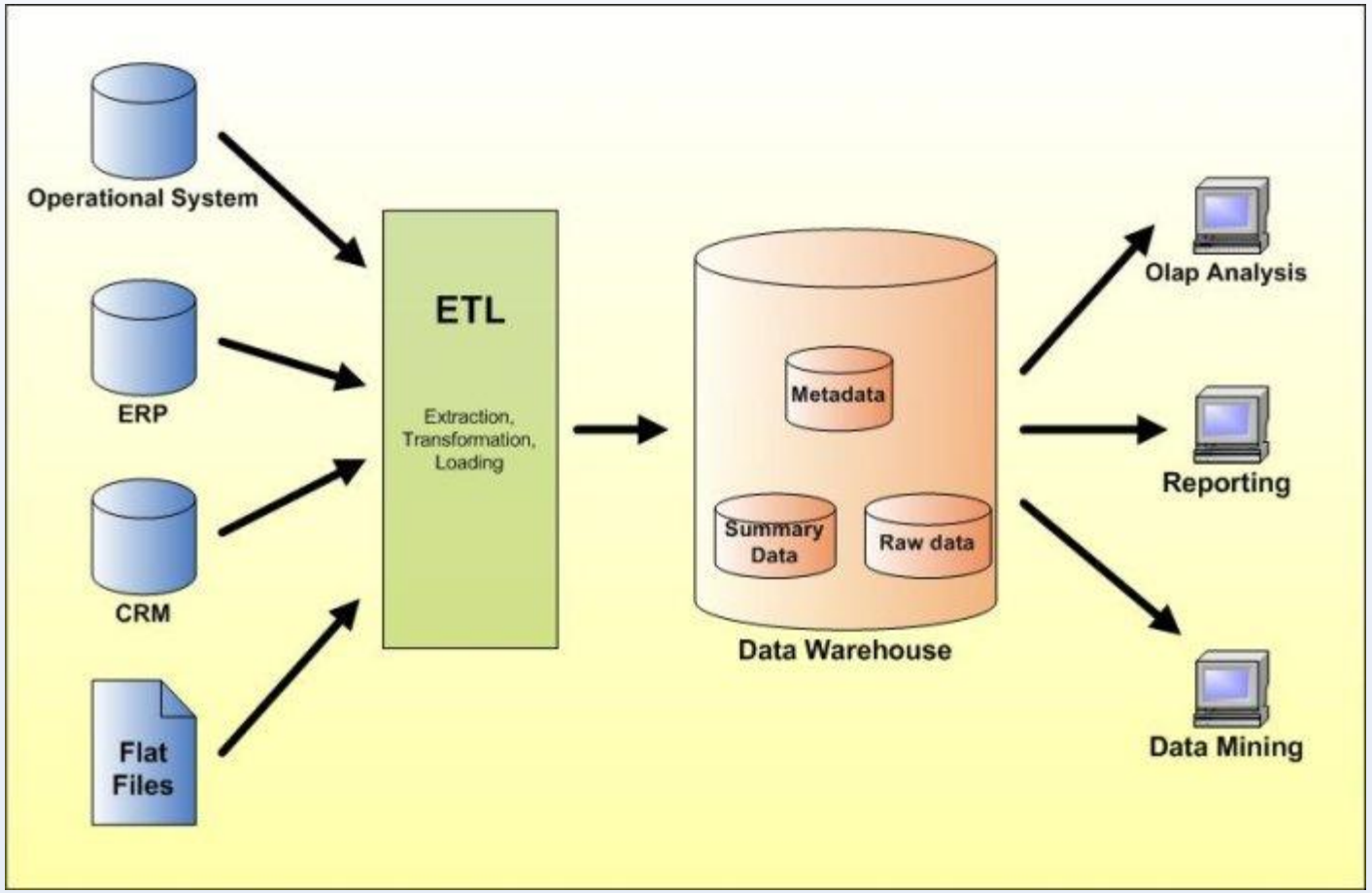
Il représente une mégabase, **thématique** le plus souvent, constituée afin d'analyser de gros volumes de données très détaillées, durables, en principe datées, et qui ont été stockées et organisées (data sourcing) sur un puissant système informatique. L'objectif est de les synthétiser de manière à en extraire l'information essentielle la plus pertinente et ainsi favoriser la prise de décision.

Le datamining constitue cette analyse permettant de passer d'une masse de détails à une synthèse exploitable. « La valeur ajoutée d'un entrepôt de données tient à la qualité des données qu'il contient. Il convient donc de ne l'alimenter qu'avec des données suffisamment fiables et cohérentes.

Entrepôts de Données(Décisionnel)



Entrepôts de Données(Décisionnel)



Décisionnel: Limites (les 3V du Big Data)

L'approche classique ne permet pas de gérer:

- Le **volume: les entrepôts sont conçus pour gérer des Go ou** To de données alors que la croissance exponentielle des données nous conduit aux Po ou Eo.
- Le **type (variety): le nombre de types, incluant les données** textuelles semi ou non structurées, augmente.
- La **vitesse (velocity): les données sont créées de plus en** plus vite et nécessitent des traitements en temps-réel.

ACID vs BASE

Systèmes distribués modernes utilisent le modèle BASE

- ***Basically Available :*** *garantie minimale pour taux de* disponibilité face à une grande quantité de requêtes.
- ***Soft-state :*** *l'état du système peut changer au cours du* temps même sans nouveaux inputs .
- ***Eventually consistent :*** *tous les réplicats atteignent le* même état, et le système devient à un moment consistant, si on stoppe les inputs

Big Data ..Pourquoi?

Pour quelles raisons ?

- Explosion de la **disponibilité des données**
- Augmentation de la **capacité de stockage**
- Augmentation de la **capacité d'analyse**

Big Data ..Pourquoi?

Disponibilité des données

En une minute sur Internet :

- Facebook : les utilisateurs éditent 3,4 millions de statuts et génèrent 4 GB de données digitales .
- Google : répond à 300000 recherches et reçoit 126 heures de vidéos. Twitter pas moins de 700 nouveaux utilisateurs rejoignent, et 350 000 Tweets sont générées.
- LinkdIn : 10 000 recherches sont exécutées .
- Alibaba : son stock de données a atteint 100 pétaoctets .
- Walmart : produit 2.5 pétaoctets par jour et traite un million de transactions par heure.

[Sakr,2016]

Big Data ..Pourquoi?

IBM

Intrinsic Property of Data ... it grows

90%
of the world's data was created in the last two years

80%
of the world's data today is unstructured

20%
of available data can be processed by traditional systems

1 in 2
business leaders don't have access to data they need

83%
of CIO's cited BI and analytics as part of their visionary plan

5.4X
more likely that top performers use business analytics

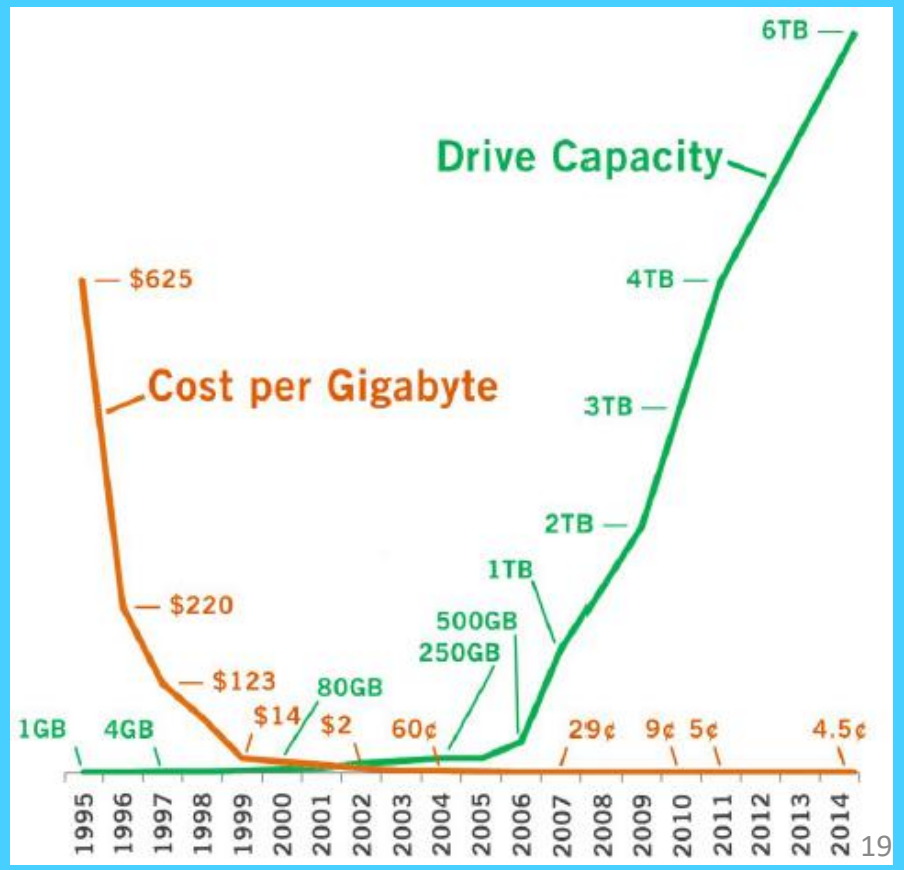
Big Data ..Pourquoi?

Data centers de quelques grands acteurs du Big Data :

- Google DataCenter :
  70000 serveurs/data center et 16 data centers, ~1M de serveurs
- Facebook :
  5 data centers
- Amazon :
  7 data centers, 450 000 serveurs
- Microsoft :
  environ 1M serveurs

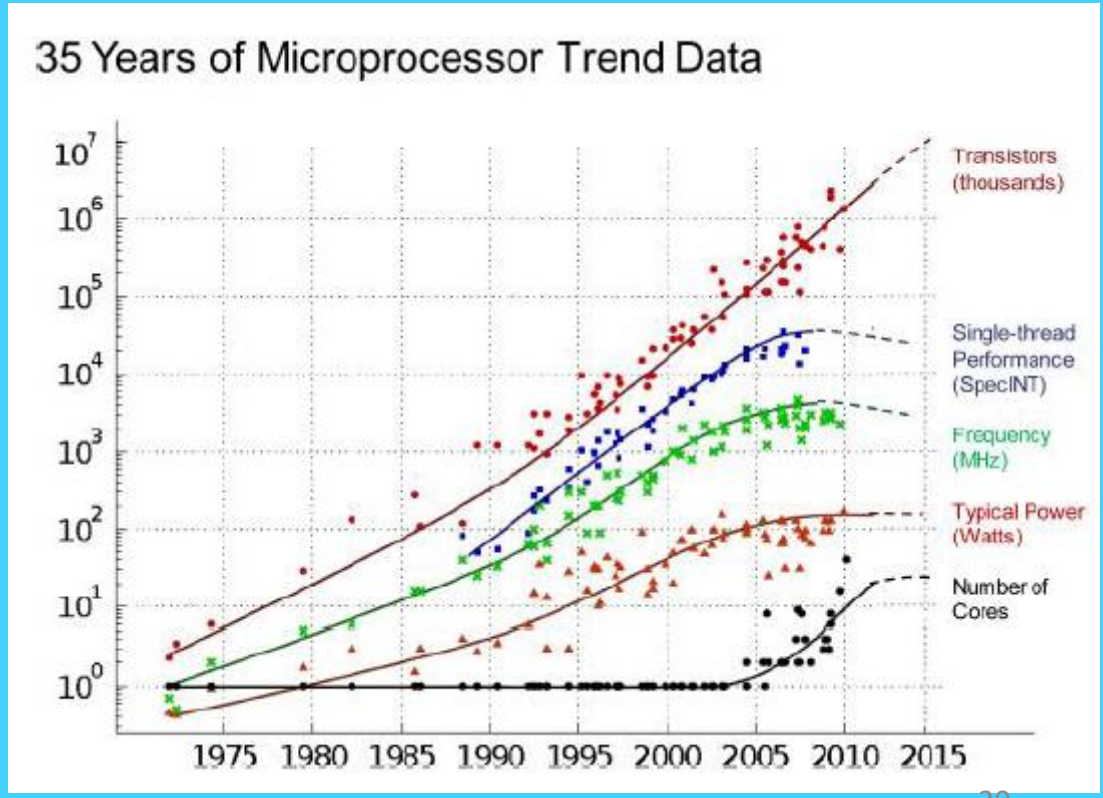
Big Data ..Pourquoi?

Capacités de stockage



Big Data ..Pourquoi?

Capacités des analyses



Big Data.. Pourquoi ?

- **Augmentation exponentielle de la quantité de données non structurées**
  - Email, chat, blog, **web, musique, photo, vidéo, capteurs,objets connectés, etc.**
- **Augmentation de la capacité de stockage et d'analyse**
  - L'utilisation de plusieurs machines en parallèle devient accessible
- **Les technologies existantes ne sont pas conçues pour ingérer ces données**
  - Base de données relationnelles (tabulaires), mainframes, tableurs (Excel), etc.
- **De "nouvelles" technologies et techniques d'analyse sont nécessaires**
  - "Google File System" - Google 2003
  - "MapReduce: Simplified Data Processing on Large Clusters" – Google,2004
  - BigTable : SGBD basé sur GFS. BigTable: A Distributed Storage System for Structured Data.
  - Hadoop: circa 2006
- **D'où le"Big Data": pas uniquement plus de data**...

Big Data – Définition 1

Big Data ou méga données, grosses données ou encore données massives, désignent un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler. En effet, on procrée environ 2,5 trillions d'octets de données tous les jours. Ce sont les informations provenant de partout : messages que nous nous envoyons, vidéos que nous publions, informations climatiques, signaux GPS, enregistrements transactionnels d'achats en ligne et bien d'autres encore.

Big Data – Définition 2

BIG DATA désigne des méthodes et des technologies « pas seulement des outils » pour des environnements évolutifs

« **augmentation** du volume de **données**, **augmentation** du nombre **d'utilisateurs**, **augmentation** de la **complexité** des **analyses**, **disponibilité rapide** des **données** » pour **l'intégration**, le **stockage** et **l'analyse** des données **structurées**, semi **structurées** et non **structurées** ».

Big Data – Définition 3

Le BIG DATA est une démarche (ou un ensemble de technologies, d'architectures, d'outils et de procédures) qui consiste à collecter puis à traiter en temps réel d'énormes volumes de données, proviennent de sources diverses, structurées et non structurées, difficilement gérables avec des solutions classiques de stockage et de traitement.

Big Data – Données

Big data =

Données

Massives

Sources de données multiples :

Big Data – Données

Big data =

Données Distribuées

Les quantités de données à stocker sont tellement importantes qu'il est inenvisageable d'utiliser une seule machine/disque

Les données sont acquises à des endroits différents. Leur transfert prendrait trop de temps

Big Data – Données

Big data =

Données peu ou non Structurées

- Données Structurées
- Données peu-structurées
- Données non-structurées

Big Data – Données

Big data =
Data Stream

- Les données arrivent en continu
- Les traiter efficacement revient à les prendre en charge quand elles arrivent
- Proposer des algorithmes qui n'ont pas besoin de faire plusieurs passes sur les données

Big Data – Données

Big data =
Visualisation

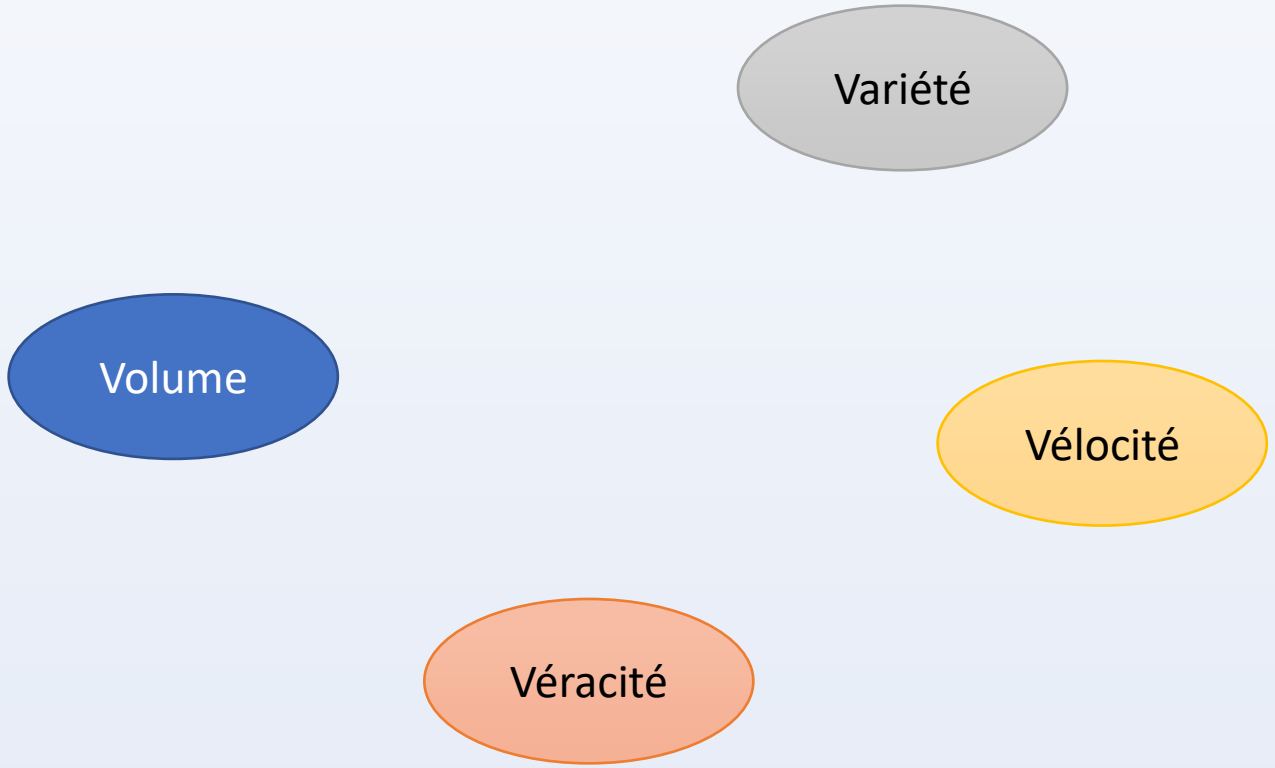
- Non seulement les données sont massives mais même les résultats le sont aussi.
- Comment appréhender/analyser/inférer la connaissance?
- Visualiser: une image vaut mille discours

Big Data – Les 4 V

Big data =

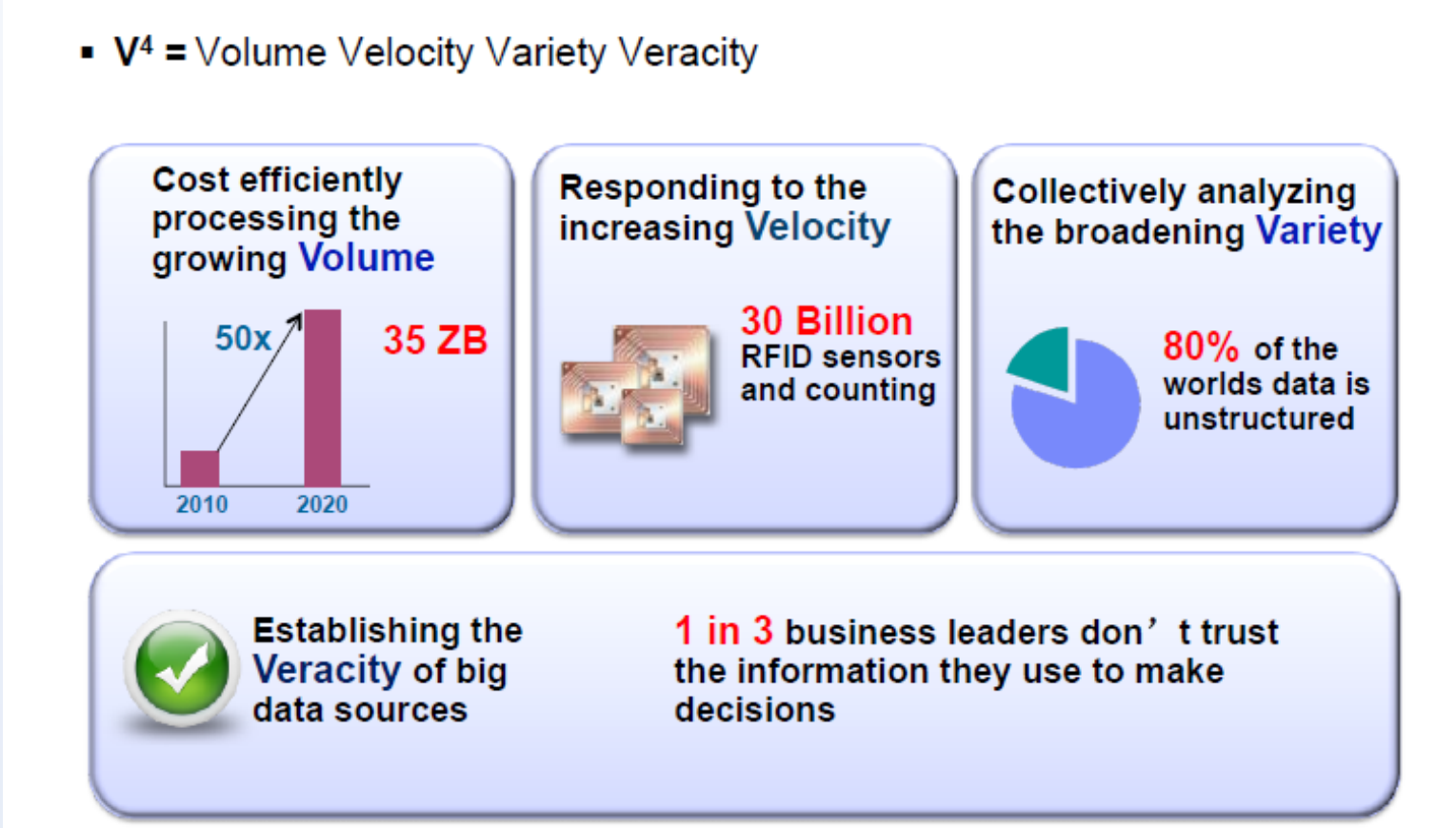
Les 4 V

V4 = Volume Velocity Variety Veracity



Big Data – Les 4 V

Big data =
Les 4 V



Source : IBM

Big Data : Les 4 V

Volume

- Le volume de données créé et géré par les entreprises est en constante augmentation.

Données d'origine brute , Térabytes à Exabytes de données sont disponibles.

Big Data : Les 4 V

Vélocité (Vitesse)

- Une des grandes forces du BIG DATA, c'est de pouvoir utiliser les données à mesure qu'elles sont collectées. Mais pour cela, il faut une puissance de calcul et des outils d'analyses très performants. Car l'immense majorité de données collectées par le BIG DATA ont une valeur ajoutée forte qui s'étiole rapidement avec le temps. Il faut les utiliser rapidement, ou elles n'ont aucune valeur.

Données dynamiques, décisions en temps réels

Big Data : Les 4 V

Variété (Type de données)

- Données structurées ou non structurées ; texte, données de capteurs, son, vidéo, données de géolocalisation, fichiers journaux ; les données collectées par le BIG DATA sont de nature très diverse. Il faut donc que les outils de traitement soient à même de prendre en charge cette diversité, pour donner du sens.

Données hétérogènes , formats non structurés, textes , images

Big Data : Les 4 V

Véracité

- La véracité fait référence à la provenance ou à la fiabilité de la source de données, à son contexte et à son importance pour l'analyse qui en découle. La connaissance de la véracité des données nous aide à mieux comprendre les risques associés aux analyses et aux décisions commerciales basées sur cet ensemble de données particulier.

Données incertaines, cohérence, fiabilité, qualité et prédictibilité des données

Big Data et Business Intelligence

Qu’est ce que le Business Intelligence ou l’Informatique Décisionnelle ?

Il s’agit de l’ensemble des techniques et des outils permettant de collecter les données brutes pour les modéliser et les restituer sous forme d’informations pertinentes et utiles pour l’analyse “business”.

Il présente une aide décisionnelle précieuse pour les gestionnaires.

Big Data et Business Intelligence

	Big Data	BI
Charge de Travail	Ad-Hoc	Préparée
Types de Données	Brutes	Structurées
Sources de Données	Externes et Opérationnelles	Opérationnelles

Big Data et Intelligence Artificielle

- La donnée est considérée comme l'or noir de l'IA. L'intelligence Artificielle est la suite logique de nos méthodes et techniques d'analyses des données, le prolongement de la business Intelligence, qui a été suivie du Big Data et de l'Advanced Analytics. La machine dite intelligente a besoin d'une quantité très importante de données analysées et croisées entre elles pour en tirer des enseignements novateurs, voire créatifs, proches du fonctionnement du cerveau humain.
- L'intelligence artificielle va être utilisée pour extraire du sens, déterminer de meilleurs résultats, et permettre des prises de décisions plus rapides à partir de sources Big Data massives.
- Le Big Data,est une source d'apprentissage pour l'apprentissage en IA.

Big Data et Data Science

- La Data Science consiste à mettre au point des séries d'algorithmes à partir de règles mathématiques et statistiques (ou de Machine Learning) afin de délivrer des solutions. Ces techniques peuvent s'appuyer sur l'analyse d'image, l'analyse de textes (text-mining), l'étude de corrélation entre capteurs, etc.
- Le lien entre les deux apparaît alors : afin d'analyser les amas de données, le Big Data repose sur les algorithmes développés par la Data Science.

Big Data : Applications

Descriptives : Que s'est-il passé ? Pourquoi ?

Prédictives : Que va-t-il se passer?

Prescriptives: Comment atteindre l'objectif ?

Big Data : Applications

Big Data Exploration

Trouver, Visualiser, comprendre toutes les données massives pour aider a la prise de décision

Vue 360° du consommateur

Comprendre les consommateurs pour une meilleure prédiction de leurs achats

Sécurité/Iintelligence

Minimiser les risques , détecter les fraudes et gérer la cyber-sécurité en temps réel

Analyses et analyses avancées

Analyser des variétés de données pour aider les décideurs

DW Augmentation

Intégrer le Big Data et le DW pour améliorer l'efficacité des opérations

Source : IBM

Big Data : Applications

Big Data et RH, les nouveaux outils de recrutement

- Les départements RH ont trouvé le moyen d'utiliser le Big Data afin de **mieux cibler leur recrutement.**
- Ils peuvent désormais **sélectionner les profils les plus intéressants grâce aux données récoltées.** Ces données sont aujourd'hui disponibles tout autour de nous. L'analyse statistique des données des réseaux sociaux, notamment LinkedIn, Google , Facebook et Twitter ainsi que les bases de données en ligne (Open Data) permettent de cibler les comportements et de repérer les candidats potentiels en amont d'un recrutement. Des outils d'analyse permettent d'**établir un profil psychologique grâce aux « like » sur Facebook** par exemple.

Big Data : Applications

Big Data et la Finance

- Les services financiers des entreprises permettent d'utiliser le Big Data **afin de réduire les risques et les coûts**, **d'identifier les opportunités et d'améliorer la précision des prévisions**. Par exemple, l'analyse des clients permettra de cibler les éléments à risque, de même avec les fournisseurs. Le Big Data fait également partie intégrante du développement des entreprises puisqu'il permet d'analyser un marché, d'établir de nouveaux business model.

Big Data : Applications

Big Data et le Marketing prédictif, connaître les attentes de sa cible

- Le Big Data est une révolution pour les services marketing, les solutions existantes commencent à un prix de départ élevé mais le retour sur investissement est quant à lui important. On parle de **marketing prédictif**, cette nouvelle méthode permet aux marketeurs une analyse approfondie de leurs clients mais également de l'**efficacité de leurs campagnes.** Cela permet notamment d'augmenter le taux d'ouverture d'une campagne d'emailing en ciblant et personnalisant une newsletter avec les données récoltées en amont.

Big Data : Applications

Big Data et la Maintenance prédictive ou le renouveau de la maintenance industrielle

- La maintenance prédictive permet de réduire les coûts de maintenance des installations industrielles grâce au Big Data. En effet, grâce à la mise en place de capteurs sur les machines, les entreprises peuvent désormais utiliser les données pour **planifier l'entretien des machines.**

Big Data : Applications

Big Data et Médecine

- Identifier des facteurs de risque de développement de certaines maladies comme les cancers ou le diabète et de mettre en place des outils de prévention cohérents et par conséquent plus efficaces ;
- Développer des systèmes d'aide au diagnostic et des outils permettant la personnalisation des traitements ;
- Vérifier l'efficacité de certains traitements, et d'effectuer une veille sanitaire, afin notamment d'identifier d'éventuelles complications ; de prédire la survenue d'épidémies.

- Le big data est donc une aide précieuse à la conduite des politiques de santé, pour l'optimisation du système de soins.

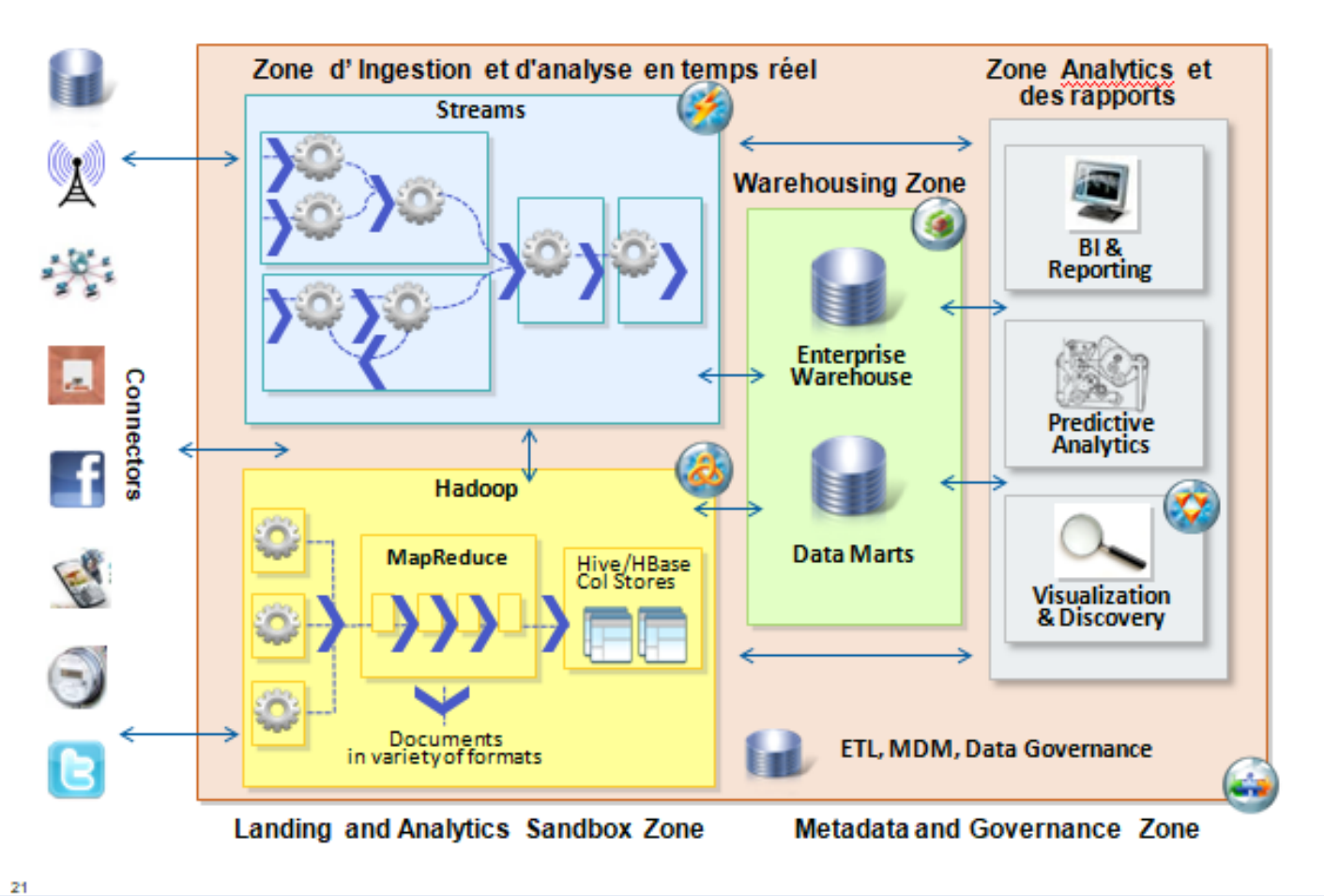
Big Data : Autres Applications

Source : IBM

Big Data ….Gains

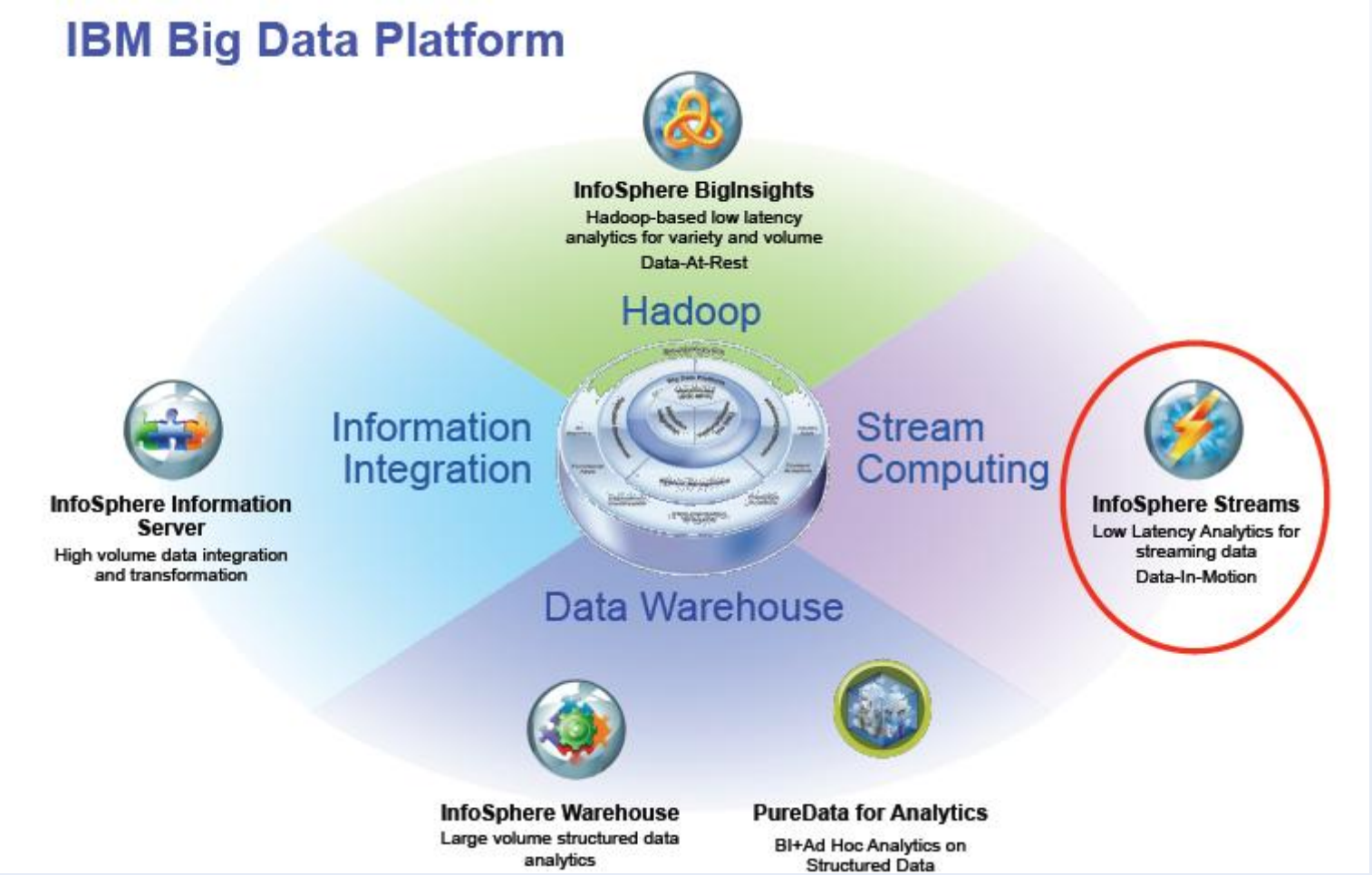
Gains	Entreprises ayant constaté un gain
Meilleure habilité à prendre des décisions stratégiques	69 %
Meilleure gouvernance opérationnelle	54 %
Meilleure connaissance et amélioration de l'expérience utilisateur	52 %
Réduction des coûts	47 %
Accélération des décisions	44 %
Développement d'un nouveau produit/service	43 %
Meilleure connaissance du marché et des concurrents	41 %
Développement d'un nouveau business model	38 %
Augmentation des revenus	35 %
Automatisation des décisions	24 %

Exemple de plateforme Big Data



Source : Hafedh FERCHICHI

Exemple de plateforme Big Data



Big Data .. Quelques solutions