Enseignant : Mohamed MANAA

IBM Big Data Trainer

«Big Data Specialist» Certified

Année Universitaire : 2020-2021

Plan

I. Concepts de base du Big Data

II. Architecture de la plateforme Big Data

III. Architecture Hadoop

IV. Système HDFS

V. Paradigmes de Traitement parallèle «MapReduce»

VI. Hadoop Query Languages

VII. Hbase

VIII. Hive

IX. Big SQL

X. JAQL

XI. Système Analytique «AQL»

XII. BigSheets

I. CONCEPTS DE BASE DU BIG DATA

I. Concepts de base du Big Data

La maturité du BigData conséquence de plusieurs disciplines

GRID Computing : Calcul parallèle et distribué, HPC (High Performance Computer), capacité de calcul haute performance

Cloud Computing : Capacité de stockage infini, réparti et sécurisé, fragmentation/réplication

Internet of Things (IoT) : Ubiquitous Computing (informatique ambiante)

Multitudes de devices connectés (plages IPV6 suffisantes)

Exemples : les web services façades de tout objet pingable (caméra, capteur, etc.), La voiture comme ordinateur ambulant, la télé-maintenance proactive, la traçabilité (RFID), le tracking par GPS, etc.

Web 3.0 (Social, Sémantique) : SNA (Social Network Analysis)

Data Management : SQL, noSQL, DWH (datawarehousing), BI (Business Intelligence)

NLP (Natural Language Processing)

I. Concepts de base du Big Data

90%

des données ont été créées au cours des sept dernières années

80%

des données actuelles ne sont pas structurés

20%

des données disponibles peuvent être traitées par des systèmes traditionnels

1 in 2

des chefs d'entreprise n'ont pas accès aux données dont ils ont besoin

83%

des responsables des technologies de l'information et de la communication citent BI et analyses des données comme part de leurs plan prévisionnel

5.4X

plus probable que les plus performants utilisent business analytics

I. Concepts de base du Big Data

Un monde interconnecté et instrumenté

500+ Million **users posting 55 Million tweets every day**

1.2 Trillion **searches**

1+ Billion **active users spending 700 Million minutes per month**

30 billion **RFID tags today (1.3B in 2005)**

4.6 billion **camera phones world wide**

100s of millions of GPS enabled **devices sold annually**

2+ billion **people on the Web by end 2011**

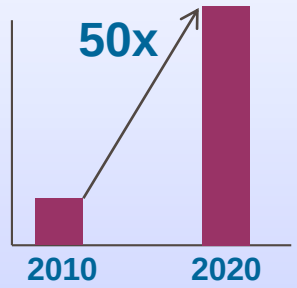
76 million **smart meters in 2009… 200M by 2014**

I. Concepts de base du Big Data

Caractéristiques du Big Data

V^4 = Volume Velocity Variety Veracity

Coûts de traitement efficace du Volume croissant



35 ZB

Répondre à la Vitesse croissante Velocity

30 Billion Capteurs RFID et comptage

Analyser collectivement l'élargissement de la Variété Variety

80% des données mondiales ne sont pas structurés

Établissement de la Véracité des grandes sources de données Veracity

1 à 3 des chefs d'entreprise ne font pas confiance à l'information qu'ils utilisent pour prendre des décisions

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

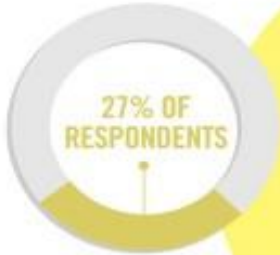
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

I. Concepts de base du Big Data

Les 5 principaux cas d'utilisation de données clés

Exploration du Big Data

Trouver, visualiser, comprendre toutes les grandes données pour améliorer la prise de décision

Vue 360 ° améliorée du client

Étendre les vues des clients existantes en intégrant des sources de données internes et externes supplémentaires

Extension de sécurité / intelligence

Risque plus faible, détection de la fraude et suivi de la cybersécurité en temps réel

Analyse d'opérations

Analyser une variété de données machine pour améliorer les résultats commerciaux

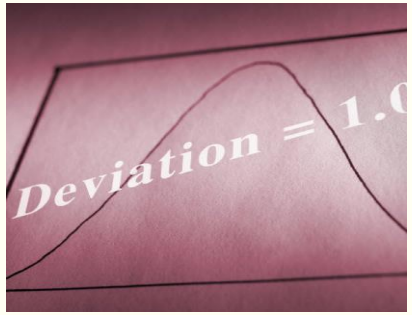
Augmentation du stockage de données

Intégrez big data et data warehouse pour accroître l'efficacité opérationnelle

I. Concepts de base du Big Data

Plus de façons & Des analyses et des techniques très variées

Spatial Analysis



Statistics

Text Analysis

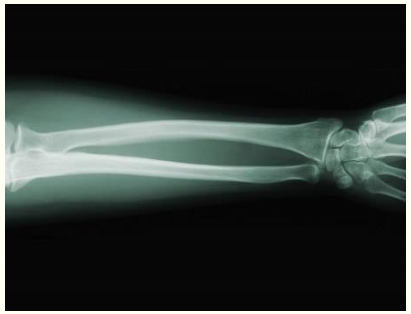


Image Analysis

Temporal Analysis

Machine Learning

Video Analysis

Audio Analysis

I. Concepts de base du Big Data

Big Data et complexité dans la santé

- Les informations médicales doublent tous les 5 ans, dont une grande partie n'est pas structurée
- 81% des médecins signalent des dépenses de 5 heures par mois en consultant des revues médicales

"Medicine has become too complex (and only) about 20 percent of the knowledge clinicians use today is evidence-based"

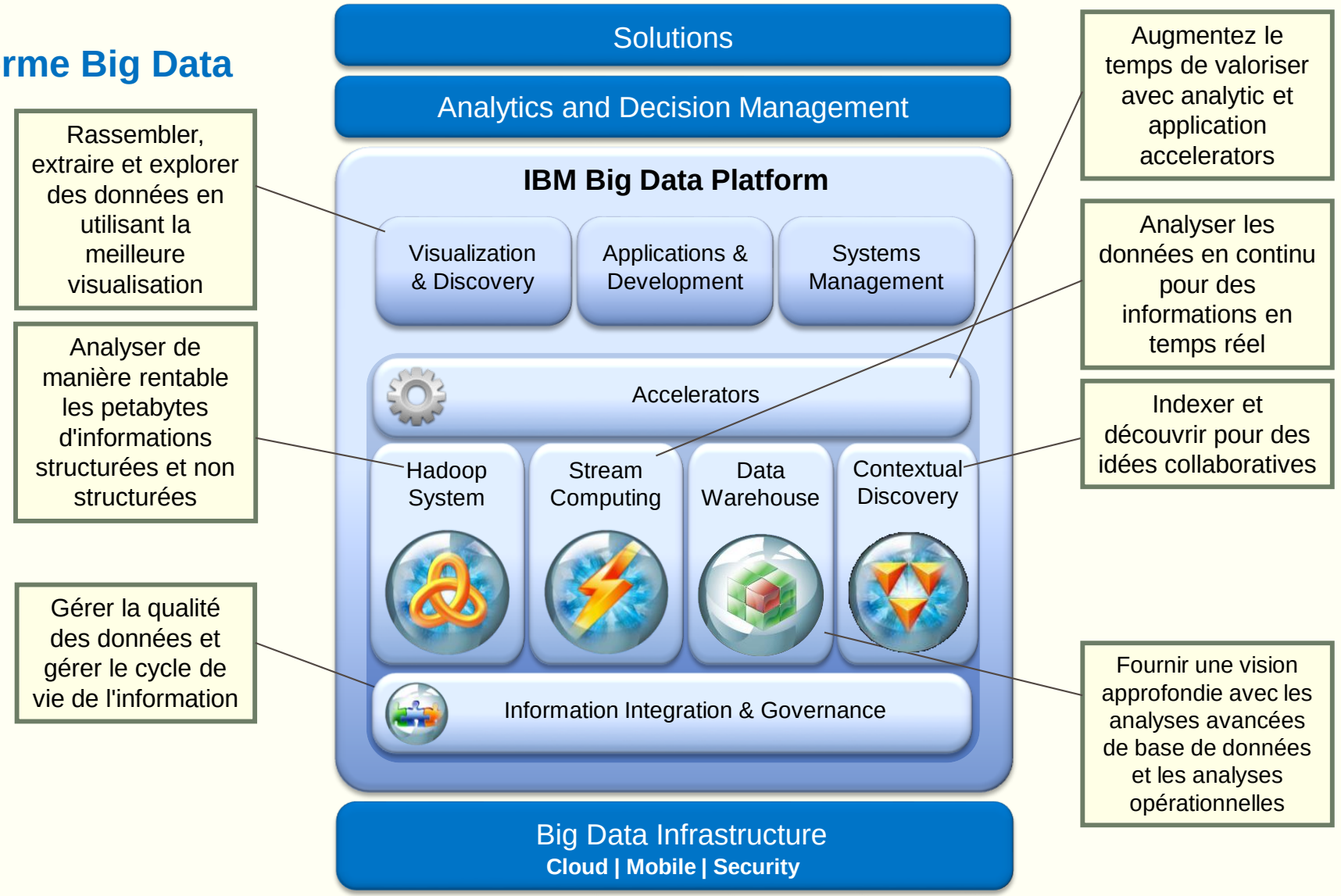
- Steven Shapiro, Chief Medical and Scientific Officer, UPMC

…to keep up with the state of the art, a doctor would have to devote 160 hours a week to perusing papers…"

The Economist Feb 14th 2013

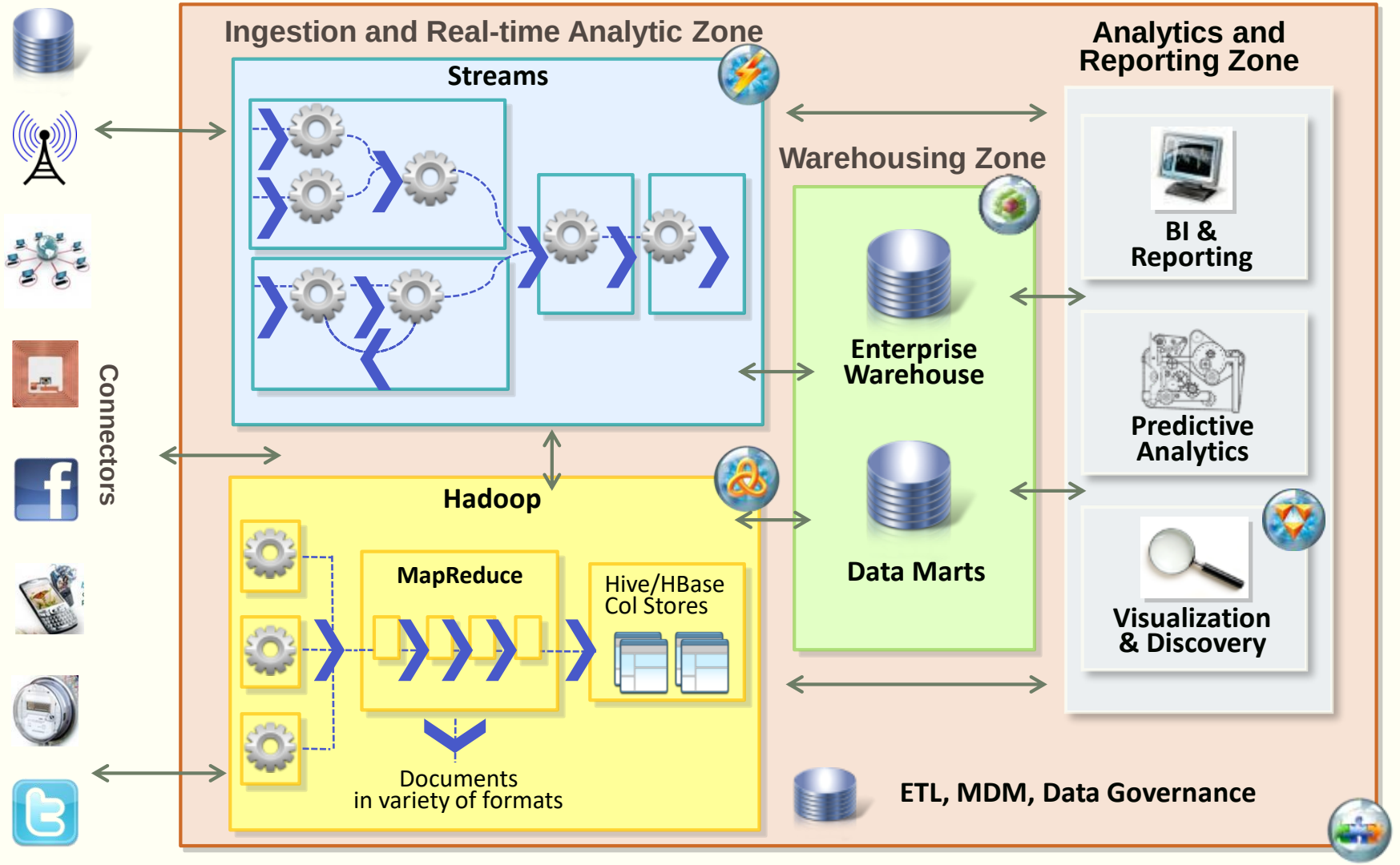
I. Concepts de base du Big Data

Plate-forme Big Data



I. Concepts de base du Big Data

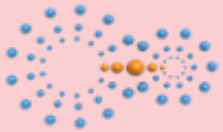
Un exemple de la plate-forme Big Data en pratique



I. Concepts de base du Big Data

Manifeste de Big Data

CONSOMMABLE
(What not How/Patterns/Expert Systems, +++)

Comprendre et naviguer dans les sources Big Data fédérées		Federated Discovery and Navigation
Gérer et stocker un énorme volume de données		Hadoop File System MapReduce
Structure et contrôle de données		Data Warehousing
Gérer les données en streaming		Stream Computing
Analyser les données non structurées		Text Analytics Engine
Intégrer et gérer toutes les sources de données		Integration, Data Quality, Security, ILM, MDM

MDM : Master Data Managment

ILM : Information Lifecycle Managment

I. Concepts de base du Big Data

Exemple d'utilisation de la plate-forme Big Data (1)

Services financiers

- Problème:
  - ✓ Gérer les plusieurs Petabytes de données qui augmente de 40 à 100% par an sous pression croissante pour prévenir les fraudes et se plaindre de la réglementation.
- Comment les grandes analyses de données peuvent-elles aider:
  - ✓ Détection de fraude
  - ✓ Gestion des risques
  - ✓ Vue à 360 ° du client

I. Concepts de base du Big Data

Exemple d'utilisation de la plate-forme Big Data (2)

Services de télécommunication

- Problème:
  - ✓ Les anciens systèmes sont utilisés pour obtenir des informations sur les données produites en interne qui font face à des coûts de stockage élevés, à un long temps de chargement des données et à un long processus d'administration.
- Comment les grandes analyses de données peuvent-elles aider:
  - ✓ Traitement CDR
  - ✓ Analyses Prédictives Fiables
  - ✓ Geomapping / marketing
  - ✓ Surveillance du réseau

I. Concepts de base du Big Data

Exemple d'utilisation de la plate-forme Big Data (3)

Services de transport

- Problème:
  - ✓ La congestion du trafic a augmenté dans le monde grâce à une urbanisation accrue et à une croissance démographique réduisant l'efficacité des infrastructures de transport et augmentant le temps de déplacement et la consommation de carburant.
- Comment les grandes analyses de données peuvent-elles aider:
  - ✓ Analyse en temps réel des flux de données de congestion météorologique et de trafic pour identifier les tendances de trafic réduisant les coûts de transport.