Introduction to BigData

OUARED Abdelkader
Teaching assistant at Ibn Khaldoun University
a_ouared@esi.dz
https://www.linkedin.com/in/abdelkader-ouared/

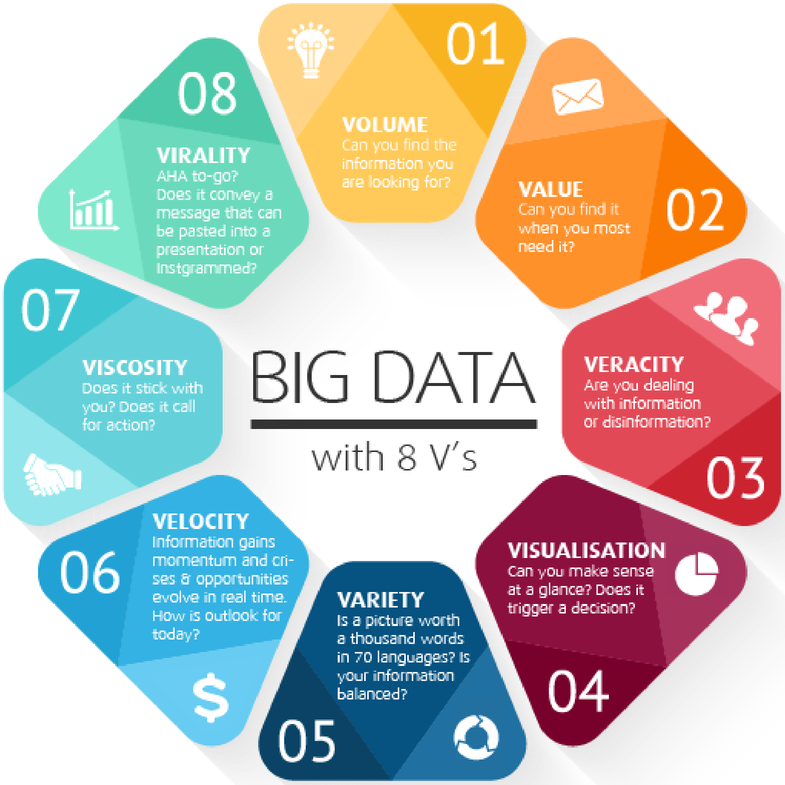
Why we need big data in modern world ?

- Every day we generate **2.5 trillion** bytes of data
- Source:
  - Sensors used to collect climate information
  - Social media posts
  - Digital images and videos published online
  - Transactional online purchase records
  - GPS signals from mobile phones
- This Data is called **BigData**

What are the challenges of big data ecosystem?

- **Bring together a large volume of varied data to find new ideas**
    - **Difficulty saving all of this data**
    - **Difficulty in processing and using this data**
    - **Data is created quickly**

The eight Vs Of big data



BIG DATA

with 8 V's

01 VOLUME
Can you find the information you are looking for?

02 VALUE
Can you find it when you most need it?

03 VERACITY
Are you dealing with information or disinformation?

04 VISUALISATION
Can you make sense at a glance? Does it trigger a decision?

05 VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?

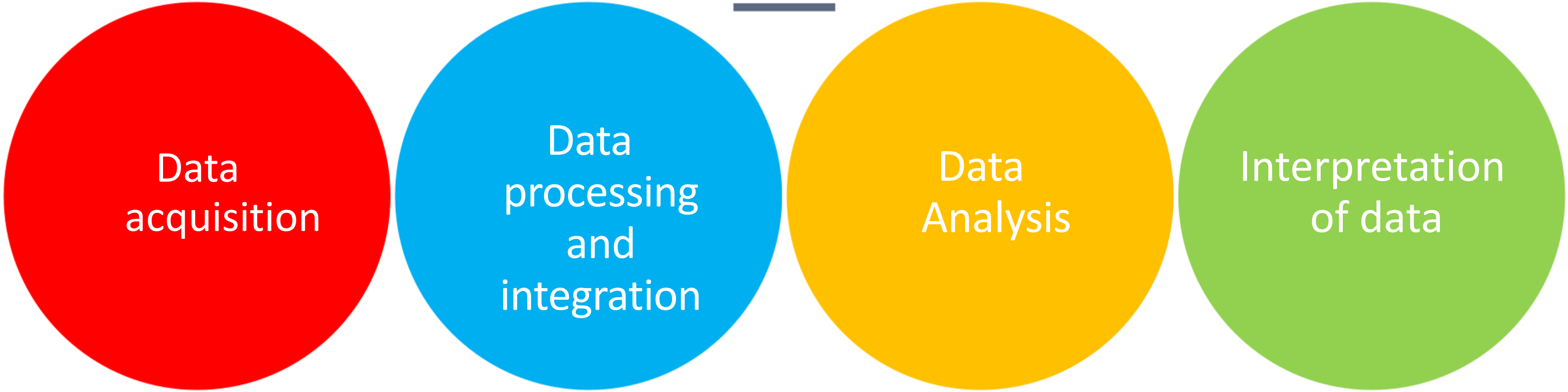
06 VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?

07 VISCOSITY
Does it stick with you? Does it call for action?

08 VIRALITY
AHA to-go?
Does it convey a message that can be pasted into a presentation or Instgrammed?

Data processing

(Agrawal, Bernstein et al. 2011)



Data acquisition

- Acquisition and filtering
- Metadata generation

Data processing and integration

- Integration
- Data aggregation

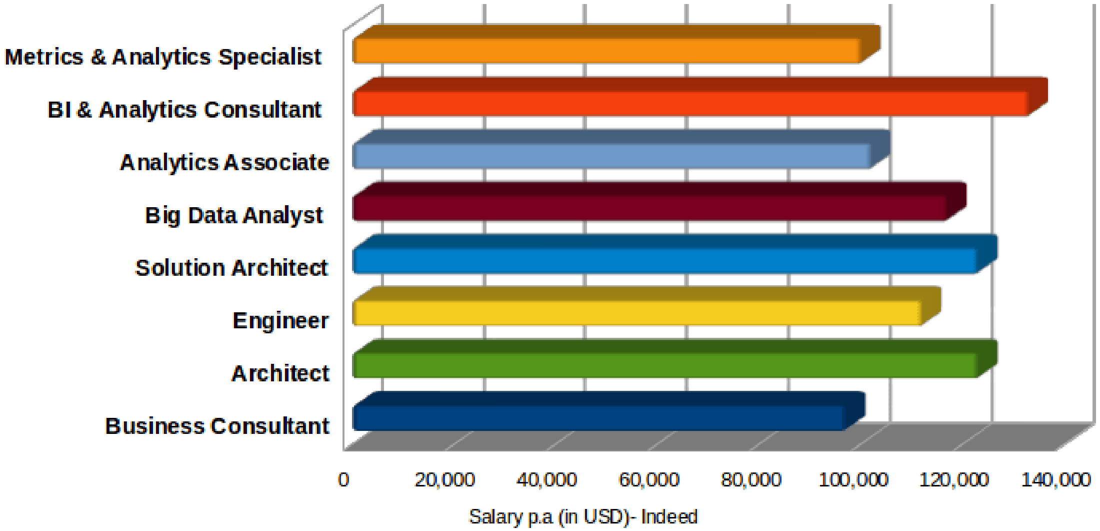
Data Analysis

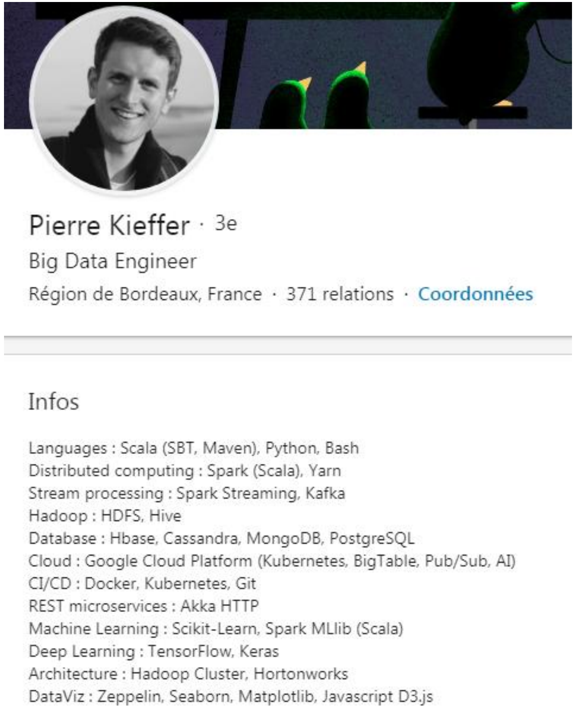
- improve the quality and data reliability
- understand the semantics

Interpretation of data

- Data visualization
- Human machine interaction

Big Data Analytics Job Titles & Salaries





Pierre Kieffer · 3e

Big Data Engineer

Région de Bordeaux, France · 371 relations · Coordonnées

Infos

Languages : Scala (SBT, Maven), Python, Bash
Distributed computing : Spark (Scala), Yarn
Stream processing : Spark Streaming, Kafka
Hadoop : HDFS, Hive
Database : Hbase, Cassandra, MongoDB, PostgreSQL
Cloud : Google Cloud Platform (Kubernetes, BigTable, Pub/Sub, AI)
CI/CD : Docker, Kubernetes, Git
REST microservices : Akka HTTP
Machine Learning : Scikit-Learn, Spark MLlib (Scala)
Deep Learning : TensorFlow, Keras
Architecture : Hadoop Cluster, Hortonworks
DataViz : Zeppelin, Seaborn, Matplotlib, Javascript D3.js

Example: job offer

Algoptis Recrute Actuellement à la recherche des profils Java EE confirmés/ **Big Data** / .Net / PHP hashtag Merci de postuler sur emna.maaoui@algoptis.fr en indiquant dans l'objet de mail : Votre Pays / Profil Exemple : Algérie / IED JAVA

Example: job offer

Madame,

Dans le cadre de développement de projets en JEE bigdata, ma société vous sollicite pour publier une annonce de recrutement via le site web de l'ESI.

Lieu de travail: Beb EZZOUAR
Pré requis: Java, JEE tomcat, spring,base de données
Salaire selon compétences

Je suis disponible sur Alger cet semaine jusqu'a jedui matin pour convenir des entretiens.

Cordialement
Ghanem BENAZZOUZ
Mobile: +213 697 563 995
Mail: gb@aigs.eu

Example: job offer

We are an international consultancy partnering with clients to chart a path through the ever-changing life sciences industry. Our people are thought leaders with a broad range of therapeutic insights and deep local market knowledge who have unique access to gold-standard data. Our evidence-based solutions and strategic insights enable life sciences leaders to readily take action and make key business decisions.

Responsibilities:

- Provides high quality, timely development and on-time input to client solutions for the pharmaceutical and related industries. Assignments typically require basic analysis and problem solving.
- Under direct supervision, assists with the review and analysis of client requirements or problems and the development of proposals of cost effective solutions.
- Assists Analysts and Consultants in developing detailed documentation and specifications. Under close supervision, performs basic quantitative or qualitative analyses to assist in the identification of client issues and the development of client specific solutions.
- Assists Analysts and Consultants in design and structures of presentations that are appropriate to the characteristics or needs of the audience.
- Proactively develops a basic knowledge of consulting methodologies and the pharmaceutical market through the delivery of consulting engagements and participation in formal and informal learning opportunities.
- Engagement based responsibilities are assigned and closely managed by consultants, engagement managers or principals.

Experience Required and background:

- 0-3 years since achieving an undergraduate degree from a recognized educational institution.
- Demonstrable analytical, interpretative and problem-solving skills
- Well-developed written and verbal communication skills including presentations, meeting and workshop facilitation
- Strong capability in juggling priorities so that deadlines are met while retaining consistently high-quality outcomes
- Must have the ability to work with team globally.
- Adjust schedule based upon projects work. Excellent interpersonal skills and ability to work effectively with others in and across the organization to accomplish team goals
- Adaptability and an ability to learn quickly and apply new knowledge
- Basic understanding of SQL, Database management, **Big Data is a plus**

Velvet Consulting | Data Scientist | Ingénieur Big Data

Velvet Consulting est un cabinet de conseil en management, spécialisé sur les domaines du **Marketing, de la Vente et de la Relation Client**.

Nous souhaitons recruter un **Ingénieur Big Data | Data Scientist** afin de renforcer nos équipes **Data Science**.

H/F diplômé(e) d'une **Grande Ecole d'Ingénieur Informatique** ou d'un **cursus universitaire en informatique / Data Science**, vous justifiez d'une première expérience (1 an minimum pour les Consultants Confirmés et 4 ans minimum pour les Consultants Seniors) sur l'analyse, la conception et la mise en œuvre de solutions **Big Data**. Vous avez mené des premiers projets de bout en bout faisant appel à votre expertise informatique et machine learning.

Dans le cadre de vos missions vous avez acquis les compétences suivantes :

Compétences techniques

Distributions : Big Insight, Cloudera
Programmation : Java / J2EE, XML, Python, R
OS : Linux / Unix
Soutions Big Data : Hadoop, Oozie, Pig, Hive, Impala, Flume, HDFS, Mahout
Frameworks Big Data : Storm, Spark
BDD : Relationnelles (Oracle) & NoSQL, Cassandra, MongoDB

Compétences mathématiques
Premières expériences en analyse statistique, machine learning, textmining…

Vous trouverez en pièce jointe une description complète du poste et des compétences recherchées.

N'hésitez pas à nous envoyer votre candidature à rh@velvetconsulting.com et nous nous ferons un plaisir de vous rappeler pour en discuter avec vous plus en détails.

Big Data Requirements

Storage Requirements

Processing Requirements

NoSQL

NoSQL systems use a distributed file system

+

MapReduce

Big data Solution

Big Data Requirements

Storage Requirements

NoSQL

Processing Requirements

MapReduce

PERSISTENCE - DATA

A Brief History…

Relational Database

Relational Database

SQL > Good for:

(ACID)

Atomicity, Consistency, Isolation, and Durability

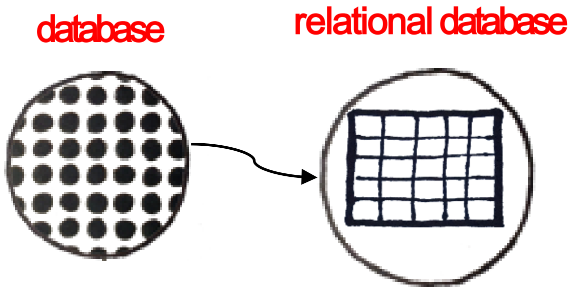
SQL> Good for:

Materials Cloud

The Law of Relational Database

By HikingArtist.com

If the only tool you have is a relational database, everything looks like a table.



database

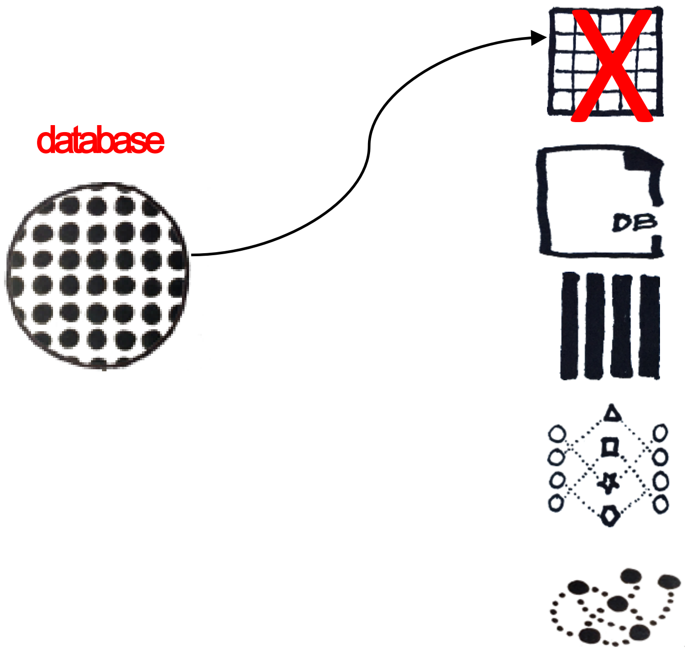
relational database

One-size-fits-all:
An Idea Whose Time Has Come and Gone

If the only tool you have is a hammer,
everything looks like a nail.

Abraham Maslow - The Psychology of Science - 1966

From Tabular Data To Other Data Model

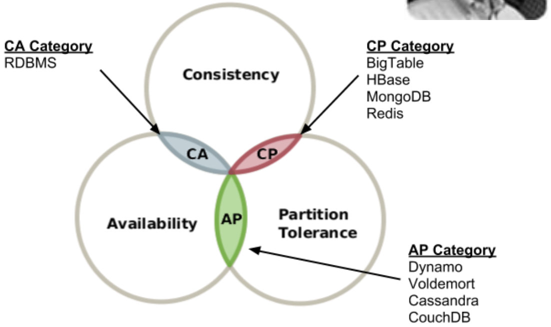


database

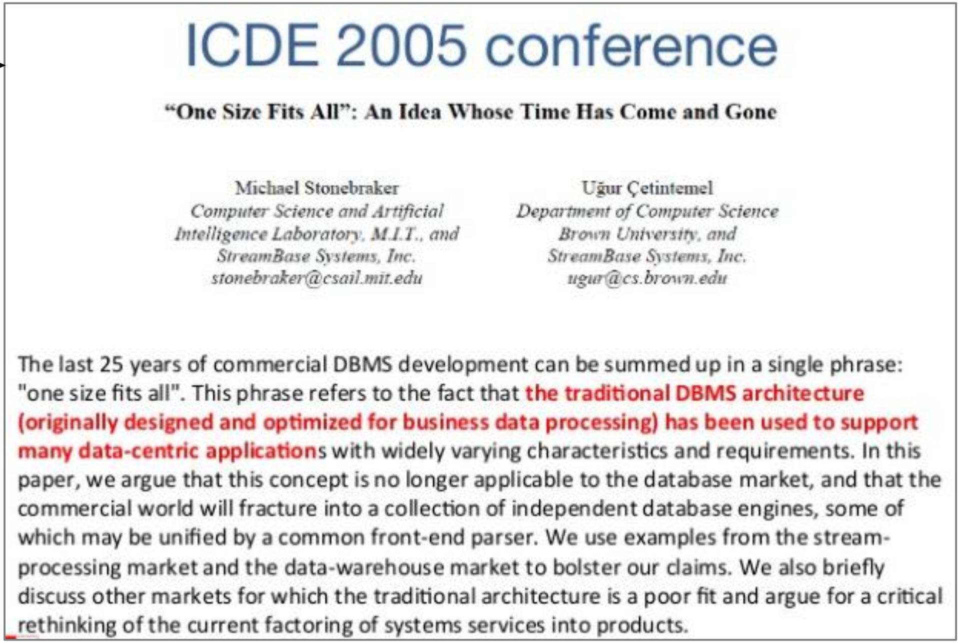
CAP theorem

NoSQL: Not Only SQL

Why NoSQL?



We can not achieve all the three items
In distributed database systems (center)



ICDE 2005 conference

"One Size Fits All": An Idea Whose Time Has Come and Gone

Michael Stonebraker
Computer Science and Artificial Intelligence Laboratory, M.I.T., and StreamBase Systems, Inc.
stonebraker@csail.mit.edu

Uğur Çetintemel
Department of Computer Science Brown University, and StreamBase Systems, Inc.
ugur@cs.brown.edu

The last 25 years of commercial DBMS development can be summed up in a single phrase: "one size fits all". This phrase refers to the fact that **the traditional DBMS architecture (originally designed and optimized for business data processing) has been used to support many data-centric applications** with widely varying characteristics and requirements. In this paper, we argue that this concept is no longer applicable to the database market, and that the commercial world will fracture into a collection of independent database engines, some of which may be unified by a common front-end parser. We use examples from the stream-processing market and the data-warehouse market to bolster our claims. We also briefly discuss other markets for which the traditional architecture is a poor fit and argue for a critical rethinking of the current factoring of systems services into products.

- Focus on high-availability & high-scalability (cap theory):
  → Schemaless (i.e., "Schema Last")
  → Non-relational data models (document, key/value, etc)
  → No ACID transactions
  → Custom APIs instead of SQL
  → Usually open-source

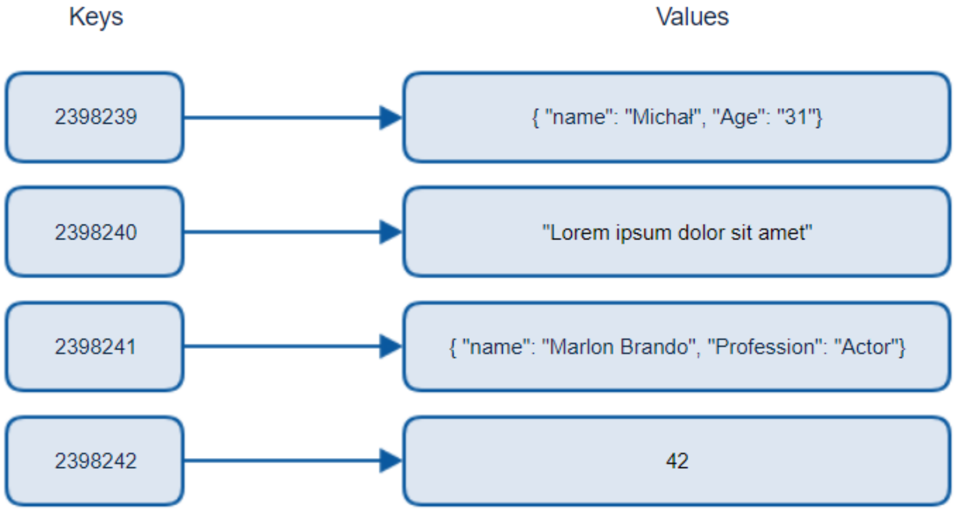
NOSQL

Different Types of NoSQL Databases

NoSQL

KEY VALUE

infinityDB, Riak…

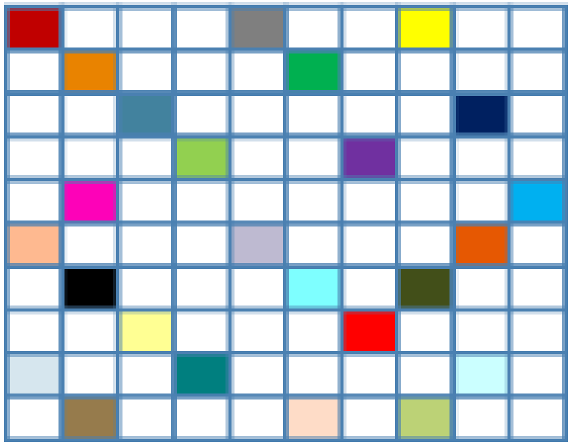


Keys	Values
2398239	{ "name": "Michał", "Age": "31"}
2398240	"Lorem ipsum dolor sit amet"
2398241	{ "name": "Marlon Brando", "Profession": "Actor"}
2398242	42

NoSQL

COLUMNS

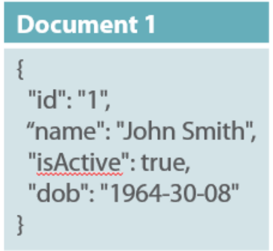
Cassandra, HTable, BigTable



NoSQL

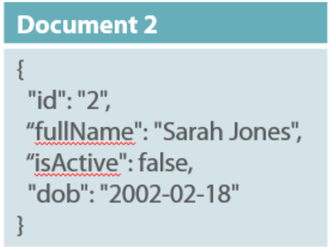
DOCUMENT

MongoDB, CouchDB, DocumentDB…



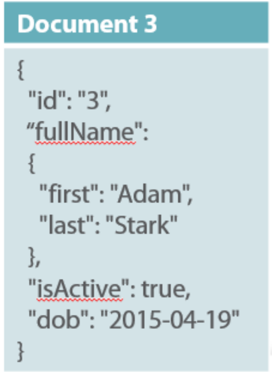
Document 1

```
{
  "id": "1",
  “name": "John Smith",
  "isActive": true,
  "dob": "1964-30-08"
}
```



Document 2

```
{
  "id": "2",
  “fullName": "Sarah Jones",
  “isActive": false,
  "dob": "2002-02-18"
}
```



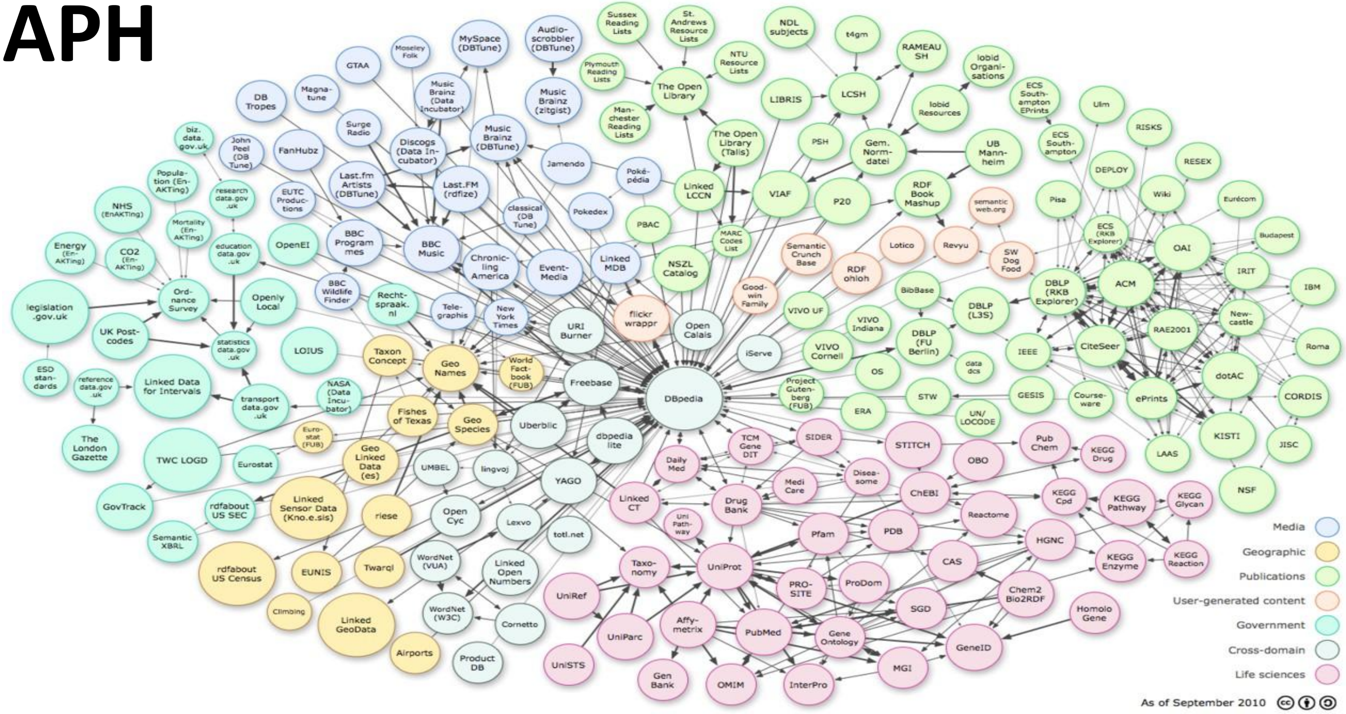
Document 3

```
{
  "id": "3",
  “fullName":
  {
    "first": "Adam",
    "last": "Stark"
  },
  "isActive": true,
  "dob": "2015-04-19"
}
```

NoSQL

GRAPH

Neo4j, OrientDB…



NoSQL: What about ACID ??

Big Data Requirements

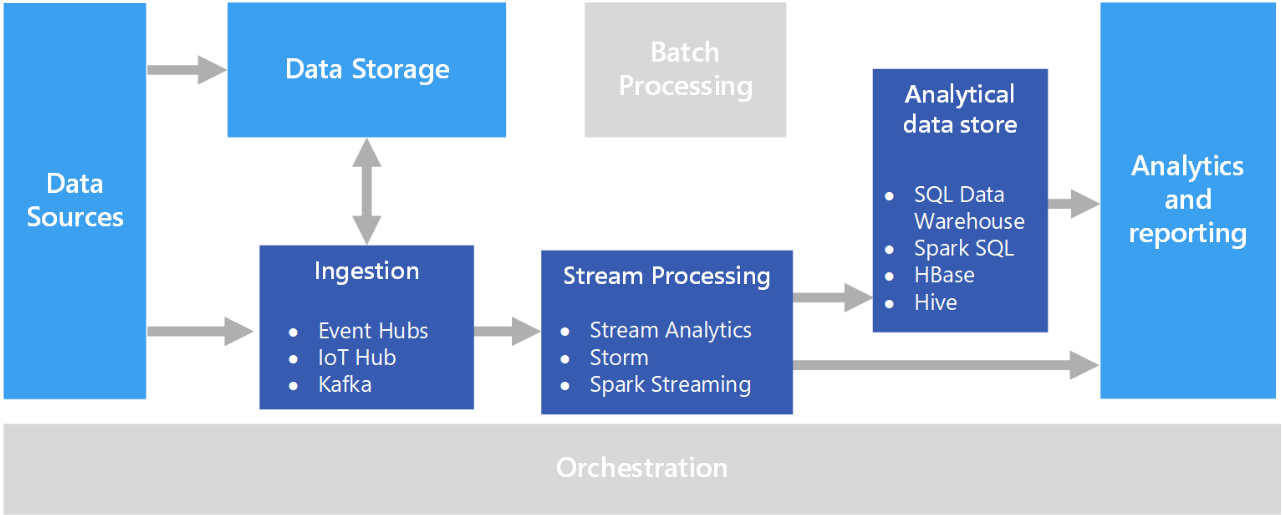
Storage Requirements

NoSQL

Processing Requirements

MapReduce

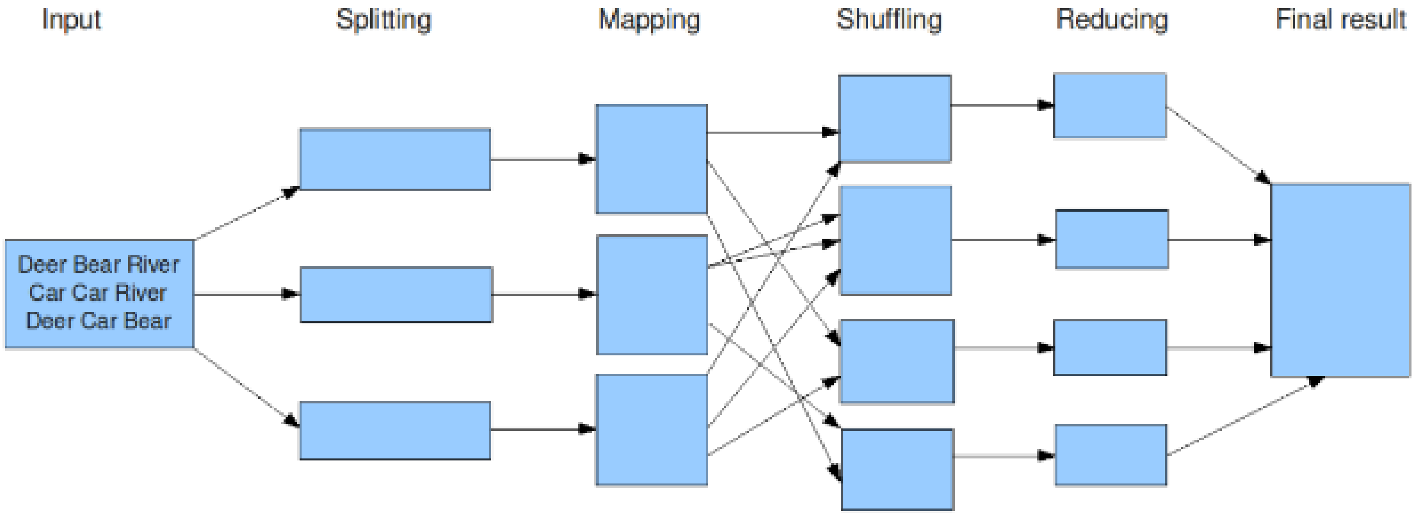
Batch vs. Stream Processing



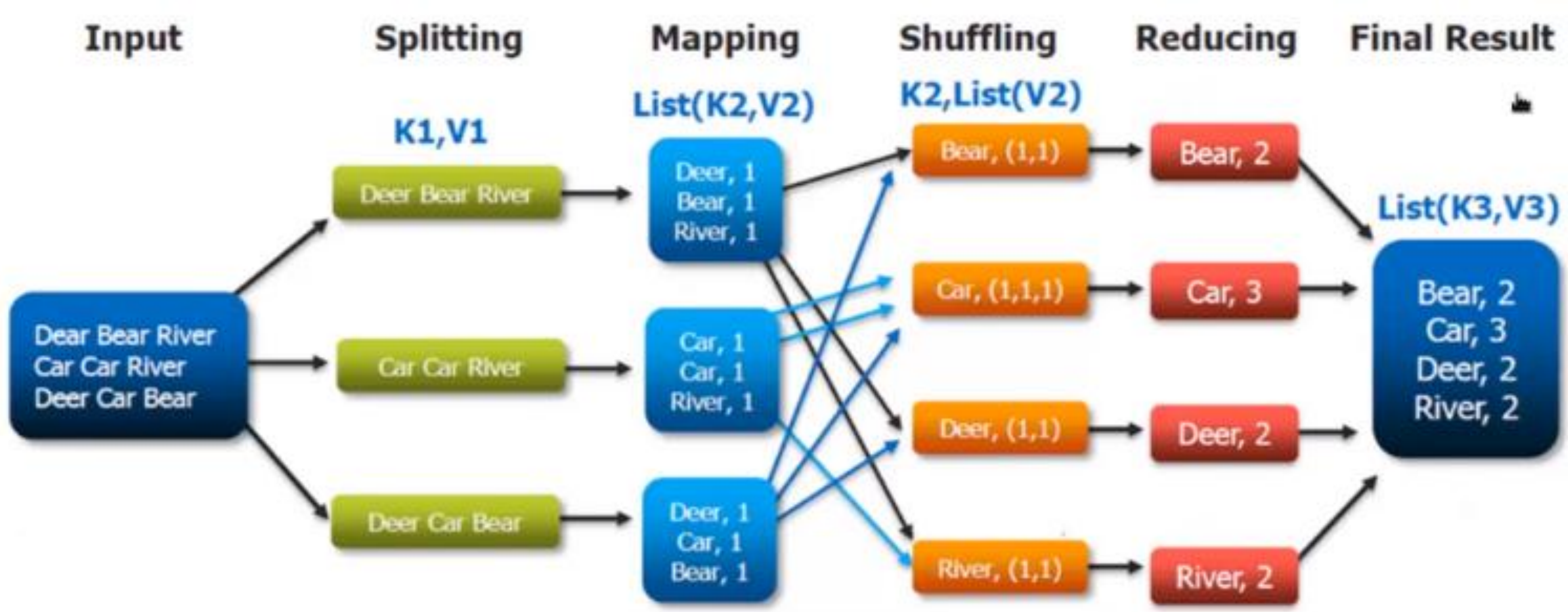
What is map reduce ?

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster

What is working principle ?

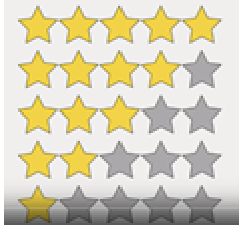


Example

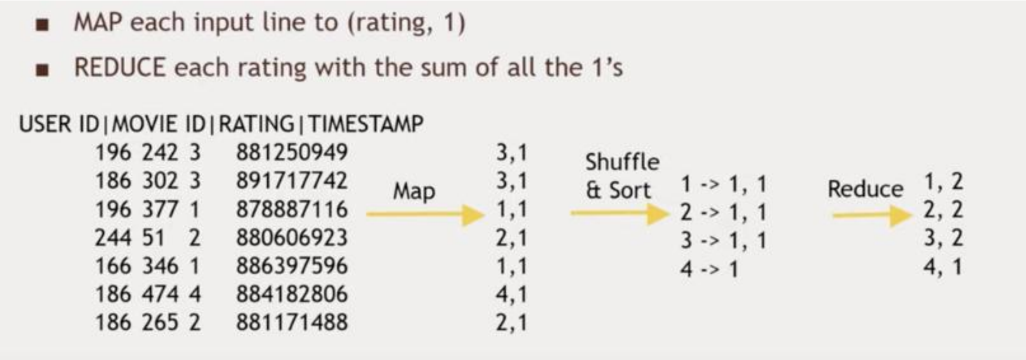

Input
Splitting
Mapping
Shuffling
Reducing
Final Result
K1,V1
List(K2,V2)
K2,List(V2)
List(K3,V3)
Dear Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, (1,1)
Car, (1,1,1)
Deer, (1,1)
River, (1,1)
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Example

- How many of each movie rating exist ?



- Making it a MapReduce problem



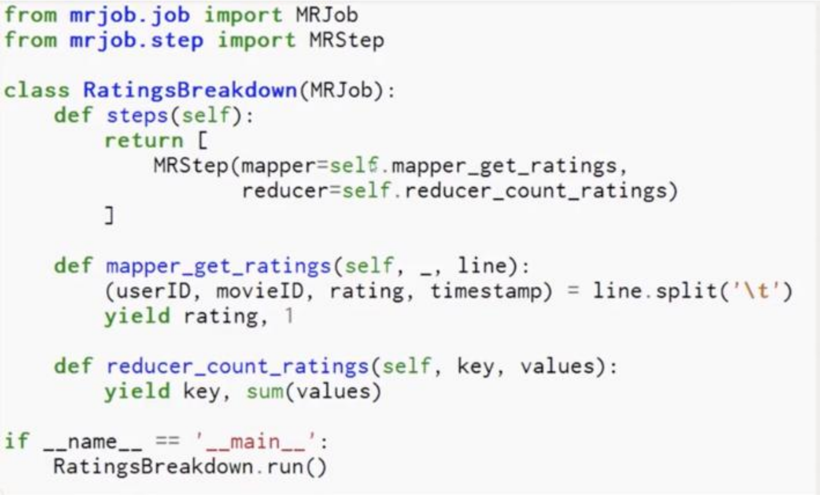
- MAP each input line to (rating, 1)
- REDUCE each rating with the sum of all the 1’s

USER ID	MOVIE ID	RATING	TIMESTAMP
196	242	3	881250949
186	302	3	891717742
196	377	1	878887116
244	51	2	880606923
166	346	1	886397596
186	474	4	884182806
186	265	2	881171488

```
3,1
3,1
1,1
2,1
1,1
4,1
2,1
```

```
1 -> 1, 1
2 -> 1, 1
3 -> 1, 1
4 -> 1
```

```
1, 2
2, 2
3, 2
4, 1
```



```
from mrjob.job import MRJob
from mrjob.step import MRStep

class RatingsBreakdown(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_ratings,
                   reducer=self.reducer_count_ratings)
        ]

    def mapper_get_ratings(self, _, line):
        (userID, movieID, rating, timestamp) = line.split('\t')
        yield rating, 1

    def reducer_count_ratings(self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    RatingsBreakdown.run()
```

Exercice

For the following task using pseudo code

- a) Explain Matrix-Vector multiplication algorithm by MapReduce ?

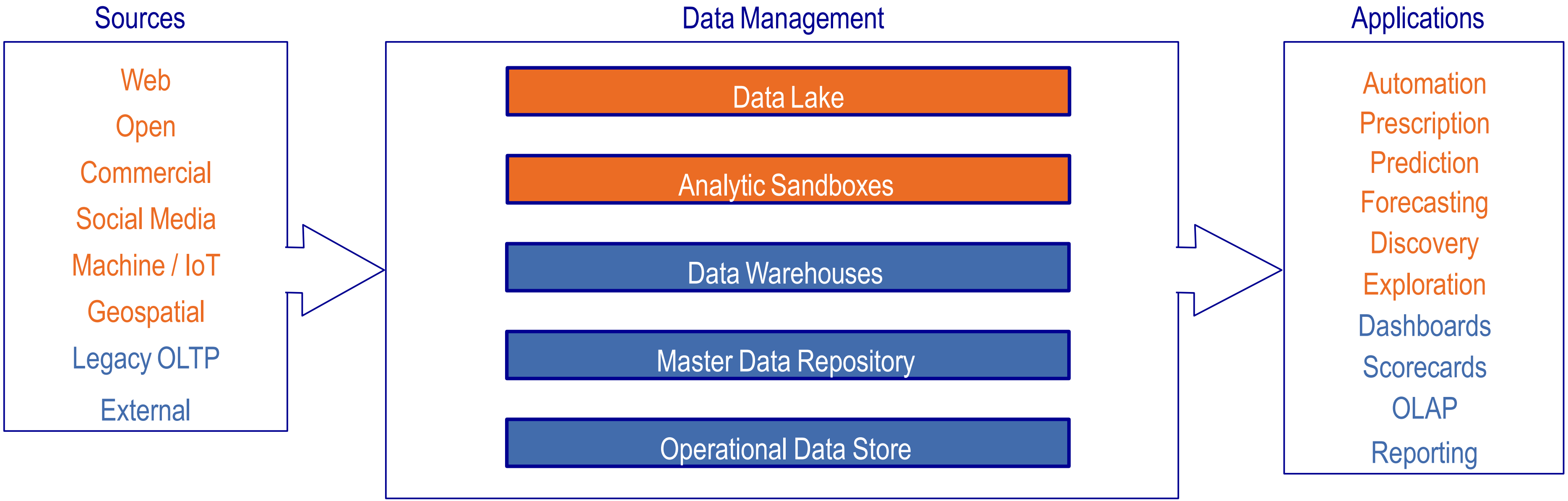
- b) Computing group by and aggregate for a relational database

Explain Issues in Data stream query processing ?

Explain:

1. Bloom Filter with the help of an example ?
2. Steps of HITS algorithm

Modern Data Integration Architecture



Support for all data use cases – from reporting to data science

Support for all data latencies – from batch to streaming

Support for hybrid ecosystem – mix of on-premises, cloud, multi-cloud

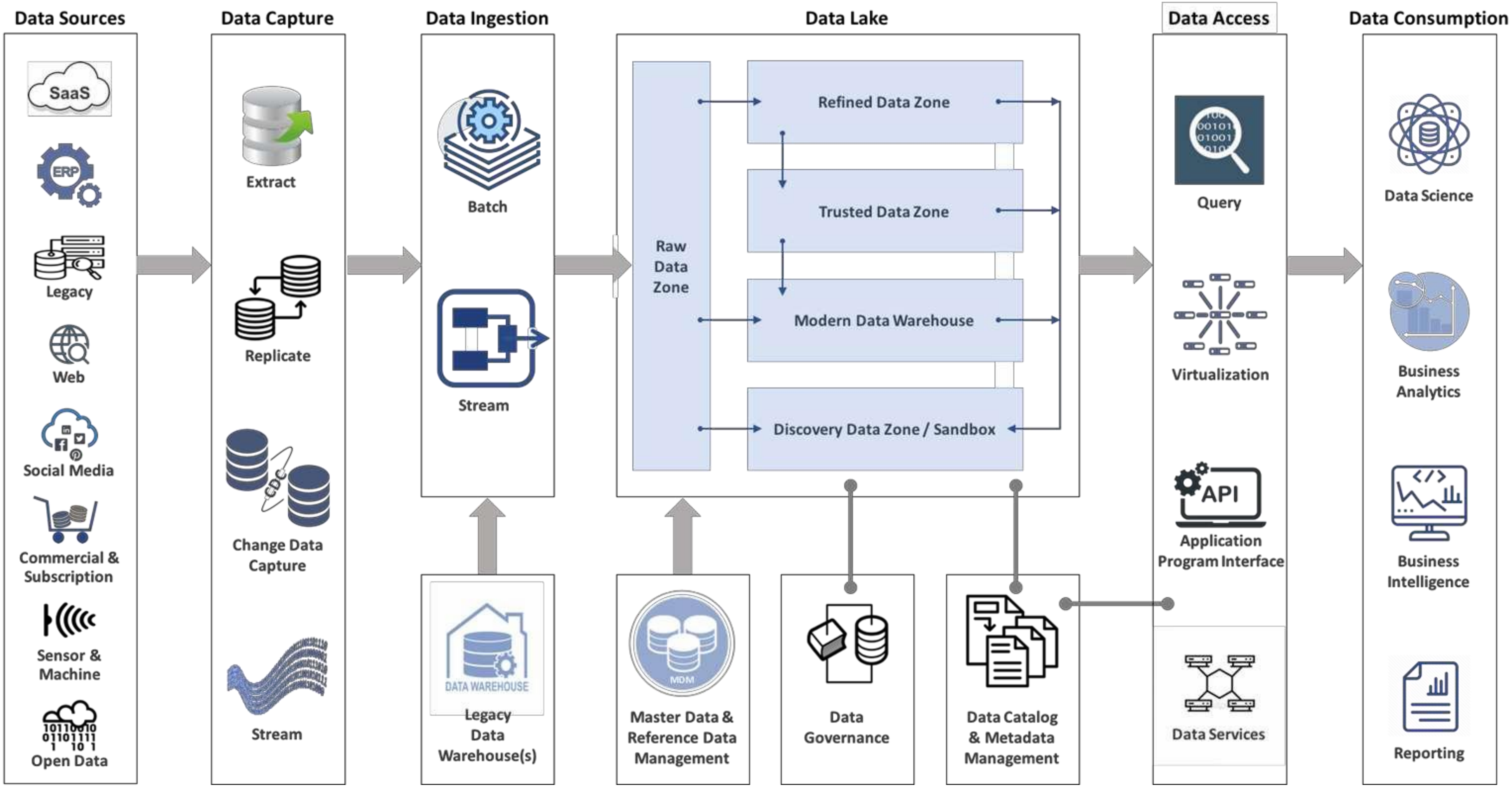
Sustaining value of legacy investments – data lake and data warehouse working together

Enabling self-service – easy access for all data consumers

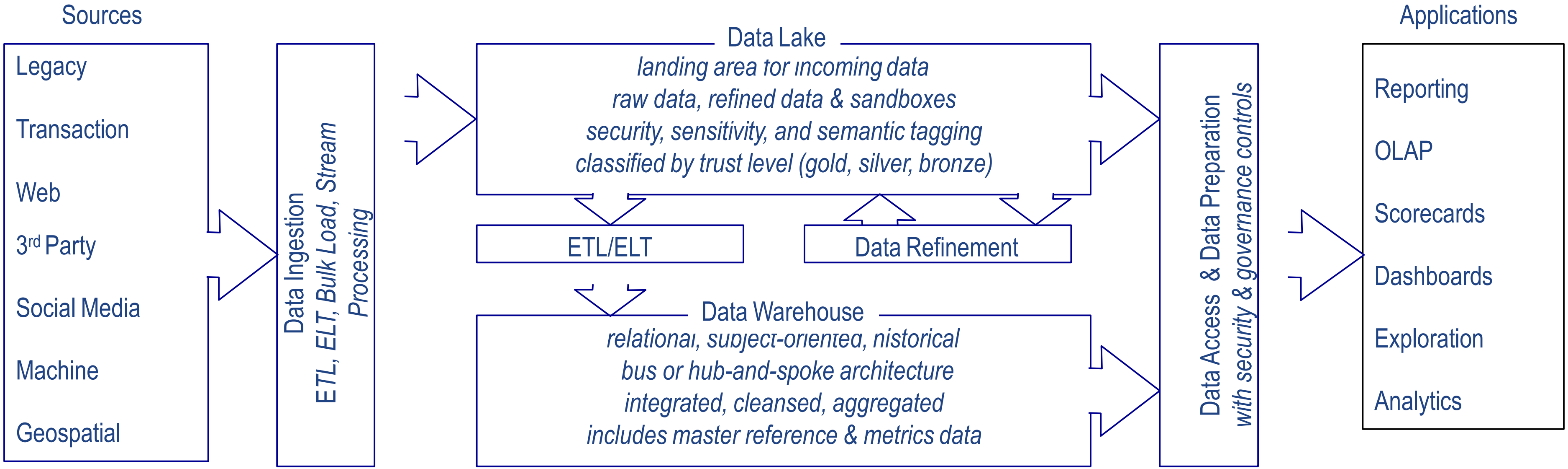
Big data capable – scalable and elastic

Sustainable – automation and operationalization

Modern Data Integration Architecture



Data Warehouse Outside the Data Lake



Sources

Legacy

Transaction

Web

3rd Party

Social Media

Machine

Geospatial

Data Ingestion

ETL, ELT, Bulk Load, Stream Processing

Data Lake

landing area for incoming data

raw data, refined data & sandboxes

security, sensitivity, and semantic tagging

classified by trust level (gold, silver, bronze)

ETL/ELT

Data Refinement

Data Warehouse

relational, subject-oriented, historical

bus or hub-and-spoke architecture

integrated, cleansed, aggregated

includes master reference & metrics data

Data Access & Data Preparation

with security & governance controls

Applications

Reporting

OLAP

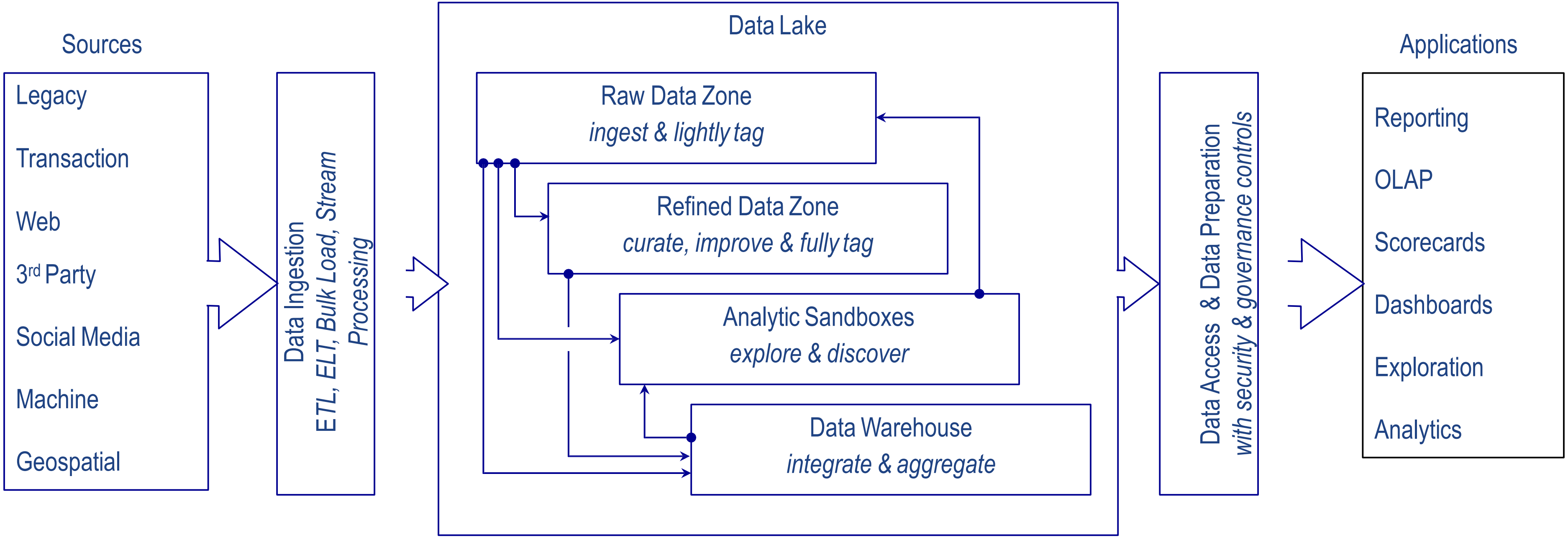
Scorecards

Dashboards

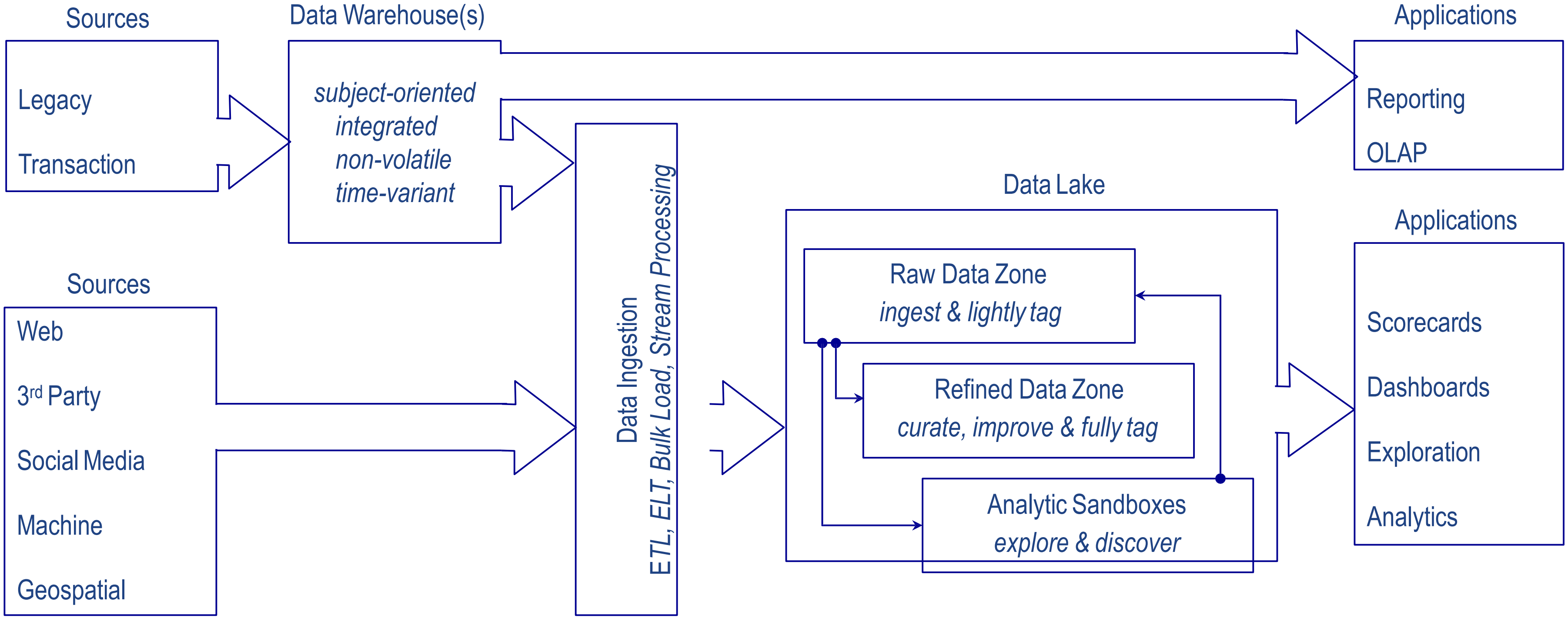
Exploration

Analytics

Data Warehouse Inside the Data Lake

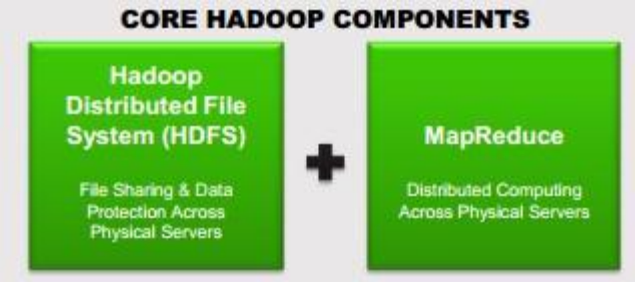


Data Warehouse In Front of the Data Lake

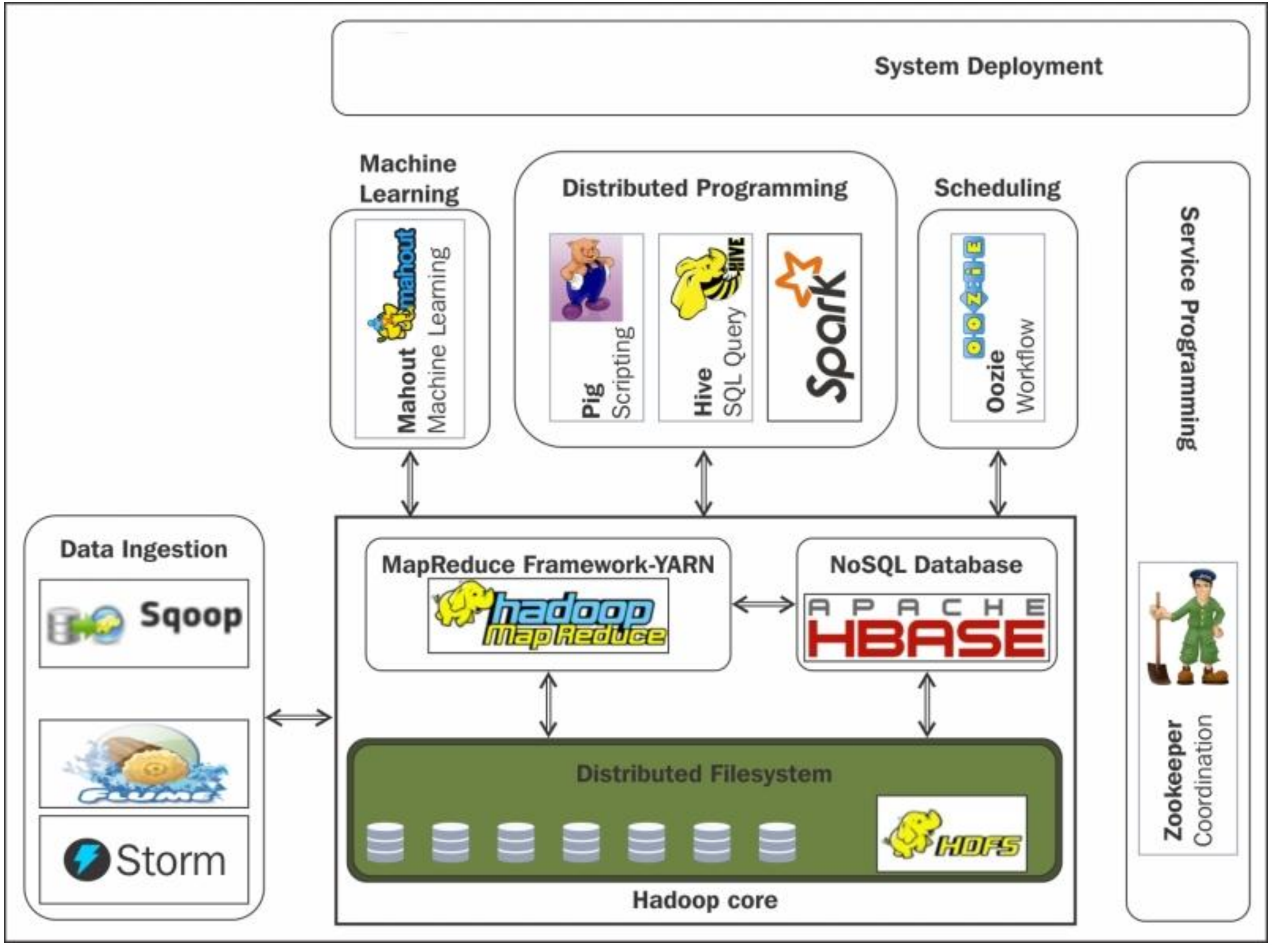


Outils Big Data - Hadoop

Hadoop:
Hadoop is a Java Framework or Software which was invented to manage huge data or Big Data. Hadoop is used for storing and processing the large data distributed across a cluster of commodity servers.
Hadoop stores the data using Hadoop distributed file system and process/query it using Map Reduce programming model.


CORE HADOOP COMPONENTS
Hadoop Distributed File System (HDFS)
File Sharing & Data Protection Across Physical Servers
+
MapReduce
Distributed Computing Across Physical Servers

Outils Big Data – The Hadoop Ecosystem



Real Time Data Processing (Storm, Yahoo S4)

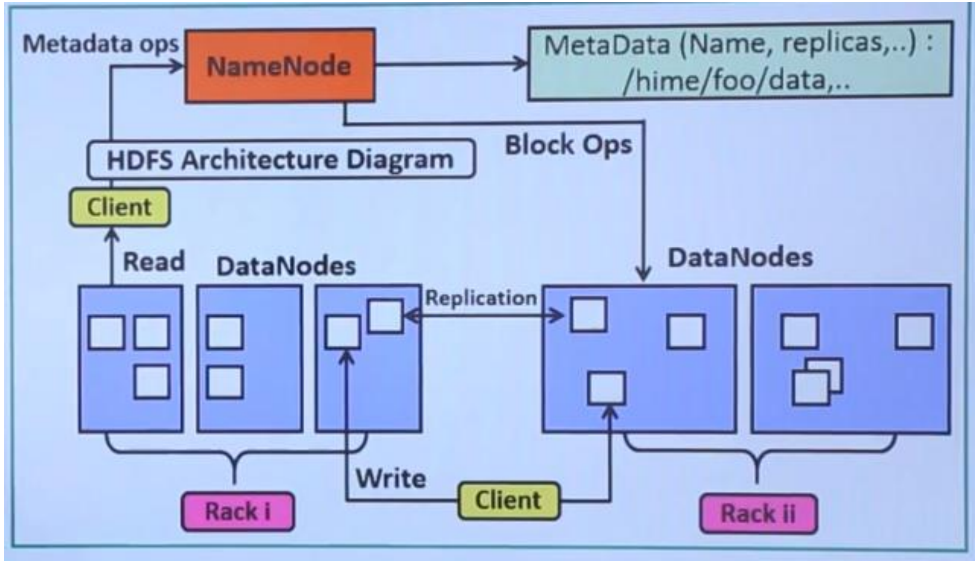
Sqoop is a tool for transferring **data** between HDFS and RDBMS.

Storm is a free and open source distributed realtime computation system.

ZooKeeper is a distributed, open-source **coordination** service for distributed applications

Outils Big Data - Hadoop

HDFS Architecture

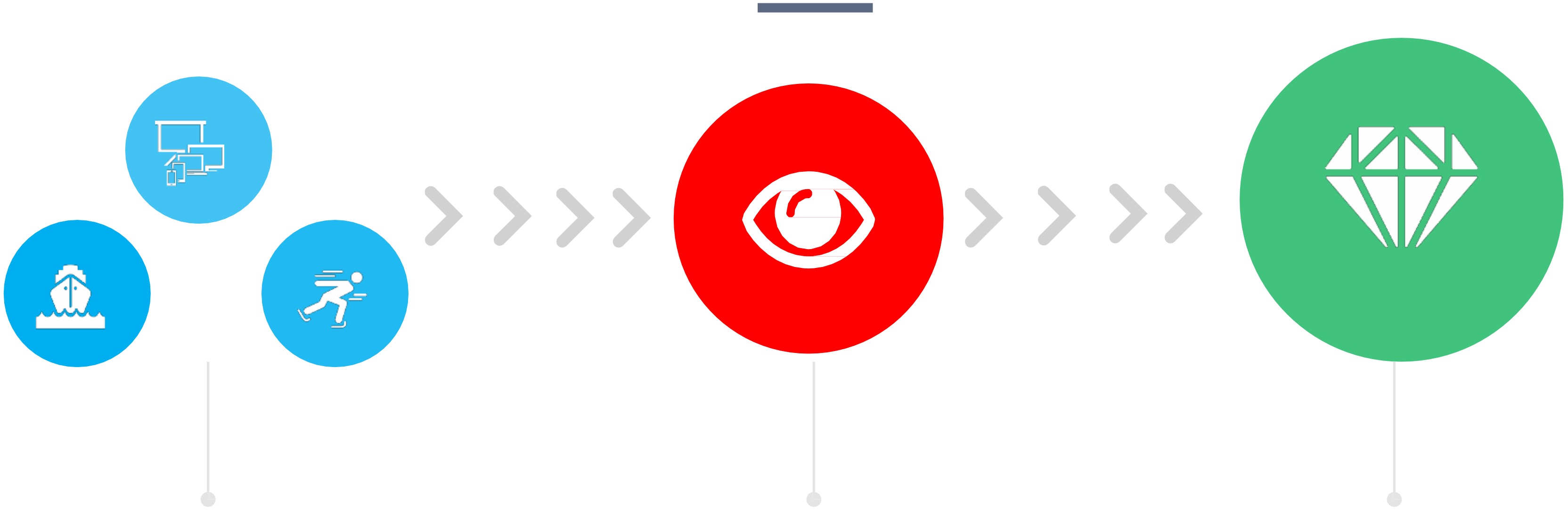


- **Hadoop Distributed File System (HDFS) :** primary data storage system used by Hadoop applications
- **HDFS** is not NoSQL
- Many NoSQL solutions in fact use HDFS for their storage.

- **HDFS :**
  - **NameNode (**store metadata, Managing FS namespace, check availability and replication**)**
  - **DataNode (**stroring data, replication creating, deleting job, send report (defaut time 3 sec) **)**

Les Challenges de visualization des Big Data

- (Gupta and Siddiqui 2014) + (Shilpa 2013)



Challenge des 3V

Le challenge est imposé parles le volume, la variété et la vélocité des Big Data.

Plus le facteur humain.

Visualisation

Représentation visuelle des données permettant des analyses visuelles.

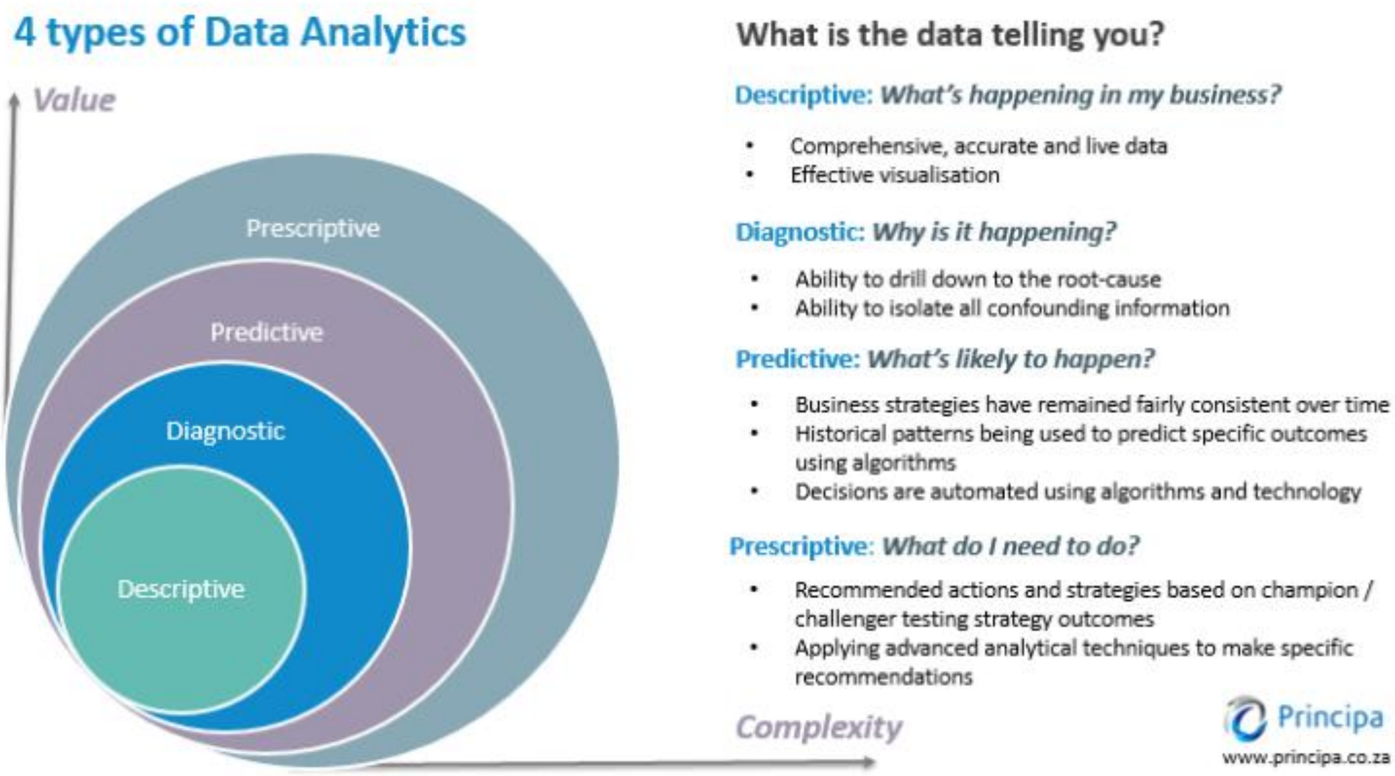
4 eme V: Valeur

Améliorer les revenus.
Etre leader de son marché
Améliorer la relation client.

Question ?

What is the difference between HBASE and HDFS in Hadoop ?

4 type of Data Analytics



4 types of Data Analytics

What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

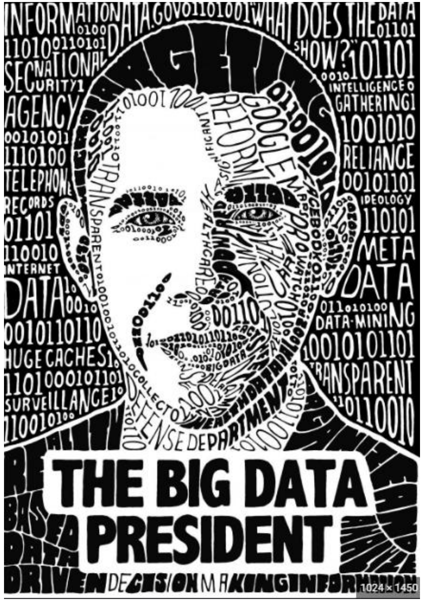
- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

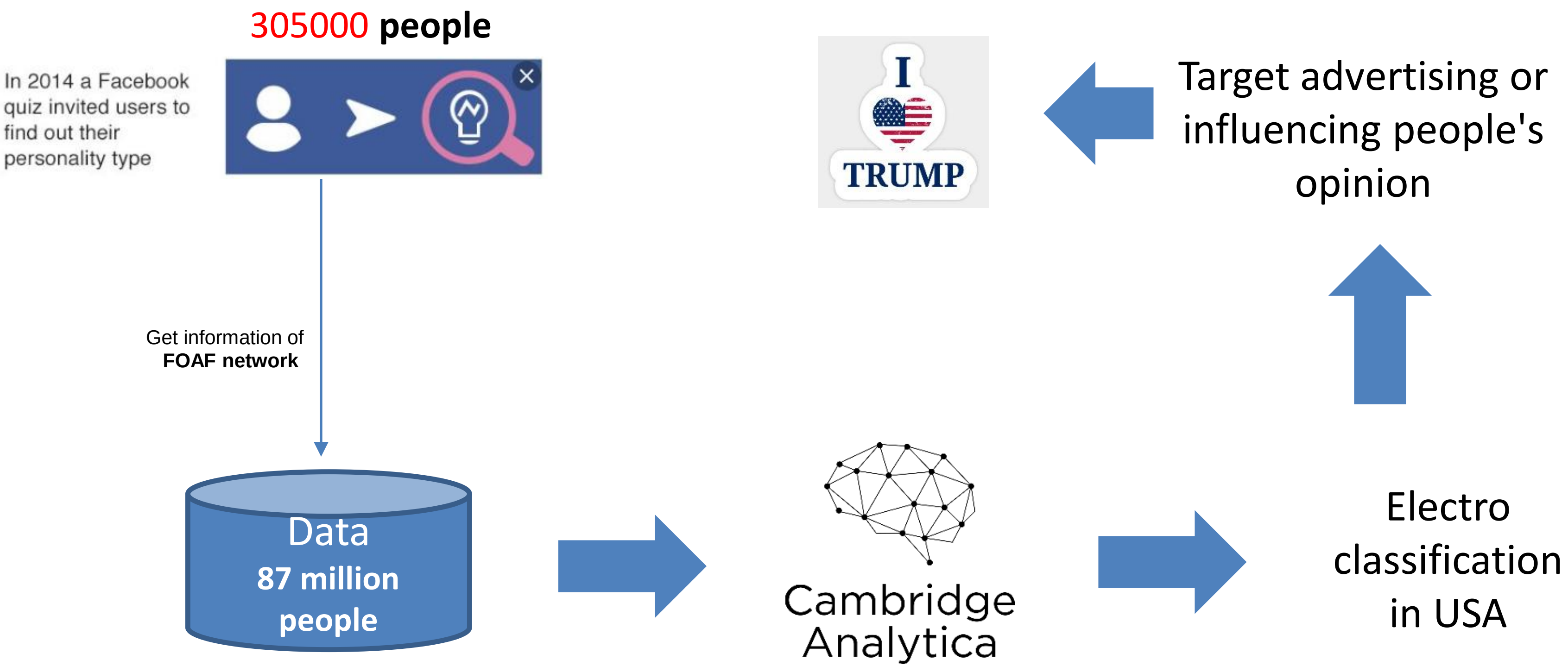
- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

Case study :Opinion Mining in Social Big Data



Case study :Opinion Mining in Social Big Data



305000 **people**

In 2014 a Facebook quiz invited users to find out their personality type

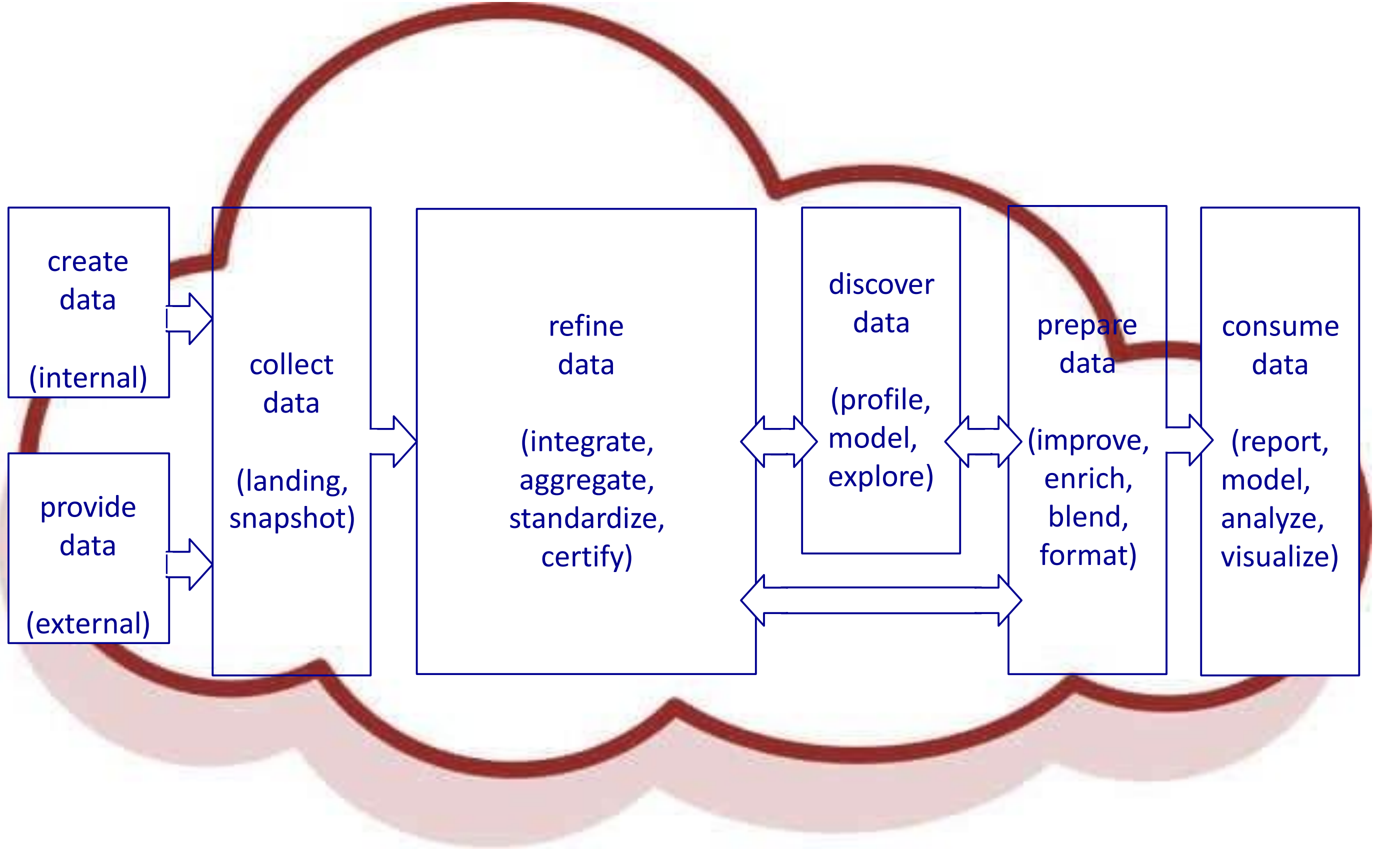
Get information of **FOAF network**

Data **87 million people**

Electro classification in USA

Target advertising or influencing people's opinion

Illustrate How Cloud and Big Data related to each other ?



create data

(internal)

provide data

(external)

collect data

(landing, snapshot)

refine data

(integrate, aggregate, standardize, certify)

discover data

(profile, model, explore)

prepare data

(improve, enrich, blend, format)

consume data

(report, model, analyze, visualize)

Exercice
Data Mining applications and bigdata

- Select one of the following industries and answer the question below

Retail industry, Banking industry, Insurance, Healthcare, Governement, Securities, Education

○ **Describe the nature of data sources in your chosen industry**

○ **Describe one possible datamining application in your industry**

○ **List and discuss the major issues that need to be tackled**

○ **Describe an example of an industry for which bigdata analytics is essential problem and issues involved**

1. How to model?
2. How to query ?

…" Suggest advanced use (**AI**)

… '' How to improve the **value**?

Project proposal example

- Student mobilization.
- Foster the emergence of participatory action

Practical Work:
Query Optimization in Spark SQL

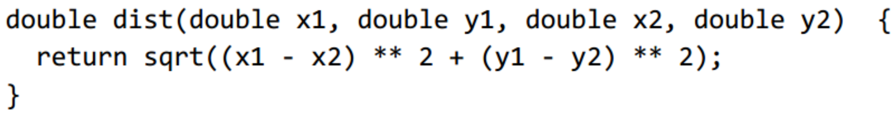
In this assignment, you will use Spark's SQL component to analyze query execution plans (Part I) and write some of your own query optimization rules (Part II). Part I involves running SQL queries in an interactive Spark shell, then writing up some analyses of the results. Part II involves writing Scala code to implement custom Spark SQL query optimization rules

- **Setup Software Dependencies** : **Spark 2.3.3**
  - Download Spark 2.3.3, prebuilt for Hadoop 2.7: spark.apache.org/downloads.html .
  - Running Spark requires Java 8 (unfortunately Java 9+ is not compatible). You'll need to download Java 8 if you do not have it already, and make sure JAVA_HOME points to your Java 8 JDK root directory (e.g. MacOS users can run the command /usr/libexec/java_home -v 1.8 to find the home dir of their Java 8 installation if you have one)

Part I: Analyze Query Plans

Part II: Write Your Own Optimization Rules

In this part, you will write a set of transforms to optimize Spark SQL logical plans that contain instances of a custom function (commonly known as a user-defined function, or UDF) we have defined called dist . This function computes the distance between two (x, y) points, and is defined as follows:



```
double dist(double x1, double y1, double x2, double y2)  {
  return sqrt((x1 - x2) ** 2 + (y1 - y2) ** 2);
}
```

Background

We strongly suggest you read/review the following resources:

- The three lectures on Query Execution and Query Optimization
- The Spark SQL paper

http://web.stanford.edu/class/cs245/readings/spark-sql.pdf

- Spark SQL programming

○ The Overview and Getting Started sections in

https://spark.apache.org/docs/2.3.3/sql-programming-guide.html

- Intro to Scala, if it's new to you

○ https://docs.scala-lang.org/tour/tour-of-scala.html

○ https://learnxinyminutes.com/docs/scala/

Questions and Answers

I Hope I Succeeded to clarify what is Big Data