

Introduction to BigData

OUARED Abdelkader

Teaching assistant at Ibn Khaldoun University

a_ouared@esi.dz

<https://www.linkedin.com/in/abdelkader-ouared/>

<https://github.com/OUARED-A>

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



Data processing

(Agrawal, Bernstein et al. 2011)



Data acquisition

- Acquisition and filtering
- Metadata generation

Data processing and integration

- Integration
- Data aggregation

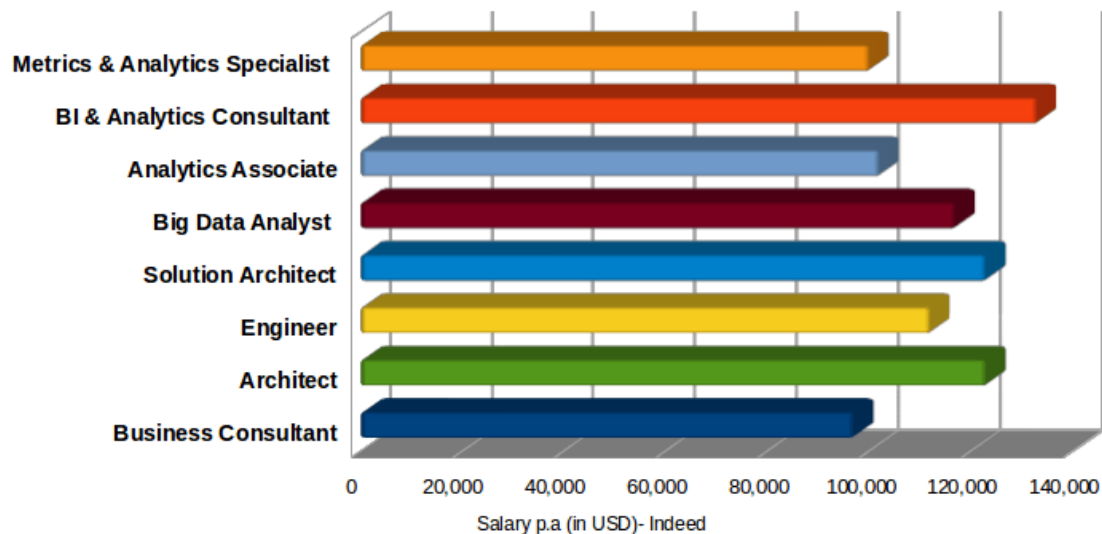
Data Analysis

- 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 j edui 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

Solutions 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**

NoSQL

NoSQL systems
use a distributed file system

+

**Processing
Requirements**

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

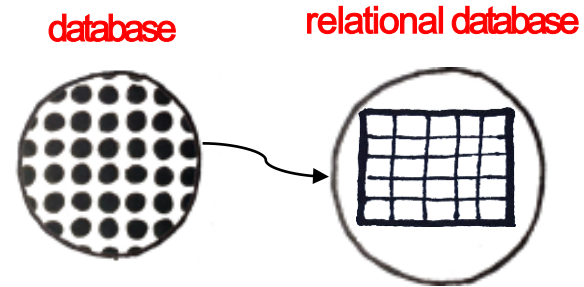


The Law of Relational Database

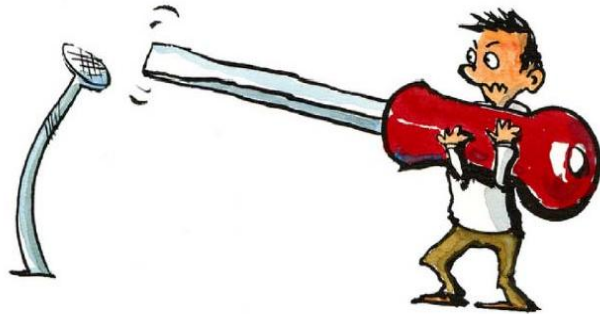


By HikingArtist.com

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



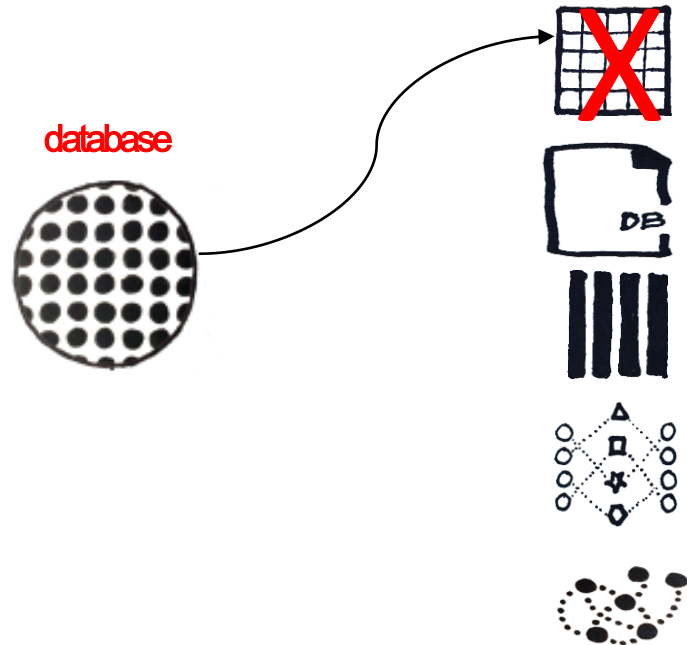
One-size-fits-all: An Idea Whose Time Has Come and Gone



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

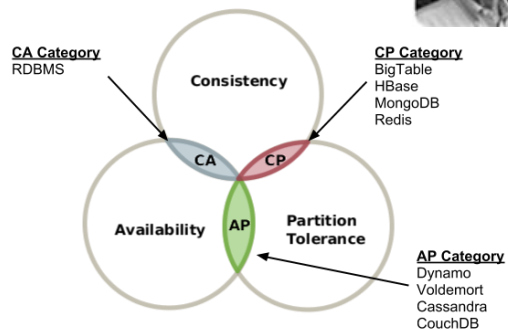
Abraham Maslow - The Psychology of Science - 1966

From Tabular Data To Other Data Model



CAP theorem

NoSQL: Not Only SQL



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



Why
NoSQL?

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

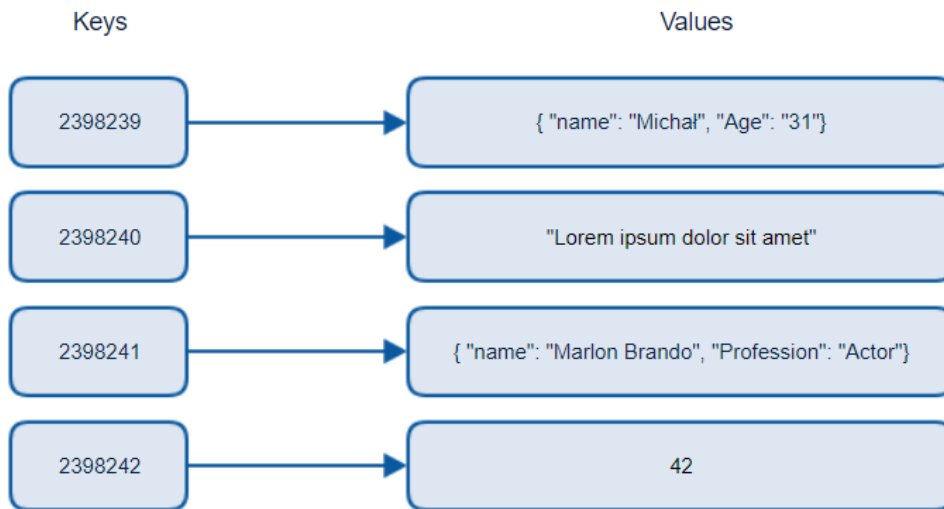
NOSQL

Different Types of NoSQL Databases

NoSQL

KEY VALUE

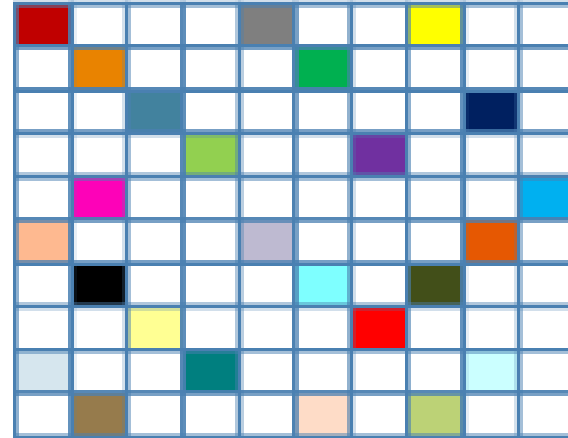
infinityDB, Riak...



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"
}
```


GRAPH

APH

As of September 2010

Legend:

- Media (Blue)
- Geographic (Yellow)
- Publications (Green)
- User-generated content (Orange)
- Government (Cyan)
- Cross-domain (Light Green)
- Life sciences (Pink)

As of September 2010 

NoSQL: What about ACID ??

Big Data Requirements

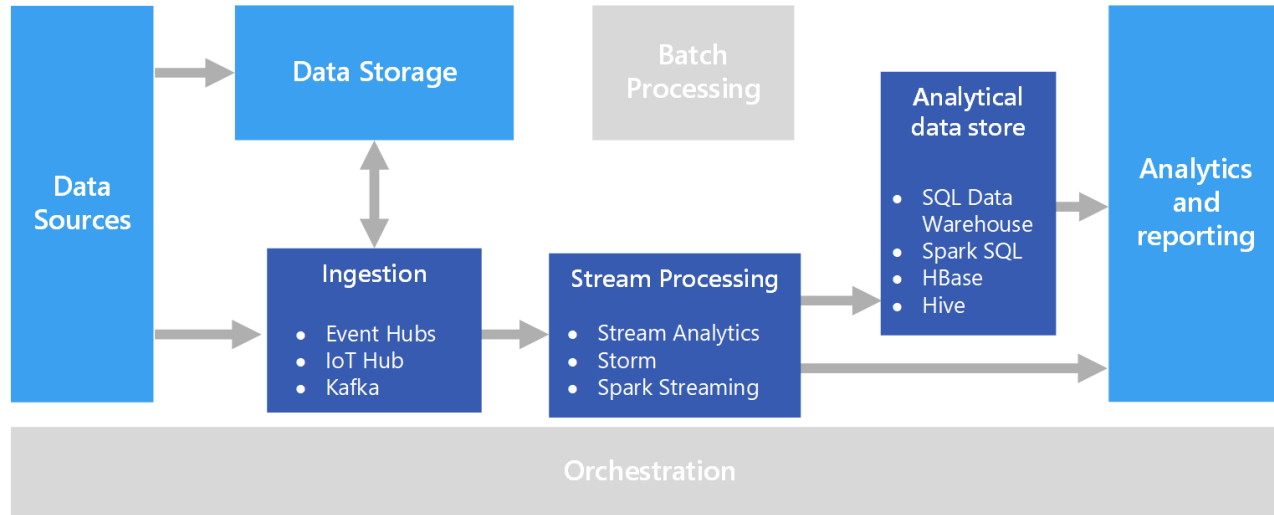
**Storage
Requirements**

NoSQL

**Processing
Requirements**

MapReduce

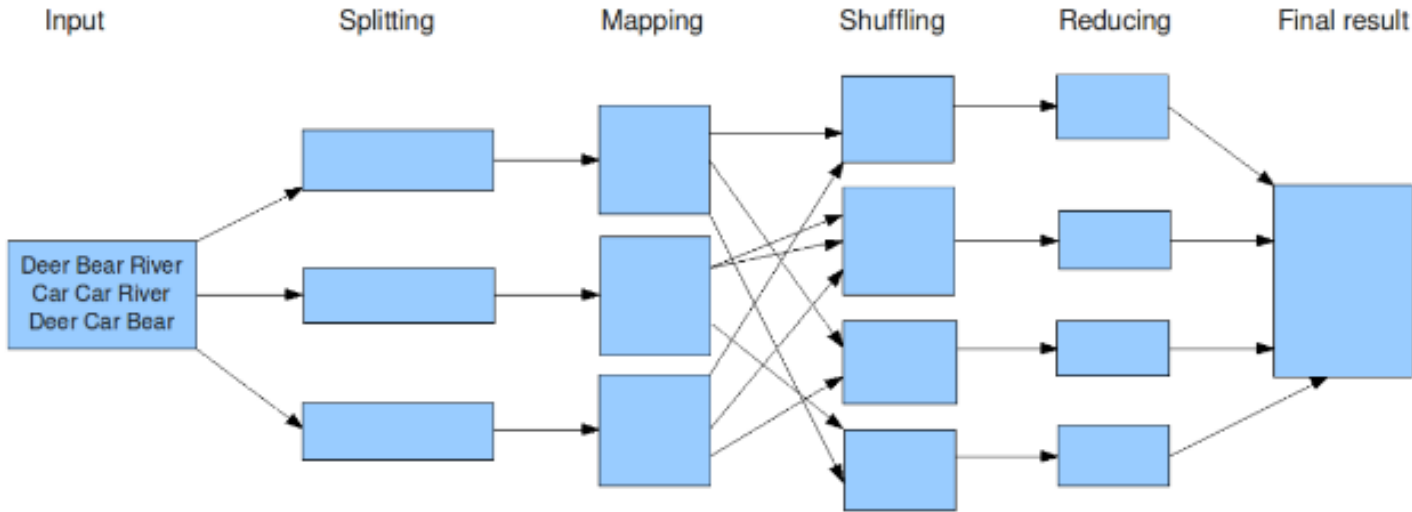
Batch vs. Stream Processing



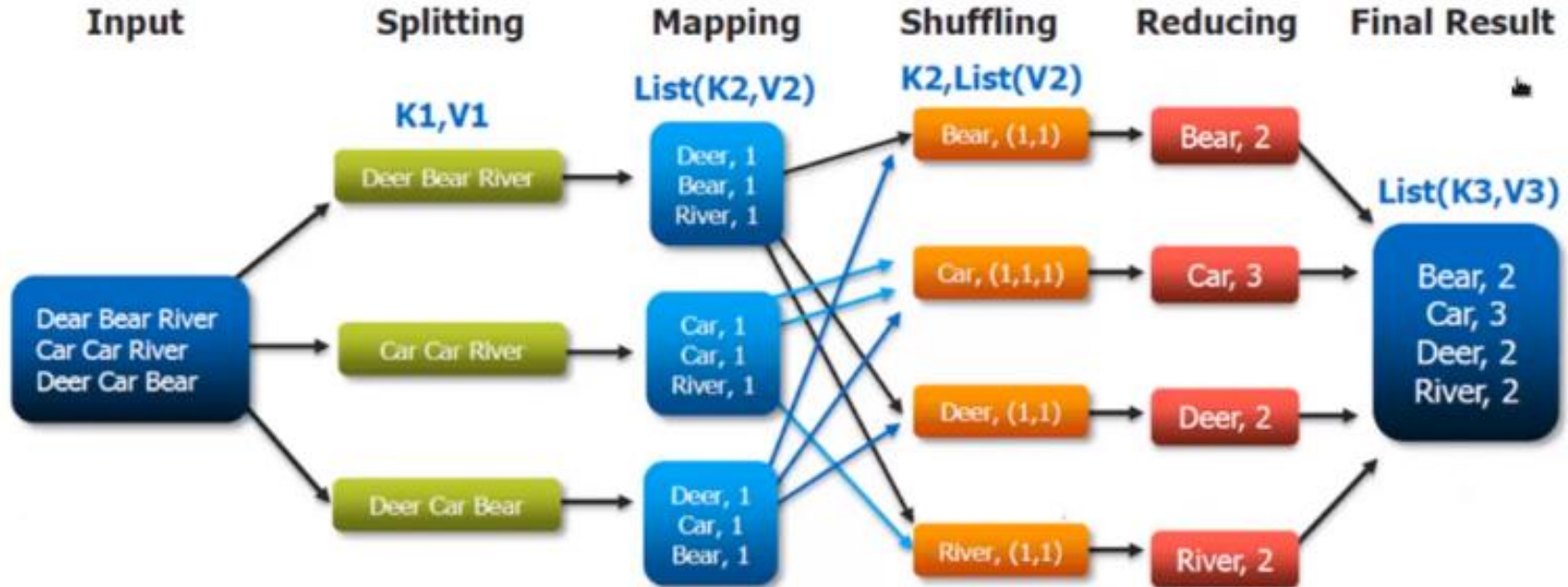
What is map reduce ?

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 ?

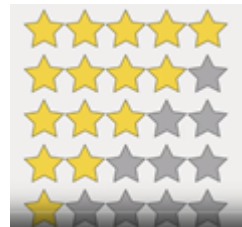


Example



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		3,1	
186	302	3	891717742		3,1	
196	377	1	878887116	Map →	1,1	
244	51	2	880606923		2,1	
166	346	1	886397596		1,1	
186	474	4	884182806		4,1	
186	265	2	881171488		2,1	

	Shuffle & Sort →		Reduce →
1 -> 1, 1		1, 2	
2 -> 1, 1		2, 2	
3 -> 1, 1		3, 2	
4 -> 1		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()
```


Exercise

For the following task using pseudo code

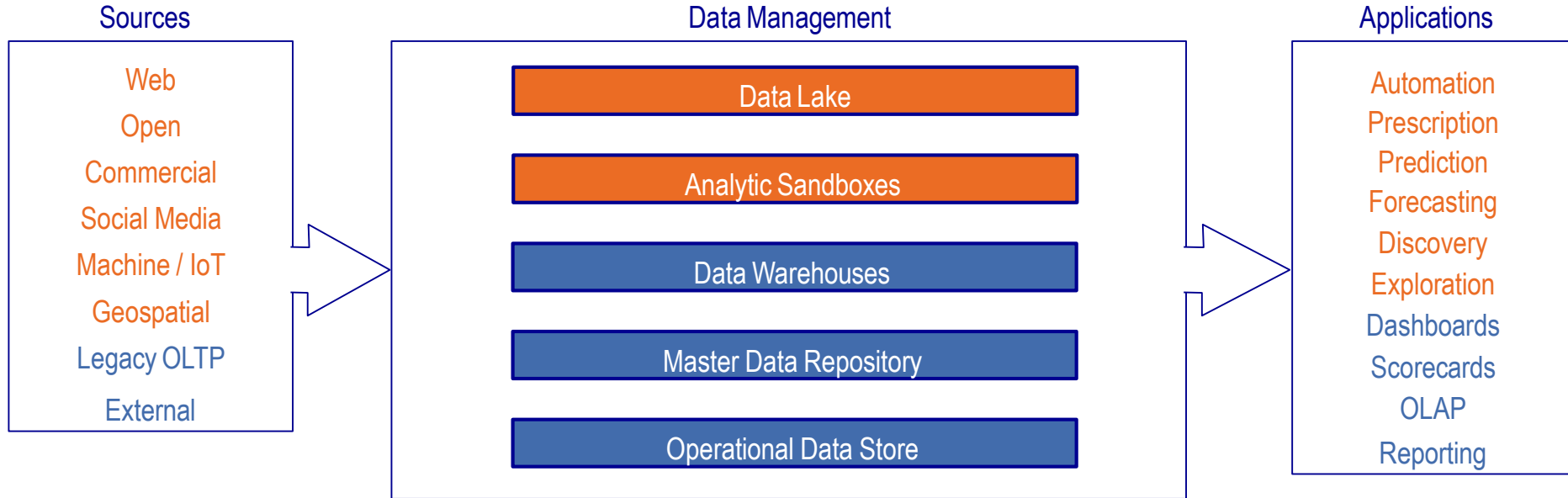
- a) Explain Matrix-Vector multiplication algorithm by MapReduce ?
- b) Computing group by and aggregate for a relational database

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

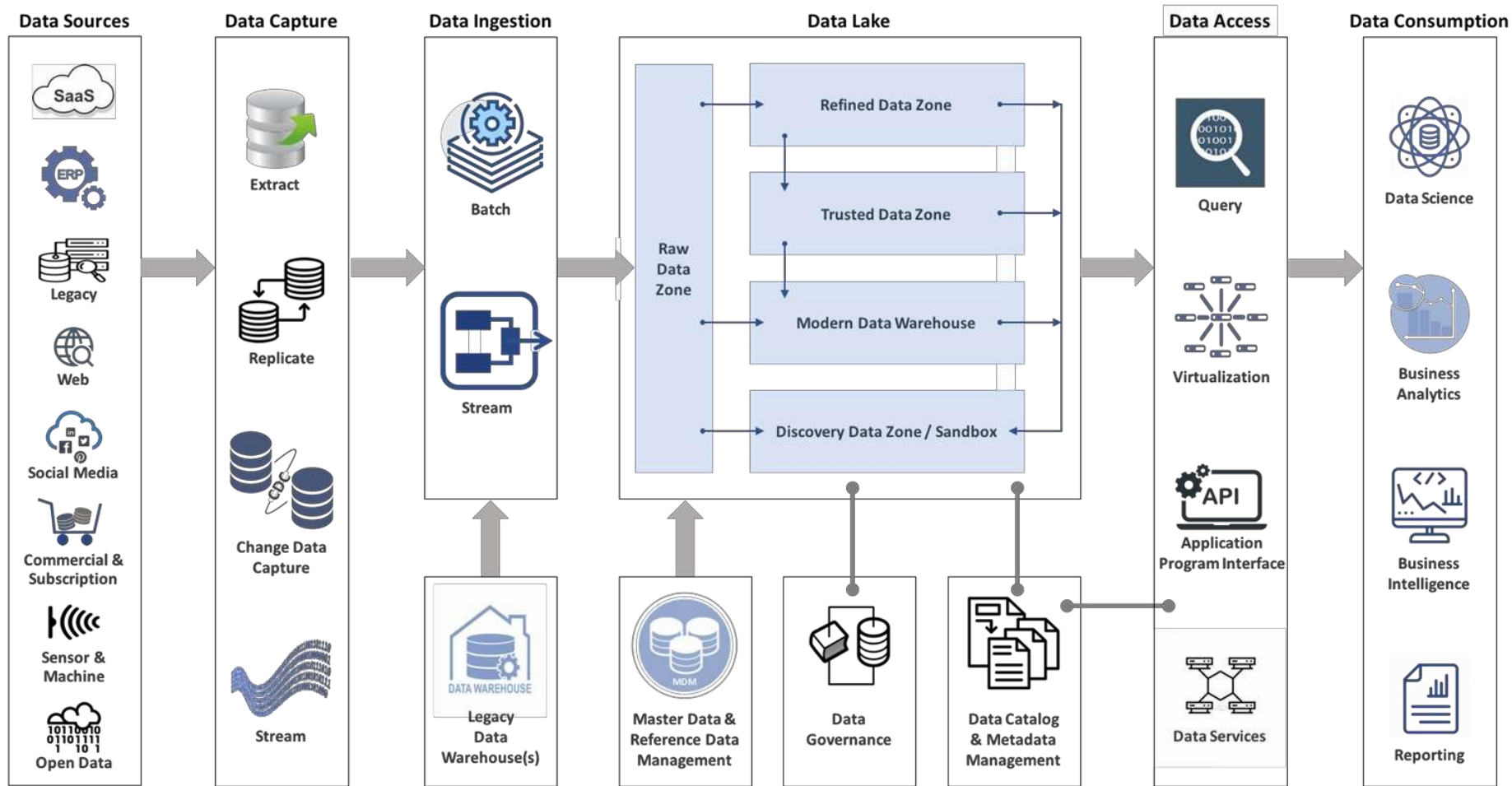
Sustaining value of legacy investments – data lake and data warehouse working together

Enabling self-service – easy access for all data consumers

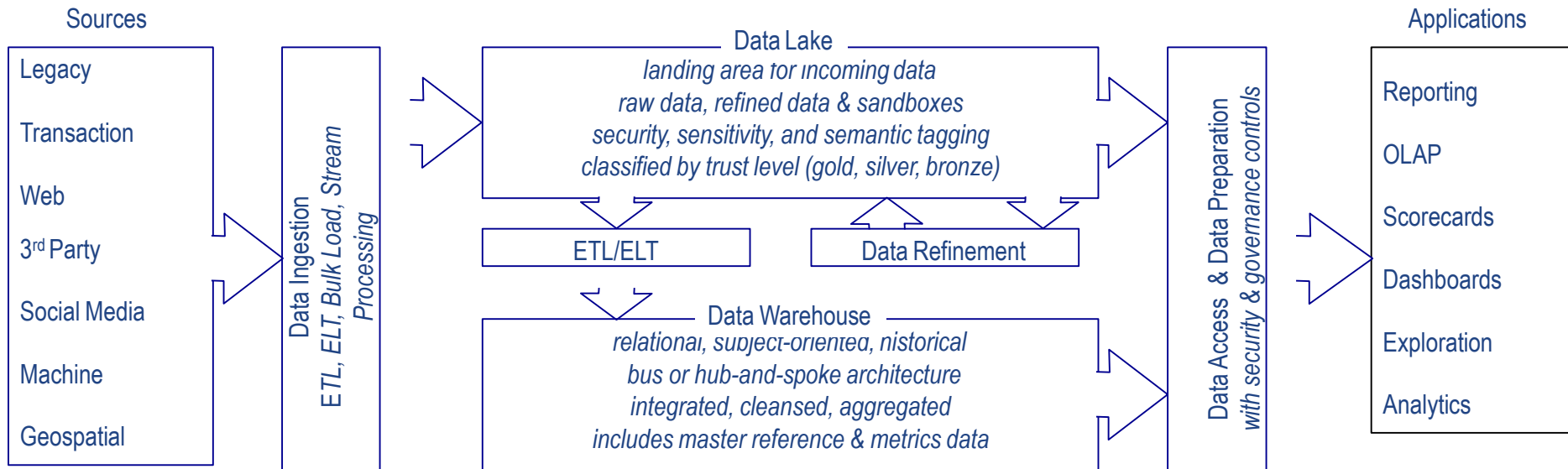
Big data capable – scalable and elastic

Sustainable – automation and operationalization

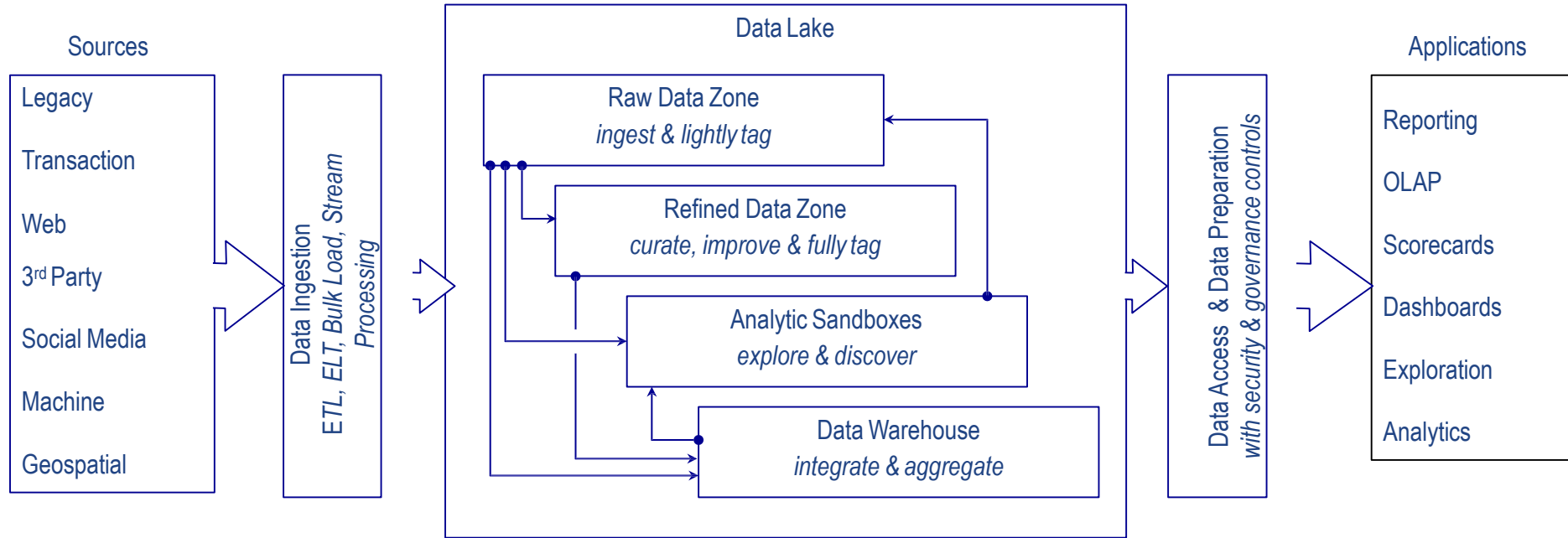
Modern Data Integration Architecture



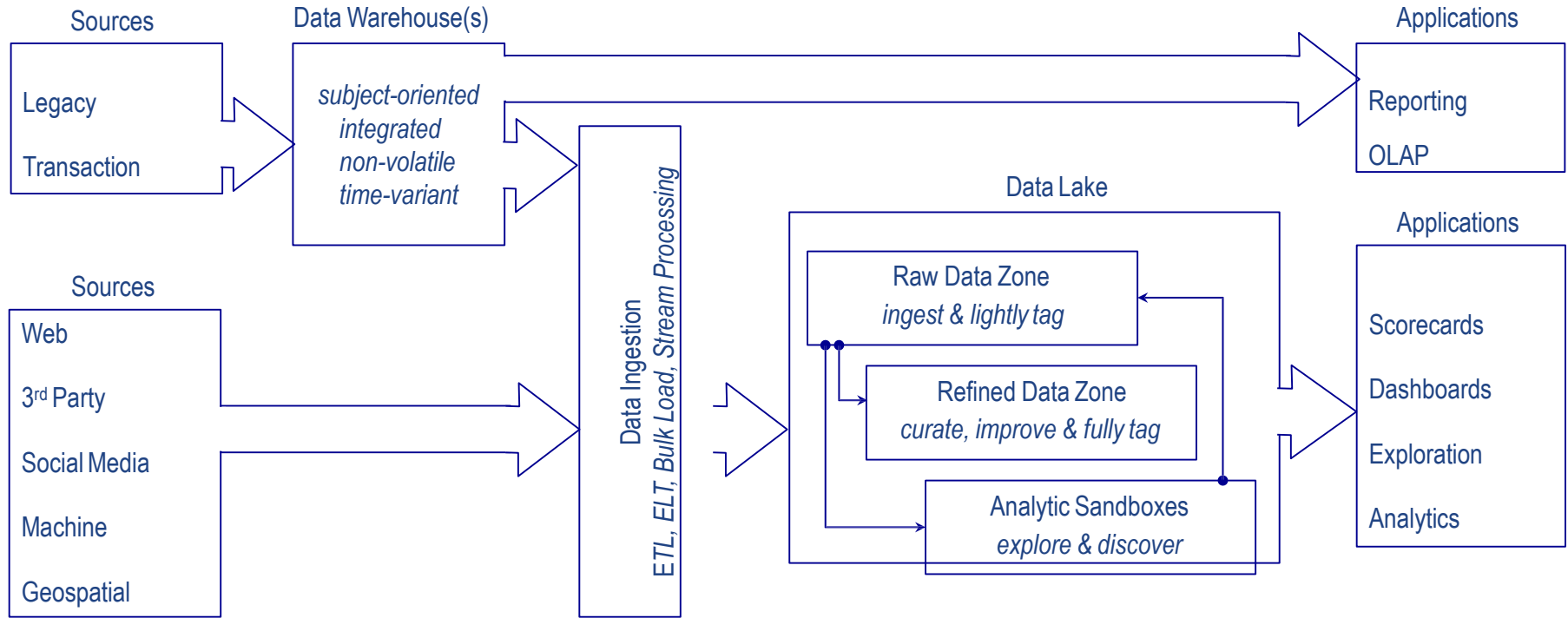
Data Warehouse Outside the Data Lake



Data Warehouse Inside the Data Lake



Data Warehouse In Front of the Data Lake

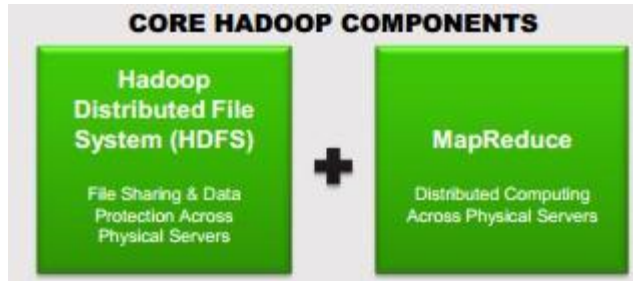


Outils Big Data - Hadoop

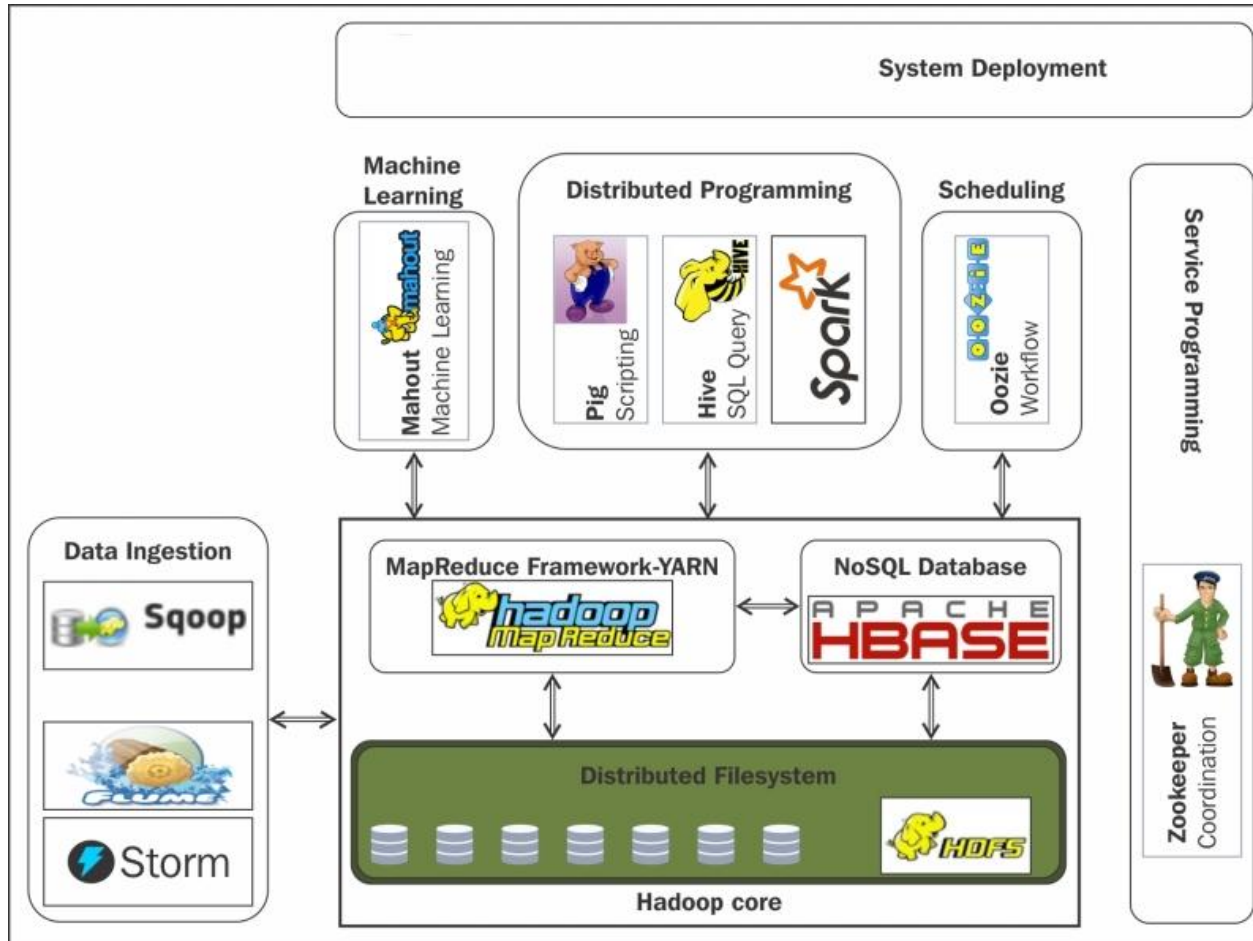
Hadoop:

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.

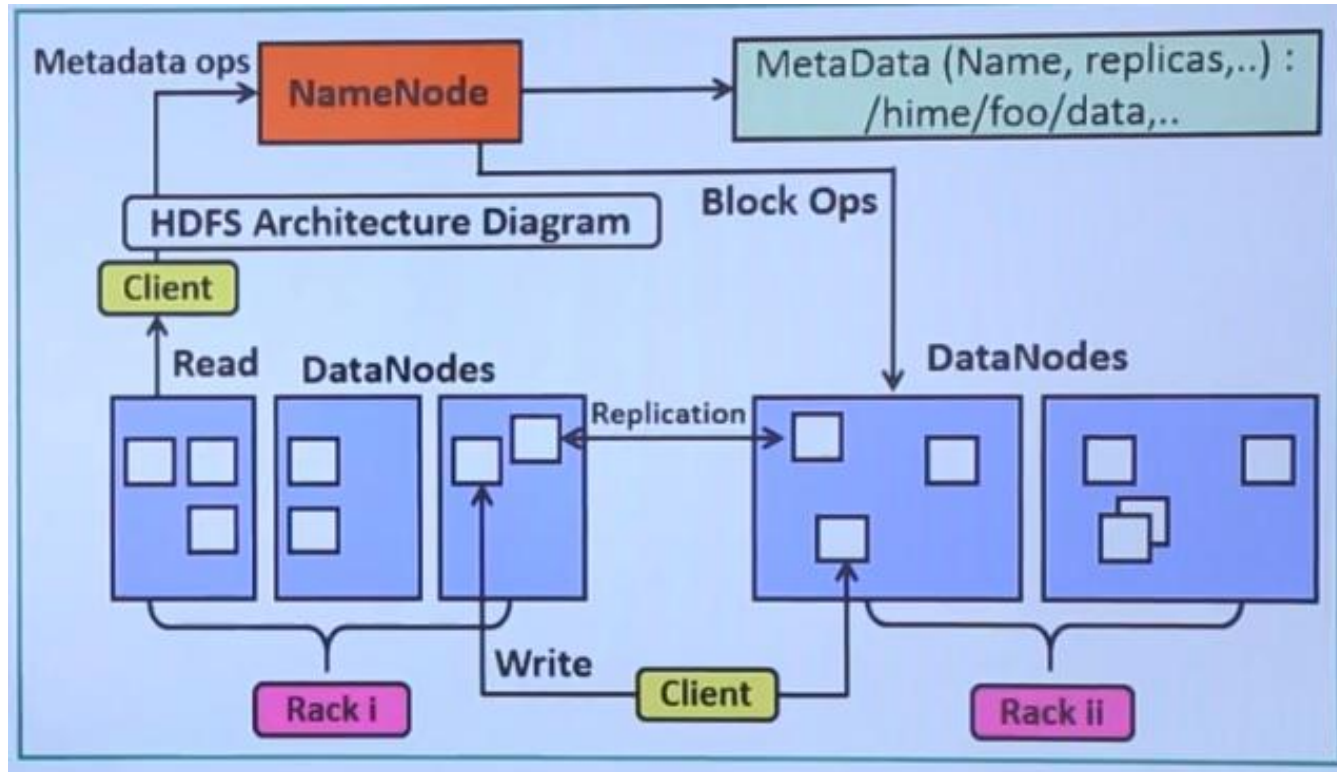


Outils Big Data – The Hadoop Ecosystem



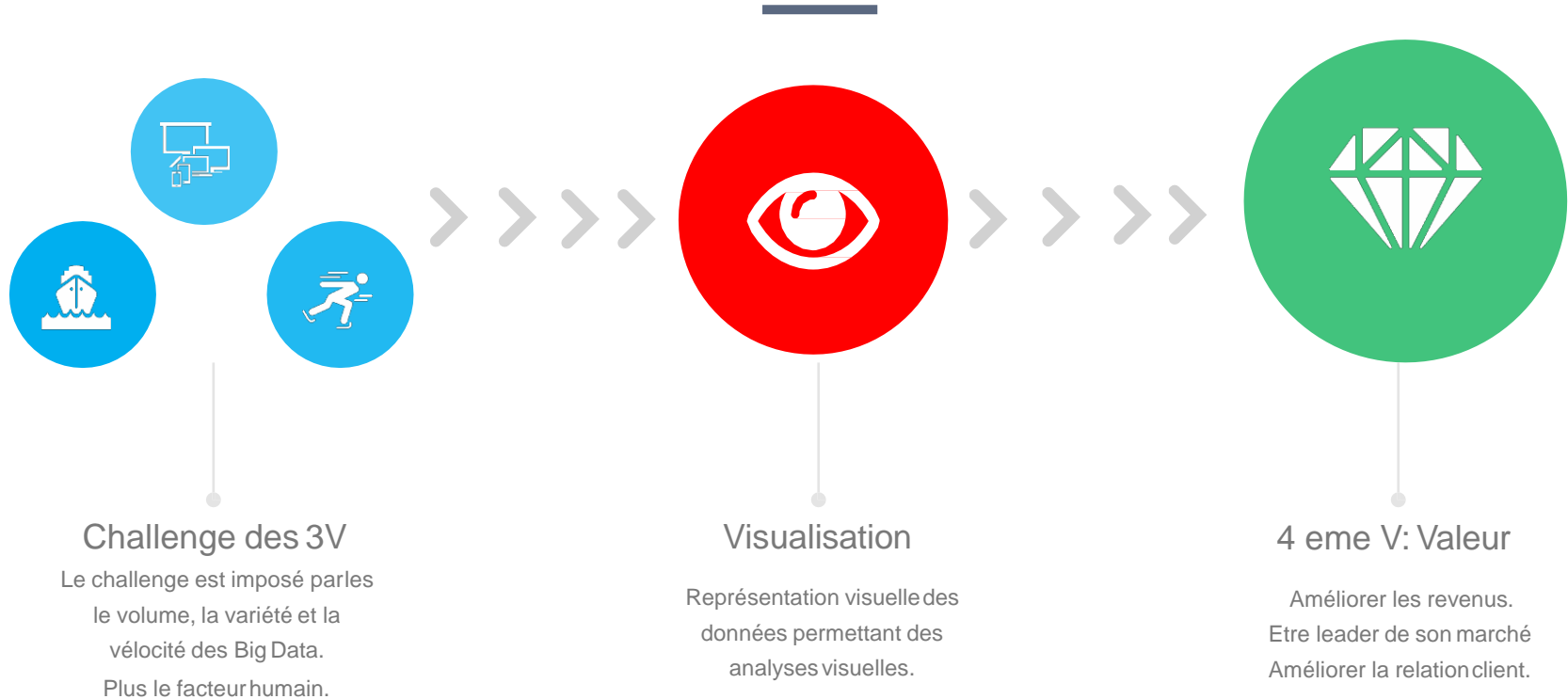
Outils Big Data - Hadoop

HDFS Architecture



Les Challenges de visualization des Big Data

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

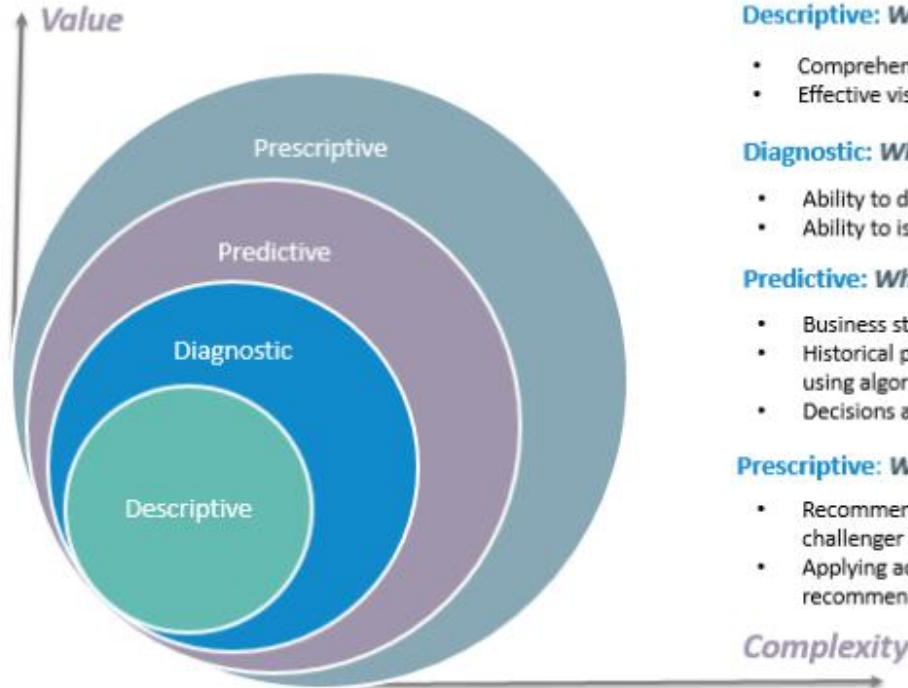


Question ?

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

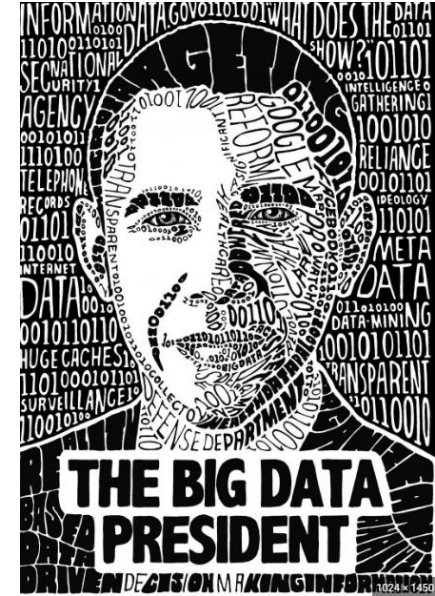
Predictive: *What's likely to happen?*

- 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

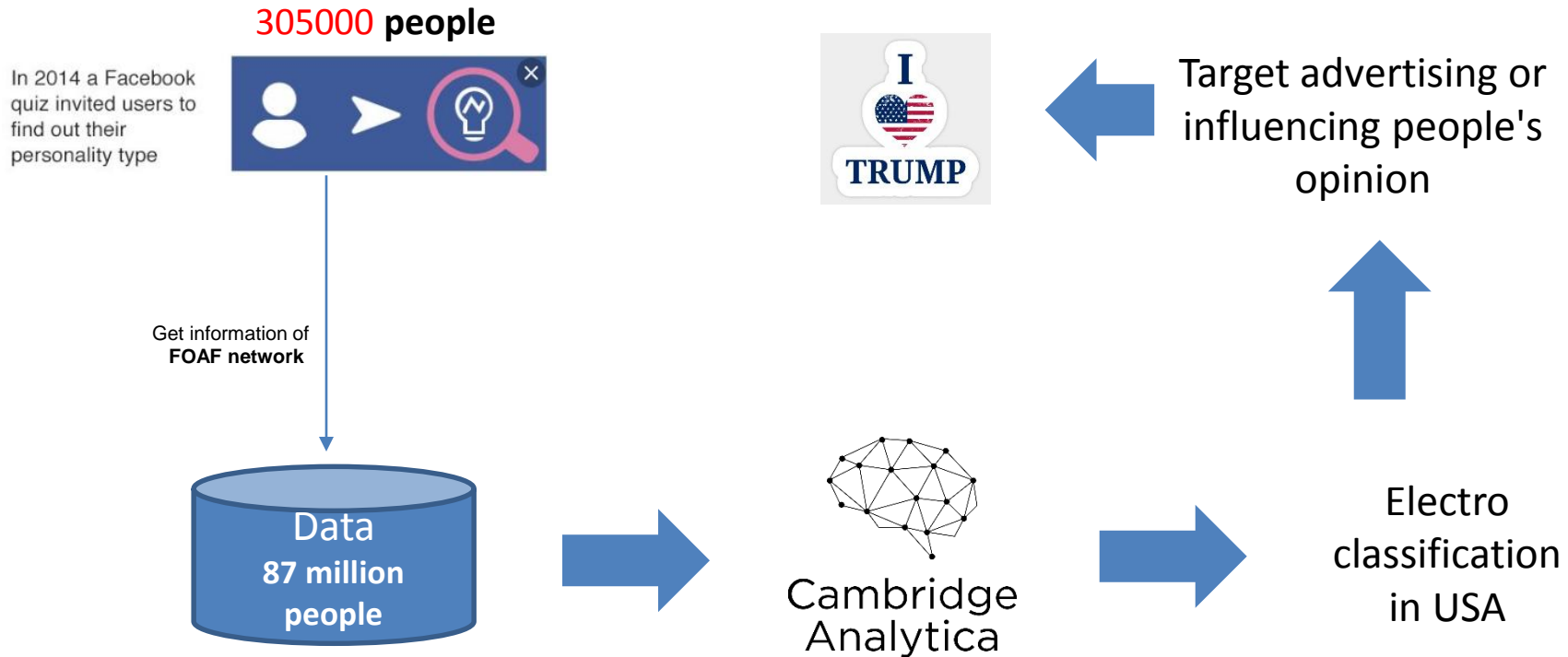
Prescriptive: *What do I need to do?*

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

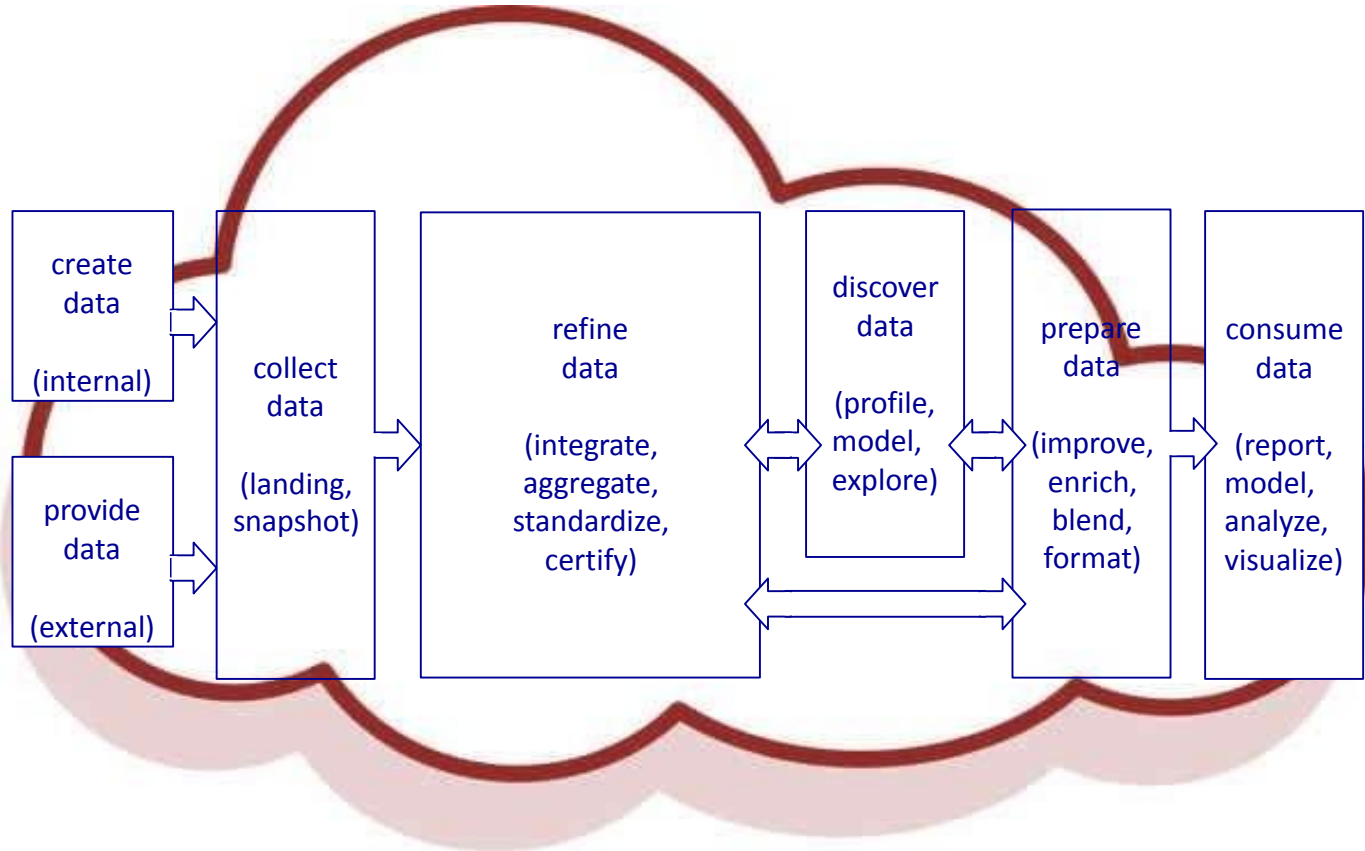
Case study :Opinion Mining in Social Big Data



Case study :Opinion Mining in Social Big Data



Illustrate How Cloud and Big Data related to each other ?



Exercise

Data Mining applications and bigdata

- Select one of the following industries and answer the question below
Retail industry, Banking industry, Insurance, Healthcare, Gouvernement, 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



Questions and Answers

I Hope I Succeeded to clarify what is Big Data