New Approach for scheduling tasks and/or jobs in Big Data Cluster

IT College

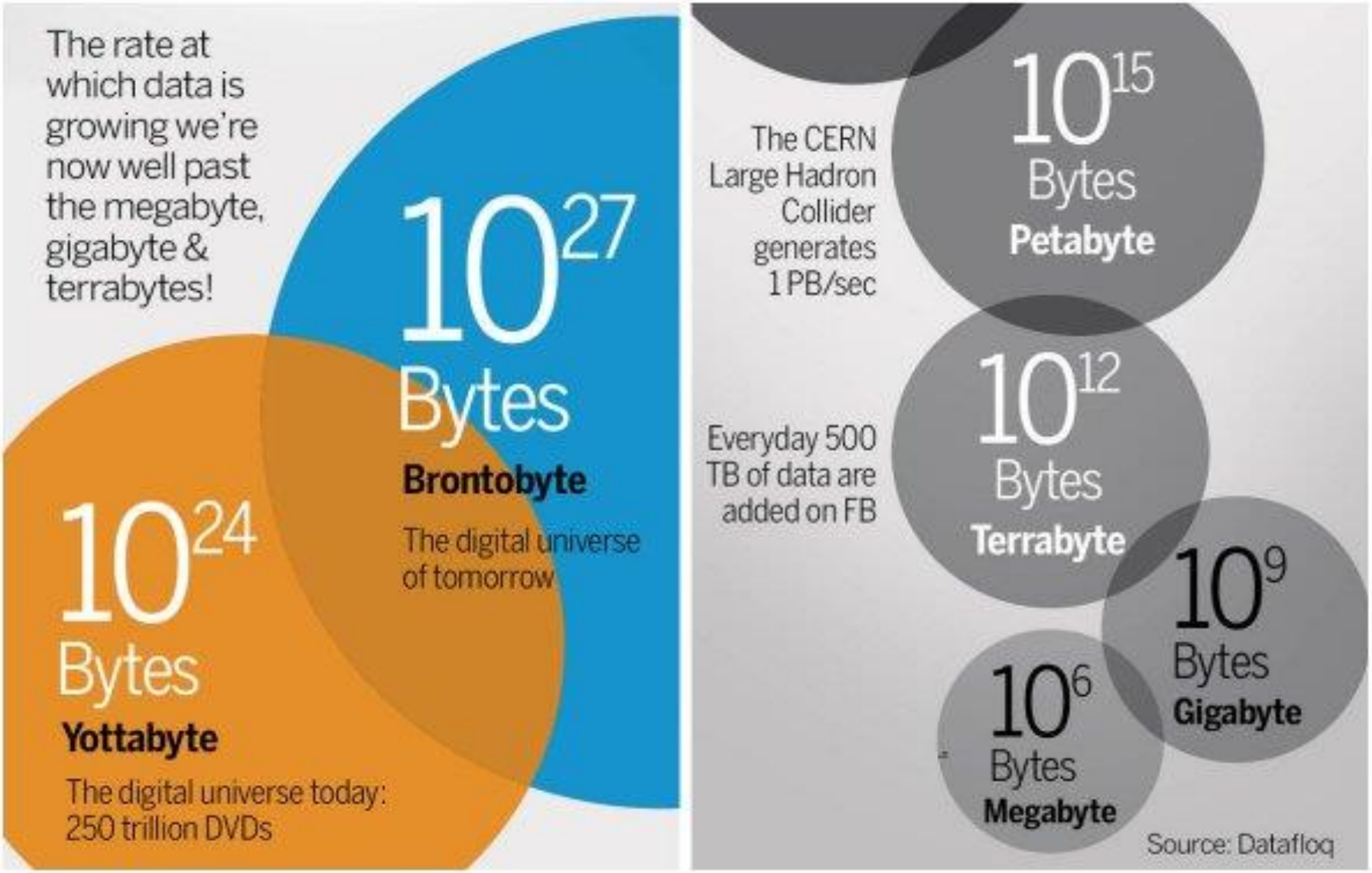
Dr. Karim Hadjar, Chairperson of MS Dept.

Agenda

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

- Introduction
- What is Big Data?
- The 4 characteristics of Big Data V4s
- Different Categories of Data
- Unstructured data is exploding
- Apache Hadoop
- Hadoop Ecosystem
- Scheduler role in Big Data
- Our new approach for scheduling tasks and/or jobs in Big Data Clusters
- Conclusion

Introduction



What is Big Data?

- Definition: “extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions”.

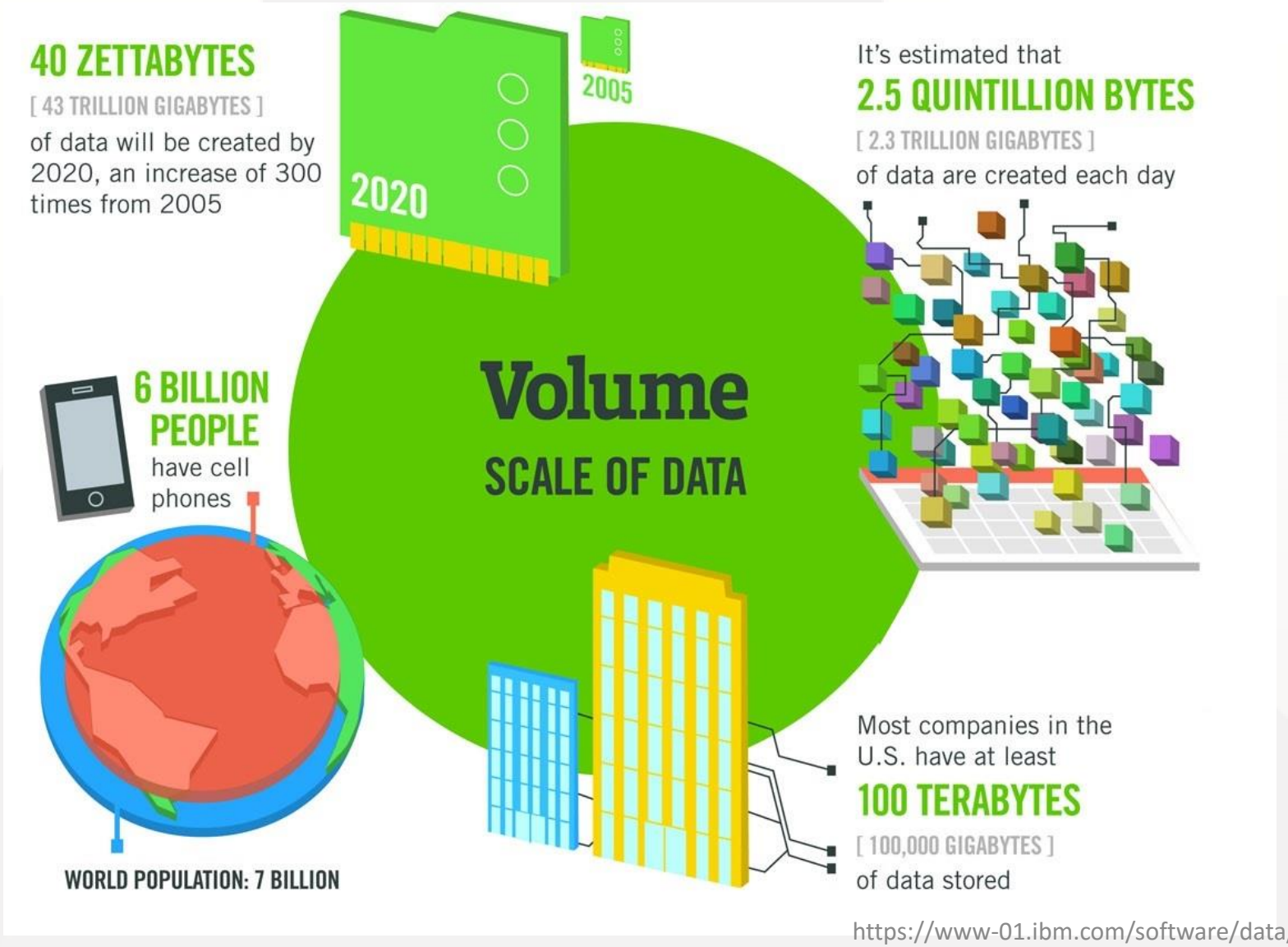
The 4 characteristics of Big Data V4s (1/5)

- Volume
  - Scale of Data
- Velocity
  - Analysis of Streaming Data
- Variety
  - Different forms of Data
- Veracity
  - Uncertainty of Data

The 4 characteristics of Big Data V4s (1/5)

- Volume
  - Scale of Data
- Velocity
  - Analysis of Streaming Data
- Variety
  - Different forms of Data
- Veracity
  - Uncertainty of Data

The 4 characteristics of Big Data V4s (2/5)



https://www-01.ibm.com/software/data/bigdata/

The 4 characteristics of Big Data V4s (3/5)

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

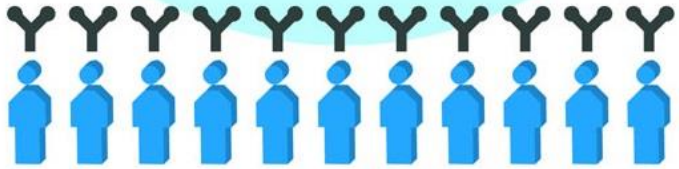
Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

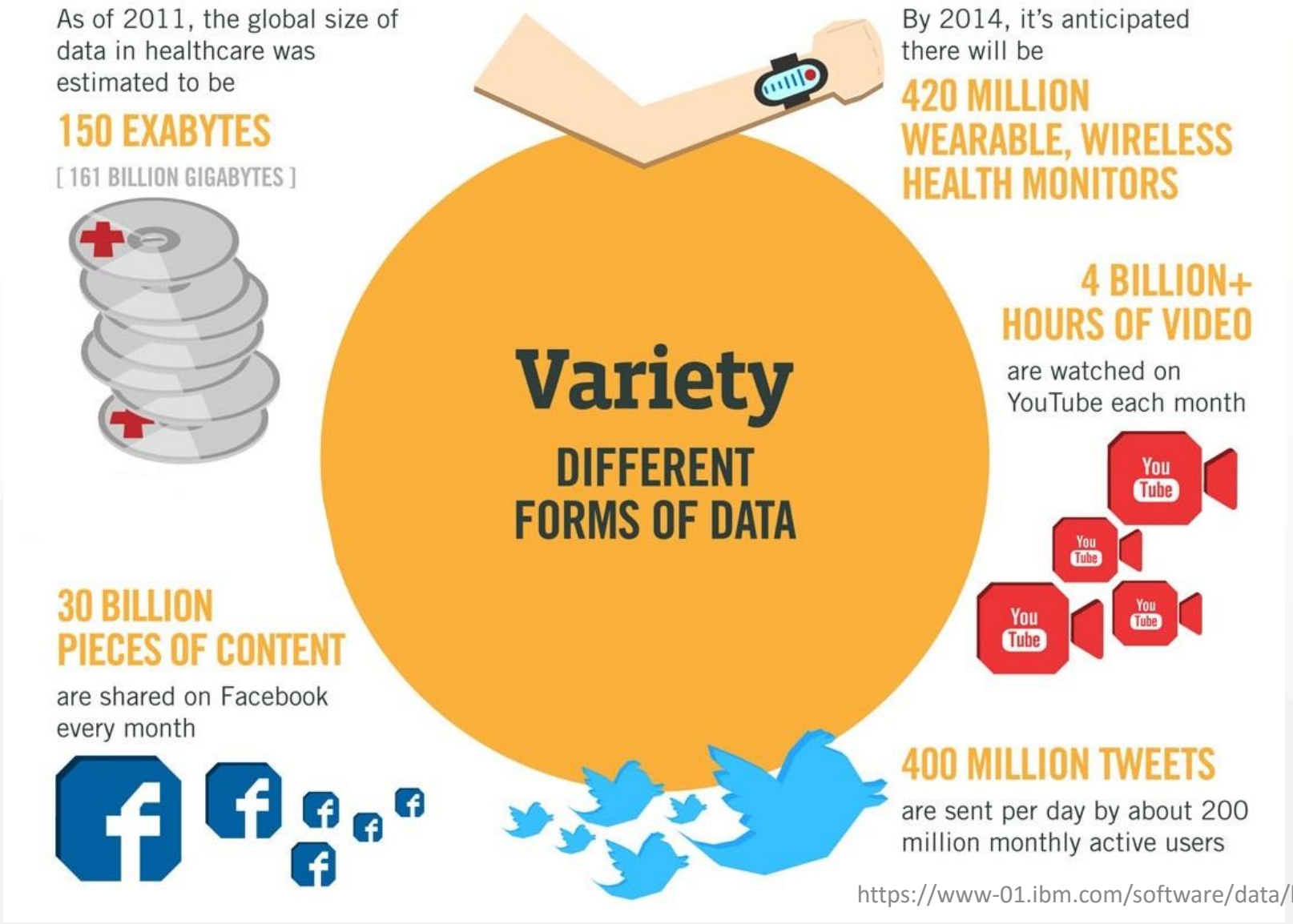
18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth



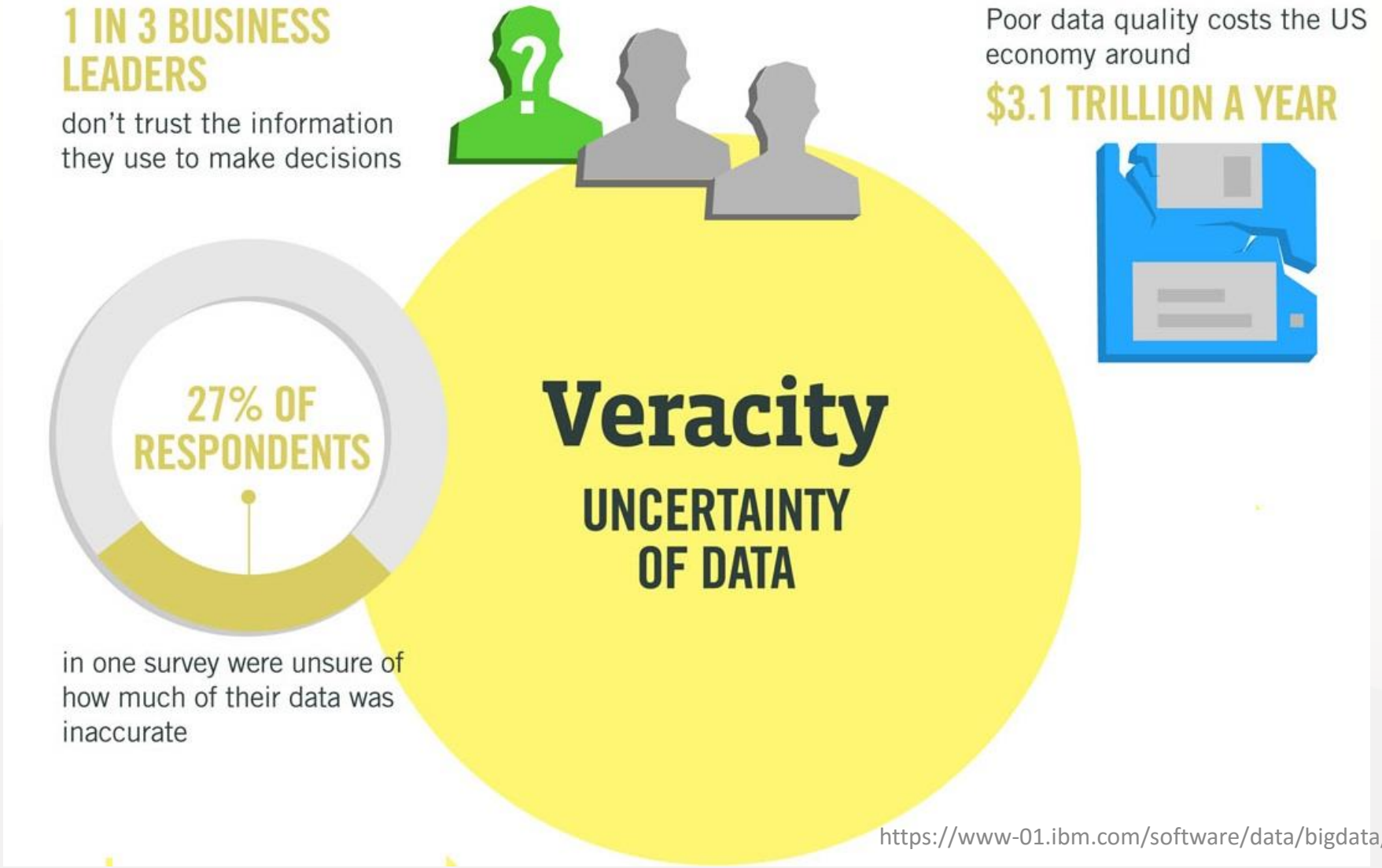
https://www-01.ibm.com/software/data/bigdata/

The 4 characteristics of Big Data V4s (4/5)



https://www-01.ibm.com/software/data/bigdata/

The 4 characteristics of Big Data V4s (5/5)



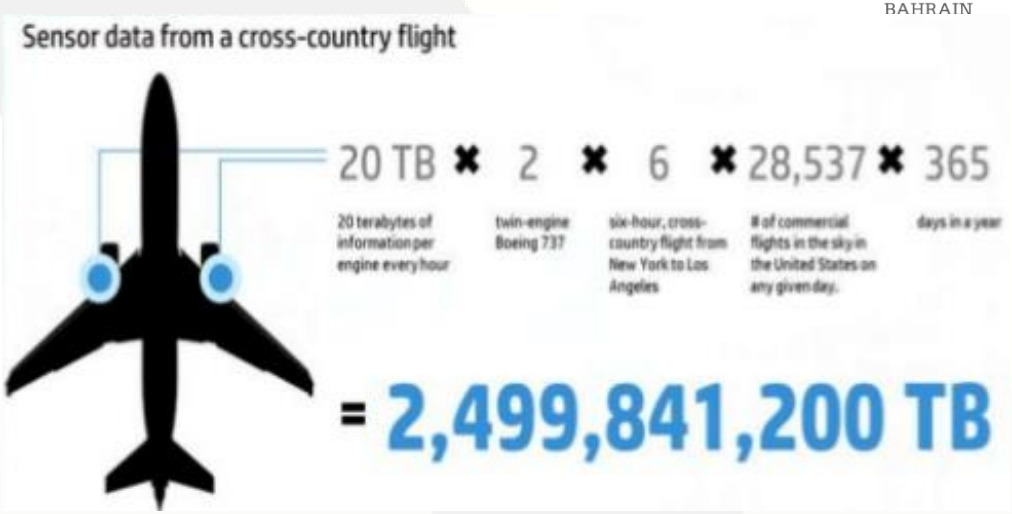
https://www-01.ibm.com/software/data/bigdata/

Different Categories of Data

- XML Files, email body
  - **Semi-structured**
- Audio, Video, Image Files, Archived documents
  - **Unstructured Data**
- Data from Entreprise Systems (ERP, CRM)
  - **Structured Data**

Unstructured data is exploding (1/2)

- IOT (number of wearable devices, Number of wireless devices (RFID, WIFI,…)



- Number of uploaded videos on social networks and on Youtube
  - 4 Billions of Hours are watched on Youtube
- Number of pieces of content exchanged on social networks
  - 400 Millions Tweets are sent per day

Unstructured data is exploding (2/2)

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

- 800% growth in data volume within the next 5 years
- Amount of unstructured data is growing 62% faster
- 80% of data will be unstructured data in 2019 (source Gartner)

Source: Gartner & IDC

Apache Hadoop

- The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing
- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage

Source: Apache Hadoop

Popular Hadoop Distributions

cloudera®

MAPR®

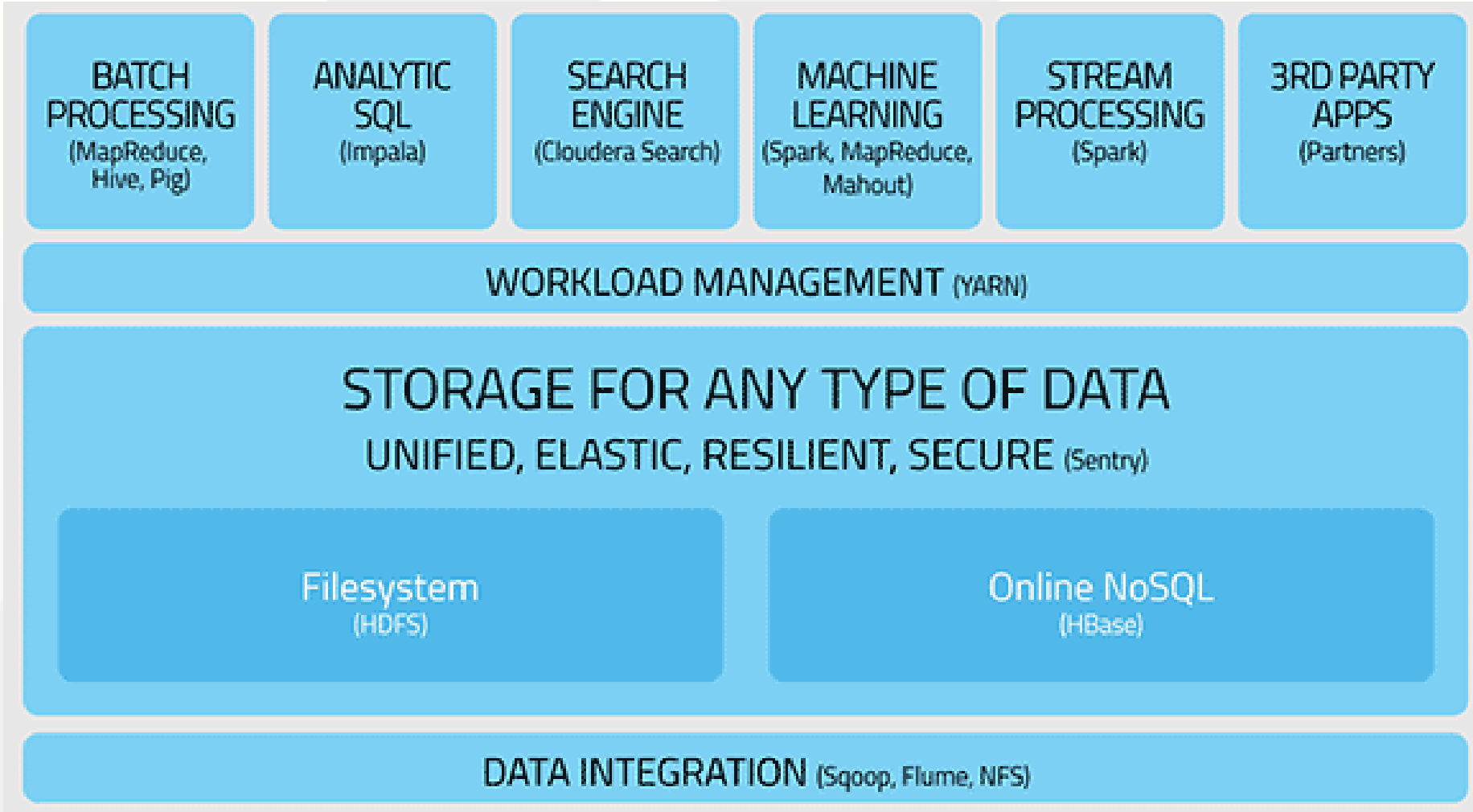
Hadoop Popular Programming Languages

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

python™

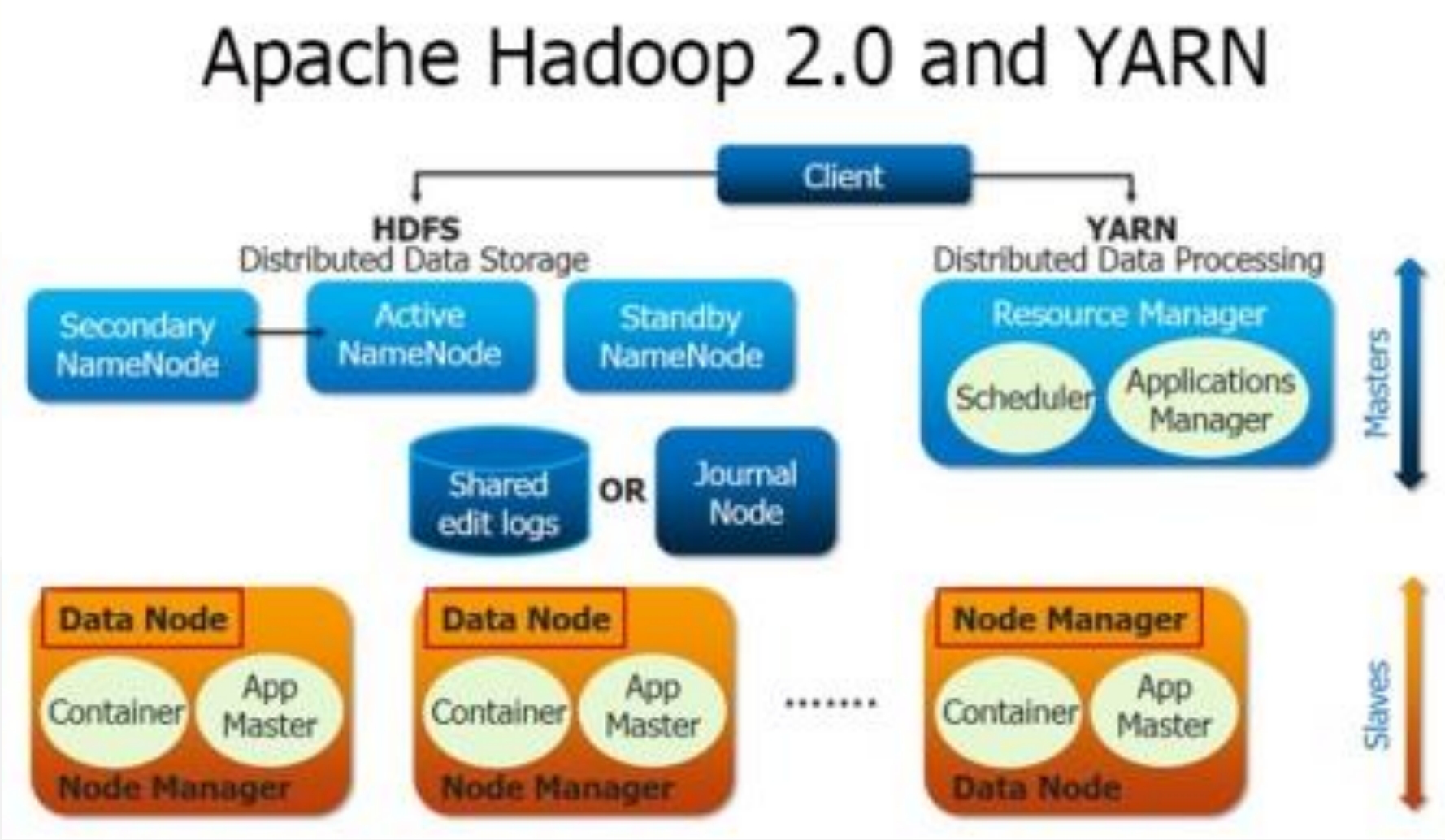
Cloudera Hadoop Ecosystem

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN



Source: Cloudera

Scheduler Role in Big Data (1/2)



Scheduler Role in Big Data (2/2)

- In order to achieve greater performance an efficient scheduler needs to be implemented
- Scheduling is a technique of assigning jobs to available resources in a manner to minimize starvation and maximize resource utilization
- Performance of scheduling technique can be improved by applying constraints
- Various scheduling algorithms are proposed in the past few years for optimal utilization of cluster resources

Our new approach for scheduling tasks and/or jobs in Big Data Cluster (1/2)

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

- Is based on resources of the Data Nodes
  - CPU Load
  - RAM Load
  - I/O Load
  - Network load
  - Type of the Job (Spark, Hbase, Impala, …)
- Job Scheduler computes the aforementioned resources load according to this formula:

$$\boldsymbol{RL = (CPU\ L)^{\alpha} + (RAM\ L)^{\beta} + (I/O\ L)^{\gamma} + (Network\ L)^{\delta}}$$

Our new approach for scheduling tasks and/or jobs in Big Data Cluster (2/2)

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

For every Data Node in the Cluster (**1 cluster**)

Get the resources load

Switch (type of job){

Case: Spark:

$$\alpha = 1\ \beta = 1\ \gamma = 0.5\ \delta = 0.5$$

Case Hbase:

$$\alpha = 0.7\ \beta = 0.7\ \gamma = 1\ \delta = 0.7$$

Case MapReduce:

$$\alpha = 0.5\ \beta = 0.7\ \gamma = 1\ \delta = 1$$

… }

End For

Sort the array of resources' load

Assign Taks and/or Jobs to the Data Nodes

Conclusion

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

- I have presented a new approach for scheduling tasks and/or jobs in Big Data clusters based on data nodes resources load

- In the future, I will introduce machine learning (Artificial Neural Networks) within the scheduler in order to efficiently assign tasks and/or jobs to data nodes