



**Your  
global  
future  
begins  
here**

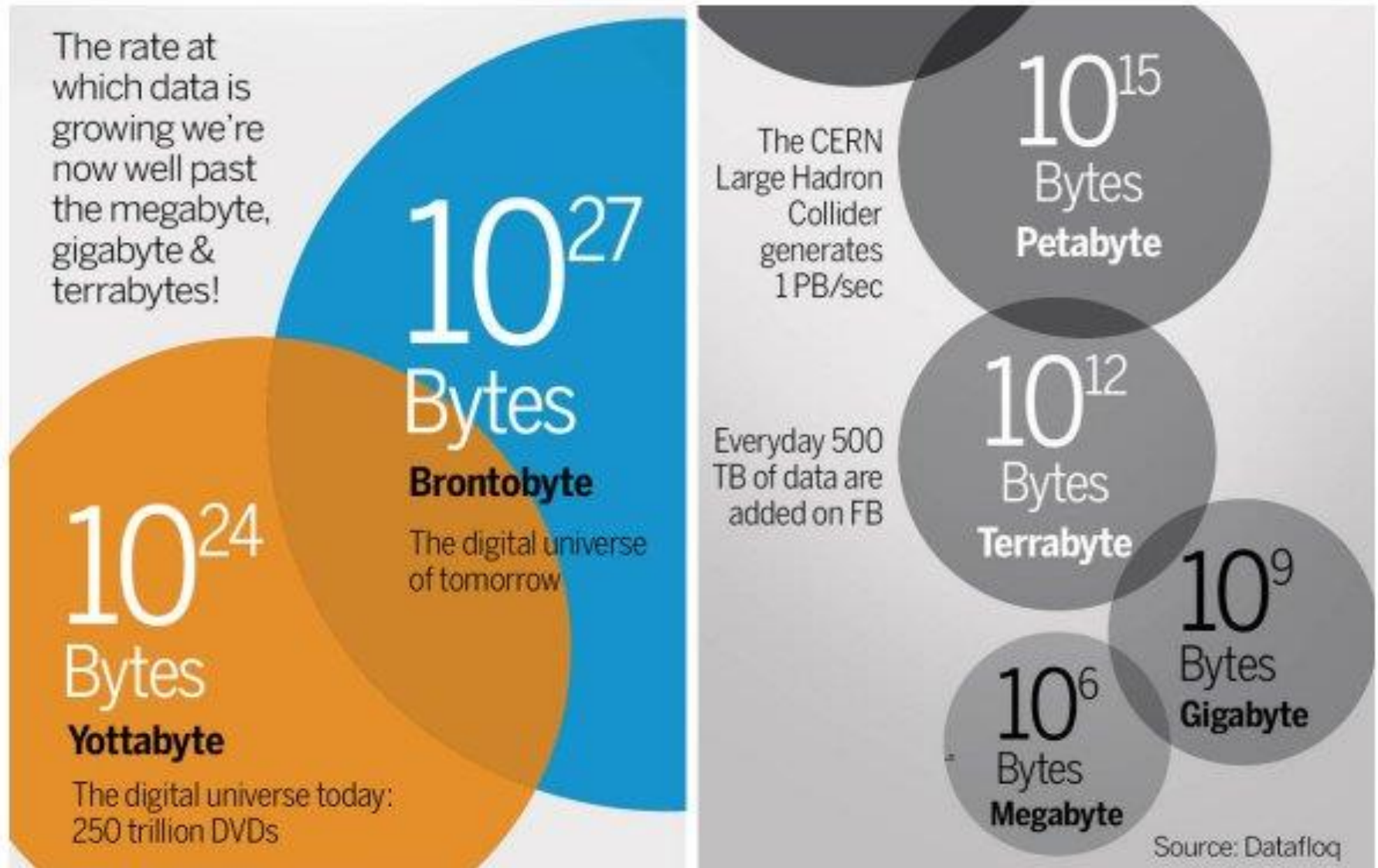
# IT College



Dr. Karim Hadjar, Chairperson of MS Dept.

- 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)

## 40 ZETTABYTES

[ 43 TRILLION GIGABYTES ]

of data will be created by 2020, an increase of 300 times from 2005



**6 BILLION PEOPLE**

have cell phones



WORLD POPULATION: 7 BILLION

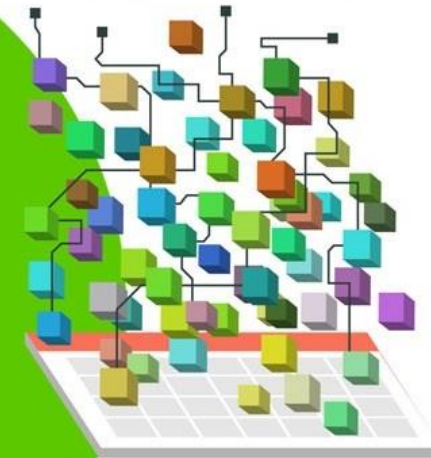
## Volume SCALE OF DATA

It's estimated that

**2.5 QUINTILLION BYTES**

[ 2.3 TRILLION GIGABYTES ]

of data are created each day



Most companies in the U.S. have at least

**100 TERABYTES**

[ 100,000 GIGABYTES ]

of data stored



<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**

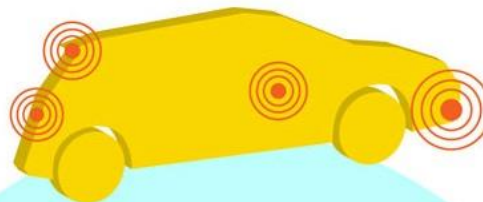
during each trading session



By 2016, it is projected there will be

**18.9 BILLION NETWORK CONNECTIONS**

– almost 2.5 connections per person on earth



Modern cars have close to

**100 SENSORS**

that monitor items such as fuel level and tire pressure

**Velocity**  
**ANALYSIS OF STREAMING DATA**

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



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

As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**

[ 161 BILLION GIGABYTES ]



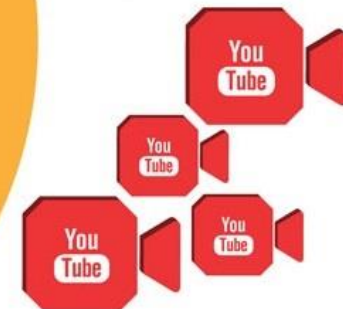
By 2014, it's anticipated there will be

**420 MILLION  
WEARABLE, WIRELESS  
HEALTH MONITORS**



**4 BILLION+  
HOURS OF VIDEO**

are watched on  
YouTube each month



**30 BILLION  
PIECES OF CONTENT**

are shared on Facebook  
every month



**Variety**  
**DIFFERENT  
FORMS OF DATA**



**400 MILLION TWEETS**

are sent per day by about 200  
million monthly active users

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

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

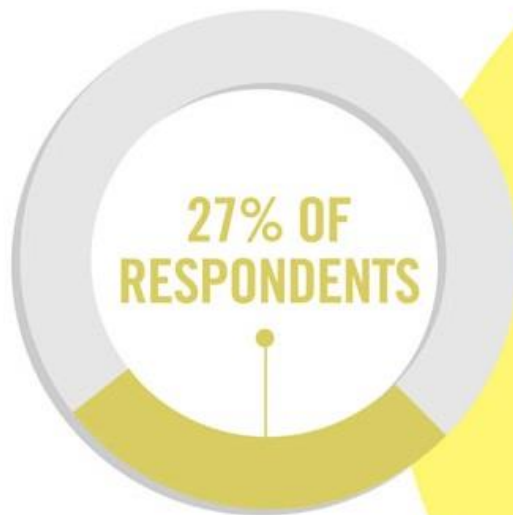
## 1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around

**\$3.1 TRILLION A YEAR**



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

## Veracity

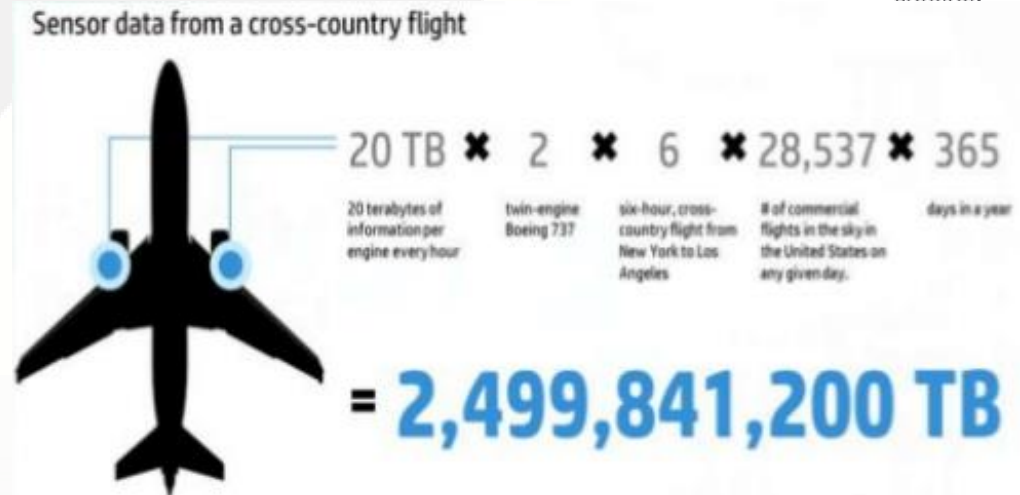
### UNCERTAINTY OF DATA

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

- XML Files, email body
  - **Semi-structured**
- Audio, Video, Image Files, Archived documents
  - **Unstructured Data**
- Data from Enterprise 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)

---

- 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





- 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

# Popular Hadoop Distributions

---



cloudera®



MAPR®

# Hadoop Popular Programming Languages

---



# Cloudera Hadoop Ecosystem



**BATCH  
PROCESSING**  
(MapReduce,  
Hive, Pig)

**ANALYTIC  
SQL**  
(Impala)

**SEARCH  
ENGINE**  
(Cloudera Search)

**MACHINE  
LEARNING**  
(Spark, MapReduce,  
Mahout)

**STREAM  
PROCESSING**  
(Spark)

**3RD PARTY  
APPS**  
(Partners)

**WORKLOAD MANAGEMENT** (YARN)

**STORAGE FOR ANY TYPE OF DATA**

UNIFIED, ELASTIC, RESILIENT, SECURE (Sentry)

**Filesystem**  
(HDFS)

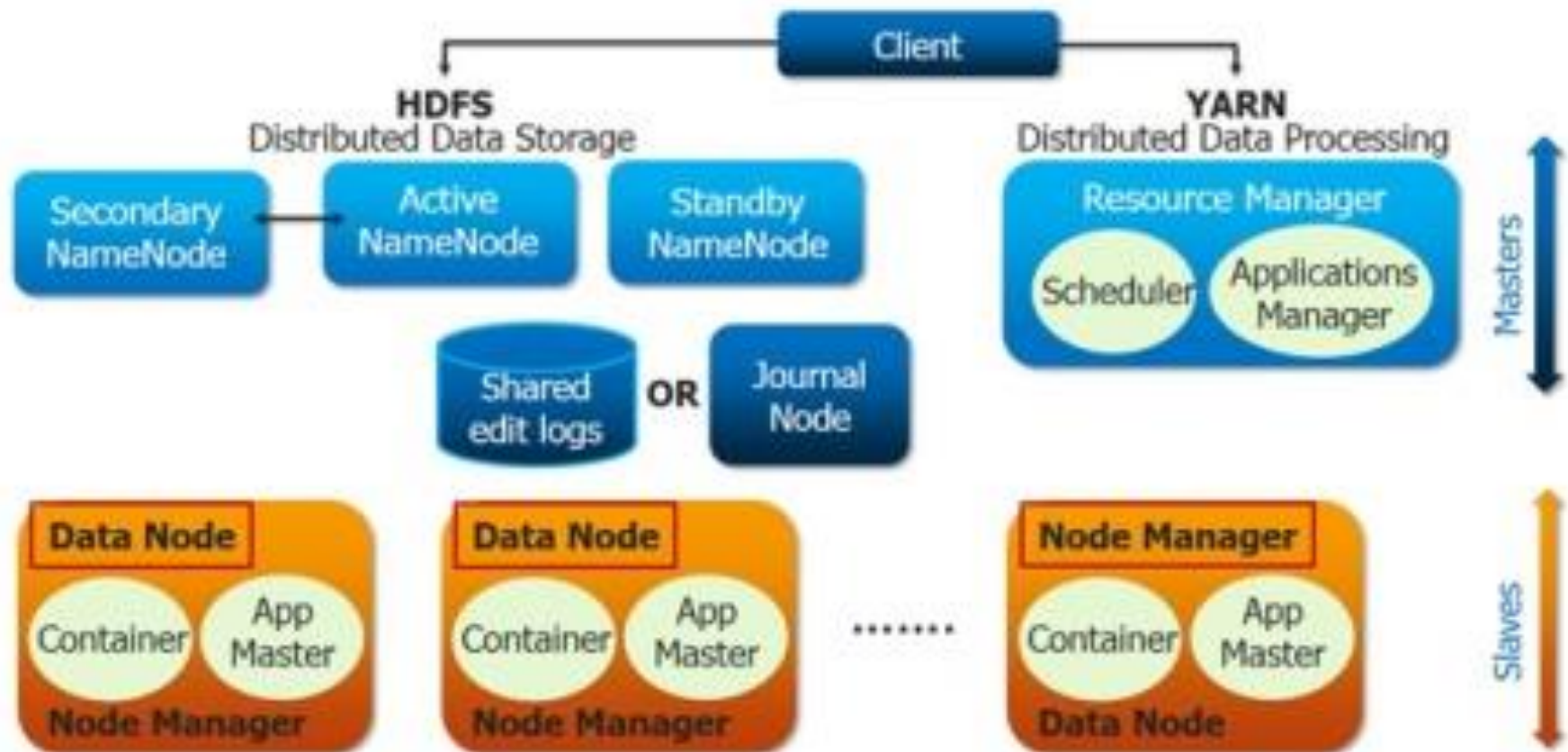
**Online NoSQL**  
(HBase)

**DATA INTEGRATION** (Sqoop, Flume, NFS)

Source: Cloudera

# Scheduler Role in Big Data (1/2)

## Apache Hadoop 2.0 and YARN





- 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)

---

- 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:

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

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

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

Get the resources load

Switch (type of job){

Case: Spark:

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

Case Hbase:

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

Case MapReduce:

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

... }

End For

Sort the array of resources' load

Assign Tasks and/or Jobs to the Data Nodes

- 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