BIG DATA

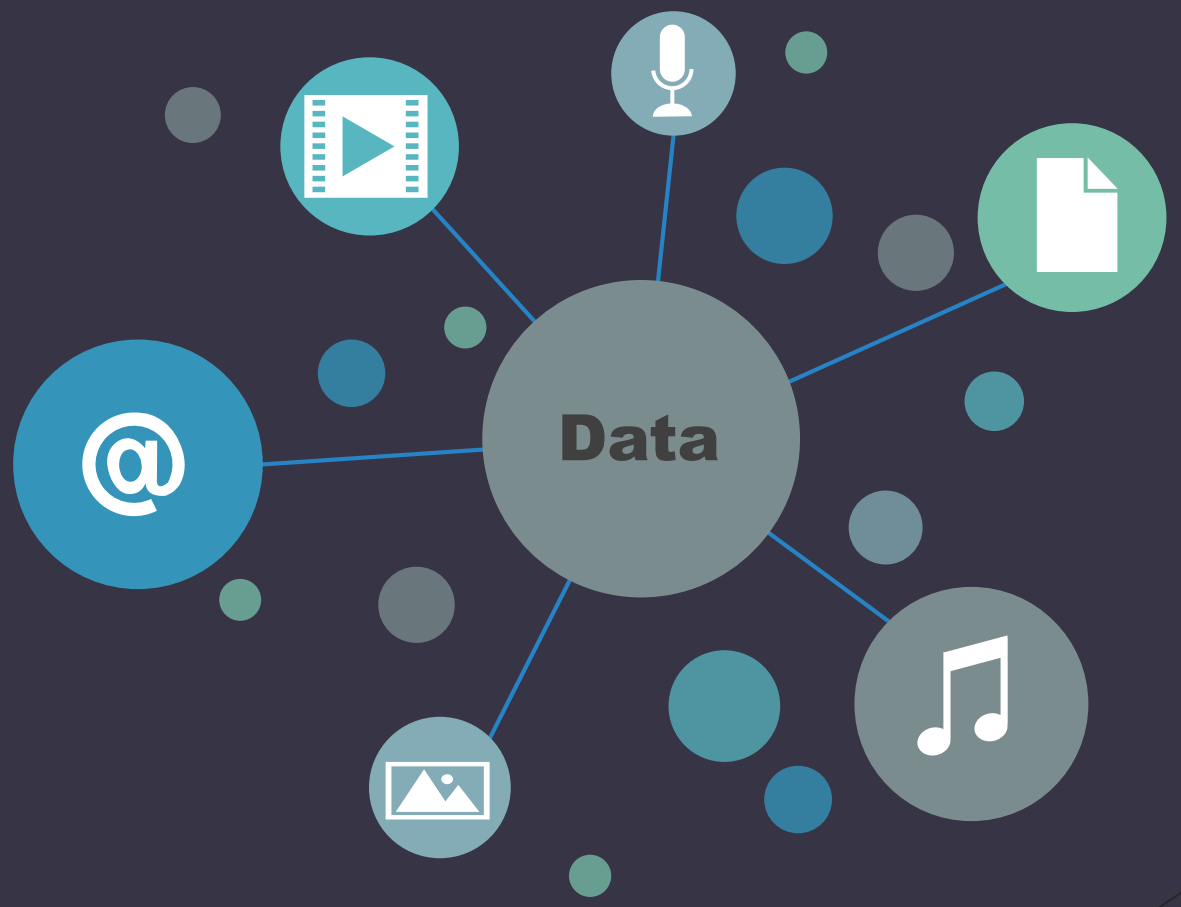
About & Usage

AGENDA

- What Big Data is
- Big Data as an Opportunity
- Big Data Usages
- What Hadoop is and it's Ecosystem
- Other Solutions

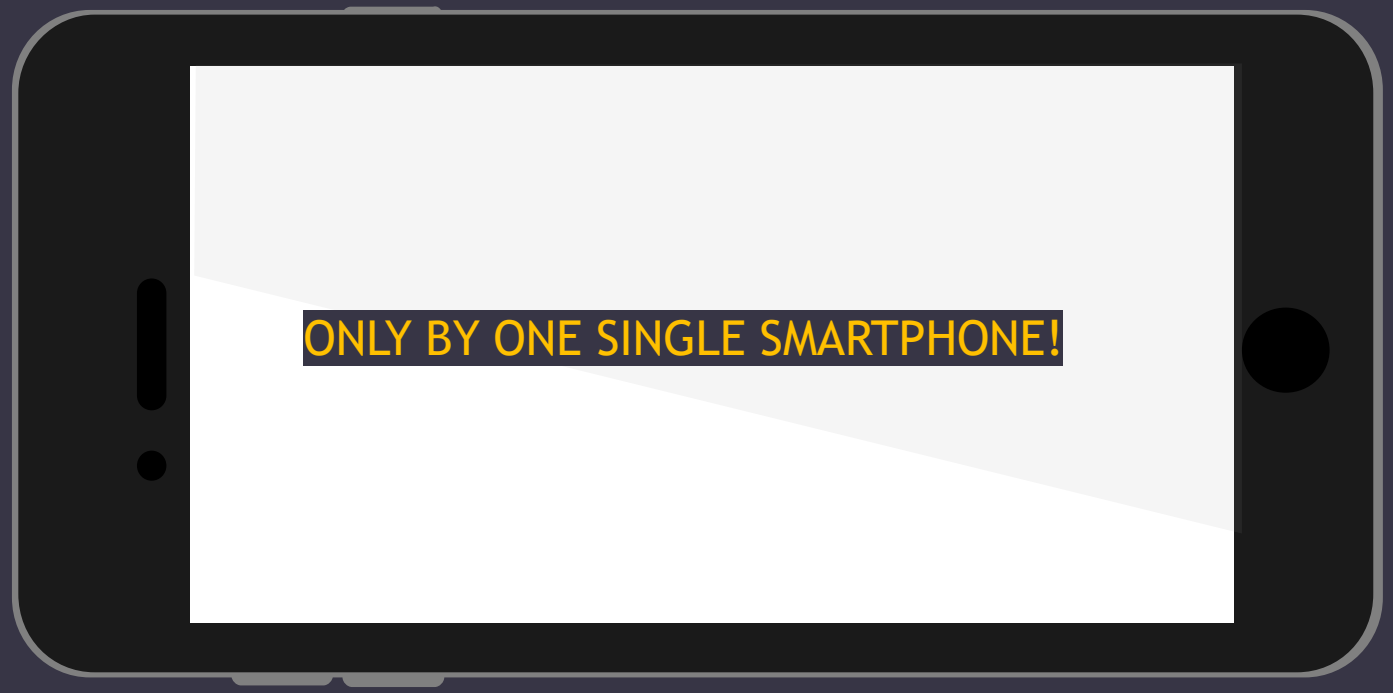
What is Data in general?

- In general, Data is any set of characters that is gathered and translated for some purpose.



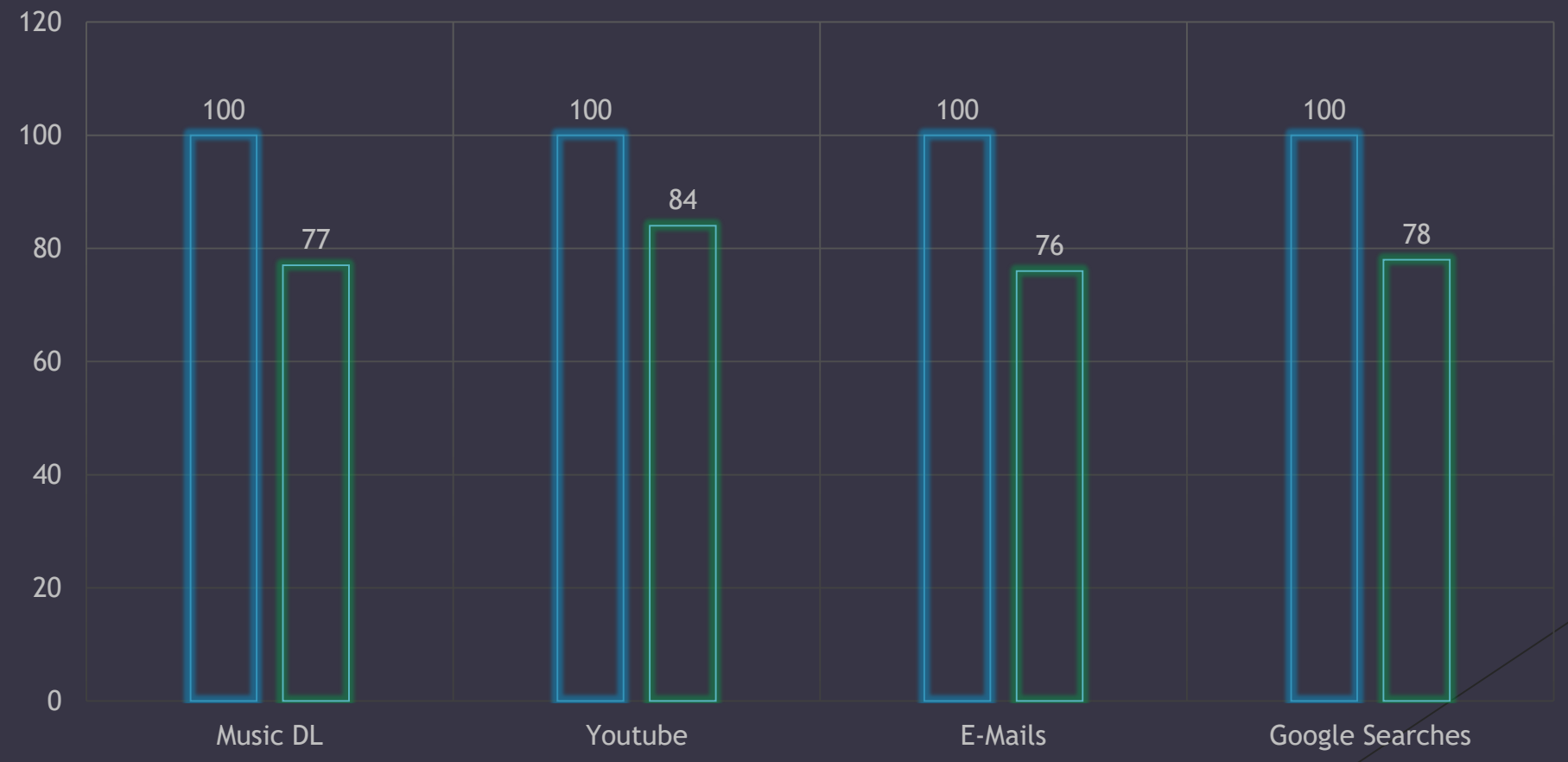
Just a quick heads-up

- About 4 ExaByte* of Data is generated each and every month.
  - * 1 ExaByte is equivalent to 1000000 TeraBytes!



A lot Can Happen in 1 Minute!

A Simple Visual Comparison



00101

But How?

- There are 50 Billion Smart Devices(iot) all around the Globe.
- Each of these devices generate Data!
- In summary:
  - In the year 2020, 1.7MB of Data is generated for each person.

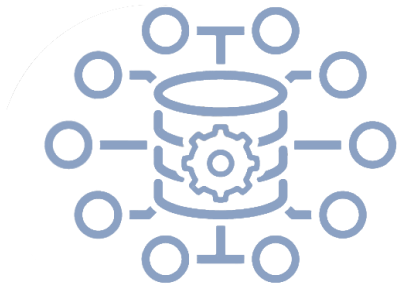
What is Big Data?

- Big Data is a term that describes the large amount of data that inundates a business on a day-to-day basis.

- **Is the amount of Data important?**

A Comparison

Small Data



- Low Volumes
- Batch Velocities
- Structured Varieties

VS

Big Data

- Extreme Volumes
- Real-time Velocities
- Multi-structured Varieties

Objectives of Big Data

Analyzing Costumer Behavior

Combining Multiple Data Sources

Generate Additional Revenue

Be More Responsive To The Market

Improving Customer Service

General usages of Big Data

1. Location Tracking: Gathers real-time Data about traffic & weather conditions and defines the best routes for optimum transportation.

2. Precision Medicine: Hospitals can improve the level of their patient care. Also, efficiency of medication could be improved by analyzing the past records of the patients.

3. Fraud Detection & Handling: Banking and transactions and failure in net banking. It uses Big Data to prevent cyber crimes, card fraud detection.

4. Advertising: One of the biggest players in Big Data are advertisers. Companies like to keep track of user behavior and provide advertisers with them.

5. Entertainment & Media: In this field, Big Data focuses on the people with the right content at the right time.

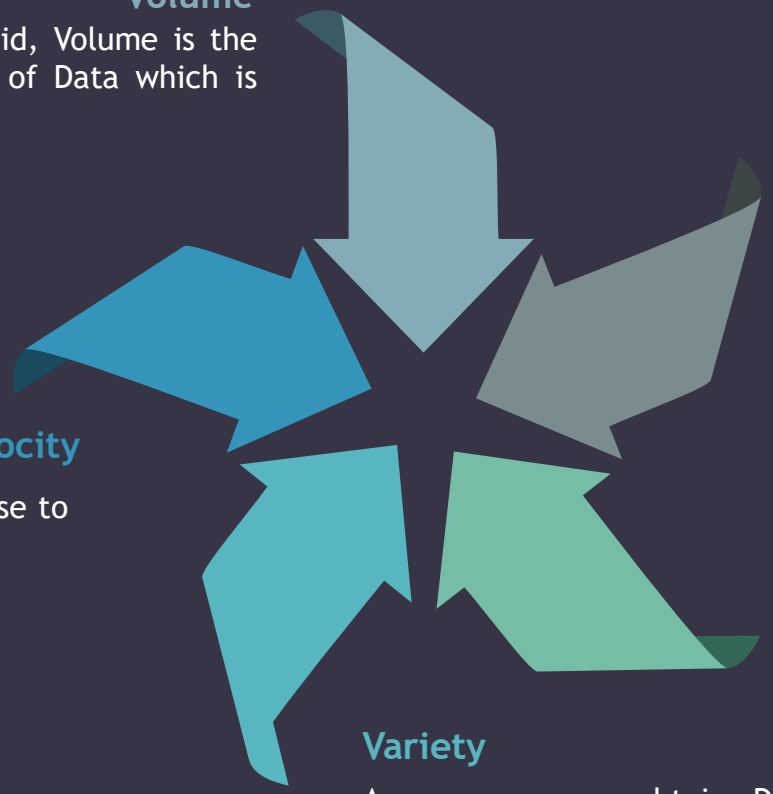
5V

Volume

if we see Big Data as a pyramid, Volume is the base. It refers to the amount of Data which is enormous.

Value

Sits on top of the Big Data pyramid. This refers do the ability to transform a tsunami of Data into business!



Velocity

Companies need Data to flow quickly, as close to real-time as possible.

Veracity

In this context it means quality. The Data needs to be clean and thorough and not be missing something.

Variety

A company can obtain Data from different sources.

01011

Types of Big Data

Unstructured

- Databases
- Data Warehouses
- Enterprise Systems

Structured

- Analog Data
- GPS Tracking Info
- Audio/Video Streaming

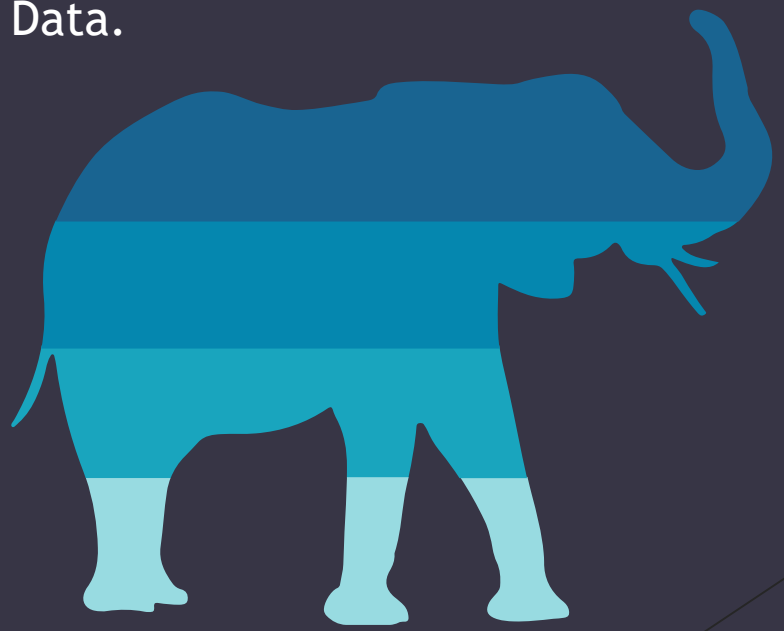
Semi-structured

- XML
- E-Mail
- HTML

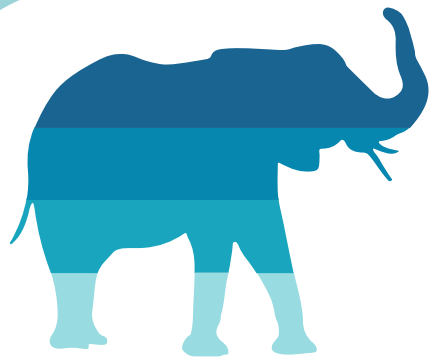
01100

“Hadoop”as a Solution

- **What is Hadoop?**
  - Hadoop is an open-source software framework for storing Data and running applications on clusters of commodity hardware. It provides massive storage for any kind of Data.



HDFS



Name Node

DataNode

DataNode

DataNode

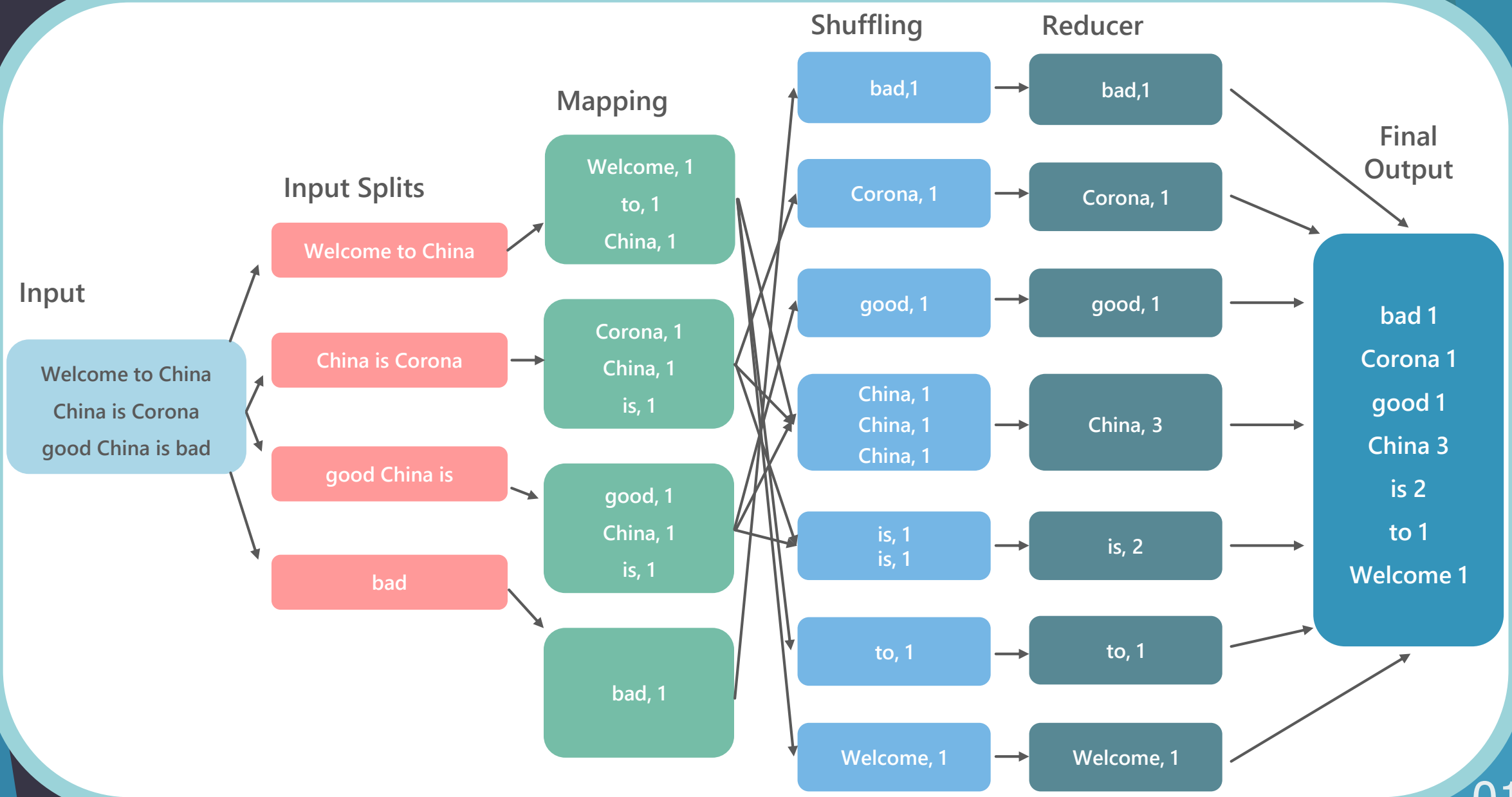
DataNode

DataNode

Data Serving

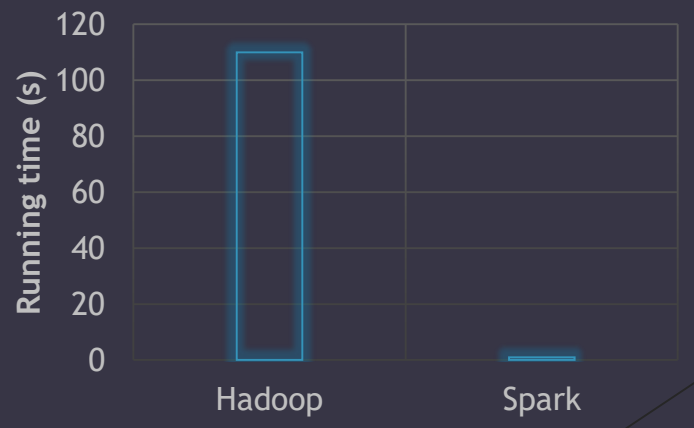
- Nodes Write to the Local Disk

MapReduce


Input
Welcome to China
China is Corona
good China is bad
Input Splits
Welcome to China
China is Corona
good China is
bad
Mapping
Welcome, 1
to, 1
China, 1
Corona, 1
China, 1
is, 1
good, 1
China, 1
is, 1
bad, 1
Shuffling
bad,1
Corona, 1
good, 1
China, 1
China, 1
China, 1
is, 1
is, 1
to, 1
Welcome, 1
Reducer
bad,1
Corona, 1
good, 1
China, 3
is, 2
to, 1
Welcome, 1
Final Output
bad 1
Corona 1
good 1
China 3
is 2
to 1
Welcome 1

Enters "Spark"

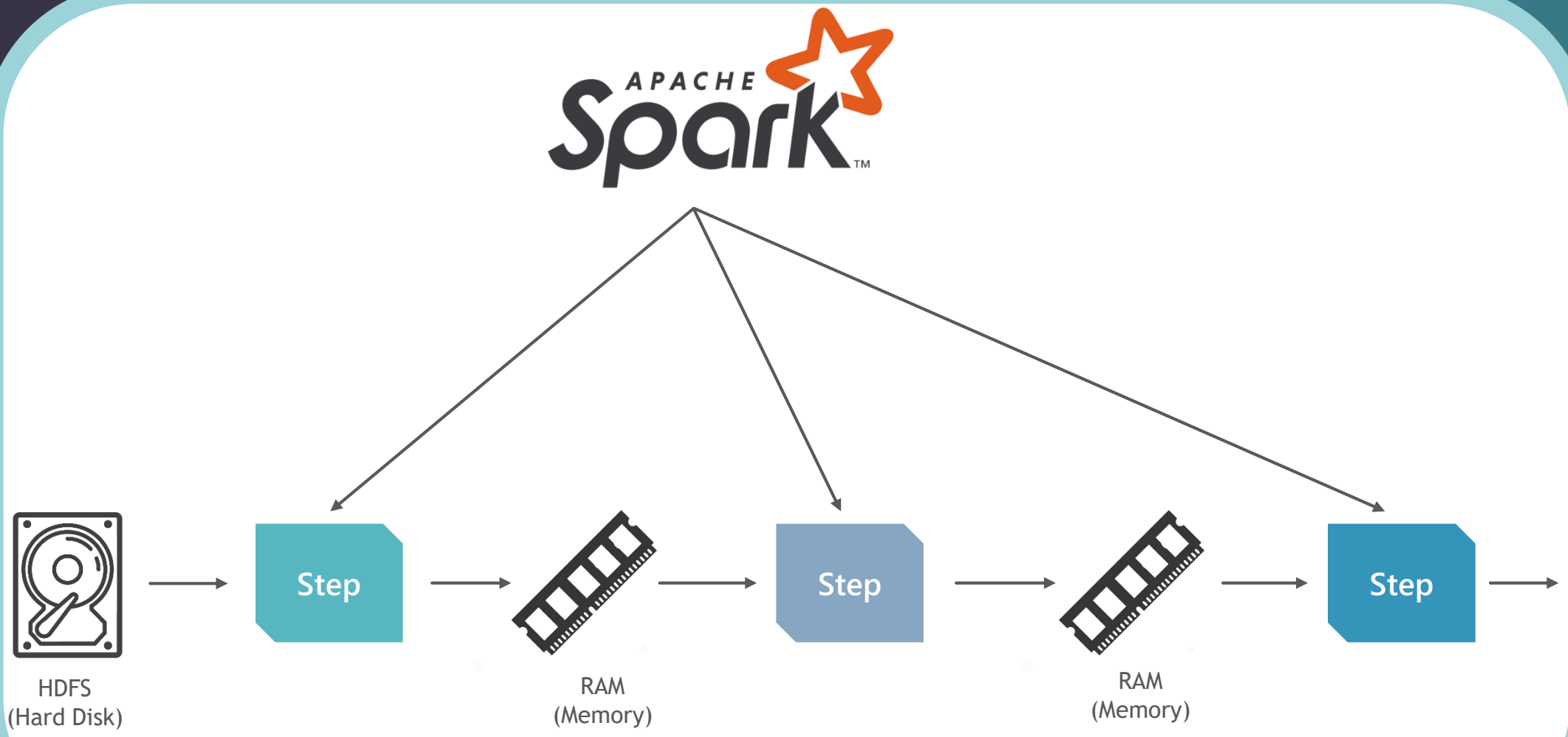
- Spark?
  - Spark is a unified analytics engine for large scale Data processing.
- Is it fast?
  - In terms of speed, it achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer and a physical execution engine.
- But how fast?
  - Just a quick demonstration:



Logistic regression in Hadoop and Spark

10000

How Spark Works



Differences between “Hadoop” and “Spark”

Factors	Spark	Hadoop
Speed	100x faster than Hadoop	Faster than traditional
Written in	Scala/Python/R/…	Java
Data Processing	Batch/Real-time/ Iterative/Interactive/Graph	Batch Processing
Ease of use	Compact & easier than Hadoop	Complex & Lengthy
Caching	Caches the Data in memory & enhances the system performance	Doesn’t support caching of Data
Cost	High cost	Low cost

Sources we Used:

- https://hadoop.apache.org/
- https://www.weforum.org/agenda/2017/08/what-happens-in-an-internet-minute-in-2017
- https://spark.apache.org/
- https://www.datasciencecentral.com/profiles/blogs/what-is-the-difference-between-hadoop-and-spark
- https://www.datamation.com/big-data/structured-vs-unstructured-data.html
- https://www.newgenapps.com/blog/5-practical-uses-of-big-data/
- https://maktabkhooneh.org/course/%D8%A2%D9%85%D9%88%D8%B2%D8%B4-%D8%B1%D8%A7%DB%8C%DA%AF%D8%A7%D9%86-%D9%BE%D8%A7%DB%8C%DA%AF%D8%A7%D9%87-%D8%AF%D8%A7%D8%AF%D9%87-%D9%BE%DB%8C%D8%B4%D8%B1%D9%81%D8%AA%D9%87-mk635/
- https://community.tealiumiq.com/t5/Customer-Data-Hub/Structured-Data-vs-Semi-Structured-Data/ta-p/15617

THANK YOU

Amir HassanEbrahimi

Ali Moshrefi

10100