

**Data Fellowship**  
**Module 3 Big Data**





## Session Outline...

### Day 1

- Understanding Big Data
- Discuss the Attributes of the 5 V's
- Discuss the technical challenges of Big Data Storage and Processing

### Day 2

- Understand Hadoop and relationship with Big Data
- Describe components of Hadoop and their roles

**...and then you are done!**

# Understanding Big Data



# What is Big Data?

# Value of Big Data

## Extracting information

Data-driven deals, selected

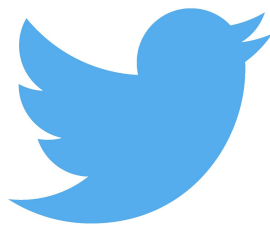
|           | Target company (Date)          | Value of deal, \$bn | Business                         |
|-----------|--------------------------------|---------------------|----------------------------------|
| facebook  | Instagram (2012)               | 1.0                 | Photo sharing                    |
|           | WhatsApp (2014)                | 22.0                | Text/photo messaging             |
| Alphabet  | Waze (2013)                    | 1.2                 | Mapping and navigation           |
| IBM       | The Weather Company (2015)     | 2.0                 | Meteorology                      |
|           | Truven Health Analytics (2016) | 2.6                 | Health care                      |
| intel     | Mobileye (2017)                | 15.3                | Self-driving cars                |
| Microsoft | SwiftKey (2016)                | 0.25                | Keyboard/artificial intelligence |
|           | LinkedIn (2016)                | 26.2                | Business networking              |
| ORACLE    | BlueKai (2014)                 | 0.4                 | Cloud data platform              |
|           | Datalogix (2014)               | 1.0                 | Marketing                        |

Source: Company reports, estimates



**Big Data is data whose scale, distribution, diversity, and / or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.**





Generates 500 million  
Tweets **every day**



Generates 1 TeraByte of  
trading data  
**every day**

Source: [New York Stock Exchange](https://www.nyse.com/)





Generates 4 petabytes of user  
data  
**every day**



# Question

**What other kinds of data sources can you think of that generates data at this scale?**

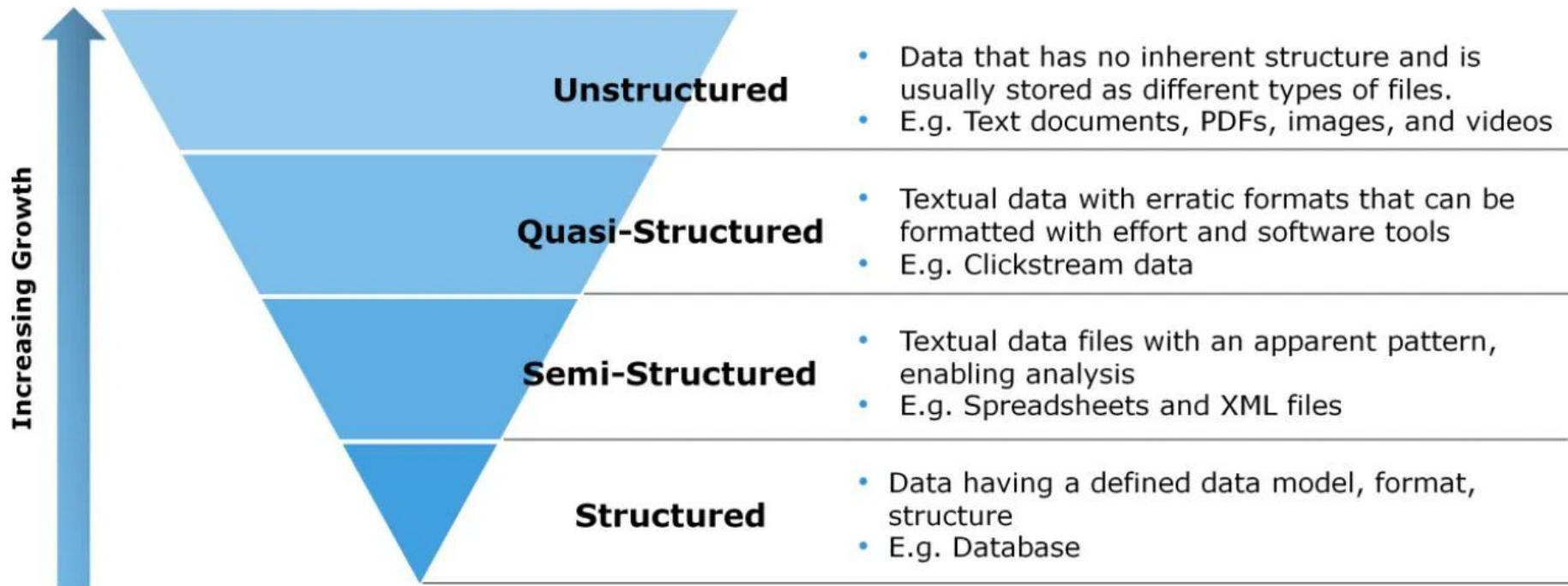


**79% of enterprise executives agree that companies that do not embrace Big Data will lose their competitive position and could face extinction**



# Data Sources

**A reason for this rapid growth of data volume is that the data is coming from different sources in various formats**



# Individual Activity

Look at the the data sources on the handout -  
how would you classify them?

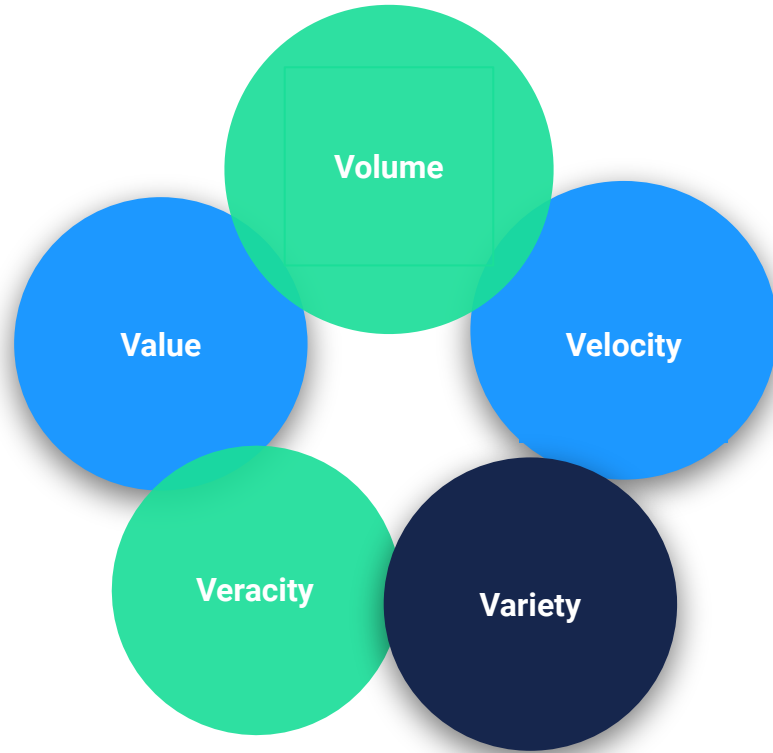
- 1) Structured
- 2) Unstructured
- 3) Semi-Structured
- 4) Quasi-Structured



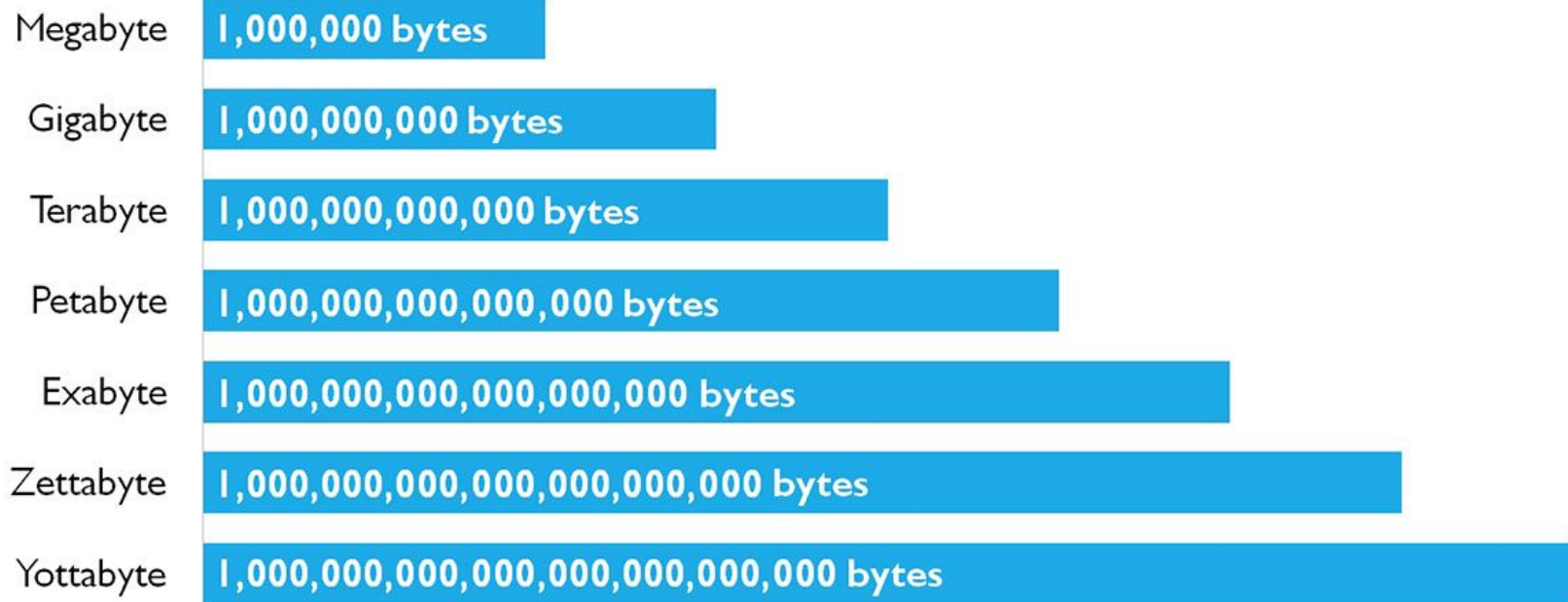
Data  
Sources

# The Five V's





- 1) **Volume** defines the huge amount of data that is produced each day by companies, for example. The generation of data is so large and complex that it can no longer be saved or analyzed using conventional data processing methods. Typically **petabytes or exabytes**.
- 2) **Variety** refers to the diversity of data types and data sources. **80 percent of the data in the world today is unstructured** and at first glance does not show any indication of relationships.
- 3) **Velocity** refers to the speed with which the data is generated, analyzed and reprocessed. Today this is mostly possible within a fraction of a second, known as real time.
- 4) **Veracity** is the authenticity and credibility of the data. Big Data involves working with all degrees of quality, since the Volume factor usually results in a shortage of quality.
- 5) **Value** denotes the added value for companies. Many companies have recently established their own data platforms, filled their data pools and invested a lot of money in infrastructure. It is now a question of generating business value from their investments.



HOW BIG ARE THEY?

1

PETABYTE



13.3 YEARS

OF HD-TV VIDEO

1.5

PETABYTES



SIZE OF THE 10 BILLION  
PHOTOS ON → **FACEBOOK**

20

PETABYTES



THE AMOUNT OF DATA **PER**  
PROCESSED BY **GOOGLE** **DAY**

50

PETABYTES



THE ENTIRE WRITTEN WORKS  
OF MANKIND, FROM THE BEGIN-  
NING OF RECORDED HISTORY,  
IN ALL LANGUAGES

Some have speculated that **5 exabytes** likely equals all of the words ever spoken by humans.



To have recorded 1 exabyte of data, you would have to have started a video call **237,823 years ago**.



That's about the time modern homo sapiens emerged on the planet.

# Group Activity

Think about a dataset you are familiar with.  
Analyse that data set against the 5 V's.

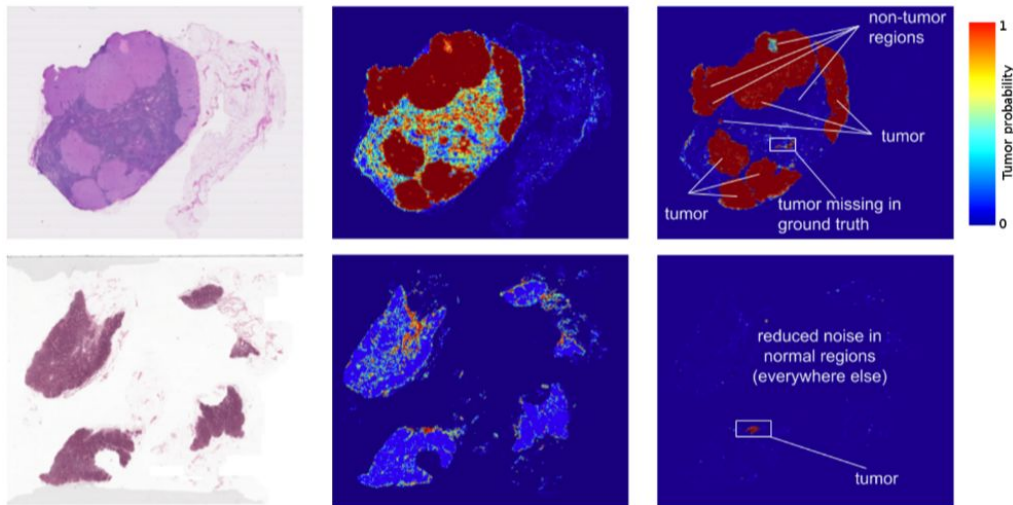
- 1) Volume
- 2) Velocity
- 3) Variety
- 4) Veracity
- 5) Value



# In the real world...

*How might machine learning help diagnoses in medicine?*

[Read more about the project](#)



Left: Images from two lymph node biopsies. Middle: earlier results of our deep learning tumor detection. Right: our current results. Notice the visibly reduced noise (potential false positives) between the two versions.

# Let's discuss NLP...

*In breakout rooms*

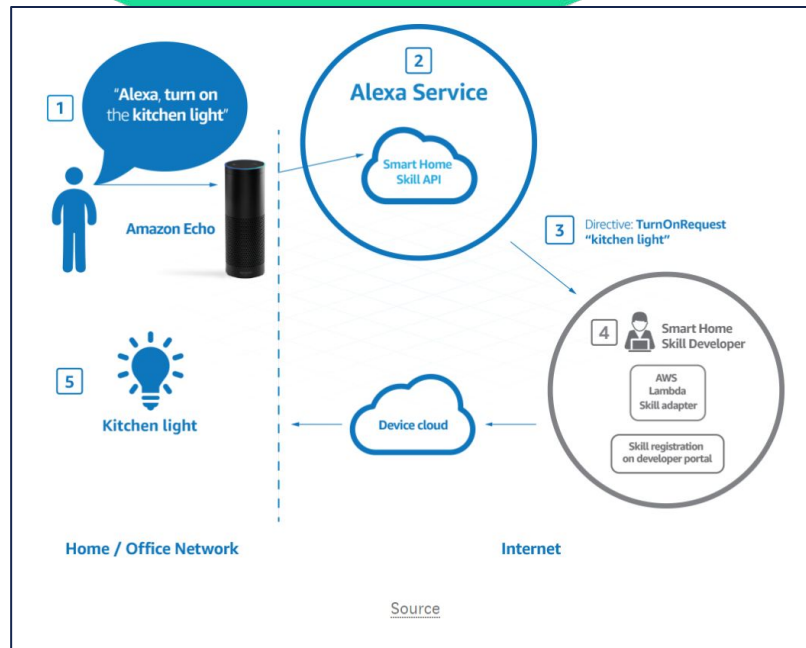
***How do you think Alexa produces a personalised recommendation for you (e.g. music, restaurant recommendations etc)?***

***Take 2 minutes to sketch the flow and share with your partner***



# Let's discuss NLP...

[Read more about this story](#)



# Your turn....

## **Assignment**

Investigate the use of big data in AI throughout your industry. Include at least one benefit and one concern in your analysis.

This will form part of your portfolio!

# Quick recap

 5 mins



# Sources

# Four Types

# Describe Big Data

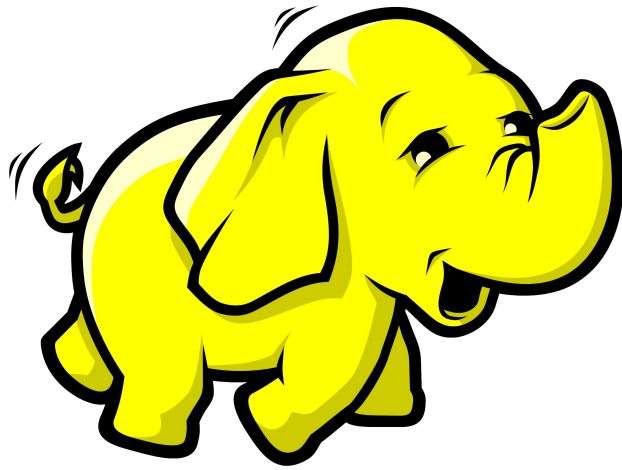
# Benefits

**5 Vs**



# Introducing Hadoop





**Hadoop is an open source software framework, maintained and development managed by the Apache Foundation, that is used to develop big data processing applications that execute in a distributed computing environment.**

**Why do we need  
Hadoop?**

**Dealing with Big  
Data is a  
challenge**

**1. Volume**

**2. Scale**

**3. Reliability**

# In the real world...

LinkedIn utilize Hadoop for the following purposes:

- Process failly production database transaction logs
- Examine the users' activities such as views and clicks
- Feed the extracted data back to the production systems
- Restructure the data to add to an analytical database
- Develop and test analytical model

**[Read more about how LinkedIn uses Hadoop](#)**



# 1. Volume

Hadoop includes a software implementation of **MapReduce**, this is small and lightweight in comparison to data.

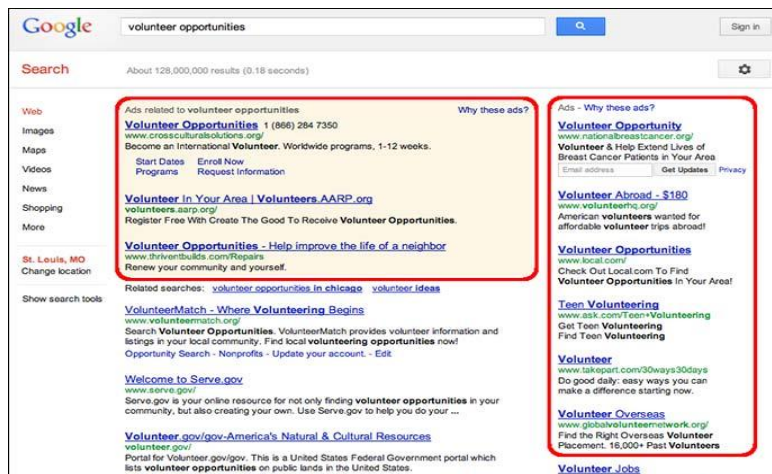
Now, don't be put off by this term. As Rajesh Kumar remarked, Map Reduce is a pretty well known concept in the field of computer science. It's just a fancy name that some PhD-types from Google published a paper on and popularized for a concept we already know quite well as **'divide and conquer'**.

# In the real world...

Prior to deploying Hadoop, it took 26 days to process three years' worth of log data. With Hadoop, the processing time was reduced to 20 minutes.

Hadoop allowed companies to start doing things like search index creation and maintenance, web page content optimization, web ad placement optimisation, spam filters, and ad-hoc analysis and analytic model development.

[Read more about this story](#)





# Activity

# Your turn...

1. We want to transfer 1TB files between 2 computers.
2. Assuming transfer speed is 100MB/s (Megabytes per second)
3. Your task is to calculate how long it will take to transfer between the two computers using
  - a. One intermediate 2TB hard-drive
  - b. 10 128GB USB Flash drive

**Assume:**

- both computers have 10 USB ports.
- Transfer into multiple USB drive simultaneously does not affect speed.

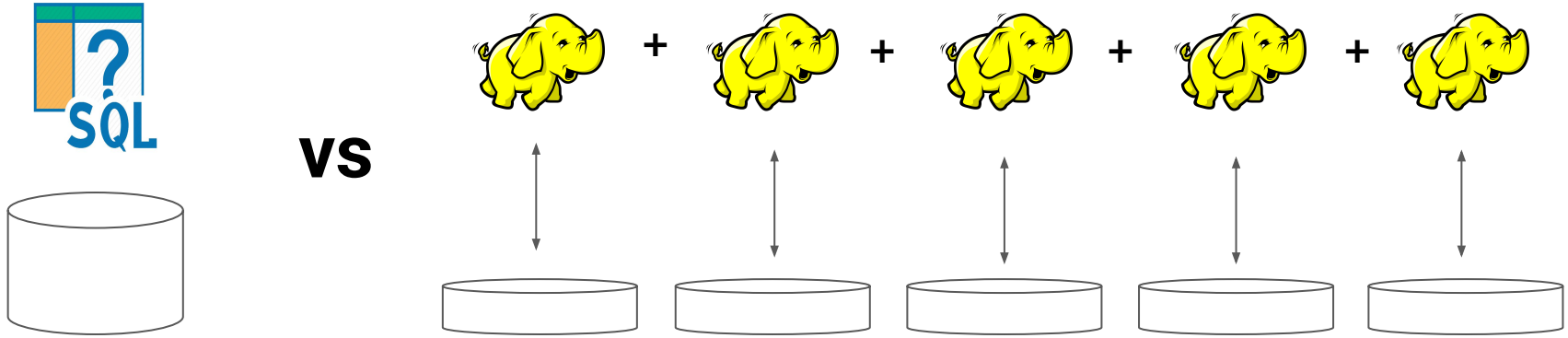
# Solution...

- **1 TB = 1 000 000 MB**
  - **It will take 1000000 Mb/100Mbps (10000s)**
  - **Equivalent to approximately 3 hours.**
  - **Both computers plus downtime (6 hours)**
- 
- **128 GB = 128000 MB**
  - **It will take 128000 Mb/100Mbps (1280s)**
  - **Equivalent to approximately 22 minutes**
  - **Both computers plus downtime (1 hour Max)**

**1.**

**2. Scale**

**3.**



- The Hadoop Distributed File system (HDFS) is a file system that provides the capability to distribute data across a cluster to take advantage of the parallel processing of MapReduce.
- Our dataset is broken down into small blocks and distributed over a cluster.
- This allows us to run large queries concurrently across multiple machines, before consolidating the results at the end.

# Activity

# Your turn...

**You want to store 1 Petabyte of Data. How would you go about storing it?**

- a. Do you expand the memory of your current machine?**
- b. Do you store in multiple machines?**
- c. What is/are the benefits and drawbacks of each method?**


1.


2.

3. **Reliability**



Block A: 

Block B: 

Block C: 

### Rack A

1.



2.



3.



### Rack B

5.



6.



7.



### Rack C

8.



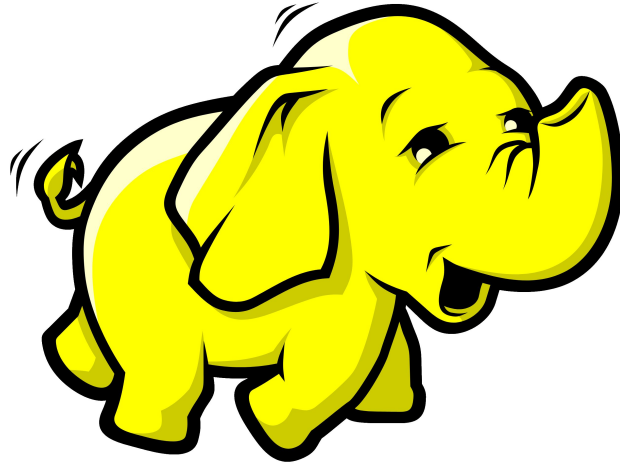
9.



10.

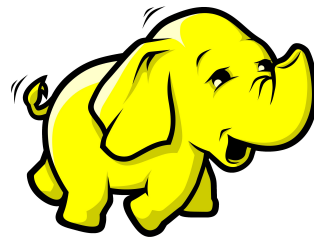


# Features of Hadoop



- **Open Source** - modifiable to business requirements
- **Distributed Computing** - used to resolve data locality issue
- **Faster than RBMS** - engineered for many readers, few writers
- **High Availability & Fault Tolerant** - data is replicated at least 3 times
- **Data Resilience** - multiple nodes can service requests if hardware fails
- **Scalability** - resources can be added with disruption to execution
- **Economic / Cost Effective** - designed to run on cheap commodity hardware

# Hadoop Architecture



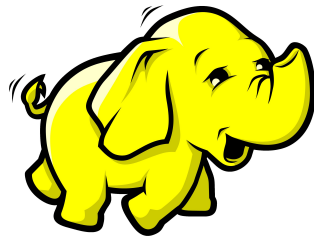
MapReduce

YARN

HDFS

Common  
Utilities

# Hadoop Architecture



MapReduce

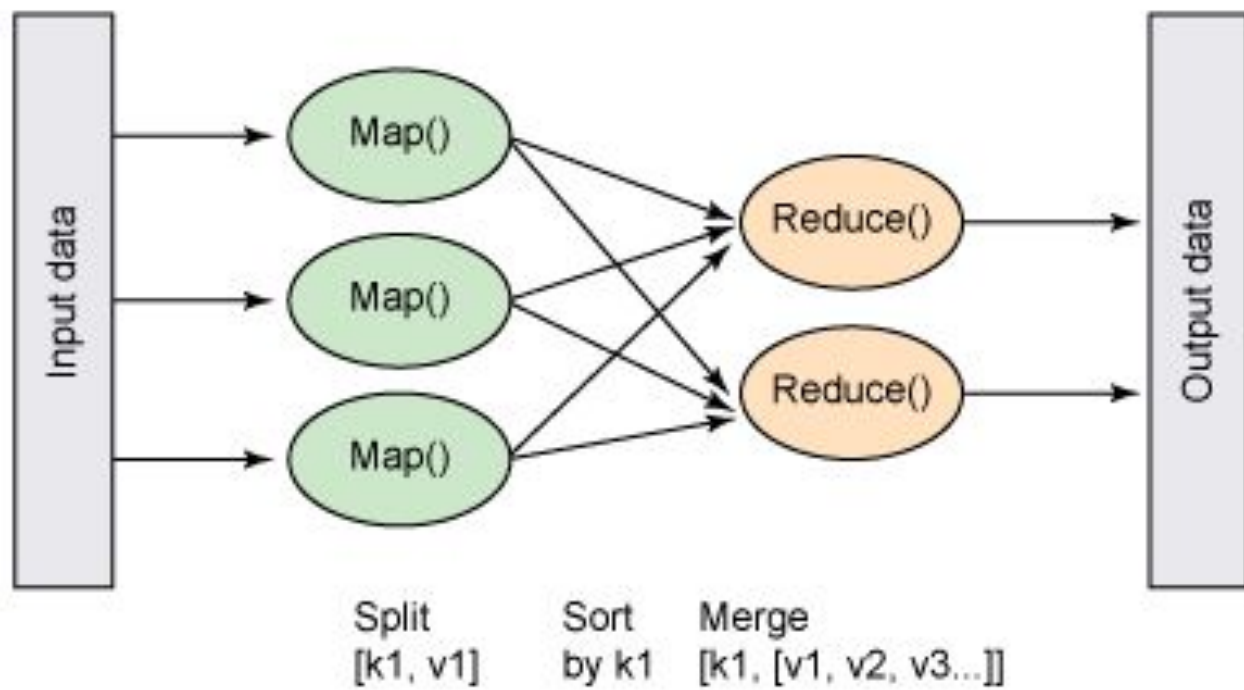
**Processing Component**

YARN

**Framework for Job Scheduling and cluster resource management (Yet Another Resource Negotiator)**

HDFS

**Data Storage framework on Hadoop (Hadoop Distributed File System)**



# Quick recap

 10 mins



# Challenges



# Features

# Fault Tolerant

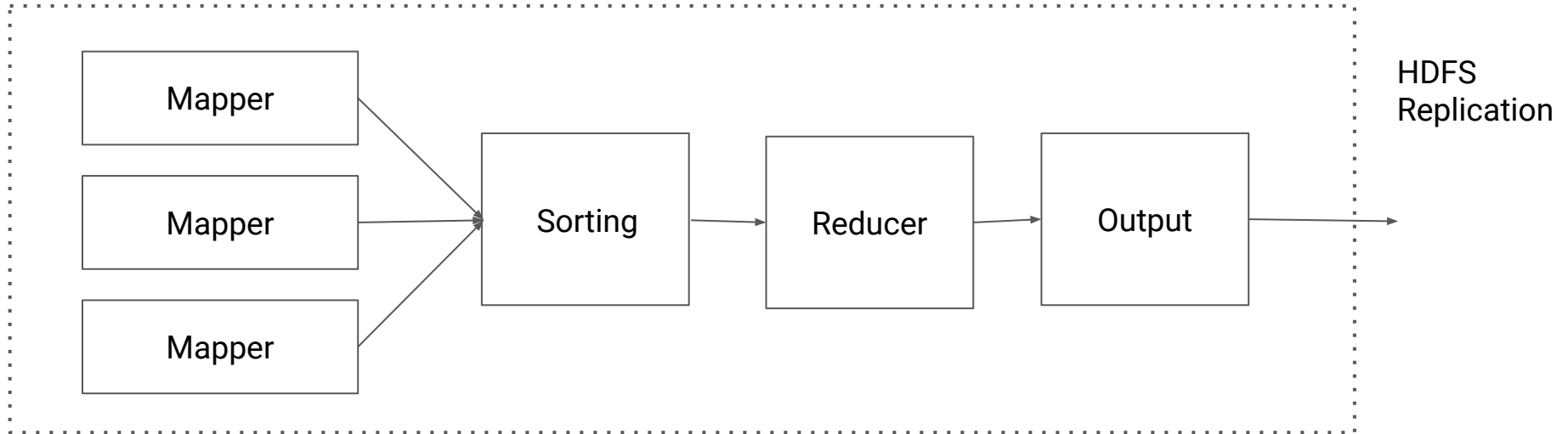
# MapReduce

# HDFS

# Hadoop's MapReduce

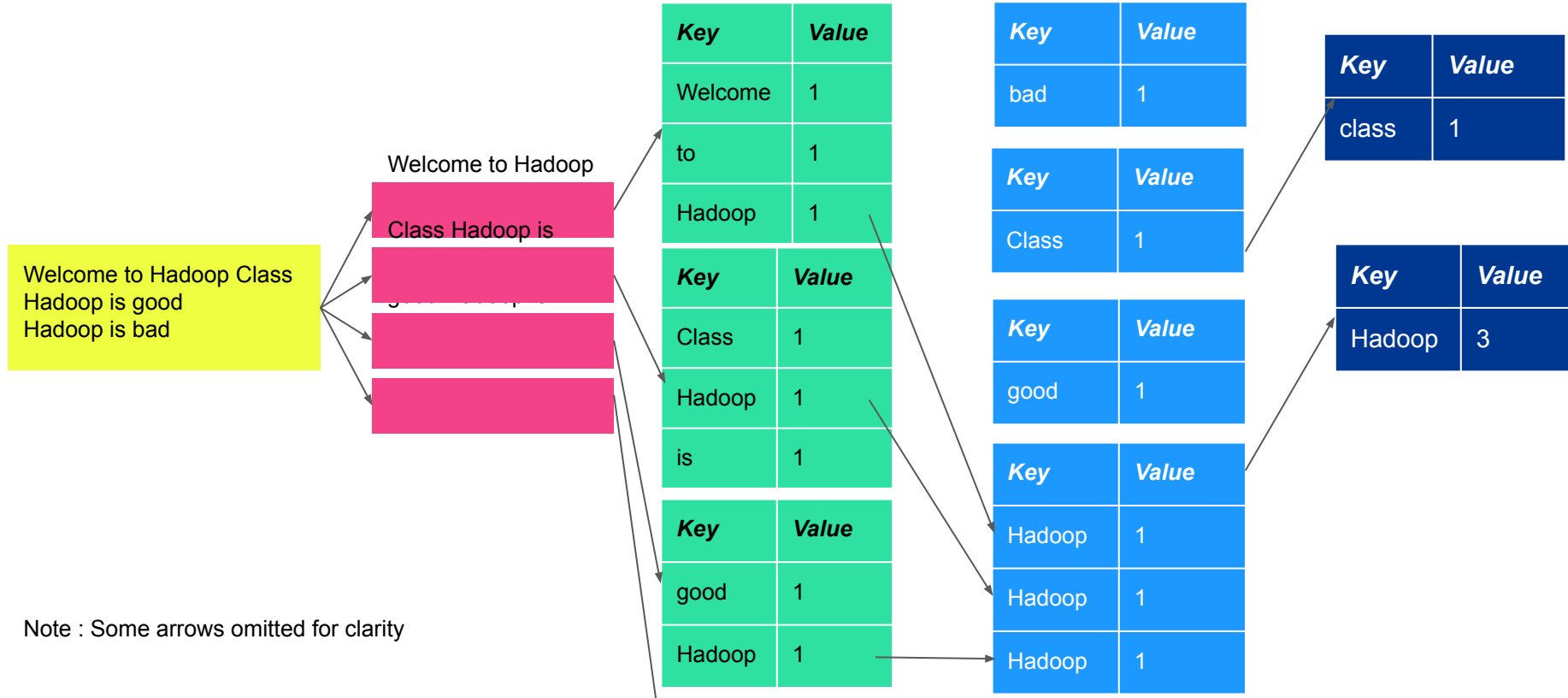


# Hadoop Execution Flow



# MapReduce Phases

| Input | InputSplit | Mapping | Shuffling | Reducing |
|-------|------------|---------|-----------|----------|
|-------|------------|---------|-----------|----------|



# Activity



# MapReduce Simulation

- **The task is the count the occurrence of the words 'THE' and 'IN' in the given document (its an image so no CTRL+F)**
- **You will be assigned to one of two groups**
  - **Group 1: Get one person to carry out the task alone while others watch.**
  - **Group 2: Split the document up (InputSplit) and get every member to count from their portion (Mapping). Then combine the counts (Reducing).**

# Quick recap

 10 mins



# Theory

# Hadoop vs RDMS

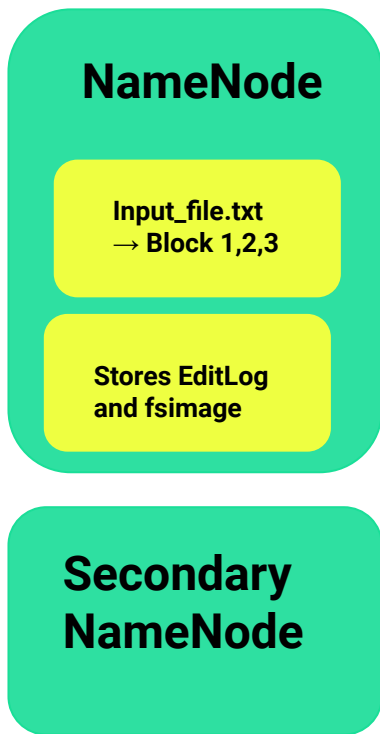
# Components

# Hadoop's HDFS

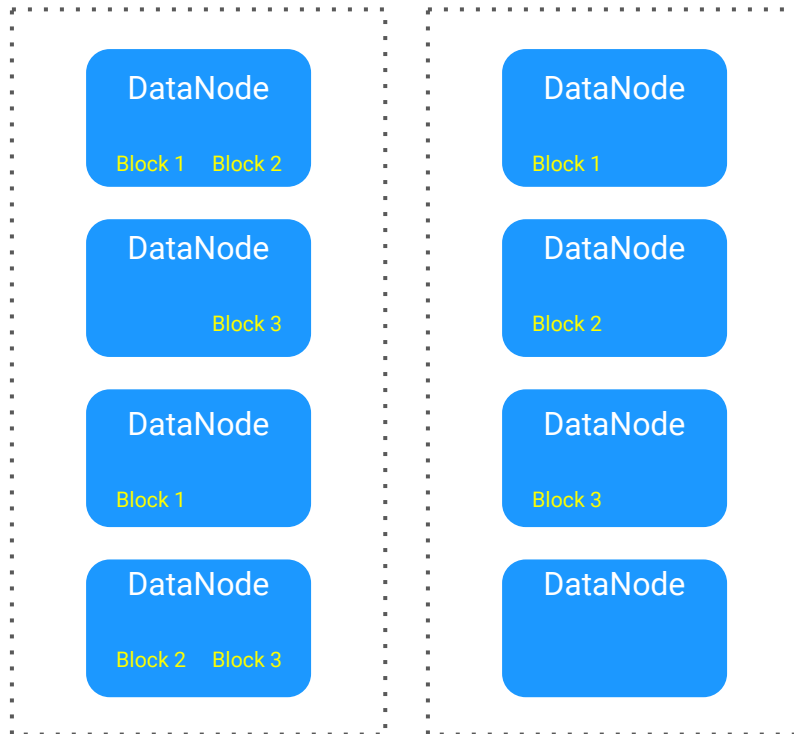


**What do we know  
so far?**

## Master Nodes

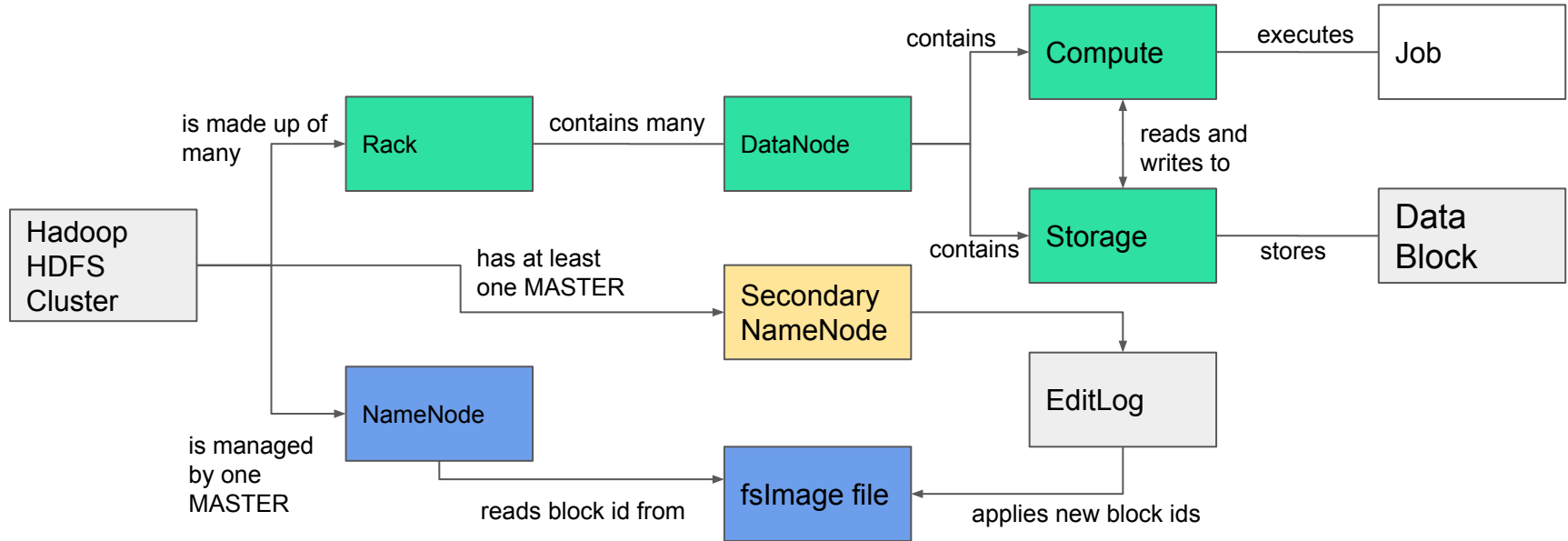


## 8 Worker Nodes across 2 Racks





# HDFS Mindmap of Concepts



# Activity

# Distribute blocks in HDFS

- You have data that has been splitted into 5 blocks
- You have 2 racks with 4 datanodes each.
- Replication factor is 3
- You need to distribute blocks into the nodes.
- A block must have presence in the 2 racks
- A datanode can only occupy two blocks
- No two datanodes in the same rack would have the same blocks

# Two Main Modes

# Rack Aware

# Secondary NameNode

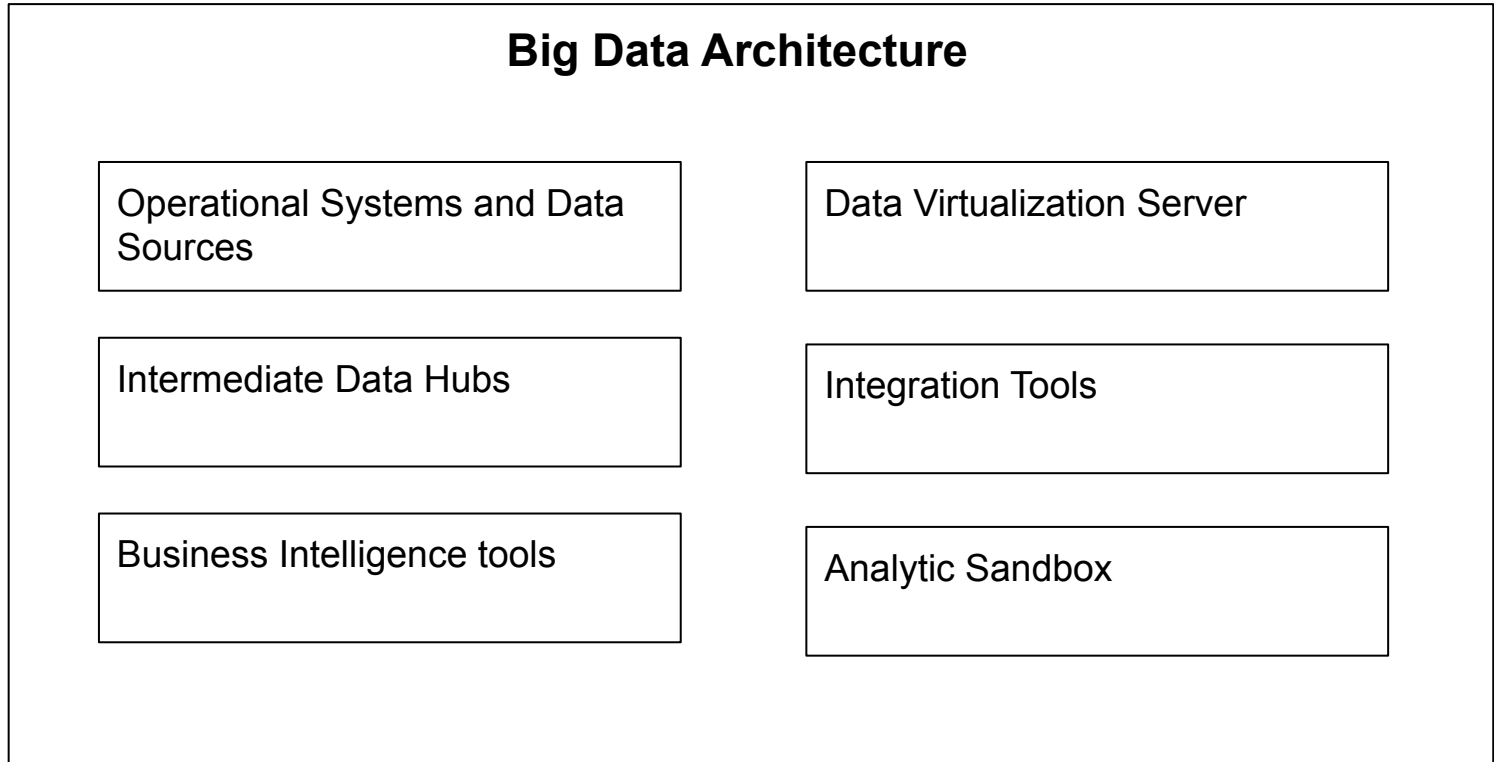
# Discussion

# Big Data Architecture





# Big Data Architecture



# Operational Systems and Data Sources

- **Unstructured data files**
- **Structured data files from relational databases**
- **Sources both internal and external to the organization**

# Intermediate Data Hubs

- Existing Data Warehouses, data marts and operational data stores.
- Usually storing structured data
- Cloud-based solutions for temporary persistence

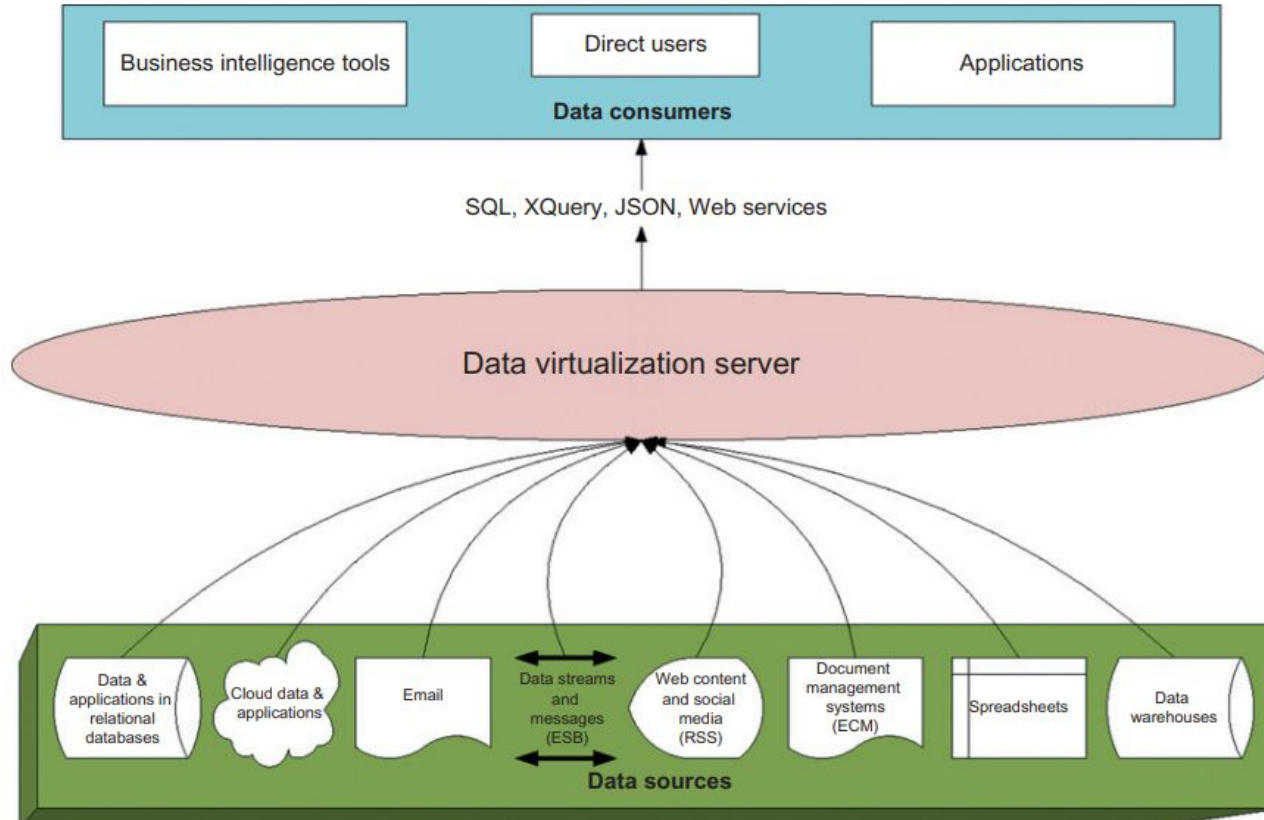
# Business Intelligence Tools

- **Structured business intelligence**
- **Hadoop MapReduce business intelligence**
- **Visualization**

# Data Virtualization

- Central and critical to in big data architecture
- Allows data to be presented in real time, from different sources and in the required format.
- Build on Data Warehouse by combining historical data with current data in real time
- Implemented in form of Data Virtualization Servers

# Data Virtualization



# Integration Tools

- **Batch data integration tools (ETL)**
- **Real-time data integration tools**

# Analytic Sandbox

- **An area for analysis of the available data sources**
- **Reporting and modelling data for decision making**



# Quick recap

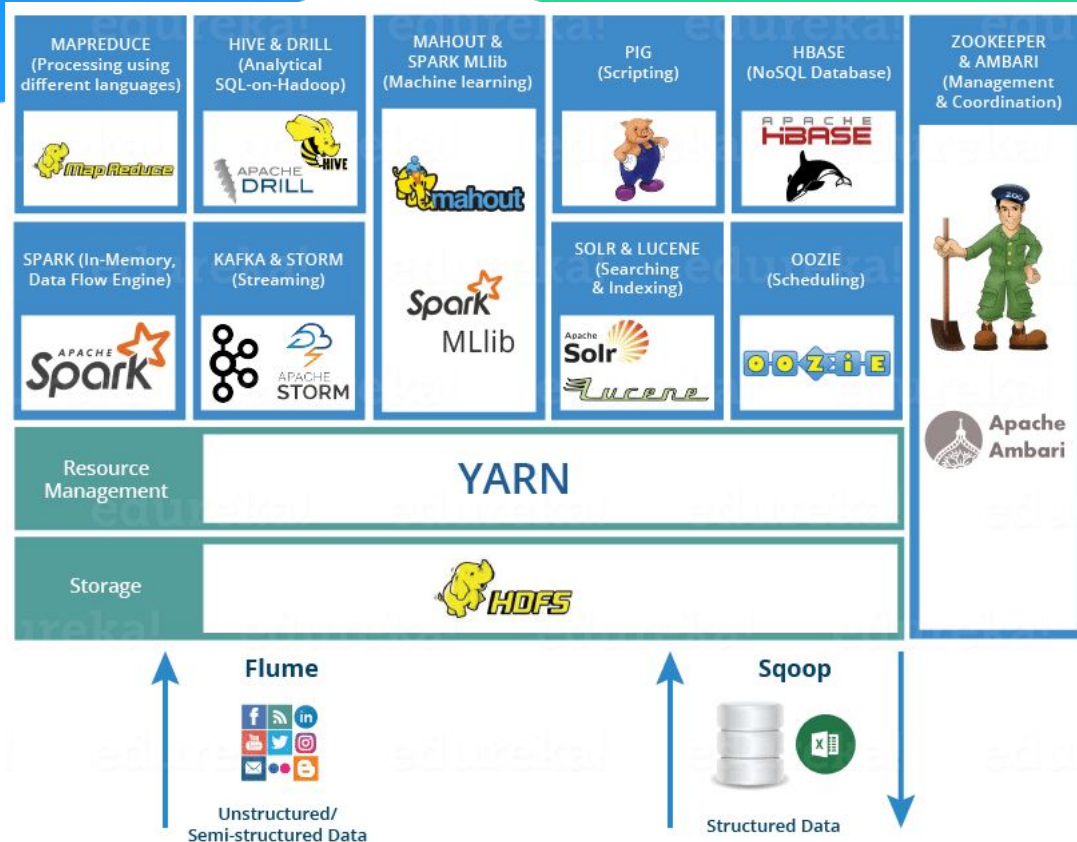
 10 mins



# The Hadoop Ecosystem



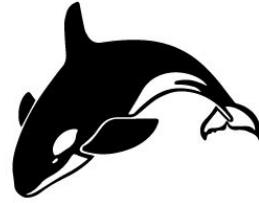
# The Hadoop Ecosystem Easy Snapshot



EXAM  
TOPIC



APACHE  
**PIG**



A P A C H E  
**HBASE**



**HIVE**

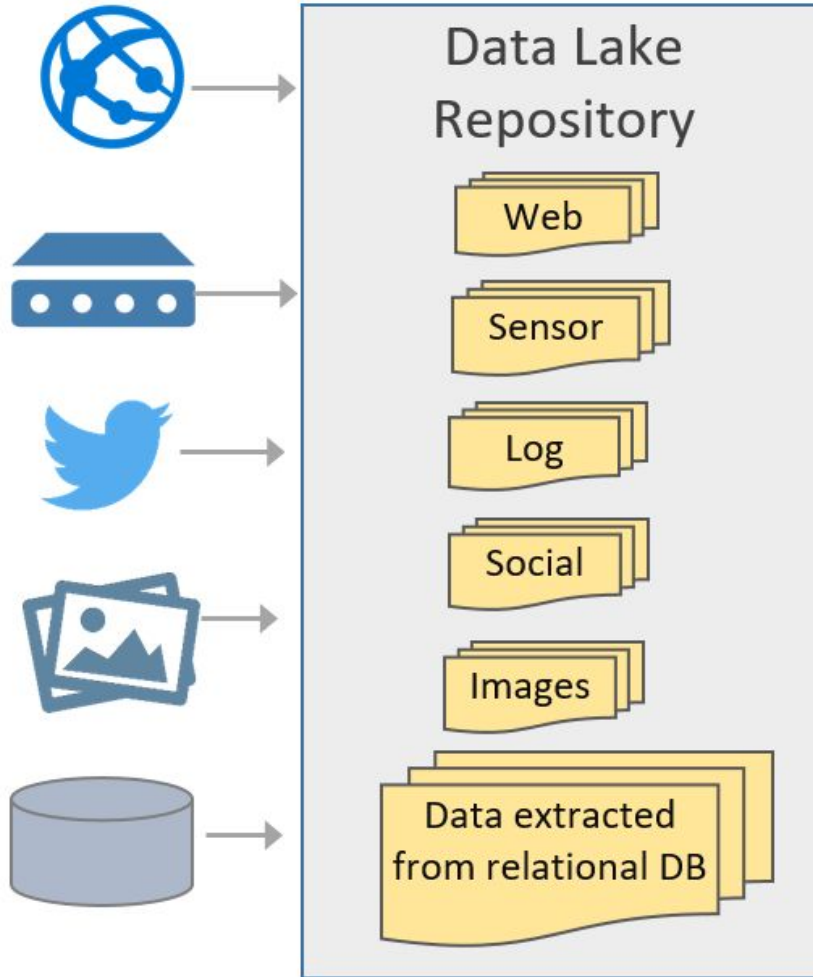


**mahout**

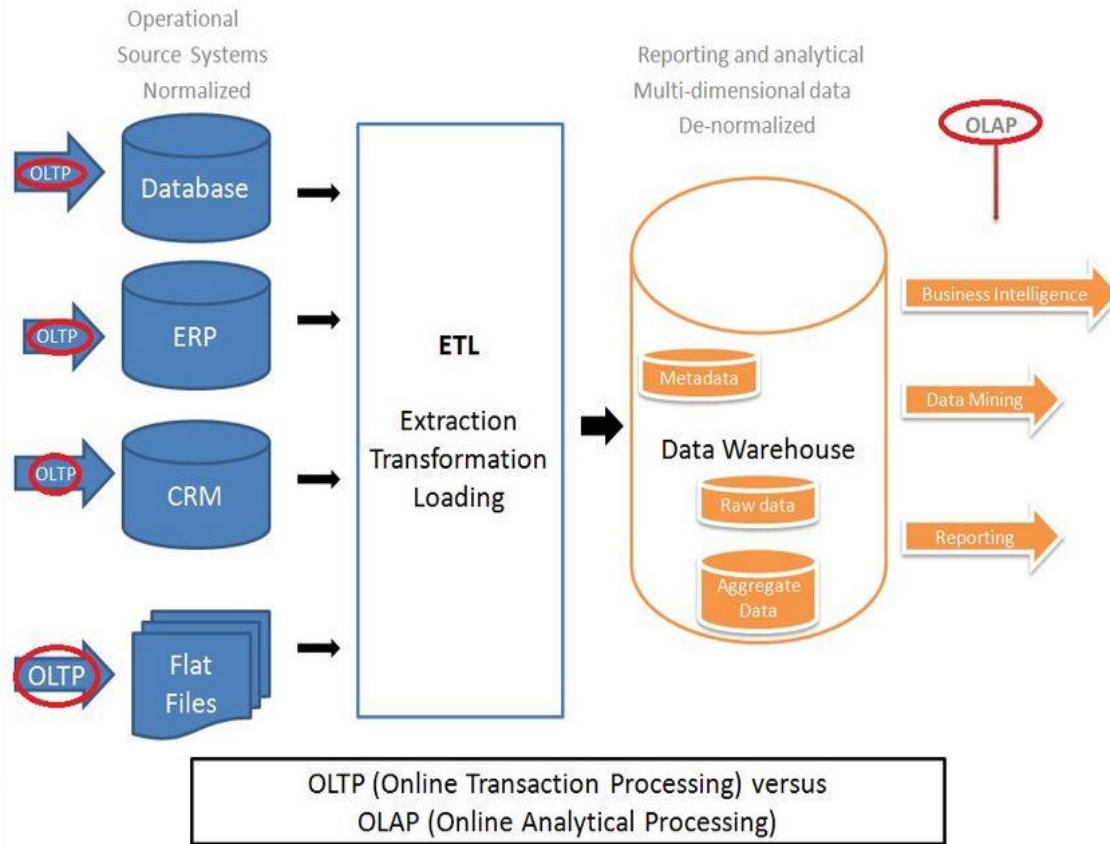
# Activity

# **Data Lakes vs. Data Warehouse**





- Storage for both structured and unstructured data from various data sources
- Cost-effective big data storage
- Storing data and big data analytics, like deep learning and real-time analytics
- Stores all data that might be used—can take up petabytes!
- No transformation needed!



- Historical data that has been structured
- Analytics for business decision
- Typically read-only queries for aggregating and summarizing data
- Only stores data relevant to analysis
- ETL process required!



# Quick recap

 5 mins



# Data Staging

# EcoSystem

**Four**

**Explain**

# Lake vs Warehouse

# Assignment

**Investigate the use of big data in AI and ML throughout your industry. I would recommend including:**

- **Potential problems for your industry**
- **Potential benefits for your industry**
- **Your opinion on whether AI/ML will help/hinder your sector**

**Those who wish to show “initiative” may decide to delve deeper into this...**

**Due Date: 30th of June 2020**

# Recap



# Pop Quiz Time



**Data Fellowship**  
**Module 3 THE END**

