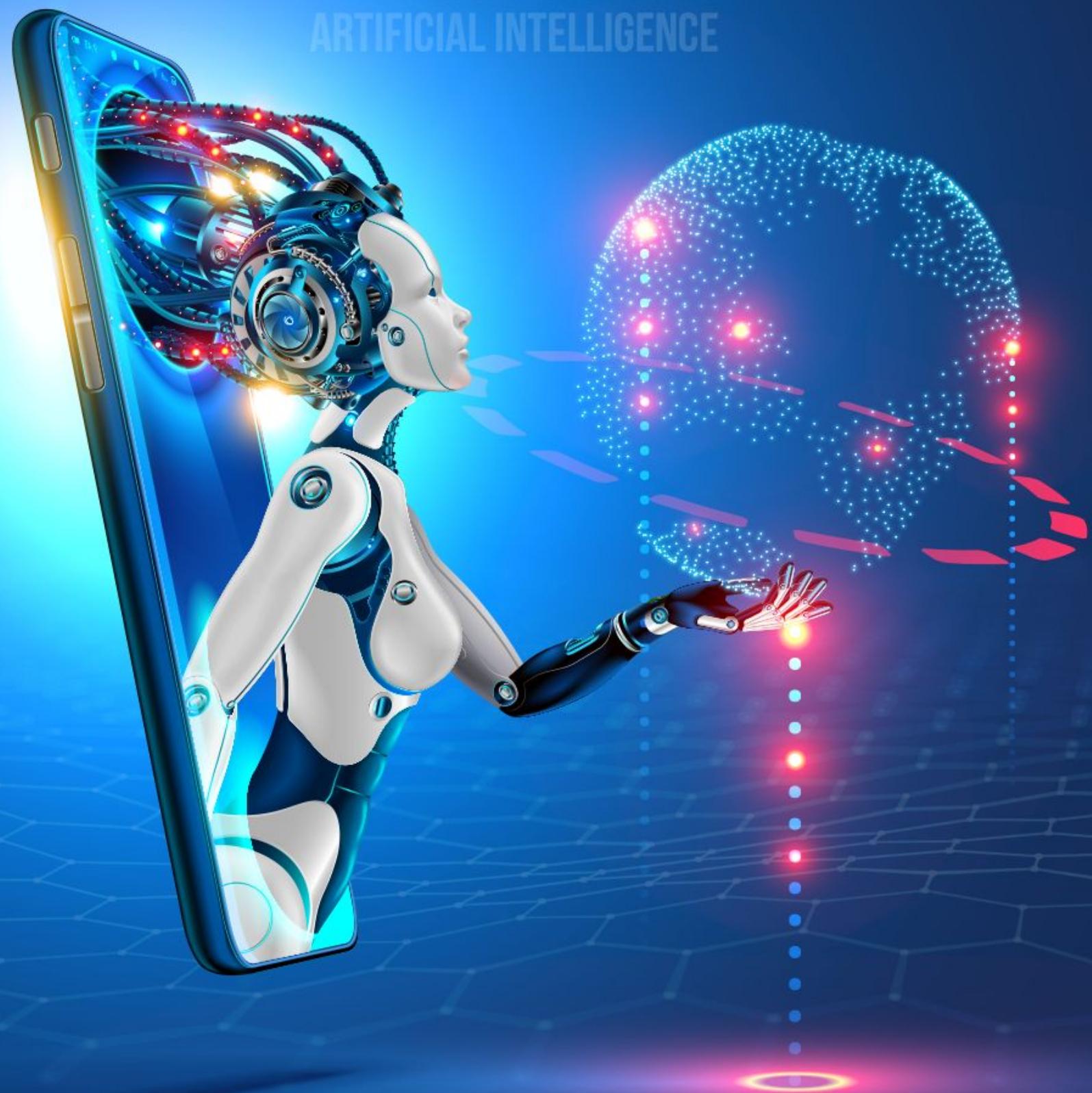


**DATA AND  
ARTIFICIAL INTELLIGENCE**



**Big Data Hadoop and Spark Developer**

# DATA AND ARTIFICIAL INTELLIGENCE



## Introduction to Big Data and Hadoop

# **Learning Objectives**

By the end of this lesson, you will be able to:

- ✓ Describe the concepts of Big Data
- ✓ Explain Hadoop and how it addresses Big Data challenges
- ✓ Describe the components of Hadoop Ecosystem



## **Introduction to Big Data**

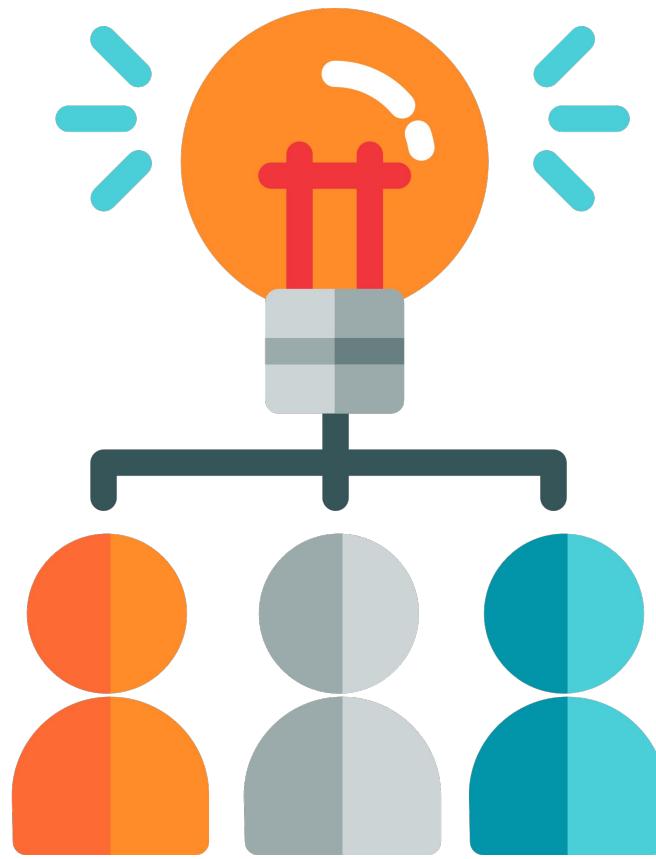
# Big Data Overview

Big Data is the data that has high volume, variety, velocity, veracity, and value.

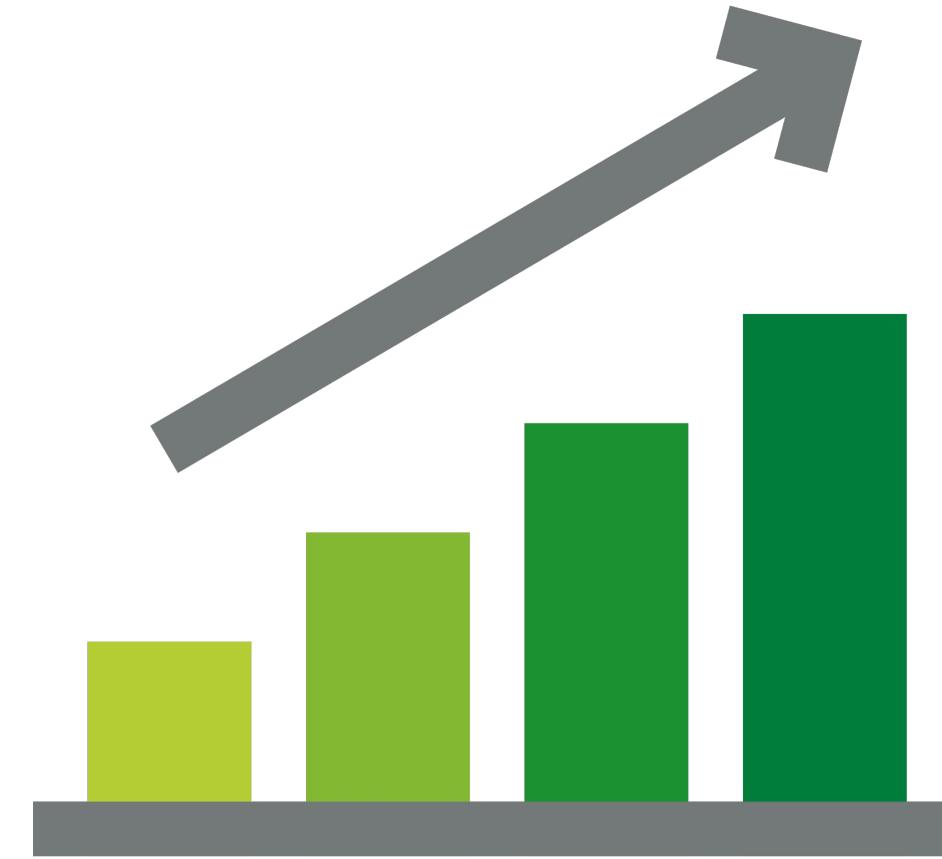


According to US Bureau of Labour Statistics, Big Data alone will fetch 11.5 million jobs by 2026.

# Traditional Decision-Making



**What We Think**



**Experience and Intuition**



**Rule of Thumb**

# Challenges of Traditional Decision-Making

Takes a long time to arrive at a decision, therefore losing the competitive advantage



Requires human intervention at various stages

Lacks systematic linkage among strategy, planning, execution, and reporting



Provides limited scope of data analytics, that is, it provides only a bird's eye view

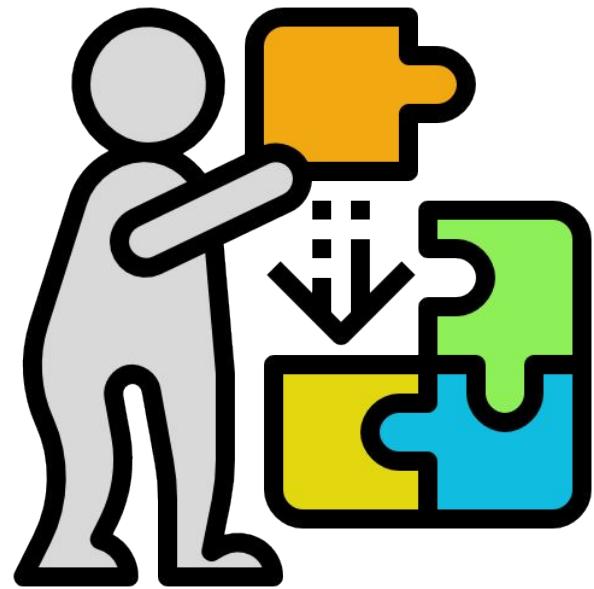
Obstructs company's ability to make fully informed decisions



# DATA AND ARTIFICIAL INTELLIGENCE

## Big Data Analytics

# The Solution: Big Data Analytics



Solution

The decision-making is based on what you know which in turn is based on data analytics.

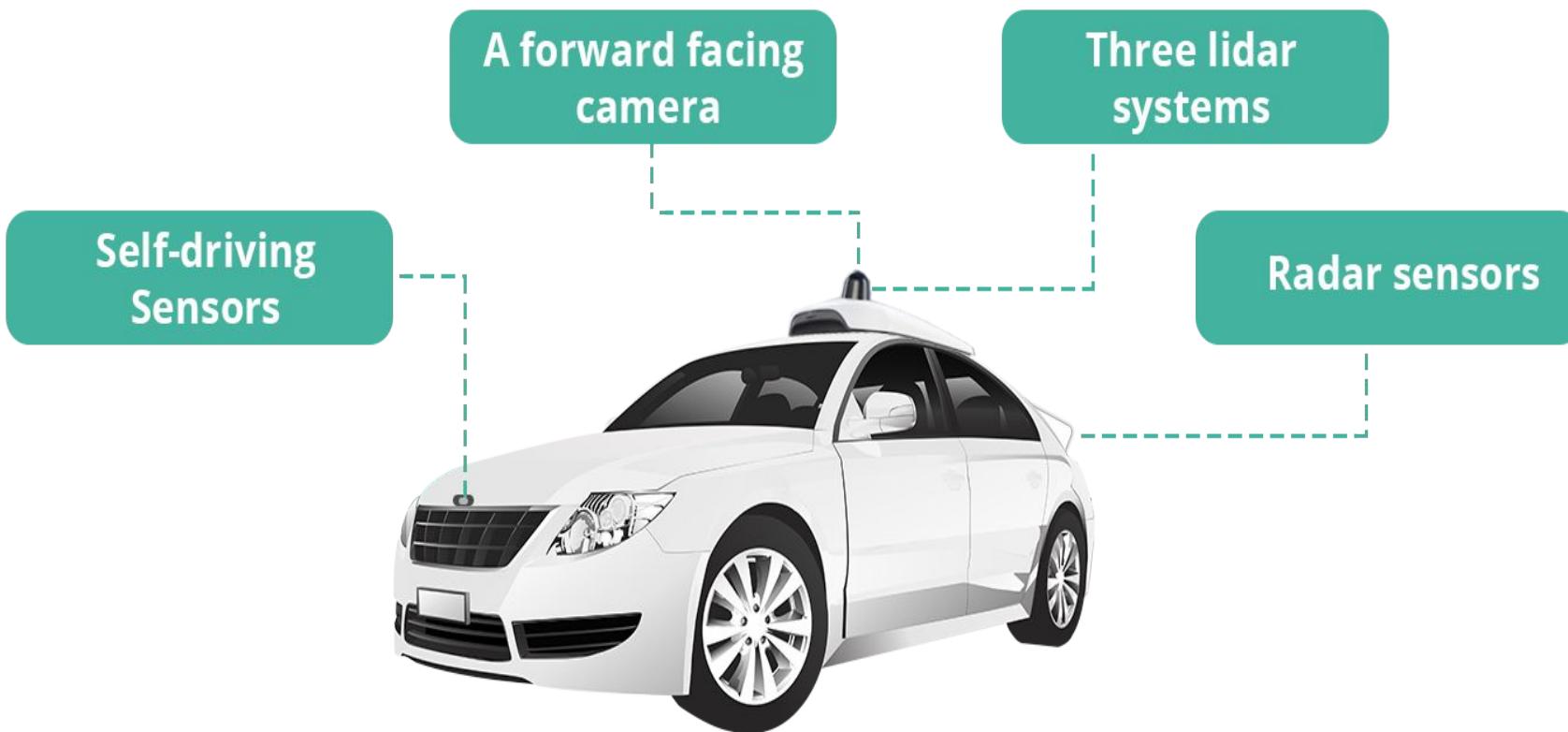
It provides a comprehensive view of the overall picture which is a result of analyzing data from various sources.

It provides streamlined decision-making from top to bottom.

Big data analytics helps in analyzing unstructured data.

It helps in faster decision-making thus improving the competitive advantage and saving time and energy.

# Case Study: Google's Self-Driving Car



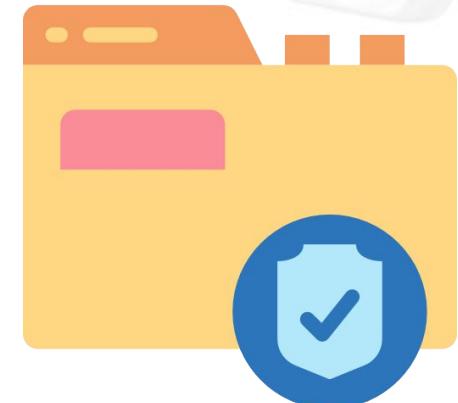
Technical Data



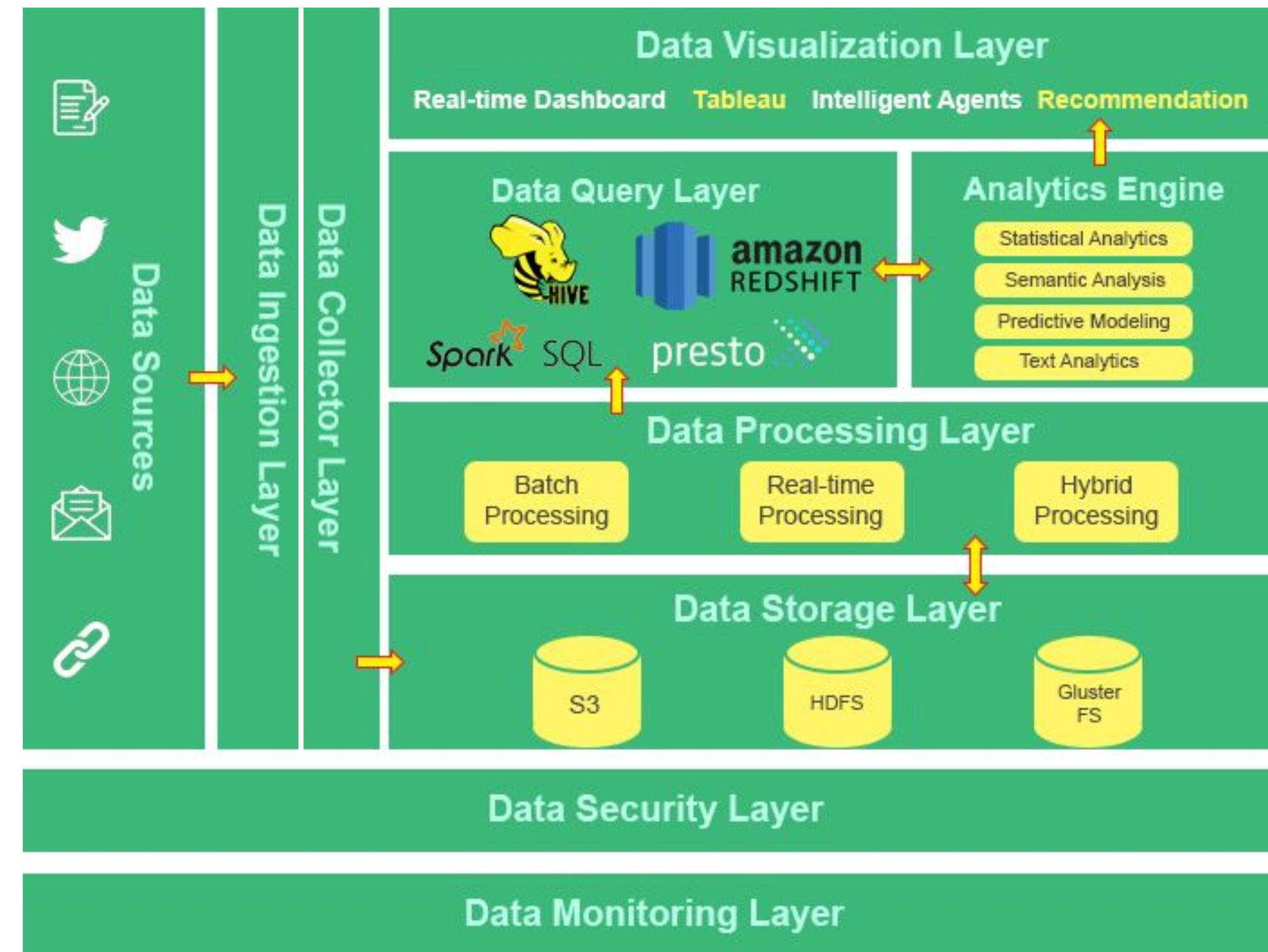
Community Data



Personal Data



# Big Data Analytics Pipeline



## **What Is Big Data?**

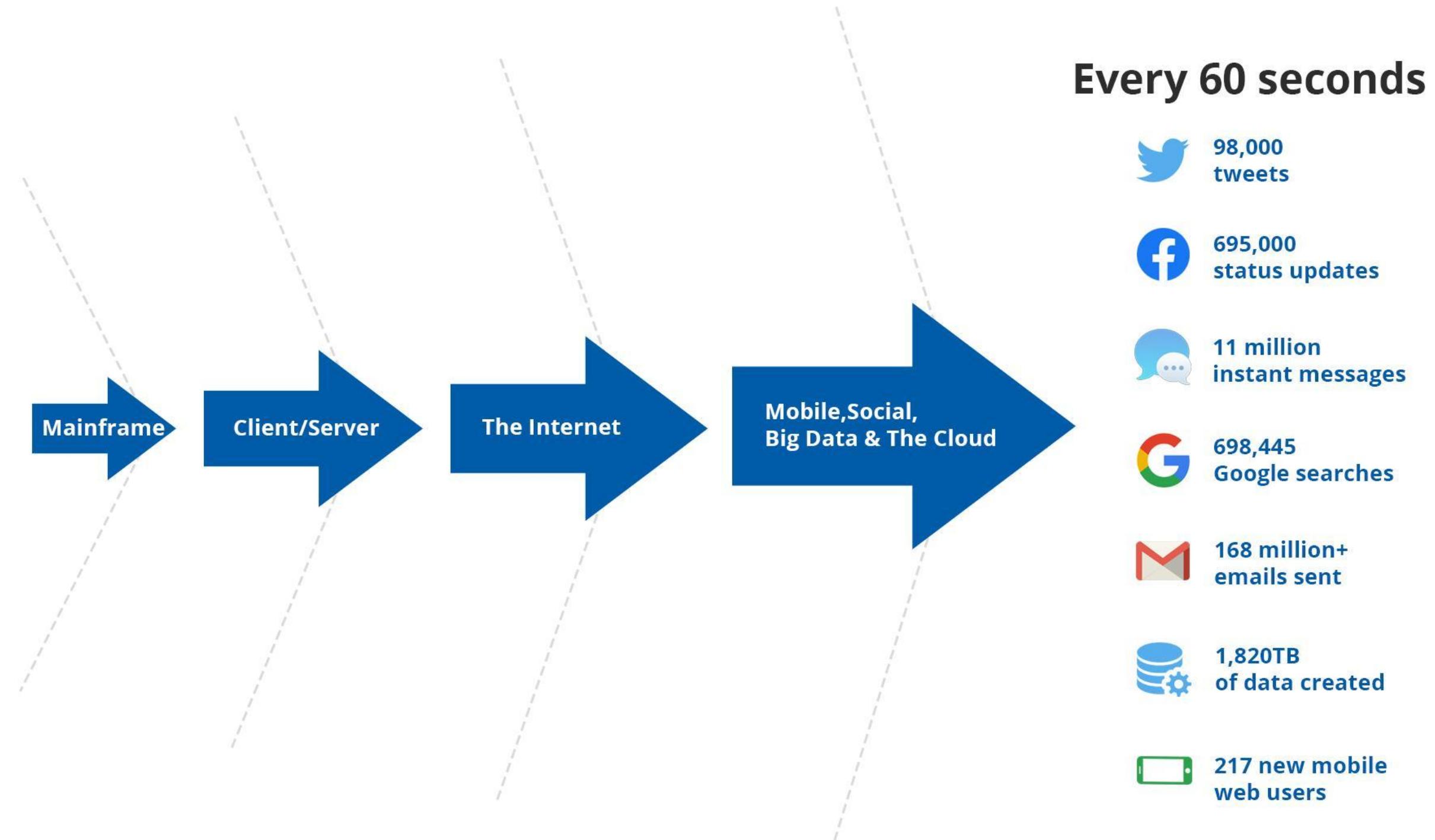
# What Is Big Data?

“

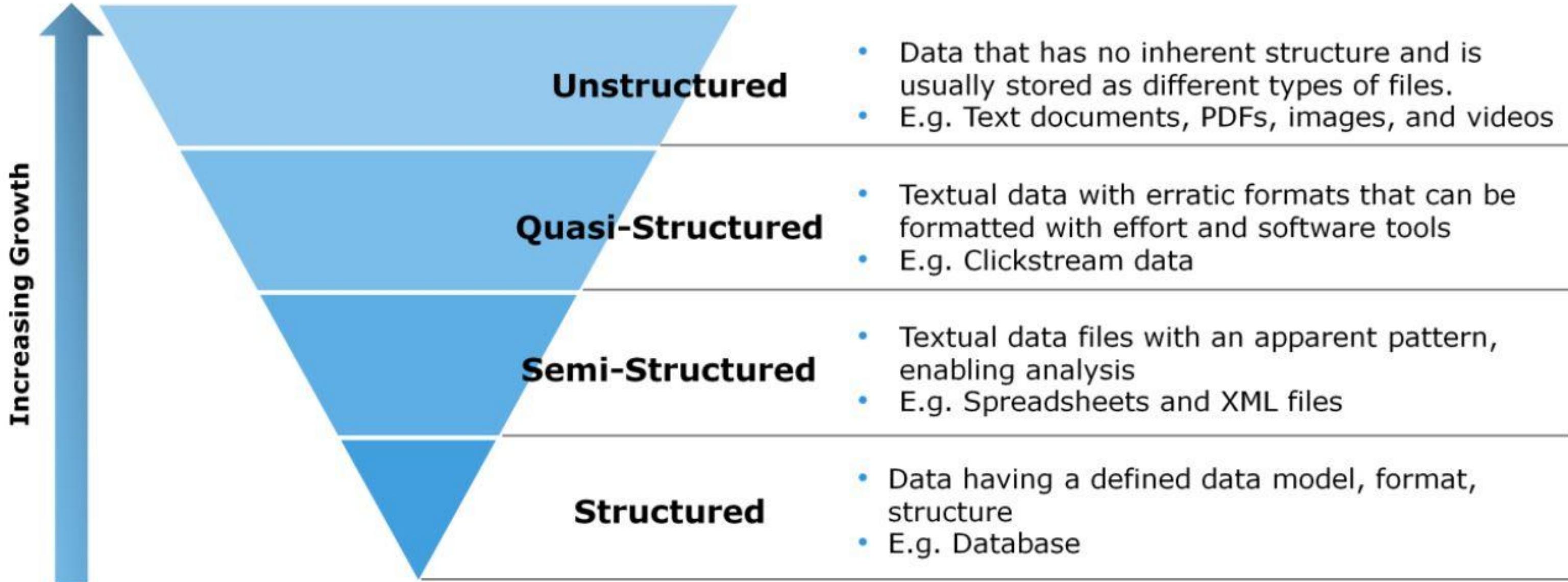
Big data refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

”

# Big Data at a Glance

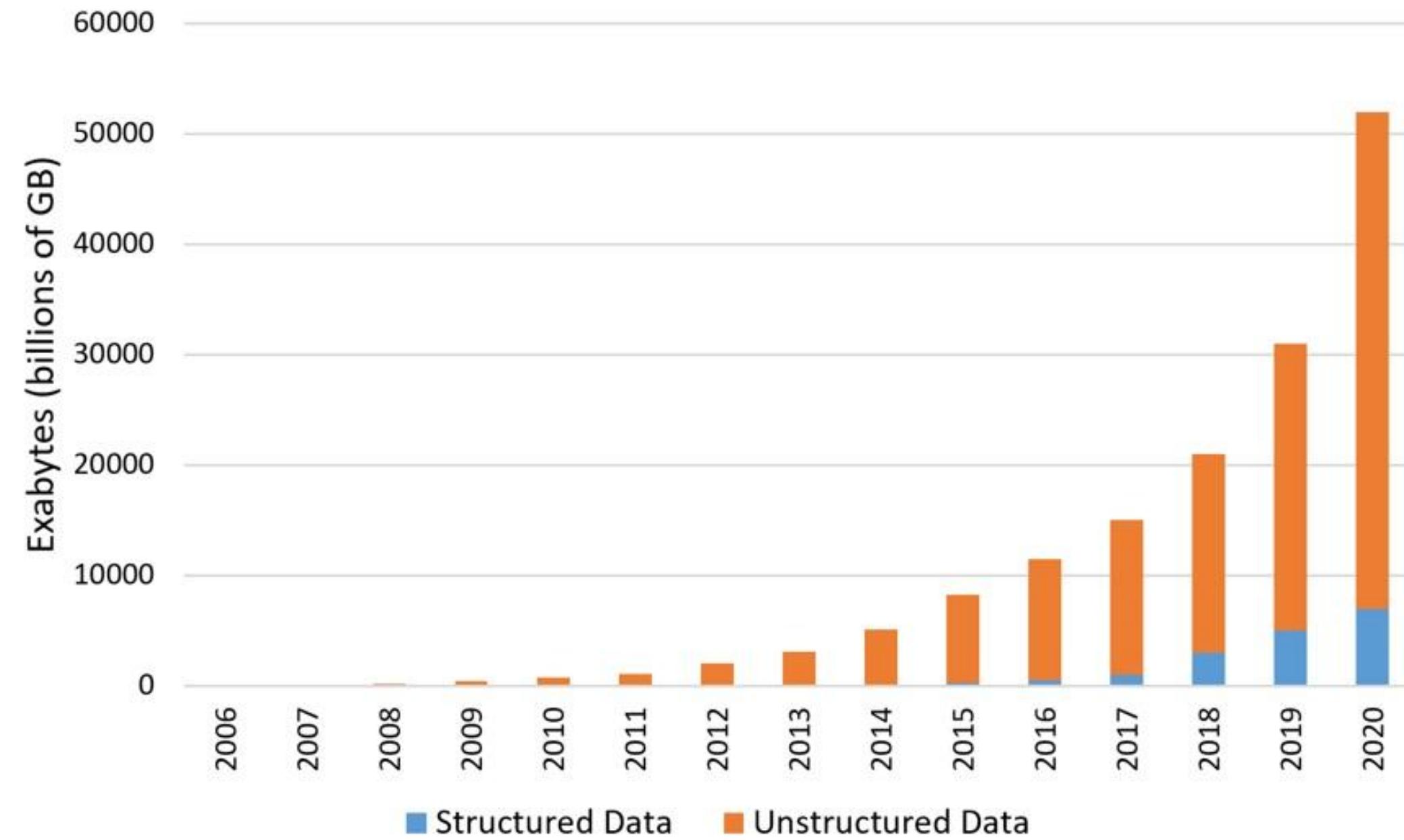


# Different Types of Data



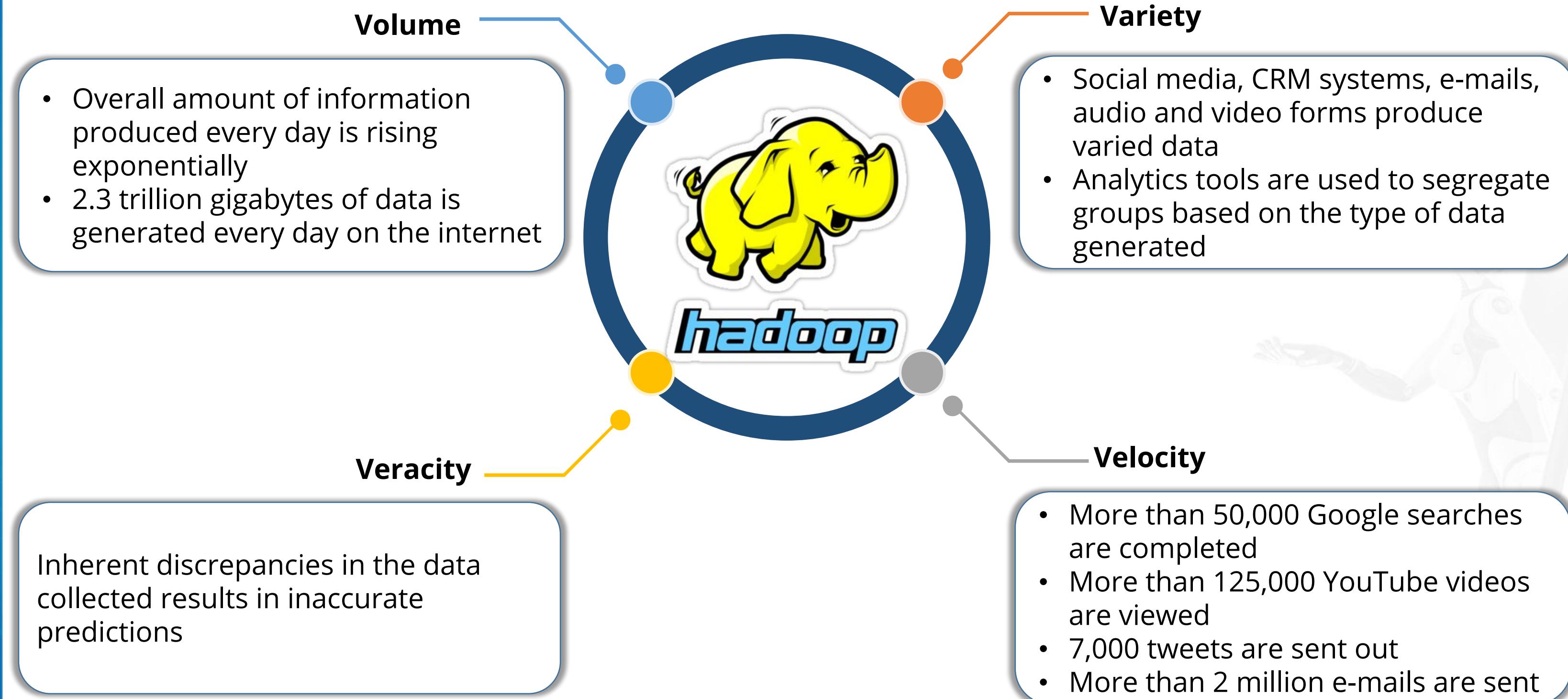
# Growth in Data

By 2020, data will show an exponential rise!



## **Four Vs of Big Data**

# Four Vs of Big Data



# Unstructured Data Conundrum

Unstructured Data



Web Logs



Social Media



Multimedia

Semi-structured Data

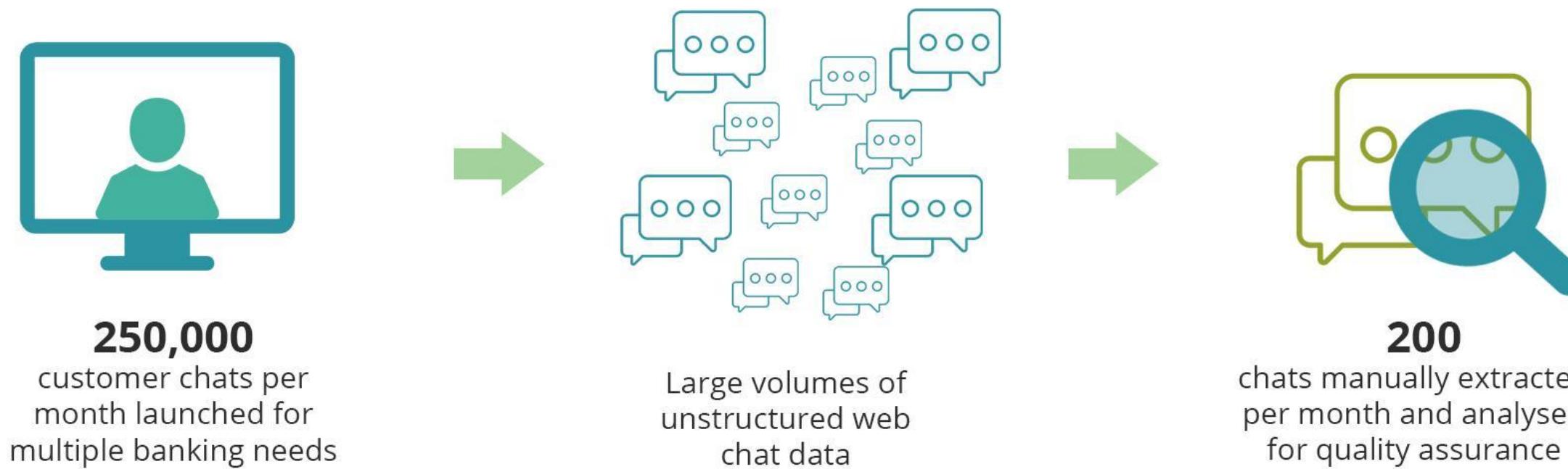


# DATA AND ARTIFICIAL INTELLIGENCE

## Case Study: Royal Bank of Scotland

# Case Study: Royal Bank of Scotland

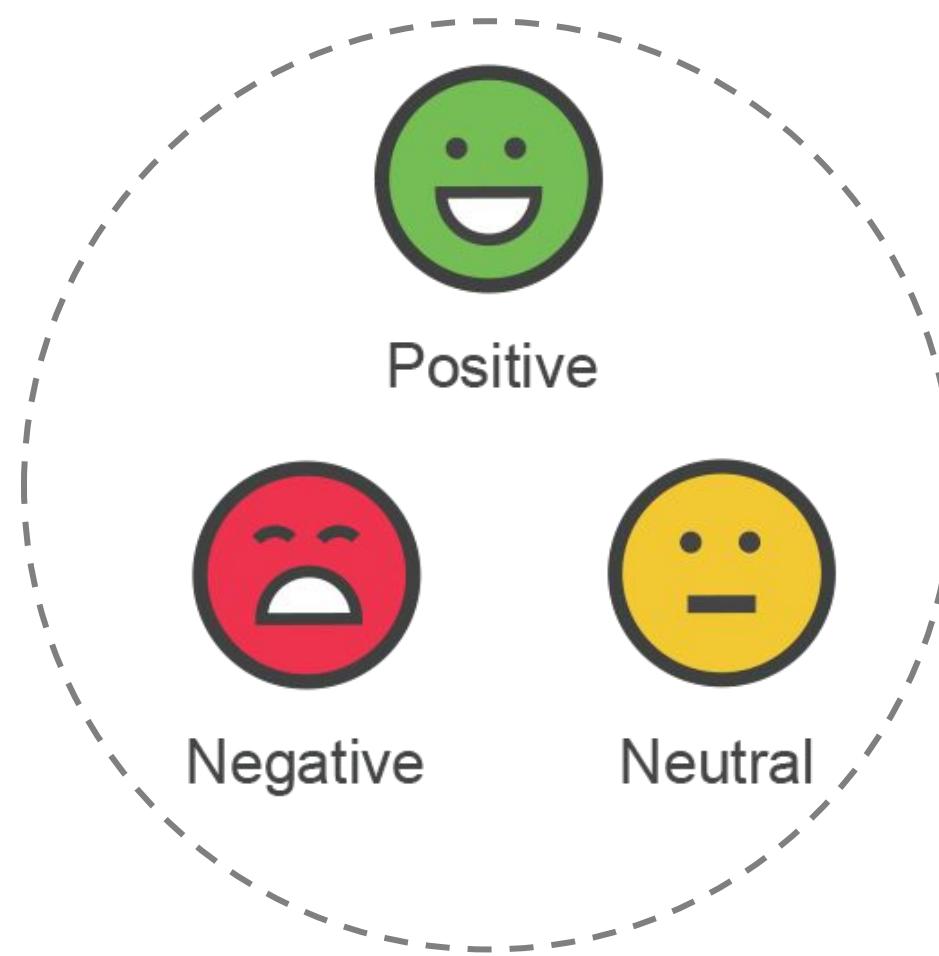
## Previous Web Chat Analysis Approach



100% of this data could be processed whereas only 3% could be processed earlier with traditional systems.

# Case Study: Royal Bank of Scotland

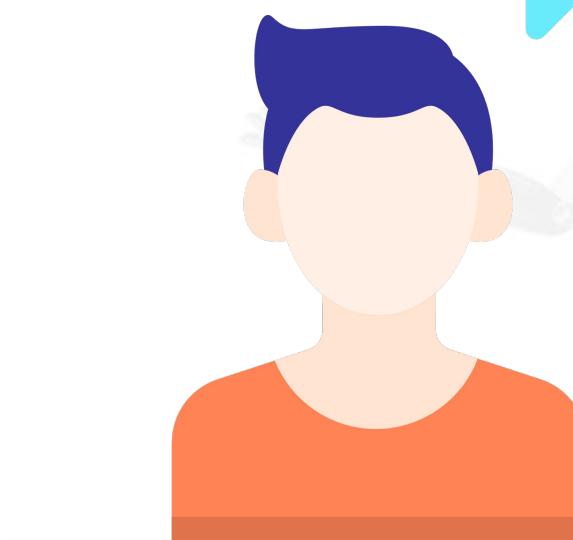
The case study of Royal Bank of Scotland gave the following three things:



Sentiment analysis



Reduced processing time



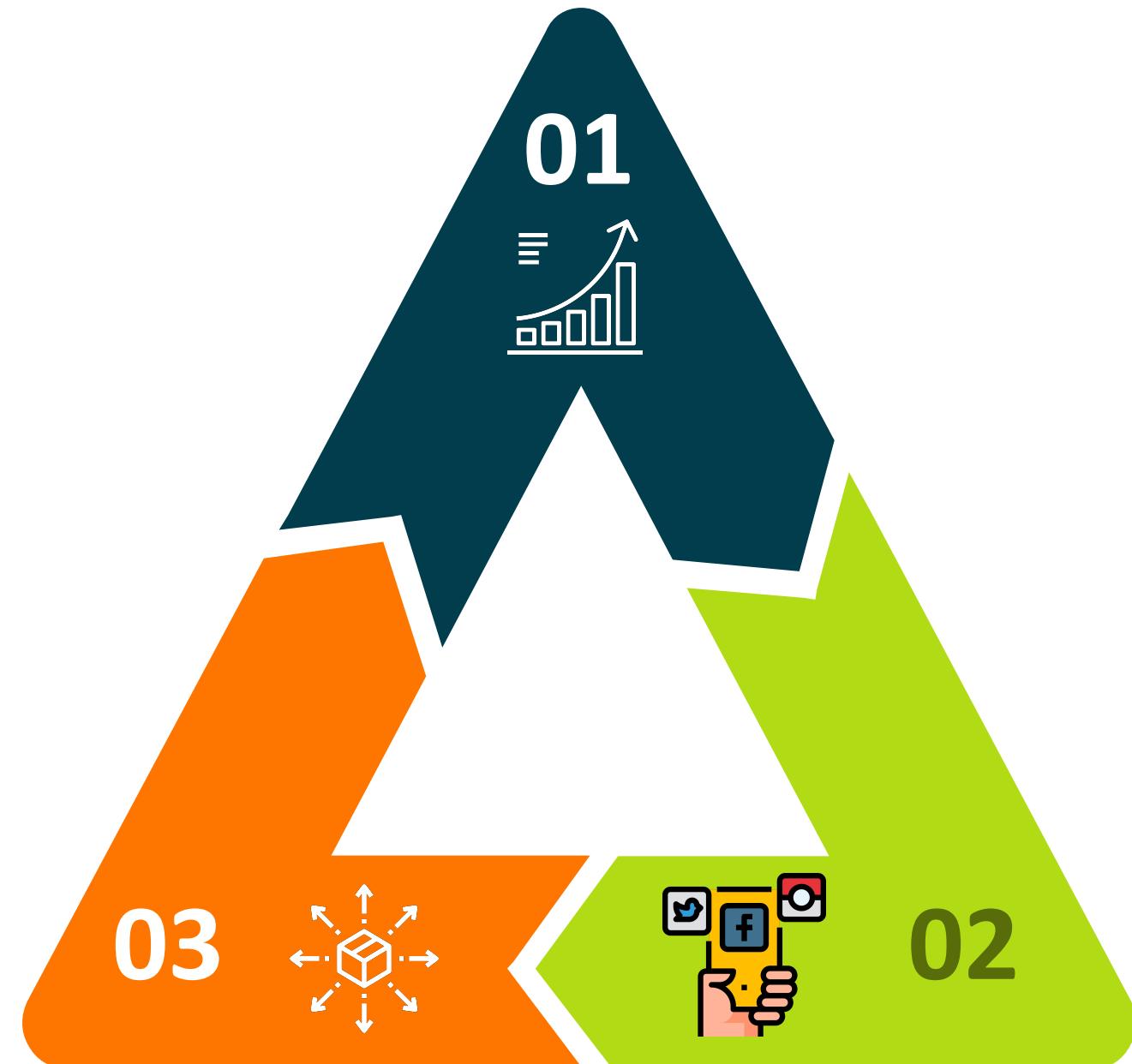
Improved customer satisfaction

## **Challenges of Traditional System**

# Challenges of Traditional Systems (RDBMS and DWH)

## GROWTH RATE

RDBMS systems are designed for steady data retention rather than rapid growth.



## DATA SIZE

Data ranges from terabytes ( $10^{12}$  bytes) to exabytes ( $10^{18}$  bytes).

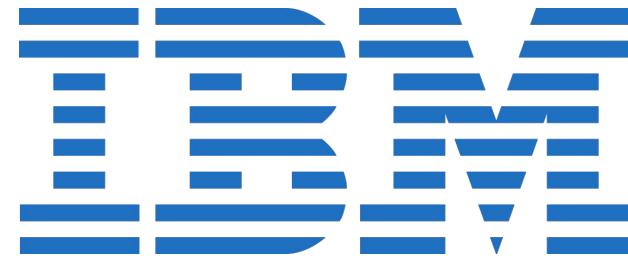
## UNSTRUCTURED DATA

Relational databases can't categorize unstructured data.

# Advantages of Big Data

- 1 Processes all types of data at scale
- 2 Processes huge data quickly in real-time
- 3 Can run anywhere and additional hardware can be added
- 4 Better decision-making, thanks to Hadoop

## Companies Using Big Data



**FUJITSU**

**Google**



**SIEMENS**

**Bloomberg**

**Coca-Cola**

# Big Data: Case Study

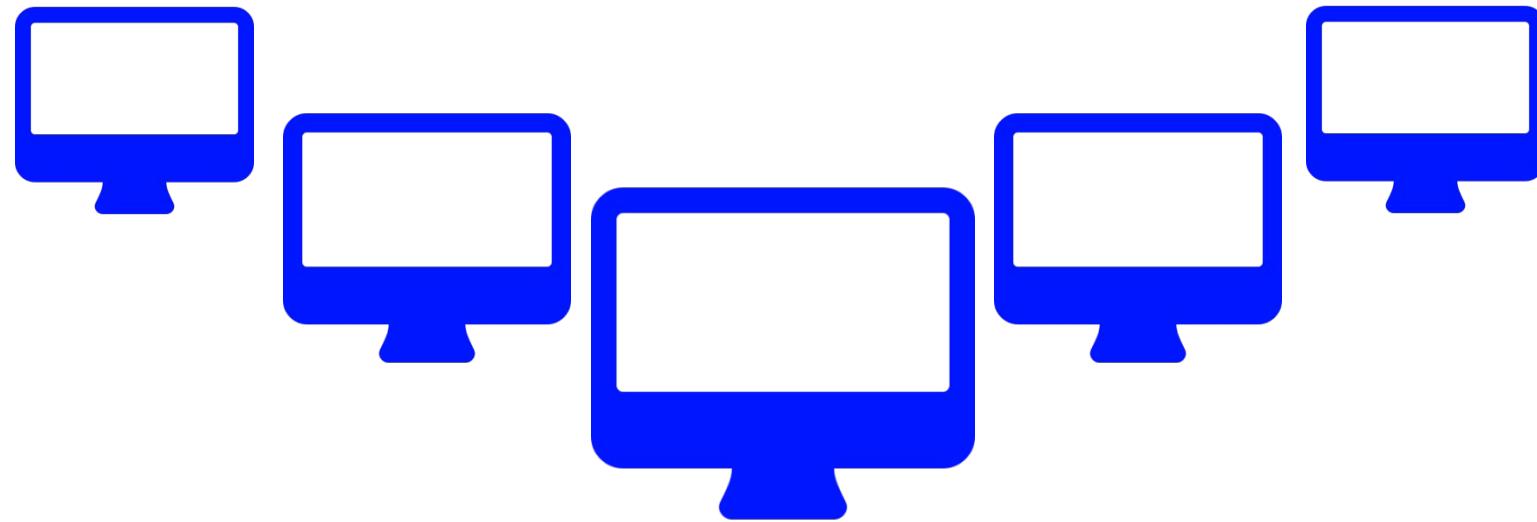


- 1 When do users watch a show?
- 2 Where do they watch it?
- 3 On which device do they watch the show?
- 4 How often do they pause a program?
- 5 How often do they re-watch a program?
- 6 Do they skip the credits?
- 7 What are the keywords searched?

# Big Data: Case Study



Solution



- Traditionally, the analysis of such data was done using a computer algorithm that was designed to produce a correct solution for any given instance.
- As the data started to grow, a series of computers were employed to do the analysis.
- They were also known as **distributed systems**.

## **Features of Big Data Analytics**

# Scalability in Big Data

- A scalable data platform accommodates rapid changes in the growth of data, either in traffic or volume.
- It utilizes and adds hardware or software to increase the output and storage of data.
- When a company has a scalable data platform, it is prepared for the potential of growth in its data needs.

# Fault Tolerance in Big Data

- **Fault tolerance** in Big data or Hadoop HDFS refers to the working strength of a system in unfavorable conditions and how that system can handle such a situation.
- HDFS also maintains the replication factor by creating a replica of **data** on other available machines in the cluster if one machine fails unexpectedly.

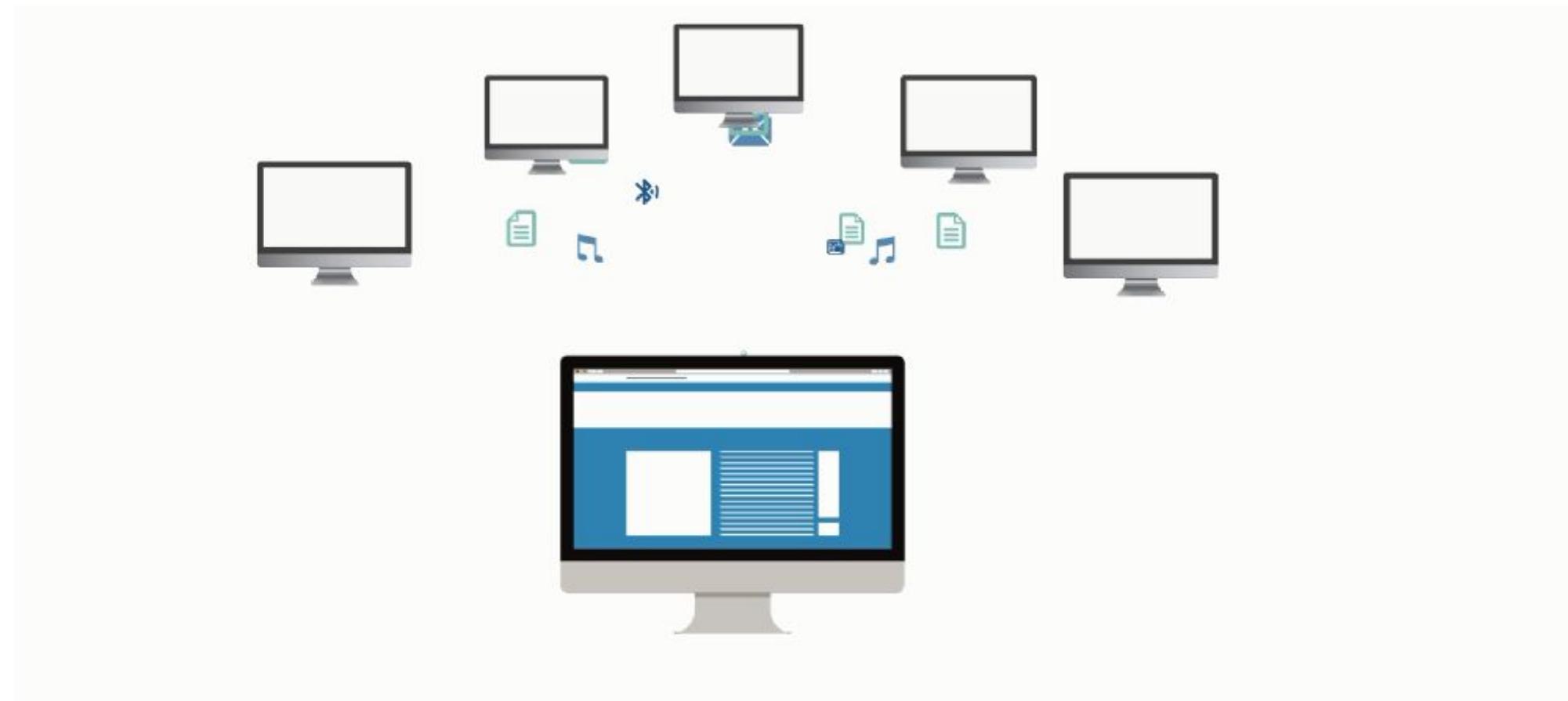
# Data Inconsistency in Big Data

- Once data is captured in **big data, inconsistent** or conflicting phenomena can occur at various granularities.
- It occurs from knowledge content, **data**, information, knowledge, meta-knowledge, to expertise, and can adversely affect the quality of the outcomes in **Big data** analysis process.

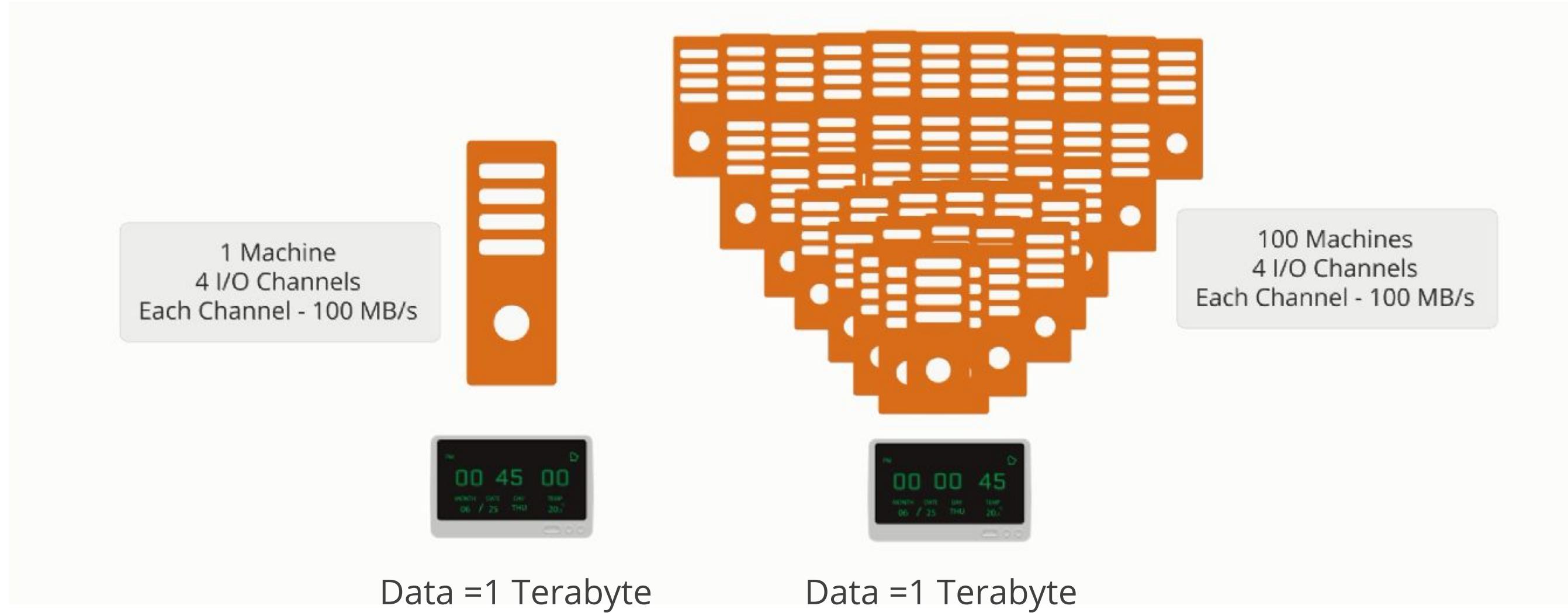
## Distributed Systems

# Distributed Systems

A distributed system is a model in which components located on networked computers communicate and coordinate their actions by passing messages.



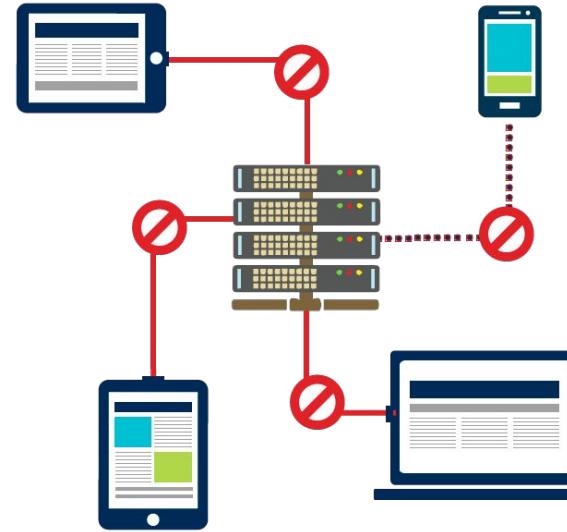
# How Does a Distributed System Work?



In recent times, distributed systems have been replaced by Hadoop.

# Challenges of Distributed Systems

Since, multiple computers are used in a distributed system, there are high chances of:



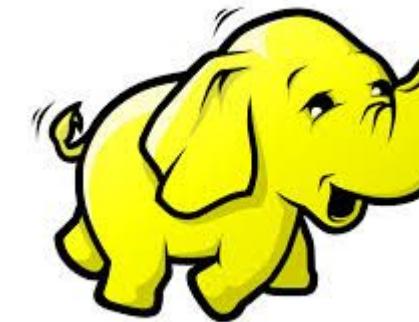
- 1 System failure

- 2 Limited bandwidth

- 3 High programming complexity



Any solution?



Hadoop

## **Introduction to Hadoop**

# What Is Hadoop?

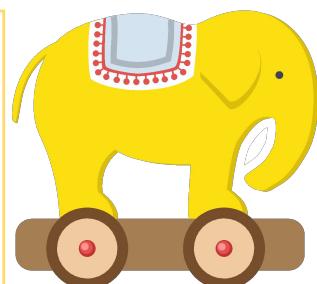
“

Hadoop is a framework that allows distributed processing of large datasets across clusters of commodity computers using simple programming models.

”



Doug Cutting discovered Hadoop and named it after his son's yellow toy elephant. It is inspired by the technical document published by Google.



# Characteristics of Hadoop

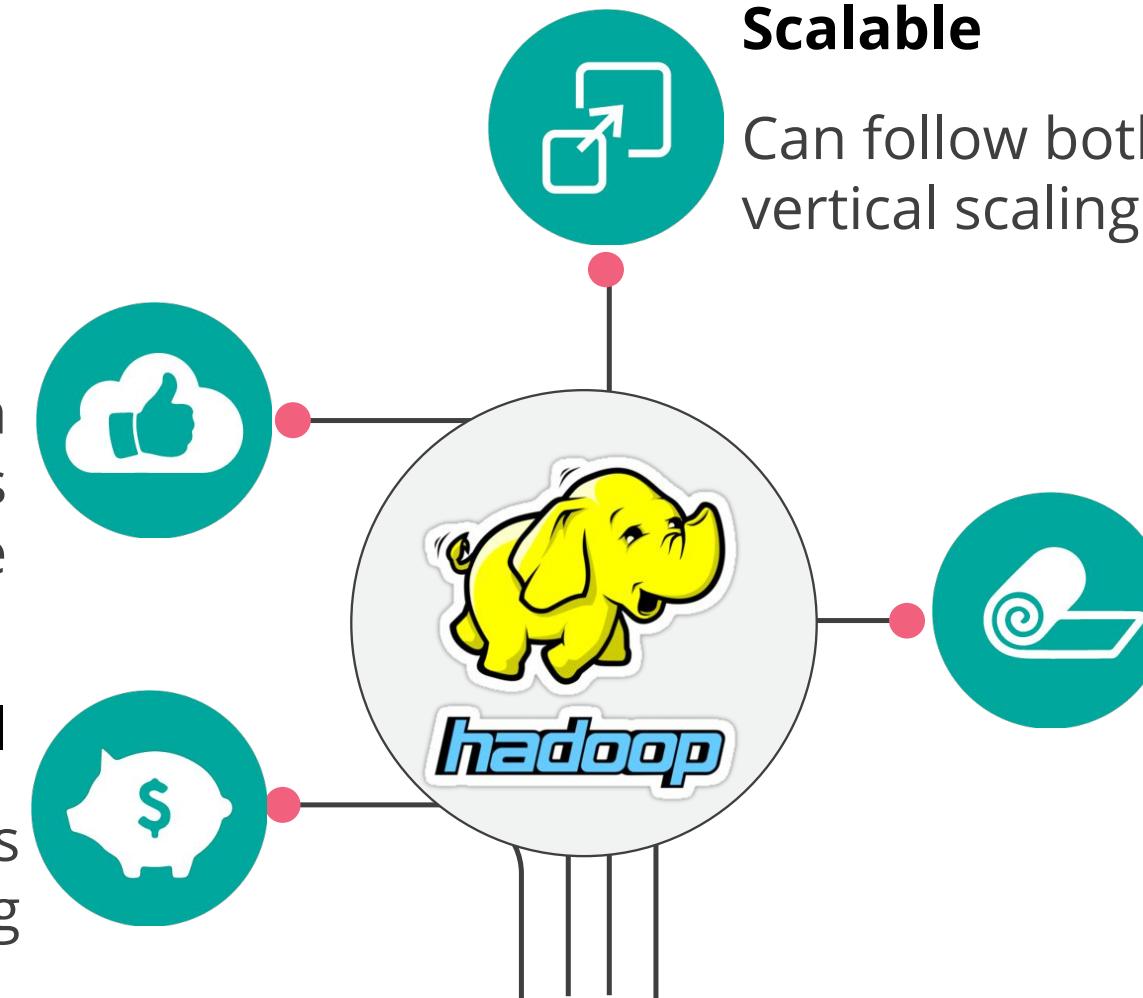
The four key characteristics of Hadoop are:

## Reliable

Stores copies of the data on different machines and is resistant to hardware failure

## Economical

Can use ordinary computers for data processing



## Scalable

Can follow both horizontal and vertical scaling

## Flexible

Can store huge data and decide to use it later

# Traditional Database Systems vs. Hadoop

**Traditional System**



Data sent to the program

**Hadoop**



Program sent to the data

# Analogy of Traditional System and Hadoop



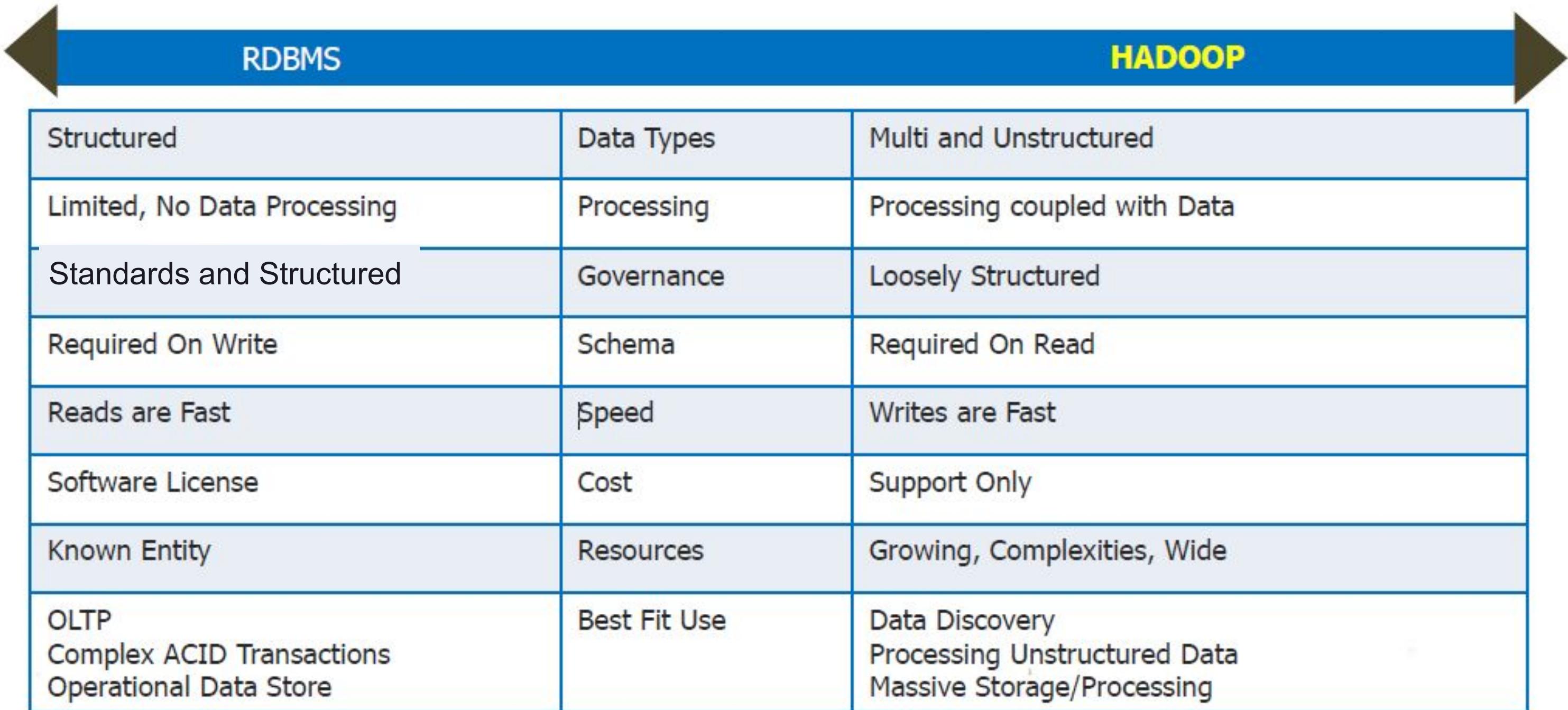
Human brings food toward the mouth

**VS.**

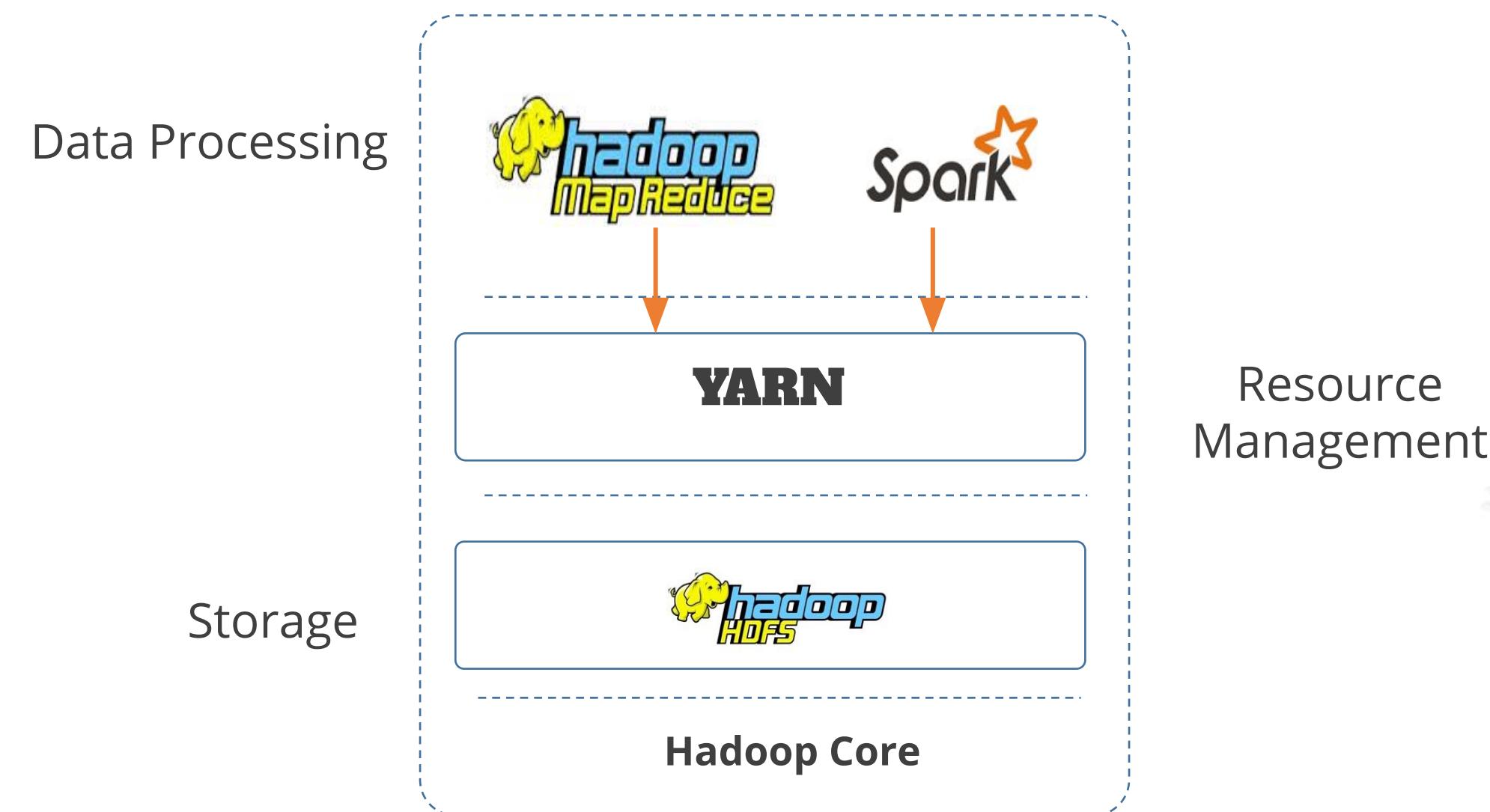


Tiger brings its mouth toward the food

# Traditional Database Systems vs. Hadoop

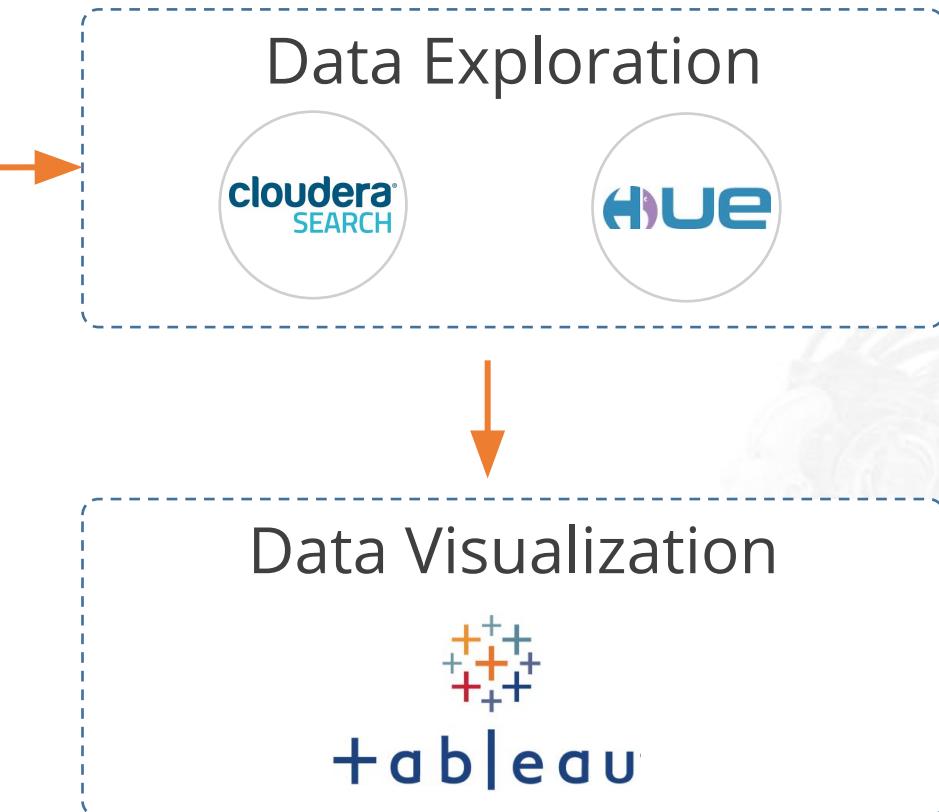
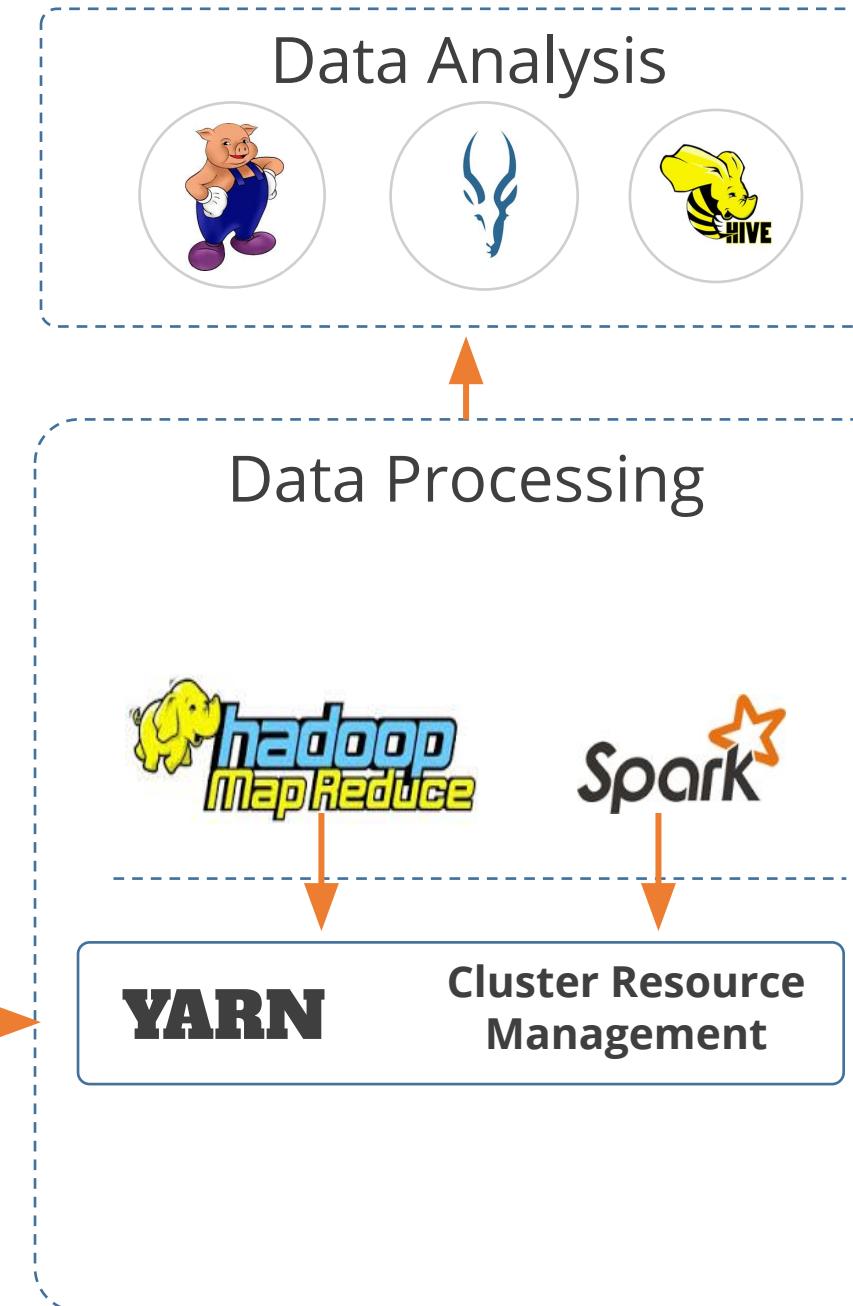
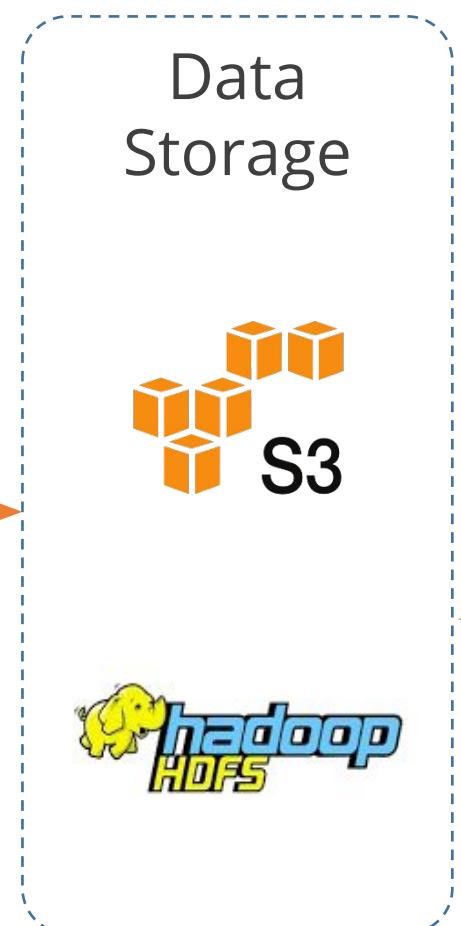


# Hadoop Core Components



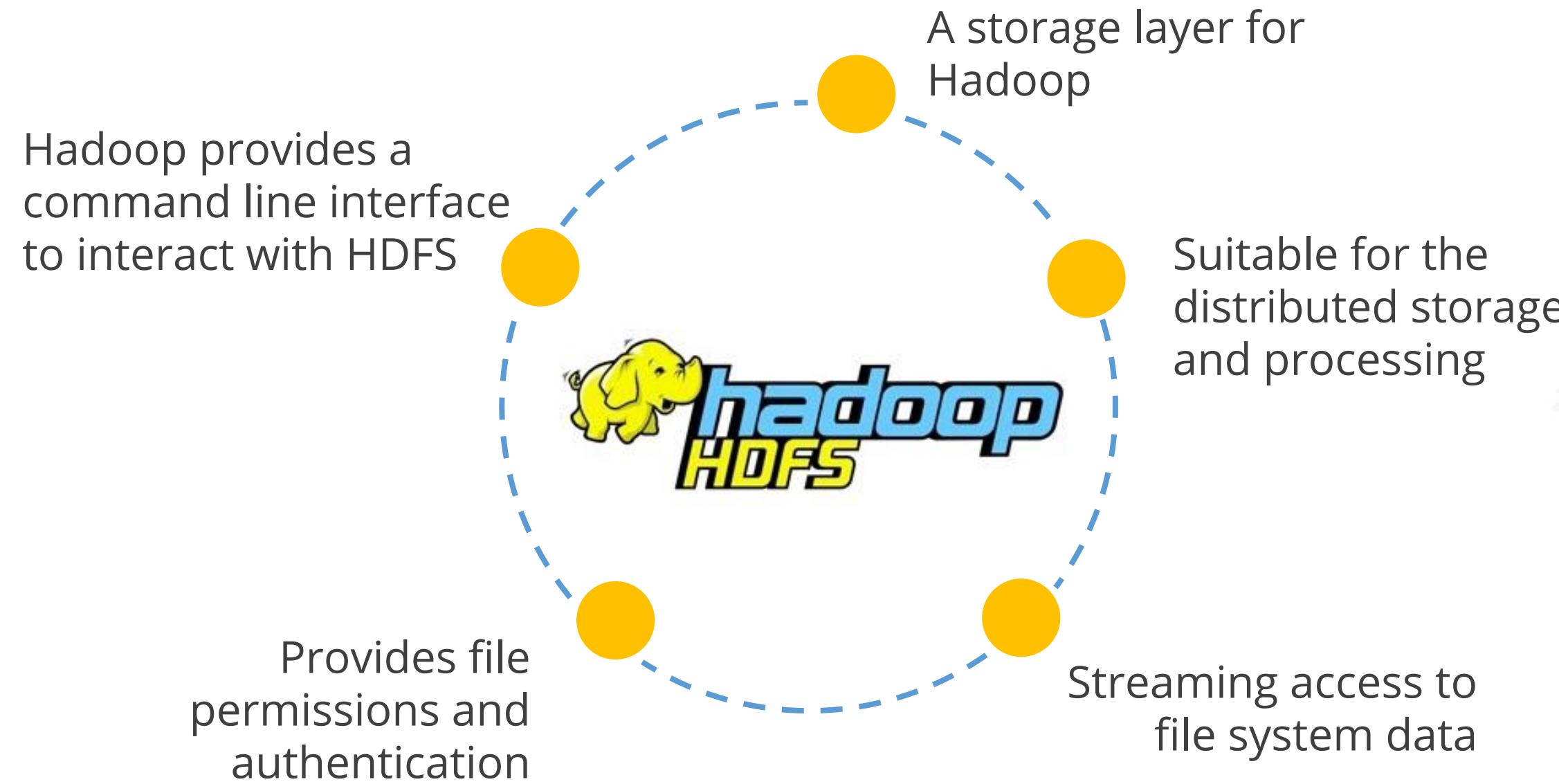
## **Components of Hadoop Ecosystem**

# Components of Hadoop Ecosystem

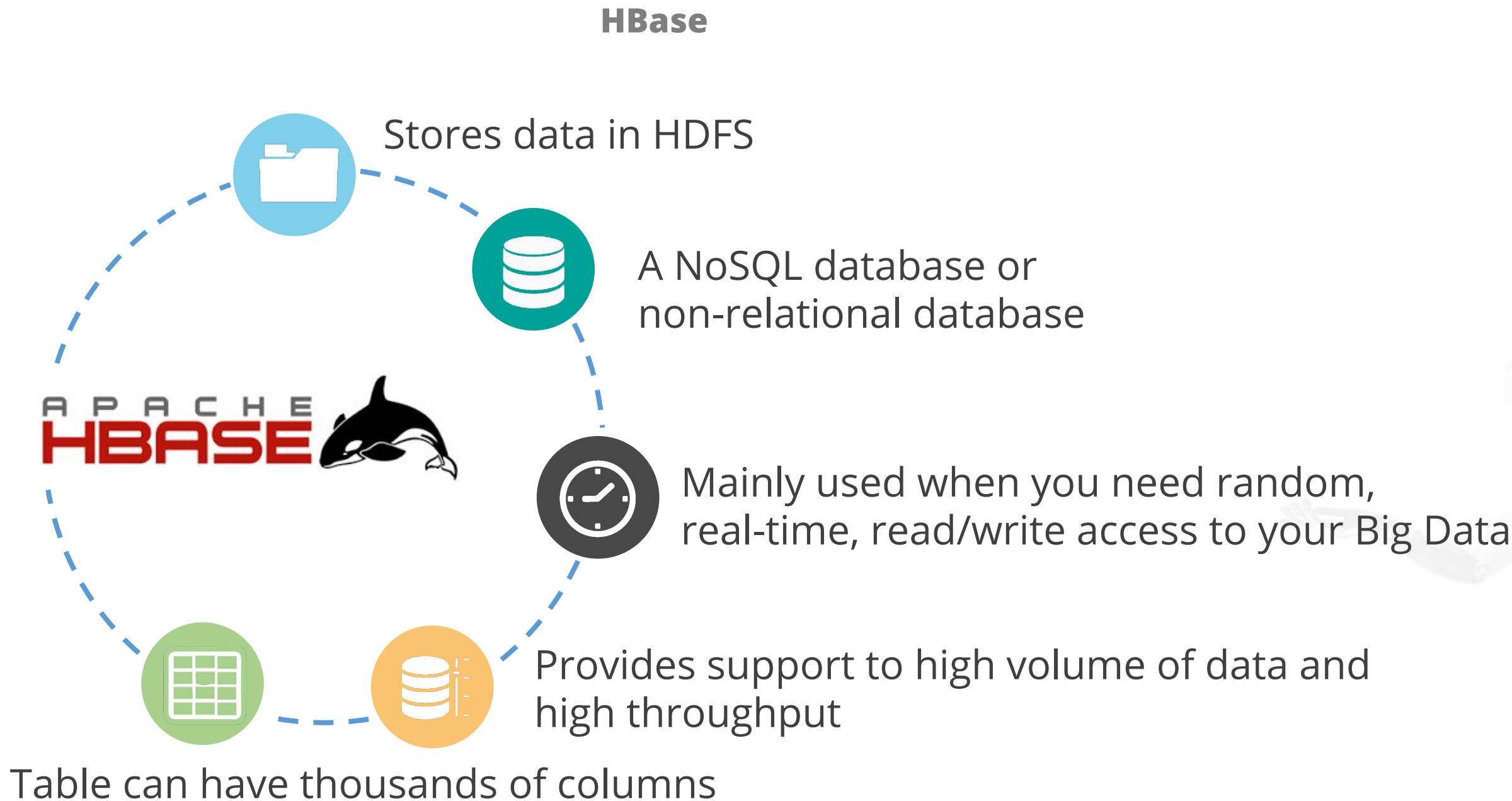


# Components of Hadoop Ecosystem

## HDFS (HADOOP DISTRIBUTED FILE SYSTEM)



# Components of Hadoop Ecosystem



# Components of Hadoop Ecosystem

SQOOP



- Sqoop is a tool designed to transfer data between Hadoop and relational database servers.
- It is used to import data from relational databases such as Oracle and MySQL to HDFS and export data from HDFS to relational databases.

# Components of Hadoop Ecosystem

## FLUME

If you want to ingest event data such as, streaming data, sensor data, or log files, then you can use Flume.



# Components of Hadoop Ecosystem

## SPARK

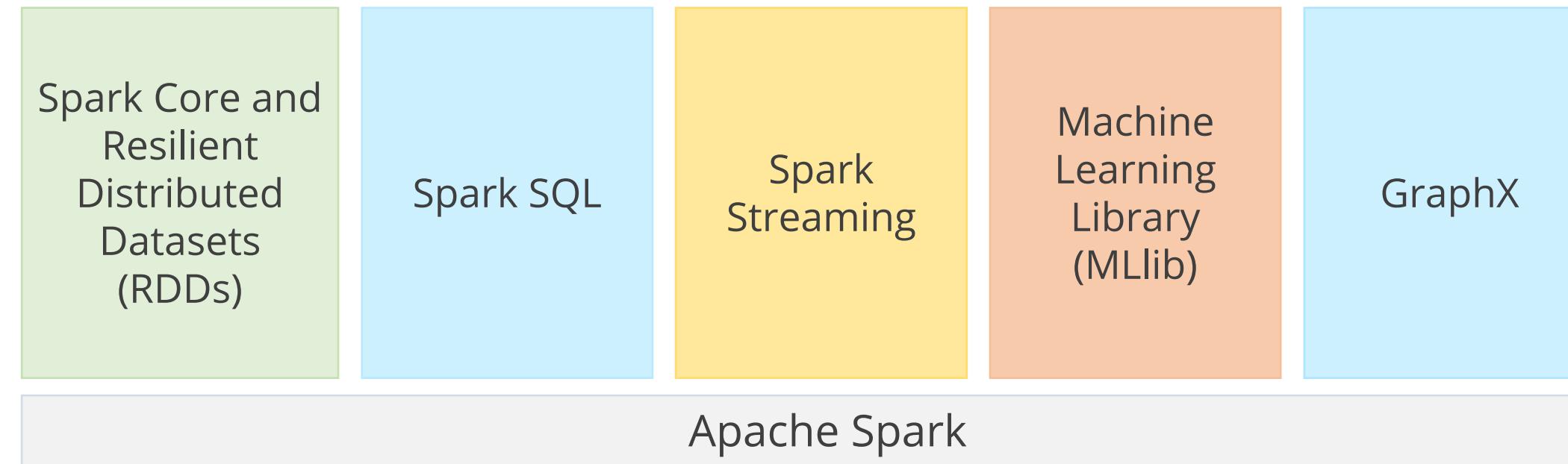
An open source cluster computing framework



Supports machine learning, business intelligence, streaming, and batch processing

# Components of Hadoop Ecosystem

## SPARK: COMPONENTS



# Components of Hadoop Ecosystem

## HADOOP MAP-REDUCE

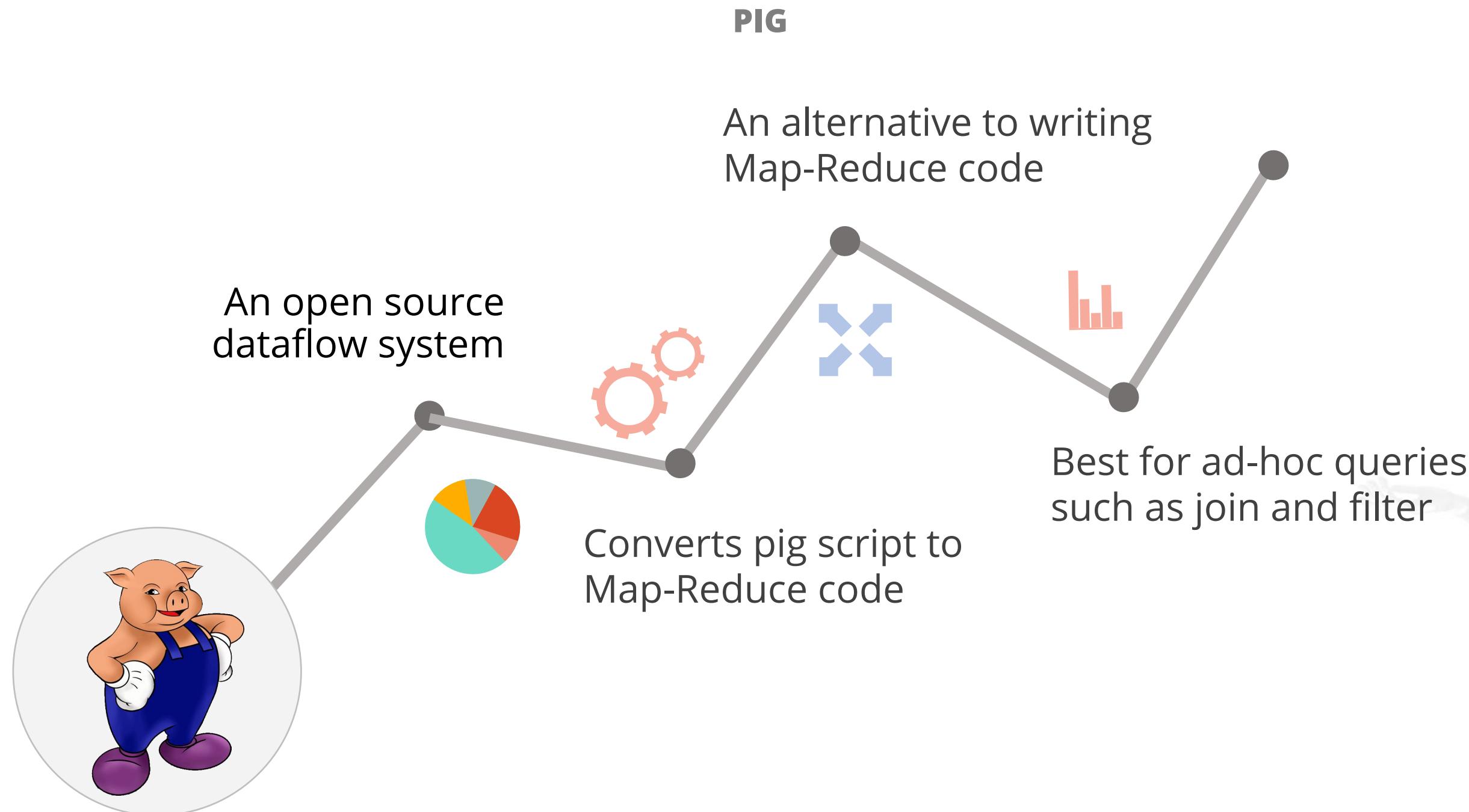
Commonly used  
An extensive and mature fault tolerance framework



The original Hadoop processing engine which is primarily Java based

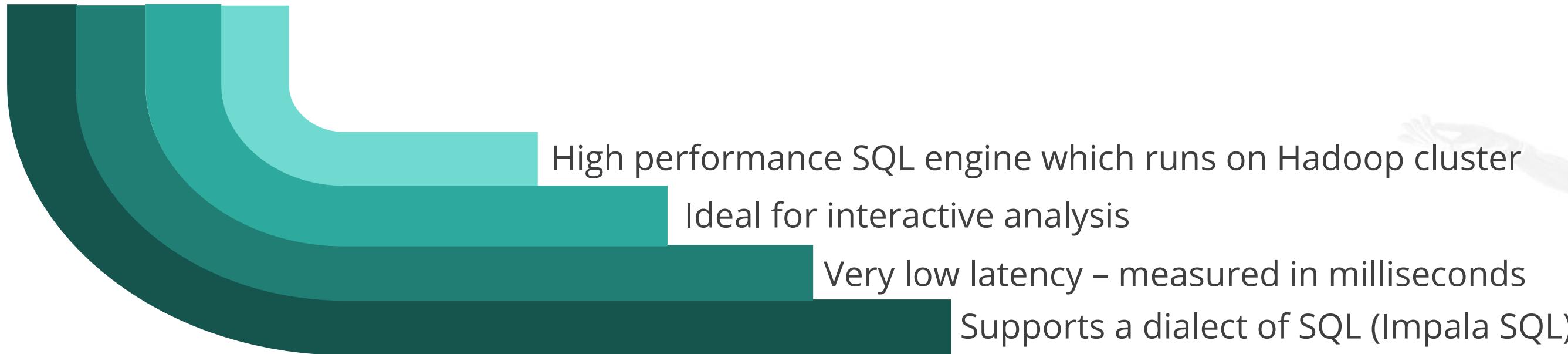
Based on the map and reduce programming model

# Components of Hadoop Ecosystem

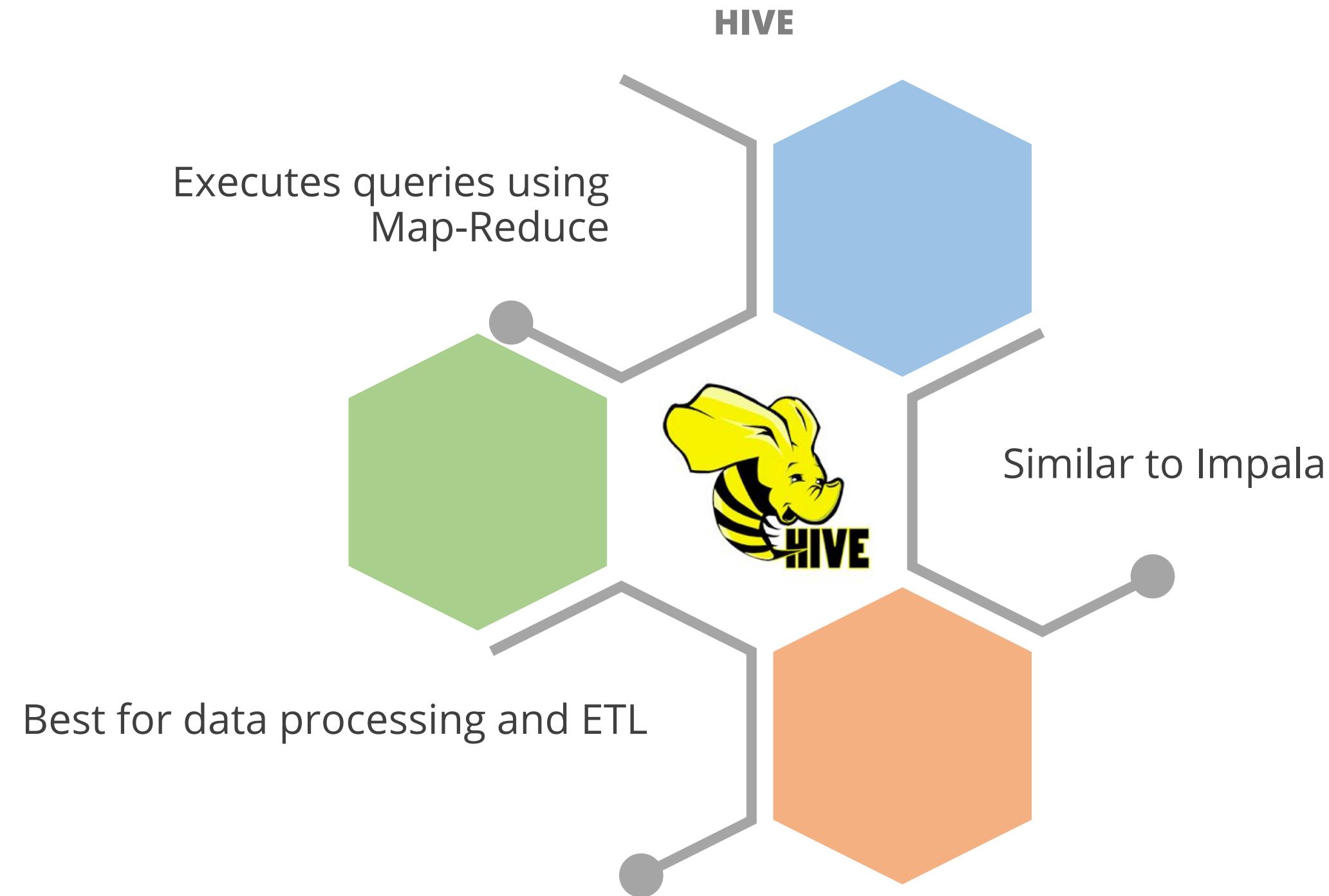


# Components of Hadoop Ecosystem

## IMPALA



# Components of Hadoop Ecosystem



# Components of Hadoop Ecosystem

## CLOUDERA SEARCH

One of Cloudera's near-real-time access products

Eliminates the need to move large datasets across infrastructures to address business tasks



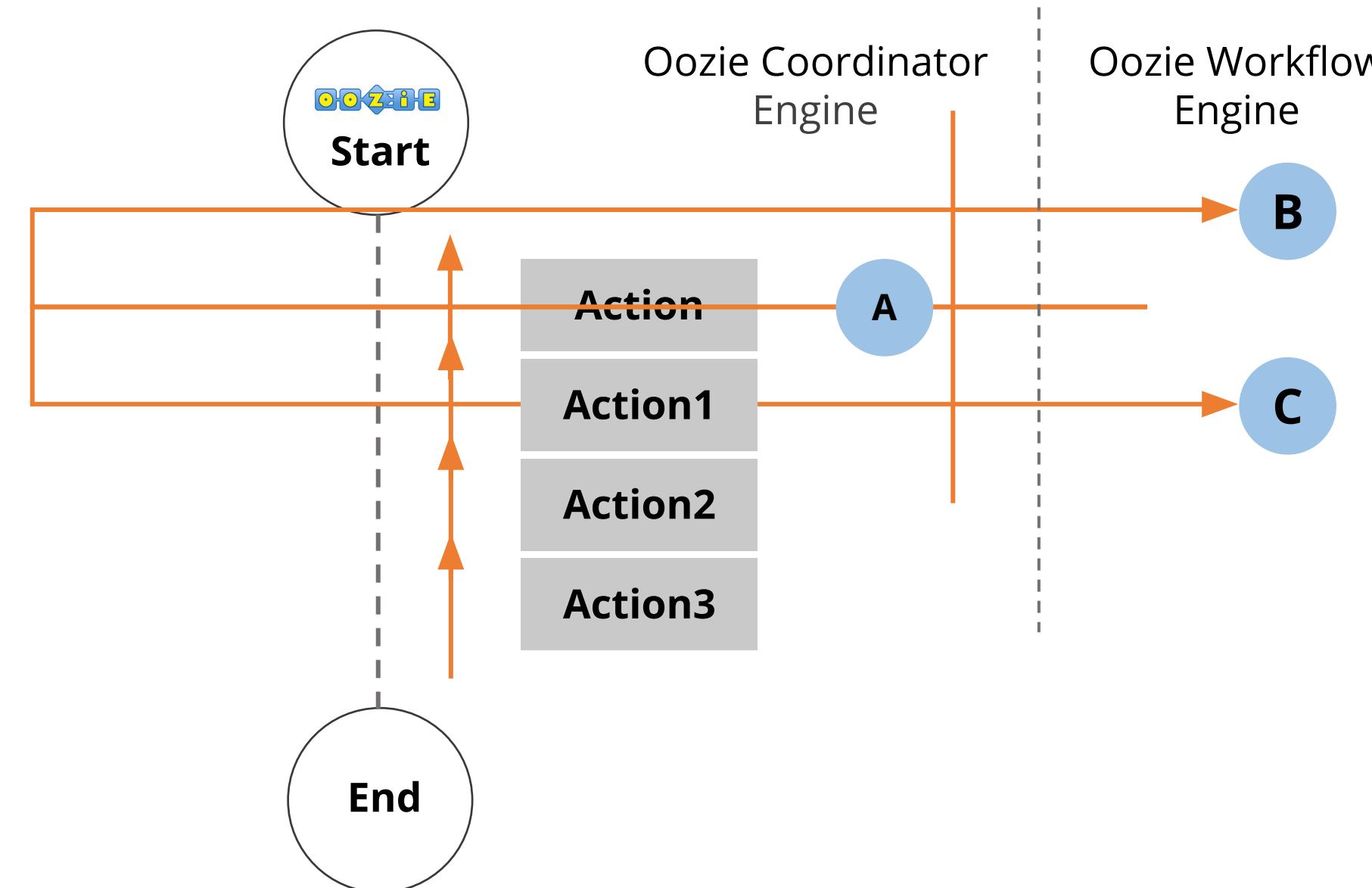
Enables nontechnical users to search and explore data stored in or ingested into Hadoop and HBase

A fully integrated data processing platform

# Components of Hadoop Ecosystem

## OOZIE

Oozie is a workflow or coordination system used to manage the Hadoop jobs



# Components of Hadoop Ecosystem

## HUE (HADOOP USER EXPERIENCE)

Hue is an acronym for Hadoop User Experience

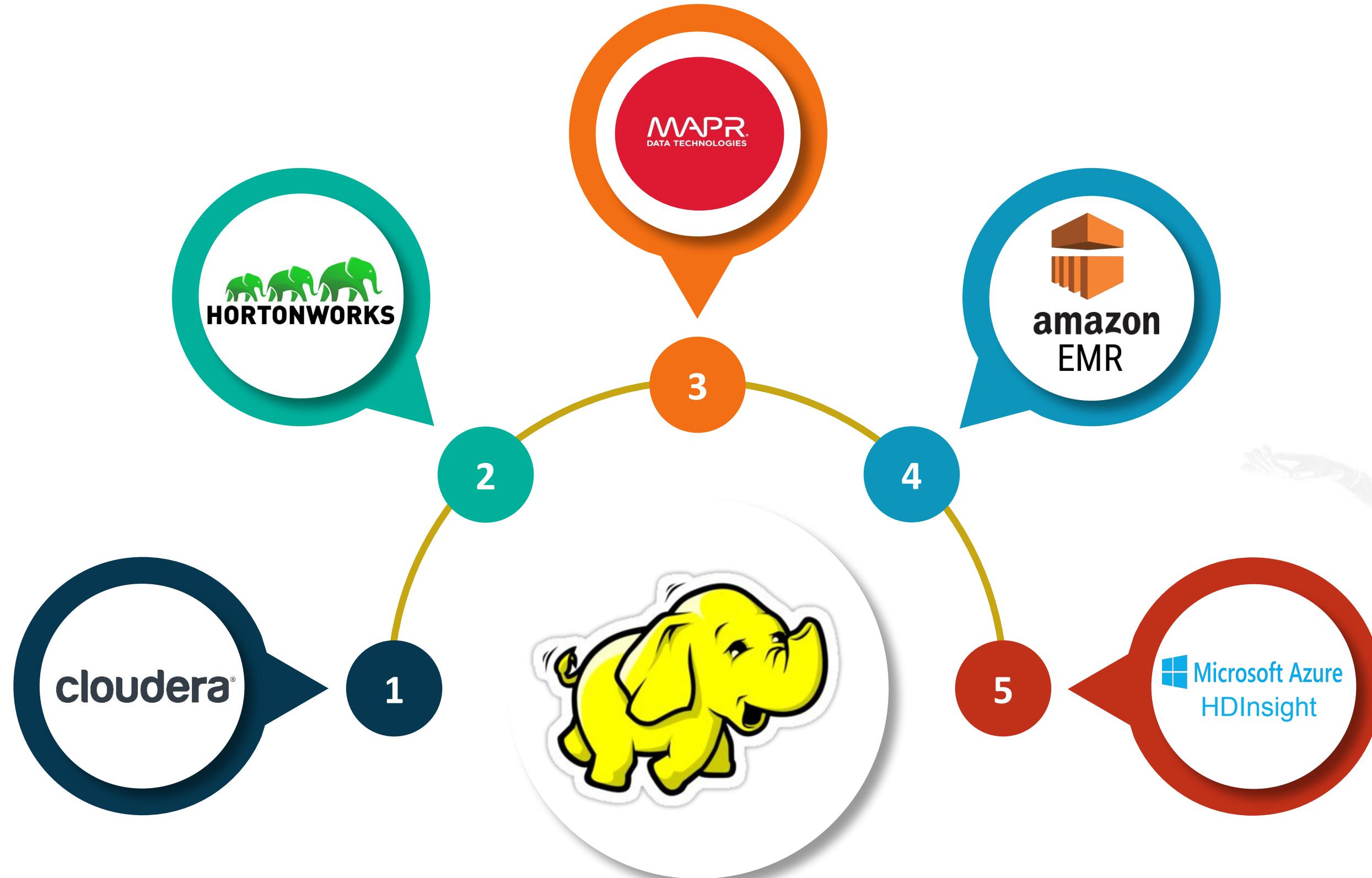
It provides SQL editors for Hive, Impala, MySQL, Oracle, PostgreSQL, Spark SQL, and Solr SQL



Hue is an open source Web interface for analyzing data with Hadoop

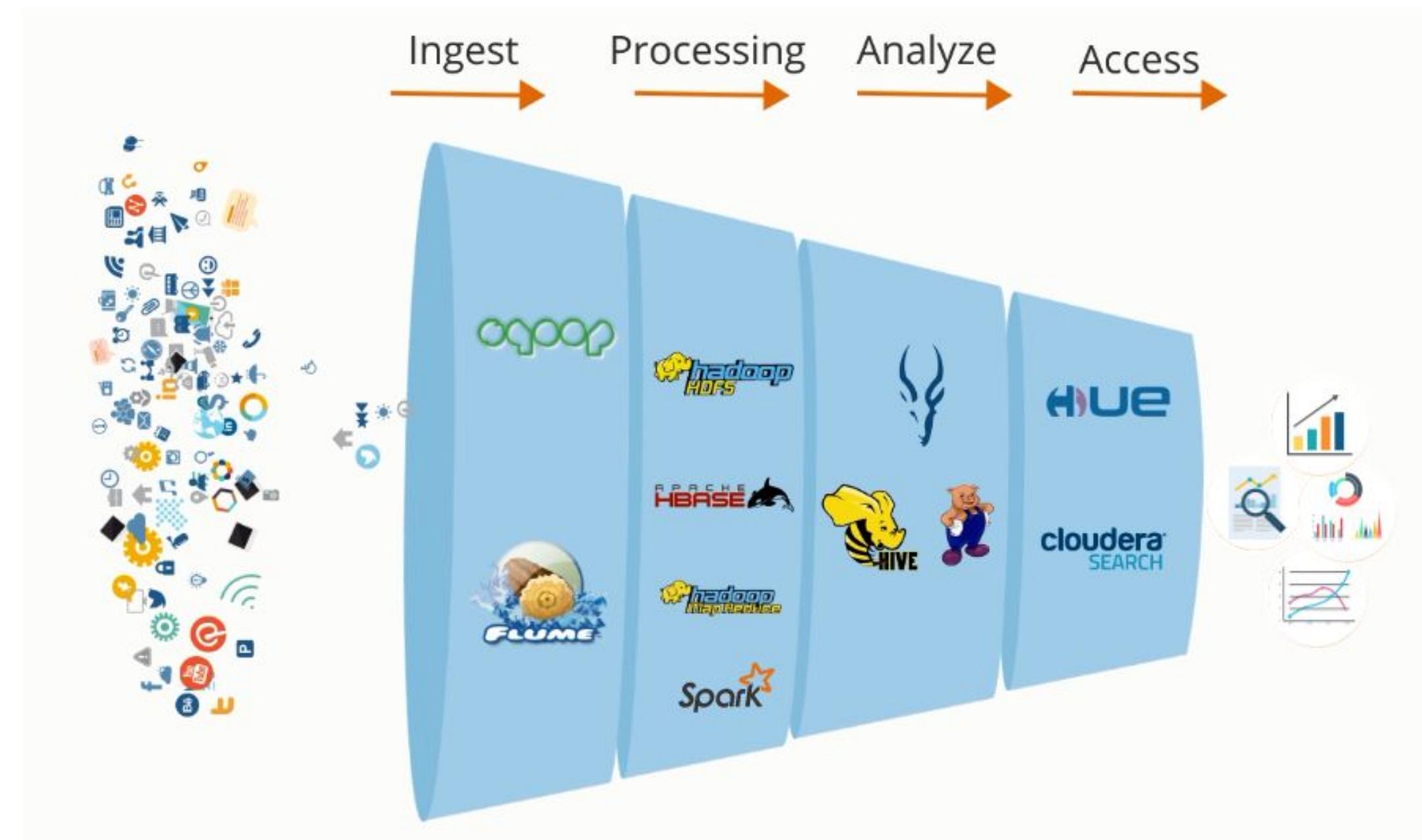
## **Commercial Hadoop Distributions**

# Various Commercial Hadoop Distributions



# Big Data Processing

Components of Hadoop ecosystem work together to process big data.  
There are four stages of big data processing:





### Walk-Through of the Simplilearn Cloud Lab

Duration: 10 mins

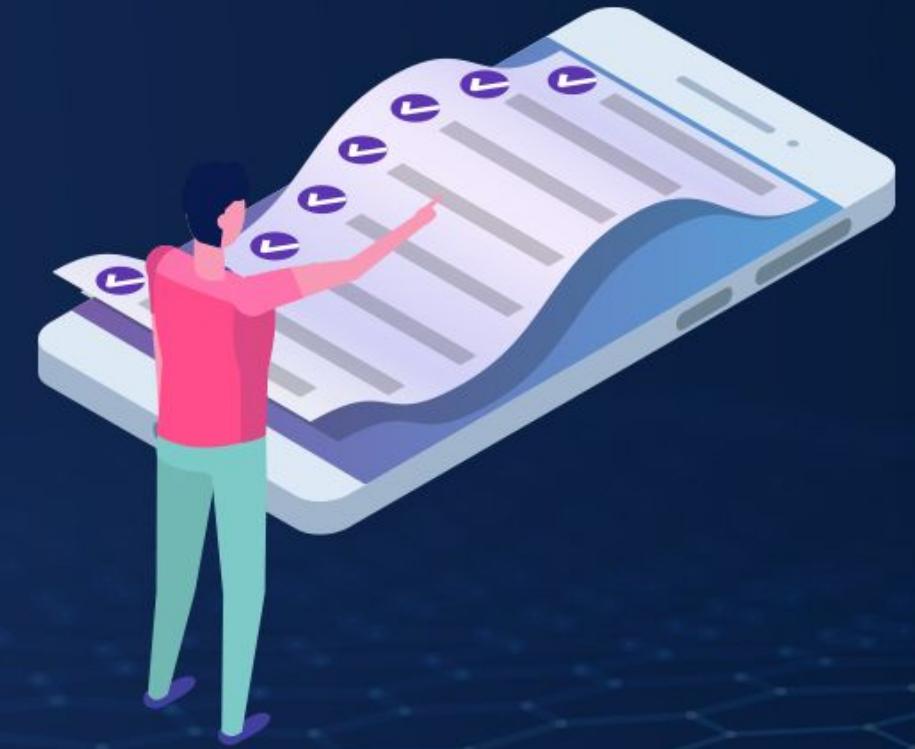
**Problem Statement:** In this demonstration, we will walk you through the Simplilearn cloud lab.

**Access:** Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

## **Key Takeaways**

You are now able to:

- ✓ Describe the concepts of Big Data
- ✓ Explain Hadoop and how it addresses Big Data challenges
- ✓ Describe the components of Hadoop Ecosystem



# DATA AND ARTIFICIAL INTELLIGENCE



## Knowledge Check

**Knowledge  
Check**

**1**

**Which of the following is a source of unstructured data?**

- a. Data from social media websites
- b. Transactional data in Amazon's database
- c. Web and server logs
- d. All of the above



**Knowledge  
Check**

**1**

## Which of the following is a source of unstructured data?

- a. Data from social media websites
- b. Transactional data in Amazon's database
- c. Web and server logs
- d. All of the above



The correct answer is **a.**

Unstructured data comprises of data that is usually not easily searchable, including formats like audio, video, and social media postings.

**Knowledge  
Check**  
**2**

A bank wants to process 1000 transactions per second.  
Which one of the following Vs reflects this real-world use case?

- a. Volume
- b. Variety
- c. Velocity
- d. Veracity



**Knowledge  
Check**  
**2**

A bank wants to process 1000 transactions per second. Which one of the following Vs reflects this real-world use case?

- a. Volume
- b. Variety
- c. Velocity
- d. Veracity



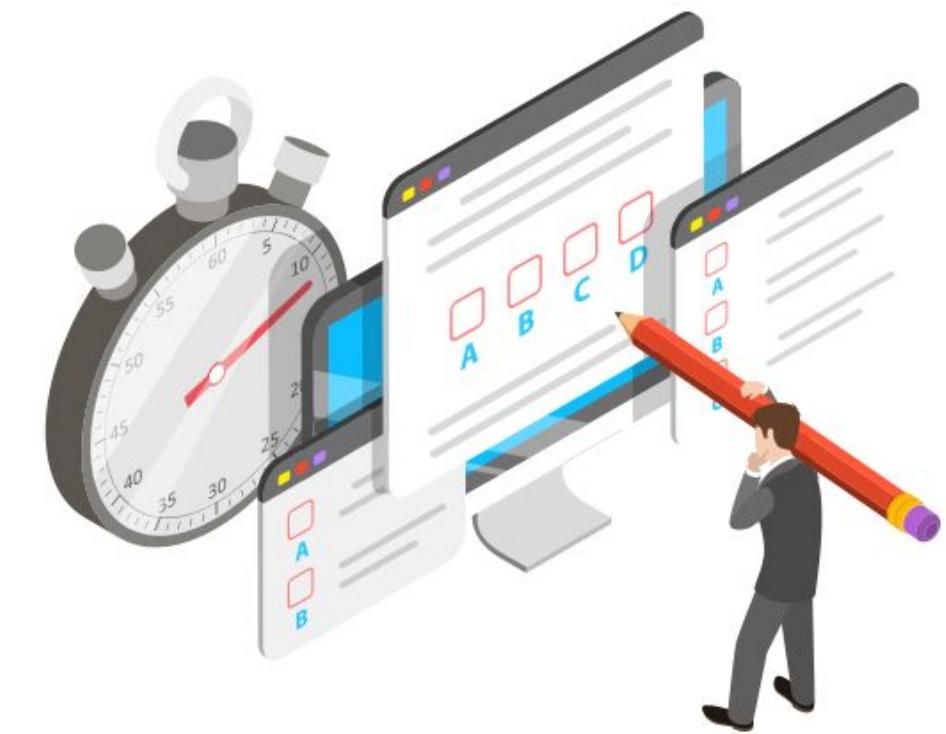
The correct answer is **C.**

Velocity is the frequency of incoming data that needs to be processed. Given use case is an example of an application that handles the velocity of data.

**Knowledge  
Check**  
**3**

**Why has popularity of big data increased tremendously in the recent years?**

- a. Due to increased volume of data
- b. Big data is an open source
- c. Abundance of unstructured data
- d. None of the above



**Knowledge  
Check**

**3**

## Why has popularity of big data increased tremendously in the recent years?

- a. Due to increased volume of data
- b. Big data is an open source
- c. Abundance of unstructured data
- d. None of the above



The correct answer is **a.**

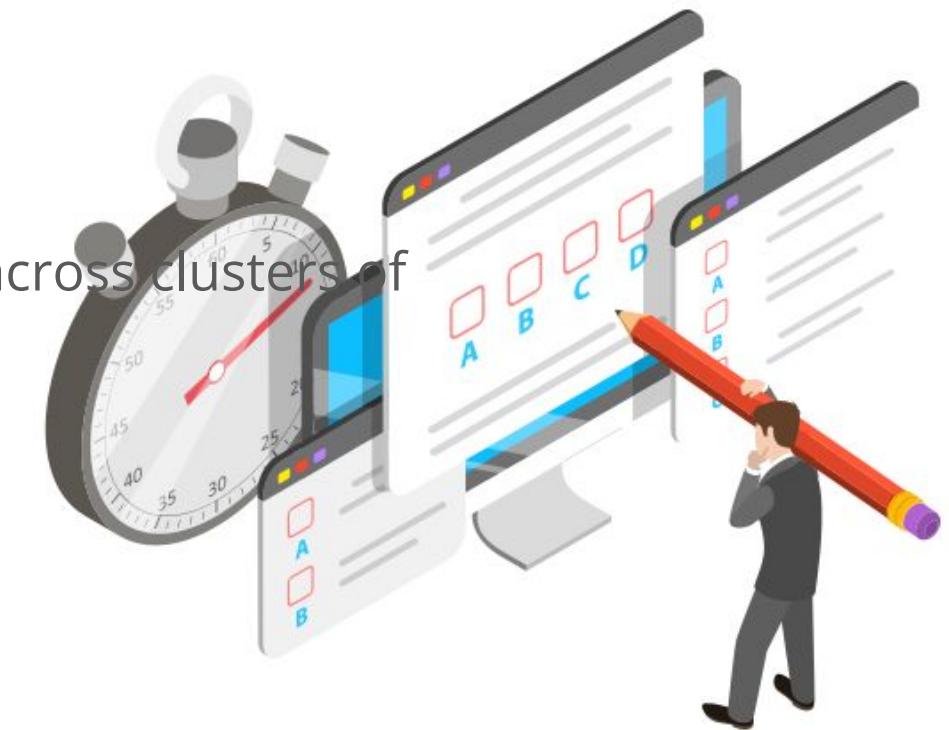
Unstructured data is growing at astronomical rates, contributing to the big data deluge that's sweeping across enterprise data storage environments.

## Knowledge Check

4

### What is Hadoop?

- a. It is an in-memory tool used in Mahout algorithm computing.
- b. It is a computing framework used for resource management.
- c. It is a framework that allows distributed processing of large datasets across clusters of commodity computers using a simple programming model.
- d. It is a search and analytics tool that provides access to analyze data.

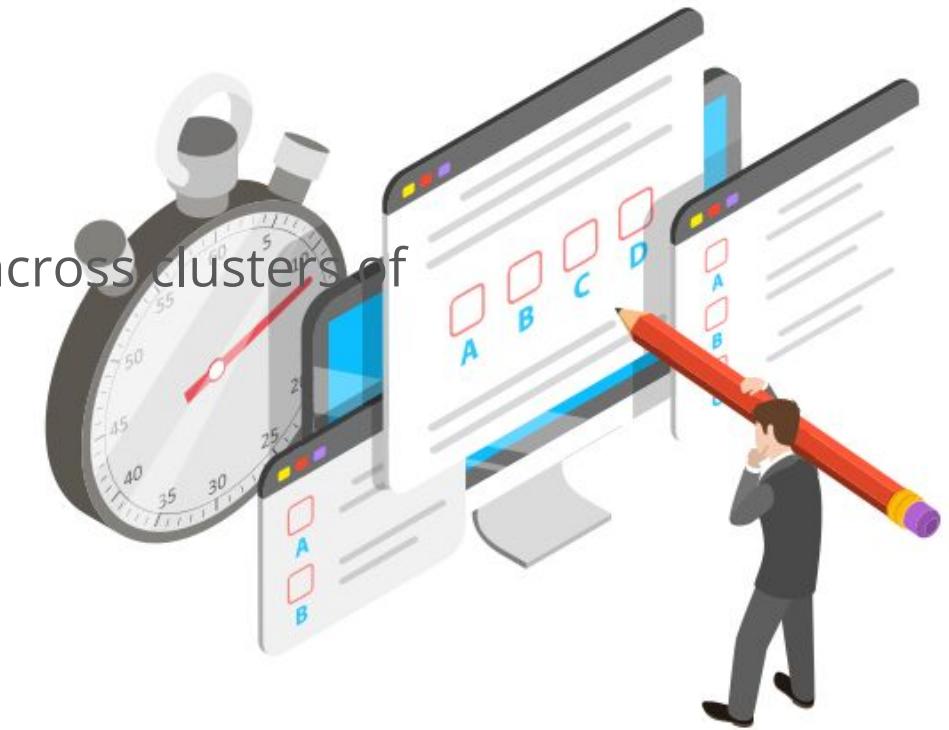


**Knowledge  
Check**

**4**

## What is Hadoop?

- a. It is an in-memory tool used in Mahout algorithm computing.
- b. It is a computing framework used for resource management.
- c. It is a framework that allows distributed processing of large datasets across clusters of commodity computers using a simple programming model.
- d. It is a search and analytics tool that provides access to analyze data.



The correct answer is **C**.

Hadoop is a framework that allows distributed processing of large datasets across clusters of commodity computers using a simple programming model.

**Knowledge  
Check**

**5**

**Which of the following is a column-oriented NoSQL database that runs on top of HDFS?**

- a. MongoDB
- b. Flume
- c. Ambari
- d. HBase



**Knowledge  
Check**  
**5**

**Which of the following is a column-oriented NoSQL database that runs on top of HDFS?**

- a. MongoDB
- b. Flume
- c. Ambari
- d. HBase



The correct answer is **d.**

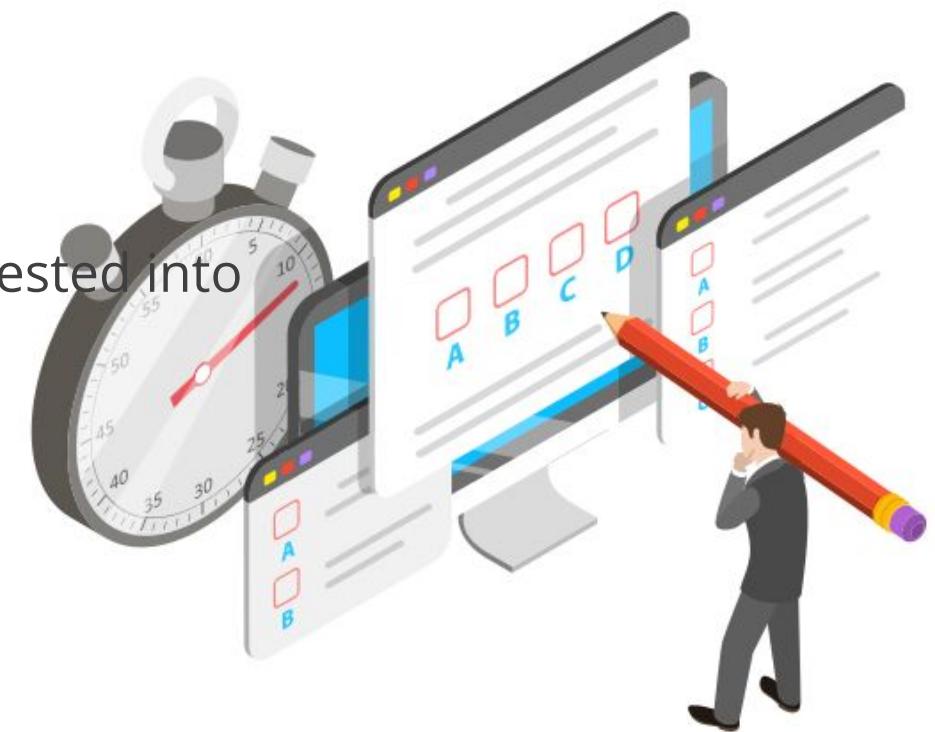
Apache HBase is a NoSQL database that runs on top of Hadoop as a distributed and scalable big data store.

**Knowledge  
Check**

**6**

**Scoop is used to \_\_\_\_\_.**

- a. Import data from relational databases to Hadoop HDFS and export from Hadoop file system to relational databases
- b. Execute queries using Map-Reduce
- c. Enable nontechnical users to search and explore data stored in or ingested into Hadoop and HBase
- d. Stream event data from multiple systems



**Knowledge  
Check**

**6**

**Scoop is used to \_\_\_\_\_.**

- a. Import data from relational databases to Hadoop HDFS and export from Hadoop file system to relational databases
- b. Execute queries using Map-Reduce
- c. Enable nontechnical users to search and explore data stored in or ingested into Hadoop and HBase
- d. Stream event data from multiple systems



The correct answer is **a.**

Scoop is used to import data from relational databases to Hadoop HDFS and export from Hadoop file system to relational databases.

# DATA AND ARTIFICIAL INTELLIGENCE

Thank You