Data Analysis Lab

Big Data concept
Problems
Challenges
Technologies

Fátima Leal
2022/2023

DCT DEPARTAMENTO CIÊNCIA E TECNOLOGIA

UPT UNIVERSIDADE PORTUCALENSE

Previous Lessons

- **Python Introduction**
- **Numpy Module** for matrixes
- **Pandas** to structure data

- Now, let us start with some concepts for data analysis

Outline

- Context
- Big Data Definition
- Problems
- Challenges
- Solutions
- Technologies

Context

- Imagine that a company asks Data Analysis Lab’s students to discover the favourite dish of Portuguese people.

- Would you accept the challenge?

?

Context

- What is the favourite dish of Portuguese people?

How do we start?

Context

We need Data!!

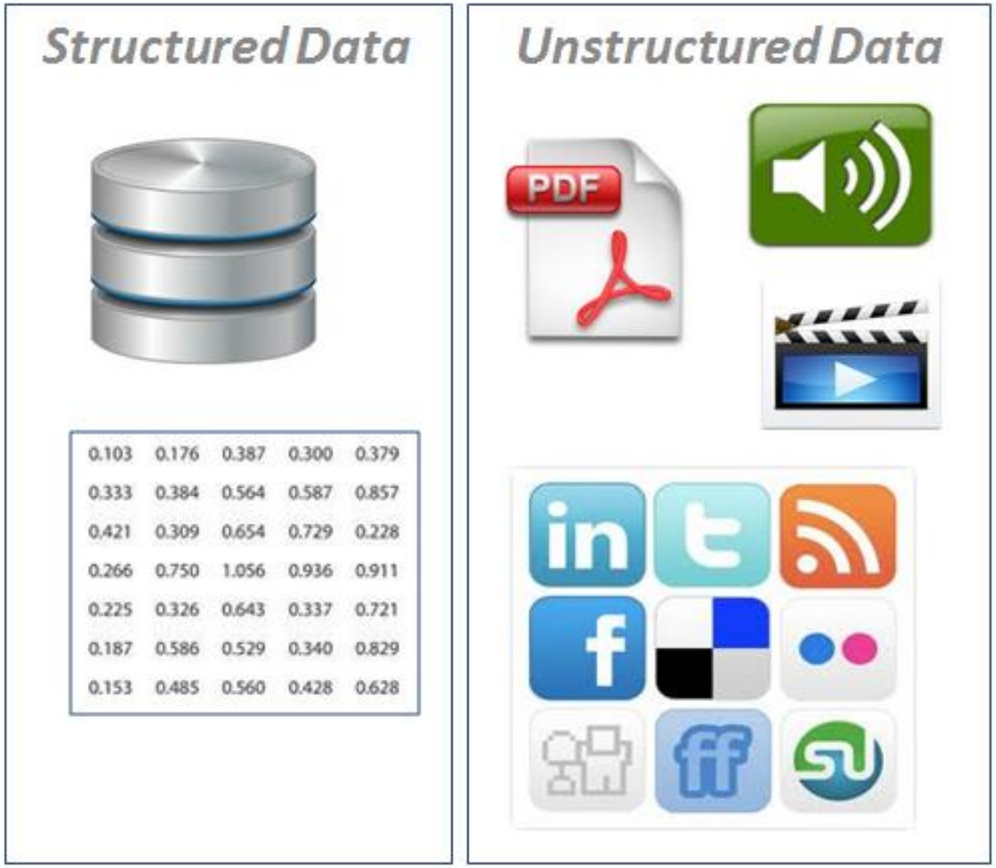
Context

What is data?

Context

How can we represent data?

Context



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Context

- Returning to the challenge…
- What is the favourite dish of Portuguese people?

- Who could provide data to answer this question?

Context

Thousands of Supermarkets!

Thousands of Suppliers!

Thousands of restaurants!

Social Networks!

….

….

….

Context

What is the result of data analysis?

Context

- What is the result of data analysis?
  - Information
  - Answers to the questions
  - Path to solve a problem
  - More questions…

Context

- This is the start of Big Data concept.
- More, more and more data
- New information
- Multiple Applications
- Extract knowledge from the data

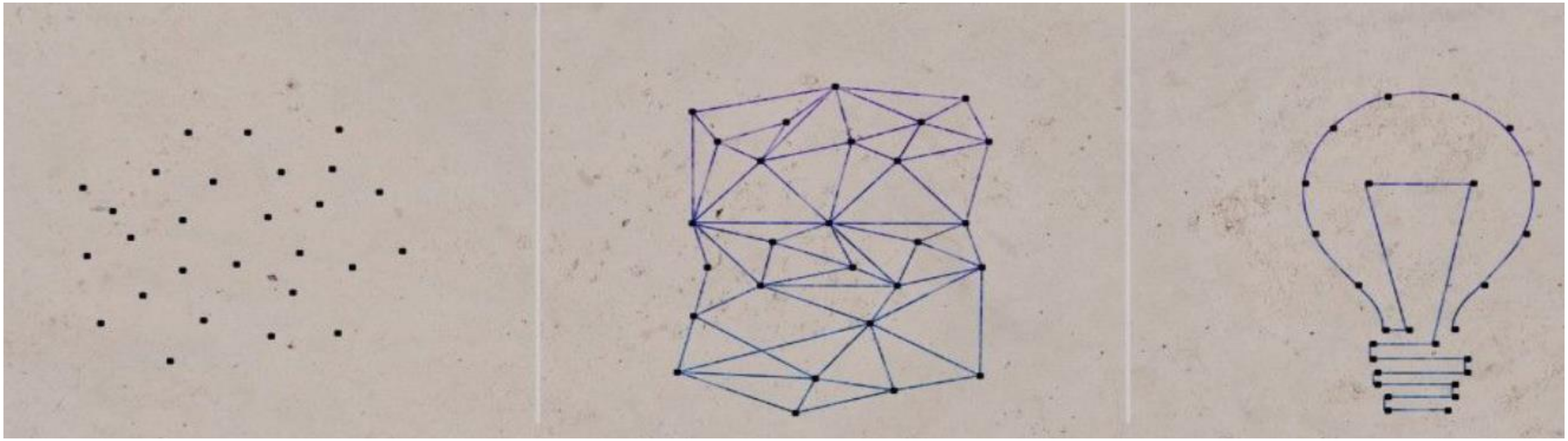
Then, we can discover a favourite dish per Portuguese region!

Context

Data

Information

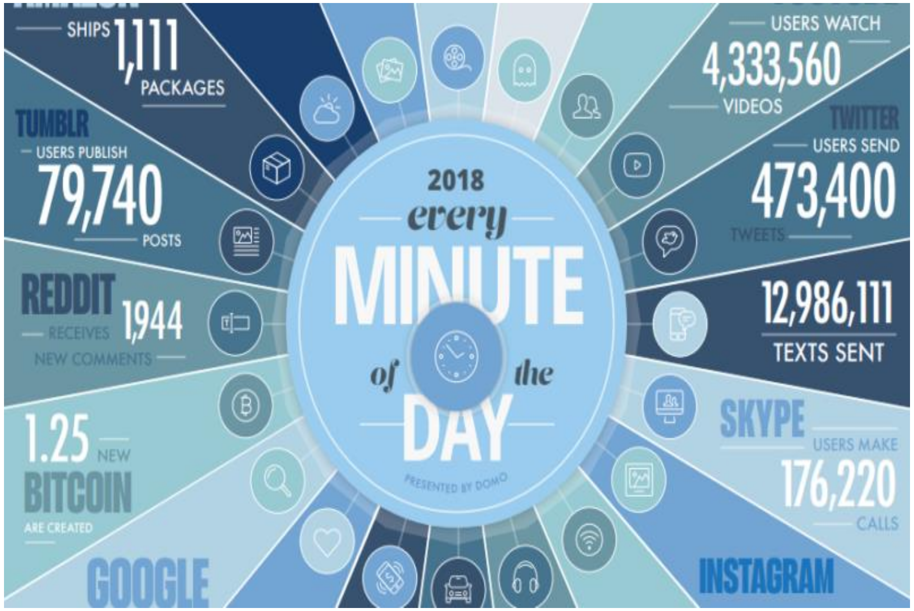
Insights



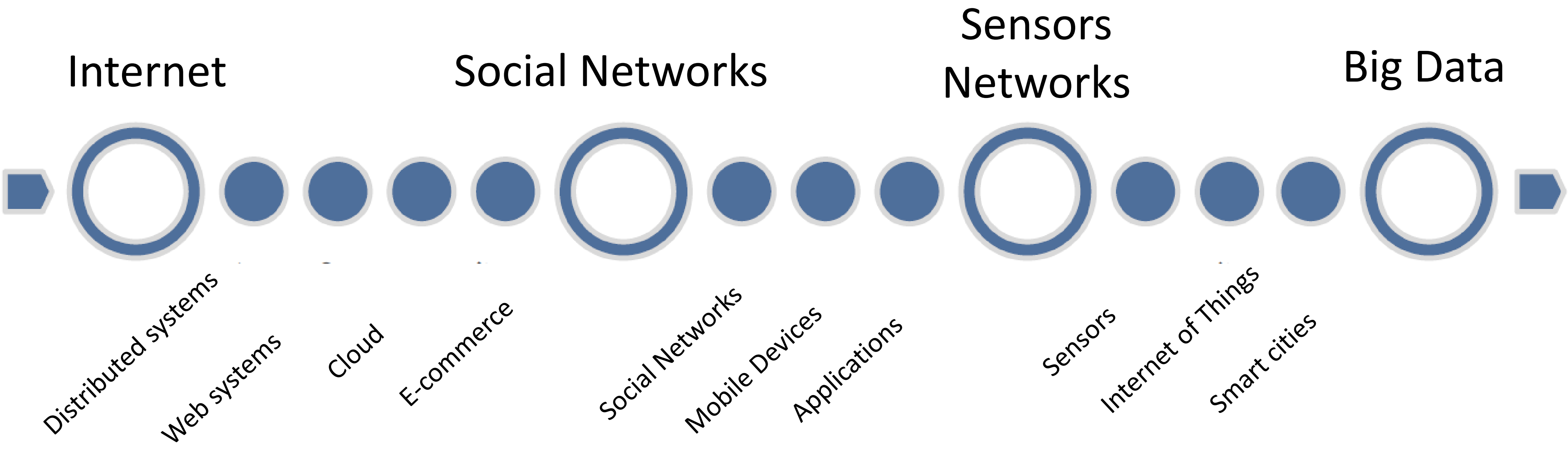
What is Big Data?

What is Big Data?

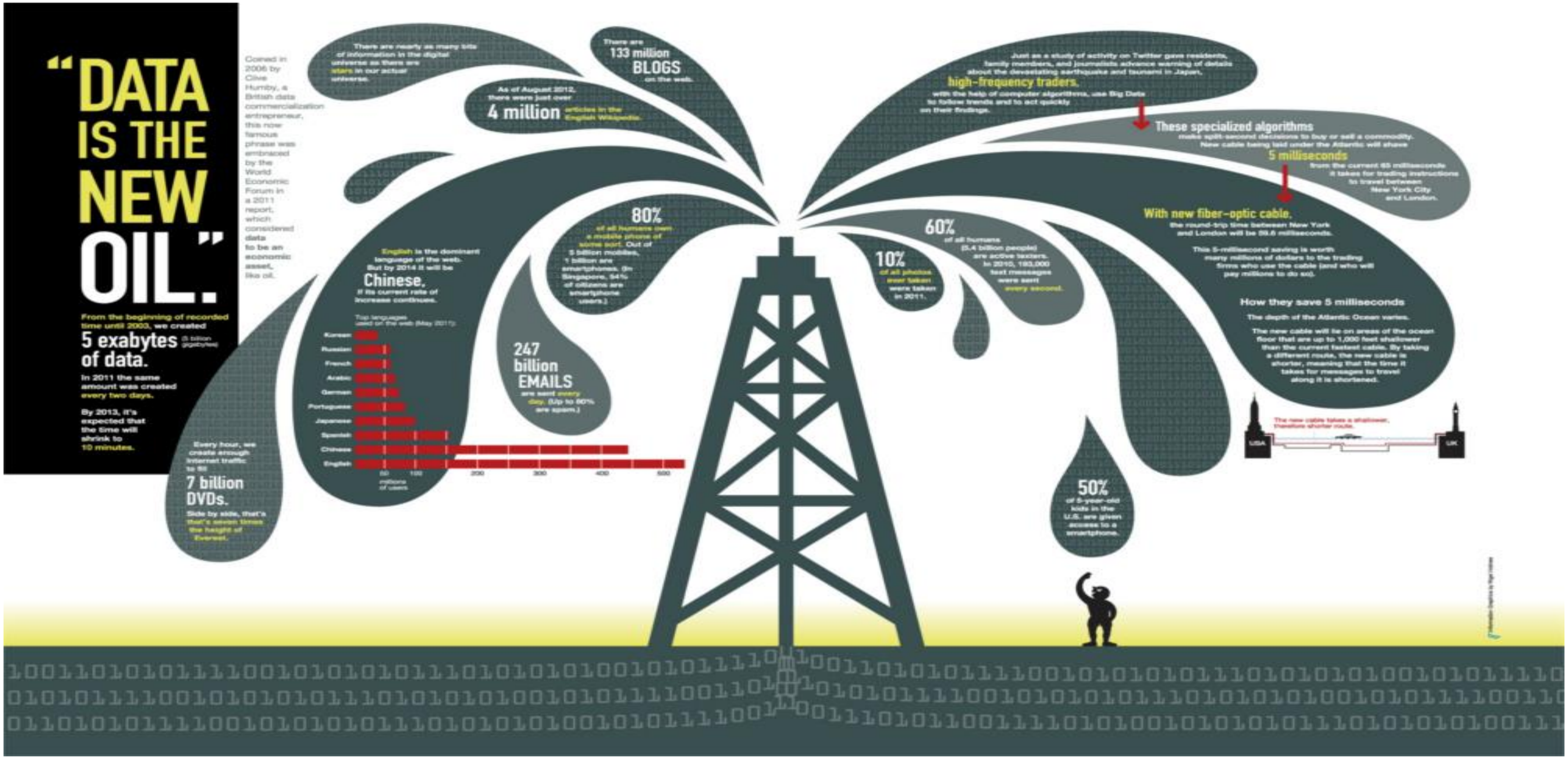
- 47 % of world population uses Internet
- Google does 3 877 140 searches per minute
- Social Networks
  - Instagram: 49 380 photos per minute
  - Youtube: 433 560 videos per minute
  - Twitter: 473 400 tweets per minute
- Communications: SMS, calls, etc.
- Services: Uber
- IoT: sensors



What is Big Data?



What is Big Data?


"DATA IS THE NEW OIL."
From the beginning of recorded time until 2003, we created
5 exabytes (5 billion gigabytes) of data.
In 2011 the same amount was created every two days.
By 2013, it's expected that the time will shrink to 10 minutes.
Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.
There are nearly as many bits of information in the digital universe as there are stars in our actual universe.
As of August 2012, there were just over
4 million articles in the English Wikipedia.
There are 133 million BLOGS on the web.
English is the dominant language of the web. But by 2014 it will be
Chinese,
if its current rate of increase continues.
Top languages used on the web (May 2011):
Korean
Russian
French
Arabic
German
Portuguese
Japanese
Spanish
Chinese
English
50
100
200
300
400
500
millions of users
80%
of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)
247 billion EMAILS
are sent every day. (Up to 80% are spam.)
Every hour, we create enough internet traffic to fill
7 billion DVDs.
Side by side, that's that's seven times the height of Everest.
Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan,
high-frequency traders,
with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.
These specialized algorithms
make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave
5 milliseconds
from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.
With new fiber-optic cable,
the round-trip time between New York and London will be 59.6 milliseconds.
This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).
How they save 5 milliseconds
The depth of the Atlantic Ocean varies.
The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.
The new cable takes a shallower, therefore shorter route.
USA
UK
60%
of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.
10%
of all photos ever taken were taken in 2011.
50%
of 5-year-old kids in the U.S. are given access to a smartphone.

UPT DCT DEPARTAMENTO CIÊNCIA E TECNOLOGIA

What is Big Data?

- Big Data is characterized by a set of Vs…

- Try to find out which and how many…

What is Big Data?

What is Big Data?

What is Big Data?

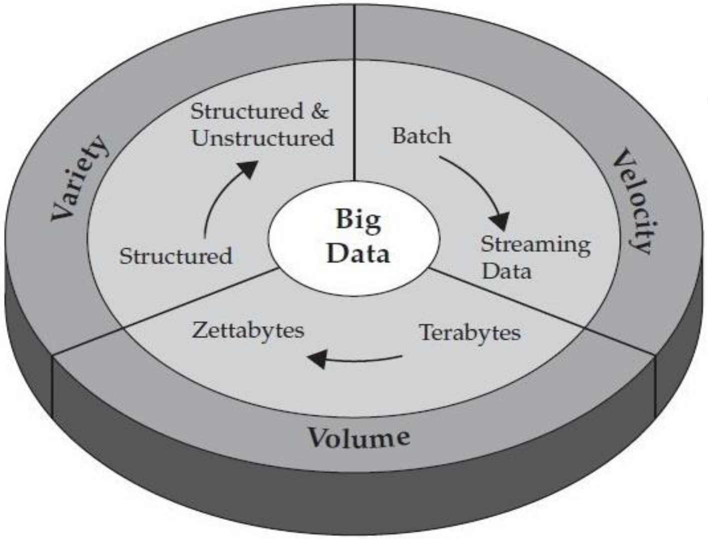
What is Big Data?

- Big data is an abstract concept: it does not involve just large amount of data

- As we could see, in “3Vs” model:
  - **Volume**: generation and collection of massive data
  - **Velocity**: data collection and analysis, *etc.*, must be rapidly and timely conducted
  - **Variety**: indicates the various types of data (semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data)

Big Data 3V’s

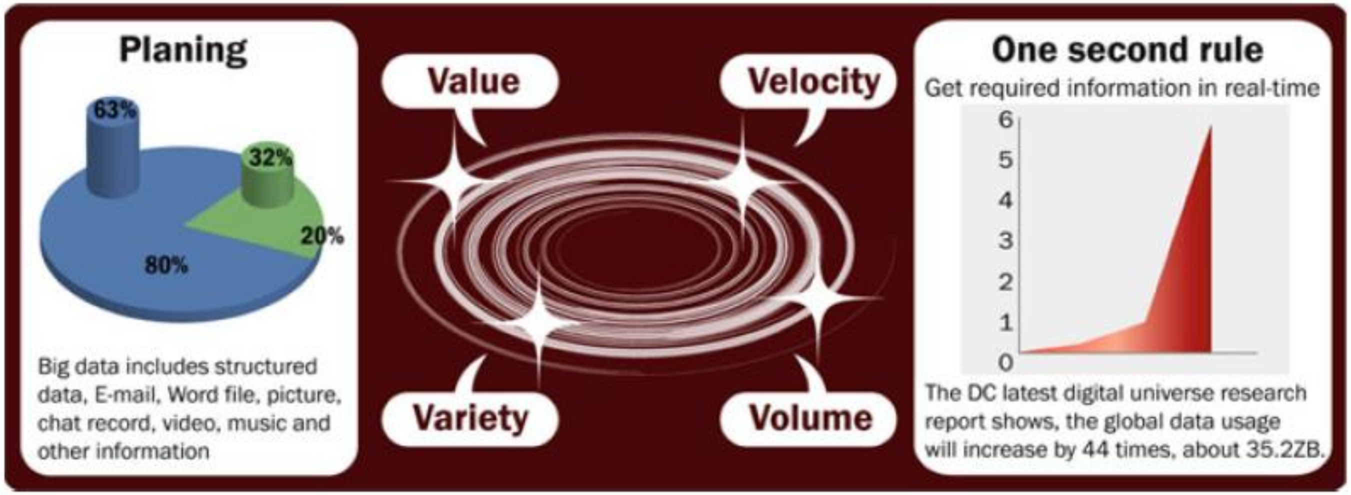
- First Big Data definition



What is Big Data?

- Value from Big Data? Is it possible?

- Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large **volumes** of a wide **variety** of data, by enabling the high-**velocity** capture, discovery, and/or analysis



Big data includes structured data, E-mail, Word file, picture, chat record, video, music and other information

Get required information in real-time

The DC latest digital universe research report shows, the global data usage will increase by 44 times, about 35.2ZB.

Discover value from datasets

What is Big Data?

- **To generate value, the high volume**, **velocity,** and **variety** of data must be processed with **advanced tools** (analytics and algorithms) to reveal **meaningful information**.

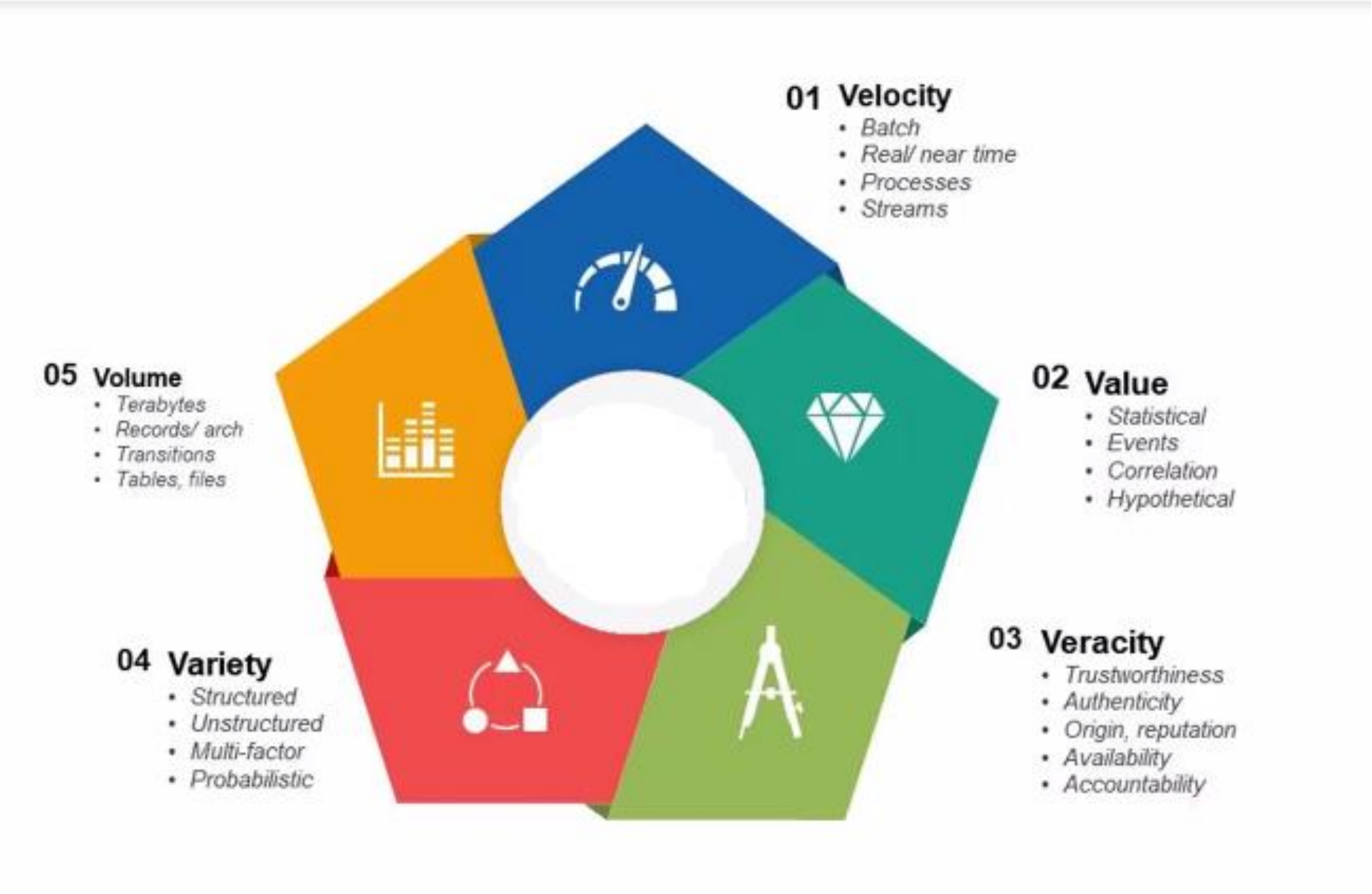
- **Veracity** refers to the quality of the data that is being analysed.

- High veracity data has many records that are valuable to analyse and that contribute in a meaningful way to the overall results.

- Low veracity data, on the other hand, contains a high percentage of meaningless data.

- The non-valuable data is noise.

What is Big Data?



Big Data 5V’s

- **Volume**: amount of data
- **Velocity**: data generation and data processing
- **Variety**: multiple nature

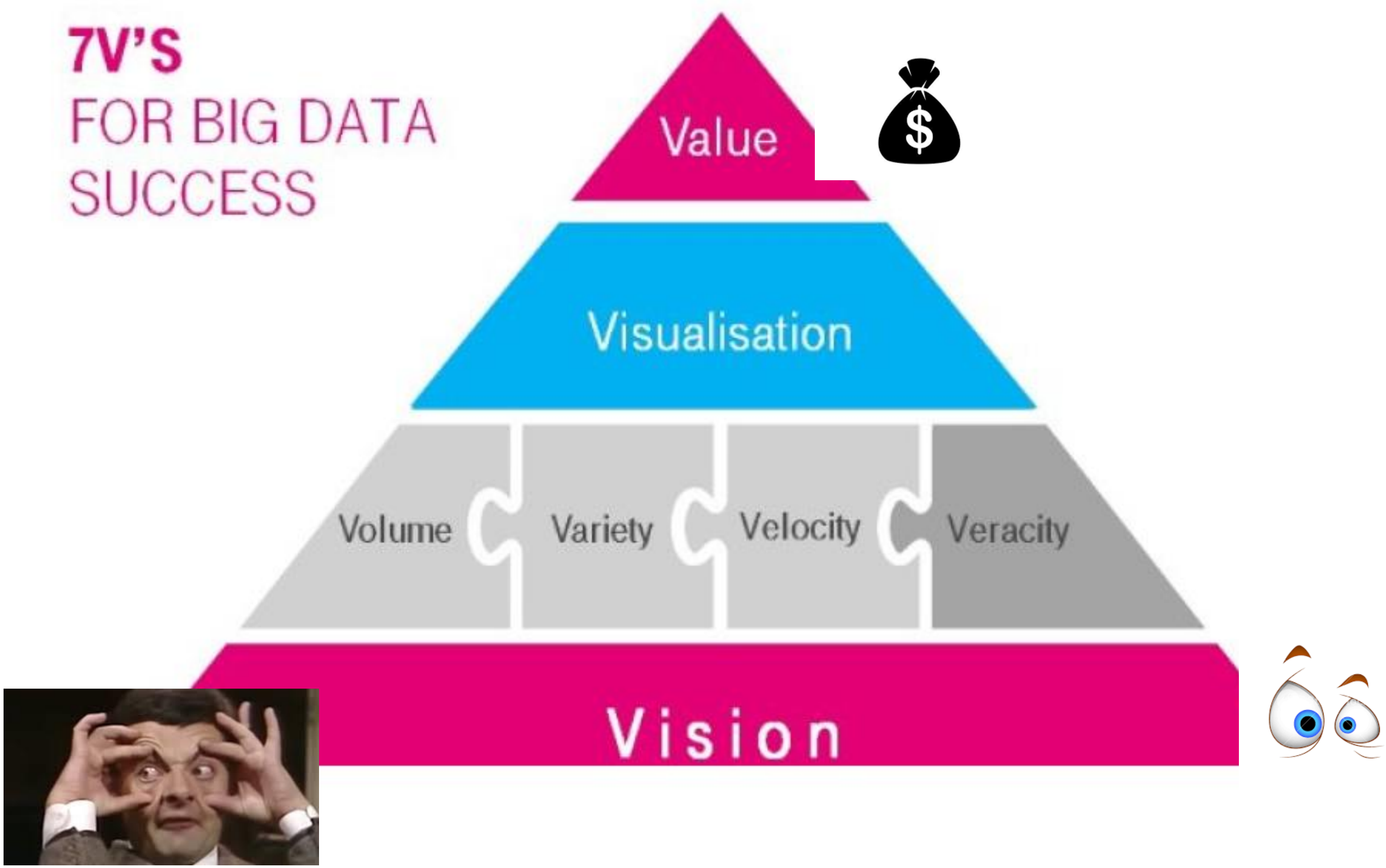
Academic Vision

- **Veracity**: data source reliability
- **Value**: its potential to generate value.

Industrial Vision

In summary: Big Data is “Vig”!

Big Data 7V's


7V'S
FOR BIG DATA
SUCCESS
Value
$
Visualisation
Volume
Variety
Velocity
Veracity
Vision

Big Data Motivations

- More CPUs to run programs faster
- More disks:
  - Modern applications require huge amounts of data
  - Many disks allow to perform I/O in parallel

- **Example**:
  - Assume that we have a single disk with 500 TB capacity.
  - This is enough to store more than 20 billion webpages (assuming an average size per page of 20KB).
  - However, to scan these 500 TB we need more than 4 months if the disk can bring 40 MB/sec.

Imagine the time required to process the data !

Big Data Challenges

- Big data era brings multiple **challenges:**
- **Storage**:
  - Traditional data management and analytics systems are based on the relational database management system (RDBMS)
  - RDBMSs only apply to **structured data**
  - RDBMSs are increasingly utilizing more and more **expensive hardware**
  - RDBMSs **cannot handle** the huge volume and heterogeneity of big data

Big Data Challenges

- **Data representation**: Data representation aims to make data more meaningful for computer analysis and user interpretation. Improper data representation will reduce the value of the original data and may even obstruct effective data analysis.

- **Data Life Cycle Management**: Pervasive sensors and computing are generating data at unprecedented rates and scales which the current storage system does not support. We must decide which data shall be stored and which data shall be discarded.

- **Analytical Mechanism:** the analytical system of big data shall process masses of heterogeneous data within a limited time. Traditional RDBMSs are strictly designed with a lack of **scalability** and **expandability**.

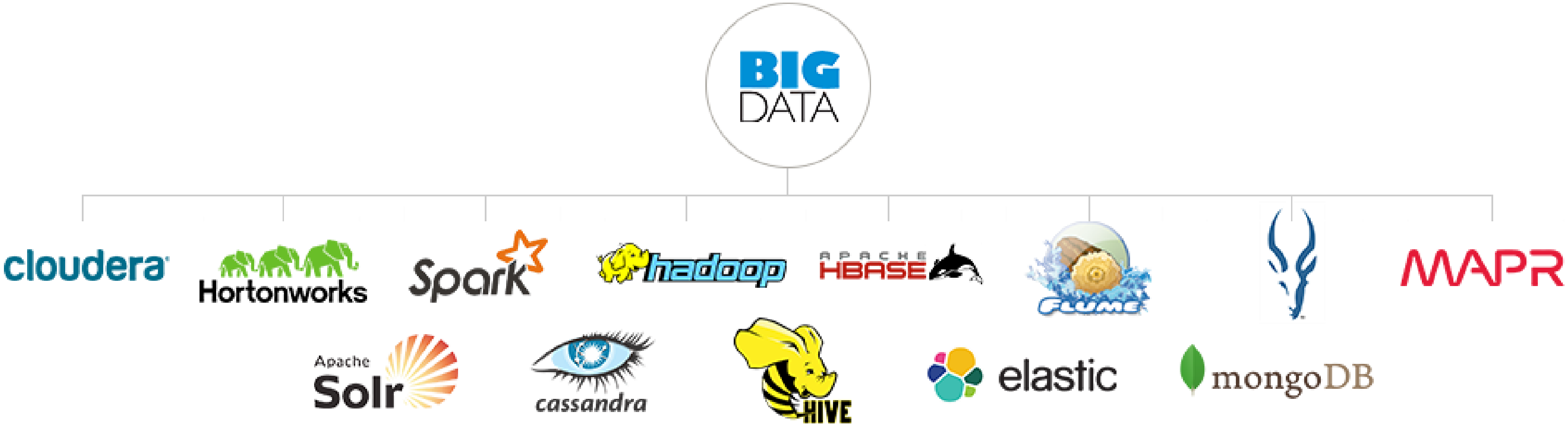
Big Data Challenges

- **Data Confidentiality:** analysis of big data challenges privacy
- **Energy Management:** the energy consumption of mainframe computing systems has drawn much attention from both economy and environment perspectives. Processing, storage, and transmission of big data will inevitably consume more and more electric energy
- **Expendability and Scalability**: analytical algorithm must be able to process increasingly expanding and more complex datasets
- **Cooperation**: analysis of big data is an interdisciplinary research

Big Data Challenges - Summary

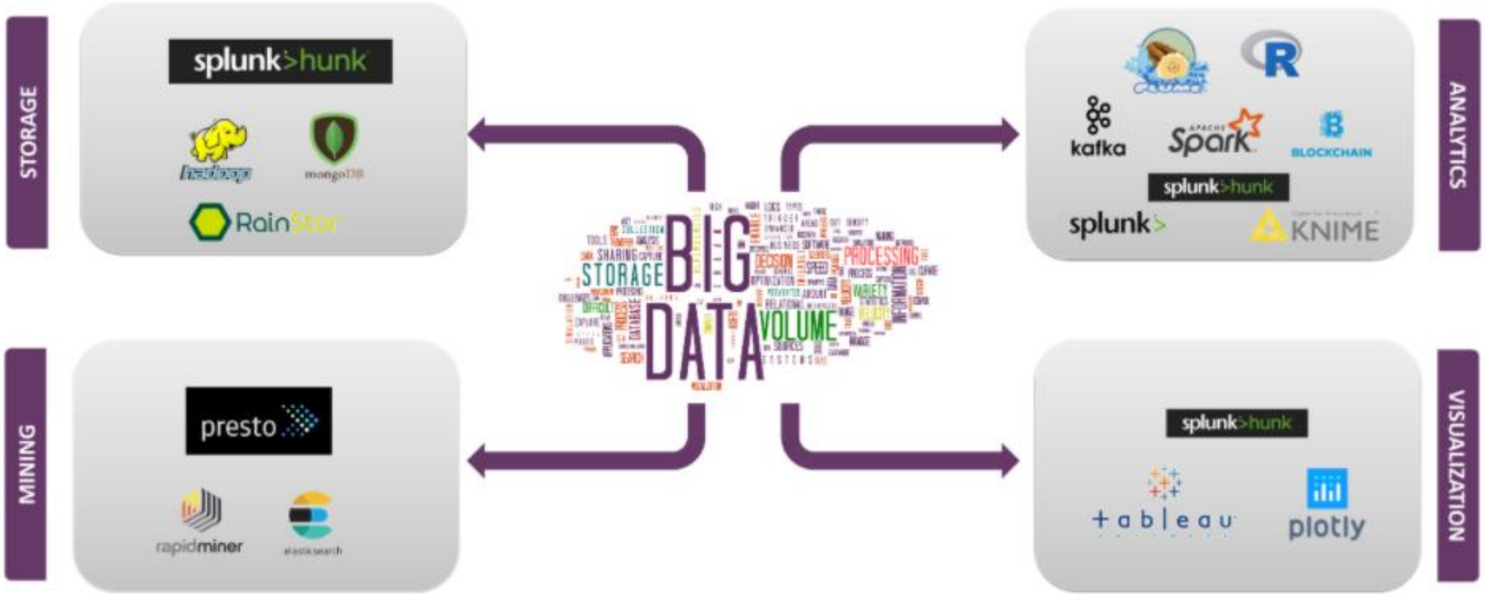
- **Scalability**: property of a system to handle a growing amount of work by adding resources to the system
- **Load balancing:** process of distributing a set of tasks over a set of resources
- **Fault tolerance:** property of a system to continue operating properly in the case of failure of some of its components
- **Efficiency:** system performance
- **Data stream processing:** real-time systems

Big Data Solutions



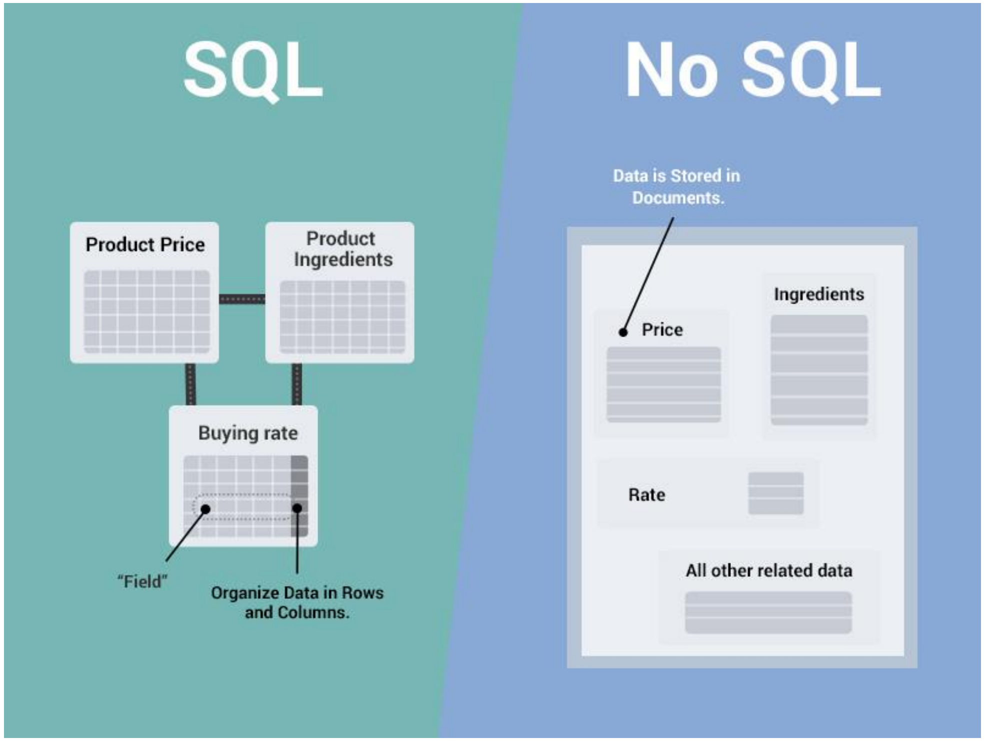
Big Data Technologies

- Data storage
- Data processing and analytics
- Data visualisation



Big Data Technologies

- **Data storage**:
- NoSQL data bases can **scale horizontally** over **many commodity servers** with high performance
- Non-relational data model and no requirements for schemas



MongoDB is a NoSQL data base which offers a direct alternative to Relational Databases. It offers flexibility to handle a wide variety of datatypes at large volumes and across distributed architectures.

Big Data Technologies

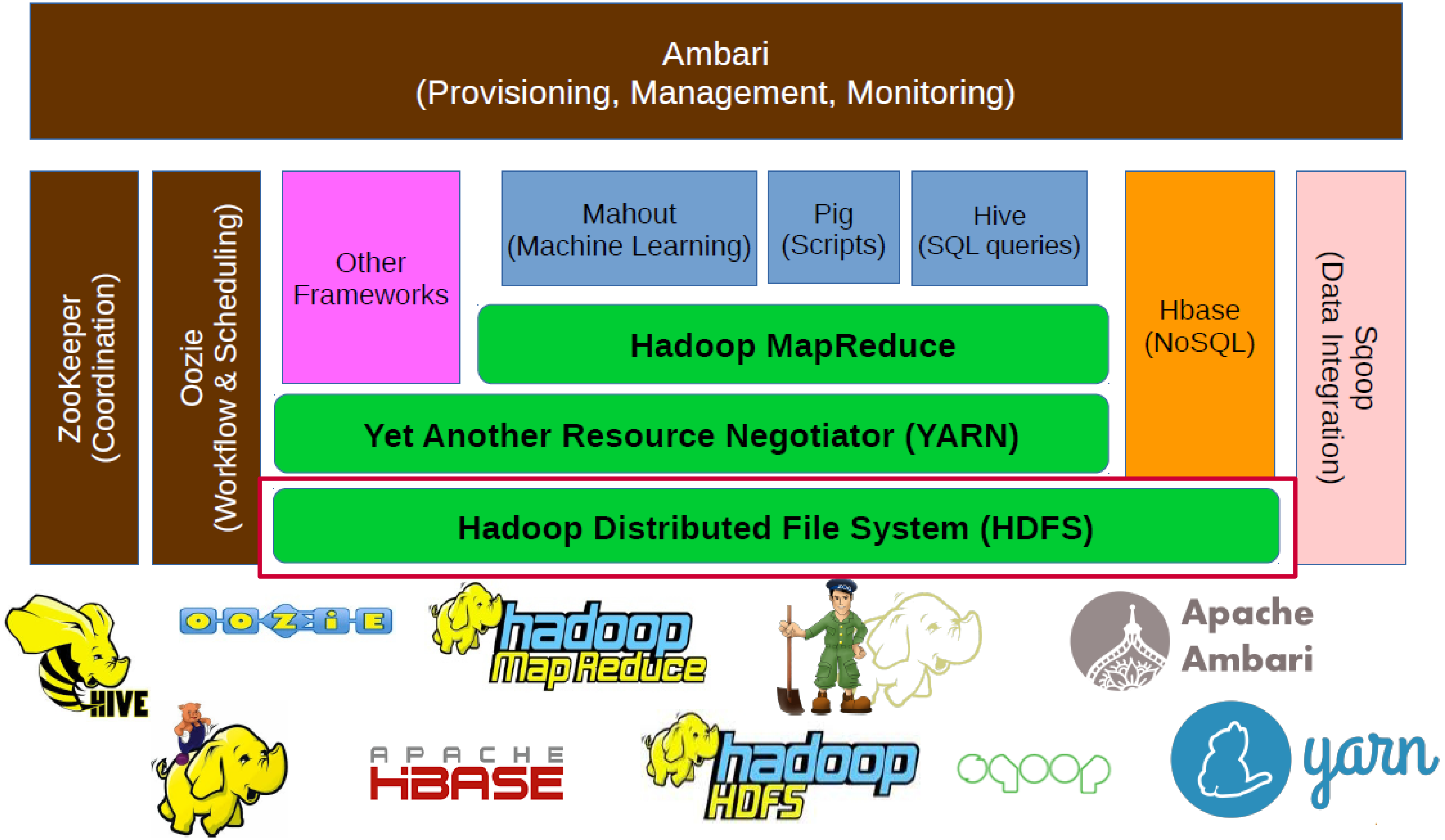
- **Data processing and analytics**

- **Hadoop:** A very successful platform to run jobs in massively parallel systems (thousands of processors and disks).

- **Hadoop** was designed to store and process data. It can store and analyse the data present in different machines with high speeds and low costs.

- It contains many different components:
  - the **Hadoop MapReduce layer**
  - the **YARN resource manager**
  - the **Hadoop Distributed File System** (HDFS)

Big Data Technologies



Ambari
(Provisioning, Management, Monitoring)

ZooKeeper
(Coordination)

Oozie
(Workflow & Scheduling)

Other Frameworks

Mahout
(Machine Learning)

Pig
(Scripts)

Hive
(SQL queries)

Hbase
(NoSQL)

Sqoop
(Data Integration)

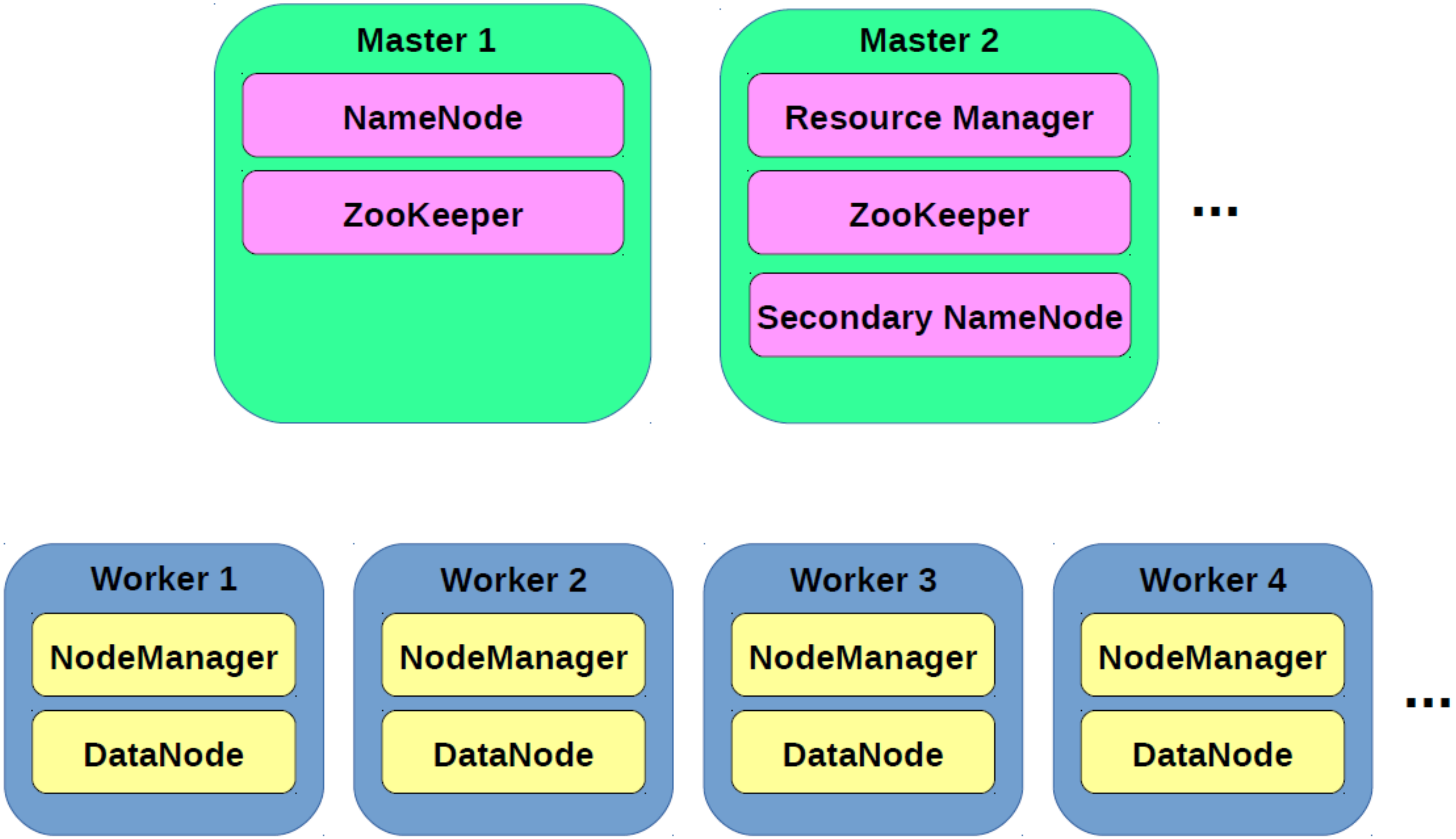
Hadoop MapReduce

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

Big Data Technologies

Hadoop Cluster Architecture



Big Data Technologies: HDFS

- HDFS stands for *Hadoop Distributed File System*

- It is a *File System* that lives across the nodes of a cluster. It stores files, each file has a filename and is in a specific directory

- It supports most of the operations supported by an ordinary File System

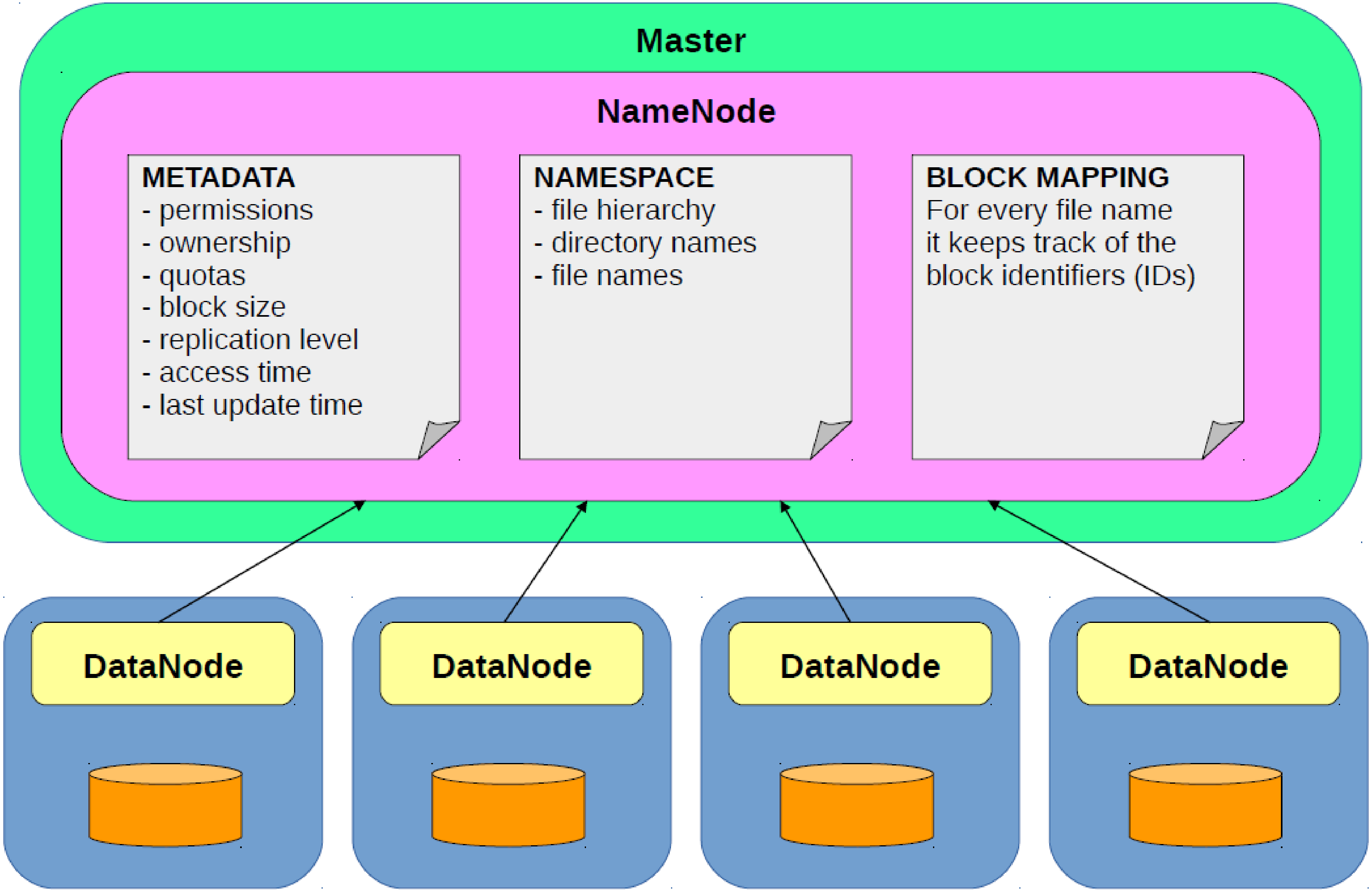
- Every HDFS cluster is comprised of:
  - one or two NameNodes and
  - many DataNodes

Big Data Technologies: HDFS

- **NameNode** (one or two per cluster)
  - Represents a single filesystem namespace
  - Is the **master service** of HDFS
  - **Determines and maintains how data is distributed** across the DataNodes
  - Actual data **never resides here**, only metadata (*e.g.*, maps of where blocks are distributed).

- **DataNode** (as many as you want per cluster)
  - Store data

Big Data Technologies: HDFS



Master

NameNode

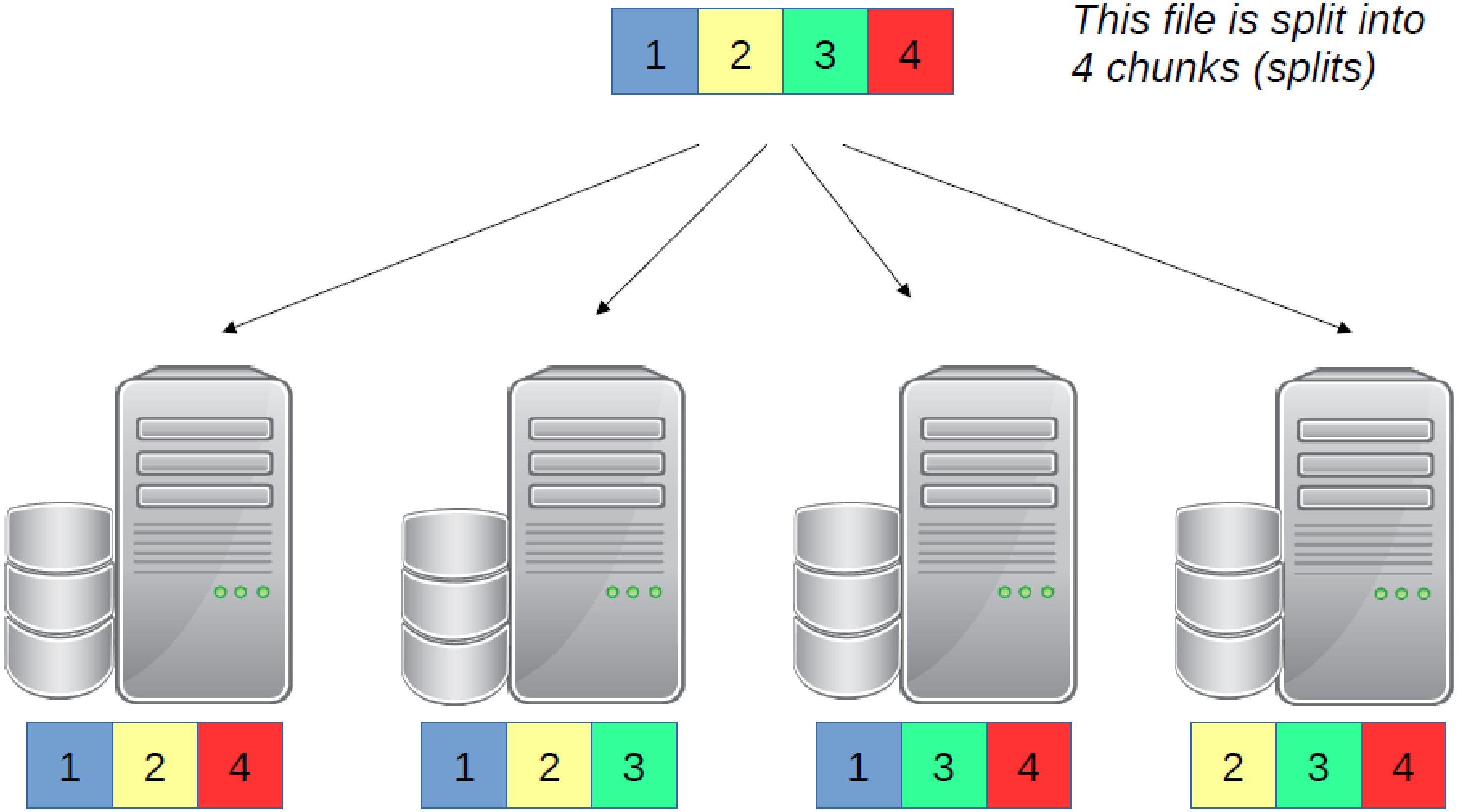
METADATA
- permissions
- ownership
- quotas
- block size
- replication level
- access time
- last update time

NAMESPACE
- file hierarchy
- directory names
- file names

BLOCK MAPPING
For every file name
it keeps track of the
block identifiers (IDs)

Big Data Technologies: HDFS

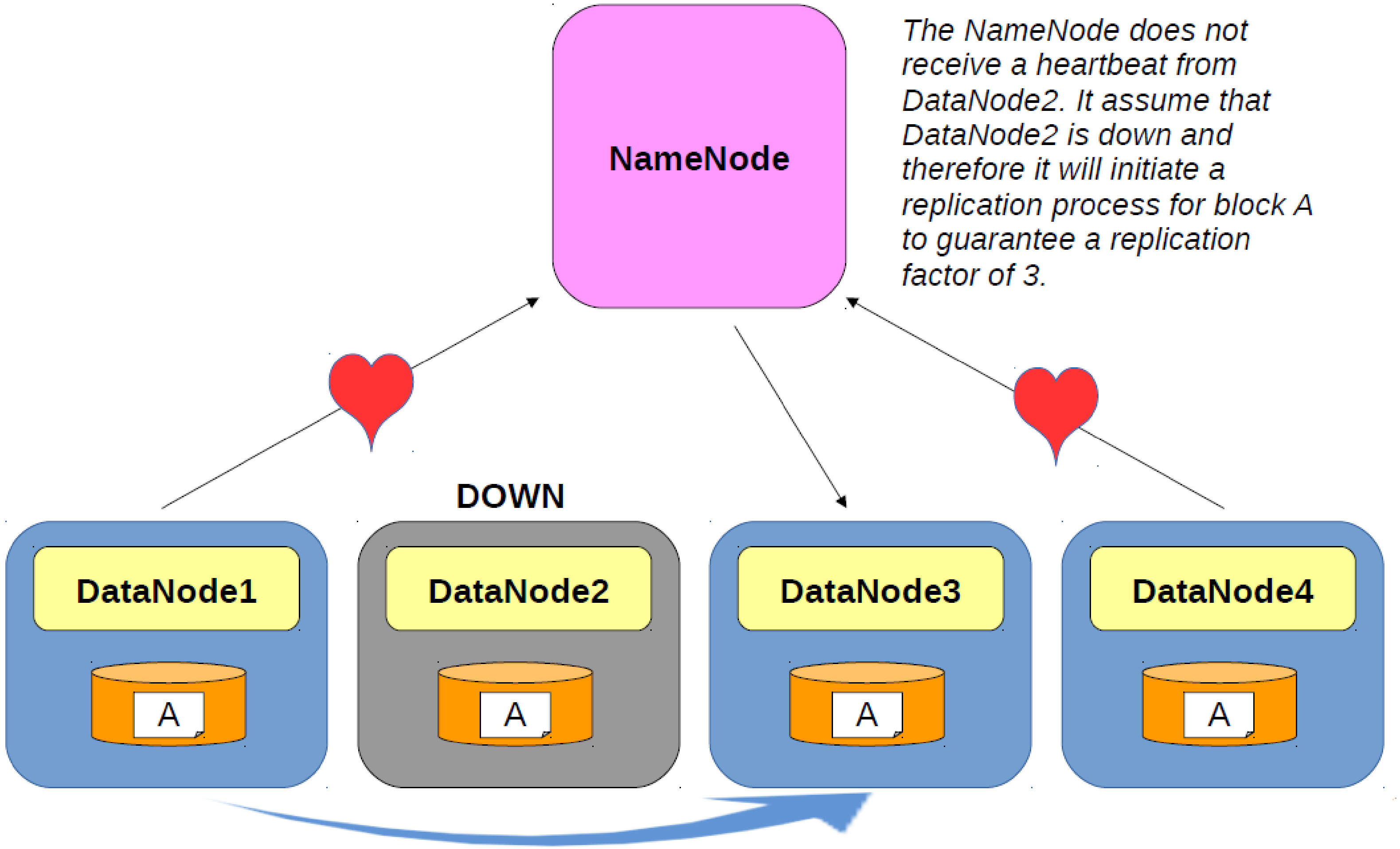
Replication in HDFS



This file is split into 4 chunks (splits)

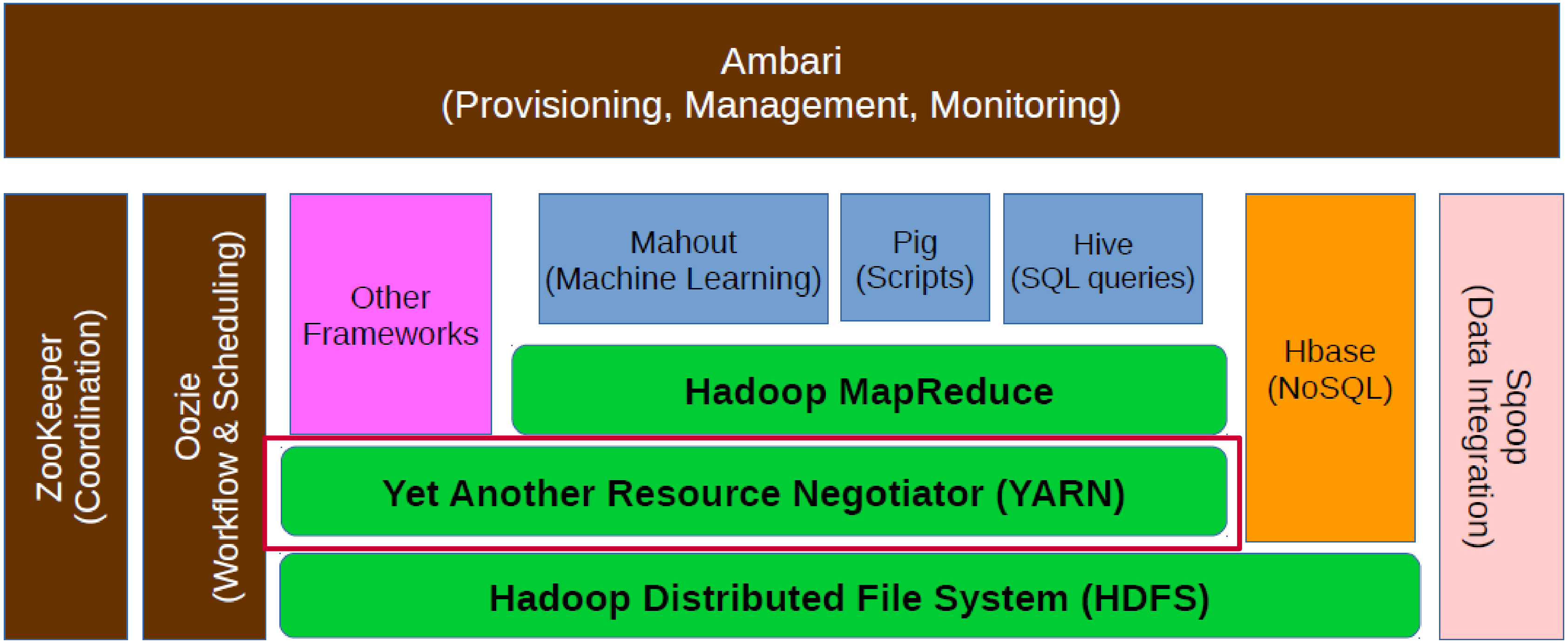
The the file is split in chunks. Each is replicated three times in this example. One of the chunk is located at a different rack for increased fault tolerance.

Big Data Technologies: HDFS



The NameNode does not receive a heartbeat from DataNode2. It assume that DataNode2 is down and therefore it will initiate a replication process for block A to guarantee a replication factor of 3.

Big Data Technologies



Ambari
(Provisioning, Management, Monitoring)

ZooKeeper
(Coordination)

Oozie
(Workflow & Scheduling)

Other Frameworks

Mahout
(Machine Learning)

Pig
(Scripts)

Hive
(SQL queries)

Hadoop MapReduce

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

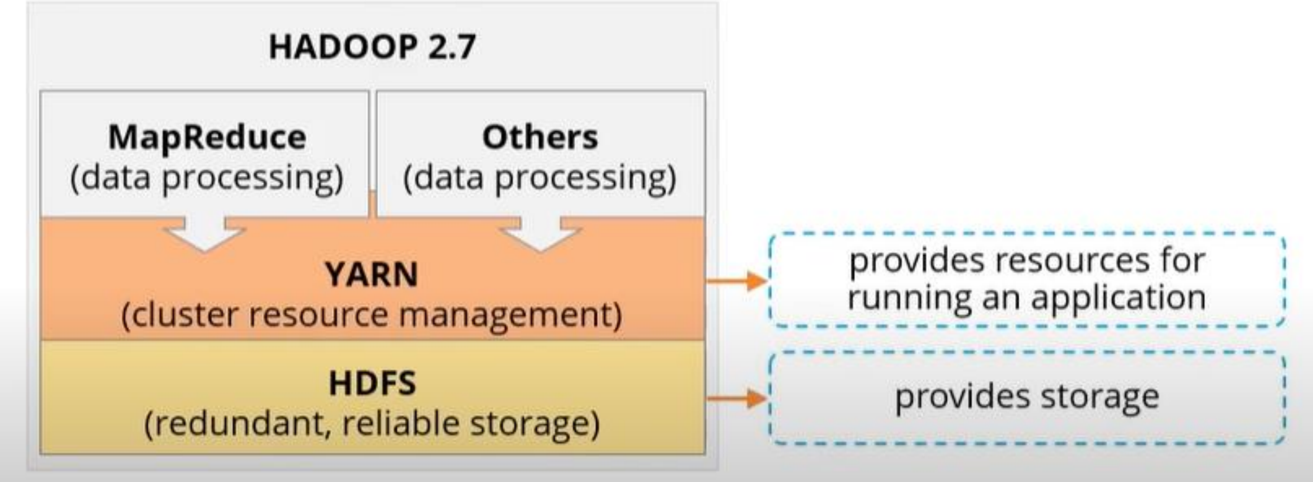
Hbase
(NoSQL)

Sqoop
(Data Integration)

Big Data Technologies: YARN

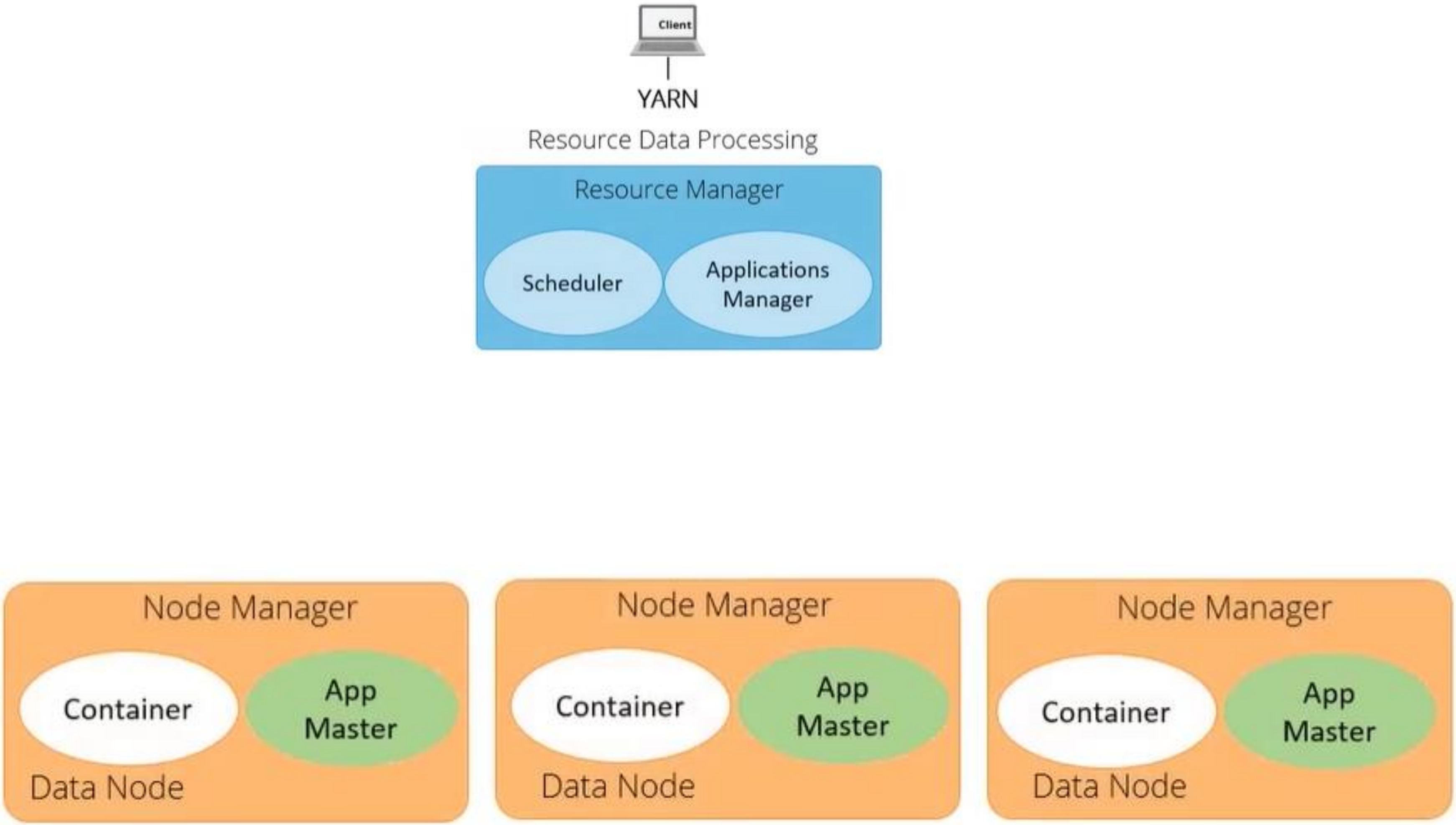
- **Apache Yarn** – “Yet Another Resource Negotiator” : A framework to provide computational resources for execution engines

- **Resource Manager** of Hadoop to improve performance

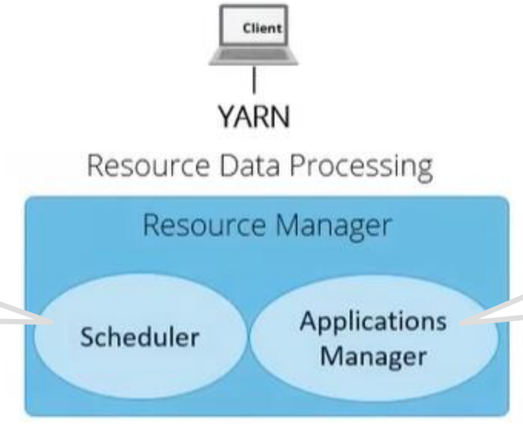

HADOOP 2.7
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)
provides resources for running an application
provides storage

- The three important elements of Yarn architecture is the **Resource Manager**, **Application Manager**, and **NodeManager**

Big Data Technologies: YARN



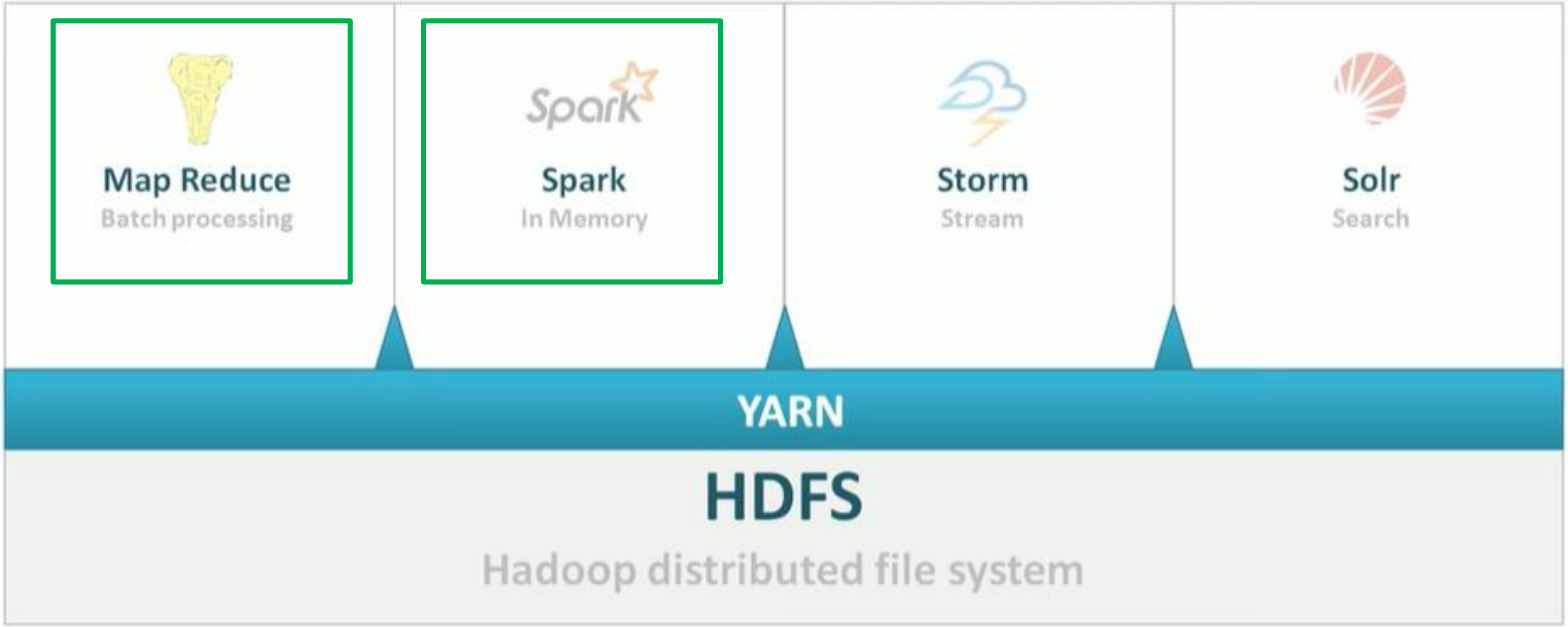
Big Data Technologies: YARN



Responsible for allocating resources to various running applications

Interface which maintains a list of applications that have been submitted, currently running, or completed

Big Data Technologies



Big Data Technologies: MapReduce

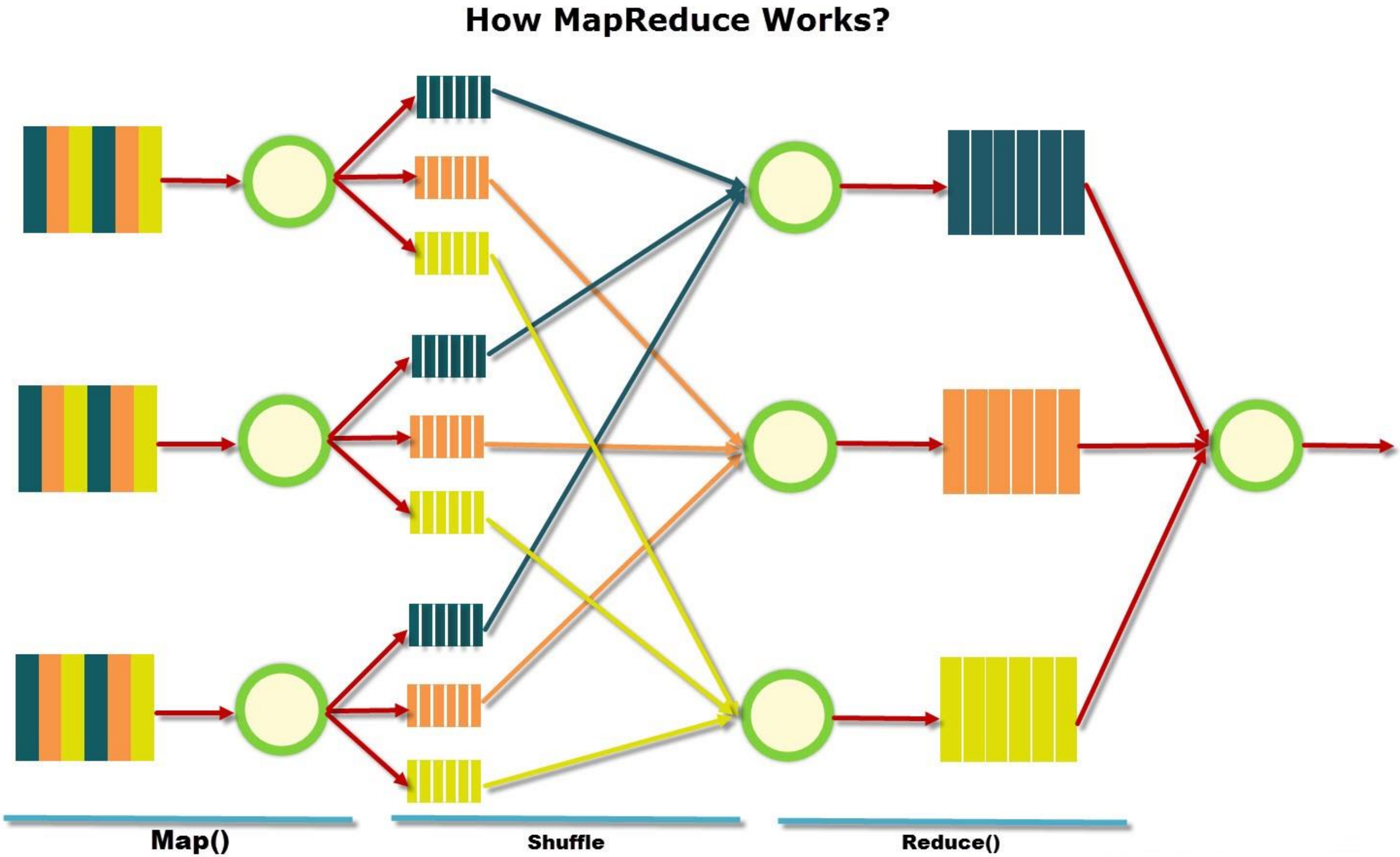
- Based on key-value pairs

- Each job is composed of one or more MR stages

- Each MR stage comprises:
  - the map phase
  - the shuffle-and-sort phase
  - the reduce phase

- The programmer **focuses on the problem**. Replication, fault tolerance, scheduling, re-scheduling and other low level procedures are handled by HDFS or YARN.

Big Data Technologies: MapReduce



Big Data Technologies: MapReduce

- The programmer must implement the following functions:
  - **map()**: accepts a set of key-value pairs and generates another list of key-value pairs.
  - **combine()**: performs an aggregation before sending the data to reducers (reduces network traffic).
  - **partition()**: uses a hash function to distribute data to reducers (load balancing, avoids hotspots).
  - **reduce()**: accepts a key and a list of values for this specific key and performs an aggregation.

- Note: combine() and partition() are optional

Big Data Technologies: MapReduce

- Let us analyse an example:

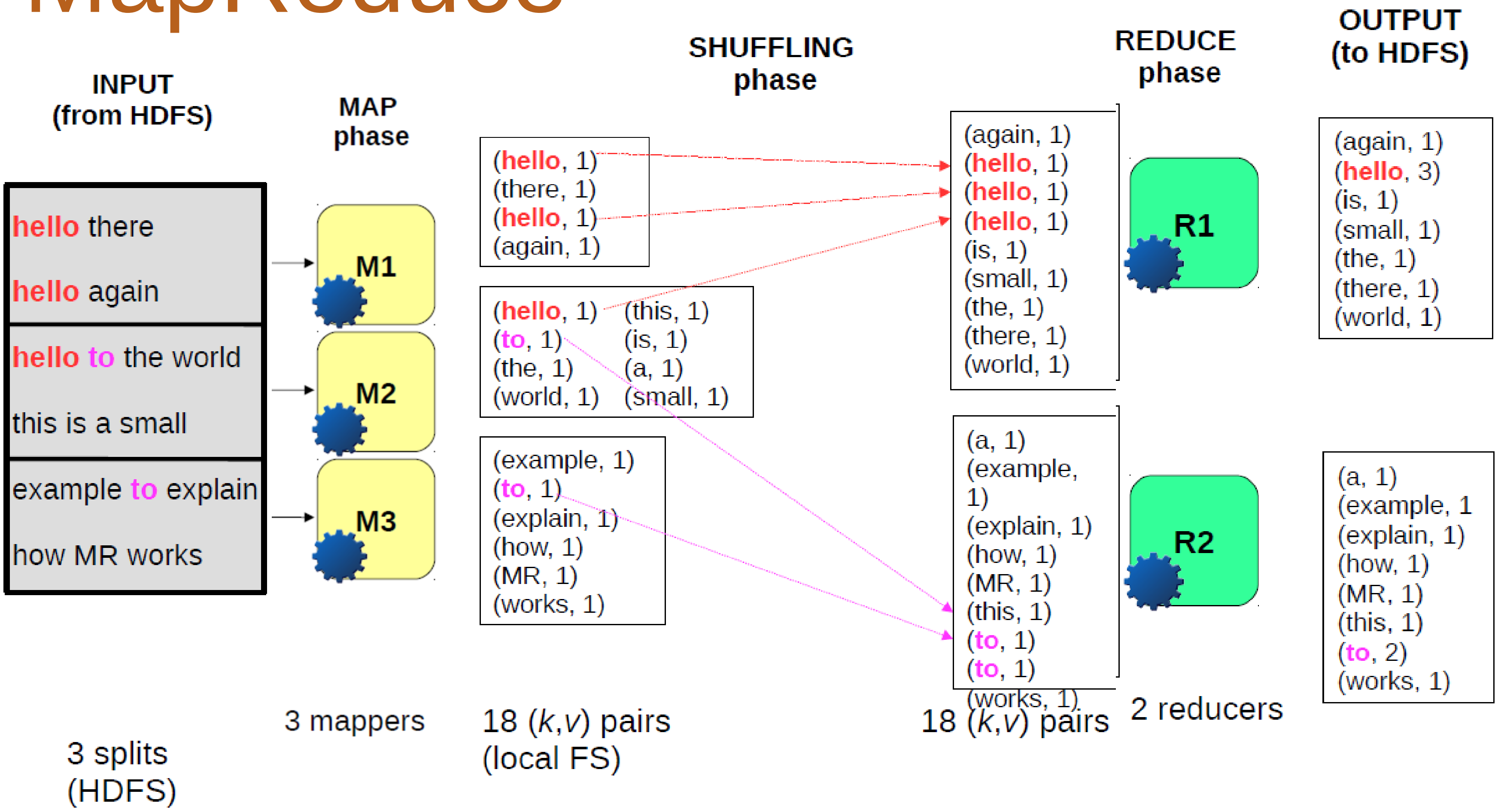
- Given a potentially massive txt file, compute the number of occurrences of every word. For every word, output a pair - (word, #occurrences)

- The number of pairs in the output equals the number of unique words in the file

e.g.,

- **input**: can you see the real me? can you? can you?
- **output**: (can,3), (you,3), (see,1), (the,1), (real,1), (me,1)

Big Data Technologies: MapReduce



INPUT (from HDFS)

hello there

hello again

hello to the world

this is a small

example to explain

how MR works

3 splits (HDFS)

MAP phase

M1

M2

M3

3 mappers

(hello, 1)
(there, 1)
(hello, 1)
(again, 1)

(hello, 1)
(to, 1)
(the, 1)
(world, 1)

(this, 1)
(is, 1)
(a, 1)
(small, 1)

(example, 1)
(to, 1)
(explain, 1)
(how, 1)
(MR, 1)
(works, 1)

18 (*k*,*v*) pairs (local FS)

SHUFFLING phase

(again, 1)
(hello, 1)
(hello, 1)
(hello, 1)
(is, 1)
(small, 1)
(the, 1)
(there, 1)
(world, 1)

(a, 1)
(example, 1)
(explain, 1)
(how, 1)
(MR, 1)
(this, 1)
(to, 1)
(to, 1)
(works, 1)

18 (*k*,*v*) pairs

REDUCE phase

R1

R2

2 reducers

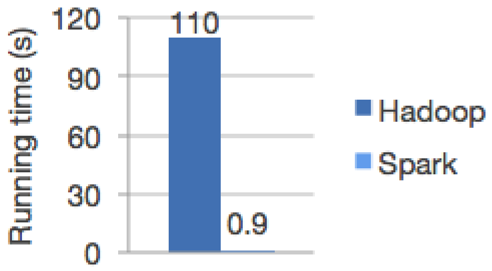
OUTPUT (to HDFS)

(again, 1)
(hello, 3)
(is, 1)
(small, 1)
(the, 1)
(there, 1)
(world, 1)

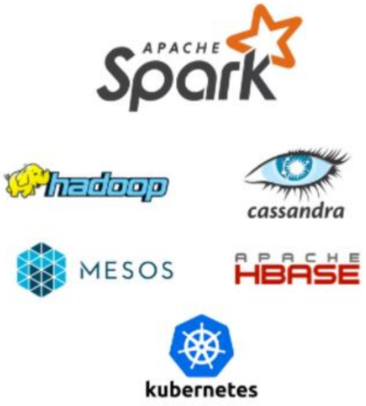
(a, 1)
(example, 1
(explain, 1)
(how, 1)
(MR, 1)
(this, 1)
(to, 2)
(works, 1)

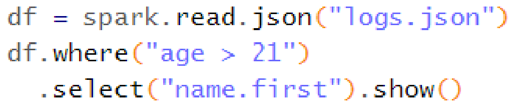
Big Data Technologies: Spark

- Apache Spark is a unified analytics engine for large-scale data processing.
- Run workload 100x faster



- Ease of use
- Generality – combines SQL, streaming and complex analytics
- Runs everywhere

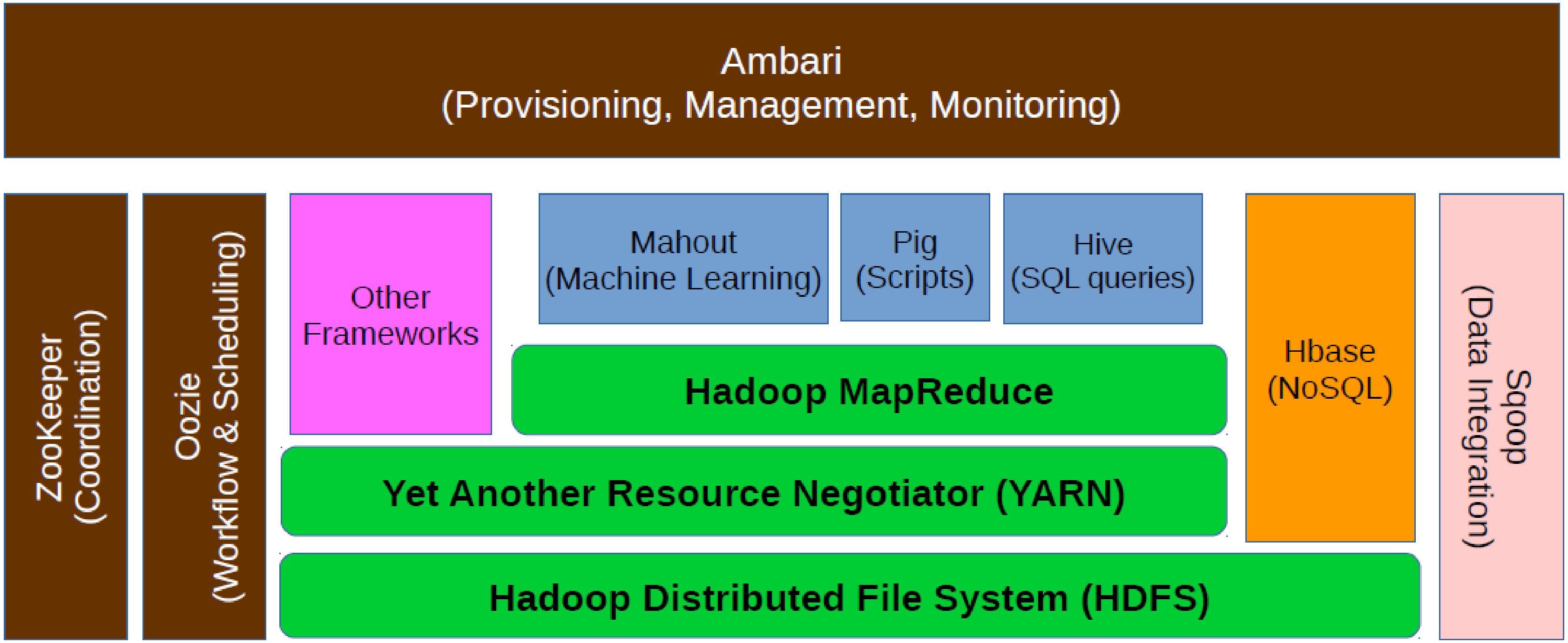




```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Big Data Technologies



Ambari
(Provisioning, Management, Monitoring)

ZooKeeper
(Coordination)

Oozie
(Workflow & Scheduling)

Other
Frameworks

Mahout
(Machine Learning)

Pig
(Scripts)

Hive
(SQL queries)

Hadoop MapReduce

Yet Another Resource Negotiator (YARN)

Hbase
(NoSQL)

Hadoop Distributed File System (HDFS)

Sqoop
(Data Integration)

Big Data Technologies: Cloud Computing

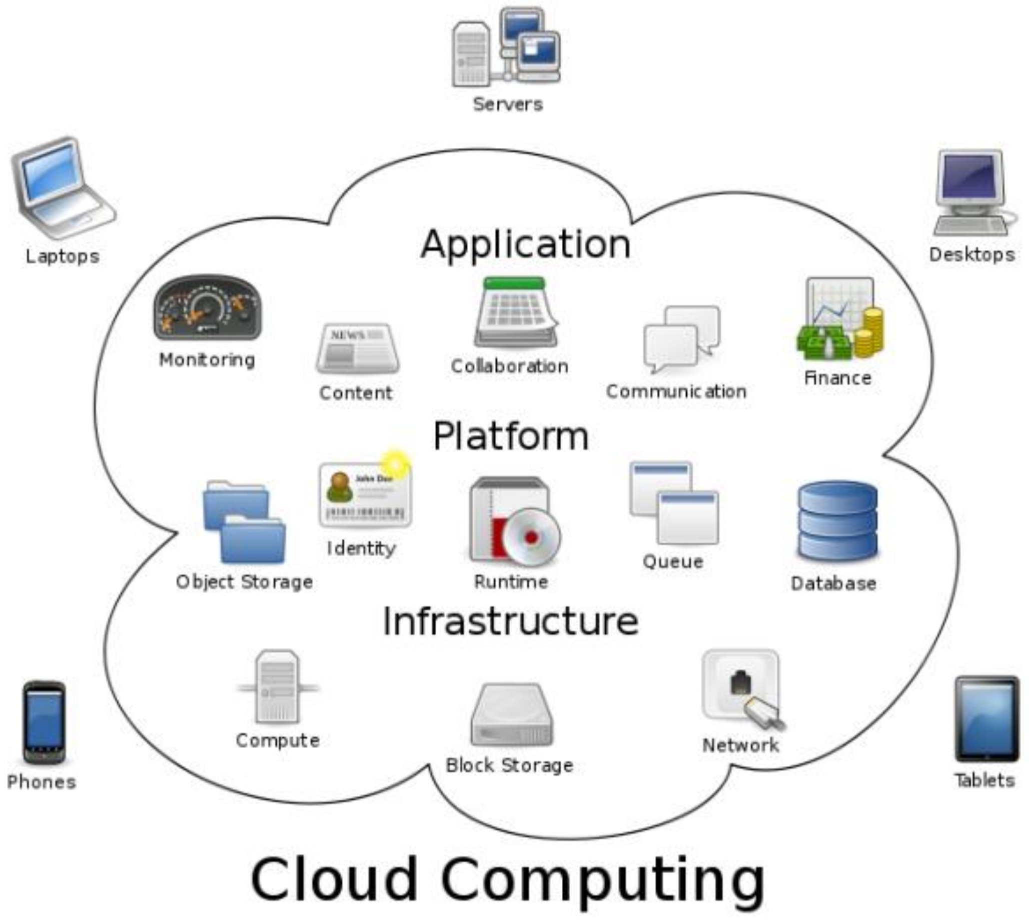
- **Cloud Computing**: computing resources and computing capacities which provide solutions for storage and processing big data

- IT **resources provided as a service**
  - Compute, storage, databases, queues, etc.

- Clouds leverage economies of **scale** of commodity **hardware**
  - Cheap storage, high bandwidth networks & multicore processors
  - Geographically distributed data centres

- Offerings from Microsoft, Amazon, Google, …

Big Data Solutions

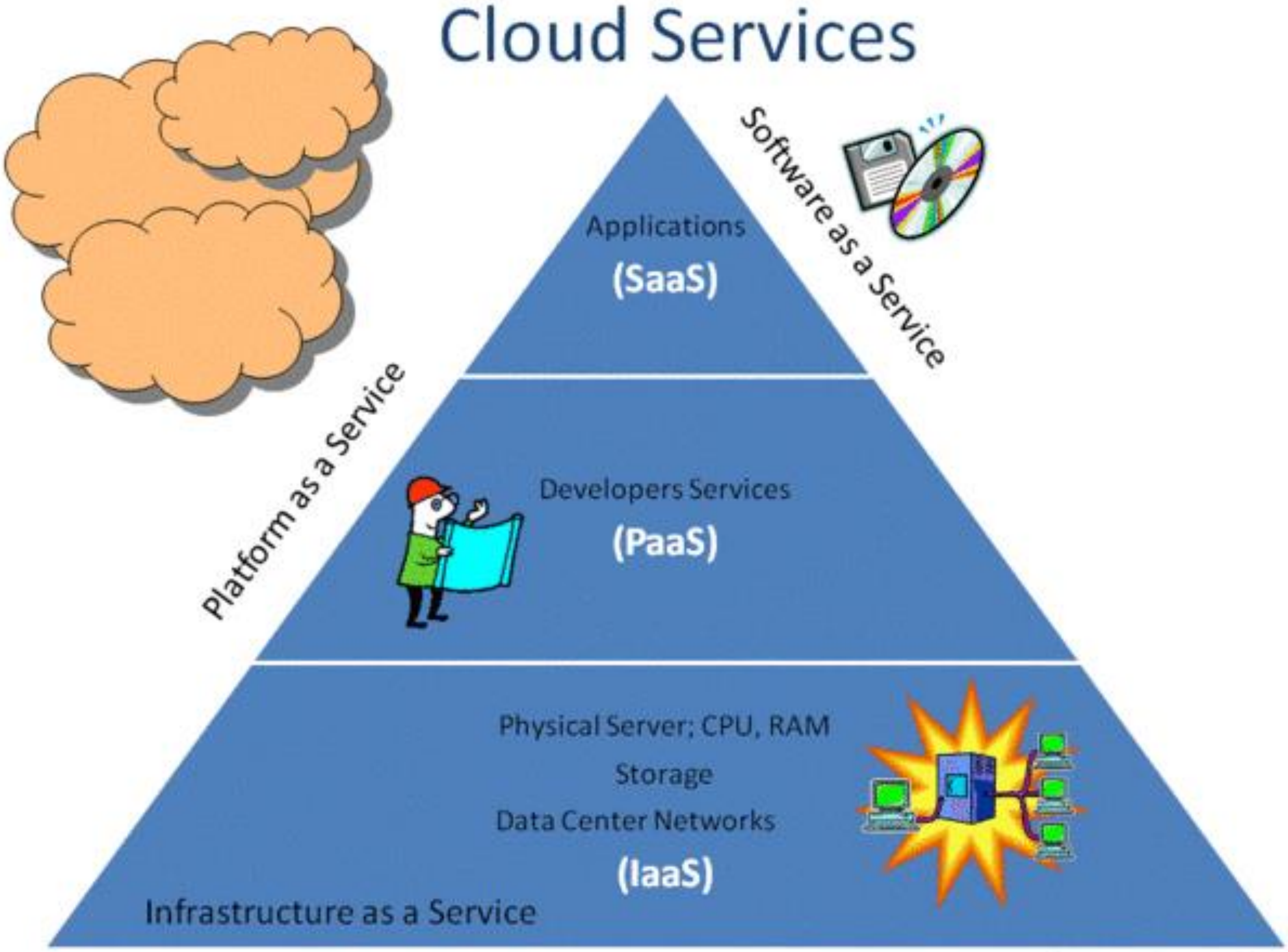


Big Data Solutions: Cloud Computing

- **Infrastructure as a service** (IaaS)
  - Offering hardware related services using the principles of cloud computing. These could include storage services (database or disk storage) or virtual servers.
  - Amazon EC2, openstack.

- **Platform as a Service** (PaaS)
  - Offering a development platform on the cloud.
  - Adds a middleware like data bases into the cloud environment

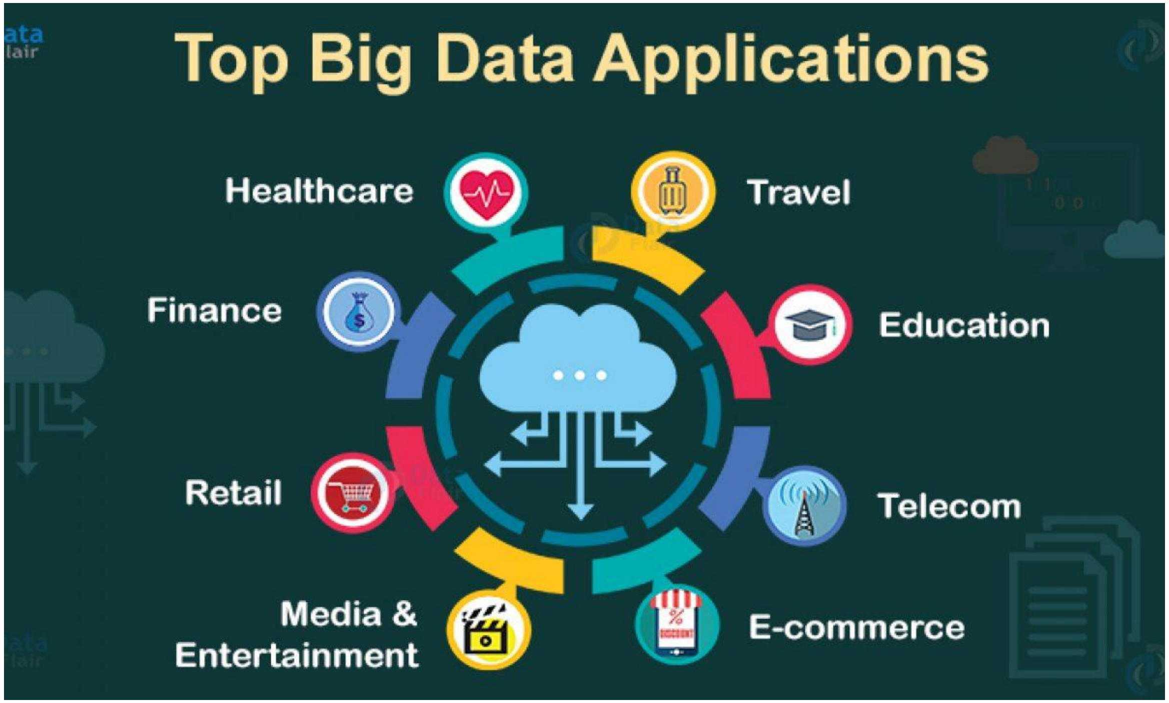
- **Software as a service** (SaaS)
  - Including a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector.
  - Dropbox, slack, Adobe creative cloud



CC-by-SA 3.0 by SD Habeger

Big Data Applications

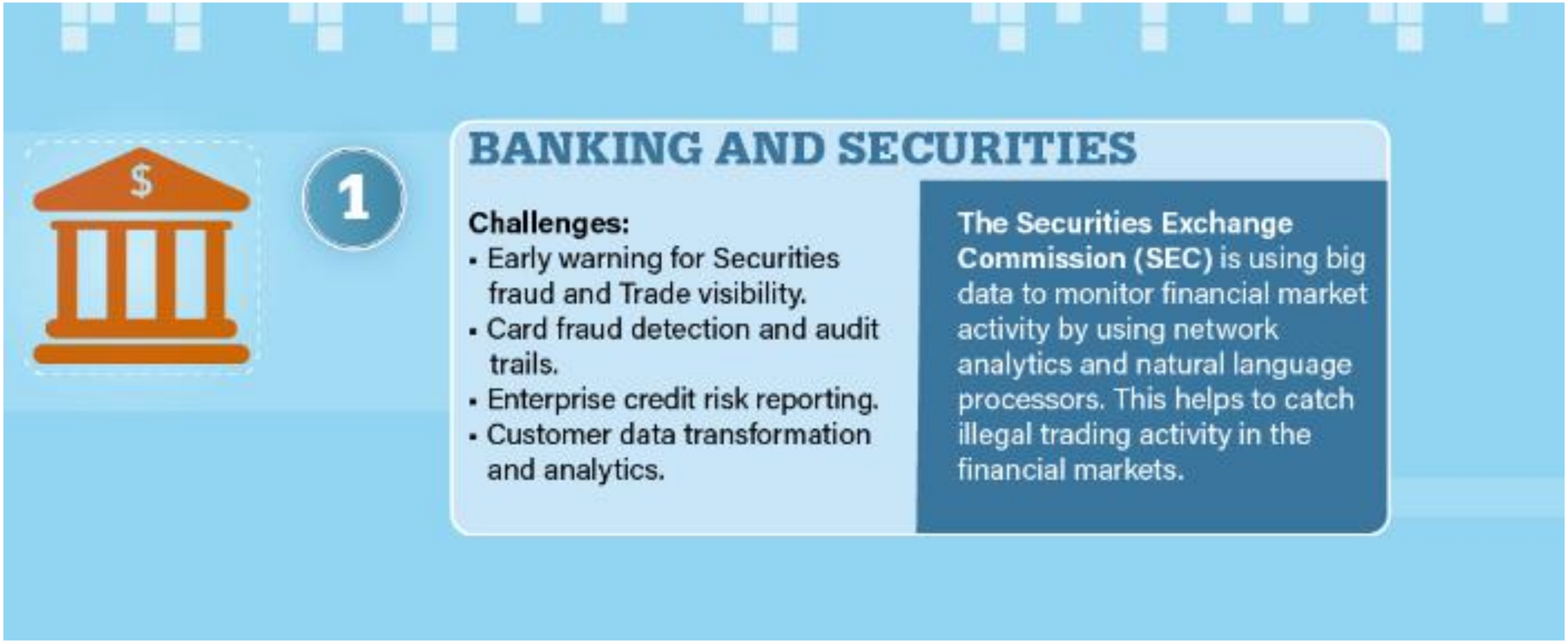
- Bank and security
- Communications
- Healthcare
- Education
- Industry
- Energy



Big Data Applications: Teamwork

- **Search for 2/3 intelligent applications for each domain:**
  - Travel
  - Education
  - Healthcare
  - Telecommunications
  - E-commerce
  - Retail
  - Finance
  - Media & Entertainment
  - Agriculture
  - Transportation
  - Manufacturing (industry processes)
- **Discover the main challenges of each domain**
- **Present the results to the class**

Big Data Applications



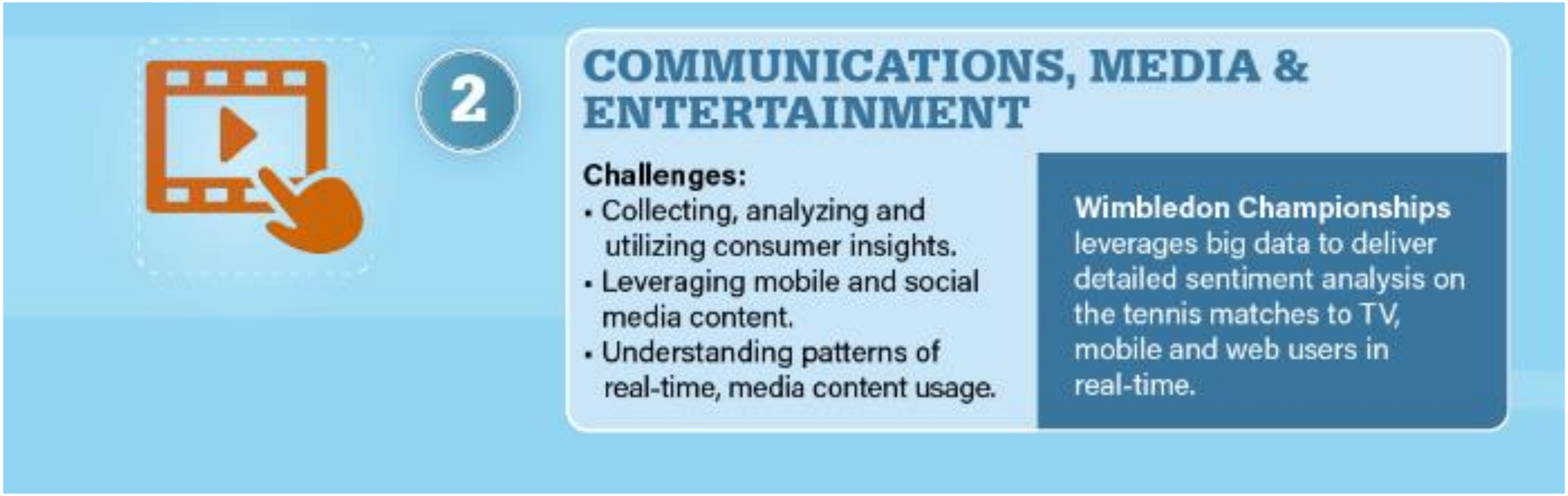
BANKING AND SECURITIES

Challenges:

- Early warning for Securities fraud and Trade visibility.
- Card fraud detection and audit trails.
- Enterprise credit risk reporting.
- Customer data transformation and analytics.

The Securities Exchange Commission (SEC) is using big data to monitor financial market activity by using network analytics and natural language processors. This helps to catch illegal trading activity in the financial markets.

Big Data Applications



2

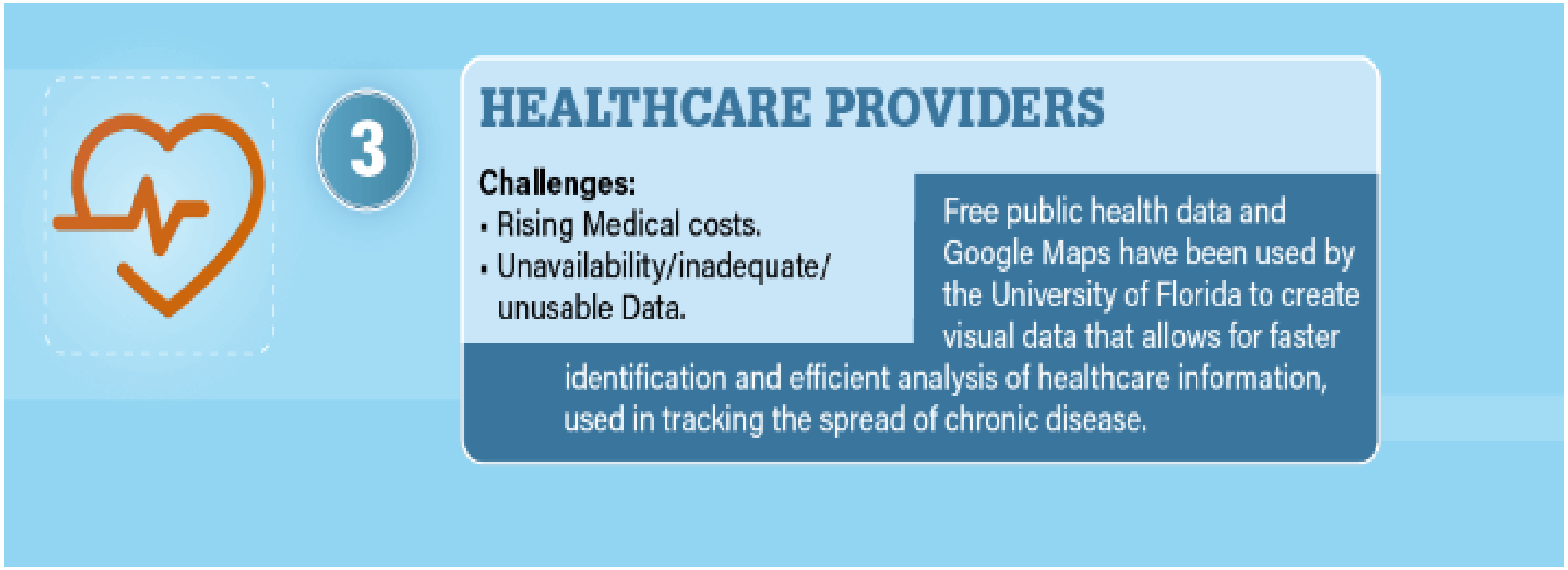
COMMUNICATIONS, MEDIA & ENTERTAINMENT

Challenges:

- Collecting, analyzing and utilizing consumer insights.
- Leveraging mobile and social media content.
- Understanding patterns of real-time, media content usage.

Wimbledon Championships leverages big data to deliver detailed sentiment analysis on the tennis matches to TV, mobile and web users in real-time.

Big Data Applications



3

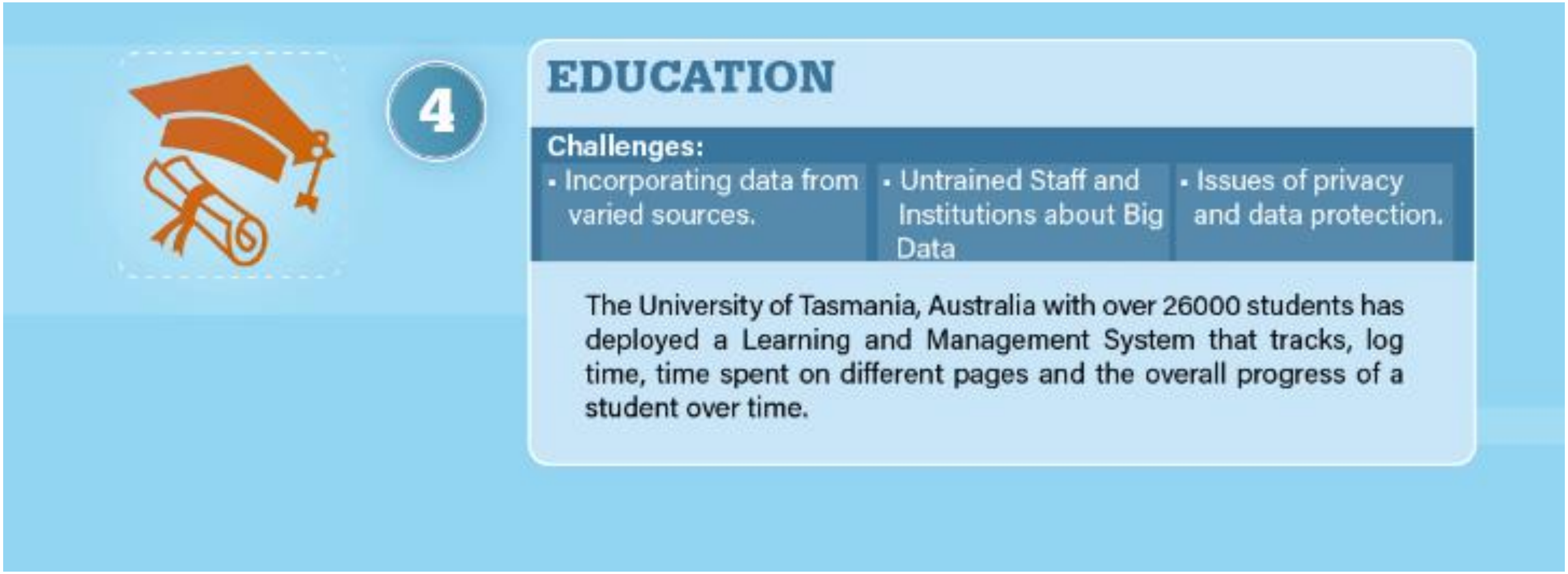
HEALTHCARE PROVIDERS

Challenges:

- Rising Medical costs.
- Unavailability/inadequate/ unusable Data.

Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

Big Data Applications



4

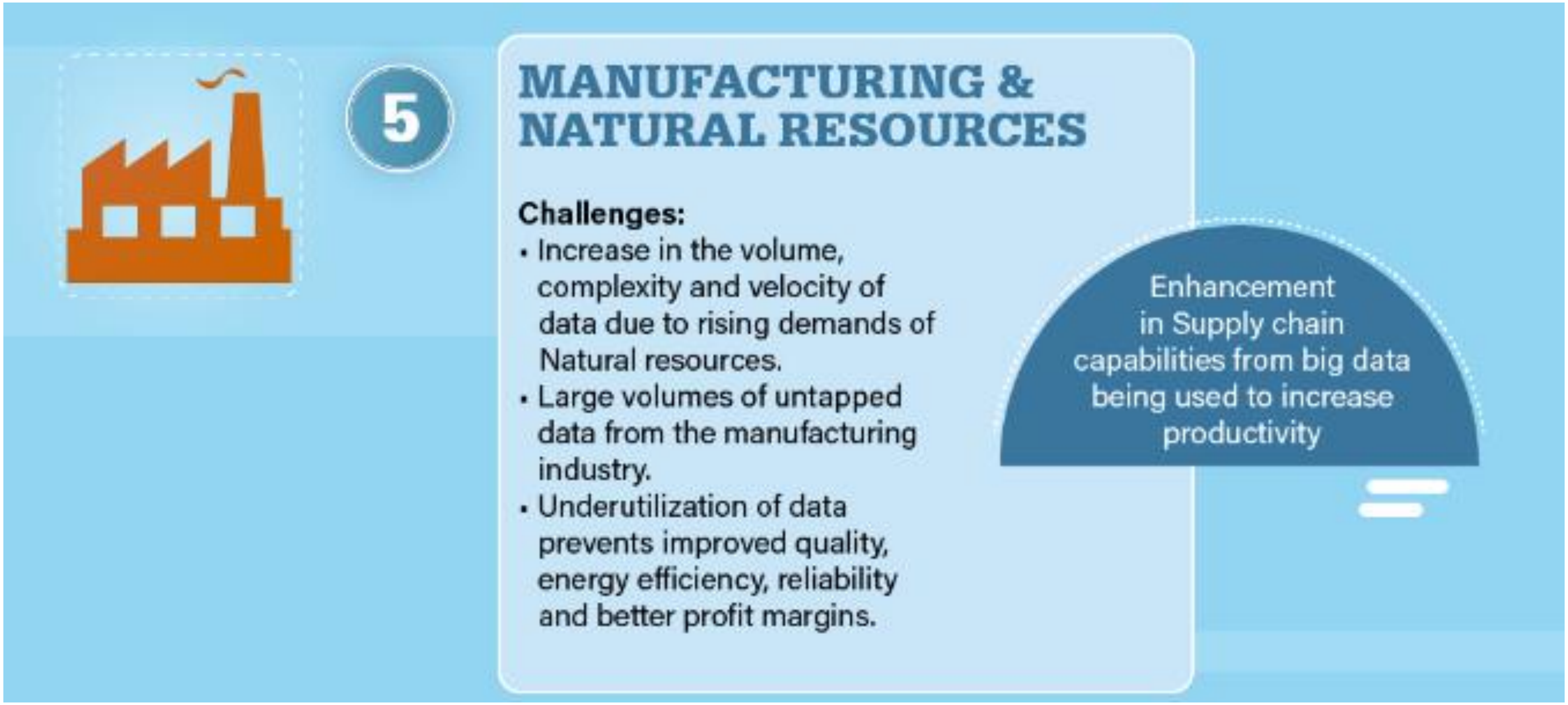
EDUCATION

Challenges:

- Incorporating data from varied sources.
- Untrained Staff and Institutions about Big Data
- Issues of privacy and data protection.

The University of Tasmania, Australia with over 26000 students has deployed a Learning and Management System that tracks, log time, time spent on different pages and the overall progress of a student over time.

Big Data Applications



5

MANUFACTURING & NATURAL RESOURCES

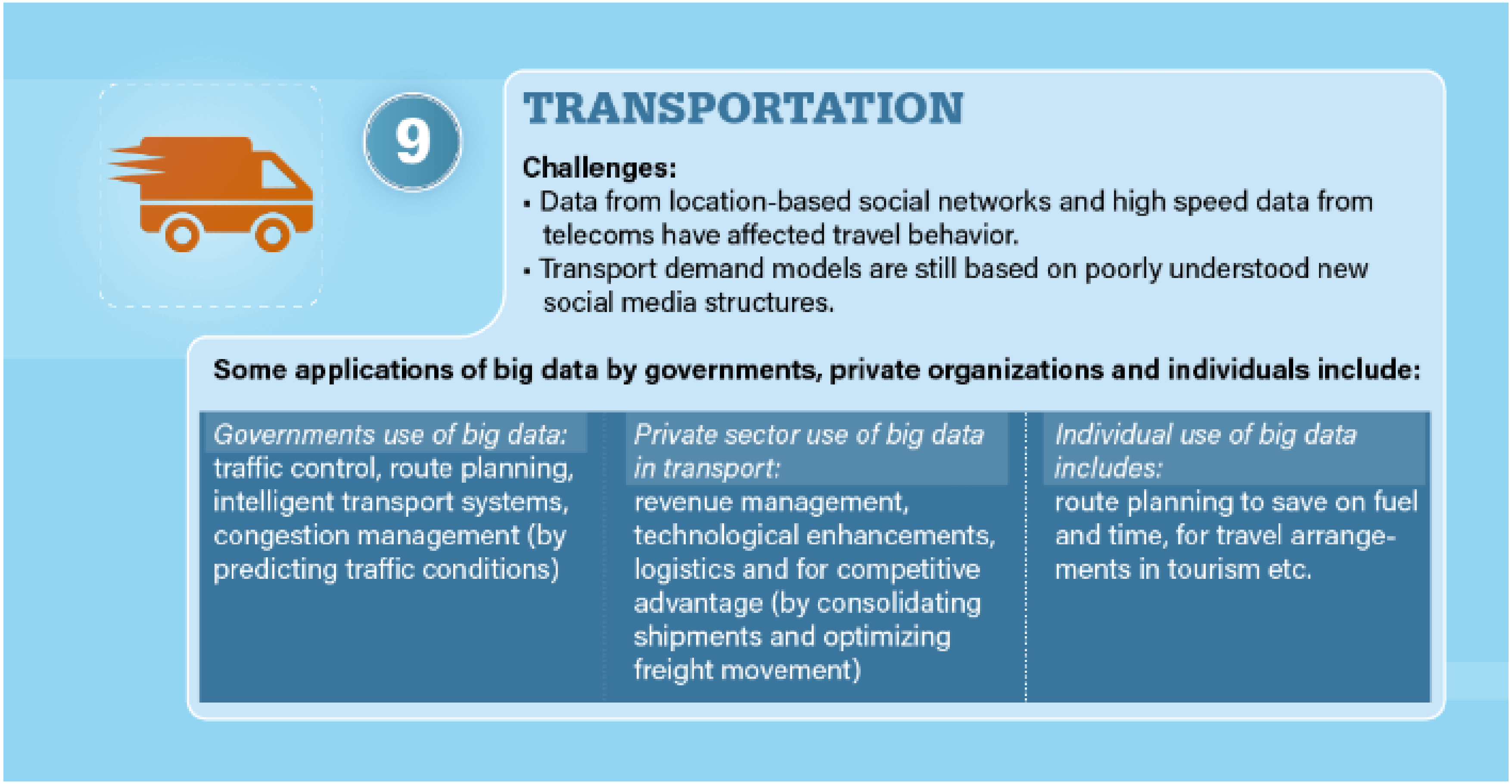
Challenges:

- Increase in the volume, complexity and velocity of data due to rising demands of Natural resources.
- Large volumes of untapped data from the manufacturing industry.
- Underutilization of data prevents improved quality, energy efficiency, reliability and better profit margins.

Enhancement in Supply chain capabilities from big data being used to increase productivity

UPT DCT DEPARTAMENTO CIÊNCIA E TECNOLOGIA

Big Data Applications



9

TRANSPORTATION

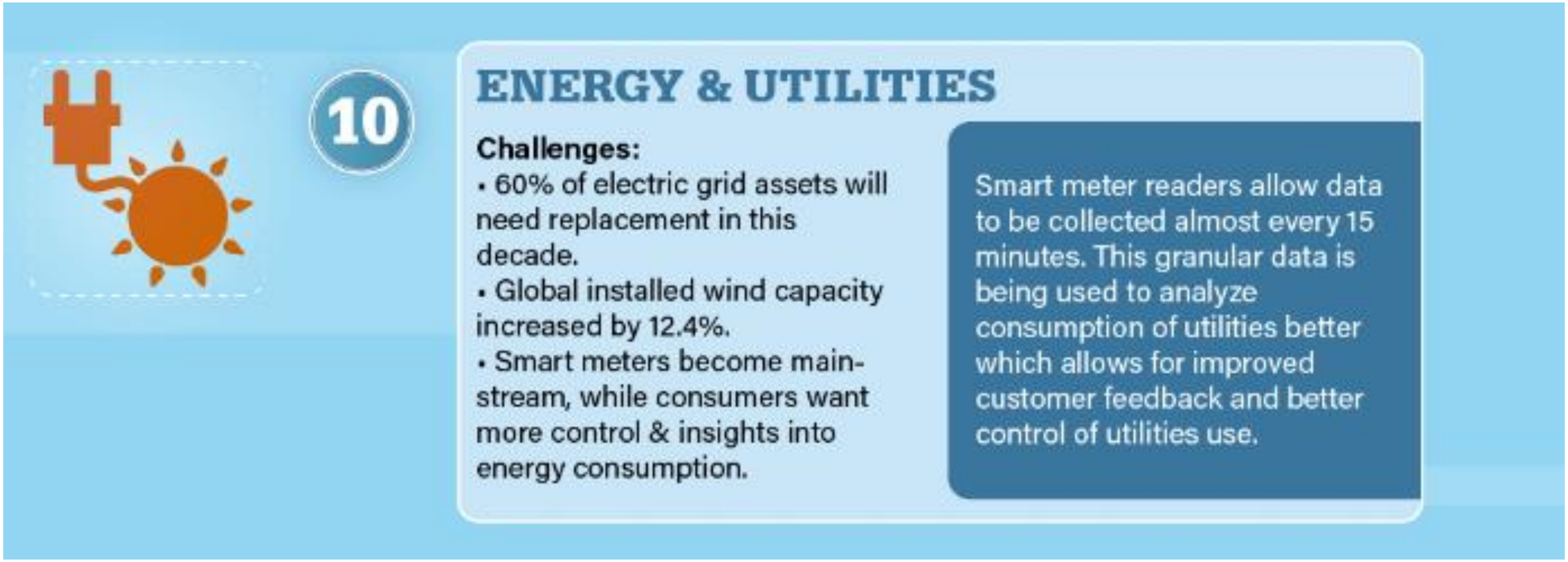
Challenges:

- Data from location-based social networks and high speed data from telecoms have affected travel behavior.
- Transport demand models are still based on poorly understood new social media structures.

Some applications of big data by governments, private organizations and individuals include:

Governments use of big data:	*Private sector use of big data in transport:*	*Individual use of big data includes:*
traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)	revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement)	route planning to save on fuel and time, for travel arrangements in tourism etc.

Big Data Applications



ENERGY & UTILITIES

Challenges:

- 60% of electric grid assets will need replacement in this decade.
- Global installed wind capacity increased by 12.4%.
- Smart meters become mainstream, while consumers want more control & insights into energy consumption.

Smart meter readers allow data to be collected almost every 15 minutes. This granular data is being used to analyze consumption of utilities better which allows for improved customer feedback and better control of utilities use.

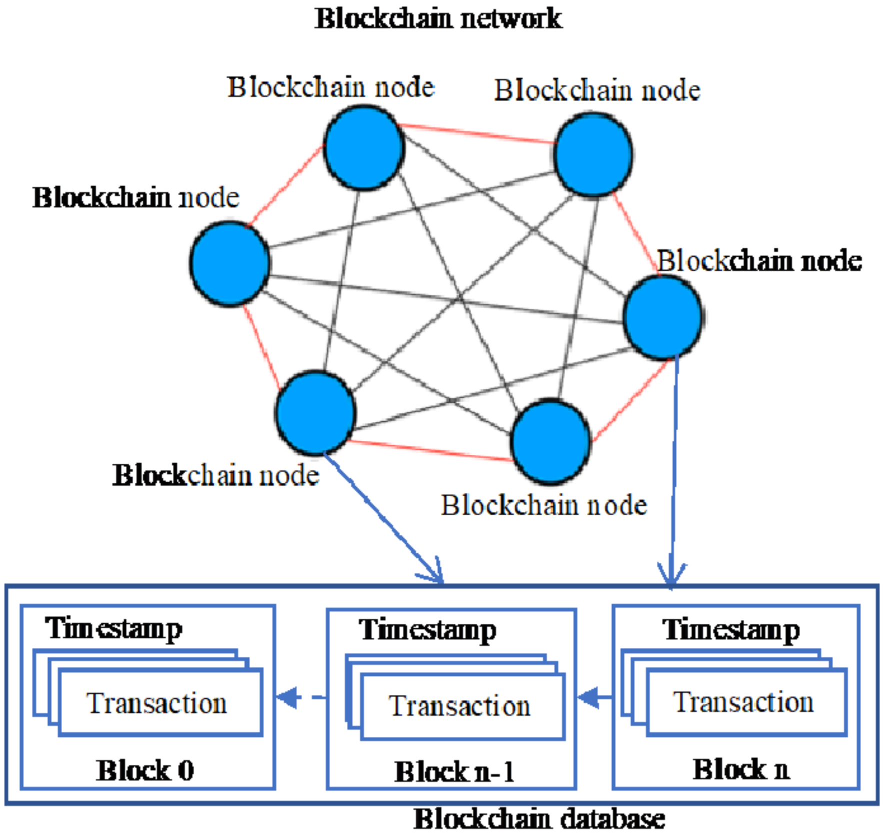
Big Data Analytics

- Big data analytics support the decision-making process by providing analytics and predictive techniques

- **Healthcare:** anti-cancer therapy, tracking patient vitals, hospital administration improvement, fraud detection and prevention for health insurance company
- **Education**: performance prediction, data visualisation, intelligent feedback, course recommendations, student skill estimation, and behaviour detection etc.
- **Agriculture:** boosting productivity, predicting yields, risk management, and food safety
- **Industry**: predictive maintenance, optimization of manufacturing line
- **Banks**: fraud detection

Blockchain and smart contracts

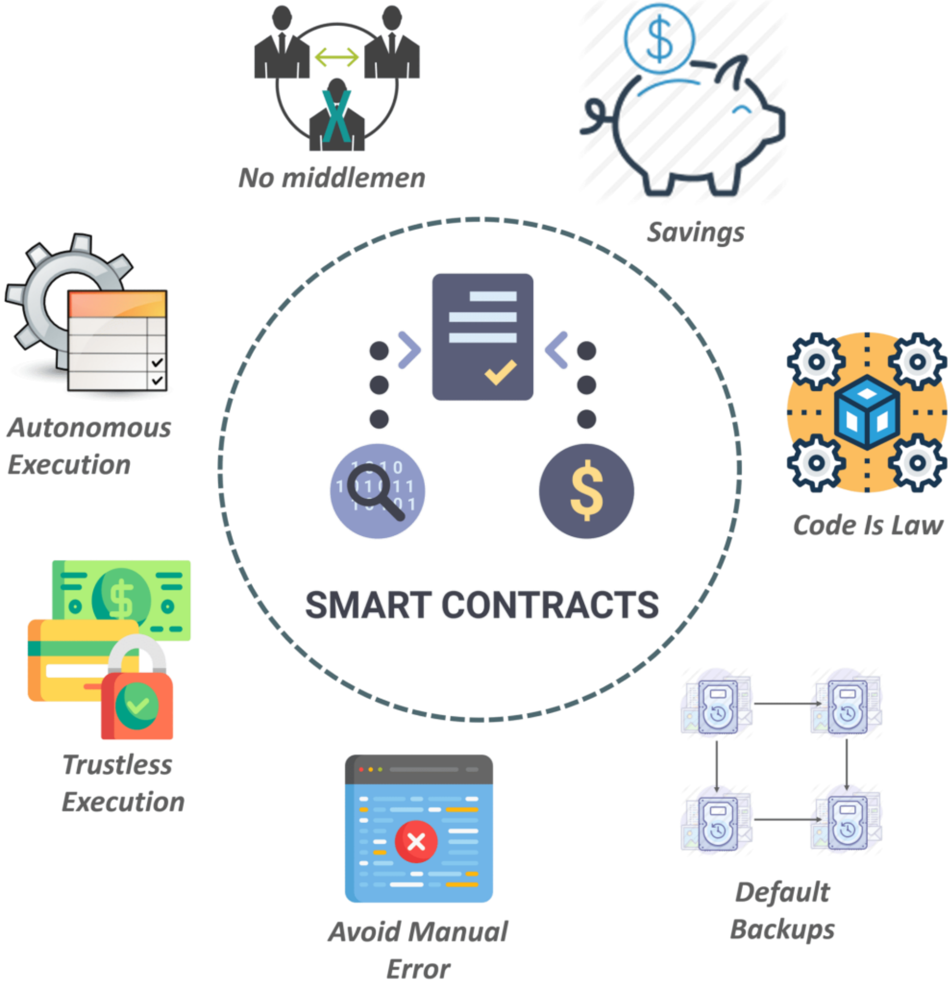
Blockchain is a distributed and immutable ledger that facilitates the process of recording transactions and tracking information



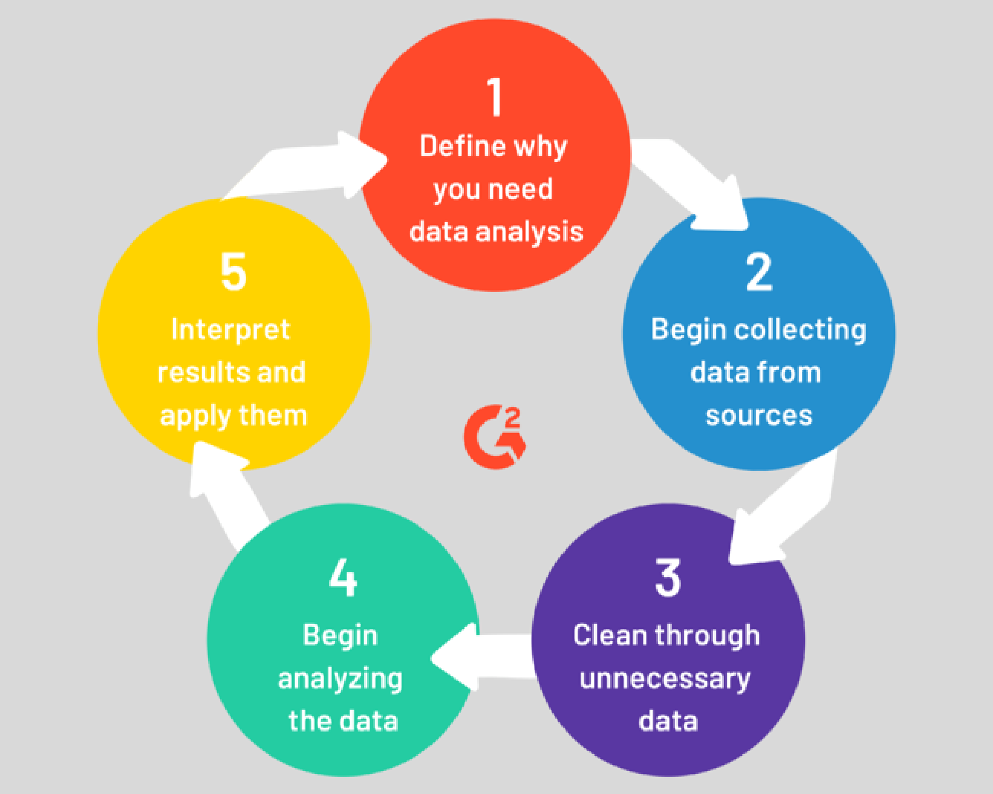
Blockchain and smart contracts

A smart contract is a computer program or a transaction protocol

It intends to automatically execute, control or document legally relevant actions according to the terms of a contract or an agreement



- Let’s go to analyse data!


1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

UNIVERSIDADE
PORTUCALENSE

Do conhecimento à prática.