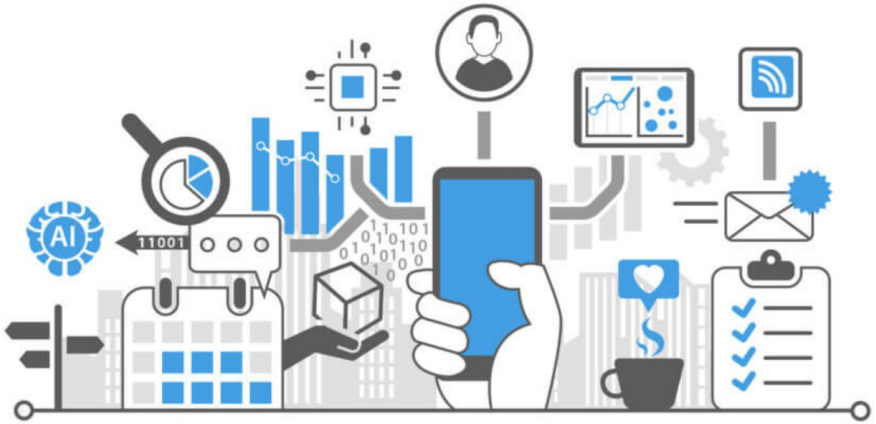
ADALTAS

1. Introduction to BIG DATA



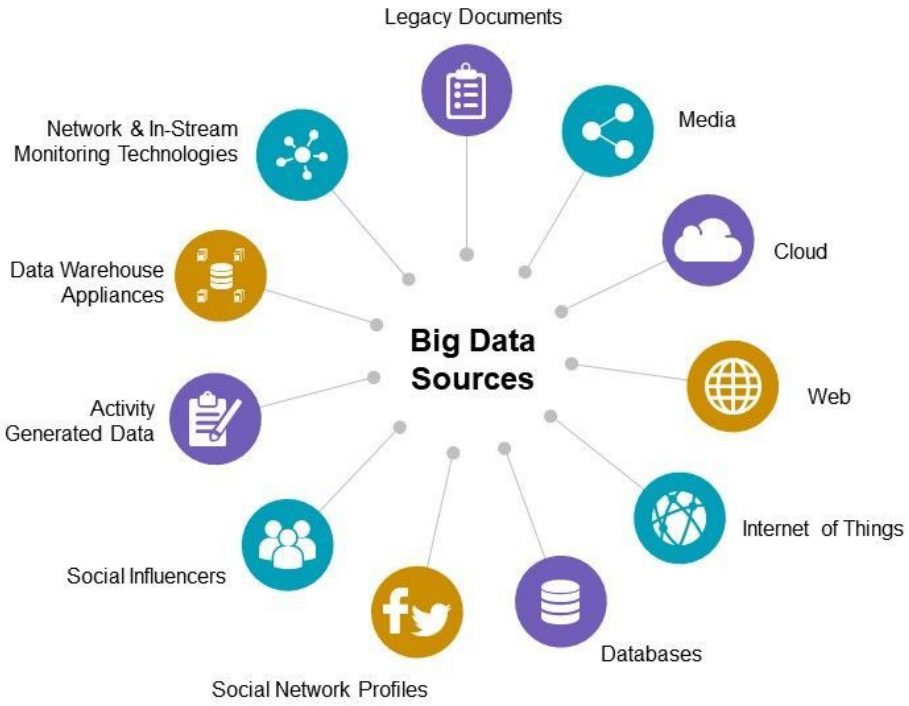
Petra KAFERLE DEVISSCHERE

Before we start...

- Data Scientist @ Adaltas
- theory, labs
- exam (5 questions, 1h)

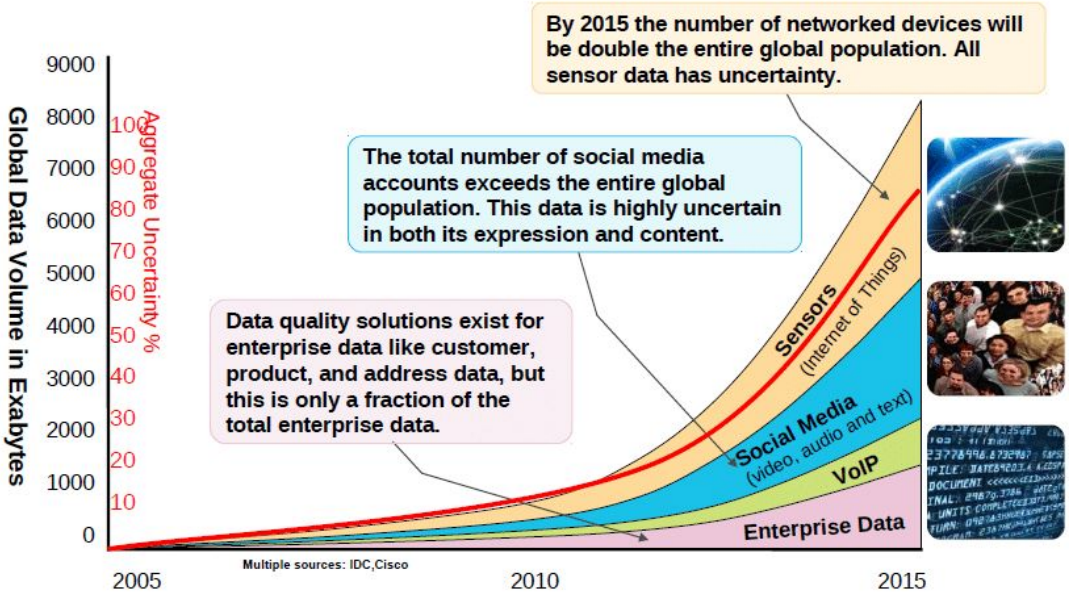
Where is data coming from?

- Machines: sensors
- People: social media
- Organisations: transactions



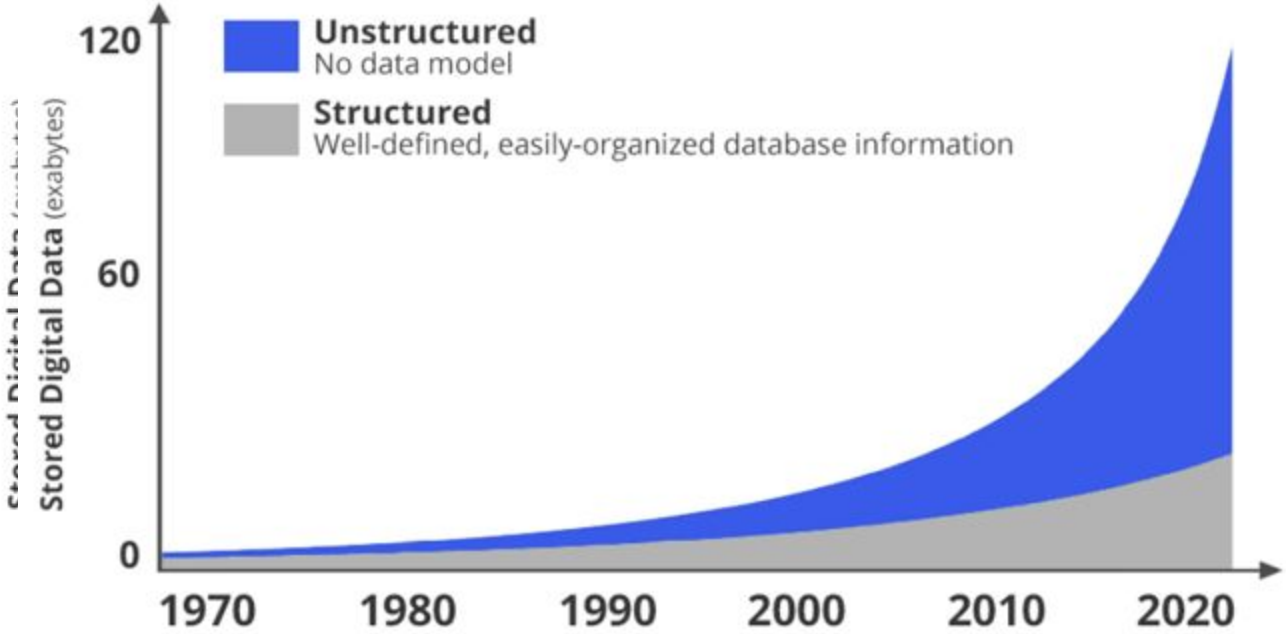
Where is data coming from?

- Machines: sensors
- People: social media
- Organisations: transactions



https://www.researchgate.net/figure/How-Data-Uncertainty-is-increasing-Source-2_fig1_329755755

Characteristics of big data



https://seekingalpha.com/article/4350544-splunk-strong-prospects-for-foreseeable-future

Characteristics of big data - 6 Vs

ADALTAS

The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.

6 Vs of big data: Volume

ADALTAS

- Size
- Several big datasets or many small data chunks
- Challenge: storage, ETL, analytics

6 Vs of big data: Variety

ADALTAS

- Complexity, heterogeneity
- Different types: photos, text, GPS…
- Coming in real-time or not
- Different media for the same data: audio of speech, transcript of the speech

6 Vs of big data: Velocity

ADALTAS

- Speed:
  - of creating
  - storing
  - analysing data
- A lot of applications are based on real-time response (Uber, recommendation systems…)

6 Vs of big data: Veracity

ADALTAS

- Quality
- Data can be noisy and biased
- Strongly depends on the data source

6 Vs of big data: Variability

ADALTAS

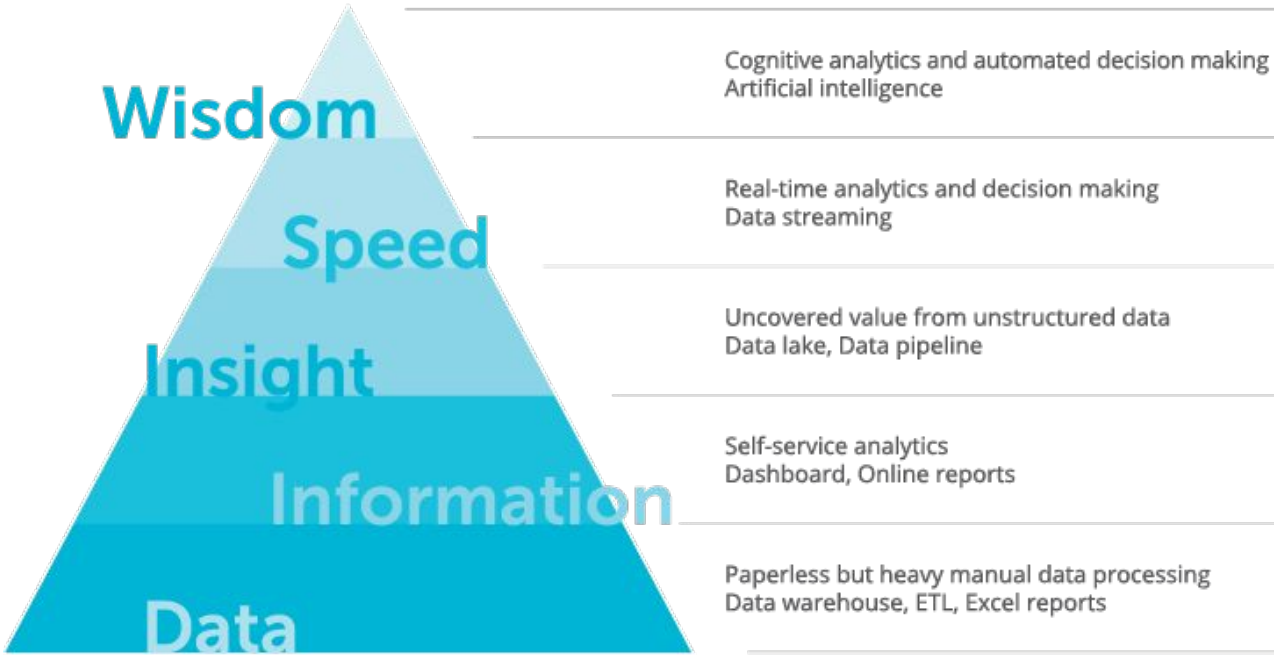
- The data is continuously changing
- Example: if you change the product, the customer's choice might change; evolution of a web page

6 Vs of big data: Value

ADALTAS

- The data is valuable, if it brings the value to the business
- To achieve that, we need a lot of different teams of people working together

Getting value from big data



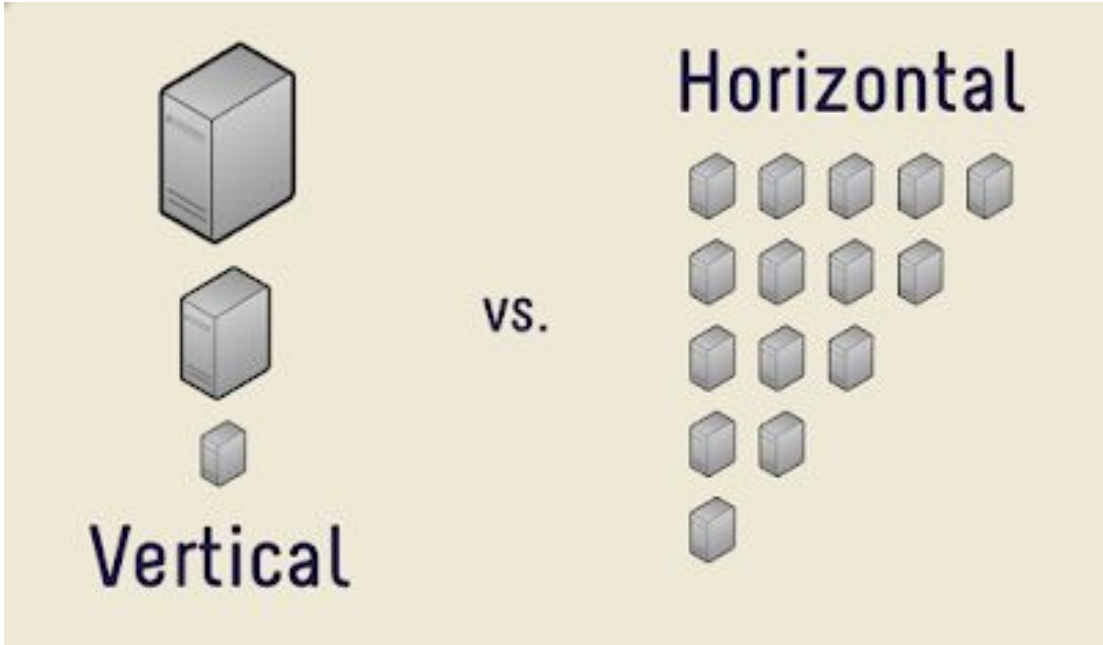
Better results with bigger datasets than with more precise ML algorithms.

How to process this data?

ADALTAS

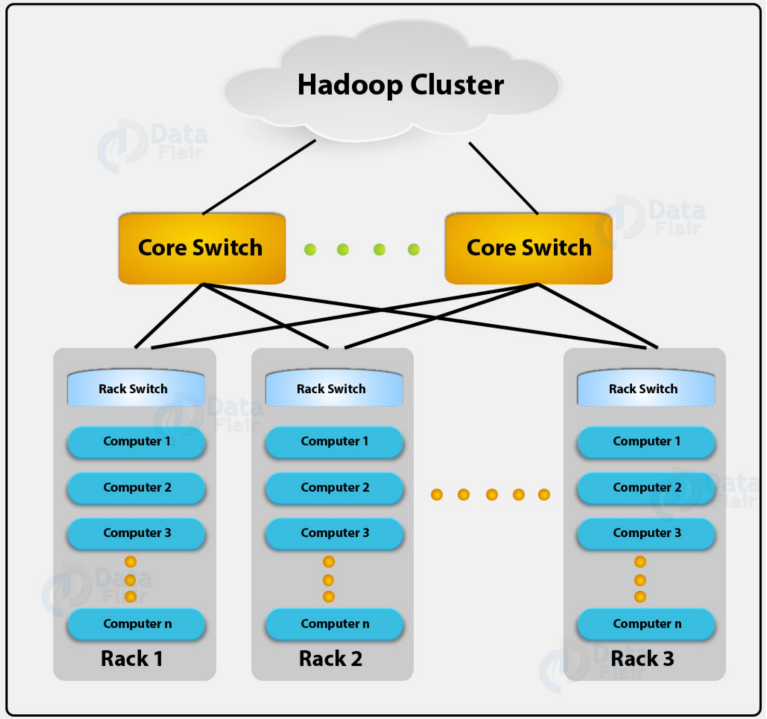
- Public cloud: AWS, Amazon, Google cloud
- Private cloud: OVH
- On-premise

Scaling

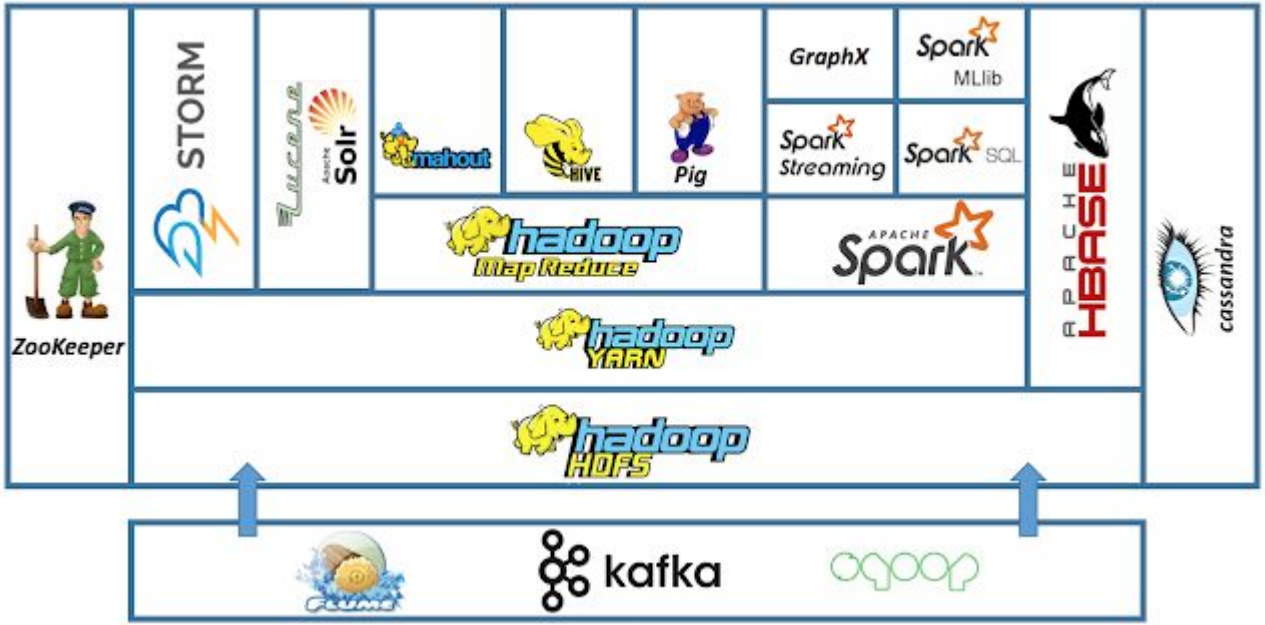


2. Hadoop Ecosystem

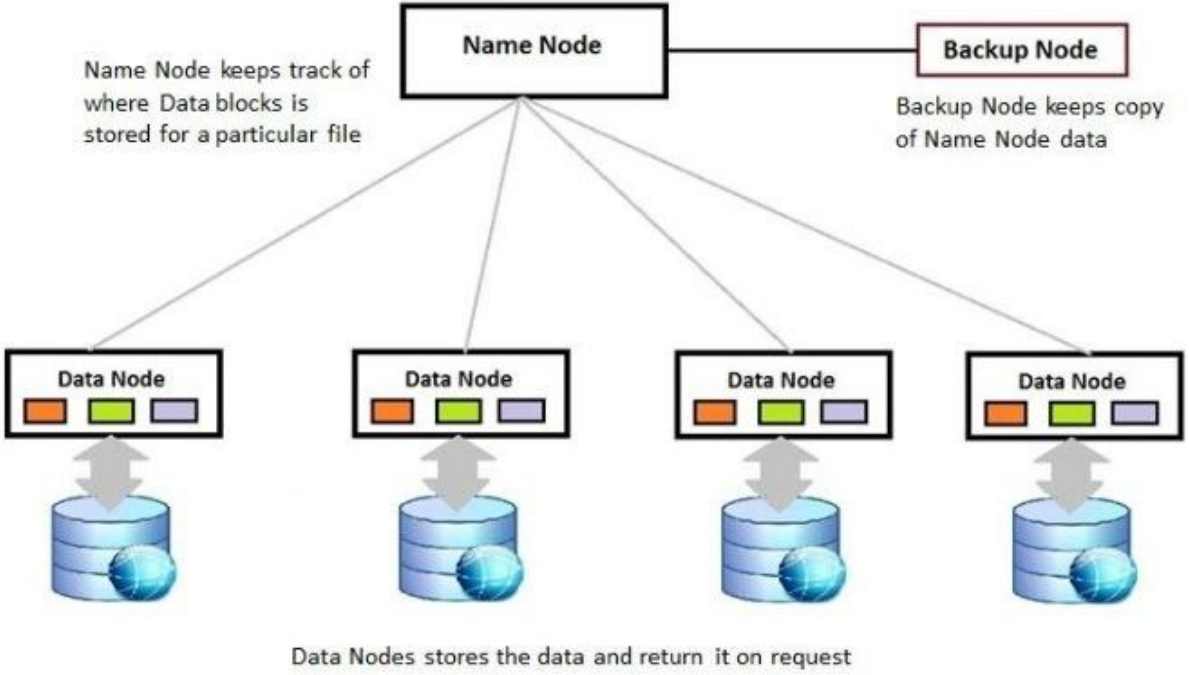
Hadoop cluster



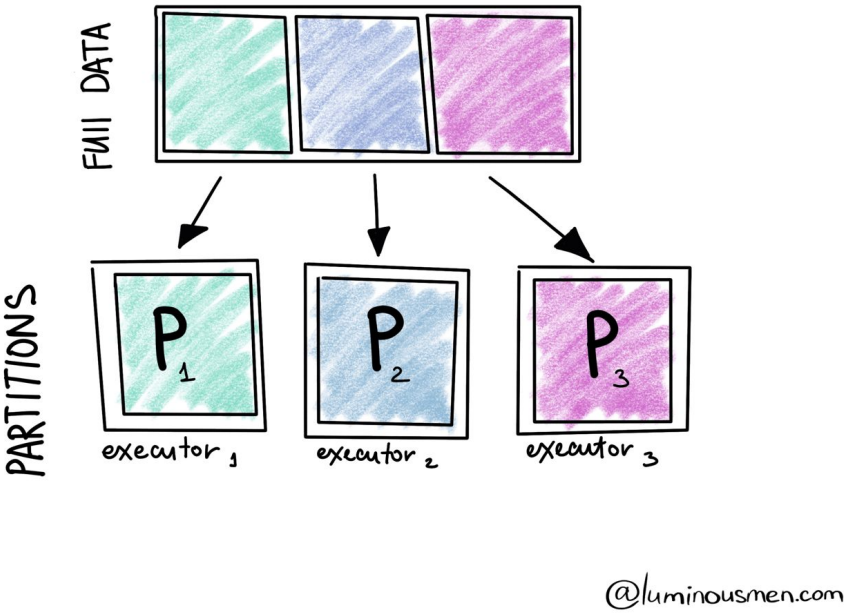
Hadoop Ecosystem



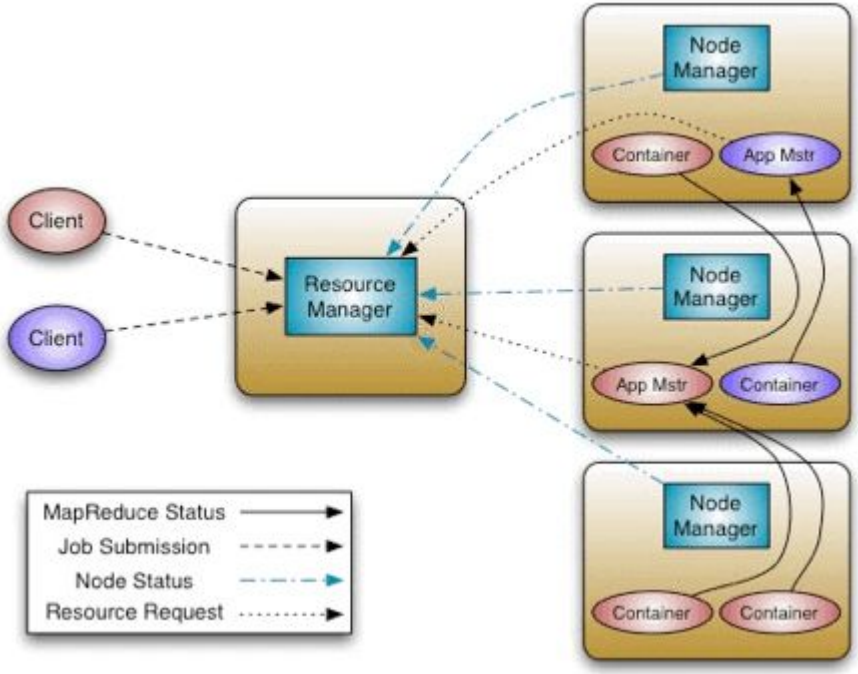
HDFS


Name Node
Backup Node
Name Node keeps track of where Data blocks is stored for a particular file
Backup Node keeps copy of Name Node data
Data Node
Data Node
Data Node
Data Node
Data Nodes stores the data and return it on request

Data parallelism

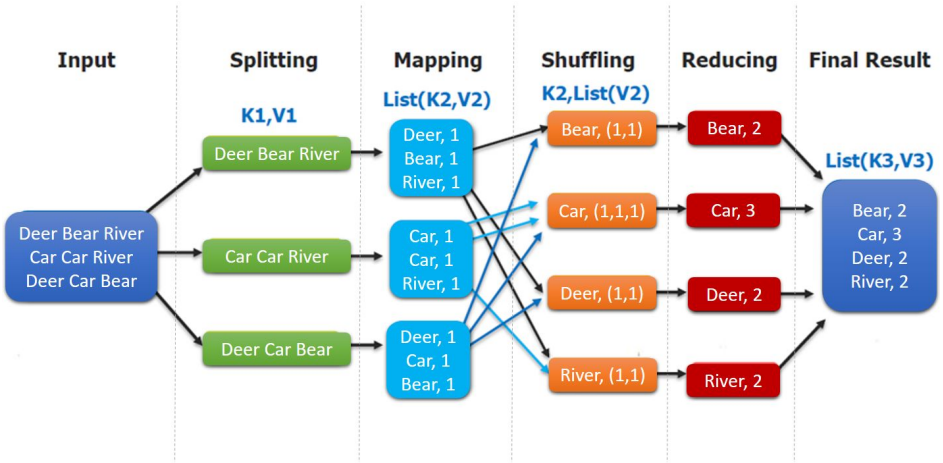


YARN


Client
Client
Resource Manager
Node Manager
Container
App Mstr
Node Manager
App Mstr
Container
Node Manager
Container
Container
MapReduce Status
Job Submission
Node Status
Resource Request

MapReduce

- Operates on key-value pairs
- Map - applies an operation to all elements
- Reduce - summarizes the results
- Writes the intermediate results to the disk -> slow
- Hive: SQL-like queries on top of MapReduce



MapReduce Word Count Process

Why Hadoop?

ADALTAS

- Scalable
- Can gracefully recover from crashes or hardware failures
- Ability to handle different data types (from 6 Vs: **V**ariety)
- Multiple jobs and/or users can use it at the same time
- Active open-source community with more than 100 projects