

Big Data, Hadoop, MapReduce

Regina Sousa and José Machado

ALGORITMI Research Center, University of Minho



CENTRO **ALGORITMI**



NoSQL Data Modeling and Management

1.1

NoSQL Concepts and Characteristics

1.2

Major Categories of NoSQL Data Models

1.3

NoSQL Features and Operations

1.4

ElasticSearch



NOSQL CONCEPTS AND CHARACTERISTICS

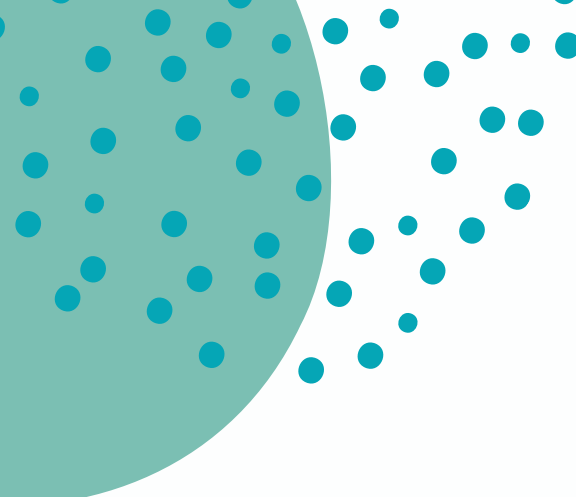
What is NoSQL?

NoSQL is a term that refers to a specific type of database model or database management system (DBMS).

• The term NoSQL is very broad, not referring to a specific database model. It refers to a whole different models that have as main feature not fitting the relational model.

• Many argue that the only thing all NoSQL databases have in common is that they do not follow the relational model. "NoREL" would be a more appropriate name.

One of the main reasons why the NoSQL approach started to be adopted was because of the big data arrival.



NOSQL CONCEPTS AND CHARACTERISTICS

Characteristics of a NoSQL Database

Non Relational

The most common situation occurs where data is unstructured or semi-structured.

Open Source

Open source is not necessarily a "NoSQL requirement", but it's a "NoSQL observation."

No Schema

In DBMS relational modeling we always have to consider the modeling effort before entering any data.

Scalable horizontally

Most NoSQL databases have excellent behavior in cluster environments.

Do not follow the principles of ACID

Without a standard query language.

**Not all NoSQL databases have these features.
However, most of these features are inherently non-existent in relational databases.**



MAJOR CATEGORIES OF NOSQL DATA MODELS

Key-value stores

Store data together as
columns instead of rows
and are optimized for
queries over large datasets

Graph databases

Are used to store
information about
networks, such as
social connections

Wide-column stores

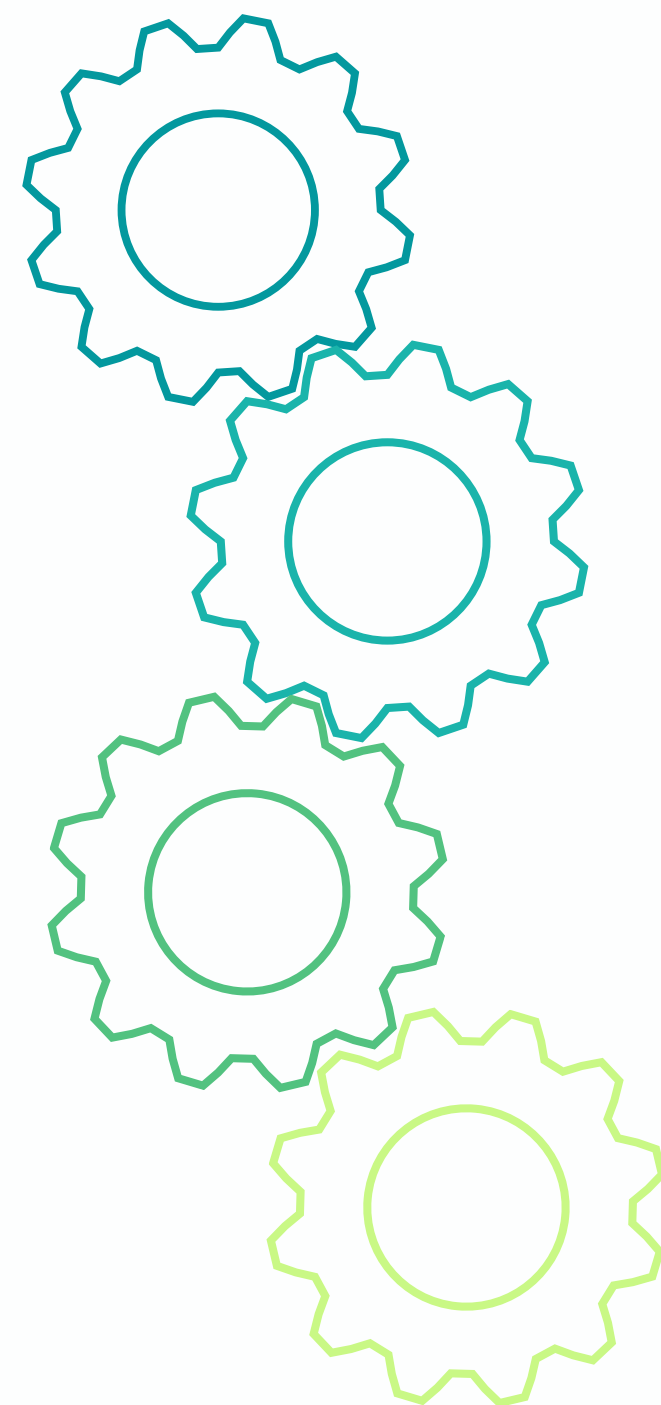
Are the simplest. Every
item in the database is
stored as an attribute
name (or "key") together
with its value.

Document databases

Pair each key with
a complex data
structure known
as a document.



NOSQL FEATURES AND OPERATIONS



01 DYNAMIC SCHEMAS

insert the data without the predefined schema

02 AUTO-SHARDING

automatically spread data across a various number of servers

03 REPLICATION

sophisticated NoSQL databases provide automated recovery
and are fully self- healing

04 INTEGRATED CACHING

keep frequently used data in system memory and remove the
need for the separate caching layer

05 SIMPLE API

offers interfaces that are easy to use for storing and querying data



ELASTIC SEARCH: DEFINITION

Search
engine and
data analysis

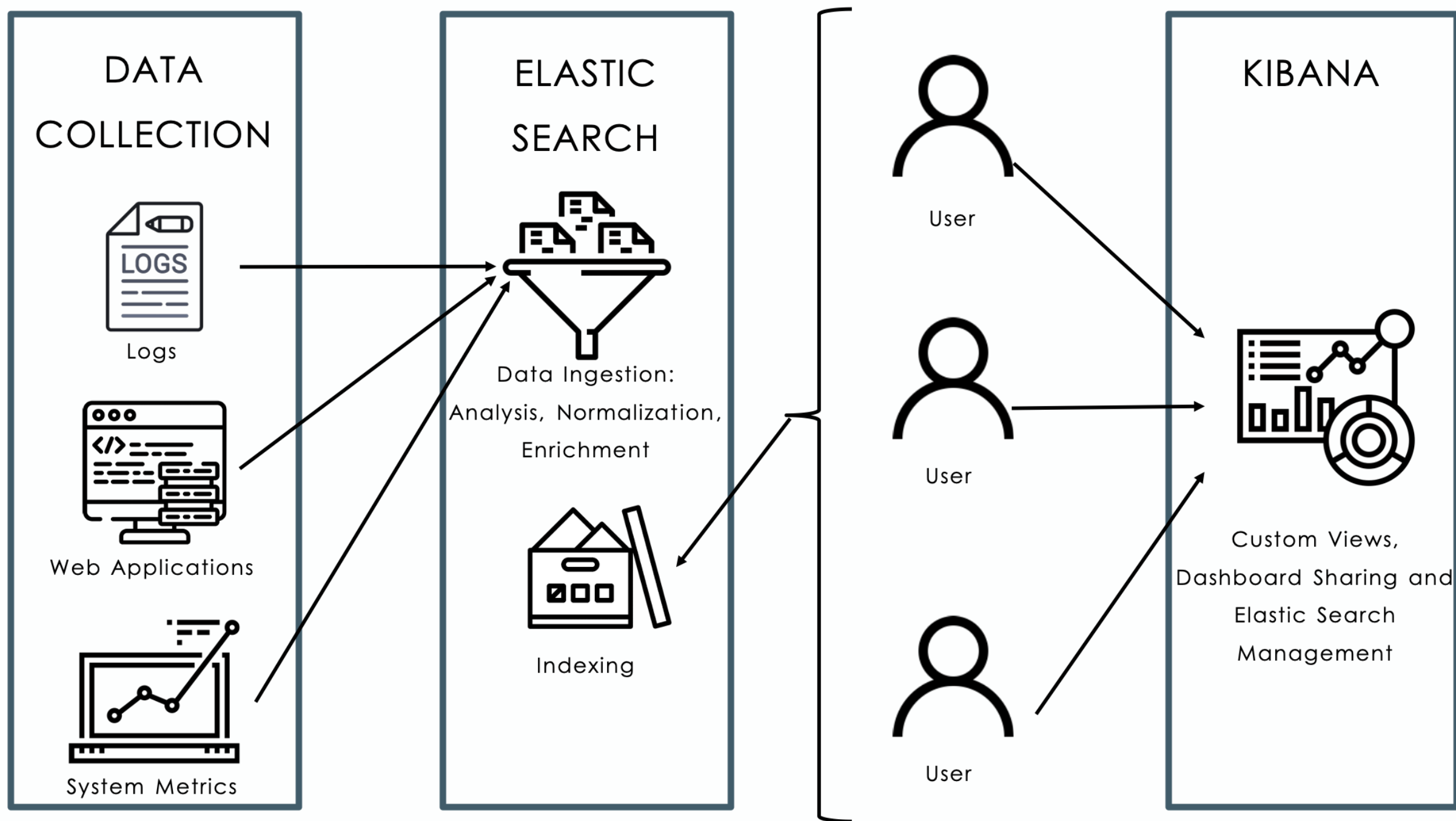
Known for its
distributed
speed and
scalability

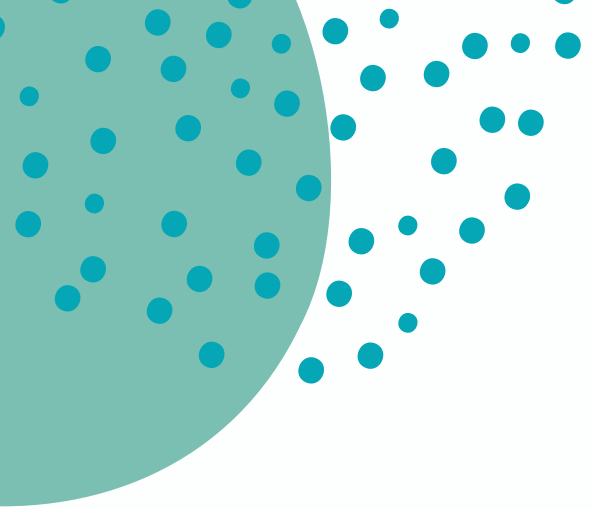
Includes several
programming
languages and
34 idioms

Is open source
and accepts
all types of
data

Collects raw data
from various
sources for later
processing and
analysis

HOW DOES ELASTIC SEARCH OPERATES





ELASTIC SEARCH: PRACTICAL APPLICATIONS

Search in application

Business Search

Logging and analysis
of log data

Search in website

Business Data Analysis

Application
Performance
Monitoring

Safety Analysis

Analysis and
Visualization of
geospacial data

Infrastructure metrics
and container
monitoring



REASONS TO USE ELASTIC SEARCH

1 SPEED

Is a nearly real-time search platform, meaning that the latency from the moment a document is indexed until it becomes searchable is very small - usually one second.

2 FACTORY DISTRIBUTION

Documents stored on Elasticsearch are distributed in various containers known as shards, which are replicated to provide backup copies of data in case of hardware failure

3 EXTENSIVE SET OF RESOURCES

In addition to speed, scalability, and resiliency, Elasticsearch has several integrated advanced features that make data storage and search even more efficient, such as data rollups and index lifecycle management.

4 SIMPLIFIES DATA INGESTION, VISUALIZATION AND REPORTING

Integration with Beats and Logstash makes it easier to process data before indexing on Elasticsearch. And Kibana provides real-time visualization of the data.

2

Large Scale Data Handling

2.1

Big Data Characteristics

2.2

Big Data Modeling and
Management

2.3

MapReduce

2.4

Hadoop



BIG DATA CONCEPT

The concept of Big Data remains so far a relative term with regard to the boundary between what is and is not considered Big Data. For a company such as Google, the concept and size of Big Data is much different from that assumed for a medium-sized company.

The most accepted definition was given by Douglas Laney. Laney observed that Big Data grew in three different dimensions:



Volume

Velocity

Variety

However, other authors have crossed these characteristics by adding several other V's to this definition, such as: Value, Veracity, Visualization, Viscosity, Virality, among others. The 4th most consensual V's is undoubtedly the **veracity**.



BIG DATA CHARACTERISTICS

VOLUME

The volume of data gives the large amount of data, mostly described in several petabytes or even more. However, not even this definition is consensual among the authors, since the definition depends on the type of data being analyzed.

VELOCITY

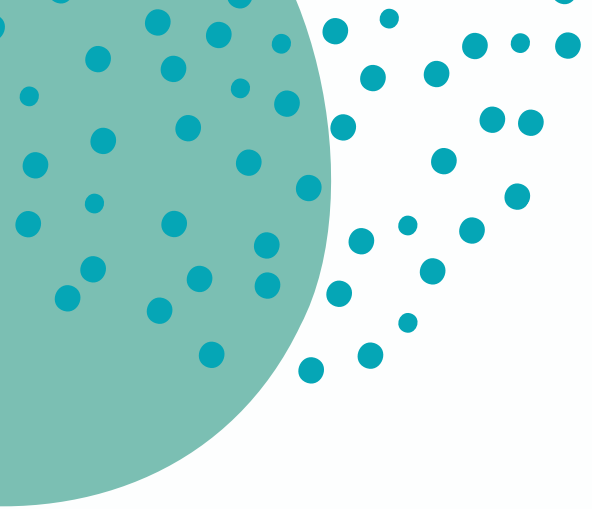
The velocity concerns both the rate of data generation and the speed of analysis they require. Big Data Velocity deals with the speed at which data flows in from sources.

VARIETY

The variety of data has increased exponentially due to the diversity of collection sources. Data can have several organizations and reach the collection point in a structured, semi-structured or even unstructured way. In addition, data formats must be taken into account.

VERACITY

Veracity encompasses the reliability inherent in some sources of data collection. For example, information taken from a social network cannot be given the same relevance as information taken from hospital software.



BIG DATA MODELING

Why Is Data Modeling Necessary?

Large amounts of data imply a system or method to keep everything in order. The process of sorting and storing data is called "data modeling". A data model is a method by which we can organize and store data.

Proper models and storage environments offer the following benefits to large data:

- **Performance:** Ensures fast query and reduces I/O output.
- **Cost:** Significantly reduces data redundancy, reducing storage and computing costs for the large data system.
- **Efficiency:** They greatly improve the user experience as well as the efficiency of data use.
- **Quality:** They make data statistics more consistent and reduce the possibility of computing errors.

6 TIPS FOR MODELING BIG DATA

01 DON'T IMPOSE TRADITIONAL MODELING

02 DESIGN A SYSTEM, NOT A SCHEMA

03 LOOK FOR BIG DATA MODELING TOOLS

04 FOCUS ON DATA THAT IS CORE TO YOUR BUSINESS

05 DELIVER QUALITY DATA

06 LOOK FOR KEY INROADS INTO THE DATA



BIG DATA MANAGEMENT

Big Data Management is a set of practices that promotes the collection, organization, administration and interpretation of large volumes of data.

The main objective is to treat the contents so that they become accessible and reliable. There are 4 terms that are essential to the definition of this method:

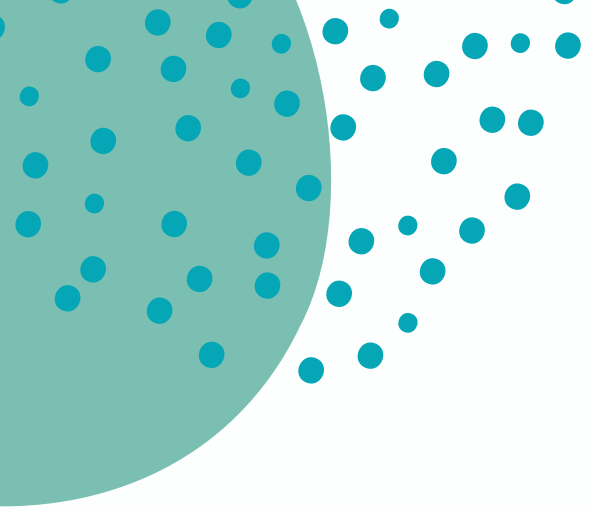


Adequacy

Integration

Migration

Management



BIG DATA MANAGEMENT

Adequacy

Ability to analyze a large amount of information, structured or not, allows the detection and correction of errors in stored information

Ability to filter and classify data so that it can later be handled assuming a standardized structure.

Integration

Migration

Ability to move data from one environment to another quickly and conveniently.

Ensure the availability and security of data, ensuring that it follows all the organization's policies and standards.

Management

BIG DATA MANAGEMENT: ADVANTAGES

**Increase in
company
revenue**

**More accurate
decision
making**

**Strategy
improvement**

**Team
productivity
and efficiency**



DATA WAREHOUSE VS DATA LAKE

Processed, Structured

DATA

Structured, Semi-Structures,
unstructured, raw

Schema on write

PROCESSING

Schema on read

Expensive for large data
volumes

STORAGE

Design for low cost storage

Fixed Configuration

AGILITY

Configure/Reconfigure as
necessary

Business Professional

SECURITY

Data Scientists/Analysts

3

Practice

3.1

Hadoop Ecosystem

3.2

MapReduce & Hadoop

3.3

Word Count Practice

3.4

Challenge

HADOOP: ECOSYSTEM



Apache Pig



mahout



HIVE



- Drill
- Zookeeper
- Ambari
- Hbase
- Solr
- Lucene
- Storm

Yarn
(Resource Manager)



HDFS
(Storage System)



Unstructured
Data



Relational
Data



MAP REDUCE

MapReduce is a **programming paradigm** that allows massive scalability across hundreds of thousands of servers in a Hadoop CLuster.

Map Reduce is **THE HEART OF HADOOP**

MAP

Takes a dataset in its raw form and converts it into another dataset, where individual elements are broken into tuples (key/value pairs)

REDUCE

Takes the Map process output and combines the data tuples into a small set of tuples



MAP REDUCE: CHARACTERISTICS

Scalable

Fault-tolerant

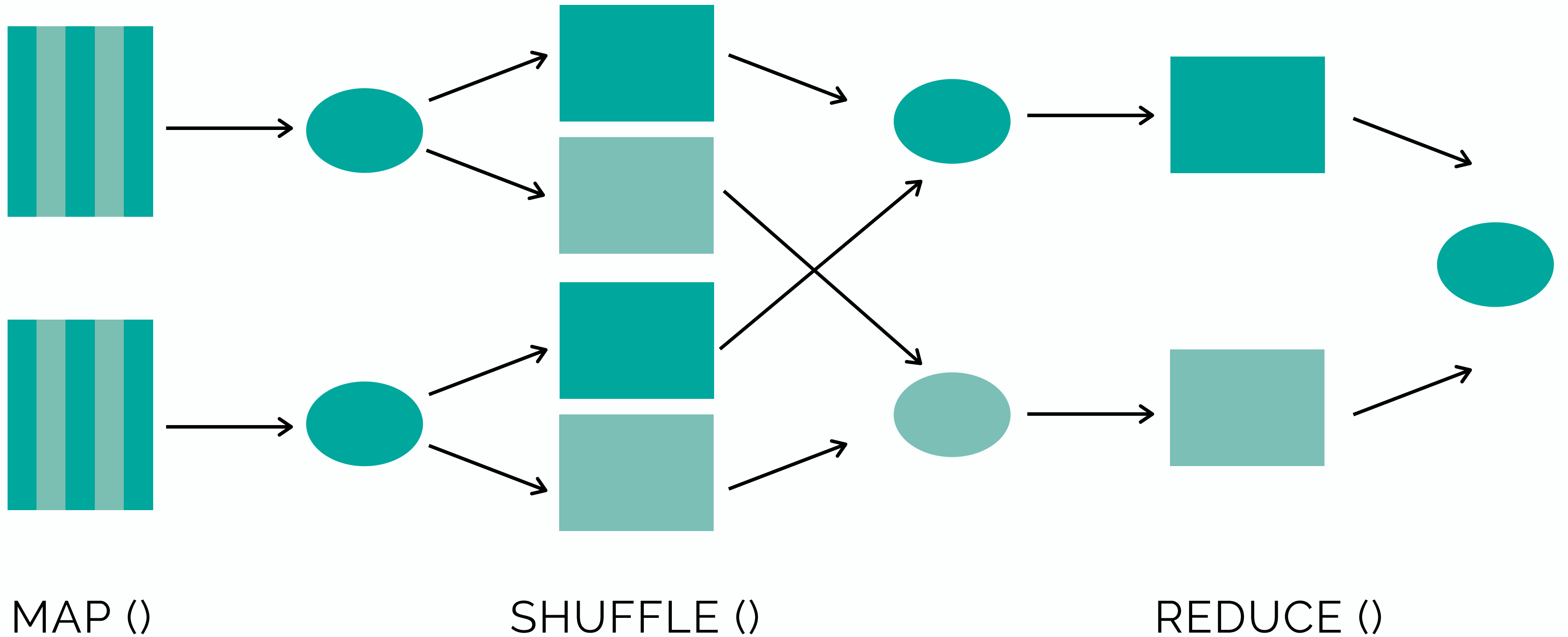
**High
availability**

Reliable

**Works on the
key value
concept**

**No traffic
congestion in
the network**

MAP REDUCE: HOW IT WORKS





HADOOP

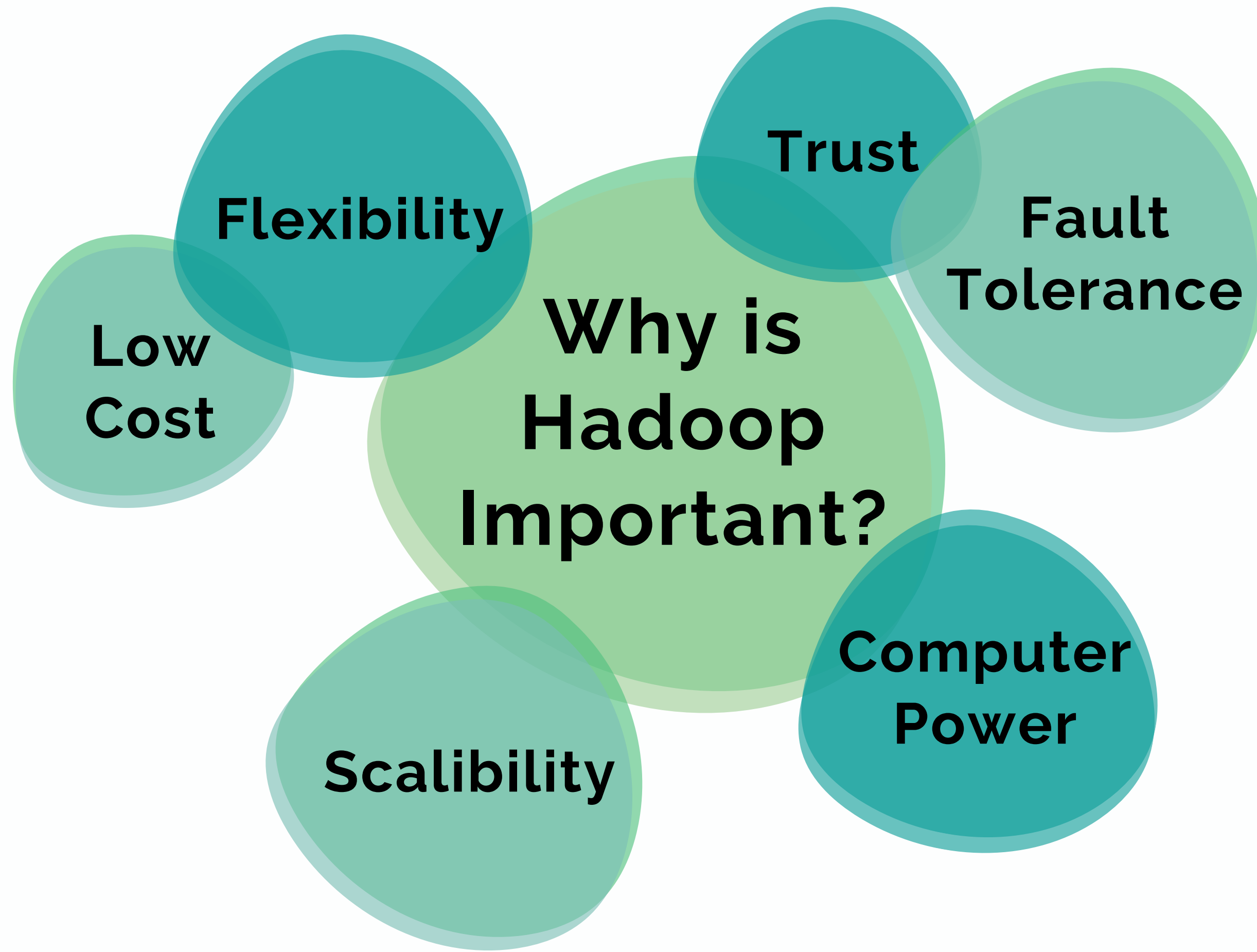
Hadoop is open source software that can handle both the storage and processing of large amounts of data, in a distributed way, using clusters of computers with commodity hardware.

WITH HADOOP, NO DATA IS TOO BIG.



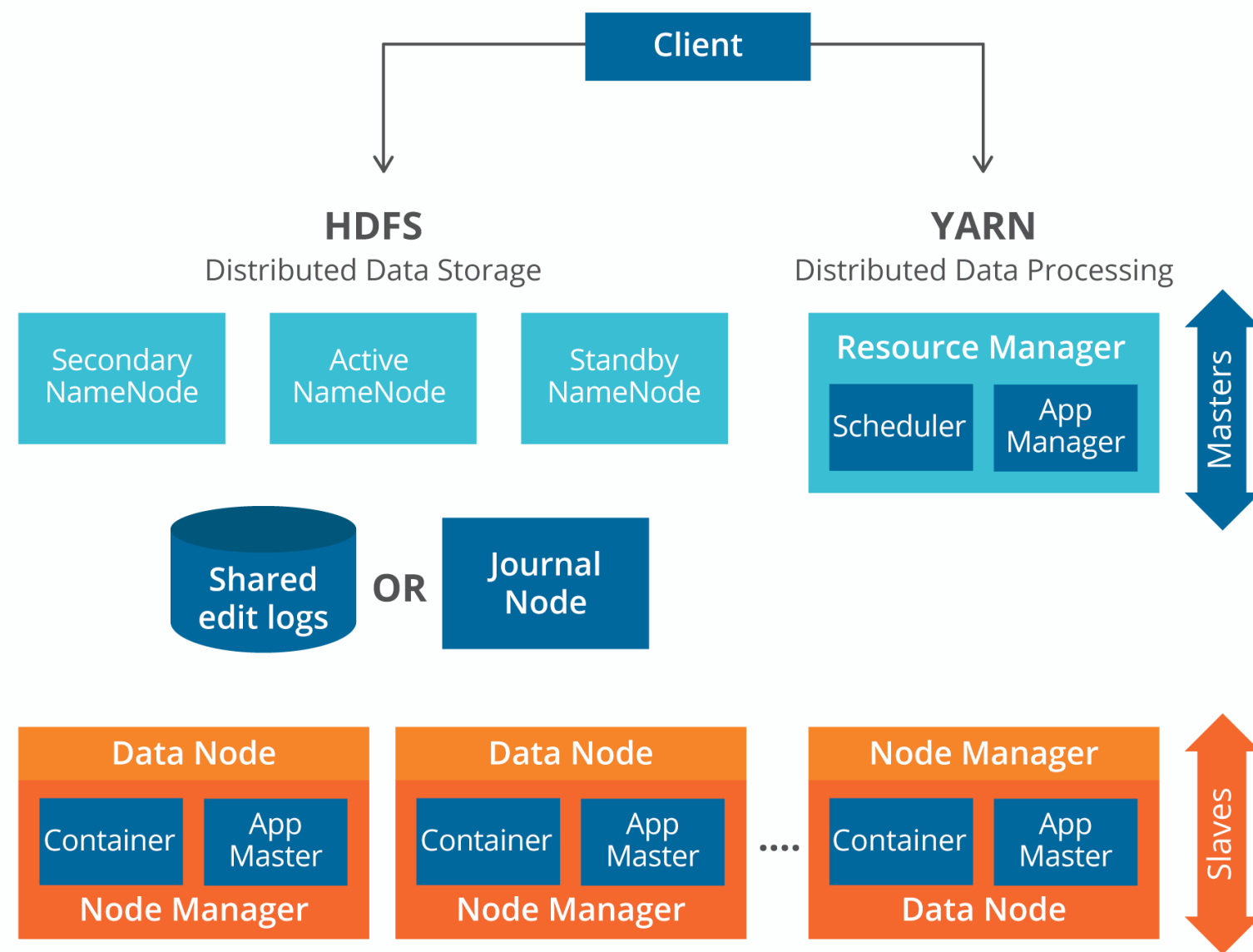
Curiosity: "Hadoop" was the name one of the creator's sons, Doug Cutting, gave his stuffed elephant.

HADOOP IMPORTANCE



HADOOP

Apache Hadoop 2.0 and YARN





HADOOP DISTRIBUTED FILE SYSTEM

HDFS is the basis of Hadoop and is therefore the most important component of the ecosystem. It is a Java software that offers features such as scalability, high availability, fault tolerance, cost-benefit, etc. It provides a distributed and robust data storage.

This component is composed by 3 other essential subcomponents:

- DataNode
- NameNode
- Secondary NameNode

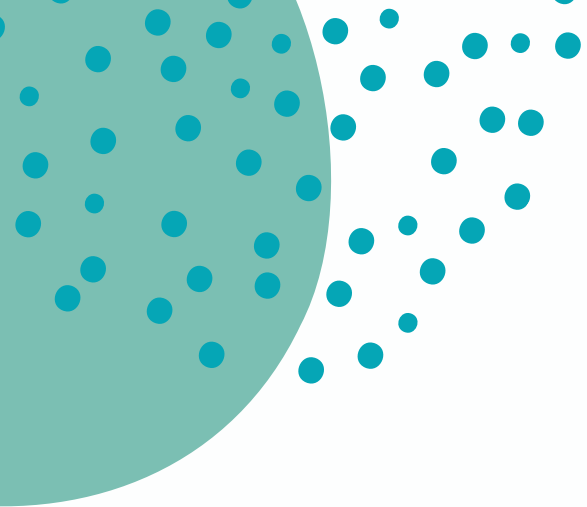


HDFS: MAIN COMMANDS

It is a different file system from the patterns we see, for example in linux

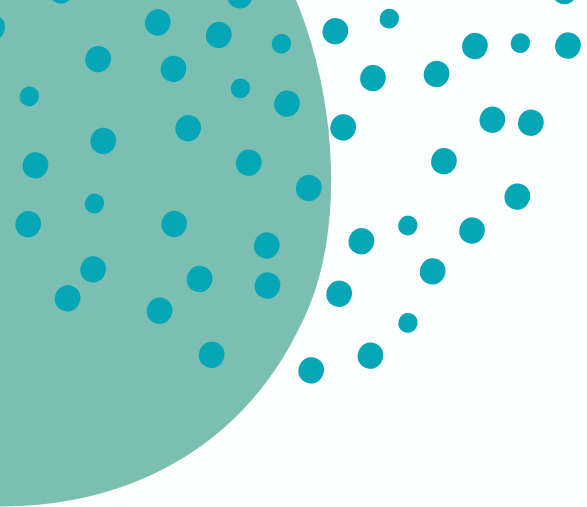
Differentiated access (no direct compatibility)

There is some similarity between commands and it is possible to share files between the two



HDFS: MAIN COMMANDS

Command	Description	Parameters	Example
-ls	List the content of the board	-d simple list -r recursive	hdfs dfs -ls -R /
-put	Copy the file from the local system to the HDFS		hdfs dfs -put name.txt /diretoria/name2.txt
-mv	Moves the file or directory from the local system to the HDFS		hdfs dfs -mv name.txt /user
-rm	Remove the file or folder	-r excludes in a recursive way	hdfs dfs -rm /user/name.txt

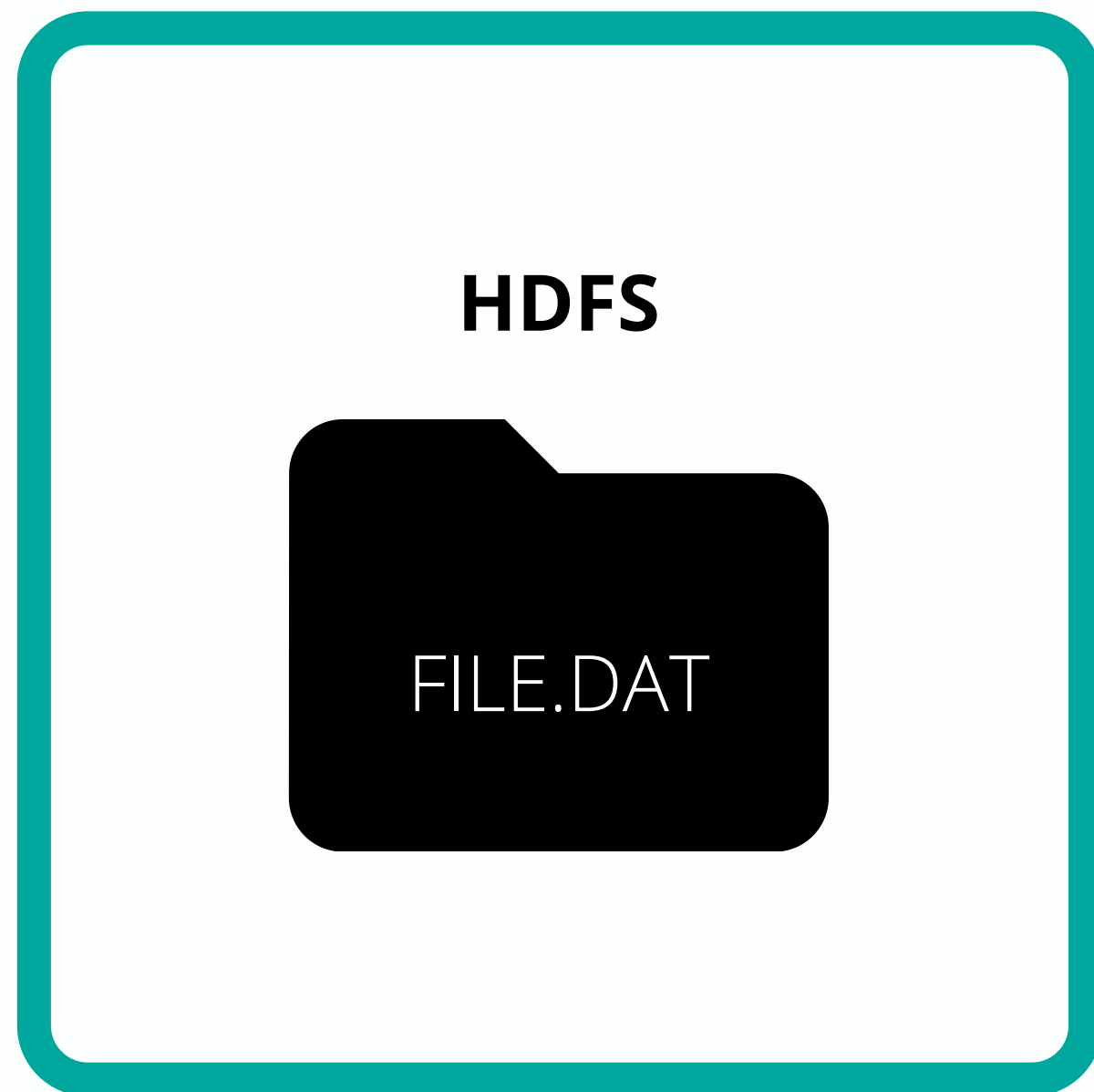


HDFS: MAIN COMMANDS

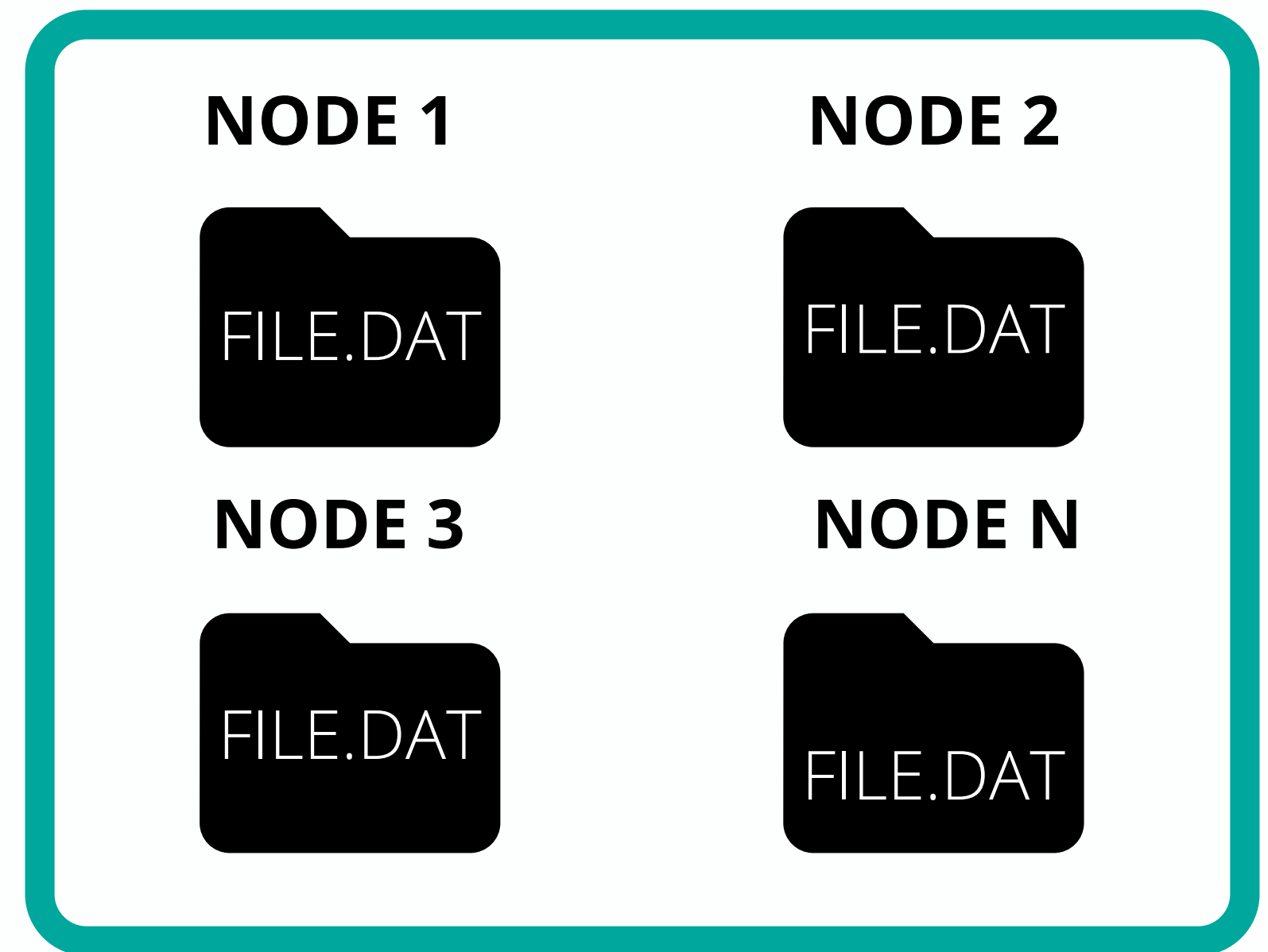
Command	Description	Parameters	Example
-du	Check file size		hdfs dfs -du /user/name.txt
-cat	Displays the contents of the file		hdfs dfs -cat name.txt
-mkdir	Create a folder	-p Creates a path	hdfs dfs -mkdir /user/diretoria
-tail	Shows the end of the file		hdfs dfs -tail /user/name.txt

HDFS: ARRANGEMENT

WHAT WE SEE



WHAT IT REALLY IS





HDFS: FILE TYPES

Text:

Standard in tools such
as HIVE

ORC:

Optimized for columns
and rows (the favorite
of the whole
ecosystem).

Parquet:

Column oriented
(Binary)

Sequence File:

Key-value
Can be easily divided or
unified

AVRO:

Binary format for
serialization. Very
useful for data
exchange

RC:

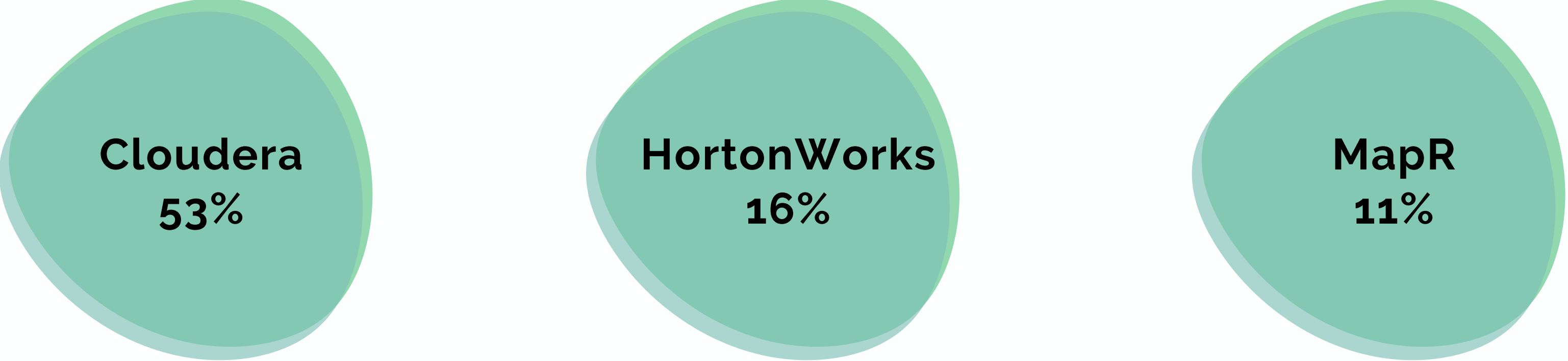
Column oriented, key-
value.



HADOOP DISTRIBUTION

IN WHAT WAYS CAN WE IMPLEMENT THIS TECHNOLOGY?

- Open source distribution, provided by Apache
- Distributed by third parties:
 - Includes the open source tool + add-ons
 - Possible support offer
 - Free version with limited number of us.



Cloudera
53%

HortonWorks
16%

MapR
11%

IMPLEMENTATION: NEEDS

- **Virtual Box (Oracle, VMWare, Docker,...):**
<https://www.virtualbox.org/wiki/Downloads>
- **Cloudera Image :** <https://www.cloudera.com/downloads/cdp-private-cloud-trial.html>
- Hadoop.zip :
 - search.txt
 - WordCount.java (Class)



VirtualBox

cloudera



PROBLEM 1: STATEMENT

Batch

A bicycle producer wants to know which model is most sought after

The producer has a website where in the search field he always keeps the model that the user searches for

Let's take the word file in its raw point and, using Hadoop count how many times each template was searched

UNDERSTANDING THE EXERCISE

HDFS

Elite
BigWhell
Elite
Elite
Elite

BigWhell
BigWhell
BigWhell
Elite
Auge55
Auge55

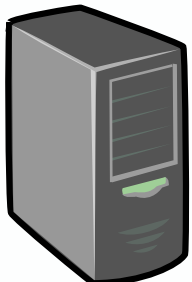
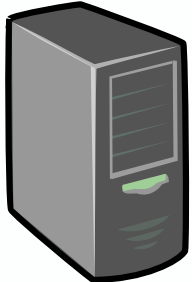
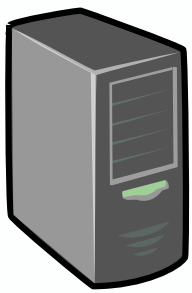
Elite
BigWhell
Auge55
Elite
Auge55

MAP

(Elite,1)
(BigWhell,1)
(Elite,1)
(Elite,1)
(Elite,1)

(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(Elite,1)
(Auge55,1)
(Auge55,1)

(Elite,1)
(BigWhell,1)
(Auge55,1)
(Elite,1)
(Auge55,1)



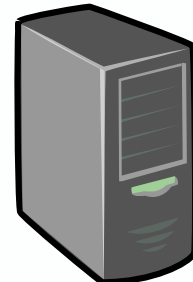
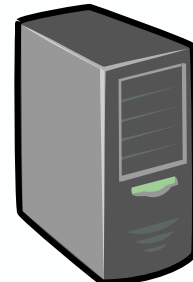
SHUFFLE

(Auge55,1)
(Auge55,1)
(Auge55,1)
(Auge55,1)
(Auge55,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(Elite,1)
(Elite,1)
(Elite,1)
(Elite,1)
(Elite,1)
(Elite,1)

REDUCE

(Auge55,1)
(Auge55,1)
(Auge55,1)
(Auge55,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)
(BigWhell,1)

(Elite,1)
(Elite,1)
(Elite,1)
(Elite,1)
(Elite,1)
(Elite,1)
(Elite,1)



RESULT

(Auge55,4)
(BigWhell,5)
(Elite,7)

PROBLEM 1: NEEDS



```
1203 Auge555
1204 Auge555
1205 StradaRacing
1206 StradaRacing
1207 Auge555
1208 Elite
1209 Auge555
1210 Auge555
1211 Auge555
1212 StradaRacing
1213 StradaRacing
1214 StradaRacing
1215 AudaxVentus
1216 AudaxVentus
1217 SL429F
1218 SL429F
1219 AudaxVentus
1220 AudaxVentus
1221 Elite
1222 Elite
```

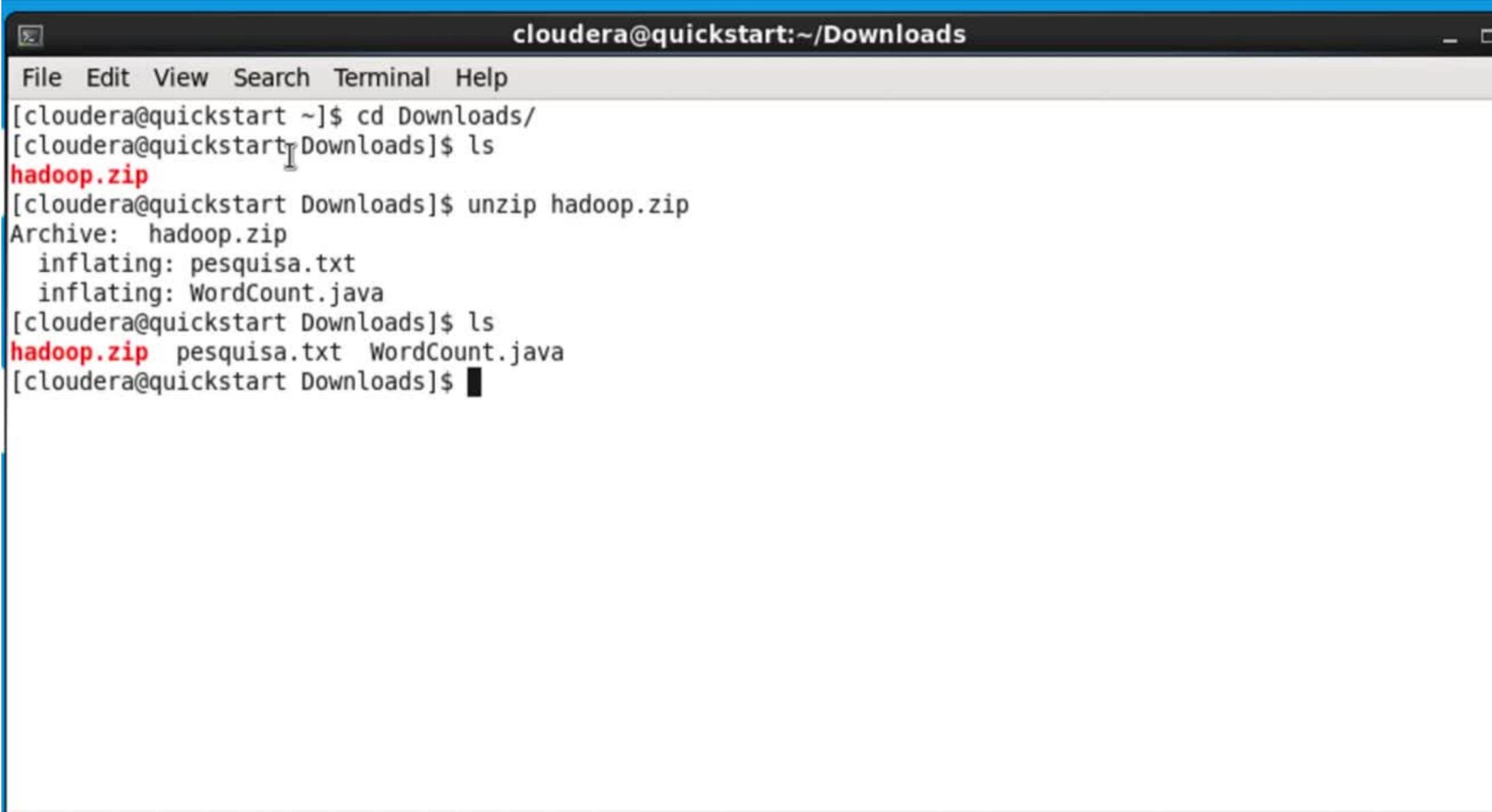
Raw File

```
WordCount.java x
1 package PackageDemo;
2
3 import ...
15
16
17 public class WordCount {
18
19     public static void main(String [] args) throws Exception
20     {
21         Configuration c=new Configuration();
22         String[] files=new GenericOptionsParser(c,args).getRemainingArgs();
23         Path input=new Path(files[0]);
24         Path output=new Path(files[1]);
25         Job j=new Job(c,"wordcount");
26         j.setJarByClass(WordCount.class);
27         j.setMapperClass(MapForWordCount.class);
28         j.setReducerClass(ReduceForWordCount.class);
29         j.setOutputKeyClass(Text.class);
30         j.setOutputValueClass(IntWritable.class);
31         FileInputFormat.addInputPath(j, input);
32         FileOutputFormat.setOutputPath(j, output);
33         System.exit(j.waitForCompletion(true)?0:1);
34     }
35 }
```

JAVA Compiler

PHASE 1

- `hdfs dfs -mkdir /count/`
- `hdfs dfs -ls /`
- `hdfs dfs -put /home/cloudera/Downloads/pesquisa.txt /count/search.txt`



```
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ cd Downloads/
[cloudera@quickstart Downloads]$ ls
hadoop.zip
[cloudera@quickstart Downloads]$ unzip hadoop.zip
Archive:  hadoop.zip
  inflating: pesquisa.txt
  inflating: WordCount.java
[cloudera@quickstart Downloads]$ ls
hadoop.zip pesquisa.txt WordCount.java
[cloudera@quickstart Downloads]$
```

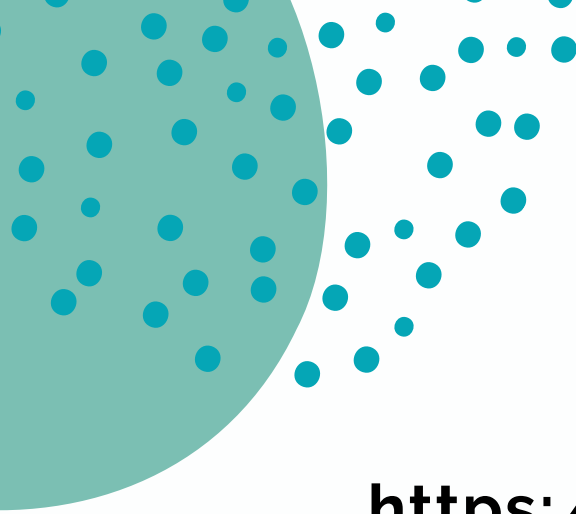
PHASE 2



PHASE 3








```
hadoop jar /home/cloudera/MRProgram.jar PackageDemo.WordCount /count/search.txt /count2  
hdfs dfs -ls /count2  
hdfs dfs -cat /count2/part-r-00000
```



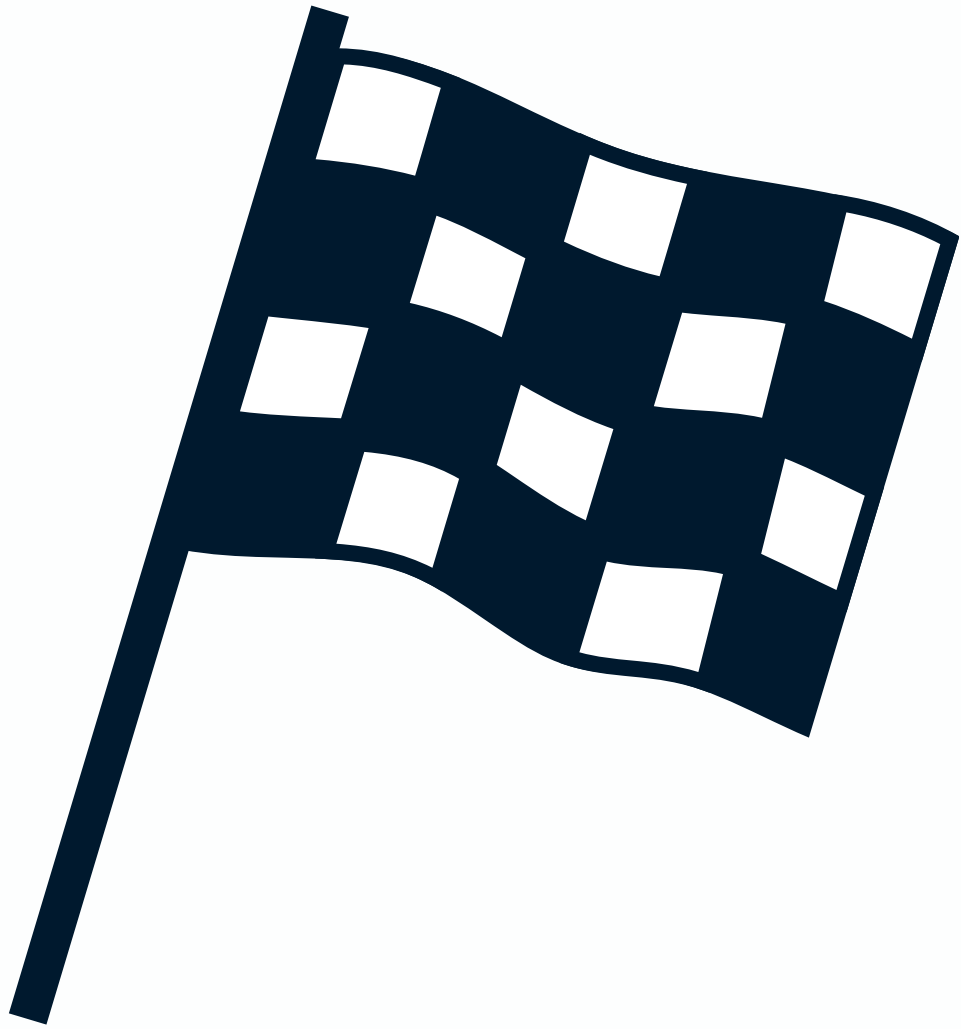


MATERIAL

<https://drive.google.com/drive/folders/1Crg5ZgoN3t86oXuUxAOPmiWlrl8bJ6YU?usp=sharing>

TITLE	LAST MODIFIED
 cloudera-quickstart-vm-5.4.2-0-virtualbox.ovf	6/9/15
 Gravação do ecrã 2020-11-06, às 21.11.01.mov	5:24 am
 Gravação do ecrã 2020-11-07, às 12.32.37.mov	5:24 am
 Gravação do ecrã 2020-11-07, às 13.11.24.mov	5:15 am
 hadoop.zip	Nov 5
 VirtualBox-6.1.16-140961-OSX-1.dmg	Nov 6
 words	5:47 am

CHALLENGE



Do the same
exercise but with
dataset words.txt

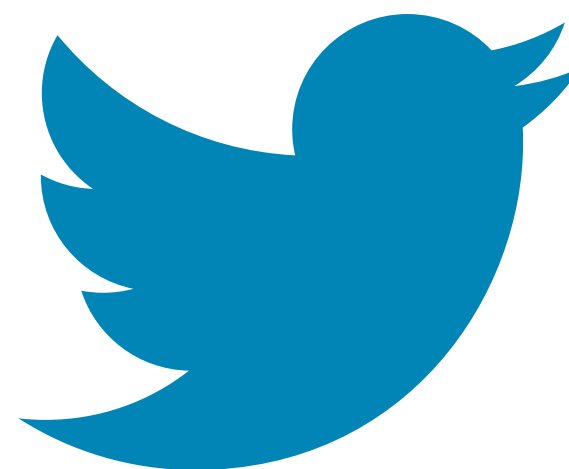


BIBLIOGRAPHY

- Luc Perkins, Eric Redmond, Jim Wilson, Seven Databases in Seven Weeks - A Guide to Modern Databases and the NoSQL Movement, Pragmatic Bookshelf, 2018.
- Connolly, T., Begg, C., Database Systems, A Practical Approach to Design, Implementation, and Management , Addison-Wesley, 6a Edição, 2014.
- <https://www.mysql.com/>
- <https://docs.microsoft.com/pt-pt/>
- <https://www.geeksforgeeks.org/data-replication-in-dbms/>
- https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781786461070
- <https://neo4j.com/blog/acid-vs-base-consistency-models-explained/>
- GIL, David; SONG, Il-Yeol. Modeling and management of big data: challenges and opportunities. 2016.
- RIBEIRO, André; SILVA, Afonso; DA SILVA, Alberto Rodrigues. Data modeling and data analytics: a survey from a big data perspective. Journal of Software Engineering and Applications, 2015, 8.12: 617.
- <https://www.informatica.com/pt/products/big-data/big-data-edition.html>
- https://www.sas.com/en_ae/insights/articles/data-management/Big-data-management-5-things-you-need-to-know.html
- <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle>



FACEBOOK



TWITTER



LINKEDIN



CENTROALGORITMI