

Engenharia de Dados com Hadoop e Spark



Data Science Academy



Data Science Academy

Bem-vindo



Engenharia de Dados com Hadoop e Spark



Engenharia de Dados com Hadoop e Spark

Cluster
Hadoop

Capítulos
2, 3 e 4

Armazenamento
de Dados
Capítulos
5, 6 e 7

Machine
Learning

Capítulo
8

Hadoop e
Spark

Capítulo
9



Data Science Academy

Engenharia de Dados com Hadoop e Spark

O que você vai aprender neste curso?

Conceitos e definições de Big Data, Hadoop, Ecosystema Hadoop e Spark

Como planejar, instalar e configurar um cluster Hadoop

Como planejar, instalar e configurar o Ecosystema Hadoop (Hive, Hbase, Zookeeper, Flume, Oozie, Ambari, Sqoop, Spark e Storm)

Configuração e utilização do HDFS e configurações avançadas do cluster Hadoop

Administração e Manutenção do Hadoop e Spark



Engenharia de Dados com Hadoop e Spark

O que você vai aprender neste curso?



Engenharia de Dados com Hadoop e Spark

E quais são os pré-requisitos?

Curso Big Data Fundamentos

Conhecimentos básicos de sistema operacional Linux (desejável)

Conhecimentos básicos de linguagem de programação (desejável)

Muita vontade de aprender e entrar no mundo do Big Data (mandatório)



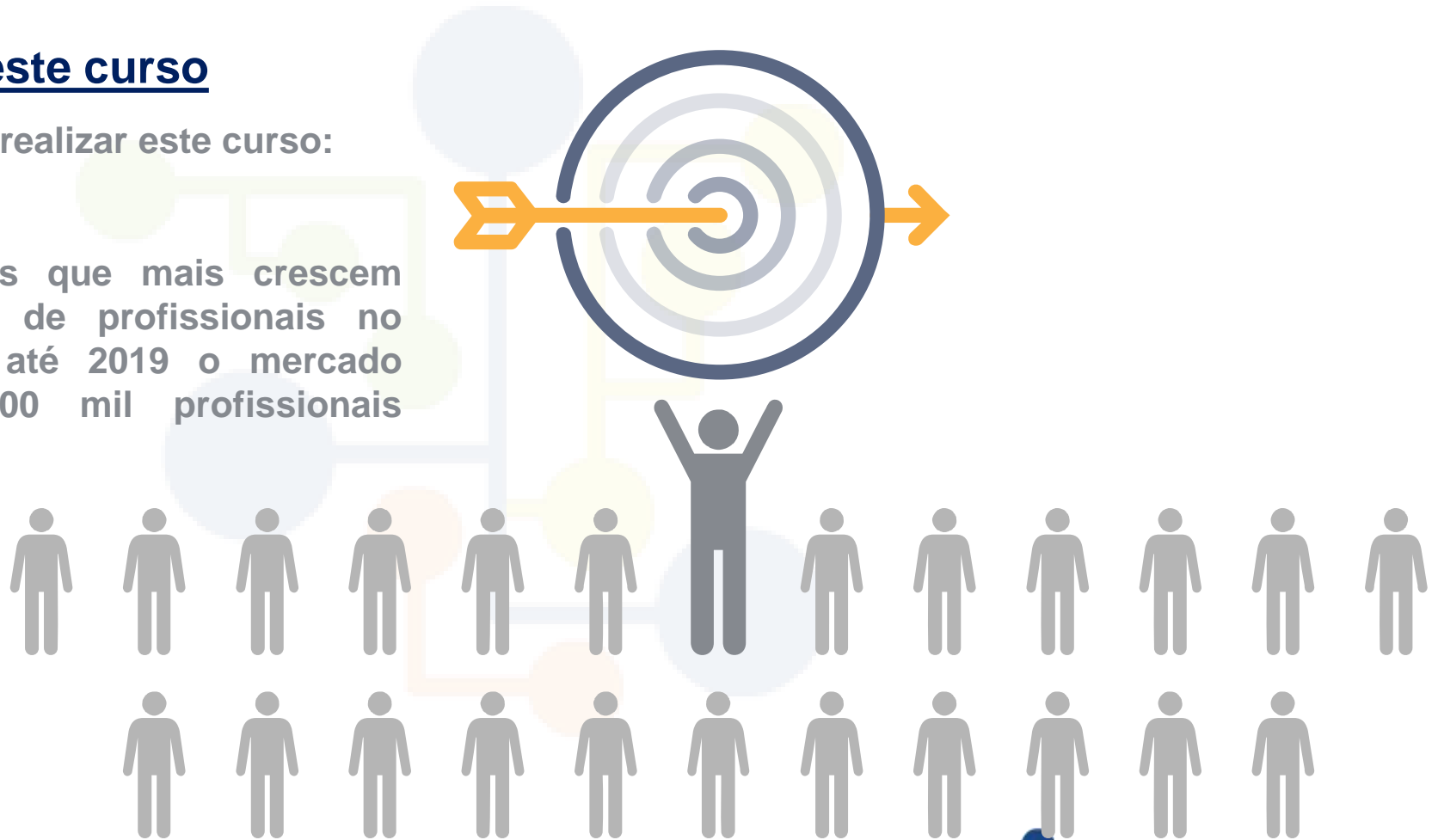
Data Science Academy

Engenharia de Dados com Hadoop e Spark

Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

Big Data é uma das áreas que mais crescem atualmente. Há um déficit de profissionais no mercado e estima-se que até 2019 o mercado precisará de mais de 200 mil profissionais habilitados em Big Data.



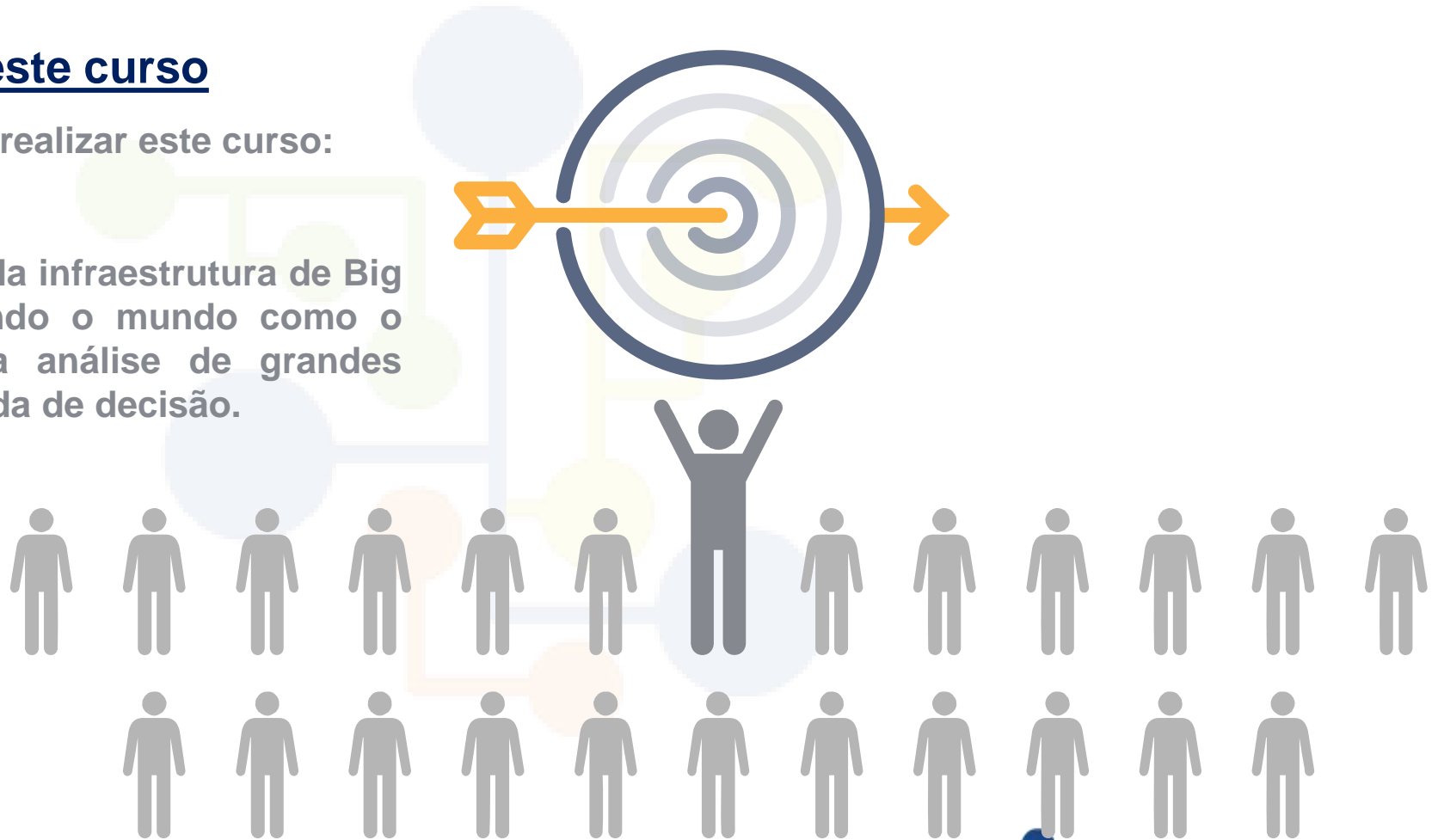
Data Science Academy

Engenharia de Dados com Hadoop e Spark

Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

Hadoop é a tecnologia base da infraestrutura de Big Data, que está revolucionando o mundo como o conhecemos. Ele permite a análise de grandes volumes de dados para tomada de decisão.



Data Science Academy

Engenharia de Dados com Hadoop e Spark

Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

Conhecimento de Hadoop é um dos skills mais procurados por recrutadores de profissionais de Big Data.



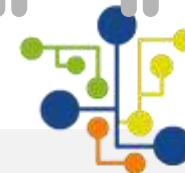
Data Science Academy

Engenharia de Dados com Hadoop e Spark

Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

O Hadoop permite a mudança de paradigma de bancos de dados tradicionais, para um framework de dados versátil, adaptável e veloz.



Data Science Academy

Engenharia de Dados com Hadoop e Spark

Estrutura do Curso

Para tornar sua experiência de aprendizagem ainda mais completa, você terá quizzes e labs ao longo de todos os capítulos.



Data Science Academy

Engenharia de Dados com Hadoop e Spark

Estrutura do Curso

Você também terá acesso e poderá fazer o download dos e-books com todo o passo-a-passo de cada lab realizado ao longo do curso.



Data Science Academy

Engenharia de Dados com Hadoop e Spark

Estrutura do Curso

Fique tranquilo se você não possui experiência em sistema operacional Linux. Tudo será explicado em detalhes.



Data Science Academy

Engenharia de Dados com Hadoop e Spark

Projetos



Data Science Academy

Engenharia de Dados com Hadoop e Spark

Avaliação Final



Data Science Academy

Engenharia de Dados com Hadoop e Spark

Bonus



Data Science Academy

Engenharia de Dados com Hadoop e Spark



Data Science Academy

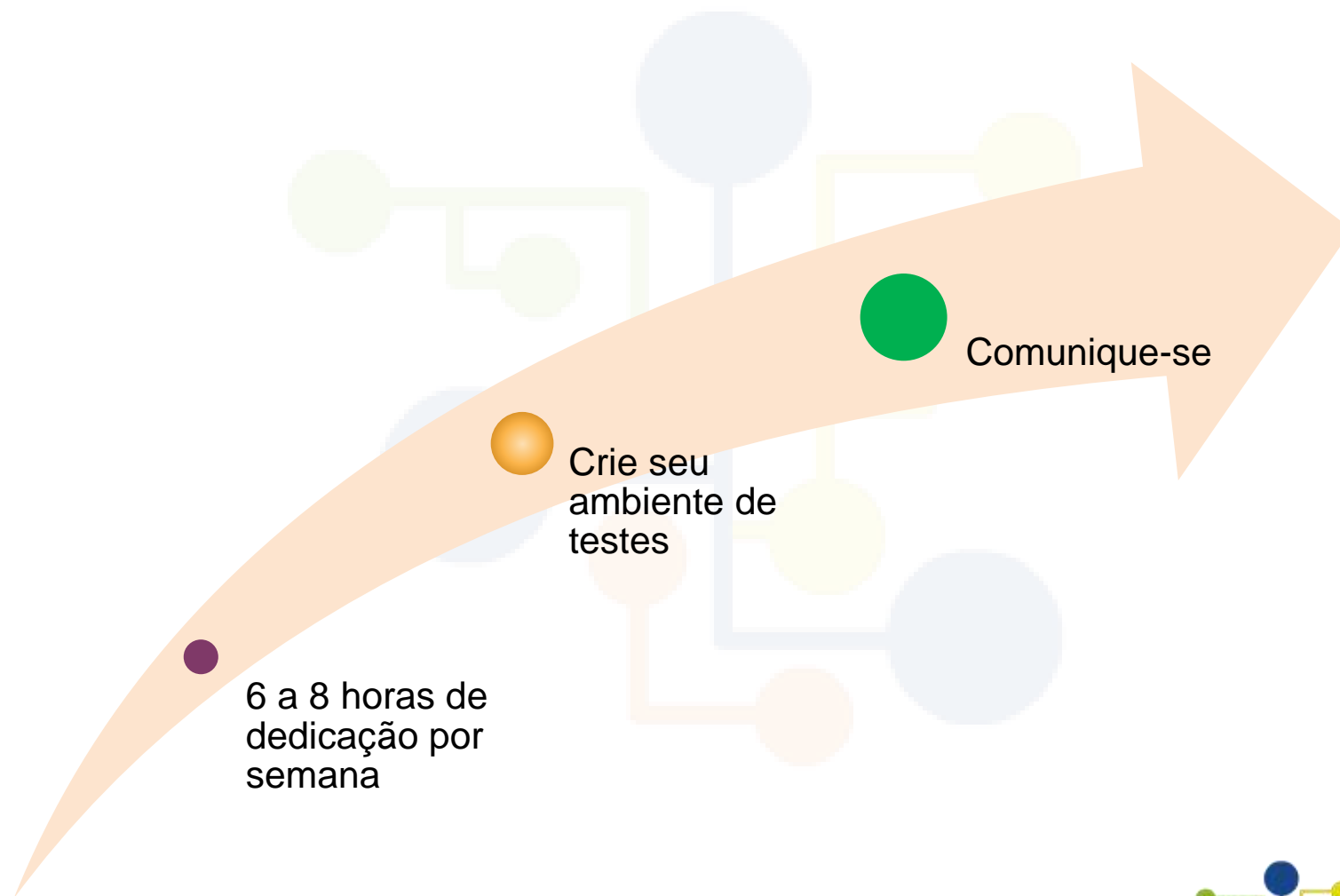
Engenharia de Dados com Hadoop e Spark

O SEGREDO
do seu sucesso
esta na constância
do seu ESFORÇO



Data Science Academy

Engenharia de Dados com Hadoop e Spark



Data Science Academy

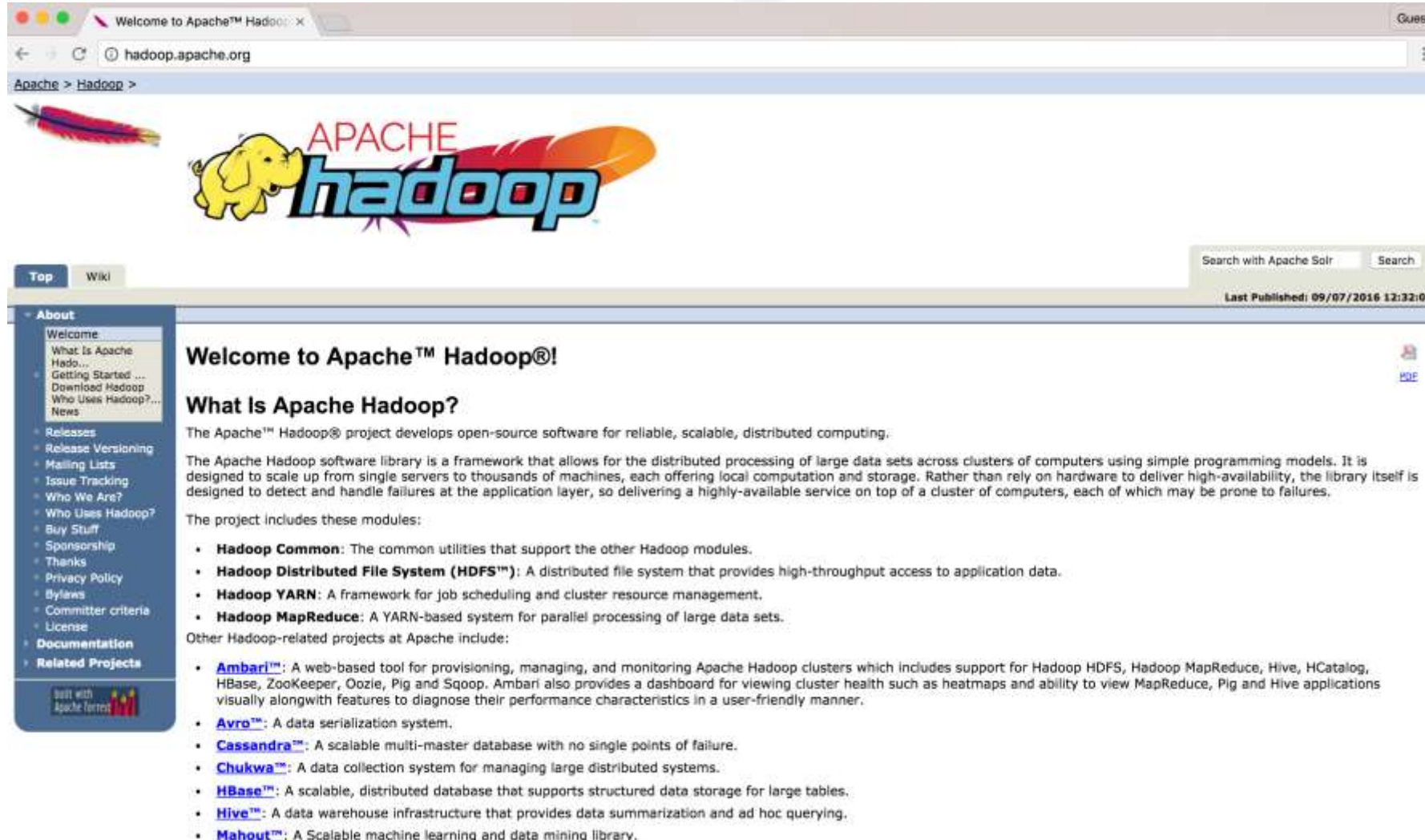
Engenharia de Dados com Hadoop e Spark

Fontes de Informação



Data Science Academy

Engenharia de Dados com Hadoop e Spark



The screenshot shows the Apache Hadoop website in a web browser. The browser's address bar displays 'hadoop.apache.org'. The page features the Apache Hadoop logo, which includes a yellow elephant and the text 'APACHE hadoop'. Below the logo, there is a search bar and a 'Last Published' timestamp of '09/07/2016 12:32:06'. The main content area is titled 'Welcome to Apache™ Hadoop®!' and 'What Is Apache Hadoop?'. It describes the project as open-source software for reliable, scalable, distributed computing. The page lists several modules: Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce. It also lists other Hadoop-related projects at Apache, including Ambari, Avro, Cassandra, Chukwa, HBase, Hive, and Mahout. A sidebar on the left contains links to 'About', 'Releases', 'Documentation', and 'Related Projects'.

Welcome to Apache™ Hadoop®!

What Is Apache Hadoop?

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects at Apache include:

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A Scalable machine learning and data mining library.

Engenharia de Dados com Hadoop e Spark

The screenshot shows the Cloudera website's product page for Apache Hadoop. The page includes a navigation bar with links to Downloads, Training, Support Portal, Partners, Developers, and Community. The main content area features the Cloudera logo and a navigation menu with links to Why Cloudera, Products, Services & Support, Solutions, and Get Started. The Apache Hadoop section contains a descriptive paragraph about the ecosystem and a 'Try now' button. To the right, a diagram illustrates the Hadoop ecosystem architecture, organized into layers: PROCESS, ANALYZE & SERVE (containing BATCH, SQL, STREAM, SEARCH, and SDK); UNIFIED SERVICES (containing RESOURCE MANAGEMENT and SECURITY); STORE (containing FILESYSTEM, RELATIONAL, and NoSQL); and INTEGRATE (containing STRUCTURED and UNSTRUCTURED data types).

Apache Hadoop - Cloudera

Guest

https://www.cloudera.com/products/apache-hadoop.html

Downloads Training Support Portal Partners Developers Community

Search Sign in Language

cloudera

Why Cloudera Products Services & Support Solutions Get Started

Apache Hadoop

Hadoop is an ecosystem of open source components that fundamentally changes the way enterprises store, process, and analyze data. Unlike traditional systems, Hadoop enables multiple types of analytic workloads to run on the same data, at the same time, at massive scale on industry-standard hardware. CDH, Cloudera's open source platform, is the most popular distribution of Hadoop and related projects in the world (with support available via a Cloudera Enterprise subscription).

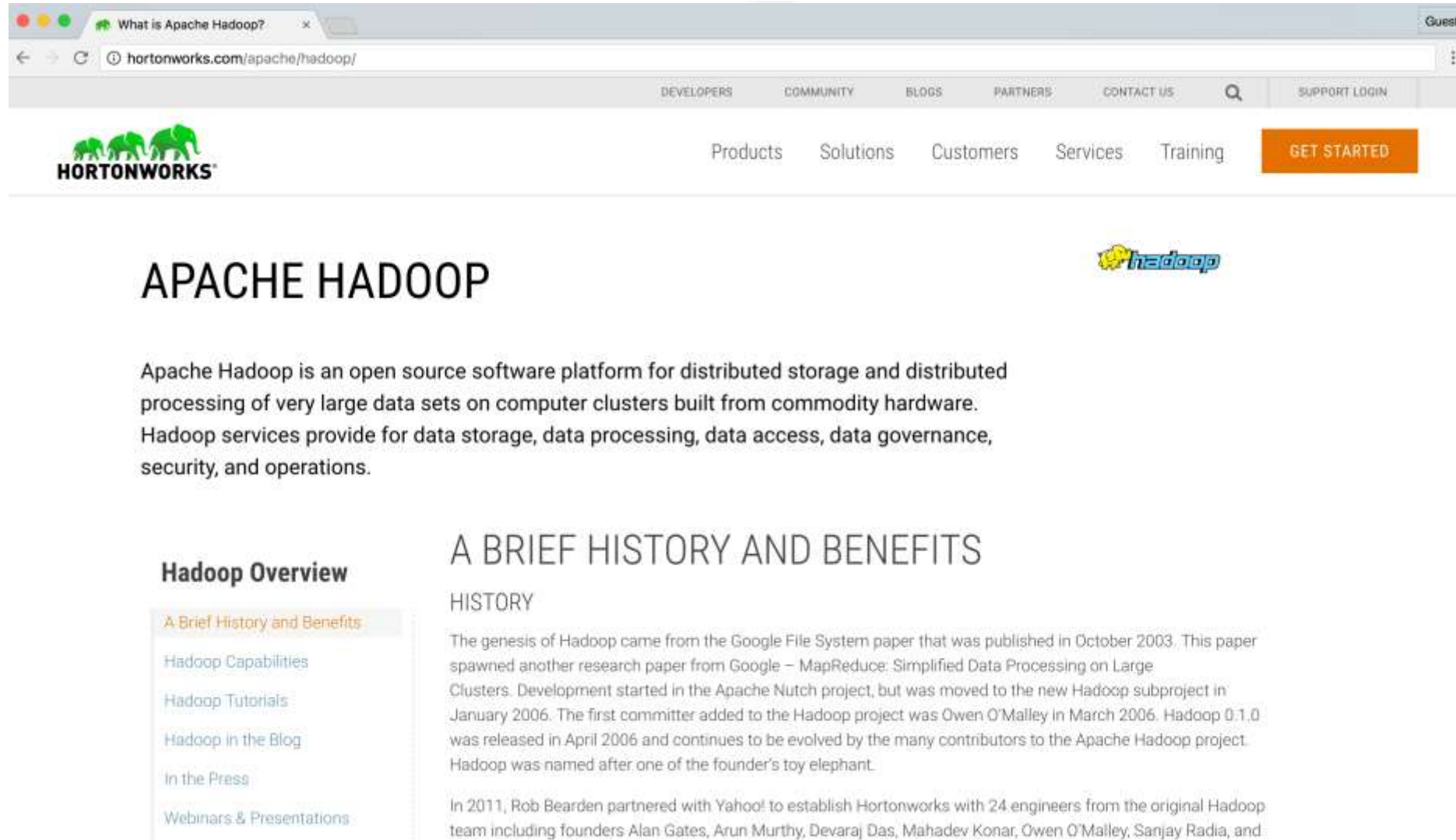
[Try now >](#)

[Hadoop ecosystem in the Engineering Blog >](#)

The diagram illustrates the Hadoop ecosystem architecture, organized into layers:

- PROCESS, ANALYZE & SERVE**
 - BATCH: Spark, Hive, MapReduce
 - SQL: Impala
 - STREAM: Spark
 - SEARCH: Solr
 - SDK: Kite
- UNIFIED SERVICES**
 - RESOURCE MANAGEMENT: YARN
 - SECURITY: Sentry, RecordService
- STORE**
 - FILESYSTEM: HDFS
 - RELATIONAL: Kudu
 - NoSQL: HBase
- INTEGRATE**
 - STRUCTURED: Sqoop
 - UNSTRUCTURED: Flume, Kafka

Engenharia de Dados com Hadoop e Spark



APACHE HADOOP

Apache Hadoop is an open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Hadoop services provide for data storage, data processing, data access, data governance, security, and operations.

Hadoop Overview

- [A Brief History and Benefits](#)
- [Hadoop Capabilities](#)
- [Hadoop Tutorials](#)
- [Hadoop in the Blog](#)
- [In the Press](#)
- [Webinars & Presentations](#)

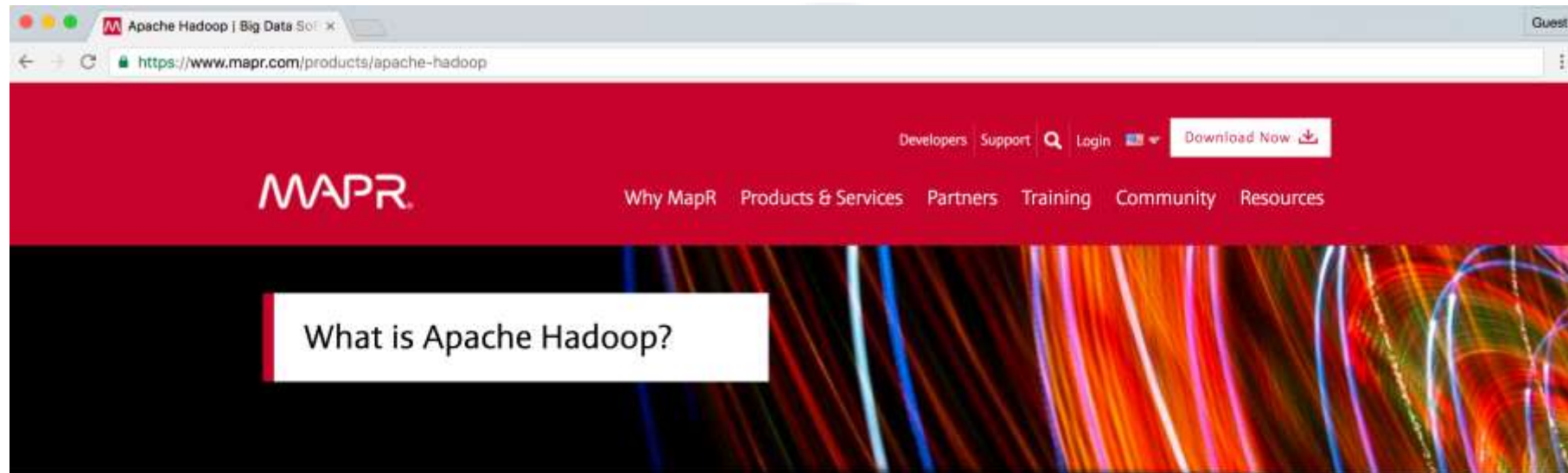
A BRIEF HISTORY AND BENEFITS

HISTORY

The genesis of Hadoop came from the Google File System paper that was published in October 2003. This paper spawned another research paper from Google – MapReduce: Simplified Data Processing on Large Clusters. Development started in the Apache Nutch project, but was moved to the new Hadoop subproject in January 2006. The first committer added to the Hadoop project was Owen O'Malley in March 2006. Hadoop 0.1.0 was released in April 2006 and continues to be evolved by the many contributors to the Apache Hadoop project. Hadoop was named after one of the founder's toy elephant.

In 2011, Rob Bearden partnered with Yahoo! to establish Hortonworks with 24 engineers from the original Hadoop team including founders Alan Gates, Arun Murthy, Devaraj Das, Mahadev Konar, Owen O'Malley, Sanjay Radia, and

Engenharia de Dados com Hadoop e Spark



Hadoop & Big Data

Apache Hadoop™ was born out of a need to process an avalanche of big data. The web was generating more and more information on a daily basis, and it was becoming very difficult to index over one billion pages of content. In order to cope, Google invented a new style of data processing known as MapReduce. A year after Google published a white paper describing the MapReduce framework, Doug Cutting and Mike Cafarella, inspired by the white paper, created Hadoop to apply these concepts to an open-source software framework to support distribution for the Nutch search engine project. Given the original case, Hadoop was designed with a simple write-once storage infrastructure.

Hadoop has moved far beyond its beginnings in web indexing and is now used in many industries for a huge variety of tasks that all share the common theme of lots of variety, volume and velocity of data – both structured and unstructured. It is now widely used **across industries**, including finance, media and entertainment, government, healthcare, information services, retail, and other industries with big data requirements but the limitations of the original storage infrastructure remain.

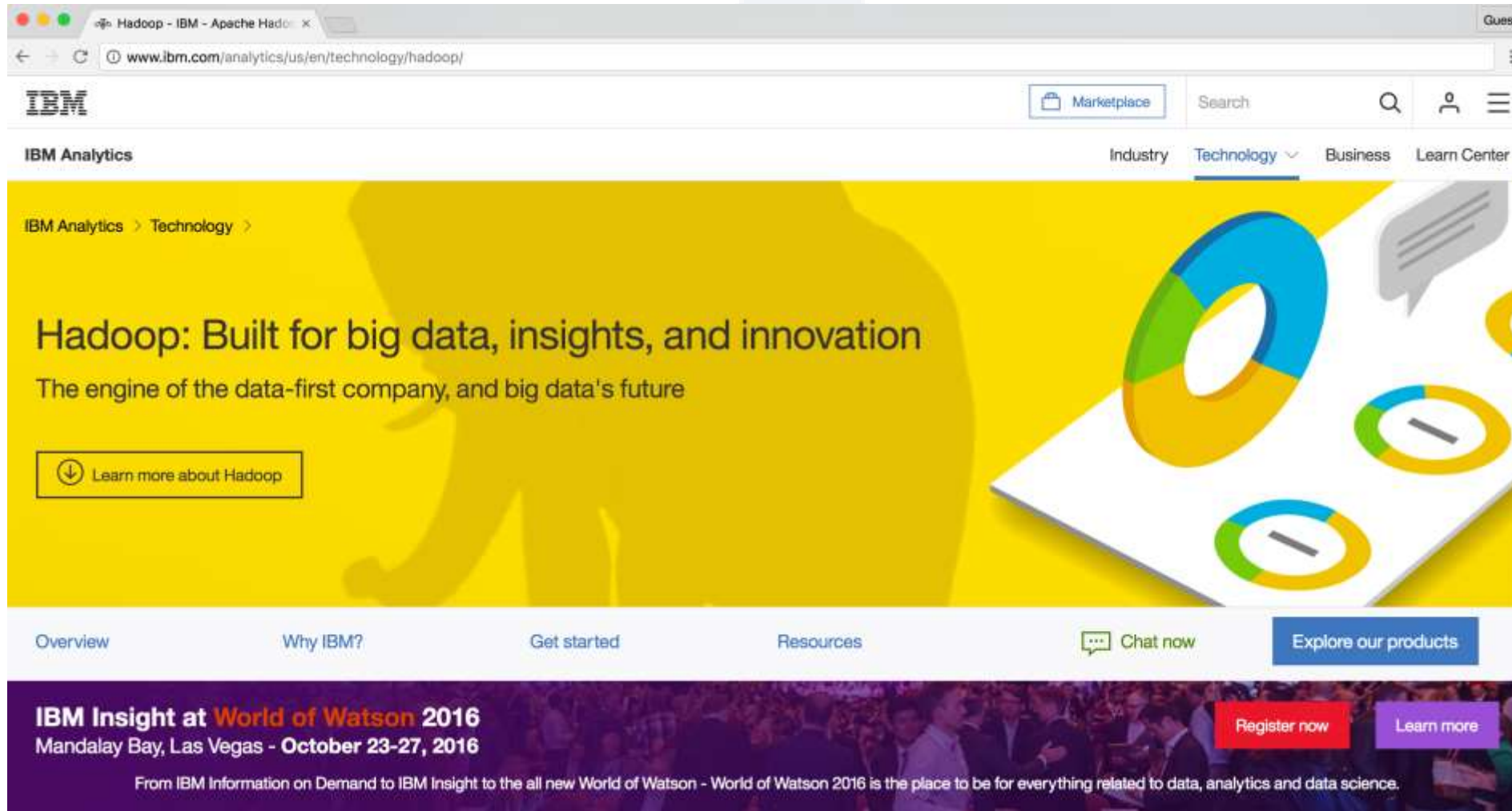
Download

Contact Us

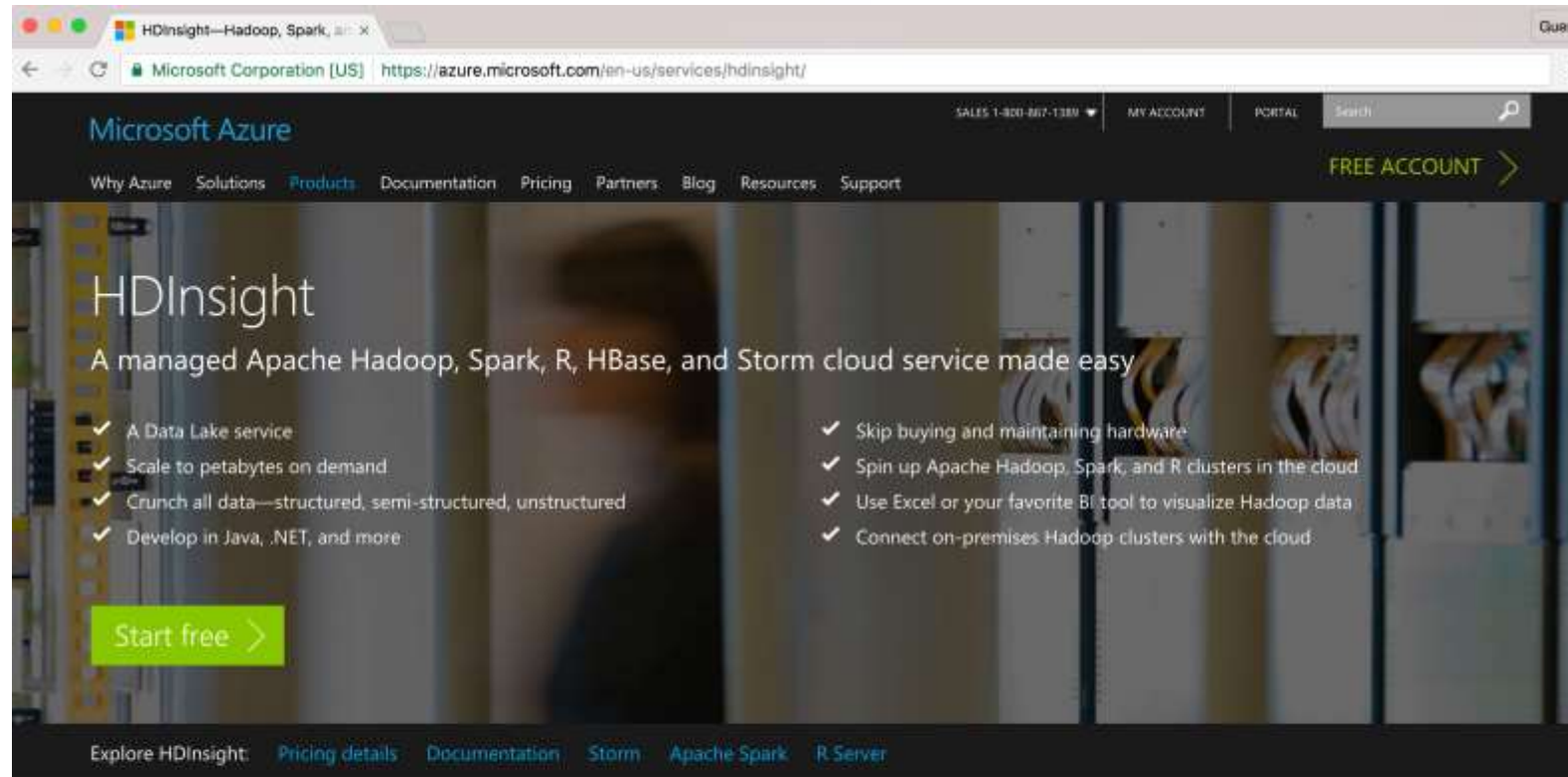
Science Academy



Engenharia de Dados com Hadoop e Spark



Engenharia de Dados com Hadoop e Spark



Comprehensive set of managed Apache big data projects

Batch	Script	SQL	NoSQL	Streaming	In-Memory	Predictive
Map Reduce	Pig	Hive	HBase	Storm	Spark	R Server



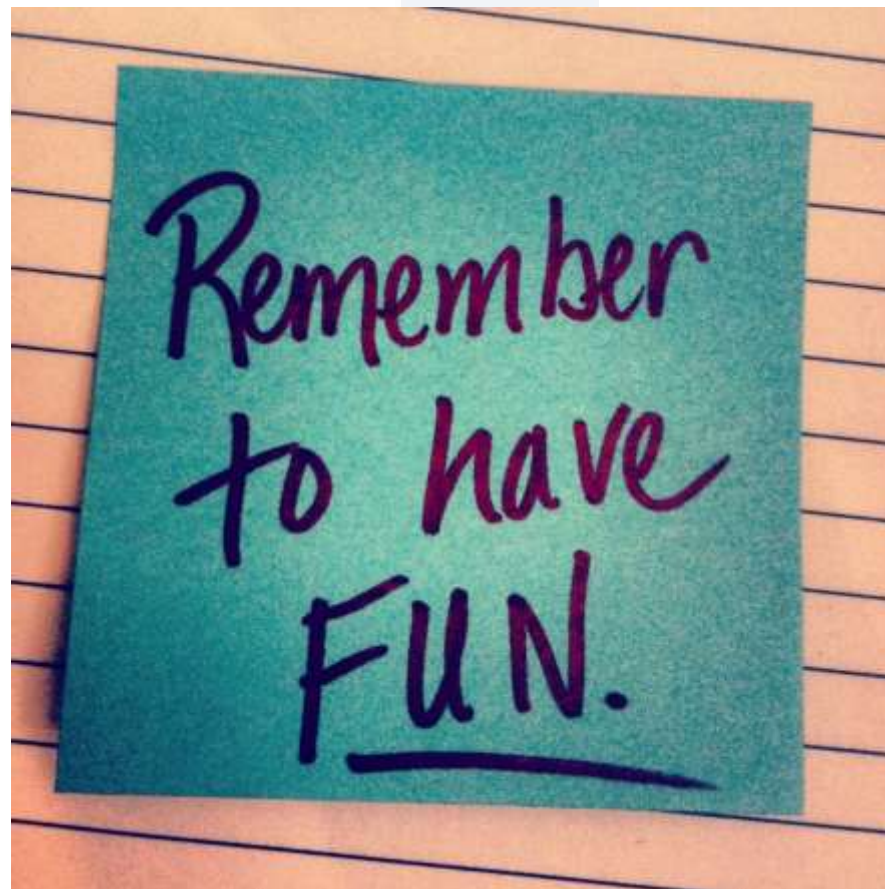
Data Science Academy

Engenharia de Dados com Hadoop e Spark



Data Science Academy

Engenharia de Dados com Hadoop e Spark



Data Science Academy

Obrigado

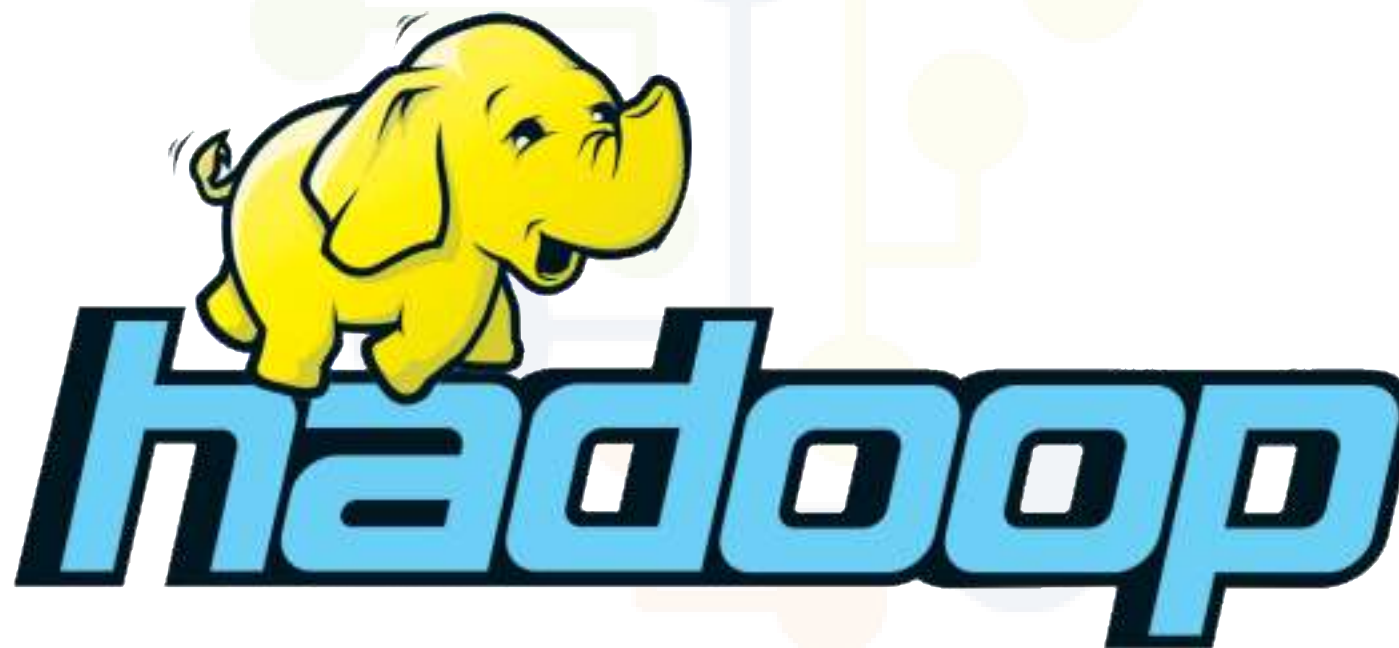




O que é Hadoop?

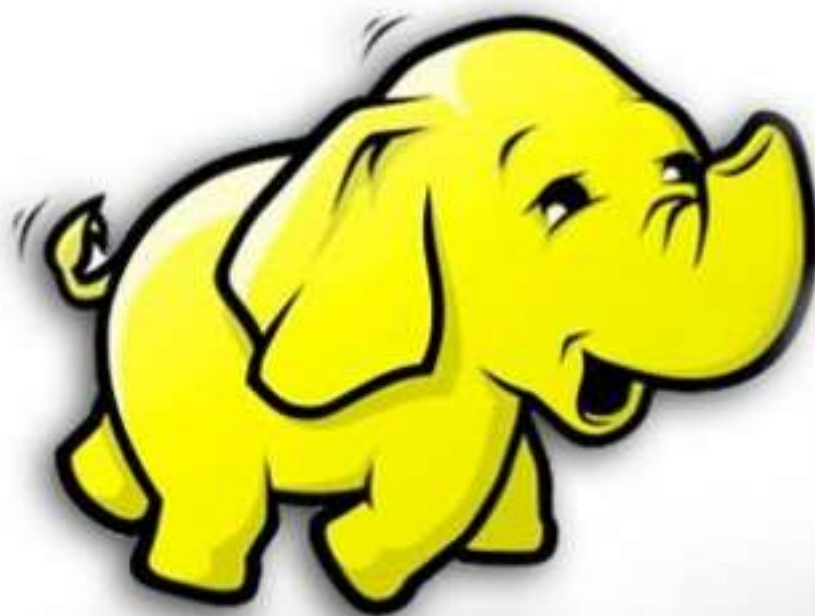


O que é o Hadoop?



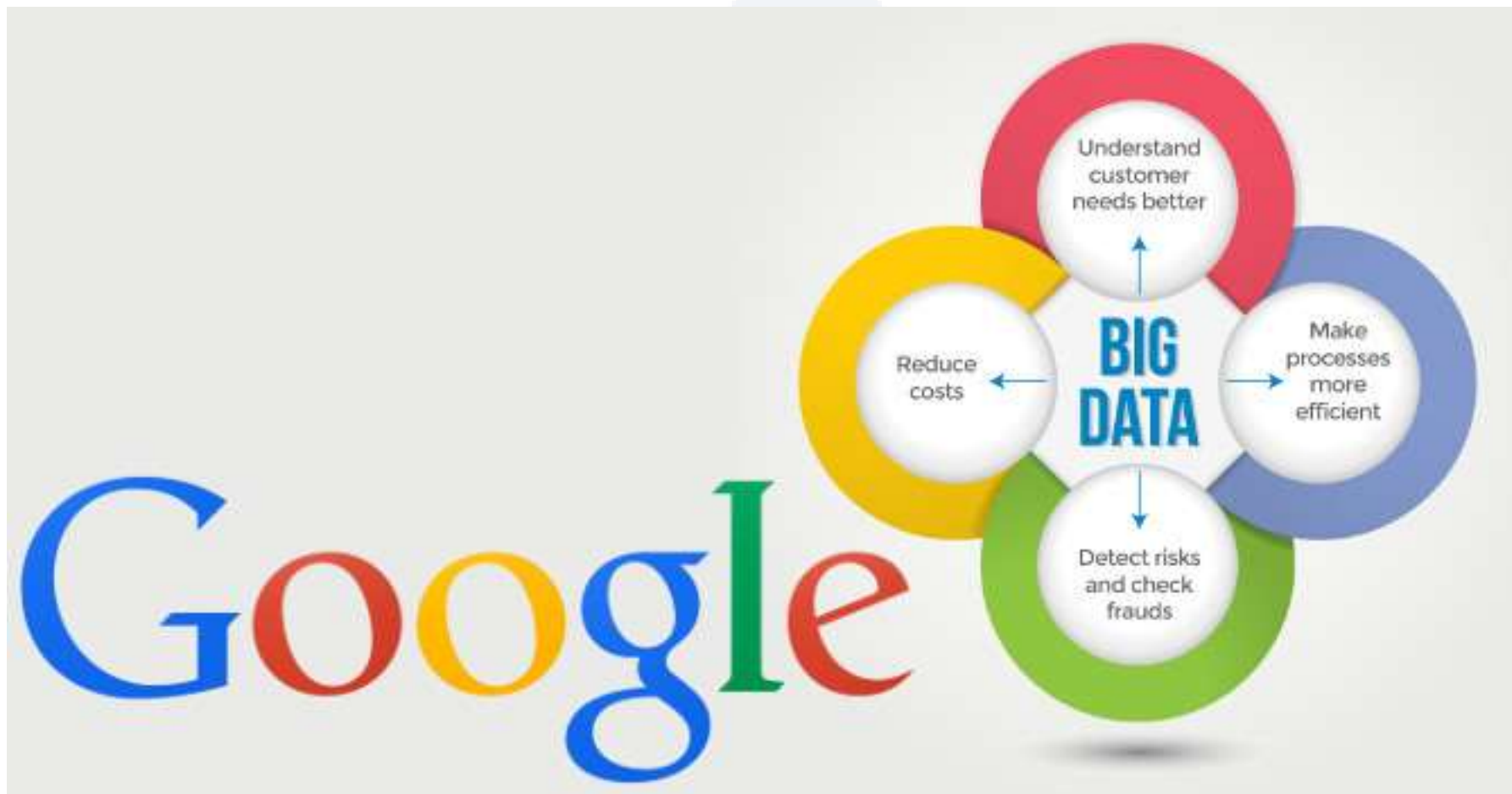
Data Science Academy

O que é o Hadoop?



Data Science Academy

O que é o Hadoop?



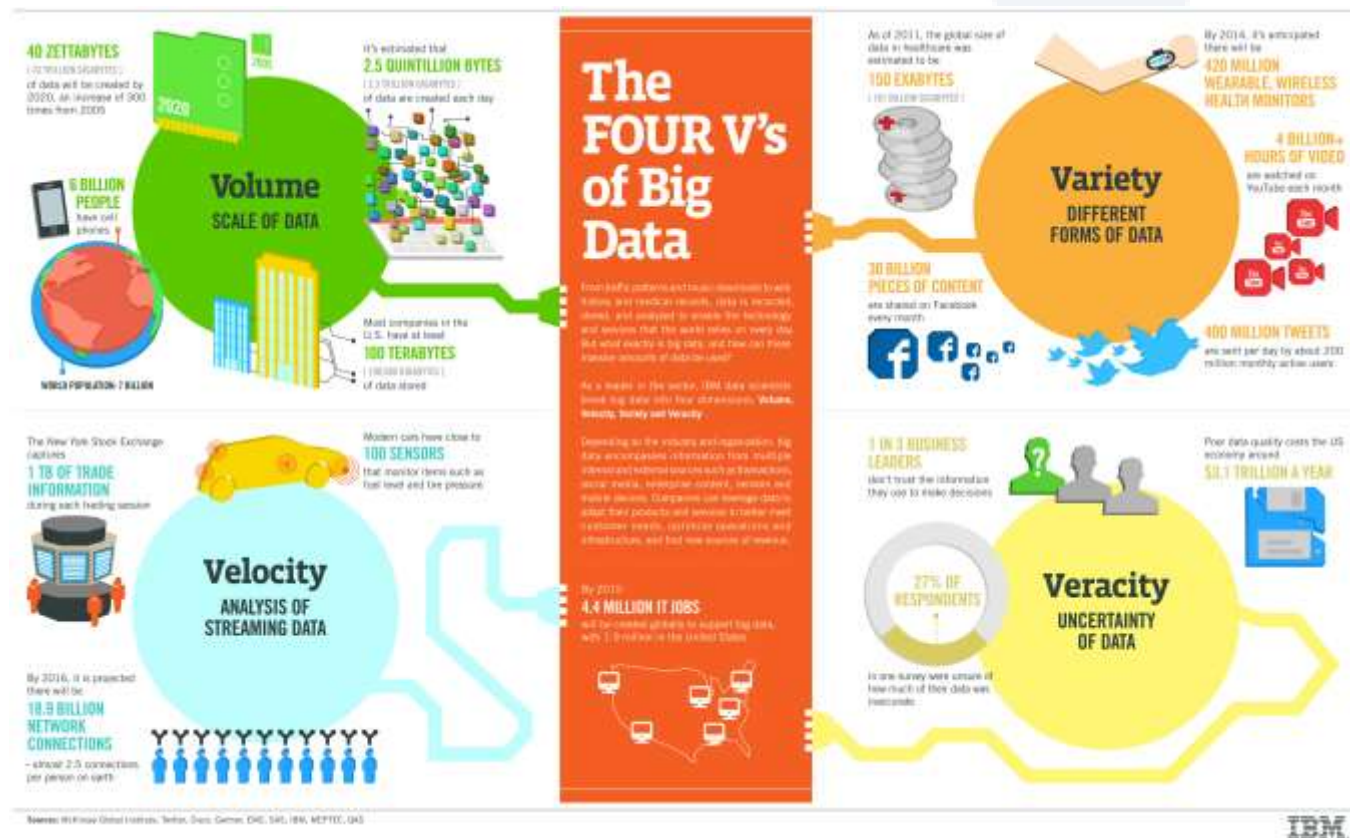
Data Science Academy

O que é o Hadoop?



Data Science Academy

O que é o Hadoop?



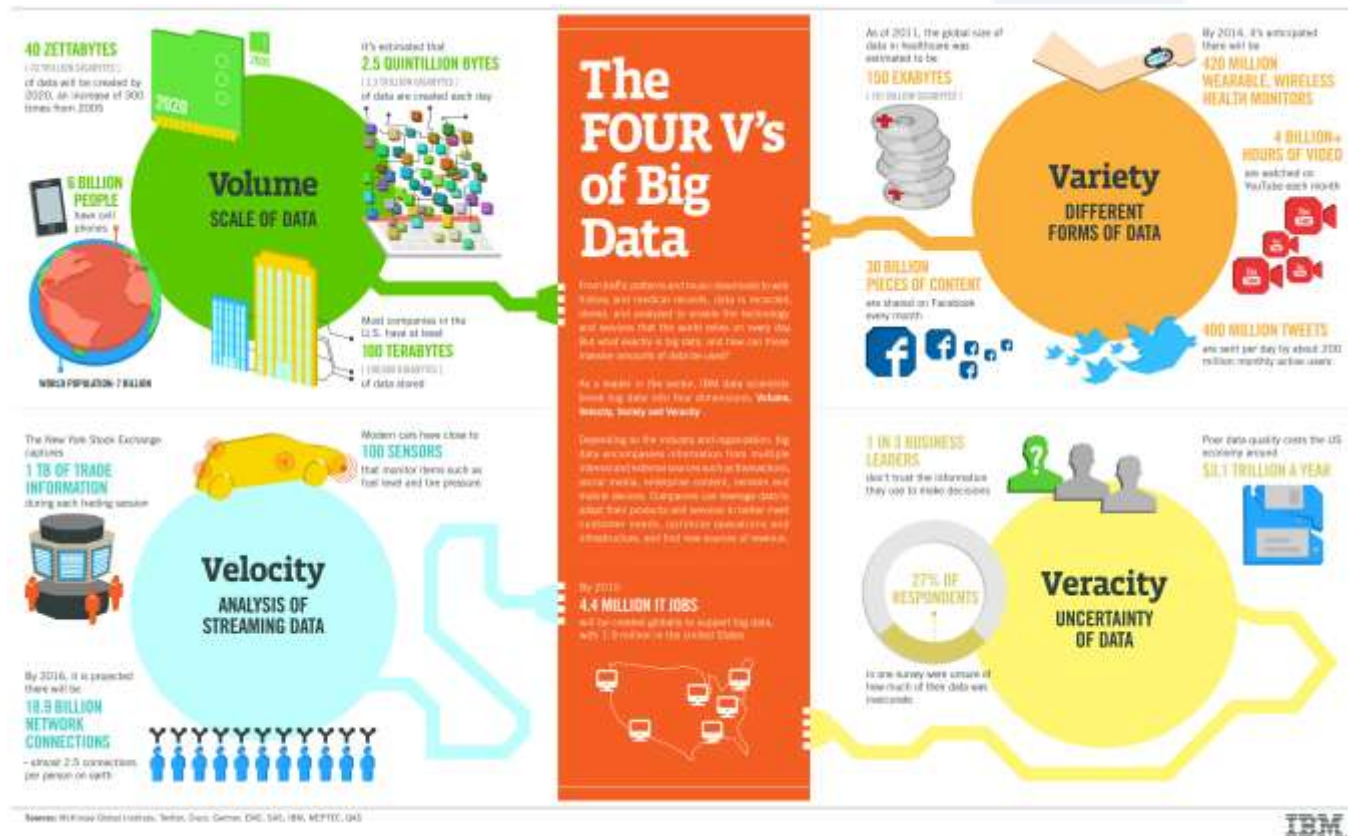
Os 4 V's do Big Data:

- Volume
- Variedade
- Velocidade
- Veracidade



Data Science Academy

O que é o Hadoop?



Os 4 V's do Big Data:

- Volume
- Variedade
- Velocidade
- Veracidade



Data Science Academy

O que é o Hadoop?



O que é o Hadoop?

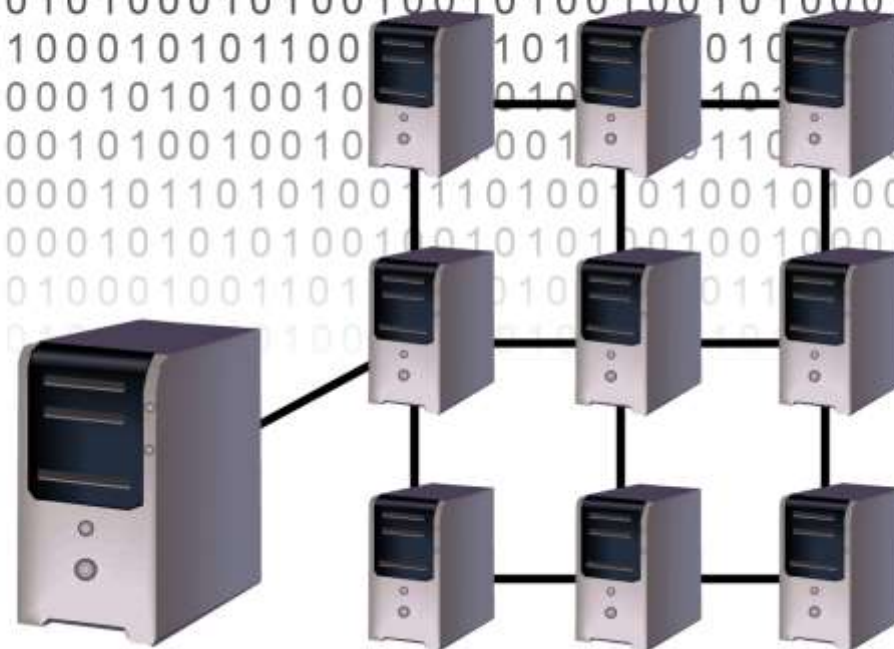
Computação Paralela



Data Science Academy

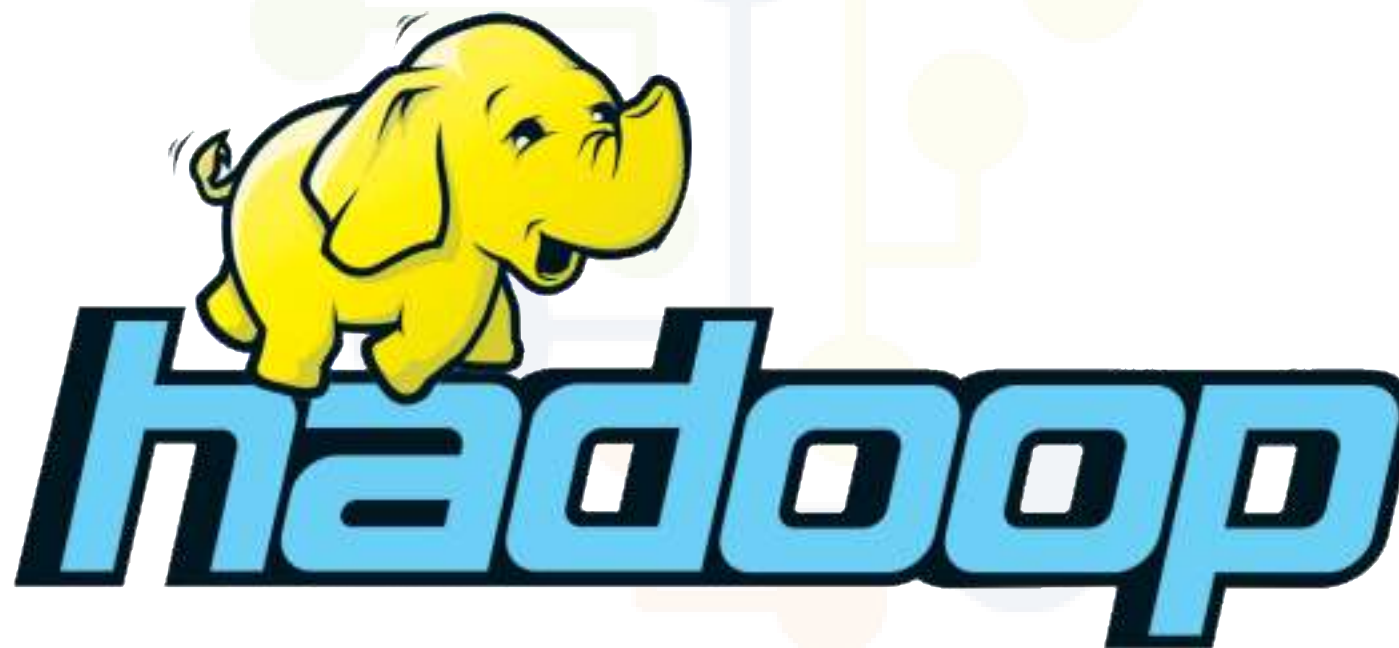
O que é o Hadoop?

Computação Paralela



Data Science Academy

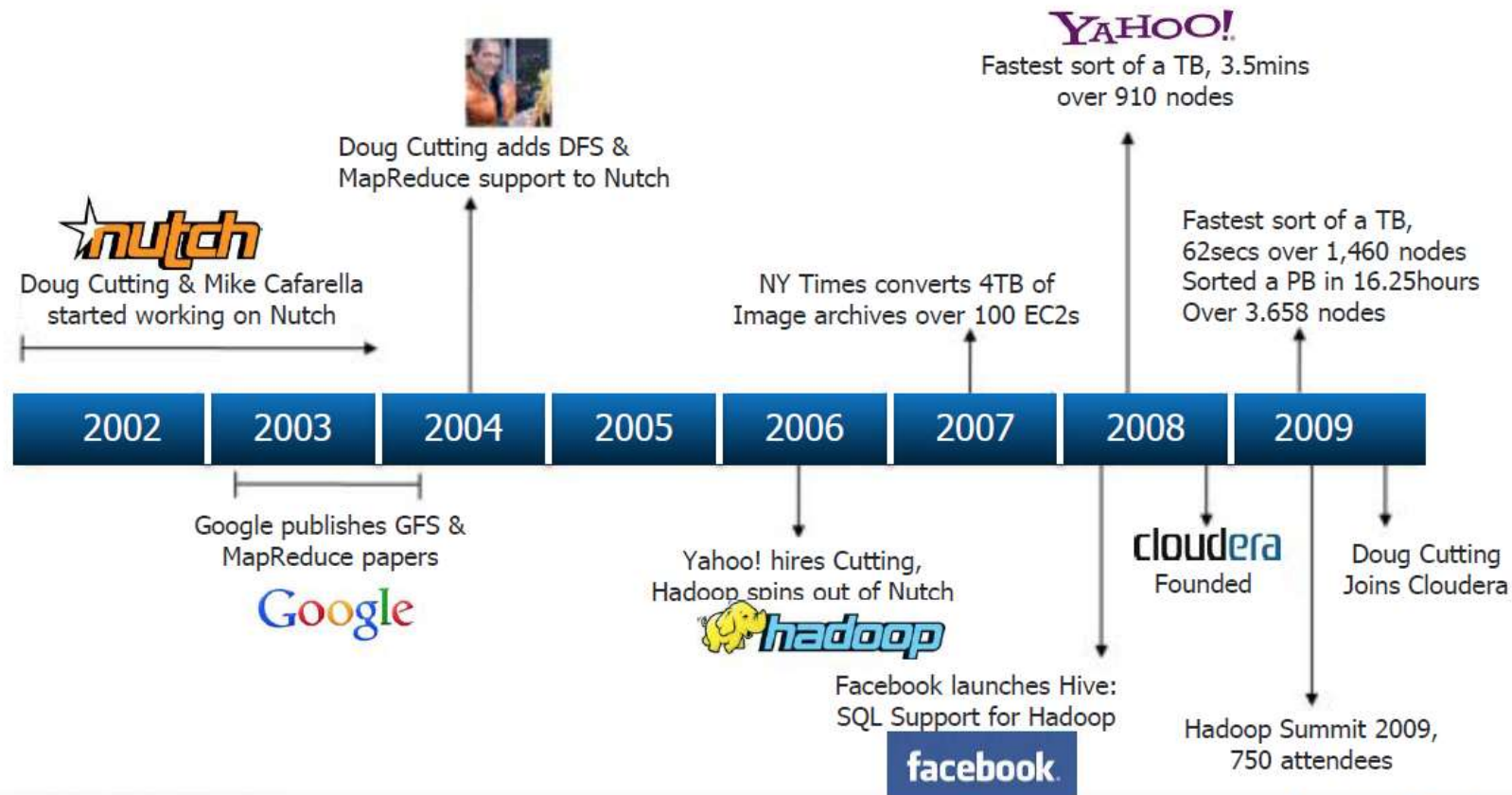
O que é o Hadoop?



Data Science Academy

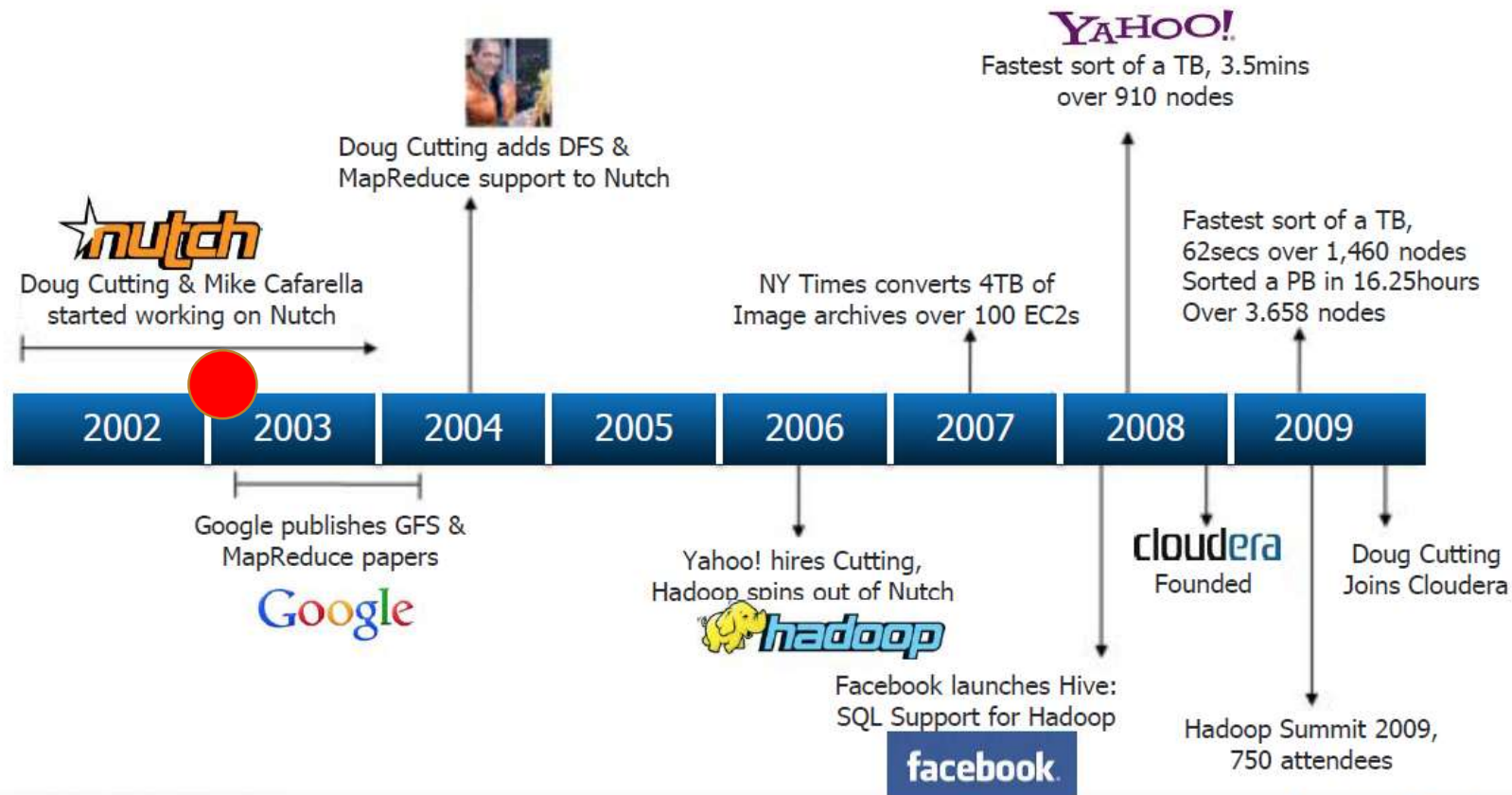
O que é o Hadoop?

Hadoop History



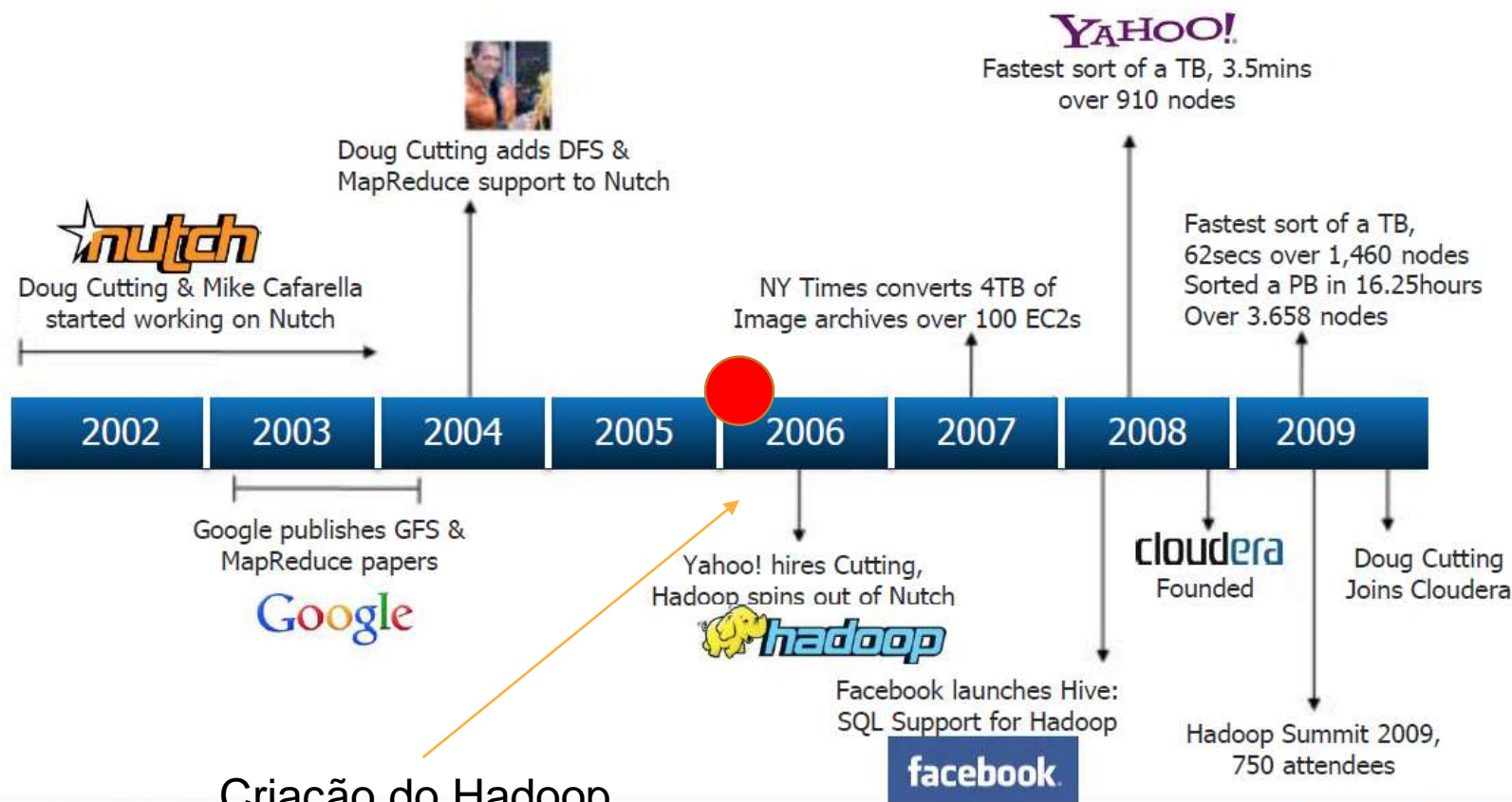
O que é o Hadoop?

Hadoop History



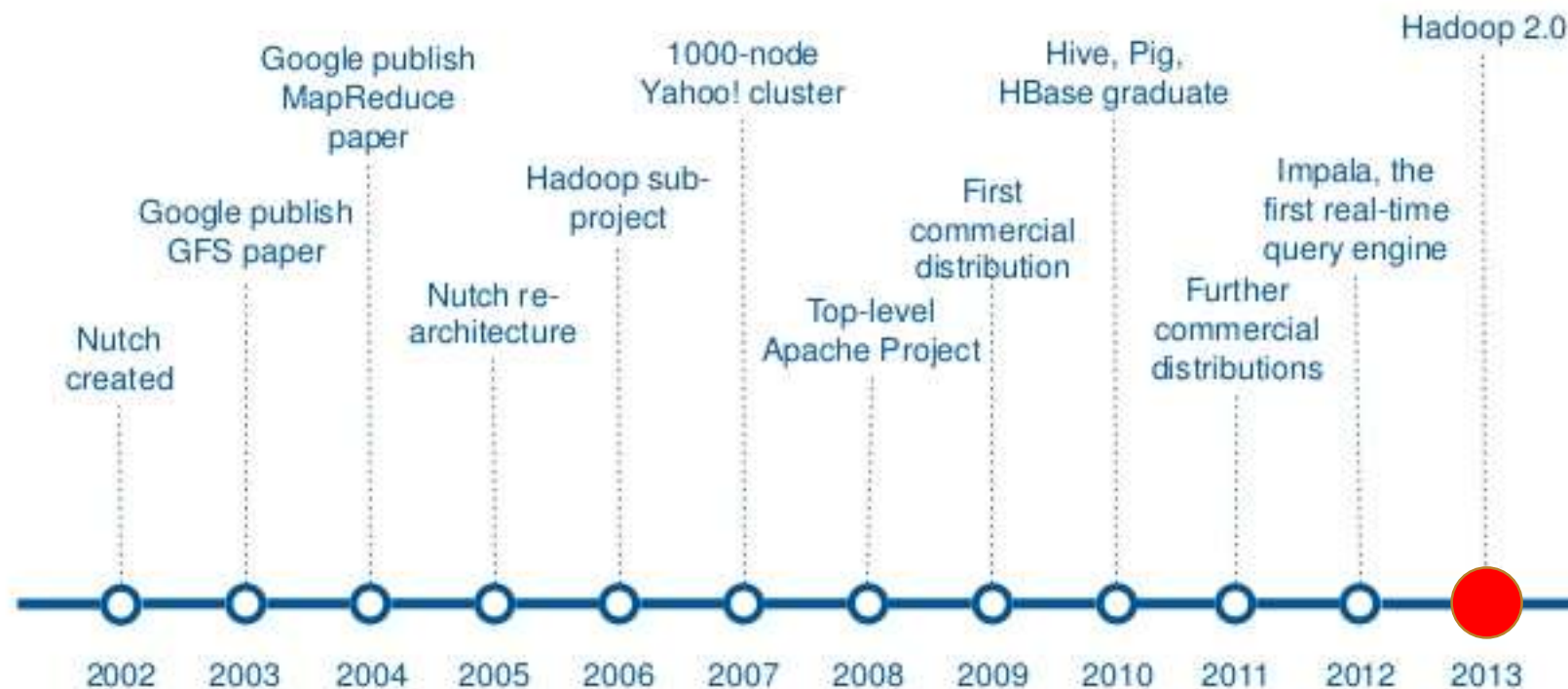
O que é o Hadoop?

Hadoop History



O que é o Hadoop?

A Brief History of Hadoop



Hadoop 2.7.3
25/08/2016

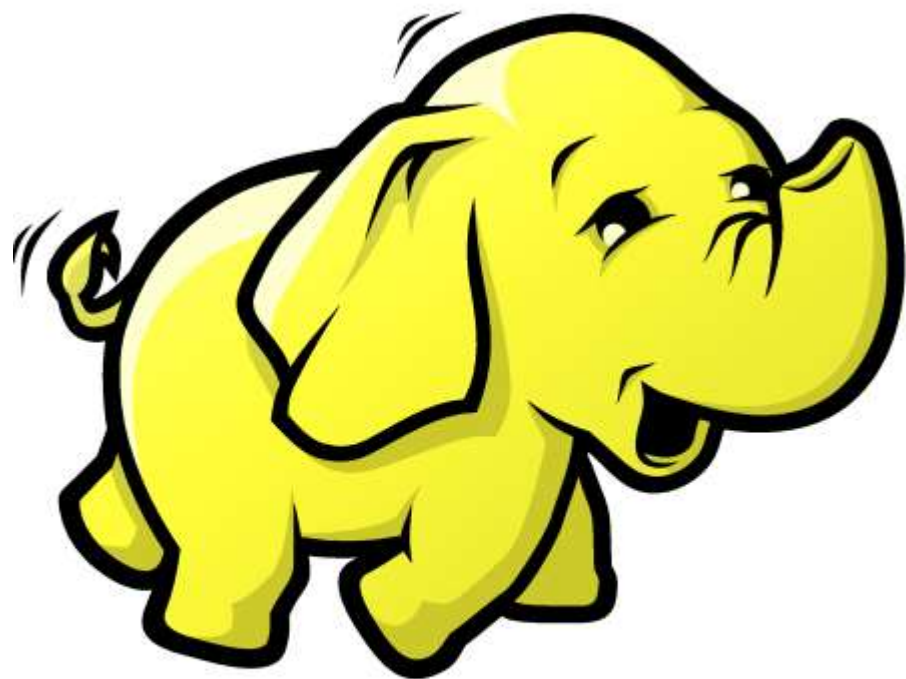
© 2013, Axialdata Systems FZ LLC

4



Data Science Academy

O que é o Hadoop?



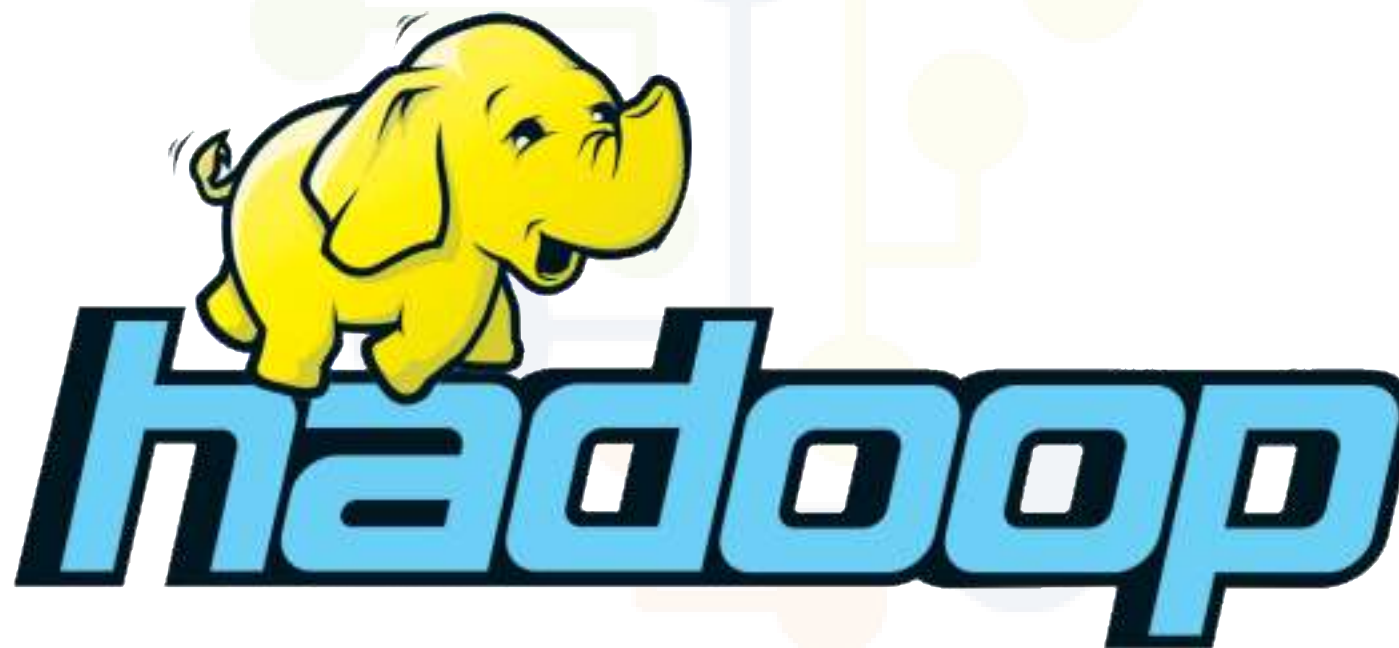
Obrigado



Quais os benefícios para as Empresas ao utilizar o Hadoop?



Benefícios do Hadoop



Data Science Academy

Benefícios do Hadoop

Open Source



Data Science Academy

Benefícios do Hadoop



Economia



Data Science Academy

Benefícios do Hadoop



SCALABILITY

Escalabilidade



Data Science Academy

Benefícios do Hadoop

Robustez



Data Science Academy

Desvantagens do Hadoop

Desvantagens

- Node Master único
- Processamento de arquivos pequenos
- Muito Processamento em Poucos Dados



Obrigado





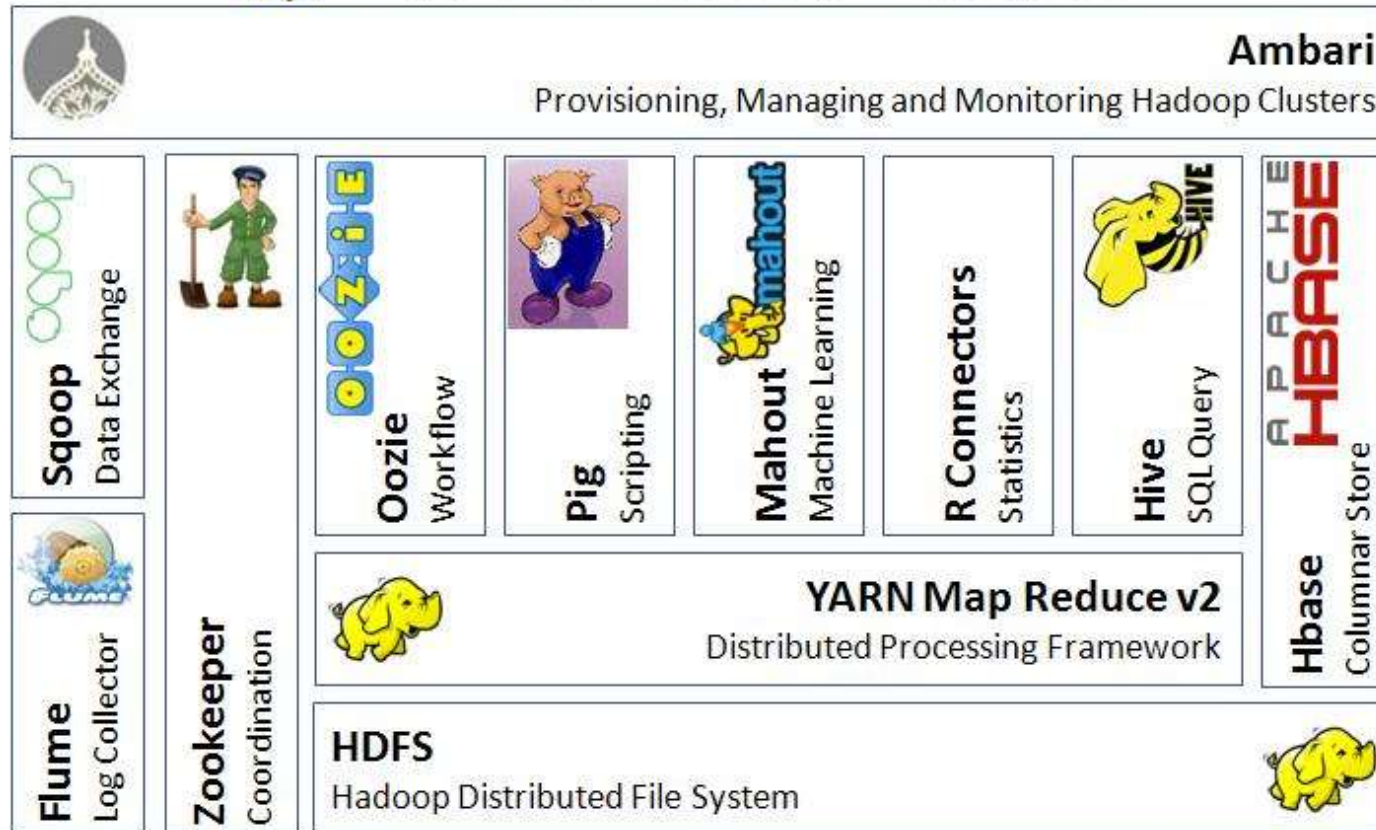
Ecosistema Hadoop



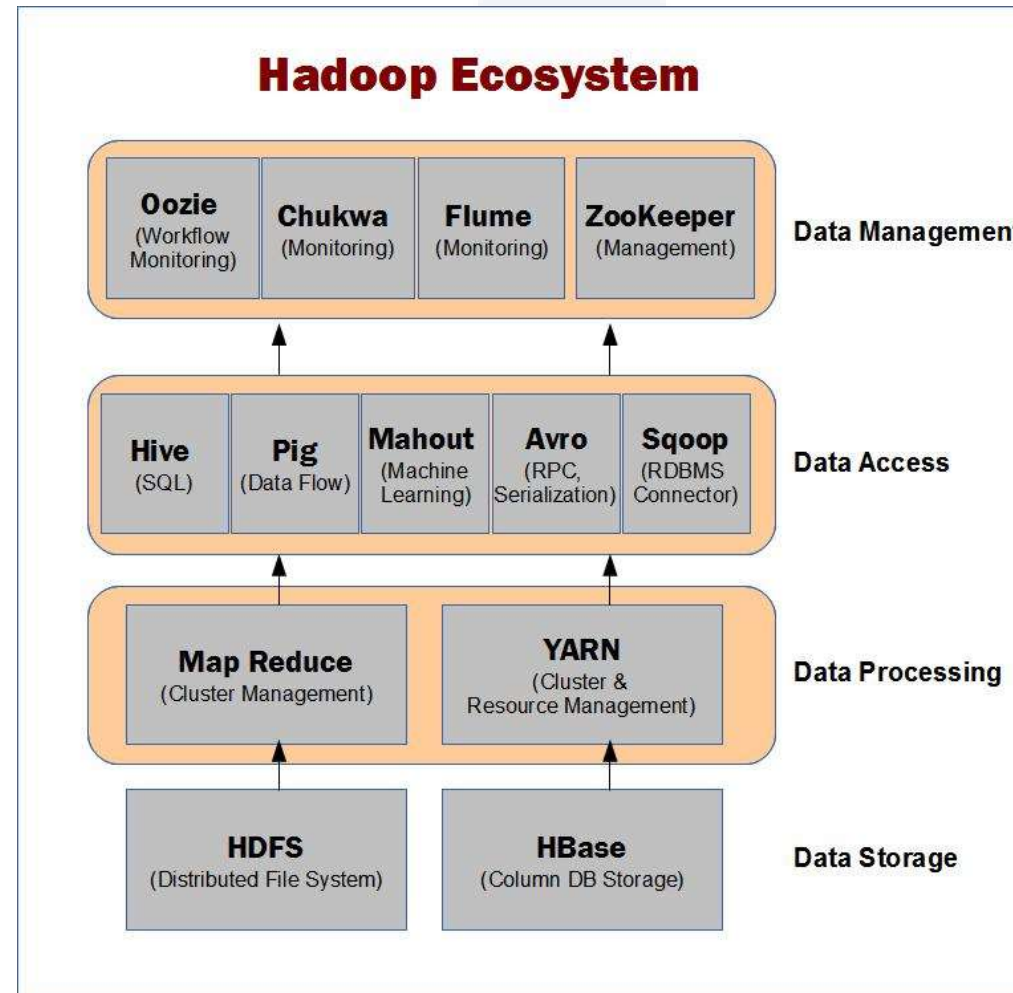
Ecosistema Hadoop



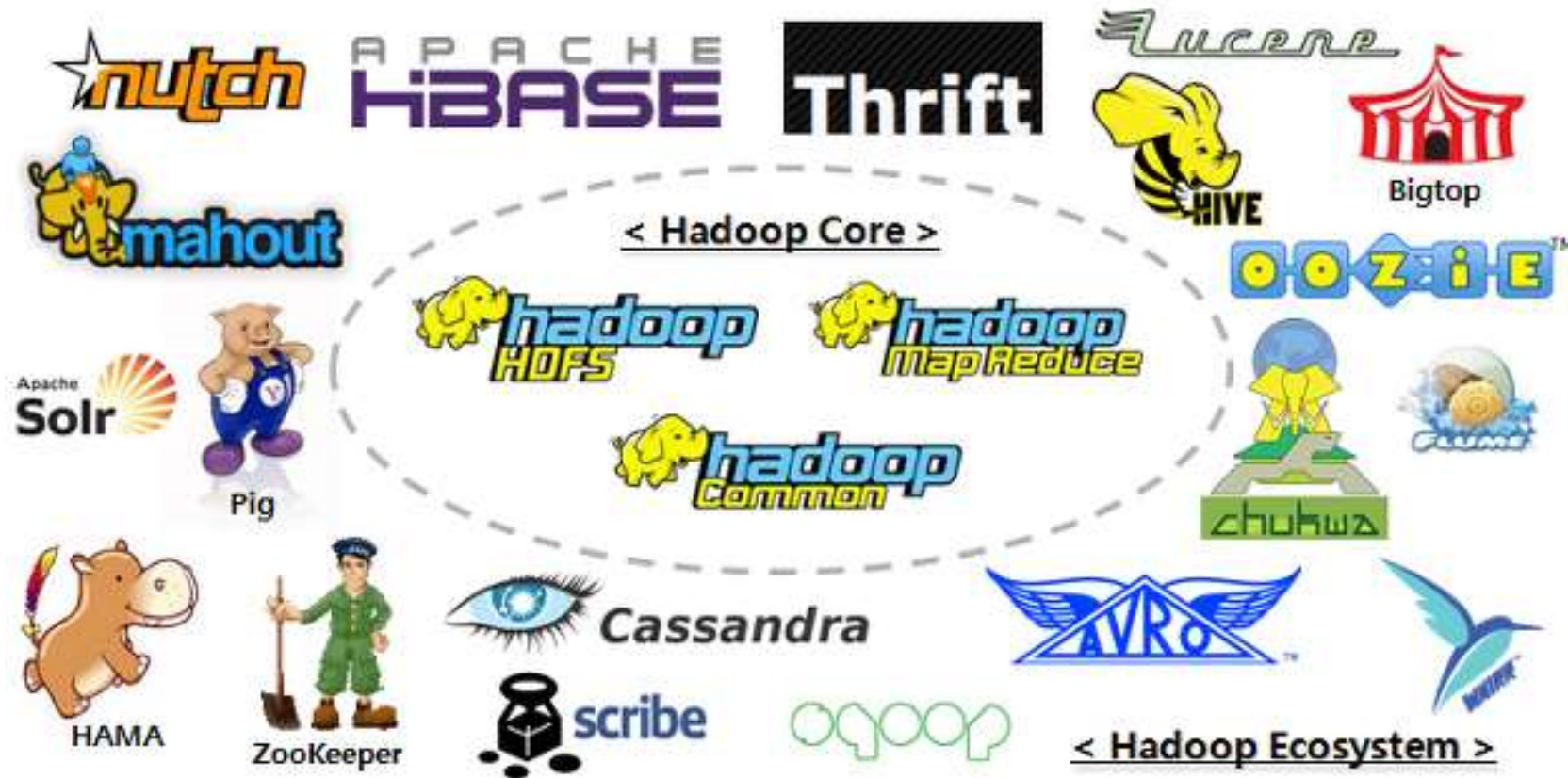
Apache Hadoop Ecosystem



Ecosistema Hadoop

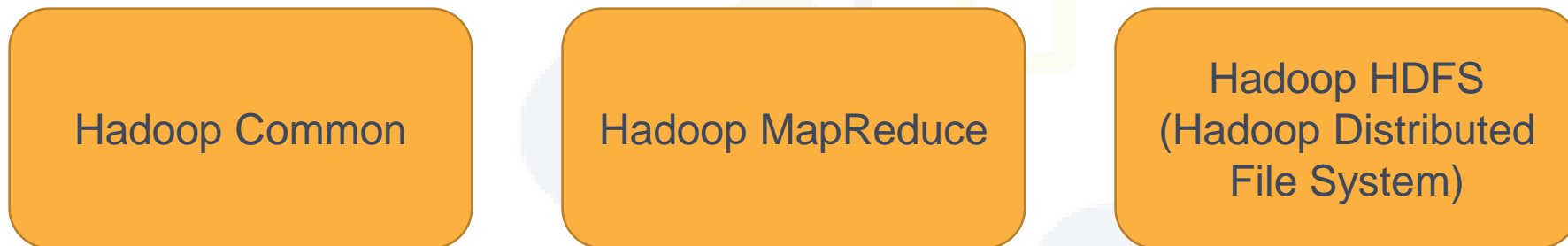


Ecosistema Hadoop



Ecosistema Hadoop

Projetos Principais do Hadoop



Ecosistema Hadoop

Outros Projetos

Zookeeper

Hive

HBase

Pig

Sqoop

Mahout

Flume

Oozie



Data Science Academy

Ecosistema Hadoop

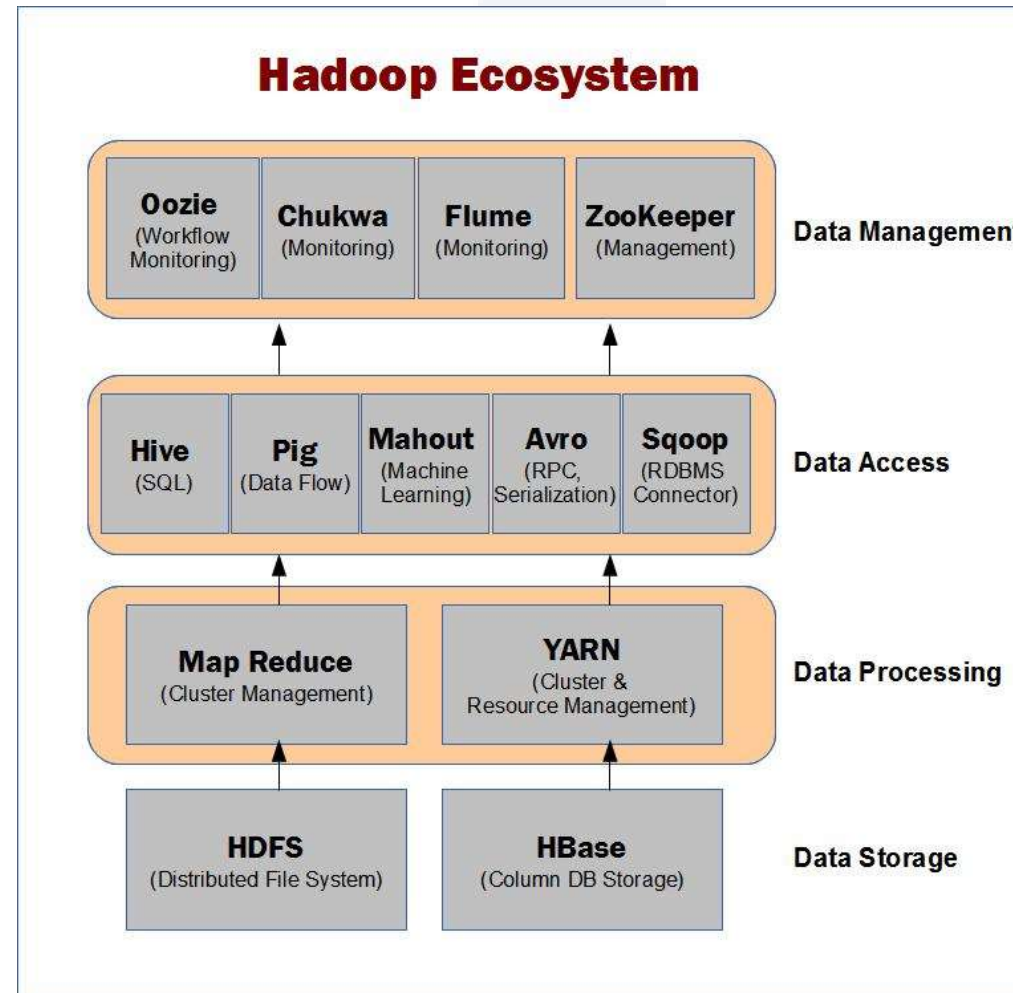
Outros Projetos

YARN



Data Science Academy

Ecosistema Hadoop



Obrigado

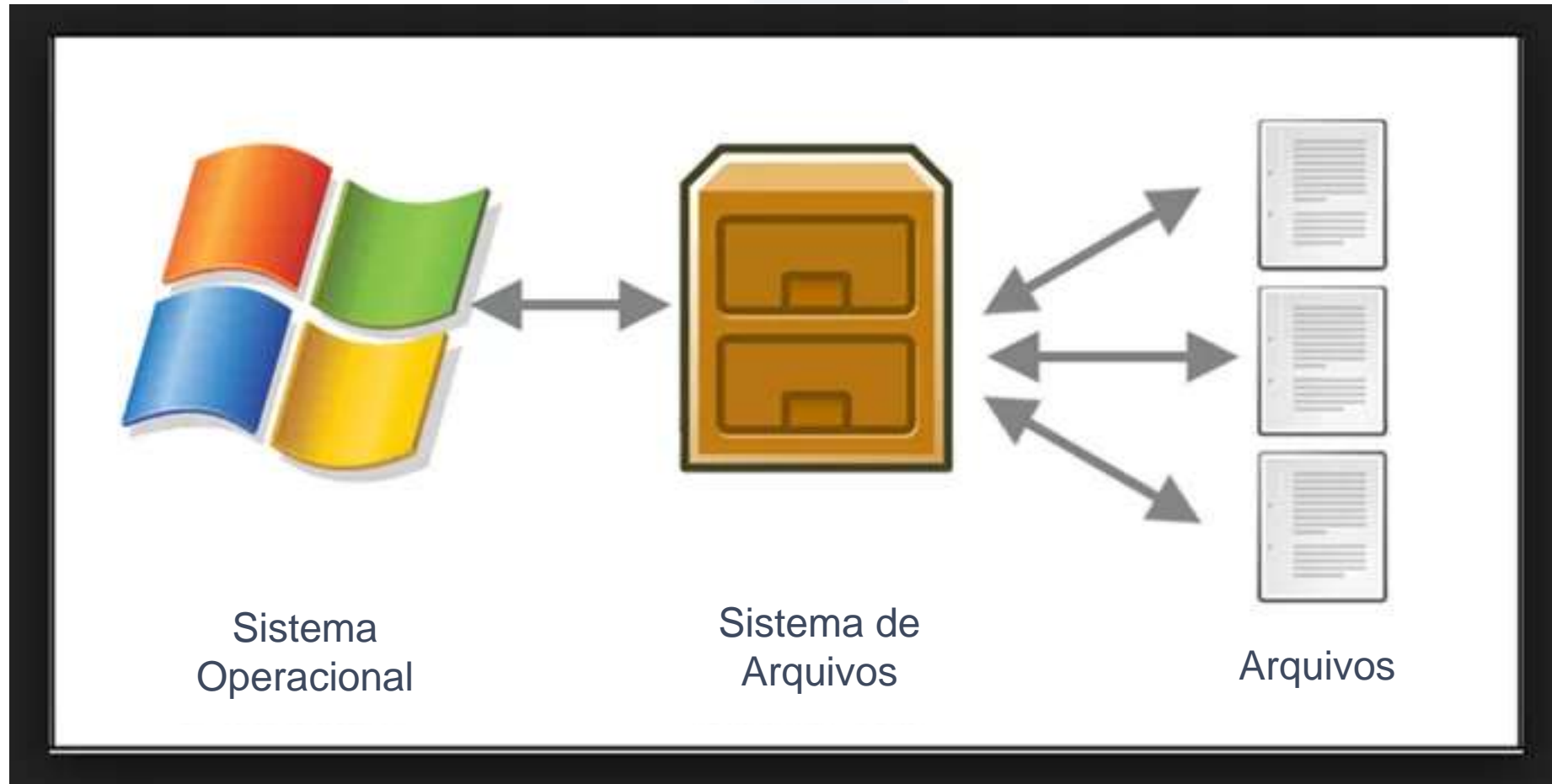


HDFS (Hadoop Distributed File System)

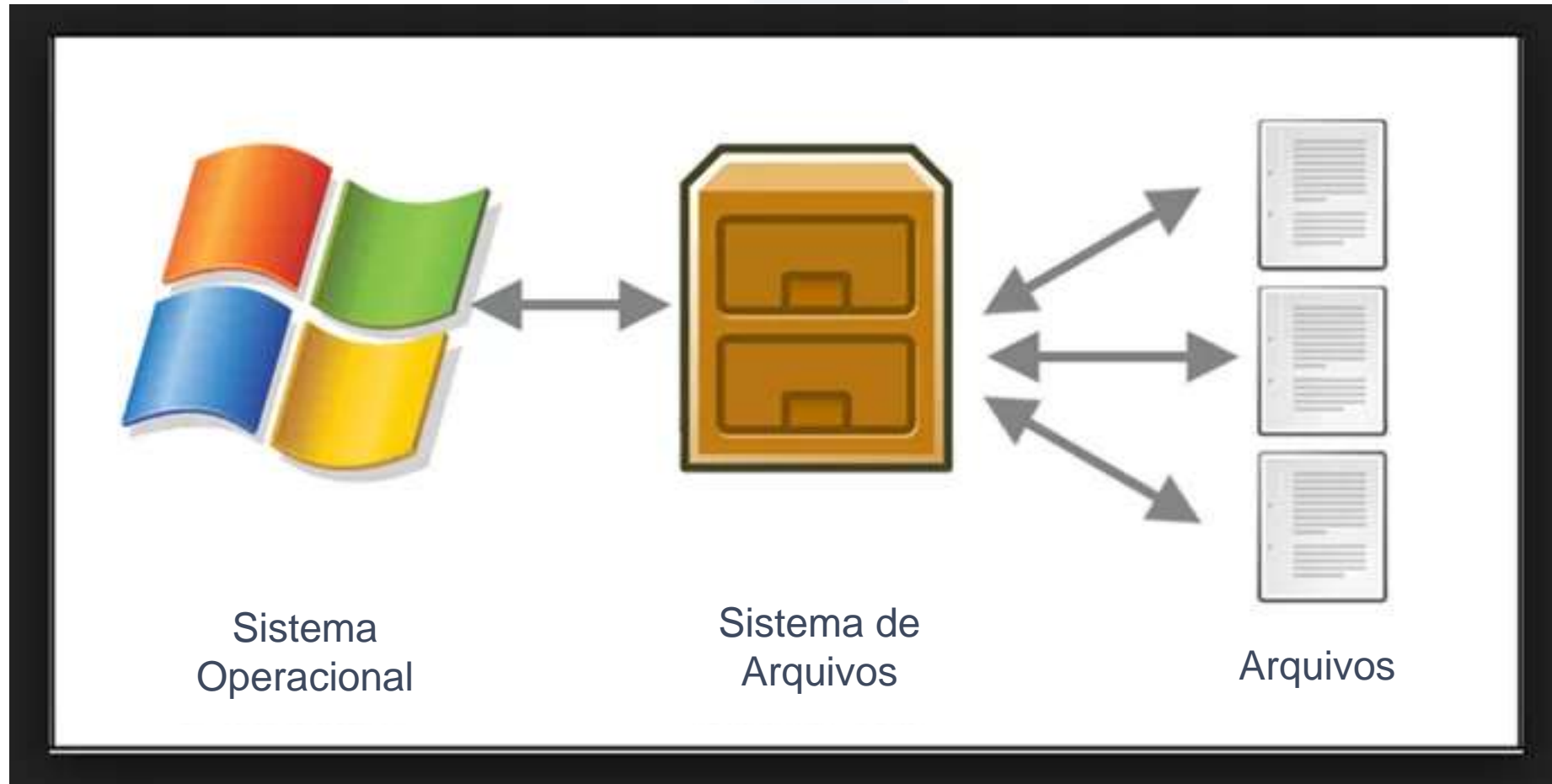
Conceito e Importância



HDFS – Conceito e Importância



HDFS – Conceito e Importância



HDFS – Conceito e Importância

Os tipos de Sistemas de Arquivos são:

Tipo	Descrição
ext2	Sistema de arquivos padrão do Linux
ext3	Sistema de arquivos ext2 melhorado
reiserfs	Sistema de arquivos do tipo Journaling
msdos	Sistema de arquivos FAT da Microsoft DOS
vfat	Sistema de arquivos FAT-32 do Microsoft Windows
iso9660	Sistema de arquivos do CD-ROM
nfs	Network File System. Usado para montar dispositivos em computadores remotos.
swap	Sistema de arquivos de troca utilizando para memória virtual.
proc	Uma janela especial dentro do Kernel do Linux. Utilizada pelos usuários, programas e utilitários para escrever ou ler parâmetros do Kernel. Geralmente montado no diretório <code>/proc</code> .



HDFS – Conceito e Importância



Data Science Academy

HDFS – Conceito e Importância



Data Science Academy

HDFS – Conceito e Importância



Sistema de Arquivos
Distribuído



Data Science Academy

HDFS – Conceito e Importância



- Tolerância a Falhas
- Integridade
- Segurança
- Desempenho
- Consistência



HDFS – Conceito e Importância

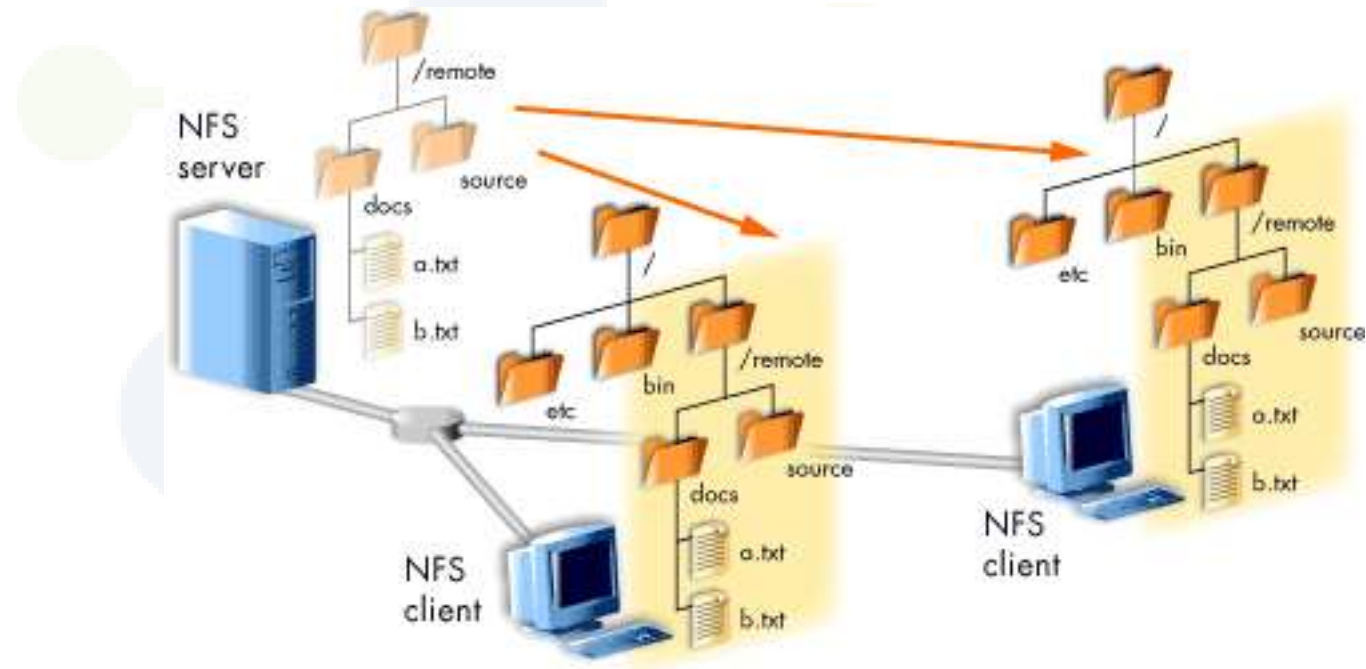
Outros Sistemas
de Arquivos
Distribuídos



Data Science Academy

HDFS – Conceito e Importância

Network File
System

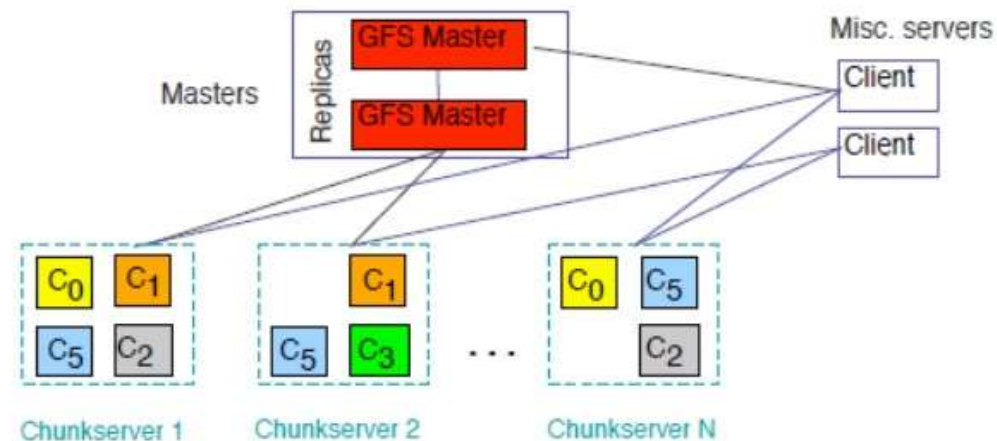


Data Science Academy

HDFS – Conceito e Importância

Google File
System

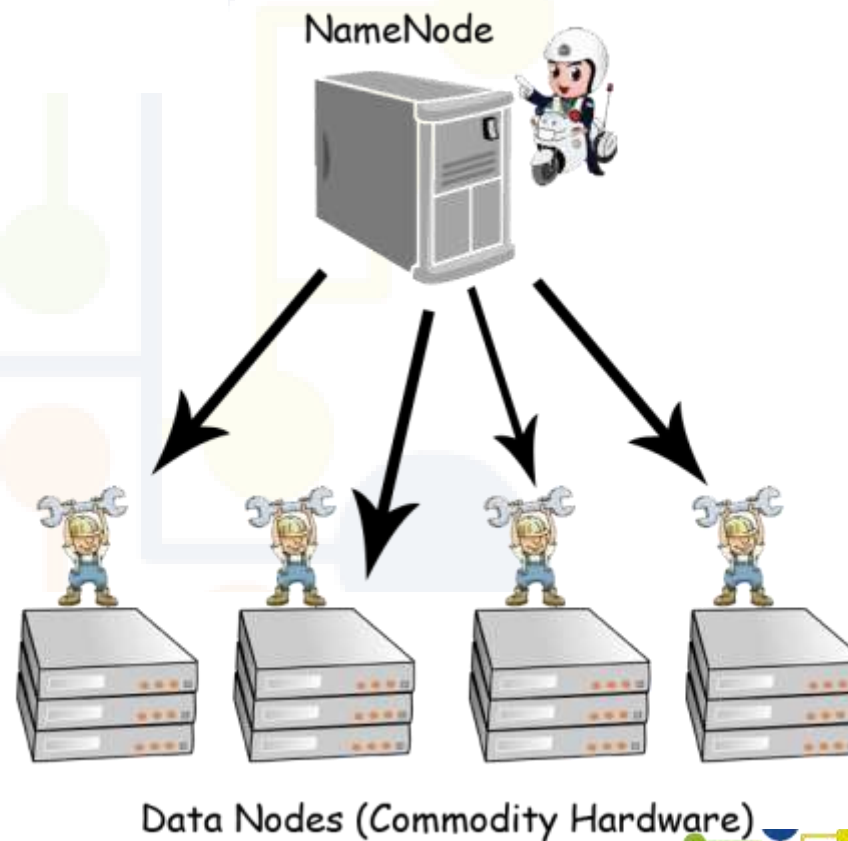
GFS (Google File System) Design



Data Science Academy

HDFS – Conceito e Importância

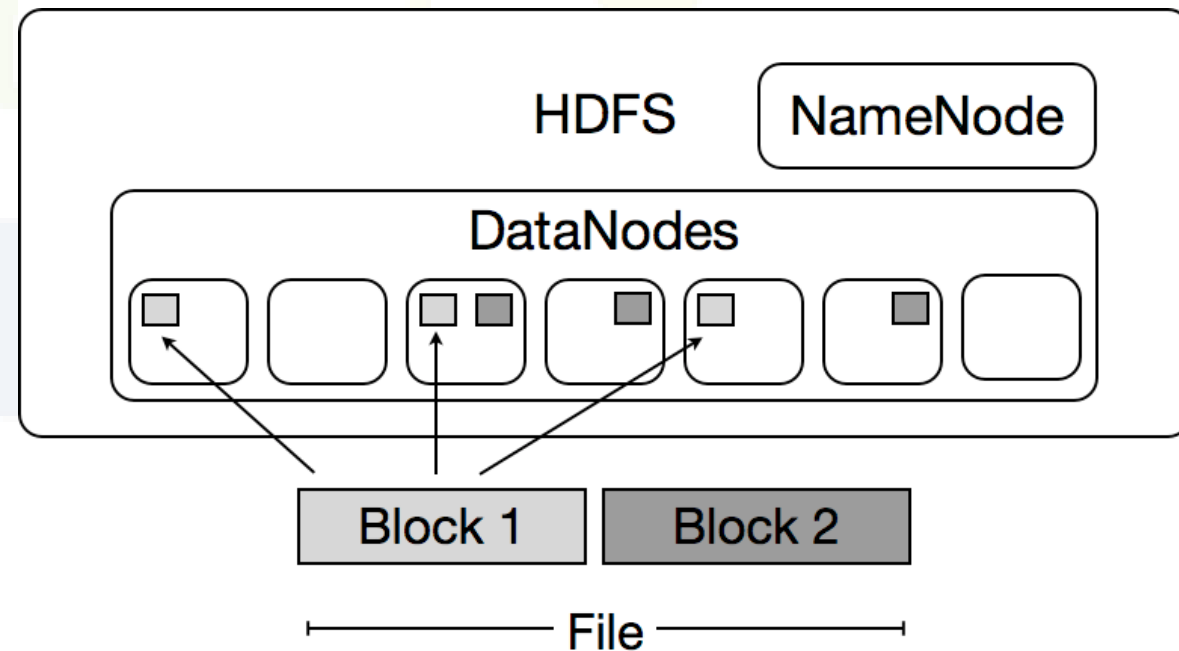
Hadoop
Distributed File
System



Data Science Academy

HDFS – Conceito e Importância

Hadoop
Distributed File
System

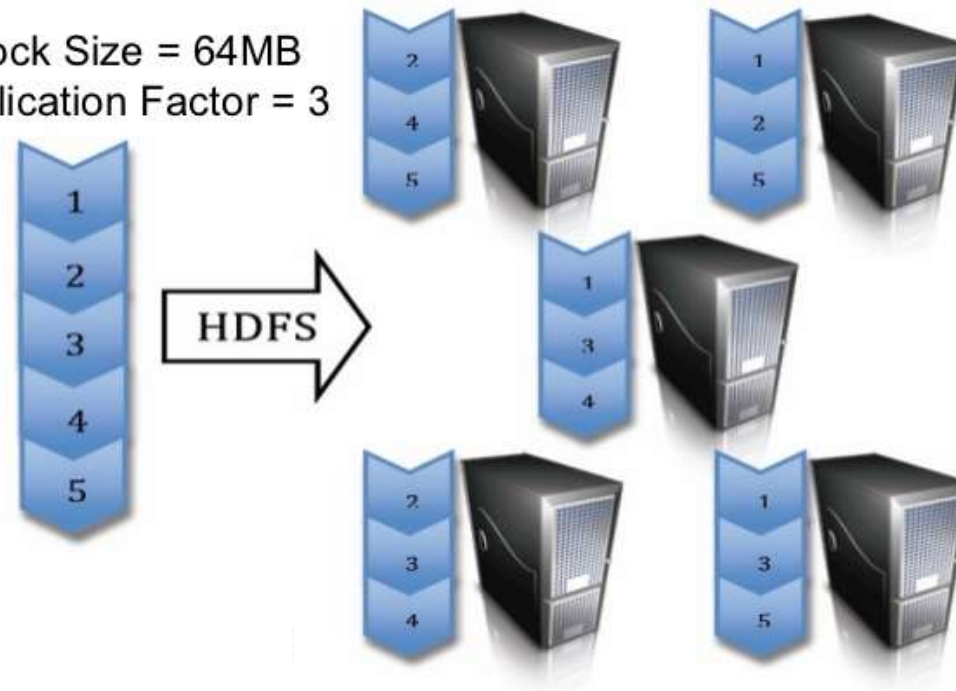


Data Science Academy

HDFS – Conceito e Importância

Hadoop
Distributed File
System

Block Size = 64MB
Replication Factor = 3



HDFS – Conceito e Importância

O HDFS foi criado para resolver "Big Problems" e por isso seu funcionamento e arquitetura são próprios para se trabalhar com grandes arquivos de dados e distribuir esses arquivos em blocos ao longo de um cluster de computadores, para que possam ser processados em paralelo.

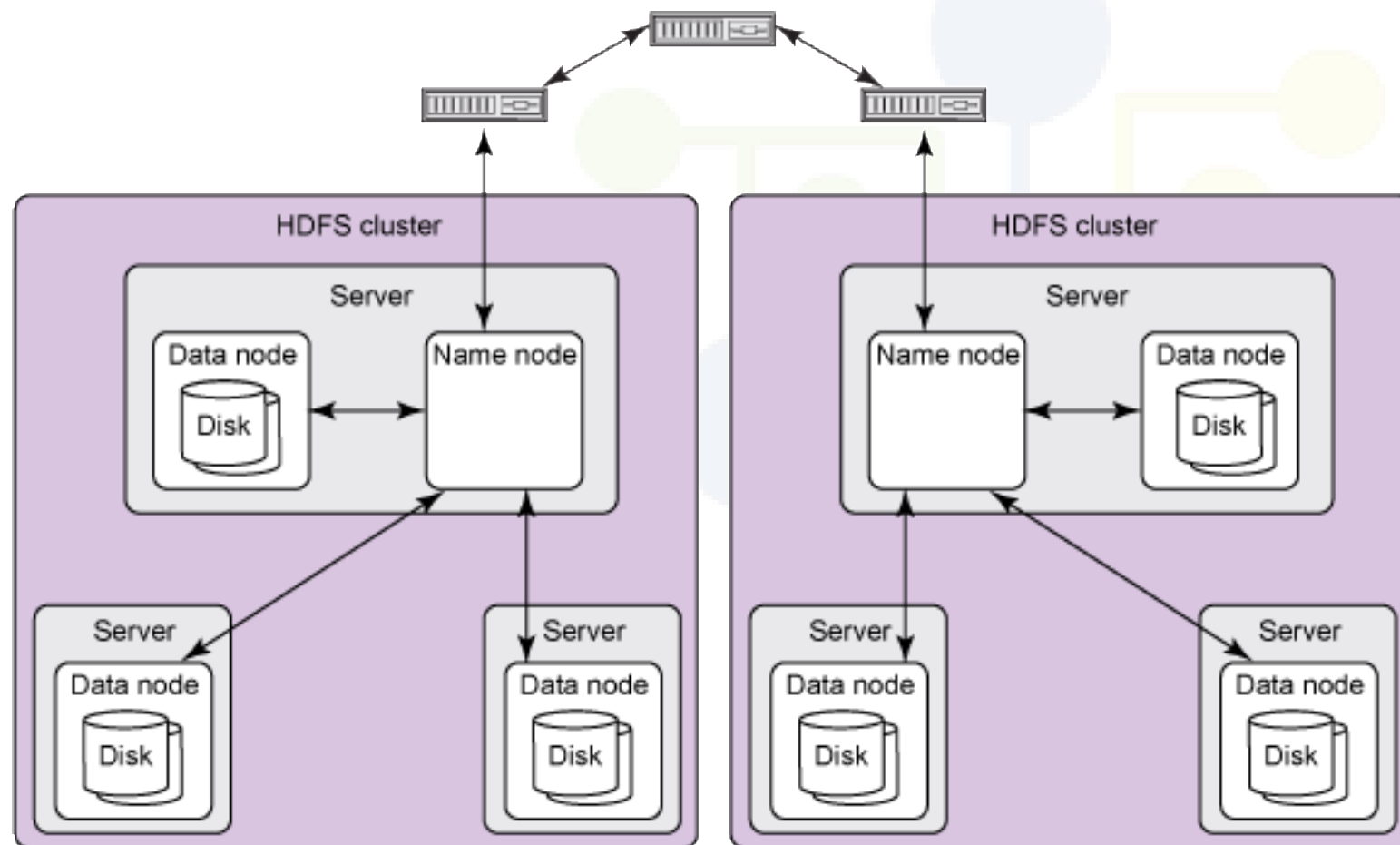
Obrigado



HDFS (Hadoop Distributed File System) Arquitetura



HDFS – Arquitetura



Arquitetura
Master/Slave



Data Science Academy

HDFS – Arquitetura

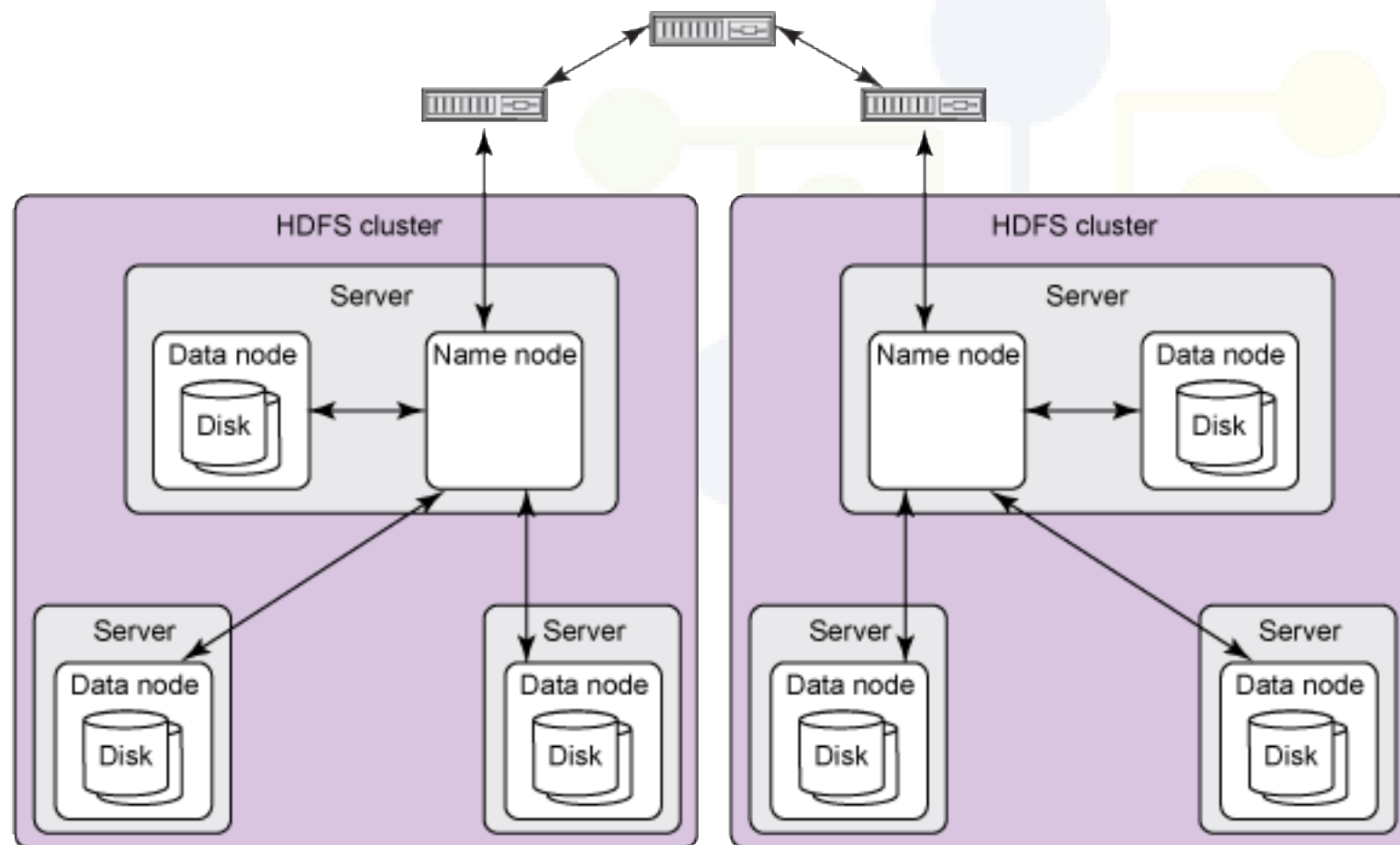


Arquitetura
Master/Slave



Data Science Academy

HDFS – Arquitetura

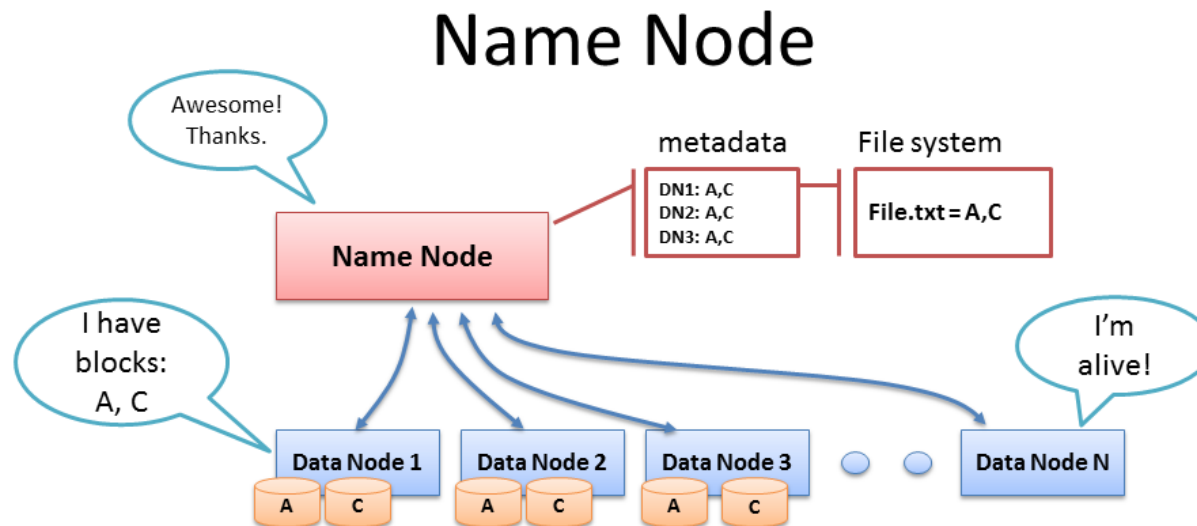


Arquitetura
Master/Slave



Data Science Academy

HDFS – Arquitetura

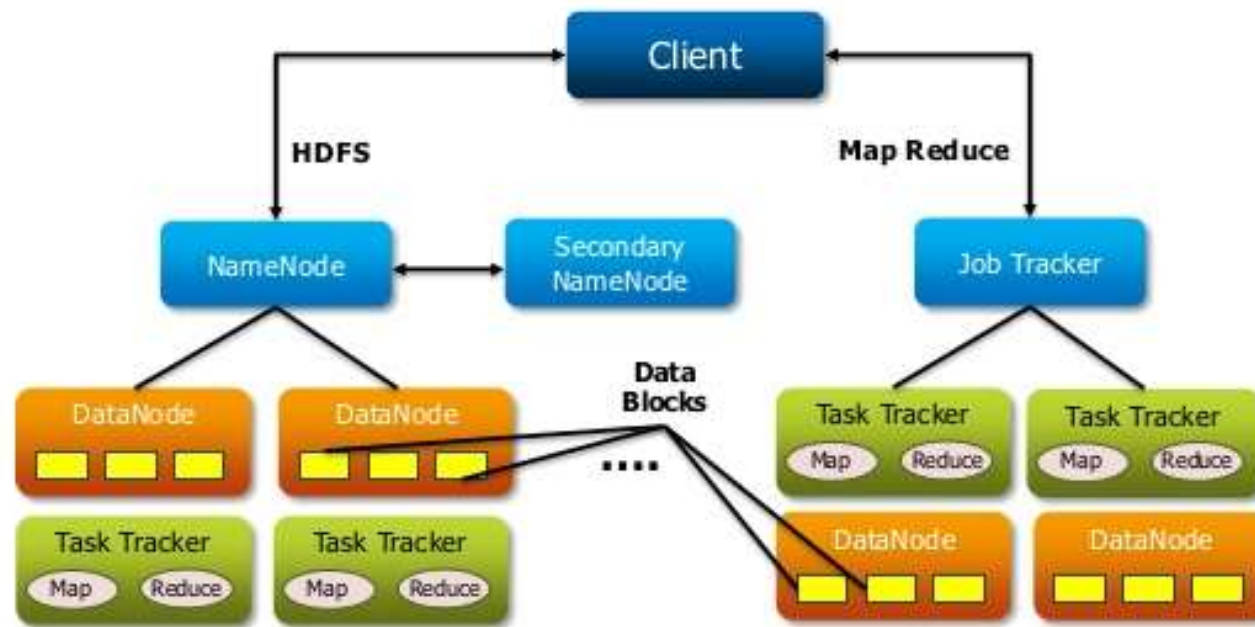


NameNode

- FsImage
- EditLog



HDFS – Arquitetura



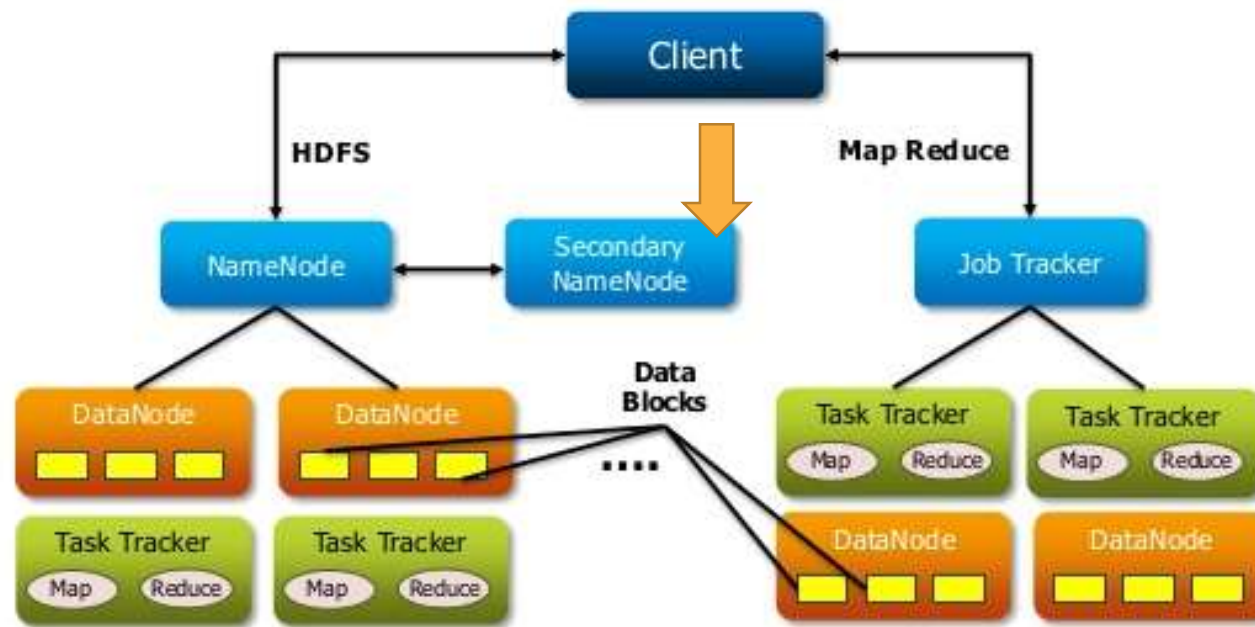
NameNode

- FsImage
- EditLog



Data Science Academy

HDFS – Arquitetura



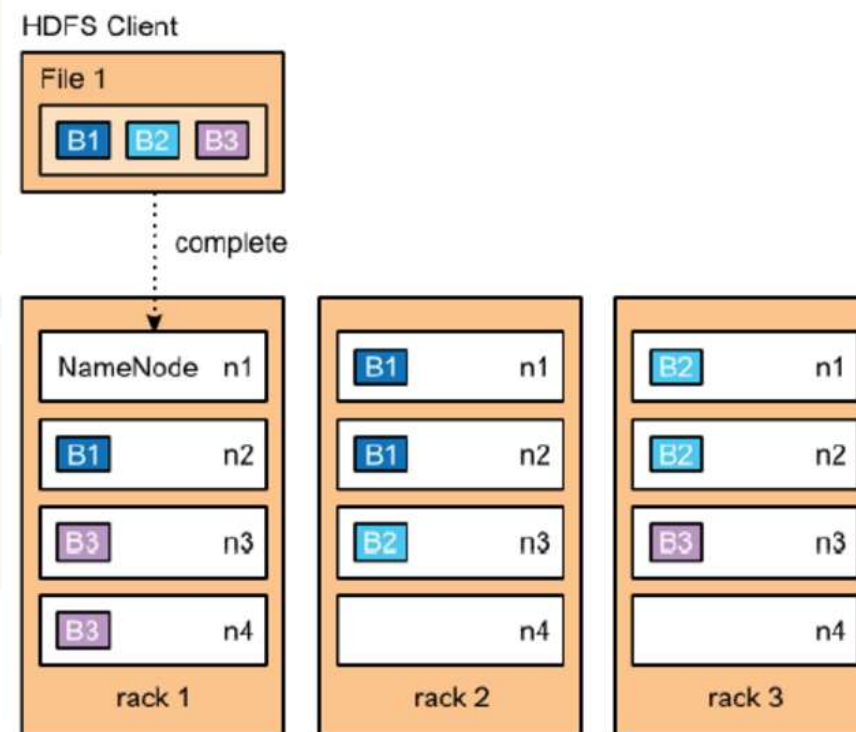
NameNode

- FsImage
- EditLog



HDFS – Arquitetura

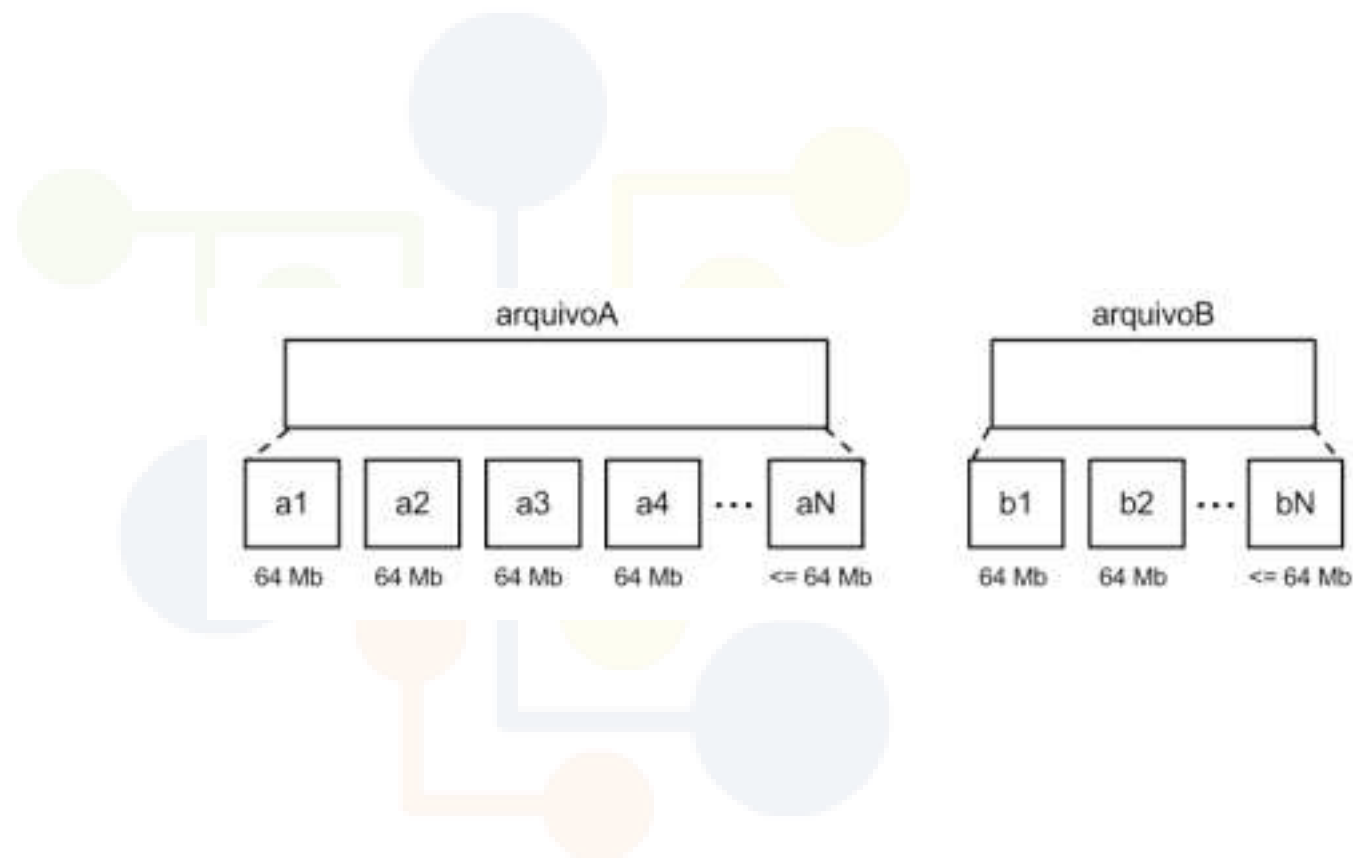
Replicação



Data Science Academy

HDFS – Arquitetura

Replicação

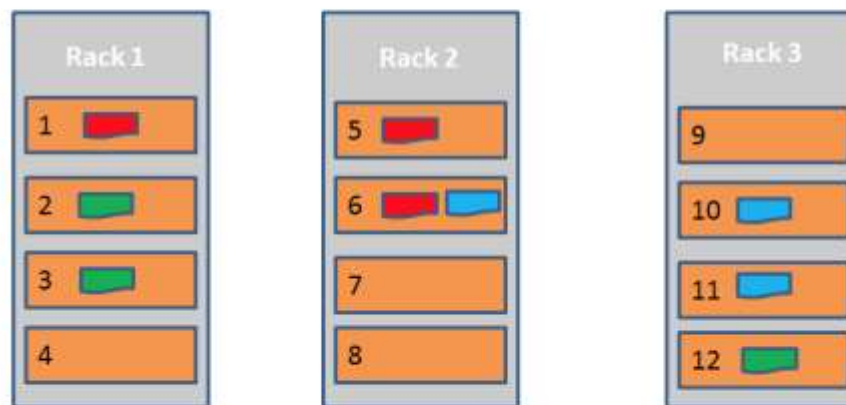


Data Science Academy

HDFS – Arquitetura

Replicação

Block A: 
Block B: 
Block C: 



Data Science Academy

Obrigado

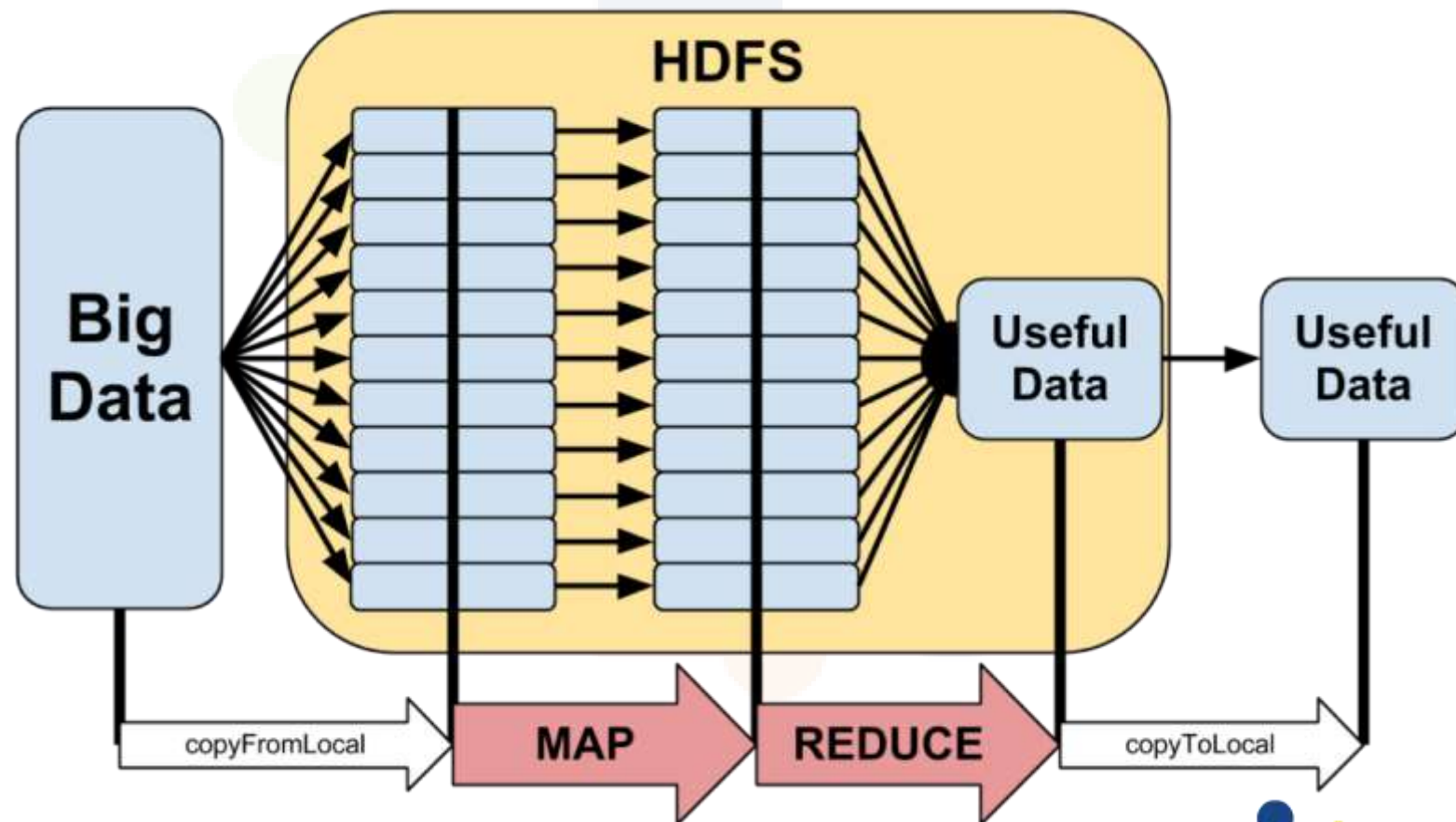




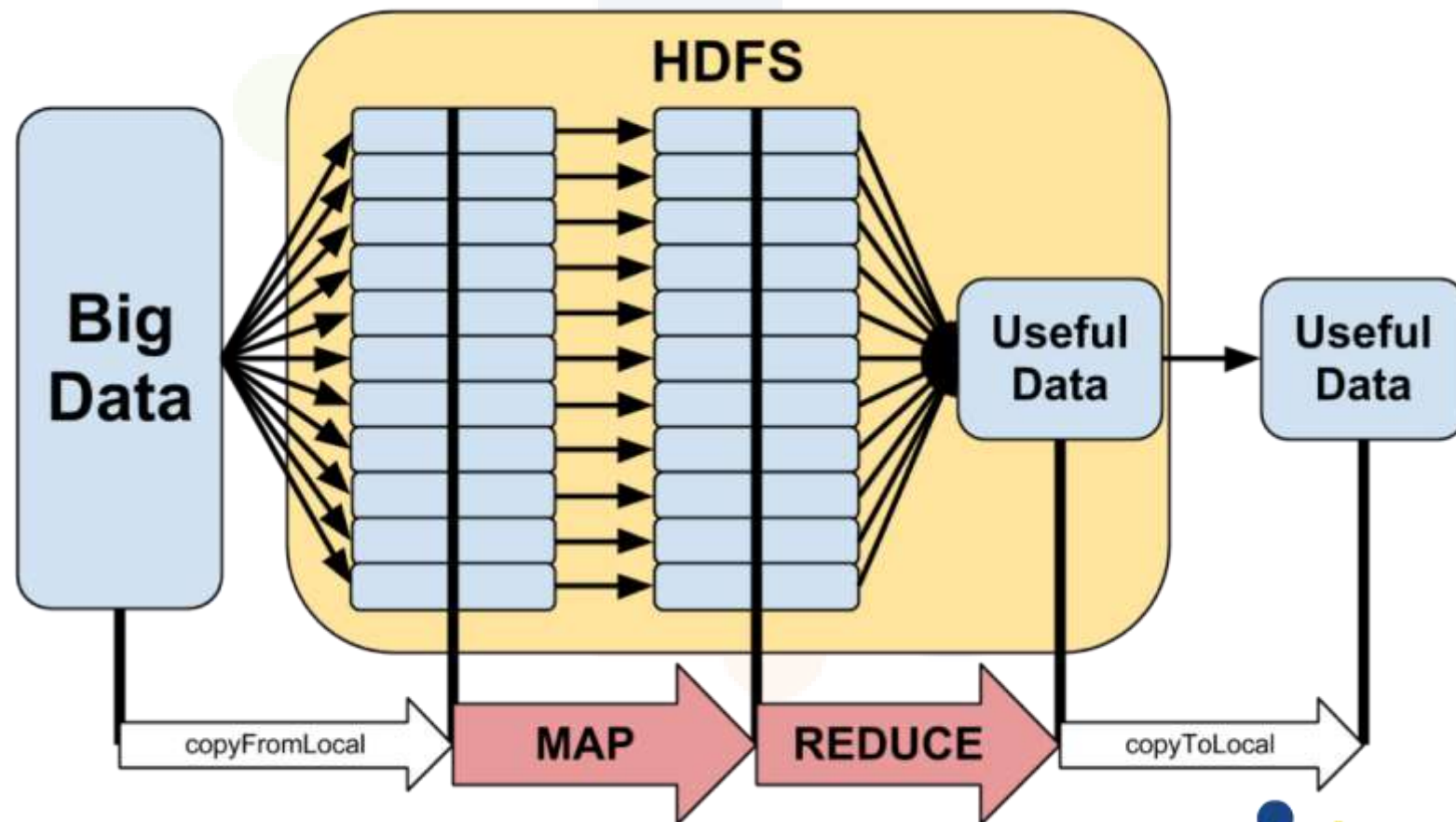
Definindo MapReduce



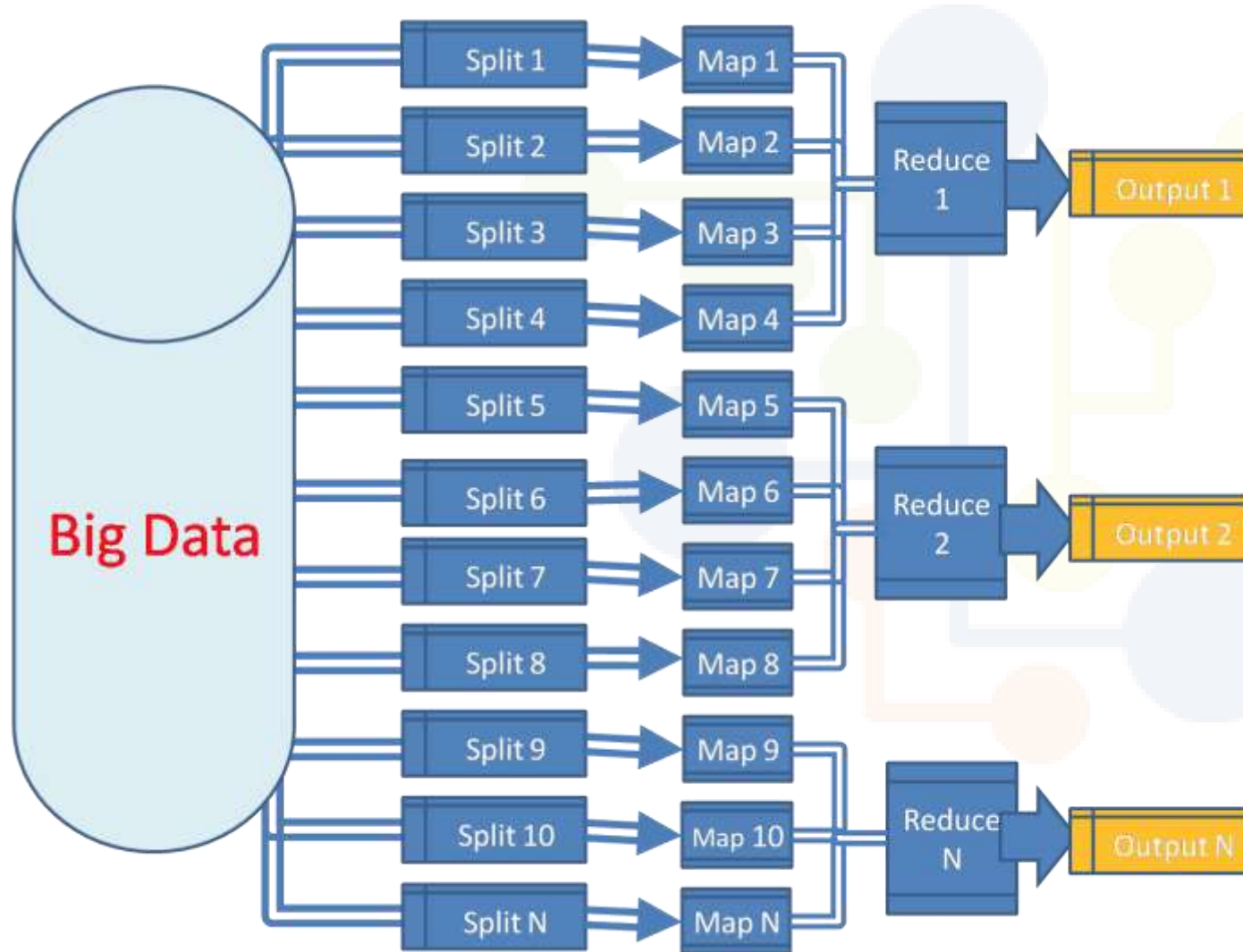
Definindo MapReduce



Definindo MapReduce



Definindo MapReduce

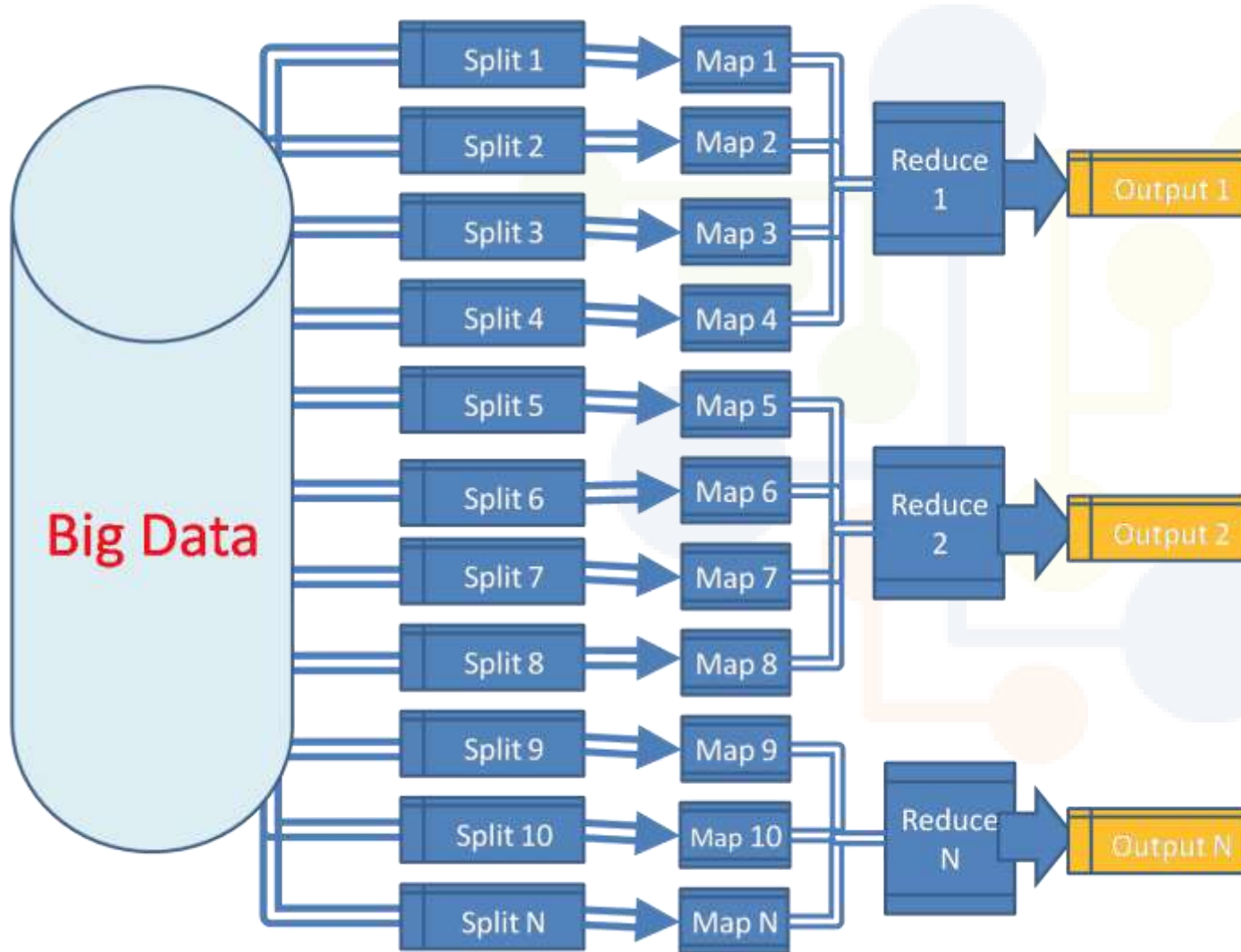


Processamento
Paralelo e Distribuído



Data Science Academy

Definindo MapReduce

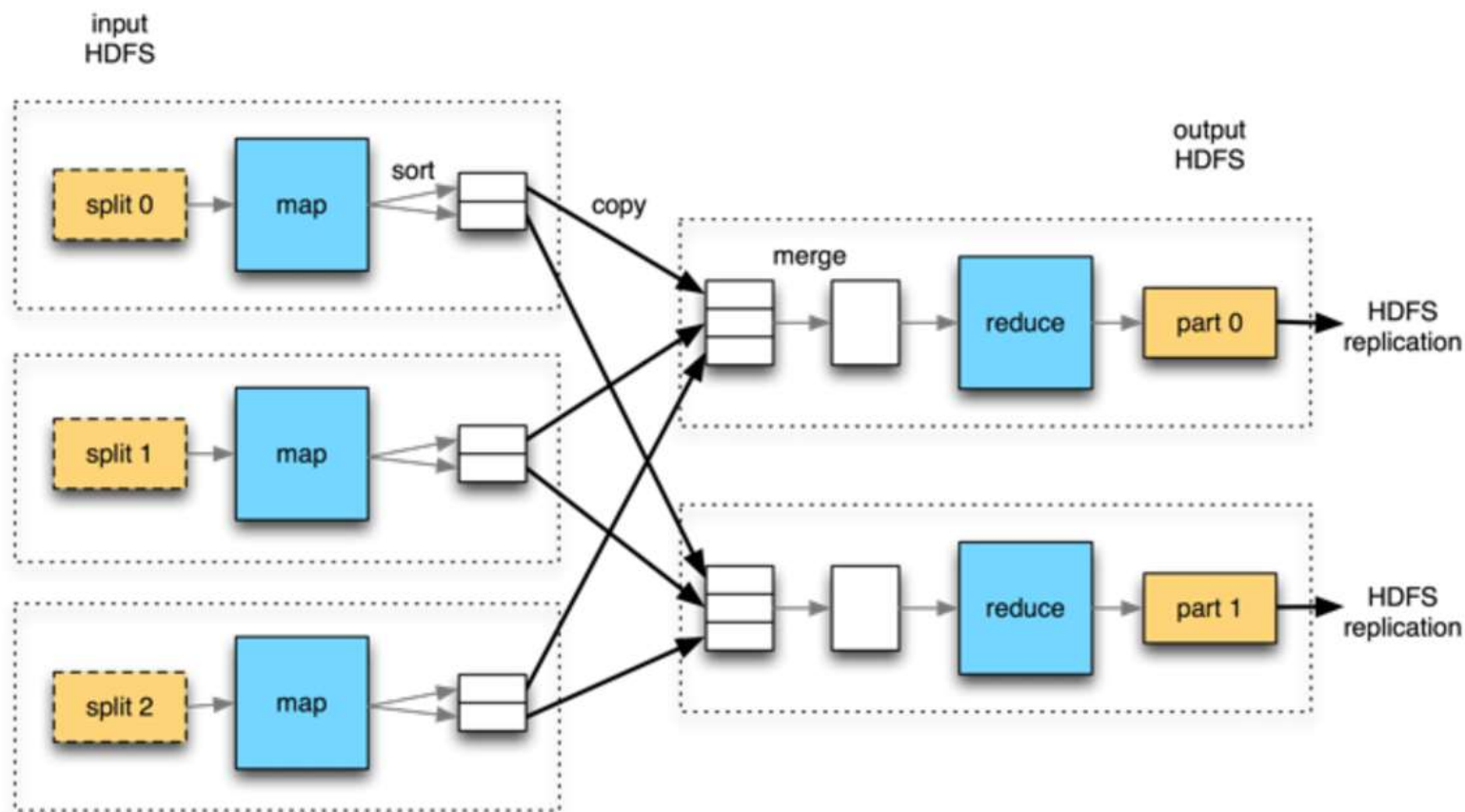


Processamento
Paralelo e Distribuído



Data Science Academy

Definindo MapReduce

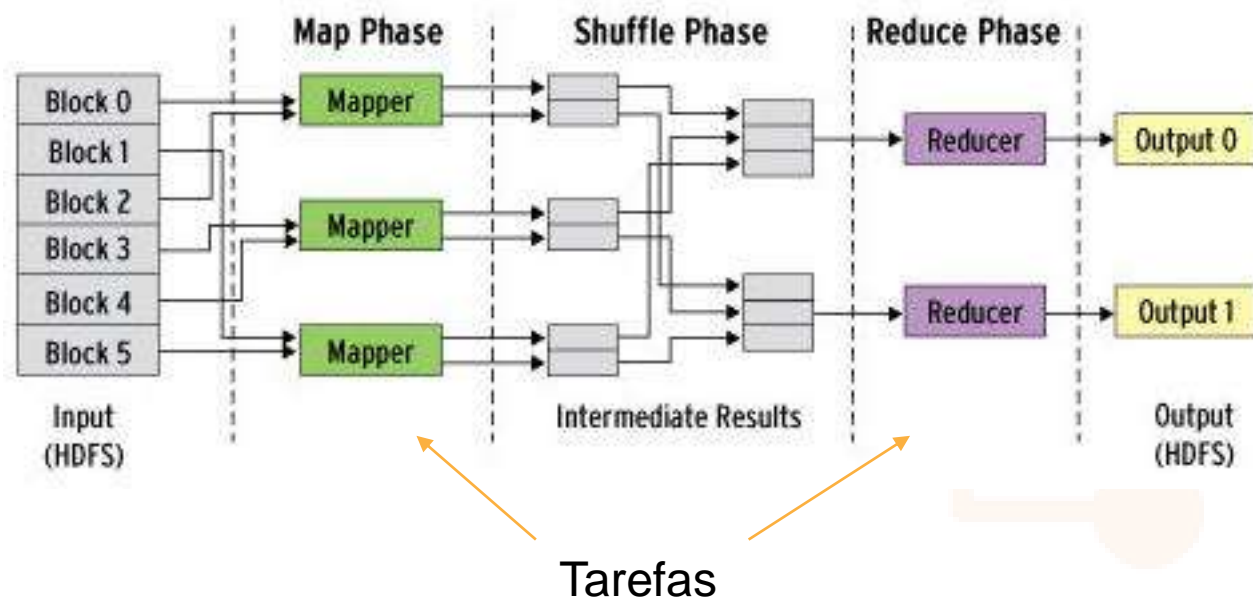


Processamento
Paralelo e Distribuído



Data Science Academy

Definindo MapReduce



Processamento
Paralelo e Distribuído



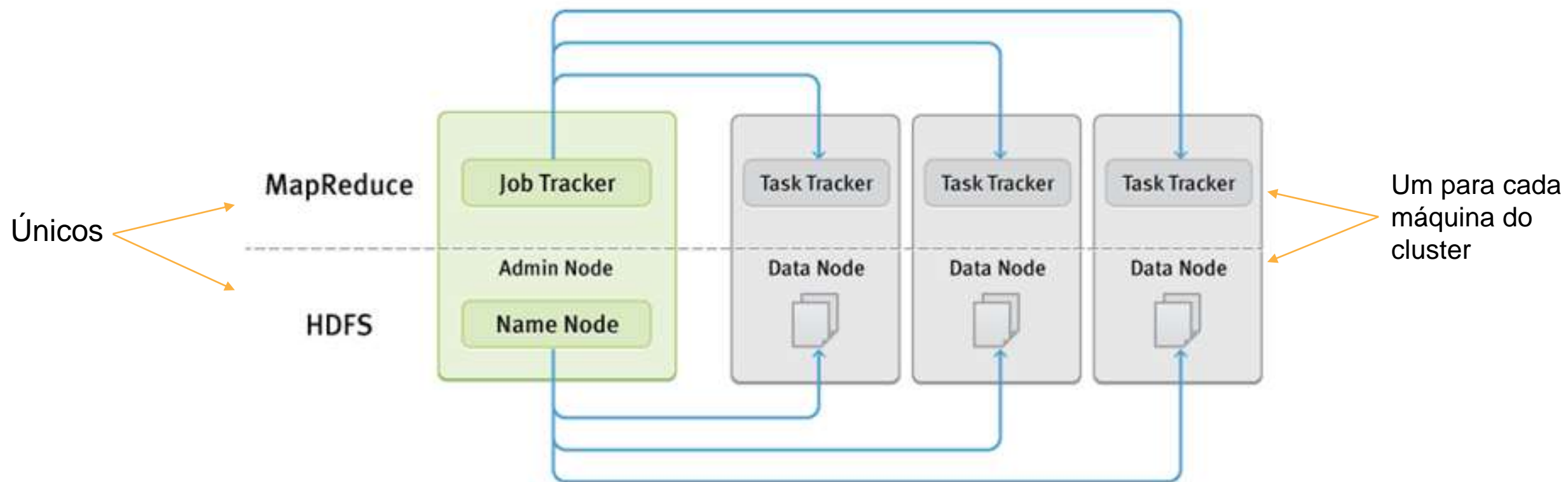
Data Science Academy

Definindo MapReduce

Arquitetura MapReduce

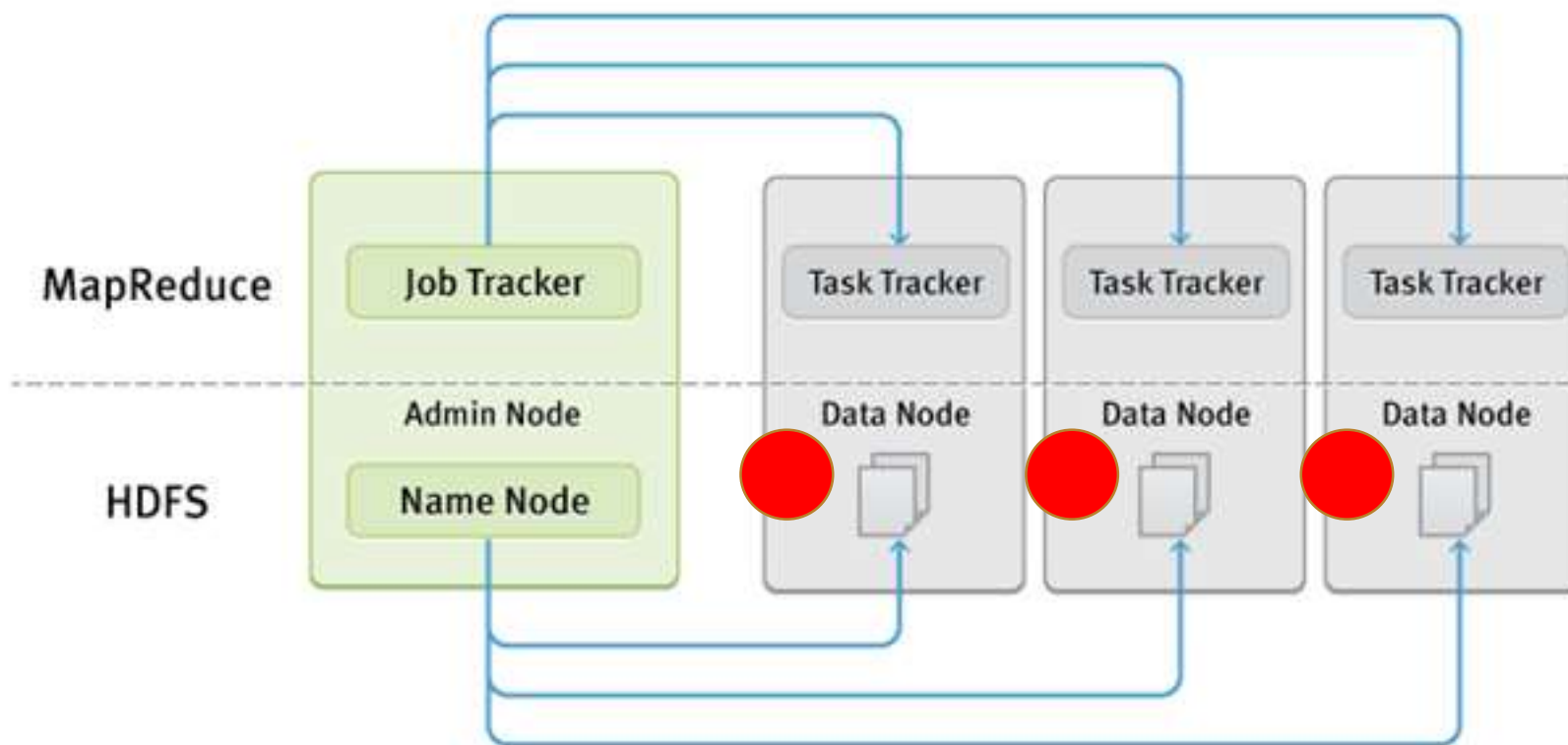


Definindo MapReduce

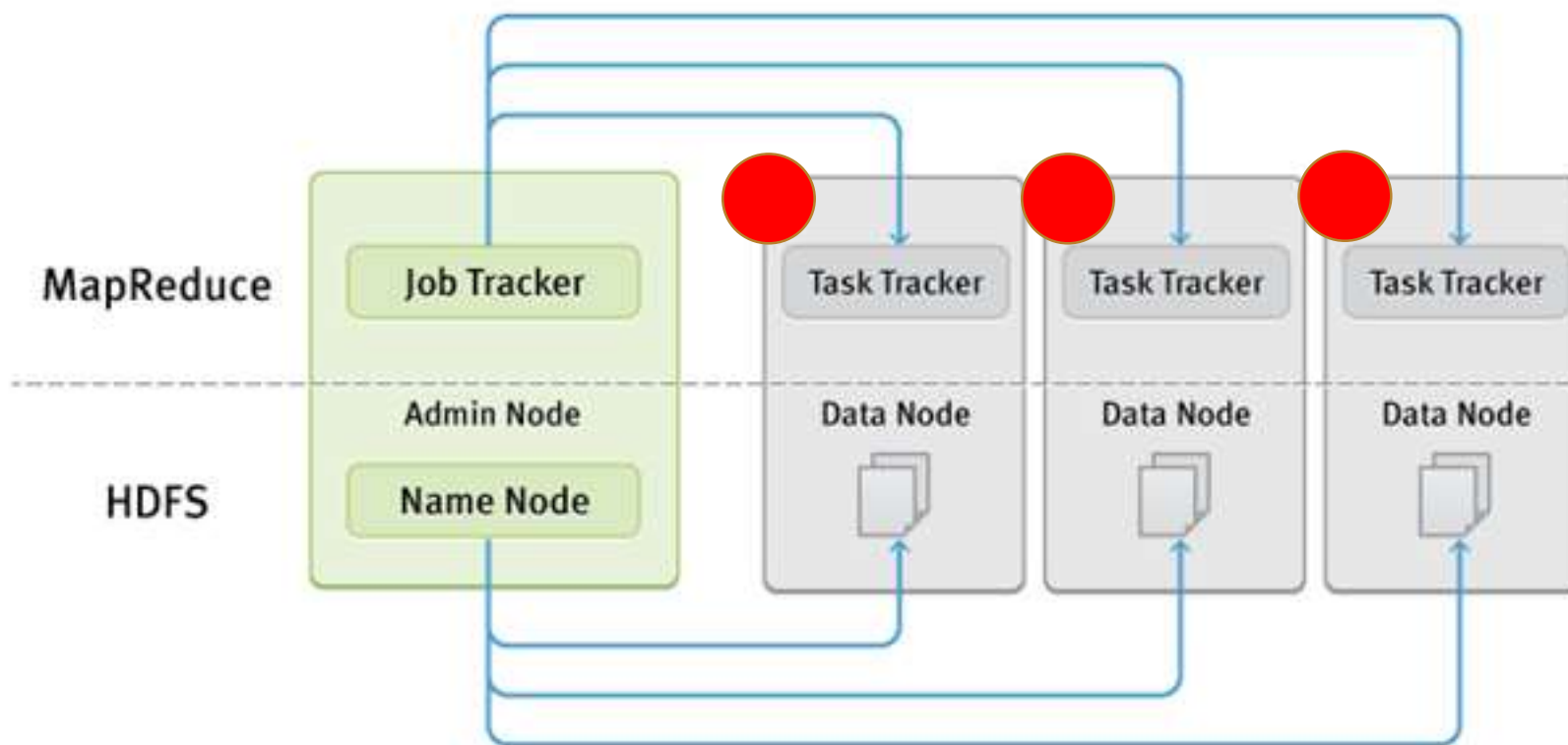


Data Science Academy

Definindo MapReduce



Definindo MapReduce



Obrigado





Hadoop x Bancos de Dados Relacionais



Hadoop x Bancos de Dados Relacionais

RDBMS
Relational Database
Management Systems

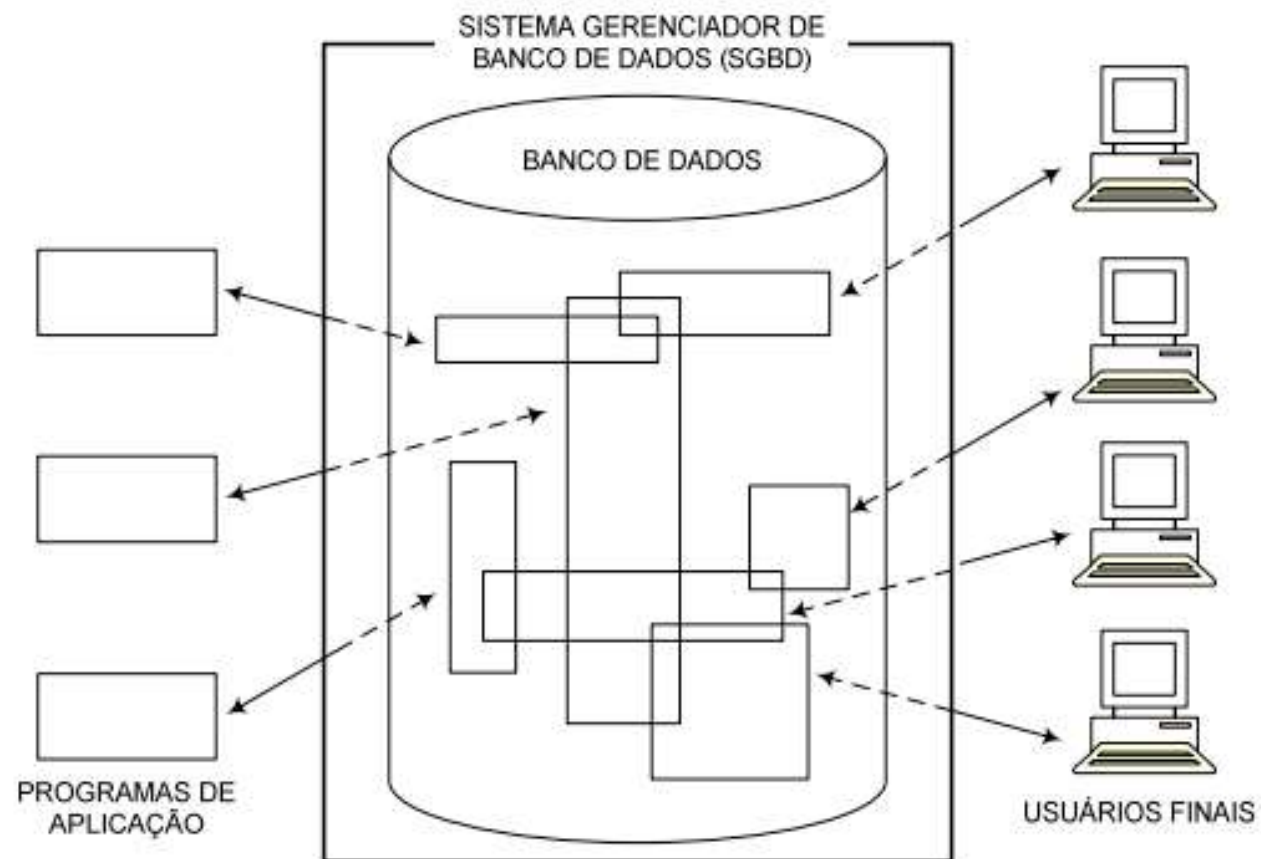


Data Science Academy

Hadoop x Bancos de Dados Relacionais

SGBD's

Gerenciam um ou mais bancos de dados



Data Science Academy

Hadoop x Bancos de Dados Relacionais

RDBMS

NoSQL

Hadoop

Durante várias décadas, os RDBMS atenderam o seu propósito e ainda o fazem muito bem, mas com o surgimento do Big Data e esta imensa quantidade de dados de diferentes categorias, gerados em diferentes velocidades, volumes e formatos, novos modelos de gestão de dados começaram a surgir. Isto levou ao crescimento por exemplo de soluções NoSQL, bancos de dados não relacionais. Com NoSQL, dados não estruturados podem ser armazenados em vários nós de processamento e não requerem schemas fixos, geralmente evitam operações de join e, normalmente, funcionam bem com escalonamento horizontal. Estima-se que existam hoje 60 bancos de dados não-relacionais e muita dessa evolução se deve ao crescimento do Big Data. O Big Data que também motivou o surgimento do Hadoop. Veja que atualmente temos diferentes opções de armazenamento para atender diferentes propósitos. Não se discute qual é melhor ou pior, mas sim qual solução deve ser usada para resolver um problema específico.



Data Science Academy

Hadoop x Bancos de Dados Relacionais

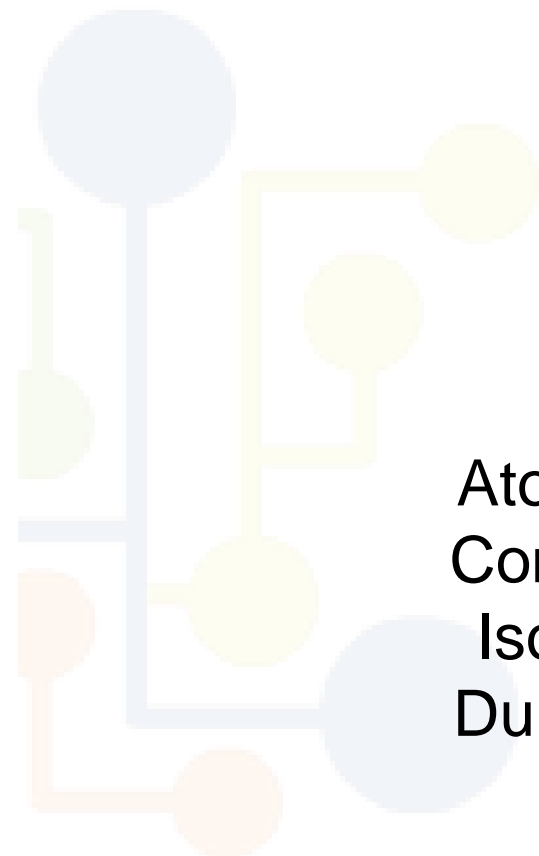


Bancos de dados relacionais usam linguagem SQL, tornando-os uma boa escolha para aplicações que envolvem a gestão de várias operações.



Data Science Academy

Hadoop x Bancos de Dados Relacionais



ACID

Atomicidade
Consistência
Isolamento
Durabilidade



Data Science Academy

Hadoop x Bancos de Dados Relacionais

ORACLE®



Data Science Academy

Hadoop x Bancos de Dados Relacionais



Hadoop

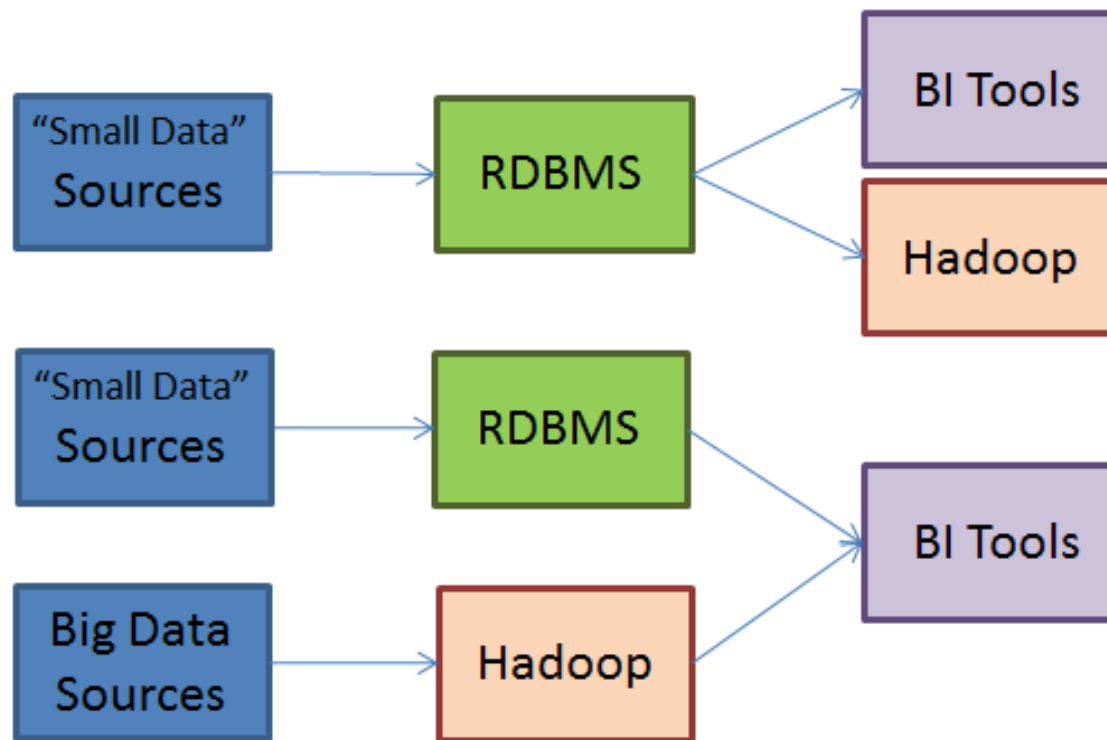


Data Science Academy

Hadoop x Bancos de Dados Relacionais

Hadoop → Grandes volumes de dados

RDBMS → Dados transacionais

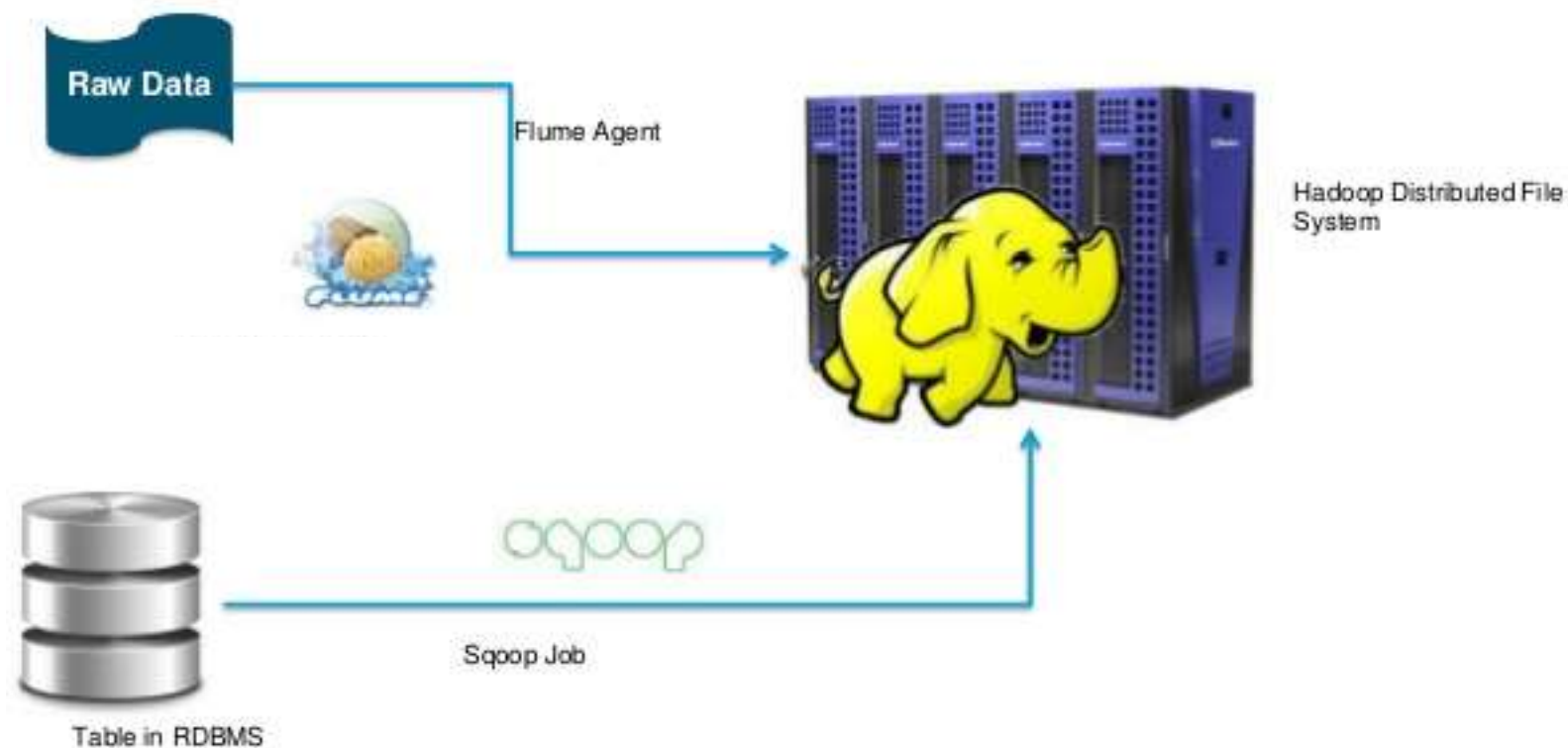


Hadoop x Bancos de Dados Relacionais



Data Science Academy

Hadoop x Bancos de Dados Relacionais



Hadoop x Bancos de Dados Relacionais

Hadoop processa dados em batch. Consequentemente, ele não deve ser usado para processar dados transacionais. Mas o Hadoop pode resolver muitos outros tipos de problemas relacionados ao Big Data.



Obrigado



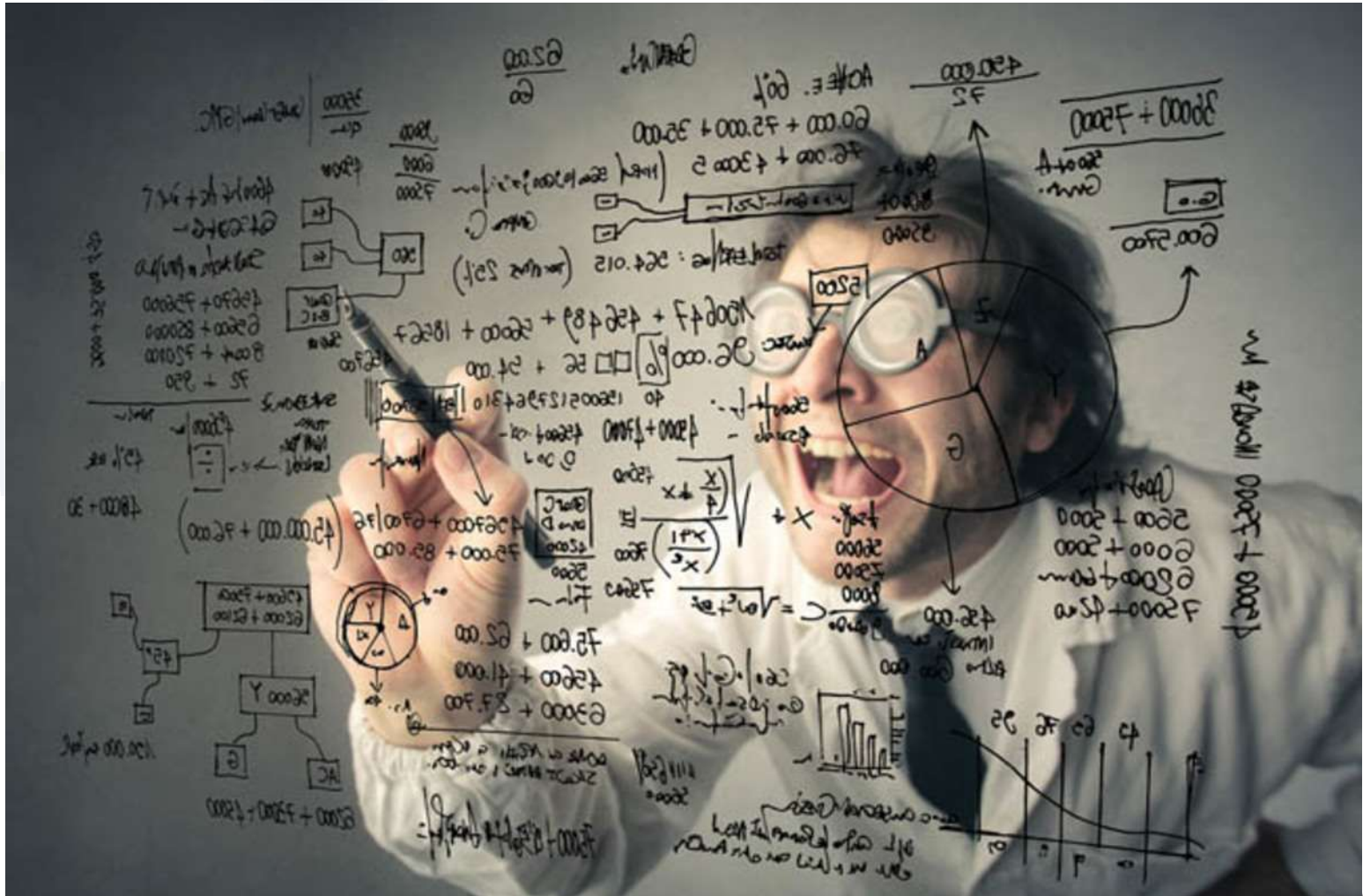


Por que Cientistas de Dados Precisam Conhecer o Hadoop?



Cientistas de Dados e Hadoop

Diferentes pessoas usam diferentes ferramentas para diferentes propósitos.



Data Science Academy

1

Hadoop é open source



2

Hadoop oferece o framework mais completo para armazenamento e processamento de Big Data



3

A líder mundial em bancos de dados relacionais, a Oracle, oferece soluções de Big Data Analytics com Hadoop



4

A líder mundial em sistemas operacionais, a Microsoft, oferece soluções corporativas em nuvem, com Hadoop



5

O Hadoop é mantido pela Apache Foundation, mas recebe contribuição de empresas como Google, Yahoo e Facebook



6

Um Cientista de Dados deve conhecer bem o paradigma de processamento MapReduce



7

Hadoop normalmente aparece como um dos skills mais procurados em um Cientista de Dados



8

Por se tratar de uma tecnologia avançada, faltam profissionais de Hadoop no mercado



9

Hadoop é usado por algumas das maiores empresas do mundo



10

O Big Data ainda está na sua infância.
Onde vamos armazenar todos esses dados?



Convencido sobre a importância
de aprender Hadoop?



Cientistas de Dados e Hadoop

Não?



Data Science Academy

Cientistas de Dados e Hadoop

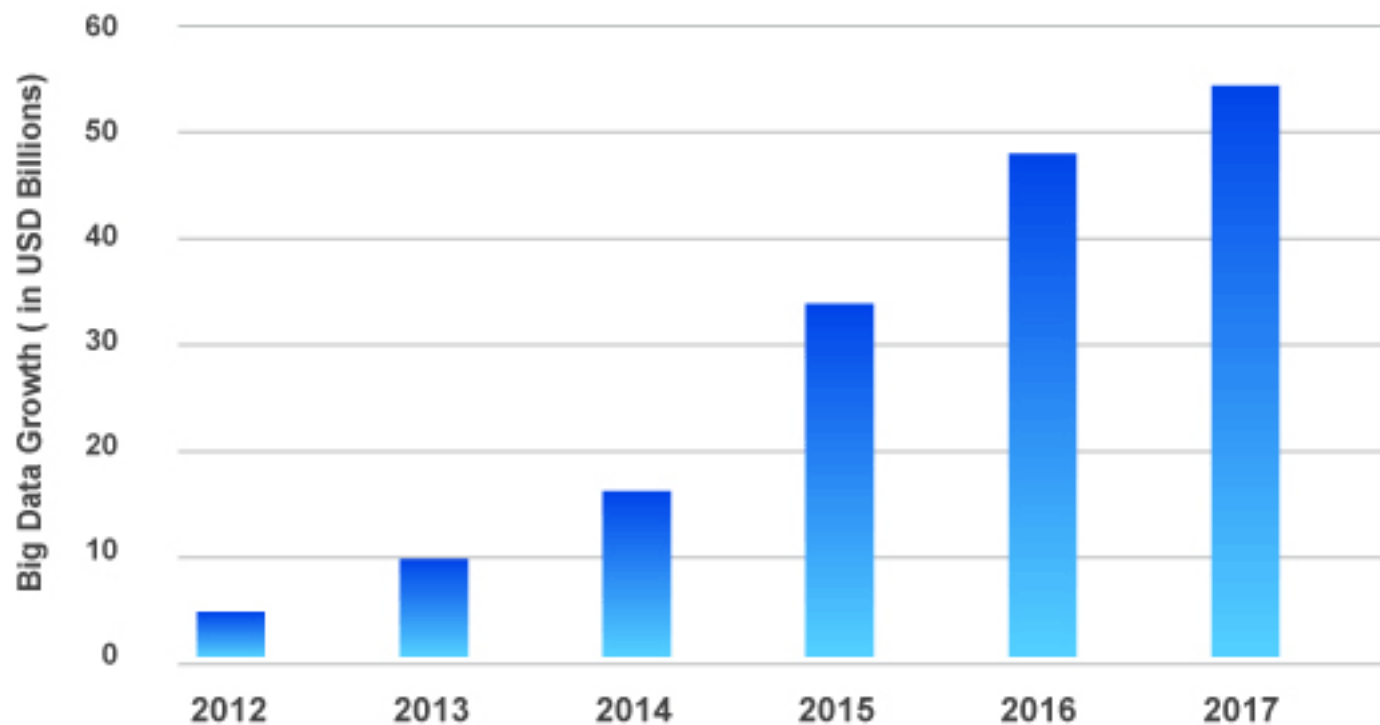
Ainda tenho mais argumentos!



Data Science Academy

Cientistas de Dados e Hadoop

Big Data Market Forecast



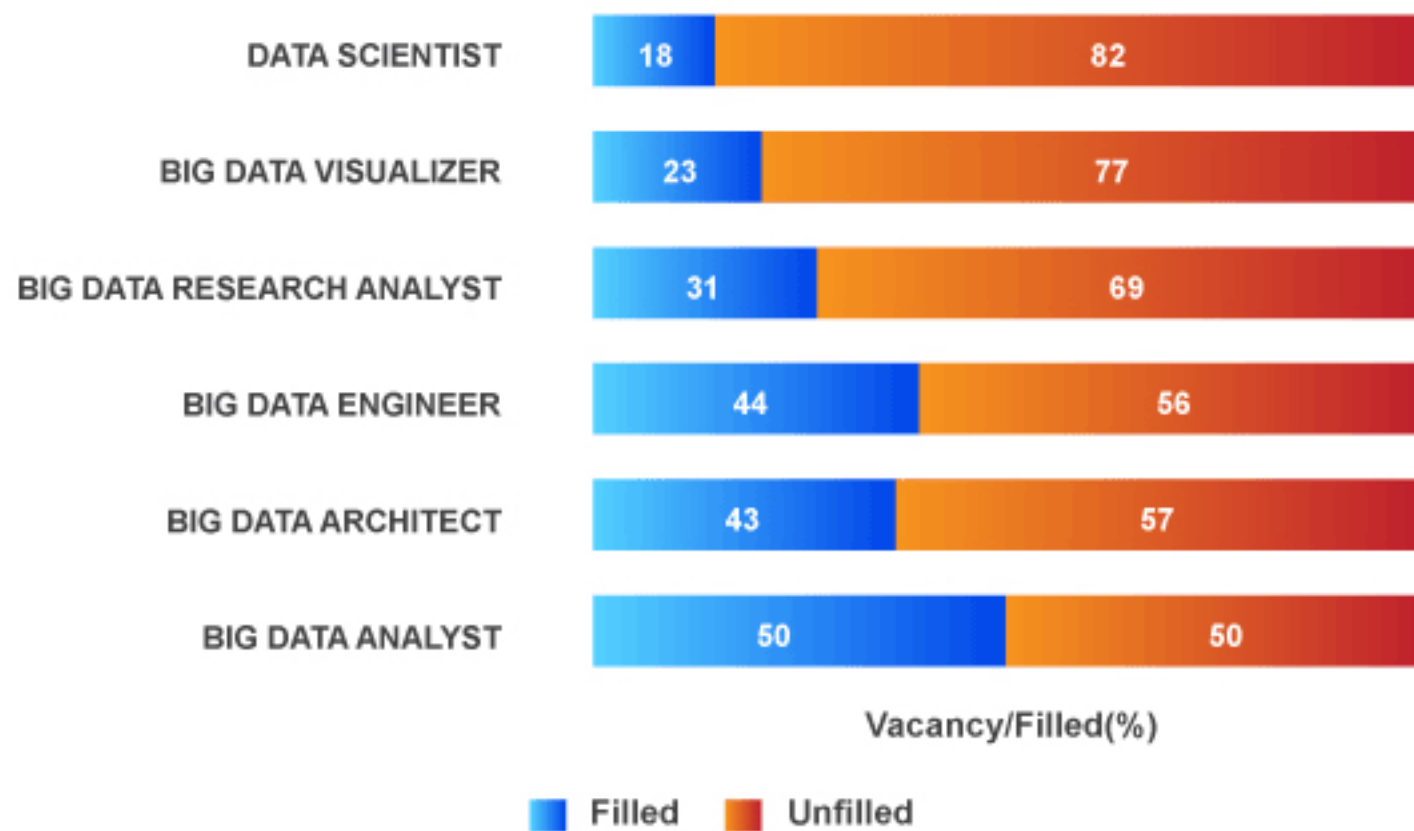
Um relatório de uma empresa de investimentos americana, a Avendus Capital, estima que o mercado de Big Data chegará a 60 bilhões de dólares em 2017, o que pode representar um incrível crescimento na busca por Cientistas de Dados que saibam coletar, armazenar e analisar big Data.



Data Science Academy

Cientistas de Dados e Hadoop

Filled job vs unfilled jobs in big data



Cientistas de Dados e Hadoop

Aprender ou não Hadoop é uma escolha sua.

Mas com certeza este conhecimento será um grande diferencial na sua carreira e na sua compreensão sobre como armazenar e analisar Big Data.



Data Science Academy

Obrigado

