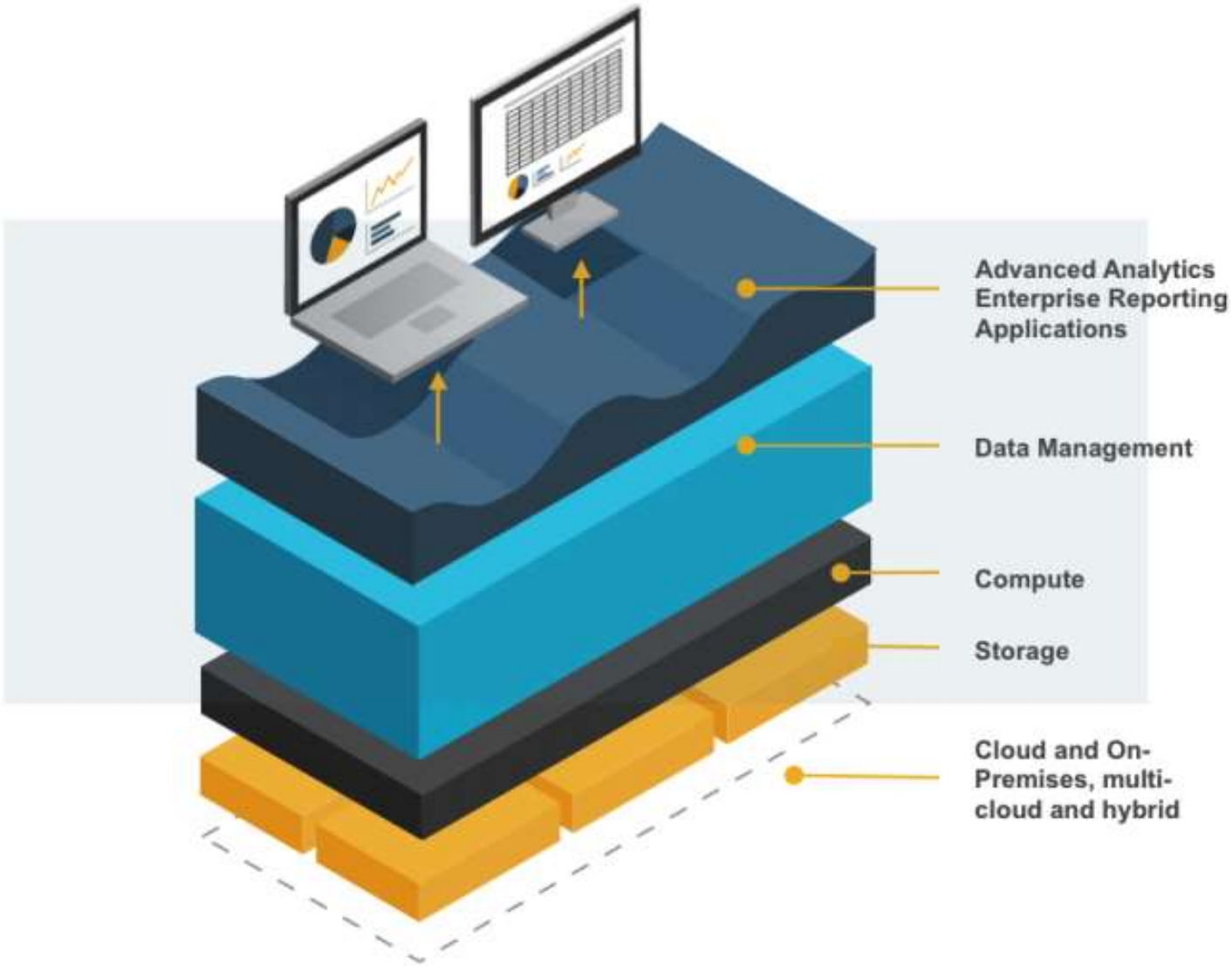
DATA LAKE

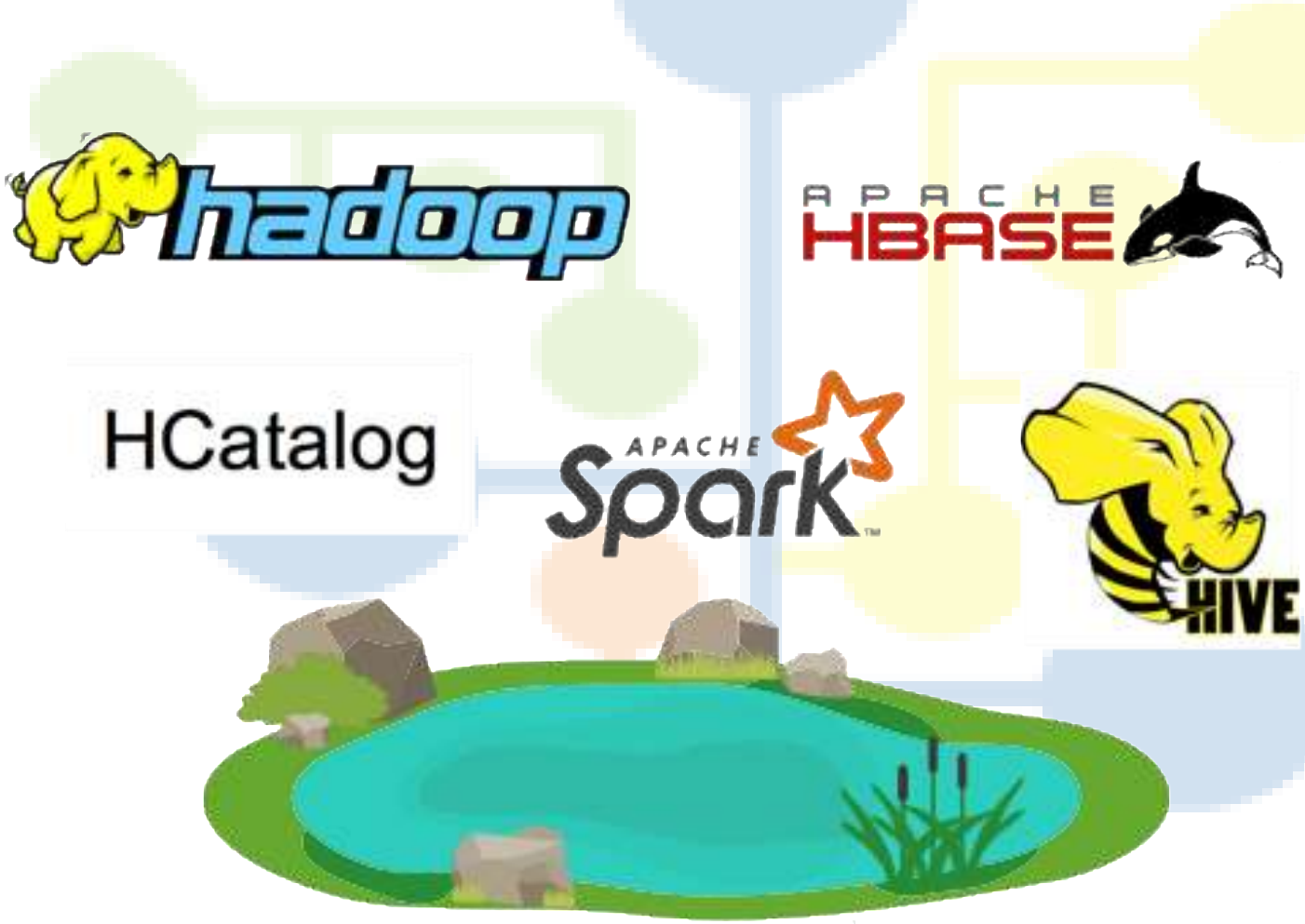
DESIGN, PROJETO E INTEGRAÇÃO

Data Lake Design

Data Lake - Design

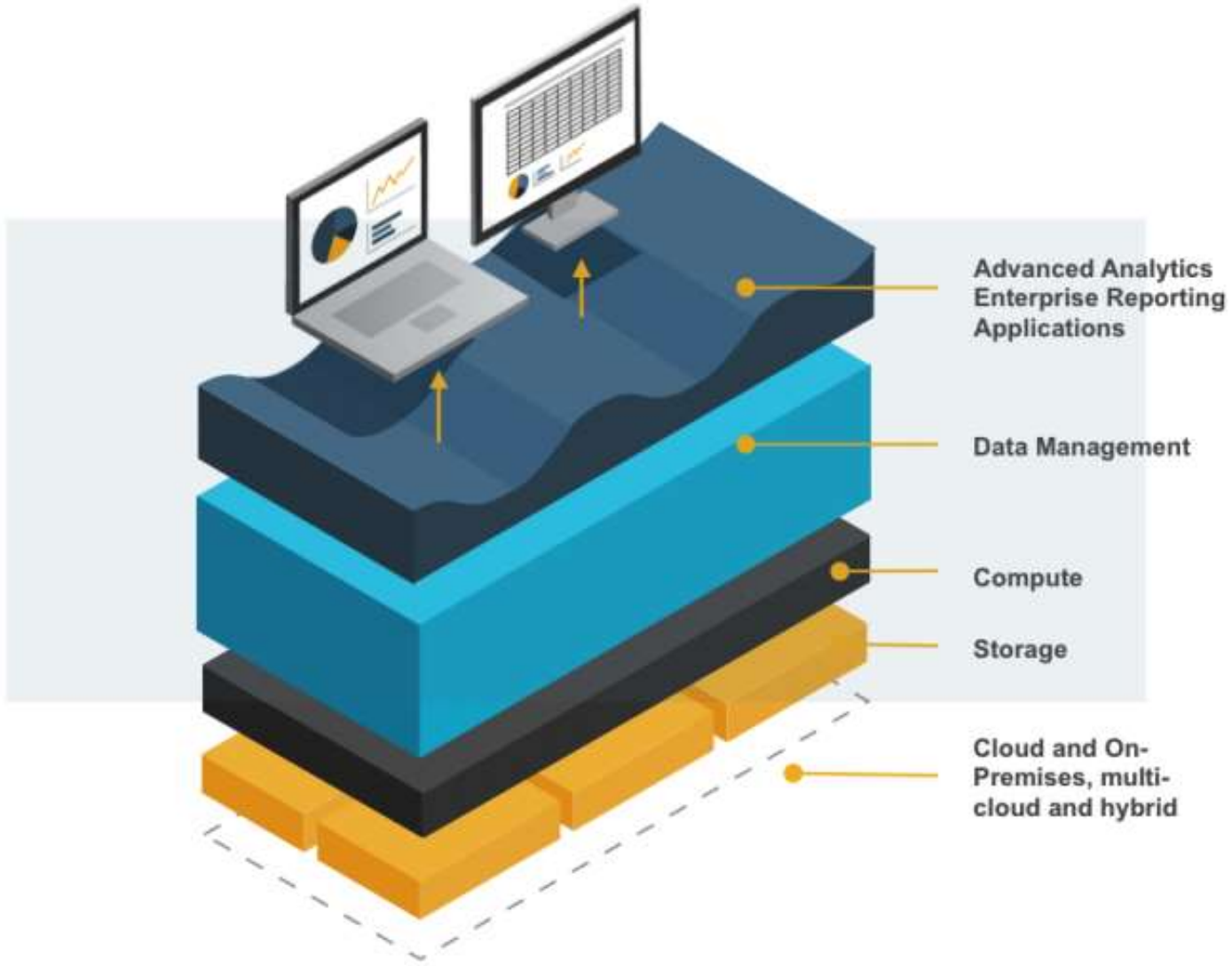


Data Lake - Design



Data Lake Stack

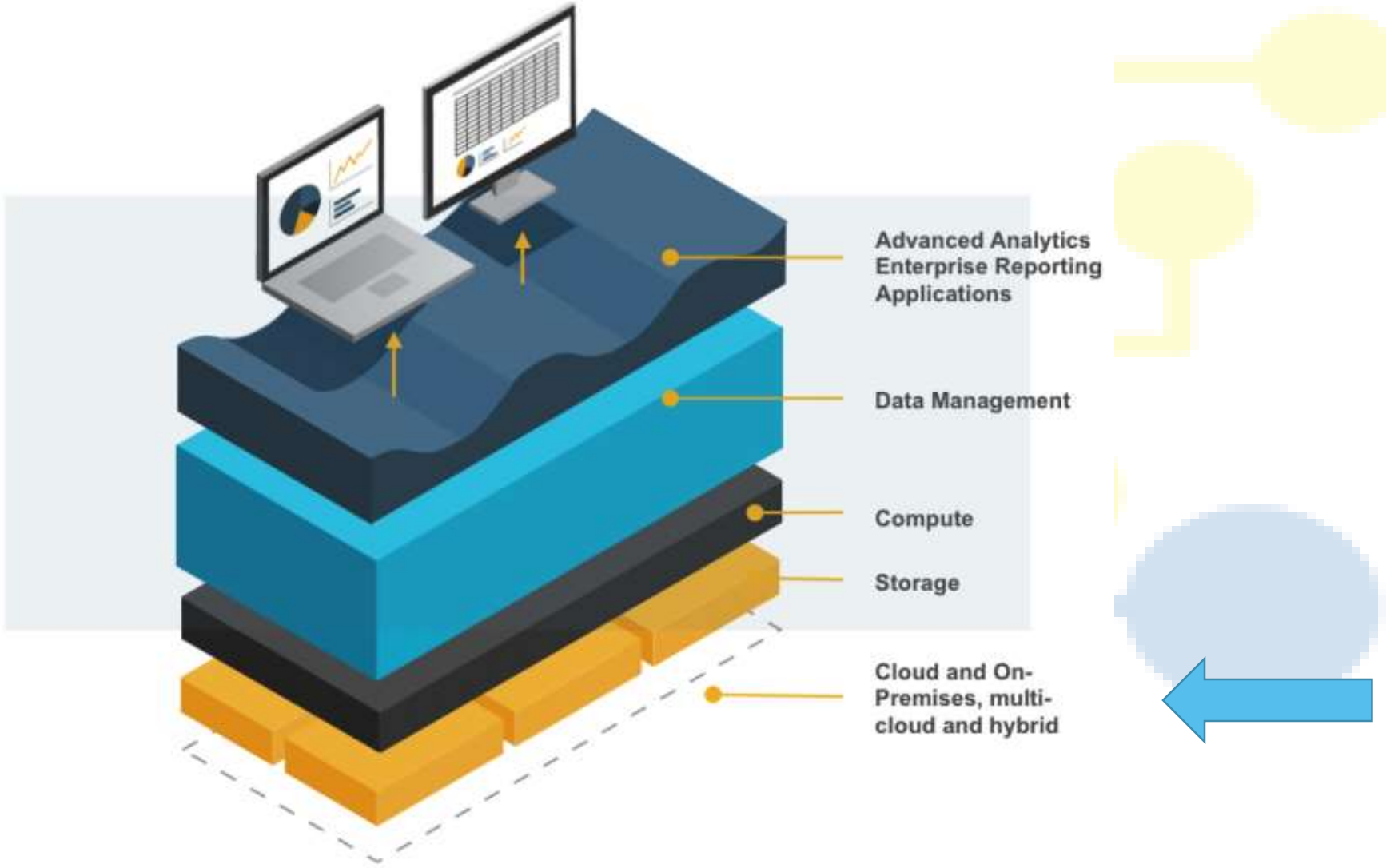
Data Lake Stack



Estes 4 componentes devem ser considerados durante a fase de Design do Data Lake e independem do tipo de implementação, que forma a base do Data Lake Stack.

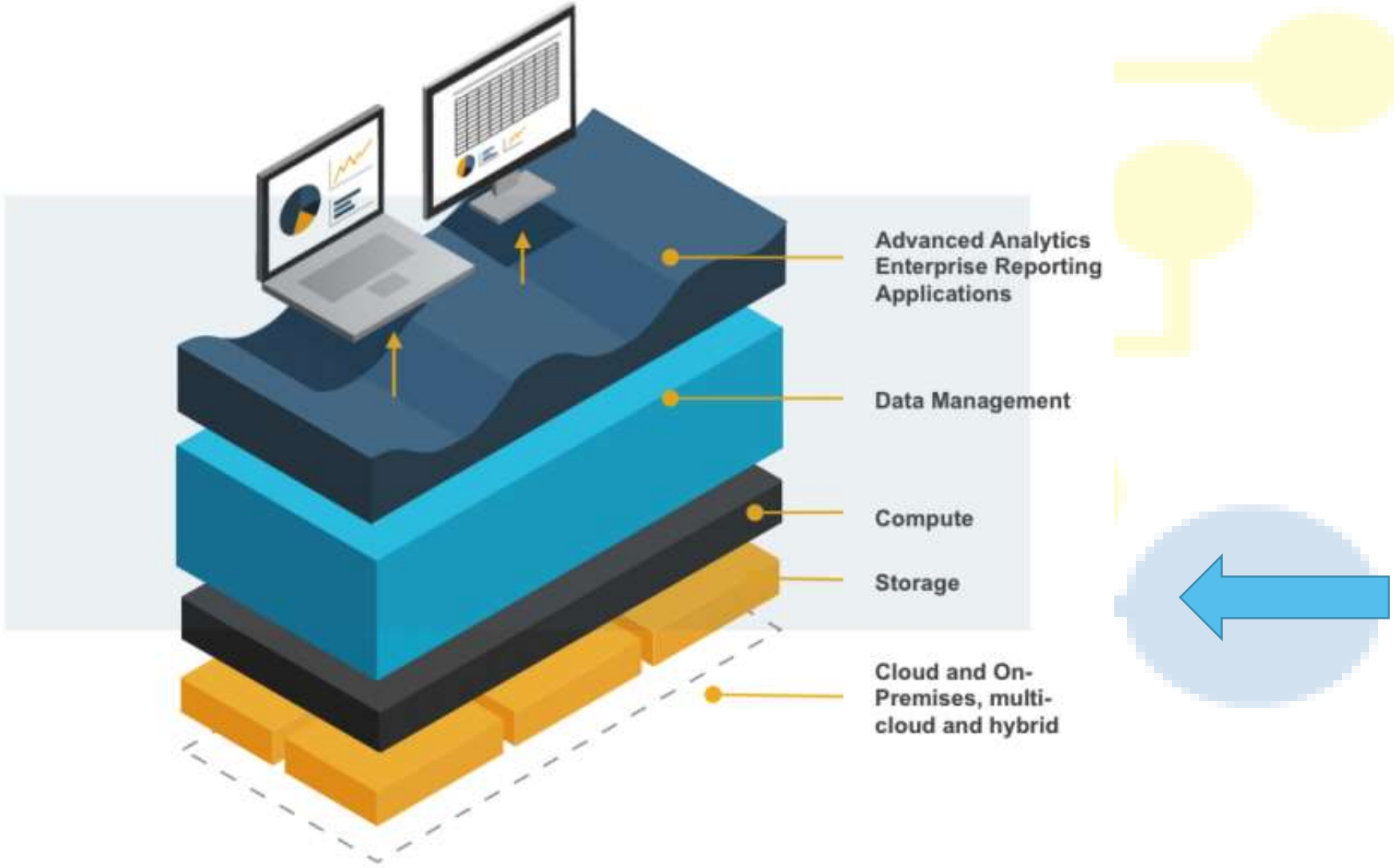
Cloud, On-Premises, Multicloud ou Hybrid

Cloud, On-Premises, Multicloud ou Hybrid



Escolha do Data Storage

Escolha do Data Storage

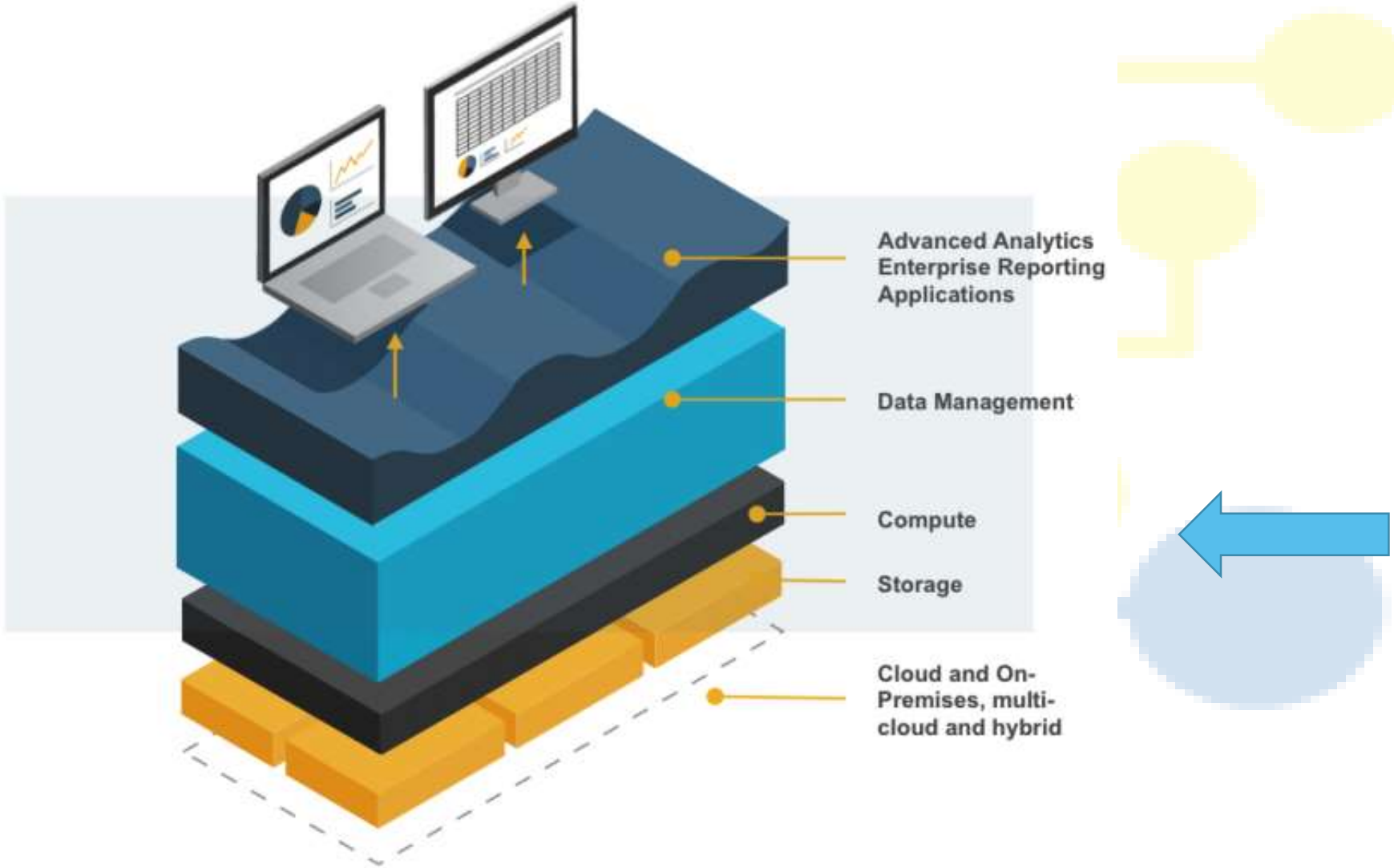


Escolha do Data Storage

- Hadoop Distributed File System (HDFS)
- Object Storage
  - Amazon Simple Storage Service - Amazon S3
  - Microsoft Azure Blob Storage
  - Google Cloud Storage
- Apache Hive Tables
- Apache Hbase
- ElasticSearch

Escolha da Solução de Data Lake Processing

Escolha da Solução de Data Lake Processing



Escolha da Solução de Data Lake Processing

Processamento em Batch

Apache MapReduce
Apache Spark*
Apache Hive
Apache Pig

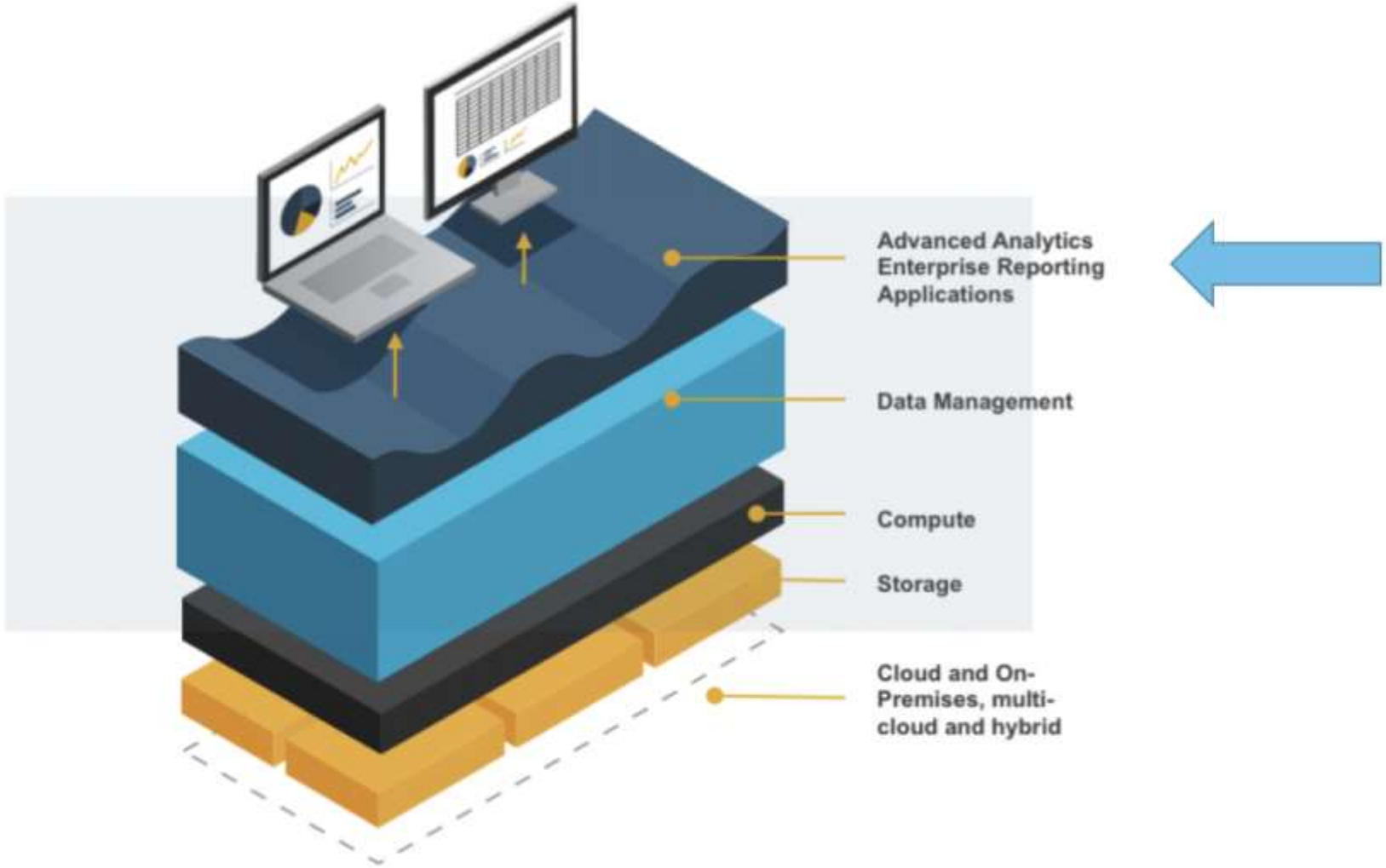
Processamento em Tempo Real (Streaming)

Apache Spark Streaming
Apache Kafka
Apache Flume
Apache Storm

Apache Drill
Apache NiFi
Apache Beam
Apache Sqoop

Advanced Analytics e Enterprise Reporting

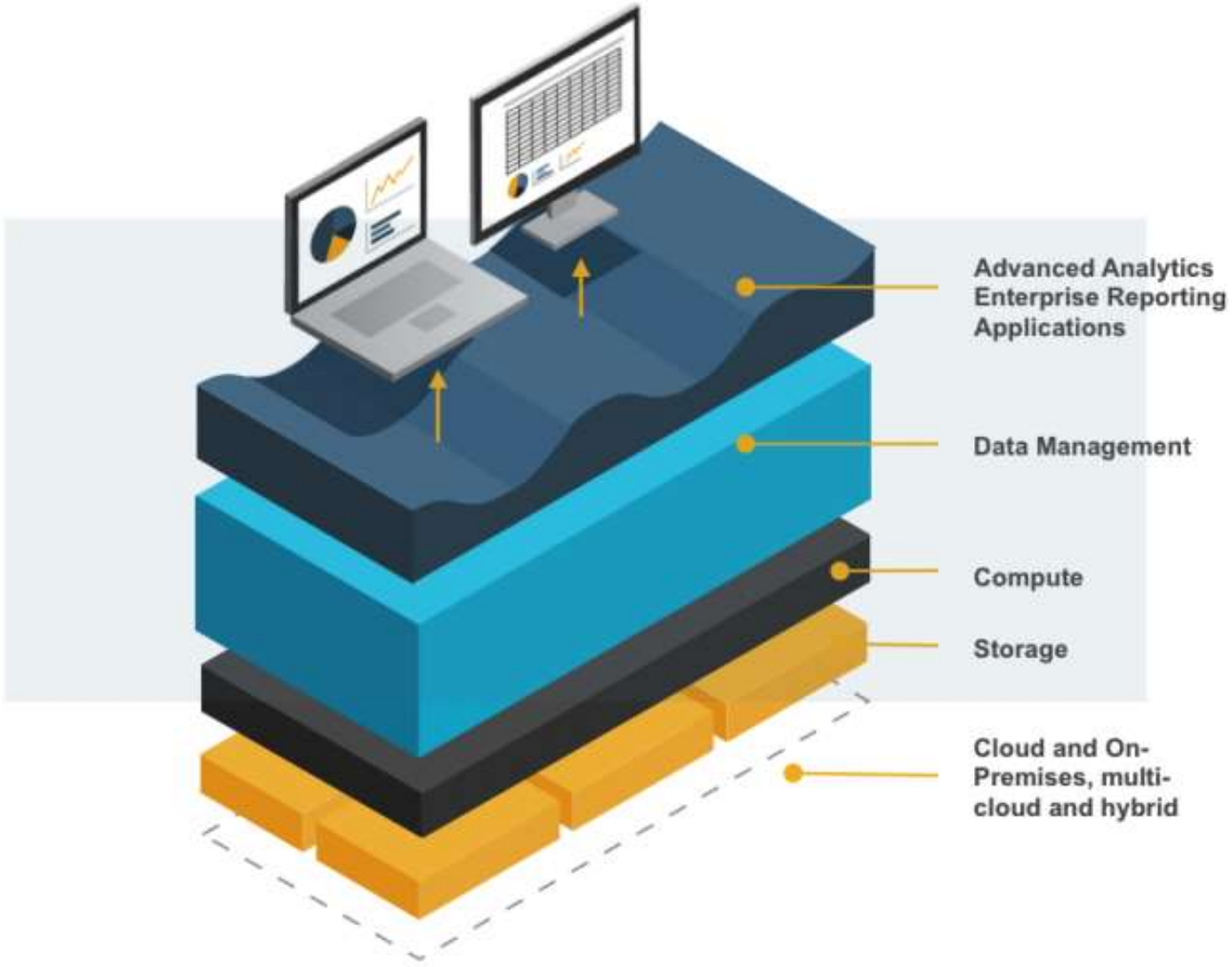
Advanced Analytics e Enterprise Reporting



Advanced Analytics e Enterprise Reporting

- Self-Service BI: Tableau, Qlik e Power BI
- Conexão JDBC
- APIs RESTful
- Conexão Ad-hoc
- Execução de Modelos de Machine Learning e IA

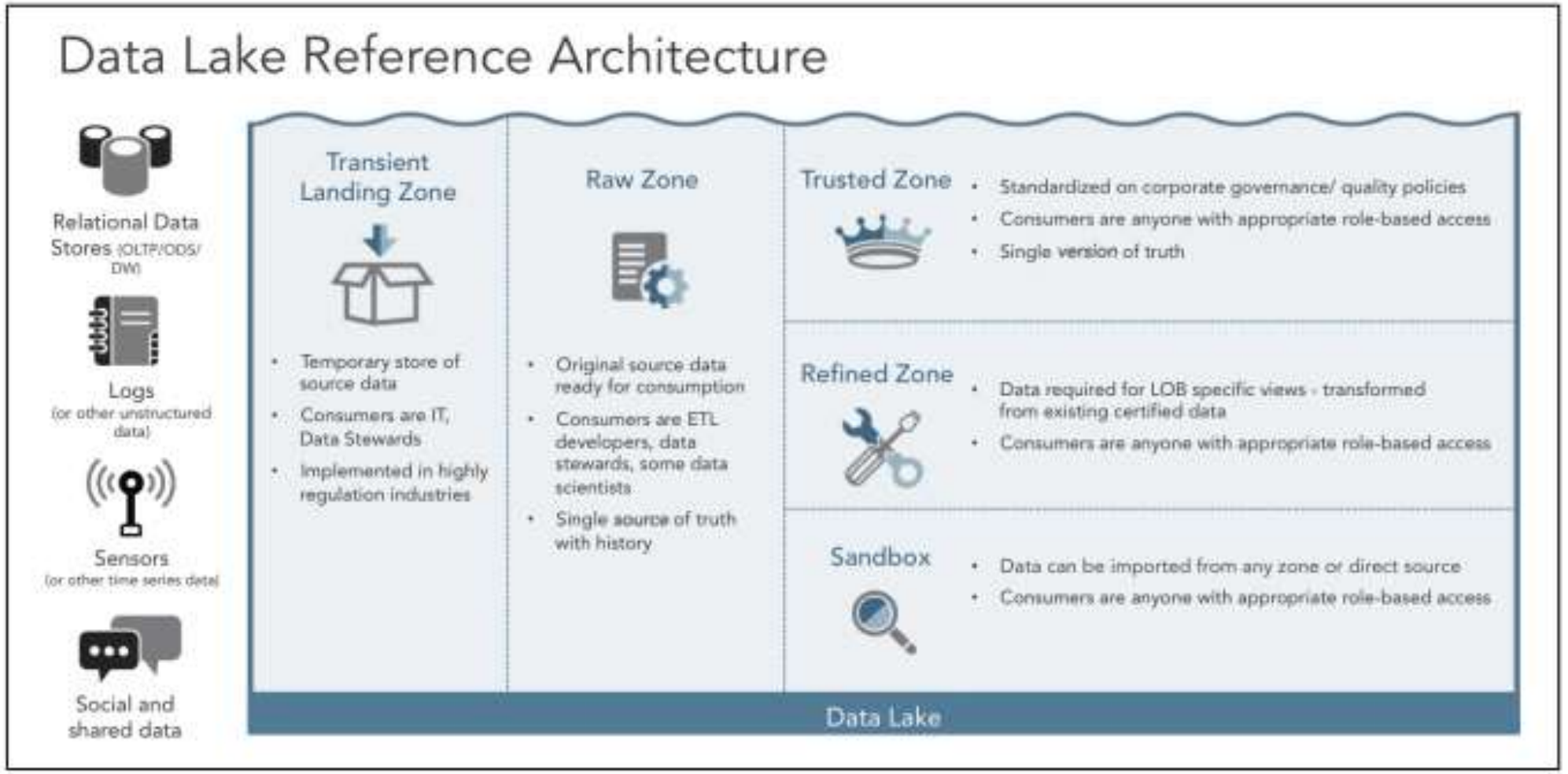
Advanced Analytics e Enterprise Reporting



Referência de Design Para o Data Lake

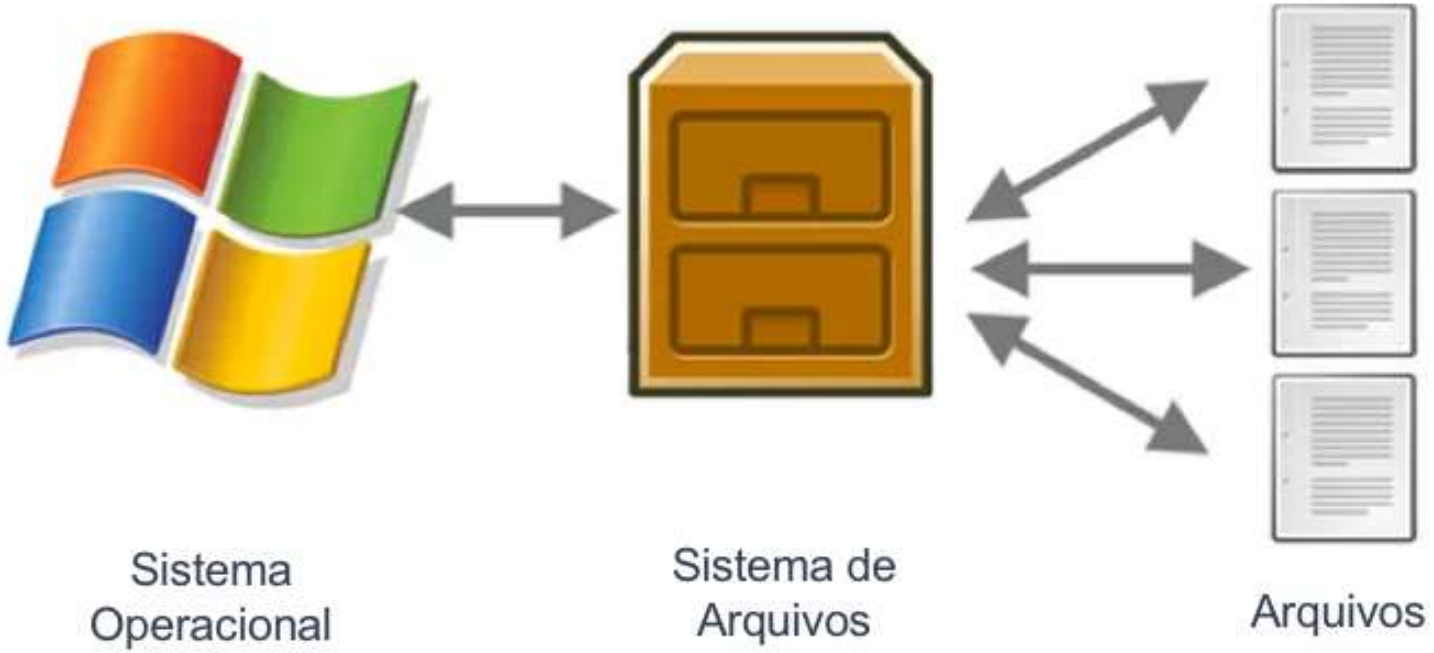
Referência de Design Para o Data Lake

Referência de Design Para o Data Lake



HDFS
Conceito e Importância

HDFS - Conceito e Importância

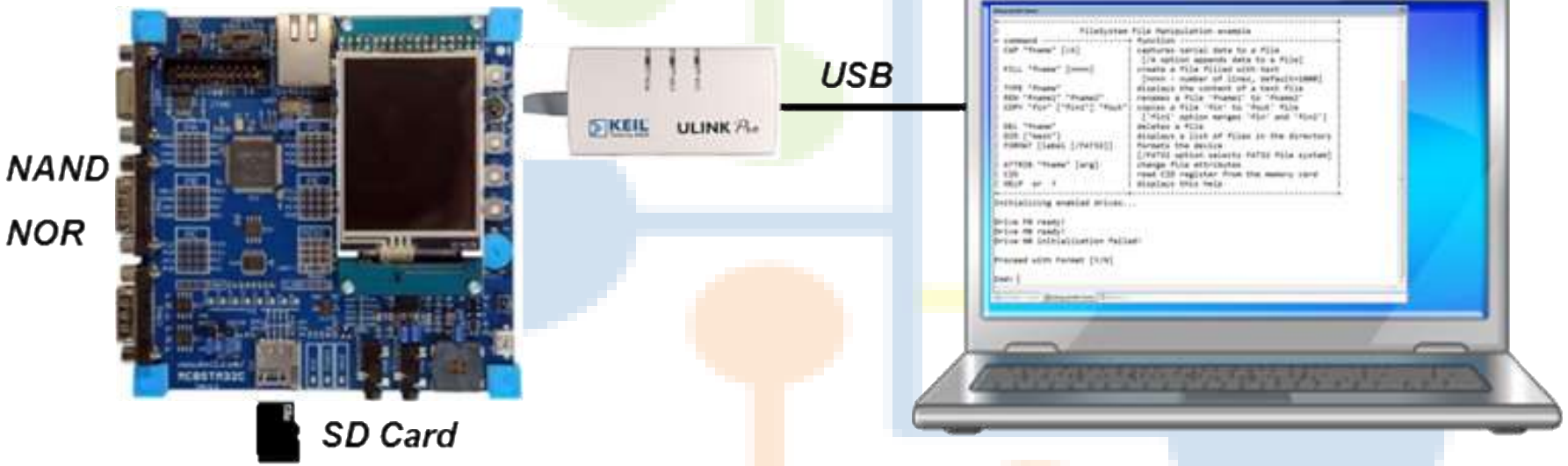


HDFS - Conceito e Importância

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 aruqivos 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 par a 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 `/proc`.

HDFS - Conceito e Importância



HDFS - Conceito e Importância

HDFS - Conceito e Importância

Sistema de Arquivos Distribuído

Principais Características do HDFS

Principais Características do HDFS

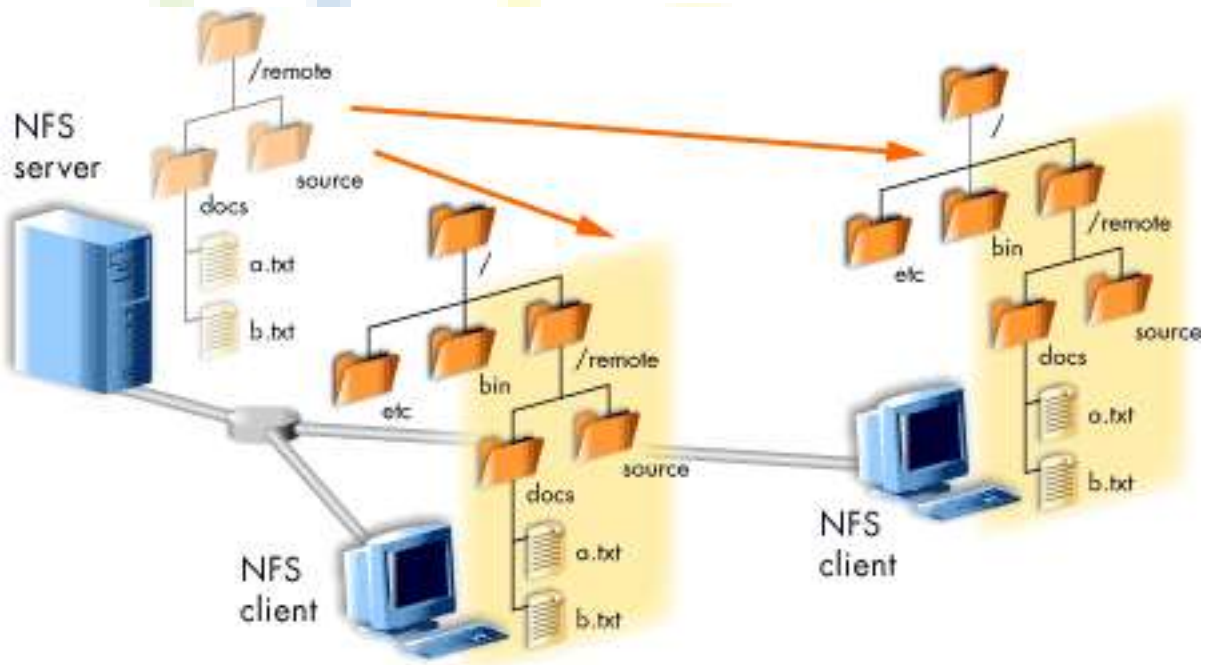
- Tolerância a Falhas
- Integridade
- Segurança
- Desempenho
- Consistência

Outros Sistemas de Arquivos Distribuídos

Outros Sistemas de Arquivos Distribuídos

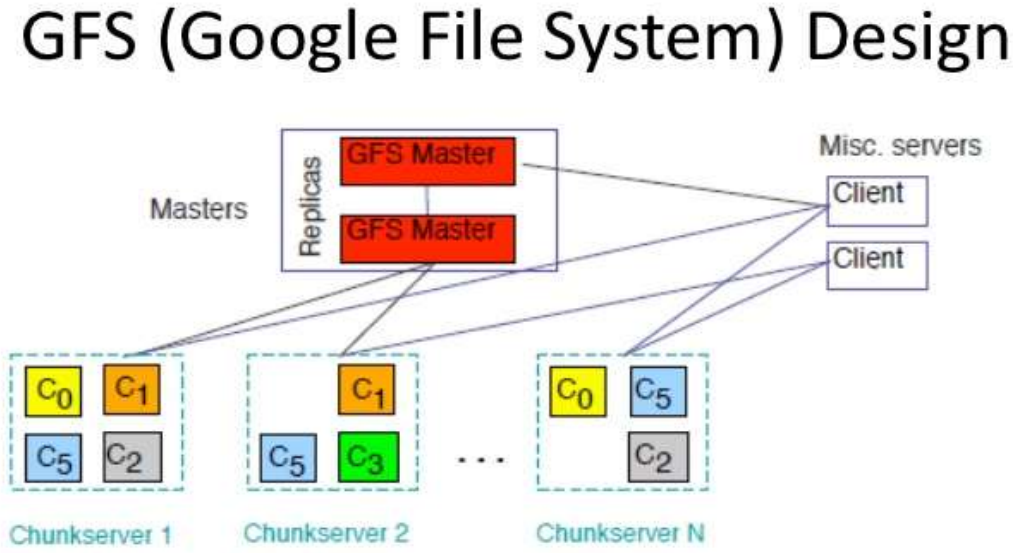
Outros Sistemas de Arquivos Distribuídos

Network File System



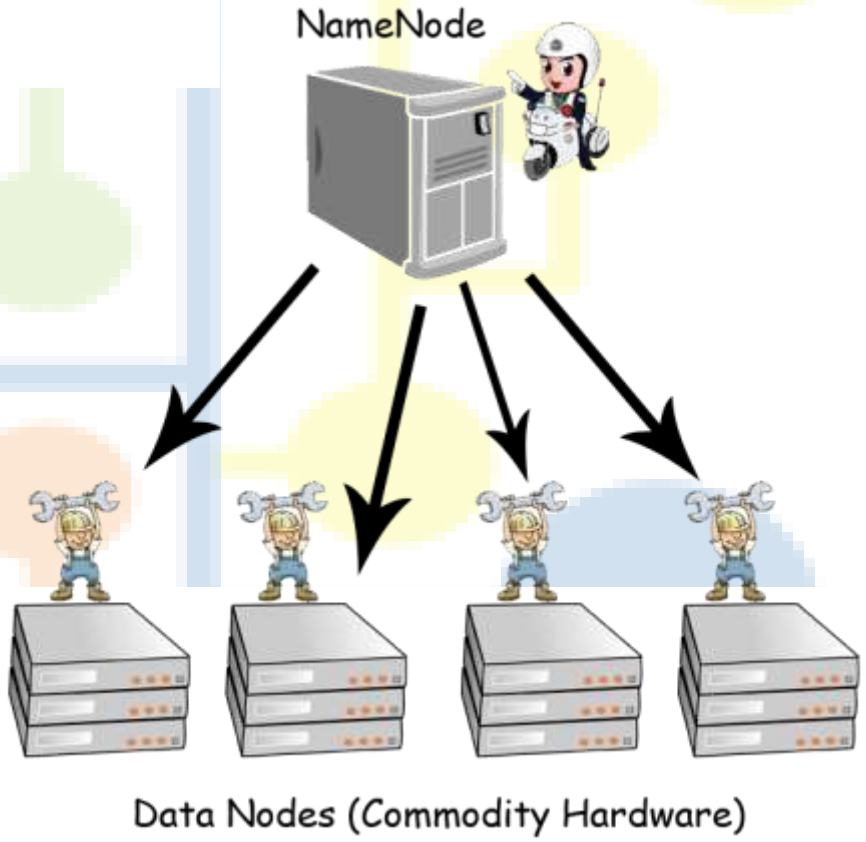
Outros Sistemas de Arquivos Distribuídos

Google File System



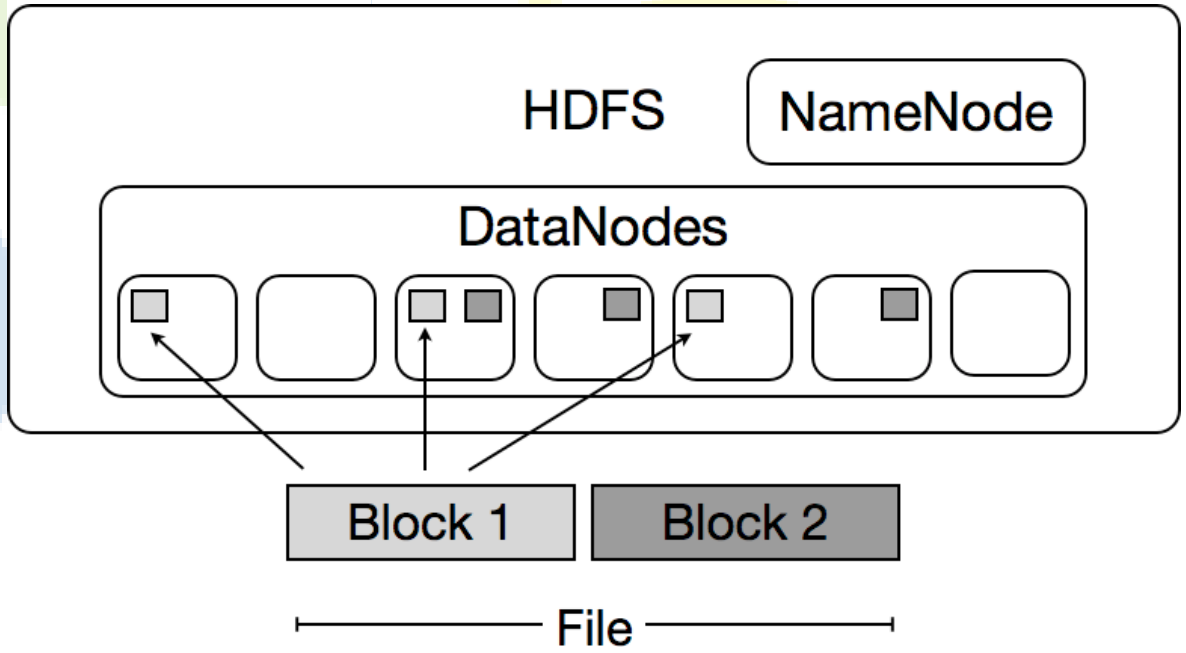
Outros Sistemas de Arquivos Distribuídos

Hadoop Distributed File System



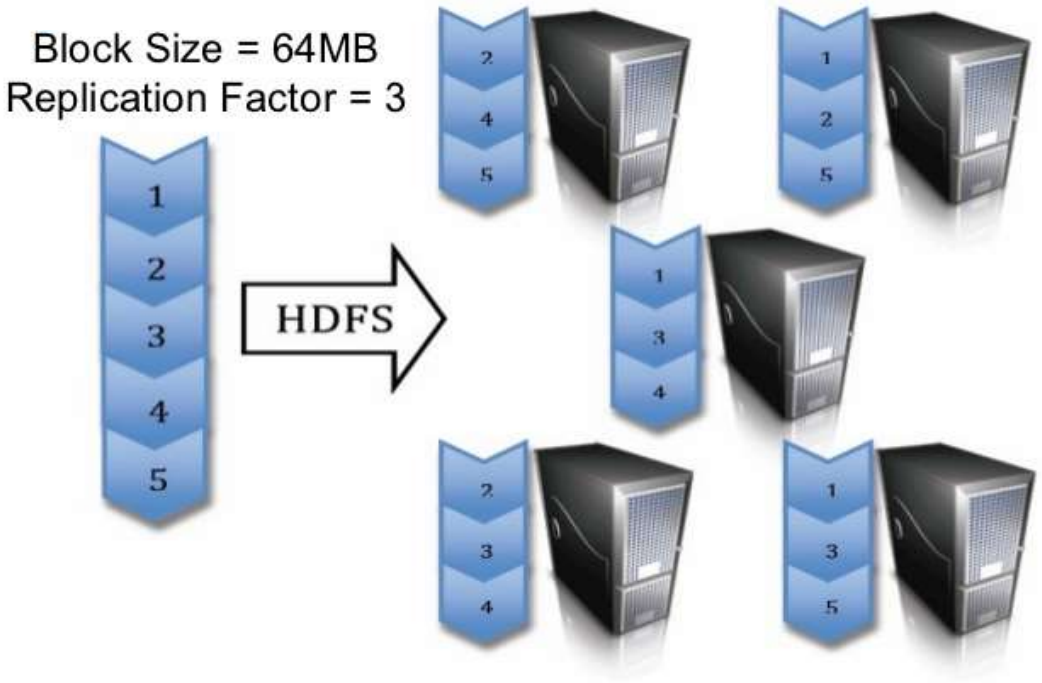
Outros Sistemas de Arquivos Distribuídos

Hadoop Distributed File System



Outros Sistemas de Arquivos Distribuídos

Hadoop Distributed File System

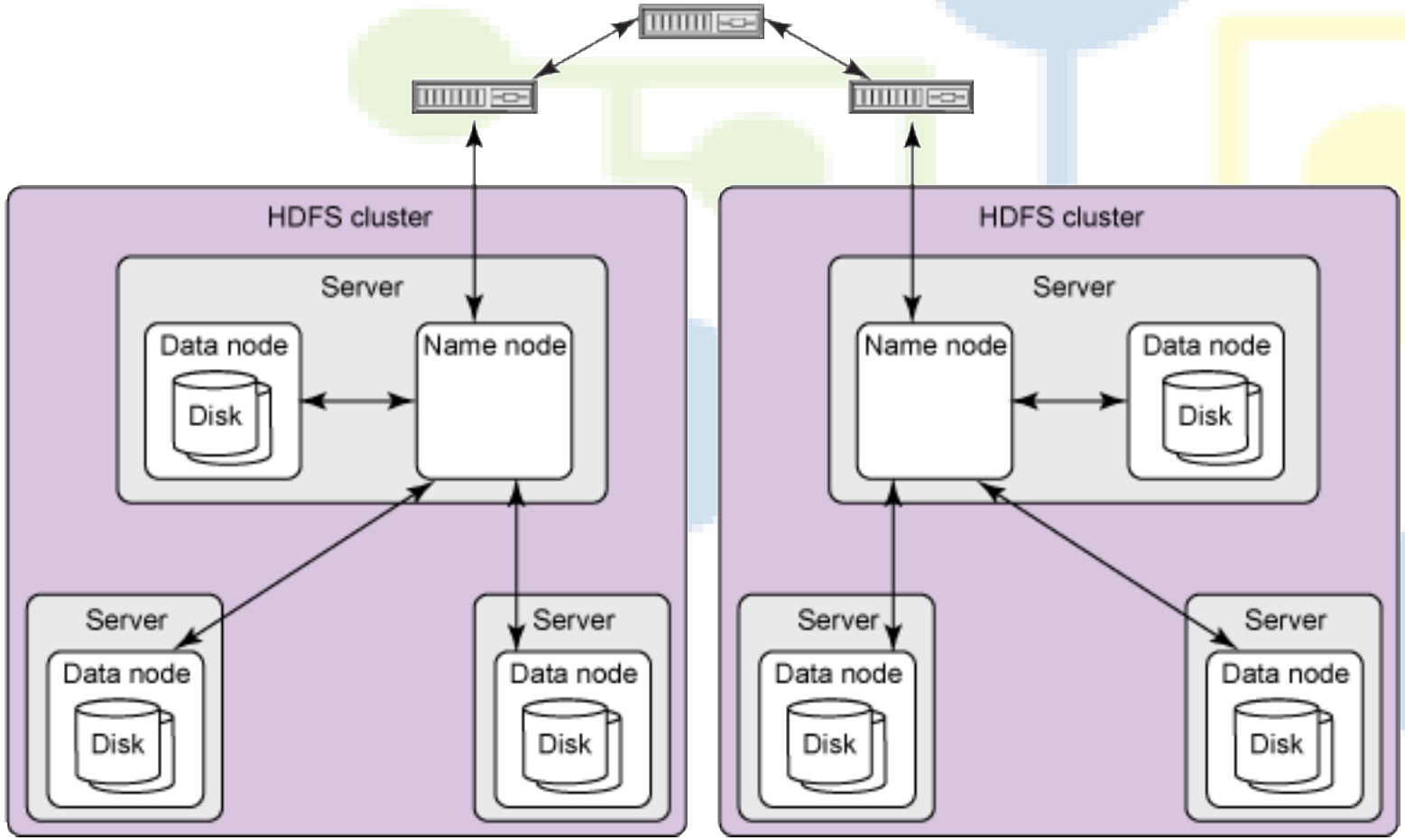


Outros Sistemas de Arquivos Distribuídos

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.

Arquitetura HDFS

Arquitetura HDFS

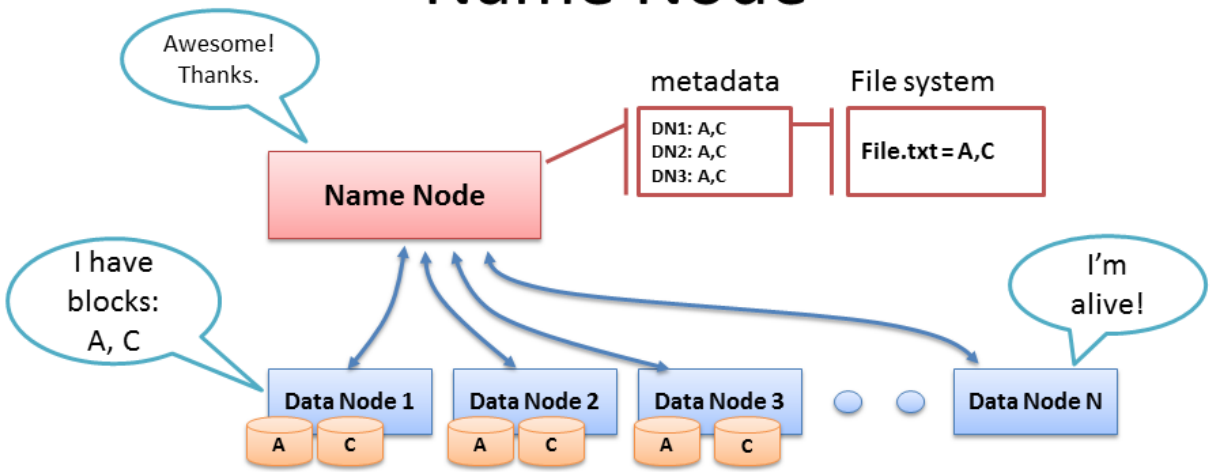


Arquitetura
Master/Slave
Master/Worker

Arquitetura HDFS

Arquitetura
Master/Slave
Master/Worker

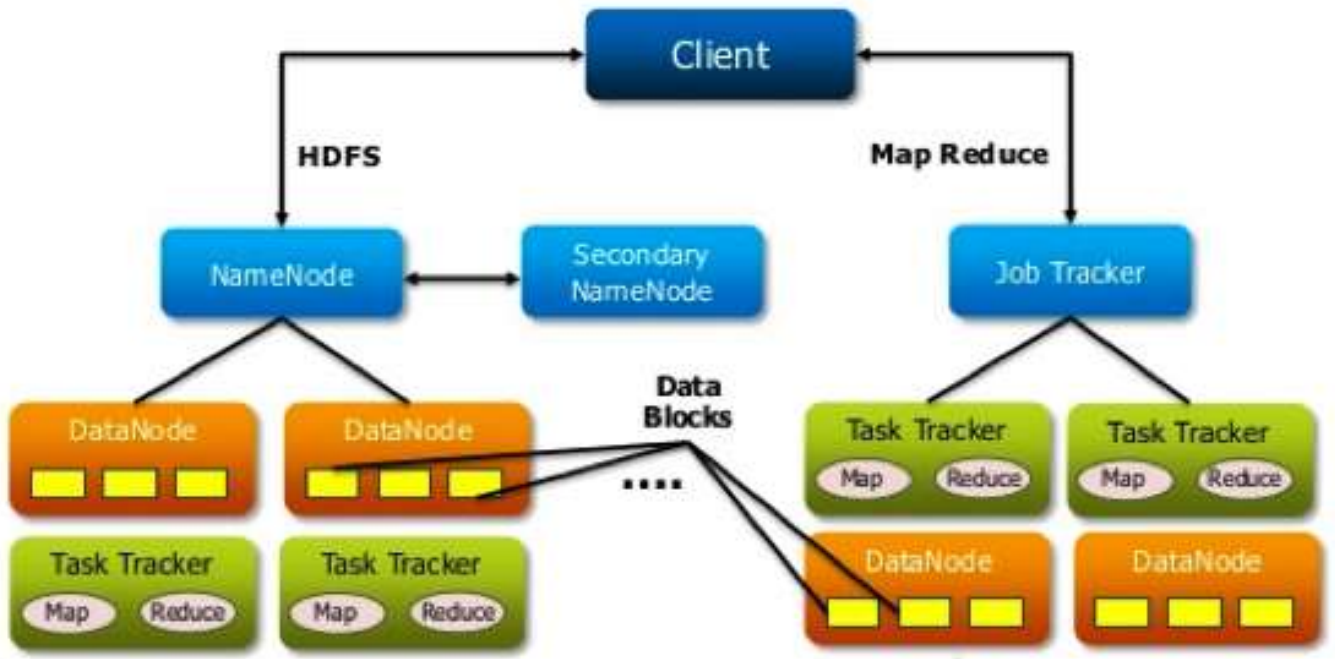
Arquitetura HDFS



NameNode

- FsImage
- EditLog

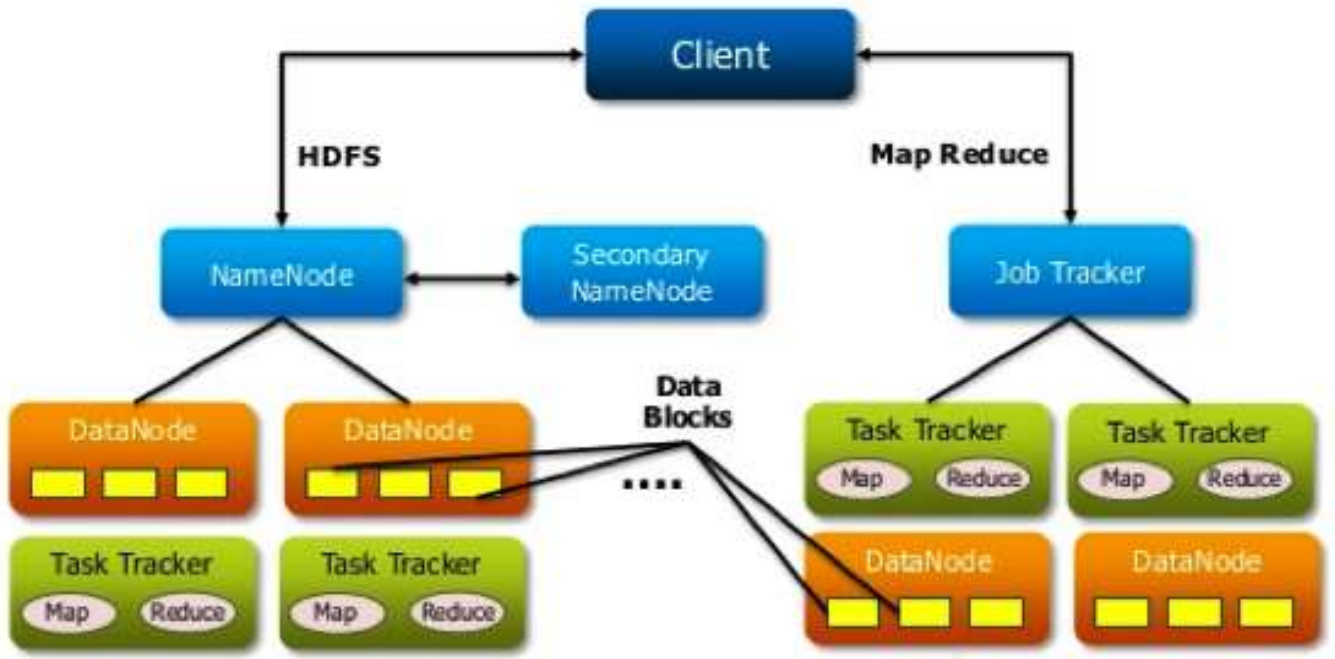
Arquitetura HDFS



NameNode

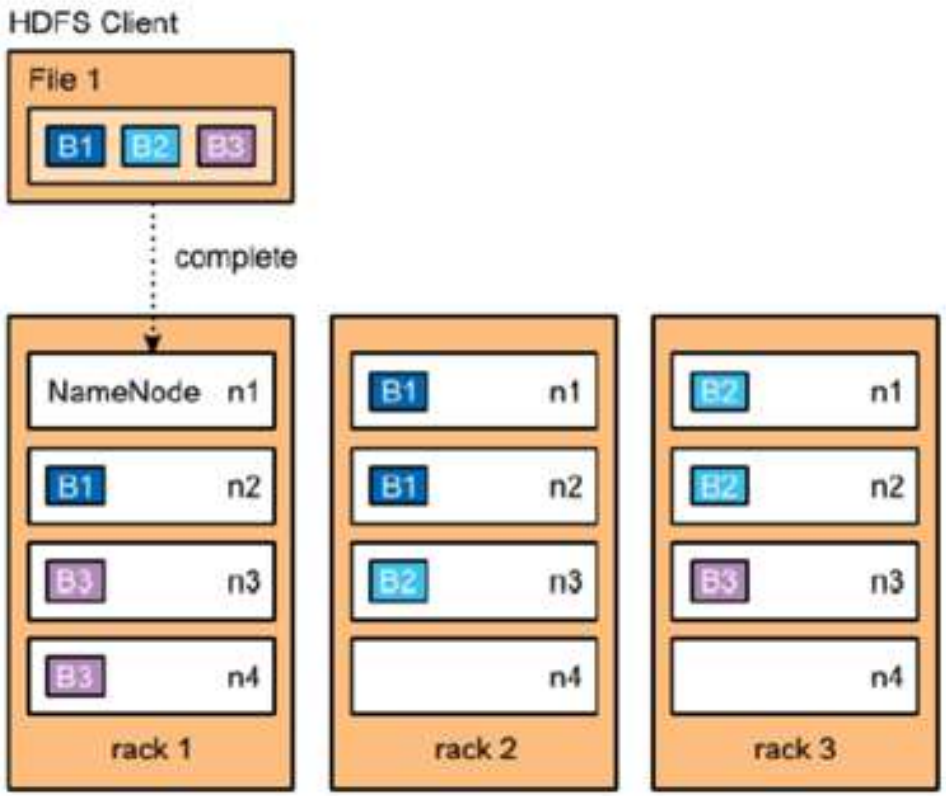
- FsImage
- EditLog

Arquitetura HDFS



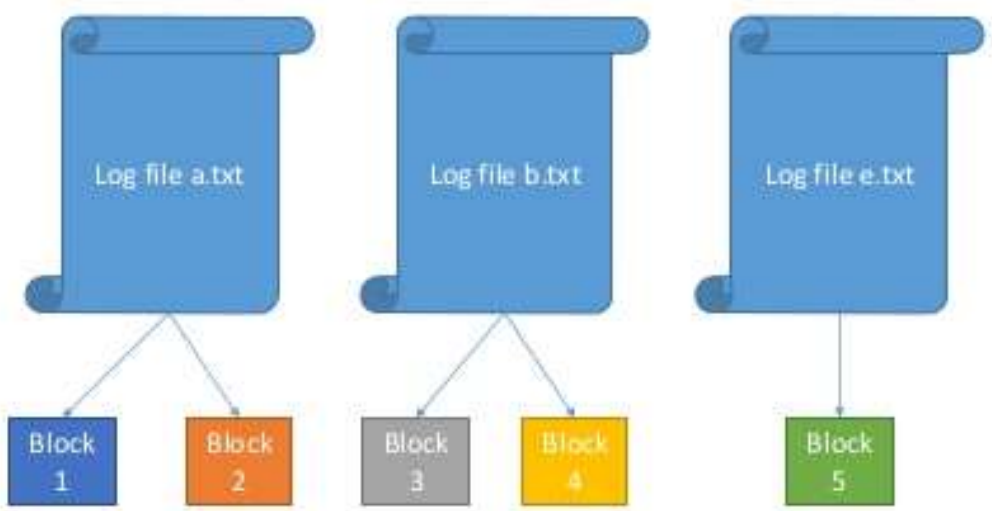
NameNode / DataNode

Arquitetura HDFS



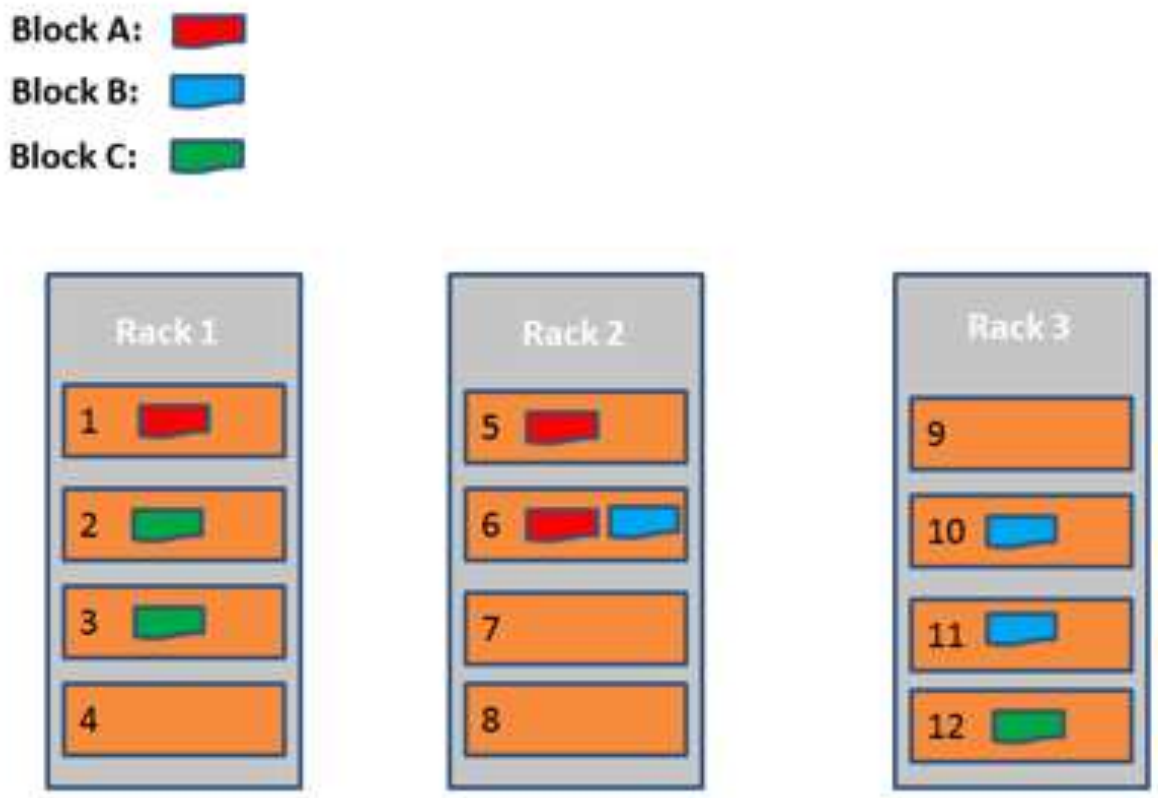
Replicação

Arquitetura HDFS



Replicação

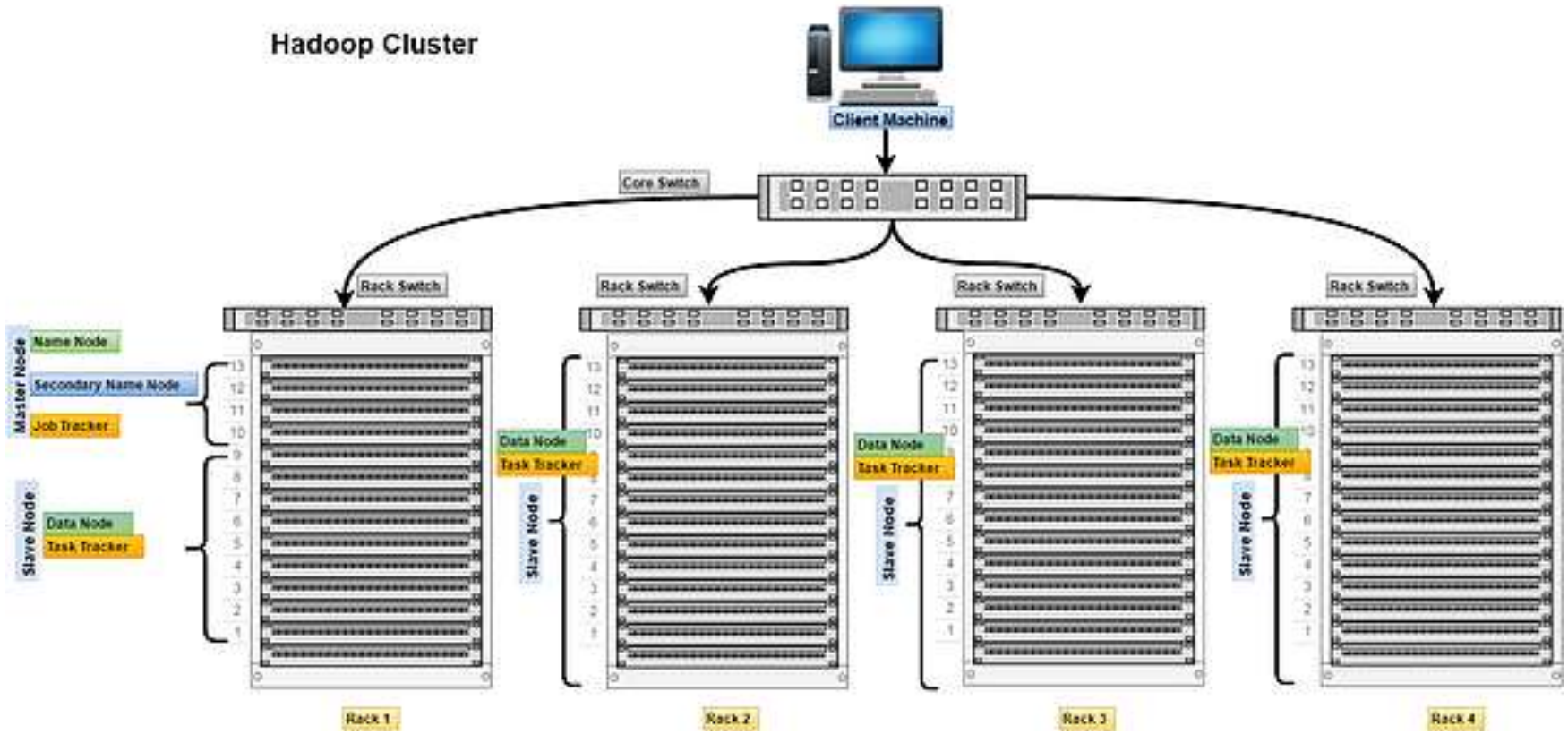
Arquitetura HDFS



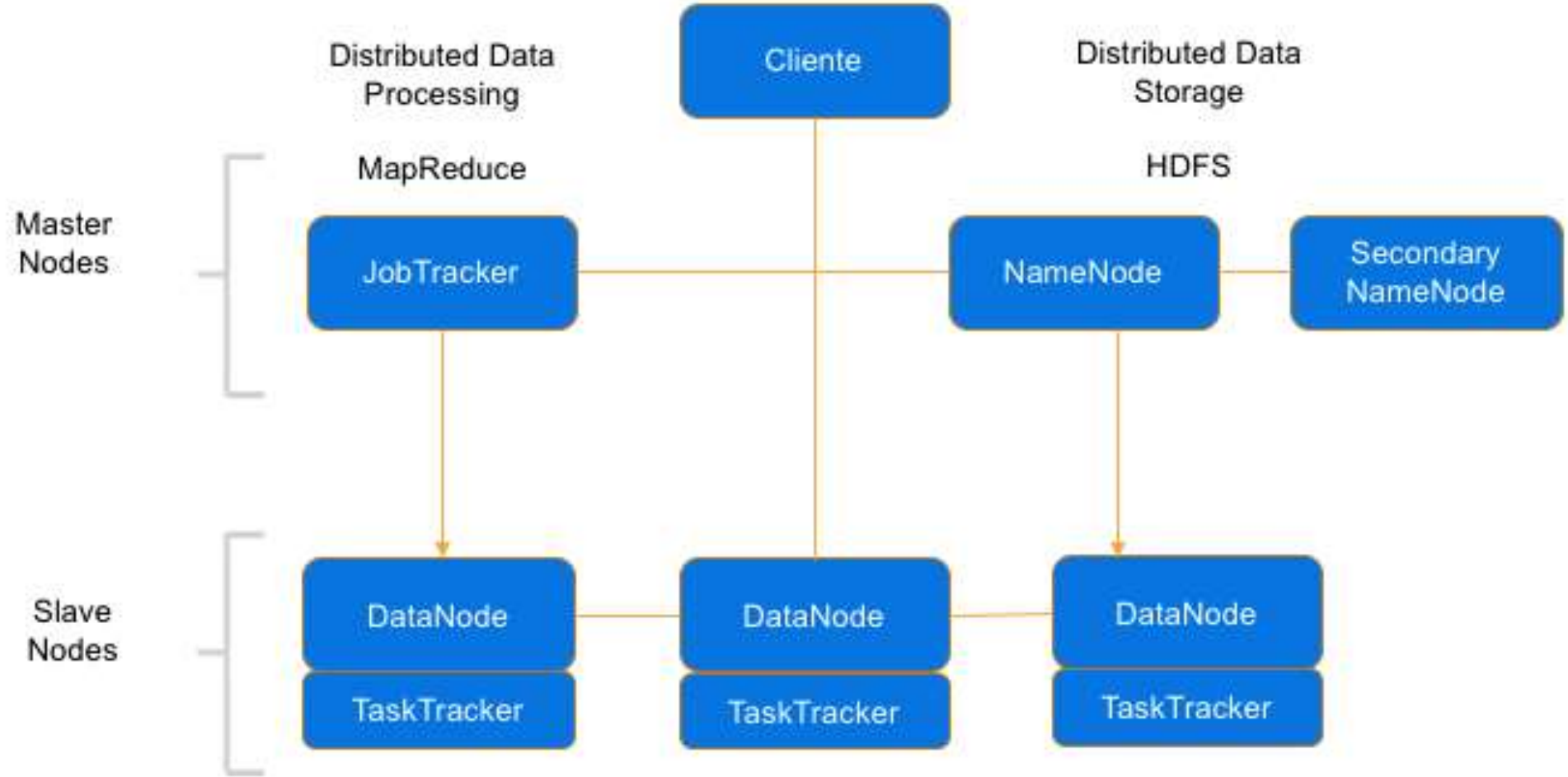
Replicação

Como Funciona o Cluster Hadoop

Como Funciona o Cluster Hadoop



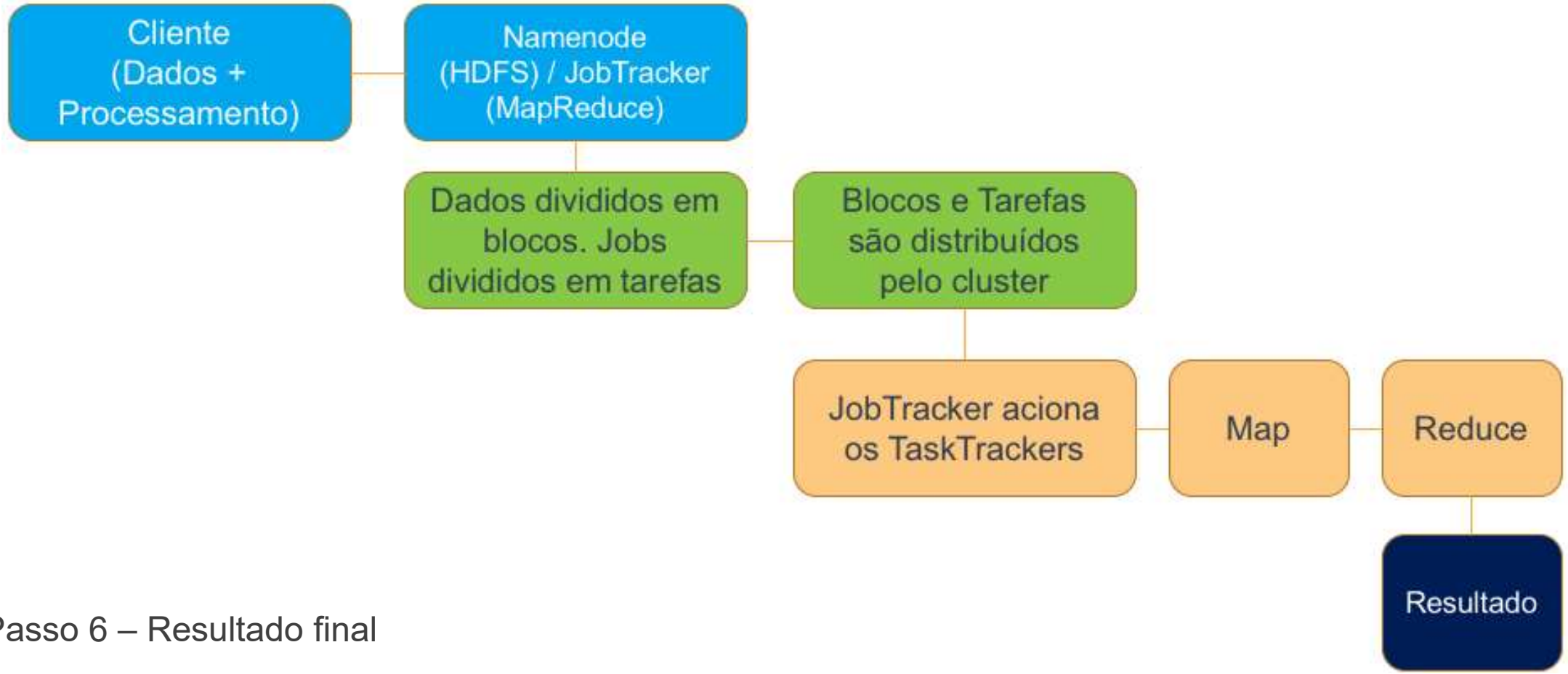
Como Funciona o Cluster Hadoop



Como Funciona o Cluster Hadoop

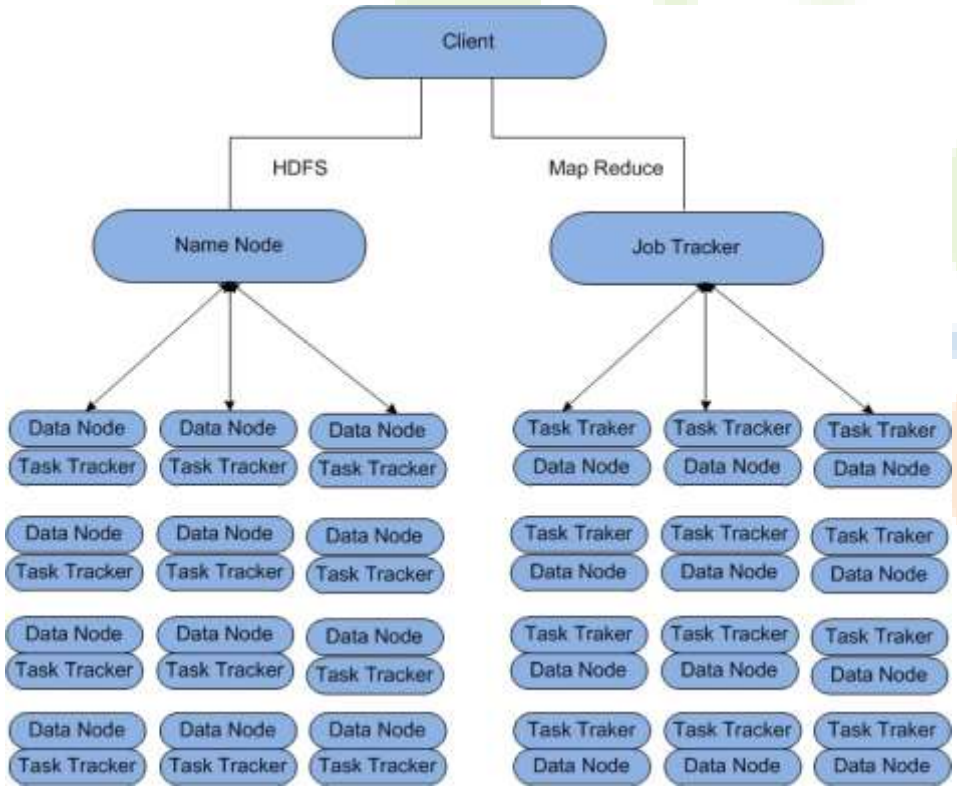
Funcionamento do Cluster

Como Funciona o Cluster Hadoop



Passo 6 – Resultado final

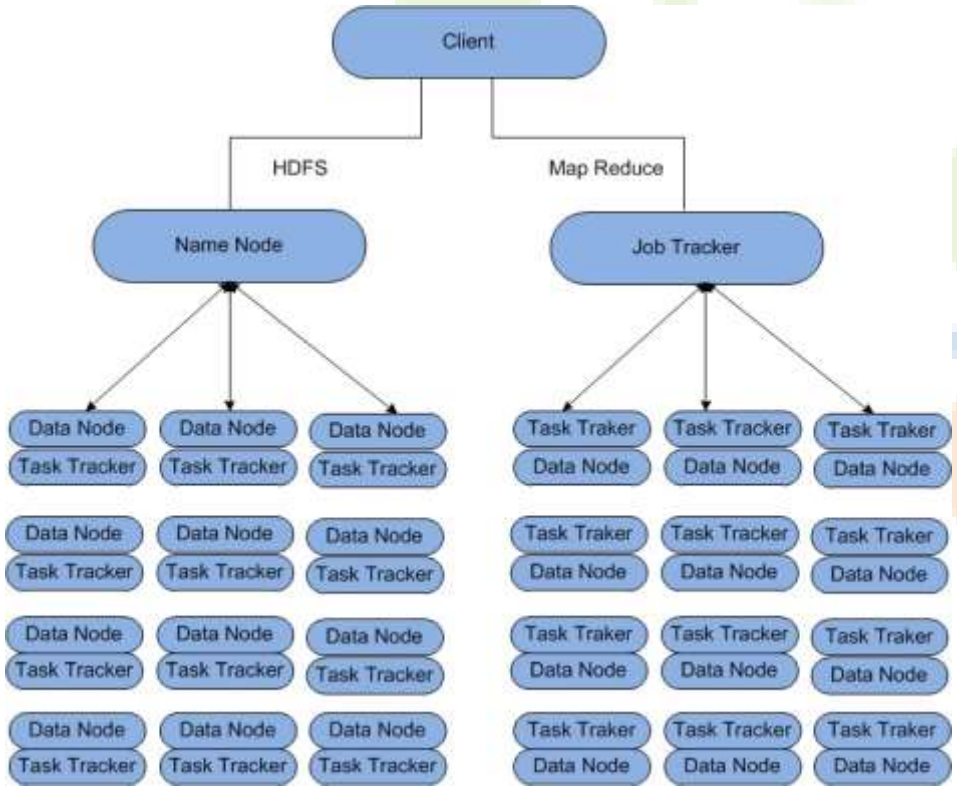
Como Funciona o Cluster Hadoop



Dados → DataNode

Metadados → NameNode

Como Funciona o Cluster Hadoop



Data Node → Armazena/Recupera Dados

TaskTracker → Executa Jobs de MapReduce

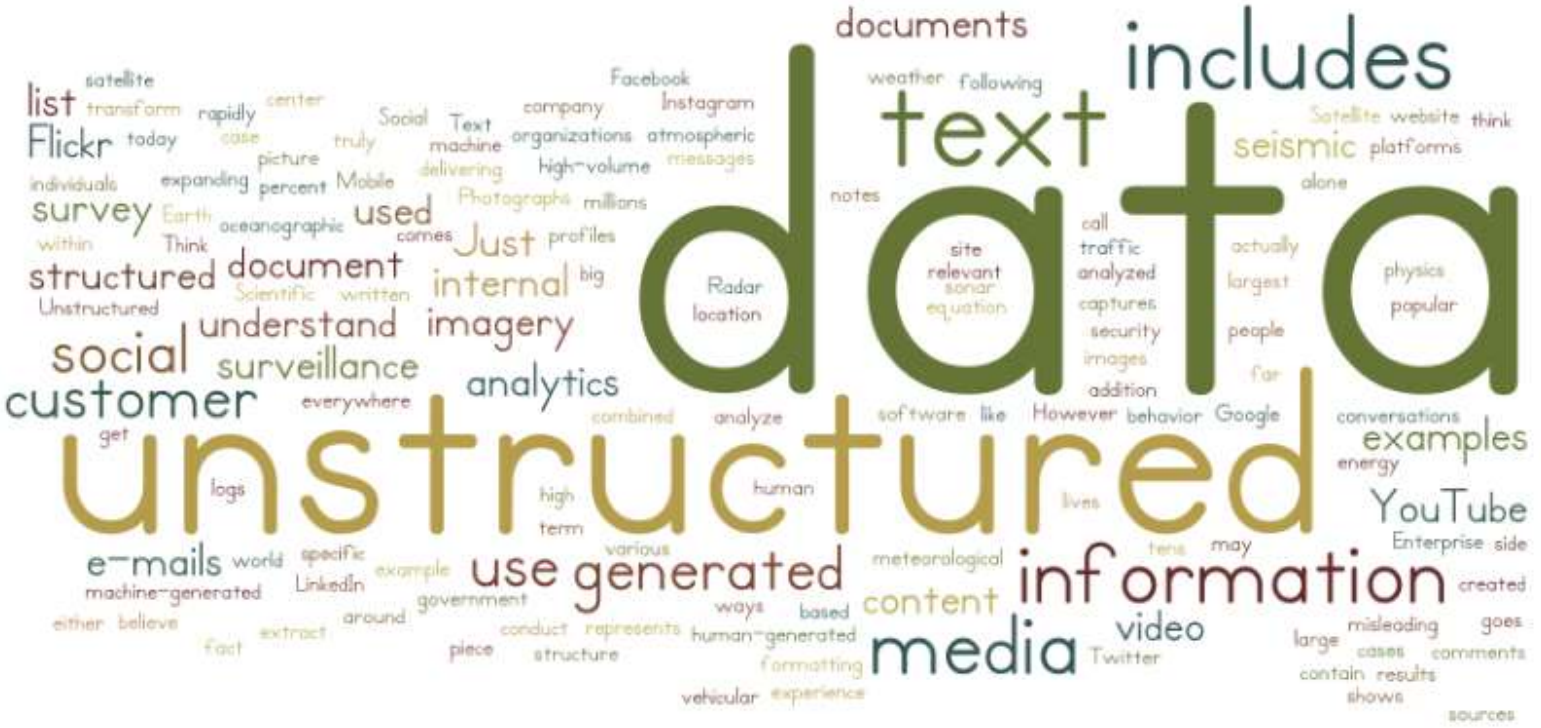
Apache Spark Framework

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Spark Framework

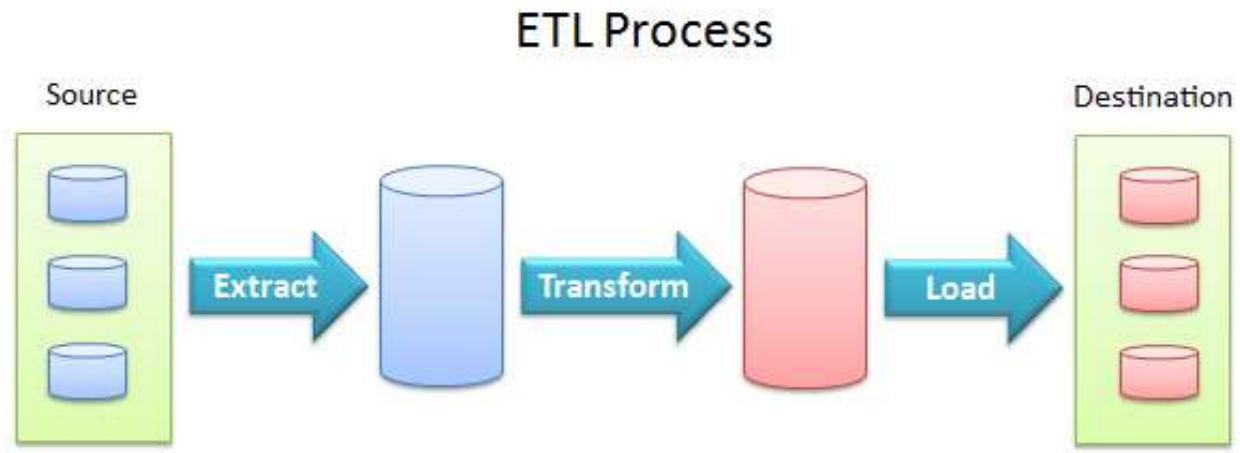
Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Spark Framework



Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Spark Framework



Apache Spark Framework

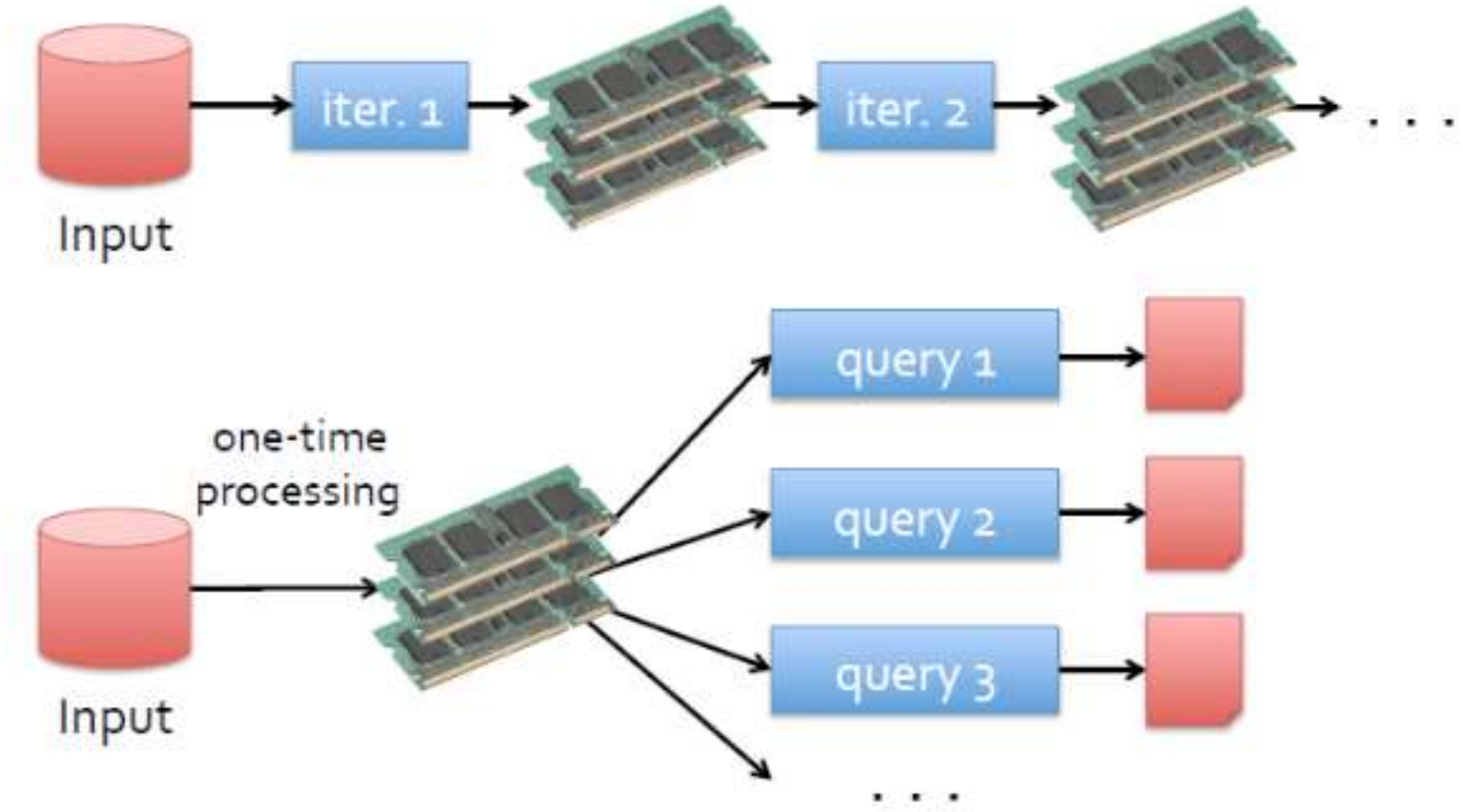
Como armazenar e processar todos esses dados, se o volume aumenta de forma exponencial?

Apache Spark Framework

Apache Spark Framework

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Spark Framework

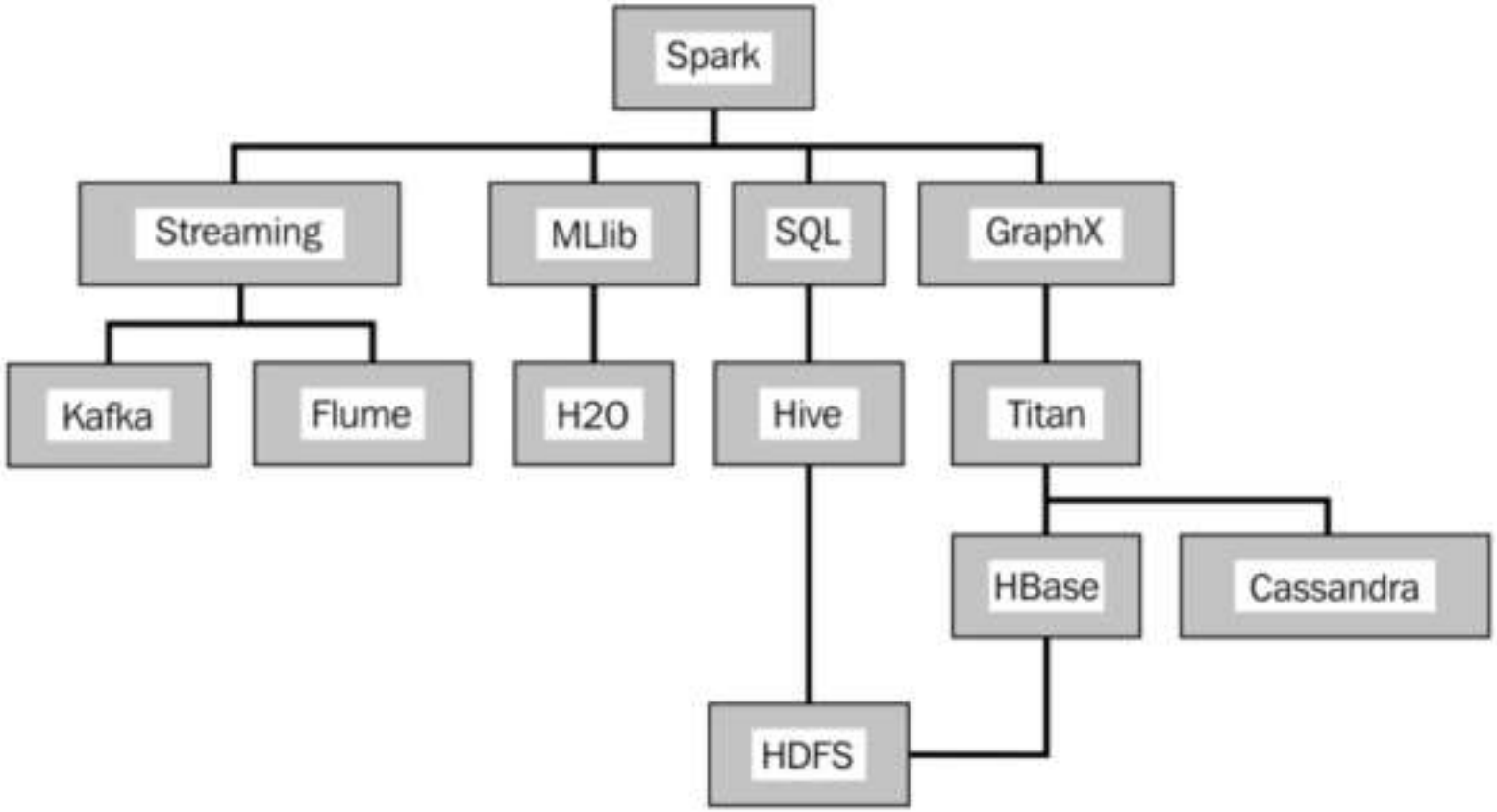


Apache Spark Framework

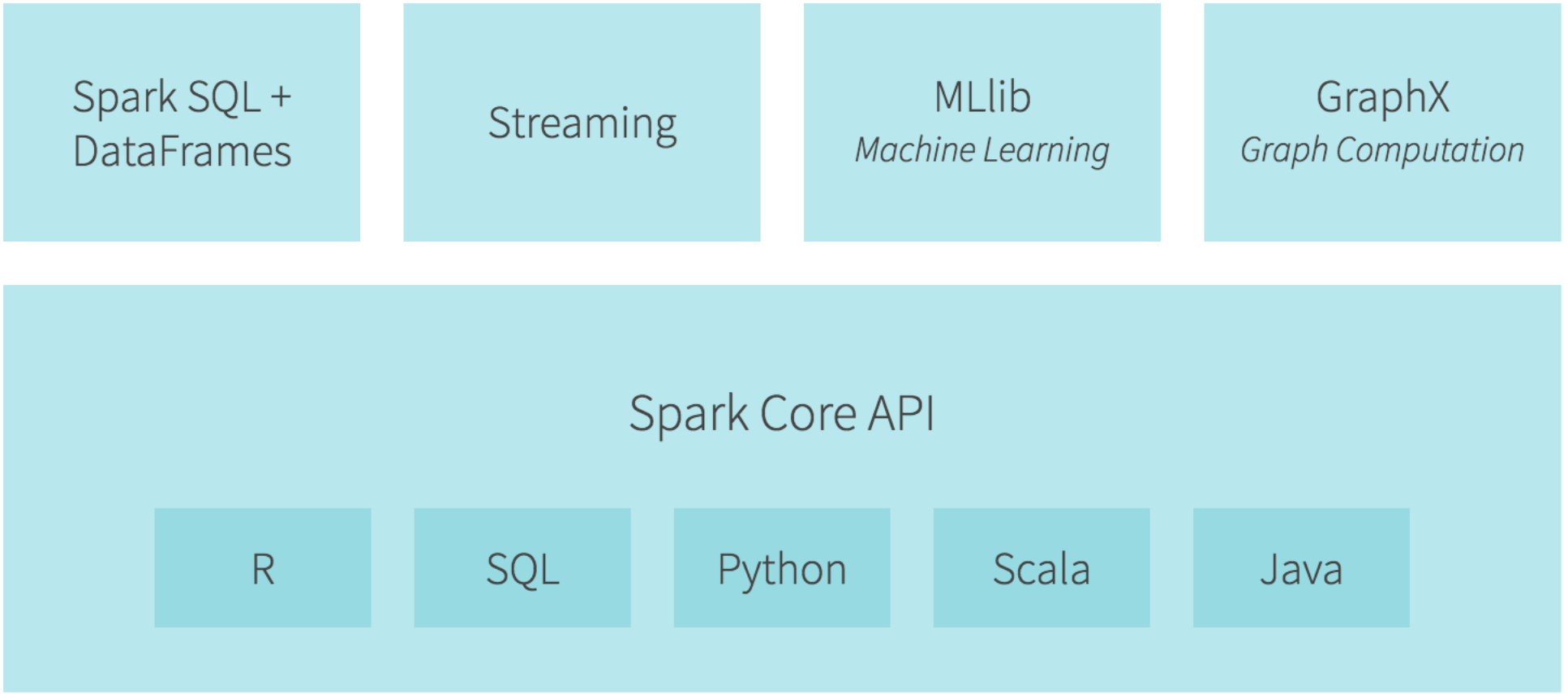
Apache Spark é um framework open-source para processamento de Big Data construído para ser veloz, fácil de usar e para análises sofisticadas.

Apache Spark Framework

Apache Spark Framework

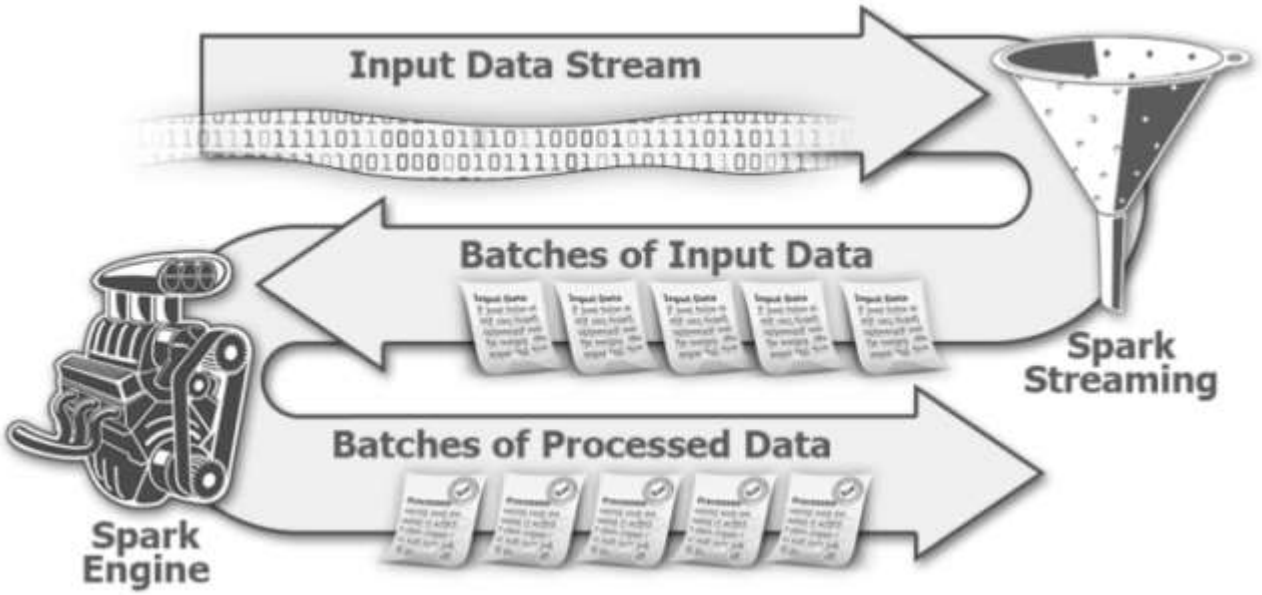


Apache Spark Framework

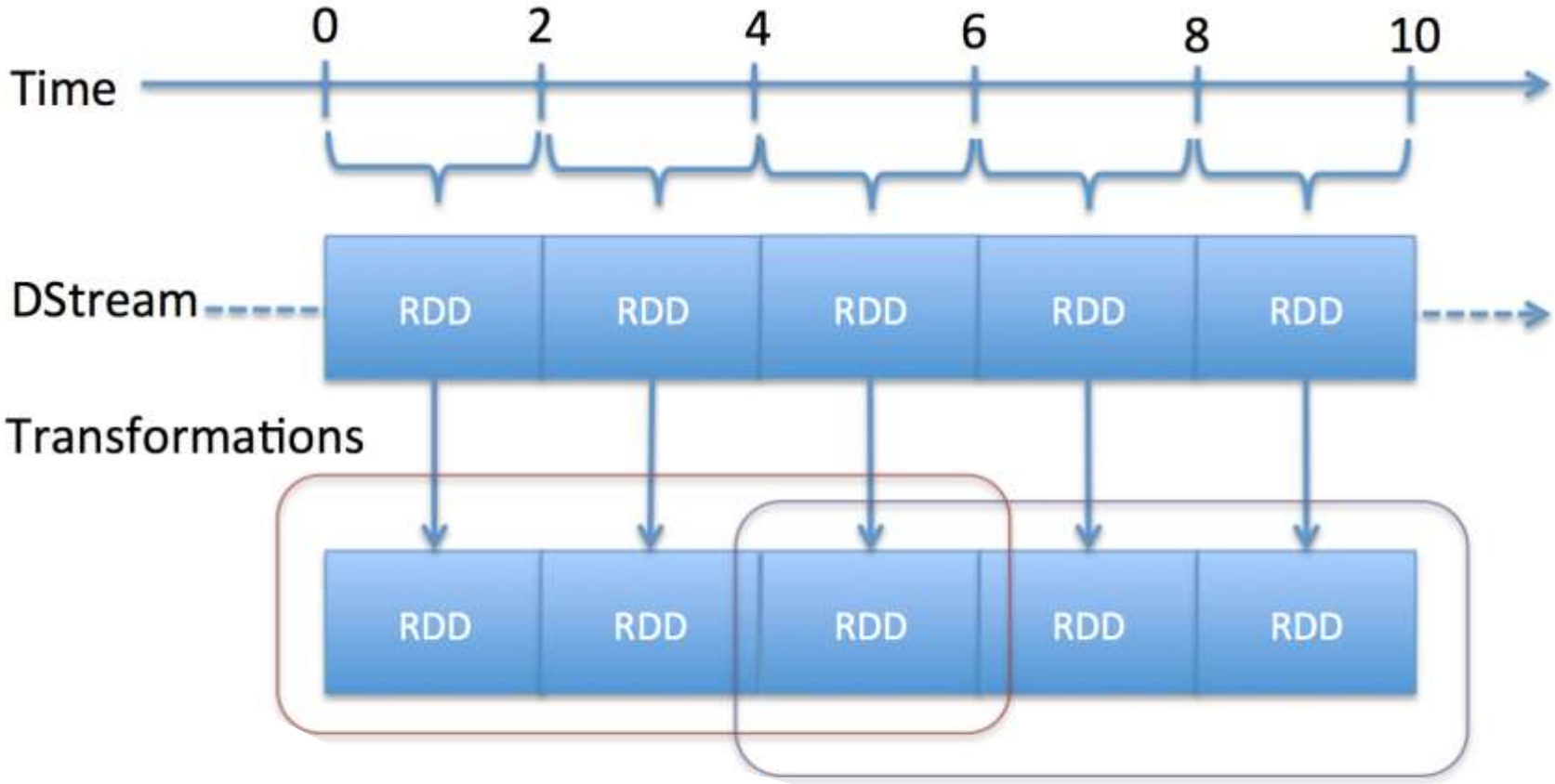


Apache Spark Framework

Spark Streaming



Apache Spark Framework



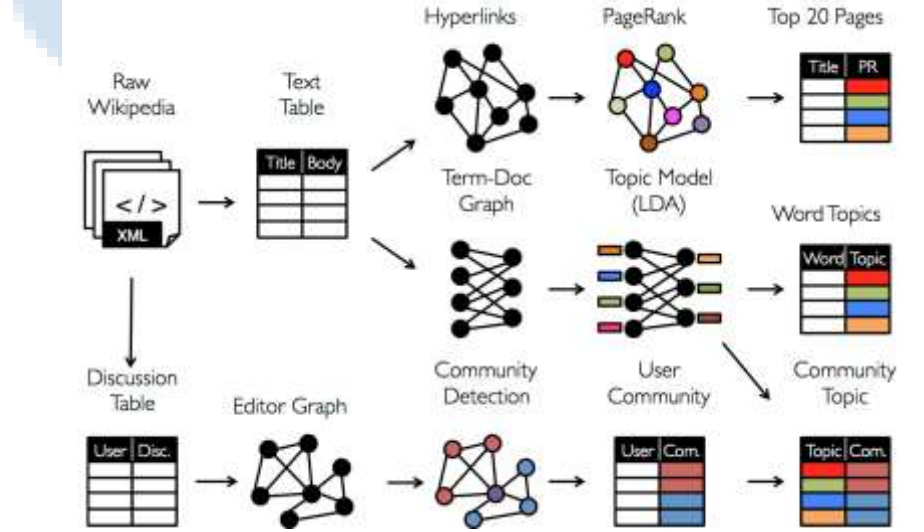
Apache Spark Framework

Apache Spark Framework

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

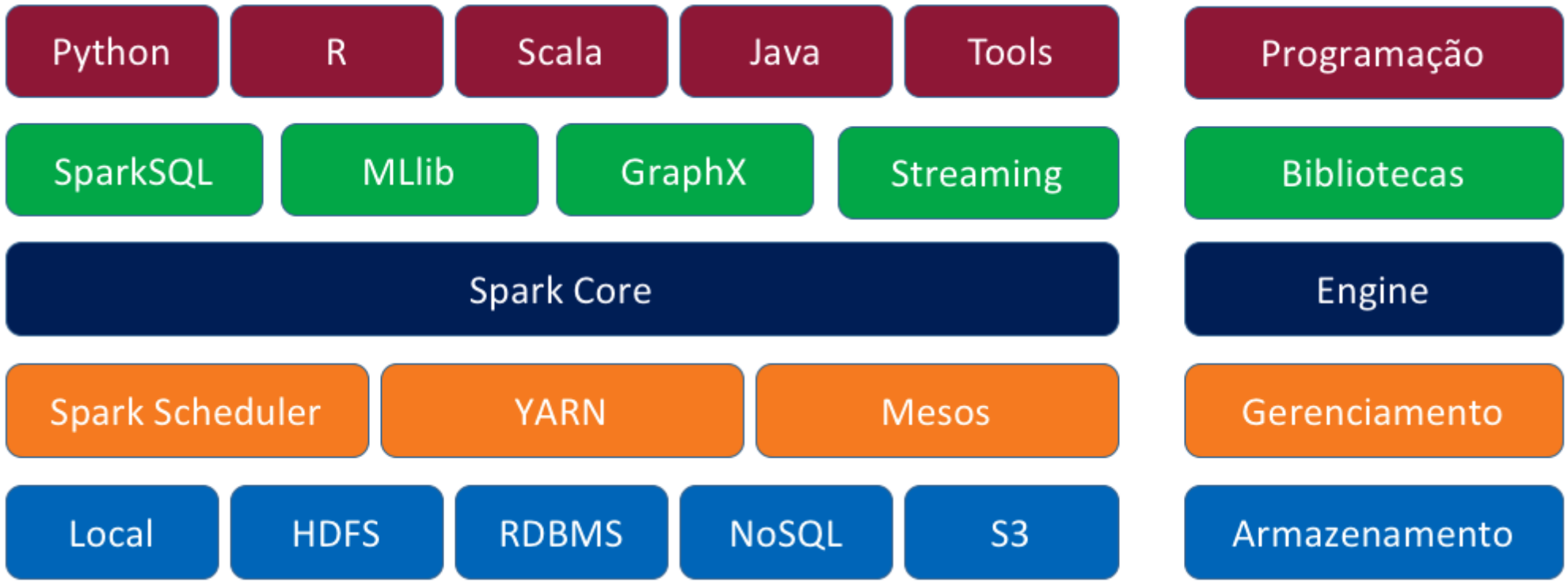
Apache Spark Framework

GraphX



Apache Spark Framework

Apache Spark Framework



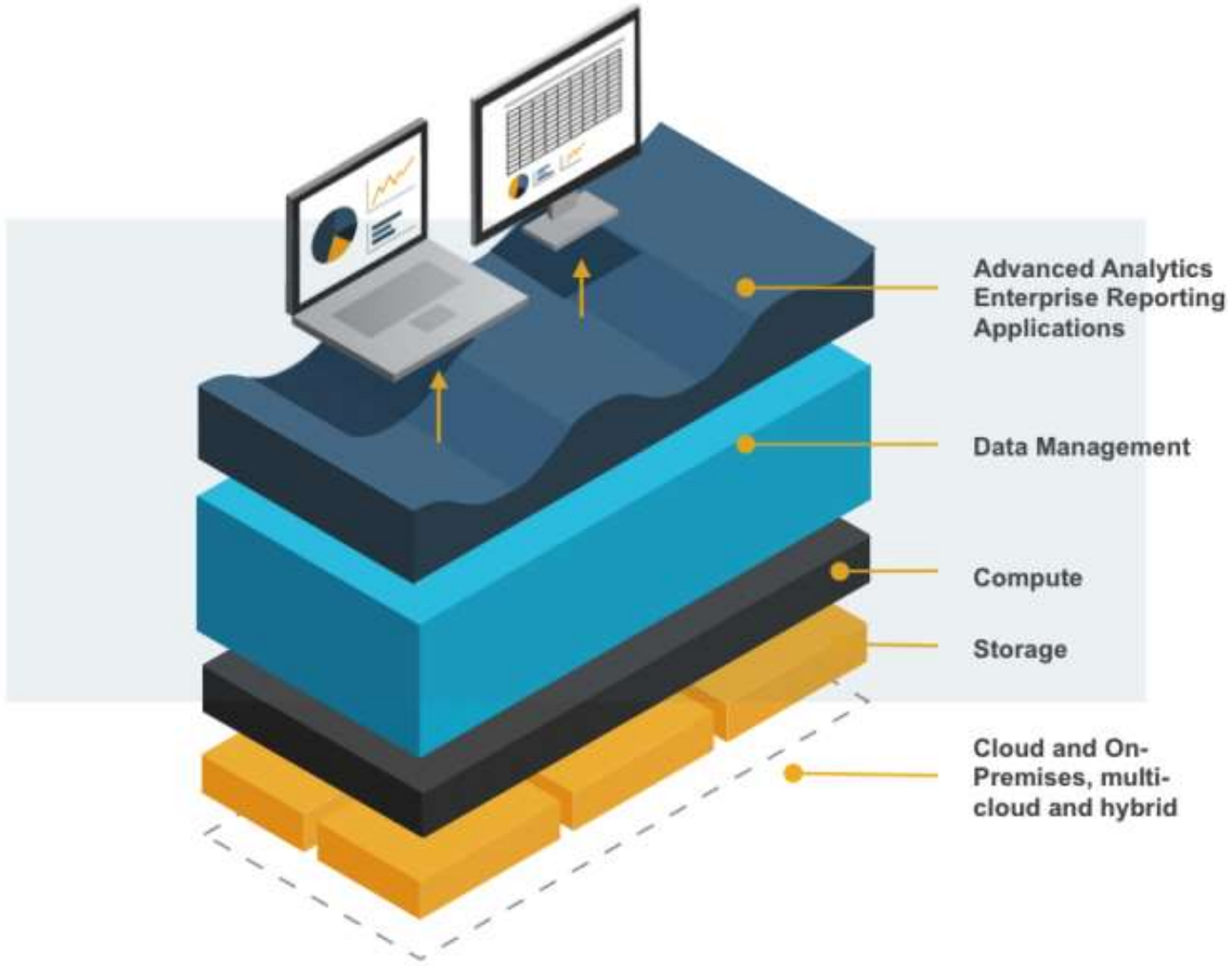
Apache Spark Framework

Quando usamos o Spark?

- Integração de Dados e ETL
- Análises Interativas
- Computação em Batch de Alta Performance
- Análises Avançadas de Machine Learning
- Processamento de Dados em Tempo Real

Definindo o Design do Data Lake

Definindo o Design do Data Lake



Não se aplica agora

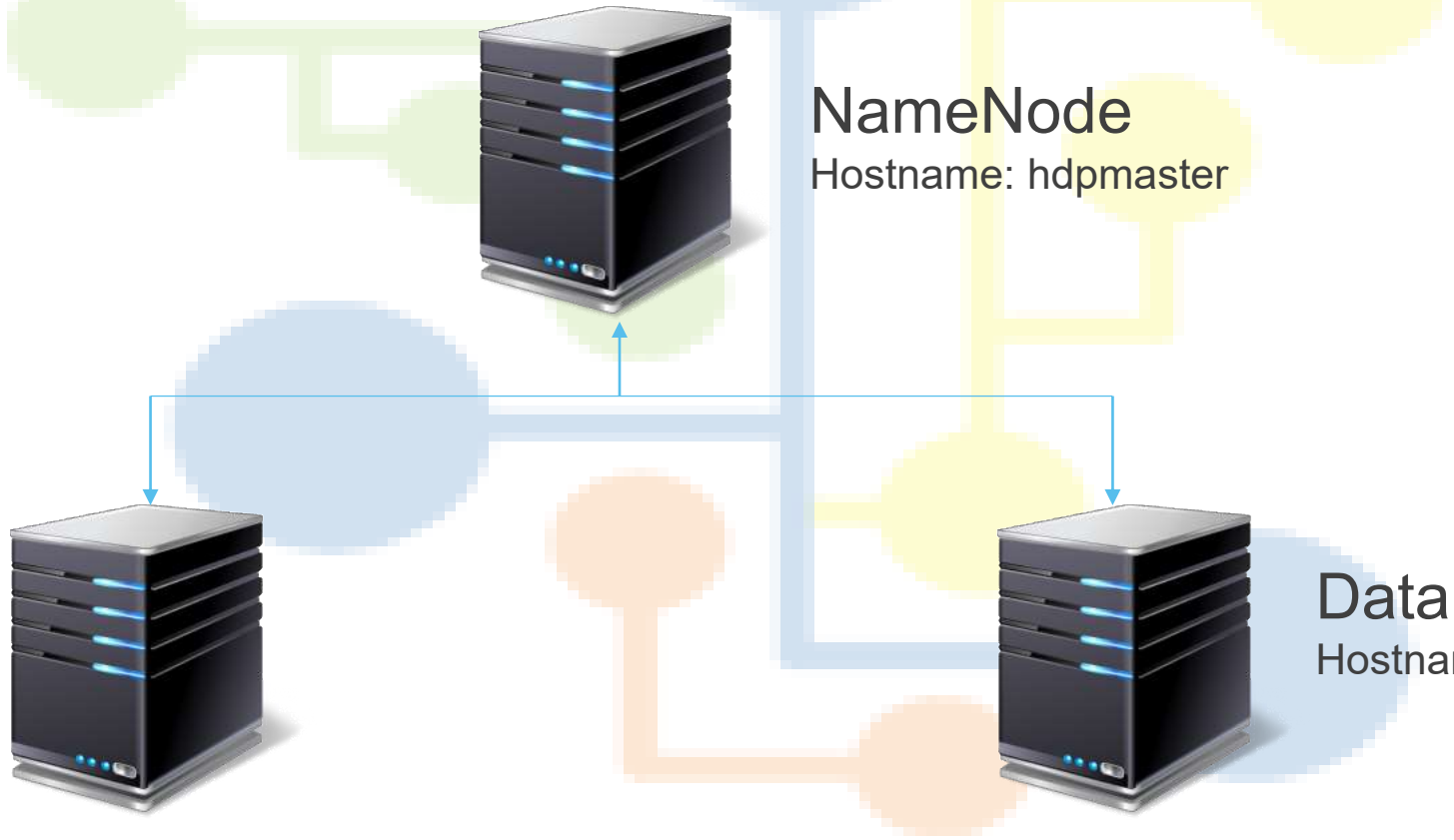
Não se aplica agora

Spark

HDFS

On-premises

Definindo o Design do Data Lake



NameNode
Hostname: hdpmaster

DataNode 2
Hostname: hdpslv2

DataNode 1
Hostname: hdpslv1

Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.