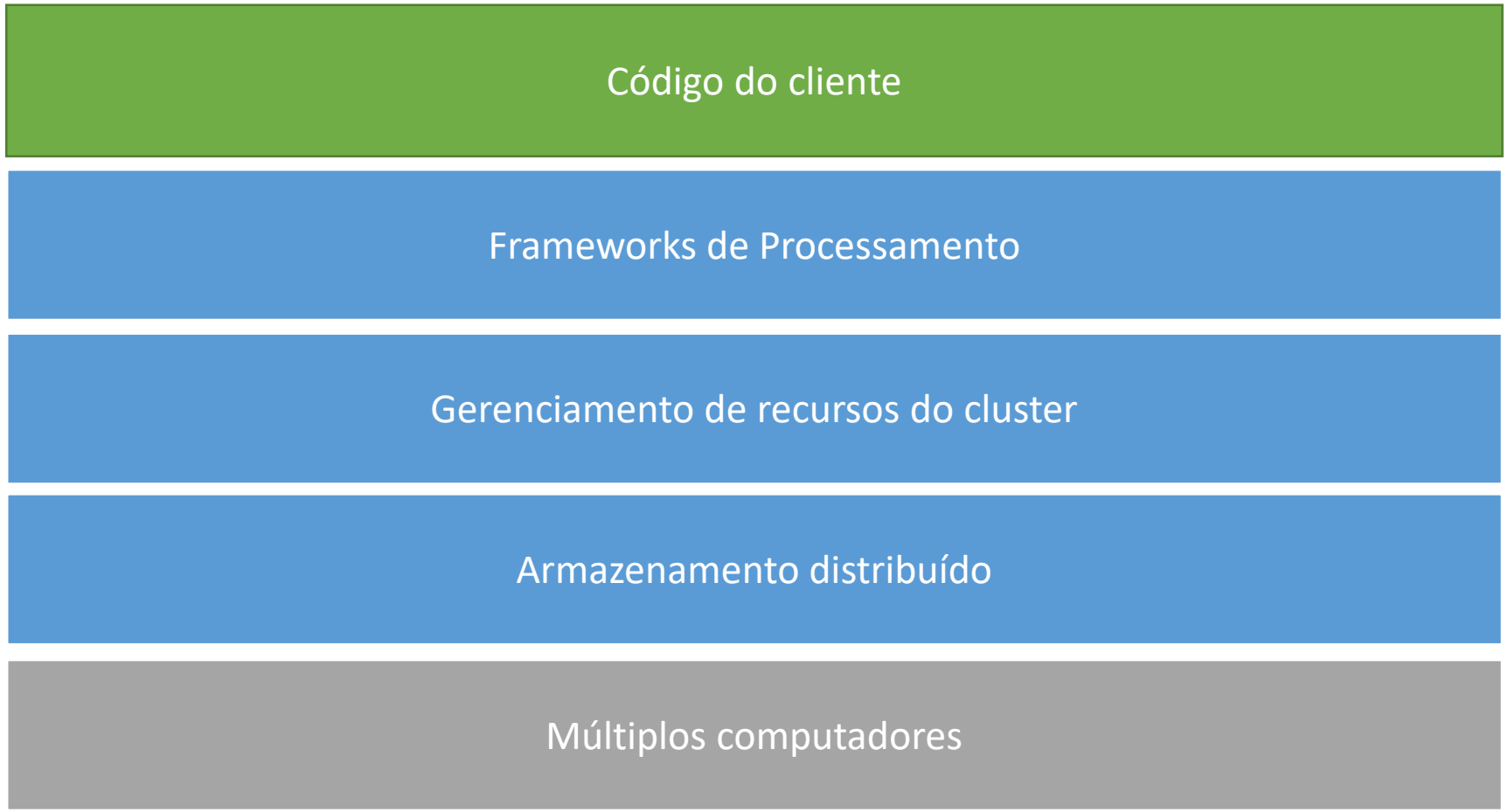
Brintell

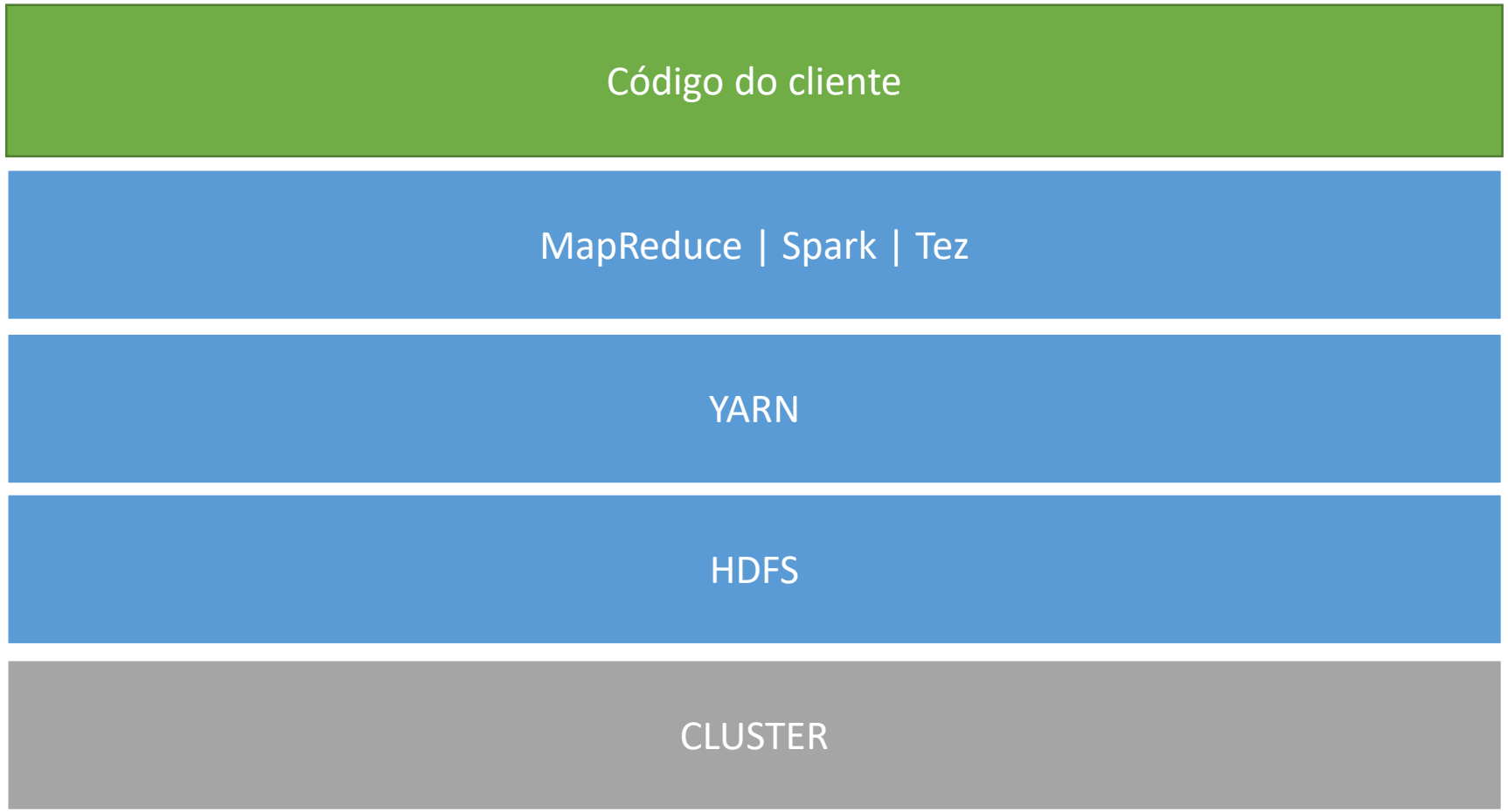
Intelligent Analytics

BIG DATA

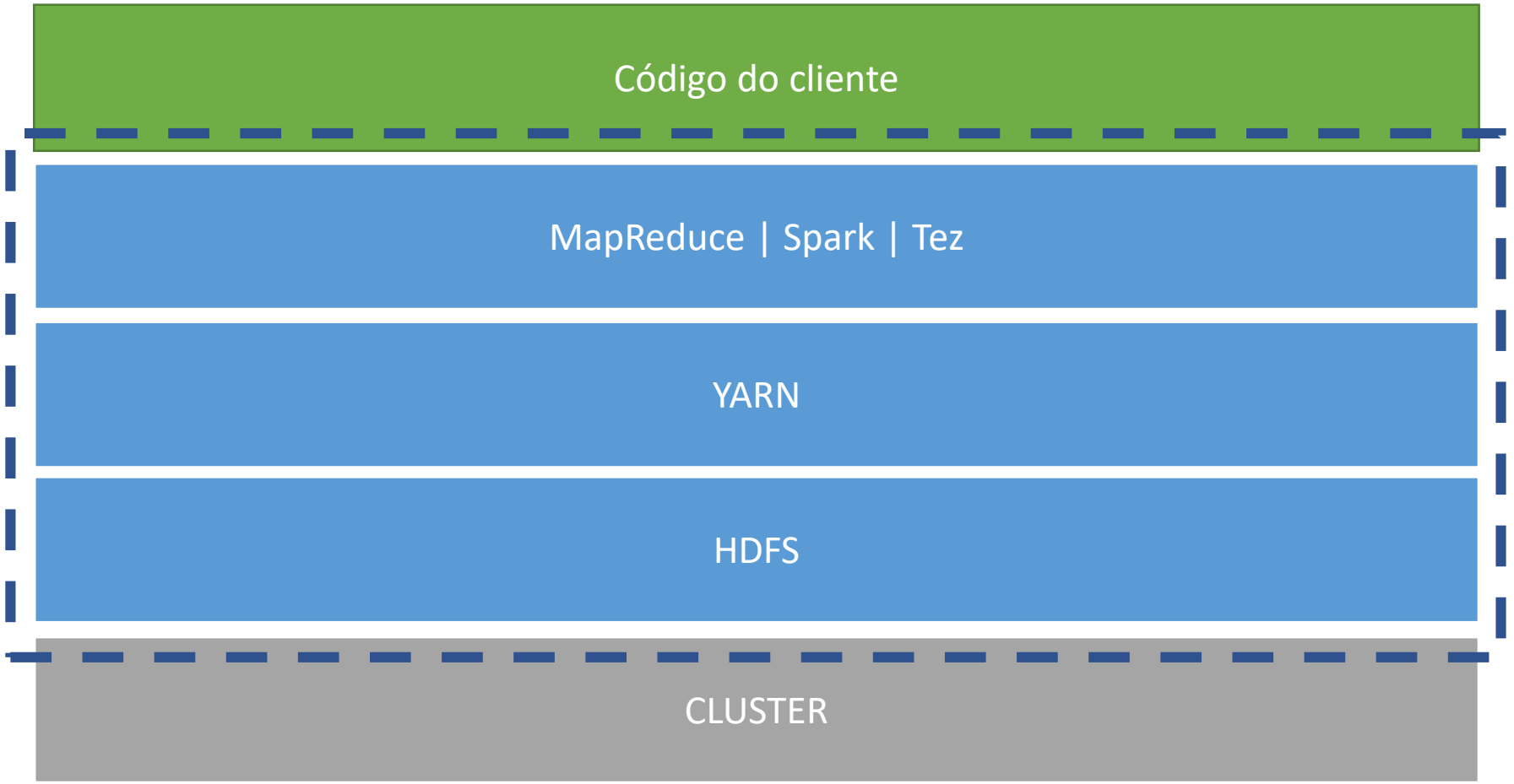
Core Hadoop

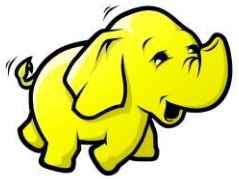

Código do cliente
Frameworks de Processamento
Gerenciamento de recursos do cluster
Armazenamento distribuído
Múltiplos computadores

Core Hadoop



Core Hadoop





Arquitetura Hadoop

HDFS - armazena os dados no cluster

MapReduce - Processa os dados no cluster

YARN - Coordena o trabalho e a divisão dos recursos do cluster

Core Hadoop

HDFS - Hadoop Distributed File System

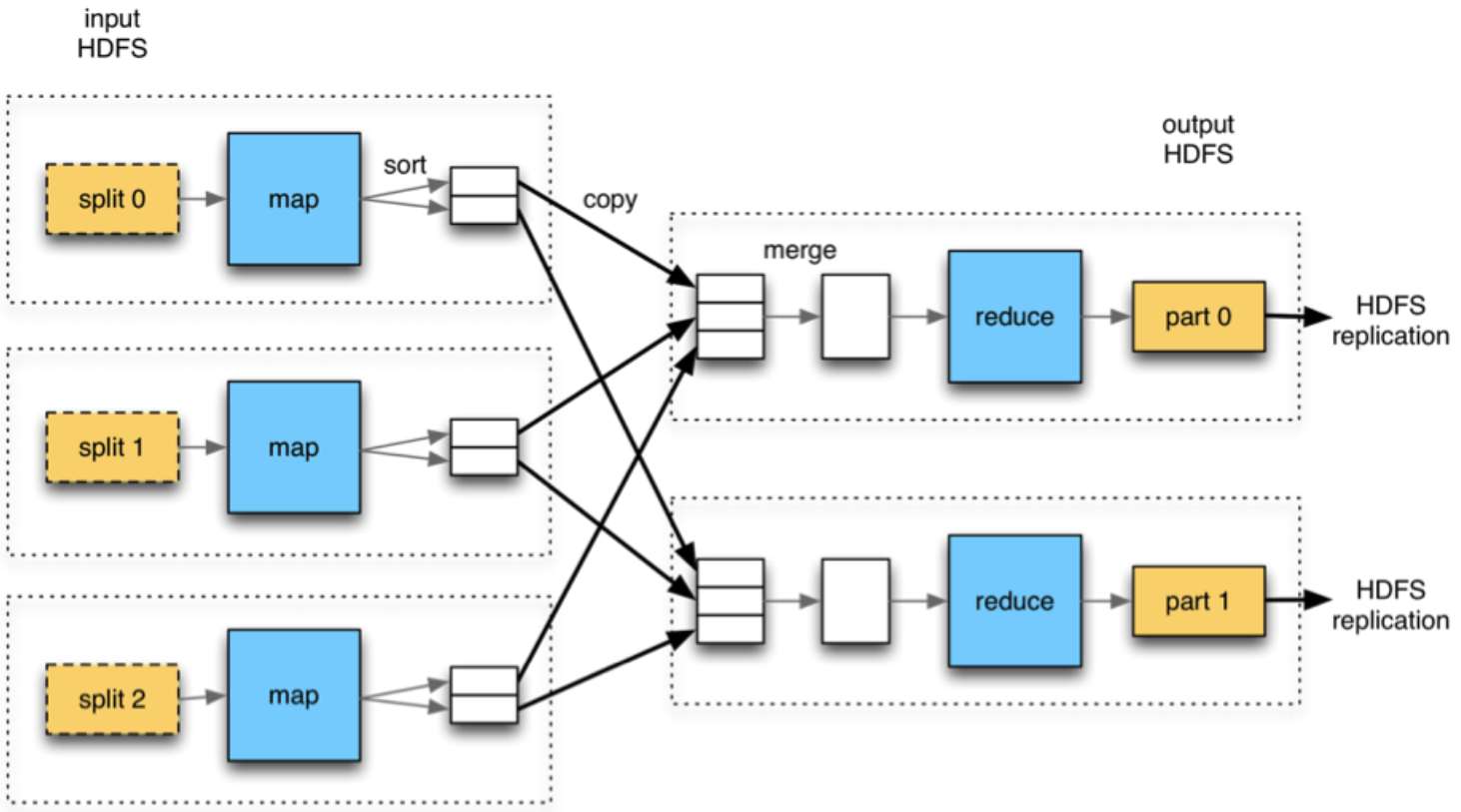
- Permite o armazenamento de qualquer tipo de dado
- Os dados são segmentados e replicados para alta disponibilidade

YARN (Yet Another Resource Nogotiator)

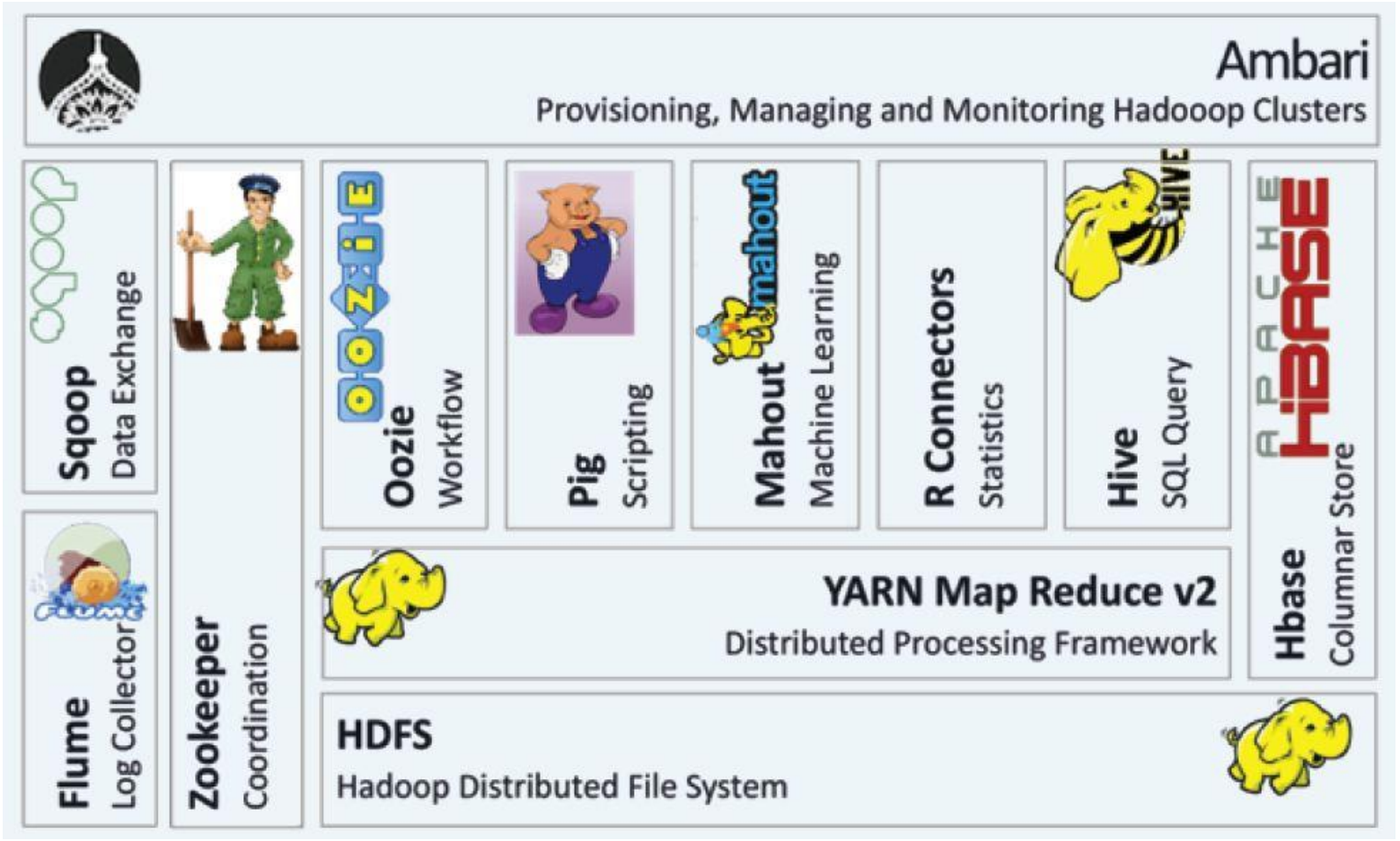
- Gerencia os recursos de processamento do cluster
- Agendador de tarefas
- Executa os frameworks de processamento

MapReduce

Framework de processamento distribuído



Ecossistema hadoop



Exemplos de projetos do ecossistema Hadoop

PIG – Editor de script com linguagem de alto nível

SQOOP – Realizar carga entre banco de dados relacionais

HIVE – SQL on Hadoop em lote

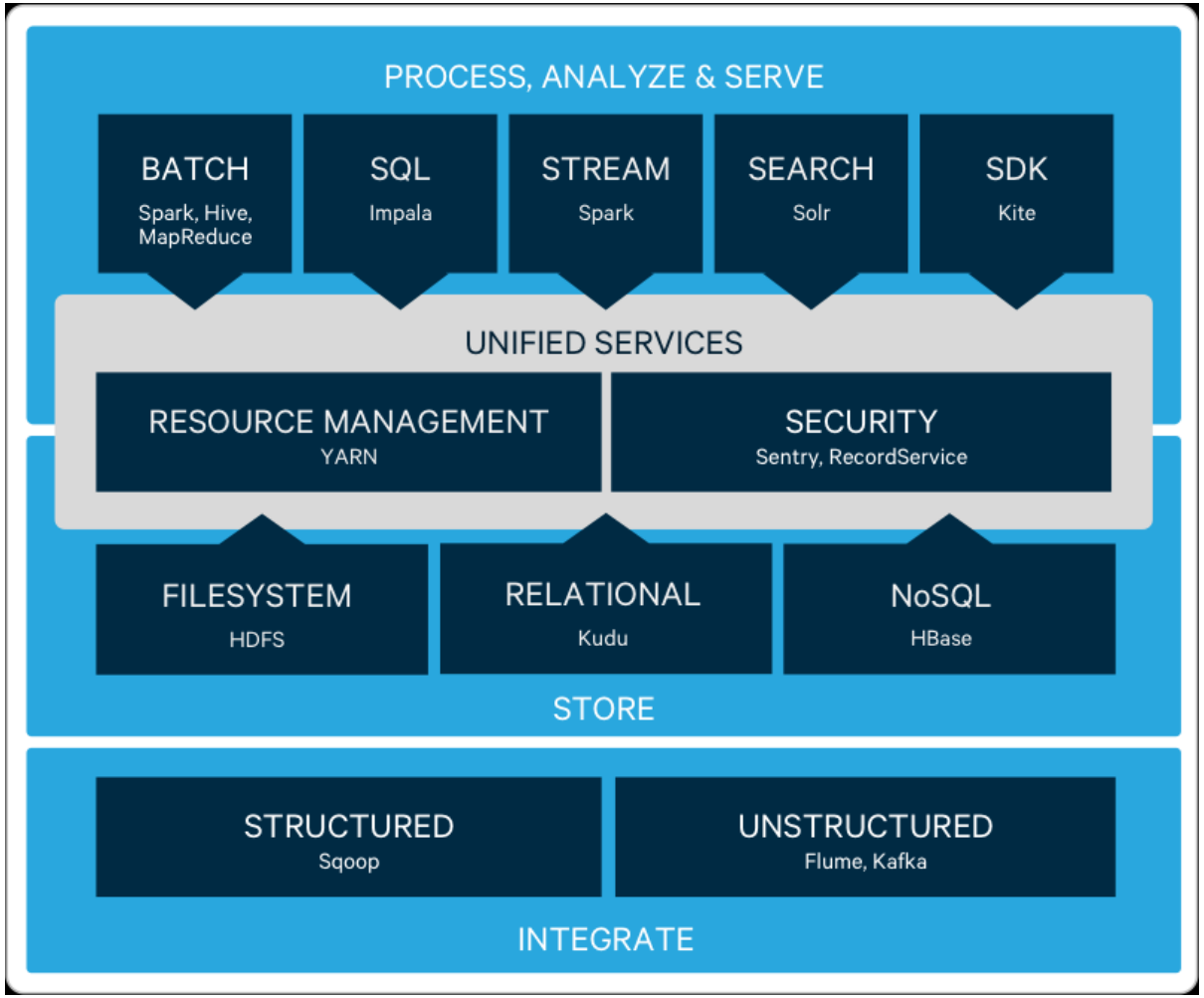
OOZIE – Coordenador de fluxo de trabalhos

IMPALA – SQL on Hadoop em tempo real

FLUME – Realizar carga de logs de servidores

...

Ecossistema hadoop



Ecossistema hadoop

Ferramentas desenvolvidas em torno do Core Hadoop

Deixar o Hadoop mais fácil de usar

Acrescenta funcionalidades

Open Source

Abrangente

Abordagem diferente para computação distribuída

Distribuir os dados quando são carregados no sistema

Executar a computação onde os dados estão armazenados

Abordagem diferente para computação distribuída

Os dados são armazenados em hardware padrão da indústria (não precisa de hardware de storage tradicional)

Adiciona capacidade com scaling out (mais maquinas), não scaling up, (maquinas maiores)

Permite que desenvolvedores foquem na aplicação, sem preocupar com restrições de computação distribuída

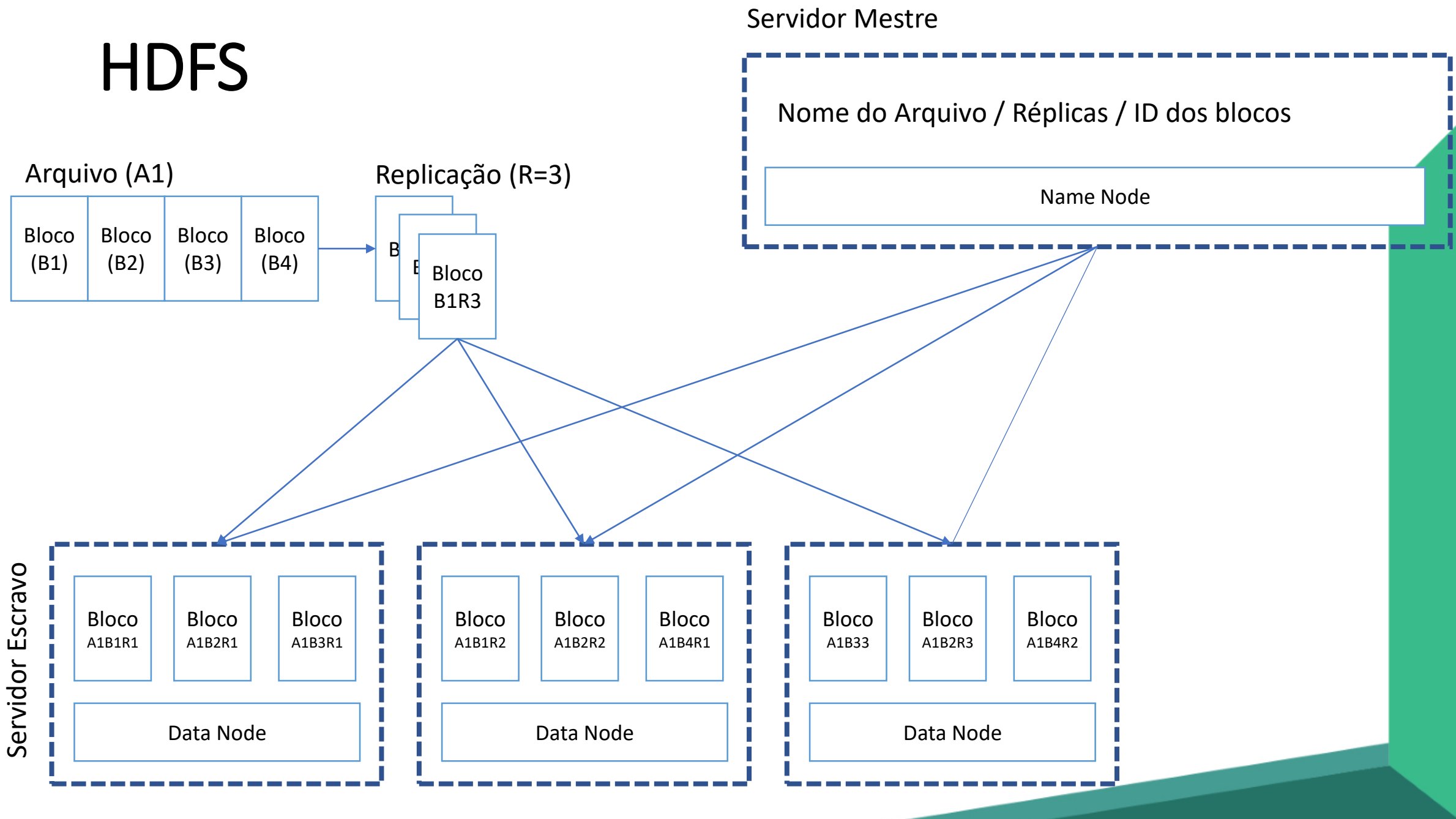
HDFS - Camada para armazenamento de dados do Hadoop

- File system para armazenar qualquer tipo de dado
- Dados replicados entre os computadores
- Performance melhor com um numero modesto de arquivos grandes milhões de arquivos, 100MB+
- Write once – append e sem acesso aleatório de escrita executado sob file system nativo do Linux
- Tolerante a falhas
- Suporta outros frameworks de processamento (Map Reduce, Spark, etc)

Como os arquivos são armazenados?

- Arquivo/blocos/replicas
- Metadata

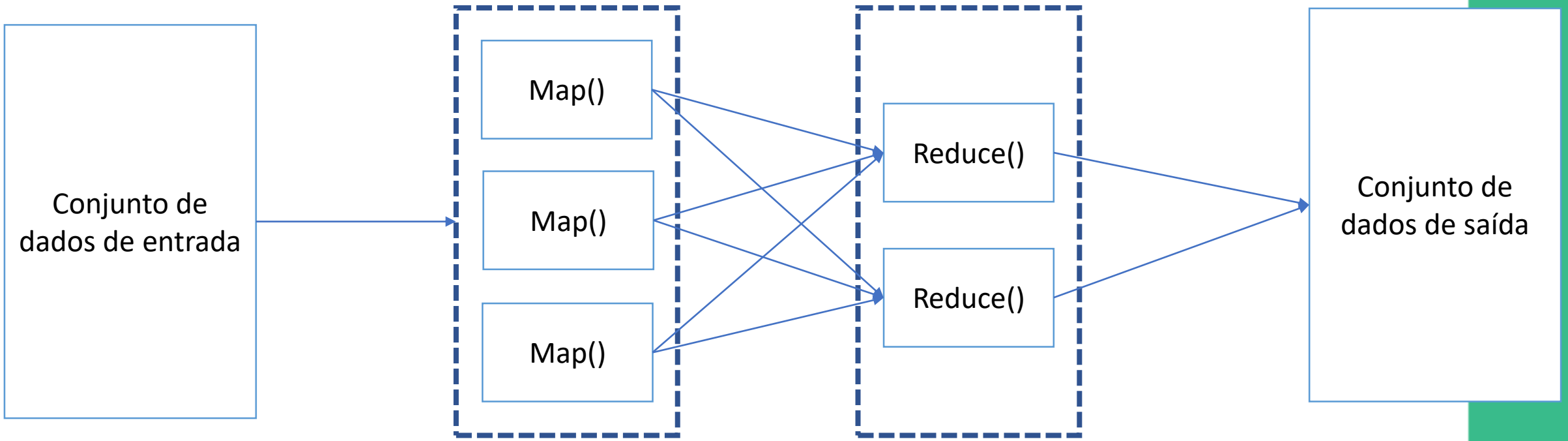
HDFS


Servidor Mestre
Nome do Arquivo / Réplicas / ID dos blocos
Name Node
Arquivo (A1)
Bloco (B1)
Bloco (B2)
Bloco (B3)
Bloco (B4)
Replicação (R=3)
Bloco B1R3
Servidor Escravo
Bloco A1B1R1
Bloco A1B2R1
Bloco A1B3R1
Data Node
Bloco A1B1R2
Bloco A1B2R2
Bloco A1B4R1
Data Node
Bloco A1B33
Bloco A1B2R3
Bloco A1B4R2
Data Node

Map Reduce

- Map Reduce é um modelo de programação
- Não é especifico de uma linguagem ou plataforma
- Orientado a registros (chave e valor)
- Facilita a distribuição de tarefas ao longo de múltiplos nós
- Foi o framework de processamento original do Hadoop
- Ainda é utilizado, porem já existem outros frameworks substitutos para outras cargas de trabalho (Tez, Spark)
- Desenvolvido em Java

Map Reduce

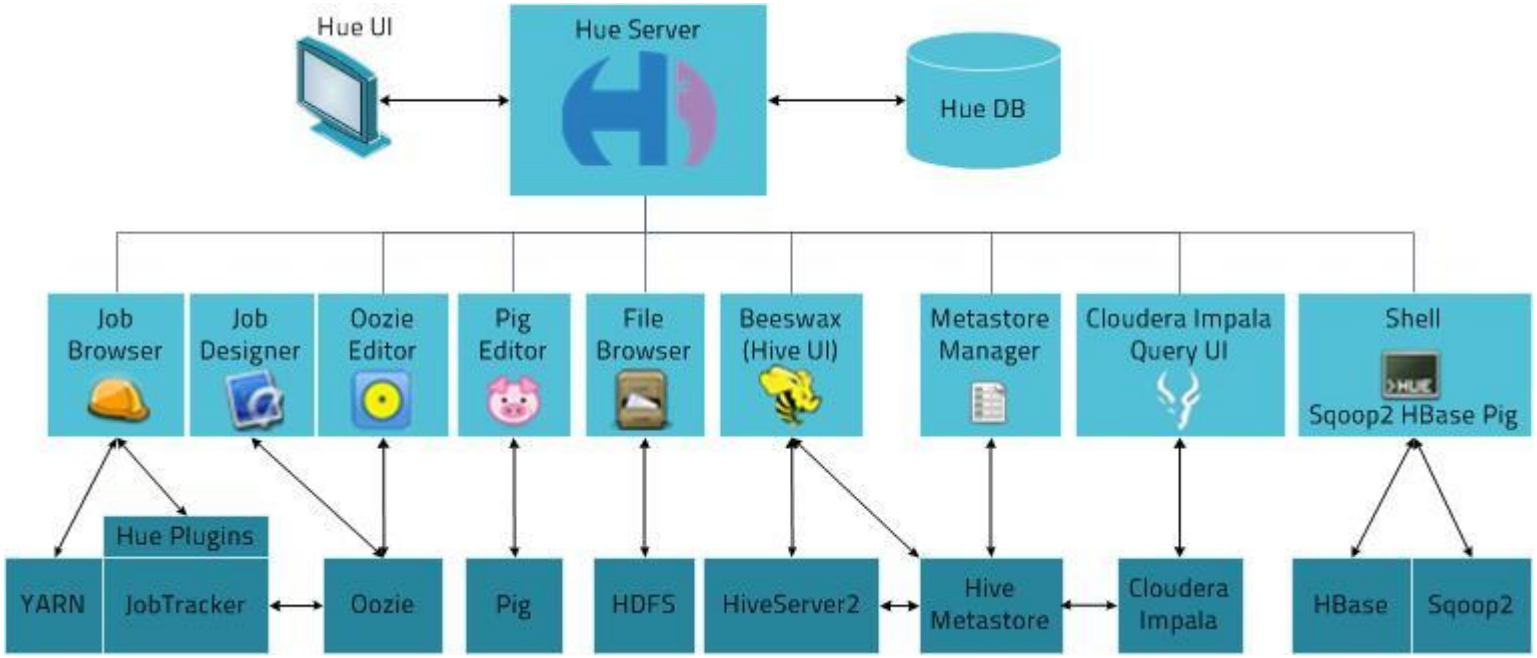

Conjunto de dados de entrada
Map()
Map()
Map()
Reduce()
Reduce()
Conjunto de dados de saída

YARN

Originalmente o Hadoop suportava apenas trabalhos em Map Reduce e não otimizava os recursos do cluster

Com o YARN o cluster pode executar diversos frameworks de processamento e pode alocar os recursos baseado na demanda

Interface com o usuário - Hadoop



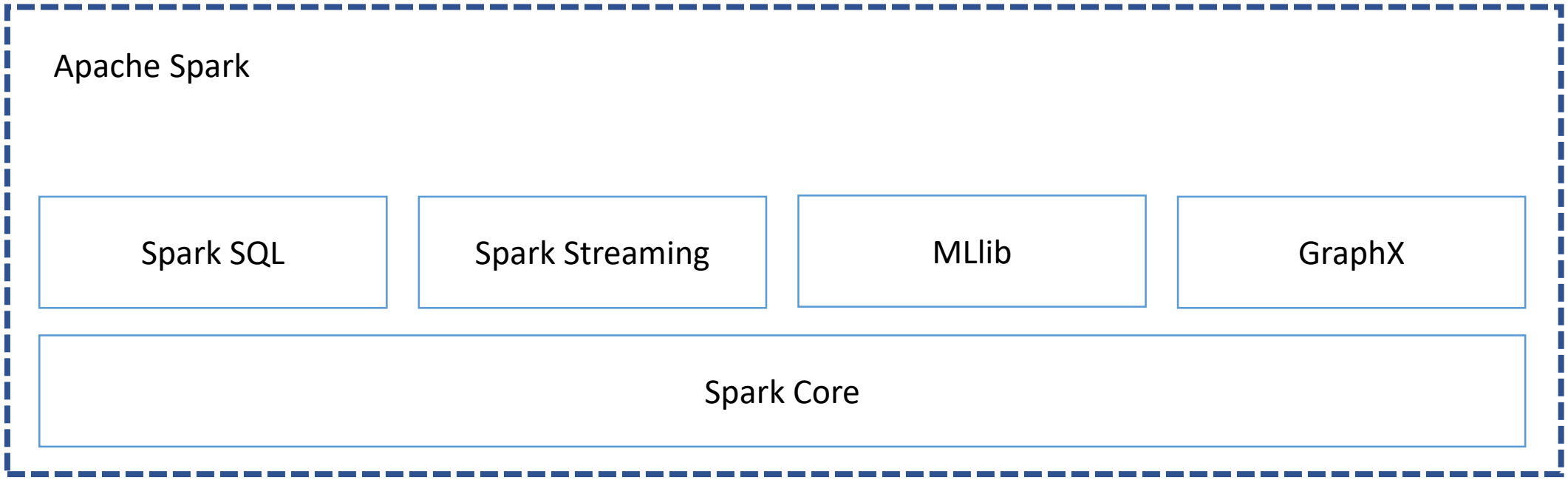
Ingestão e recuperação de dados

- Hadoop (API,CLI)
- Flume
- Sqoop
- BI/BA

Spark

- ✓O Spark é um framework para processamento de Big Data construído com foco em velocidade para a realização de análises complexas
- ✓Permite que aplicações em clusters Hadoop executem até 100 vezes mais rápidas em memória
- ✓Permite a execução de Map/Reduce, suporta consultas SQL, streaming de dados, aprendizado de máquina e processamento de grafos

Spark


Apache Spark
Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core

Banco de Dados NoSQL

- ✓Termo aplicado para classificar bancos de dados não relacionais de alto desempenho.

- ✓Modelos de dados alternativos: Orientados a documentos, chave-valor, colunares e etc.

Hbase

- ✓Banco NoSQL distribuído
- ✓Orientado a coluna
- ✓Inspirado pelo Google BigTable

Sqoop

- ✓Executa a transferência bidirecional de dados entre o Hadoop e diversos serviços de armazenamento externo de dados estruturados.

- ✓Utiliza um driver JDBC (Java Database Connectivity) para se conectar aos bancos