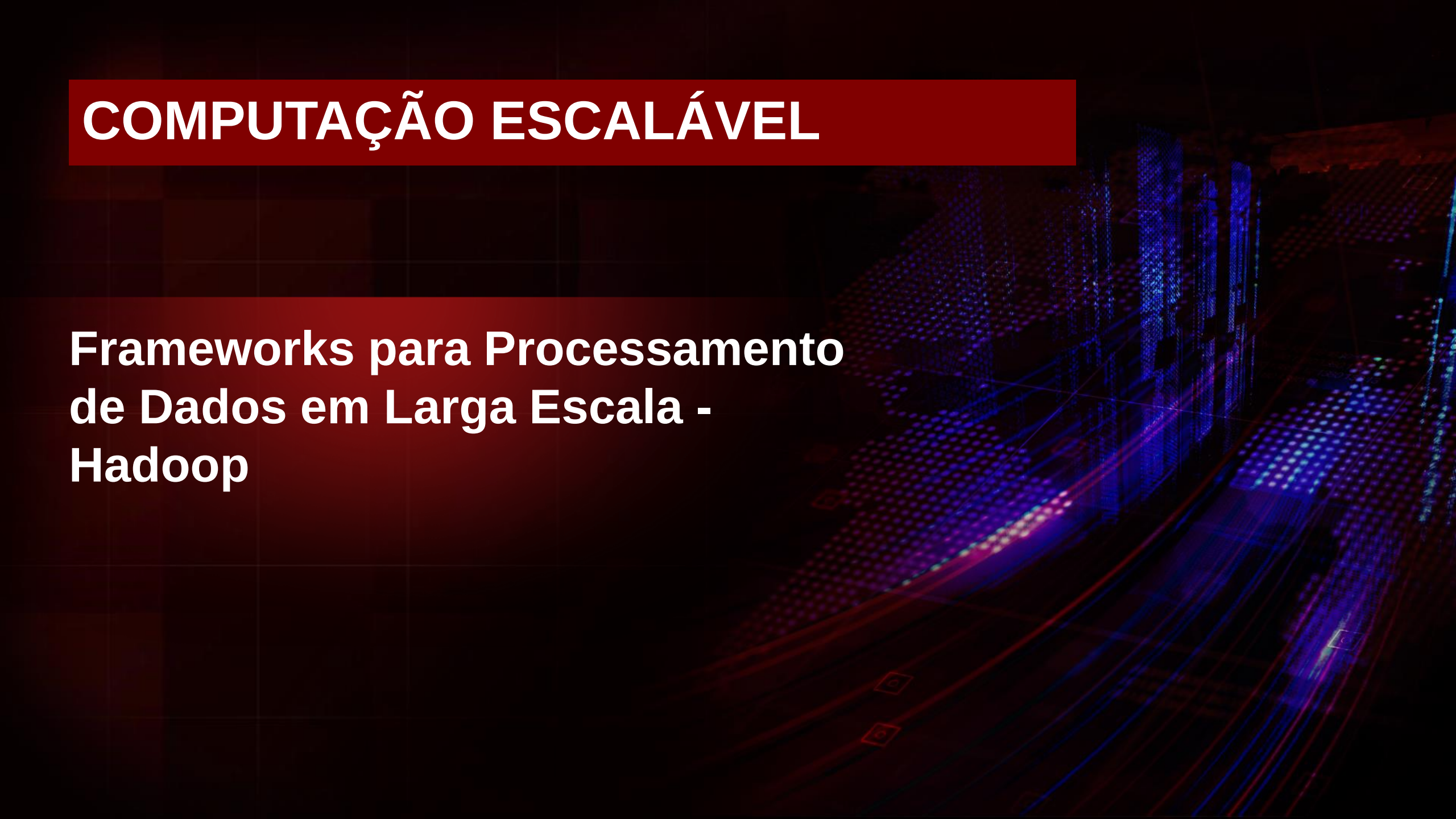
COMPUTAÇÃO ESCALÁVEL

Frameworks para Processamento de Dados em Larga Escala - Hadoop



ROTEIRO

- Características
- O que é Apache Hadoop?
- Arquitetura do Hadoop
- Ecossistema Hadoop
- Aplicações

CARACTERÍSTICAS

- Linguagem Java
- Modelo de programação Map Reduce
- Arquitetura Mestre/Escravo
- Processamento Distribuído
- Hadoop Distributed Filesystem (HDFS)
- Tolerante a Falhas
- Memória Secundária (Disco)
- Escalável

CARACTERÍSTICAS

- Não é uma linguagem de programação
- Não resolve tudo sobre processamento de grandes volumes de dados
- Não é processamento em tempo real
- Não é uma lib de aprendizado de máquina

O QUE É APACHE HADOOP?

É uma implementação de código aberto de estruturas para armazenamento de dados e computação confiável, escalável e distribuído

É uma arquitetura flexível e altamente disponível para computação em larga escala e processamento de dados em uma rede de hardware comum

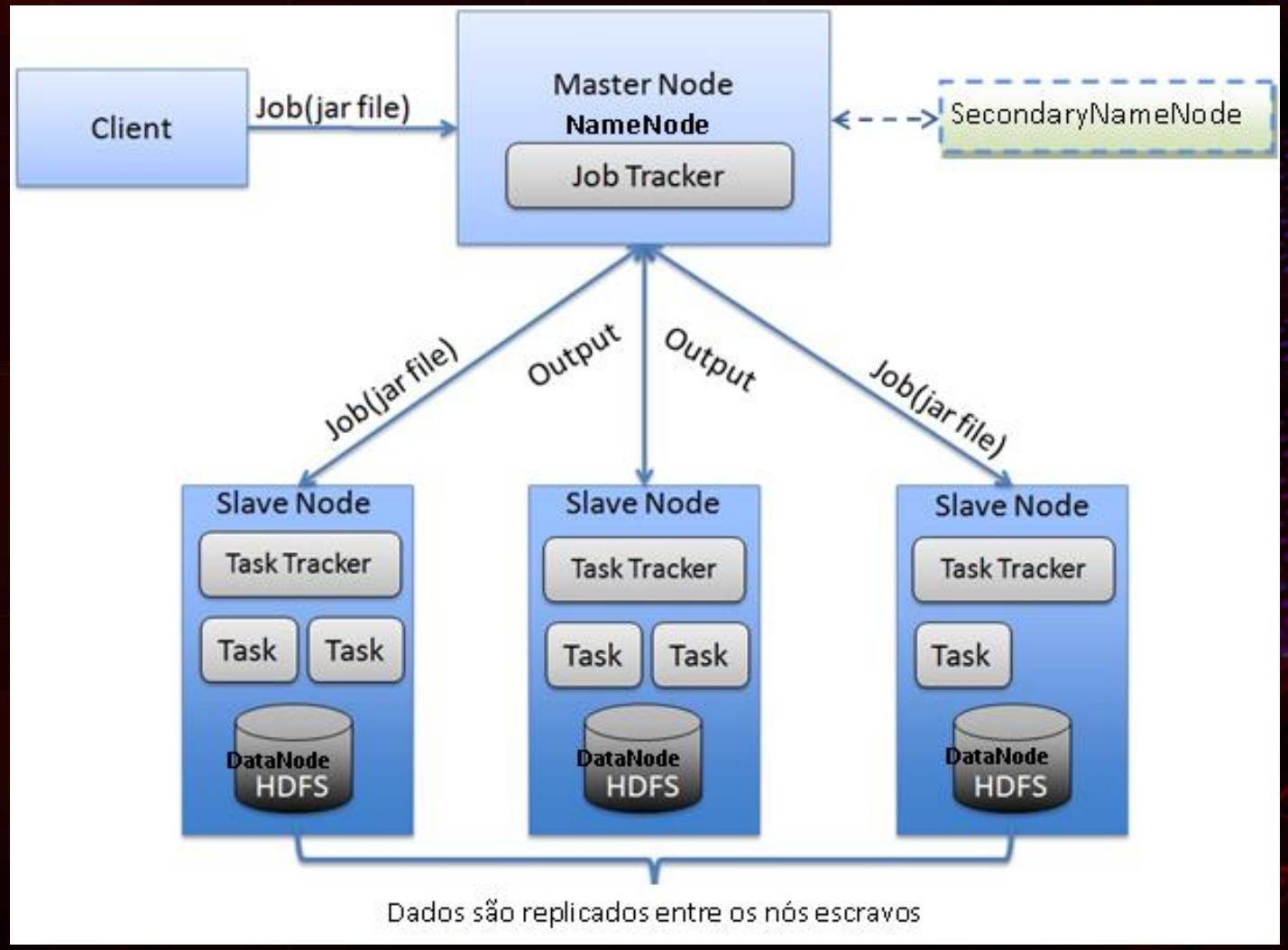
O QUE É APACHE HADOOP?

Projetado para responder à pergunta: “Como processar big data com custo e tempo razoáveis?”

Objetivos/Requisitos

- Abstrair e facilitar o armazenamento e o processamento de conjuntos de dados grandes e/ou em rápido crescimento
- Dados estruturados e não estruturados
- Modelos de programação simples
- Alta escalabilidade e disponibilidade
- Use hardware comum (barato!) com pouca redundância
- Tolerância ao erro
- Mover computação em vez de dados

ARQUITETURA APACHE HADOOP



ARQUITETURA APACHE HADOOP

Distribuída com alguma centralização

- A maior parte do poder computacional e de armazenamento do sistema está nos nós principais do cluster
- Os nós principais executam o TaskTracker para aceitar e responder às tarefas do MapReduce e também o DataNode para armazenar os blocos necessários o mais próximo possível
- O nó de controle central executa o NameNode para acompanhar os diretórios e arquivos do HDFS e o JobTracker para despachar tarefas de computação para o TaskTracker
- Escrito em Java, também suporta Python e Ruby

ARQUITETURA APACHE HADOOP

Sistema de Arquivos Distribuídos Hadoop (HDFS)

- Adaptado às necessidades do MapReduce
- Direcionado para muitas leituras de fluxos de arquivos
- Gravações são mais caras
- Alto grau de replicação de dados (3x por padrão)
- Não há necessidade de RAID em nós normais
- Tamanho de bloco grande (64 MB)
- Reconhecimento de localização de DataNodes na rede

ARQUITETURA APACHE HADOOP

NameNode

- Armazena metadados para os arquivos, como a estrutura de diretórios de um FS típico
- O servidor que contém a instância NameNode é bastante crucial, pois existe apenas um
- Log de transações para exclusões/inclusões de arquivos, etc. Não usa transações para blocos inteiros ou fluxos de arquivos, apenas metadados.
- Lida com a criação de mais blocos de réplica quando necessário após uma falha do DataNode

ARQUITETURA APACHE HADOOP

DataNode

- Armazena os dados reais em HDFS
- Pode ser executado em qualquer sistema de arquivos subjacente (ext3/4, NTFS, etc.)
- Notifica NameNode de quais blocos ele possui
- NameNode replica blocos 2x no rack local, 1x em outro lugar

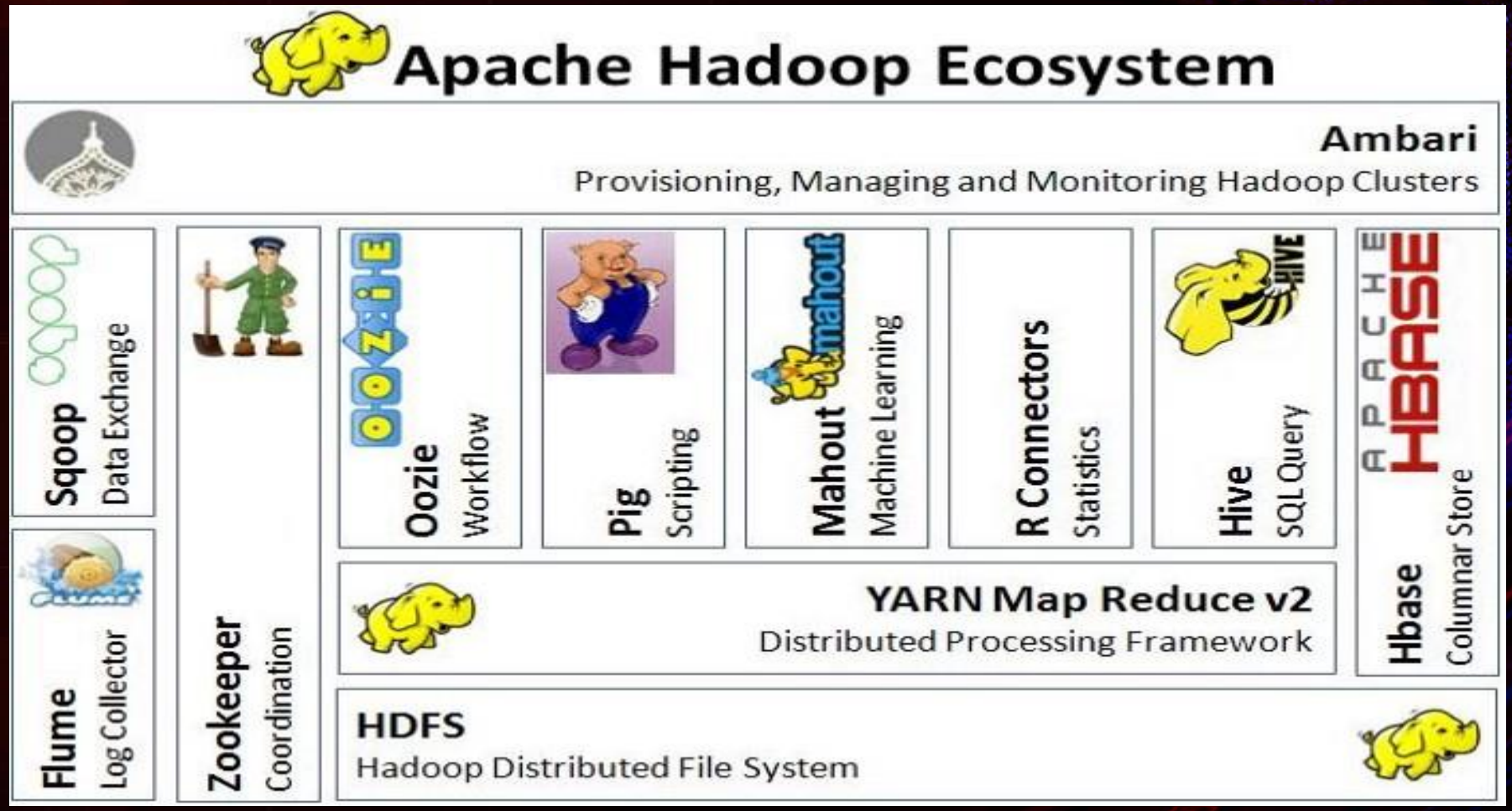
ARQUITETURA APACHE HADOOP

MapReduce Engine

- O JobTracker divide os dados em tarefas menores (“Mapa”) e os envia para o processo TaskTracker em cada nó

- O TaskTracker reporta de volta ao nó JobTracker e relata o progresso do trabalho, envia dados (“Reduzir”) ou solicita novos trabalhos

ECOSSISTEMA HADOOP



APLICAÇÕES

Publicidade (comportamento do usuário de mineração para gerar recomendações)

Pesquisas (documentos relacionados ao grupo)

Segurança (busca por padrões incomuns)

Soluções Comerciais

- Dell
- Amazon
- Microsoft
- Cloudera

BIBLIOGRAFIA

1. https://www.ime.usp.br/~ipolato/JAI2012-Hadoop-Slides.pdf
2. http://www.each.usp.br/dc/papers/erad-hadoop-DanielCordeiro.pdf
3. https://www.devmedia.com.br/hadoop-fundamentos-e-instalacao/29466
4. https://johnosd.medium.com/hadoop-seus-componentes-principais-e-sua-evolu%C3%A7%C3%A3o-cf125c99fadd
5. https://cetax.com.br/apache-hadoop-tudo-o-que-voce-precisa-saber/
6. https://itforum.com.br/noticias/28-bilhoes-de-gigabytes-de-storage-foram-distribuidos-no-primeiro-trimestre/
7. https://www.ime.usp.br/~ipolato/JAI2012-Hadoop-Slides.pdf