JUSTIÇA 4.0

PNUD

CNJ CONSELHO NACIONAL DE JUSTIÇA

CURSO DE CIÊNCIA DE DADOS APLICADA AO PODER JUDICIÁRIO

SPARK PARA CIÊNCIA DE DADOS

Semana 1 - Apache Hadoop

PROF. CARLOS M. D. VIEGAS

Semana 1 - Apache Hadoop

- Conteúdo

1. Introdução ao Ecossistema Hadoop
2. Sistema de arquivos HDFS
3. Modelo de programação MapReduce
4. Gerenciamento de recursos com Yarn
5. Instalação e configuração do Hadoop
6. Análise de logs para diagnóstico e resolução de problemas

Introdução ao Ecossistema Hadoop

- O que é o Apache Hadoop?
  - Framework de código aberto criado em 2005 por Doug Cutting e Mike Carafella
    - Desenvolvido em linguagem Java
  - Projetado para armazenar e processar grandes volumes de dados em larga escala
    - Solução escalável, distribuída e de baixo custo
  - Aplicações comuns do Hadoop:
    - Processamento de texto em larga escala
    - Aprendizado de máquina e mineração de dados
    - Análise de dados em larga escala

  - Mas a computação paralela não seria suficiente?
    - Complexidade
    - Divisão e escalonamento de tarefas
    - Balanceamento de carga
    - Sincronismo entre tarefas

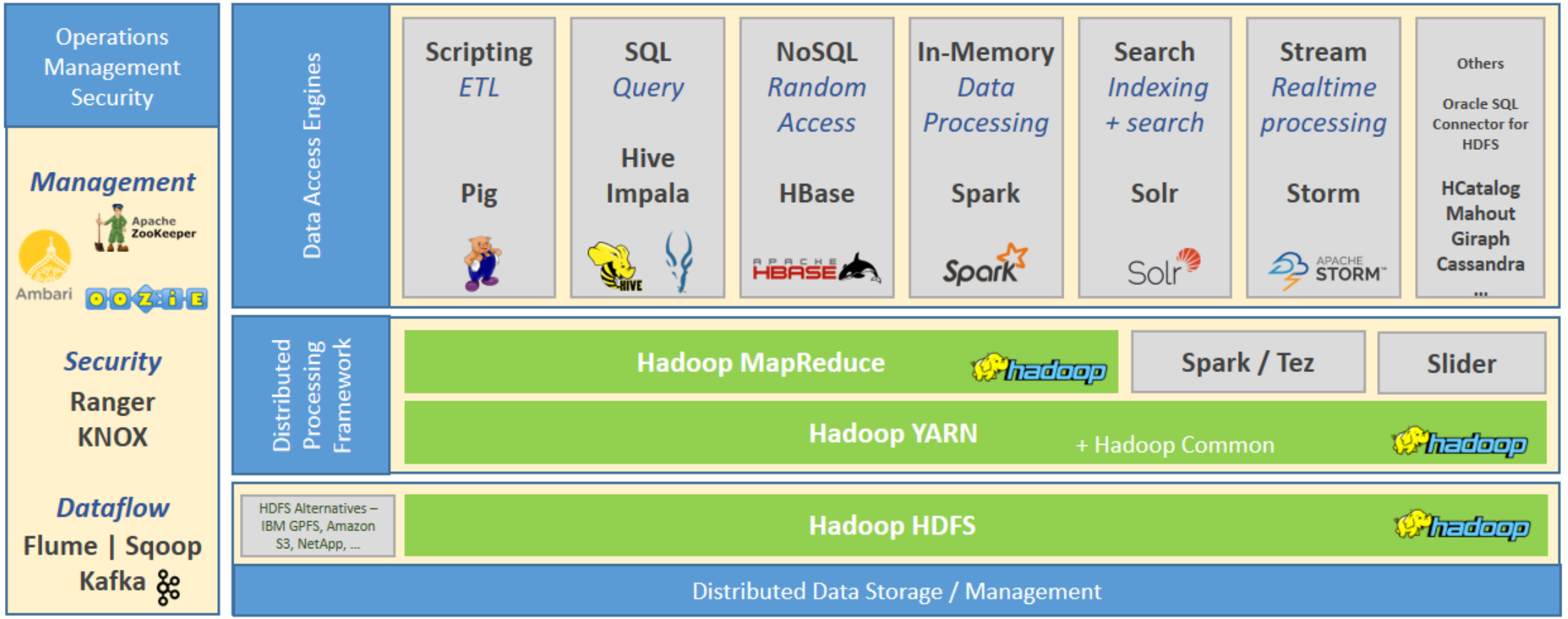
Introdução ao Ecossistema Hadoop

- Fatores para o sucesso do Hadoop:
  - Gratuito e de código aberto
  - Permite utilizar computadores de baixo custo
  - Não exige alterações na infraestrutura da rede (rede comum)
  - Tolerância a falhas
  - Facilidade de uso
  - Escalável

- O Apache Hadoop é amplamente utilizado em uma variedade de setores, incluindo finanças, comércio, saúde, transporte, e pesquisa científica

Introdução ao Ecossistema Hadoop

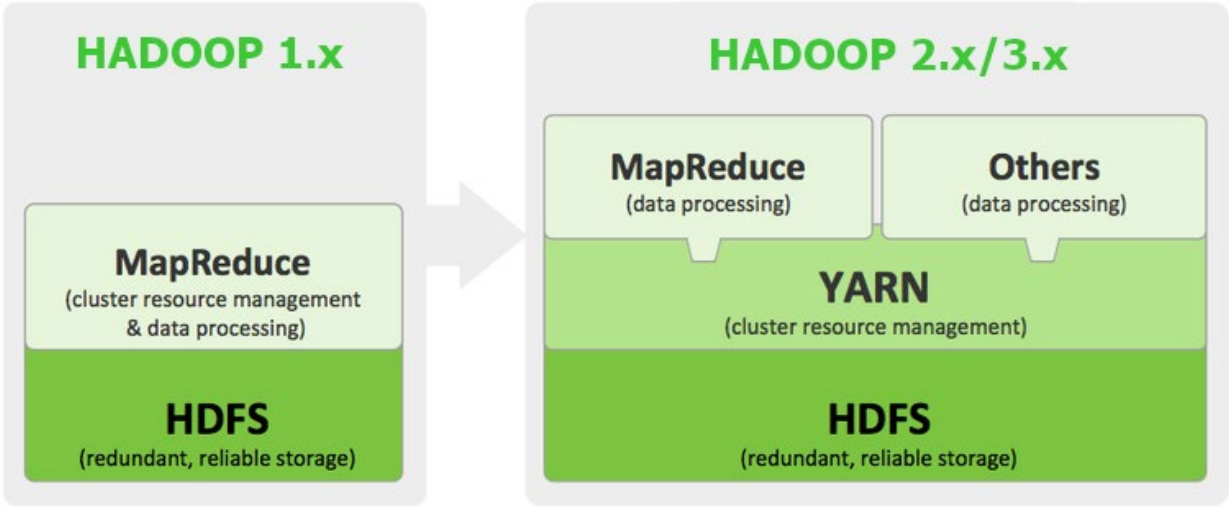
- Framework Hadoop e ferramentas complementares



Fonte da imagem: https://blogs.sap.com/2017/07/19/bridging-two-worlds-integration-of-sap-and-hadoop-ecosystems/

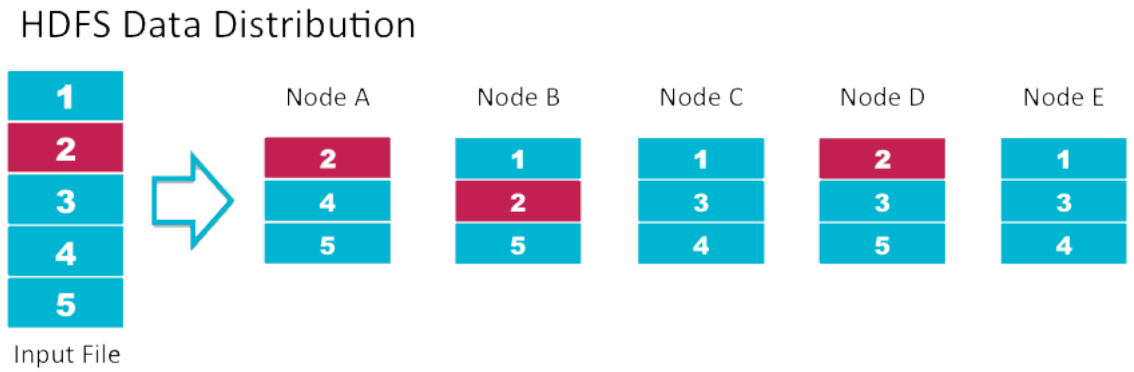
Introdução ao Ecossistema Hadoop

- Componentes base (core) do Apache Hadoop
  - HDFS (*Hadoop Distributed File System*)
    - Armazenamento distribuído
  - MapReduce
    - Computação distribuída
  - Yarn
    - Escalonamento de tarefas e gerenciamento de recursos



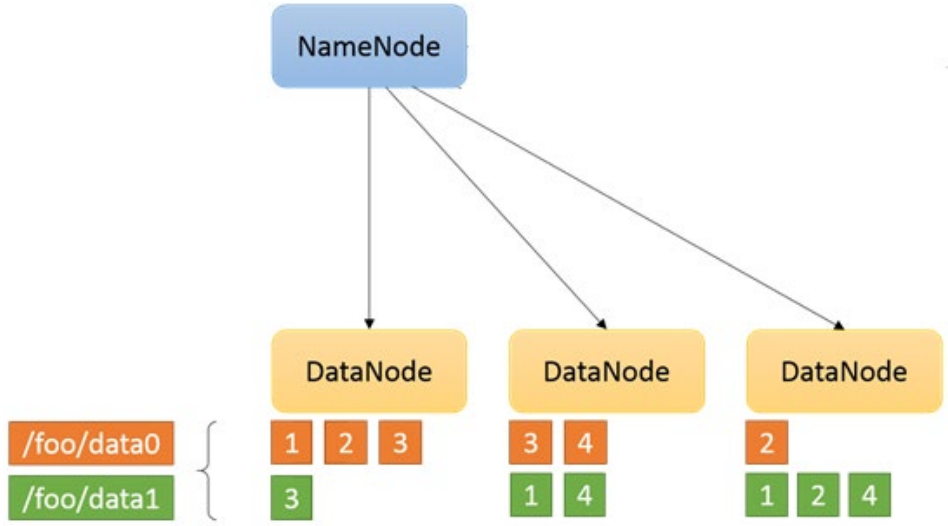
Sistema de arquivos HDFS

- HDFS (Hadoop Distributed File System)
  - Armazenamento distribuído
    - Otimizado para arquivos grandes
    - Princípio WORM (Write Once, Read Many Times)
  - Opera no modelo Mestre/Escravo
  - Os dados são armazenados em blocos (tipicamente de 128 MBytes)
    - São criadas várias réplicas dos dados e espalhadas nos nós do cluster
    - Tolerante a falhas (confiável): nós podem falhar
      - Recuperação automática: dados são redistribuídos



Sistema de arquivos HDFS

- HDFS (Hadoop Distributed File System)
  - Um cluster HDFS possui basicamente dois nós:
    - Mestre (NameNode)
      - Mantém e gerencia informações sobre blocos de dados
    - Escravo (DataNode)
      - Armazenam os dados em blocos
    - Pode ainda existir um segundo Mestre
      - Secondary NameNode

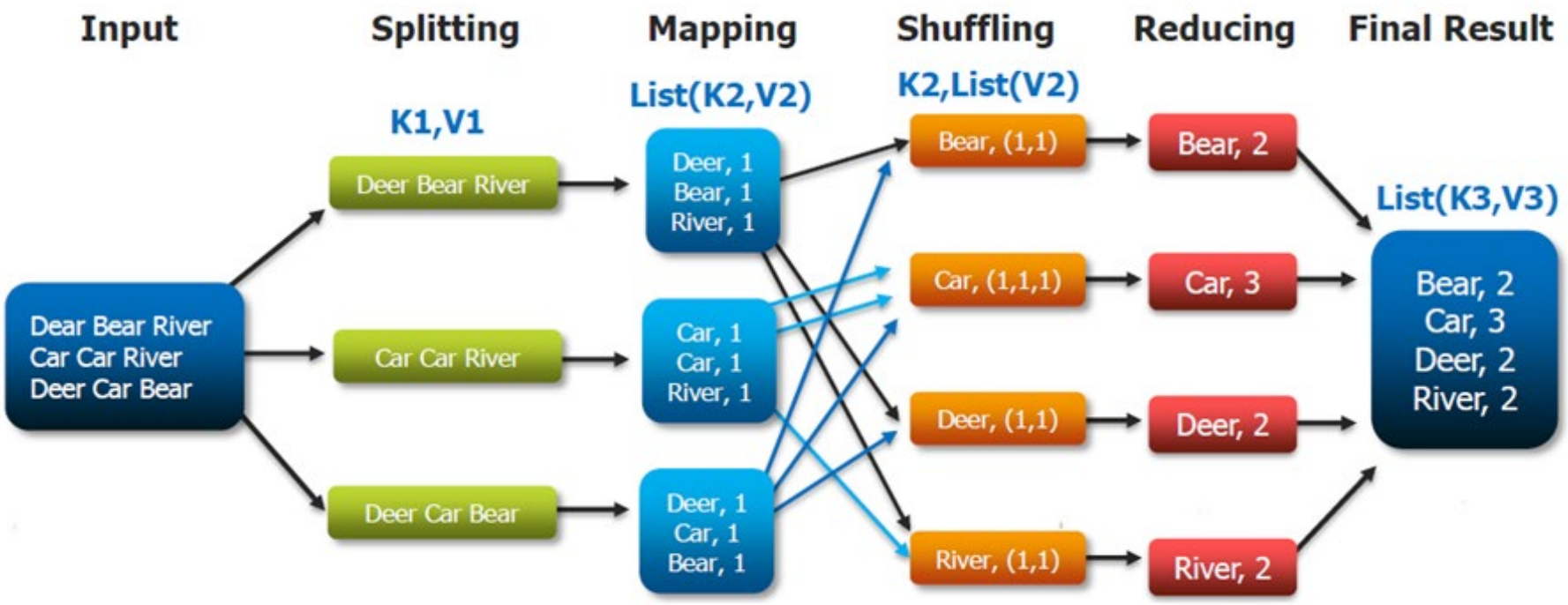


Modelo de programação MapReduce

- MapReduce
  - Computação distribuída
  - Programação para processamento de grandes conjuntos de dados
    - Pode ser programado em várias linguagens
  - Processa de forma paralela e distribuída os dados armazenados no HDFS
    - São criadas tarefas para processamento em lote
  - Lê os dados como pares chave/valor: `map(K1, V1)`
  - Basicamente funciona fazendo mapeamento e redução
    - Mapeia os dados
    - E os reduz (classificando)
    - Vários processos são disparados para realizar essas funções
      - Job trackers: gerenciam as tarefas do MapReduce
      - Task Trackers: monitoram individualmente cada tarefa de mapeamento e redução

Modelo de programação MapReduce

- MapReduce
  - Exemplo ilustrativo de funcionamento

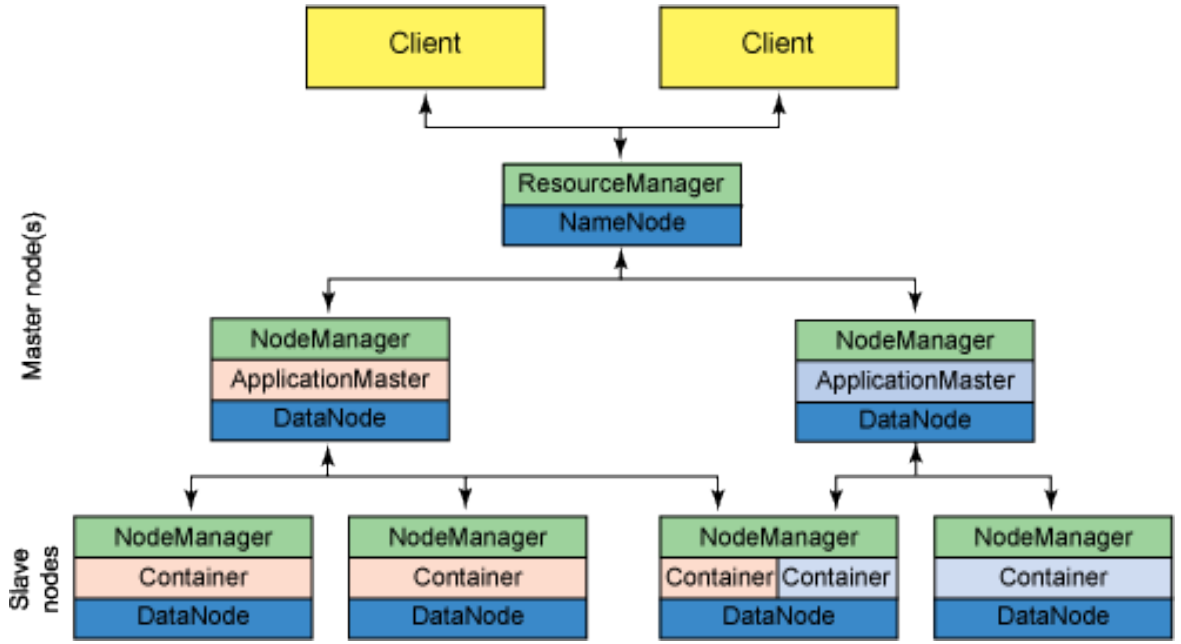


Modelo de programação MapReduce

- **Mapeamento (map)**
  - A função `mapper()` processa uma série de pares chave-valor (sequencialmente e individualmente), e produz na saída também pares de chave-valor
- **Embaralhamento e ordenação/classificação (shuffle and sort)**
  - À medida que o mapeamento é finalizado, o resultado da saída é enviado para a função de redução `reducer()`. Este processo é chamado de embaralhamento (shuffle).
  - E na ordenação/classificação todos os pares chave-valor da saída são ordenados para serem enviados para a função `reducer()`
- **Redução (reduce)**
  - A função `reducer()` agrega os valores de cada chave e produz pares de saída chave-valor

Gerenciamento de recursos com Yarn

- Arquitetura do Hadoop Yarn
  - ResourceManager: gerencia as tarefas do Hadoop Yarn, sendo responsável pelo escalonamento e execução de processos nos nós escravos
  - NodeManager: gerencia a execução de tarefas em cada nó

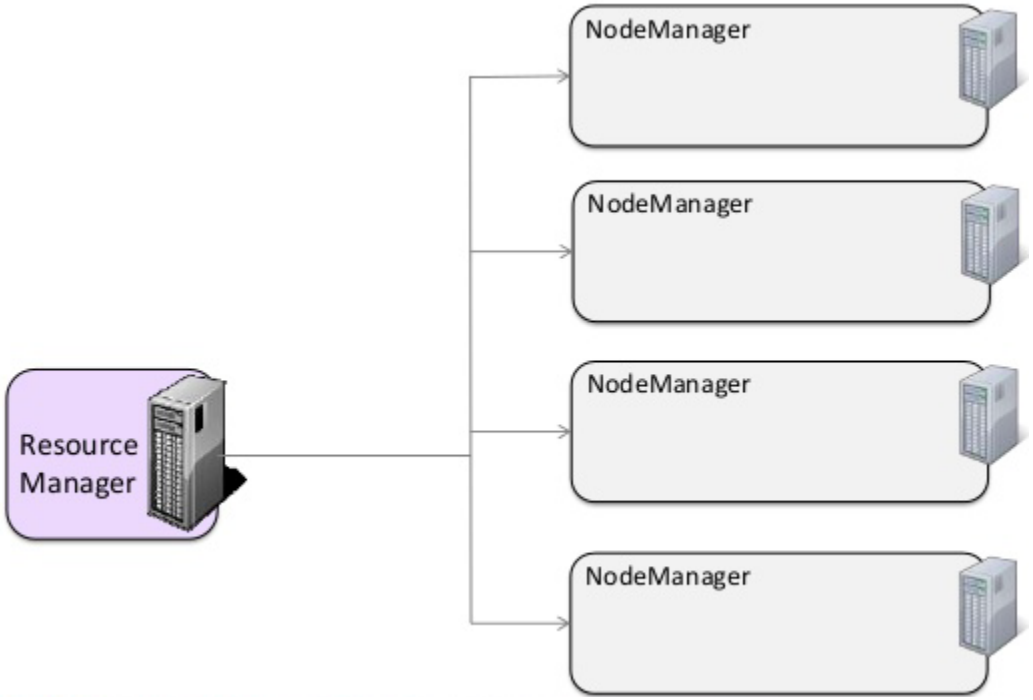


Gerenciamento de recursos com Yarn

- Funcionamento do Hadoop e Yarn
  - Cada tarefa possui um Application Master (AM) para monitorar a aplicação e coordenar a execução no cluster
  - O AM controla de que forma uma tarefa será executada
  - A execução ocorre em containers nos nós escravos
    - Cada nó escravo executa o NodeManager
    - O cluster todo é gerenciado pelo ResourceManager (executado no nó mestre)

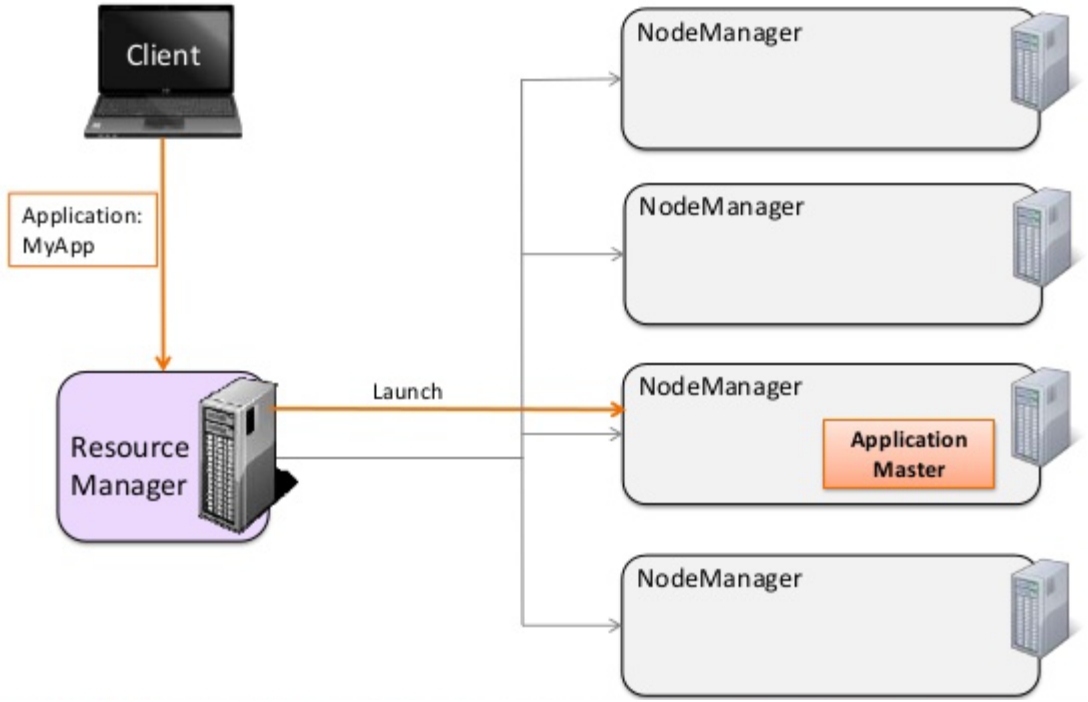
Gerenciamento de recursos com Yarn

- Funcionamento do Hadoop e Yarn



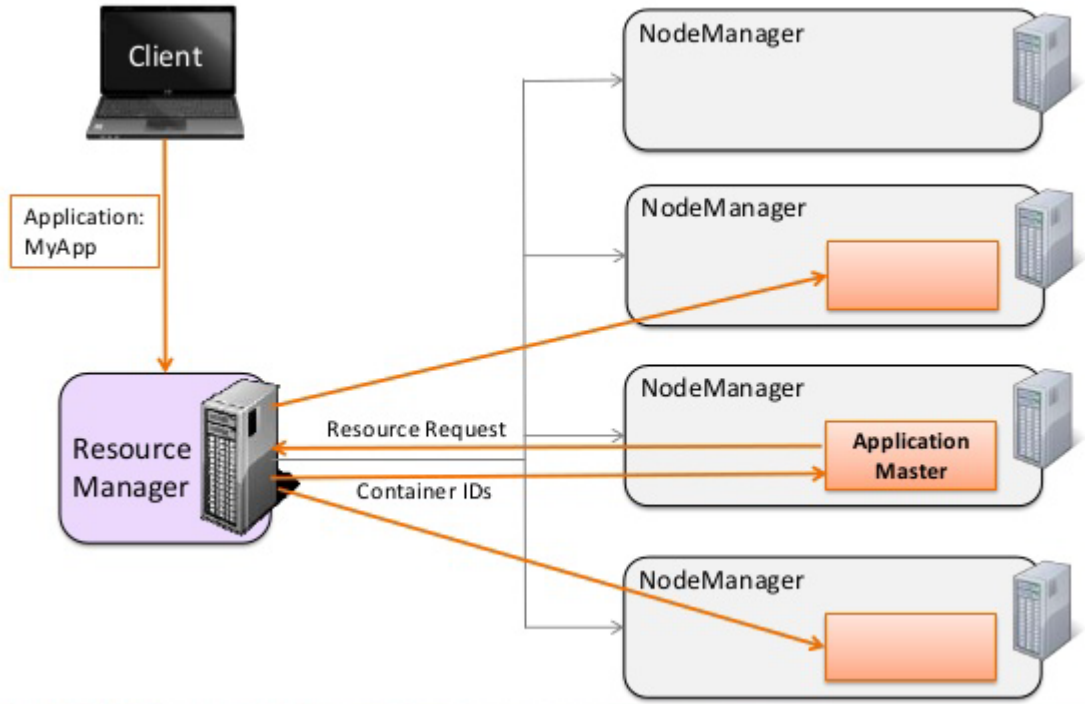
Gerenciamento de recursos com Yarn

- Funcionamento do Hadoop e Yarn



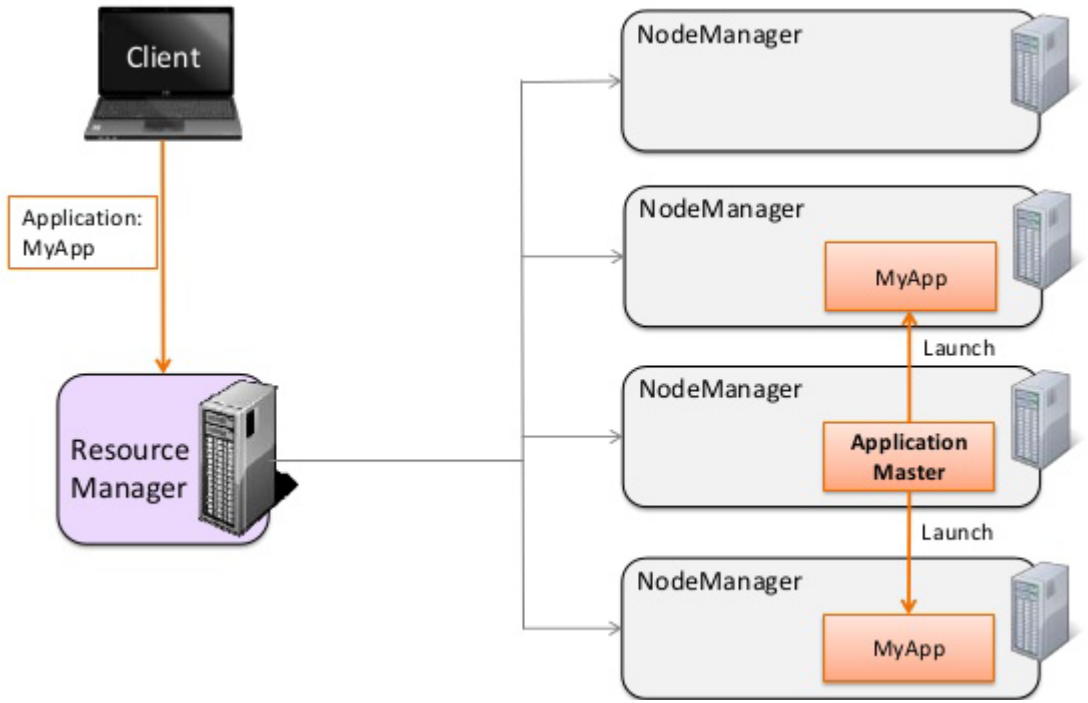
Gerenciamento de recursos com Yarn

- Funcionamento do Hadoop e Yarn



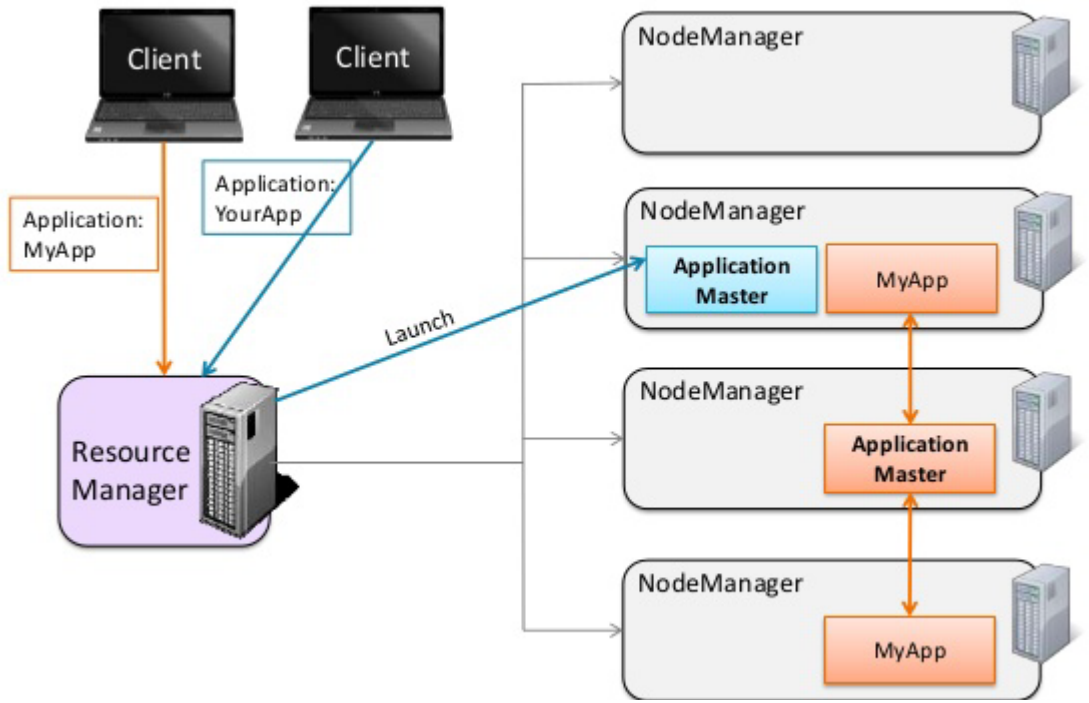
Gerenciamento de recursos com Yarn

- Funcionamento do Hadoop e Yarn



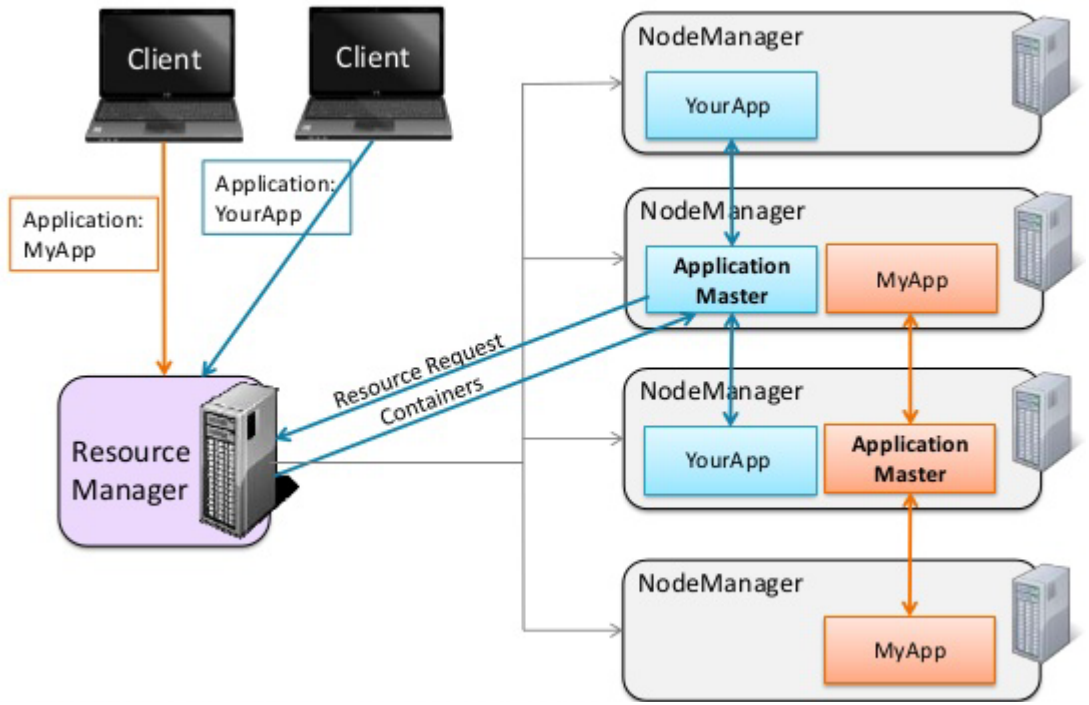
Gerenciamento de recursos com Yarn

- Funcionamento do Hadoop e Yarn

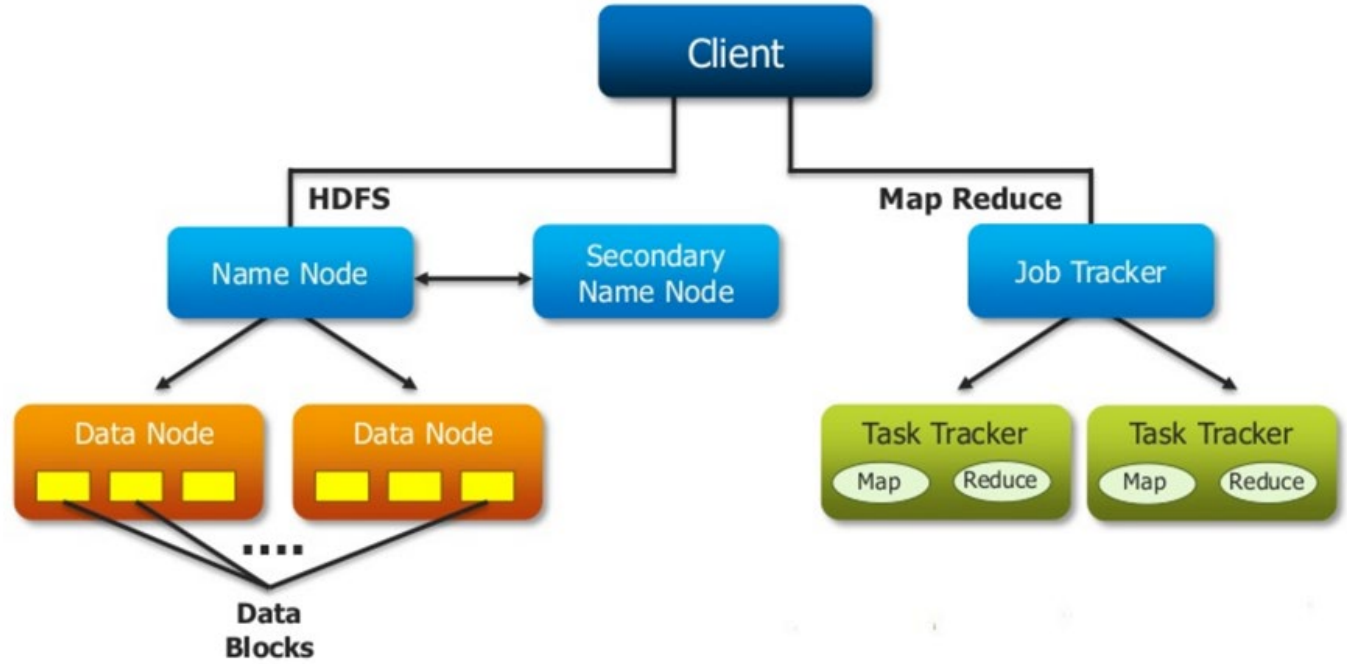


Gerenciamento de recursos com Yarn

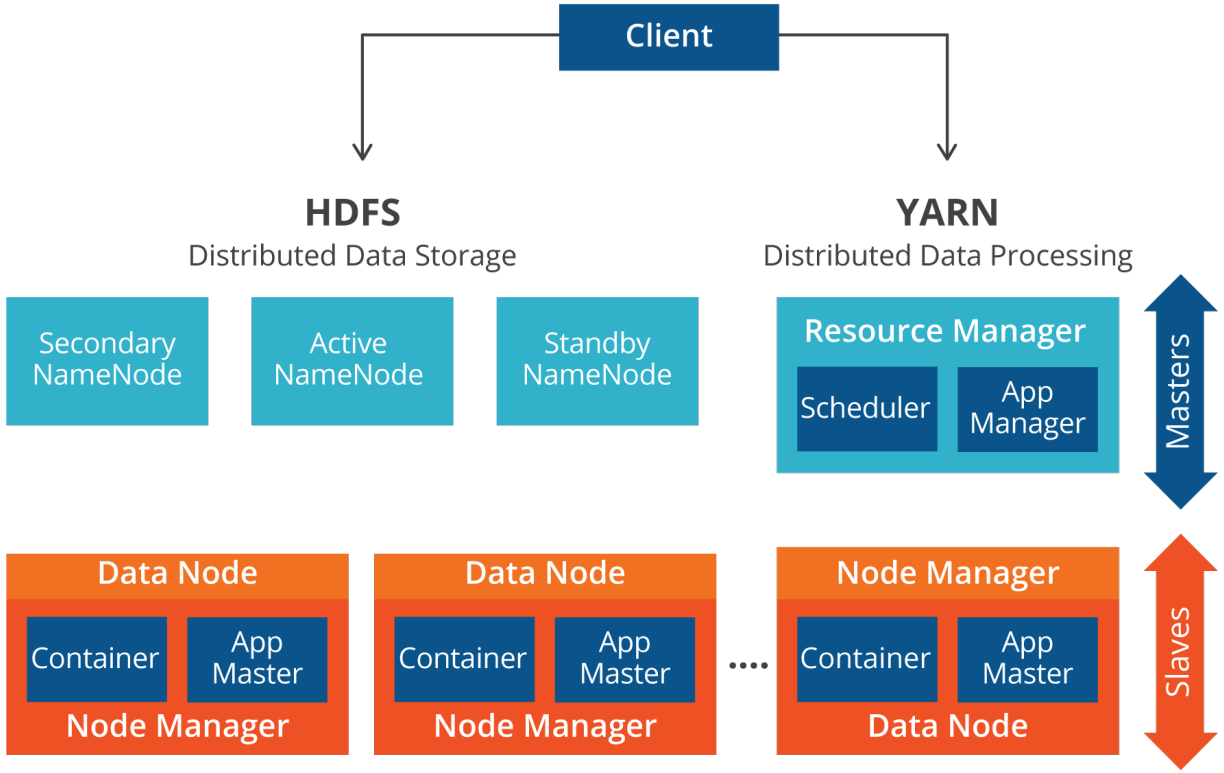
- Funcionamento do Hadoop e Yarn



Arquitetura Apache Hadoop 1.x



Arquitetura Apache Hadoop 2.x e 3.x



Fonte da imagem: https://www.bmc.com/blogs/hadoop-architecture/

Instalação e configuração do Hadoop

- Fatores importantes para a instalação ser bem sucedida:
  - Sistema Operacional com as variáveis ambiente corretamente definidas
  - Versão da máquina java (JVM)
    - Hadoop 3.x suporta Java 11

A instalação, configuração e análise de logs do Hadoop seguirá por meio dos roteiros disponibilizados na plataforma Moodle da CEAJUD.

OBRIGADO

CONTATO: viegas@dca.ufrn.br

CURSO DE CIÊNCIA DE DADOS APLICADA AO PODER JUDICIÁRIO

JUSTIÇA 4.0

PNUD

CNJ CONSELHO NACIONAL DE JUSTIÇA