PUC Minas

Introdução ao Hadoop

Diego Roberto Gonçalves de Pontes

O Crescimento Exponencial dos dados

- Geração de Dados: Nos últimos anos, a quantidade de dados gerados aumentou exponencialmente.
  - Empresas: Registros de transações, dados de clientes, logs de servidores.
  - Redes Sociais: Postagens, curtidas, compartilhamentos, comentários.
  - Dispositivos IoT: Sensores, câmeras, dispositivos inteligentes gerando dados continuamente.
- YouTube: Mais de 500 horas de vídeo carregadas por minuto.
- Facebook: Mais de 4 petabytes de dados gerados diariamente.
- Twitter: Aproximadamente 500 milhões de tweets por dia.

Introdução ao hadoop

- Desenvolvido em 2003 pelo Google Labs.
- Necessidade de processar e analisar grandes volumes de dados gerados pela web e outras fontes.
- Principal ferramenta: MapReduce para processamento distribuído, HDFS para armazenamento distribuído

Histórico do Hadoop

- 2003: Google desenvolve MapReduce e GFS (Google File System) para lidar com grandes volumes de dados.
- 2005: Doug Cutting, inspirado no GFS e MapReduce, cria o Hadoop.
- 2006: Yahoo! adota Hadoop e se torna um grande contribuidor do projeto.
- 2008: Hadoop se torna um projeto principal da Apache.
- 2011: Lançamento da versão 1.0.0 do Hadoop, consolidando sua posição no mercado.

O que é Hadoop?

- Definição: Framework de código aberto para armazenamento e processamento de grandes volumes de dados em clusters de computadores.
- Componentes Principais: HDFS (Hadoop Distributed File System) e MapReduce, que juntos permitem o processamento eficiente de Big Data.
- Uso: Amplamente utilizado por empresas como IBM, Facebook, Yahoo!, entre outras, para processar e analisar grandes volumes de dados.

Para Que Serve o Hadoop?

- Processamento de Big Data: Hadoop é ideal para processar grandes volumes de dados que excedem a capacidade dos sistemas tradicionais.
- Análise de Dados: Facilita a análise de dados provenientes de diversas fontes, como logs, sensores, mídias sociais, etc.
- Escalabilidade: Permite adicionar novos nós ao cluster facilmente, aumentando a capacidade de processamento conforme necessário.

Conceitos Básicos do Hadoop

- MapReduce: Modelo de programação que divide o processamento em duas fases: Map (mapeamento), que processa e gera pares chave-valor, e Reduce (redução), que agrega esses valores.
- HDFS: Sistema de arquivos distribuído que divide os dados em blocos e os distribui entre os nós do cluster para armazenamento eficiente e tolerância a falhas.
- Nó Mestre e Nós Escravos: O NameNode (nó mestre) gerencia o HDFS, enquanto os DataNodes (nós escravos) armazenam os dados e executam as tarefas de processamento.

Funcionamento do MapReduce

- Fase Map: Processa os dados de entrada, dividindo-os em pares chave-valor intermediários.
- Shuffle and Sort: Agrupa e ordena os pares chave-valor por chave, preparando-os para a fase de redução.
- Fase Reduce: Agrega os valores associados a cada chave, produzindo o resultado final do processamento.

Arquitetura do HDFS

- HDFS (Hadoop Distributed File System): Sistema de arquivos distribuído que armazena dados de forma redundante em um cluster.
- NameNode: Nó mestre que gerencia o namespace do sistema de arquivos e regula o acesso aos arquivos pelos clientes.
- DataNodes: Nós escravos que armazenam efetivamente os dados. Responsáveis por ler, escrever e replicar blocos de dados conforme instruções do NameNode.

Divisão de Arquivos em Blocos

- Blocos de Dados: Os arquivos são divididos em blocos de tamanho fixo (padrão de 128 MB).
- Vantagem: Facilita o armazenamento distribuído e o processamento paralelo.
- Distribuição: Cada bloco é armazenado em diferentes DataNodes, garantindo redundância e tolerância a falhas.

Função do NameNode

- Gerenciamento de Metadados: NameNode mantém informações sobre os arquivos, como localização dos blocos, réplicas e estrutura do diretório.
- Operações de Cliente: Quando um cliente solicita um arquivo, o NameNode informa quais DataNodes contêm os blocos desse arquivo.
- Tolerância a Falhas: NameNode não armazena os dados reais, mas os metadados; portanto, sua alta disponibilidade é crucial.

Função dos DataNodes

- Armazenamento de Blocos: DataNodes armazenam os blocos de dados conforme instruções do NameNode.
- Relatório ao NameNode: DataNodes enviam relatórios regulares ao NameNode sobre os blocos armazenados.
- Execução de Operações: Leitura e escrita de dados, replicação de blocos para manter a redundância.

Processo de Leitura de Dados

- Solicitação ao NameNode: O cliente solicita ao NameNode a leitura de um arquivo.
- Localização dos Blocos: O NameNode retorna a localização dos blocos do arquivo nos DataNodes.
- Leitura dos Blocos: O cliente lê diretamente os blocos dos DataNodes.
- Montagem do Arquivo: Os blocos lidos são montados novamente para formar o arquivo original.

Tolerância a Falhas e Replicação

- Replicação: Cada bloco é replicado em múltiplos DataNodes para garantir redundância.
- Detecção de Falhas: O NameNode detecta falhas nos DataNodes através de relatórios periódicos.
- Re-replicação: Em caso de falha de um DataNode, o NameNode instrui outros DataNodes a replicar os blocos afetados para manter o nível de replicação.

PUC Minas