Hadoop vs Spark

Integrantes:

Lucas Santos Nogueira
Matheus Barbosa Souza

Rafael Sidnei Alves

E-mail: mbscbsjbs@gmail.com

Sumário

mbscbsjbs@gmail.com

Introdução

- A história da computação e marcada por períodos de inovação disruptiva que muda completamente o cenário da tecnologia.
- Os gerenciadores de banco de dados pareciam imunes a mudanças de paradigma. Mas desde 2004, o cenário começou a mudar e agora a mudança parece inevitável.
- A prova disso, e que os grandes players do mercado, Oracle, IBM e Microsoft, estão revendo seus produtos, que juntos representam mais de 90% do mercado de bancos de dados.

mbscbsjbs@gmail.com

Hadoop versus Spark

- Hadoop e Spark são tecnologias diferentes, com diferentes casos de uso.
- Berço das duas tecnologias, a própria Apache Software Foundation as aloca em categorias diferentes:
- Hadoop é um banco de dados, o Spark é uma ferramenta de Big Data.
- O Hadoop e usado para processamento em lote, enquanto o Spark pode ser usado para ambos.
- A esse respeito, os usuários do Hadoop podem processar usando tarefas MapReduce em que o processamento em lote é necessário.
- Em teoria, o Spark pode executar tudo o que o Hadoop pode e muito mais.

mbscbsjbs@gmail.com

O Spark tem melhor desempenho que o Hadoop quando:

- O tamanho dos dados varia de GBs a PBs;
- Há uma complexidade algorítmica variada, do ETL ao SQL ao aprendizado de máquina;
- Trabalhos de fluxo de baixa latência para trabalhos em lote longos processamento de dados, independentemente do meio de armazenamento, seja discos, SSDs ou memória;
- Além destes, o Hadoop ultrapassa o Spark.

mbscbsjbs@gmail.com

Hadoop

- O Hadoop e um framework open source, escalável e tolerante a falhas escrito em Java.
- Ele processa com eficiência grandes volumes de dados em um cluster de hardware de commodity.
- O Hadoop não é apenas um sistema de armazenamento, mas também uma plataforma para armazenamento de dados e processamento de dados.
- O Hadoop começou como um projeto do Yahoo em 2006, tornando-se um projeto de código aberto do Apache de nível superior mais tarde.

mbscbsjbs@gmail.com

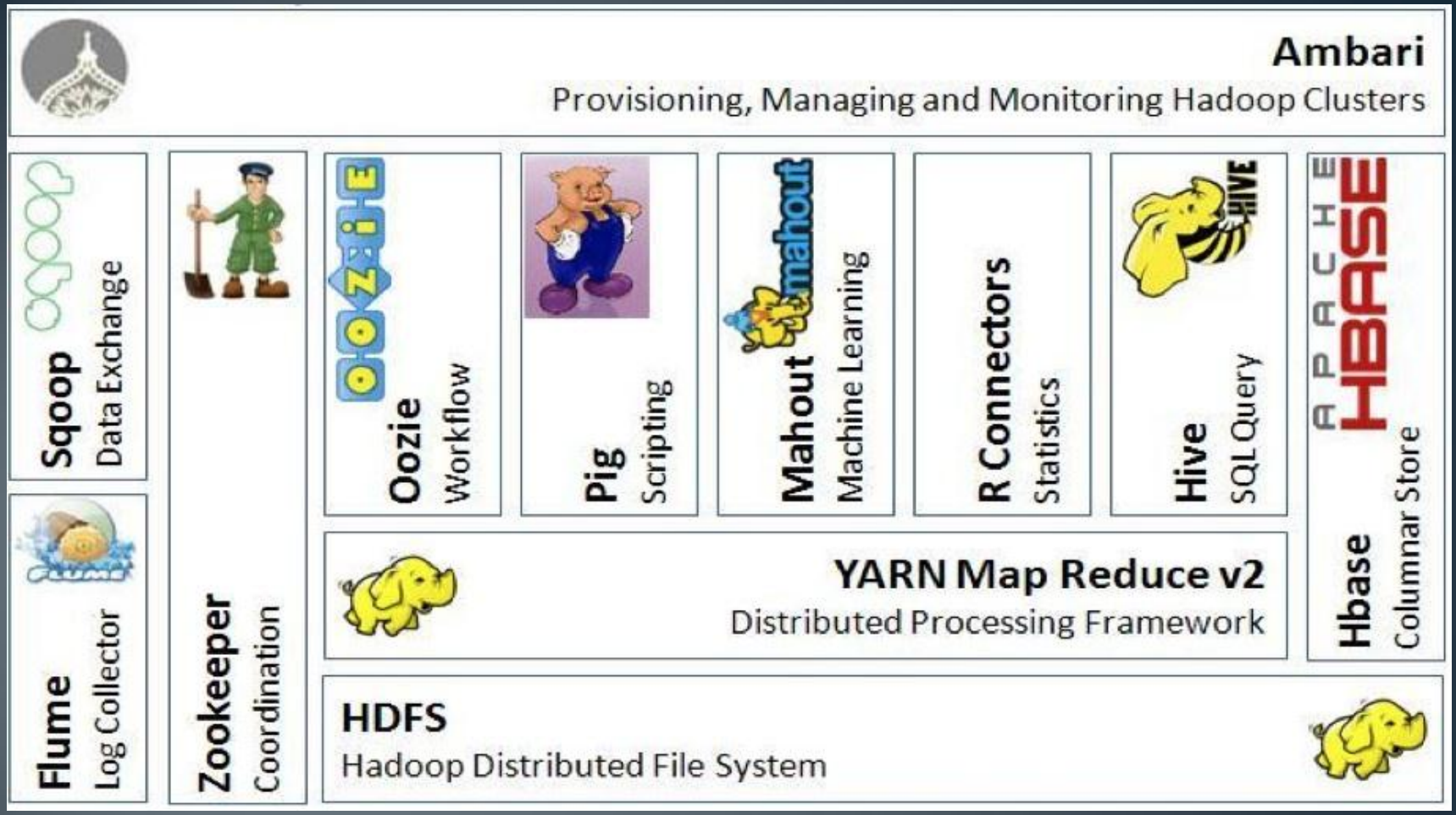

Figura 1 – Arquitetura de Hadoop.

Spark

- O Spark é um projeto mais recente, desenvolvido inicialmente em 2012, no AMPLab, na UC Berkeley.
- É também um projeto Apache de nível superior focado no processamento de dados em paralelo em um cluster.
- Mas a maior diferença e que ele funciona na memoria.
- O Spark é estruturado em torno do Spark Core, o mecanismo que impulsiona o agendamento, as otimizações e a abstração do RDD.
- Existem varias bibliotecas que operam sobre o Spark Core, incluindo o Spark SQL, que permite executar comandos semelhantes a SQL em conjuntos de dados distribuídos.

mbscbsjbs@gmail.com

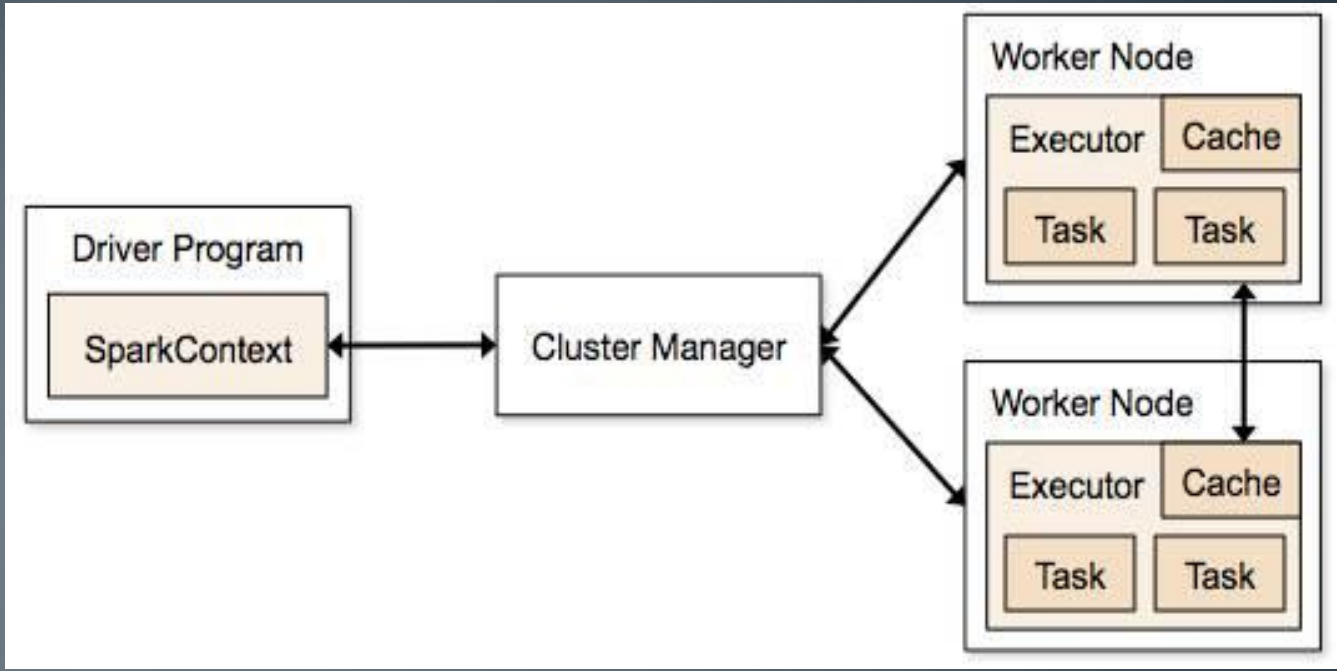

Figura 2 – Arquitetura do Spark.

Diferenças

Arquitetura:

Hadoop

- Todos os arquivos passados para o HDFS são divididos em blocos.
- Cada bloco é replicado um número especificado de vezes no cluster com base em um tamanho de bloco configurado e em um fator de replicação.
- MapReduce fica no topo do HDFS e consiste em um JobTracker.
- Depois que um aplicativo é gravado em um dos idiomas, o Hadoop aceita o JobTracker, o seleciona e aloca o trabalho para TaskTrackers ouvindo em outros nós.

mbscbsjbs@gmail.com

Spark

- Funcionam de maneira semelhante ao Hadoop, exceto pelo fato de que as computações são realizadas na memória e armazenadas lá, até que o usuário as persista ativamente.
- Conforme o RDD e as ações relacionadas estão sendo criados, o Spark também cria um DAG, ou Directed Acyclic Graph, para visualizar a ordem das operações e o relacionamento entre as operações no DAG. Cada DAG tem etapas e etapas. Dessa maneira, e semelhante a um plano de explicação no SQL.
- Uma nova abstração no Spark é o DataFrames, que foi desenvolvido no Spark 2.0 como uma interface complementar aos RDDs.
- Os dois são extremamente semelhantes, mas os DataFrames organizam os dados em colunas nomeadas, semelhantes aos pandas ou pacotes R do Python.

mbscbsjbs@gmail.com

Custos

- O Spark e o Hadoop estão disponíveis gratuitamente como projetos Apache de código aberto.
- A regra geral para instalações no local é que o Hadoop requer mais memória no disco e o Spark requer mais RAM.
- Como o Spark é o sistema mais novo, os especialistas nele são mais raros e mais caros.
- Um cluster EMR otimizado para computação para o Hadoop, o custo para a instância mais pequena, c4.large, e de US $0,026 por hora.
- E o menor cluster otimizado para memória do Spark custaria US $0,067 por hora.

mbscbsjbs@gmail.com

Tolerância e Falha de Segurança

- O Hadoop é altamente tolerante a falhas porque foi projetado para replicar dados em muitos nós. Cada arquivo é dividido em blocos e replicado inúmeras vezes em varias maquinas, garantindo que, se uma única máquina ficar inativa, o arquivo possa ser reconstruído a partir de outros blocos em outro lugar.
- A tolerância a falhas do Spark é obtida principalmente por meio de operações de RDD. Inicialmente, o data-at-rest é armazenado no HDFS, que é tolerante a falhas por meio da arquitetura do Hadoop.

mbscbsjbs@gmail.com

Semelhanças

Os componentes do Hadoop podem ser usados juntamente com o Spark das seguintes maneiras:

- HDFS: O Spark pode ser executado sobre o HDFS para aproveitar o armazenamento replicado distribuído.
- MapReduce: O Spark pode ser usado junto com o MapReduce no mesmo cluster do Hadoop ou separadamente como uma estrutura de processamento.
- YARN: Os aplicativos Spark podem ser executados no YARN (Hadoop NextGen).
- Processamento em lote e em tempo real: MapReduce e Spark s~ao usados juntos, onde MapReduce é usado para processamento em lote e Spark para processamento em tempo real.

mbscbsjbs@gmail.com

Referências

- Lei Gu and Huan Li. Memory or time: Performance evaluation for iterative operation on hadoop and spark. In *High Performance Computing and Communicati-ons & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC EUC), 2013 IEEE 10th International Conference on*, pages 721{727. IEEE, 2013.
- Afshan K. What is the dierence between hadoop and spark?, 2017.
- Amir K. How do hadoop and spark stack up?, 2018.