Introdução ao Hive

O termo 'Big Data' é usado para coleções de grandes conjuntos de dados que incluem grande volume, alta velocidade e uma variedade de dados que cresce dia a dia. Usando sistemas tradicionais de gerenciamento de dados, é difícil processar muitos dados. Portanto, a Apache Software Foundation introduziu uma estrutura chamada Hadoop para resolver o gerenciamento de Big Data de e desafios de processamento.

Hadoop

Hadoop é um framework open-source para armazenar e processar grandes volumes de dados em um ambiente distribuído. Composto por 2 módulos, um é o MapReduce e o outro é o Hadoop Distributed File System (HDFS).

MapReduce: É um modelo de programação paralelo para processar grandes quantidades de dados, semi-estruturados e não estruturados em grandes clusters de hardware de commodities.

HDFS: Hadoop Distributed File System é um parte do framework Hadoop, usado para armazenar e processar conjuntos de dados. Ele fornece um sistema de arquivos tolerante a falhas, para ser executado em hardware de commodities.

O ecossistema Hadoop contém diferentes subprojetos (ferramentas) como o Sqoop, Pig e Hive que são usados para ajudar os módulos do Hadoop.

- **Sqoop**: é usado para importar e exportar os dados de e para entre HDFS e RDBMS.
- **Pig**: É uma plataforma de linguagem procedural usada para desenvolver scripts para operações MapReduce.
- **Hive**: é uma plataforma usada para desenvolver scripts tipo SQL para fazer operações MapReduce.

Nota: Existem várias formas de executar operações MapReduce:

- A abordagem tradicional usando o programa Java MapReduce para dados estruturados, semi-estruturados e não estruturados.
- A abordagem de script para MapReduce para processar dados estruturados e semi-estruturados usando Pig.
- O Hive Query Language (HiveQL ou HQL) para MapReduce para processar dados estruturados usando Hive.

O que é Hive

Hive é uma ferramenta de infraestrutura de data warehouse para processar dados estruturados no Hadoop. Ele reside no topo do Hadoop para sumarizar os dados, fazer a consulta e análise fácil.

Inicialmente Hive foi desenvolvido pelo Facebook, mais tarde a Apache Software Foundation assumiu e desenvolveu-o ainda mais como open-source, sob o nome Apache Hive. É usado por companhias diferentes. Por exemplo, a Amazon no Amazon Elastic MapReduce.

Hive não é

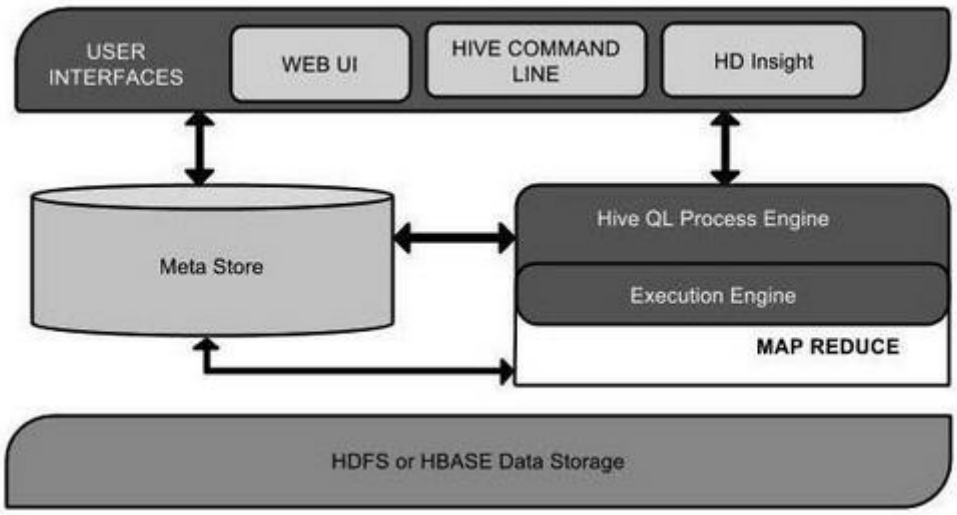
- Uma banco relacional;
- Um design para OnLine Transaction Processing(OLTP);
- Uma linguagem para consultas em tempo real e atualizações row-level;

Características Hive

- Armazena o esquema em um banco de dados e processas dads no HDFS;
- É projetado para OLAP;
- Fornece linguagem tipo SQL para consulta, chamada HiveQL ou HQL;
- É familiar, rápido, escalável e extensível;

Arquitetura Hive

O diagrama de componentes a seguir descreve a arquitetura do Hive:

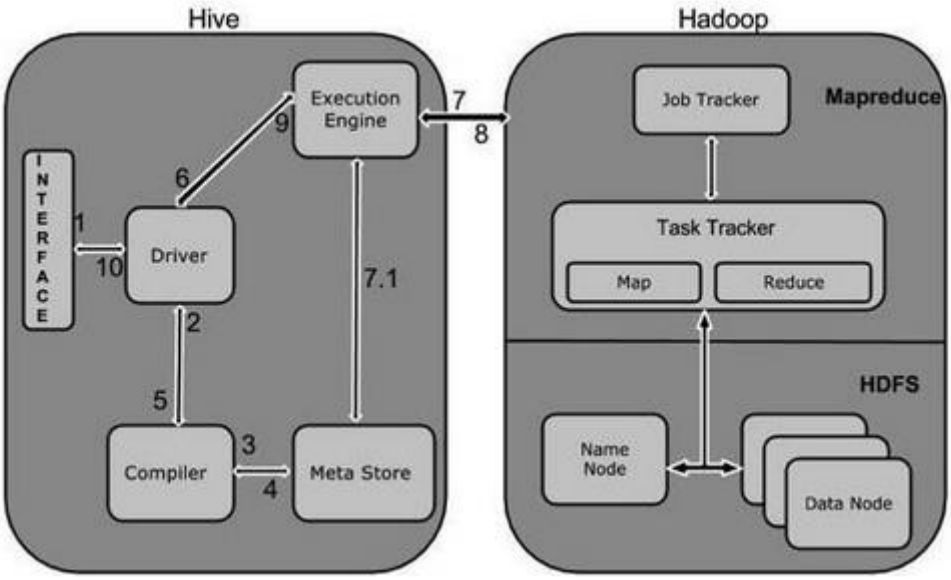


Este diagrama de componentes contém diferentes unidades. A tabela a seguir descreve cada unidade:

Nome da Unidade	Operação
Interface do Usuário	Hive é um software de infraestrutura de data warehouse que pode criar interação entre usuário e HDFS. As interfaces de usuário que o Hive suporta são Hive Web UI, linha de comando Hive e Hive HD Insight (no servidor Windows).
Meta Store	O Hive escolhe os respectivos servidores de banco de dados para armazenar o esquema ou Metadados de tabelas, bancos de dados, colunas em uma tabela, seus tipos de dados e mapeamento HDFS.
HiveQL Process Engine	HiveQL é semelhante ao SQL para consultar informações de esquema no Metastore. É uma das substituições da abordagem tradicional para o programa MapReduce. Em vez de escrever o programa MapReduce em Java, podemos escrever uma consulta para o job MapReduce e processá-la.
Execution Engine	A parte de junção do mecanismo de processo HiveQL e MapReduce é Hive Execution Engine. O mecanismo de execução processa a consulta e gera resultados igual aos resultados do MapReduce. Ele usa o sabor do MapReduce.
Hbase ou HDFS	O sistema de arquivos distribuído ou Hbase são técnicas de armazenamento de dados no sistema de arquivos.

Trabalhando com o Hive

O diagrama a seguir descreve o fluxo de trabalho entre o Hive eo Hadoop.



A tabela a seguir define como o Hive interage com o framework Hadoop:

Step Nº	Operation
1	**Executa Consulta**: A interface do Hive, como linha de comando ou UI Web, envia consulta ao Driver (qualquer driver de banco de dados, como JDBC, ODBC, etc.) para executar.
2	**Obtém o plano**: O Driver, com ajuda do compilador, válida a sintaxe e o plano de consulta ou requerimento de consulta.
3	**Obtém metadados:** O compilador envia os dados requisitados para a Meta Store (Qualquer banco de dados).
4	**Enviar metadados**: Metastore envia metadados como uma resposta para o compilador.
5	**Envia Plano**: O compilador verifica o requisito e reenvia o plano para o driver. Até aqui, a análise e compilação da consulta está concluída.
6	**Executa o plano:** O Driver envia o plano de execução para o mecanismo de execução.
7	**Executar Job:** Internamente,o processo de execução do job é um job MapReduce; O mecanismo de execução envia a tarefa para o JobTracker, que está no name node e atribui essa tarefa a TaskTracker, que está no nó de dados. Aqui, a consulta executa o job MapReduce.
7.1	**Operações de Metadados:** Enquanto isso, na execução, o mecanismo de execução pode executar operações de metadados com o MetaStore.
8	**Resultado de Busca:** O mecanismo de execução recebe os resultados dos nós de dados.
9	**Enviar Resultados:** O mecanismo de execução envia os valores resultantes para o Driver.
10	**Enviar Resultados:** O driver envia os resultados para a interface Hive

Fonte: https://www.tutorialspoint.com/hive/hive_introduction.htm