Semantix

Hive - Básico

Aula 4

Quem sou eu?

Eu sou Rodrigo Augusto Rebouças.

Engenheiro de dados da Semantix
Instrutor do Semantix Academy

Você pode me encontrar em:
rodrigo.augusto@semantix.com.br

Introdução

Hive

Semantix

Análise de dados

- Hive
- Impala
- Presto
- Spark

Semantix

Análise Hive

- Apache Hive
  - Criado pelo Facebook em 2007
    - Processamento lento para as consultas diárias
    - Moveram Data warehouse para o Hadoop
    - Criar tarefas MapReduce consumia tempo
  - Ferramenta para permitir fácil acesso aos dados via SQL
- Data warehouse construído em cima do Hadoop
- Camada de acesso a dados armazenados no HDFS
- Não é um SGBD
  - Criar tabelas no Hive
  - Dados são armazenadas no HDFS

Análise Hive

- Realizar consultas em grandes volumes de dados
  - Recursos avançados de particionamento
    - Subdividir os dados
      - Organizar através de colunas
  - Não é usada para fornecer respostas em tempo real (Impala)

Componentes

Semantix

Componentes Hive

- HCatalog
  - Camada de gerenciamento de armazenamento para o Hadoop
  - Permite que usuários com diferentes ferramentas de processamento de dados leiam e gravem os dados
- WebHCat
  - Servidor web para se conectar com o Metastore Hive
- HiveServer2 (HS2)
  - Serviço que permite aos clientes executar consultas no Hive

Componentes Hive

- Metastore
  - Todos os metadados das tabelas e partições do Hive são acessados através do Hive Metastore
  - Existem diferentes maneiras de configurar o servidor metastore
    - Embedded Metastore
    - Local Metastore
    - Remote Metastore
- Beeline
  - Cliente Hive
  - Faz uso de JDBC para se conectar ao HiveServer2

Formato e Estrutura dos Dados

Semantix

Formato de Arquivos

- Não existe um formato Hive
- Conector para vários formatos
  - Arquivos de texto com valores separados por vírgula e tabulação (CSV / TSV)
  - Parquet
  - ORC
  - AVRO
  - JSONFILE
  - Outros ...

Estrutura dos Dados

- Dados estruturados e semi-estruturados
- Hierarquia dos dados
  - Database
    - Table
      - Partition - Coluna de armazenamento dos dados no sistema de arquivo (diretórios)
        - Bucket - Dados são divididos em uma coluna através de Hash

- Exemplo de caminho

/user/hive/warehouse/banco.db/tabela/data=010119/000000_0

Linguagem Hive

- Hive Query Language
- HiveQL
- HQL
  - Instruções SQL são transformadas internamente em Jobs de MapReduce

Banco de Dados e Tabelas

Semantix

Informações BD e Tabelas

- Listar todos os BD
  - show database;
- Estrutura sobre o bd
  - desc database <nomeBD>;
- Listar as tabelas
  - show tables;
- Estrutura da tabela
  - desc <nomeTabela>;
  - desc formatted <nomeTabela>;
  - desc extended <nomeTabela>;

Criação Banco de Dados

- Criar BD
  - create database <nomeBanco>;
- Local diferente do conf. Hive
  - create database <nomeBanco> location “/diretorio”;
- Adicionar comentário
  - create database <nomeBanco> comment “descrição”;
- Ex
  - create database test location “/user/hive/warehouse/test” comment “banco de dados para treinamento”
  - default
    - /user/hive/warehouse/test.db

Tabelas Hive

- Tipo
  - Internas
  - Externas
- Partição
  - Não particionada
  - Particionada
    - Dinâmico
    - Estático

Tabela Interna e Externa

- Tabela interna
  - create table user(cod int, name string);
  - drop table
    - Apaga os dados e metadados
- Tabela externa
  - create external table e_user(cod int, name string) location '/user/semantix/data_users';
  - drop table
    - Usar para compartilhar os dados com outras ferramentas
    - Apaga apenas os metadados
    - Dados ficam armazenado no sistema de arquivos

Atributos para Criação de Tabelas

Semantix

Tipos Dados Simples

- INT
- SMALLINT
- TINYINT
- BIGINT
- BOOLEAN
- FLOAT
- DOUBLE
- DECIMAL
- STRING
- VARCHAR
- CHAR

Tipos Dados Complexos

- ARRAY
  - Lista de Elementos [‘Seg’, ‘Ter’, ‘Qua’, ‘Qui’, ‘Sex’]
- MAP
  - Tipo Chave-valor ‘nome’ ->’Rodrigo’
- STRUCT
  - Define os campos dos seus tipos de dados
  - STRUCT<col_name
    : data_type, ...
- UNION
  - Armazenar diferentes tipos de dados no mesmo local de memória
  - UNIONTYPE<data_type,
    data_type, ...>

Opções Leitura de dados

- Definir delimitadores
  - row format delimited
  - fields terminated by ‘<delimitador>'
  - lines terminated by ‘<delimitador>’
  - Delimitadores: ‘qualquer coisa’, \b (backspace), \n (newline), \t (tab)

- Pular um número de linhas de leitura do arquivo
  - tblproperties("skip.header.line.count"=“<número de linhas");

- Definir Localização dos dados (Tabela externa)
  - location '/user/cloudera/data/client’;

Exemplo Criação de Tabela

- Tabela Externa

```
create external table user(
        id int,
        name String,
        age int
)
row format delimited
fields terminated by '\t'
lines terminated by ‘\n’
stored as textfile
location '/user/cloudera/data/client’;
```

Laboratório

Resolução de exercícios

Semantix

Exercícios Criação de Tabela Raw

1. Enviar o arquivo local “/input/exercises-data/populacaoLA/populacaoLA.csv” para o diretório no HDFS “/user/aluno/<nome>/data/populacao”
2. Listar os bancos de dados no Hive
3. Criar o banco de dados <nome>
4. Criar a Tabela Hive no BD <nome>
   a. Tabela interna: pop
   b. Campos:
      - zip_code - int
      - total_population - int
      - median_age - float
      - total_males - int
      - total_females - int
      - total_households - int
      - average_household_size - float
   c. Propriedades
   d. Delimitadores: Campo ‘,’ | Linha ‘\n’
   e. Sem Partição
   f. Tipo do arquivo: Texto
   g. tblproperties("skip.header.line.count"="1")’
5. Visualizar a descrição da tabela pop

Inserir e Carregar Dados

Semantix

Inserção Dados

- Inserir dados
  - insert into table <nomeTabela> partition(<partition>=’<value>’) values(<campo>,<value>), (<campo>,<value>), (<campo>,<value>);
- Ex
  - insert into users values(10, ‘Rodrigo’),(11,’Augusto’);
  - insert into users partition(data=now()) values(10, ‘Rodrigo’),(11,’Augusto’);
  - insert into users select * from cliente;

Carregamento Dados

- Carregar dados no sistema de arquivos local
  - hive> load data inpath <diretório> into table <nomeTabela>;
- Ex.
  - load data local inpath ‘/home/cloudera/data/test’ into table alunos
  - load data inpath ‘/user/cloudera/data/test’ overwrite into table alunos partition(id)

Seleção de Dados

Semantix

Seleção Dados

- select * from <nometable>
  <where ...>
  <group by … >
  <having … >
  <order by … >
  <limit n>;

- Ex
  - hive> **select** * from client **where** state=sp **group by** city **having** population > 100 **order by** client **limit** 10;

Tipos Join

- Aceita apenas ANSI JOINS
- Inner Join, Left Outer, Right Outer, Full Outer
  - select * from a join b on a.valor = b.valor
  - select * from a,b where a.valor = b.valor
    - erro

View Consulta

- Salvar consultas
- Tratar como tabelas
- Objetos Lógicos
  - Esquema é fixo quando criado a View
  - Alterar tabela não altera a view
- Comando
  - create view <nomeView> as select * from nome_table;

Laboratório

Resolução de exercícios

Semantix

Exercícios Inserir Dados na Tabela Raw

1. Visualizar a descrição da tabela pop do banco de dados <nome>
2. Selecionar os 10 primeiros registros da tabela pop
3. Carregar o arquivo do HDFS “/user/aluno/<nome>/data/população/populacaoLA.csv” para a tabela Hive pop
4. Selecionar os 10 primeiros registros da tabela pop
5. Contar a quantidade de registros da tabela pop

Semantix

Obrigado!

Alguma pergunta?

Você pode me encontrar em:

rodrigo.augusto@semantix.com.br

GET SMARTER