Semantix

Hive - Básico

Aula 5

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

Particionamento

Conceitos e Comandos Básicos

Semantix

Particionamento Hive

- Tabela não particionada
  - Consultas precisam varrer todos os arquivos no diretório
    - Processo lento para big table
- Tabela particionada
  - Consultas podem varrer os arquivos de uma partição
    - Campo com registros duplicados (grupos)
      - estado, gênero, categoria, classe, etc...
  - Fatias horizontais de dados, separadas por partes

Tabela Particionada

- Definir parâmetros na criação da tabela
  - Partição
    - Um campo que não está na estrutura da tabela
    - Organizar os dados
    - Consultas SQL interpretarão como coluna
      - Select * from user where **cidade**=sp
    - Comando
      - partitioned by (<campo> <type>)
  - Buket
    - Quantidade que os dados serão divididos
    - Campo precisa estar na estrutura da tabela
    - Comando
      - clustered by (<campo>) into <qtd> buckets

Exemplo Criação de Tabela

- Tabela Particionada

```
create table user(
        id int,
        name String,
        age int
)
partitioned by (data String)
clustered by (id) into 4 buckets;
```

Tipos de Particionamento

Estático e Dinâmico

Semantix

Particionamento Estático

- Você pode inserir os arquivos individualmente em uma tabela de partição
- Criar novas partições manualmente
- Comando
  - hive> alter table <nomeTabela> add partition(<partição>=’<valor>’);
- Ex.
  - Criar uma partição para cada dia
  - hive> alter table logs add partition(data=’2019-21-02’);

Particionamento Dinâmico

- Partições são criadas automaticamente nos tempos de carregamento
- Novas partições podem ser criadas dinamicamente
  - Baseada no valor da última coluna
    - Se a partição não existir, ela será criada
    - Se existir, os dados serão adicionados na partição
  - Ex.
    - hive> insert overwrite table user_cidade partition (cidade) select * from user;
- Por padrão, o particionamento dinâmico está desativado, para ativar
    - SET hive.exec.dynamic.partition = true;
    - SET hive.exec.dynamic.partition.mode = nonstrict ;

Opções com Partições

- Visualizar partições de uma tabela
  - hive> show partitions user;

- Excluir partições de uma tabela
  - hive> alter table user drop partition (city='SP');

- Alterar nome da partição de uma tabela
  - hive> alter table user partition city rename to partition state;

Reparar Tabela

Semantix

Reparar Tabela

- Reparar partições na tabela Hive
  - Quando a tabela não encontra a partição
  - Sincronizar a tabela com o metastore
- Comando
  - msck repair table <nomeTabela>

Laboratório

Resolução de exercícios

Semantix

Exercícios Criação de Tabela Particionada

1. Criar a pasta “/user/aluno/<nome>/data/nascimento” no HDFS
2. Criar e usar o Banco de dados <nome>
3. Criar uma tabela externa no Hive com os parâmetros:
   a) Tabela: nascimento
   b) Campos: nome (String), sexo (String) e frequencia (int)
   c) Partição: ano
   d) Delimitadores:
      a) Campo ‘,’
      b) Linha ‘\n’
   e) Salvar
      a) Tipo do arquivo: texto
      b) Local: '/user/aluno/<nome>/data/nascimento’
4. Adicionar partição ano=2015
5. Enviar o arquivo local “/input/exercises-data/names/yob2015.txt” para o HDFS no diretório /user/aluno/<nome>/data/nascimento/ano=2015
6. Selecionar os 10 primeiros registros da tabela nascimento no Hive
7. Repita o processo do 4 ao 6 para os anos de 2016 e 2017.

Exercícios Seleção de Tabelas

1. Selecionar os 10 primeiros registros da tabela nascimento pelo ano de 2016
2. Contar a quantidade de nomes de crianças nascidas em 2017
3. Contar a quantidade de crianças nascidas em 2017
4. Contar a quantidade de crianças nascidas por sexo no ano de 2015
5. Mostrar por ordem de ano decrescente a quantidade de crianças nascidas por sexo
6. Mostrar por ordem de ano decrescente a quantidade de crianças nascidas por sexo com o nome iniciado com ‘A’
7. Qual nome e quantidade das 5 crianças mais nascidas em 2016
8. Qual nome e quantidade das 5 crianças mais nascidas em 2016 do sexo masculino e feminino

Tipos de arquivos e compressão

Semantix

Tipos de Arquivo para Salvar

- Adicionar o parâmetro stored na criação da tabela
- stored as <formatoArquivo>
  - TEXTFILE (Padrão)
  - SEQUENCEFILE
  - RCFILE
  - ORC (Hortonworks)
  - PARQUET (Cloudera)
  - AVRO
  - JSONFILE

Tipos de Compressão para Salvar

- Se o arquivo mapred-site.xml não estiver configurado pode setar manualmente
  - hive> SET hive.exec.compress.output=true;
  - hive> SET mapred.output.compression.codec= **codec**;
  - hive> SET mapred.output.compression.type=BLOCK;
- Codec:
  - gzip: org.apache.hadoop.io.compress.GzipCodec
  - bzip2: org.apache.hadoop.io.compress.BZip2Codec
  - LZO: com.hadoop.compression.lzo.LzopCodec
  - Snappy: org.apache.hadoop.io.compress.SnappyCodec
  - Deflate: org.apache.hadoop.io.compress.DeflateCodec

Tipos de Compressão para Salvar

- Adicionar o parâmetro compress em tblproperties na criação da tabela
- stored as <formatoArquivo> tblproperties(' formatoArquivo.compress’=‘<CompressaoArquivo>’);
  - GZIP
  - BZIP2
  - LZO
  - SNAPPY
  - DEFLATE

Exemplo Criação de Tabela

- Tabela com partição e compressão

```
create table user(
        id int,
        name String,
        age int
)
partitioned by (data String)
clustered by (id) into 256 buckets
stored as parquet tblproperties('parquet.compress'='SNAPPY');
```

Laboratório

Resolução de exercícios

Semantix

Exercícios Criação de Tabelas Otimizadas

1. Usar o banco de dados <nome>
2. Selecionar os 10 primeiros registros da tabela pop
3. Criar a tabela pop_parquet no formato parquet para ler os dados da tabela pop
4. Inserir os dados da tabela pop na pop_parquet
5. Contar os registros da tabela pop_parquet
6. Selecionar os 10 primeiros registros da tabela pop_parquet
7. Criar a tabela pop_parquet_snappy no formato parquet com compressão Snappy para ler os dados da tabela pop
8. Inserir os dados da tabela pop na pop_parquet_snappy
9. Contar os registros da tabela pop_parquet_snappy
10. Comparar as tabelas pop, pop_parquet e pop_parquet_snappy no HDFS.

Semantix

Obrigado!

Alguma pergunta?

Você pode me encontrar em:

rodrigo.augusto@semantix.com.br

GET SMARTER