Semantix

Sqoop - Básico

Aula 7

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

Otimizar Importação

Semantix

Paralelismo

Semantix

Quantidade Mapeadores

- Por padrão, o número de mapeadores é 4
- -m, ou –num-mappers: Quantidade de mapeadores
  - $ sqoop import … -m 2
- Divisão aplicada a coluna com chave primária
  - Se existir
    - -num-mappers 8
  - Se não existir
    - -num-mappers 1
    - –auto-reset-to-one-mapper
      - manipular tabelas automaticamente
    - -num-mappers > 1 = erro
      - Solução split

Divisão Colunas não Chave

- –split-by: Dividir mapeadores em uma coluna sem chave
  - $ sqoop import … --plit-by: id

- Valores nulos na coluna
  - Registros correspondentes da tabela serão ignorados
- Dados na coluna de divisão não precisam ser exclusivos
  - Pode haver uma distorção nos dados durante a importação

Valores Nulos

Semantix

Valores Nulos

- Por padrão o Sqoop importa os dados null como string null
  - Valor escrito para um campo nulo de número
    - --null-non-string <valor nulo>
    - $ sqoop import … --null-non-string '-1'
  - Valor escrito para um campo nulo de string
    - --null-string <valor nulo>
    - $ sqoop import … --null-string ‘NA’
  - OBS:
    - $ sqoop import… --null-non-string '\\N' --null-string ‘\\N’
    - Importação compatível com Hive e Impala

Formato e Compressão de Dados

Semantix

Formato Dados

- Por padrão o Sqoop armazena os dados no formato de arquivo de texto
  - Formato de arquivo de texto (Default)
    - $ sqoop import ... --as-textfile
  - Formato de arquivo de Parquet
    - $ sqoop import ... --as-parquetfile
  - Formato de arquivo de Avro
    - $ sqoop import ... --as-avrodatafile
  - Formato de arquivo de Sequência
    - $ sqoop import ... --as-sequencefile

Compressão Dados

- Por padrão, o Sqoop comprimi os dados por gzip
- Codecs de compactação
  - Gzip - org.apache.hadoop.io.compress.GzipCodec
  - Bzip2 - org.apache.hadoop.io.compress.BZip2Codec
  - Snappy - org.apache.hadoop.io.compress.SnappyCode
  - Others (deflate, lz4, ...)
- Codecs suportados
  - /etc/hadoop/conf/core-site.xml
- Comandos
  - Habilitar compressão: --compress
  - Escolher compressão: --compression-codec <codec>

```
$ sqoop import … --compress  --compression-codec org.apache.hadoop.io.compress.SnappyCode
```

Laboratório

Resolução de Exercícios

Semantix

Exercícios Sqoop – Importação BD Employees - Otimização

MySQL

1. Criar a tabela cp_titles_date, contendo a cópia da tabela titles com os campos title e to_date
2. Pesquisar os 15 primeiros registros da tabela cp_titles_date
3. Alterar os registros do campo data para nulo da tabela cp_titles_date, quando o título for igual a Staff

Sqoop - Realizar as importações no warehouse /user/hive/warehouse/db_test e visualizar no HDFS

4. Importar a tabela salaries com 8 mapeadores no formato avro
5. Importar a tabela titles com 8 mapeadores no formato parquet e compressão snappy
6. Importar a tabela cp_titles_date com 4 mapeadores (erro)
   - Importar a tabela cp_titles_date com 4 mapeadores divididos pelo campo título
   - Importar a tabela cp_titles_date com 4 mapeadores divididos pelo campo data
   - Qual a diferença dos registros nulos?

Exercícios Bônus Importação BD Employees – Formato e Compressão

- Importar a tabela employees com 1 mapeador no warehouse /user/hive/warehouse/db_format
  - Formato
    - --as-textfile (padrão)
    - --as-parquetfile
    - ---as-avrodatafile
    - --as-sequencefile
  - Compactação
    - sem compactação (Padrão)
    - --compress (Padrão Gzip)
    - --compress --compression-codec org.apache.hadoop.io.compress.GzipCodec
    - --compress --compression-codec org.apache.hadoop.io.compress.BZip2Codec
    - --compress --compression-codec org.apache.hadoop.io.compress.SnappyCodec
    - --compress --compression-codec org.apache.hadoop.io.compress.DeflateCodec
    - --compress --compression-codec com.hadoop.compression.lzo.LzoCodec

Jobs

Semantix

Sqoop Job

- Salvar os comandos de importação e exportação
- Especifica parâmetros para identificar e recuperar o job salvo
- Importação ou exportação incremental
  - Importar/exportar as linhas atualizadas da tabela do RDBMS/HDFS.
- Comando
  - $ sqoop job --<atributo>

Atributos Job

- --create <job-id>: Criar job
  - $ sqoop **job --create myjob** --import --connect jdbc:mysql://database/db \
    --username root –password secret --table employee --m 1
- --list: Verificar jobs salvos
  - $ sqoop job --list
- --show <job-id>: Ver detalhes do job
  - $ sqoop job --show myjob
- --exe <job-id>: Executar job
  - $ sqoop job --exec myjob
- --delete <job-id>: Deletar jov
  - $ sqoop job --delete myjob

Carga Incremental

Semantix

Carga Incremental - Append

- Anexar dados em um conjunto de dados existentes no HDFS
  - Anexar todos os dados
    - $ sqoop import … **--append** --where ‘id_venda >10’
  - Anexar apenas os novos dados (Incremental)
    - Sem sobrescrever os dados, em relação à uma coluna e um valor exclusivo crescente
    - $ sqoop impot … **--incremental** append \
      --check-column id_venda \
      --last-value 50

Carga Incremental - Lastmodified

- Inserir dados em um conjunto de dados existentes no HDFS
  - ~~Anexar~~ Atualizar apenas os novos dados (Incremental)
  - Sobrescrever os dados, em relação à uma coluna e um valor de data e hora
    - Adicionar o atributo –merge-key
      - Coluna e um valor exclusivo crescente
    - $ sqoop import … --incremental lastmodified \
      --merge-key data_id \
      --check-column data_venda \
      --last-value ‘2021-01-18'

Laboratório

Resolução de Exercícios

Semantix

Exercícios Sqoop – Importação BD Sakila – Carga Incremental

MySQL

1. Criar a tabela cp_rental_append, contendo a cópia da tabela rental com os campos rental_id e rental_date
2. Criar a tabela cp_rental_id e cp_rental_date, contendo a cópia da tabela cp_rental_append

Sqoop - Realizar as importações no warehouse /user/hive/warehouse/db_test e visualizar no HDFS

3. Importar as tabelas cp_rental_append, cp_rental_id e cp_rental_date com 1 mapeador no formato parquet e compressão snappy
4. Mysql: Executar o sql /db-sql/sakila/insert_rental.sql no container do database
5. Atualizar a tabela cp_rental_append no HDFS anexando os novos arquivos
6. Atualizar a tabela cp_rental_id no HDFS de acordo com o último registro de rental_id
7. Atualizar a tabela cp_rental_date no HDFS de acordo com o último registro de rental_date

Importar Dados no Hive

Semantix

Padrão Tabela Hive

- Caminho padrão das Tabelas Hive
  - /user/hive/warehouse/
- Bom desempenho
  - Formato Parquet
  - Compressão Snappy

Tipo Coluna

- Sqoop é pré-configurado para mapear a maioria dos tipos SQL para Java ou Hive
  - --map-column-java: Mapeamento para java
    - $ sqoop import ... --map-column-java id=String,value=Integer
  - --map-column-hive: Mapeamento para hive
    - $ sqoop import ... --map-column-hive id=String,value=Integer

Importação Tabela Hive

sqoop import ...

- --hive-import
  - Importar tabela para o Hive
- --hive-overwrite
  - Sobrescrever os dados se a tabela hive existir
- --create-hive-table
  - O job irá falhar se uma tabela hive existir
- --hive-table
  - Especificar o nome da tabela hive
  - Comando:
    - -- hive-table <db_name>. <table_name>

Exemplo Imporação Tabela Hive

- $ sqoop import --table employees \
  --connect jdbc:mysql://database/employees \
  --username=root \
  --password=secrect \
  --warehouse-dir=/user/hive/warehouse/teste.db \
  --hive-import \
  --create-hive-table \
  --hive-table teste.user

Exportar Dados

Semantix

Exportar Dados do HDFS para o RDBMS

- Exportar
  - Qual diretório do HDFS
  - Qual JDBC?
  - Qual usuário e senha?
  - Qual database?
  - Quais tabelas?
  - Quais dados?
- Comando
  - export
  - ex
    - $ sqoop **export** --connect jdbc:mysql://database/log \
      --username root --password secret ...

Comandos Básicos Exportação

- Definir o diretório de leitura no HDFS
  - --export-dir <diretório>
- Definir o nome da Tabela no SGBD
  - --table <nome_tabela>
- Opção de Atualização:
  - --update-mode
    - updateonly (default)
      - Acrescenta novas linhas na tabela
      - Cada registro de entrada é transformado em um INSERT
    - allowinsert
      - Atualizar as linhas se existirem na tabela
      - Inserir linhas se não existirem na tabela

Exportar Dados

- A tabela precisa ser criada no SGBD antes da exportação do Sqoop

  - MySQL: create table product_recommendations(...)

  - sqoop **export** \
    --connect jdbc:mysql://database/employees\
    --username root --password secrect \
    --export-dir /user/root/recommender_output \
    --update-mode allowinsert \
    --table product_recommendations

Laboratório

Resolução de Exercícios

Semantix

Exercícios Sqoop – Importação para o Hive e Exportação - BD Employees

1. Importar a tabela employees.titles do MySQL para o diretório /user/aluno/<nome>/data com 1 mapeador.
2. Importar a tabela employees.titles do MySQL para uma **tabela Hive** no banco de dados seu nome com 1 mapeador.
3. Selecionar os 10 primeiros registros da tabela titles no **Hive**.
4. Deletar os registros da tabela employees.titles do MySQL e verificar se foram apagados, através do Sqoop
5. Exportar os dados do diretório /user/hive/warehouse/<nome>.db/data/titles para a tabela do MySQL employees.titles.
6. Selecionar os 10 primeiros registros registros da tabela employees.titles do MySQL.

Semantix

Obrigado!

Alguma pergunta?

Você pode me encontrar em:

rodrigo.augusto@semantix.com.br

GET SMARTER