Semantix

Sqoop - Básico

Aula 6

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

Sqoop

Semantix

Ingestão de Dados

- Processo de Enviar/Receber os dados locais para o sistema distribuído
- Data Lake
  - Batch
    - Sqoop
  - Stream
    - Flume
    - Katka

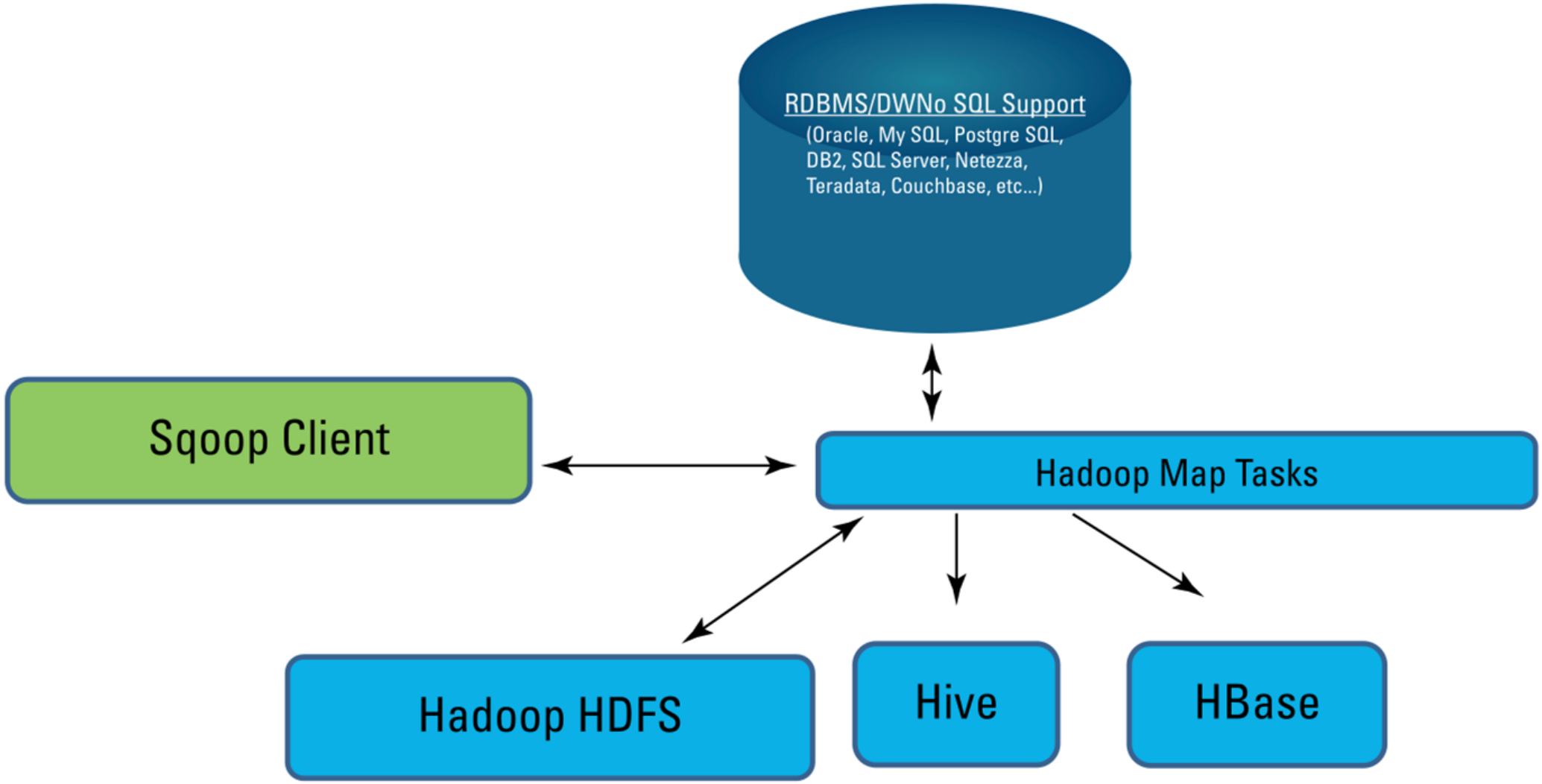
Apache Sqoop

- Ferramenta para transferir dados entre o Hadoop e banco de dados relacionais ou mainframes

Sqoop = “SQL to Hadoop”

- Importar dados
  - Banco de dados relacional (RDBMS)
    - MySQL, SQL Server, Oracle
  - Para o HDFS, Hive ou HBase
    - Transformar dados em MapReduce
      - Execução em paralelo
      - Tolerância a falhas
- Exportar dados
  - Armazenamento do Hadoop para um RDBMS

Funcionamento Sqoop



Hadoop For Dummies, 2014

Semantix

Laboratório – Preparar BD

Resolução de Exercícios

Semantix

Exercícios Verificar e Instalar os Banco de Dados de testes

- Copiar os dados do local para o contêiner database

$ docker cp input/exercises-data/db-sql/ database:/

- Acessar o contêiner database

$ docker exec -it database bash

- Instalar Banco de Dados de testes
  - Diretório /db-sql - BD employees (Já existe)

    $ cd /db-sql

    $ mysql -psecret < employees.sql
  - Diretório /db-sql/sakila - BD sakila

    $ cd /db-sql/sakila/

    $ mysql -psecret < sakila-mv-schema.sql

    $ mysql -psecret < sakila-mv-data.sql

Comandos Sqoop

Semantix

Comandos Sqoop

- $ sqoop <comando>
  - help
  - version
  - import
  - import-all-tables
  - export
  - validation
  - job
  - metastore
  - merge
  - codegen
  - create-hive-table
  - eval
  - list-databases
  - list-tables

Comandos Básicos

- Verificar a versão
  - $ sqoop version
- Listar todos os comandos
  - $ sqoop help
- Ajuda de um comando
  - $ sqoop help import
    - --connect, --connect-manager, --driver, --hadoop-mapred-home, --help, --password-file, -P, --password, --username, --verbose, ...
- Importar todas as tabelas de um banco de dados
  - $ sqoop import-all-tables ...

Acessar o Sqoop

Semantix

Listar Banco de Dados e Tabelas

Semantix

Conexão Banco de Dados

- Informações para conexão
  - Database type (MySQL, Oracle etc)
  - Hostname
  - Port number
  - Database Name (list-databases)

- Parâmetros
  - --connect <conexão> \
    --username usuario \
    --password senha

DB	String de conexão
HSQLDB	jdbc:hsqldb:*//
MySQL	jdbc:mysql://
Oracle	jdbc:oracle:*//
PostgreSQL	jdbc:postgresql://
CUBRID	jdbc:cubrid:*

Listar BD e Tabelas

- Listar Banco de Dados
  - $ sqoop list-databases \
    --connect jdbc:mysql://database \
    --username usuario \
    --password senha

- Listar tabelas
  - $ sqoop list-tables \
    --connect jdbc:mysql://database/employees \
    --username usuario \
    --password senha

Consultar Tabelas

Semantix

Consultas Tabelas

- Executar consultas em bancos de dados remotos
  - Usar para carregar ou consultar tabelas de log
    - Execução de tarefas do Sqoop
- Comando
  - eval
  - Ex.

```
$ sqoop eval \
--connect jdbc:mysql://database/employees \
--username=root \
--password=secret \
--query "SELECT * FROM employees limit 15"
```

Exemplos Consulta

- Criar tabela
  - $ sqoop eval --connect … --query "create table setor(cod int(2), name varchar(30))"
  - $ sqoop eval --connect ... --query "describe setor“
- Inserir linhas na tabela:
  - $ sqoop eval --connect … --query "insert into setor values(1,'vendas')"
- Consultar tabela
  - sqoop eval --connect … --query "select * from setor"
  - sqoop eval --connect … --query "select * from employees where first_name like ‘A’"

Laboratório

Resolução de Exercícios

Semantix

Exercícios Sqoop – Pesquisa e Criação de Tabelas

1. Mostrar todos os databases
2. Mostrar todas as tabelas do bd employees
3. Inserir os valores ('d010', 'BI') na tabela departments do bd employees
4. Pesquisar todos os registros da tabela departments
5. Criar a tabela benefits(cod int(2) AUTO_INCREMENT PRIMARY KEY, name varchar(30)) no bd employees
6. Inserir os valores (null,'food vale') na tabela benefits
7. Pesquisar todos os registros da tabela benefits

Importar Dados

Semantix

Importar Dados do RDBMS para o HDFS

- Importar
  - Qual JDBC?
  - Qual usuário e senha?
  - Qual database?
  - Quais tabelas?
  - Quais dados?
- Comando
  - import
  - Ex
    - $ sqoop import --connect jdbc:mysql://database \
      --username root –password secret

Importar tabela, coluna e linha

- --table: Importar de apenas uma tabela
  - $ sqoop import --table employees \
    --connect jdbc:mysql://database/employees \
    --username root \
    --password secret

- --columns: Importar colunas específicas
  - $ sqoop import ... --columns "id,last_name"

- --where: Importar linhas correspondentes
  - $ sqoop import ... --where "state='SP'"

Importar Dados - Diretórios

Semantix

Armazenar Diretório Diferente

- Por padrão, o Sqoop armazena os dados no diretório home do HDFS
  - ex. /user/<username>/<tablename>
- --target-dir: Armazenar em um diretório específico
  - $ sqoop import … --target-dir /user/root/db
- --warehouse-dir: Armazenar em um diretório base
  - $ sqoop import … --warehouse-dir /user/root/db

- Diferença
  - Importar tabela departments
    - --target-dir /data = /data
    - --warehouse-dir /data = /data/departments

Armazenar Diretório Existente

- Por padrão, o Sqoop falha a importação se o diretório de destino já existir

- -delete-target-dir: Sobrescrever o diretório
  - $ sqoop import … --warehouse-dir /user/cloudera/db -delete-target-dir

- -append: Anexar os dados no diretório existente
  - $ sqoop import ... --warehouse-dir /user/cloudera/db -append

Importar Dados - Delimitadores

Semantix

Diferentes Delimitadores

- Por padrão, o Sqoop gera arquivos de texto
  - Campos delimitados por vírgula
  - Linhas terminadas por quebra de linha \n
- Comandos
  - --fields-terminated-by <delimitador>
  - --lines-terminated-by <delimitador>
  - Especificar o delimitador
    - ‘qualquer coisa’, \b (backspace), \n (newline), \t (tab), \0 (NUL)
  - Ex.

    $ sqoop import … --fields-terminated-by ’\t’ --lines-terminated-by ‘&’

Laboratório

Resolução de Exercícios

Semantix

Exercícios Sqoop – Importação BD Employees

1. Pesquisar os 10 primeiros registros da tabela employees do banco de dados employees
2. Realizar as importações referentes a tabela employees e para validar cada questão, é necessário visualizar no HDFS*
   A. Importar a tabela employees, no warehouse /user/hive/warehouse/db_test_a
   B. Importar todos os funcionários do gênero masculino, no warehouse /user/hive/warehouse/db_test_b
   C. importar o primeiro e o último nome dos funcionários com os campos separados por tabulação, no warehouse /user/hive/warehouse/db_test_c
   D. Importar o primeiro e o último nome dos funcionários com as linhas separadas por " : " e salvar no mesmo diretório da questão 2.C

* Dica para visualizar no HDFS:

```
$ hdfs dfs -cat /.../db_test/nomeTabela/part-m-00000 | head -n 5
```

Semantix

Obrigado!

Alguma pergunta?

Você pode me encontrar em:

rodrigo.augusto@semantix.com.br

GET SMARTER