Sqoop理论与代码文件

Sqoop概述

- 功能
  - 主要用于在Hadoop(Hive)与传统的数据库(mysql、postgresql...)间进行数据的传递，可以将一个关系型数据库（例如： MySQL ,Oracle ,Postgres等）中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。
- 说明
  - Sqoop1与2的区别
    - Sqoop1.x
      - Sqoop1.x中，仅仅使用了一个Sqoop客户端，它是单用户的、架构部署简单。客户端发送命令到Sqoop，Sqoop转换为MapReduce作业运行在Hadoop集群环境上，从而实现RDBMS和Hadoop之间相互导入导出。Sqoop1.x只一个mapreduce作业，只有map没有reduce。
    - Sqoop2.x
      - Sqoop2.x 中，引入了sqoop server集中化管理Connector，支持多种交互方式：命令行、Web UI、Rest API，所有的链接安装在sqoop server上，完善了权限管理机制（可配置管理员、使用者等角色），Connector规范化（不再包含数据传输，格式转换、与Hive、Hbase交互等功能仅负责数据读写）。Sqoop2.x中的MapReduce作业既有Map也有Reduce。
    - 图示

功能	Sqoop1	Sqoop2
客户端连接RDBMS的数据源	支持	不支持 解决方法：使用已安装了数据库上执行语句的机器作为客户端，安装Sqoop客户端。 Postgresql、Netezza、DB2 Connector：此类数据库官方不提供 JDBC 要求的数据库上运行，导致：导致不能在 Sqoop 所在节点安装
Kerberos安全集成	支持	不支持
直接从RDBMS导入到Hive表	支持	不支持 解决方法：先导入到HDFS，再创建Hive表，LOAD到Hive 中。Sqoop 1 中是直接创建 Hive 表，LOAD DATA 进行。，并且使用直接入 Hive 的 Hbase
直接从Hive导入到RDBMS	不支持 解决方法：先导出文件到HDFS，再执行 sqoop export 命令。Sqoop 1 中是直接读取 Hive 表的 HDFS 文件，并导入 RDBMS。Sqoop 2 中，可以直接从 Hive 表导入 RDBMS。	不支持 同样有 Sqoop 1 现存的解决方法选项

- 原理
  - 将导入或导出命令翻译成mapreduce程序来实现
  - 在翻译出的mapreduce中主要是对inputformat和outputformat进行定制

Sqoop使用案例

- 导入数据
  - RDBMS到HDFS
    - 在mysql中建表
      - mysql -uroot -p000000
      - mysql> create database company;
      - mysql> create table company.staff(id int(4) primary key not null auto_increment, name varchar(255), sex varchar(255));
      - mysql> insert into company.staff(name, sex) values('Thomas', 'Male');
      - mysql> insert into company.staff(name, sex) values('Catalina', 'FeMale');
    - 导入数据
      - bin/sqoop import \
      - --connect jdbc:mysql://hadoop102:3306/company \
      - --username root \
      - --password 000000 \
      - --table staff \
      - --target-dir /user/company \
      - --delete-target-dir \
      - --num-mappers 1 \
      - --fields-terminated-by "\t"
    - 查询导入
      - bin/sqoop import \
      - --connect jdbc:mysql://hadoop102:3306/company \
      - --username root \
      - --password 000000 \
      - --target-dir /user/company \
      - --delete-target-dir \
      - --num-mappers 1 \
      - --fields-terminated-by "\t" \
      - --query 'select name,sex from staff where id <=1 and $CONDITIONS;'
    - 使用关键字查询导入数据
      - bin/sqoop import \
      - --connect jdbc:mysql://hadoop102:3306/company \
      - --username root \
      - --password 000000 \
      - --target-dir /user/company \
      - --delete-target-dir \
      - --num-mappers 1 \
      - --fields-terminated-by "\t" \
      - --table staff \
      - --where "id=1"
  - RDBMS到Hive
    - bin/sqoop import \
    - --connect jdbc:mysql://hadoop102:3306/company \
    - --username root \
    - --password 000000 \
    - --table staff \
    - --num-mappers 1 \
    - --hive-import \
    - --fields-terminated-by "\t" \
    - --hive-overwrite \
    - --hive-table staff_hive
  - RDBMS到HBase
    - bin/sqoop import \
    - --connect jdbc:mysql://hadoop102:3306/company \
    - --username root \
    - --password 000000 \
    - --table company \
    - --columns "id,name,sex" \
    - --column-family "info" \
    - --hbase-create-table \
    - --hbase-row-key "id" \
    - --hbase-table "hbase_company" \
    - --num-mappers 1 \
    - --split-by id
- 导出数据
  - HDFS/HIVE到RDBMS
    - bin/sqoop export \
    - --connect jdbc:mysql://hadoop102:3306/company \
    - --username root \
    - --password 000000 \
    - --table staff \
    - --num-mappers 1 \
    - --export-dir /user/hive/warehouse/staff_hive \
    - --input-fields-terminated-by "\t"
- 脚本打包
  - mkdir opt
  - vim opt/job_HDFS2RDBMS.opt
    - export
    - --connect
    - jdbc:mysql://hadoop102:3306/company
    - --username
    - root
    - --password
    - 000000
    - --table
    - staff
    - --num-mappers
    - 1
    - --export-dir
    - /user/hive/warehouse/staff_hive
    - --input-fields-terminated-by
    - "\t"
- 脚本使用
  - bin/sqoop --options-file opt/job_HDFS2RDBMS.opt

尚硅谷大数据研究院出品
www.atguigu.com