



Sqoop

Grishko Stanislav, stanislav.grishko@bigdatateam.org

Big Data Instructor, <http://bigdatateam.org/>

Data integration manager at X5 Retail Group



- ▶ Для чего нужен *Sqoop*?
- ▶ Import
 - ▶ Пример загрузки таблицы из базы в *Hive*
 - ▶ Как это устроено?
 - ▶ Задания
- ▶ Export
 - ▶ Пример загрузки таблицы из базы в *Hive*
 - ▶ Как это устроено?
 - ▶ Задания
- ▶ Разбор основных кейсов
- ▶ *Sqoop 2.0*



Как расшифровывается Sqoop?



Как расшифровывается Sqoop?

SQL + HADOOP

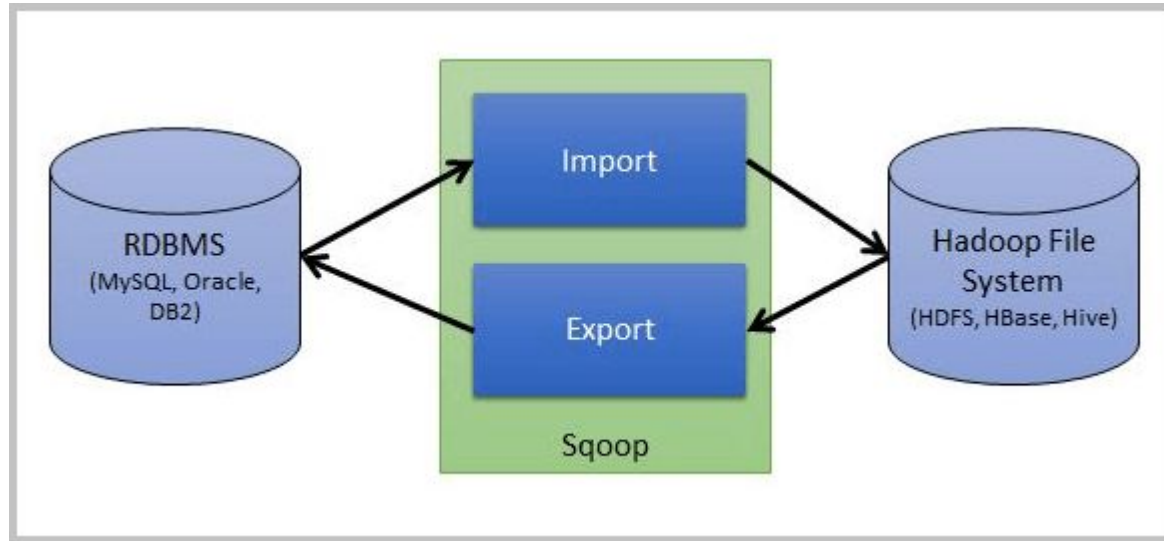


Для чего нужен Sqoop?



Для чего нужен *Sqoop*?

Создан для двунаправленной передачи данных между Hadoop и почти любым внешним структурированным хранилищем данных.





Import



Import

Вопрос: за какое минимальное количество действий можно загрузить данные таблицы из РСУБД в HDFS?



Sqoop import принцип работы

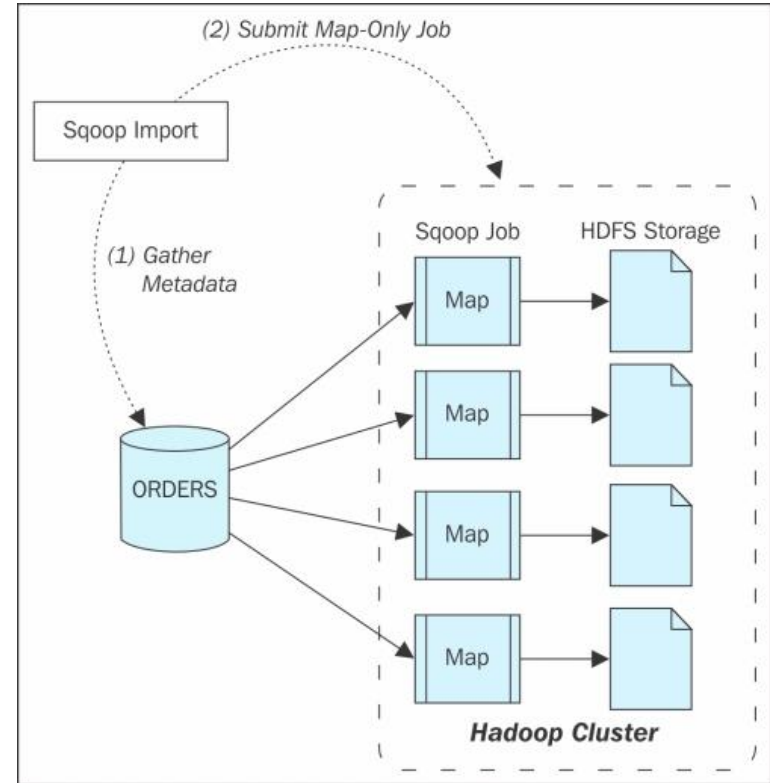


Запускаем sqoop import



Sqoop:

- ▶ Получает мета-информацию с базы;
- ▶ Создает на стороне клиента MapReduce.java файл;
- ▶ Запускает задачу выгрузки данных на узлах кластера;
- ▶ Полученные данные складывает в hdfs;



*



```
sqoop import \  
--connect "jdbc:mysql://some-host:3306/database_name" \  
--username user \  
--password pass \  
--table table_name_in_DB \  
--target-dir /user/data
```




```
sqoop import \  
--connect "jdbc:mysql://some-host:3306/database_name" \  
--username user \  
--password pass \  
--table table_name_in_DB \  
--target-dir /user/data \  
--hive-import \  
--null-string '\\N' \  
--null-non-string '\\N' \  
--hive-overwrite \  
--hive-table sqoop_db.person \  
--hive-drop-import-delims
```



Скрипт sqoop import для загрузки в Hive

```
sqoop import \  
--connect "jdbc:mysql://some-host:3306/database_name" \  
--username user \  
--password pass \  
--table table_name_in_DB \  
--target-dir /user/data  
--hive-import \  
--null-string '\\N' \  
--null-non-string '\\N' \  
--hive-overwrite \  
--hive-table sqoop_db.person \  
--hive-drop-import-delims
```





Задание 1

Выгрузить через `sqoop-import` таблицу (person) в директорию HDFS (/user/<пользователь>/sqoop_person_data)



Задание 2

Выгрузить через `sqoop-import` таблицу (`person_details`) в директорию HDFS (`/user/<пользователь>/sqoop_person_data_details`) и создать таблицу в hive (`sqoop_db.person_<ваши инициалы>`)



Export



Export

Вопрос: за какое минимальное количество действий можно загрузить данные таблицы из HDFS в РСУБД?



Sqoop export принцип работы

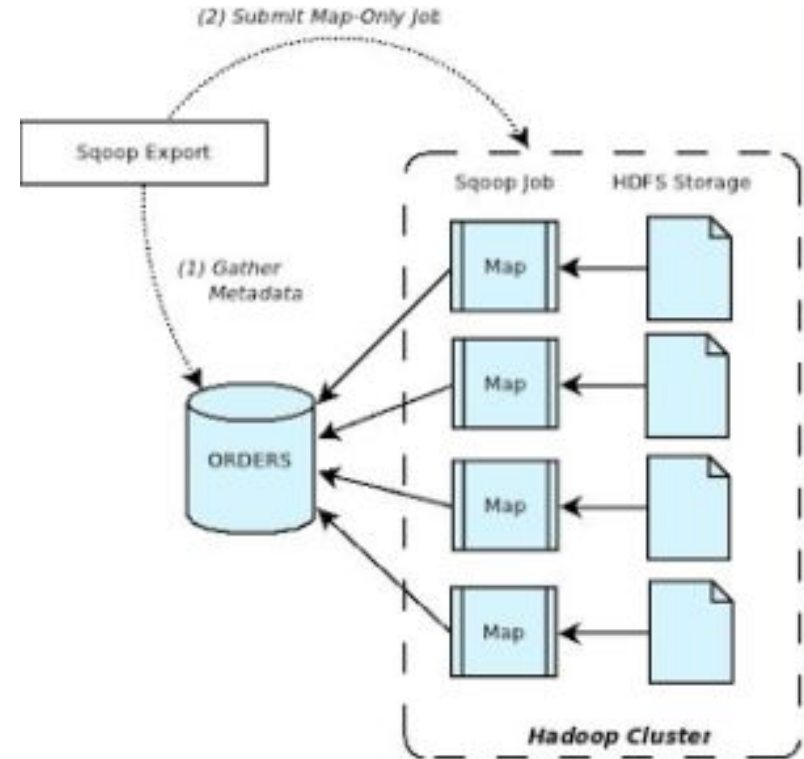


Запускаем sqoop export



Sqoop:

- ▶ Получает мета-информацию с базы;
- ▶ Создает на стороне клиента MapReduce.java файл;
- ▶ Запускает задачу выгрузки данных на узлах кластера;
- ▶ Полученные данные добавляет в РСУБД





```
sqoop export \  
--connect "jdbc:mysql://some-host:3306/database_name" \  
--username user \  
--password pass \  
--table table_name_in_DB \  
--export-dir /user/data/part-m-00000
```



Скрипты sqoop import/export

```
sqoop import \  
--connect "jdbc:****" \  
--username user \  
--password pass \  
--table table_name_in_DB \  
--target-dir /user/data \  
--num-mappers 1 \  
--fields-terminated-by '\'  
--hive-import \  
--null-string '\\N'\  
--null-non-string '\\N'\  
--hive-overwrite \  
--hive-table sqoop_db.person_export \  
--hive-drop-import-delims
```

```
sqoop export \  
--connect "jdbc:****" \  
--username user \  
--password pass \  
--table table_name_in_DB \  
--num-mappers 1 \  
--null-string '\\N'\  
--null-non-string '\\N'\  
--input-fields-terminated-by ','\  
--export-dir /user/data
```



Задание

- 1) Выгрузить через `sqoop-import` таблицу (`person_export`) в директорию HDFS (`/user/<пользователь>/sqoop_person_data_details`) и **создать** таблицу в `hive` (`sqoop_db.person_<ваши инициалы>`)
- 2) Загрузить данные через `sqoop-export` в таблицу (`person_export_from_hdfs`)



Разбор основных кейсов работы со Sqoop



Вместо импорта всей таблицы необходимо передать только подмножество строк на основе различных условий, которые можно выразить в виде инструкции SQL с предложением WHERE.

`--where "person_id = 14"`



Вместо импорта всей таблицы необходимо передать только подмножество строк на основе уникального SQL запроса.

```
--query 'SELECT <перечисление всех нужных полей> FROM person p \\  
LEFT JOIN person_details pd on pd.person_id = p.id WHERE $CONDITIONS' \
```



Ввод пароля в интерфейс командной строки небезопасен. Он может быть легко извлечен из списка запущенных процессов операционной системы.

-P

или

--password-file my-sqoop-password



Разделенный вкладками CSV-файл, который Sqoop использует по умолчанию, по каким-либо причинам не подходит для вашего варианта использования. Вы предпочитаете двоичный формат по сравнению с обычным текстом.

--as-sequencefile

--as-avrodatafile



Сопоставление типов по умолчанию, которое Sqoop предоставляет между реляционными базами данных и Hadoop, обычно работает хорошо. У вас есть варианты использования, требующие переопределения сопоставления.

```
--map-column-java id=Long
```

или

```
--map-column-java f1=Float,f2=String,f3=String
```



Sqoop кодирует значения NULL базы данных с помощью строковой константы null. Ваша последующая обработка (запросы Hive, пользовательское задание MapReduce или скрипт Pig) использует другую константу для кодирования пропущенных значений. Вы хотели бы переопределить значение по умолчанию.

```
--null-string '\N'  
--null-non-string '\N'
```



Вы хотите импортировать все таблицы из своей базы данных сразу, используя одну команду, а не импортировать таблицы по одному.

```
sqoop import-all-tables \  
--connect jdbc:mysql://mysql.example.com/sqoop \  
--username sqoop \  
--password sqoop \  
--exclude-tables cities,countries
```



Вы ранее экспортировали данные из Hadoop, после чего вы запустили дополнительную обработку, которая изменила его. Вместо того, чтобы стирать существующие данные из базы данных, вы предпочитаете просто обновлять любые измененные строки.

```
--update-key person_id
```



У вас есть данные в Hadoop, которые вам нужно экспортировать. К сожалению, соответствующая таблица в вашей базе данных содержит больше столбцов, чем данные HDFS.

```
--columns country,email,person_id
```



Sqoop 2.0



Sqoop 2 is being deprecated. Customers are advised to use Sqoop1 instead.*



- ▶ Для чего нужен *Sqoop*?
- ▶ Как работает *Sqoop import*?
- ▶ Как работает *Sqoop export*?
- ▶ Что нужно использовать для загрузки данных *Sqoop 2.0* или *Sqoop 1.0*?



Thank you! Questions?

Stanislav Grishko

BigData Instructor, <http://bigdatateam.org/>

Data integration manager at X5 Retail Group