Hadoop Data ingestion Workshop - Keys

Apache Sqoop

Exercise 1:

```
hdfs dfs -du -h /user/cloudera/sqoop-mysql/employees
```

Exercise 2:

```
-- hive
create database emp_mysql;


-- Sqoop
sqoop import --connect jdbc:mysql://localhost/employees --username
cloudera --password cloudera --table salaries --direct -m 1 --hive-
import --create-hive-table --hive-table emp_mysql.salaries


-- Hive
show tables in emp_mysql;

select * from emp_mysql.salaries;


-- hdfs

hdfs dfs -ls -R /user/hive/warehouse/emp_mysql.db

hdfs dfs -cat /user/hive/warehouse/emp_mysql.db/salaries/part-m-
00000 | wc -l
```

Exercise 3:

```
mysql –u root -p

-- Mysql

select gender, count(*) from employees group by gender;

-- Sqoop

sqoop import --connect jdbc:mysql://localhost/employees --username
cloudera --password cloudera --query 'select
```

```
emp_no,birth_date,first_name,last_name,hire_date from employees
where gender="M" AND $CONDITIONS' --direct -m 1 --split-by emp_no --
hive-import --create-hive-table --hive-table
emp_mysql.employees_part --hive-partition-key gender --hive-
partition-value 'M' --target-dir
'/user/hive/warehouse/emp_mysql.db/employees_part/gender=M'


-- hdfs

hdfs dfs -ls -R /user/hive/warehouse/emp_mysql.db

-- Hive

select gender, count(*) from employees_part group by gender;
```

Exercise 4:

```
-- Mysql

CREATE TABLE stations_hive_export(code varchar(6),name
varchar(50),latitude DOUBLE,longitude DOUBLE);


-- Sqoop

sqoop export --connect jdbc:mysql://localhost/employees --username
cloudera --password cloudera --table stations_export_hive --direct -
-export-dir /user/hive/warehouse/stations


-- Mysql

select count(*) from stations_export_hive;


where (year=2015 and month=03) limit 10;
```

Exercise 5:

```
sqoop import --connect jdbc:mysql://localhost/employees --username
cloudera -password cloudera --table salaries -m 1 --target-dir
/sqoop-avro-import --as-avrodatafile --compression-codec snappy


hdfs dfs -cat /sqoop-avro-import/part-m-00000.avro | head --bytes
10K > sample_file.avro
```

```
avro-tools getschema sample_file.avro > salary.avsc


hdfs dfs -put salary.avsc /tmp


CREATE EXTERNAL TABLE avro_snappy_salaries_table

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.avro.AvroSerDe'

STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat'

OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat'

LOCATION '/sqoop-avro-import/'

TBLPROPERTIES ('avro.schema.url' = 'hdfs:///tmp/salary.avsc');


select * FROM avro_snappy_salaries_table LIMIT 10;
```