Assessment
Data at Scale
YCBS-257

Assignment 2

30% due on 23:59 Sunday Feb 24 2019

Hadoop Data Ingestion

Apahe Sqoop is a tool used to extract data from structured data sources into Hadoop. It can be used to transfer huge data between Hadoop and other relational database systems. We can import the data from Relational databases such as MySQL, Oracle into Hadoop or HBase for processing purpose and after completing processing we can again export the data back into RDBMS. The processing can be done with Map-Reduce programs or Hive.

Apache Flume acts as a system used to write data to Apache Hadoop and Apache HBase in a reliable and scalable fashion and we can write data into it in any format and language such as MapReduce, Hive, Pig, and Impala.

In this assignment, we are going to ingest data from different sources into different destinations.

To start, run your Cloudera virtual machine. You will use Linux command-line to interact with Sqoop, Flume and CentOS operating system.

You are highly recommended to read the online documentation while doing this assignment:

https://sqoop.apache.org/docs/1.4.6/SqoopUserGuide.html
https://flume.apache.org/FlumeUserGuide.html
http://hbase.apache.org/0.94/book/ops_mgt.html

Provided files:

The employees sample database, installed while working the last workshop.

Sqoop Data import 50%

You will use Apache Sqoop to import data from MySQL server into Apache Hive and Apache HBase.

Question 1:

Write a Sqoop command to import the **employees** table from MySQL employees database into a Hive partition. This partition should contains only rows with 'gender' = 'F' (female) from the original table. This partition belongs to an existing Hive managed table employees_part in the emp_mysql database.

Note: (employees_part table was created while working on the workshop and already contains a Hive partition with 'gender'='M')

Question 2:

Write a HiveQL script to count the rows in each partition of employees_part

Question 3:

Write a Sqoop command to import the employees table from MySQL employees database into HBase.

The Sqoop command should create a new HBase table employees_hbase, column family infos and use emp_no as rowkey

Flume Data ingestion 50%

Fan out is the process of delivering events form one source to multiple sinks through multiple channels. We have 2 modes for fan out, they are replicating and multiplexing. In the replicating flow, the event is sent to all the configured channels. In the multiplexing flow, the event is sent to only a subset of channels.

Question 1:

The data source is a financial stream. We want to send data to a particular HDFS directory based on the currency symbol (EUR, USD).
You will simulate the data source using telnet by typing the currency symbol manually.

Write a Flume agent configuration file that:

- Collect data from a netcat source
- Event with 'EUR' symbol is routed to the 'EUR' directory on HDSF
- Event with 'USD' symbol is routed to the 'USD' directory on HDFS
- All other events are routed to the 'GNL' directory on HDFS

Question 2:

Write the Flume command to run the agent

What to submit:

One MS Word file <your name> - W2019A2.docx

The file should containing two separated and named sections

- Section one: Sqoop commands
- Section two: Flume agent configuration

✓ Comments can be submitted in French or in English in the text file