Datalab - Docker

Simone Perego 807209

May 26, 2020

Layer Analytics

docker-compose run -p 8888:8888 -p 9000:9000 -p 9866:9866 -p 10000:10000 -p 50070:50070 -p 50090:50090 –name hadoop-master analytics
Init Hadoop:
/usr/local/hadoop/sbin/start-all.sh
Create hdfs folder
hdfs dfs -mkdir -p ./dati
Go to /home and copy file into hdfs
hdfs dfs -put ft_document_en.out./dati/
hdfs dfs -put ft_skill_analysis_en.out ./dati/
hdfs dfs -put ft_skill_professional_relevance.out ./dati/

- **JUPYTER**

Go to / and then init jupyter
jupyter notebook –allow-root –no-browser –ip='*'
Go to browser ip:8080 and copy the token

- **HIVE**

Init metastore and hiveserver2
/usr/local/apache-hive-3.1.2-bin/bin/schematool -initSchema -dbType derby
/usr/local/apache-hive-3.1.2-bin/bin/hiveserver2
Open another shell and open hadoop-master container
docker exec -i -t hadoop-master /bin/bash
Init beeline
/usr/local/apache-hive-3.1.2-bin/bin/beeline
!connect jdbc:hive2://hadoop-master:10000/default
Now is possibile create database and create table to connect with

- Dbeaver
- Jupyter

Layer Processing

docker-compose run -p 2222:2222 -p 4040:4040 –name spark-worker processing
Create another folder into hdfs:
hdfs dfs -mkdir /user/vertica
hdfs dfs -mkdir /user/vertica/staging
Init Spark-Shell or PySpark
/usr/local/spark/bin/spark-shell
To load data
var ds = spark.read.parquet(“hdfs://hadoop-master:9000/user/root/dati/ft_document_en.out”)
Write data into hdfs
var ds1 = ds.select(”general_id”, ”country”).dropDuplicates().groupBy(”country”).agg(count(”general_id”) as ”numero”)
ds1.write.parquet(”hdfs://hadoop-master:9000/vertica/staging/NOMEFILE.parquet”)

Layer Presentation

docker-compose run -p 5433:5433 -p 5434:5434 -p 5450:5450 –name vertica-host presentation
Init Vertica DB:
/etc/bootstrap.sh
Open another shell and open vertica-host container
docker exec -i -t vertica-host /bin/bash
Use vsql client
./vsql -U dbdmin
Change Export Address (Modify 172.18.0.* with CONTAINER_IP)
CREATE SUBNET kv_subnet with ’172.18.0.0’;
ALTER DATABASE database EXPORT ON kv_subnet;
CREATE NETWORK INTERFACE kv_node2 on v_database_node0001 with ’172.18.0.*’;
ALTER NODE v_database_node0001 export on kv_node2;
Create table from hdfs
CREATE EXTERNAL TABLE prova3 (country VARCHAR, numero INT) AS COPY FROM ’hdfs://hadoop-master:9000/user/vertica/staging/NOMEFILE.parquet’ PARQUET;