Hive Queries

Enter 'Hive' in linux shell to enter in hive environment.

CREATE TABLE IF NOT EXISTS bitcoin_tweets.bitcoin_tweets (user_name String, user_created String, text String) COMMENT 'Bitcoin tweets' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';

CREATE TABLE IF NOT EXISTS bitcoin_tweets.bag_of_words (word String) COMMENT 'Bag of words' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\n' ;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/bag_of_words.csv' INTO TABLE bag_of_words;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00000' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00001' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00002' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00003' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00004' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00005' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00006' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00007' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00008' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00009' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00010' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00011' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00012' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00013' INTO TABLE bitcoin_tweets.bitcoin_tweets;

LOAD DATA INPATH 'gs://dataproc-staging-us-central1-744395099175-nyv31aa8/bitcoin-tweets/pig_Output/part-m-00014' INTO TABLE bitcoin_tweets.bitcoin_tweets;

CREATE TABLE IF NOT EXISTS bitcoin_tweets.spam_data (user_name String, user_created String, text String) COMMENT 'Spam dataset' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';

insert into spam_data select user_name,user_created,text from bitcoin_tweets bt JOIN bag_of_words bw ON (upper(bt.text) like upper(concat('%',bw.word,'%')));

CREATE TABLE IF NOT EXISTS bitcoin_tweets.ham_data (user_name String, user_created String, text String) COMMENT 'Ham dataset' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';

insert into ham_data select user_name, user_created, text from bitcoin_tweets where text not in (select text from spam_data);

select user_name,count(user_name) from ham_data group by user_name,user_created order by count(user_name) desc limit 10;

select user_name,count(user_name) from spam_data group by user_name,user_created order by count(user_name) desc limit 10;

CREATE TABLE IF NOT EXISTS bitcoin_tweets.ham_data_top10 (user_name String, user_created String, text String) COMMENT 'Ham dataset top 10' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';

CREATE TABLE IF NOT EXISTS bitcoin_tweets.spam_data_top10 (user_name String, user_created String, text String) COMMENT 'Spam dataset top 10' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';

insert into ham_data_top10 select * from ham_data where user_name in (select user_name from ham_data group by user_name,user_created order by count(user_name) desc limit 10);

insert into spam_data_top10 select * from spam_data where user_name in (select user_name from spam_data group by user_name,user_created order by count(user_name) desc limit 10);

CREATE TABLE spam_csv_export_data ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED as textfile AS select 'user_name' as user_name, 'user_created' as user_created, 'text' as text;

CREATE TABLE ham_csv_export_data ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED as textfile AS select 'user_name' as user_name, 'user_created' as user_created, 'text' as text;

insert into ham_csv_export_data select user_name,user_created,text from ham_data_top10;

insert into spam_csv_export_data select user_name,user_created,text from spam_data_top10;

quit;

Enter 'hadoop fs -cat 'hdfs://assignment-1-cluster-m/user/hive/warehouse/bitcoin_tweets.db/spam_csv_export_data/*' > ~/spam_hive_output.csv' in linux command line.

Enter 'hadoop fs -cat 'hdfs://assignment-1-cluster-m/user/hive/warehouse/bitcoin_tweets.db/ham_csv_export_data/*' > ~/ham_hive_output.csv' in linux command line.