Big Data on AWS v3.0 - Lab 4: Processing Server Logs with Hive on Amazon EMR

==================================================================================================================

Using this command reference.

==================================================================================================================

1. Locate the step you need.

2. Replace items in angle brackets - < > - with appropriate values. For example, in this command you would replace the value - <JobFlowID> - (including the angle brackets) with the parameter indicated in the lab instructions:

elastic-mapreduce --list <JobFlowID>. You can also use find and replace to change bracketed parameters in bulk.

3. Do NOT enable the Word Wrap feature in Windows Notepad or the text editor you use to view this file.

==================================================================================================================

==================================================================================================================

Task 3: Running Hive Interactively

==================================================================================================================

+++ Step 3.1.2

sudo chown hadoop -R /var/log/hive

mkdir /var/log/hive/user

mkdir /var/log/hive/user/hadoop

+++ Step 3.1.3

hive -d SAMPLE=s3://us-west-2-aws-training/awsu-ilt/AWS-200-BIG/v3.0/data/lab-4-hive/data -d DAY=2009-04-13 -d HOUR=08 -d NEXT\_DAY=2009-04-13 -d NEXT\_HOUR=09 -d OUTPUT=s3://hive-bucket<YourInitials>/output/

+++ Step 3.1.4

ADD JAR ${SAMPLE}/libs/jsonserde.jar;

+++ Step 3.2.1

CREATE EXTERNAL TABLE impressions (requestBeginTime string, adId string, impressionId string, referrer string, userAgent string, userCookie string, ip string) PARTITIONED BY (dt string) ROW FORMAT serde 'com.amazon.elasticmapreduce.JsonSerde' with serdeproperties ('paths'='requestBeginTime, adId, impressionId, referrer, userAgent, userCookie, ip') LOCATION '${SAMPLE}/tables/impressions';

+++ Step 3.2.2

ALTER TABLE impressions ADD PARTITION (dt='2009-04-13-08-05');

+++ Step 3.2.3

MSCK REPAIR TABLE impressions;

+++ Step 3.2.4

CREATE EXTERNAL TABLE clicks (impressionId string) PARTITIONED BY (dt string) ROW FORMAT SERDE 'com.amazon.elasticmapreduce.JsonSerde' WITH SERDEPROPERTIES ('paths'='impressionId') LOCATION '${SAMPLE}/tables/clicks';

+++ Step 3.2.5

MSCK REPAIR TABLE clicks;

+++ Step 3.2.6

show tables;

==================================================================================================================

Task 4: Joining Tables Using Hive

==================================================================================================================

+++ Step 4.1.1

CREATE EXTERNAL TABLE joined\_impressions (requestBeginTime string, adId string, impressionId string, referrer string, userAgent string, userCookie string, ip string, clicked Boolean) PARTITIONED BY (day string, hour string) STORED AS SEQUENCEFILE LOCATION '${OUTPUT}/tables/joined\_impressions';

+++ Step 4.1.2

CREATE TABLE tmp\_impressions (requestBeginTime string, adId string, impressionId string, referrer string, userAgent string, userCookie string, ip string) STORED AS SEQUENCEFILE;

+++ Step 4.1.3

INSERT OVERWRITE TABLE tmp\_impressions SELECT from\_unixtime(cast((cast(i.requestBeginTime as bigint) / 1000) as int)) requestBeginTime, i.adId, i.impressionId, i.referrer, i.userAgent, i.userCookie, i.ip FROM impressions i WHERE i.dt >= '${DAY}-${HOUR}-00' and i.dt < '${NEXT\_DAY}-${NEXT\_HOUR}-00';

+++ Step 4.1.4

CREATE TABLE tmp\_clicks ( impressionId string ) STORED AS SEQUENCEFILE;

+++ Step 4.1.5

INSERT OVERWRITE TABLE tmp\_clicks SELECT impressionId FROM clicks c WHERE c.dt >= '${DAY}-${HOUR}-00' AND c.dt < '${NEXT\_DAY}-${NEXT\_HOUR}-20';

+++ Step 4.1.6

INSERT OVERWRITE TABLE joined\_impressions PARTITION (day='${DAY}', hour='${HOUR}') SELECT i.requestBeginTime, i.adId, i.impressionId, i.referrer, i.userAgent, i.userCookie, i.ip, (c.impressionId is not null) clicked FROM tmp\_impressions i LEFT OUTER JOIN tmp\_clicks c ON i.impressionId=c.impressionId;

+++ Step 4.1.7

MSCK REPAIR TABLE joined\_impressions;

+++ Step 4.1.8

select \* from joined\_impressions limit 10;

Â© 2017 Amazon Web Services, Inc. or its affiliates. All rights reserved.