Prof. Sarada Samantaray

Big Data Technology

M. Tech Data Science

NMIMS Deemed-to-be-University

Processing Big Data with HIVE in Hadoop

Lab – Processing Big Data with Hive

Overview

In this lab, you will process data in web server log files by creating Hive tables, populating them with data, and querying them.

Preparing for Data Processing



Data file is attached in zip format. Extract the data into your local hard disk (into a directory called “iislog”).

View Source Data

1. In the **iislog** folder where you extracted the lab files, open any of the text files in a text editor. Each file in this folder is an Internet Information Services (IIS) log file from a web server, and there is a file for each month between January and June 2017. When you have viewed the information in the file, close it without saving any changes.
2. View the contents of the **iislogs\_gz** folder. This contains the same log files compressed using the gzip compression algorithm, significantly reducing the size of the files.

Upload the Log Files to HADOOP

Now you are ready to upload the source data to Hadoop DFS storage for processing. The instructions here assume you will use hdfs / dfs to do this.

1. Start hadoop, by signing into your Hadoop system.
2. Upload all the 6 files from iislog directoryfolder to a new folder named **data/logs** in root of the container.

Creating, Loading, and Querying Hive Tables

Now that you have uploaded the source data, you can create Hive tables and use them to process the data. To do this, you will need to open an SSH console that is connected to your cluster.

1. Take an terminal window

OR

1. Open PuTTY, and in the **Session** page, paste the host name into the **Host Name** box. Then under **Connection type**, select **SSH** and click **Open**. If a security warning that the host certificate cannot be verified is displayed, click **Yes** to continue.
2. When prompted, enter the SSH username and password you specified when provisioning the cluster (not the cluster login).

View the Uploaded Data Files

Now that you have a remote command console for your cluster, you can verify that the log files have been uploaded to the shared cluster storage.

1. In the SSH console for your cluster, enter the following command to view the contents of the **/data/logs** folder in the HDFS file system.

hdfs dfs -ls /data/logs

1. Verify that the folder contains six log files.

Create a Hive Table for the Raw Log Data

While there are many Hive editors available, and you can use any one that you prefer. The instructions here assume that you will use the Hive command line interface to work with Hive in this lab.

1. In the SSH console for your cluster, enter the following command to start the Hive command line interface:

hive

1. In the Hive command line interface, enter the following HiveQL statement to create a table named **rawlogs**:

CREATE TABLE rawlog

(log\_date STRING,

log\_time STRING,

c\_ip STRING,

cs\_username STRING,

s\_ip STRING,

s\_port STRING,

cs\_method STRING,

cs\_uri\_stem STRING,

cs\_uri\_query STRING,

sc\_status STRING,

sc\_bytes INT,

cs\_bytes INT,

time\_taken INT,

cs\_user\_agent STRING,

cs\_referrer STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' ';

1. Wait for the job to complete.
2. Enter the following command to verify that the table has been created (there may also be an existing table named **hivesampletable**):

SHOW TABLES;

1. Enter the following command to query the table, and verify that no rows are returned:

SELECT \* FROM rawlog;

Load the Source Data into the Raw Log Table

1. In the Hive command line interface, enter the following HiveQL statement to move the log files you previously uploaded into the folder for the **rawlogs** table:

LOAD DATA INPATH '/data/logs' INTO TABLE rawlog;

1. Wait for the job to complete.
2. Enter the following command to return the first 100 rows in the **rawlogs** table:

SELECT \* FROM rawlog LIMIT 100;

1. View the output, noting that the query retrieved the first 100 rows from the table in tab-delimited text format. Note that the output includes rows containing comments from the source document (the first column value for these rows is prefixed with a **#** character, and null values are used for columns in the table schema for which there are no values in the source data).

Clean the Log Data

1. In the Hive command line interface, enter the following HiveQL statement to create an external table named **cleanlog** based on the **/data/cleanlog** folder:

CREATE EXTERNAL TABLE cleanlog

(log\_date DATE,

log\_time STRING,

c\_ip STRING,

cs\_username STRING,

s\_ip STRING,

s\_port STRING,

cs\_method STRING,

cs\_uri\_stem STRING,

cs\_uri\_query STRING,

sc\_status STRING,

sc\_bytes INT,

cs\_bytes INT,

time\_taken INT,

cs\_user\_agent STRING,

cs\_referrer STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '

STORED AS TEXTFILE LOCATION '/data/cleanlog';

1. Wait for the job to complete.
2. Enter the following command to extract rows that do not contain comments from the **rawlogs** table, and insert them into the **cleanlog** table:

INSERT INTO TABLE cleanlog

SELECT \*

FROM rawlog

WHERE SUBSTR(log\_date, 1, 1) <> '#';

1. Wait for the job to complete.
2. Enter the following command to retrieve the first 100 rows from the **cleanlog** table:

SELECT \* FROM cleanlog LIMIT 100;

1. View the results returned by the query, noting that the **cleanlog** table does not include any rows prefixed with a **#** character.

Create a View

1. In the Hive command line interface, enter the following command to create a view named **vDailySummary**:

CREATE VIEW vDailySummary

AS

SELECT log\_date,

COUNT(\*) AS requests,

SUM(cs\_bytes) AS inbound\_bytes,

SUM(sc\_bytes) AS outbound\_bytes

FROM cleanlog

GROUP BY log\_date;

1. Wait for the job to complete.
2. Enter the following command to query the **vDailySummary** view:

SELECT \* FROM vDailySummary

ORDER BY log\_date;

1. Wait for the job to complete and then view the output returned by the query.

Partitioning Data

Partitioning data can improve performance when queries commonly filter on specific columns, or when the job can be effectively split across multiple reducer nodes.

Create and Load a Partitioned Table

1. In the Hive command line interface, enter the following command to create a partitioned table (you can copy and paste this from **Create Partitioned Table.txt** in the **HDILabs\Lab02** folder):

CREATE EXTERNAL TABLE partitionedlog

(log\_day int,

log\_time STRING,

c\_ip STRING,

cs\_username STRING,

s\_ip STRING,

s\_port STRING,

cs\_method STRING,

cs\_uri\_stem STRING,

cs\_uri\_query STRING,

sc\_status STRING,

sc\_bytes INT,

cs\_bytes INT,

time\_taken INT,

cs\_user\_agent STRING,

cs\_referrer STRING)

PARTITIONED BY (log\_year int, log\_month int)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '

STORED AS TEXTFILE LOCATION '/data/partitionedlog';

1. Enter the following command to load data from the **rawlog** table into the partitioned table:

SET hive.exec.dynamic.partition.mode=nonstrict;

SET hive.exec.dynamic.partition = true;

INSERT INTO TABLE partitionedlog PARTITION(log\_year, log\_month)

SELECT DAY(log\_date),

log\_time,

c\_ip,

cs\_username,

s\_ip,

s\_port,

cs\_method,

cs\_uri\_stem,

cs\_uri\_query,

sc\_status,

sc\_bytes,

cs\_bytes,

time\_taken,

cs\_user\_agent,

cs\_referrer,

YEAR(log\_date),

MONTH(log\_date)

FROM rawlog

WHERE SUBSTR(log\_date, 1, 1) <> '#';

1. Wait for the query job complete. By partitioning the data in this way, queries filtered on year and month will benefit from having to search a smaller volume of data to retrieve the results.
2. Enter the following command to query the partitioned table, and when the job has completed, view the results:

SELECT log\_day, count(\*) AS page\_hits

FROM partitionedlog

WHERE log\_year=2008 AND log\_month=6

GROUP BY log\_day;

View Folders for a Partitioned Table

1. In the Hive command line interface, enter the following command to exit Hive:

exit;

1. At the command prompt, enter the following command to view the **/data/partitionedlog** folder:

hdfs dfs -ls /data/partitionedlog

1. Note that the folder contains a subfolder for each value in the first partitioning key (in this case, a single folder named **log\_year=2017**).
2. Enter the following command to view the contents of the **log\_year=2017** subfolder (note that quotation marks are required because the folder name includes a “=” character):

hdfs dfs -ls "/data/partitionedlog/log\_year=2017"

1. Note that the subfolder contains a further subfolder for each value in the second partitioning key (in this case, a folder for each **log\_month** value in the source data).
2. Enter the following command to view the contents of the **log\_month=6** subfolder:

hdfs dfs -ls "/data/partitionedlog/log\_year=2008/log\_month=6"

1. Note that the subfolders for each month contain one or more data files. In this case, a file with a name similar to **000000\_0** contains all of the log records with a year of 2008 and a month of 6.
2. Enter the following command to view the last few kilobytes of data in the file (modifying the file name if necessary):

hdfs dfs -tail "/data/partitionedlog/log\_year=2008/log\_month=6/000000\_0"

1. Note that the first field in the data file contains the day of the month (the last few rows will contain the value **30**.) The year and month are not stored in the data file, but are determined from the partitioning subfolder in which the file is stored - this file contains only the records for June 2008.
2. Close all open command windows and Hadoop system (power off saving last state of the machines).

Cleaning Up

Now that you have finished this lab, you can delete the HDInsight cluster and storage account. This ensures that you avoid being charged for cluster resources when you are not using them. If you are using a trial Azure subscription that includes a limited free credit value, deleting the cluster maximizes your credit and helps to prevent using it all before the free trial period has ended.

**Note**: If you are proceeding straight to the next lab, omit this task and use the same cluster in the next lab. Otherwise, follow the steps below to delete your cluster and storage account.

Delete Cluster Resources

1. In the Microsoft Azure portal, click **Resource Groups**.

2. On the **Resource groups** blade, click the resource group that contains your HDInsight cluster, and then on the **Resource group** blade, click **Delete**. On the confirmation blade, type the name of your resource group, and click **Delete**.

3. Wait for your resource group to be deleted, and then click **All Resources**, and verify that the cluster, and the storage account that was created with your cluster, have both been removed.

4. Close the browser.