Big Data Class Project - Part 1

Class: BUAN 6346.502 - MIS 6346.502

Semester: Spring 2020

The class project report is individual work. If you choose to consult anyone, please list names of every person you have discussed project with. The Due date for the project will be published in the eLearning.

Project Description

The project is an essential part of this class. It will allow you to practice and demonstrate your Big Data (BD) and Analytics skills. It can also be a valuable addition to your projects portfolio that you can demonstrate to prospective employers.

The project is divided to two parts. First part is described in this document.

In part 1 you will be able to practice concepts you have learned in the first part of the class:

1. HDFS
2. Hive
3. AWS Big data technologies: S3 and Athena

Project Requirements

Part 1 - Amazon Reviews Analysis

For the project, you will perform analysis of the provided dataset using Big Data techniques.

Project Environment: AWS EMR - AWS Educate Class account.

Project tools:

1. AWS S3
2. AWS EMR - Hive and HDFS
3. AWS Athena - AWS alternative to Hive for the files stored in S3

Data Set

Amazon reviews dataset: https://registry.opendata.aws/amazon-reviews/

Documentation: https://s3.amazonaws.com/amazon-reviews-pds/readme.html

We will using parquet version of the dataset in order to reduce storage requirements on AWS EMR and to speed up processing and analytics.

The dataset is partitioned by product category, here is the subset of the partitions. Note the shell command you can execute from AWS EMR master machine shell:

```
aws s3 ls s3://amazon-reviews-pds/parquet/
PRE product_category=Apparel/
PRE product_category=Automotive/
PRE product_category=Baby/
PRE product_category=Beauty/
PRE product_category=Books/
PRE product_category=Camera/
PRE product_category=Digital_Ebook_Purchase/
.........................................
.........................................
PRE product_category=Video_Games/
PRE product_category=Watches/
PRE product_category=Wireless/
```

Each category folder folder contains parquet files with the reviews for the category. Example:

```
aws s3 ls s3://amazon-reviews-pds/parquet/product_category=Automotive/
2018-04-09 06:35:42   84703245 part-00000-495c48e6-96d6-4650-aa65-
3c36a3516ddd.c000.snappy.parquet
.................................................................
2018-04-09 06:35:46   85203786 part-00009-495c48e6-96d6-4650-aa65-
3c36a3516ddd.c000.snappy.parquet
```

Requirements

Use the following product categories:

```
- Wireless
- Automotive
- Music
- Digital_Music_Purchase
- Sports
- Toys
- Digital_Video_Games
- Video_Games
```

Start your analysis from year 2005. Exclude multiple reviews by the same users for the same product. Each user should be allowed to review the product only once. To improve performance of your queries, create external table to point to HDFS/S3 file that will include all review-ids to be excluded.

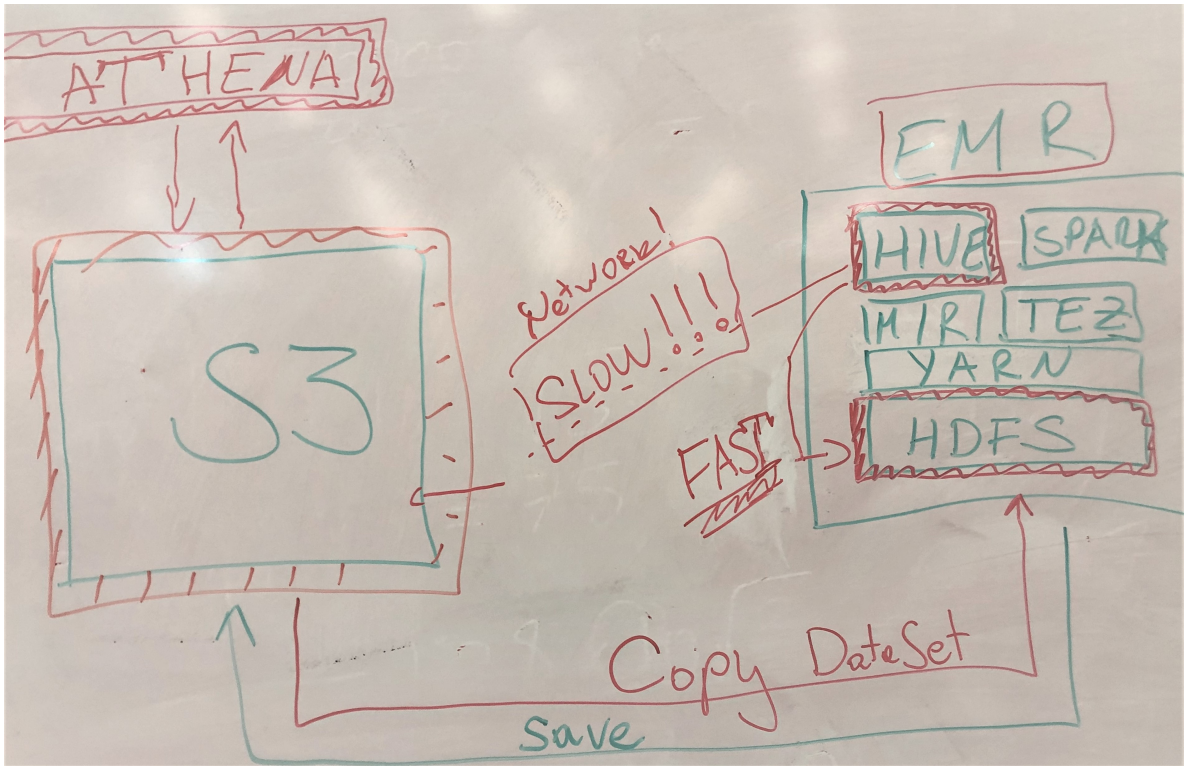
Using AWS EMR Hive and AWS Athena, answer the following questions:

1. Explore the dataset and provide basic exploratory analysis:
    1. Number of reviews
    2. Number of users
    3. Average review stars
    4. Average length of the review
    5. Number of verified versus unverified reviews
    6. At least two more additional metrics
    7. Provide trending (over time) analysis of each of the metrics above

2. Provide detailed analysis of Music/Digital_Music_Purchase and Digital_Video_Games/Video_Games over time.
    1. Do you see correlation (maybe negative) between the categories over time?
    2. Are there same users reviewing in both categories?
    3. Can you identify similar items in both categories? Do they get same rating?
    4. You should cover additional questions and not limit yourself to the above questions
3. You should demonstrate your ability to use Hive advanced functions:
    1. Window functions: moving average, rank, aggregation functions using relevant ordering and partitioning
    2. Analytical Aggregate functions: percentile, min, max, average, standard deviation, correlation

Technical details on working with the Dataset

High level architecture



AWS Athena

You should use AWS Athena to perform majority of the analysis. This is to simplify access and avoid charges associated with EMR. However, Athena might not support certain functions Hive supports, for example `percentile_approx` function.

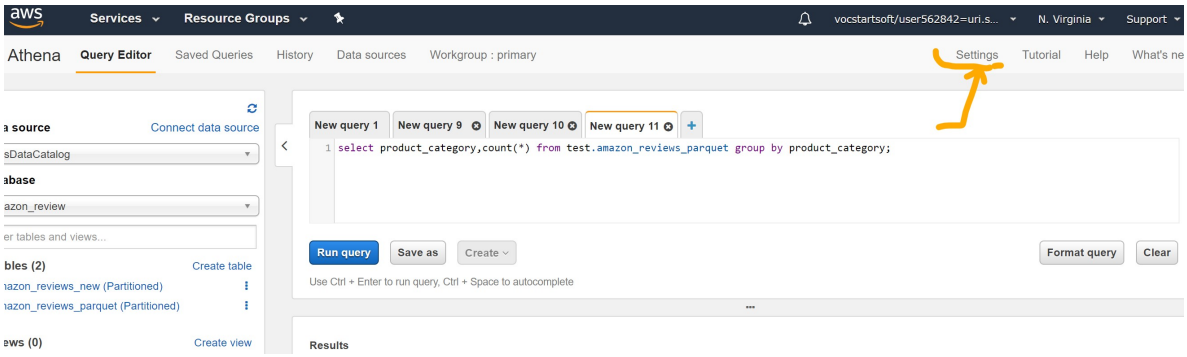
Once analysis has been completed, you should execute final queries using AWS EMR Hive.

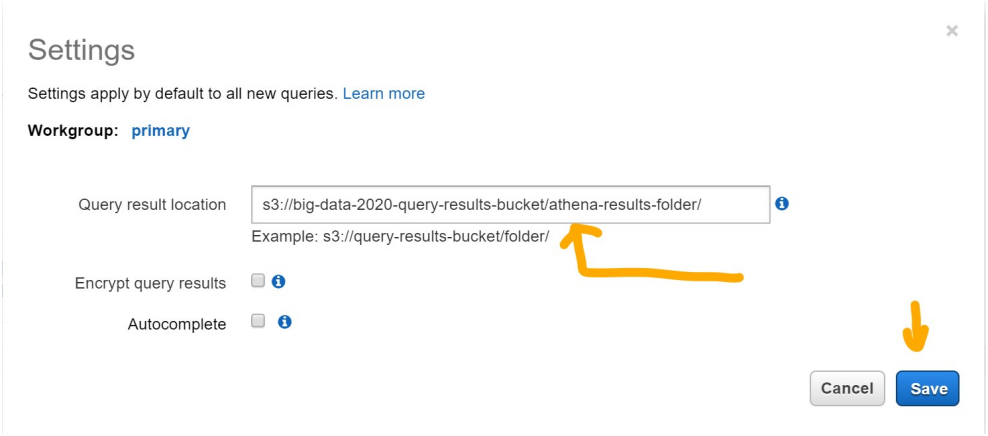
AWS Athena is AWS service and can be accessed from AWS console by selecting "Athena" from services list.

Athena is similar to Hive, and designed to provide SQL interface for the files stored in S3.

To start with Athena, first you need to create bucket in S3. To access S3, go to "Services"->"S3" from the AWS console.

In the example below, I have created S3 bucket with name `big-data-2020-query-results-bucket` and folder `athena-results-folder/` inside the bucket. You will need to create bucket with different names, since S3 buckets are unique. The full path the location to save Athena query results is `s3://big-data-2020-query-results-bucket/athena-results-folder/`.





To create table in Athena for the review dataset, run following commands. Those commands are same as commands we have executed to create external Hive table:

```
create database amazon_review;

CREATE EXTERNAL TABLE amazon_review.amazon_reviews_parquet(
  marketplace string,
  customer_id string,
  review_id string,
  product_id string,
  product_parent string,
  product_title string,
  star_rating int,
  helpful_votes int,
  total_votes int,
  vine string,
  verified_purchase string,
  review_headline string,
  review_body string,
  review_date DATE,
  year int)
```

```
PARTITIONED BY (product_category string)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  's3://amazon-reviews-pds/parquet/';

MSCK REPAIR TABLE amazon_review.amazon_reviews_parquet;
```

AWS EMR

The final project deliverables should be produced using AWS EMR Hive. Here are the steps to get you going:

1. Provision EMR cluster. You might want to provision 2 Slave servers to speed up processing
2. Copy Amazon reviews to EMRs HDFS, you should run similar command for the relevant product categories:

```
hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Apparel/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Apparel/ \
--dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Apparel/
```

Create an external table in Hive by pointing it to the HDFS folder (`/hive/amazon-reviews-pds/parquet/`) you just populated with parquet files containing the reviews:

```
create database amazon_review;
drop table amazon_review.amazon_reviews_parquet;

CREATE EXTERNAL TABLE amazon_review.amazon_reviews_parquet(
  `marketplace` string,
  `customer_id` string,
  `review_id` string,
  `product_id` string,
  `product_parent` string,
  `product_title` string,
  `star_rating` int,
  `helpful_votes` int,
  `total_votes` int,
  `vine` string,
  `verified_purchase` string,
  `review_headline` string,
  `review_body` string,
  `review_date` DATE,
  `year` int)
PARTITIONED BY (
  `product_category` string)
--ROW FORMAT DELIMITED
--STORED AS PARQUET
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
```

```
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  'hdfs:///hive/amazon-reviews-pds/parquet/'
TBLPROPERTIES (
  'transient_lastDdlTime'='1583454851');

Msck repair table amazon_review.amazon_reviews_parquet;
```

Example of running Hive SQL to calculate percentiles for the reviews:

```
hive> select
product_category,percentile_approx(star_rating,array(0.10,0.25,0.50,0.6,0.7,0.8)
),avg(star_rating) from amazon_review.amazon_reviews_parquet group by
product_category;

----
Wireless        [1.0,3.0,4.336336492690657,5.0,5.0,5.0] 3.892277254145134
Apparel [2.0,3.165215352960072,4.552996758064953,5.0,5.0,5.0]
4.105233930306817
```

Project report

You will create a well formatted project report that would include the following sections. You can use any editor, and your report will be delivered in PDF file format.

Introduction and problem description

Formulate your own problem description and describe it in detail.

Formulate a list of tasks that you are about to perform to solve this problem.

Provide in detail your approach to the problem at hand and the techniques you will be utilizing to tackle the problem.

Submit technical scripts and SQL queries

Provide code (SQL and shell) and screenshots from EMR Hive to show your work.

Visualization

You can enhance your findings by utilizing graphs and plots. You can use any tool and libraries to do so.

Conclusion

Provide conclusions to any metric and or finding. You should provide your own assessment and not just rely on the report user to make his own conclusions.

References

Provide relevant references. These may include a description of data, previous analysis of the data available on the internet, code references etc.

References are very important, any code that you reuse should be recognized and mentioned.