Cloud Technologies – Group L Assignment 2

Application Overview:

We decided to make a web map Interface for locating vegetarian restaurants in the US, filtering them by State, Cuisine and sorting by top/bottom 10/20/500/100 based on a rating. We were unable to complete the functionality of the app on time.

Data Source:

The data has been collected from Kaggle website. The dataset consists of a list of over 18,000 restaurant in USA that serves vegetarian food. The dataset includes address, city, state, phone number, latitude, longitude and many more columns that are used to determine vegan- friendly cities in the USA. It can be found at:

https://www.kaggle.com/datafiniti/vegetarian-vegan-restaurants

Processing:

Extract, Transform and Load:

The dataset consisted of different columns with misspelled words, blank spaces which required lots of cleaning.

Pig/Mapreduce/hive were used to perform the ETL operations and the data were further processed and used for analysis.

The below steps have been followed:

Creation of HDFS directory :

Step 1:

A HDFS directory is created with the name ‘ apprest ‘ (**hdfs dfs -mkdir /apprest)**

Step 2:

The csv file is uploaded to HDFS directory.

(**hdfs dfs -put home/ca675/downloads/restaurantsfinal.csv /apprest**)

PIG Operations:

The raw data is processed using pig now. Code is outlined in *pig_code.pig.txt*

Hive Operations:

The processed data is loaded into a table in a database in Hive.

Step 1:

A table named **restsc** is created using the below query.

```
CREATE TABLE IF NOT EXISTS restsc ( Name String, Address String , Cuisine String, Score Int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY "\t"
LINES TERMINATED BY "\n"
STORED AS TEXTFILE;
```

Step 2:

The processed data from pig is loaded into hive using the below query.

```
LOAD DATA INPATH '/apprest/tb/part-m-00000' INTO TABLE restsc;
```

Analysis using Hive queries:

1. **To find the top 10 restaurants by score:**
   Select Name, Score from restsc order by score desc limit 10;
2. **To find the top 10 cuisines by score:**
   Select Cuisine, Score from restsc order by score desc limit 10;
3. **To find the list of restaurants with chinese cuisines:**
   Select name, cuisine from restsc where Cuisine=’Chinese’ limit 10;
4. **To find the list of restaurants in Atlanta:**
   Select name, Address from restsc where Address=’Atlanta’ limit 10;

Related Work-any similar systems/app

Yelp - It was the first app to find the restaurant on iphone and its continues update keep it on the top list.

LocalEats - It is a great app that entirely focuses on the small restaurant that are located in the local communities.

Urbanspoon - It is a restaurant finder app based on the feature of a slot machine-esque method of randomly finding a local restaurant.

Dublin Bus – Dublin Bus mobile app makes use of maps to outline routes.

Challenges and lessons learned

We tried to work on the AWS. A number of EMR clusters were started and terminated. Quick deployment of clusters would have helped in executing pig and hive script at a faster rate. Unfortunately, after loading the dataset on the Amazon EC2 cluster, we faced problem in connecting the application to the cluster and loading the data as per requirement. By doing that, AWS has charged us that's why we decided to go for the open sources.

We also realised how difficult it was to connect our front-end to our back end. We required an FTP client, namely Filezilla, to upload our processed CSV file to our localhost, which could then be accessed by Javascript to produce the web UI. We used http://www.convertcsv.com/csv-to-json.htm to convert our CSV to a JSON file (required for plotting map markers).

Responsibility statement:

Specific tasks and general roles

- **Research (linking front-end to back-end) –** Nitin, Kevin, Shivam
- **Mid-Way Report –** All
- **Peer Feedback -** All
- **ETL process using PIG/Hive** – Kanmani
- **Processing and Manipulation of data in the Amazon EMR -** Nitin
- **UI design**- Kevin, Nitin, Shivam
- **Final report**- All
- **Screencast**- Kanmani and Kevin
- **Processing of data on the Google Cloud -** Shivam

We used http://www.belbin.com/about/belbin-team-roles to identify roles as below:

Kanmani-**Implementer**

- I worked on the ETL processing using pig and analysed the objective questions using hive queries.
- I also contributed significantly to most of the written aspects of the assignment as well as the screencast

Kevin- **Implementer, Completer**

- I developed the UI and linked it to Kanmani's completed CSV. I was unable however to complete the functionality of the app in due time.
- I also contributed heavily to the written aspects of the assignment as well as the screencast.

Nitin- **Team Worker, Specialist, Implementer**

- Data cleansing of the CSV file.
- Created basic User Interface to set query parameters using J-Swing components. It is accessible through the github repository.
- Created an instance of S3 bucket and EC2, also managed to connect with User Interface using Netbeans cloud services, but due to cost of service couldn't use the service anymore.
- Contributed to the written aspects of the assignment.

Source codes : https://github.com/NitinYadav20/Cloud-Technologies-App-Development-/tree/master/RestaurantApplication

Shivam - **Implementer, Specialist**

- Use Google Cloud Platform
    - create Compute Engine with i masternode and 2 worker nodes
    - Storage - use to store the initial data, csv/txt format
    - SQL - link storage with the SQL
    - DataProc - link it with storage and computing engine to use Hive(Hive/Beeline)
- Contribute towards application making on HTML and Java
- Solely tried till the end time of this assignment to connect Cloud SQL and DataProc with the Application and Storage will be automatically gets accessed.

Google Cloud Platform Credentials:

ID - shivam.kalra4@gmail.com

Password - 9953228625

(Note: There are charges so, I would request you to start the compute engine and stop it after use.

You can see the outcome of our work through presentation)

Include individual scores allocated to each member

Kanmani-25%

Kevin-25%

Nitin-25%

Shivam-25%

Response to peer feedback:

Overall it was a very good analysis of our proposal, however in terms of clearly outlining future steps, we didn't feel that we were any further than when we wrote the proposal, seeing as we were still assessing different technologies. Some useful technologies were flagged, but nothing specific to our idea.

Appendix :

Youtube URL: https://youtu.be/w2CAZnUd0zc

Github repository: https://github.com/meehankevin/CA675_Cloud_Technologies

Presentation: https://github.com/NitinYadav20/Cloud-Technologies-App-Development-/blob/master/cloud-PPT.pdf