Application of Cloud Technologies

A report submitted to Dublin City University by
Sujith Galla, Harold Hainsworth, Monica Holmes, JJ Kelly, Billy Kilduff, Paul Ryan

Student ID	Name	Email
18214293	Sujith Galla	sujith.galla3@mail.dcu.ie
18212453	Harold Hainsworth	harold.hainsworth2@mail.dcu.ie
19215442	Monica Holmes	monica.holmes5@mail.dcu.ie
19215416	JJ Kelly	jj.kelly253@mail.dcu.ie
19213493	Billy Kilduff	billy.kilduff4@mail.dcu.ie
57668945	Paul Ryan	paul.ryan34@mail.dcu.ie

School of Computing, Dublin City University, Glasnevin, Dublin 9

9th December 2019

Lecturer: Dr. Long Cheng

Video Demo:
https://youtu.be/5AzJ-NU6pl0

Midway Report:
https://github.com/AlBi11y/CloudApp/blob/master/Documentation/CA675_A2_Mid
Way_Report_FINAL.pdf

Introduction

The sharing economy is growing and growing. According to a recent internal economic report, Airbnb estimates a contribution of more than €700 million to the Irish economy (a combination of host income and guest spending) from over 1.8 million visitors staying in an Airbnb listed accommodation. Hosts are pocketing almost €168 million of this figure.

Our goal is to create a tool for Airbnb hosts in Ireland to better understand their peer hosts, other properties in their neighbourhood and the optimal pricing strategy for their property, utilising a custom built model that we have created through Google Cloud that advises hosts of suggested prices based upon their properties' characteristics.

This will be done by taking a publicly available dataset from Inside Airbnb featuring over 9,000 properties in the Dublin area. The inspiration for this idea is based upon a similar website that is available here.

Technologies Used

- **Google Cloud Platform** (GCP) suite offers a number of cloud computing services. We used this to configure, create and run clusters, and upload our data onto a remote server.
- **Apache Hadoop** open source framework was then used for storing data and running our application on a cluster, making the most of the resources available in the network.
- **Apache Pig (Pig Latin)** platform is suitable for analysing large data sets, this allowed us to load and cleanse the data, and to reduce the large dataset to only the relevant columns.
- **Apache Hive (HQL)** data warehousing software is suitable for managing large datasets residing in distributed storage, which we used to create and query our database residing in Hadoop.
- **Docker** was used to run our model. The output of this instance was then pushed back to Git for later use in our visualisation tool. The model code was scripted in Python.
- **Qlikview** was chosen to show the visualisation of our data. This was due to team experience with the tool and the application's ability to manage large datasets with minimal processing time.
- **Movavi** screen capture software was used to create the screencast/video to demonstrate our application.
- **GitHub**, providing software development version control hosting with Git, was used to manage the source code, allowing us to collaborate and control distributed versions.
- **Slack** was also used, mainly for group communication.

Data Source

As per above, the data which we chose to use for our app was Airbnb listings data for Dublin and is publicly available from Inside Airbnb.

A detailed listings file was downloaded as the main dataset for this project and stored on GitHub (file available here). The file contains 106 columns related to listings, such as price per night, longitude, latitude, property features, and much more. In the next stage, we needed to wrangle this data for use in our Qlikview dashboard.

Data Processing

Dataproc

Data extraction, transformation, load and querying were run on GCP, utilizing dataproc - a fast, relatively easy to use and fully managed cloud service for running Hadoop clusters.

Extract, Transform, Load (ETL) Process

The ETL work was done using Apache Pig. The raw data downloaded (listings) consisted of 106 columns. This entire dataset was loaded into Pig, a number of columns were cleansed, indicators created, and the transformed Pig output was stored in HDFS directory for further processing (cleansed dataset is available here).

The Pig code that was used in this process is also available on GitHub here.

Hive Querying

The reduced dataset was subsequently picked up in Hive to populate the newly created reduced_listings table. A number of queries were written in HiveQL and the output of these were used within our dashboard. The HiveQL code can be found in on GitHub here.

The cleansed and reduced output was saved as a file on Hadoop cluster and placed under Data directory on the HDFS. This was done using the following Hive queries found here.

Model Creation

A hedonic pricing model was used to predict the price of a stay at a property based upon property characteristics such as location, number of bedrooms, etc. We ran correlation functions to determine what affects pricing and found location, accommodation type & amenities to be contributing factors on the price of a property listing.

There were more data points available per listing such as host type, reviews, ratings, number of days property however these did not strongly correlate towards the final price and hence not considered for final user input. It was also necessary to keep user input simple and seamless for predicting price of their listing. Due to the nature of prediction, a regression technique was deemed suitable for predicting the price. Before prediction, attributes such as beds, room type that presented with null values were either excluded or filled with median values. Outliers based on price were also excluded to improve model accuracy. Several different regression techniques were tested on the selected features such as linear, ridge & lasso regressions.

In the end the lasso regression technique was used to model the prediction function as it produced the best RMSE results. After the model was created, using SQL and cross joins, every possible combination of user input required for any listing was created and respective prices predicted. This data was then stored as a file to be uploaded to Qlikview.

The model was run using a Docker instance on Google Cloud Platform (see Docker code here). The project Git was pulled into the container. We then used python code and cleansed data files used to build the image. The output of the model was saved to the repo and it was then pushed back the Git. Pricing model code used can be found on GitHub here.

Connecting Data to Visualisation

In order to visualise our data, we initially used BigQuery to create a connection. A bucket in Google Cloud Storage was created within the cluster and the previously cleansed CSV file was then stored into the Bucket. Code for this is available on GitHub here.

Google BigQuery then connected to the Bucket to access the file and create a new table in BIgQuery using the data from our CSV file. These new files were envisaged to be connected directly to Tableau.

However, after testing the functionality of Tableau, we made the decision to use Qlikview instead. We decided to do this as Qlikview was better for loading such large datasets. Equally we had prior experience using this visualisation tool.

Qlikview was not compatible with BigQuery. Therefore in order to load the data into Qlikview, we downloaded our datasets from the cluster and uploaded directly into Qlikview. Our Qlikview dashboard code can be viewed on GitHub here.

Challenges and Lessons Learned

Overall we're pleased with the output to the project.

However we did encounter challenges that we would approach differently for future work. We spent a lot of time trying to connect our cluster directly to our visualisation tool. We finally successfully connected to Tableau via BigQuery after much difficulty.

When we finally had our connection, we opted to change visualisation tools from Tableau to Qlikview for 'fit' reasons. For future projects, it's advisable to fit your problem/dataset to the visualisation tool first as this will influence the connection type between your cluster and visualisation tool.

Responsibility statement

The general roles for the group were outlined in the Midway Report (Appendix A).

Below summary illustrates the contribution of each of the team members throughout the project:

Sujith - Satisfactory

- Idea Generation with Paul
- Prediction Model Build in Python with JJ

Harold - Satisfactory

- Midway Report Preparation with Paul & Monica
- Data Visualisation in Qlikview
- Presentation of the Demo Video

Monica - Satisfactory

- Midway Report Preparation with Paul & Harry
- Set up Airbnb Account, downloaded data and carried out initial data exploration.
- Set up a new project on Google Cloud Console and granted the IAM roles to team members to facilitate group collaboration. Created, configured and started the project cluster.
- Set up a shared CloudApp repository on GitHub, along with a project and a basic Kanban style board (an aid to work through the tasks outlined in the Midway Report)
- Loaded data into PIG, transformed it and stored the output in HDFS.
- Wrote HiveQL code to create a table and populate it with the reduced dataset.
- Wrote a number of HiveQL queries to produce figures for the UI.
- Planned to work on the visualisation but that was not possible as Tableau was dropped
- Final Report Preparation with Paul & Billy

JJ - Satisfactory

- Prediction Model Build in Python with Sujith
- Loaded data into PIG and wrote Pig Latin statements to transform it

Billy - Satisfactory

- Initial work on cluster connection to visualization tool
- Assisted with preparation of the demo video, by having the steps outlined
- Final Report Preparation with Paul & Monica

Paul - Satisfactory

- Idea Generation with Sujith
- Data Visualisation in Qlikview
- Midway Report Preparation with Monica & Harry
- Final Report Preparation with Billy & Monica

Appendix A

Tasks planned as per Midway Report:

Task	Main Contributor	Complete	Notes
Choose dataset	Team	Yes	Airbnb
Midway report	Paul, Harry, Monica	Yes	
Download data	Billy	Yes	
Airbnb account setup & Data Exploration	Monica	Yes	
Load the data into PIG and query it in Hive or Spark	JJ / Monica	Underway	
Build a model in Python	JJ & Sujith	Underway	
Attempt to automate data processes, set up 2 VM's for handling traffic from different regions	Sujith	To Do	
Connect Hadoop on cloud to Tableau	Billy	Underway	
Visualisation in Tableau	Harry, Paul, Monica	To Do	
Demonstration Video	Team	To Do	
Final Report	Team	To Do	