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

November 2019

Lecturer: **Dr. Long Cheng**

Assignment 2

Application of Cloud Technologies using Dublin data from Airbnb

Proposal

Ireland's love for Airbnb continues to grow and grow. 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. We want to explore this trend by taking a look at Airbnb data from the Dublin market, that is publicly available from Inside Airbnb.
The project will explore the data through 3 lenses:

1. What is the current landscape?

We want to take the listings data available and map it out on a map of Dublin. This will give our users the opportunity to explore the data at a granular level, view listings on the market, their prices and some of the key features of the accommodation available on Airbnb in Dublin.

- Listings placed on Tableau map using long/lat coordinates
- Hover over and allow users to see more info about the listing
    - URL link to the listing on Airbnb site
    - Price
    - Listing Availability
    - Review Score

2. Neighbourhood level view

We will take this data and give the user an aggregated view of listings based on the postcode and neighbourhood. We wish to explore the impact of neighbourhoods on listing prices.

- Boundary view of Airbnb listings by their defined neighbourhoods
- Introduce an additional boundary level which uses postcode/town data and present this view also
- With these filters, provide some descriptive statistics
    - Average price
    - Max/min price
    - The average number of amenities
    - Average review score
    - Total number of reviews
    - The aggregate number of property/room types

3. Listing price predictor

Lastly, having completed the data exploration and cleansing, we would like to create a model utilizing different listing features and giving users the opportunity to find out how much money they could make by listing their property with Airbnb - a feature that's not currently available on the platform. The model will display a predicted price for the user's property after they have input a few answers with regard to their property.

Specific Tasks (high-level description)

We plan to collaborate on each task but we have divided the high-level tasks and assigned them to main contributors responsible for the delivery.

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	

Choice of Technologies

Apache Pig (Pig Latin) platform is suitable for analysing large data sets, therefore we can use it 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, meaning we could use it to create and query our database residing in Hadoop.

OR

Apache Spark is an analytics engine suitable for large-scale data processing, achieving high performance, using a query optimizer and a physical execution engine to name but a few. It is compatible with Hadoop data, can process data in HDFS and Hive, and is designed to perform interactive queries and machine learning, therefore we are considering it instead of Pig and Hive.

We plan to explore both options and will document what and why we chose in the final report.

Google Cloud Platform (GCP) suite offers a number of cloud computing services, meaning we can configure, create and run clusters, and upload our data onto a remote server.

Apache Hadoop open source framework will be used for storing data and running our application on a cluster, making the most of the resources available in the network.

Hadoop Streaming will be used to deploy our Python machine learning module that will calculate an output the price estimation of all possible listings.

Tableau is our chosen data visualisation software, mainly because it connects to almost everything, is very interactive and offers lots of great features, which we hope will help us to attract users.

Movavi screen capture software will be used to create the screencast/video to demonstrate our application.