Flight Price Prediction

HARSHIT KRISHNAKUMAR[1,*] AND KARTHIK ANBAZHAGAN[2]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[2] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors: harkrish@iu.edu, kartanba@iu.edu

project-001, March 27, 2017

This project aims at tracking live flight status and flight pricing in the US. Live flight data streams are obtained using Python APIs and stored in Big Data Hadoop Distributed File Systems. This paper explores the use of Apache Hive to store data streams and analyse the data. The analyses will be presented in real time using d3.js.

Keywords: big data, apache hive

https://github.com/cloudmesh/sp17-i524/project/S17-IR-P002/report/report.pdf

CONTENTS

INTRODUCTION

Air travel is getting increasingly popular with the airlines providing cheaper fares and better services. More often, customers tend to look for flights in the last minute, which is exploited by third party vendors who look to gain more profits in the rush hour. Skyscanner is a travel fare aggregator website and travel metasearch engine which helps users find the lowest rates from multiple travel sites, as well as instant comparisons for hotels and car hire removing the need for customers to search across different airlines for prices [1].

A metasearch engine (or aggregator) is a search tool that uses another search engine's data to produce their own results from the Internet [2]. Metasearch [3] engines take input from a user and simultaneously send out queries to third party search engines for results. Sufficient data is gathered, formatted by their ranks and presented to the users. The Skyscanner Live Pricing allows developers to access live pricing information on prices for different flights, by making requests to the Live Pricing API.

In this project, we would be querying the Skyscanner Live Pricing API using Apache HIVE and deploying the data on cloud (1-TBD & 2-TBD). Cloudmesh would be used for cloud management and the software stack deployment would be done through Ansible. We would benchmark performance of our analysis across mutilple clouds. We would be presenting a real-time visualization of the cheapest air fare and the most likely travel destination analysis in D3.js.

WORKFLOW

The project will make use of Python APIs to retrieve live flight prices information from Skyscanner and dump it in Apache HIVE database [4]. SQL Analyses are performed on this data and the results of analyses are stored in HIVE and presented in an interactive dashboard or website. The dashboard will take the onward and return journey locations, and the date of travel as inputs from users and show different price ranges for different dates commencing from the next available flight, for a period of three months. This aims to provide the users an idea as to when is the safe time to book flight tickets and beyond which date will the prices shoot up.

EXECUTION SUMMARY

The schedule for completion of this project has been outlined below:

1. Mar 06-Mar 12, 2017 Creating virtual machines on Chameleon cloud using Cloudmesh and coming up with a project proposal

2. Mar 13-Mar 19, 2017 using cloudmesh to set up Hadoop clusters and installing the required software packages

3. Mar 20-Mar 26, 2017 Fetching the data from Skyscanner API and adding it to our HIVE database

4. Mar 27-Apr 02, 2017 Running few data mining/time series models to predict the ticket prices

5. Apr 03-Apr 09, 2017 Review the work done and find out scopes for improvement and creating a benchmark report
6. Apr 10-Apr 16, 2017 Presenting the work in D3.Js in real-time as a visualization of the analysis
7. Apr 17-Apr 23, 2017 Complete the Project Report

ACKNOWLEDGEMENTS

The author thanks Professor Gregor Von Lazewski for providing us with the guidance and topics for the Project. The author also thanks the AIs of Big Data Class for providing the technical support.

REFERENCES

[1] B. J. Jansen, A. Spink, and C. Ciamacca, "An analysis of travel information searching on the we," Pennsylvania State University, Paper 10(2), 101-118, 2018, accessed: 2017-3-14. [Online]. Available: https://faculty.ist.psu.edu/jjansen/academic/jansen_travel_searching.pdf

[2] S. Berger, *Sandy Berger's Great Age Guide to Online Travel*, 1st ed. Greenwich, CT, USA: Que Publishing, 2005. [Online]. Available: https://www.amazon.com/Great-Guide-Internet-Sandy-Berger/dp/0789734427

[3] E. J. Glover, S. Lawrence1, W. P. Birmingham, and C. L. Giles, "Architecture of a metasearch engine that supports user information needsarchitecture of a metasearch engine that supports user information needs," in *Eigth International Conference on Information Knowledge Management*. ACM New York, NY, USA, 1999, pp. 210–216. [Online]. Available: http://www.researchgate.net/publication/2596239_Architecture_of_a_Metasearch_Engine_that_Supports_User_Information_Needs/file/d912f5131046ecac21.pdf

[4] L. Leverenz, "Gettingstarted with hive, apache software foundation," Web Page, Sep. 2016. [Online]. Available: https://cwiki.apache.org/confluence/display/Hive/GettingStarted

AUTHOR BIOGRAPHIES

Harshit Krishnakumar is pursuing his MSc in Data Science from Indiana University Bloomington

Karthik Anbazhagan is pursuing his MSc in Data Science from Indiana University Bloomington