Analysis of H-1B Temporary Employment-Based in Data Science Profession

JIMMY ARDIANSYAH[1,*]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *jardians@indiana.edu - S17-IR-2002*

Project Proposal, April 1, 2017

This project aims to analyze The H-1B temporary employment-based visa for Data Science related jobs in the United States. We are trying to answer the number of questions related to Data Science related jobs in America's workforce based on H-1B visa.

Keywords: Apach, Hadoop, H1B, Data Science

https://github.com/jardians/sp17-i524/blob/master/project/S17-IR-2002/report/report.pdf

INTRODUCTION

Every day tech industry executive bemoan the lack of data scientists—the people who theoretically know how to look at the data your company generates, and delve into it to derive the all-important insights we keep hearing about. It's no secret that there's a shortage of data scientists in America's workforce. Many companies look to hire overseas to help ease the domestic talent shortfall (in fact, one in three data scientists are born outside the U.S.) so understanding the ins and outs of visas is rapidly becoming a business necessity [1]. To accomplish the goals, I would like to answer question like the following:

- What is the number of petitions with Data Engineer or Scientist jobs title increasing over time?
- Which part of the US has the most Data Engineer or Scientist jobs?
- what year petitions with Data Engineer or Scientist jobs granted the most between 2011 to 2016?
- Which employers file the most petitions with Data Engineer or Scientist jobs title each year?

PLAN

Following table gives a breakdown of tasks in order to complete the project. Assuming week1 starts after submission of the proposal. These work items are high level breakdown on the tasks and may changes if needed.

Time	Work Item	Status
Week-1	Ansible Playbook Deployment	Planned
Week-2	ETL and Analysis	Planned
Week-3	Performance Measurement	Planned
Week-4	Report Creation	Planned

Fig. 1. Planned Schedule

DESIGN

I break the high-level design of the technologies used into 3 main sectios– storagee, ingestion, processing and analyzing.

- Storage refers to decision around the storage system such as HDFS or HBase
- Ingestion refers to getting data from source and loading it into Hadoop for processing.
- Analyzing refers to running various analytical queries on processed dataset to find answer and insight to the questions presented.

Technology	Purpose
Hadoop	Hadoop File System
Spark	Analysis Tool
Kafka	Data Ingestion Tool
Ansible	Platform Deployment

Fig. 2. Planned Technologies Deployment [2] [3] [4] [5]

In the design implementation, We will use HDFS [5] to store data, Kafka [3] for ingesting dataset from Kaggle.com [6] into Hadoop. We will user Spark [4] for processing and analyzing.

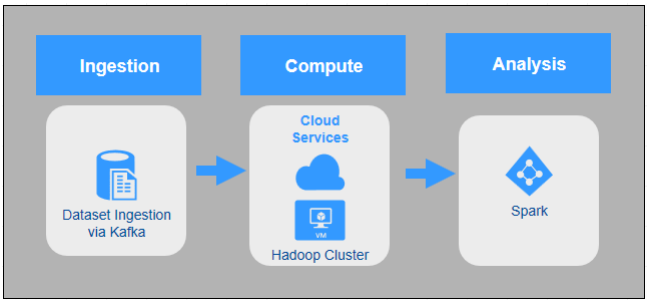

Fig. 3. Design Overview [3] [4] [5]

DATASET METADATA DESCRIPTION

The columns included in the dataset download from Kaggle [6] site are followed :

- `CASE_STATUS`: Status associated with the last significant event or decision.
- `EMPLOYER_NAME`: Name of employer submitting labor condition application.
- `SOC_NAME`: the occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.
- `JOB_TITLE`: Title of the job
- `FULL_TIME_POSITION`: Y = Full Time Position; N = Part Time Position
- `PREVAILING_WAGE`: Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer's minimum requirements for the position. YEAR: Year in which the H-1B visa petition was filed
- WORKSITE: City and State information of the foreign worker's intended area of employment
- LON: longitude of the Worksite
- LAT: latitude of the Worksite

DEPLOYMENT

Solution will be deployed using Ansible [2] playbook. Automated deployment should happen on two or more nodes cluster. Deployment script should install all necessary software along with the project code to the cluster nodes.

BENCHMARKING

We will access the performance of the Hadoop/Spark clusters deployed on difference usage, storage size and IO througput.

RESULT

Result of data analysis and benchmarking will be showcased in this section.

CONCLUSION

Using this H-1B Visa Petitions 2011-2016 data from Kaggle, we should be able build statiscal report regarding to Data Science Jobs related.

ACKNOWLEDGEMENT

This work was done as part of the course "I524: Big Data and Open Source Software Projects" at Indiana University during Spring 2017. We acknowledge our Professor Gregor Von Laszewski and all Associate Instructors for helping us and guiding us throughout this project.

REFERENCES

[1] M. Li, M. J. Wildes, and A. W. Moses, "Hiring data scientists from outside the u.s.: A primer on visas," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: https://hbr.org/2016/09/hiring-data-scientists-from-outside-the-us-a-primer-on-visas
[2] Wikipedia, "Ansible," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: https://en.wikipedia.org/wiki/Ansible
[3] Apache Kafka, "Apache kafka," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: https://kafka.apache.org/
[4] Apache Spark, "Apache spark," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: https://spark.apache.org/
[5] Wikipedia, "Apache hadoop," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hadoop
[6] S. Naribole, "H-1b visa petitions 2011-2016," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: https://www.kaggle.com/nsharan/h-1b-visa