**PROJECT PROPOSAL**

Juhi Deshpande CWID: A20430435

***Project Topic:*** Big Data Analysis on H1-B Visa Petitions 2011-16 using Hive and MapReduce

***Data Set Source:***

<https://www.kaggle.com/nsharan/h-1b-visa/download>

***Introduction:***

The H1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1B visa, an US employer must offer a job and petition for H1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/ higher education (Masters, Ph.D.) and work in a full-time position. The dataset consists of the H1-B data of the employees between the year 2011-16 with approximately 3 million records. It includes columns like case status, employer name, SOC name, job title, fulltime position status, prevailing wage, year filed, worksite.

***Test Cases needed to be implemented:***

1. Maximum, Minimum, Average based on the columns.
2. Which worksite has the popular job title?
3. Which job title has the case status as “Certified”, “Certified -withdrawn”, “Denied”, “Withdrawn”?
4. How many H1-B filers are “Certified” having the fulltime status as “Yes”?
5. What is the general trend of the employer’s H1-B status in the period 2011-16?
6. What is the general trend of the employer name with respect to the year?
7. Is the number of petitions with Data Engineer job title increasing over time?
8. Which part of the US has the most Data Engineer jobs?
9. Which industry has the most number of Data Scientist positions?
10. Which employers file the most petitions each year?

***Tools and Technologies to be used for the implementation:***

Hive, Tableau

***References:***

1. <https://github.com/yugokato/Spark-and-Kafka_IoT-Data-Processing-and-Analytics#2-format-of-sensor-data>
2. <https://nycdatascience.com/blog/student-works/h-1b-visa-petitions-exploratory-data-analysis/>
3. <https://github.com/sharan-naribole/H1B_visa_eda>

Comments:

There is no need to use the full 3 million record database. If it becomes a performance or storage problem, use only a subset of the data, say 100,000 to 500,000 records.﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿

I would like you to try various versions of the queries for each test case as follows:

(a) Try using Hive with both the Tez and the MapReduce options and record any performance differences.

(b) Do the same queries using Spark SQL and note the performance of that.

(c) See what happens as you increase the number of EMR data nodes from, say, 2 to 3 to 5

Also, as your 3 page review of the literature compare Hive with other SQL-on-Hadoop alternatives such as Impala, Presto and Apache Phoenix.