Big Data Analysis using Hadoop on H-1B_VISA

Purba Mondal

H1b-Visa

- 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 duration of stay is three years, extendable to six years. An exception to maximum length of stay applies in certain circumstance.
-

Columns in the dataset

CASE-STATUS (“Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”)

- EMPLOYEE-NAME
- SOC-NAME
- JOB-TITLE
- FULL-TIME-POSITION (Y,N)
- YEAR(2011,2012,2013,2014,2015,2016)
- WORK-SITE
- LONGITUDE
- LATITUDE

The dataset has nearly 3 million records.

- How can we manage such a big amount of data?
- How to store?
- Does it called Big Data??

What is BIG DATA?

- ‘Big Data’ is similar to ‘small data’, but bigger in size
- but having data bigger it requires different

  Approaches:– Techniques, tools and architecture
- an aim to solve new problems or old problems in a

  better way.
- Three characteristis of BigData

  Volume Velocity Variety
- To store divided into many parts and evaluate parallely.

Processing tools for analysis

- Map-reduce
- Hive
- Pig
- Sqoop

Map-reduce

- Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner
- Map takes a data in the form of pairs and returns a list of <key, value> pairs.
- Using the output of Map, sort and shuffle are applied.
- Output of sort and shuffle will be sent to reducer phase.
- In Reducer calculation part is done on list of values for unique keys
- Final output will<key, value> will be stored/displayed.

Hive

- Hive is a data ware house system for Hadoop. It runs SQL like queries called HQL (Hive query language)
- Hive Query Language is similar to SQL and gets reduced to map reduce jobs in backend.
- Hive's default database is derby.
- Hive's metastore is used to persist schema i.e. table definition(table name, columns, types), location of table files, row format of table files, storage format of files.

Pig

- Apache Pig is a high level data flow platform for execution Map Reduce programs of Hadoop.
- Pig executes in two modes: Local Mode and MapReduce Mode.
- Local Mode:
- $ pig-x local
- It executes in a single JVM and is used for development experimenting and prototyping.Local mode works on local file system.
- MapReduce Mode:
- $ pig or $ pig-xmapreduce
- The MapReduce mode is also known as Hadoop Mode. In this Pig Latin into MapReduce jobs and executes them on the cluster.

1 a) Is the number of petitions with Data Engineer job title increasing over time?

- Ussing map-reduce
- Result:

```
2011 increasing  60
2012 increasing  81
2013 increasing  151
2014 decreasing 38
2014 increasing  211
2015 increasing  394
2016 increasing  786
```

1.b) Find top 5 job titles who are having highest avg growth in applications.[ALL]

- Using pig
- Result:

```
SENIOR SYSTEMS ANALYST JC60,4255.4
SOFTWARE DEVELOPER 2,3480.8
PROJECT MANAGER 3,3233.4
SYSTEMS ANALYST JC65,2985.0
MODULE LEAD,2917.2
```

2 a) Which part of the US has the most Data Engineer jobs for each year?

- Use map-reduce
- Result:

```
SEATTLE, WASHINGTON     2011,20
SEATTLE, WASHINGTON     2014,45
SEATTLE, WASHINGTON     2015,61
```

2b) find top 5 locations in the US who have got certified visa for each year.[certified]

- Use hive
- Result:

```
2011 NEW YORK, NEW YORK23172
2011 HOUSTON, TEXAS 8184
2011 CHICAGO, ILLINOIS   5188
2011 SAN JOSE, CALIFORNIA    4713
2011 SAN FRANCISCO, CALIFORNIA  4711
```

3)Which industry(SOC_NAME) has the most number of Data Scientist positions? [certified]

- Use map reduce
- Result:

STATISTICIANS,649

COMPUTER AND INFORMATION RESEARCH SCIENTISTS,500

OPERATIONS RESEARCH ANALYSTS,426

Computer and Information Research Scientists,208

COMPUTER OCCUPATIONS, ALL OTHER,179

4)Which top 5 employers file the most petitions each year? - Case Status - ALL

- Use map-reduce
- Result:

```
TATA CONSULTANCY SERVICES LIMITED  2011,5416
MICROSOFT CORPORATION       2011,4253
DELOITTE CONSULTING LLP       2011,3621
WIPRO LIMITED     2011,3028
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION       2011,2721
INFOSYS LIMITED 2012,15818
WIPRO LIMITED     2012,7182
TATA CONSULTANCY SERVICES LIMITED  2012,6735
DELOITTE CONSULTING LLP       2012,4727
IBM INDIA PRIVATE LIMITED2012,4074
```

5) Find the most popular top 10 job positions for H1B visa applications for each year?
a) for all the applications
b) for only certified applications.

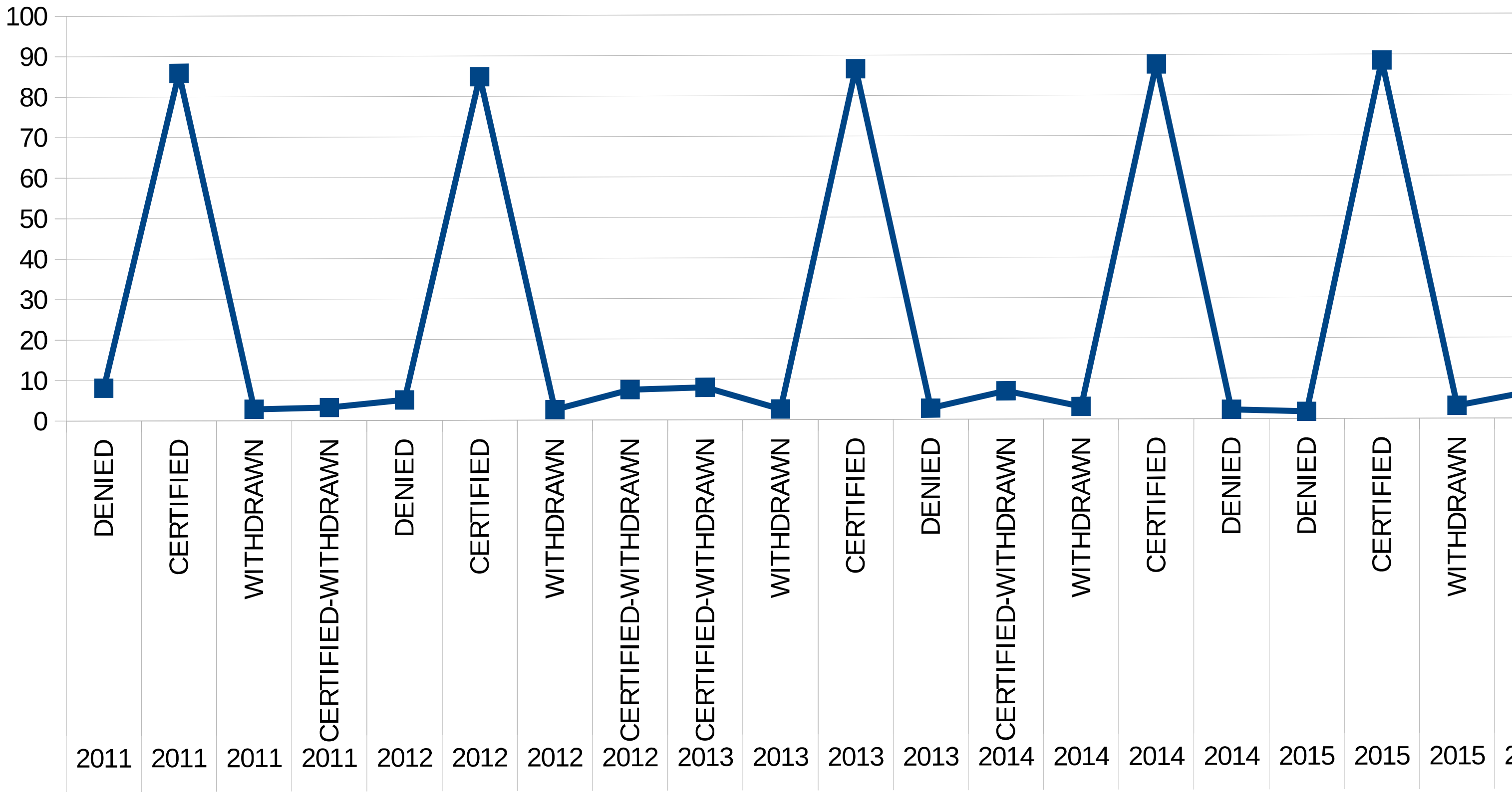
- Use hive
- Result:

```
PROGRAMMER ANALYST	2011	31799
SOFTWARE ENGINEER	2011	12763
COMPUTER PROGRAMMER	2011	8998
SYSTEMS ANALYST	2011	8644
BUSINESS ANALYST	2011	3891
COMPUTER SYSTEMS ANALYST	2011	3698
ASSISTANT PROFESSOR	2011	3467
PHYSICAL THERAPIST	2011	3377
SENIOR SOFTWARE ENGINEER	2011	2935
SENIOR CONSULTANT	2011	2798
```

6) Find the percentage and the count of each case status on total applications for each year. Create a line graph depicting the pattern of All the cases over the period of time.

- Use pig
- Result

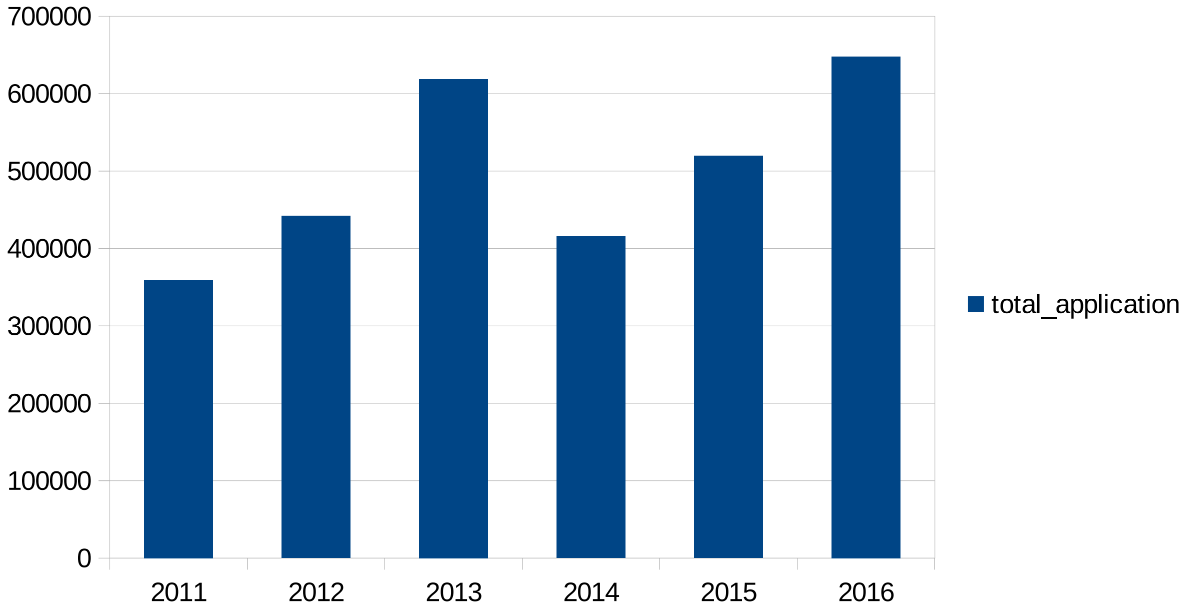
```
2011 DENIED                      8.119476
2011 CERTIFIED                   85.83175
2011 WITHDRAWN                  2.8165913
2011 CERTIFIED-WITHDRAWN        3.2321813
2012 DENIED                      5.075949
2012 CERTIFIED                  84.856125
2012 WITHDRAWN                  2.5805628
2012 CERTIFIED-WITHDRAWN         7.487362
2013 CERTIFIED-WITHDRAWN         8.014222
2013 WITHDRAWN                  2.6214957
2013 CERTIFIED                   86.61816
2013 DENIED                     2.7461243
2014 CERTIFIED-WITHDRAWN         6.998096
2014 WITHDRAWN                   3.086863
2014 CERTIFIED                  87.624245
2014 DENIED                     2.2907934
2015 DENIED                      1.765399
2015 CERTIFIED                  88.452255
2015 WITHDRAWN                  3.1443594
2015 CERTIFIED-WITHDRAWN        6.6379843
2016 CERTIFIED                   87.93507
2016 WITHDRAWN                  3.3791137
```



7) Create a bar graph to depict the number of applications for each year [All]

- Use hive
- Result:

year	total_application
2011	358767
2012	442114
2013	618727
2014	415607
2015	519427
2016	647803



8) Find the average Prevailing Wage for each Job for each Year (take part time and full time separate). Arrange the output in descending order - [Certified and Certified Withdrawn.]

- Using hive
- Result:
  TENNIS INSTRUCTOR Y 2011 CERTIFIED 15980.0
- ASSISTANT WOMAN'S SOCCER COACH Y 2011 CERTIFIED 15840.0
- HOCKEY COACH Y 2011 CERTIFIED 15810.0
- SPORTS PERFORMANCE TRAINING DIRECTOR Y 2011 CERTIFIED 15710.0
- TEMPORARY INSTRUCTOR OF INFORMATION TECHNOLOGY Y 2011 CERTIFIED 15680.0
- TEMPORARY INSTRUCTOR OF CIVIL ENGINEERING/TECHNOLO Y 2011 CERTIFIED 15680.0
- HEAD INSTRUCTOR Y 2011 CERTIFIED 15590.0
- SENIOR COACH AND DEPUTY ACADEMY ADMINISTRATOR Y 2011 CERTIFIED 15520.0
- PROFESSIONAL SKILLS COACH Y 2011 CERTIFIED 15520.0
- PROFESSIONAL SKILLS COACH Y 2011 CERTIFIED-WITHDRAWN 15520.0
- ASSISTANT DIRECTOR OF COACHES Y 2011 CERTIFIED 15420.0
- COSMETOLOGISTS Y 2011 CERTIFIED 15308.0
- TEACHER ASSISTANT, ELL (LEAD) Y 2011 CERTIFIED 15240.0
- CERTIFIED STRENGTH CONDITIONING SPECIALIST/ATHELTI Y 2011 CERTIFIED-WITHDRAWN 15170.0
- LEGISLATIVE SECRETARY Y 2011 CERTIFIED 15110.0
- LECTURER IN PHYSICS Y 2011 CERTIFIED 15080.0
- PURE TALENT TRAINING STYLIST Y 2011 CERTIFIED 15080.0
- SPECIAL CLASS ASSISTANT Y 2011 CERTIFIED 15070.0

9) Which are the employers along with the number of petitions who have the success rate more than 70% in petitions. (total petitions filed 1000 OR more than 1000) ?

- Using pig
- Result:

```
INFOSYS LIMITED 99.54055 130592
ACCENTURE LLP  99.39307 33447
TATA CONSULTANCY SERVICES LIMITED  99.337204   64726
HCL AMERICA, INC.   99.26801 22678
RELIABLE SOFTWARE RESOURCES, INC.99.14658 1992
NTT DATA, INC.  99.13251 4611
ERP ANALYSTS, INC.99.10364 1785
PATNI AMERICAS INC.  99.07907 3149
KFORCE INC. 99.06015 1596
GENPACT LLC    98.852776   1046
SMARTPLAY, INC.  98.83805 1377
SYNTEL CONSULTING INC.  98.8317   3167
CREDIT SUISSE SECURITIES (USA) LLC   98.82168 2546
MASTECH, INC., A MASTECH HOLDINGS, INC. COMPANY   98.81408 5228
GENESIS ELDERCARE REHABILITATION SERVICES, INC.   98.78788 1320
HORIZON TECHNOLOGIES INC   98.78683 1731
SYNTEL INC   98.7667   1946
```

10) Which are the job positions along with the number of petitions which have the success rate more than 70% in petitions (total petitions filed 1000 OR more than 1000)?

- Using pig
- Result:

```
MARKETING SPECIALIST	86.558136	2150
BUSINESS OPERATIONS SPECIALIST	85.97679	1034
ATTORNEY	85.90476	1050
INTERIOR DESIGNER	85.45187	1361
CIVIL ENGINEER	84.980064	2257
MARKET RESEARCH ANALYST	84.45265	8934
LAW CLERK	83.90872	1709
SALES MANAGER	83.68507	1232
OPERATIONS MANAGER	83.47339	1785
ACCOUNTANT	83.470955	14048
GRAPHIC DESIGNER	83.14741	5020
PUBLIC RELATIONS SPECIALIST	82.70326	1931
FINANCIAL MANAGER	82.40741	1080
MARKETING MANAGER	80.807175	2230
CHIEF EXECUTIVE OFFICER	80.63927	1095
GENERAL MANAGER	78.48665	1348
```

11) Export result for question no 10 to MySql database.

- Creatre a table in mysql
- Import tha data in mysql table from hadoop by sqoop
- Result:

```
 GRAPHIC DESIGNER                              |      83.1474 |      5020 |
| PUBLIC RELATIONS SPECIALIST                       |      82.7033 |
1931 |
| FINANCIAL MANAGER                             |      82.4074 |      1080 |
| MARKETING MANAGER                              |      80.8072 |      2230 |
| CHIEF EXECUTIVE OFFICER                         |      80.6393 |      1095 |
| GENERAL MANAGER                               |      78.4866 |
```

Thank you