

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

JIMMY ARDIANSYAH<sup>1,\*</sup>

<sup>1</sup>School of Informatics and Computing, Bloomington, IN 47408, U.S.A.

\*jardians@indiana.edu

Project: S17-IR-2002

This project aims to analyze The H-1B temporary employment-based visa for Data Science related occupations 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.

© 2017 <https://creativecommons.org/licenses/>. The authors verify that the text is not plagiarized.

**Keywords:** Cloud, I524

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

## 1. INTRODUCTION

The H-1B non-immigrant classification is a vehicle through which a qualified alien may seek admission to the United States on a temporary basis to work in his or her field of expertise. An H-1B petition can be filed for an alien to perform services in a specialty occupation. Prior to employing an H-1B temporary worker, the U.S. employer must first file a Labor Condition Application (LCA) [1] with Department of Labor Certification [2] and then file an H-1B petition with United States Citizenship and Immigration Services(USCIS). The LCA specifies the job, salary, length, and geographic location of employment. The employer must agree to pay the alien the greater of the actual or prevailing wage for the position [3].

To qualify as a specialty occupation, the position must meet one of the following requirements: (1) a bachelor's or higher degree or its equivalent is normally the minimum entry requirement for the position; (2) the degree requirement is common to the industry in parallel positions among similar organizations or, in the alternative, the position is so complex or unique that it can be performed only by an individual with a degree; (3) the employer normally requires a degree or its equivalent for the position; or (4) the nature of the specific duties is so specialized and complex that the knowledge required to perform the duties is usually associated with attainment of a bachelor's or higher degree

In the past 6 years, 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 [4]. To accomplish the goals, I would like to answer question like the following:

- Is it 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?

## 2. 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

### 3. DESIGN

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

- Storage refers to decision around the storage system such as HDFS or HBase [5]
- 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.

### 4. 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

### 5. DEPLOYMENT

Solution will be deployed using Ansible [7] ad-hoc commands and Linux commands. Driver script called `cc_main_driver.sh` should install all necessary software and project codes to the cluster nodes. The `cc_main_driver.sh` will copy both Python script called `cc_analyze_data.py` which analyzes and generates graphs/tables and shell script called `cc_etl_data.sh` into clusters. The `cc_main_driver.sh` will trigger `cc_etl_data.sh` to pull dataset from the web as well executes `cc_analyze_data.py` to analyze a dataset.

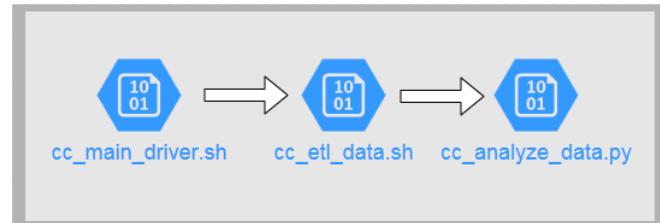


Fig. 2. Deployment Schema

### 6. BENCHMARKING

The original input dataset with approximately 3,000,000 rows (h1b\_3mRows) split into two smaller datasets: 1,000,000 rows (h1b\_1mRows) rows and 2,000,000 rows (h1b\_2mRows). Then, I executed Python script with Linux time function ( i.e: `time python ./cc_analyze_data.py`) against each of the input dataset mentioned above in order to measure both the storage size and elapsed time during the execution.

The benchmark testing on Chameleon Cloud environment revealed in the Figure-4 that elapsed processing time decreased when the number of rows in the dataset reduced. In the Figure-5, similar trend applied to disk space usage that it decreased linearly as the less number of rows need to be stored.

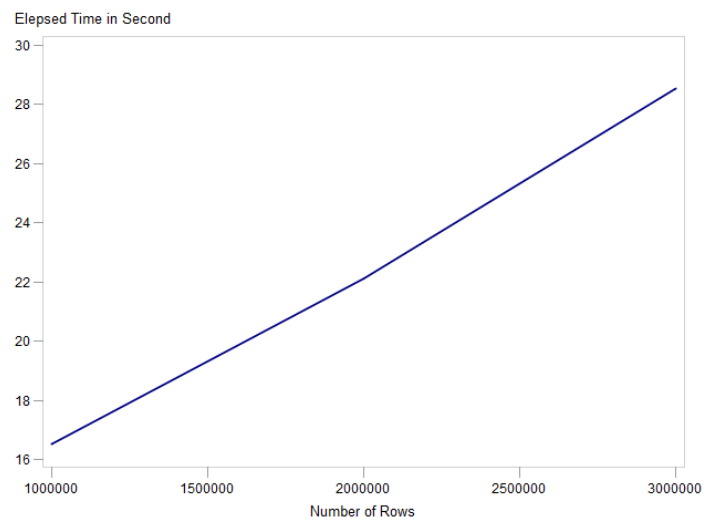


Fig. 3. Benchmark Testing - Number of Rows Vs. Elapsed Time

***** BENCHMARK *****				
DATASET	REAL	USER	SYS	DISK
3000000 (h1b_3mRows)	0m28.528s	0m15.520s	0m0.384s	470 MB
2000000 (h1b_2mRows)	0m22.112s	0m09.671s	0m0.255s	312 MB
1000000 (h1b_1mRows)	0m16.528s	0m05.528s	0m0.201s	156 MB

Fig. 4. Benchmark Testing - Number of Rows Vs. Disk Storage

### 7. DATA REPORT

General petition distribution between Fiscal Year(FY) 2011 to FY 2016, United States Citizenship and Immigration Services (USCIS) approved 2,615,623 petitions submitted by the employer on behalf of alien workers as indicated in the Figure-5.

Of the petitions approved during FY 2011-2016, a total 10,132 petitions, or .38 % were Data Science related occupations (i.e: Data Scientist, Data Analytics, Data Science Engineer, Statistician and Data Modelling) as shows in the Figure-6.

***** CASE STATUS DISTRIBUTION *****	
CERTIFIED	2615623
CERTIFIED-WITHDRAWN	202659
DENIED	94346
WITHDRAWN	89799
PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED	15
REJECTED	2
INVALIDATED	1

**Fig. 5.** General Distribution of Petition - All Jobs

***** CASE STATUS DISTRIBUTION *****	
CERTIFIED	10132
CERTIFIED-WITHDRAWN	1009
WITHDRAWN	391
DENIED	245

**Fig. 6.** General Distribution of Petition - Data Science Related Occupations

As Figure-7 indicated, petitions submitted regardless of the CASE\_STATUS and all JOB\_TITLE increased approximately 5 to 7 percent. For Data Science related petitions also increased especially in metropolitan areas such as San Francisco, New York and Menlo Park. The highest number of petition related to Data Science petitions to acquire H1-B visa was in the Fiscal Year 2016 as shown on the Figure-8.

***** PETITION PER STATE PER YEAR *****							
YEAR	2011.0	2012.0	2013.0	2014.0	2015.0	2016.0	TOTAL
STATE							
ALABAMA	1487	1572	1487	1781	1873	2053	10253
ALASKA	205	273	260	246	213	199	1396
ARIZONA	4391	5488	6389	7306	8746	9734	42054
ARKANSAS	1680	1890	2442	2329	3015	3406	14762
CALIFORNIA	65690	76402	83852	98512	115743	119741	559940
COLORADO	3630	4378	4889	5811	6827	6502	32037
CONNECTICUT	5885	7827	7447	8917	10142	10035	50253
DELAWARE	2152	2348	3172	3184	3760	3522	18138
DISTRICT OF COLUMBIA	3491	3570	3687	3727	4099	4134	22708
FLORIDA	15227	16368	15283	17644	20401	20850	105773
GEORGIA	10829	12733	13994	17728	23026	24857	103167
HAWAII	655	725	615	602	598	557	3752
IDAHO	638	644	635	609	778	887	4191
ILLINOIS	18595	22350	24510	27407	32768	35184	160814
INDIANA	3837	4340	4281	5509	6150	6399	30516
IOWA	2308	2513	2607	3168	3207	2940	16743
KANSAS	1713	2046	2233	2424	2598	2768	13782
KENTUCKY	1600	2017	1889	2170	2403	2623	12702
LOUISIANA	1615	1661	1662	1838	2702	2191	11669
MAINE	541	586	672	714	718	687	3918
MARYLAND	8544	8350	8132	9601	10891	10738	56256
MASSACHUSETTS	14720	16556	16898	19913	23488	24891	116466
MICHIGAN	8305	9918	11535	13918	18318	20970	82964
MINNESOTA	5683	6900	7194	8996	9975	9937	48685
MISSISSIPPI	648	645	668	678	792	839	4270
MISSOURI	3756	4714	4988	6200	7182	7973	34813
MONTANA	163	137	156	134	205	191	986
NEBRASKA	1089	1242	1388	1708	1815	2014	9256
NEVADA	1129	1223	1119	1231	1350	1396	7448
NEW HAMPSHIRE	1185	1526	1558	1676	2078	1906	9929
NEW JERSEY	23611	27856	29794	36783	47662	48370	214076
NEW MEXICO	782	953	854	908	1005	1039	5541
NEW YORK	41769	44512	42565	48877	55017	58670	291410
NORTH CAROLINA	7783	10411	11668	13550	17413	18847	79672
NORTH DAKOTA	403	446	469	490	575	544	2927
OHIO	8582	10426	11642	13515	16066	16344	76575
OKLAHOMA	1457	1656	1577	1846	2046	2015	10597
OREGON	2859	3103	3712	4595	4803	4718	23790
PENNSYLVANIA	12896	15552	16779	19150	22202	23380	109959
PUERTO RICO	309	311	207	209	214	202	1452
RHODE ISLAND	1038	1323	1792	2225	2881	2458	11717
SOUTH CAROLINA	1628	1795	1672	2084	2801	2952	12932
SOUTH DAKOTA	328	286	261	281	398	348	1902
TENNESSEE	3463	4544	4268	4584	5161	5652	27672

**Fig. 7.** H1-B Petition Per Year Per State - All Jobs

***** PETITION PER STATE PER YEAR *****							
YEAR	2011	2012	2013	2014	2015	2016	TOTAL
STATE							
ALABAMA	8	8	9	6	4	2	37
ARIZONA	14	8	7	14	18	24	85
ARKANSAS	4	3	1	6	25	34	73
CALIFORNIA	183	219	301	508	733	1003	2947
COLORADO	5	4	6	17	18	17	67
CONNECTICUT	36	21	25	26	37	61	206
DELAWARE	9	13	14	9	20	17	82
DISTRICT OF COLUMBIA	12	12	16	8	25	25	98
FLORIDA	23	16	22	28	59	46	194
GEORGIA	19	19	27	40	84	105	294
HAWAII	NaN	3	3	3	5	4	18
IDAHO	NaN	NaN	NaN	1	2	1	4
ILLINOIS	66	60	66	100	123	173	588
INDIANA	14	21	26	18	28	28	135
IOWA	5	7	9	9	7	11	48
KANSAS	12	15	9	11	18	16	81
KENTUCKY	6	4	2	1	4	9	26
LOUISIANA	2	1	NaN	1	5	3	12
MARYLAND	53	60	41	63	50	56	323
MASSACHUSETTS	51	78	92	123	193	249	786
MICHIGAN	15	18	24	25	40	64	186
MINNESOTA	18	15	20	21	26	29	129
MISSISSIPPI	NaN	4	1	2	2	2	11
MISSOURI	15	17	11	17	18	38	116
NA	1	NaN	NaN	NaN	NaN	NaN	1
NEBRASKA	8	5	2	6	18	9	48
NEVADA	3	9	4	5	4	11	36
NEW HAMPSHIRE	4	2	4	5	6	6	27
NEW JERSEY	96	124	142	150	168	223	903
NEW MEXICO	NaN	1	NaN	NaN	3	NaN	4

**Fig. 8.** H1-B Petition Per Year Per State - Data Science Related Occupations

As revealed in Figure-9, New Jersey, California, Massachusetts and Illinois are top locations that hires Data Science talents. Almost all technology based companies are now aware that data-driven decision making is critical if they want to succeed. As showed in the Figure-10, some of the biggest and well-known technology companies are the biggest driving force in hiring talent pool with Data Science skills.

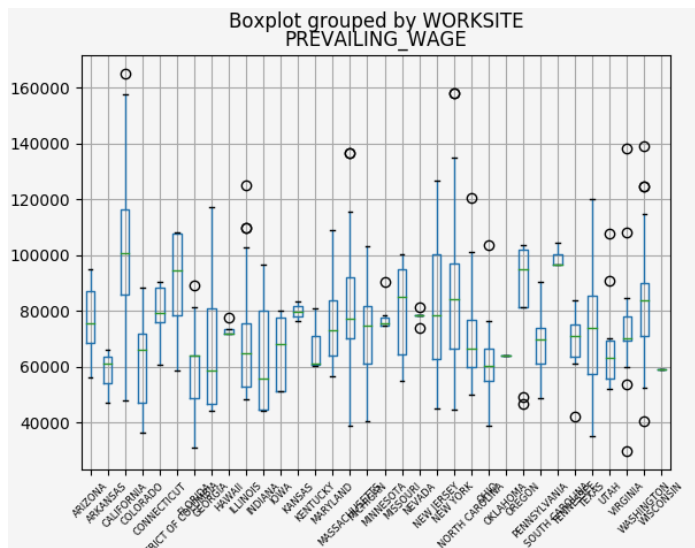
***** TOP 25 LOCATION HIRING DATA SCIENTIST *****	
SAN FRANCISCO, CALIFORNIA	332
NEW YORK, NEW YORK	224
MENLO PARK, CALIFORNIA	103
MOUNTAIN VIEW, CALIFORNIA	101
REDMOND, WASHINGTON	78
PALO ALTO, CALIFORNIA	71
SAN JOSE, CALIFORNIA	55
SUNNYVALE, CALIFORNIA	52
BOSTON, MASSACHUSETTS	45
BELLEVUE, WASHINGTON	44
CHICAGO, ILLINOIS	41
CAMBRIDGE, MASSACHUSETTS	36
SEATTLE, WASHINGTON	34
SAN MATEO, CALIFORNIA	27
AUSTIN, TEXAS	25
ATLANTA, GEORGIA	25
REDWOOD CITY, CALIFORNIA	21
SANTA MONICA, CALIFORNIA	17
HOUSTON, TEXAS	16
SANTA CLARA, CALIFORNIA	15
SAN DIEGO, CALIFORNIA	13
WASHINGTON, DISTRICT OF COLUMBIA	12
BURLINGTON, MASSACHUSETTS	12
LOS ANGELES, CALIFORNIA	12
CHARLOTTE, NORTH CAROLINA	11

**Fig. 9.** Top 25 Location Hiring Data Scientist

***** TOP 25 COMPANY HIRING DATA SCIENTIST *****	
MICROSOFT CORPORATION	139
FACEBOOK, INC.	98
UBER TECHNOLOGIES, INC.	48
TWITTER, INC.	31
AIRBNB, INC.	25
GROUPON, INC.	21
LINKEDIN CORPORATION	20
AGILONE, INC.	19
IBM CORPORATION	16
WAL-MART ASSOCIATES, INC.	15
INTUIT INC.	14
RANG TECHNOLOGIES, INC.	13
PAYPAL, INC.	12
SCHLUMBERGER TECHNOLOGY CORPORATION	11
APPLE INC.	11
STITCH FIX, INC.	10
TRIPADVISOR LLC	10
INTEL CORPORATION	9
THE NIELSEN COMPANY (US), LLC	9
LYFT, INC.	8
GOOGLE INC.	7
AMERICAN EXPRESS COMPANY	7
CLOUDWICK TECHNOLOGIES INC.	7
ICUBE CONSULTANCY SERVICES, INC	7
ZILLOW, INC.	7

**Fig. 10.** Top 25 Companies Hiring Data Scientist

As shown in Figure-11, for occupations in Data Science field, the median annual compensation reported by employers of H-1B workers between FY 2011 to FY 2016 was ranged from a low of \$40,000 to a high \$110,000 which depends on geological location.



**Fig. 11.** Data Scientist Wage Across States

## 8. CONCLUSION

Overall, there is compelling evidence that the H-1B visa program is helping to alleviate acute shortages in Data Science occupations since the number of petitions submitted increased linearly from FY 2011 to FY 2016. Armed with such information, as well as indicators presented above, Data Science occupation mostly concentrated in large metropolitan areas. Well-known technology companies has indicated hired professional with Data Science skill sets.

## 9. 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] Wikipedia, "Labor condition application," Web Page, Apr. 2017, accessed: 2017-04-20. [Online]. Available: [https://en.wikipedia.org/wiki/Labor\\_Condition\\_Application](https://en.wikipedia.org/wiki/Labor_Condition_Application)
- [2] USCIS, "Labor certification," Web Page, Apr. 2017, accessed: 2017-04-20. [Online]. Available: <https://www.uscis.gov/tools/glossary/labor-certification>
- [3] Wikipedia, "H-1b visa," Web Page, Apr. 2017, accessed: 2017-04-20. [Online]. Available: [https://en.wikipedia.org/wiki/H-1B\\_visa](https://en.wikipedia.org/wiki/H-1B_visa)
- [4] 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>
- [5] Wikipedia, "Apache hadoop," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: [https://en.wikipedia.org/wiki/Apache\\_Hadoop](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>
- [7] Wikipedia, "Ansible," Web Page, Mar. 2017, accessed: 2017-03-20. [Online]. Available: <https://en.wikipedia.org/wiki/Ansible>