Issues in Informing Science + Information Technology

An Official Publication of the Informing Science Institute InformingScience.org

IISIT.org

Volume 15, 2018

CHANGING PARADIGMS OF TECHNICAL SKILLS FOR DATA ENGINEERS

Robert T Mason Regis University, Denver, USA RMASON@REGIS.EDU

ABSTRACT

Aim/Purpose	This paper investigates the changing paradigms for technical skills that are needed by Data Engineers in 2018.
Background	A decade ago, data engineers needed technical skills for Relational Database Management Systems (RDBMS), such as Oracle and Microsoft SQL Server. With the advent of Hadoop and NoSQL Databases in recent years, Data Engineers require new skills to support the large distributed datastores (Big Data) that currently exist. Job demand for Data Scientists and Data Engineers has increased over the last five years.
Methodology	This research methodology leveraged the Pig programming language that used MapReduce software located on the Amazon Web Services (AWS) Cloud. Data was collected from 100 Indeed.com job advertisements during July of 2017 and then was uploaded to the AWS Cloud. Using MapReduce, phrases/words were counted and then sorted. The sorted phrase / word counts were then leveraged to create the list of the 20 top skills needed by a Data Engineer based on the job advertisements. This list was compared to the 20 top skills for a Data Engineer presented by Stitch that surveyed 6,500 Data Engineers in 2016.
Contribution	This paper presents a list of the 20 top technical skills required by a Data Engineer.
Keywords	data engineer, data scientist, data science

INTRODUCTION

The Technical Committee on Data Engineering (TCDE, 2018) of the IEEE Computer Society focuses on the variety topics that include data design, data development, data management, and utilization of information systems. Data topics can range from data security, databases, cloud computing, data models, data integration, and data quality. There are peer-reviewed, international open-access journals that focus on Data Engineering. The Data Science and Engineering (DSE, 2018) journal focuses on four main areas: 1) big data, 2) information extraction from big data, 3) theory behind

Accepting Editor: Eli Cohen | Received: November 30, 2017 | Revised: March 18, 2018 | Accepted: April 2, 2018.
Cite as: Mason, R. T. (2018). Changing paradigms of technical skills for data engineers. *Issues in Informing Science and Information Technology, 15,* 35-42. https://doi.org/10.28945/4033

processing large volumes of data, and 4) big data analytics. A decade ago, Data Engineers relied heavily on the technology of Relational Database Management Systems (RDBMS). For example, Grisham, Krasner, and Perry D. (2006) described an Empirical Software Engineering Lab (ESEL) that introduced Relational Database concepts to students with hands-on learning that they called "Data Engineering Education with Real-World Projects." However, as seismic improvements occurred for the processing of large distributed datasets, big data analytics has moved into the forefront of the IT industry. As a result, the definition for Data Engineering has broadened and evolved to include newer technology that supports the distributed processing of very large amounts of data (e.g., Hadoop Ecosystem and NoSQL Databases). This paper examines the technical skills that are needed to work as a Data Engineer in today's rapidly changing technical environment. Research is presented that reviews 100 job postings for Data Engineers from Indeed (2017) during the month of July 2017 and then ranks the technical skills in order of importance. The results are compared to earlier research by Stitch (2016) that ranked the top technical skills for Data Engineers in 2016 using LinkedIn (2018) to survey 6,500 people that identified themselves as Data Engineers. Data Scientists and Data Engineers are in high demand according to a list of the 50 best jobs in America published by Glassdoor (2018). Data Scientist is ranked as the number one best job in America for 2018 and Data Engineer is ranked as the 33rd best job with a medium base salary of $100,000 and 2,816 job openings.

INCREASING DEMAND FOR DATA ENGINEERS

The number of available jobs for Data Scientists and Data Engineers has been increasing as shown in Figure 1. Although the number of available jobs dipped slightly in 2016, the trend appears to be rebounding for 2017. IBM estimates that the jobs for data engineers, data scientists and data developers will reach nearly 700,000 openings by 2020 (Columbus, 2017).

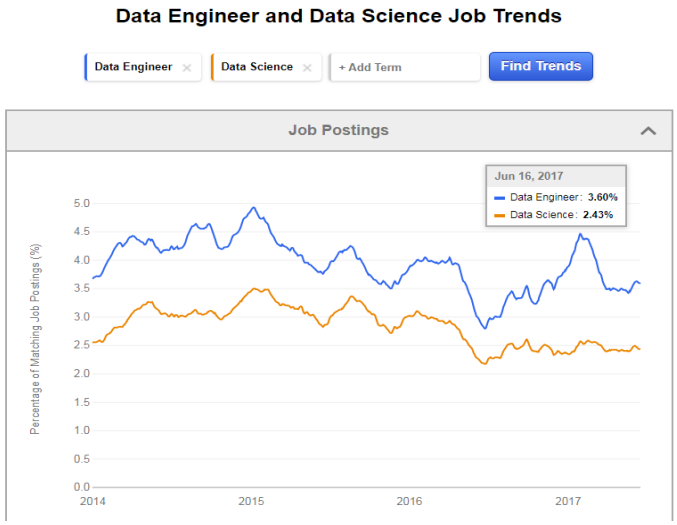

Figure 1. Job Trends from Indeed.com for Data Science and Data Engineering.

The Stitch (2016) report that was published in March of 2016 shows a rapid increase in Data Engineering jobs over the five years from 2010 to 2015 as shown in Figure 2. As mentioned previously, Stitch (2016) ranked the top technical skills for Data Engineers in 2016 using a LinkedIn (2018) survey of 6,500 people that identified themselves as Data Engineers in publicly visible personal and company profiles, skills, and professional experiences. LinkedIn is a popular business-oriented networking site, the emphasis of the site is to connect people that share work related interests. Stitch is a technology company that facilitates the building of data pipelines using various software products and conducts research on the topics of Data Science and Data Engineering. Stitch sent surveys to 6,500 participants and asked them to rank the top skills needed by Data Engineers. The survey re-

sults included 30,000 professional experiences and the respondents worked at 3,400 different companies throughout the world. The results were combined to a create a list of the top 20 skills needed by a Data Engineer. According to the Stitch Report, the top 5 skills needed by Data Engineers were SQL, Java, Python, Hadoop, and Linux.

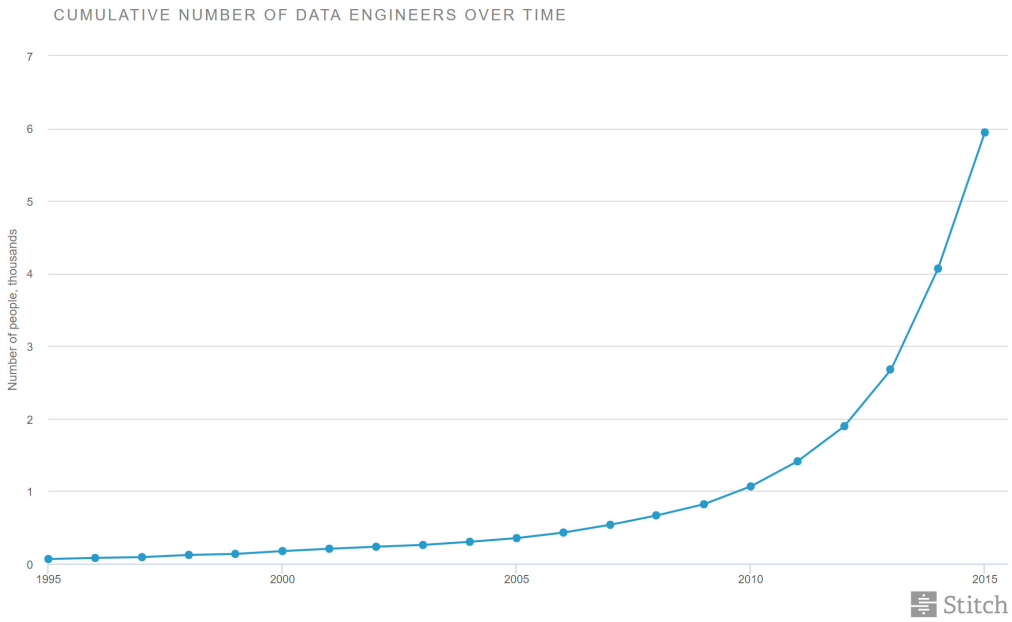

Figure 2. Cumulative growth in Data Engineers over the last two decades.

What Are the Job Responsibilities and/or Technical Skills Needed by Data Engineers?

Although the number of jobs for Data Engineering is increasing, there is very little up-to-date peer-reviewed research available that defines the job responsibilities and/or technical skills of a Data Engineer (Saltz, Yilmazel, & Yilmazel, 2016). Below is a list of the job responsibilities for a Data Engineer by Tamir, Miller, and Gagliardi (2015):

- Extract, clean, and integrate data (wrangling)
- Bridge between the data science models and production systems
- Implement machine learning & computational algorithms at scale (e.g., Data Science)
- Put the right data system to work for the job at hand. Meaning they need a deep understanding of transactional ACID (relational) databases along with a growing variety of NoSQL (BASE) databases including JSON document, graph, column stores, and partitioned row.
- Demonstrate a deep understanding of distributed computing (e.g. Hadoop ecosystem and NoSQL databases) considerations for consistency, scalability, and security.
- Protect customer privacy and anonymity.

Another list of job responsibilities for a Data Engineer is provided by Degree Prospects (2017) that hosts a website dedicated to Master of Science degree programs in Data Engineering.

- Design, construct, install, test and maintain highly scalable data management systems
- Ensure systems meet business requirements and industry practices
- Build high-performance algorithms, prototypes, predictive models, and proof of concepts

- Research opportunities for data acquisition and new uses for existing data
- Develop data set processes for data modeling, mining, and production
- Integrate new data management technologies and software engineering tools into existing structures
- Create custom software components (e.g., specialized UDFs) and analytics applications
- Employ a variety of languages and tools (e.g., scripting languages) to marry systems together
- Install and update disaster recovery procedures
- Recommend ways to improve data reliability, efficiency and quality
- Collaborate with data architects, modelers and IT team members on project goals

This list of job responsibilities doesn't specifically mention the necessity for knowledge about distributed computing. However, the website (Degree Prospects, 2017) does provide more details about the technical skills that are needed by Data Engineers. As shown below, NoSQL and Hadoop-based technologies are two types of technologies listed by Degree Prospects and use distributed computing.

Technical skills required by a Data Engineer *(DEGREE PROSPECTS, 2017)*

- Statistical analysis and modeling
- Database architectures
- Hadoop-based technologies (e.g. MapReduce, Hive and Pig)
- SQL-based technologies (e.g. PostgreSQL and MySQL)
- NoSQL technologies (e.g. Cassandra and MongoDB)
- Data modeling tools (e.g. ERWin, Enterprise Architect and Visio)
- Python, C/C++ Java, Perl
- MatLab, SAS, R
- Data warehousing solutions
- Predictive modeling, NLP and text analysis
- Machine learning
- Data mining
- UNIX, Linux, Solaris and MS Windows

RESEARCH FOR RANKING TECHNICAL SKILLS FOR DATA ENGINEERS

Stitch (2016) surveyed 6,500 people that identified themselves as Data Engineers on the Linkedin website which is a business networking social media site. Based on the survey results, Stitch then ranked the top 20 technical skills that are needed by Data Engineers. This is similar to the list of skills shown above by Degree Prospects (2017), however there is no indication that the Degree Prospect list is ranked (e.g., in a particular order). The Stitch top 20 technical skills are used to compare to the research results discussed in this paper.

To verify the Stitch ranked list of technical skills, an alternative data source was selected for this research project. The Indeed (2017) website was used a data source to collect and analyze 100 job advertisements for Data Engineering positions during July of 2017. Indeed is a job site aggregator that was launched from Austin, Texas in 2004. Indeed contains job listings from thousands of job boards, associations, company career pages, etc. Indeed is available in over 60 countries and 28 languages and it has become the highest-traffic job site in the USA (Schonfeld, 2010).

RESEARCH METHODOLOGY

One hundred Data Engineering job postings were collected and analyzed from Indeed during July 2017. The job postings were pasted into a text file and then related words were grouped together to make phrases. For example, the word "data" was put into context with other related words to form phrases such as "Big Data", "Data Architecture" and "Data Engineering". A text editor was used for

this task and the find/replace functionality of the text editor proved to be very useful for this project. After making phrases, the text file was uploaded to the Amazon cloud (AWS). The AWS cloud is a "pay as you go" environment that allows registered users to run programs inexpensively. For example, the cost of running the program (for this research) in the AWS cloud cost less 1 US dollar. A simple Apache Pig program (2 lines of code) was written which called MapReduce software to count the phrases and words within the text file. Apache Pig is a high-level platform for Apache Hadoop that can be used to create programs. These Pig programs can use MapReduce which is a programming library written with the Java Programming language and is used for processing big datasets on a cluster of server machines using a parallel, distributed algorithm. In this case, the size of the of data file was small and only took a few seconds to process.

The resulting phrases/words with occurrence counts were download to a Personal Computer (PC) and then were loaded into an Excel spreadsheet. Using a spreadsheet enabled the phrases/words to be sorted by occurrence count and then allowed the filtering out of irrelevant words. Another task to prepare the data involved the combination phrases or words that were synonymous. For example, the occurrence count for the acronym ELT and the occurrence count for the acronym ETL were added together to make an overall ELT/ETL occurrence count. ETL is a Data Warehousing acronym for Extracting, Transforming and Loading data. This task required knowledge of the subject area. Also, some words were counted in lower case and then the same word was also counted in mixed or upper case, thus producing two or three occurrence counts for the same word. These different counts were added together to make an overall occurrence count for the word (e.g., word occurrence counts for Python and python were added together). Finally, the Indeed occurrence counts were sorted to allow for the identification of a list of the top 20 technical skills needed by a Data Engineer.

RESEARCH RESULTS

Table 1 shows a comparison of the Stitch (2016) report and the Indeed (2017) research. Matching words are shown in bold text.

Table 1. Comparison of the Indeed research to the Stitch report.

Ranking Order	*Stitch Report (2016) -Ranked Technical Skills*	*Indeed Research (2017) - Ranked Technical Skills*
1	**SQL (Relational RDBMS)**	**Hadoop/HDFS**
2	**Java (programming)**	**Big Data**
3	**Python ((programming)**	Design Skills
4	**Hadoop/HDFS**	**SQL (Relational RDBMS)**
5	Linux	**Python (programming)**
6	Databases	**Java (programming)**
7	MySQL	**ETL/ELT (Extract Transform and Load data)**
8	**Data Warehousing**	Spark (Fast Processing Engine for Hadoop)
9	**Javascript**	Amazon Web Services (AWS)
10	**C++**	**Programming**

Ranking Order	*Stitch Report (2016) -Ranked Technical Skills*	*Indeed Research (2017) - Ranked Technical Skills*
11	Business Intelligence	**Hive**
12	Oracle	Data Modeling
13	Microsoft SQL Server	Kafka (move and load data)
14	Data Analysis	Scala
15	**ETL/ELT (Extract Transform and Load data)**	**Data Warehouse**
16	**Big Data**	Cloud Computing
17	**Software Development**	Data Pipelines
18	Unix	Application Program Interfaces (APIs)
19	**C**	AWS Redshift Data Warehousing
20	**Hive**	**Scripting languages**

DISCUSSION

Eleven of the twenty phrases/words that are highlighted in bold (Table 1) match in both lists. The author considered C, C++ and Java a match to the broader category of Programing in the job postings data. Although the ranked order of the two lists did not match, the top 5 ranked technical skills for both lists are almost the same. Therefore, the reader of this paper might consider the skills of SQL, Python, Hadoop/HDFS to be very important technical skills for a Data Engineer. Although the programming language of R is very popular with Data Scientists and was on the job postings result list, it did not make the top 20 skills for Data Engineering in the Stitch report. The R programming language is oriented towards analytical processing (e.g., used by Data Scientists), whereas the Python language is a scripting and object-oriented language that facilitates the creation of Data Pipelines (e.g., used by Data Engineers).

Because the data was collected one year apart and from very different data sources, the timing of the data collection and the different data sources can account for some of the differences in the ranked lists. The source of the data for the Stitch report was much richer since it was based on a survey of 6,500 Data Engineers. Obviously, a much larger sample size lends more credibility to results of the Stitch report and less credibility to results from this research. However, although the sample size for the job postings research was much smaller, the research results are useful because they add to the overall body of knowledge about required skills for Data Engineers.

It is worth mentioning that the job postings research introduced the words/phrases of *Design Skills, Spark, AWS (Amazon Web Services), Data Modeling, Kafta, Scala, Cloud Computing, Data Pipelines, APIs and AWS Redshift Data Warehousing*. These words may have been included in the Stitch research raw data, however they were not included in the final report.

CONCLUSION

Available jobs for Data Engineers and Data Scientists has grown rapidly over the last five years. Data Scientists and Data Engineers are listed in the Best 50 Jobs in America. IBM predicts more job

growth for these two positions over the next few years. A goal of this research was to add to the body of research for the skills needed by Data Engineers. Skills were compared to the research presented in the Stitch report which was based on a survey of 6,500 Data Engineers. For this research, one hundred job advertisements were collected from the Indeed job aggregator website in July of 2017. Phrases/words were counted using an AWS Cloud Pig program that used MapReduce. Although the list results matched with 11 technical skills from both lists, new words/phrases were introduced from the job postings ranked word list. The data was collected one year apart and from very different data sources. Therefore, the timing of the data collection and the different data sources may account for some of the differences in the two lists. The research from Stitch has more credibility because of large sample size. However, this job postings research provides additional information that is worth considering when reviewing skills for Data Engineers.

Future Work

Future work for this research could involve using a larger sample size of job postings and the collection of data from other job listing websites. Although the process of making phrases from words in this sample data was done manually using a text editor, this process could be automated with a program that would use a data dictionary (ontology) of Data Engineering, Data Science and Information Technology terms. A program could provide a more consistent and faster data preparation process. Alternative analysis tools could be applied to the research data, perhaps statistical analysis to determine the consistency of the results. A follow-up survey to Data Engineers could provide confirmation for this research methodology (collecting job postings) and the findings.

References

Columbus, L. (2017). *IBM predicts demand for data scientists will soar 28% by 2020.* Retrieved from https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#3edba2f47e3b

Degree Prospects. (2017). *Data engineer responsibilities.* Retrieved from http://www.mastersindatascience.org/careers/data-engineer/

DSE. (2018). *Data science and engineering.* Retrieved from https://link.springer.com/journal/41019

Glassdoor. (2018). *50 best jobs in America.* Retrieved from https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Grisham, P., Krasner, H., & Perry D. (2006). Data engineering education with real-world projects. *SIGCSE Bull.* 38(2), 64-68. https://doi.org/10.1145/1138403.1138435

Indeed. (2017). *Data engineering job advertisements.* Retrieved from https://www.indeed.com/

LinkedIn. (2018). *Business networking site.* Retrieved from https://www.linkedin.com/

Schonfeld, E. (2010). *Indeed slips past Monster, Now largest job site by unique visitors.* Retrieved from https://techcrunch.com/2010/11/17/indeed-monster-largest-job-site/

Saltz, J., Yilmazel, S., & Yilmazel, O. (2016). Not all software engineers can become good data engineers. *Proceedings from the 2016 IEEE International Conference on Big Data.* Washington, DC, USA: IEEE. https://doi.org/10.1109/BigData.2016.7840939

Stitch. (2017). *The state of data engineering.* Retrieved from https://www.stitchdata.com/resources/reports/the-state-of-data-engineering/?thanks=true

Tamir, M., Miller, S., & Gagliardi, A., (2015). *The data engineer.* Retrieved from http://dx.doi.org/10.2139/ssrn.2762013

TCDE. (2018). *IEEE Technical Committee on Data Engineering.* Retrieved from https://www.computer.org/web/tandc/tcde

BIOGRAPHY

Bob Mason is an associate professor at Regis University College of Computer & Information Sciences (CC&IS). He is currently developing a Specialization in Data Engineering for the M.S. in Data Science degree. Prior to accepting this position with Regis as a full-time professor, Bob was employed by various Fortune 500 companies for 25 years as a Data Architect, Database Administrator, Manager and Software Engineer. He worked as a part-time affiliate faculty member at Regis for 10 years and primarily taught database technologies courses. Bob completed his Ph.D. in 2011 from Nova Southeastern University located in Davie, FL, USA.