# DATA ENGINEERING SKILLS AND COMPETENCIES BASED ON INDUSTRY DEMAND

*Ashraf Shirani, Lucas College & Graduate School of Business, San Jose State University, One Washington Square, San Jose, CA 95192, ashraf.shirani@sjsu.edu*

## ABSTRACT

The objective of this study is to identify skillsets and competencies expected of data engineers from the current industry demand and other sources. Recent position announcements for data engineering jobs posted on LinkedIn by major US organizations were collected and analyzed. Text analytics techniques were then applied to the gathered position descriptions in order to discern foundational and core skillsets in the field. The findings of this study would hopefully provide guidance for information systems programs and faculty to augment and update their data related curricula, especially in data engineering.

**Keywords:** Data Engineering; ETL; Data Science; Business Analytics

## INTRODUCTION

Estimates of the amount, variety, and speed at which data is being generated, ingested, and consumed are astoundingly large today and increasing at increasing rates. It is estimated that 2.5 quintillion bytes of data is created every day, and about 90% of the total worldwide data today was created in the past two years [5]. The enabling technologies and business processes and practices behind data creation, analysis, and use are also evolving and morphing rapidly. These trends, in turn, require that information systems and technology workforce must engage in continuous up-skilling. And in order to be job-ready, information systems graduates be trained to meet industry expectations, not just in the foundational data skills but also the skills and competencies necessary for one or more emerging data specializations. One such data specialization is that of data engineering.

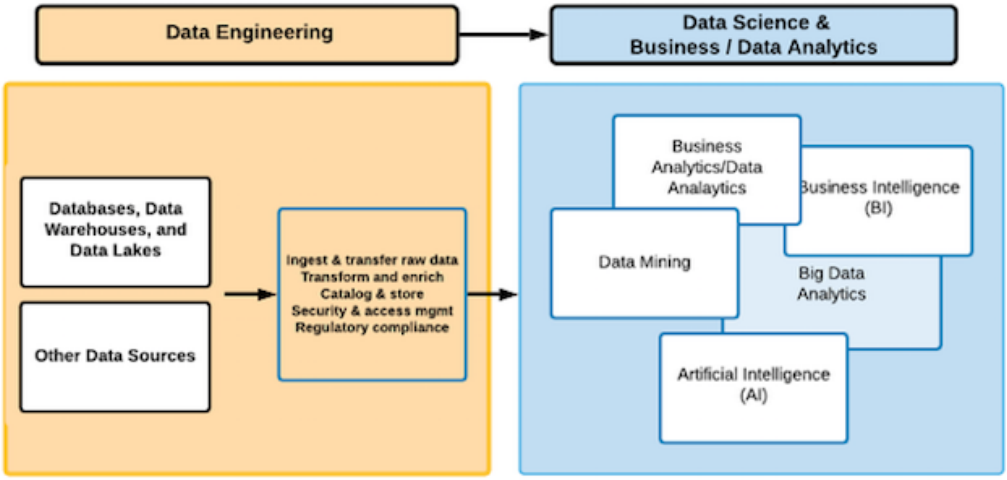
Data engineering skills are becoming increasingly essential in supporting the work in multiple facets of data science and business analytics. While data scientists and business analysts find insights and actionable information from data, "data engineers are concerned with the production readiness of that data and all that comes with it: formats, scaling, resilience, security, and more" [4].

Today, besides the traditional database administration (DBA) specialization, a number of other specialized data roles exist –including those of a data scientist, business analyst, data analyst, business intelligence analyst, data warehouse administrator, and more. *Data engineering* is a relatively new and emerging addition to this list. In this paper, the author reports the results of a study that aims to identify specific skillsets associated with the role of a data engineer. Primary focus of this study is on data engineering *technical skills* rather than the entire set of competencies that may be expected of a data engineer. The findings are expected to be of practical use in curriculum development and to provide pedagogical guidance in undergraduate information systems and technology programs.

### The Data Engineering Landscape

As is often the case with emerging technologies and disciplines, data engineering has been defined variously. The role of a data engineer is typically explained in terms of its relationship with that of a data scientist or data analyst in that the former helps extract raw data, often from diverse and disparate sources, and clean and prepare it for the latter to build and deploy machine learning models and perform analytics [8]. While we use the term *data engineering* (DE) consistently throughout this paper, it should be noted that there are other (though less common) synonymous terms associated with the same set of activities as those of a data engineer, including *analytics engineering* and *machine learning engineering* [2]. For example, Carroll [3] describes the role of analytics engineers as this: "While a data analyst spends their time analyzing data, an analytics engineer spends their time transforming, testing, deploying, and documenting data." Figure 1 depicts a high-level view of the data engineering landscape.

**Figure 1: The Data Engineering Landscape**




Data engineering, however, is not an entirely new discipline. In general terms, the same set of activities have been performed with respect to data warehousing since the 1990s and have been known collectively as *extract, transform, and loa*d (ETL) [9]. Data warehouses store integrated current and historic data from multiple operational data sources and their schema design is optimized for query processing and data visualization. ETL is the process of extracting raw data from an original source, transforming it into the format that matches the data warehouse schema, and uploading it to the warehouse [7]. Data warehouses have traditionally stored structured data, primarily from internal transactional databases. In recent years, however, massive growth in the production and use of non-relational, less structured (NoSQL) data has prompted organizations to efficiently store and process such data along with structured data. This, in turn, has led to data repositories known as *data lakes* along with a variation of the ETL process –*extract, load, and transform*, or ELT. The ELT process allows ingesting raw data regardless of its structure, and transforming it later as needed by the end users. Both ETL and ELT co-exist today, typically in cloud data warehouses. These and other trends in data production and consumption, along with market forces and rapidly evolving technology require a corresponding set of data engineering skills to support the ETL/ELT processes.

Methodology and data sources for this study are described in the next section, followed by discussion of the results, and concluding remarks.

## METHODOLOGY

### Data

Data for this study was obtained from LinkedIn, a professional career and employment-oriented networking company and website. Descriptions for the data engineering jobs posted on the LinkedIn website during December 2021 were copied and pasted into a text document. In order to ascertain data engineering technical skills, the focus of data collection was on job descriptions, qualifications, experience, responsibilities, and required skills posted for each job listing. A total of 25 listings from 22 companies were gathered (three companies had two different position announcements). Also, the focus was on postings by major technology and business enterprises as these companies are often leaders in defining the toolset and practices of an emerging field such as data engineering. Table 1 below shows the companies whose data engineering job postings were analyzed for this study.

**Table 1: Sources of Data**

| | | | |
|---|---|---|---|
| Amazon | Google | Moody's Analytics | TikTok |
| Accenture | HCL Technologies | Netflix | Twitch |
| Anheuser Busch | Jefferson Frank | NIO | Visa |
| Apple | League | Optello | Zoom |
| Electronic Arts | LinkedIn | Rappi | |
| Facebook | Lyft | Tesla | |

### Data Analysis

Text processing was don on the text document that contained copies of all of the data engineering job descriptions. Specifically, the following text processing tasks were performed using a data mining software, *RapidMiner*:

- *Tokenization*: the entire text document was split into tokens (words in this case).
- *Stopwords*: words such as articles, prepositions or conjunctions, like ‘a’, ‘the’, ‘and’, etc. were removed from the document. *Stopwords* are needed for grammatical correctness but they do not add much to overall understanding of the text.
- *Case conversion*: tokens were all converted into lower case letters so the same word may not be interpreted and counted multiple times due to its case.
- *n-Grams*: two-word n-grams were created in order to capture skill-relevant terms such *data science* and *data pipeline* rather than the words data, science, or pipeline alone.
- *TF-IDF (*term frequency – inverse document frequency): TF-IDF is meant to signify the importance of a word to a document. Its value increases proportionately to the frequency of a word in the document but is offset by the count of documents in which the word appears. In the present study, though, all the text was in a single document and so TF-DIF is simply a term frequency (TF) matrix.
- TF was sorted in descending order.
- TF was then filtered to remove words that had a count of five or less. Doing so was deemed appropriate since a a word that appears at most in five times out of 25 job postings is not a good indicator of the overall skill requirements.
- Finally, the terms that did not refer to a skill were removed manually.

## RESULTS AND DISCUSSION

Table 2 shows top 23 skills by count in the overall job postings.

**Table 2: Top 23 Skills by Frequency**

| Skill | Count | Skill | Count | Skill | Count |
|---|---|---|---|---|---|
| Data pipelines | 35 | Apache Hadoop | 10 | Apache Hive | 7 |
| Analytics | 35 | Apache | 9 | Streaming | 7 |
| Python | 28 | Cloud | 9 | Warehouse | 7 |
| Data science | 24 | Distributed systems | 9 | Batch processing | 6 |
| ETL | 24 | Frameworks | 9 | Java | 6 |
| Apache Spark | 18 | Programming language | 8 | Relational | 6 |
| Databases | 12 | Machine learning | 8 | Snowflake | 6 |
| Apache Airflow | 11 | scripting | 8 | | |

As expected, the term-frequency matrix alone does not portray a complete picture of the data engineering skillset. However, term-frequency information augmented with a reading by the author of the full text of job descriptions and industry blogs on the subject, revealed the following trends.

### Data Engineering Core Skills

*Traditional ETL Skillset*: Despite the advent of more recent platforms and technology in the data ecosystem, the traditional ETL skills remain relevant and in use by the industry. These include, for example, dimensional modeling, designing; building, and working with data warehouses and data marts; and working with the proprietary and open-source data transformation and integration tools.

*Big Data Engineering*: Among other big data technologies in the Hadoop ecosystem, today Apache Spark is a popular framework for big data processing. Some of the components of Apache Spark include Spark SQL, Spark streaming, machine learning library (MLlib), graph computation (GraphX), and Spark R.

*Streaming Data Processing*: The "velocity" part of the popular big data definition relates to the ability to capture and process streaming data in real- or near real-time. Apache Kafka is used to build real-time streaming data pipelines and applications that adapt to the data streams. [Amazon AWS, 2022]

*Data Pipelines*: The process of moving data from one repository to another can be orchestrated and scheduled with tools such as Apache Airflow, Luigi, Oozie, and more. Skills in using these and other such technologies are important in the ETL/ELT process.

*Proprietary Data Warehousing and Data Lake Platforms*: In addition to the open-source software and frameworks, a number of major data warehousing and data lake vendors are used by many companies. Major vendors in this space include Snowflake, Amazon AWS, Microsoft Azure, and Google GCP.

### Foundational Knowledge and Skills

Data engineering is a fast-evolving field and its core skillsets are also in flux. Foundational knowledge and skills, on the other hand, are relatively stable and serve as prerequisites for the more advanced core skills. The following knowledge areas, skills, and tools would provide a good set of foundational skills for data engineering education.

- Relational and non-relational (NoSQL) database concepts and data modeling
- Data warehousing and dimensional modeling; data lakes
- Programming languages including SQL and Python (and optionally, Java)
- Cloud computing essentials
- Tools: Shell scripting and essential Linux commands

## CONCLUSION

Data engineers are in demand and are highly paid data professionals [6]. This paper reported results of a study that analyzed job descriptions in 25 job ads posted on LinkedIn. The objective of the study was to identify skillsets required for the role of a data engineer. A text analysis of the job postings revealed a set of core and foundational skills. This, however, provided a partial view of the data engineering landscape. Reading the full text of the job postings and a review of literature on the current industry practices provided a more complete view of the skillsets reported in the last section of the paper.

## REFERENCES


[1] Amazon AWS.What is apache Kafka? Retrieved from https://aws.amazon.com/msk/what-is-kafka/

[2] Ayan, G. (2021). Skills of a data engineer. Retrieved from https://medium.com/slalom-australia/skills-of-a-data-engineer-241c4f615990

[3] Carroll, C. (2019). What is analytics engineering? Retrieved from https://www.getdbt.com/what-is-analytics-engineering/

[4] Furbush, J. (2018). Data engineering: A quick and simple definition. Retrieved from https://www.oreilly.com/content/data-engineering-a-quick-and-simple-definition/

[5] Johnson, C. (2021). How much data is produced every day 2021? Retrieved from https://www.the-next-tech.com/blockchain-technology/how-much-data-is-produced-every-day-2019/

[6] Indeed Editorial Team. Top 8 skills you need to become a data engineer. Retrieved from https://www.indeed.com/career-advice/resumes-cover-letters/data-engineer-skills

[7] Krevitt, D. A love letter to ETL tools. Retrieved from https://www.getdbt.com/analytics-engineering/etl-tools-a-love-letter/

[8] Oram, A. (2020). *The evolving role of the data engineer* O'Reilly Media.

[9] Shirani, A. & Roldan, M. L. (2009). Data warehousing and business intelligence skills for information systems graduates: Analysis based on marketplace demand . *Issues in Information Systems, 10*(2), 333.