Issues in Information Systems
Volume 23, Issue 3, pp. 221-230, 2022

DOI: https://doi.org/10.48009/3_iis_2022_118

Towards advanced data skills for information systems graduates

Ashraf Shirani, *San Jose State University, ashraf.shirani@sjsu.edu*

Abstract

This study aims to explore advanced data skills –in particular technical skills expected of a data engineering (DE) professional. Descriptions and requirements for the DE jobs, one each from 50 US companies, were collected from three job sites. An automated text analysis was performed on the job descriptions to extract a list of technical skills and competencies required for the DE job in the industry. The extracted information was then manually synthesized and categorized by the author. Findings of the study are intended to be of use by information systems graduates looking to pursue advanced data education and for curricula committees in offering advanced data courses and specializations.

Keywords: data engineering, ETL, cloud data engineering, skills development

Introduction

Data engineering (DE) is an emerging field in the data analytics landscape. Data engineers work closely with data scientists, analysts, and artificial intelligence (AI) and machine learning (ML) professionals. They prepare data for analytics or business operations and typically build data pipelines to extract and combine data from multiple, often varied sources (Greet, 2022). Their job frequently involves cleansing, transforming, and consolidating data for use in analytical applications.

Even though the extent to which digital technology supports their business processes, products, and services varies from company to company, practically every company today is a technology company (Mims, 2018). Continuing and accelerating inroads of information technology into organizations large and small have rendered data as the primary resource for enabling enterprise operations and decision support. Without timely and relevant data most organizations would hardly function while some would cease to exist in their current form. New and re-engineered business models and practices, in turn, are changing the nature of the data itself – especially how, when and where it is generated, stored, and processed. Relational databases, for instance, now share their decades-long prominence with a variety of non-relational data sources, as do data warehouses with data lakes. On-premises and local data storage are being replaced or augmented with cloud storage. And a growing number of business applications require real- or near-real time data to perform their functions. Many companies are eager to use their massive datasets and deploy emerging technologies such as cloud computing, machine learning, and artificial intelligence to enhance their current operations or offer new products and services. As a result, the demand for graduates with advanced data and information technology skills is high and growing fast (Rzeznikiewicz, 2022).

Most management information systems (MIS) programs in the US and elsewhere do a great job providing foundational information systems education and practical skills. Besides the core business subjects and elective courses, undergraduate information systems curricula typically include courses in database management, systems analysis and design, networking, and a programming language. Faculty and

department administration often keep such curricula current and updated. That, however, is often not the case when it comes to more *advanced* knowledge and skills that the industry requires. The gap between the supply and demand of advanced information technology knowledge and skills is large and growing. According to a survey of 283 large private organizations conducted by McKinsey Global Institute, about two-third of the executives in the US and Europe believe that due to digitization and automation, they would need to replace or retrain more than one-fourth of their workforce by 2023 (Illanes et al., 2018).

Changing and often disruptive new business models have resulted in skill gaps; number of skills required for a single job on average has been increasing at the rate of 10% every year since 2017 (Baker & Zuech, 2021). And at the same time, hiring employees to fill the gap has been getting harder to do. A practical strategy to meet their hiring needs, employers should identify their current employees with closely matching skills for the job and train them to close any gaps. This approach is generally known as *skill adjacency*, and it may arguably be the most effective approach to fill the increasing skill gaps (Baker & Zuech, 2021; Dapena, 2019). Having prior prerequisite kills makes it easier for a worker to train for the next adjacent or higher level skill. And at the same time, it is more cost-effective and time-efficient for the employers to train new hires or existing employees for the skills they need. An example of the skill adjacency approach in practice might be to train an employee who has Python programming and statistical analysis skills for machine learning and natural language processing jobs. According to a study for the World Economic Forum by LinkedIn (Wiles, 2020), about half of the employees who moved into data science and artificial intelligence jobs came from unrelated industries –an endorsement of the skill adjacency strategy.

Beyond the foundational database knowledge, adjacent advanced data skills have traditionally been in database administration and security –and they continue to be so. In recent years, however, a number of additional data-related roles have emerged, particularly those that support data science and analytics teams. The focus of this study is on the latter –cloud data engineering in particular.

With the advent and popularity of cloud data warehousing platforms and services such as Snowflake, Databricks, Google BigQuery, and those by Amazon AWS and Microsoft Azure, *cloud data engineering* is at the forefront of such new roles. These roles are essential to effectively scaling data and analytics by building data pipelines, managing extract, transform, and load (ETL) operations, and handling cloud data infrastructure (Moses, 2021). *Cloud data engineering* is the practice of data engineering on the cloud, i.e., performing data engineering activities that involve data, all or part of which may be stored and processed on a cloud platform (Snowflake, 2022). In this paper the terms data engineering and cloud data engineering are used synonymously.

The objective of this study is to identify knowledge and skills expected of a cloud data engineer in the industry. The larger goal, however, is to learn industry practices and expectations of the data engineering role to help information systems graduates and relevant curricula committees make informed decisions in order to stay current in the field. Specifically, this study addresses the following research question: *What technical skills and competencies are required for the data engineering job in business and industry today?*

Methodology

Data collection

It is essential to clarify that no individuals were involved or contacted to obtain the survey data –it was therefore not necessary to obtain consent of any individuals or conduct human subjects reviews. The data were collected from online job postings for data engineering positions by following the steps below.

First, the following three prominent job boards for professional jobs (Garner, 2018) were selected by the author: *Indeed.com*, *Glassdoor.com*, and *Linkedin.com*. The author then signed-up to receive emails as soon as new data engineering job announcements became available on their websites. Creating an account and signing up for emails allowed the author to obtain detailed job description, which in some cases were not available through publicly available job postings. This, in turn, allowed the author to perform more fine-grained analysis and better understanding of the job requirements such names of the specific technologies and platforms the employers were using. The author then selected 50 job descriptions using the following selection criteria:

1. ***Qualifications*:** The position announcement specifically stated that the minimum academic qualifications were not just a degree in Computer Science but also a related field. This was to ensure that MIS graduates would be eligible to apply for the job.
2. ***Employer*:** Another consideration was that the data sample included a mix of large, midsize, and smaller employers, as well as informational technology and non-technology companies.
3. ***Mode of work*:** Due to the Covid-19 pandemic, some jobs were either online or in-person only, while others were hybrid, i.e., partly in-person and partly online. Position announcements were selected with a view to have a balanced mix of both modes of work.

One job announcement from each of the companies listed in Table 1 was included in the data sample.

Table 1: Sources of Data

Amazon	ComTec	InfoLogitech	O'Neil Global Advisors	Team Symru Inc.
Accenture	Drive Centric	Jefferson Frank	Optello	Tesla
Accenture	DSI Systems	Kroger	PepsiCo	The Center for Health Affairs
Agama Solutions	Electronic Arts	Laconic Recruiting	Pratt & Whitney	The Mom Project
AllState	Facebook	League	Rappi	TikTok
Anheuser Busch	GitLab	LinkedIn	Smart Warehousing	Twitch
Apple	Google	Lyft	SmartSource	United Health Group
Bayer	Grammarly	Moody's Analytics	Snap Finance	Vaco
CloudQnect	HCL Technologies	Netflix	Snapidea Systems	Visa
Coca-Cola	Infillion	NIO	StarsHR	Zoom

Data preparation

As a first step, all descriptions containing qualifications, work experience, and other requirements were combined into a single text document. An exploratory data analysis (described in the following subsection) was then performed on this document. In the second step, the compressive text document was split into two subsets, one containing technical skill requirements and the other non-technical (or "soft") skills. According to Indeed.com (2022), "Soft skills are personality traits and behaviors that will help candidates get hired and succeed in their work. Unlike technical skills or “hard” skills, soft skills are interpersonal and behavioral skills". Soft skills enable a person to fit in at a workplace and include traits such as personality, attitude, flexibility, motivation, and manners (Doyle, 2021). "Hard" skills, on the other hand, are often known as technical skills. Soft skills such as teamwork and written and oral communication skills are usually complimentary to technical skills required for a job, and either set of skills alone is not sufficient for a

successful career in most jobs. This study, however, focuses on identifying technical skills, hence discussion of the soft skills is omitted from the study though it should be emphasized that soft skills are a prerequisite.

Text analysis

Since the focus of this study is on technical skills only, these skills were then analyzed using text analysis in Python. The following tasks were performed in an iPython (Jupyter) notebook:

- ***Case conversion*:** the text document containing the entire collection of job descriptions was first converted into lower case letters so the same word or term may not be considered as a different word due to a different case.
- ***Tokenization*:** the text document was then split into tokens (or words).
- ***Stopwords*:** *Stopwords* are needed for grammatical correctness but they do not add much value to overall understanding of the text. Hence words such as articles of speech, prepositions or conjunctions, like 'a', 'the', 'and', etc. were removed from the document.
- ***n-Grams*:** two-word n-grams were created to capture skill-relevant terms such *data science* and *data pipeline* rather than the words data, science, or pipeline alone as they would convey different meanings.
- ***Term frequency (TF):*** *a term frequency matrix was generated and sorted in the descending order of word count*

After the automated text analysis process using Python, the author manually reviewed the word frequency matrix to identify technical terms, and for grouping the technologies based on their use in various stages of the ETL process. The resulting list was then used to perform an exploratory data analysis, described next.

Exploratory data analysis

Before conducting more detailed analysis of the data (i.e., job descriptions), an exploratory data analysis (EDA) was performed. The author used a Jupyter notebook with Python libraries including natural language toolkit (*nltk*), *pandas* for tabular data analysis, *plotly* for data visualization, and others to create a word frequency matrix, and then a word cloud visulaization(shown in Figure 1) in order to see the prevalence of terms in the entire set of job postings.

Figure 1: Exploratory Data Analysis with a Word Cloud

The font size of a term in Figure 1 is directly proportional to its frequency in the entire collection of job descriptions. As we see in this visualization (and as expected), *data* skills and *experience* are among the

most commonly cited and thus primary job requirements. Among the secondary and tertiary level skills and attributes, include *SQL*, *communication* skills, *teamwork*, and *remote work*. A more detailed and thorough analysis of the data was then performed. Findings are discussed in the following section.

Findings, Discussion, and Recommendations

Since all of the data used in this study was text-based rather than numeric, a descriptive data analysis and synthesis was performed by the author. Specifically, skills were synthesized into themes and skill categories to help provide actionable curricula guidance so the findings can be mapped to courses or course modules.

The review of the data engineering job descriptions revealed two distinct themes: (1) Data engineering jobs on specific commercial platforms popular in the field, including Amazon AWS, Microsoft Azure, Google GCP, and Snowflake. (2) Jobs that do not require training or work experience on specific platforms but rather look for knowledge and expertise in specific data engineering skills and technologies. Although a large part of data engineering work is performed on the aforementioned specific cloud computing platforms, many organizations still perform part or all of these activities on-premises. Certain government agencies, the military, and utilities have their own data centers and they prefer their data not to be on public clouds (Lee, 2022). Also, increasingly more and more organizations are in the process of moving their data to the cloud and hence need the expertise in both on-premises and cloud data environments. These two different modes of preparing for data engineering skills are further discussed below.

Prerequisites for Training in Data Engineering

Before discussing the two separate paths to DE training and jobs, prerequisites for the training regardless of the path must be clearly noted:

All DE positions require a bachelor's degree in Computer Science, Engineering, Software Engineering, or a related field. Assuming that an MIS degree would be considered a related field for a given position, the candidates must ensure that they have *intermediate* to *advanced* level knowledge of the following:

- SQL
- At least one programming language such as Python
- Relational databases
- NoSQL databases including key-value, document, and wide-column databases
- Data warehousing, data lakes, and OLAP concepts

Key-value and document databases are among the most popular types of data stores to support web, mobile, and social media applications. In key-value databases, data is retrieved using a unique key or a combination of unique and the values can be simple data types like strings and numbers or complex objects such as arrays or documents. The term document refers to data structures that follow the JavaScript Object Notation (JSON) format to store semi-structured data. these types of NoSQL data stores are particularly suitable for storing and querying user-generated content such as data from chat sessions, tweets, blog posts, ratings, and comments.

Data Engineering on Cloud Data Platforms

Major cloud data platforms offer streamlined, well-structured, and professionally curated paths to their DE services. Courses, training, and certification offerings by Amazon AWS, Microsoft Azure, Google GCP,

and Snowflake are perhaps the best entry point into the field for most candidates. Since practically every major DE project involves some interaction with the cloud, knowledge of cloud computing fundamentals is essential to a professional DE job –and major cloud data platforms seamlessly integrate such training in their DE training offerings. Further details about each of the aforementioned four platforms are available on the major cloud providers' websites, and can also be easily found through an online search.

Non Platform-Specific Data Engineering

A generic path to DE training is proposed in Figure 2 below. As shown, it recommends knowledge of cloud computing fundamentals and hands-on experience with one or more of cloud computing vendor platforms before proceeding to train for specific ETL skills. The cloud computing fundaments knowledge typically includes general concepts and specific topics on storage, security, and data services.

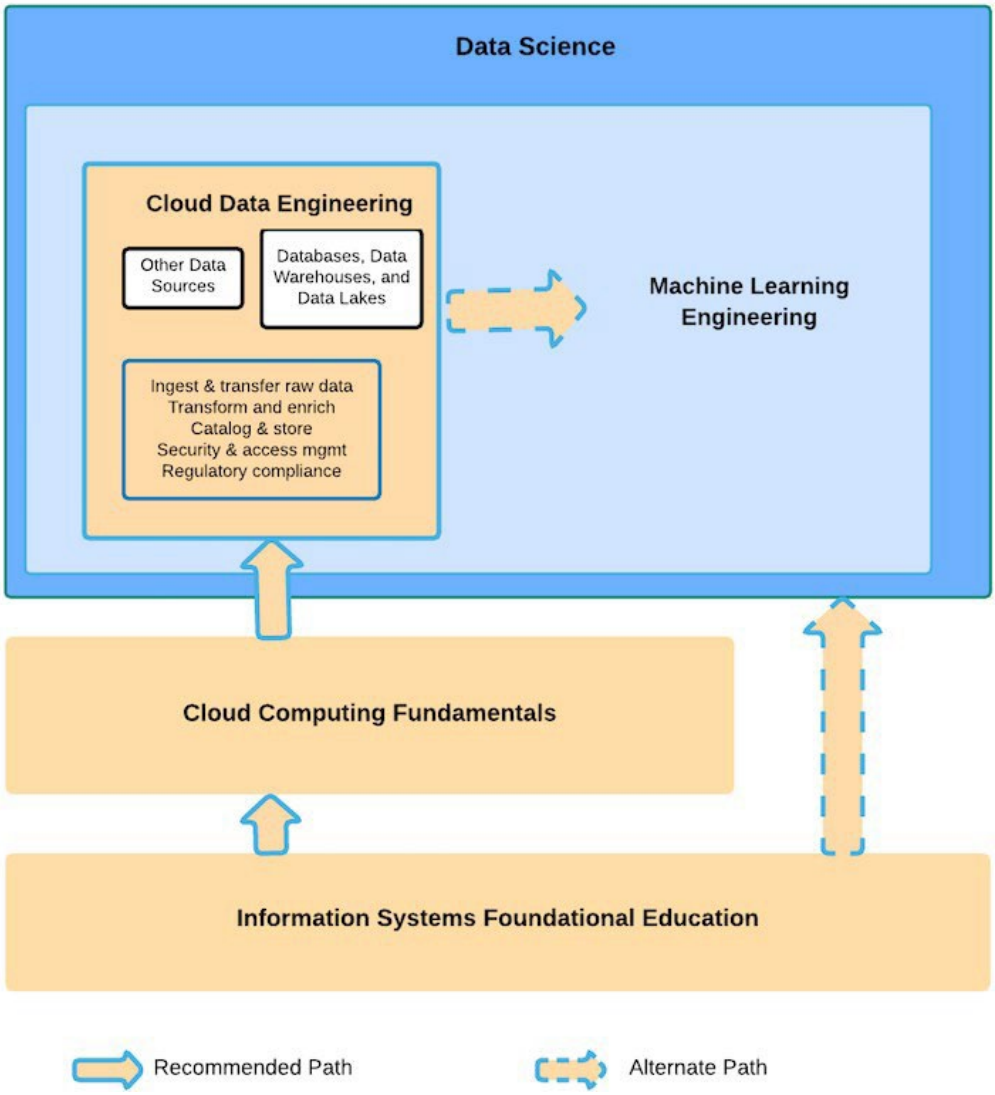

Figure 2: Paths to Data Engineering Training

Specific DE software tools and cloud-based services discovered though the analysis of job postings and review of relevant literature were categorized by their use in different stages of the ETL stages. A list of such tools and services appears in Table 2 below.

Table 2: Data Engineering Tools and Services

Cloud Database and Data Warehousing (Data Sources & Sinks)	Data Extraction, Loading/Ingestion, and Transformation	Data Workflow Management	User-Facing BI and Analytics
Amazon Athena and Redshift	Apache Nifi, Sqoob, Hive	Apache Airflow	Tableau
Azure Synapse	Azure Data Factory	Luigi	Power BI
Google BigQuery, Bigtable	AWS Glue	Apache Oozie	Qlik
Databricks	Airbyte	Dagster	Looker
Oracle	Alteryx	Apache Nifi	MicroStrategy
MySQL, Postgres, SQL Server	dbt	AWS Step Functions	IBM Cognos
MongoDB, HBase, Cassandra	Talend	Perfect	
Snowflake	Trifacta	Argo, AirFlow	

The alternate, non-platform specific, path to DE education is generally fragmented and often can be accomplished piecemeal. It is therefore highly recommended that candidates should follow a structured path that comprises of professionally developed set of courses and labs. Conventional academic institutions as well as commercial online outlets such as Udemy.com (Udemy, 2022) and Coursera.org (Coursera, 2022) offer data engineering specializations, projects, and courses.

Data Governance and Regulatory Compliance

Together with technical know-how, knowledge and methods of monitoring and complying with relevant regulatory requirements are essential for data engineers. Most countries, jurisdictions, and organizations have privacy and security policies, and ETL processes at various stages must comply with those policies. Major cloud data platforms provide services to facilitate the compliance process though data engineers are expected to be knowledgeable in the relevant policies and regulations. Table 3 below lists some of the common data privacy and security laws and regulations.

Table 3: Selected Data Laws and Regulations

Law or Regulation	Jurisdiction
California Consumer Privacy Act (CCPA)	California, USA
General Data Protection Regulation (GDPR)	European Union (EU)
Gramm–Leach–Bliley Act (GLBA)	USA
Health Insurance Portability and Accountability Act of 1996 (HIPAA)	USA
Any existing data use contracts	Potentially all jurisdictions

The above list is not exhaustive but is indictive of DE knowledge expectations with respect to regulatory compliance found in the job descriptions.

Conclusion

With the popularity of academic disciplines under the broad umbrella of data science and analytics, as well as wide-ranging use of related data products and technologies in the industry, the need for the data engineering role has grown proportionately in recent years. In this study, a sample of 50 recent job postings by organizations of varying size and in multiple industries in the US were analyzed to find commonly expected education and technical skill requirements for the DE role. The larger objective of the study was to provide actionable information to the information systems (MIS) graduates and academic curricula committees for training and up-skilling in this emerging field. The concept of '*skill adjacency'* suggests that it is easier and time-efficient for individuals to upgrade their skills and education to the next adjacent level; data engineering is one of such adjacent advanced data skill for MIS graduates.

One limitation of this study is that the selection of the three job boards and the 50 job announcements were selected by the author, which may introduce selection bias. These selections, however, were not random but were professionally curated by the author based on specific criteria described in the methodology section of the paper.

References

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

Anand, S. (2018, March 7). Test of blockchain for real estate is readied. *The Wall Street Journal*, pp. B8.

AWS.Amazon.com. (2022). AWS training and certification blog: Data engineer. Retrieved from https://aws.amazon.com/blogs/training-and-certification/tag/data-engineer/

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

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

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

Baker, M. & Zuech, T. (2021). Gartner HR research finds 58% of the workforce will need new skill sets to do their jobs successfully. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2021-02-03-gartner-hr-research-finds-fifty-eight-percent-of-the-workforce-will-need-new-skill-sets-to-do-their-jobs-successfully

Bessen, J. (2014). Employers Aren't just whining – the "Skills gap" is real. *Harvard Business Review,* (August 25.)

Bureau of Labor Statistics, U.S. Department of Labor. (2017). Occupational outlook handbook, database administrators. Retrieved from https://www.bls.gov/ooh/computer-and-information-technology/database-administrators.htm

Burning glass technologies. blurring lines: How business and technology skills are merging to create high opportunity hybrid jobs. Retrieved from http://burning-glass.com/wp-content/uploads/Blurring_Lines_Hybrid_Jobs_Report.pdf

Cappelli, P. (2019). Your approach to hiring is all wrong. *Harvard Business Review,* (May-June)

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

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

Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Chiang, R., Grover, V., Liang, T., & Zhang, D. (2018). Special issue: Strategic value of big data and business analytics. *Journal of Management Information Systems, 35*(2), 383-383-387.

Coindesk. (2016). US government awards $600k in grants for blockchain projects. Retrieved from https://www.coindesk.com/us-government-grants-blockchain-projects/

Columbus, L. (2017). IBM predicts demand for data scientists will soar 28% by 2020. *Forbes,*

Coursera.org. (2022). Career learning paths: Data engineering. *Https://www.Coursera.org/learning-paths/data-Engineering,*

Dapena, K. (2019, 04/20). *The Wall Street Journal*, pp. B6-B7.

Databricks. (2022). Data engineering on databricks. Retrieved from https://databricks.com/solutions/data-engineering

Davenport, T., & Patil, D. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review,*

The definition and selection of key competencies. (2005). ().Organization for Economic Cooperation and Development (OECD). Retrieved from https://www.oecd.org/pisa/35070367.pdf

Delventhal, Z. (2018). Hyperledger sawtooth blockchain application Retrieved from https://github.com/delventhalz/transfer-chain-js

Diplock, T., Meier, P., Jordan, D., & Ek, N. (2018). How to actually put your data analysis to good use. *Harvard Business Review,* , May 1, 2019.

Doyle, A. (2021). Top soft skills employers value with examples. Retrieved from https://www.thebalancecareers.com/list-of-soft-skills-2063770

edX.org, & Linux Foundation. (2018). Blockchain for business - an introduction to hyperledger technologies. Retrieved from https://www.edx.org/course/blockchain-business-introduction-linuxfoundationx-lfs171x-0

Egloff, C., Sanders, U., Riedl, J., Mohottala, S. & Georgaki, K. (2018). The digital imperative in container shipping. Retrieved from https://www.bcg.com/publications/2018/digital-imperative-container-shipping.aspx

Fayyad, U., Wierse, A., & Grinstein, G. G. (2002). *Information visualization in data mining and knowledge discovery*. San Francisco, CA: Morgan Kaufmann.

Fraher, E., & Ricketts III, T. (2016). Building a value-based workforce in north carolina. *North Carolina Medical Journal, 77*(2), 94-94-98. Furbush, J. (2018). Data engineering: A quick and simple definition. Retrieved from https://www.oreilly.com/content/data-engineering-a-quick-and-simple-definition/