Market Demand Analysis for Data Engineering Skills using Data Modeling and Text Mining

Cesar P. Malenab Jr.[1], Robert Kerwin C. Billones[2]

[1] Department of Civil Engineering
[2] Department of Manufacturing Engineering and Management
De La Salle University
2401 Taft Avenue, Manila 0922, Philippines
[1,2] E-mail: {cesar_malenabjr, robert.billones}@dlsu.edu.ph

Abstract **— The age of information exchange has been in an unprecedented phase which created a significant gap in the demand for professionals equipped with digital competencies. In this study, a data-driven approach to understanding the demand for data engineering was performed to highlight the most desirable hard and soft skills and identify skillset combinations that would satisfy the needs of the current market. Online job advertisements were collected using web scraping tools and were modeled in a relational database. Topic modeling using Non-negative Matrix Factorization (NMF) clustered the job descriptions into skillsets based on their topic probability distribution. The analysis showed that SQL, Python, and ETL are the most in-demand hard skills while communication, analytical thinking, and leadership qualities are important soft skills. Data architecture and analytics emerged as the skillset combination that is valued the most in the labor market.**

Keywords — Analytics, Data Architecture, Non-negative Matrix Factorization, Topic Modeling, Web Scraping

I. INTRODUCTION

Digital transformation has greatly impacted business workflows as well as the daily life of consumers due to the utilization of intelligent data. Industries around the world are transitioning into the digital environment and International Data Corporation (IDC) has predicted that 175 zettabytes of data will be generated annually by 2025 [1]. The potential to leverage the vast amount of data to gain market advantage in today's competitive landscape has transformed companies to become data-oriented. Furthermore, businesses that identified themselves as data-driven were reported to be more profitable than their competitors [2], and this has inspired enterprises to explore big data technologies to keep up with the current dynamics.

The demand for professionals equipped with the necessary skills to handle data and transform them to yield tangible benefits has created a significant gap in the labor market. The chronic shortage is becoming a growing concern such that productivity gains from Big Data are being restricted. Employers are struggling to acquire the right talent while the labor market demand for both Data Scientists and Data Engineers will continue to grow by 39% [3]. In the Philippine context, there is an uneven use of digital technologies with larger firms being more proficient with the use of technologies [4], which indicates a shortage of talent in the country to progress its technological advancement.

Due to the diversified and interdisciplinary nature of a data role, there is no single job title that can fit all the essential skills required within organizations. Studies have shown that 'Data Scientist' has been used as an umbrella term to loosely classify the different job roles that companies need [5]. The overlap of job descriptions between data professionals is attributable to the heterogeneous nature of required skillsets particularly mathematics, programming languages, machine learning algorithms, data visualization, business understanding, and communication skills. In addition, technological trends such as moving to cloud-based platforms and automation have affected the core skill requirements within a short time frame. The lack of a common framework for qualifications because of the recent emergence of data roles has caused employability mismatch and a thorough understanding of these roles is needed to overcome the gap between supply and demand for data professionals.

Educational institutions are starting to remodel their curriculum to include data-centric skills as part of the academic discipline [6]. This action is due to the cultural shift of traditional education on account of skill demonstration dictating one's employability instead of university degrees in the digital industry [7]. The urgency to equip the future workforce with digital competence is also reflected in various learning facilities providing training related to Industry 4.0 technologies.

The compounding problem of unclear job roles and talent shortage can be addressed by extracting insights from the labor market as reflected in job advertisements. This study attempts to identify the most in-demand technical and soft skills and the interdisciplinary skillsets that satisfy the needs of the Data Engineering job market. Online job advertisements will be collected using web scraping tools which will be stored in a relational database. Skillsets will be identified using text mining techniques on the contents of the job descriptions to

classify skills based on frequently occurring keywords. The results of the study aim to delineate the core competencies that would define the role of a data engineer in the current market. Using this data-driven approach, present and future professionals in the data industry can start acquiring the necessary credentials and this study can aid educators, policymakers, and business leaders in developing programs that best reflect the needs of the labor market.

II. Review of Related Literature

The age of digitalization and information exchange has been in an unprecedented phase over the last two decades [8]. As such, companies have been adopting new technologies that can handle a large volume of unstructured data and process them for better business decision-making. Moreover, big data frameworks and scalable cloud technologies have presented cost-reduction opportunities for businesses and high availability for data consumers [9]. The utilization of these technologies has given firms greater competitiveness which leads to larger market shares.

The Philippine digital economy continues to display an upward trend from USD 20 billion in 2022 to USD 35 billion by 2025 [10]. For the first two quarters of 2022, Bangko Sentral ng Pilipinas (BSP) reports 348 million transactions in PESONet and InstaPay digital payment platforms, equivalent to a 24% increase compared to the prior year [11]. The growing digital market is due to the intensified adoption of technologies by firms in various sectors. These technologies include online sales, data analytics for market trends, supply chain management, resource planning, and more advanced utilizations such as virtual collaboration through cloud platforms and automation [4].

The rapid pace of data growth has consequently required digital competency to mobilize the workforce in the age of technological advancement. Labor Market Information Report shows occupations like Data Development Engineer and Database Manager as one of the most in-demand occupations for Key Employment Generators (KEGs) industry groups [12]. Thus, digital skills development has received increased attention both from the education sector and training institutions whereby a series of guidelines and curriculums are dedicated to the strengthening of digital proficiency. Furthermore, the Philippine Development Plan includes an accelerated effort to incorporate developing technologies such as big data, cloud services, and artificial intelligence in the Inclusive Innovation and Industrial Strategy (i3s) [13].

In order to identify the expected skillsets in the current digital market's demands, researchers have employed a data-driven approach of collecting online job postings and performing text mining techniques to map qualifications sought by employers. Chunmian et al. [14] investigated the demand for blockchain-related occupations and identified skillsets through topic modeling analysis. In addition, the researchers identified how these skillsets affect the salary offers which distinguishes the top skills valued in the blockchain industry. Overall, the study demonstrates that the skillsets with the most impact on salary levels are Blockchain Development and Cryptocurrency. On the other hand, Smaldone et al. [15] provided insights into thematic areas that arise in data scientists' jobs in the American market. The researchers clarified the primary competencies that job seekers must possess in this interdisciplinary field. Their framework involved parsing through online job advertisements and extracting information through linguistic queries. The unstructured data in job advertisements were subjected to a series of pre-processing pipelines before applying topic modeling using Latent Dirichlet Allocation (LDA). Big Data came out as the highly desirable skill based on its term frequency while problem-solving and data processing had strong correlations among the skillsets identified. Furthermore, De Mauro et al. [5] appropriated Big Data skillsets to individual job families that are often loosely represented as a data science position. A substantial amount of online job posts was collected and categorized into four job families through text mining techniques and expert judgment. The researchers mapped the skillsets obtained from LDA to the job families which revealed that job roles like Business Analyst and Data Scientist require a combination of skills from both fields.

The data-driven approach of text-mining job portals has also been used to improve the curricula of educational institutions. Zimmermann and Brandtner [16] pointed out pieces of training that should be included in the university to be competent in the Supply Chain Management Domain of the Austrian Labor Market. The researchers listed out subject focus areas to prepare future professionals in various job profiles. Likewise, Phaphuangwittayakul and Sarangwong [17] addressed the education-job mismatch in the Thai labor market by performing keyword extraction on recruitment websites. The authors used Rapid Automatic Keyword Extraction (RAKE) algorithm on job descriptions to obtain a

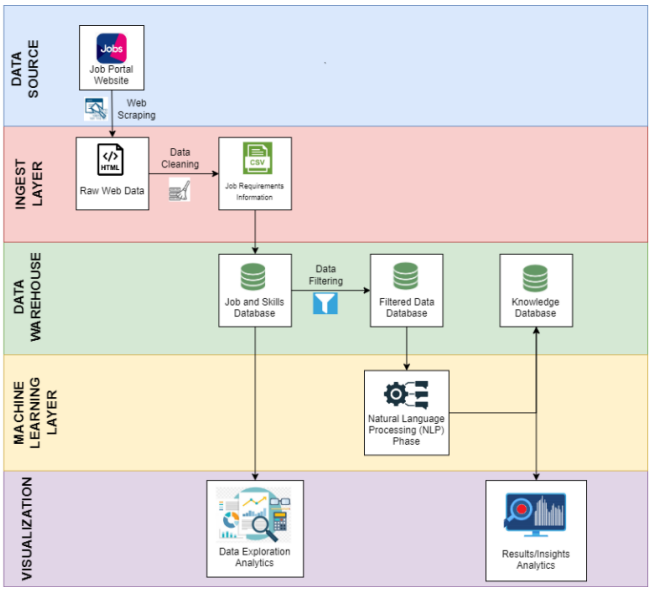

Fig. 1. Data flow diagram

set of common words in various job categories. The keywords extracted served as job functions within each job category which raises skill mismatch with those present in college courses. Similarly, Rahhal et al. [18] emphasized the requirements of the Morocco Cybersecurity labor market to reshape university training and meet the skill gap due to its increasing demand. Networking and Risk analysis stood out as the most sought expertise in the cybersecurity industry. Sozykin et al. [19] developed educational programs for Information Technology (IT) specialists by collecting data from job portals and employing machine learning algorithms to analyze the joint occurrence of skills as signaled in job descriptions. The study aims to increase the marketability of graduating students in the web development and data science field. In a similar manner, Orchirbat et al. [20], created a Book Recommender System (BRS) to accompany students in their development in the field of IT. Instead of revising education curricula, the authors extracted keywords from IT job markets and computer science curricula to generate book recommendations based on the results of text mining. Stanton [21] analyzed LinkedIn job postings to prepare business students in business analytics by listing the necessary credentials for entry-level talents. The researchers concluded that the traditional curriculum is not adequate to satisfy the needs of employers for analytics professionals.

The citations above have shown that job advertisements can be a source of research data for defining the market demands concerning human resources. The research model often includes gathering job postings from recruitment websites through web scraping or Application Programming Interfaces (APIs). Subsequently, text mining techniques are performed to handle the unstructured text data to highlight the skillsets or job categories that may emerge based on the frequencies of keywords. The current study aims to build on these works and identify the needs of the Philippine labor market for data engineering skills. Understanding the market demand can (1) help present and future professionals build their portfolio with the most marketable skills, (2) supplement the frameworks of educational institutions and training centers to better prepare students in the digital industry, and (3) give an insight to the general landscape of data engineering industry with the Philippines as the geographical focus.

III. Methodology

Figure 1 shows a high-level view of the data flow to extract the in-demand skillsets from online job portals. The data flow architecture is composed of five layers including: (1) data sources, (2) data ingest layer, (3) data warehouse, (4) machine learning, and (5) data visualization. The framework was adopted from Billones et al. [22] with the addition of a machine learning layer that would refer to the implementation of unsupervised clustering of job advertisements.

A. Data Collection

Indeed platform aggregates job openings from local and international sources and displays a semi-structured collection of these advertisements. Job postings were collected from Indeed.com with ‘data engineer’ as the search term and web scraping tools extracted features such as job title, company, salary, location, job description, company size, and industry among others. Web scraping is the method of extracting raw details from websites and a web crawler can mine data from web pages such as a large corpus of text [23]. In this study, Selenium [24] served as the web driver for internet browser automation, and raw HTML data was transformed into structured data using the Beautiful Soup library [25]. The two libraries were combined to develop a web crawler in Python which eventually obtained 1,017 job postings from the online platform.

B. Data Preprocessing

Duplicate job postings were discarded to avoid the overrepresentation of terms in the dataset. Moreover, job titles that do not directly refer to Data Engineering roles were removed since the search term used contained the word ‘engineer’ that resulted in the inclusion of jobs from other engineering industries. The job posts to be processed was trimmed down to 293 job postings.

The data extracted from online job portals primarily contain text that needs to be cleaned and filtered for further analysis. At this stage, the spaCy [26] open-source library was used to

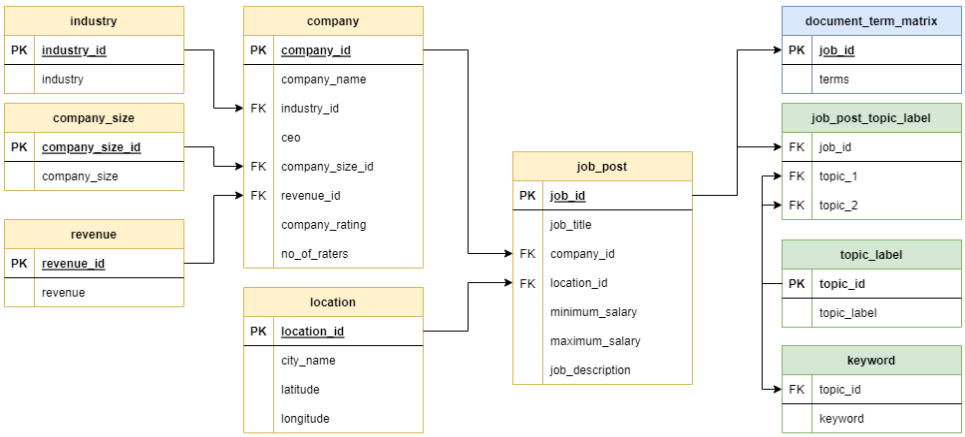

Fig. 2. Database schema

perform text pre-processing techniques namely, tokenization, lowercasing, removal of stop words and punctuations, and lemmatization. Tokenization is the process of splitting text into units, called tokens. The tokens serve as an element describing a document which can be considered as a vector representation of the document. To remove the case sensitivity of words, tokens are converted into its lowercase. Stop words are words that commonly exist in a document without adding meaning to the topic. These are mainly composed of prepositions, articles, and conjunctions that are irrelevant to text classifications [27]. In this study, custom stop words were included to remove text that is being assessed as important due to their frequency but holds no value for data interpretation. Lemmatization aims to remove the inflections of a set of words and group them into a single word, called lemmas [28]. The text-cleaning pipeline aims to minimize the dimensionality of the corpus thereby increasing the efficiency of text mining algorithms.

After the pre-processing of documents, each job post will be defined in terms of the frequency of its individual words. These term frequencies serve as the vector equivalent of the dataset and combining all these vectors will create a two-dimensional matrix called Document-Term Matrix (DTM). The rows represent the job advertisements and the columns are the features or words in each document [16]. The order of words is not considered in DTM since it follows a bag-of-words approach [29]. In the present study, the unique number of words was reduced from 3,752 to 3,111 after text preprocessing and the DTM consisted of 293 rows (job ads) and 3,111 columns (words). The DTM can be extended to determine the most occurring bigrams, trigrams, or four-grams which are combinations of two, three, or four words in sequence, respectively [30].

C. Topic Modeling

Topic modeling methods are unsupervised models that identify a set of underlying topics for a set of documents using the distribution of words [31]. The 'topics' signify hidden themes throughout the corpus and individual documents can be classified according to these topics [32]. Two widely used generative probabilistic models in the field of topic modeling are Latent Dirichlet Allocation (LDA) and non-negative matrix factorization (NMF). NMF-based models decomposed the document-term matrix into two low-rank nonnegative matrices to learn topics [33]. These models give semantically meaningful results that are comprehensible for thematic clustering of high-dimensional data [34].

In this study, NMF models are used to draw out topics that refer to the skillsets that are likely to appear in job descriptions. Apart from the document-term matrix, the parameter *k* is an input that indicates how many topic classifications should be performed in the corpus [35]. Several studies have considered perplexity and topic coherence to select the suitable number of topics [15], but there are no specific guidelines for this parameter. If *k* is set to be too low, the number of topics will be generic, otherwise,

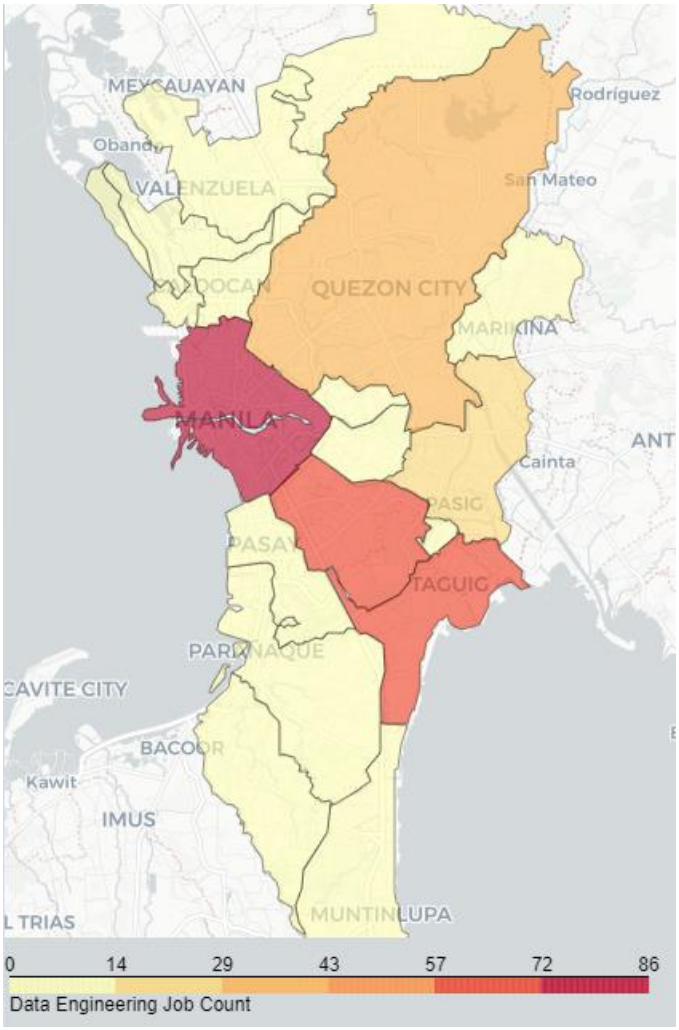

Fig. 3. Geographical distribution of Data Engineering jobs in Metro Manila

meaningful topics may be difficult to identify if set to too many. The number of topics in the study was selected by evaluating multiple outputs of NMF based on k ranging from 2 to 10 to determine the acceptable value. The value of k chosen was 4 which provides the most interpretability on the topic results of the model.

D. Data Modeling

Data collected from online job portals and data products from text mining algorithms are stored in the PostgreSQL database [36]. Figure 2 shows the relational data model employed in the database. Psycopg [37] and SQLAlchemy [38] libraries were used to store scraped data from Python to the relational database and likewise, obtain the results of queries from the database for the application of topic modeling algorithms.

The job post table serves as the central fact table where dimension attributes and measures are located. The smaller dimension tables such as the company, location, document term matrix, and topic label around the job post table present an organization of data in the configuration of a star schema.

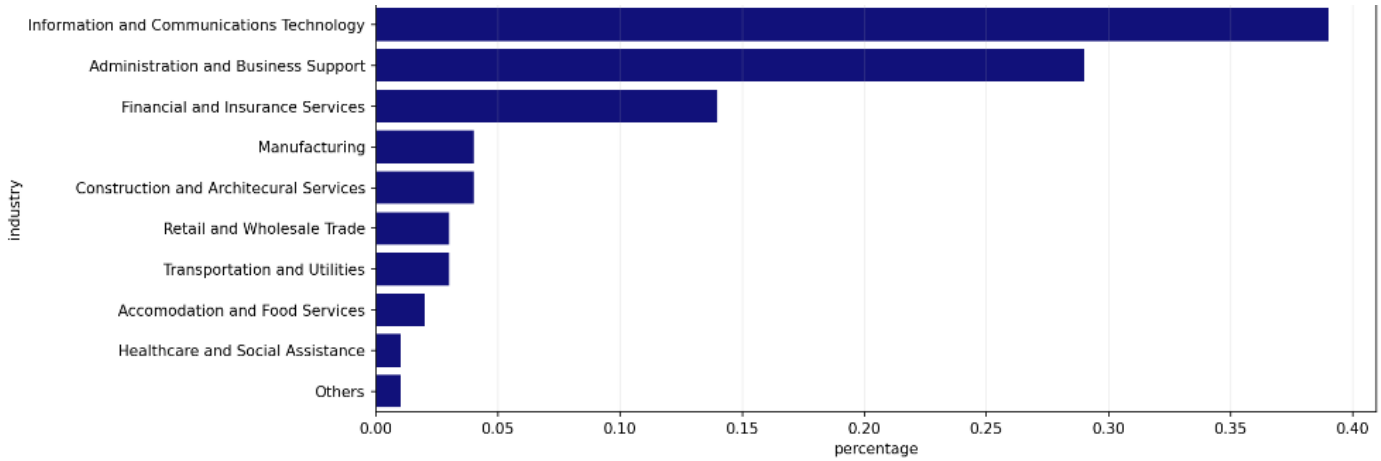

Fig. 4. Industry demand distribution of Data Engineering jobs

Relationships between entities can identify the primary key suitable for a specific table. The relationship between a job post and a company is a one-to-many relationship, wherein a job advertisement can come from only one company but a company can have multiple job openings. In a one-to-many relationship, the primary key comes from the *many* side as this will be unique in each instance which dictates that job_id is the primary key for the fact table.

An example of a many-to-many relationship is the job post and topic label table where a topic label can refer to multiple job posts and a job post can be composed of multiple topic labels as a result of topic modeling techniques. In this kind of relationship, both instances of the attributes determine the primary key, commonly referred to as the composite key [39]. Hence, the two foreign keys, job_id and topic_id make up the primary key of the table.

IV. Results and Discussion

A. Geographical Distribution

The demand for data engineering jobs in Metro Manila is shown in Figure 3. The distribution was obtained per city to identify the locations where data engineering professionals are most needed at the time of the study.

From the figure, the leading cities for data engineering jobs include Manila, Taguig, Makati, Quezon, and Pasig. Correspondingly, these cities have the highest reported locally sourced revenues for 2021 [40], which is an indication that these are centers of economic activities where the expansion of business is likely to concentrate. As a result, products and services are extended to a wider customer base which requires more data professionals to optimize data processes and scale data frameworks. It can also be noted that the major business districts are situated in these cities where leading firms and enterprises centralized their operations. Overall, it can be inferred that the demand for data engineers is correlated with the expanding domestic markets and economic activities of local cities.

B. Industry Demand

Information and Communications Technology (ICT), Administration and Business Support, and Financial and Insurance services are the top industries for data engineers as shown in Figure 4. The results are consistent with the World Bank survey that identified sectors that are classified as 'technology-intensified' [4]. With the Philippines being a global leader in the Information Technology-Business Process Management (IT-BPM) sector, the steady growth of employment of data engineers in this sector are likely to continue in the coming years. Along with this, banking and finance also need data professionals to supervise online banking services due to the surge of e-commerce during the pandemic. Risk management, fraud detection, and investment options analysis are some of the key responsibilities of data engineers.

Industries such as manufacturing, construction, retail and trade, and transportation are seeing intensified use of digital technologies which suggests that the demand for data engineers is likely to increase for these sectors in the near future.

C. Hard Skills and Soft skills

The most important hard and soft skills across the job descriptions examined are shown in Tables I and II, respectively.

The primary hard skills required of data engineers are SQL, Python, ETL, analytics, and cloud which are not far from studies conducted in other geographical settings [41]. In the 2022 study of top programming languages by IEEE, SQL is the most popular language in Jobs ranking with Python in third place [42]. Databases are the foundation of technological infrastructures which a data engineer primarily handles. Though cloud and big data technologies are the trends of the present digital environment, the result of the study shows that SQL is still a primary query language for most employers in the Philippines. Though this might be the case, knowledge of cloud computing services is also an in-demand skill in the

Table I. *Top Ten Data Engineering Hard Skills*

No.	Hard skills	Occurrence rate (%)
1	SQL	6.59
2	Python	4.82
3	ETL	4.44
4	Analytics	4.12
5	Cloud	3.92
6	AWS	3.41
7	Big data	3.41
8	Pipelines	3.28
9	Excel	3.22
10	Scala	3.15

present market. The top five skills comprise the main tools a data engineer utilizes to extract data from multiple sources and combine them in a consistent data storage that is readily available for analysis.

The most desirable soft skills critical to the success of a data engineer are communication, analytical thinking, leadership, innovative, and verbal skills. The result highlights the importance of communicating the processes and results to stakeholders that are paramount to the success of a data-driven decision. Collaboration between data professionals can be better achieved if the needs and expectations within a business are properly expressed. Analytical skills allow a data engineer to dissect a business problem down to its most fundamental needs and use the tools available to address the issue. Likewise, a data engineer can also adopt proper tools among numerous technologies available and continually find new ways to solve business problems. Ultimately, leadership constitutes teamwork and effective feedback necessary to complete a series of tasks towards a common goal.

The study gives clarity on highly desirable prerequisites that are advantageous to candidates that are planning to expand their skillsets considering the evolving nature of data engineering.

Table II. *Top Ten Data Engineering Soft Skills*

No.	Soft skills	Occurrence rate (%)
1	Communication	21.31
2	Analytical	15.12
3	Leadership	11.11
4	Innovative	7.65
5	Verbal	7.10
6	Independent	6.01
7	Collaboration	5.65
8	Project management	3.64
9	Organizational	3.46
10	Proactive	3.28

D. Skillsets

After identifying individual hard and soft skills, NMF was employed to cluster skills to topics referring to generic skillsets, and each job description is assigned with its corresponding topics based on its probability distribution.

Table III. *Top Ten Keywords of the Skillsets Discovered using Non-negative Matrix Factorization*

No.	Skillset	Top ten keywords
1	Architecture	data pipeline, azure data, big data, design implement, data warehouse, computer science, business requirement, data engineering, data source, process improvement
2	Big Data	file format, cloud platform, solution architecture, premise cloud, real time, data lake, format cloud, file xml, api file, xml json
3	Business Impact	end enterprise, business application, integration business, develop state, state etl, processing integration, source processing, data depend, etl transform, application develop
4	Analytics	master data, big data, data technology, data scientist, advanced analytic, supply chain, oracle sql, data source, manipulate merge

Table III shows the top ten keywords identified for the four skillsets including the interpreted skillset titles. The first topic refers to the overall capability of creating and facilitating data infrastructures from the source up to storage as signaled from keywords such as *data pipeline, big data, data warehouse, and design implement.* Job descriptions classified with Architecture contain responsibilities such as: *'implement data pipelines to bring data at reach to platforms and analyze source data and data flows for data consumption'.* The second topic contains keywords like *cloud platform, real time, data lake, xml* and *json.* These refer to Big Data technologies where large volumes of real-time data are constantly processed requiring scalable storage. It covers skill requirements such as: *'execute jobs to import data periodically/(near) real-time from an external source'* and experience in big data tools such as *Hadoop, Kafka, Hive,* and *Spark.* The third topic is about the business impact that the dataset will bring into the organization. It requires industry knowledge and may include responsibilities including: *'data ingestion jobs for Proof-of-Concept (POC) leaning towards business solutions and support the overall end to end enterprise projects.* Lastly, Analytics skillset refers to the awareness of a data engineer in the overall process of drawing insights from data up to implementation of data products. Key responsibilities include: '*data expert for datasets used by data scientists and experience in visualization tools like Tableau and Power BI'.*

Among the four skillsets, Architecture is the most in demand for data engineers with over 90% of job advertisements classified as such. The three other topics are almost equally distributed with the remaining 10%. This is expected as infrastructure architecture is the primary function of a data engineer in the data environment. Since a job post can be

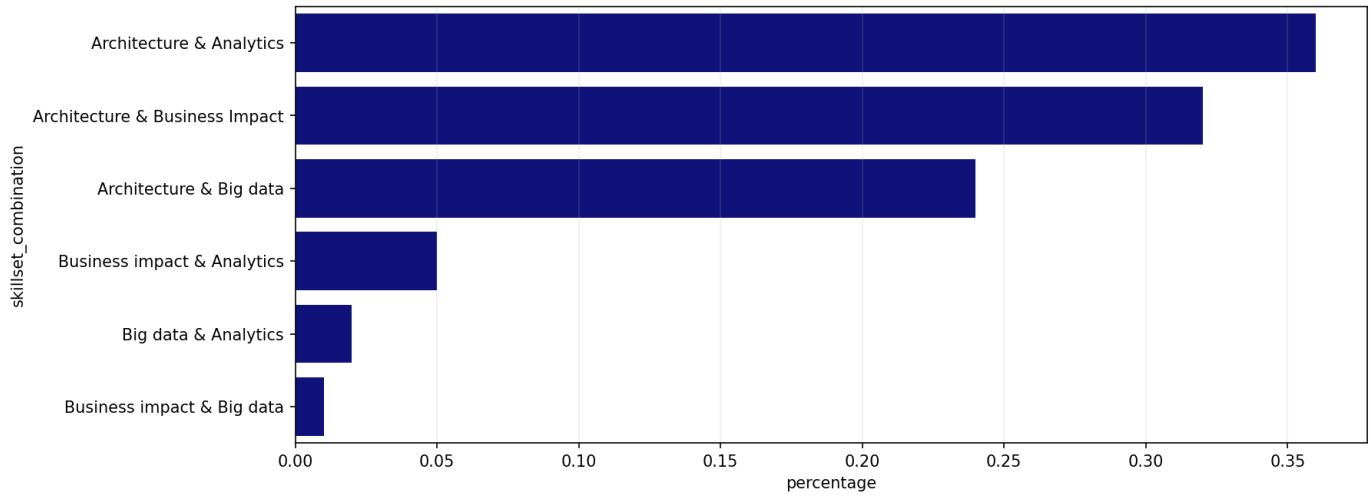

Fig. 5. Skillset combination distribution

described as a combination of the probabilities of topic distribution, a deeper understanding of the market requirements can be gained by obtaining the most occurring skillset combinations. From Figure 5, the most in-demand combinations in order are Architecture and Analytics, Architecture and Business Impact, and Architecture and Big Data. The percentage of occurrence reflects the interdisciplinary nature of data engineering where knowledge of other roles such as data scientists and data analyst is also important. This implies that skills such as descriptive statistics, data visualization, machine learning algorithms, and data storytelling are important repertoire that a data engineer should also possess.

V. Conclusion

The rapid transition of businesses and organizations to adopt the digital environment has led to a surge in demand for data professionals. The continuous growth of the data industry has created a significant gap in the labor market and educational institutions and training centers are slowly incorporating digital competencies to keep up with the overwhelming demand. This study used a data-driven approach of collecting job advertisements to characterize the labor market demand in terms of geospatial and industry-level distribution, desirable hard and soft skills, and skillset combinations sought after by employers.

The findings of the study show that the primary skills identified are not far from studies performed in other geographical locations with SQL, Python, and ETL as the leading hard skills and communication, analytical thinking, and leadership as the coveted soft skills. In addition, knowledge in both data architecture and analytics emerges as the skill combination that would give a competitive advantage to data professionals. These insights can help build development frameworks and guidelines based on demand-driven skills. The systematic approach in this study clarified the status of the data engineering job market which can be used to avert the increasing skill gap and talent shortage in this evolving profession.

References

[1] Reinsel, D., Gantz, J., & Rydning, J. (2018). *The digitization of the world from edge to core*. Retrieved November 25, 2022, from https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] McAfee, A., & Brynjolfsson, E. (2014). *Big Data: The management revolution*. Harvard Business Review. Retrieved November 25, 2022, from https://hbr.org/2012/10/big-data-the-management-revolution

[3] Markow, W., Braganza, S., & Taska, B. (2017). *The quant crunch - burning glass technologies*. Retrieved November 25, 2022, from https://www.burning-glass.com/wp-content/uploads/The_Quant_Crunch.pdf

[4] World Bank. (2021). Philippines Economic Update, December 2021. https://doi.org/10.1596/36874

[5] De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human Resources for Big Data Professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, *54*(5), 807–817. https://doi.org/10.1016/j.ipm.2017.05.004

[6] Overly, S. (2013, September). *As demand for big data analysts grows, schools rush to graduate students with necessary skills*. The Washington Post. Retrieved November 25, 2022, from https://www.washingtonpost.com/business/capitalbusiness/as-demand-for-big-data-analysts-grows-schools-rush-to-graduate-students-with-necessary-skills/2013/09/13/afbafb3e-1a66-11e3-82ef-a059e54c49d0_story.html

[7] Rometty, G. (2021). *Hiring skills, not diplomas: How to ignite the next generation of talent*. Ethisphere Magazine. Retrieved November 25, 2022, from https://magazine.ethisphere.com/ibm-ptech-hiring-skills-not-diplomas/

[8] Hilbert, M. (2015). Big Data for Development: A Review of Promises and challenges. *Development Policy Review*, *34*(1), 135–174. https://doi.org/10.1111/dpr.12142

[9] Davenport, T. H. (2014). *Big data work: Dispelling the myths, uncovering the opportunities*. Harvard Business Pr.

[10] Google. (n.d.). *E-conomy sea 2022 report*. Google. Retrieved November 17, 2022, from https://economysea.withgoogle.com/home/
[11] Neil. (2022, September 15). *Instapay, PESONET transactions grow*. BusinessWorld Online. Retrieved November 17, 2022, from https://www.bworldonline.com/banking-finance/2022/09/16/474916/instapay-pesonet-transactions-grow/
[12] Department of Labor and Employment. (2021). JobsFit COVID-19 Labor Market Information ReportL How the Pandemic Is Reshaping the Philippine Labor Market.
[13] National Economic and Development Authority. (2021). Updated Philippine Development Plan: 2017- 2022
[14] GE, C., SHI, H., JIANG, J., & XU, X. (2022). Investigating the demand for blockchain talents in the recruitment market: Evidence from topic modeling analysis on job postings. *Information & Management*, *59*(7), 103513. https://doi.org/10.1016/j.im.2021.103513
[15] Smaldone, F., Ippolito, A., Lagger, J., & Pellicano, M. (2022). Employability skills: Profiling data scientists in the digital labour market. *European Management Journal*, *40*(5), 671–684. https://doi.org/10.1016/j.emj.2022.05.005
[16] Zimmermann, R., & Brandtner, P. (2022). Job profiles in the field of data-driven supply chain management an analysis of the Austrian Job Market. *Procedia Computer Science*, *204*, 706–713. https://doi.org/10.1016/j.procs.2022.08.085
[17] Phaphuangwittayakul, A., Saranwong, S., Panyakaew, S.-ngapong, Inkeaw, P., & Chaijaruwanich, J. (2018). Analysis of skill demand in Thai labor market from online jobs recruitments websites. *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. https://doi.org/10.1109/jcsse.2018.8457393
[18] Rahhal, I., Makdoun, I., Mezzour, G., Khaouja, I., Carley, K., & Kassou, I. (2019). Analyzing cybersecurity job market needs in Morocco by mining job ads. *2019 IEEE Global Engineering Education Conference (EDUCON)*. https://doi.org/10.1109/educon.2019.8725033
[19] Sozykin, A., Koshelev, A., Bersenev, A., Shadrin, D., Aksenov, A., & Kuklin, E. (2021). Developing educational programs using Russian IT job market analysis. *2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*. https://doi.org/10.1109/usbereit51232.2021.9454998
[20] Ochirbat, A., Shih, T. K., Chootong, C., Sommool, W., & Gunarathne, W. K. T. M. (2019). Automatic book generation by using ICT job-skills and computing curricula. *2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media)*. https://doi.org/10.1109/ubi-media.2019.00023
[21] Stanton, W. W., & Stanton, A. D. A. (2020). Helping business students acquire the skills needed for a career in analytics: A comprehensive industry assessment of entry-level requirements. *Decision Sciences Journal of Innovative Education*, *18*(1), 138–165. https://doi.org/10.1111/dsji.12199
[22] Billones, R. K., Guillermo, M. A., Lucas, K. C., Era, M. D., Dadios, E. P., & Fillone, A. M. (2021). Smart Region Mobility Framework. *Sustainability*, *13*(11), 6366. https://doi.org/10.3390/su13116366
[23] Sundaramoorthy, K., Durga, R., & Nagadarshini, S. (2017). NewsOne — an aggregation system for news using web scraping method. *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*. https://doi.org/10.1109/ictacc.2017.43
[24] Selenium. (2022). Retrieved November 25, 2022, from https://www.selenium.dev/
[25] Beautiful Soup 4.9.0 documentation. (2020). Retrieved November 25, 2022, from https://www.crummy.com/software/BeautifulSoup/bs4/doc
[26] Spacy, Industrial-strength Natural Language Processing in Python. (2022). Retrieved November 25, 2022, from https://spacy.io/
[27] Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006
[28] Cambridge University Press. (2008). *Stemming and lemmatization*. Introduction to information retrieval. Retrieved November 25, 2022, from https://nlp.stanford.edu/IR-book/
[29] Gurcan, F., & Cagiltay, N. E. (2019). Big Data Software Engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access*, *7*, 82541–82552. https://doi.org/10.1109/access.2019.2924075
[30] Michalczyk, S., Nadj, M., Maedche, A., & Gröger, C. (2021). *Demystifying job roles in data science: A text mining approach*. AIS Electronic Library (AISeL). Retrieved November 25, 2022, from https://aisel.aisnet.org/ecis2021_rp/115/
[31] Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2016). Topic modelling for Qualitative Studies. *Journal of Information Science*, *43*(1), 88–102. https://doi.org/10.1177/0165551515617393
[32] Tong, Z., & Zhang, H. (2016). A text mining research based on LDA Topic Modelling. *Computer Science & Information Technology (CS & IT)*. https://doi.org/10.5121/csit.2016.60616
[33] Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. https://doi.org/10.1145/3178876.3186009
[34] Kuang, D., Choo, J., & Park, H. (2014). Nonnegative matrix factorization for interactive topic modeling and document clustering. *Partitional Clustering Algorithms*, 215–243. https://doi.org/10.1007/978-3-319-09259-1_7
[35] Jacobi, C., van Atteveldt, W., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271
[36] *PostgreSQL*. (2022). Retrieved November 25, 2022, from https://www.postgresql.org/
[37] Di Gregorio, F., & Varazzo, D. (2021). *PSYCOPG 2.9.5 documentation*. Psycopg. Retrieved December 2, 2022, from https://www.psycopg.org/docs/
[38] *The Python SQL Toolkit and Object Relational Mapper*. SQLAlchemy. (2022). Retrieved November 25, 2022, from https://www.sqlalchemy.org/
[39] Teorey, T. J. (2011). *Database modeling and design: Logical design*. Elsevier.
[40] De Leon, S. (2022). *Manila ranks no. 3 in highest locally sourced revenues in 2021*. PIA. Retrieved December 2, 2022, from https://mirror.pia.gov.ph/news/2022/08/03/manila-ranks-no-3-in-highest-locally-sourced-revenues-in-2021
[41] Jalis, A. (2016). *The state of Data Engineering: Stitch benchmark report*. Stitch. Retrieved December 2, 2022, from https://www.stitchdata.com/resources/the-state-of-data-engineering/
[42] Cass, S. (2022). *Top programming languages 2022*. IEEE Spectrum. Retrieved December 2, 2022, from https://spectrum.ieee.org/top-programming-languages-2022