Available Online : https://proceeding.researchsynergypress.com/index.php/cset/index
RSF Conference Series: Engineering and Technology
ISSN 2809-6843 (Online) | 2809-6878 (Print)
Volume 2 Number 1 (2022): 37-42

Exploration of Data Scientist's Current Expertise and Qualification Gap in Indonesia (Case Study: Jakarta Metropolitan Area)

Feliks P. Sejahtera Surbakti[1], Christine Natalia[2], Nicolette Kezia[3]

[1,2,3] Industrial Engineering Department, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia

Abstract

Data is a crucial asset for an organization to compete in the 21st Century, and Indonesia is no exception. Data scientists are relatively a new profession that is one of the most needed todays. As far as we are concerned, no research shows the skills and capabilities needed as a data scientist in Indonesia. This research aims to determine the skill and capability factors of data scientists, identify the actual skills and capabilities of data scientists in Indonesia, and identify gap skills and capabilities from data scientists in Indonesia. This research was conducted using a mixed-method approach, where interview results were used to design a questionnaire and determine the question indicators. This study has shown four skill and capability factors for data scientists; primary capabilities, data management, personal skills, and sources. The gap analysis results show that all indicators have a negative value, where data analysis-related training is most needed for a data scientist.

Keywords: *Data Scientist, Skills, Capabilities, Mixed-Method, Gap Analysis*

INTRODUCTION

The fourth industrial revolution or known as Industry 4.0 mainly focus on automation and cyber technology, hence this actively demonstrates that data plays a huge part in industry 4.0. To compete with other organizations, one needs to have the ability to maximized data usage. Data is the new oil. Data science is known as a study of the generalizable extraction of knowledge from data (Dhar and Mazumdar, 2014). Data scientists are generally expected to be in control of everything in terms of skills possessed. A good data scientist must gain insight that will have significant impacts on the organization's business problems. In other words, data scientists must act as problem solvers in their daily job. According to a Harvard Business Review article, data scientists were named "The Sexiest Job of the 21st Century" (Davenport and Patil, 2012).

According to Harris (2015), there are four types of data scientists: data business people, data creatives, data developers, and data researchers. Baumeister et al. (2020) has conducted thorough study and summarise six data scientists' roles according to technical skills, business skills, experience and educational demands. Based on previous research and the real condition in practice, data scientists have a lot of roles. Moreover, there is some findings that there is skill gap between what the industry need and what the data scientists have. Mikalef and Krogstie (2019) has conducted researched to identify gap skill of data scientist in Norway and found that while technical skills continue to be important, soft and managerial skills are greater importance. Meanwhile, Della Volpe and Esposito (2020) has conducted study in Italia and find data scientist's skills gap between universities provided and companies needed. According to this research, analytical skills which included information retrieval, data storage, cloud computing, business intelligence, sentiment analysis, text mining and predictive analysis has a significant gap.

In contrast to those articles, Indonesia itself has just developed the data scientist profession lately. This delay certainly brings several impacts for the data scientist profession in Indonesia. One of the impacts is the gap in skills between data scientists in Indonesia and industry needs. Data scientist demand in Indonesia has continuingly increased, yet the supply of this profession is still inadequate. The unavailability of specialized education programs related to data science is one of the factors of inadequacy. The following research questions have been conducted this research, firstly, what are the current skills and capabilities set that is considered important as a data scientist in Indonesia? Secondly, what are the

Corresponding author
feliks.prasepta@atmajaya.ac.id; chrisnatalia@atmajaya.ac.id; Nicolet.201704530077@student.atmajaya.ac.id
Research Synergy Foundation
DOI : 10.31098/cset.v2i1.506

actual capabilities of data scientists in Indonesia with the capabilities needed by data scientists generally by industry? Thirdly, what are the gaps of data scienntist's capability in Indonesia?

LITERATURE REVIEW

Data analysis exist earlier before data science, wherein John W. Tukey explained that there was a new type of data analysis that focuses on science rather than methodology (Tukey, 1962). Furthermore, the latest type of data processing emerged and was known as data mining in 1980. Data mining itself is known as one aspect of science supporting data science itself. The conference titled "Data Science, Classification, and Related Methods" by the International Federation of Classification Societies in 1996 was the first conference to discuss about data science. The term data scientist first appeared in the literature with "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century" by the National Science Board in 2005 (Zhu and Xiong, 2015). Later on, a dedicated research center for data science was first established in Shanghai, China in 2007 (Zhu and Xiong, 2015). Since then, data science has continued to develop until it becomes the knowledge we know today. As a profession with a wide range of expertise, each data scientist has different and unique skills. The data scientist combines business skills, the ability to build data products in stages, the ability to explore and iterate solutions (Zhang et al., 2020). Since then, data science has continued to develop until it becomes the knowledge we know today. As a profession with a wide range of expertise, each data scientist has different and unique skills.

There are several skills that a data scientist needs to have. These skills are needed to support the daily work faced by data scientists. Zhu and Xiong (2015) stated that in dataology or what we know today as data science, there are three basic theories, namely, data acquisition, data analysis, and data awareness. However, the data scientist theory requires the addition of new technologies such as data exploration, experimental data, data visualization, and so on. As data science develops, it is known that many new technologies and sciences are applied in data science. According to Voulgaris (2014), these skills include model building, planning, problem solving, fast learning, adaptability, teamwork, flexibility, research ability, attention to detail, and reporting. Provost and Fawcett (2013) state that there are analytical techniques that are important for a data scientist to have, including statistics, database querying, data warehousing, regression analysis, machine learning, and data mining. The data scientist profession requires the right expertises, characteristics, and ways of thinking. According to Voulgaris (2014), a data scientist requires traits such as curiosity, experimentation, creativity, systematic work, and communication. A data scientist with a curious nature will observe data and try to solve existing problems. The process of solving these problems will be carried out through statistical and mathematical analysis. Experimental nature is a bold attitude in experimenting with new ideas, developing ideas, testing hypotheses, until the stage of implementing these ideas.

A number of researchers have reported the study on data science skills gap. Li et al. (2021) conducted research to identify data science qualification gap in the United States for manufacturing sector. This study summarised the gaps between demand and supply of qualifications and proposed opportunities for training to bridging the qualification gap. Similar with this research, Dolezel and McLeod (2021) studied the data scientists skills gap in healthcare sector. They conducted a national survey in United States and found a big gap between the demand and supply of skills data science. Organizational needs for data scientists continue to increase, but the availability of this profession is inadequate (Mikalef and Krogstie, 2019). One of the causes of these two impacts is the unavailability of special education programs related to data science. This delay itself needs to be pursued so that there is no big gap in the skills and qualifications of data scientists in Indonesia and globally. According to our knowledge, there is no research to explore the current capabilities and the gaps of data scienntist's capability in Indonesia.

RESEARCH METHODOLOGY

Our study applies the mixed-method approach since the method ensure that research findings are grounded in participants' experiences (Creswell and Creswell, 2017) and no previous research has identified data scientists' capabilities in Indonesia. The data collection for qualitative approach was made through interviews, and questionnaires were used for the quantitative approach. The interview guide was designed in such a way through the primary references of to Voulgaris (2014) and (Mikalef et al., 2017) research. The interviews were done from April 2020 until December 2021. The criteria used to select

research interviewees are: has worked or is currently working as a data scientist for a minimum of 3 years; has worked in the Jakarta metropolitan area, Indonesia and have got training related to data scientist. Interviews were conducted using video conferencing tools such as Google Meet and Zoom Meeting.

We interviewed a total of 15 participants from the 15 organizations. Respondents' criteria from this research are data scientists with a minimum of 3 years of work experience and experience working in the Greater Jakarta area. The interviews' average duration was 60 minutes. We found that saturation happened within the thirteenth interviews based on the interviews data set. The data analysis was done in an iterative manner, as recommended by the grounded theory methodology. The coding procedure began with what is known as microanalysis (Strauss and Corbin, 1998), which is a line-by-line analysis of each semi-structured interview transcript to identify initial codes. The author utilized NVivo 12 to organize and analyse the data in order to produce and label nodes/codes in accordance with Saldana's (Saldaña, 2015) guidelines. A dual coder strategy was used for the coding procedure. All transcripts were coded by the author, resulting in a node structure of indicators and text coding. Following that, two investigators worked with the author to review and modify the coding and related node structure. Then, having the node structure and transcripts in hand, an independent second coder was assigned to code all of the interviews. This coding was compared to the authors' coding, and the reliability of the coding was determined using Cohen's Kappa (Stemler, 2001), yielding an inter-rater agreement of 87.5 percent (Stemler, 2001).

Following the qualitative approach, the quantitative approach was conducted using a closed-type questionnaire, a 5-point Likert scale to answer the research questions. Research indicators for questionnaire were obtained from the results of qualitative approach. Two types of questionnaires were designed: the actual capability and the importance capability questionnaire. The actual capability questionnaire is a questionnaire that discusses the current capabilities of respondents as data scientists. Furthermore, the capability importance level questionnaire is a questionnaire that outlines how important the capabilities of data scientists. The number of respondents was counted by multiplied the number of indicators by 5 (Hair et al., 2007). A total of 65 participants were participated in this research.

Exploratory factor analysis was used to identify the factor structure of data scientists' capabilities in Indonesia. This analysis was applied as we have no hypothesis regarding to the underlying factors of data scientists' capability in Indonesia. Furthermore, the relative competence and gap analysis was done to identify the comparison and gap between the actual capabilities of data scientists and its importance. There are 3 criteria to assess the gap; negative, adjustment margin, and positive. If the discrepancy is negative, then it is necessary to provide the immediate actions to close the gap for the capability as soon as possible, for example provide training to them. The adjustment margin gap indicates the need for improvement for these capabilities but not urgent, while the positive criteria conclude that there is no capability gap, so no need to take actions for these capabilities (Hansson, 2001).

FINDING AND DISCUSSION

The demographic data of the 15 interviewees from qualitative study shows that most of interviewees, 60% of total respondents, work as data scientists in their organizations. The rest of respondents have a title as Advisory Team, Commissary, Data Scientist Manager, Data Scientist Assistant Manager, Senior Data Analyst and Customer Facing Data Scientist. It represents any related position and rank regarding data scientist in an organization. Participants of this research came from various organization to increase the heterogeneous of the respondents in order to achieve more generalisation of research findings. The highest participants (34%) were from technology companies, followed by financial institution (27%). The rest of respondents were divided equally (13%) from start up, insurance and consulting companies.

Following qualitative study, the research continued with quantitative research. The research indicators for questionnaire were obtained from the results of qualitative approach. Questionnaires were distributed to 65 respondents who had met the selection criteria. The questionnaire distribution was carried out using Google Form from January 28, 2021, to March 10, 2021. Validity and reliability tests were carried out to determine the question instruments used in the questionnaire. The questionnaire is

declared valid if the Pearson correlation value is higher than 0.361. There are 13 indicators for both questionnaires that validity score are more than the 0.361 value. For reliability analysis, the questionnaire is reliable if the Cronbach Alpha value more than 0.7. The Cronbach Alpha value obtained for both questionnaires are 0.730 and 0.753; thereby, both questionaries are reliable.

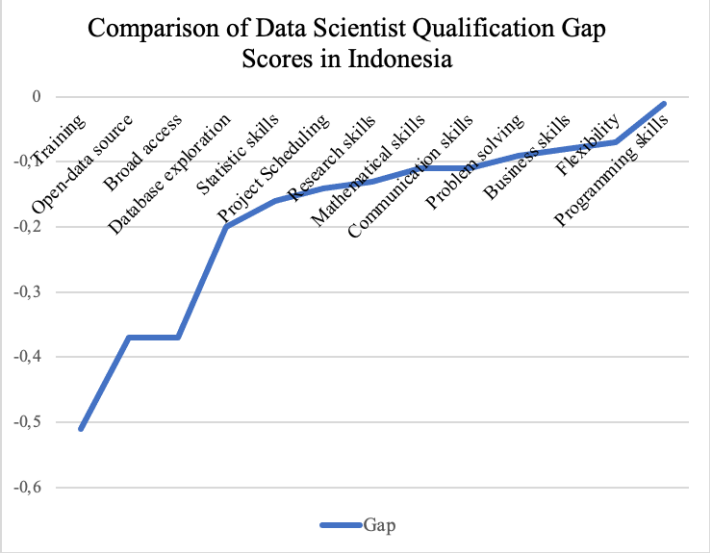

Figure 7. Gap Comparison of Data Scientist in Indonesia

Figure 7 shows the comparison of data scientist capabilities gap scores. The capability gap scores were ordered from the smallest value to the largest value. The smallest capability gap was found in training indicator; therefore, it's concluded that organization in Jakarta metropolitan area has implemented adequate training for data scientists. Although there are no negative gap criteria, providing training and adjusting capabilities for data scientists in the Jakarta metropolitan area is still needed. Training can be provided by the organization or obtained individually by data scientists. Based on the interview data, the organization generally provides training in the form of online courses. The online course can be accessed through various platforms. Data scientists, fresh graduates, and students can develop their capabilities by utilizing online courses platforms.

Based on the interviews, participants stated that there are still some misunderstandings regarding the capabilities of data scientists. As a result, numerous fresh graduates are considered incapable of working in terms of their capabilities. Interviewees stated that the essential skills that data scientists need to have include mathematics, statistics, computers, and business. From the results of quantitative data, it is known that the four capabilities are included in factor 1. Factor 1 can also be referred to as the primary capability of the data scientist. The skill sets include statistics, mathematics, programming, communication, problem-solving, project scheduling, and business.

A mixed-method was applied to assess whether the data from the qualitative approach can be generalized to a sample population in the quantitative approach. Based on the results of the research questionnaire, seven primary capabilities are known needed by data scientists: statistical skills, mathematics, computers, communication, problem-solving, project scheduling, and business. Divided based on the types of skills, there are several indicators in technical skills and general skills. The technical skills required are statistical, mathematical, computer, and business skills. Statistical skills are needed by a data scientist, such as regression, correlation, and so on. Mathematical capabilities such as mathematical logic are used to support the ability to analyse business problems. Programming skills are used to support the tools needed by data scientists, such as SQL, Python, R Language, data visualization, and others. Business skills are no less critical than other technical abilities; data scientists can find solutions with a large and profitable impact or influence on the organization by knowing business problems in detail.

Meanwhile, the general skills required by data scientists based on data processing are communication skills, problem-solving, and project scheduling. As a data scientist, of course, their work activities involve

various parties such as stakeholders, other divisions, and business users. Communication skills play an essential role in this case, wherein if there is no good communication, the task will not be optimally carried out. Communication is also needed when insight has been obtained from business problems and a machine learning model has been designed; data scientists are expected to convey the results obtained to business users and stakeholders properly.

Problem-solving ability achieved the highest value for data scientists' skills. Data scientists are expected to solve business problems by using data as a support for solving these problems. If a data scientist does not have the problem-solving ability, the organization would not significantly impact data scientists. This is contrary to the primary function of a data scientist who acts as a problem solver in the organization. Project scheduling is also essential; data scientists are expected to schedule and complete the project on time with a specific timeframe. With poor project scheduling skills, the project will be delayed, hampering the organization's productivity.

The second factor, data management, is known to support data scientists in their daily work. Indicators that consist of data management include broad access, database exploration, and training. With general access, data scientists can obtain data from various sources, and the opportunity to provide accurate and effective solutions is higher. Database exploration needs to be done to find out the appropriate database for multiple problems. Interviewees reckon that sometimes different issues require different databases. This is because there are additional features in each database. By doing exploration, data scientists can immediately work on points with the correct database. Training aspect needs to be continuously done by data scientists. As previously explained, training can be provided by organizations or individuals. Training serves to improve the ability and knowledge of data scientists at work.

Some indicators of ability are not included in the primary capabilities; among others, these indicators are flexibility and research skills. The two indicators are grouped into factor 3 and given the name of personal skills. Personal skills are the ability to interact at work. Flexibility in working as a data scientist is needed to support communication skills. Having good flexibility will facilitate collaboration in work. In addition, complex work causes data scientists to need to solve problems quickly and agilely, wherein flexibility is required to run effectively. Based on interviews, it is known that data scientists need research skills because data science is still developing. Hence, data scientists must always know the latest news or research related to data science. The resource persons stated that their research abilities positively impacted them because sometimes they found solutions to the problems they faced at work. The fourth factor is given the name of sources. It is known that open-data sources get a low score because most organizations in Indonesia do not implement data retrieval from open sources. Still, there are certain specifications for data sources.

CONCLUSION AND FURTHER RESEARCH

There are four capability factors for data scientists in Indonesia obtained from the exploratory factor analysis method. The four factors are Main Capability, Data Management, Personal Skills, and Source. The primary capabilities needed by data scientists are problem-solving, statistics, mathematics, computers, business, communication, and project scheduling skills for project-based jobs. The value of current expertise in the gap analysis method shows that the average value of the 13 indicators has a relatively high range of values, namely 3.78 to 4.83. This value indicates that the actual conditions of data scientists in Indonesia are equivalent, and there are no significant differences. The gap analysis method shows that all research indicators have an adjustment margin type, so it is necessary to provide data analysis related-training for data scientists.

One of the benefits of this research is universities or higher educations can design a curriculum that is in accordance with the main capabilities needed by a data scientist. By providing appropriate lecture material, students can better master data science capabilities. We suggest for further research to compare the capabilities of data scientists in Indonesia with other countries that have previously applied data science in business, so that new capabilities and knowledge can be learned to improve the performance of data scientists in Indonesia.

REFERENCES

Baumeister, F., Barbosa, M. W. & Gomes, R. R. 2020. What Is Required To Be A Data Scientist?: Analyzing Job Descriptions With Centering Resonance Analysis. *International Journal Of Human Capital And Information Technology Professionals (Ijhcitp),* 11**,** 21-40.

Creswell, J. W. & Creswell, J. D. 2017. *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches*, Sage Publications.

Davenport, T. H. & Patil, D. 2012. Data Scientist. *Harvard Business Review,* 90**,** 70-76.

Della Volpe, M. & Esposito, F. 2020. How Universities Fill The Talent Gap: The Data Scientist In The Italian Case. *African Journal Of Business Management,* 14**,** 53-64.

Dhar, S. & Mazumdar, S. Challenges And Best Practices For Enterprise Adoption Of Big Data Technologies. Technology Management Conference (Itmc), 2014 Ieee International, 2014. Ieee, 1-4.

Dolezel, D. & Mcleod, A. 2021. Big-Data Skills: Bridging The Data Science Theory-Practice Gap In Healthcare. *Perspectives In Health Information Management,* 18.

Hair, J. F., Money, A. H., Samouel, P. & Page, M. 2007. Research Methods For Business. *Education+ Training*.

Hansson, B. 2001. Competency Models: Are Self‐Perceptions Accurate Enough? *Journal Of European Industrial Training*.

Harris, H. D., Murphy, S. P., And Vaisman, M. 2015. Analyzing The Analyzers: An Introspective Survey Of Data Scientists And Their Work.

Li, G., Yuan, C., Kamarthi, S., Moghaddam, M. & Jin, X. 2021. Data Science Skills And Domain Knowledge Requirements In The Manufacturing Industry: A Gap Analysis. *Journal Of Manufacturing Systems,* 60**,** 692-706.

Mikalef, P., Framnes, V. A., Danielsen, F., Krogstie, J. & Olsen, D. H. Big Data Analytics Capability: Antecedents And Business Value. Pacific Asia Conference On Information Systems (Pacis), 2017. 1-13.

Mikalef, P. & Krogstie, J. Investigating The Data Science Skill Gap: An Empirical Analysis. 2019 Ieee Global Engineering Education Conference (Educon), 2019. Ieee, 1275-1284.

Provost, F. & Fawcett, T. 2013. Data Science And Its Relationship To Big Data And Data-Driven Decision Making. *Big Data,* 1**,** 51-59.

Saldaña, J. 2015. *The Coding Manual For Qualitative Researchers,* London, Sage.

Stemler, S. 2001. An Overview Of Content Analysis. *Practical Assessment, Research & Evaluation,* 7**,** 137-146.

Strauss, A. & Corbin, J. 1998. *Basics Of Qualitative Research: Techniques And Procedures For Developing Grounded Theory,* London, Sage Publications, Inc.

Tukey, J. W. 1962. The Future Of Data Analysis. *The Annals Of Mathematical Statistics,* 33**,** 1-67.

Voulgaris, Z. 2014. *Data Scientist: The Definitive Guide To Becoming A Data Scientist*, Technics Publications.

Zhang, A. X., Muller, M. & Wang, D. 2020. How Do Data Science Workers Collaborate? Roles, Workflows, And Tools. *Proceedings Of The Acm On Human-Computer Interaction,* 4**,** 1-23.

Zhu, Y. & Xiong, Y. 2015. Towards Data Science. *Data Science Journal,* 14.