Not All Software Engineers Can Become Good Data Engineers

Jeffrey S. Saltz
Syracuse University
Syracuse, USA
e-mail: jsaltz@syr.edu

Sibel Yilmazel
Anadolu University
Eskisehir, Turkey
e-mail: syilmazel@anadolu.edu.tr

Ozgur Yilmazel
Anadolu University
Eskisehir, Turkey
e-mail: ozgur@anadolu.edu.tr

Abstract— **The amount of data that businesses collect and analyze has been rapidly increasing, which has triggered an increase in big data teams. With the growth of both the number and size of big data teams, specialized roles are starting to be defined. One such role is the data engineer, who focuses on ensuring that the data is easily available for advanced analytics. Via a case study, this paper explores the role of the data engineer and the key characteristics that enable someone to be a good data engineer. The paper also explores if good software engineers could become good data engineers. Our findings show that the knowledge and skills required to be a data engineer are significantly different from those required to be a software engineer. Hence, not surprisingly, we found that that not all software engineers could become good data engineers.**

Keywords-Data Science; Big Data; Process Methodology.

I. INTRODUCTION

There has been a tremendous increase in the amount of data that businesses produce. Along with the size of data, the complexity and variety are also increasing, as demonstrated by the significant growth in the collection and analysis of unstructured data. This substantial growth of big data can also be observed via the increase in the number of data science programs offered by universities [1,2].

Naturally, there has been significant focus on big data analytics, and the role of the data scientist to leverage big data [3]. However, another role typically required within a big data team is that of a data engineer. Much less has been written about data engineers. In fact, little has been written about skills and responsibilities required of a typical data engineer. The role of a data engineer seems similar to that of a software engineer. In both roles, the person is focused on writing software that will be used by others, or software that will generate results that others will then leverage. Also, for both data projects and more traditional software projects, the methodologies and project lifecycle are similar in terms of the need to design the system, write the code and then execute the test scripts.

In order to understand the role of a data engineer we conducted a case study at a big data consulting organization. In particular, we focused on the data engineers hired at that organization during the past five years. The research questions that we pursued were:

RQ1. What was the role of a data engineer?

RQ2. How do data engineers compare with Software Engineers?

RQ3. Do good software engineers make good data engineers?

It is our belief that understanding the role of a data engineer can have a positive impact on teams doing big data work, in that those teams will benefit from having a clearly defined role, working on a key aspect of the big data challenge that is often neglected.

Hence, to help start the dialog of possible skills and responsibilities required by a data engineer, the rest of this paper first documents what has previously been published within this domain. Then, in section three, we describe the methodology used within our case study. Section four describes our findings and section five discusses our results. Finally, in section six, we present our conclusions and possible next steps.

II. RELATED WORK

Certainly much has been written about the use of data science to generate useful insight from data. In fact, within a corporate context, data is increasingly being viewed as a strategic resource for the organization [4]. To help teams successfully execute big data projects there has been some work exploring the critical success factors that can help a team do a big data project [5,6] as well as some research on agile methodologies one could use to execute a big data project [7,8]. However, there is still much to be done. In fact, it has been observed that there are significant challenges in trying to leverage data in a strategic manner [9,10].

One of the challenges that has been noted is the lack of focus on understanding the key roles required to actually execute a big data project [11] and that the required skill set for dealing with big data has not yet been studied empirically [12]. While there has not been much research on big data roles, in one analysis that has been done of the skills required within a big data team, a competency taxonomy for the required big data skills was developed by doing a latent semantic analysis on job advertisements harvested from an online employment platform [12]. Not surprisingly, this analysis found that business knowledge was as important as technical skills for working successfully on big data initiatives and that people filling big data roles needed both strong software development and statistical skills.

Another challenge is that, to date, the big data talent discussion has largely focused on a single role – the data scientist. However, another key role required within the data science team is the data engineer [13]. A data engineer has been described as the "bridge between raw technical

computing and storing and IT infrastructure, and your data scientists, who are spending time in R and other languages to build models" [14]. In other words, a data engineer connects the data scientist to the IT data management group by making data more accessible to the data scientist and providing the tools a data scientist would want to use [14]. This enables the data scientist to focus more on developing models (i.e., developing the insight), and not on the data munging required for the data modeling.

While the data engineer job is one of the fastest growing job categories [15,16], there has been little research to understand this role. In fact, with respect to the data engineering role, there was only one article found in recent literature. That analysis explored the role of the data engineer [17], and focused on the skills required to do data engineering. In summary, the data engineer needed the ability to:

- Extract, clean, and integrate data (wrangling)
- Bridge between data models and production systems
- Implement computational algorithms at scale
- Leverage both relational and noSQL databases
- Protect customer privacy and anonymity

Since data engineering is new and data engineers are in high demand, a natural avenue to explore in the hiring of a data engineer is to hire a software engineer and then "convert" that person into a data engineer. However, it is not clear how effective an organization could be at actually converting traditional IT professionals into data engineers [14]. One way to gain a better understanding of the roles within a data science team is to document case studies of how teams are actually doing data science [11, 18]. Hence, we performed a case study aiming to gain a better understanding with respect to an organization being able to convert software engineers into data engineers.

III. DATA COLLECTION

A. Background

A case study on the hiring and use of data engineers was conducted within a big data consulting company. This organization consults mainly telecommunication, retail and financial companies with their big data projects. The projects vary from ETL to data analytics software platforms and from security integration to recommendation systems within the customers' big data ecosystem. The company often partners with one of the leading distribution providers of Apache Hadoop for the enterprise. The team's average number of employees ranges between twelve and eighteen depending on the project contract sizes.

During the past five years, the organization has significantly increased their data engineering team. To achieve this growth, the organization explored many avenues to hire data engineers. This was required because, since this is a new field with significant demand, there were a limited number of data engineers available for hire.

The company believed that only searching for engineers who were experienced with data analytics tools and had Hadoop-like programming experience might result in missing inexperienced but potentially good employees. Since there was (and continues to be) high demand for data engineers, the organization needed to explore these alternative candidates. In fact, the company did hire this type of employee (i.e., that did not have experience with Hadoop technologies, but were capable of understanding real world business problems and solving those challenges with the tools and programming languages with which they were already familiar). In these situations, the organization trained their new employees via an internal big data training program.

To hire Data Engineers, the organization typically posted job ads on local university communication boards and national job search websites. However, for the past two years, they have relied mostly on LinkedIn, university communication boards, big data focus groups and the organization's big data trainee references. The university communication boards were and still are one of the most effective ways to reach college students/graduates to hire. However, national job search websites are not perceived to be as effective as they were five years ago. LinkedIn and big data specialized groups seem to be more effective than national job search websites. It is believed that the reason driving this change is that there are many resumes with outdated content, references that we are not familiar, etc.

Like many organizations, this organization is more likely to interview a candidate who's been referred to the organization or whose connections/references are already known. After the initial screening, the interview process starts with one-on-one interviews in their office. The candidates are asked technical questions about their programming experience, database skills, Linux, algorithms and network. The organization then provides the candidate with take home questions and the candidate has one day to email the solution back to the organization. That homework usually includes a data engineering problem where the candidate is provided a small dataset along with an "assignment". Some example assignments include:

- There is one TB of data on disk. The data is a set of tuples, and one of the elements (columns) in the data set is the first name of the person. The task is to sort and count the unique first names in the dataset, and then produce an output dataset that will have the first name and the count (i.e. a list of tuples with each tuple consisting of <first name> and then <count>).
- There is a paperback dictionary available for use. Describe the algorithm, and appropriate pseudo-code needed to find the word 'mingle' in the dictionary.
- When working with a giant online electronics store, how might one discover that their customers are about to leave the online store?
- There is a sorted linked list with the number 2, 6, 8, 10, 12, 13, 16, 19, 20, 35, 37, 43 and 45. Write the code to insert 17 into that sorted linked list.

- There is a sorted array with x number of elements. Write the code to search for a given element in the array.
- Describe what happens when you type 'www.linkedin.com' on the browser and hit enter.

Hence, as one can see from these interview challenges, the role of the data engineer, at least at this big data consulting organization, requires knowledge of both basic algorithm coding (as shown in the sorted list question) as well as basic data manipulation skills (as shown in the tuples question). The candidate also needs to understand basic web architecture (as shown in the browser question). This skillset is consistent with Tamir [17], and demonstrates the breadth of skills required to be a proficient data engineer.

B. Data Collection

Information was collected for employees who have worked at the consulting firm during the past five years. Data collected included documented performance reviews, exit interviews, internal emails and other types of team communication. In addition, information was collected via informal discussions and semi-structured employee interviews.

Specifically, during the past five years, the company hired twenty-two data engineers. The roles, responsibilities, and skills required were investigated for each of those twenty-two data engineers. Out of the twenty-two engineers, thirteen are still working at the company. Three of the engineers left the firm for personal reasons and the remaining six engineers did not work out, and were asked to leave the organization. All of the employees worked in Europe, typically at the central office of the consulting organization.

IV. FINDINGS

A. Skills Needed

At this company, data engineers were typically grouped by business domain (such as telecom, retail and finance). Although the project managers stay the same throughout the project, most of the engineers rotate between teams so that they can become familiar with different domains, explore different sets of problems, and allow the organization to optimize resource utilization. With this in mind, data engineers need both hard and soft skills.

With respect to specific technical skills (sometimes noted as "hard skills"), it was important to have a solid foundation in basic programming. In particular, languages such as Java and Python were deemed to be very important. As previously mentioned, the organization did not expect new data engineers to have any experience with MapReduce or Spark related technologies, since there were very few people who actually knew how to program and use these tools. In fact, the organization believed that having programming experience in Python, Java and Linux was enough to get started learning the big data technology ecosystem.

In addition, basic knowledge of Linux was also perceived to be very useful. Furthermore, with respect to data structures and algorithmic techniques, newly hired data engineers were expected to have knowledge of typical data structures and algorithms that might be useful within a big data environment. Finally, in terms of analytical skills, it was expected that data engineers would be able to identify relationships between different types of data, using basic data science / data mining techniques.

With respect to softer skills (such as communication skills that are often less technical in nature), since problem solving was at the heart of the role, it was perceived to be very important that data engineers had creative problem solving skills. In addition, being able to work diligently with data, such as collecting data from different sources and performing data transformation were also a key aspect of the role. Finally, system design skills were perceived to be very important in that a big data engineer needed to be able to help construct system architectures.

B. Responsibilities

The main responsibility of a data engineer was to get data from different data sources and convert the data into the appropriate / desired format. Note that the role involved the full life-cycle of development, including the design, build, test and maintenance of highly scalable data management systems. This also included supporting the end users with the right analytics tools and capabilities.

A focus on data quality was deemed very important, such as understanding relevant data standards and ensuring the converted / imported data was correct.

While not required at the start, when a data engineer was hired, that person was expected to learn the Hadoop ecosystem and its related components. In this regard, data engineers were responsible to know the Hadoop ecosystem and define system architectures that leverage the appropriate Hadoop ecosystem components. In addition, the data engineer needed to be able to integrate new data management tools into a customer's existing infrastructure.

Finally, data engineers needed to develop custom software components and analytics applications as needed. The type of coding ranged from writing a software component to import a new data set to implementing a new data mining algorithm.

C. Types of Data Engineers

There were two levels of big data engineering positions. A *big data solution engineer* was responsible for translating business requirements into system requirements, developing big data software applications and test scripts and implementing and supporting big data analytical solutions. A *big data architect* usually works on multiple projects and

architects big data infrastructure. This role translates functional and technical requirements into a design and then leads the big data solution engineers to implement the data solution in accordance with the architectural design. In other words, the big data architect selects the big data solution software, defines the hardware requirements and leads the big data solution engineers. Typically, at this company, successful big data solution engineers with 3-5 years experience become big data architects.

Note that this is progression is similar to software engineers (who write document per a specification) and senior software engineers/solution architects (who design the software system to be built).

D. *Four use cases of an Unsuccessful Data Engineer*

Below, we document the four uses cases where a good software engineer was not a successful data engineer. The goal was to understand some of the key drivers for success (or key drivers for a person not being a good data engineer).

Person A had a graduate computer engineering degree with work experience using Java and database design. This person demonstrated problem solving skills with complex data structures during their interview and worked as a data engineer for approximately one year. During the exit interview, Person A stated that being a data engineer required exposure and knowledge of different business domains, and this person was not happy with this construct – in that this person wanted to develop a software product for a particular business sector. In short, Person A was not happy in a service oriented consulting role. Hence, Person A would also not have succeeded as a traditional software engineer within a broad-based consulting firm.

Person B had a computer engineering graduate degree with object oriented work experience and as well as experience using the Python programming language. Person B had experience in several different business domains, via previous experience working to provide onsite customer support. However, Person B only lasted eight months within the organization. During the exit interview, Person B stated the customer (of the project) often changed the project requirements. These changes required changes to the analytics algorithm and the business rules associated with the analytical code. This scenario is often encountered within a big data project. There is currently no "one size fits all" type of approach to big data projects. Even when a company has a clear vision and strong leadership with its data strategy, it is likely that the requirements will change as the project evolves. The data engineer needs to adapt to this fluidity in requirements. Person B was not equipped with a desire to find solutions to the changing requirements. The ability to think your way through any situation is a must for a data engineer. Note that this is different than most software engineers. For most software engineers, requirements are well defined. An agile scrum methodology might allow requirements to be defined incrementally, but the change in requirements for data engineers is driven by the need to try different solutions and find "insight in the data".

Person C also had a computer engineering graduate degree, with work experience using Java and SQL. Person C lasted about one year as a data engineer. Person C was trained to use the Hadoop ecosystem, with a focus on using MapReduce. Person C had a difficult time adapting to the evolving Hadoop technology stack. For example, as the big data ecosystem evolved and, for example, MapReduce was replaced with Spark, this person struggled. Person C had a hard time adjusting to Spark. Although Person C's had experience with general purpose programming languages, this person was not able (or did not have the desire) to keep learning the new Hadoop technologies while working with the existing technologies within a project. This is different than most software engineers, in that the technical stack used by software engineers is often fairly stable (ex. Java development). While there are certainly new programming languages and platforms that are used (such as Ruby), the pace of change is much higher within the big data technology ecosystem (and this high rate of change seems likely to continue for some time).

Person D had a computer science graduate degree. This person had .Net programming experience. Person D only lasted as a data engineer for three months. This person had a very hard time adjusting to new datasets and data transformations, particularly gaining the ability to leverage scripting languages. Being able to work with scripting languages to manipulate data is a must for data engineers. This speaks to another difference between software engineers and data engineers, in that data engineers need to be comfortable with a wider range of "tools" (such as leveraging scripting languages and programming languages and database scripting).

V. DISCUSSION

A. *Key Attributes of Successful Data Engineers*

Table 1 compares the key attributes of software and data engineers, which are discussed in more detail in this section.

Attribute	SW Engineer	Data Engineer
Passion to Learn	✓-	✓+
Finding Practical Solutions	✓	✓+
Strong Technical Competence	✓+	✓+
Excellent Communication Skills	✓	✓+
Works well in a team and as an Individual	✓	✓+
Ownership Through Lifecycle	✓-	✓+
Approach problem from Business Perspective	✓-	✓+
Adaptability	✓-	✓+

Table 1: Comparing Software & Data Engineers

Key: ✓+ (very important), ✓(important), ✓- (not as important)

Passion to Learn - Big data technologies are constantly evolving. As a result, a data engineer has to be able to monitor new technologies and familiarize themselves with these emerging technologies. A data engineer has to be able evaluate all the technologies in the ecosystem and decide which is best for the customer, based on the client's environment and requirements. Though a software engineer can be specialized as a Java programmer, mobile developer or SQL database manager, being able to adapt to new technologies is important, but occurs at a slower pace. In addition, the big data engineer has to have a deep knowledge of all the technologies in the field. This includes the ability to gain knowledge of new technologies and the interest to learn some of these new technologies on the person's own time (such as using internal training materials, leveraging external MOOCs, or in general, leveraging any relevant resources that they can find).

Finding practical solutions – The engineer needs to be able to develop practical solutions (on time, on budget). Since many customers have their own use cases and almost all have legacy databases and applications, this means that the data engineer needs to understand the data and the data flow of the client. Being creative when finding practical solutions for data capture and transform was identified as a critical success factor. For example, being able to describe the big data problem and possible big data solutions. The creativity is required since there are no defined programs/rules in big data (as there may be in other parts of the software engineering world).

Strong technical competence – specifically using data structures in coding solutions, understanding physical database design principles when developing solution designs and working with open source software technologies for both the design and development of client solutions.

Excellent communication skills – ability to understand data and data transformation needs. Incorporates customer needs into technical solutions.

Works well in a team and as an individual – Sometimes the engineer needs to work in a one-on-one environment, but other times, the person needs to perform well in a team environment (with different types of customers and/or other technical team members). Most software engineers or software developers work within project teams. Though our focus is comparing software engineers who work within teams with data engineers who also work within teams, one difference is that the data engineer might be the only data engineer within that team, as compared to software developers, where there are typically many developers withina team. ~~as most of the time data engineer is the only one on the team.~~ Therefore a data engineer needs to be able to perform well in both individual and team environment.

Ownership Through Life Cycle - Unlike most software engineers, a data engineer has to have ownership throughout the data life cycle, including requirements gathering, solution development and ongoing support. Most 'traditional' software projects consist of different roles within the project. In contrast, most of the time, a big data engineer assumes the full range of responsibilities. In other words, the data engineer typically does what is done by multiple roles within a software project, such as a business analyst who gathers requirements, a software architect who converts the requirements to a system design, and a developer who actually develops the code (which, for a data engineer, would include tasks such as extracting, transforming and loading data).

Approaching the problem from a Business Perspective - A data engineer needs to approach the problem from a business perspective. In other words, being solution driven. This is different than a software engineer that is good at programming and typically has other technical skills such as SQL proficiency, but might not be required to provide an end-to-end solution for the client. Hence, data engineers need strong business problem solving skills as well as strong technical skills, but a software engineer typically needs to focus on just having strong technical expertise (i.e. programming skills).

Adaptability - software engineers are typically trained to produce software in a more controllable situation. This is actually part of the definition of the "software engineering" term - to make the production of software more quantifiable and controllable. Hence software engineers try to use reusable techniques and methods to create their solutions. This contrasts with the daily tasks of a data engineer, who might have problems that need to be solved once (or run on a daily basis). In the case of a "one-off" situation, sustainability or repeatability of the process/code is not a big concern. In these situations, a creative data engineer who can solve the problem at hand with an existing tool will be more appreciated than an engineer who creates a nicely designed and documented code module.

VI. CONCLUSION

This paper reports on the results of a case study within a big data consulting company. Based on the analysis of the use of data engineers over a five year period, we identified several key characteristics of the data engineer.

With respect to our first research question (the role of a data engineer), it is clear that data engineers work across the life cycle of a project to extract, transform, load, manipulate and analyze large data sets. This work requires in depth computing and data knowledge, as well as excellent "soft skills" to interact with clients.

Due to these broader skills and characteristics required of good data engineers, not all excellent software engineers

make excellent data engineers (hence answering the last research question - do good software engineers make good data engineers). In fact, we can leverage an object oriented programming concept called 'inheritance' (or IS-A relationship). Our study revealed that there seems to be an IS-A relationship between a data engineer and a software engineer, in that the data engineer IS-A software engineer. But just as the IS-A relationship is unidirectional we also believe that a software engineer not necessarily is a data engineer. Hence, we addressed research question 2 (comparing data engineers with software engineers), in that both data and software engineers need technical excellence, but that data engineers often require a broader set of skills.

On a similar note, in the early days of the programming era most software developers wrote COBOL programs for reporting purposes and data analysis. Most of these developers were not graduates of a formal computer science program, but over time, it became clear that a robust computer science degree and set of processes was essential for the field. Similar to this, we see that more structured and formal processes (and likely educational programs) need to be established for data engineers, *vis-à-vis* they also need to be agile to accommodate the fast paced evolution of the data analysis requirements.

Finally, the most pressing next step is to explore the role of data engineers across a broader spectrum of organizations. For example, based on our initial results, it would be interesting to validate or refine our results via a survey across a broad array of organizations executing big data projects.

REFERENCES

[1] O'Neil, M., (2014), "As Data Proliferate, So Do Data-Related Graduate Programs," The Chronicle of Higher Education, DOI= http://m.chronicle.com/article/As-Data-Proliferate-So-Do/144363

[2] Violino, B., (2014), "The Hottest Jobs In IT: Training Tomorrow's Data Scientists," Forbes, DOI=http://www.forbes.com/sites/emc/2014/06/26/the-hottest-jobs-in-it-training-tomorrows-data-scientists/

[3] Van der Aalst, Wil MP. (2014), "Data scientist: The engineer of the future." *Enterprise Interoperability VI*. Springer International Publishing, pp 13-26.

[4] Wade, M., and Hulland, J., (2004), "Review: The Resource Based View and Information Systems Research: Review, Extension, and Suggestions for Future Research," MIS Quarterly, 28(1), pp. 107–142.

[5] Gao J., Koronios A., Selle S. (2015). "Towards A Process View on Critical Success Factors in Big Data Analytics Projects", Twenty-first Americas Conference on Information Systems (AMCIS), Puerto Rico

[6] Saltz, J., and Shamshurin, I., (2016). "Big Data Team Process Methodology: A Literature Review and the A Literature Review and the Identification of Critical Factors for a Project's Success", in *Big Data (Big Data), 2016 IEEE International Conference on*, in press.

[7] Dharmapal S., Sikamani T. (2016). Big Data Analytics Using Agile Model. International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).

[8] Saltz, J., Shamshurin, I., Crowston. K., (2017). "Comparing Data Science Project Management Methodologies via a Controlled Experiment", Hawaii International Conference on System Science (HICSS), in press.

[9] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A., (2015), "Big Data: The Next Frontier for Innovation, Competition, and Productivity", McKinsey & Company.

[10] Chen, H., Chiang, R., and Storey, V., (2012),"Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly Executive, 36(4), pp. 1165-1188.

[11] Saltz, J., (2015), "The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness," 2015 IEEE International Conference on Big Data.

[12] Debortoli, S., Müller, O. & vom Brocke, J. Bus Inf Syst Eng (2014) 6: 289. doi:10.1007/s12599-014-0344-2

[13] Miller, S., (2014), "Collaborative Approaches Needed to Close the Big Data Skills Gap", Journal of Organization Design JOD, 3(1): 26-30 DOI: 10.7146/jod.9823

[14] Fitzgerald, M. (2015). "Gone Fishing--for Data". MIT Sloan Management Review, 56(3).

[15] Data Engineer Job Trends. (2014). Indeed.com. http://www.indeed.com/jobtrends?q=data+engineer

[16] Data Engineer Jobs. (2014). LinkedIn. https://www.linkedin.com/job/q-data-engineer-jobs

[17] Tamir, M., Miller, S., and Gagliardi, A., (2015), "The Data Engineer". http://dx.doi.org/10.2139/ssrn.2762013

[18] Saltz, J., and Shamshurin, I., (2015),"Exploring the process of doing data science via an ethnographic study of a media advertising company," in *Big Data (Big Data), 2015 IEEE International Conference on*, pp. 2098-2105: IEEE.