A Study of Various Roles in a Successful Data Organization

Vishal Srivastava

[#]*Graduate, Computer Science, Columbia University*
New York City, USA

Abstract : *Apart from the fact that Data Science is a fast evolving field, it is also a term that combines many roles and functions into one. For large enterprises and governments often the question is : Where to start?. The key challenge that derives most inertia towards having a successful data organization within a large company is the unknown components and how they structure and operate together. Further, which components are an absolute requirement v/s optional or good to have in the start. As the field of AI (Artificial Intelligence) is quickly evolving, it is even more necessary for enterprises to build a strong data organization that is nimble to adapt to the technological advancements. Keeping these challenges in mind, this paper attempts to provide various roles, their definition and a framework that can be used to structure a well equipped data organization.*

Keywords — *Data Science, Organizational structure, Enterprise,Artificial Intelligence, Machine Learning.*

I. INTRODUCTION

IBM predicts that the demand for data scientists will soar 28% by 2020[1]. The market for data scientist and enterprise appetite for data science is not slowing down. It is increasingly more important for data science organizations to be well structured, efficient and advanced to be successful.

Data Science is generally a relatively new organization unlike the vital orgs of marketing, IT, etc in a company. Hiring and retaining the talent is extremely critical for implementing a successful data strategy. Implementing a good hiring strategy also takes understanding of how the org operates. On top of that a company's recruitment department have divided attention across other departments like sales, marketing and IT. Due to lack of enough talent, the pay-scale is also very high. Outsourcing and hiring contract workers is one strategy that many employers choose but even that requires a good understanding of how the org functions.

So, the challenges that come with executing a successful data strategy generally starts with understanding the various roles in a data organization. A single term is a misuse to define all the roles that come together for a complete data strategy. This paper attempts to define the various roles and provide context on them by doing research on the importance of these roles and strategies and tools that they use.

II. THE ROLES IN A DATA ORGANIZATION

Data science starts with storing large amounts of data efficiently. Managing the life cycle of big data can be challenging and is comprised of data generation, data acquisition, data storage and data analysis[5]. Table 1 below shows the approximately average number of monthly jobs currently open for the above roles from top job sites indeed.com and monster.com. It clearly shows the strong demand for each of the roles on their own (Limited to searching with terms).

Title/ Opening	DE	DS	RS	MLE	DSM	DevOps	DV
Indeed	45	75	25	30	55	85	35
Monster	33	50	20	27	42	67	29
Total	78	125	45	57	97	152	55

Table 1. The summary of the job openings (in thousands) from two top sites for roles in a data organization

Some of these roles can be consolidated. Broadly classifying the roles can be following:

A. Data Engineer (DE:

The data engineer is responsible to develop, tests and maintains architectures, such as databases and large-scale processing systems similar to Extraction-Transformation-Load (ETL). The data engineer is also responsible to build and maintain a data platform that other teams like analytics, data scientists, managers can utilize. It is important to note that data engineering function can also be called as information engineering, information technology engineering or simply software engineering. But the core responsibility is to collect the data that the organization needs, use tools related to big data storage like Hadoop, HDFS,Kafka, etc. and orchestrate data pipelines using tools like Luigi, Airflow, etc. A typical data engineer is a skilled programmer in languages like python, scala or java.

B. Data Scientist (DS):

The data scientist utilizes the data sometimes in raw format or processed format and applies statistical techniques to cleanse the data and mine information out of it. The data scientist needs to be well skilled in statistics, probability, linear algebra and at least one programming language such as R or Python or Julia. It is important for the data scientist to understand the domain and the business problem to be able to look at the datasets from a business perspective. In today's world data scientists also need to be well skilled engineers as is seen in the research conducted by van der et al. [4].

C. Machine Learning Engineer (MLE):

A machine learning engineer is usually building the platforms where models can be developed and deployed. A machine learning engineer should have skill sets of both software engineer and machine learning algorithms. Usually practitioners come from a computer science education background. This is relatively a new job in the field but one of the fastest growing demands is for machine learning engineers.

D. Data Science Manager/Executive (DSM):

A Data Science Manager or Executive leads the team and ties the work to the key strategy of the business. The manager should have a deep understanding of the technologies and methodologies for a quick feasibility analysis as well as project scoping. The manager also helps budget and recruit on the team and evaluate any tools that might be interesting.

In addition to that, one key responsibility of a data science executive/manager is to evangelize the use of data science across the firm and educate other divisions within the firm to look for data science use cases in their day to day work.

E. DevOps Engineer (Machine Learning):

The dev ops engineer is a role that works with data engineers and machine learning engineers to build capabilities for setting up the infrastructure. The processing, training and testing of data is largely dependent on infrastructure availability and scalability and the role of devops is critical in enabling that. The devops engineer also creates automated scripts for continuous deployment and delivery of models and algorithms developed by data scientists or machine learning engineers.

F. Data Visualization Engineer (DVE):

Modeling is only 20% of any data science work. A lot of it is to find useful insights and communicate it to the key stakeholders. The data visualization engineer is able to create charts and dashboards that can be understood by a wider audience. These charts generally showcase insights from underlying complex algorithm or model. The data visualization engineer should be expert in tools like tableau, d3js, matplotlib, etc.

G. Research Scientist (ML/AI/NLP/Vision focused) (RS):

A Research scientist is not a position in every company but is increasingly critical at companies that are competing on the basis of Artificial Intelligence technological advancement. These research scientists are in high demand in companies like Google, Facebook, Uber, etc. A research scientist is usually a doctorate holder and is focused on a narrow area of AI like machine learning, deep learning, computer vision etc. The research scientists works in an R&D environment generating intellectual property and new product innovation for the firm.

H. Other roles:

There are other roles like data governance, data modeler, data architect, data project/product manager, data steward, etc. that are not covered in this study. The study excluded them because of the following reasons:

i. These roles are generally blended into the roles mentioned above.

ii. To limit the scope of the study to data organization and a lot of these roles might be within a technology organization.

iii. Lastly, these other roles generally are only applicable to companies that have special requirements due to regulations, their size, etc.

.

III. BUILD V/S BUY

Although the above roles are key, a factor that needs to be considered is an organization's appetite towards Build v/s Buy. A software acquisition can offset some or all of the need for a particular role.

Build v/s Buy decisions come with a set of challenges. The biggest one is "Decision". There are many factors that Factors mainly applying to large organizations were strategic role of the software, intellectual property concerns, and risk, Factors particularly relevant to SMEs (ubiquitous systems, availability of free download, and customizable to specific government/tax regulations) [2].

IV. PROPOSED SIMPLISTIC TEAM STRUCTURE

A study done by McKinsey analyzed organizational structure based on 7s : Shared Values, Systems, Styles, Staff, Skills, Strategy and Structure [3]. Although most of these parameters are self explanatory and might seem obvious, it is important to note from the study that these are harder to apply and maintain in practice. For a successful data organization it is key to understanding roles as described in this paper but also key to understand relationships between them.

Organizational structures can be very complex as many factors like culture, size, budgets, etc can play an important role. One way of understanding a problem is to look at it from an abstract simplified view. The org structure needs to evolve with the underlying business in an iterative way. Many organizations have started with something and iterations multiple times before finally settling for an org structure.

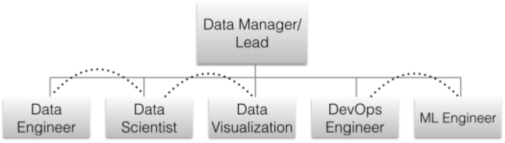

Fig. 1 A Simplistic data org.

In Fig.1, we take an example of a simplistic data organization. The solid lines show the hierarchy of the org but the dotted line represent the shared skills and more frequent collaboration between the roles. Hence if the roles are co-located, they can share a lot in common.

V. CONCLUSIONS

Building a successful data org in a company is a challenging task]. Without the understanding of various roles and tools, it gets even harder. In this paper we looked at various roles in a data org and some strategic viewpoints that go into deciding buy v/s build. The understanding of the roles should be widely communicated among various stakeholders.

REFERENCES

[1] Forbes, “IBM Predicts Demand For Data Scientists Will Soar 28% By 2020” , May 2017.

[2] Farhad Daneshgar, Graham C. Low, and Lugkana Worasinchai. 2013. An investigation of 'build vs. buy' decision for software acquisition by small to medium enterprises. Inf. Softw. Technol. 55, 10 (October 2013), 1741-1750. DOI=http://dx.doi.org/10.1016/j.infsof.2013.03.009

[3] Mohammad Mehdi Ravanfar. Analyzing Organizational Structure Based on 7s Model of Mckinsey," 2015 Global Journal of Management and Business Research: A Administration and Management,Vol 15, Issue 10.

[4] van der Aalst W.M.P. (2014) Data Scientist: The Engineer of the Future. In: Mertins K., Bénaben F., Poler R., Bourrières JP. (eds) Enterprise Interoperability VI. Proceedings of the I-ESA Conferences, vol 7. Springer, Cham

[5] K.Indira Gandhi, Sri.C.Sreedhar, “Survey on Big Data: Management and Challenges,” 2015, https://ijcttjournal.org/2015/Volume20/number-1/IJCTT-V20P106.pdf