IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 10, October 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910019 |

The Necessity for Data Science and the Roles of Data Science Teams: An Analysis of Various Cases

Sheikh Md Zubair Md Zahoor

Former Research Scholar, OPJS University, Churu, Rajasthan, India

ABSTRACT: The paper examines the advantages of well-known positions before discussing the lack of defined roles in the data science community, which could be due to the field's novelty.
The study delves into five case studies, each of which examines a different attempt to construct a uniform set of responsibilities.
The outcomes of these case studies are then used in the study to discuss the utility of online job postings for data science positions. While some roles, such as data scientist and software engineer, were frequently used in all five case studies, no role was used on a regular basis. However, the study continues by emphasizing the need to create a data science workforce framework that students, businesses, and academic institutions can use.
This paradigm would help companies to better match their data science teams to the capabilities they need.

KEYWORDS: Data Science, BigData, Data Science Roles and Responsibilities, Project Management.

I. INTRODUCTION

Big data indicates a noteworthy move in utilizing information-intensive computations methods and technologies. The transition to parallelism has complicated big data techniques and necessitated a variety of specialized competences. As a result, the term "information science" has become ubiquitous, pertaining to any incident that has an effect on data. This is exacerbated by the fact that the data science professional transforms from an employee to a data science team. In this novel situation, we lack a vocabulary to set apart the roles and abilities required for functioning data science team or big data team. Many issues arise as a result of a lack of lexicon (For example, identifying the suitable candidate for a certain position in a data science team). This paper provides a solution by citing examples and specifications of the data science workforce.

In Section [2] author provided some context and purpose for creating appropriate categories and defining the required skills and capabilities. A variety of case studies on the use of various job titles are shown in Section [3]. Section [4] compares and contrasts the case studies with our research on online job postings that employ those positions. The future developments that will affect the data science workforce are discussed in Section [5]. Lastly, Section [6] sums up conclusions and provides future scope.

II. RELATED WORK

The mid-2000s saw the emergence of new systems for storing massive data files (Hadoop Distributed File System [HDFS]), physically distributed logical datasets (Hadoop), parallel processing of distributed data (MapReduce), and the paradigm change known as big data.

Horizontal scaling is used in the Hadoop ecosystem to distribute data across independent nodes, with scalability coming from the addition of nodes. As a result, parallelism is increasingly being used in the development of scalable, data-intensive applications. While various conceptual definitions have been ascribed to it, Big data is not "bigger data" than what conventional techniques can manage; nevertheless, it is data which necessitates parallel processing to fulfill time limits for end-to-end analytical performance at a reasonable price.

CHARACTERISTICS OF WORKFORCE

The major objective for job adverts is to have a standard vocabulary to classify and explain the kinds of data science work which has to be performed in order to identify, recruit, train, develop, and manage an appropriately qualified workforce. While data science has its own functions and lexicon, the necessity for workforce explanations is not exclusive to data science.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910019 |

To put it another way, data science isn't the only subject that requires to define roles and responsibilities. Another area in which there is a necessity for this is Cyber-security. The National Institute of Science and Technology (NIST) in the United States established the National Initiative for Cyber-security Education (NICE) Cyber-security Workforce Framework for data protection, which describes the categories, specialized areas, and work positions for cyber-security professionals. They also gave descriptions of tasks, knowledge, skills, and abilities, as well as a mapping of these to job responsibilities.

The US Department of Defence Cyber Workforce Framework [5] is another effort to establish workforce definitions. This development is still in progress, and it includes revisions to a related document, a role-based paradigm for federal Cyber-security or information technology training. The following are some of the benefits highlighted in the NICE report that uniformly apply well to the realm of data science:

- Employers: should keep track of their employees' skills, training, and qualifications; enhance job descriptions; create career paths; and monitor efficiency.
- Educators: design curriculum and organize training for specific roles through programs, courses, and seminars.
- Technology Providers: determine the work opportunities, responsibilities, and knowledge, skills, and abilities related to their products.

SKILLS AND ROLES

While there are certain skills which are similar among all types of data science roles, certain skills might be exclusive to the particular role. It will be critical to ensure that each function of data science work is defined in the same manner that the NICE workforce framework contains knowledge, skills, and abilities that could be applicable to various work roles. Many generalist professionals may perform a range of jobs, but to maintain consistency, job descriptions should not coincide.

DIFFICULTIES BECAUSE OF INADEQUATE PROCESS MODEL

There are a variety of challenges in developing a framework for data science practitioners, but arguably the most important is that there is no agreed-upon data science process template: In the late 1990s, the model (CRISP-DM) Cross-Industry Standard Process for Data Mining Framework was designed. This framework is still used by the majority of professionals, but it precedes the Internet, big data, machine learning, agile, the Internet of things (IoT) and other technologies, and it did not address management processes or system development.

LIFECYCLES SOFTWARE DEVELOPMENT

With the help of reporting or business intelligence, the majority of analytical system development leads to situational awareness. The software development lifecycles (SDLCs) for requirement-driven analytics systems are designed for this kind of development. Advanced analytics systems, on the other hand, are results-oriented and require creativity when it comes to picking data and software aspects, developing models, and validating and enhancing models. Many of the skillets used in data science would coincide with which are used in SDLCs, but to be completed, tasks would need to be explicitly defined. However, it is important to integrate the data science and SDLC models, specifically with fast and advance standard methodologies.

III. METHODOLOGY

As a result, giving job designations and job requirements that further accurately define tasks, knowledge, skills, and capabilities could serve the data science industry while also removing the phrase "data scientist" from overusing. To ensure that an accurate view of ongoing ideas and application within the area of data science can be captured, a cross section of organizations is being reviewed. Two regulatory agencies, two business bodies, and one advisory / consulting agency were specifically chosen for our case studies. The assessment of the designated responsibilities was relied upon written documents from each company to every case study, and interactions with representatives from the specified companies have also been conducted wherein paperwork wasn't quite as robust.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910019 |

CASE STUDIES

Different situations in data science technology will be explored in this section.

i) NIST

The NIST Big Data Public Working Group (NBD-PWG) intends to facilitate development in big data by establishing clarity on important, core concepts. The results were published in volume series in the NIST Big Data Interconnect System. The establishment of a big data reference architecture (RA) that classify big data systems parts is one of the key activities of NBD- PWG.

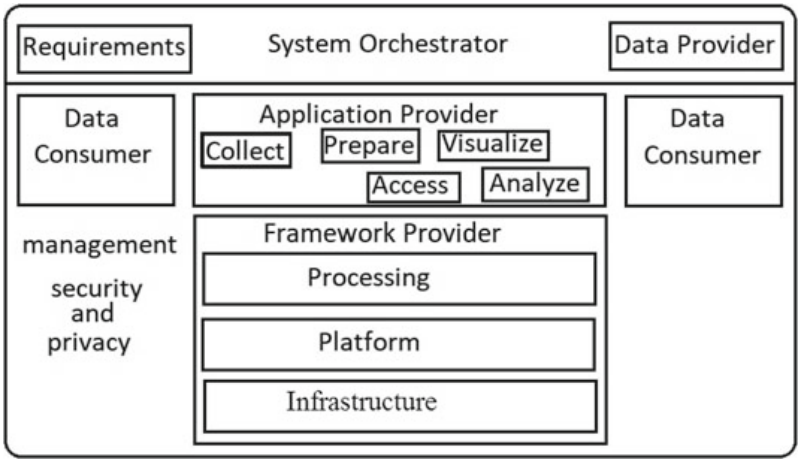

Figure 1: NIST Big Data Reference Architecture

It also specifies at a greater extent the responsibilities of individuals whose tasks are covered in that section. The RA is made up of five parts, each of which has a specific role to perform, as displayed in Figure. 1 and also described below:

- System Orchestrator - analyzes and integrates the essential data application tasks into a vertically functional system.
- Data Provider - enters the big data system with new data or information feeds.
- Big Data Application Provider - implements a life cycle to ensure that security and privacy concerns, and also framework orchestrator-defined prerequisites, are met.
- Big Data Framework Provider — creates a computer environment in which certain transformation applications can be run while maintaining data privacy and integrity.
- Data Consumer - end-users or other systems that employ the big data application provider's results are referred to as data consumers.
- Security and Privacy - for policies, requirements, and audits, interacts with the system orchestrator, as well as the big data app and platform vendors for development, deployment, and operation.
- Management - Big data system administration should include both system and data-related components of the big data environment, i.e., system and big data lifecycle administration. Deployment, customization, configuration management, software management, storage management, capability management, resources management, and efficiency management are all part of system management. The operations encompassing the data lifecycle of collecting, compilation, analysis, presentation, and accessibility are referred to as big data lifecycle management.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910019 |

ii) EDISON

The EDISON project [12] is an EU-funded initiative to "boost the quantity of skilled and professional data scientists throughout Europe and even beyond." The EDISON Data Science Framework (EDSF), that includes the proficient Framework, Data Science Professional (DSP) Profiles, and the mock-up core curriculum, is the centre of the data gathering.

Data science infrastructural administrators, data science experts, data science technology professionals, and data and resource input and accessibility are the four key occupational groups outlined by the DSP Profiles. Particular tasks inside every occupational group are described in any of these profiles. The data science professional, who plays the following roles, is of distinct significance.

- Data Scientist - identify and comprehend abundant sources of data, handle enormous volumes of data, combine sources of data, ensure dataset consistency, and generate visuals to aid in data comprehension. Create mathematical models, exhibit and interpret data insights and conclusions to experts and researchers, and suggest applications for the data.
- Data Science Researcher - This position uses scientific discovery research/processes, such as hypothesis and hypothesis testing, to get meaningful insights about a scientific issue, a business procedure, or to uncover underlying relationships among various processes.
- Data Science Architect - It creates and manages data science application and facility framework. It generates suitable data models and performs workflow processing.
- Data Science Programmer - This position creates, builds, and codes huge data (science) analytics solutions to serve scientific or industrial processes.
- Data / Business Analyst - mines data related to system, services, or organizational efficiency from a broad range of data and presents it in a useful format.

iii) SPRINGBOARD

Data engineer, data scientist, and data analyst are the three roles described by Springboard, an online data science educational platform. As shown in Figure 2, each of these roles necessitates software engineering, math/statistics, and data communication competencies. Based on springboard's descriptions, we've listed several potential roles below:

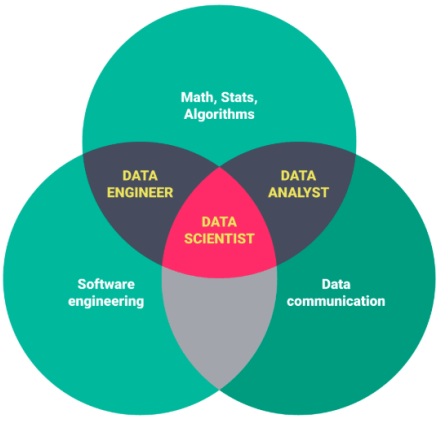

Figure 2: Springboard Overview in Software Engineering

- Data Engineer - To manage enormous amounts of data at volume, it primarily depends on his or her software engineering expertise: normally concentrates on coding, wiping up data, then executing requests by data scientists, and is familiar with a wide range of programming languages, from Java to Python. Whenever

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910019 |

someone takes a data scientist's predictive model and turns it into code, they are often acting as a data engineer.

- Data Scientist - It fills the space among programming and deployment of data science, data science theory, and data business concerns. It can convert a business problem into a data question, develop predictive models to address the issue, and then deliver a narrative regarding the outcomes.
- Data Analyst - It analyzes the data and prepares reports and visuals to convey what information the data conceals. The data analyst responsibility is achieved when someone helps others from all around the firm comprehend particular queries using visuals. A data architect, who specializes on organizing the technology which handles data models, but this is not presented in the picture rather than noted by springboard.

iv) SAIC

SAIC is a systems administrator that mainly serves the federal govt, including clients varying from civilians to military and Intel. SAIC created an intrinsic process model referred as Data Science Edge TM [8] to enhance productivity through development and deployment of BDA systems, as demonstrated in Figure 3; the model expanded the previous inadequate data mining process of CRISP-DM to encompass big data, systems development, and data-driven decision-making factors, as well as integration with flexible process models.

It is a very comprehensive process model corresponds to SAIC's generic data science roles. SAIC offers distinctive role and responsibility for data science, big data platforms, and data management in addition to the usual roles for software and systems development.

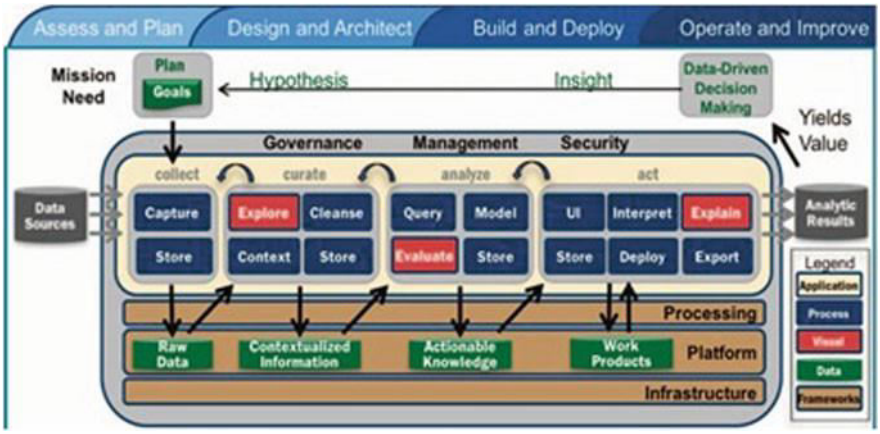

Figure 3: SAIC's Data Science Edge

Information Architect: It builds shared information environments based on information models or concepts. It designs data models for efficient database performance. It develops data standards and transforms data to regulated vocabularies, as well as designing data structures for data exchange.

Data Scientist: It interacts with data in cross-functional teams at all phases of the analysis lifecycle to produce invaluable insights, and it uses a scientific methodology to produce insights from data, with outcomes verified at every phase.

Metrics and Data: It creates, examines mines, organizes, and analyzes data in order to enhance performance, make better decisions, and achieve a competitive edge. It controls all parts of final data processing and performs statistical data analysis and information to support precise predictive forecasting or classification.

Knowledge and Collaboration Engineer: It creates and deploys tools and technology to help companies manage their information and collaborate more effectively.

Big Data Engineer: It creates parallel information-intensive systems employing big data technology as well as works with Hadoop's entire open-source stack, from cluster administration to data warehousing and scheduling analytics technology.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910019 |

SAIC also provides managerial roles for these positions. While DSE distinguishes three forms of visualization, there are no distinct roles for visualization or business intelligence professionals, who are typically discovered within software engineers or data scientists. The relevance of identifying the origin of data in a specific domain is recognized by SAIC. BDA systems are created by teams that bring together the whole necessary talents to develop an agile BDA system.

	Data science researcher	Data scientist	Data architect	Data analyst	Data science programmer	Data engineer
NIST	No	No	No	No	No	No
EDISON	Yes	Yes	Yes	Yes	Yes	No
SAIC	No	Yes	No	No	No	Yes
SPRINGBOARD	No	Yes	Yes	Yes	No	Yes

Table 1 Comparative analysis of all the cases mentioned in the paper

IV. DISCUSSION

Table 1 shows which positions were utilized in several case studies. The positions employed are not used on a regular basis, which we clearly observe. Information architects and software engineers were among the positions mentioned, and so were data suppliers and source system specialists, both of which are atypical.

V. CONCLUSION

Giving high accuracy and coherence for the term data scientist will assist in the skills required to train and recruit professionals. A collaboration of partners from government, enterprise, and research would be one method to get a complete summary of the data science personnel. Before attempting to establish the skills and job positions needed to data scientists, the consortium's first goal is to build a comprehensive data science process model which represents all participants' agreement on what constitutes a data science engagement. Employing the NICE model, this research was able to give categories, specializations, job roles, and assignments in order to define the distinctions in roles. Yet another potential roadblock would be that traditional analytics systems requiring simple summary statistics, analysis, or business intelligence are almost entirely built by computer and process engineers, with traditional roles of data modelers, database analysts, and database administrators thrown in the mix. Considering the coevolution of big data and data science, we observe that present usage may not reflect how the market is developing, therefore a future (maybe every six months) evaluation of role utilization in the field to establish trends over time may be a suitable further move.

REFERENCES

1. Saltz JS, Grady NW (2017) The ambiguity of data science team roles and the need for a data science workforce framework. In: IEEE international conference on big data, pp 2355–2361.
2. Framework (2015) DRAFT NIST big data interoperability framework, volume 7, standards roadmap. NIST special publication 1500-7.
3. Saltz JS (2015) The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In: IEEE international conference on big data, pp 2066–2071.
4. Shearer C (2000) The CRISP-DM: the new blueprint for data mining. J Data Wareh 5(4).
5. Saltz J, Shamshurin I, Connors C (2017) Predicting data science sociotechnical execution challenges by categorizing data science projects. J Assoc Inf Sci Technol 68(12):2720–2728.
6. Saltz J, Crowston K (2017) Comparing data science project management methodologies via a controlled experiment. In: Proceedings of the 50th Hawaii international conference on system sciences.
7. Toth P, Klein P (2013) A role-based model for federal information technology/cyber security training. NIST special publication 800-16:1–152.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910019 |

8. Newhouse W, Keith S, Scribner B, Witte G (2017) National initiative for cybersecurity education (NICE) cybersecurity workforce framework. NIST special publication 800:181.
9. Newhouse WD (2017) Nice cybersecurity workforce framework: national initiative for cybersecurity education. No. special publication (NIST SP)-800-181.
10. Chang WL, Grady N (2015) NIST big data interoperability framework: volume 1, big data definitions. No. special publication (NIST SP)-1500-1.
11. Saltz JS, Shamshurin I (2016) Big data team process methodologies: a literature review and the identification of key factors for a project's success. In: 2016 IEEE international conference on Big Data, pp 2872–2879.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details