STM JOURNALS

Journal of
Advanced Database
Management & Systems

ISSN: 2393-8730
Volume 11, Issue 3, 2024
September—December
DOI (Journal): 10.37591/JoADMS

https://journals.stmjournals.com/joadms

Review JoADMS

Leveraging Full Stack Data Science for Healthcare Transformation: An Exploration of the Microsoft Intelligent Data Platform

Prashant D. Sawant*

Abstract

The rapid progress of the Fourth Industrial Revolution has been largely driven by the evolution of artificial intelligence (AI), with notable contributions from technologies such as Generative Pre-trained Transformers (GPT). This revolution has seen the convergence of physical, digital, and biological technologies, leading to transformative impacts across various sectors. Data science, serving as a crucial enabler, has enabled the development of intelligent value chains. However, the application of data science in this domain often remains limited to isolated elements of the data pipeline. The concept of Full Stack Data Science (FSDS) addresses this limitation by providing a comprehensive approach to managing all aspects of a data science project. This paper explores the challenges faced by Full Stack Data Scientists, such as data integration, quality, scalability, security, privacy, tool complexity, stakeholder communication, and regulatory compliance. It further discusses how the Microsoft Intelligent Data Platform (MIDP) offers a suite of tools and services that support FSDS, particularly in the healthcare industry. The roles of Full Stack Data Scientists in healthcare, including data integration, analysis, visualization, model deployment, and governance, are examined. The paper also highlights the significance of data processing frameworks and machine learning services provided by Azure, emphasizing their utility in the data science workflow.

Keywords: Artificial intelligence, generative pre-trained transformers, full stack data science, microsoft intelligent data platform, healthcare

INTRODUCTION

The Fourth Industrial Revolution, marked by artificial intelligence (AI), particularly Generative Pre-trained Transformers (GPT), which facilitate the convergence of physical, digital, and biological technologies, has ushered in significant transformations across various sectors [1–3]. A key catalyst of this revolution is data science, which has played a pivotal role in creating intelligent value chains [4]. However, the application of data science in the value chain is often confined to specific elements of the data pipeline [5, 6]. This is where the concept of Full Stack Data Science (FSDS) [5, 6] emerges, offering a holistic approach to managing all facets of a data science project [7–9].

***Author for Correspondence**
Prashant D. Sawant
E-mail: prasdsaw@gmail.com

AI Research Director, Ai-Discovery Consultancy, Melbourne, Australia

Received Date: April 27, 2024
Accepted Date: June 03, 2024
Published Date: August 31, 2024

Citation: Prashant D. Sawant. Leveraging Full Stack Data Science for Healthcare Transformation: An Exploration of the Microsoft Intelligent Data Platform. Journal of Advanced Database Management & Systems. 2024; 11(3): 1–8p.

Figure 1 depicts the components of FDSD, such as fundamentals of math, Python, data structure and SQL, machine learning (ML) (deep learning, neural network, and ML algorithms), data analytics (hypothesis testing, EDS, and visualization), data engineering (storage and processing), Docker, Kubernetes, and web development, and finally, the deployment of the software or applications. Full stack scientists should have knowledge and skills in

all these components to perform their duties. Full Stack Data Scientists encounter several challenges during the implementation of a data platform, which are discussed below.

- *Data integration:* A frequent challenge is the integration of data from various sources. Each data source may use a different format or structure, and transforming all these data into a unified format can be a complex task.
- *Data quality:* Another prevalent problem is the low quality of data. This can include missing, inconsistent, or incorrect data. Data scientists often need to spend a significant amount of time cleaning and preprocessing data before they can be used.
- *Scalability:* As data volume expands, ensuring scalability becomes crucial. The platform must effectively manage larger datasets while maintaining performance standards.
- *Security and privacy:* Safeguarding data security and preserving user privacy are paramount, particularly when handling sensitive information. This involves establishing robust access controls, encryption measures, and techniques for anonymizing data.
- *Complexity of tools and technologies:* The landscape of data science tools and technologies is vast and constantly evolving. Keeping up with the latest tools, learning how to use them, and deciding which are the best fit for a particular task can be challenging.
- *Communication with stakeholders:* Data scientists must frequently convey intricate findings to individuals without technical expertise, necessitating the skill of articulating results in a comprehensible manner.
- *Regulatory compliance:* Particularly in industries such as healthcare and finance, data scientists must ensure that their work complies with all relevant regulations. This may encompass rules concerning data security and privacy, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), along with regulations specific to particular industries.

These challenges highlight the breadth and depth of the skills required to be a successful Full Stack Data Scientist. It is not just about technical skills but also about understanding the business context, communicating effectively, and navigating the regulatory environment.

The Microsoft Intelligent Data Platform (MIDP) offers a comprehensive suite of tools and services [10, 11] that can greatly assist Full Stack Data Science (FSDS) in their work.

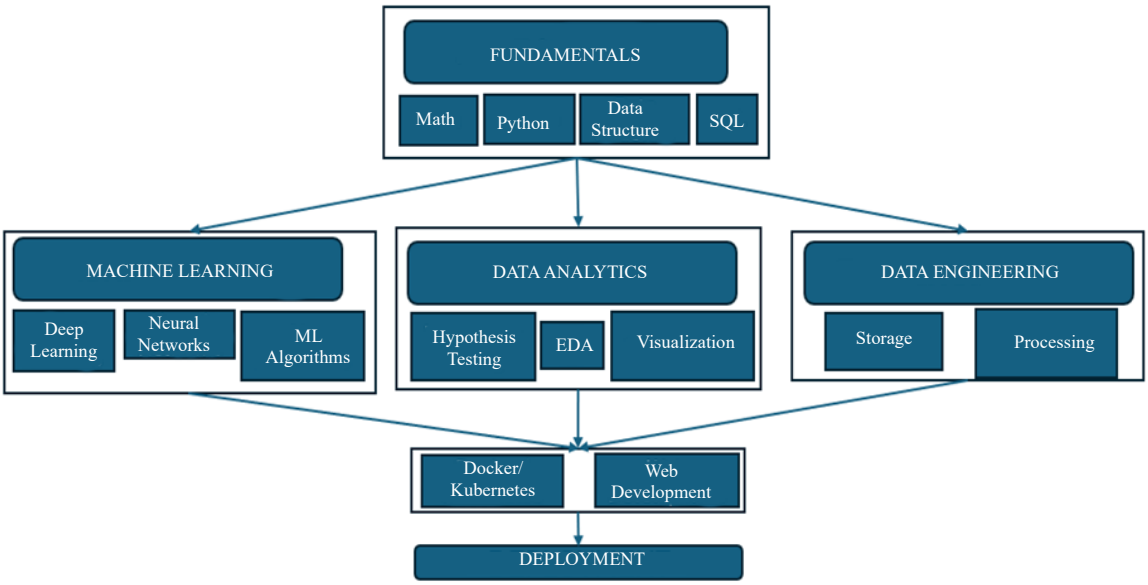

Figure 1. Components of Full Stack Data Science.

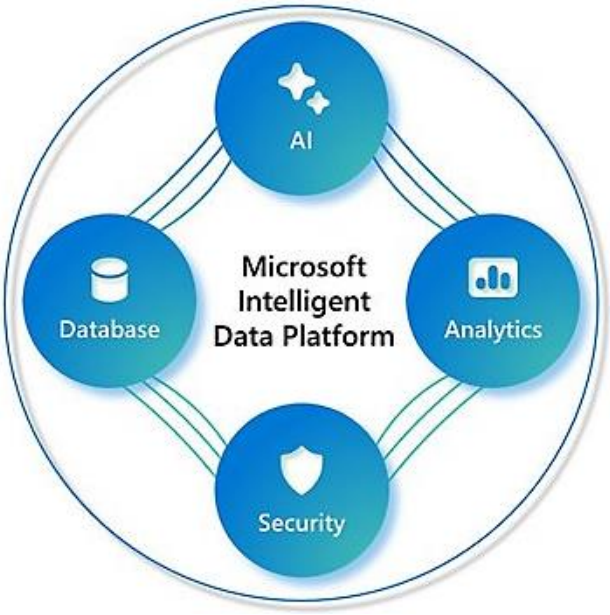

Figure 2. Microsoft Intelligent Data Platform [11].

The MIDP provides AI, databases, analytics, and security, as depicted in Figure 2. Some of the important components and services of MIDP that are useful for FSDS are as follows.

- *Unified data and analytics:* MIDP integrates databases, analytics, and governance into a seamless data platform [12]. This enables data scientists to concentrate on generating value rather than on dealing with the integration and management of their data estate [10, 11].
- *Azure Synapse link for SQL:* This service offers a smooth data pipeline to Azure Synapse Analytics and facilitates nearly real-time analytics for SQL Server 2022 and the Azure SQL Database [10]. This capability is now available for over 10 million Azure SQL databases worldwide [10].
- *SQL Server 2022:* SQL Server 2022 represents the most Azure-integrated release to date, featuring ongoing advancements in performance, security, and availability [10]. Through seamless disaster recovery to Azure SQL Managed Instance, SQL Server 2022 ensures genuine resilience by connecting the SQL Server to Azure [10].
- *Azure Arc-enabled SQL managed instance:* Azure Arc introduces its latest innovation with the general availability of the business-critical tier of the Azure Arc-enabled SQL Managed instance. This offering enables customers to perform their most challenging mission-critical tasks in hybrid and multi-cloud setups [10].
- *Azure AI:* A key part of the Intelligent Data Platform, Azure AI, makes building data-informed ideas and experiences easy [12].

This article examines the function of the Microsoft Intelligent Data Platform (MIDP) in conjunction with a Flexible Service Delivery System (FSDS) within the healthcare sector.

UNIFIED DATA AND ANALYTICS

MIDP integrates databases, analytics, and governance into a seamless data platform [12]. This enables data scientists to dedicate more time to generating value rather than handling the integration and management of their data assets [12].

The MIDP, a comprehensive data management system that amalgamates databases, analytics, and governance, is a singular, cohesive data platform. This integration facilitates a streamlined workflow for data scientists, enabling them to focus their efforts on extracting value from data rather than expending time on the integration and management of their data estates. The application of the MIDP and FSDS in the healthcare sector can be explained as follows.

Hospitals produce large volumes of data on a daily basis ranging from patient records to diagnostic outcomes. The MIDP can integrate these disparate data sources into a unified platform, allowing data scientists to easily access and analyze these data. This may result in enhanced diagnostic precision, better patient treatment, and increased operational efficiency within the hospitals. A Full Stack Data Scientist plays several key roles in the context of implementing MIDP in healthcare as follows.

Data Integration

Full Stack Data Scientists are responsible for integrating various healthcare data sources into a unified system [13]. This could involve working with databases, APIs, and other data sources and ensuring that the data are correctly ingested into the platform [10]. For instance, in a hospital setting, this could involve integrating data from electronic health records, diagnostic results, and patient demographics into a unified perspective [10].

Data Analysis

Once the data are integrated, Full Stack Data Scientists analyze it to derive insights [12]. This can involve statistical analysis, machine learning, predictive modeling, and other data science techniques [13]. For instance, they can utilize machine learning algorithms to forecast disease outbreaks or detect patients vulnerable to specific conditions [9].

Data Visualization

Full Stack Data Scientists create visualizations to present findings from the data [13]. This could involve using tools such as Microsoft Power BI to create dashboards and reports that make the data understandable to nontechnical stakeholders [13].

Model Deployment

If the analysis involves building predictive models, Full Stack Data Scientists are also responsible for deploying these models into production [13]. This could involve writing code, working with APIs, and integrating models into the existing IT infrastructure [13].

Data Governance

Full Stack Data Scientists are also involved in data governance, ensuring the ethical use and regulatory compliance of the data [14, 15]. This could involve implementing data privacy measures, such as data masking for sensitive fields [16–18], as well as ensuring compliance with regulations like GDPR and HIPAA [19].

Collaboration

A Full Stack Data Scientist works closely with other teams, including IT, business intelligence, and healthcare professionals, to ensure that the data platform meets the needs of the organization [19].

These roles highlight the breadth and depth of the skills required to be a successful Full Stack Data Scientist in a healthcare setting. They not only need technical skills but also an understanding of the business context, the ability to communicate effectively, and the capability to navigate the regulatory environment [20, 21]. Thus, a Full Stack Data Scientist plays a crucial role in every stage of the data lifecycle in a healthcare setting, from data integration and analysis to visualization, model deployment, and governance. Their work enables healthcare organizations to leverage their data to improve patient care, enhance operational efficiency, and make data-driven decisions.

DATA PROCESSING FRAMEWORKS

Full Stack Data Scientists may find Microsoft Azure's support for data processing frameworks such as Pandas, PySpark, and Dask [22] as these frameworks provide robust tools for the efficient manipulation, transformation, and aggregation of large datasets.

Pandas

This is a Python library that provides flexible data structures, making it easier to handle relational or labeled data. This proves to be particularly beneficial in handling numerical tables and time-series data.

PySpark

The Python library for Apache Spark, PySpark, is an open-source distributed computing system employed for large-scale data processing and analytics. It allows Full Stack Data Scientists to process large volumes of data across a distributed network, which is crucial when dealing with big data scenarios.

Dask

Another adaptable parallel computing library designed for analytics seamlessly integrates with the current Python ecosystem and manages sizable datasets that exceed memory capacity.

These frameworks are instrumental in helping Full Stack Data Scientists work with complex data, enabling them to extract meaningful insights from large, diverse datasets. They are particularly useful in scenarios where data scientists must process large volumes of data quickly and efficiently, making them an integral part of the data science toolkit on Microsoft Azure.

AZURE MACHINE LEARNING SERVICES

Azure Machine Learning, a key component of Microsoft data platforms, provides a comprehensive environment that enables Full Stack Data Scientists to build, train, and deploy machine learning models efficiently [23]. This service aims to simplify the machine learning process, enabling data scientists to concentrate on their data science tasks rather than on managing the underlying infrastructure. This offers a no-code or low-code approach for model development, which is particularly beneficial for rapid prototyping. However, this is not only about building models. Azure Machine Learning also provides robust tools for data preparation, which is a critical step in any data science project. It offers support for multiple data formats and includes functionalities for data cleaning, transformation, and feature engineering. Moreover, Azure Machine Learning is not limited to model building and data preparation. They also provide services for model deployment, management, and monitoring. This means that data scientists can deploy their models as web services and monitor their performance in real time. Thus, Azure Machine Learning is a powerful tool for Full Stack Data Scientists, offering a wide range of capabilities that cover the entire data science workflow, from data preparation and model building to deployment and monitoring.

DATA VISUALIZATION

As a Full Stack Data Scientist, one of the key steps in the data science pipeline is data visualization, and tools such as Microsoft Power BI are instrumental in this process [24]. Power BI comprises a set of analytical tools tailored for businesses, enabling data scientists to analyze data comprehensively and efficiently to convey insights. It facilitates seamless connection to numerous data sources, simplifies data preparation, and supports spontaneous analyses. Additionally, Power BI empowers data scientists to generate visually appealing reports, share them across the organization via the web and mobile devices, and create interactive visualizations independently without relying on IT personnel or database administrators. Furthermore, Power BI seamlessly integrates with existing applications. This includes embedding analytics in the app or website, automating Power BI reports and dashboards, and developing custom visuals for unique presentations. Thus, Microsoft Power BI is a powerful tool for Full Stack Data Scientists, offering a wide range of capabilities that cover the entire data visualization process, from data preparation and exploration to report generation and presentation.

SCALABILITY AND FLEXIBILITY

Microsoft data platforms offer significant scalability and flexibility, which are essential characteristics for Full Stack Data Scientists. Scalability ensures that the system can accommodate increasing data volumes and complex computations effectively. For instance, Azure services such as

the Azure SQL Database and Azure Cosmos DB offer virtually limitless cloud scalability. This means that whether data scientists are dealing with gigabytes or petabytes of data, the system can be scaled up to meet the needs. Conversely, flexibility denotes the capability to adjust to evolving needs or circumstances. Microsoft data platforms support a wide range of data types, structures, and sources, from structured SQL data to unstructured big data. They also support various programming languages and frameworks, allowing data scientists to use tools with which they are most comfortable. Microsoft Azure provides options for both Infrastructure as a Service alias IaaS and Platform as a Service alias PaaS, offering flexibility in how data scientists deploy and manage their applications. In real-world scenarios, a Full Stack Data Scientist might need to analyze streaming data from IoT devices one day, and high-dimensional genomic data the next. The scalability and flexibility of Microsoft data platforms make it possible to handle diverse data science tasks effectively. Thus, the scalability and flexibility of Microsoft data platforms are key enablers for Full Stack Data Scientists, allowing them to handle large volumes of data, complex computations, and diverse data science tasks with ease.

SECURITY AND COMPLIANCE

A Full Stack Data Scientist plays a vital role in ensuring data security and compliance through various means.

- *Understanding regulations:* They need to understand the data protection and privacy regulations relevant to their work, such as the GDPR or HIPAA. This entails understanding the types of data deemed sensitive, proper handling procedures, and the necessary consent required for their usage.
- *Implementing security measures:* While a data platform may provide security features, it is often up to the data scientist to implement them correctly. This can include setting up appropriate access controls, encrypting sensitive data, and anonymizing the data for use in analytics.
- *Data governance:* Full Stack Data Scientists often play a role in data governance, which includes defining who has access to what data, how data quality is ensured, and how data is classified and cataloged. Good data governance is the key to both security and compliance.
- *Monitoring and auditing:* They may also be involved in monitoring data usage to detect any unusual or suspicious activity. In addition, they may need to prepare for audits by documenting how data are used and ensuring that all data processing activities comply with relevant regulations.
- *Incident response:* During a security breach, Full Stack Data Scientists might participate in the incident response process. This could include identifying what data were affected, assessing the impact, and helping to prevent future incidents.
- *Training and awareness:* Full Stack Data Scientists can also play a role in training other staff members on data security and compliance issues. This guarantees that all members of the organization grasp the significance of safeguarding data and are proficient in handling data with care.

While a Full Stack Data Scientist ensures data security and compliance, its role can vary depending on the organization and the individual's specific role. From a FSDS perspective, the robust security features and compliance capabilities of Microsoft data platforms are of paramount importance. These are discussed below.

Security

Microsoft's data platforms have built-in security measures at multiple levels. At the data level, features such as encryption at rest and in transit, data masking, and access control help to protect sensitive data. At the network level, features such as firewalls and virtual networks help to protect against external threats. In addition, sophisticated threat detection capabilities assist in swiftly identifying and addressing potential security incidents. For example, the Azure SQL Database provides advanced security features such as Always Encrypted, which protects sensitive data both at rest and in transit, and Azure Active Directory Integration, which provides identity and access management capabilities.

Compliance

Microsoft's data platforms are structured to aid organizations in adhering to diverse data protection and privacy regulations. This includes regulations such as GDPR and many others. Microsoft provides comprehensive documentation and guidance to help organizations understand how to use their services in a compliant manner.

For instance, Azure Purview, a unified data governance service, can help organizations maintain a comprehensive data catalog, classify data, and track data lineages, which are important aspects of regulatory compliance.

In summary, the security and compliance features of Microsoft data platforms not only protect sensitive data but also simplify the task of regulatory compliance for Full Stack Data Scientists, allowing them to focus on extracting value from the data.

CONCLUSION

The integration of FSDS within the healthcare industry represents a paradigm shift in the management and utilization of data. Microsoft Intelligent Data Platform (MIDP) has emerged as a comprehensive solution that addresses the multifaceted challenges encountered by data scientists. This paper has delineated the roles and responsibilities of Full Stack Data Scientists in ensuring data security, compliance, and effective data governance within healthcare settings. It underscored the importance of understanding regulations, implementing security measures, and fostering collaboration among various teams. The capabilities of MIDP in providing unified data and analytics, scalable data processing frameworks, and robust machine learning services have been demonstrated. These features enable Full Stack Data Scientists to focus on extracting value from data, thereby enhancing patient care, operational efficiency, and data-driven decision-making in healthcare. The study concludes that the synergy between FSDS and MIDP holds the potential to revolutionize healthcare data management, paving the way for innovative solutions and improving patient outcomes.

REFERENCES

1. Sawant PD. Artificial Intelligence: The era of new Industrial Revolution [eBook]. Kindle Store. Amazon.com.au; 2022.
2. Sawant PD. A real-time visualization framework to enhance prompt accuracy and result outcomes based on the number of tokens. J Artif Intell Res Adv. 2024;11:45-53.
3. Schwab K, World Economic Forum (2016). The Fourth Industrial Revolution: what it means and how to respond. [online] World Economic Forum. Available from: https://www.weforum.org/stories/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.
4. ODSC Community. (2020). Data science in manufacturing: An overview. [online]. Open Data Sci. Available from: https://opendatascience.com/data-science-in-manufacturing-an-overview/.
5. The Click Reader. (2021). Full-stack data science: A complete overview. [online]. Medium. Available from: https://medium.com/@theclickreader/full-stack-data-science-a-complete-overview-9d66df1ac771.
6. Perfect eLearning. (2023). Building end-to-end data solutions: Full stack data science. [online]. Available from: https://perfectelearning.com/blog/building-end-to-end-data-solutions-full-stack-data-science.
7. StrataScratch. (2020). How to become a full stack data scientist. [online]. Available from: https://www.stratascratch.com/blog/how-to-become-a-full-stack-data-scientist/n.
8. Shopify. (2022). What is a full Stack data scientist? [online]. Available from: https://shopify.engineering/what-is-a-full-stack-data-scientist.
9. Institute of Data. (2024). What is full stack data science? [online]. Available from: https://www.institutedata.com/us/blog/what-is-full-stack-data-science/.
10. Kumar R. (2022). Introducing the Microsoft intelligent data platform. Microsoft Azure Blog. [online]. Microsoft Website. Available from: https://azure.microsoft.com/en-us/blog/introducing-the-microsoft-intelligent-data-platform/.

11. Microsoft intelligent data platform. (2023). Microsoft Cloud. [online]. Microsoft.com. Available from: https://www.microsoft.com/en-us/microsoft-cloud/solutions/intelligent-data-platform.
12. Hawk J. (2024). The Microsoft Intelligent Data Platform—Unleash your data and accelerate your transformation. Microsoft Azure Blog. Microsoft Website. [online]. Available from: https://azure.microsoft.com/en-us/blog/the-microsoft-intelligent-data-platform-unleash-your-data-and-accelerate-your-transformation/.
13. Thomas S. (2022). Microsoft Cloud for Healthcare: Transforming patient care through data and insights. Microsoft Industry Blogs. Microsoft Website. [online]. Available from: https://www.microsoft.com/en-us/industry/blog/healthcare/2022/04/05/microsoft-cloud-for-healthcare-transforming-patient-care-through-data-and-insights/.
14. Health. (2024). About health data in Australia. Australian Government Department of Health and Aged Care. [online]. Available from: https://www.health.gov.au/topics/health-data-and-medical-research/about-health-data.
15. OAIC. (2019). Guide to health privacy. [online]. oaic.gov.au. Available from: https://www.oaic.gov.au/__data/assets/pdf_file/0011/2090/guide-to-health-privacy.pdf.
16. Orion Health. (2023). Data masking vs de-identification: Understanding the differences and relevance in healthcare. [online]. Orion Health. Available from: https://orionhealth.com/us/blog/data-masking-vs-de-identification-in-healthcare/.
17. Kadam S. (2019). Data masking: Concept, tools, masking policies & healthcare data masking. Medium. Petabytz. [online]. Available from: https://medium.com/petabytz/data-masking-concept-tools-masking-polices-healthcare-data-masking-3c28ca03a30a.
18. Jain K. (2023). Data anonymization and data masking in healthcare. [online]. Medium. Available from: https://karen-j.medium.com/data-anonymization-and-data-masking-in-healthcare-db2650f04213.
19. HealthIT.gov. (2019). Privacy, Security, and HIPAA. [online]. Available from: https://www.healthit.gov/topic/privacy-security-and-hipaa.
20. Bhatia J. (2024). The role of data science in healthcare industry in 2024. [online]. PW Skills Blog. Available from: https://pwskills.com/blog/the-role-of-data-science-in-healthcare-industry-in-2024/.
21. SQLPad. (2024). Guide to becoming a full stack data scientist in 2024. [online]. Available from: https://sqlpad.io/tutorial/guide-to-becoming-a-full-stack-data-scientist-in-2024/.
22. Samuel Westby S. (2023). Full-Stack Data Science Roadmap 2023. [online] Samuel Westby. Available from: https://www.samwestby.com/2023/data-science-roadmap.
23. Sinha V. (2020). Building a data science tech stack: A comprehensive guide. Hevo Data. [online]. Available from: https://hevodata.com/learn/building-a-data-science-tech-stack/.
24. Microsoft. (2024). What is data science? Become a data scientist. Microsoft.com. [online]. Available from: https://azure.microsoft.com/en-ca/resources/cloud-computing-dictionary/what-is-data-science/.

