Chase Rawson

Dr. Sulpizio

DATA 610

April 25, 2023

As a data analyst in the financial industry, this feasibility proposal aims to explain how we should implement machine learning technology to handle the large amount of personal and sensitive data we acquired from our acquisition. The current technology used to hold the data we currently analyze and maintain will not be able to handle the volume of data held by our acquisition. This proposal will detail the steps and provide the knowledge needed to accomplish this goal.

As defined, “Artificial intelligence, or AI, is the ability of a computer or machine to mimic or imitate human intelligent behavior and perform human-like tasks that require human intelligence such as thinking, reasoning, learning from experience, and most importantly, making its own decisions.” (David, 2021). AI technology is commonly used in a way to mimic human behavior in a way to perform different tasks. AI is commonly found in industrial robots at a car manufacturing plant or as personal assistants such as Alexa and Cortana in Windows. Machine learning is vastly similar, a subset of AI that can “automatically learn from the data without explicitly being programmed or assisted by domain expertise.” (David, 2021). ML technologies are used in product recommendations on e-commerce websites, using historical data to recommend the right product for the consumer, or in email spam filters, using historical data to analyze and delete emails that are known to be fraudulent. Both of these technologies can be integrated into deployment and service models, yet these technologies are not used the same depending on the type of deployment model. The first deployment model that is used is on-premises, where the hardware and software components of the AI and ML system are managed on-site at our company’s headquarters. This model is appropriate to be used and managed by the IT team we employ, and, depending on how much sensitive data we have, it would be best for us to control instead of this being controlled by a third party. Since we propose to set up a cloud environment (the second deployment model is the cloud model), the AI and ML system is located and accessed through the cloud. The technology used to host this model is not located at our headquarters, relieving any worries about managing this technology. This model is appropriate to be used if we want a flexible, affordable, scalable way to manage the data. AI and ML systems can also exist in the three service models available. For the Software-as-a-Service (SaaS) model, the AI and ML system is accessed over the internet, usually after we pay a subscription fee. We are not responsible for maintaining the technology used. This would be the better choice if we want to spend minimal time setting up a model. For the Platform-as-a-Service (PaaS) model, the AI and ML system is accessed over the internet with much more flexibility to build models used to maintain the influx of data coming from our acquisition without being responsible for managing the infrastructure. Lastly, for the Infrastructure-as-a-Service (IaaS) model, the AI and ML system provides the most flexibility to build models and develop an infrastructure to manage the data, with the vendor not responsible for managing any of the technologies we use, best used to manage and process large amounts of data like in our situation. (What Is IaaS (Infrastructure as a Service)?, n.d.).

Deploying a cloud deployment model comes with many benefits, including scalability to address our growing company, affordability to eliminate the need to configure an expensive infrastructure, and accessibility to access data anywhere with an internet connection. Nevertheless, drawbacks are also common such as security risks where “most cloud providers stick to the so-called “shared responsibility model” when they are responsible for securing the cloud computing environment itself. At the same time, it is the client who needs to ensure the actual security in the cloud.” (Hazdun, 2023). Also, the cloud infrastructure may get compromised while under the service provider’s control. It is essential to determine whether our provider has proven they have proper security controls to protect our data. For an on-premises model, we, the company, have control over security, customization, the ability to control the performance, and the ability to add flexibility and allow offline availability to our data. However, we are investing more money in setting up an AI and ML infrastructure instead of paying the subscription fee and hiring qualified IT staff to maintain and scale our headquarters’ AI and ML system.

Data analytic models that integrate Artificial Intelligence and Machine Learning technologies provide more advanced analytics that provides insights into data than a typical data analyst would discover and present. It uses sophisticated quantitative methods to produce these insights to support predictive and perspective models (What Is Data and Analytics?, n.d.). They can handle large amounts of data, like in our situation, to make these informed decisions, which can be crucial to our customers as we are more aware of their priorities when consulting with any financial issue they may have. These technologies also provide a competitive advantage over our rivals, as AI and ML technology is relatively new and not used yet by many different companies across many industries. However, by adopting these technologies, we need to be aware of any privacy and security issues that may arise. However, most importantly, there has been an increased concern with AI and ML technologies regarding bias, discrimination, incomplete data, and other ethical concerns with the knowledge generated, as well as developing a set of rules that our analysts must follow to prevent misuse of this crucial technology. Henceforth, suppose we are questioned by our clients responsible for these decisions with their financial accounts and how they are made. In that case, we risk the lack of accountability for our actions because we are unaware of how AI and ML technology chose how they assisted with their money.

As proposed by the company, our goal is to set up a cloud environment to host, maintain, and analyze the large volume of information that has been collected. Since we are switching to implementing a cloud environment, organizational and technical issues will arise. First, do we trust our cloud vendor to protect our data from security risks, and do they have the ability to provide us with technology that determines if we comply with data security regulations? Whether the data is stored on the cloud or on-premises, they each experience the same number of threats equally. However, the cloud service provider also needs to prove that their software is free of any vulnerabilities at all times. (Morrow, 2018). Second, we should consider whether our IT team has the skills necessary to implement and maintain these cloud systems. Maintenance of the infrastructure is performed at the vendor’s end, but if we have any issues on our end, will our IT team be able to understand the problem, and are they capable of communicating with the vendor? Lastly, this change to our company will impact the organizational structure, with our employees possibly concerned depending on how they will use this technology; we need to successfully adopt this technology with minimal concerns and teach our employees how to use the cloud technology.

The cloud environment we plan to host and maintain the data relies on properly configured infrastructure to perform this task. This infrastructure must handle unstructured and structured data, but what is the difference? Structured data is highly organized and easily decrypted by AI and ML technologies, managed by the SQL programming language. Unstructured data is the opposite; it cannot be processed and analyzed by conventional data tools and is best managed by non-relational SQL databases. (IBM Cloud Education, 2021). Structured data is collected with a data format predefined. In contrast, unstructured data involves various techniques to collect text data (such as data mining) and audio, visual, and image data (such as microphones and cameras). With different methods used to collect structured and unstructured data, storing the data is also different, “structured data is stored in tabular formats (e.g., excel sheets or SQL databases) that require less storage space. It can be stored in data warehouses, which makes it highly scalable. Unstructured data, on the other hand, is stored as media files or NoSQL databases, which require more space. It can be stored in data lakes, which makes it difficult to scale.” (IBM Cloud Education, 2021).

Lastly, when managing our data and the data we are acquiring from our acquisition, it is vital to understand the three Vs of big data and its relationship to the cloud infrastructure:

1. The volume of data is how much we possess or the size of the data in gigabytes, terabytes, or larger. The cloud infrastructure we propose to migrate our data must handle large volumes of data on our vendor’s infrastructure.
2. The velocity of data refers to how quickly we can access our data from the cloud infrastructure. It is vital to have a proper internet connection to and from the vendor’s servers to process and manage real-time data.
3. The variety of data refers to the different data types as described above, structured and unstructured data.

Whether our potential cloud vendor can process and manage these different data types should be asked. Because with our acquisition, our data set size will increase, managing much more data (volume of data) than we have been managing. The infrastructure we rely on to manage our data in the cloud must not allow longer processing times and provide us with reduced efficiency (velocity of data) (Cai & Zhu, 2015). Also, the cloud infrastructure must allow us to analyze structured and unstructured data using advanced tools such as SQL (variety of data).

**Citations**

Cai, L., & Zhu, Y. (2015, May 22). *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*. Data Science Journal. Retrieved April 27, 2023, from <https://datascience.codata.org/articles/10.5334/dsj-2015-002>

David, D. (2021, July 12). *AI vs ML – What’s the Difference Between Artificial Intelligence and Machine Learning?* freeCodeCamp.org. Retrieved April 27, 2023, from <https://www.freecodecamp.org/news/ai-vs-ml-whats-the-difference/>

Hazdun, N. (2023, March 28). *Cloud Versus On Premises: Advantages And Disadvantages Of Both Models*. Forbes. Retrieved April 27, 2023, from <https://www.forbes.com/sites/forbestechcouncil/2023/03/27/cloud-versus-on-premises-advantages-and-disadvantages-of-both-models/?sh=42b6d75c5df6>

IBM Cloud Education. (2021, June 29). *Structured vs. Unstructured Data: What’s the Difference?* . IBM. Retrieved April 27, 2023, from <https://www.ibm.com/cloud/blog/structured-vs-unstructured-data>

Morrow, T. (2018, March 5). *12 Risks, Threats, & Vulnerabilities in Moving to the Cloud* . SEI Blog. Retrieved April 27, 2023, from <https://insights.sei.cmu.edu/blog/12-risks-threats-vulnerabilities-in-moving-to-the-cloud/>

*What Is Data and Analytics?* Gartner. (n.d.). Retrieved April 27, 2023, from <https://www.gartner.com/en/topics/data-and-analytics>

*What Is IaaS (Infrastructure as a Service)?* Amazon. (n.d.). Retrieved April 27, 2023, from <https://aws.amazon.com/what-is/iaas/>