**Lab 01: The Evolution and Adoption of Big Data in Modern Business**

Steven Simonsen

School of Engineering and Computational Sciences, Merrimack College

DSE6220 Big Data: Hadoop and Spark

Professor Jeremiah Lowhorn

May 12, 2024

**The Evolution and Adoption of Big Data in Modern Business**

To adequately discuss the advent of big data platforms and their adoption in business today, I have examined the financial industry with respect to this topic. According to Ahmadi (2024, as cited in Chen, 2012), “Big Data plays a crucial role in the financial sector, integrating AI algorithms with vast datasets to provide actionable insights, enhance decision-making, and optimize operational efficiency” (p.4). As the world continues to evolve, more data is generated as a result. Many factors, including the IoT (Internet of Things) and continued expansion of the ability to access the internet worldwide, have contributed to the ever-increasing collection of data, thus leading to big data. For this reason, the financial industry, and many others, have adopted big data platforms to facilitate feasible data ingestion, wrangling, and visualizations to lead to successful business outcomes.

For businesses investing in data science, big data platforms are needed to efficiently act upon very large amounts of data. Big data platforms can significantly increase a business’s productivity by structuring large amounts of information within unstructured data sets, thus leading to increased efficiency, performance, and agility (Pamucar, D et al., 2024). Much of the data that is collected today is captured and in an unstructured format. Therefore, to efficiently utilize time and money, it is practical for businesses to invest in big data platforms. From my own experience and class discussion, this is made obvious through the example of social media platforms. Pictures, posts, and comments are collected in an unstructured format, and must be standardized in a structured manner for meaningful analysis and decision-making. For these reasons, it’s vital for businesses, especially large businesses with large amounts of data collected, to invest in big data platforms.

Returning to the financial industry as an example, big data has helped to facilitate change to business models. By utilizing Big Data and AI, customers can identify tailored solutions to their problems based on the collection of data and determination of their banking preferences (Ahmadi, 2024). This ultimately provides businesses with new methods for fostering customer relationships through the examination of such solutions. In conclusion, financial firms and banks can change the way they approach customer relationships by structuring their business models to meet the ever-evolving expectations and needs of their consumers.

**Data Lake Discussion**

**Data Lakes Defined**

“Despite their popularity, a notable gap exists—there is currently no formal and universally accepted definition for ‘data lakehouse’” (Cherradi, 2024, p.2). However, data lakes typically function as a central repository to an organization by executing tasks such as SQL queries, employing ACID (Atomicity, Consistency, Isolation, and Durability) transactions, and visualizing incredibly large petabyte-scale data sets on cloud storage (Cherradi, 2024). I found the lack of a universally accepted definition for a data lake to be surprisingly useful. While practical business applications and uses are formally defined, the lack of a definition intrinsically implies the variability in scope of data lakes. To elaborate, data lakes can vary widely depending on the needs and applications across the business.

**Data Lake Adoption**

To illustrate the need for organizations to adopt data lakes, Cherradi (2024) states, “As organizations grapple with escalating data volumes and diverse data types, the need for a comprehensive solution that seamlessly integrates the strengths of data warehouses and data lakes becomes increasingly apparent” (p.3). Since big data ingestion is proving to be necessary across a large spectrum of industries, a solution is needed to combine the traditional ETL (Extract, Transform, Load) capabilities with the collection of unstructured data. Many companies are turning to data lakes to accomplish this need, since a data lake can support and process both structured and unstructured data types. However, the implication can be made to assume that challenges exist surrounding data quality within a data lake. While the traditional data warehouse often contains traditional structured datasets, the introduction of unstructured data into a data lake can present issues in ensuring data quality in terms of consistency and integrity of the data. As a solution, big data platforms exist to perform data driven tasks for standardization, cleansing, and maintenance of the data.

**Data Lakes in Data Science**

Data lakes can be incredibly useful tools in the data science field. It is the job of data scientists to convert data insights into action with statistical models, predictive analytics, and visualizations. Often, data scientists gather data from a variety of sources, and the data can be in both structured and unstructured formats. A data lake provides a centralized place to access this information. Additionally, a data lake provides an environment to perform ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) in the context of big data. Big data platforms can be used to perform this functionality, and data lakes can also provide data governance for each member within an organization that may need access to the data at different points in the data lifecycle.

**Big Data Four Vs**

The four Vs of big data consist of the following: Volume, Variety, Velocity and Veracity. According to Dasgupta (2018), the four Vs can be defined as follows:

Volume is the amount of data being generated. Variety is the different types of data, such as textual, media, and sensor or streaming data. Velocity is the speed at which data is being generated, such as millions of messages being exchanged at any given time across social networks. Veracity has been a more recent addition to the three Vs and indicates the noise inherent in data, such as inconsistencies in recorded information that requires additional validation. (p.52)

Each of these concepts help to illustrate the larger theme that data, especially in the context of big data, is layered across a rich landscape. For example, while it is important to understand the volume of the data, it can be equally important to understand the velocity. A very large data set with a low velocity could indicate a very different application throughout its data lifecycle, as opposed to a data set that is very large with a very high velocity.

Over the last 10-15 years, the four Vs of big data have evolved dramatically. In 2010, the amount of data created, captured, copied, and consumed globally was two zettabytes worldwide (Taylor, 2023), whereas in 2025 the amount of data is expected to increase to 181 zettabytes. Considering a zettabyte is a trillion gigabytes, this is quite a large increase in a 15-year period. The velocity of data has also grown dramatically, given the vast leap made in industries such as telecommunications services. Just 15 years ago, there were dramatically fewer mobile devices interacting with one another. New social media networks, such as TikTok, have continued to contribute to the increase in velocity with respect to overall data generation. The variety of the data produced has also increased. As previously discussed, the introduction of big data platforms to handle large amounts of unstructured data has contributed to the ability to collect and manage a larger variety of data in various formats. Finally, veracity has also increased over the last 10-15 years. Just from the increase in volume alone, data inconsistencies and noise has also risen. As an example, imagine the complexity of the data that the social media company Instagram receives daily. The variation of devices creating the data, the formats used to create the data (pictures, likes, comments, etc.), and versions of software each device is running could all contribute to the veracity of the data. It very quickly becomes clear that veracity has increased over the last 10-15 years due to volume, but also variation in how data is generated.

Technology has evolved as data collection has grown. Two of the more recent areas where technology has grown notably have been in the areas of AI (Artificial Intelligence) and ML (Machine Learning). According to Anurag (2024), “Our lives depend on AI, from virtual assistants like Siri, Alexa, and Google Assistant to driverless cars and disease diagnosis” (p.2). Anurag (2024) goes on to state:

The true breakthrough came with the emergence of Machine Learning (ML). It enabled systems to learn from past data and make predictions without explicit programming. This shift began a new era where AI systems could understand and improve their performance over time. (p.2)

A prime example of the above is in reference to driverless cars. The use of driverless automobiles has become very common over the last decade. As new roads are built, and unfortunately as automobile accidents occur, the use of AI and ML can be utilized to collect data and continuously improve the performance of the vehicles. Every day, there are arguments in news and media outlets questioning the safety and reliability of self-driving vehicles. However, as more data is collected, and better models are created and fitted to the data sets with ML, the AI capabilities of self-driving vehicles will continue to improve upon safety and reliability.

In conclusion, the adoption of big data across numerous industries continues to expand. As such, the use of big data platforms also continues to also grow, as does the reliability and use-cases for Machine Learning and Artificial Intelligence. Data lakes will continue to be critical for data processing and storage of data across industries. As such, data scientists and machine learning experts will continue to serve a vital role as our society continues to push the boundaries of big data into the future.

**References**

Ahmadi, S. (2024). A Comprehensive Study on Integration of Big Data and AI in Financial Industry and its Effect on Present and Future Opportunities. *International Journal of Current Science Research and Review*, *7*(01), 66-74.

Anurag, A. S. (2024). The Evolution of AI and Data Science. In *The Ethical Frontier of AI and Data Analysis* (pp. 295-312). IGI Global

Cherradi, M. (2024, April). Data Lakehouse: Next Generation Information System. In *Seminars in Medical Writing and Education* (Vol. 3, pp. 67-67).

Dasgupta, N. (2018). *Practical Big Data Analytics*. Packt Publishing.

Pamucar, D., Simic, V., Görçün, Ö. F., & Küçükönder, H. (2024). Selection of the best Big Data platform using COBRAC-ARTASI methodology with adaptive standardized intervals. *Expert Systems with Applications*, *239*, 122312.

Taylor, P. (2023, November, 16). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/