Assignment 2: Who Defines Big Data?

Akhilesh Keerthi

G01353729

Data Analytics Engineering, MS

George Mason University

AIT-580 Analytics: Big Data to Information

Prof. Myeong Lee

February 23, 2022

Who Defines Big Data?

There is no “authority” or designated body as to the exact meaning of “Big Data” and its solutions. However, the word was coined to describe how we can process large amounts of data, which used to be a big deal to analyze using traditional technologies. The buzzword big data deals with the study of extremely large data sets. The volume, speed, and complexity of data is increasing at a faster rate than science and technology develops in the areas of data analysis, management, transportation, and use. There is broad consensus among business, academic, and government leaders about the potential of big data to drive innovation, drive commerce, and accelerate progress. With Big Data, new scalable architectures can be used to handle the volume, speed, and variety of data. Despite the broad consensus on the inherent benefits of Big Data and the existing limitations, the lack of understanding on some of the key fundamentals continues to confuse potential consumers and hinder them. become development. Here are some questions:

1. What is Big Data, and how is it defined?

2. What characteristics characterize Big Data solutions?

3. What's new in the Big Data world?

Google, Amazon, IBM, Oracle, Yahoo and other commercial companies have their own, often self-service definitions that lead directly to their specific solutions and products. Here are some of the benefits these companies' big data services offer to their customers: reduced complexity, easier scalability, better flexibility, potential cost savings. Features and improved security are just a few of the benefits. However, we'll take a look at three companies to see how they define big data and how they solve problems with it.

Big Data on AWS

According to Amazon Web Services (AWS) by Amazon (Amazon Web Services, n.d.), big data can be defined as data management challenges that cannot be handled by a traditional database due to increased volume. volume, speed and variety of data. In a relatively short time - from daily to real-time - data must be collected, stored, processed, and analyzed. AWS provides a comprehensive, fully integrated suite of cloud computing services to help you develop, secure, and deploy big data applications. There's no hardware to buy, no infrastructure to maintain, and no worries about scale with AWS, so you can focus your efforts on finding new insights. AWS offers several big data services, including Amazon EMR, AWS Lambda, and Amazon Kinesis. Amazon EMR is a cloud-based big data platform that uses open source analytics frameworks such as Apache Spark, Apache Hive, and Presto to perform large-scale distributed data processing tasks, interactive SQL queries and machine learning (ML) analytics applications. Amazon Kinesis makes it easy to collect, process, and analyze streaming data in real time, so you get timely insights and react quickly to new data. Amazon Kinesis provides the functionality needed to process streaming data at cost-effective scale, along with the freedom to choose the tools that best meet your application needs.

Big data, according to Amazon Web Services (AWS) by Amazon (Amazon Web Services, n.d.), can be defined as data management challenges that cannot be managed with traditional databases due to rising volume, velocity, and variety of data. Within relatively short time frames – ranging from daily to real-time – data must be collected, stored, processed, and analyzed.

AWS offers a comprehensive and completely integrated set of cloud computing services to assist you in developing, securing, and deploying big data applications. There's no hardware to buy, no infrastructure to maintain, and no scaling to worry about with AWS, so you can focus your efforts on finding new insights. AWS offers several Big Data services, including Amazon EMR,

AWS Lambda, and Amazon Kinesis.

Amazon EMR is a cloud big data platform that uses open-source analytics frameworks like Apache Spark, Apache Hive, and Presto to conduct large-scale distributed data processing jobs, interactive SQL queries, and machine learning (ML) applications.

Amazon Kinesis makes it simple to gather, process, and analyze real-time, streaming data, allowing you to gain timely insights and respond rapidly to new data. Amazon Kinesis provides crucial capabilities for cost-effectively processing streaming data at any scale, as well as the freedom to select the tools that best meet your application's needs.

Big Data on Microsoft

Microsoft Azure (Microsoft Azure, n.d.) provides a big data architecture for ingesting, processing, and analyzing data that is too large or complex for traditional database systems. A typical big data solution workload includes one or more of the following:

- Batch processing of large data sources at rest.
- Processing big data in motion in real time.
- Interactive discovery of big data.
- Predictive analytics and machine learning.

Here are some big data tools: Data Lake Store, HDInsight, Cosmos DB and other similar services are available. Data Lake Storage - Scale data lake archiving to terabytes and petabytes of optimized data. Data can be structured, semi-structured, or unstructured, and often comes from several heterogeneous sources. Storage and processing are included in a complete data lake solution. Fault tolerance, unlimited scalability, and the ability to deliver data of any shape and size with high throughput are all hallmarks of data lake storage. Data lake processing involves the use of one or more processing engines that have been designed with these goals in mind and

can process large amounts of data stored in the data lake. Apache HBase is an open source Hadoop-based NoSQL database inspired by Google BigTable. In non-mathematical databases organized by column families, HBase provides random access and excellent consistency for large amounts of unstructured and semi-structured data.

Microsoft Azure (Microsoft Azure, n.d.) offers a big data architecture for ingestion, processing, and analysis of data that is too massive or complicated for traditional database systems. Typical workloads in big data solutions include one or more of the following:

- Batch processing of big data sources at rest.
- Real-time processing of big data in motion.
- Interactive exploration of big data.
- Predictive analytics and machine learning.

The following are some big data tools: Data Lake Store, HDInsight, Cosmos DB, and other similar services are available.

Data lake stores – The scaling of data lake stores to terabytes and petabytes of data is optimized. Data can be structured, semi-structured, or unstructured, and it often comes from several heterogeneous sources. Both storage and processing are included in a complete data lake solution. Fault-tolerance, limitless scalability, and high-throughput intake of data of all forms and sizes are all features of data lake storage. Data lake processing involves the use of one or more processing engines that were designed with these goals in mind and can handle large amounts of data stored in a data lake.

Apache HBase is a Hadoop-based, open-source NoSQL database that is modeled after Google BigTable. In a schema-less database organized by column families, HBase allows random access and excellent consistency for vast amounts of unstructured and semi-structured data.

Big Data on Google

Big data differs from traditional data assets in terms of volume complexity and the requirement for powerful business intelligence tools to process and evaluate it, according to Google Cloud Platform (Google Cloud Platform, n.d.). Volume, variety, velocity, and variability are the characteristics that define big data. Data can be an asset for a business. Understanding the areas that affect business—from market conditions and client purchasing patterns to your business processes—can be aided by using big data to reveal insights. BigQuery, Dataflow, Data Studio, and others are some of the big data technologies available on GCP. BigQuery is a fully managed serverless enterprise data warehouse solution. You can easily start data analysis and accelerate your digital transformation journey as there is no infrastructure to set up or manage. Dataflow can be used for batch autoscaling, dynamic job rebalancing, flexible scheduling, and flexible pricing. Dataflow's real-time AI capabilities provide real-time human-like responses to large volumes of data.

References

SKIENA, S. T. E. V. E. N. S. (2018), Data Science Design Manual. SPRINGER INTERNATIONAL PU.

Group., N. I. S. T. B. D. P. W. National Institute of Standards and Technology. Retrieved from https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions.

Amazon Web Services. (n.d.). What is Big Data? - Amazon Web Services. Amazon. Retrieved February 22, 2022, from https://aws.amazon.com/big-data/what-is-big-data/

Google Cloud Platform. Dataflow. Retrieved February 22, 2022, from https://cloud.google.com/dataflow

Microsoft. (n.d.). Data Lakes - Azure Architecture Center. Retrieved February 22, 2022, from https://docs.microsoft.com/en-us/azure/architecture/data- guide/scenarios/data-lake