Big Data Analytics

Dr Sirintra Vaiwsri | Email: sirintra.v@itm.kmutnb.ac.th

Big Data History and What is Big Data?

History of Big Data

1944

- Rider et al. (1944) raised the problem of data growth.
- Data would have been increased by doubling in size every sixteen years.

1984

- Tilly (1984) first used the term Big Data in a social history paper.

1997

- Cox and Ellsworth (1997) first used the term Big Data in computer science.

NOTE

- Mashey (1999) has been credited as the first who used the term Big Data in computer science.
- This is because he used this term in his various speeches.

History of Big Data

Yahoo and Google were the first who have the scalability problems to process and store indexes of documents on the internet (Aven, 2018).

2003

- Google published a whitepaper “The Google File System (GFS)”.
- It shows the management of the file storage system for large distributed data (Buyya et al., 2016)

2004

- Google published a whitepaper “MapReduce: Simplified Data Processing on Large Clusters” (Buyya et al., 2016).

2005

- Doug Cutting and Mike Cafarella applied GFS and Google MapReduce along with the web crawler project to produce Hadoop (Buyya et al., 2016).

2006

- Yahoo hired Doug Cutting to work on Hadoop which was then joined the Apache Software Foundation (Buyya et al., 2016).

History of Vs in Big Data

Douglas Laney

Douglas Laney (2001) introduced Volume, Velocity, and Variety.

IBM

IBM added Veracity to the first 3Vs (Buyya et al.,2016).

Demchenko et al.

Demchenko et al. (2014) added Value to the 4Vs introduced by IBM.

Microsoft

Microsoft added Veracity, Variability, and Visibility to the 3Vs introduced by Douglas Laney (Buyya et al.,2016).

Characteristics of Big Data (4Vs)

Volume	Volume refers to a large amount of data. It requires high performance techniques, large data storage, and high computing resources (Vaiwsri, 2023).
Velocity	Velocity refers to the speed of data management including data creation, processing, and analysis (Kaisler et al., 2013).
Variety	Variety refers to the heterogeneous nature of data and different data types such as structured, unstructured, and semi-structured (Bahga and Madisetti, 2019; Vaiwsri, 2023).
Veracity	Veracity refers to data quality and accuracy of data (Bahga and Madisetti, 2019; Vaiwsri, 2023). Therefore, the data uncertainties must be taken into account (Vatsalan et al., 2017).

What is Big Data?

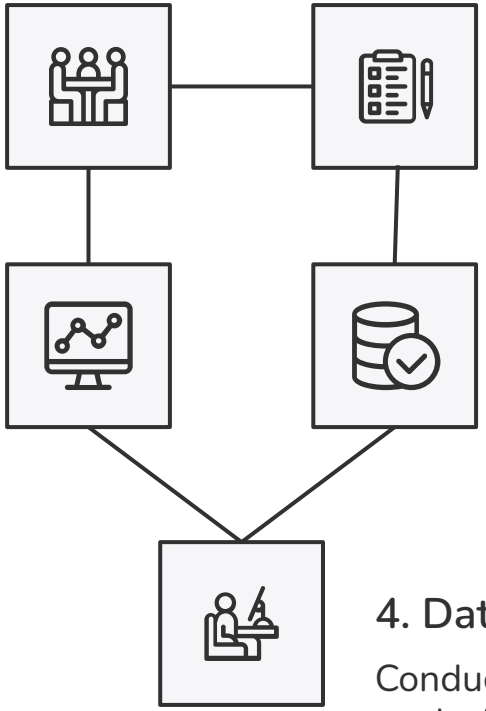
- Big Data refers to a collection of large volumes of data that cannot be efficiently processed by traditional database methods and tools (Kaisler et al., 2013).
- Big Data has become well-known in the last decade because the data size has increased, and various organisations and application domains have to involve Big Data in their analytics (Vatsalan et al., 2017).
- Data analytics often refers to processes, technologies, frameworks, and algorithms to extract and create information from raw data (Bahga and Madisetti, 2019).
- Big Data Analytics must concern the process, storage, and analysis of large data collection (Bahga and Madisetti, 2019).

Big Data Analytics Life Cycle

Big Data Analytics Life Cycle (Dietrich et al., 2015)

1. Discovery

Learn business domain, required resources, and identify a problem



2. Data Collection

Consider types of data, ingestion mechanism, and collection systems

3. Data Preparation and Storage

Extract, transform, and load data into database or file systems

4. Data Processing and Analysis

Conduct process and analysis

5. Visualisation

Visualise analysed data for decision support or further analysis

Discovery (Dietrich et al., 2015)

- Discovery is a process to identify business problems and expected outcomes and plan the overall requirements of a project.
- The processes such as :
  - Learn and understand the business domain
  - Collect requirements from the project sponsor and the business user
  - Brainstorm to consider appropriate resources such as types of data, technologies, and systems
  - Identify problems and make hypotheses for testing outcomes

Data Collection (Dietrich et al., 2015; IBM, 2023)

- **Types of data** :
  - **Structured** : data that is well-structured, organised, labelled, and has a predefined data model that conforms to a tabular format.
  - **Unstructured** : data that has not been structured, images, audio, video, sensor data, text messages, and social media posts.
  - **Semi-structured** : includes components of both structured and unstructured data, where these data are mostly unstructured, but include metadata that identifies certain characteristics such as XML files.
- **Types of data sources** :
  - **Batch data** : a data set that was collected over a period of time.
  - **Stream data** : a data set that is collected and processed at almost the same time (the time difference between data collecting and processing is in milliseconds)

Data Preparation and Storage (Dietrich et al., 2015; Bahga and Madisetti, 2019)

- Conducting extraction (E), transformation (T), and loading (L) data into the database or file systems.
- The sequence of processes can be ETL or ELT depending upon the data sources and processes required.
- The collected data often contain errors, missing values, duplicates, and inconsistencies such as inconsistent abbreviations, units and formats.
- Data pre-processing is required.

Data Preparation and Storage

- **Types of data storage:**
  - **Distributed file systems (DFS)** - distribute data across multiple locations to allow multiple users access from different locations (Arel, 2023; Cohesity, 2023).
  - **Non-relational databases (NoSQL)** (Bahga and Madisetti, 2016; Edward and Sabharwal, 2015) -
    - does not have a strict schema and formal definition
    - was designed to manage a large amount of data that cannot be managed using the traditional relational database management system (RDBMS)

Data Processing and Analysis

- The process and analysis are depended upon the types of data sources, identified problems, and expected outcome.
- The types of processes can be categorised into two groups (Janev et al., 2020) which are:
  - **Batch processing** - has high performance to process a large volume of data that was collected over some period of time.
  - **Stream processing** - also has high performance and a high ability to process a large volume of data that was collected at nearly the same time.

Visualisation

- Visualisation is a report of the results of Big Data Analytics (Janev et al., 2020)
- The visualise should deliver meaningful insight and be understandable.
- The report can be a graph, plot, chart, table, or a combination of (some of) them.

References

- Rider, F., et al. (1944). Scholar and the future of the research library.
- Tilly, C. (1984). The old new social history and the new old social history. Review (Fernand Braudel Center), 7(3), 363-406.
- Cox, M., & Ellsworth, D. (1997, October). Application-controlled demand paging for out-of-core visualization. In Proceedings. Visualization (Cat. No. 97CB36155) (pp. 235-244). IEEE.
- Mashey, J. R. Big data and the next wave of {InfraStress} problems, solutions, opportunities. In 1999 USENIX annual technical conference (USENIXATC 99) (1999).
- Aven, J. (2018). *Data Analytics with Spark Using Python*. Addison-Wesley Professional.
- Buyya, R., Calheiros, R. N., & Dastjerdi, A. V. (Eds.). (2016). *Big data: principles and paradigms*. Morgan Kaufmann.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.
- Demchenko, Y., De Laat, C., & Membrey, P. (2014, May). Defining architecture components of the Big Data Ecosystem. In *2014 International conference on collaboration technologies and systems (CTS)* (pp. 104-112). IEEE.
- Vaiwsri, S. (2023). Privacy-preserving record linkage for high linkage quality.

References

- Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In 2013 46th Hawaii international conference on system sciences (pp. 995-1004). IEEE.
- Bahga, A., & Madisetti, V. (2019). Big Data analytics: A hands-on approach.
- Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. Handbook of big data technologies, 851-895.
- Dietrich, D., Heller, B., & Yang, B. (2015). *Data science & big data analytics: discovering, analyzing, visualizing and presenting data*. Wiley.
- IBM. https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop#about. Accessed: 2023-06-05.
- Arel, R. https://www.techtarget.com/searchstorage/definition/distributed-file-system-DFS. Accessed: 2023-07-04.
- Cohesity. https://www.cohesity.com/glossary/distributed-file-system/. Accessed: 2023-07-04.
- Edward, S. G., & Sabharwal, N. (2015). *Practical MongoDB: Architecting, Developing, and Administering MongoDB*. Apress.
- Janev, V., Graux, D., Jabeen, H., & Sallinger, E. (2020). Knowledge graphs and big data processing (p. 209). Springer Nature.