Big Data Concepts

Subjects: Computer Science, Information Systems

Contributor: Diana Martinez-Mosquera, Rosa Navarrete, Sergio Lujan-Mora

The entries consolidates the main concepts about Big Data modeling and management. The characterization and the definitions about structured, semi-structured and unstructured data.

Keywords: Big Data ; modeling ; databases

1. A Brief History of Big Data

The production and processing of large volumes of data began to be of interest to researchers many years ago. By 1944, estimations for the size of libraries, which increased rapidly every year, were made in American universities[1]. In 1997, at the Institute of Electrical and Electronics Engineers (IEEE) Conference on Visualization, the term “Big Data” was used for the first time during the presentation of a study about large datasets’ visualization[2].

Big Data is the buzzword of recent years, that is, a fashionable expression in information systems. The general population relates the term Big Data to its literal meaning of large volumes of data. However, Big Data is a generic term used to refer to large and complex datasets that arise from the combination of famous Big Data V’s that characterize it[3].

2. Big Data Characterization

As mentioned before, Big Data does not refer only to high volumes of data to be processed. At the beginning of the Big Data studies, their volume, velocity and variety were considered as fundamental characteristics, which were known as the three Vs of Big Data. After advances in the research, new Vs, such as value and veracity, were established. Currently, there are authors who propose up to 42 characteristics needed to consider data as Big Data, therefore, they define 42 Vs for Big Data[4]. For the purposes of our study, we will mention only ten Vs of Big Data, that are presented in a scientific study[5]. Table 1 summarizes each characteristic, along with a brief description.

Table 1. Ten Vs Big Data.

Characteristic	Brief Description
Volume	Large data sets
Velocity	High data generation rate
Variety	Different type of data formats
Variability	Consistent data
Viscosity	Data velocity variations
Virality	Data transmission rate
Veracity	Accuracy of data
Validity	Assessment of data

Visualization	Data symbolization
Value	Useful data to retrieve info

3. Volume and Velocity

To deal with the Volume and Velocity characteristics of Big Data, ecosystems and architectural solutions, such as lambda and kappa, have been created. Both architectures propose a structure of layers to process Big Data; the main difference between them is that lambda proposes a layer for batch data processing and another for streaming data, while kappa proposes a single layer for both batch and streaming processing[6]. This SLR focuses on data modeling, a concept related to the Variety characteristic, which is explained next.

4. Variety

Variety is a characteristic referring to the different types of data and the categories and management of a big data repository. As per this characteristic, Big Data has been classified into structured, semi-structured and unstructured data[7][8]. The next subsections explain in detail each data type.

4.1. Structured Data

In Big Data, structured data are represented in tabular form, in spreadsheets or relational databases[8]. To deal with this type of data, widely developed and known technologies and techniques are used. However, according to the report presented by the CISCO company, this type of data only constituted 10% of all existing data in 2014[9]. Therefore, it is very important to analyze the 90% of the remaining data, corresponding to the semi-structured data and unstructured data that will be described below.

4.2. Semi-Structured Data

Semi-structured data are considered to be data that do not obey a formal structure, such as a relational database model. However, they present an internal organization that facilitates its processing; for instance, servers' logs in comma-separated values (csv) format, documents in eXtensible Markup Language (XML) format, JavaScript Object Notation (JSON) and Binary JSON (BSON) and so forth. Some authors may consider XML and JSON as structured[8].

4.3. Unstructured Data

Unstructured data are considered those that have either no predefined schema or no organization in their structure[10]. Within this type of data are text documents, emails, sensor data, audio files, images files, video files, data from websites, chats, electronic health records, social media data and spatio-temporal data, among others[7]. According to CISCO, the volume of unstructured data between 2017 and 2022 is expected to increase up to twelvefold[11].

To support the Variety, Volume and Velocity of Big Data, non-relational, distributed and open source data storage systems have been created. These systems include horizontal scalability, linearization, high availability and fault tolerance. Usually, these databases are known as NoSQL.

References

1. Rider, F.. The Scholar and the Future of the Research Library: A Problem and Its Solution; Hadham Press: New York, NY, USA, 1944; pp. 98–100.
2. Cox, M.; Ellsworth, D. Application-controlled demand paging for out-of-core visualization. In Proceedings of the 8th IEEE Conference on Visualization, Phoenix, AZ, USA, 24 October 1997; pp. 235–244.
3. André Ribeiro; Afonso Silva; Data Modeling and Data Analytics: A Survey from a Big Data Perspective. *Journal of Software Engineering and Applications* **2015**, *8*, 617-634, 10.4236/jsea.2015.812058.
4. Shafer, T. The 42 V's of Big Data and Data Science. Available online: https://www.elderresearch.com/blog/42-v-of-big-data (accessed on 23 August 2019)
5. Manogaran, G.; Thota, C.; Lopez, D.; Vijayakumar, V.; Abbas, K.M.; Sundarsekar, R.. Big Data Knowledge System in Healthcare; Springer International Publishing: Cham, Switzerland, 2017; pp. 133–157.

6. Valerio Persico; Antonio Pescapè; Antonio Picariello; Giancarlo Sperlí; Benchmarking big data architectures for social networks data processing using public cloud platforms. *Future Generation Computer Systems* **2018**, *89*, 98-109, 10.1016/j.future.2018.05.068.
7. Costa, C.; Santos, Y.; Big Data: State-of-the-art Concepts, Techniques, Technologies, Modeling Approaches and Research Challenges. *Int. J. Comput. Sci.* **2017**, *44*, 1–17, .
8. Ali Davoudian; Liu Chen; Mengchi Liu; A Survey on NoSQL Stores. *ACM Computing Surveys (CSUR)* **2018**, *51*, 1-43, 10.1145/3158661.
9. CISCO. Big Data: Not Just Big, but Different—Part 2. Available online: https://www.cisco.com/c/dam/en_us/about/ciscoitatwork/enterprise-networks/docs/i-bd-04212014-not-just-big-different.pdf (accessed on 10 September 2019).
10. P. O'sullivan; G. Thompson; A. Clifford; Applying data models to big data architectures. *IBM Journal of Research and Development* **2014**, *58*, 18:1-18:11, 10.1147/jrd.2014.2352474.
11. CISCO VNI. Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper. Available online: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.htm (accessed on 10 September 2019).
