Review Paper on Big Data Analytics

Vivek Chauhan1, Kiran Jot Singh2,

1Student, Department of Mechatronics Engineering Chandigarh University, Gharuan

2Assistant Professor, Chandigarh University, Gharuan

Abstract

In this technological era, there is a large amount of data from social media, website analytics, online surveys. The uniform generation of this large amount of data needs data management and analysis. To handle this huge amount of data require special tools and techniques. Big data is not only big in the data but also big in velocity and variety, which are difficult to deal with old-school tools and techniques. So big data is here to help as an important part of our lives. By collecting a large amount of data and analyze this data we can predict the customer behaviour for selling things online. This paper gives the basic introduction of big data and analytic methods and some real-life examples to understand the use of big data analytics.

Keywords

Data management, data storage, big data, analytics, HDFS.

Introduction

Today, we are using the internet for long hours. For buying things online we visit an online website and search for the product and we find the list of products that are stored in their database. We can not think of the world without data storage. Now, think we are logged in to our facebook account and post a picture if Facebook does not have data storage or database the picture that you posted on your timeline is lost directly after posting. So to see the posts shared by others organization need a database. Every time when we go online we generated a huge amount of data which have to be stored and analyzed. This type of data is high in volume, velocity and of a different variety like on a single day facebook generated more than 500 TB of data which have to be managed and analyzed. This data is uniformly increased and become complex and can not be processed by old tools. To manage and analyze such a large dataset we need big data and analytics tools.

Literature Review

Characteristics Of Big Data

There are five Vs of big data Volume, Variety, Velocity, Variability, Veracity. The volume means the amount of data generated and stored. The variety of the big data is the type and nature of data. The velocity of data is identified by the rate of data generation and processing. The variability characterizes the need to get significant information thinking about every single conceivable situation. The veracity confirms that the information is reasonable for its planned reason and usable inside the scientific model. The information is to be tried against an arrangement of characterized criteria. Big data can be measured by estimate in TBs or PBs, and even the quantity of records, exchanges, tables, or documents. something that makes enormous information huge is

that it's originating from a more noteworthy assortment of sources than any time in recent memory, including logs, clickstreams, and social media. Utilizing these hotspots for investigation implies that regular organized data is presently joined by unstructured data, for example, content and human dialect, and semi-structured data, for example, eXtensible Markup Language. There's additional information, which is difficult to order since it originates from sound, video, what're more, different gadgets.

Tools For Analytics

Such data can never again be effortlessly broke down with the conventional data management and examination methods. Accordingly, there emerges a requirement for new instruments and strategies particular to enormous information investigation, and additionally the required architectures for putting away and overseeing such information.[1] As needs are, the development of enormous information affects everything from the information itself and its gathering, to the handling, to the last removed choices.

Big Data analytics framework

Apache Hadoop

Apache Hadoop is an open source, a scalable framework which is written in Java. It can efficiently process large volumes of data from a cluster. Hadoop is a platform for storing large volumes of data and also process it. Hadoop are composed of several components.HDFS(Hadoop Distributed File System) is the essential data storage framework utilized by Hadoop applications.[10] It utilizes a NameNode and DataNode architecture to execute a conveyed document framework that gives superior access to information crosswise over highly scalable Hadoop groups.HDFS also confirms that the data remain available in spite of a host failure. YARN(Yet Another Resource Negotiator) is the cluster coordinating and managing the resources. MapReduce is a programming model intended for preparing vast volumes of data in parallel by separating the work into an arrangement of autonomous undertakings.[7]

Apache Spark

Apache Spark is a quick, in-memory information handling motor with exquisite and expressive advancement APIs to enable data workers to proficiently execute spilling, machine learning or SQL remaining tasks at hand that require quick iterative access to datasets. With Spark running on Apache Hadoop YARN, engineers wherever would now be able to make applications to abuse Spark's influence, infer bits of knowledge and enhance their information science outstanding tasks at hand inside a solitary, shared dataset in Hadoop.

Apache Storm

Apache Storm is a real-time system for processing high volume and velocity of data. Apache Storm is fast, with the ability to process a million records per second per node on a cluster of humble size.

Big Data Management

Big data management to the productive taking care of, association or utilization of substantial volumes of structured and unstructured data having a place with an organization.analytics and risk discovery.[6] Big data management enables an organization to comprehend its clients better, grow new items and settle on imperative money-related choices in view of the examination of a lot of corporate information.

Big data for decision making

Big data is turning into an undeniably critical resource for decision makers. A large amount of data from different sources, for example, scanners, portable telephones, dedication cards, the web, and online life stages give the chance to convey huge advantages to associations. This is conceivable just if the information is legitimately broke down to uncover profitable experiences, taking into account chiefs to underwrite upon the subsequent open doors from the abundance of noteworthy and constant information produced through supply chains, generation forms, client practices.[8]

Customer intelligence

Big data analytics holds to a large extent for customer knowledge, and can very advantage businesses, for example, retail, managing an account, and broadcast communications. Enormous information can make straightforwardness, and make applicable information all the more effectively available to partners in a convenient manner.[2] This can permit them to make more educated promoting choices, and market to various fragments based on their inclinations alongside the acknowledgement of offers and showcasing openings.

Risks and Fraud Detection

Businesses, for example, investments or retail saving money, and protection can profit by enormous information investigation in the zone of hazard administration. Since the assessment and direction of hazard is a basic perspective for the monetary administration's area, enormous data analytics can help in choosing ventures by breaking down the probability of increases against the probability of misfortunes. Also, inner and outer enormous information can be broken down for the full and dynamic evaluation of hazard exposures. For fraud detection particularly in the administration, keeping the money, and protection ventures, big data analytics can be utilized to identify and avert extortion. Client behaviour can be utilized to show ordinary client behaviour, what's more, recognize suspicious or dissimilar exercises through the exact hailing of exception events. Moreover, giving frameworks huge information about winning extortion examples can enable these frameworks to take in the new sorts of cheats and act appropriately, as the fraudsters adjust to the old frameworks intended to distinguish them.

Conclusion

we have analyzed the creative subject of huge information, which has as of late picked up bunches of enthusiasm because of its apparent exceptional chances and advantages. In the data period, we are at present living in, voluminous assortments of high-speed information are being delivered day by day, and inside them lay inborn points of interest and examples of concealed learning which ought to be extricated and used. Consequently, enormous information the investigation can be connected to use business change and improve basic leadership, by applying progressed systematic strategies to enormous information and uncovering concealed bits of knowledge and profitable information

References

1. Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011)
2. Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)
3. Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C.: Text Knowledge Mining: An Alternative to Text Data Mining. In: IEEE International Conference on Data Mining Workshops,pp. 664–672 (2008)
4. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and SpaceefficientData Placement Structure in MapReduce-based Warehouse Systems. In: IEEEInternational Conference on Data Engineering (ICDE), pp. 1199–1208 (2011)
5. EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508(2012)
6. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)
7. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New AnalysisPractices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492(2009)
8. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data:The Big Data Revolution! In: Proceedings of the ACM International Workshop on DataWarehousing and OLAP, pp. 101–104 (2011)
9. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: ASelf-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011)
10. Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Y smart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011)