Sailee Mohan Ingle, 2021, 9:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Big Data Analytics for Industries

Miss. Sailee Mohan Ingle, Asst. Prof. Lokesh S. Khedekar
Department of Computer Science and Engineering,
Rajarshi Shahu College of Engineering,
Sant Gadge Baba Amravati University,
Buldana-443001, India.
saileeingle@gmail.com, lokeshkhedekar@gmail.com

Abstract- Big data is a new driver of the world economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. While the data complexities are increasing including data's volume, variety, velocity and veracity, the real impact hinges on our ability to uncover the `value' in the data through Big Data Analytics technologies.

Keywords- Big data, data mining, analytics, decision making.

I. INTRODUCTION

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use.

Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity.

1. Big Data Analytics:

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zetta bytes.

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency.

Big data has one or more of the following characteristics: high volume, high velocity or high variety. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data.

For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media — much of it generated in real time and at a very large scale.

Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable.

Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

2. Characteristics of Big Data:

Fig 1. Big Data.

3. Big Data Analytics Tools:

1. Hadoop	11. Storm
2. Xplenty	12. Rapidminer
3.Cdh (Cloudera Distribution For Hadoop)	13. Qubole
4. R	14. Tableau
5. Cassandra	15. Samoa
6. Knime	16. Openrefine

II. BIG DATA STORAGE AND MANAGEMENT

One of the first things organizations have to manage when dealing with big data is where and how this data will be stored once it is acquired. The traditional methods of structured data storage and retrieval include relational databases, data marts, and data warehouses.

The data is uploaded to the storage from operational data stores using Extract, Transform, Load (ETL), or Extract, Load, Transform (ELT), tools which extract the data from outside sources, transform the data to fit operational needs, and finally load the data into the database or data warehouse.

Thus, the data is cleaned, transformed, and catalogued before being made available for data mining and online analytical functions. However, the big data environment calls for Magnetic, Agile, Deep (MAD) analysis skills, which differ from the aspects of a traditional Enterprise Data Warehouse (EDW) environment. First of all, traditional EDW approaches discourage the incorporation of new data sources until they are cleansed and integrated.

III. BIG DATA ANALYTIC PROCESSING

After the big data storage, comes the analytic processing. There are four critical requirements for big data processing. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time.

The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase.

Additionally, the third requirement for big data processing is the highly efficient utilization of storage space. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how to store the data so that space utilization is maximized be addressed.

Finally, the fourth requirement is the strong adaptively to highly dynamic workload patterns. As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns.

Map Reduce is a parallel programming model, inspired by the "Map" and "Reduce" of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions.

According to EMC, the Map Reduce paradigm is based on adding more computers or resources, rather than increasing the power or storage capacity of a single computer; in other words, scaling out rather than scaling up. The fundamental idea of Map Reduce is breaking a task down into stages and executing the stages in parallel in order to reduce the time needed to complete the task.

The first phase of the Map Reduce job is to map input values to a set of key/value pairs as output. The "Map" function accordingly partitions large computational tasks into smaller tasks, and assigns them to the appropriate key/value pairs.

Thus, unstructured data, such as text, can be mapped to a structured key/value pair, where, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the "Reduce" function.

Reduce then performs the collection and combination of this output, by combining all values which share the same key value, to provide the final result of the computational task. The Map Reduce function within Hadoop depends on two different

nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results. The Map Reduce job starts by the Job- Tracker assigning a portion of an input file on the HDFS to a map task, running on a node.

On the other hand, the Task Tracker nodes actually run the jobs and communicate results back to the Job Tracker. That communication between nodes is often through files and directories in HDFS, so inter-node communication is minimized.

Figure 1 shows how the Map Reduce nodes and the HDFS work together. At step 1, there is a very large dataset including log files, sensor data, or anything of the sorts. The HDFS stores replicas of the data represented by the blue, yellow, beige, and pink icons, across the Data Nodes.

In step 2, the client defines and executes a map job and a reduce job on a particular data set, and sends them both to the Job Tracker. The Job Tracker then distributes the jobs across the Task Trackers in step 3.

The Task Tracker runs the mapper, and the mapper produces output that is then stored in the HDFS file system. Finally, in step 4, the reduce job runs across the mapped data in order to produce the result.

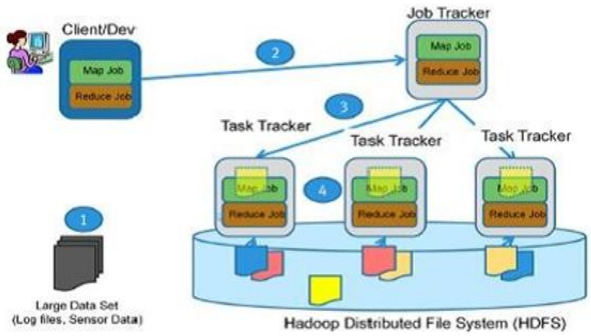

Fig 2. Text Here Your Fig Title.

After big data is stored, managed, and processed, decision makers need to extract useful insights by performing big data analyses. In the subsections below, various big data analyses will be discussed, starting with selected traditional advanced data analytics methods, and followed by examples of some of the additional, applicable big data analyses.

IV. BIG DATA ANALYTICS AND DECISION MAKING

The first phase of the decision making process is the intelligence phase, where data which can be used to identify problems and opportunities is collected from internal and external data sources.

In this phase, the sources of big data need to be identified, and the data needs to be gathered from different sources, processed, stored, and migrated to the end user. Such big data needs to be treated accordingly, so after the data sources and types of data required for the analysis are defined, the chosen data is acquired and stored in any of the big data storage and management tools previously discussed.

After the big data is acquired and stored, it is then organized, prepared, and processed, This is achieved across a high-speed network using ETL/ELT or big data processing tools, which have been covered in the previous sections.

The next phase in the decision making process is the design phase, where possible courses of action are developed and analyzed through a concept-ualization, or a representative model of the problem.

The framework divides this phase into three steps, model planning, data analytics, and analyzing. Here, a model for data analytics, such as those previously discussed, is selected and planned, and then applied, and finally analyzed.

Consequently, the following phase in the decision making process is the choice phase, where methods are used to evaluate the impacts of the proposed solutions, or courses of action, from the design phase. Finally, the last phase in the decision making process is the implementation phase, where the proposed solution from the previous phase is implemented.

As the amount of big data continues to exponentially grow, organizations throughout the different sectors are becoming more interested in how to manage and analyze such data. Thus, they are rushing to seize the opportunities offered by big data, and gain the most benefit and insight possible, consequently adopting big data analytics in order to unlock economic value and make better and faster decisions.

Therefore, organizations are turning towards big data analytics in order to analyze huge amounts of data faster, and reveal previously unseen patterns, sentiments, and customer intelligence.

This section focuses on some of the different applications, both proposed and implemented, of big data analytics, and how these applications can aid organizations across different sectors to gain valuable insights and enhance decision making.

V. QUALITY MANAGEMENT AND IMPROVEMENT

Especially for the manufacturing, energy and utilities, and telecommunications industries, big data can be used for quality management, in order to increase profitability and reduce costs by improving the quality of goods and services provided.

For example, in the manufacturing process, predictive analytics on big data can be used to minimize the performance variability, as well as prevent quality issues by providing early warning alerts. This can reduce scrap rates, and decrease the time to market, since identifying any disruptions to the production process before they occur can save significant expenditures.

VI. RISK MANAGEMENT AND FRAUD DETECTION

Industries such as investment or retail banking, as well as insurance, can benefit from big data analytics in the area of risk management. Since the evaluation and bearing of risk is a critical aspect for the financial services sector, big data analytics can help in selecting investments by analyzing the likelihood of gains against the likelihood of losses. Additionally, internal and external big data can be analyzed for the full and dynamic appraisal of risk exposures. Accordingly, big data can benefit organizations by enabling the quantification of risks.

High-performance analytics can also be used to integrate the risk profiles managed in isolation across separate departments, into enterprise wide risk profiles. This can aid in risk mitigation, since a comprehensive view of the different risk types and their interrelations is provided to decision makers.

As for fraud detection, especially in the government, banking, and insurance industries, big data analytics can be used to detect and prevent fraud.

Analytics are already commonly used in automated fraud detection, but organizations and sectors are looking towards harnessing the potentials of big data in order to improve their systems. Big data can allow them to match electronic data across several sources, between both public and private sectors, and perform faster analytics.

VII. CONCLUSION

In this research, we have examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge which should be extracted and utilized.

Hence, big data analytics can be applied to leverage business change and enhance decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and valuable knowledge.

REFERENCES

[1] Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010).

[2] Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D.: Visual Analytics for the Big Data Era—A Comparative Review of State-of-the-Art Commercial Systems. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182 (2012).

[3] Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010).

[4] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010).

[5] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012).

[6] Cebr: Data equity, unlocking the value of big data. In: SAS Reports, pp. 1–44 (2012).

[7] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009).