ESP-JETA
ESP Journal of Engineering & Technology Advancements
Volume 3 / Special Issue Paper Id: ICI2ETM-M118 / Page Count: 28-30

Original Article

Comparison Tools for Big Data Analytics

Deepali N. Kende[1], Samina H. Khan[2]

[1,2]Assistant Professor, Department of Computer Science, ICEEM, Maharastra, India.

Abstract*: Big data analytics is the process of collecting, examining, and analysing large amounts of data to discover market trends, insights, and patterns that can help companies make better business decisions. This information is available quickly and efficiently so that companies can be agile in crafting plans to maintain their competitive advantage. . Across different business segments, increasing efficiency leads to overall more intelligent operations, higher profits, and satisfied customers. Big data analytics helps companies reduce costs and develop better, customer-centric products and services. This paper gives more information about different types of data, different types of software to handle big data and its applications.*

Keywords: *Data Analytics, Crafting Plans.*

I. LITERATURE SURVEY

Various techniques are there to handle big data, In this paper we studied various techniques to handle big data according to its specifications so that readers can easily understand and choose any one technique.

II. TYPES OF BIG DATA ANALYTICS

There are four main types of big data analytics that support and inform different business decisions.

A. Descriptive Analytics

Descriptive analytics refers to data that can be easily read and interpreted. This data helps create reports and visualize information that can detail company profits and sales.

Example: During the pandemic, a leading pharmaceutical company conducted data analysis on its offices and research labs. Descriptive analytics helped them identify unutilized spaces and departments that were consolidated, saving the company millions of dollars.

B. Diagnostics Analytics

Diagnostics analytics helps companies understand why a problem occurred. Big data technologies and tools allow users to mine and recover data that helps dissect an issue and prevent it from happening in the future.

Example: A clothing company's sales have decreased even though customers continue to add items to their shopping carts. Diagnostics analytics helped to understand that the payment page was not working properly for a few weeks.

C. Predictive Analytics

Predictive analytics looks at past and present data to make predictions. With artificial intelligence (AI), machine learning, and data mining, users can analyze the data to predict market trends.

Example: In the manufacturing sector, companies can use algorithms based on historical data to predict if or when a piece of equipment will malfunction or break down.

D. Prescriptive Analytics

Prescriptive analytics provides a solution to a problem, relying on AI and machine learning to gather data and use it for risk management.

Example: Within the energy sector, utility companies, gas producers, and pipeline owners identify factors that affect the price of oil and gas in order to hedge risks.

III. DIFFERENT DATA ANALYTIC TOOLS

A. Hadoop

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Hadoop consists of four main modules:

- Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.
- Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.
- MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.
- Hadoop Common – Provides common Java libraries that can be used across all modules.

R – Programming:

R is the leading analytics tool in the industry and is widely used for statistics and data modeling. It can easily manipulate data and present it in different ways. It has exceeded SAS in many ways like capacity of data, performance and outcome. R compiles and runs on a wide variety of platforms viz -UNIX, Windows and macOS. It has 11,556 packages and allows you to browse the packages by category. R also provides tools to automatically install all packages as per user requirements, which can also be well assembled with Big data.

B. Tableau Public:

Tableau Public is a free platform to explore, create, and publicly share data visualizations (or vizzes, as we affectionately call them) online. Anyone can create vizzes using our in-platform web authoring or Tableau Desktop Public Edition for free. Those with Tableau Desktop Professional Edition can also publish to Tableau Public for free. With millions of inspiring data visualizations to discover and learn from, Tableau Public makes it easy to develop your own data skills and create an online portfolio of work. Join the Tableau Public community where you can grow and learn from each other while making data a part of your everyday life.

C. Python:

Python provides a huge number of libraries to work on Big Data. Python is very scalable in handling large amounts of data which is a necessity where Big Data is concerned. Python is also extremely flexible. and efficient. It allows developers to complete more work using fewer lines of code. Python provides features for identifying and processing unstructured data which can include voice, text, and image data as well. Python can also handle data processing when the data is in different files such as CSV, XML, HTML, SQL, and JSON, etc

D. SAS:

SAS (Statistical Analysis System) introduced the SAS Business Intelligence and Analytics Solution for helping large enterprises explore their large datasets in a visually appealing format. SAS analytics is a data analytics tools that is used increasingly in Data Science, Machine Learning, and Business Intelligence applications. Not only it equips

organizations with all necessary tools to monitor the key BI metrics but also produces powerful analytics and comprehensive reports for their decision makers to take well-informed decisions.

E. Apache Spark:

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. Spark was created to address the limitations to MapReduce, by doing processing in-memory, reducing the number of steps in a job, and by reusing data across multiple parallel operations. With Spark, only one-step is needed where data is read into memory, operations performed, and the results written back—resulting in a much faster execution

F. MangoDB:

MongoDB is an open-source document-oriented database that is designed to store a large scale of data and also allows you to work with that data very efficiently. It is categorized under the NoSQL (Not only SQL) database because the storage and retrieval of data in the MongoDB are not in the form of tables. The data model that MongoDB follows is a highly elastic one that lets you combine and store data of multivariate types without having to compromise on powerful indexing options, data access, and validation rules.

IV. CONCLUSION

Choosing analytic tool depends on my factors like business objective, pricing, user interface and visualization, advanced analytics, integration, multiple source of data, Customization, collaboration and security.

V. REFERENCES

[1] Sameera Siddiqui, Deepa Gupta," Big Data Process and Analytics : A Survey", International Journal Of Emerging Research in Management & Technology, ISSN: 2278-9359, Volume 3, Issue 7, July 2014.
[2] Han Hu, Yongyang Nen, Tat Seng Chua, Xuelong Li," Towards Scalable System for Big Data Analytics: A Technology Tutorial", IEEE Access, Volume 2, Page No 653, June 2014.
[3] Bharti Thakur, Manish Mann," Data mining for big data: A Review", International journal of advanced Research in Computer Science and Software Engineering, ISSN: 2277 128x, Volume 4, Issue 5, May 2014.
[4] Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions", International Conference On Cloud, Big Data and Trust 2013, Nov 2013.
[5] Albert Bifet, "Mining Big Data in Real Time", informatica, 2013.
[6] Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money," Big Data: Issues and Challenges Moving Forward", Hawaii International Conference on System Science, IEEE Computer Society, Page No. 995, 2013.