BIG DATA ANALYSIS METHODS

Seitnazarov Kuvanshbay Kenesbayevich
Doctor of technical sciences, associate professor of TATU
Nukus named after Muhammad al-Khorazmi
Haitbayev Azizbek Pirnazarovich
Master of the Urganch branch of TATU named
after Muhammad al-Khorazmi
azizbekkhaitbayev93@gmail.com , +998 88 567 56 56
https://doi.org/10.5281/zenodo.7851412

Abstract: Nowadays, the volume of big data is increasing dramatically, and recent developments in technology have resulted in enormous amounts of data. Traditional databases do not have the ability to process this variety of data and therefore have led to the application of new technologies, methods and tools. This article sheds light on big data, available big data analytics tools, the need to use big data analytics, and its benefits and challenges.

Key words: Technology, trend, messenger, model, smart technologies, big data, big data, data analysis, sensor, cluster, analytics.

Enter. All aspects of our lives are based on data and it is analyzed every second, which includes a lot of information, such as messages on Internet-connected messengers on your phone, reports on money transfers, purchase histories on online shopping platforms. Billions of people in the world have access to the Internet. Every move they make online generates new data. For today's business development, data statistics and analytics are very important for its survival and profitability. With this in mind, data analytics has become a highly valuable field today. The reason is that the most convenient and easy way to know the level of demand and needs for any business is the statistics of big data collected using modern technologies. All areas of business are using this area. When organizations use data, they benefit customers and businesses by developing new data-driven services, developing new business models and strategies, and selling data-driven products and utilities. The drive to invest and implement data analytics tools and techniques is huge, and businesses need to adapt, innovate and strategize for the evolving digital marketplace.

Data is being created at such a rapid rate that we are forced *to invent new units to measure it.* Hosted by Abby McCain conclusion that's it shows that each person creates average 1.7 MB information per second [1]

Users use search engines to find relevant information for their information needs on the Internet. The search engine Google receives over 63,000 searches per second on any given day [2]

New technologies such as big data analytics are considered smart technologies because they involve the risk of disruption caused by the ability to become autonomous, learn and operate, in addition to automating existing processes and providing information. [3]

But this is not a sufficient solution for analyzing large volumes of data. The reason is that most of the data available today is unstructured, that is, unanalyzed data.

currently more than 44 zettabytes of data in the entire digital universe, and 70% of that data is user-generated. [4]

If we look at the data explosion, we can see that 90% of all available data in the world was created in the last two years. However, previous predictions did not even reach half of these indicators.

Data extraction, data cleaning, data integration, data transformation, and data reduction operators can be considered as preliminary data analysis processes. [5]

In 2023, managing unstructured data will become an increasingly large challenge for enterprises. To quantify the problem; The global data industry is expected to more than double from 2022 to 2026, and unstructured data now accounts for 80% to 90% of new enterprise data. This significant growth is due to digital transformation, which in the last two years has created the need to work in a hybrid way, to collaborate in different locations, which has created additional file sharing needs. The demand for actionable insights has also grown – data analytics has increased to drive efficiency and innovation. [6]

Big data analytics is the process of collecting, examining, and analyzing large amounts of data to uncover insights, a trend that helps companies make better decisions. This information can be generated quickly and efficiently to help companies plan to maintain their competitive advantage. Big data has the 5V characteristics of volume, variety, velocity, and accuracy, which necessitates big data analytics.

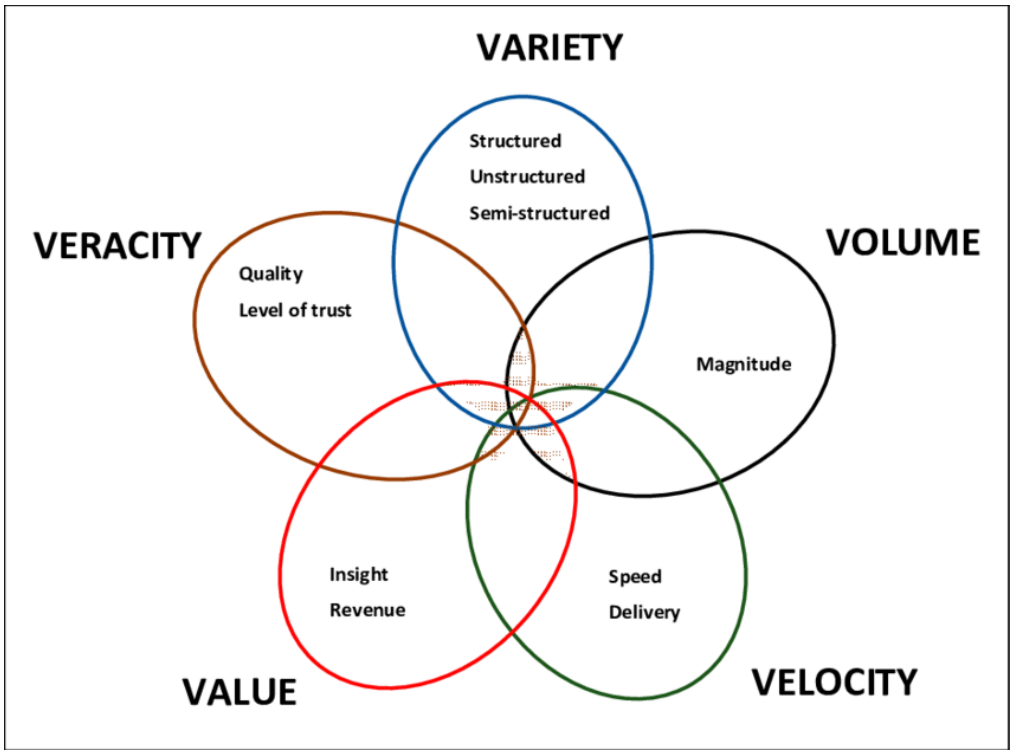

Figure 1. 5 characteristics of big data.[7]

There are two main types of data processing: batch and stream. Batch processing occurs on blocks of data stored for a certain period of time. Usually the data processed in bulk is large, so it takes more time to process it. Hadoop MapReduce is the best framework for processing data in batches. This approach works well in situations where there is no need for real-time analysis and when it is important to process large amounts of data to obtain more detailed information. [8]

On the other hand, stream processing is the key to real-time data processing and analysis. Stream processing allows data to be processed as it arrives. This data is immediately fed into the analysis tools, so the results are generated immediately. There are many scenarios where such an approach can be useful, such as fraud detection, where anomalies that signal fraud are detected in real time. Another use case would be online retailers, where real-time processing allows them to collect a large history of customer interactions so that additional purchases can be recommended for customers in real-time.

Data analysis is carried out step by step.

1. Experts gather information from various sources **.** Often , it will be consists from half compiled and not structured data Although each organization uses different data streams, some common sources include:
   - internet streaming data;
   - web server logs;
   - cloud applications;
   - mobile applications;
   - social network content;
   - text from customer emails and survey responses;
   - mobile phone records ;
   - Machine data captured by IoT-connected sensors.
2. Data is prepared and processed . After data is collected and stored in a data warehouse, data professionals need to properly organize, configure, and segment the data for analytical queries. Careful preparation and processing of data increases the efficiency of analytical queries. Stream processing processes small batches of data at a time, reducing the lag time between collection and analysis for faster decision making. Stream processing is more complex and expensive.
3. Data is cleaned to improve quality **.** Data professionals clean data using scripting tools or data quality software. They look for any errors or inconsistencies, such as duplication or formatting errors, and organize the data.
4. Making big data usable takes time. Once it's ready, advanced analytics processes can turn big data into big insights.
   - Data mining through large data sets to identify metrics and relationships by identifying anomalies and creating data clusters .
   - Predictive analytics uses an organization's historical data to make predictions about the future and identify upcoming risks and opportunities.
   - Deep learning mimics human learning by using artificial intelligence and machine learning to layer in algorithms and identify patterns in the most complex and abstract data.

When we look at the main technologies and tools of big data analysis, these technologies give us several advantages.

- Making faster and better decision

Businesses can access vast amounts of data and analyze a variety of data sources to gain new insights and take action.

- Cost reduction and operational efficiency

Flexible data processing and storage tools help organizations save on the costs of storing and analyzing large amounts of data. Discovers patterns and insights that help identify ways to make the business more efficient.

— Market access based on improved data

Analyzing data from sensors, devices, video, logs, transactional applications, the web, and social media enables an organization to become data-driven. Helps assess customer needs and potential risks and develop new products and services.

Dynamic capabilities theory provides useful analytical tools for uncovering the capabilities and processes required to exploit big data. Through these capabilities, organizations can acquire and use valuable, unique, nonrepeatable, and organized (VRIO) resources [9]

Big data analytics cannot be boiled down to a single technology. Several types of tools work together in parallel to help collect, process, clean, and analyze big data. Below are some of the key tools of the big data ecosystem.

- Hadoop is an open-source paradigm that efficiently stores and processes large data sets on clusters of commodity hardware. This form is free and capable of handling large amounts of structured and unstructured data, making it a valuable backbone for any big data operation.
- NoSQL databases are non-relational data management systems that do not require a conditional schema, making them an excellent option for large, unstructured data. NoSQL stands for non-binding SQL and these databases can handle different data models.
- MapReduce is an important component of Hadoop form that performs two functions. The first is a mapping that filters the data to different nodes within the cluster. The second is reduction, which organizes and reduces the results of each node to answer the query.
- YARN stands for "Another Resource Negotiator". It is another component of second generation Hadoop. Cluster management technology helps in scheduling work and resource management in the cluster.
- Spark is an open-source cluster computing framework that uses implicit data parallelism and fault tolerance to provide a programming interface to entire clusters. For fast computing, Spark is capable of both batch and stream processing.
- Tableau is a data analytics platform that lets you prepare, analyze, collaborate, and share your big data insights. Tableau excels at self-service visual analytics, enabling people to ask new questions about managed big data and easily share those insights across the organization.

CONCLUSION

Since the problems of processing and analyzing large and complex input data are always present in data analysis, several effective analysis methods have been presented to reduce the memory cost to speed up the calculation time. In addition to the well-known improved methods for these analysis methods , a large number of studies have developed their own efficient methods based on mining algorithms or specific characteristics of the problem. These types of improved methods are usually designed to address the shortcomings of mining algorithms or to apply different methods to solve the mining problem. Such situations can be found in many association rules and sequential sample form problems, since the original assumption of these problems is to analyze large data sets. Because the previously common template algorithm had to scan the entire dataset many times, it was computationally

expensive. How to reduce the number of scans of the entire data set to save computational costs is one of the most important things in all frequent case studies. A similar situation exists in data clustering and classification studies.

References:

1. https://www.zippia.com/advice/big-data-statistics/
2. https://doi.org/10.1016/j.dim.2023.100038
3. https://doi.org/10.1016/j.lrp.2022.102290
4. https://explodingtopics.com/blog/big-data-stats
5. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0030-3#ref-CR20
6. https://www.statista.com/statistics/871513/worldwide-data-created/
7. http://dx.doi.org/10.3390/info11010017
8. Akhtar SMF (2018) Big Data Architect's Handbook, Packt
9. https://doi.org/10.1016/j.techfore.2023.122402