Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2019, 6(8):51-55

Research Article ISSN: 2394 - 658X

Exploring Large Datasets: Techniques for Managing and Using Data Efficiently

Chakradhar Avinash Devarapalli

Software Engineer
*avinashd7[at]gmail.com

ABSTRACT

With the evolution of the digital world, more and more data is being generated which is making a commodity. Most of this data is still raw and not being utilized to date. In other words, only a limited amount of data is being utilized and there is a need for a proper system that allows interesting individuals and organizations to efficiently convert this data into meaningful form where insights can be generated. There are numerous challenges associated with the use of this data which mainly include, storage limitations, displaying data properly, scalability issues, security concerns, and performance of algorithms based on provided data. However, the latest developments are allowing the participants to efficiently handle large datasets by employing appropriate techniques. This writing presents appropriate solutions to the challenges of managing large datasets in addition to the latest techniques. Therefore, using these solutions and techniques, one can store, retrieve, and process large datasets more easily.

Key words: Data Handling, Dataset, Large Datasets, Data Visualization, Big Data, Data Management

---

INTRODUCTION

The data is the new oil in this digital age of a fast-moving world. It is mainly because it helps in better decision-making in various fields ranging from health care to finance. The evolution of the online world provided a large amount of data as almost every field is digitalized making its relevant data available to the service providers [1]. The large datasets are helpful as they provide the aggregation of related information. These datasets may contain text files, images, and videos. Therefore, a large space may be needed for the storage of this big data before processing it. Efficient management is needed to properly store and retrieve data to get useful insights that can help in making more informed decisions.

Moreover, there are certain challenges associated with handling large datasets. Specialized tools and techniques such as parallel computing and distributed computing are needed to efficiently process the data after successfully storing it. These valuable techniques help organizations to get more interesting insights and therefore make more informed decisions. Most businesses are dependent on end-user data to enhance customer experience and improvement of products or services [2].

The article is focused on providing the appropriate techniques needed to manage large datasets. It is focused on providing solutions to the problems being faced in the industry in terms of managing large amounts of data efficiently. The objective is to equip individuals as well as organizations with appropriate methods to save efforts in terms of cost and time. The factors under consideration are volume, complexity, velocity, and variety of data. The goal is to minimize the challenges and present appropriate techniques for large datasets that have increased volume and are more complex.

LITERATURE REVIEW

The advancement in the field of data management is an ongoing process as different researchers are contributing to better utilize the largely available data. The researchers are mainly focused on suitable compression techniques and visualization tools for complex data [3]. The visualization of data is significant as proper insights can only be extracted when it is displayed properly without missing important information. The compression techniques help in decreasing the size of large datasets without losing important attributes which allows the data scientists to access the data graphically.

Distributed and cloud computing played a vital role in the advancement of data handling. These techniques have provided more efficient processing of data and storage facilities [4]. Cloud computing services made it possible for small business owners to store, retrieve, and process data using online services like Azure and AWS. It also allows to reduce the extra costs by paying for the shared services that are being managed by the dedicated organizations.

PROBLEM STATEMENT

Despite the regular efforts in the field of data handling, it is still challenging to manage large datasets. The ordinary techniques are only helpful in case of managing a limited amount of data but these contribute less when the size and complexity of data increases. The efficiency and performance of algorithms are dependent on the quality of data and therefore appropriate strategies are needed to manage large datasets. Additionally, solutions are needed for the challenges standing in the way of handling big data.

CHALLENGES IN LARGE DATASETS

4.1 Irregular Data

The given data can have missing values or inconsistent values which are too difficult to be identified with regular techniques [5], [6]. These errors in the data may disrupt the performance of the algorithm.

4.2 Maintaining Quality

Data quality should not be compromised while applying pre-processing operations as these techniques can lead to the loss of useful information. This can in turn generate inaccurate results and unexpected outcomes as the algorithms are sensitive to data on which they are trained.

4.3 Data Access

The data can be huge in volume, complex in structure, and can have variance which is why it is challenging to regularly access it. This process is time and resource-consuming. The data transfer between the nodes can be slow and sensitive. Therefore, compression and transferring protocols are needed.

4.4 Storage Limitations

Large firms can have their hardware-accelerated systems like large servers to meet their storage requirements but small organizations cannot afford the physical systems because it requires special infrastructure. The cost of storage is the biggest factor and therefore an alternative way is needed for comparatively small organizations.

4.5 Data Visualization

Displaying data can be challenging when it is equipped with a complex structure and large size. The more complex structure requires more appropriate tools for displaying data. Moreover, the problems like visual noise, and information loss can occur while displaying data for the insights.

4.6 Scalability

The presented system as a solution to data handling should be capable of handling more extended datasets in future use. Problems like memory limitations and computational complexity may occur in the future.

4.7.Data Security

The data needs to be protected from both internal and external attacks. The datasets may have various types like structured, semi-structured, and unstructured data. This makes the task even more challenging.

4.8.Performance

The data preprocessing techniques need to be efficient enough to not suffer from common problems. To operate on large files, the required data needs to be loaded from the memory and it in turn decreases the speed when the size of the data increases.

OVERCOMING CHALLENGES

5.1.Cloud Computing

Cloud computing is the most suitable for small organizations unable to afford the complete infrastructure for data storage [4]. Large organizations like Amazon Web Services (AWS) and Azure provide the services to analyze large datasets.

5.2.Data Management Policies

The standard rules can help in data management for maintaining the overall quality of data and applying standard security measures. The list of policies contains data collection, data storage, data quality, security, compliance, and data governance.

5.3.Data Visualization

Data visualization can be problematic but it can be easier by using the data correctly. The insights from the given data can be drawn with techniques like sampling to visualize samples randomly for a generalized view, aggregation of data to reduce data points, and use of interactive tools to automatically explore the subsets of large datasets. The tools like heatmaps can help to easily understand data and extract useful insights visually.

5.4.Data Management Platforms

For the efficient handling and analysis of large datasets, platforms like Hadoop, Apache Spark, and NoSQL can be helpful. Hadoop Distributed File System for data storage helps in parallel processing. Apache Spark is for big data analytics as it offers implicit memory for data processing. The NoSQL database can handle structured to unstructured data and offer diversity in data types.

5.5.Data Encryption and Access Control

To avoid the security concerns, encryption assists in restricting the unauthorized access. The technique of Advanced Encryption Standard is used to secure data. The access and manipulation of data can be controlled with methods like Role-based and Attribute-based Access Controls for large datasets.

TECHNIQUES FOR DATA HANDLING

Apart from the above solutions, the following techniques can be applied for efficient data handling [7].

6.1.Data Preprocessing

Cleaning and transformation of data before its actual use help in the analysis and avoid wastage of resources in later stages [8]. The common preprocessing techniques are data cleaning, handling missing data, feature scaling, managing outliers, encoding, dataset splitting, normalization, handling unbalanced data, and dimensionality reduction. The methods for improved data preprocessing are:

6.1.1.Sampling

Instead of a complete dataset, a sample of data can be used for initial analysis to save the computations. This speeds up the process but increases the bias factor [9].

6.1.2.Parallel Processing

For improved performance, the data processing tasks can be divided into chunks to simultaneously process them.

6.1.3.Feature Engineering

The use of relevant features increases the efficiency of the algorithm by employing the techniques of dimensionality reduction, grouping, and normalization.

6.1.4.Quality Checks

Regular checks with the use of automated tools can assist in finding the accuracy of the dataset.

6.2.Distributed Computing

Distributed file systems like Apache Spark and databases like NoSQL can help store large datasets. It allows the processing of big data in a cluster which helps in better performance. The fault tolerance is improved when a large dataset is distributed across clusters. Thus, the time in processing activities can be saved for distribution.

6.3.Data Indexing

The indexes can help in quick data retrieval operations. The frequently accessed queries can be indexed like the example code below:

```
db.StudentCollection.createIndex({ student_id: 1 }) // Creating index
db.StudentCollection.updateOne(
 { student_id: 50 }, // Filter
 { $set: { student_id: 51 } } //Update
```

```
)
```

6.4 Data Compression

The data compression technique can help reduce the same data size without losing useful information. It also helps in transferring data as less size allows to speed up the process. The compression techniques such as gzip and LZ4 can effectively reduce the size and make it easy to manage the storage of large datasets.

6.5 Data Partitioning

```
%%time
#reading directly
dataset_original=pd.read_csv('dataset.csv')
#reading in chunk
data_chunk=pd.read_csv('dataset.csv',chunksize=100000)
```

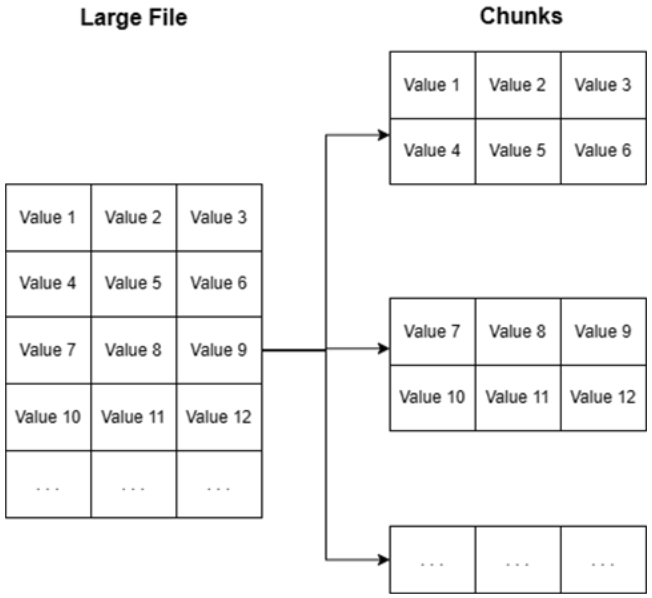

Figure 1: A large CSV File is divided into Chunks

RESEARCH IMPACT

This writing presents appropriate techniques to handle large datasets more efficiently. The presented solutions against the possible challenges of managing large amounts of data can assist in effective data handling. It has the potential to facilitate the industry's efficient management of large datasets. The organizations will be able to reduce costs on storage, improve processing speed, and increase accessibility of data. Responsible individuals can easily visualize, process, store, and manage large data with the use of appropriate techniques. Thus, decision-making will be improved for both individuals and organizations.

FUTURE DEVELOPMENT

Even with the advancement of technology, only a limited amount of information from the largely available is being utilized in today's world. The main reason for this is the lack of effective tools and techniques to handle large amounts of data. Although, the current techniques are playing a vital role in this field but future potentially holds in more optimized data compression algorithms, efficient visualization, and enhanced data processing. This is an ongoing evolution and each day is collaborating with effective strategies from researchers to extract more insights from widely available data. The future trends in the analysis of large datasets are Edge Computing, Artificial Intelligence, Blockchain Integration, and the Internet of Things.

CONCLUSION

As a cessation, the digital world is running based on data, and more optimized algorithms are being designed to get the most out of this vast information available. However, the extraction of useful insights is not only

dependent on efficient algorithms. These algorithms are useful only if the data being fed meets the requirements and is consistent. So, there is a need for efficient management of large datasets to make more informed decisions in this era of evolution.

The presented solutions against the challenges faced during data management will contribute to the progress of the industry. The given appropriate techniques when properly applied can assist both individuals and organizations to generate optimized results from the data effectively

REFERENCES

[1]. A. Siddiqa, I. A. T. Hashem, I. Yaqoob, M. Marjani, S. Shamshirband, A. Gani and F. Nasaruddin, "A survey of big data management: Taxonomy and state-of-the-art," *Journal of Network and Computer Applications,* vol. 71, pp. 151-166, Aug. 2016.

[2]. J. Spiess, Y. T'Joens, R. Dragnea, P. Spencer and L. Philippart, "Using big data to improve customer experience and business performance," *Bell Labs Technical Journal,* vol. 18, no. 4, pp. 3-17, Mar. 2014.

[3]. M. H. Gross, L. Lippert and O. G. Staadt, "Compression methods for visualization," *Future Generation Computer Systems,* vol. 15, no. 1, pp. 11-29, Feb. 1999.

[4]. C. Yang, Q. Huang, Z. Li, K. Liu and F. Hu, "Big Data and cloud computing: innovation opportunities and challenges," *International Journal of Digital Earth,* vol. 10, no. 1, Nov. 2016.

[5]. C. M. Sanders, S. L. Saltzstein, M. M. Schultzel, D. H. Nguyen, H. S. Stafford and G. R. Sadler, "Understanding the Limits of Large Datasets," *Journal of Cancer Education,* vol. 27, pp. 664-669, Jun. 2012.

[6]. S. Kotsiantis, D. Kanellopoulos and P. Pintelas, " Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering,* vol. 30, 2006.

[7]. Z. Baraka, "Opportunities to manage big data efficiently and effectively," Dublin Business School, 2014. [Online]. Available: http://hdl.handle.net/10788/2275. [Accessed 18 May 2019].

[8]. W. A. Malik and V. D. Hut, "Data Cleaning of Large Datasets: New Methods and Techniques," Germany, 2010.

[9]. C. Ferrari, G. Foca and A. Ulric, "Handling large datasets of hyperspectral images: Reducing data size without loss of useful information," vol. 802, pp. 29-39, Oct. 2013.