

LARGE & COMPLEX DATA STREAMS USING BIG DATA

Surabhi Swarnkar¹, Virendra Singh²

¹M. Tech Scholar Dept. of Computer Science Engineering, Babulal Tarabai Institute of Research and Technology Sagar, M.P. India

²Asst. Prof. Dept. of Computer Science Engineering, Babulal Tarabai Institute of Research and Technology Sagar, M.P. India

Abstract - The emerging large datasets have made efficient data processing a much more difficult task for the traditional methodologies. Invariably, datasets continue to increase rapidly in size with time. The purpose of this research is to give an overview of some of the tools and techniques that can be utilized to manage and analyze large datasets. We propose a faster way to catalogue and retrieve data by creating a directory file – more specifically, an improved method that would allow file retrieval based on its time and date. This method eliminates the process of searching the entire content of files and reduces the time it takes to locate the selected data. We also implement the nearest search algorithm in an event where the searched query is not found. The algorithm sorts through data to find the closest points that are within close proximity to the searched query.

We also offer an efficient data reduction method that effectively condenses the amount of data. The algorithm enables users to store the desired amount of data in a file and decrease the time in which observations are retrieved for processing. This is achieved by using a reduced standard deviation range to minimize the original data and keeping the dataset to a significant smaller dataset size.

1. INTRODUCTION

Data has attained the form of continuous data streams rather than finite stored data sets, posing barriers to users that wish to obtain results at a preferred time. Data prescribed in this manner displays no bounds or limitations; thus, a delay in the retrieval of data can be expected. For this reason, a search for the selected data in vast amounts of unsorted data is a time-consuming process. Furthermore, the size of the data itself becomes part of the problem.

Generally, three main issues are involved in the data retrieval process. First, is the decision as to which type of information is worth retrieving? In essence, the application must be able to differentiate between relevant data and data that are not essential to the user. In other words, the frequency of a term is not enough to infer the quality of the file that contains it. However, the best way to accomplish this goal is by pulling information that follows the user's preference. The second issue that is involved in the data retrieval process is the methods that will be used to acquire that data. When deciding on the methods that one can use to retrieve data, time must be considered. Indeed, effective data

retrieval methods are in demand because scientists are in need of algorithms that are sensitive to time. In addition, the relevance of data rests on accessing it on time. The third issue that must be accounted for lies in the methods that will be used to analyse the data. Data must be analysed fairly quickly in order to provide users with the ability to pull data in a reasonable amount of time.

2. BIG DATA PROCESSING

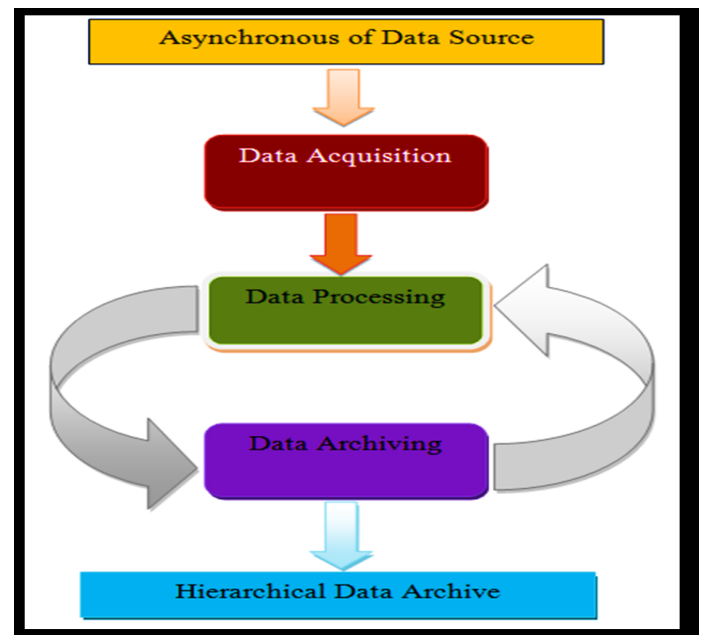


Fig -1: Data Processing

Data has to be effectively processed in order to convert raw data into meaningful information. See Figure below for details.

Converting raw data into an easily usable form involves a great deal of data processing (Member, Ortega & Shen, 2010). Computers conduct data processing, which accepts the raw data as input and then provides information as output. The systems that perform this task are a vital component of satellite operations.

3. RESULTS

Often, many scientists and researchers find themselves combating vast amounts of data without an effective or a fast

algorithm to process the data efficiently and in a timely manner. However, scientists are still required to effectively acquire desired results in a reasonable amount of time to conduct their studies properly. In this chapter, we provide examples of data that are constantly used by scientists as well as results of the proposed algorithm.

The graph below shows the instances in which a voltage dropped below a desired threshold. More specifically, on 09/08/31 at 22:00:00, all of the voltages are below the threshold; however, from 09/08/31 23:00:00 to 09/09/01 02:00:00, few of the voltage readings are below the threshold.

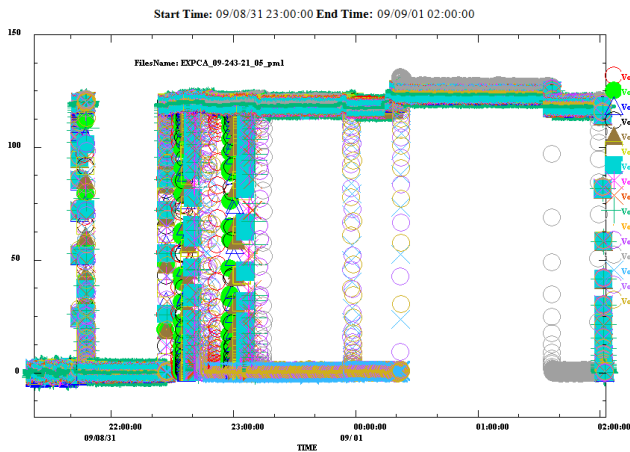


Fig -2: Diagram Shows the Voltages That Are Below the Threshold

Figure 2 shows how actual data is plotted and the specific voltage readings. However, detecting anomalies is often difficult without an improved algorithm that is used to manage such data as seen in Figure 2, Figure 3 shows the voltage measurements. Based on this graph, there is an indication that some of the measurements are below a minimum. This specific diagram reveals that voltages 14, 15, and 16 are out of family.

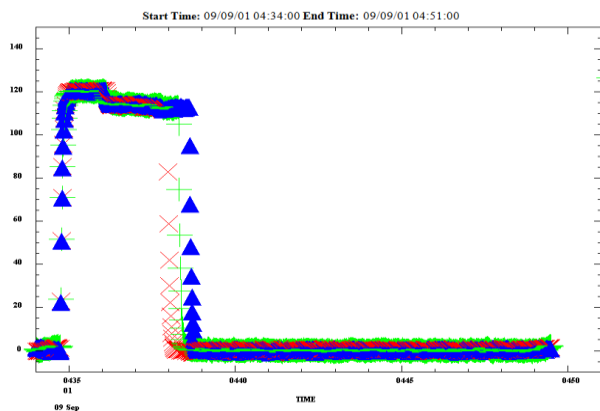


Fig -3: This Particular Data Exemplifies a Sample Of Out Of Family Ranges for a Typical

Spacecraft

The detection of these types of anomalies can easily be located and can be made easily retrievable by using the proposed methods. Figure 4 is an example where such methods can be effective.

Figure 4 is an example of a plotted data, which shows nearly a year's worth of data. In this plot, a problem in September 1999 can be seen. The temperature showing for this reading is out of range. However, it is unclear which data file in September 1999 has the problem. A whole year's worth of data would have to be plotted to find this anomaly. Instead, we can use the algorithm that we have proposed which would allow researchers to search in the directory file by time and locate the file that has the problem.

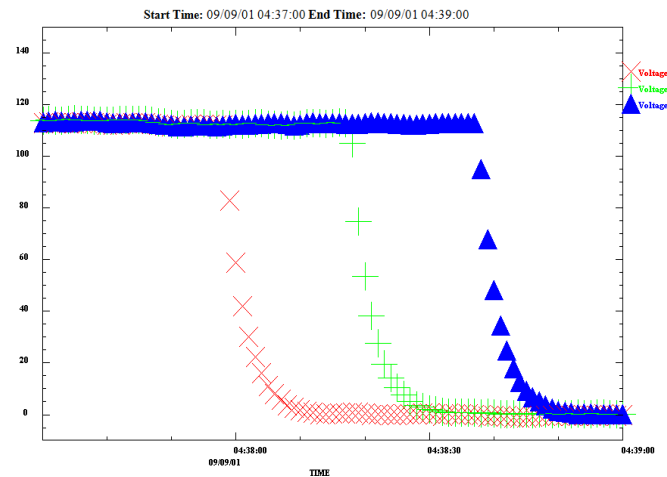


Fig -4: An Expanded View of the Data That Contains More Specific Detail than Figure 3.

The graph below shows the September 1, 1999 file data. It is evident that in this particular file lies voluminous data which makes it problematic for the user to obtain the anomaly.

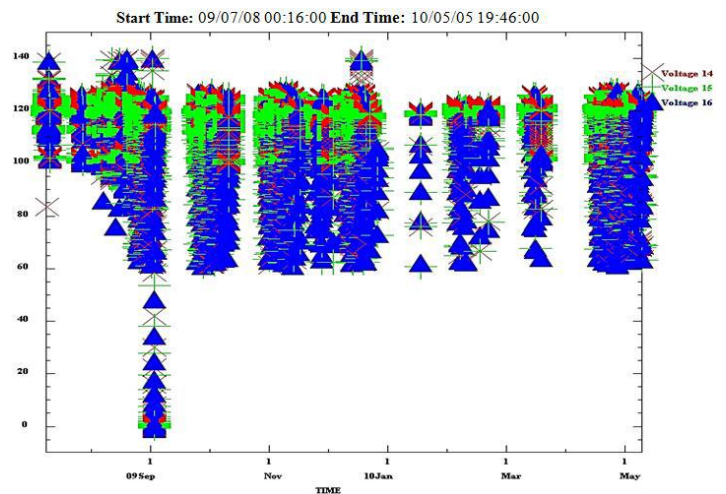


Fig -5: The Data from the Sept. 1, 1999 File Which Contains the Anomaly

It is understood that scientists must be able to analyze the generated data fairly quickly in order to ensure its integrity by identifying anomalies. One of the most common space anomalies is temperature variances. Most often scientists require global temperature readings to view the changes that transpired so that they can effectively identify temperature anomalies. Acquiring temperature anomalies is quite important; for instance, global warming is always measured using temperature anomalies. As another example, Figure 6. shows the temperatures variance as the annual and five-year running mean temperature changes with the base period 1951-1980.

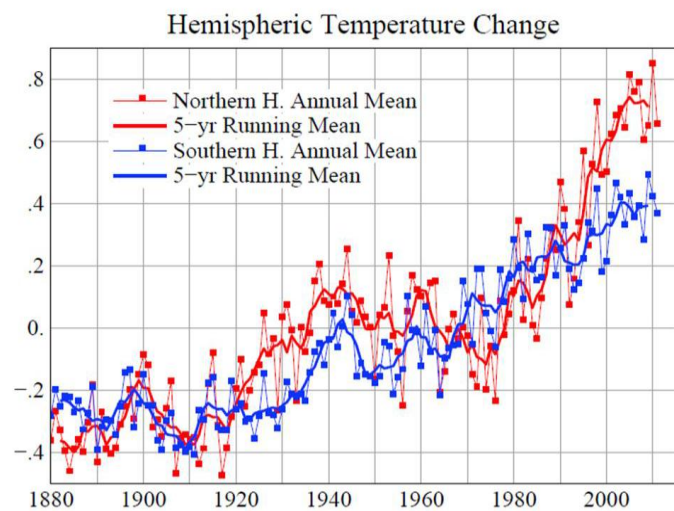


Fig -6: Hemispheric Temperature Change

This temperature analysis has been conducted by scientists at NASA's Goddard Institute for Space Studies (GISS). This plot shows that the average global temperature on earth increased by about 0.8°Celsius (1.4°Fahrenheit) since 1880. Two-thirds of the warming has occurred since 1975 at a rate of roughly 0.15-0.20°C per decade. Figure 7 shows an annual and five-year running mean surface air temperature in the contiguous 48 United States relative to the 1951-1980 mean.

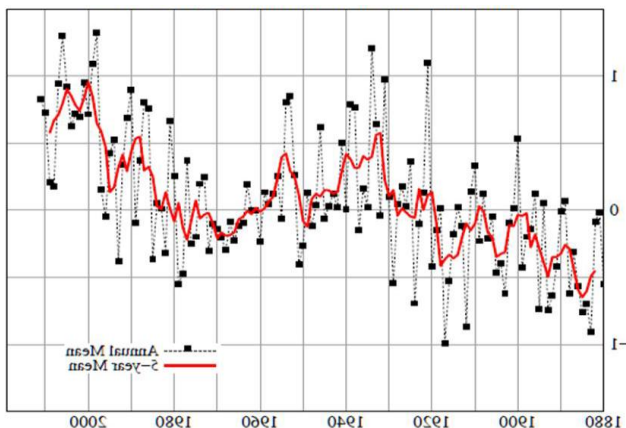


Fig -7: Temperature

Figure 7 shows temperature readings from 1880 to present with the base period 1951-1980. The black line is the annual mean and the red line is the five-year running mean. See Figure 8 to view the actual plotted data of the specified anomaly. In the early 1920's, there is an indication of a low average global temperature occurring between 1880 and 2013. High global temperature average has increasingly predominated, with the ratio now about two-to-one for the 48 states as a whole.

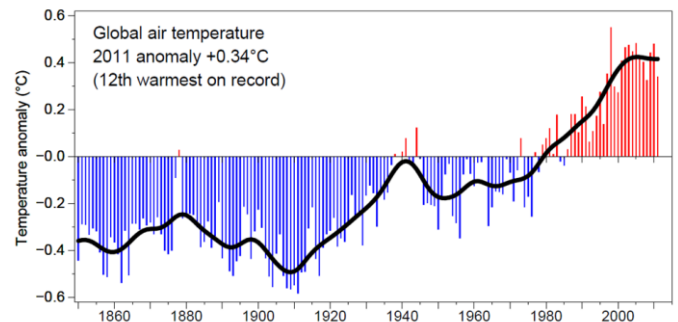


Fig -8: Temperature Anomaly Plot

Such tasks cannot be executed successfully without an advanced algorithm that is able to retrieve data in a significantly short amount of time. The relevancy of data lies with researcher's ability to access such results in a timely manner. Therefore, the proposed algorithm is quite suited for such a case because it is fast and efficient. The retrieval process of spacecraft data is very crucial in determining the accuracy of the results; however, this task was complicated by the extremely large amount of the available data. Nevertheless, the proposed algorithm is able to present accurate results even when operating with such complexities. Researchers now have the ability to locate the desired file from large datasets at a preferred time. The practical value of the proposed algorithm is presented through empirical results.

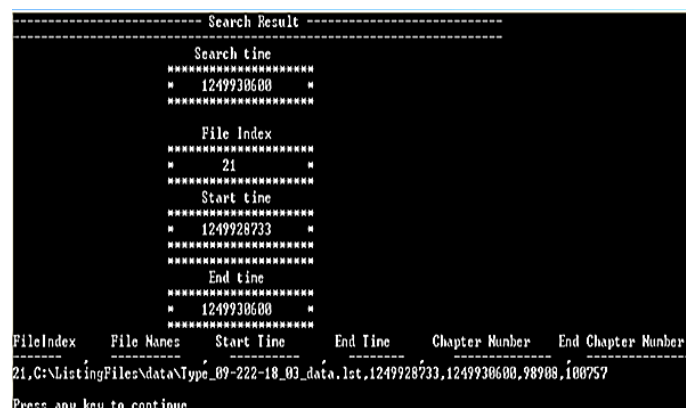


Fig -9: Result of Algorithm

This algorithm was able to locate the selected file from large datasets. Essentially, keying the desired time will provide the

user with the file index and its corresponding start and end time and the file name as shown in above figure. However, if the time is not found, then the algorithm locates a time that is within close proximity to the searched query, which is achieved with a nearest search algorithm.

4. CONCLUSIONS

The existing algorithms were carefully observed in order to identify their weaknesses. The sole purpose is to ensure that the proposed algorithm does not share the same weaknesses. One of the main and common weaknesses in the majority of the existing algorithms is the length of time required to search and retrieve the data of interest. Time is crucial to scientists and researchers; therefore, suitable methods to reduce the processing time are in high demand. Hence, this algorithm when compared to the existing algorithms has proven to be the most promising for scientists as well as other researchers in the field of computing. Because time is lessened and accurate results are presented, scientists can further investigate anomalies and arrive at a desirable outcome.

The suggested algorithm generates accurate and reliable results. The advantages/benefits of this algorithm:

- ✓ Data is extracted, transformed, and loaded onto the system at an ideal time.
- ✓ Stores and manages data in a database system effectively.
- ✓ Provides data access to researchers and scientists.
- ✓ Analyzes data by the application software.
- ✓ Presents data in a useful format.
- ✓ Condenses a large amount of data to a significant smaller size; typically by one to two orders of magnitude.
- ✓ Processes incomplete data effectively.
- ✓ Sorts through unsorted data.
- ✓ Fast paced and prompt access to data along with data retrieval processing techniques.

5. FUTURE ENHANCEMENTS

Future research can apply the methods presented in this dissertation to develop algorithms for searching and retrieving scientific data. The experimental and theoretical work presented will help researchers to develop a range of tools for searching, retrieving, and processing data. Due to a significant reduction of processing time achieved by the proposed algorithm, researchers can manage and obtain the desired data at a preferred time. This algorithm is not

limited to studies conducted by NASA or scientists in general. It can also be utilized in several data centers as well as in the medical field. For instance, in the field of medicine, the processing of medical data is playing an increasingly important role, e.g. computer tomography, magnetic resonance imaging, and so forth. These data types are produced persistently in hospitals and are increasing at a very high rate. Therefore, the need for systems that can provide efficient retrieval of medical data that is of a particular interest is becoming very high. The suggested algorithm can be utilized, in this case, to ease the burden of data retrieval and to assess the relevant data retrieval process. The algorithm can manage data in all of its aspects, including data in ASCII formats, binary codes, compressed data, uncompressed data, and so forth. Researchers who desire to advance this study should be aware of several suggestions that were noted in this particular study. The first is reducing space limitations. This algorithm uses dynamic arrays. The array occupies an amount of memory that is proportional to its size, independent of the number of elements that are actually of interest; therefore, determining ways to constraint this condition can further improve this study. Those that desire to further advance this study should also consider the type of systems that are compatible with conducting this study. In other words, this stipulates that the machine needed to test or perform these methods must be up to speed and updated to the current database system. The use of basic machines will not suffice and will not generate accurate results.

Optimizing directory of large data files using directory files to compress into multiple directories. For very large directory file jobs, using multiple machines to simultaneously build directory files on different portions of a data collection is generally much faster than creating directory files on a single machine. Splitting up the directory file creation job (parallel processing), is also a good strategy if disk space is insufficient to create the directory files all at once. Furthermore, in the future we surely need to expand this research to quantify the algorithm's time complexity using an analytical approach. We may also consider data in higher dimensions. The analytical approach seeks to reduce a system to its elementary components in order to study the system in detail and understand the types of interactions that exist between these components, we will perform additional tests for the proposed algorithms to demonstrate the, time complexity working with big, complex data.

Finally, index pre-processing is required for the implementation of these families of algorithms. The one-time offline file pre-processing overhead is a small tradeoff that significantly reduces the real-time search complexity by typically two orders of magnitude.

REFERENCES

- [1] Pulkit Sharma, Komal Mahajan, Dr. Vishal Bhatnagar "Analyzing Click stream Data using Hadoop" year 2016 IEEE
- [2] Sowmya R, Suneetha K R "Data Mining with Big Data" year 2017 IEEE
- [3] Beebe, R. (2010). Data Management, Preservation and The Future Of Pds.
- [4] Esfandiari, M., Ramapriyan, H., & Sofinowski, E. (2007). Earth Observing System (EOS) Data and Information System (EOSDIS Evolution Update and Future. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4423727
- [5] Query by image and video content: the QBIC system. Authoria, B., & James, B. (2005). Staffing Strategies: Can You Find, Recruit, and Retain the Talent You Need?
- [6] Bengio, S., & Poh, N. (2005). How Do Correlation and Variance of Base-Experts Affect Fusion in Biometric Authentication Tasks?
- [7] Abolhassani, M., Fuhr, N., G'over, N., & Großjohann, K. (2003). Content oriented XML retrieval with HyREX.
- [8] Benitez, A. B. (2002). Multimedia Knowledge Integration, Summarization and Evaluation.
- [9] Ashley, J., Dom, B., Flickner, M., Gorkani, M., Hafner, J., Huang, Q., Lee, D.,
- [10] Bhuptani, R., & Shettar, R. (n.d.). A Vertical Search Engine – Based on Domain Classifier.

BIOGRAPHIES



Surabhi Swarnkar is an M.Tech Scholar & currently researching on LARGE & COMPLEX DATA STREAMS USING BIG DATA. She is a Manager at BrainDesk company Sagar (MP).