BIG DATA ANALYSIS AND STORAGE OPTIMIZATION

Simran Syal[1], Ishdeep Singla[2]
[1]M.Tech Student, CGC Group of Colleges
[2]Assistant Professor, Chandigarh University
Department of Computer Science and Engineering
Gharuan, India.

Abstract: This paper explains the basic concept, management and trust on big data. The use of metadata in big data has been explained. An algorithm is proposed to analysis and optimization of storage in big data.

Keywords: Big data, Metadata, Storage optimization, and Data repositories.

I. INTRODUCTION

Big data is the data which is so bulky, diverse and messy that it's exponential growth and its availability makes it difficult to be processed using traditional database management tools and software techniques. Here, the data can be both structured and unstructured. Structured data is the data which is fixed in value. For example, numbers, dates, gender etc. Whereas unstructured data has various varieties like text, video, audio, posts etc.

The big data definition is incomplete without defining the four V's: Volume, Velocity, Variety and Veracity.

1. **Volume :** Volume is the scale or size of data. Big data is large in volume. It does not have a certain specified size of petabytes of data.
2. **Velocity :** Velocity is the analysis of streaming data. Velocity is also known as speed. Here, the speed not only refers to how fast the data is collected but also how fast you can analyze and utilize it.
3. **Variety :** Variety is the different forms of data. Big data has large variety of data which can be both structured and unstructured. Because big data use unstructured data and NoSQL additionally, many kinds of attribute could apply to the analysis for creating value [1].
4. **Veracity :**V eracity is the uncertainty of data. The uncertainty can be due to data inconsistency, incompleteness, ambiguity, and latency.

II. MANAGEMENT OF BIG DATA

The management of data is required in order to ensure whether big data is fit for certain purpose and for the business purpose. Big data [2] is about turning unstructured, imperfect, invaluable, complex data into useful information. Different types of data may require different governance. For example, machine generated data usually have lots of replication in data which needs to be normalized, sometimes data may be missing due to any failure like sensor failure or network connection failure for which certain activities like data masking or data cleaning is required.

The big data model implies that almost every type of information eventually can be derived from sufficiently large datasets [3].

Management and governance for different types of data include :

- Big data systems should be integrated with the relevant data that resides in environment. It is very often that big data resides in isolation. So, different approaches may be required in different situations so the integration environment needs to be flexible.
- Data should be trustworthy. This means data must be secure as well as there must be data privacy. Data should be reliable based on which decision making can be done.
- Data should be suitable in context.

III. BIG DATA TRUST

If the data is not trustworthy, the decisions made will be biased or ineffective. Machine generated data is not usually prone to errors until or unless there is any failure. Various modern devices can detect the fault arisen so that the result generated can be ignored. However, machine generated data is subject to replication as in case of unstructured data. So, the data needs to be of good quality.

In recent years, big data security research has been actively conducted through data loss protection or access control, etc. [4]. But data management and classification for security are more difficult than current information security due to data volume [5].

IV. NEED FOR METADATA IN BIG DATA

Data provenance means recording information about data at its birth. This can be achieved through metadata. In general form, it means how data is related to the existing sources of data. Big data should generate the right metadata through tells how data is recorded, what data is recorded and how the data can be measured.

V. BIG DATA ANALYSIS AND STORAGE OPTIMIZATION

As the development of human race progresses, there is a large explosion of data day to day which needs to be maintained and stored properly and efficiently. The development of high processing computers and availability of large datacenters have led to faster computability and high end storage functionality but there is an urge to develop an algorithm to maintain and efficiently store the much larger data sets in day to day life. The storage budget for IT administrations remains flat, if not decreasing, and hence it is strongly desirable and imperative to design storage management solutions that are both effective and cost-effective [6].

Various organizations are facing the problem of dealing with the exponential growth of data. In order to cope up with this problem, the organizations need to minimize the volume of data being stored and exploit new techniques which can further improve performance and storage utilization [7].

The algorithm needed to be developed for analyzing the large incoming data should pass through the following general flow stream :

a) Data Input

The data does not come out of anywhere. It is taken from a data generating source. For example, consider the ability to sense and observe the world, from the heart rate of an older citizen, and existence of toxins in the air inhaled, to a specific square kilometer array telescope, that will produce up to 1 million terabytes of raw data each day [8]. In the same way, petabytes of data is generated from the scientific experiments and simulations. The data which is of no interest can be filtered and compressed by orders of magnitude.

b) Filtering out information

The data taken directly from generating source may not be in the format in which it is required for analysis. For example, consider the range of electronic health records in a hospital that comprises transcribed dictations from different physicians, structured data from sensors and measurements which possibly have some associated uncertainty [8]. Thus, the data cannot be left as such and be effectively analyzed. Rather some adequate filtering process is needed that takes the required information from the raw collected data and store it in a structured form that is suitable for analysis.

c) Structuring the data

For the various forms of the large amount of data, it is not only enough to record the data and store it into a repository. Consider, for example, data collected by a range of scientific experiments [8]. If there are only a few groups of data sets in a repository, it is unusual that anyone will ever be able to find and use any of this data. However, it is somehow possible in the presence of some adequate metadata but still there will be challenges. Data analysis is more challenging than simply understanding, identifying, locating, and citing data. For large-scale analysis, this has to be done in a completely automated manner. There are many ways in which the same information can be stored. Some designs may have advantages over others due to some purposes, and possibly certain disadvantages for other purposes.

d) Implementing faster query methodology

The difficult part posed while querying when dealing with big data is fast query methodology which can be done only when the starting part of analysis is done structurally and in a particular fashion keeping in mind of the later fast query implementation. The way to do a faster query is described later in this paper which is taken care of in this algorithm. The query should be done in such a way that no part of important data is left out in the result and the query is done in minimal time. This is a very big challenge in big data world.

e) Interpretation

There can be many important parts of data which can be left out while filtering out the important data from the raw incoming data. As such, interpretation of data to bring out the important data out of it is very important in analysis phase. A detailed research has to be done while structuring out the data sets for a particular field in a particular sector. Hence, the analytics differ from one sector to another like the healthcare sector will have different filters for data than weather reporting field of scientific sector.

Below is the schematic representation of flow of algorithm for big data analytics :

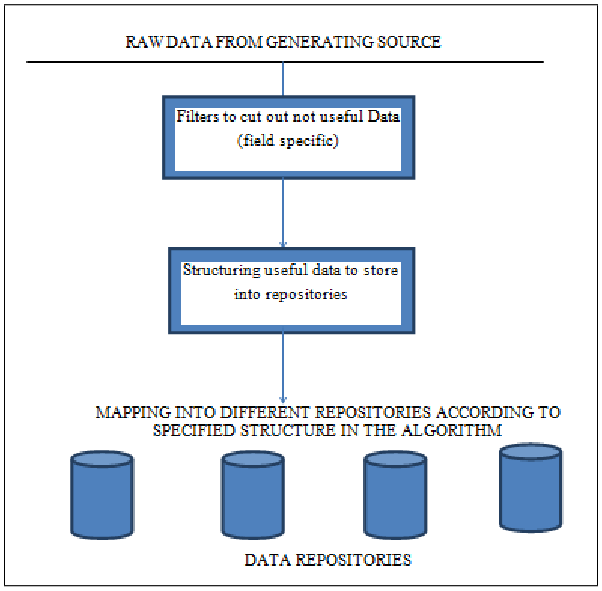
RAW DATA FROM GENERATING SOURCE
Filters to cut out not useful Data (field specific)
Structuring useful data to store into repositories
MAPPING INTO DIFFERENT REPOSITORIES ACCORDING TO SPECIFIED STRUCTURE IN THE ALGORITHM
DATA REPOSITORIES

Figure 1: Flow of algorithm for big data analytics

VI. ALGORITHMIC PERSPECTIVE

There are various sectors in which big data can be applied and each sector should have a different filter algorithmic perspective according to its needs some of which is discussed below :

1) Healthcare Sector

Taking the general example of collecting heartbeat data of a patient. The data is continuous in nature coming from the heart monitor and it is very crucial to note the heartbeat nature of patient for analysis by surgeons or physicians. However, continuous note of patient's heartbeat would be waste of data space. Instead, a filter can be applied to curb out the unnecessary data.

```
void filter(String input)
{
Read input
Condition:
Collect data when heartbeat changes and note down the time of changed heartbeat and heartbeat data.
Filters for above condition :
i) Note data for 72 >= heartbeat < 80
ii) Note data for 80 <= heartbeat < 90
iii) Note data for heartbeat >= 90
iv) Note data for 60 <= heartbeat < 72
v) Note data for 50 <= heartbeat < 60
vi) Note data for heartbeat < 50
Write filtered data storage repository (choose repositories randomly meant for storage)
}
```

Choosing a random repository will enable a security base so that if anyone creeps into anyone of the data storage, won't be able to crack out the data.

The data will store into something following manner :

Suppose, the data input is :

Patient name : Ravi Kumar

Heartbeat monitor status after converting to digital format noted per second :

Heartbeat	Time
76	2:01 p.m.
76	2:02 p.m.
73	2:03 p.m.
73	2:04 p.m.
73	2:05 p.m.
73	2:06 p.m.
73	2:07 p.m.
73	2:08 p.m.
73	2:09 p.m.
65	2:10 p.m.

After running the algorithm, the data that will go into repositories will be :

Ravi	Heartbeat	Time
	76	2:01 p.m.
	65	2:10 p.m.

As it can be seen that a lot of space is saved by applying the filters and hence the data storage problem can highly be resolved by applying this algorithm.

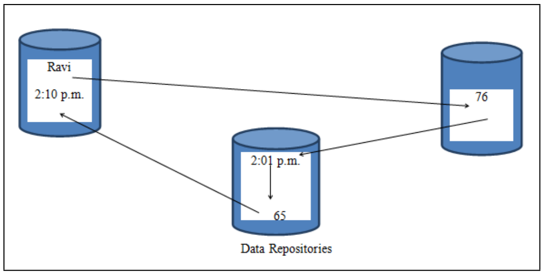

Figure 2: Pictorial representation of mapping of data on data repositories

However, the question arises as to how to retrieve back the data when queried as the query will take a lot of time by applying search in each and every data repository. The answer is simple to this randomness and it will require just another structuring of data which will enhance the query search and hence faster search.

Now the data is filtered and discrete data sets are obtained. A hash method on each of the data can be applied as follows :

```
int hash(char* data)
{
Convert data into some specified mapped value
}
void Store(int hash)
{
Store the data into corresponding data repository to which hashed value points to.
}
```

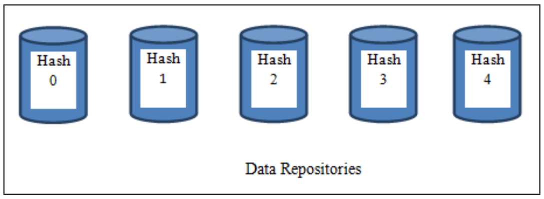

Figure 3: Naming conventions of Data repositories.

Since each repository is marked by a particular hash value, hence the data when computed upon by hash function will point to a particular repository and gets stored in it together with the information of its next data location.

Example: Ravi will hold the information of location of 76 and so on till the recording is stopped.

This will facilitate faster querying. Since when the patient's information is queried, the first data is hash data and is to be looked for that data into known repository. After the first discrete data is found, the location of next data is known and hence next data can be retrieved in O(1) time.

2) **Scientific Sector**
Take the case of weather report collection. In this case, filters can be used like whenever there is change in a particular range of temperature or condition of weather like cloudy, rainy, hot, change in a particular gradient of humidity then there value and time can be noted and similarly the data can be stored and retrieved. The segregation of data can also be done in a structured manner. Also, astronomical data can also be filtered and structurally stored in a way like if there is some pictorial difference in the pictures being captured which can be identified by comparing the change in two pictorial data which is captured in form of 1s and 0s in memory.

3) **Financial Sector**
The increase or decrease in the foreign exchange can be noted down within a range by applying the specific filters by doing appropriate research and finding out the filter conditions which does not apply to this part of paper. It has been found that big data is giving effective results in the field of: healthcare, manufacturing, banking, insurance, government, science, natural resource, retail, public sector administration, personal-location data and other services [9][10][11].

VII. CONCLUSION

This paper, tries to solve two big challenges faced in big data. Also, an algorithm for Healthcare sector has been proposed which can also be used in other fields by doing little modifications particularly when dealing with filters.

References

[1] Mohania M Gupta R, Gupta H. Cloud computing and big data analytics: What is new from databases perspective? *In. Anonymous Big Data Analytics, Springer, 2012*, pages 42–61, 2012.

[2] Big data for development: Challenges and opportunities. *Global Pulse*, 2012.

[3] Jensen, Meiko. Challenges of privacy protection in big data analytics. *IEEE International Congress on Big Data*, 2013.

[4] Tankard, C. Big data security. *Network Security*, pages 5–8, 2012.

[5] Gantz, J. and Reinsel, D. Extracting value from chaos. *White Paper, IDC*, 2011.

[6] Song, Yang, Alatorre, Gabriel, Mandagere, Nagapramod, and Singh, Aameek,. Storage mining: Where it management meets big data analytics. *IEEE International Congress on Big Data*, 2013.

[7] Singh, Sachchidanand, and Singh, Nirmala. Big data analytics. *International Conference on Communication, Information and Computing Technology (ICCICT)*.

[8] Challenges and opportunities with big data. *White Paper*.

[9] Smith, Matthew, Szongott, Christian, Henne, Benjamin, Voigt, Gabriele Von. Big data privacy issues in public social media. *IEEE*, 2013.

[10] Big data: The next frontier for innovation, competition, and productivity. *McKinsey and Company*, 2011.

[11] 2012 big data survey results. *Treasure Data*, 2012.