IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

Fundamentals of Big Data-A Survey

Chirag[1], Shivam[2]
B.Tech, Department of CSE
Dronacharya College of Engineering, Gurugram, India

Abstract: *The age of big data is now coming. But the traditional data analytics may not be able to handle such large quantities of data. The question that arises now is, how to develop a high-performance platform to efficiently analyse big data and how to design an appropriate mining algorithm to find the useful things from big data. To deeply discuss this issue, this paper begins with a brief introduction to data analytics, followed by the discussions of big data analytics. Some important open issues and further research directions will also be presented for the next step of big data analytics.*

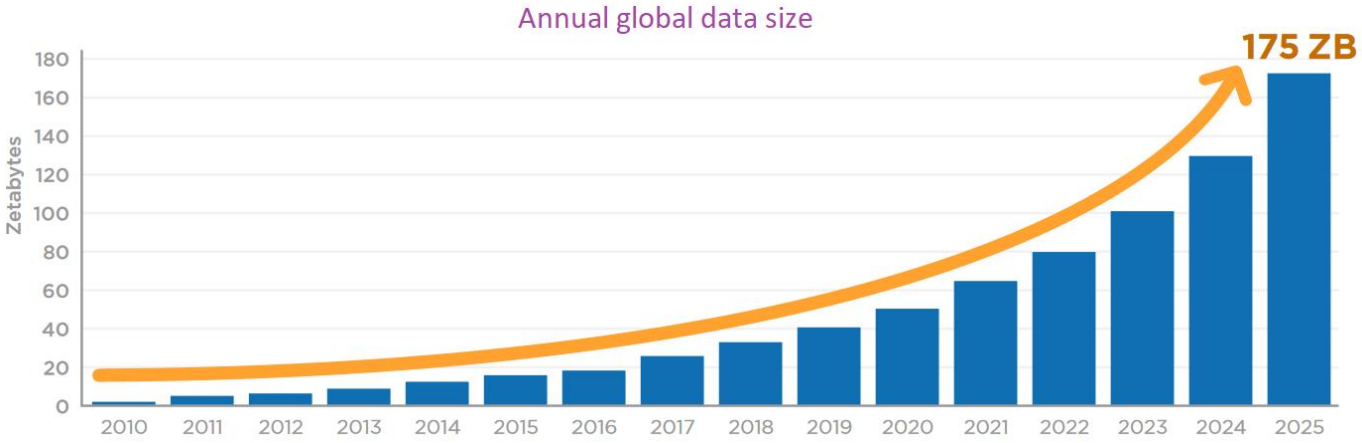
Keywords: Big data.

I. INTRODUCTION

Imagine a world without data storage; an area where every detail a couple of persons or organization, every transaction performed, or every aspect which may be documented is lost directly after use. Organizations would thus lose the power to extract valuable information and knowledge, perform detailed analyses, furthermore as provide new opportunities and advantages. Anything starting from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is that the building block upon which any organization thrives.

Now think about the extent of details and therefore the surge of knowledge and knowledge provided now a days through the advancements in technologies and therefore the internet. With the rise in storage capabilities and methods of knowledge collection, huge amounts of knowledge have become easily available. Every second, more and more data are being created and needs to be stored and analysed so as to extract value. Furthermore, data has become cheaper to store, so organizations have to get the maximum amount value as possible from the huge amounts of stored data.

The size, variety, and rapid change of such data require a replacement kind of big data analytics, furthermore as different storage and analysis methods. Such sheer amounts of hugedata have to be properly analysed, and pertaining information should be extracted.



The contribution of this paper is to produce an analysis of the available literature on big data analytics. Accordingly, a number of the varied big data tools, methods, and technologies which might be applied are discussed, and their applications and opportunities provided in several decision domains are portrayed. Our corpus mostly includes research from a number of the highest journals, conferences, and white papers by leading corporations within the industry. because of long review process of journals, most of the papers discussing big data analytics, its tools and methods, and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

its applications were found to be conference papers, and white papers. While big data analytics is being researched in academia, several of the economic advancements and new technologies provided were mostly discussed in industry papers.

Here are some real-world examples of Big Data in action:

- Consumer product companies and retail organizations are monitoring social media like Face book and Twitter to get an unprecedented view into customer behaviour, preferences, and product perception.
- Manufacturers are able to monitor minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to replace or maintain. Replacing it too soon wastes money and replacing it too late triggers an expensive work stoppage.
- Manufacturers are also monitoring social networks, but with a different goal than marketers: They are using it to detect aftermarket support issues before a warranty failure becomes publicly detrimental.
- The government is making data public at the national level, state level, and city level for users to develop new applications that can generate public better.
- Financial Services organizations are taking data mined from customer interactions to slice and dice their users into finely tuned segments and enables these financial institutions to create increasingly relevant and sophisticated offers.
- Advertising and marketing agencies are tracking social media to Insurance companies are using Big Data analysis to see which home insurance applications can be immediately processed, and which ones need a validating in-person visit.
- Retail organizations are engaging brand advocates, changing the perception of brand antagonists, and even enabling enthusiastic customers to sell their products. All these things are doing by embracing social media.
- Hospitals predict those patients that are likely to seek readmission within a few months of discharge by analysing medical data and patient records. The hospital can then prevent another costly hospital stay.
- To offer more appealing recommendations and more successful coupon programs the, Web based businesses are developing information products that combine data gathered from customers.
- Sports teams are using data for tracking ticket sales and are using bigdata for tracking team strategies also.

II. DATA ANALYTICS

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.
Technologies such as business intelligence (BI) tools and systems help organizations take the unstructured and structured data from multiple sources. Users (typically employees) input queries into these tools to understand business operations and performance. Big data analytics uses the four dataanalysis methods to uncover meaningful insights and derive solutions.

2.1 5 V's of Big Data:

- **Volume:** the size and amounts of big data that companies manage and analyze.
- **Value:** the most important "V" from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits.
- **Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data.
- **Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time.
- **Veracity:** the "truth" or accuracy of data and information assets, which often determines executive-level confidence.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

The additional characteristic of variability can also be considered:

- **Variability:** the changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases.

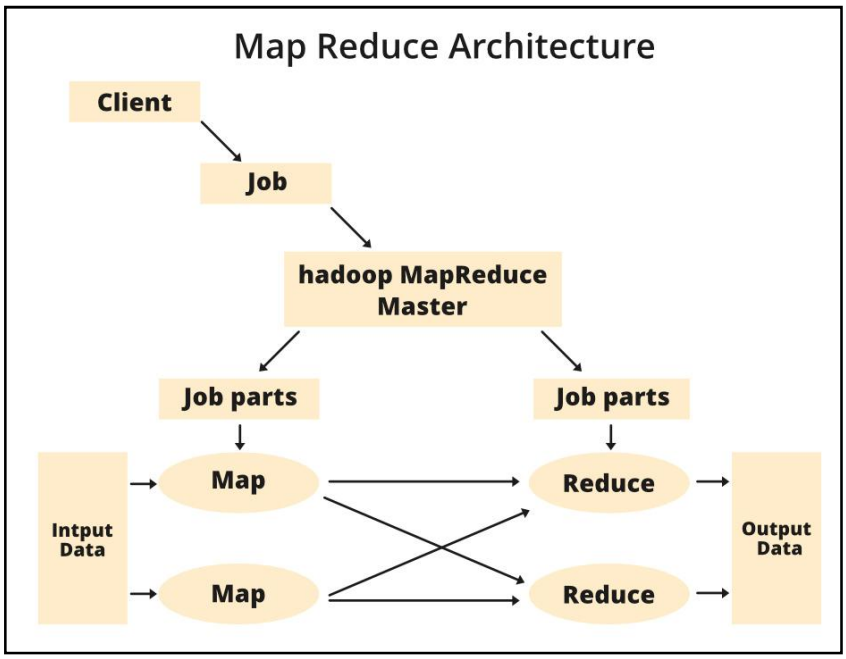
III. PARALLEL PROGRAMMING & MAP REDUCE

Data analysis software parallelizes fairly naturally. Many programmers are interested to building programs on the parallel model. The parallel research had the most success in the field of parallel databases. Rather than requiring the programmer to unknot an algorithm into separate threads to be run on separate cores, parallel databases let them break up the input data tables into pieces, and pump each piece through the same single-machine program on each processor. This "parallel dataflow" model makes parallel programming as easy as programming a single machine. And it works on "shared-nothing" clusters of computers in a data centre: The machines involved can communicate via simple streams of data messages, without a need for an expensive shared RAM or disk infrastructure.

Famous big data analysis tool is Hadoop. Apache Hadoop is an open-source software framework. it is written in Java for distributed storage and distributed processing of big data on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

The heart of Hadoop is MapReduce. It is this programming paradigm that allows for massive scalability across thousands of servers in a Hadoop cluster. It is useful for batch processing on petabytes or zeta bytes of data stored in Apache Hadoop. If we are familiar with clustered scale-out data processing solutions. Then the MapReduce concept is simple to understand. MapReduce programming model has twisted a new page in the parallelism story. The MapReduce framework is a parallel dataflow system that works by dividing data across machines. Each of which runs the same single-node logic. MapReduce asks programmers to write traditional code, in languages like C, Java, Python and Perl. In addition to its familiar syntax, MapReduce allows programs to be written to and read from traditional files in a file system, rather than requiring database schema definitions.

MapReduce refers to two separate and distinct tasks. The first is the job of map, which takes a set of data and converts it into another set of data. Individual elements are broken down into value pairs. The reduce job takes the output from a map as input and combines those data values into a smaller set of values. The reduce job is always performed after the map job. So, the sequence of the name MapReduce.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

IV. BEST BIG DATA ANALYTICS USE CASES

- **Sentiment Analysis:** Sentiment analysis offers powerful business intelligence to enhance the customer experience, revitalize a brand, and gain competitive advantage. The key to successful sentiment analysis lies in the ability to dig for multi-structured data pulled from different sources into a single database.
- **360-Degree View of Customer:** A 360-degree customer view offers a deeper understanding of customer behaviour and motivations. Obtaining a 360-degree customer review requires analysis of data from different sources like social media, data collecting sensors, mobile devices etc. From there, more effective micro-segmentation and real-time marketing are getting as result.
- **Ad Hoc Data Analysis:** Ad-hoc analysis only looks at the data requested or needed, providing another layer of analysis for data sets that are becoming larger and more varied. Big data ad-hoc analytics can help in the effort to gain greater insight into customers by analyzing the relevant data from unstructured sources, both external and internal.
- **Real-Time Analytics:** Systems that offer real-time analytics quickly decipher and analyze data sets, providing results even as data is being generated and collected. This high-velocity method of analytics can lead to immediate reaction and changes. It allows for better sentiment analysis, split testing, and improved targeted marketing.
- **Multi-Channel Marketing:** Multi-channel marketing creates a seamless across different types of media like company websites, social media, and physical stores. During all stages of the buying process multi-channel marketing requires an integrated big data approach.
- **Customer Micro-Segmentation:** Customer micro-segmentation provides more tailored and targeted messaging for smaller groups. This personalized approach requires analysis of big data collected through sources like customers' online interactions, social media etc.
- **Ad Fraud detection:**Ad fraud detection requires data analysis of fraud strategies by recognizing patterns and behaviours. Data that shows irregularity of group behaviour make it so ad fraud is find out and blocked before it is spread.
- **Click stream analysis:** Click stream analysis helps to grow the user experience by optimizing company websites, and offering better insight into customer segments. click stream analysis helps to personalize the buying experience, getting an improved return on customer visits with big data.
- **Data Warehouse Modernization:** Integrate big data and data warehouse capabilities to boost operational efficiency. Optimize your data warehouse to enable fresh types of analysis. Use big data technologies to set up a staging area or landing zone for your new data before formative what data should be moved to the data warehouse. divest infrequently accessed or aged data from warehouse and application databases using in sequence integration software and tools.
- **Big Data and Predictive Modelling:** The most common uses of big data by companies are for tracking business processes and outcomes, and for building a wide array of predictive models. Amazon and Netflix recommendations rely on predictive models of what book or movie an individual might want to purchase. Google's search results and news feed rely on algorithms that predict the significance of particular web pages or articles. Apple's auto- complete function tries to forecast the rest of one's text or email based on past convention patterns. Online advertising and marketing rely greatly on automated predictive models that aim individuals who might be particularly likely to answer to offers.

The application of predictive algorithms extends well ahead of the online world. In health care, it is now common for insurers to adjust payments and quality measures based on "risk scores," which are resulting from predictive models of human being health expenses and outcomes. An individual's risk score is naturally a weighted sum of health indicators that recognize whether an individual has different persistent conditions, with the weights chosen based on a statistical analysis. Credit card companies use predictive models of default and repayment to guide their underwriting, pricing, and marketing actions.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

V. TOP 6 BIG DATA CHALLENGES

1. **Lack of knowledge Professionals:** Companies need skilled data professionals to run these modern technologies and large Data tools. These professionals will include data scientists, analysts, and engineers to work with the tools and make sense of giant data sets. One of its challenges that any Company face is a drag of lack of massive Data professionals. This is often because data handling tools have evolved rapidly, but in most cases, the professionals haven't. Actionable steps got to be taken to bridge this gap.

   Solution: Companies are investing extra money in the recruitment of skilled professionals. They even have to supply training programs to the prevailing staff to urge the foremost out of them. Another important step taken by organizations is purchasing knowledge analytics solutions powered by artificial intelligence/machine learning. These Big Data Tools are often suggested by professionals who aren't data science experts but have the basic knowledge. This step helps companies to save tons of cash for recruitment.

2. **Lack of proper understanding of Massive Data:** Companies fail in their Big Data initiatives, all thanks to insufficient understanding. Employees might not know what data is, its storage, processing, importance, and sources. Data professionals may know what's happening, but others might not have a transparent picture. For example, if employees don't understand the importance of knowledge storage, they cannot keep a backup of sensitive data. They could not use databases properly for storage. As a result, when this important data is required, it can't be retrieved easily.

   Solution: Its workshops and seminars must be held at companies for everybody. Military training programs must be arranged for all the workers handling data regularly and are a neighbourhood of large Data projects. All levels of the organization must inculcate a basic understanding of knowledge concepts.

3. **Data Growth Issues:** One of the foremost pressing challenges of massive Data is storing these huge sets of knowledge properly. The quantity of knowledge being stored in data centres and databases of companies is increasing rapidly. As these data sets grow exponentially with time, it gets challenging to handle. Most of the info is unstructured and comes from documents, videos, audio, text files, and other sources. This suggests that you cannot find them in the database. Data and analytics fuels digital business and plays a major role in the future survival of organizations worldwide.

   Solution: Companies choose modern techniques to handle these large data sets, like compression, tiering, and deduplication. Compression is employed to reduce the number of bits within the data, thus reducing its overall size. Deduplication is the process of removing duplicate and unwanted data from a knowledge set. Data tiering allows companies to store data in several storage tiers. It ensures that the info resides within the most appropriate storage space. Data tiers are often public cloud, private cloud, and flash storage, counting on the info size and importance. Companies also are choosing its tools, like Hadoop, NoSQL, and other technologies.

4. **Confusion while Big Data Tool selection:** Companies often get confused while selecting the simplest tool for giant Data analysis and storage. Is HBase or Cassandra the simplest technology for data storage? Is Hadoop MapReduce ok, or will Spark be a far better data analytics and storage option? These questions bother companies, and sometimes they cannot seek the answers. They find themselves making poor decisions and selecting inappropriate technology. As a result, money, time, effort, and work hours are wasted.

   Solution: You'll either hire experienced professionals who know far more about these tools. Differently is to travel for giant Data consulting. Here, consultants will recommend the simplest tools supporting your company's scenario. Supporting their advice, you'll compute a technique and select the simplest tool.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

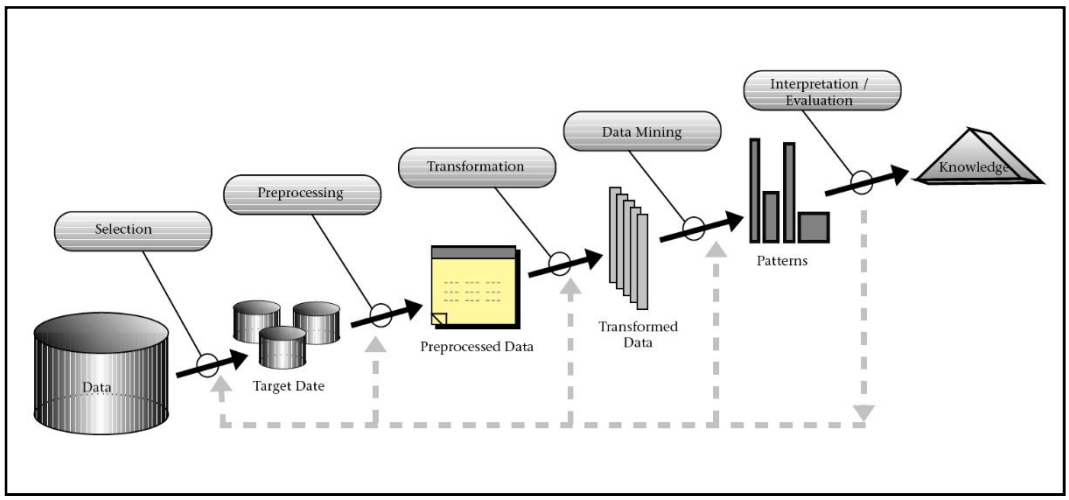
5. **Integrating Data from a Spread of Sources:** Data in a corporation comes from various sources, like social media pages, ERP applications, customer logs, financial reports, e-mails, presentations, and reports created by employees. Combining all this data to organize reports may be a challenging task. This is a neighbourhood often neglected by firms. Data integration is crucial for analysis, reporting, and business intelligence, so it's perfect.

   Solution: Companies need to solve their Data Integration problems by purchasing the proper tools. A number of the simplest data integration tools are mentioned below:
   - Talented Data Integration
   - Centerprise Data integrator
   - ArcESB
   - IBM InfoSphere
   - Xplenty
   - Informatica PowerCenter
   - Clover DX
   - Microsoft SQL QlikView

6. **Securing Data:** Securing these huge sets of knowledge is one of the daunting challenges of massive Data. Often companies are so busy understanding, storing, and analyzing their data sets that they push data security for later stages. This is often not a sensible move, as unprotected data repositories can become breeding grounds for malicious hackers. Companies can lose up to $3.7 million for stolen records or knowledge breaches.
   Solution: Companies are recruiting more cybersecurity professionals to guard their data. Other steps to Securing it include Data encryption, Data segregation, Identity, and access control, Implementation of endpoint security, and Real-time security monitoring. Use its security tools, like IBM Guardian.

VI. DATA MINING

Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data. The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

The KDD process in data mining typically involves the following steps:

1. **Selection:** Select a relevant subset of the data for analysis.
2. **Pre-processing:** Clean and transform the data to make it ready for analysis. This may include tasks such as data normalization, missing value handling, and data integration.
3. **Transformation:** Transform the data into a format suitable for data mining, such as a matrix or a graph.
4. **Data Mining:** Apply data mining techniques and algorithms to the data to extract useful information and insights. This may include tasks such as clustering, classification, association rule mining, and anomaly detection.
5. **Interpretation:** Interpret the results and extract knowledge from the data. This may include tasks such as visualizing the results, evaluating the quality of the discovered patterns and identifying relationships and associations among the data.
6. **Evaluation:** Evaluate the results to ensure that the extracted knowledge is useful, accurate, and meaningful.
7. **Deployment:** Use the discovered knowledge to solve the business problem and make decisions.

After the data mining problem was presented, some of the domain specific algorithms are also developed. An example is the apriori algorithm which is one of the useful algorithms designed for the association rules problem. Although most definitions of data mining problems are simple, the computation costs are quite high. To speed up the response time of a data mining operator, machine learning, metaheuristic algorithms, and distributed computing were used alone or combined with the traditional data mining algorithms to provide more efficient ways for solving the data mining problem. One of the well-known combinations can be found in, Krishna and Murty attempted to combine genetic algorithm and *k*-means to get better clustering result than *k*-means alone does.

As most data mining algorithms contain the initialization, data input and output, data scan, rules construction, and rules update operators.

```
1 Input data D
2 Initialize candidate solutions r
3 While the termination criterion is not met
4    d = Scan(D)
5    v = Construct(d, r, o)
6    r = Update(v)
7 End
8 Output rules r
```

Data mining algorithm

D represents the raw data, *d* the data from the scan operator, *r* the rules, *o* the predefined measurement, and *v* the candidate rules. The scan, construct, and update operators will be performed repeatedly until the termination criterion is met. The timing to employ the scan operator depends on the design of the data mining algorithm; thus, it can be considered as an optional operator. Most of the data algorithms can be described. It also shows that the representative algorithms—*clustering*, *classification*, *association rules*, and *sequential patterns*—will apply these operators to find the hidden information from the raw data. Thus, modifying these operators will be one of the possible ways for enhancing the performance of the data analysis.

Clustering is one of the well-known data mining problems because it can be used to understand the "new" input data. The basic idea of this problem is to separate a set of unlabelled input data to *k* different groups, e.g., such as *k*-means. Classification is the opposite of clustering because it relies on a set of labelled input data to construct a set of classifiers (i.e., groups) which will then be used to classify the unlabelled input data to the groups to which they belong. To solve the classification problem, the decision tree-based algorithm, naïve Bayesian classification, and support vector machine (SVM) are widely used in recent years.

Unlike clustering and classification that attempt to classify the input data to *k* groups, association rules and sequential patterns are focused on finding out the "relationships" between the input data. The basic idea of association rules is finding all the co-occurrence relationships between the input data. For the association rules problem, the apriori algorithm is one of the most popular methods. Nevertheless, because it is computationally very expensive, later studies have attempted to use different approaches to reducing the cost of the apriori algorithm, such as applying the genetic algorithm to this problem. In addition to considering the relationships between the input data, if we also consider the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

sequence or time series of the input data, then it will be referred to as the sequential pattern mining problem. Several apriori-like algorithms were presented for solving it, such as generalized sequential pattern and sequential pattern discovery using equivalence classes.

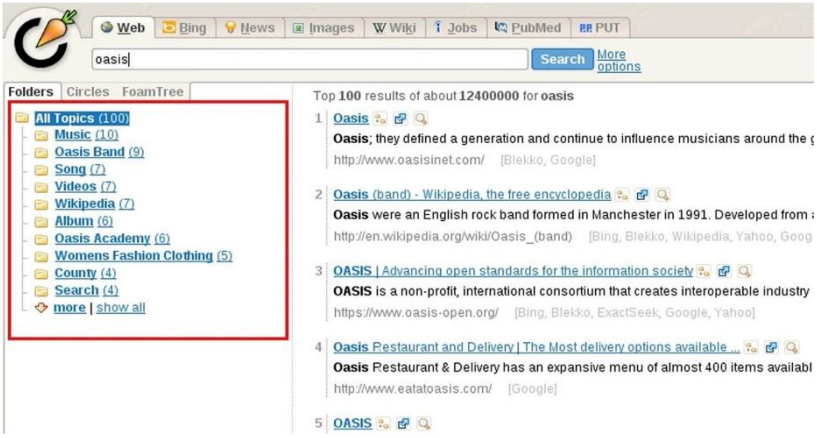
Output the result: Evaluation and interpretation are two vital operators of the output. Evaluation typically plays the role of measuring the results. It can also be one of the operators for the data mining algorithm, such as the sum of squared errors which was used by the selection operator of the genetic algorithm for the clustering problem. To solve the data mining problems that attempt to classify the input data, two of the major goals are: (1) cohesion—the distance between each data and the centroid (mean) of its cluster should be as small as possible, and (2) coupling—the distance between data which belong to different clusters should be as large as possible. In most studies of data clustering or classification problems, the sum of squared errors (SSE), which was used to measure the cohesion of the data mining results.

VII. EFFICIENT DATA ANALYTICS METHODS FOR DATA MINING

- **Unscalability and centralization** Most data analysis methods are not for large-scale and complex dataset. The traditional data analysis methods cannot be scaled up because their design does not take into account large or complex datasets. The design of traditional data analysis methods typically assumed they will be performed in a single machine, with all the data in memory for the data analysis process. For this reason, the performance of traditional data analytics will be limited in solving the volume problem of big data.
- **Non-dynamic** Most traditional data analysis methods cannot be dynamically adjusted for different situations, meaning that they do not analyze the input data on-the-fly. For example, the classifiers are usually fixed which cannot be automatically changed. The incremental learning is a promising research trend because it can dynamically adjust the classifiers on the training process with limited resources. As a result, the performance of traditional data analytics may not be useful to the problem of velocity problem of big data.
- **Uniform data structure** Most of the data mining problems assume that the format of the input data will be the same. Therefore, the traditional data mining algorithms may not be able to deal with the problem that the formats of different input data may be different and some of the data may be incomplete. How to make the input data from different sources the same format will be a possible solution to the variety problem of big data.

VIII. CLUSTERING

Since the problems of handling and analyzing large-scale and complex input data always exist in data analytics, several efficient analysis methods were presented to accelerate the computation time or to reduce the memory cost for the KDD process, as shown in Table 2. The study of shows that the basic mathematical concepts (i.e., triangle



Screenshot of the results of clustering search engine

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

inequality) can be used to reduce the computation cost of a clustering algorithm. Another study shows that the new technologies (i.e., distributed computing by GPU) can also be used to reduce the computation time of data analysis method. In addition to the well-known improved methods for these analysis methods (e.g., triangle inequality or distributed computing), a large proportion of studies designed their efficient methods based on the characteristics of mining algorithms or problem itself, which can be found in, and so forth. This kind of improved methods typically was designed for solving the drawback of the mining algorithms or using different ways to solve the mining problem. These situations can be found in most association rules and sequential patterns problems because the original assumption of these problems is for the analysis of large-scale dataset. Since the earlier frequent pattern algorithm (e.g., apriori algorithm) needs to scan the whole dataset many times which is computationally very expensive. How to reduce the number of times the whole dataset is scanned so as to save the computation cost is one of the most important things in all the frequent pattern studies. The similar situation also exists in data clustering and classification studies because the design concept of earlier algorithms, such as mining the patterns on-the-fly, mining partial patterns at different stages, and reducing the number of times the whole dataset is scanned, are therefore presented to enhance the performance of these mining algorithms. Since some of the data mining problems are NP-hard or the solution space is very large, several recent studies have attempted to use metaheuristic algorithm as the mining algorithm to get the approximate solution within a reasonable time.

Abundant research results of data analysis show possible solutions for dealing with the dilemmas of data mining algorithms. It means that the open issues of data analysis from the literature usually can help us easily find the possible solutions. For instance, the clustering result is extremely sensitive to the initial means, which can be mitigated by using multiple sets of initial means, most data analysis methods have limitations for big data.

IX. BIG DATA INPUT

The problem of handling a vast quantity of data that the system is unable to process is not a brand-new research issue; in fact, it appeared in several early approaches [2, 21, 72], e.g., marketing analysis, network flow monitor, gene expression analysis, weather forecast, and even astronomy analysis. This problem still exists in big data analytics today; thus, pre-processing is an important task to make the computer, platform, and analysis algorithm be able to handle the input data. The traditional data pre-processing methods (e.g., compression, sampling, feature selection, and so on) are expected to be able to operate effectively in the big data age. However, a portion of the studies still focus on how to reduce the complexity of the input data because even the most advanced computer technology cannot efficiently process the whole input data by using a single machine in most cases.

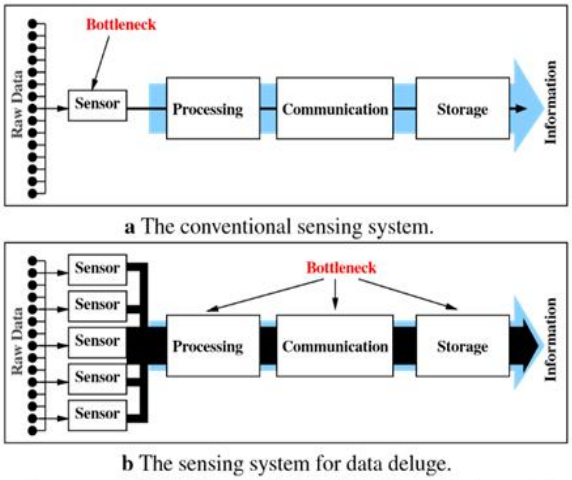


Fig. 6 The comparison between traditional data analysis and big data analysis on wireless sensor network

By using domain knowledge to design the pre-processing operator is a possible solution for the big data. In, Ham and Lee used the domain knowledge, *B*-tree, divide-and-conquer to filter the unrelated log information for the mobile web log analysis. A later study considered that the computation cost of pre-processing will be quite high for massive logs, sensor, or marketing data analysis. Thus, Dawelbeit and McCrindle employed the bin packing partitioning method to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

divide the input data between the computing processors to handle these high computations of pre-processing on cloud system. The cloud system is employed to pre-process the raw data and then output the refined data (e.g., data with uniform format) to make it easier for the data analysis method or system to perform the further analysis work.

Sampling and compression are two representative data reduction methods for big data analytics because reducing the size of data makes the data analytics computationally less expensive, thus faster, especially for the data coming to the system rapidly. In addition to making the sampling data represent the original data effectively, how many instances need to be selected for data mining method is another research issue because it will affect the performance of the sampling method in most cases.

To avoid the application-level slow-down caused by the compression process, in, Jun et al. attempted to use the FPGA to accelerate the compression process. The I/O performance optimization is another issue for the compression method. For this reason, Zou et al. Employed the tentative selection and predictive dynamic selection and switched the appropriate compression method from two different strategies to improve the performance of the compression process. To make it possible for the compression method to efficiently compress the data, a promising solution is to apply the clustering method to the input data to divide them into several different groups and then compress these input data according to the clustering information. The compression method described in is one of this kind of solutions, it first clusters the input data and then compresses these input data via the clustering results while the study also used clustering method to improve the performance of the compression process.

In summary, in addition to handling the large and fast data input, the research issues of heterogeneous data sources, incomplete data, and noisy data may also affect the performance of the data analysis. The input operators will have a stronger impact on the data analytics at the big data age than it has in the past. As a result, the design of big data analytics needs to consider how to make these tasks (e.g., data clean, data sampling, data compression) work well.

X. CONCLUSION

In this paper, we reviewed studies on the data analytics from the traditional data analysis to the recent big data analysis. From the system perspective, the KDD process is used as the framework for these studies and is summarized into three parts: input, analysis, and output. From the perspective of big data analytics framework and platform, the discussions are focused on the performance-oriented and results-oriented issues. From the perspective of data mining problem, this paper gives a brief introduction to the data and big data mining algorithms which consist of clustering, classification, and frequent patterns mining technologies. To better understand the changes brought about by the big data, this paper is focused on the data analysis of KDD from the platform/framework to data mining. The open issues on computation, quality of end result, security, and privacy are then discussed to explain which open issues we may face. Last but not least, to help the audience of the paper find *solutions* to welcome the new age of big data, the possible high impact research trends are given below:

For the computation time, there is no doubt at all that parallel computing is one of the important future trends to make the data analytics work for big data, and consequently the technologies of cloud computing, Hadoop, and map-reduce will play the important roles for the big data analytics. To handle the computation resources of the cloud-based platform and to finish the task of data analysis as fast as possible, the scheduling method is another future trend.

Using efficient methods to reduce the computation time of input, comparison, sampling, and a variety of reduction methods will play an important role in big data analytics-

Because these methods typically do not consider parallel computing environment, how to make them work on parallel computing environment will be a future research trend. Similar to the input, the data mining algorithms also face the same situation that we mentioned in the previous section, how to make them work on parallel computing environment will be a very important research trend because there are abundant research results on traditional data mining algorithms.

How to model the mining problem to find *something* from big data and how to display the knowledge we got from big data analytics will also be another two vital future trends because the results of these two researches will decide if the data analytics can practically work for real world approaches, not just a theoretical stuff.

The methods of extracting information from external and relative knowledge resources to further reinforce the big data analytics, until now, are not very popular in big data analytics. But combining information from different resources to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

add the value of output knowledge is a common solution in the area of information retrieval, such as clustering search engine or document summarization. For this reason, information fusion will also be a future trend for improving the end results of big data analytics.

Because the metaheuristic algorithms are capable of finding an approximate solution within a reasonable time, they have been widely used in solving the data mining problem in recent years. Until now, many state-of-the-art metaheuristic algorithms still have not been applied to big data analytics. In addition, compared to some early data mining algorithms, the performance of metaheuristic is no doubt superior in terms of the computation time and the quality of end result. From these observations, the application of metaheuristic algorithms to big data analytics will also be an important research topic.

Because social network is part of the daily life of most people and because its data is also a kind of big data, how to analyze the data of a social network has become a promising research issue. Obviously, it can be used to predict the behaviour of a user. After that, we can make applicable strategies for the user. For instance, a business intelligence system can use the analysis results to encourage particular customers to buy the goods they are interested.

The security and privacy issues that accompany the work of data analysis are intuitive research topics which contain how to safely store the data, how to make sure the data communication is protected, and how to prevent someone from finding out the information about us. Many problems of data security and privacy are essentially the same as those of the traditional data analysis even if we are entering the big data age. Thus, how to protect the data will also appear in the research of big data analytics.

REFERENCES

[1]. Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. Available: http://www2.sims.berkeley. edu/research/projects/how-much-info-2003/printable_report.pdf.

[2]. Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.

[3]. Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the Twenty-first International Conference on Machine Learning, 2004, pp 1–9.

[4]. Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15(5):1170–87.

[5]. Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9.

[6]. Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Control ling-Data-Volume-Velocity-and-Variety.pdf.

[7]. van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigDataStartups, Tech. Rep. 2013. [Online]. Available: http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/.

[8]. Borne K. Top 10 big data challenges a serious look at 10 big data v's, Tech. Rep. 2014. [Online]. Available: https:// www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.

[9]. Press G. $16.1 billion big data market: 2014 predictions from IDC and IIA, Forbes, Tech. Rep. 2013. [Online]. Available: http://www.forbes.com/sites/ gilpress/2013/12/12/16-1-billion-big-data-market-2014-predictions-from-idc-and-iia/.

[10]. Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013. [Online]. Available: http://www.idc. com/prodserv/FourPillars/bigData/index.jsp.

[11]. Taft DK. Big data market to reach $46.34 billion by 2018, EWEEK, Tech. Rep. 2013. [Online]. Available: http://www. eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html.

[12]. Research A. Big data spending to reach $114 billion in 2018; look for machine learning to drive ana lytics, ABI Research, Tech. Rep. 2013. [Online]. Available: https://www.abiresearch.com/press/ big-data-spending-to-reach-114-billion-in-2018-loo.

[13]. Furrier J. Big data market $50 billion by 2017—HP vertica comes out #1—according to wikibon research, SiliconANGLE, Tech. Rep. 2012. [Online]. Available: http://siliconangle.com/blog/2012/02/15/ big-data-market-15-billion-by-2017-hp-vertica-comes-out-1-according-to-wikibon-research/

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

[14]. Kelly J, Vellante D, Floyer D. Big data market size and vendor revenues, Wikibon, Tech. Rep. 2014. [Online]. Available: http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues.

[15]. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart. 2012;36(4):1165–88.