THE ANALYSIS OF BIG DATA MANAGEMENT OF CORPORATE COMPANIES BY USING PREDICTIVE MACHINE LEARNING APPROACH

A. Vaniprabha and T. Kiruthiga

Department of Electronics and Communication Engineering, Vetri Vinayaha College of Engineering and Technology, India

Abstract

The Big Data is a combination of part-structured, fully structured or unstructured data collected by organizations that can be used for a variety of applications such as machine learning, forecast modeling, fraud detection, sentimental analysis, and other advanced analysis. The number of companies, organizations, and institutes that use large data solutions is increasing exponentially in recent times. Some estimates put the total amount of data thus generated daily at 2.5 trillion bytes. Using them aside it is very difficult to understand such gigantic numbers, but various companies have enthusiastically accepted super contract analyzes for their lofty goals. In this paper a smart big data management and enhance the data type analysis model for predicting the business analysis for corporate companies. On that basis we are only just beginning to understand a little bit how revolutionary super marketing can be, and as it really grows higher, we can expect many changes depending on how business is done in the coming years.

Keywords:

Big Data, Fully Structured, Unstructured Data, Part-Structured, Machine Learning, Sentimental Analysis

1. INTRODUCTION

There are various tools available to analyze superfluous data. These tools allow us to collect a wide variety of data from a variety of sources, including hexadecimal multimedia, Internet services, business applications, and machine registration data [1]. Big Data refers to the vast, diverse range of information that is available and growing every day. This includes the amount of data, the speed at which it is generated, and the type of data points collected [2]. It comes from many sources and comes in many formats (csv, tsv, html, json, parquet, avro). There was even a misconception that Big Data should be the size of a terabyte, jetbyte, or ex-byte - but in fact, Big Data was just that [3]. That is, depending on where the particular data is used, it will be known as Big Data or Normal Data. For example, suppose we have a 50MB file and want to send it as an attachment in the email, but we cannot attach it to the email as it is too big for the email [4]. In this case, the '50MB' file is referred to as the Big Data (Email) of the email. The growing number of such super data providers has made it easier to use this super data [5]. Today we all live in the age of Big Data, in which various developments and revolutionary changes take place at equal intervals [6].

Companies use this huge data to improve their operations, make better decisions, provide better customer service, create customized marketing campaigns based on specific customer preferences, and ultimately maximize profits [7]-[9]. Companies that use it have an advantage over companies that do not, which means they are better informed and can make their own business decisions faster and bigger [10]. Big Data allows companies to focus more on the customer. Current data and historical data can be used to evaluate changing consumer preferences, allowing companies to improve their marketing strategies and be more responsive to customer needs and requirements [11]. In addition, it provides electronic information to healthcare and government agencies about electronic medical records, social media, the Internet, data from other sources, threats of infection or the spread of an infection [12]. In the energy sector, Big Data helps oil companies and gas companies identify potential drilling rigs and monitor the operation of pipelines; similarly, it is used to monitor electrical structures. Manufacturers and logistics companies rely on it to manage their supply chains and improve supply chains. Other applications of the government include emergency assistance, crime prevention, and effective urban initiatives.

Data means unprocessed facts or images. The student's name, marks obtained by the students in certain subjects, personal details of the students, factual details and pictures are related to the student [6]. There is no exact benefit to be gained from unprocessed data. If names and marks are relocated without being sorted, they will not give the correct result. The second major benefit of using data is economics. Someone else has already collected the data, so the researcher does not have to spend money, time, energy and resources to explore this stage. Sometimes you have to buy secondary data, but the cost is always lower than the cost of collecting such data from scratch, which includes scratch, travel and transportation, office space, equipment, and other overhead costs. In addition, since the data has already been collected and cleaned up and collected electronically, the researcher spends most of the time analyzing the data instead of having the data ready.

The second big advantage of using second data is the width of the data. The federal government conducts a large, national number of studies, and individual researchers have to gather hard time. Many of these datasets are long-lived, meaning that the same data is collected from many different times. It allows researchers to see events and changes over time. A third major benefit of using secondary data is that.

2. LITERATURE REVIEW

Computerized data processing is the process of producing information and making it easier for the user to customize. Its speed is an important parameter. So people like data processing. The main purpose of data processing is to handle large amounts of information and to assist the process in answering a variety of questions [1]. Data can only be retrieved by storing it in a specific form. Larger calculations lead to data processing. Information should inform the user of all important aspects of it. These must be submitted in such a way that the corporation is aware of its profit and loss and its status and makes accurate decisions and should lead to accurate decision making [2].

Installing the system requires a great deal of labor and expense. There is a shortage of computer professionals to do this.

Some devices become redundant before installation because the system requires a lot of time to install and the hardware technology is on a very advanced path. Alternative arrangements need to be made when computer systems do not work [3]. Human methods are slow but highly flexible. For example, when changing the form of the report if it is to be done in readiness the necessary instructions in this regard should be communicated to the appropriate employee. If the system is otherwise systemic it is necessary to tell the computer systems inspection stages before designing [5]. The primary data and the secondary data have a common similarity. Primary data are collected by researchers or by a team of researchers for a specific purpose or analysis. Here, a research team considers and develops a research project, collects data designed to answer specific questions, and makes own analyzes of the collected data [7]. In this case, the people involved in the data analysis are familiar with the research design and data collection process. The secondary data analysis is the use of data collected by someone else for another purpose [11]. In this case, the researcher raises questions addressed through the analysis of a set of data not involved in a collection. He did not collect the data to answer the researcher's specific research questions, but instead collected them for another purpose. So, the same data set can actually be a primary data for a researcher and a secondary data set to a different one [12]. A related problem is that the variables vary or are classified more than what the researcher has chosen. For example, it may not be a continuous variant, but may have been subdivided into categories to be defined as "white" and "other" rather than having a genus for each major species. Another significant drawback to using secondary data is that the researcher does not know how the data collection process was done and how well it went. Information about data being affected by issues such as low response rates or reflection misinformation of specific research questions is not usually well known to the researcher. Sometimes this information is readily available. However, many other secondary data sets are not linked to this type of information, and the researcher must learn to read between the lines and consider how to observe problems in the data collection process.

3. PROPOSED MODEL

In today's world a lot of data has to be stored and processed. Regardless of the type of organization i.e. office, college, school, hospital, bank, train booking, factories, theaters, etc., they process a large amount of data. Data processing can be defined as: "Data processing is the process of submitting unprocessed data into a computer and adapting it to the user's instructions. Also provides the required feedback with the right topics in beautiful form.

- **Single User**: If you are using this single user then only one user can use it at a time.
- **Multi User**: In this case, the computer uses the same microprocessor as a single-user mode but here two or more users complete their task by sharing the primary processor.
- **Batch Processing**: Here all kinds of work are input. And is collected in bulk and processed in part. The submitted function is performed sequentially. The disadvantage of this is that it is difficult to replicate a single task and is to repeat the whole process.
- **Multi Programming**: It is necessary to look at several programs at once. That means implementing several independent programs. It must be done simultaneously by inserting it inwards or stacking one on top of the other. This is called setting up a simultaneous command line.
- **Multi-Processing**: Dental processing is done simultaneously by a personal computer webinar in two or more instruction sequences.
- **Online and Real-Time Processing**: This operating system is also called direct processing. Real-time modes are processes that process data as soon as it receives input.
- **Time Sharing Concept**: Time sharing method is a processing method. It has more than one arbitrary computer system. These computers work online and are capable of accessing the processor directly.
- **Main Processor**: A processor is given a specific time for each mode of operation. This is called time slice. The processing is done very fast and the user imagines that the processor is always working according to his request.
- **Distributed System**: It was processed the metrology-related information using the computers and other devices. These are connected by an interstitial web or a spacious web or diffusion web. The methods here are controlled by a centralized server. This is considered an extension of the time distribution system.

A computer network is a network of computers and communication equipment that lets users delivers messages and share resources with other users. The wide boundary network connects the interstitial web separated by distance.

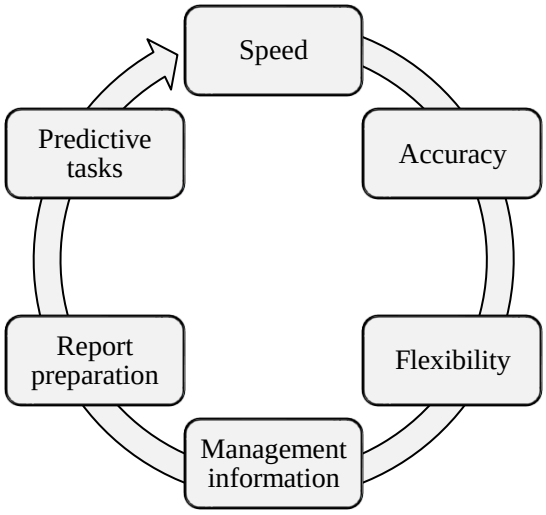

Fig.1. Primary focused for the proposed model

It connects one landscape boundary with another. Broadband enables data transfer by enabling the database to connect to remote locations. Its range of aggression is vast. Its communication rate goes from a few hundred units per second to a few thousand units. These will also be operated by satellites and microwaves regulated by government agencies. The primary focused for the proposed model shown in Fig.1.

- **Speed**: The computer processes data from multiple data files in a short amount of time. This is made possible by computers that run very fast.
- **Accuracy**: The data processed by the computer is of high accuracy. Systems written programs control data before and after processing. It detects incorrect data and greatly confirms the accuracy. And verifies the reliability of the reports issued.
- **Flexibility**: Modern digital computers can be used for a variety of reasons.
- **Management Information**: These provide useful information to control management and support decision making.
- **Facilitates report preparation**: The computer facilitates the preparation of various reports by the organizational officers required for decision making and control.
- **Increased ability to perform predictive tasks**: The computer performs predictions with greater speed. It enhances efficiency.

The Internet is a term used to describe a technology that connects computers in more than 65 countries worldwide. These computers are connected to each other based on the philosophy of the network and allow a person to communicate with another person in a corner that is very far from one end of the earth. This website is always changeable and subject to redesign and redesign. The secondary focused areas for the proposed model shown in Fig.2,

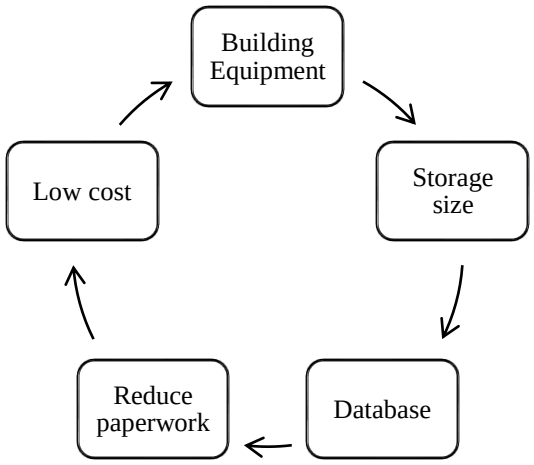

Fig.2. Secondary Focused Areas for the Proposed Model

- **Selection of Building Equipment**: There are a large number of equipment required for computer systems. This allows businesses to select the equipment that best suits their processing needs.
- **Storage Size**: Large amounts of data can be conveniently stored, retrieved and transferred. Computers store large amounts of data and are able to capture the required information quickly.
- **Database**: Supports computer database creation. Integrates data records through such databases and reduces data redundancy.
- **Reduce paperwork**: Using the computer for data processing will increase the management of businesses. Helps to reduce paper work.
- **Low cost**: The first investment in installing a system reduces the cost of a subsequent event, even if it is more expensive. Processing more data and keeping records can reduce costs.

Being a machine that makes millions of predictions frees the computer from frustration, dissolution, or inattention. It makes predictions with the same accuracy and speed. The data is sent in packets to a separate computer. A package contains the data from which the data is to be transferred, the address from which the data is being sent, and the address to which it should be sent. A router is a specialized device.

4. RESULTS AND DISCUSSION

The proposed Predictive machine learning approach (PMLA) was compared with the existing Predictive big data analytics (PBDA), Advance predictive big data analysis (APBDA), Semantics in predictive big data analytics (SPBDA) and big data and predictive analytics (BDPA)

4.1 COMPARATIVE ANALYSIS

This involves examining user behavior metrics and comparing the awareness of a company's products, services, and brands with its competitors and observing customer engagement in real time. This allows the processing of large datasets distributed across clusters of computers. It is also one of the best supersonic devices designed to measure thousands of computers from a single server. It is based on the concept of 'data placement'. This allows for faster data processing in Table.1.

Table.1. Comparison of Comparative Analysis

Data	PBDA	APBDA	SPBDA	BDPA	PMLA
100	78.31	62.76	62.53	79.56	88.27
200	76.65	60.90	63.37	79.15	88.37
300	74.99	59.04	64.21	78.74	88.47
400	73.33	57.18	65.05	78.33	88.57
500	71.67	55.32	65.89	77.92	88.67
600	70.01	53.46	66.73	77.51	88.77
700	68.35	51.60	67.57	77.10	88.87

4.2 SOCIAL NETWORKS REQUEST

Information about what people say on social media about a particular company or product, which goes beyond what a survey can provide. This data can be used to help identify targeted audiences for marketing campaigns by observing action songs related to specific topics from a variety of sources. The database is widely used today to effectively manage large amounts of data. It supports multiple data center copying. Data is automatically copied to multiple nodes in Table.2.

Table.2. Comparison of social network request

Data	PBDA	APBDA	SPBDA	BDPA	PMLA
100	76.20	58.04	70.66	72.66	88.44
200	76.78	56.90	72.80	69.42	88.49
300	77.36	55.76	74.94	66.18	88.54
400	77.94	54.62	77.08	62.94	88.59
500	78.52	53.48	79.22	59.70	88.64
600	79.10	52.34	81.36	56.46	88.69

4.3 MARKETING ANALYSIS

This information can be used to make new products, services, and initiatives informed and innovative. It Allows programmers to analyze large datasets. It helps to quickly query and manage large data sets. It supports partitioning, compilation, and tables. It is filtered to manage and query only structured data. It provides the Java Database Connection (JDBC) interface.

Table.3. Comparison of Marketing Analysis

Data	PBDA	APBDA	SPBDA	BDPA	PMLA
100	81.28	56.02	71.23	70.14	88.53
200	81.12	54.82	69.61	70.27	88.56
300	80.38	53.17	67.81	69.00	88.56
400	80.03	51.82	66.13	68.66	88.58
500	79.58	50.40	64.42	68.09	88.60
600	79.13	48.97	62.71	67.52	88.61

4.4 CUSTOMER SATISFACTION

All information gathered can reveal what customers think of a company or brand, how it can protect brand loyalty in the event of potential problems, and how to improve customer service efforts.

Table.4. Comparison of customer satisfaction

Data	PBDA	APBDA	SPBDA	BDPA	PMLA
100	79.77	50.20	63.26	74.34	85.62
200	80.10	51.70	63.85	76.21	86.66
300	81.44	52.81	64.83	77.04	86.79
400	82.11	54.18	65.55	78.56	87.53
500	82.94	55.49	66.34	79.91	88.11
600	83.78	56.79	67.12	81.26	88.70

4.5 UNDERSTANDING OF BIG DATA

Generally, data professionals may be aware of what is going on, while others may not know any obvious details about it. Data (Big Data) attempts fail. The better option for data streaming and storage, annoy companies, sometimes leaving them unable to find answers. So, they make the wrong decisions and choose the wrong technology.

Table.5. Comparison of understanding of big data

Data	PBDA	APBDA	SPBDA	BDPA	PMLA
100	82.58	53.19	66.04	77.95	87.75
200	83.63	54.20	67.18	78.87	87.32
300	84.34	55.13	68.29	80.20	88.56
400	85.64	56.13	68.99	81.07	88.67
500	86.52	57.10	70.12	82.20	89.08
600	87.51	58.08	71.11	83.26	89.48

4.6 DATA PROTECTION

Securing such a large amount of data is also a big challenge. Because companies are busy understanding, storing, and analyzing their databases, they are delaying action on data protection and being vulnerable to malicious attackers.

Table.6. Comparison of data protection

Data	PBDA	APBDA	SPBDA	BDPA	PMLA
100	63.60	54.92	61.53	73.35	85.43
200	63.93	56.42	62.12	75.22	86.47
300	65.27	57.53	63.10	76.05	86.60
400	65.94	58.90	63.82	77.57	87.34
500	66.77	60.21	64.61	78.92	87.92
600	67.61	61.51	65.39	80.27	88.51

4.7 QUALITY-DEPENDENT

If data acquire the properties of objects, they are called quality-dependent. The interrelated data is formatted as a record. The collection of these records becomes a file. A database is a system in which data is organized and arranged vertically across rows. Each cross row is recognized as a record and each long row is taken as a file.

Table.7. Comparison of Quality dependent

Data	PBDA	APBDA	SPBDA	BDPA	PMLA
100	66.41	57.91	64.31	76.96	87.56
200	67.46	58.92	65.45	77.88	87.13
300	68.17	59.85	66.56	79.21	88.37
400	69.47	60.85	67.26	80.08	88.48
500	70.35	61.82	68.39	81.21	88.89
600	71.34	62.80	69.38	82.27	89.29

5. CONCLUSION

The data collection process often maintains professional and professional levels that do not involve individual researchers or small research projects. The data collection for multiple federal databases is often done by members of staff who specializes in a particular task, and have many years of experience in that particular area and that particular area. Many small research projects do not have a level of expertise because a lot of data is

collected from part-time students. There is a major drawback in using secondary data that the researcher may not be able to answer specific research questions or contain the specific information that the researcher wants. It should not be collected in the geographical area or in the desired years or in specific populations of interest to the researcher. Since the researcher does not collect the data, he has no control over what is in the database. Often this analysis can reduce or alter the original questions the researcher is trying to answer.

REFERENCES

[1] M. Seyedan and F. Mafakheri, "Predictive Big Data Analytics for Supply Chain Demand Forecasting: Methods, Applications, and Research Opportunities", *Journal of Big Data*, Vol. 7, No. 1, pp. 1-22, 2020.

[2] A.R. Reddy and P.S. Kumar, "Predictive Big Data Analytics in Healthcare", *Proceedings of 2nd International Conference on Computational Intelligence and Communication Technology*, pp. 623-626, 2016.

[3] S. Seo and K. Obermayer, "Self-Organizing Maps and Clustering Methods for Matrix Data", *Neural Networks*, Vol. 17, No. 8-9, pp. 1211-1229, 2004.

[4] Y. Rani and D. Rohil, "A Study of Hierarchical Clustering Algorithm", *International Journal of Information and Computation Technology*, Vol. 3, No. 10, pp. 1225-1232, 2013.

[5] C. Selvi and E. Sivasankar, "A Novel Optimization Algorithm for Recommender System using Modified Fuzzy C-means Clustering Approach", *Soft Computing*, Vol. 78, pp. 1-16, 2017.

[6] M.V. Nural, M.E. Cotterell and J.A. Miller, "Using Semantics in Predictive Big Data Analytics", *Proceedings of IEEE International Congress on Big Data*, pp. 254-261, 2015.

[7] M.A. Waller and S.E. Fawcett, "Data Science, Predictive Analytics, and Big Data: A Revolution that will Transform Supply Chain Design and Management", *Journal of Business Logistics*, Vol. 34, No. 2, pp. 77-84, 2013.

[8] S. Smys, "Survey on Accuracy of Predictive Big Data Analytics in healthcare", *Journal of Information Technology*, Vol. 1, No. 2, pp. 77-86, 2019.

[9] A. Gunasekaran and S. Akter, "Big Data and Predictive Analytics for Supply Chain and Organizational Performance", *Journal of Business Research*, Vol. 70, pp. 308-317, 2017.

[10] B. Heller, S. Seetharaman and P. Mahadevan, "ElasticTree: Saving Energy in Data Center Networks", *Proceedings of USENIX Conference on Networked Systems Design and Implementation*, pp. 1-17, 2010.

[11] A. Callado, C. Kamienski, S.N. Fernandes and D. Sadok, "A Survey on Internet Traffic Identification and Classification", *IEEE Communications Surveys and Tutorials*, Vol. 11, No. 3, pp. 37-52, 2009.

[12] X. Li, J. Chen, G. Zhao and M. Pietikainen, "Remote Heart Rate Measurement from Face Videos under Realistic Situations", *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 4264-4271, 2014.