Big data analytics and machine learning in logistics and supply chain management: analysis, optimization and future scope

Vaibhavi Golgire

Department of Computer Engineering, Pimpri Chinchwad College of Engineering , savitribai phule pune university, Pune, Maharashtra, India.

Abstract : Big data refers to vast quantities of data from multiple sources in various formats. Because of the versatility of the supply chain industry to volume, data generation speed and a variety of data, Data analysis and knowledge using traditional methods is difficult. Data analysis includes the discovery of useful knowledge from large data sets to support decision-making including data review, cleaning, transformation and modelling. Many big data and analysis tools have made it possible for people to gain precious insights from this vast volume of data. Data analytics will help forecast the supply chain industry that is our leading customers, future consumer demand, product design insight, attractive functions, system failure prediction, and much more. This thesis explores the application of Big Data Analytics in logistics and management of supply chains. Importance of new technologies for database management, including MongoDB, DB2, MarkLogic and Apache Cassandra, various ETL tools, such as SSIS, Informatica, Talend, Ab Initio, data replication and recovery. Conventional relation databases are not appropriate for storing this vast amount and variety of data, so scalable databases such as Apache HBase that accommodate large volumes of data using clusters and NoSQL systems have to be used. The use of these advanced technology enables the company to make effective and efficient decision-making that reduces its operating costs. Open research issues in this area includes data protection and safety problems, big data efficiency and scalability, lack of resources.

IndexTerms **- supply chain management, logistics, Big data.**

I. INTRODUCTION

With the immense amount of data generated in almost every industry every day, accurate decisions can not only be taken on the basis of such data. The enormous deployment of connected devices, such as GPS data, enterprise resource planning systems, feeds from social media, different IoT devices such as mobile devices, RFID readers, webcams and sensor and transport systems adds enormous amounts of autonomous information to these network systems [7]. Big data technologies can be used to identify the nuances and importance of the underlying data [5]. Unstructured or time-sensitive data or simply very broad data can not be processed by conventional database engines. This form of data needs a different approach to processing, called big data, which uses huge parallels on readily available hardware [6].

In an increasingly competitive and globalized economy , companies need to thrive on the market to make continual improvements in efficiency, productivity, quality and flexibility, and to maintain and improve customer service levels. Analytics' insights help to make the customer's demand-centric supply chain, which is linked at each point to support the customer better, smarter and effective behind the scenes. Supply chain analysis plays a key role in improving supply chain efficiency through increased supply chain visibility, demand management, uncertainty, and cost fluctuation. Statistical time series analysis can be used for demand prediction and inventory planning and aspects such as forecast intervals can be used in the inventory planning process [4].

The analyses of associations are used for merchandising, product development and cross-sales .In promotional planning, multiple regression analyses can be used to see how variables such as "price" affect demand. Multiple regression analysis can also be used when conducting multi-variable experiments to assess the efficacy of different improvements that are implemented at the same time. This is important in industries that deal with perishable products (e.g. supermarkets, industries where cash flow is usually the key constraint (and stock is one of the closest metrics). Perhaps the best known use of this is when ecommerce sites are updated.

The rest of the paper is structured accordingly – Section II presents the corresponding work. Section III explains the solution suggested. In section IV, the report and the document will be concluded. Finally, the references used in the review paper are available.

II. Related Work

Comparative analysis of reference articles is shown below in Table 1.

Table 1. Comparative Analysis

Paper	Description	Advantages	Limitation
Distribution cost optimization using Big Data Analytics, Machine Learning and Computer Simulation for FMCG Sector	Sri Lankan FMCG Company's data is used for analysis and quantitative model is developed to reflect cost structure. Main aim behind developing this model is to minimize transportation cost of goods by proper vehicle utilization, reducing distribution cost. orders, vehicles. cost are taken as input, and best optimum route is found to cover all the orders to minimize the cost	Dijkstra algorithm, floyd warshall is widely used to obtain the shortest path of distribution. Apart from Dijkstra algorithm, big data analytics to identify the constraints for the distribution points. Warshall algorithm are used to obtain least cost routes	only centralized distribution model and only FMCG sector is considered and need to explore more factors that are affecting cost
Big Data Analytics Adoption in Warehouse Management: A Systematic Review	This paper presents overview of Big data analytics (BDA)technologies in warehouse management	Role of Genetic algorithms, heuristic algorithm, fuzzy logic, time series algorithm, routing Algorithm described in the field of analytics for SCM .	Need to consider some other important aspects also like Monitoring warehouse performance in real time, risks and accidents predictions, warehouse human resources planning and monitoring, warehouse cost saving
Quality Analytics in a Big Data Supply Chain Commodity Data Analytics for Quality Engineering	Layered architecture of analytics is presented in which Descriptive analytics , Predictive analysis, Perspective analysis discussed in detail	Use of Analytic tools like IBM Cognos Framework Manager, IBM Cognos Report Studio, Oracle Business Intelligence Foundation Suite or SAP Business Objects, Word clouds discussed	Diagnostic analytics did not mention. In supply chain management this analysis also plays important role it takes current and previous data and provides deeper analysis
Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice	BDA capabilities identified and evaluated and presented in 4 quadrant structure. challenges faced by organization during whole BDA process is discussed in which author mentioned time and resource constraints availability, privacy and security issues. Best practices of BDA mentioned what infrastructure should organization have ,technologies ,how to avoid bad data so that one can save time and money, correct approach of information sharing.	Descriptive analysis is done to summarize existing research. challenges faced by organization during whole BDA process is discussed.	Risk management related issues did not mention, BDA can help to minimize risk and its potential impact, traceability issues, also issues related to agility of SCM world did not discuss.
Big data analytics and application for logistics and supply chain management	Detailed analysis is performed on papers published in period 2010 to 14 March 2018. application of big data is discussed through rigorous review process. author has selected 5 papers for this analysis	In this article, author informed researcher about issues associated with big data and its application which will facilitate work in this field. Researcher will inspire from above information and will contribute more and more.	
Unstructured big data analytics for air cargo logistics management	Air cargo logistics plays a crucial role in speedy distribution enabling products to be promptly shipped between locations, but few	This study proposes hybrid data mining model to investigate	Furthermore, integrated innovation diffusion theory and resource

	studies have explored the business knowledge of air cargo logistics by unstructured big data. Hence, this study develops a hybrid data mining model to tackle critical issues of air cargo logistics as well as to generate operational strategies for those who would like to manage aviation logistics. Eight vital themes of air cargo logistics from the analytical results of the proposed hybrid data mining model are runway expansion, the air cargo platform, special logistics and certification, e-commerce logistics, cooperation, network expansion, market observation, and airplane conversion.	beneficial insights of air cargo logistics management. The procedure of unstructured data analytics comprises data collection, data pre-processing, cluster, and association rules.	dependence theory with the analytical findings of the hybrid data mining model, this study creates air cargo logistics strategies involving "develop innovative aviation logistics in a suitable manner" and "establish strategic cooperation to enhance aviation logistics performance." Finally, the proposed strategies of logistics management of air cargo carriers can be adopted for those who are interested in running aviation logistics.
Application of Big Data in a Multi-Category Product-Service System for Global Logistics Supports	Determining how a product-service system (PSS) combines tangible products with intangible services in order to fulfill customer needs has been a study issue in this field (Mahut et al., 2017; Boehm and Thomas, 2013; Geng et al., 2010). For example, IKEA does not "sell furniture;" rather it claims that it "creates a home" (Isaksson et al., 2009). A PSS focuses on being both a product-and-service provider and can be divided into five categories: products/services/combinations/substitutions, services at the point of sale, different product use concepts, maintenance services, and revalorization services (Mont, 2002). The applications of PSS can be seen in the following two cases: Toyota's material handling group has traditionally sold forklift trucks and also serves as a warehouse transportation solution provider; ITT Flygt offers cleaning solutions instead of selling pumps (Isaksson et al., 2009). However, these PSS cases belong only to a specific class, not to a complete combination of all PSS classes. Few researchers have explored ways to implement a multi-category PSS or how to combine product and service offerings.	Recent study of big data analytics technology has been applied on cloud computing with data that indicates that firms are increasingly adopting big data in cloud solutions such as software-as-a-service (SaaS) because it offers an attractive, low-cost alternative (Wang et al., 2016).	Although applying BDA applications to improve a company's competence is a well-known concept, considering global logistics with BDA applications is an innovative concept in which to implement multi-category PSSs. There are some implications based on the experiences of case company. First, after the implementation of the GLSV system, the KPIs of the case company showed that global logistics with BDA applications could indeed help the case company with PSS implementation. The business model of the case company could, therefore, shift from being production-oriented to knowledge-service-oriented (Mont, 2004).
Near real-time big data analytics for NFC-enabled logistics trajectories	The system uses RFID tags to optimize logistics processes. Knowing in real time the location of products, avoid unnecessary movement, make inventories in real time, avoid errors such as check suppliers receptions on the platform and register without delay, check storage positions, control the order preparation, check the loads and against theft use RFID technologies[4,5, 12, 13]. Product's RFID data is in high volume which requires efficient system to shape and form the trajectories of products in order to respond to real-time queries. Also, it allows to avoid errors such as suppliers check receptions on the platform and register without delay, verify storage positions, control the order preparation, check the loads and against theft.	RFID data is in high volume with low costs. Also, there is a need to optimize logistics processes such as knowing in real time the location of products, to analyze and decision making in near real time. In this paper we have proposed the conceptual Meta model for products trajectories to present trajectories from different facets: raw, structured, semantic and composite region of interests. To deal with products trajectories big data, we proposed a Data warehousing conceptual	For a given moving object and time interval, RFID allows collecting tremendous amounts of spatio-temporal messages (tagid, location, readtime, observation). Modelling trajectories with structured presentation accelerated processing of application that only referring to the information about the spatio-temporal position of the beginning B, end E and sequence of stops S and moves M. SpaceTimeEvent class presents an event as an occurrence that happens in a small space and lasts

		schema using composite documents, and we created the NoSQL trajectories data warehouse.	a short time, e.g. changing position. From temporal point of view, SpaceTimeEvent class is a composition of TemporalObject class according to specification of TimeReference class. From spatial point of view, it is a composition of SpatialObject according to SpatialReference class specification.

III. PROPOSED SOLUTION

Big data is an area in which information is systematically collected or processed by data sets that are too broad or complex to be managed by conventional data processing tools. Initially, big data were connected to three major concepts: number, variety and speed. Machine learning algorithms delivers unprecedented value to supply chain operations: from cost reduction through minimized operational overhead and risk mitigation, to enhanced supply chain forecasting, fast deliveries, and improved customer service. They also provides insights into where automation can deliver the most significant scale advantages

Algorithm

In order to extract more value from your results, use these five algorithm categories.

a) Linear regression

Linear regression is one of advanced analytics' most common algorithms. It is also one of the most frequently used. People can easily see how the input data works and how it is connected to the output data.

Linear regression uses two sets of continuous quantitative measures. The first set is the independent variable or predictor. The other is the response or variable based. The objective of linear regression is to define the relationship as a formula which describes the dependent variable according to the independent variable. The dependent variable can be estimated for any case of an independent variable when this relation is quantified.

Time is one of the commonly used independent variables. If your independent variable is income, costs, clients, use or efficiency, you can predict a value with linear regression if your relationship to time is established.

b) Logistic regression

Logistic regression sounds like linear regression but is actually focused on categorisation rather than quantitative forecasting problems. In this context, the output variable values are not continuous but discrete and finite, with infinite values as with linear regression.

The aim of logistic regression is to determine whether or not an instance of an input variable fits in a group. The logistic regression output is between 0 and 1. Returns closer to 1.
demonstrate that the input variable matches the category more precisely. Results closer to 0 how that the input variable actually does not fit into the category.

Logistic regression is also used to respond to clearly defined questions, yes or no. Is a customer going to shop again? Is a credit worthy of a buyer? Is the potential a customer? The response to these questions will cause a sequence of actions in the business process that can lead to potential revenues.

c) Trees of Classification and Regression

Classification and regression trees make use of data categorisation decisions. Any decision is based on one of the input variables. With each question and answer, the instance of data is pushed closer to being clearly classified. This collection of questions , answers and corresponding data divisions establish a tree-like structure. A group is at the end of each line of questions. This is called the classification tree leaf node.

These trees can become very large and complex. One way of managing complexity is by cutting the tree or purposely eliminating the interrogation levels to match the exact fit with abstraction. A model which works well with input values of all instances, both known and unknown, is paramount. Preventing this model from over fitting involves a delicate balance between fitness and abstraction.

Random forests are a type of classification and regression trees. Instead of creating a single tree with several logic branches, a random forest is the accumulation of several small and simple trees that each analyse the data instances and categorise. After all these basic trees have completed their data assessment, the method fuses the results into a final category prediction based on the composition of the smaller categories. This is generally called an ensemble form. These random forests also combine accurate fit and abstraction and have been successfully implemented in many business cases.

In comparison to practical regression based on a yes or no categorisation, multivalue categorization can be predicted by classification and regression trees. They are often easier to imagine and see the definitive route to a certain categorization.

d) Neighbors K-Nearest

K-nearest neighbour is also an algorithm for classification. It is called a "lazy learner," because the training stage is very small. The learning process consists of the data stored in the training set. With the evaluation of new situations, the distance to each data point in a training set is assessed and a consensus is reached as to which category the new data instance is based on its proximity to the training instances.

This algorithm can be costly based on the size and complexity of the training set. Because each new instance needs to be compared with all instances in the training data set and a derived distance, several computer resources are available to this process each time it is running.
This algorithm of categorization enables multi-value data categorizations. Moreover, noisy training data appears to distort classifications.

K-nearest neighbours are often selected because the results are easy to use , easy to train and easy to interpret. It is also used when searching for related objects in search applications.

e) Clustering of K-means

Clustering K-means focuses on the formation of similar attributes classes. These groups are known as clusters. If these clusters are formed, other instances can be assessed to see where they match.

This technique is also used for data scanning. In the beginning, the analyst indicates the number of clusters. The K-mean clustering method divides the data into the amount of clusters based on the identification of data points with similarities around a common center, the centroid. These clusters are not the same as groups because they have no business sense at all. They are associated instances of input variables. When defined and evaluated, these clusters can be translated into a category and given a name that has business significance.

K-means clustering is commonly used because it is easy to use , easy to explain and efficient. One observation is that the clustering of k-means is highly susceptible to outliers. These outliers can dramatically change the existence and meaning of these clusters and consequently analytical results.

These are some of the most common algorithms in advanced analysis. Each has advantages and disadvantages and various ways to generate market value effectively. The goal of these algorithms is to further refine the data to such an extent that the knowledge that results can be used in business decisions. It is this process of providing downstream processes with more detailed and higher value knowledge that is necessary for businesses to really maximise the value of their data and achieve the desired results.

Proposed System Specification:

BACK END

There will be database systems which will store the information related to various processes of supply chain management. These datasets includes database for Invoices, Audit Compliance, Material Consumption Forecasting, Factory Inventory, Purchase order, Deliver Record ,Material quality, Contract Specification[3].

Software Requirement

Front End UI- SAPAND HANA, SQL Server, Excel and Access, Microstrategy
Simulation- Arena, FICO, AnyLogic
Predictive modelling -SPSS, SAS, Revolution analytics
Business Intelligence- Tablue, SSIS

IV. SCOPE OF PROPOSED SYSTEM

Now, let's go over how the future looks for a Big Data career.

a) Increasing demand for Data Analytics:

Not long ago, the importance of data processing in the modern world Peter Sondergaard of Gartner Research highlighted:

"Information is the 21st century oil and combustion engine analytics."

The amount of data we churn every minute is increasing every day. While Data Science gurus vows the importance of data, what use will the data be for us if there are insufficient professionals in the field of data analytics? Who will analyse these huge amounts of data and turn them into a useful corporate resource? Since businesses worldwide realise the true data value, the demand for professional data analytics is growing. Big data analysis is necessary to process this data.

b) Increasing enterprise adoption of Big Data:

Big data analytics was found to be one of the top priorities of the participating organisations, according to the Peer Review report, – they agree that their overall efficiency can be enhanced.

The inference here is that more and more companies worldwide use Big Data technology and machine learning to boost their efficiency.

V. CONCLUSION

Every established supply chain process generate humungous data in day-to-day process of moving products from point-A to point-B. Data analytics can help logistics and supply chain industry to predict the demand of customer, to predict demand of product, predicting machine failure, Warehousing management and to optimize the route. In a globalized and increasingly competitive economy, companies require to survive in the market, to make continuous improvements in levels of efficiency, productivity, quality and flexibility, maintaining and improving the level of service to customers. One of the most promising technology that can help companies is big data analytics, professional management of the supply chain becomes a vital competitive tool for companies to provide differential value to their customers, and to be profitable.

VI. FUTURE SCOPE

Due to the technical and competitive obstacles and emerging business prospects posed by the Information Technology transition, big data analytics has arrived as the next frontier. Organizations that use big data analytics in real time know what's happening in the data in real time, and then use the knowledge to better decisions have a strategic advantage. Fully reviewing the tools, strategies, measures, and approaches for big data stream analysis, as well as numerous big data topics to provide a systematic view of where future research could lead.

REFERENCES

[1] Adikari, A. and Amalan, T., 2019. Distribution cost optimization using Big Data Analytics, Machine Learning and Computer Simulation for FMCG Sector. 2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)

[2] Ghaouta, A., bouchti, A. and Okar, C., 2018. Big Data Analytics Adoption in Warehouse Management: A Systematic Review. 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)

[3] Tan, J., Ang, A., Lu, L., Gan, S. and Corral, M., 2016. Quality Analytics in a Big Data supply chain: Commodity data analytics for quality engineering. 2016 IEEE Region 10 Conference (TENCON)

[4] Arunachalam, D., Kumar, N. and Kawalek, J., 2020. Understanding Big Data Analytics Capabilities In Supply Chain Management: Unravelling The Issues, Challenges And Implications For Practice

[5] Govindan, K., Cheng, T., Mishra, N. and Shukla, N., 2018. Big data analytics and application for logistics and supply chain management. Transportation Research Part E: Logistics and Transportation Review, 114, pp.343-349.

[6] C. Lin and M. Lin, "Application of Big Data in a Multicategory Product-Service System for Global Logistics Support," in IEEE Engineering Management Review, vol. 47, no. 4, pp. 108-118, 1 Fourthquarter,Dec. 2019

[7] L. Karim, A. Boulmakoul and A. Lbath, "Near real-time big data analytics for NFC-enabled logistics trajectories," 2016 3rd International Conference on Logistics Operations Management (GOL), Fez, Morocco, 2016

[8] S.-R. Hamid,"Benchmarking Key Success Factors for the Future Green Airline Industry," Procedia - Soc. Behav. Sci., vol. 224,pp. 246–253, Jun. 2016.

[9] K. Huang and H. Lu, "A Linear Programming-based Method for the Network Revenue Management Problem of Air Cargo," Transp. Res. Procedia, vol. 7, pp. 459–473, Jan. 2015.

[10] F. Heinitz, M. Hirschberger, and C. Werstat, "The Role of Road Transport in Scheduled Air Cargo Networks," Procedia - Soc. Behav. Sci., vol. 104, pp. 1198–1207, Dec.2013.