INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 7, Issue - 11, Nov – 2021
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.719
Received Date:30/10/2021 Acceptance Date: 14/11/2021 Publication Date: 30/11/2021

Big Data Analytics is evolution not Revolution

[1]Nagendra R., [2]Dr. Nagendra M.
[1]Research Scholar, [2]Professor and HOD
[1 & 2]Department of computer Science
[1]Rayalaseema University, Kurnool, Andrapradesh, INDIA - 518007,
[2]Sri Krishnadevaraya University, Ananthapuram, Andrapradesh, INDIA - 515003,
Email: [1] Profnagendra@yahoo.com, [2]nagendra_m@rediff.com.

Abstract: *Information is now available in an overabundance, so much so, that distinguishing the noise from the signal has become very problematic. In the past, the collection and storage of information was the primary issue. Currently, there are massive amounts of data both structured and unstructured, that need to be analyzed in an iterative, as well as in a time sensitive manner. In response to this need, data storage and retrieval systems saw changes in any folds. File Processing System was first to replace non-computer based approach for maintaining records. It was a successful System of its time and still there are many organizations that are using File Processing System to maintain their data and information. But it is just not suitable for handling data of big firms and organizations. It has many drawbacks and disadvantages that made it out of date. Now as we know that File Processing System uses different files to store data. DBMS systems came in to picture which is able to solve many problems of file system like Data integrity, Data redundancy, Data Isolation, Security etc. Every system has its own pro's and con's so DBMS. Data warehouse lead to reduce cost of storage, complexity, size, performance and was able to address many of the problems of DBMS. Data mining came into picture. Next concept emerged in the field of data Analytics is Data mining. Data mining is an important part of knowledge discovery process that we can analyze enormous set of data and get hidden and useful knowledge. Data mining is applied effectively not only in the business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government etc. The phenomena of Big Data*

and Analytics bring a new life to the discipline of data mining. Big Data is defined by its size, comprising a large, complex and independent collection of data sets, each with the potential to interact. In addition, an important aspect of Big Data is the fact that it cannot be handled with standard data management techniques due to the inconsistency and unpredictability of the possible combinations. This study intends to provide the insight of journey of data from raw concepts to Big data Analytics.

Key Words: *Big data Analytics, File processing system, Database management system, Data mining, Data warehousing.*

1.INTRODUCTION:

The evolution of Big data and analytics has brought a new life to the data mining. This topic has chosen to define and trace the origin of big data and answer whereabouts of Big data Analytics. Big Data Analytics meant more than Data mining (DM). The vast amount of data mandates novel algorithmic approaches to Big Data Analytics. But there is more to come: Big Data often has a significant crowd sourcing aspect and now places a heavy emphasis on data cleansing, outlier detection, and the cloud. The structuring of the data is a big challenge due to its nature (often text and images). This necessitates an engineering approach for data driven science and engineering applications of Big Data Analytics.

The explosion of the Internet, social media, technology devices and apps is creating a tsunami of data. The patterns, trends and associations related to the human behavior and interactions can be revealed by collecting and analyzing extremely large datasets. Better understanding of the consumer habits, target marketing campaigns, improved operational efficiency, lower costs and reduced risk can be extracted by using Big data analytics. Market intelligence and information technology advisory services are often provided by International Data Corporation (IDC).

The importance of using more data in order to support the decision on strategies has been realized by various companies. The cases studies have shown that more data usually gives better algorithms. Hence the companies are changing decision to invest more in processing the larger sets of data than investing in expensive algorithms. Larger amount of correlations can be provided by the large quantity of data than if they are analyzed in separate sets or on a smaller set. The processing

limitation is a bottle neck even though large data gives a better output. This project intends to provide the insight of journey of data from raw concepts to Big data Analytics.

2. LITERATURE REVIEW:

The maintenance of records was initially based on non-computer based approaches and effectively replaced by the file processing system. It was used by many organizations since it was a successful System of its time to maintain their data and information. But this approach has a limitation and not suitable for handling the data of big firms and organizations. This approach uses different files to store data.

2.1 Disadvantages of File Processing System

- **Duplicate Data**
  The data may be stored more than once resulting in duplication of the saved data since all the files are independent on each other.
- **Inconsistency**
  A number of copies of same data containing different values can occur in file processing system leading data inconsistency. If data items need change all the files containing that data needs to be modified and may create a risk of outdated values of data.
- **Accessing Anomalies**
  The difficulty in accessing the data in a desired or efficient way results in accessing anomalies making the supervision very difficult. It compels the user to create a program for it in order to extract information in a specific manner.
- **Poor Data Integrity**
  The data meeting certain consistency constraints can give integrated data by its collection. Programmer adds some codes to ensure these constraints. Addition of new constraints at that time is difficult in data processing system resulting in poor data integrity.
- **Poor Data Security**
- File processing system threatens the data security due to poor data security. Any person can easily access and modify the data stored in files. A level of security is required in order to restrict all the users for accessing the data.
- **Atomicity Problem**
  Saving the data values demand atomicity meaning that information is completely entered or cancelled at all. Any system of data storing can fail at any times but the data must be consistent. Any system may fail at any time and at that time it is desired that data should be in a consistent state.
- **Wastage of Labor and Space**
  No organization can afford wastage of the precious labor in this labor costly era. The file processing system needs multiple copies of data resulting in wastage of labor and space.
- **Data Isolation**
  Isolated data in file processing system is stored in different files in different formats. It may create confusion when the data has to be extracted in two files which may not meet the needs and may lack relationship to each other.

2.2 Database Management System

The creation, definition and manipulation of the data can be handled by DBMS software. This software has advantages over routine file processing system and eliminates them.

The data base approach also has disadvantages as follows,

- **Complexity:** The provision of the functionality that is expected of a good DBMS makes the DBMS an extremely complex piece of software. A thorough understanding is required by the Database designers, developers, database administrators and end-users to take full advantage of it. The failures to understand leads to bad design decisions and may have serious consequences for an organization.
- **Size:** DBMS is an extremely large piece of software due to its complexity and breadth of functionality and occupies many megabytes of disk space and demands substantial amount of memory to run efficiently.

- **Performance:** A specific application such as invoicing is being written for a file based system. The performance may be good in such a system. A DBMS is written to more general in order to cater many applications than just one. But those applications may not run as fast as expected.
- **Higher impact of a failure:** The vulnerability of the system is increased by centralization of the resources. The failure of any component can bring operations to a halt since all users and applications rely on variability of DBMS.
- **Cost of DBMS:** The environment and functionality of DBMS decides its cost. It also demands annual maintenance cost.
- **Additional Hardware costs:** The additional space requirements due to disk storage of DBMS necessitate the purchase of additional storage space. Purchase of larger machine may be required to achieve the required performance. It may also demand a dedicated machine for maintaining DBMS. Further expenditure has to be incurred for procurement of additional hardware.
- **Cost of Conversion:** The cost of DBMS and extra hardware can be insignificant when compared to the cost of converting existing applications to new DBMS and hardware. The installation also demands additional training of the staff to use new systems and possibly the employment of specialist staff for conversion and running the system. This is the main reason for some organizations feel tied to their current systems and prevents them to switch for modern technology

2.3 Data warehousing

Data warehouse is developed to provide solution for some of problems of DBMS. Data warehouse is defined as "A subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

In this definition the data is:

• Subject-oriented as the warehouse is organized around the major subjects of the enterprise (such as customers, products, and sales) rather than major application areas (such as customer invoicing, stock control, and product sales). The main purpose of data warehouse is to support decision than the application oriented data.

• Integrated since the source of data may be from different enterprises and different application systems. The source data is often inconsistent using, for example, different formats. It demands the consistency of data in a unified view to the user by using integrated data.

• Time-variant because data in the warehouse is only accurate and valid at some point in· time or over some time interval.

• The data will be non-volatile since the data may not be updated in real time but it is refreshed at regular intervals from a number of data sources. The new data will be added as supplement than a replacement. The new data is continually absorbed in to the database and incrementally integrates with the previous data.

Fig 1: Needs of Data warehouse

2.4 Problems of Data Warehousing

The problems of developing and maintaining the DMS data are as follows

- **Underestimation of resources of data loading**
  The time required to extract, clean and load the data is often underestimated into the data warehouse. This process may consume lot of total development time but some tools have been designed to reduce the time and effort spent on this process.
- **Hidden problems with source systems**
  The hidden data feeding problems associated with the source systems may be identified after years of feeding the data warehouse. For example, the staff may enter the null incomplete property data even when available applicable when entering the details of a new property.
- **Required data not captured**
  The required data may not be captured by the source systems in some circumstances which may be important for the data warehouse purpose. For example, the registration date of a property may not be captured which is important for the subsequent analysis.
- **Increased end-user demands**
  The request for support from staff may increase after satisfying the end users queries than decreasing. The increase in awareness on the capabilities and value of the data warehouse is the main reason for it. Other reason can be due to increase in demands once a warehouse is online the users and queries increase together with the requests for more and more complex queries.
- **Data homogenization**
- Important value of data may be missed since the concept of data warehouse deals with similarity of the data formats.
- **High demand for resources**
  Large amount of data may be required by data warehouse.
- **Data ownership**
  The ownership of the data may change the with the change in attitudes of end users by using Data warehouse. The decision making requires to load the sensitive data owned by one department in data warehouse which may result in reluctance of that department to load due to data breach.
- **High maintenance**
  High maintenance system is the requirement of a data warehouse. The data warehouse is affected by the reorganization of the business processes and the source systems resulting in high cost of maintenance.
- **Long-duration projects**
  It may require three years to build a data warehouse making the organizations reluctant in investigating in to the data warehouse. The historical data captured by department in the data warehousing may end in resultant data marts. These data marts may fulfill the requirements of one department and limited to the functionality of that department or area only.
- **Complexity of integration**
  The most important area for the management of a data warehouse is the integration capabilities. The organization to spend lot of time for determination of wellness of integration of data warehousing tools for overall solution needed. This process makes the task difficult as there is availability of a number of tools for every operation of the data warehouse.

2.5 Data mining

Next concept emerged in the field of data Analytics is Data mining. The data mining helps us to analyze enormous set of the data and get hidden and useful knowledge. The data mining is not only used in the field of business but also helpful in weather forecast, medicine, transportation, healthcare, insurance, government. etc. Data mining is advantageous when used in specific industry but also equally carries disadvantages including breach of privacy, security and misuse of information which are detailed below.

2.6 Disadvantages of data mining.

- **Privacy Issues**
  With the booming of internet along with social networks, e – commerce, forums, blogs, the concerns about the personal privacy are increasing enormously. The people are afraid of furnishing the personal information since they may be used in an unethical manner and resulting in lot of troubles. The purchase behavior trends can be collected

by many ways by the business. But most of business doesn't last longer. Hence one can suspect the sale or leak of personal information.

- **Security issues**
  The security of the data collected is top most important aspect and an important issue. The information of the employees and customers of a business including social security number, birthday, payroll etc. and their storage in a secured manner is important. There are cases including the hackers accessing and stealing the big data of customers from Ford Motor credit company, Sony etc. The situation may get worsen if the data contains much personal and financial information like credit card stole or identity theft becomes a big problem.

- **Misuse of information/inaccurate information**

  The information collected for ethical purposes by using data mining can be misused. This can be exploited by unethical persons or business to reap benefits of vulnerable people or discriminate against a group of people. The data mining is not a perfectly accurate technique which may also contain inaccurate information which can be used for decision making leading to serious consequences.

2.7 Big Data Concept

The term 'Big data' was initially coined by Roger Magoulas from O'Reilly media in 2005 in order to define a great amount of data which cannot be imagined by traditional data management techniques which fails to manage and process due to its complexity and size. A study on its evolution of big data as a research and scientific topic had cited since 1970s and comprises of publications in 2008. Different points of view are being treated regarding Big Data concept by including its implications in many fields. MIKE 2.0 which is open source for information management defines Big data by its size, comprising a large, complex and independent collection of data sets, each with the potential to interact. The inconsistency and unpredictability of the possible combinations makes unable to handle Big data.

In IBM's view Big Data has four aspects:
1. Volume: refers to the quantity of data gathered by a company. This data must be used further to obtain important knowledge;
2. Velocity: refers to time taken in processing Big Data. The efficiency can be maximized by identifying very important and immediate responses activities.
3.Variety: Refers to the type of data that Big Data can comprise where it can be either structured as well as unstructured;
4.Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future.

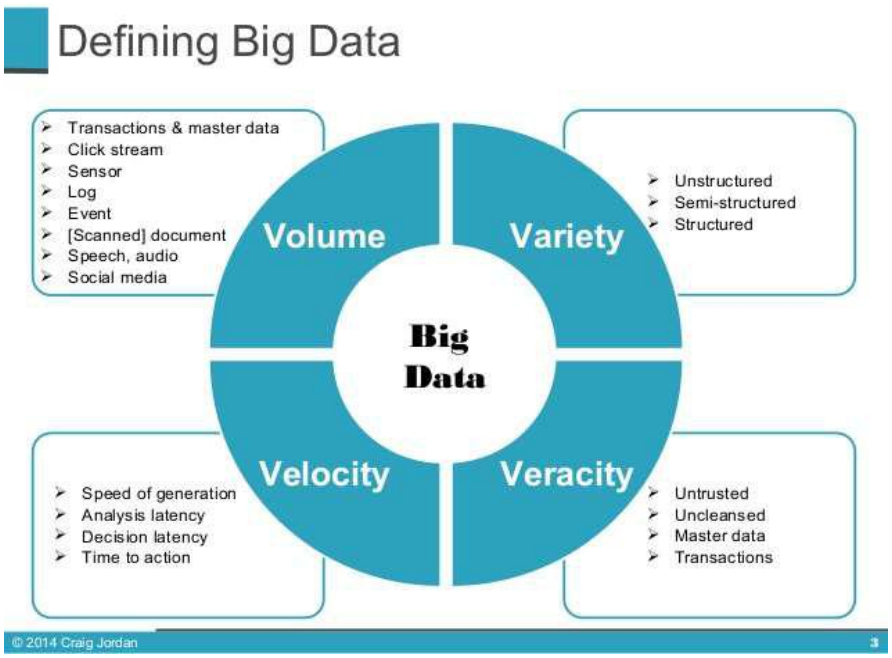

Fig 2: Big data aspects

The Gartner's IT glossary defines the big data as high volume, velocity and variety information assets which demand the cost effective, innovative forms of information processing for enhanced insight and decision making. Chairperson of O' Reilly, Ed Dumbill described big data Strata conference as "data which exceeds the processing capacity conventional database systems. The data is too big, moves too fast or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it" In a simple way, Big data is an expression comprising different data sets of very large, highly complex, unstructured, organized, stored and processed using specific methods and techniques used for business purposes.

2.8 PRODUCTION OF BIG DATA

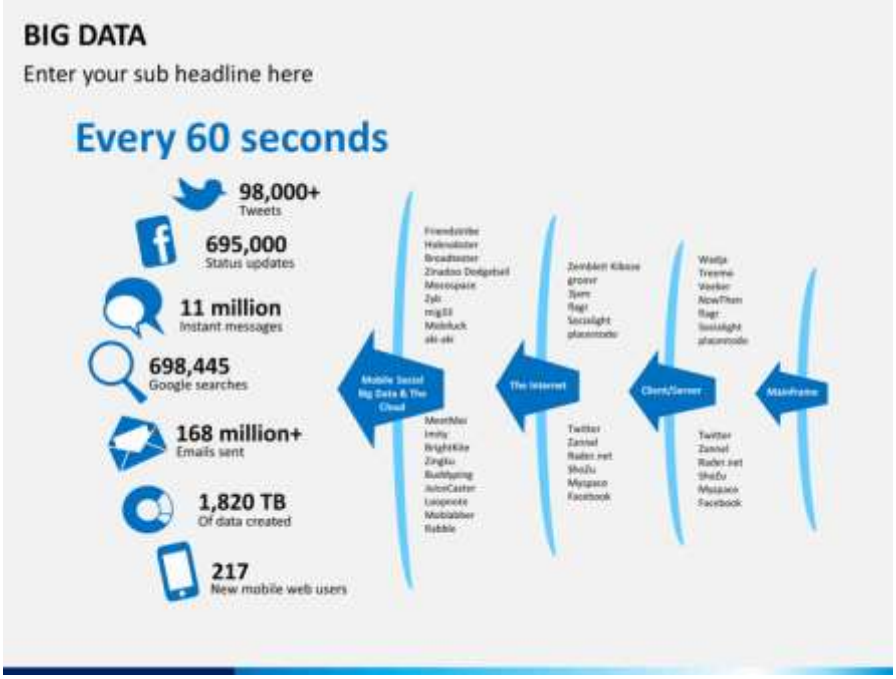

Fig 3: Data generation per minute

Big data is being generated by everything around us at all times including every digital process and social media exchange is producing it. The systems, sensors, mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. We need optimal processing power, analytic capabilities to extract meaningful value from big data. The traditional RDBMS fails to store large volume of data nowadays. So Big data is the problem and Hadoop is the solution. In other words, it can be told as Big data is the issue, Hadoop is the implementation. For example, Google is producing every day data up to more than 12 PB like Face book providing 10 PB, eBay producing 8 PB per day. For storing and processing of large volume of data we need use Hadoop as Framework. Hadoop is a framework used for storing and processing of large volume of data. Whereas traditional RDBMS can only store data, not able to process the data. For this we need to write more complex logic by following any programming Language. It's too tedious to write code for the same.

2.9 Challenges

The main challenge for businesses nowadays is to make best use of wealth of information. The experts working in the field classified big data in to three sub categories.

• Smart data — Information is useful and actionable if it can be organized and segmented according to a company's needs. This type of data can be combined with a number of sources and can be customized to address particular business challenges.

• Identity data — The marketing companies can make best use of profile data combined with social media data, purchasing habits and other behavioral analytics for their marketing campaigns in a precise manner.

• People data — The social media data sets provides such data to help companies to better understanding of their customers as individuals and develop programs to address and anticipate their needs. It seeks to create a shared community of customers with mutual likes, ideas and sentiments.2 Big data sets are so large that traditional processing methods often are inadequate. The challenges include the data analysis, its capture, management, search, sharing, storage, transfer, visualization and privacy protection for big data. As companies work through these data processing and management issues, the focus is shifting to the areas of data strategy and data governance.

3. METHODOLOGY :

I think that certainly is a start. One of the challenges we have is the fact that we haven't been in a position to do this before, as we didn't think it was possible. We always thought it was too expensive. And in many situations, that was absolutely true. It is no longer expensive and making the difference. One need not go out in order to by a multimillion dollar specialized system and we can do this at close to real time. The data warehouse was the traditional system used store data for many years. A separate process can make to mine that data and make assumptions about it. With big data, we can do a lot of things in parallel.

You have to see the data you have as an asset, do some level of intelligent transformation, which is heavily analytical and data-centric, and turn it into something that customers want. That's the start. The visitors can be turned as audiences or sites and properties can be used as inventory. That is what data monetization is all about. A creative analysis can be done with the data in order to create something users want to buy. Whether it's on the cloud or cheap scale-out architectures on commodity microprocessors, one can be regional retailer with 100 stores and build big data platform using x 86 processors, Hadoop, and MPP Postures architecture. This venture does not cost much money. One can spend an evolutionary amount of money in order to capitalize on it. Since day one of commercial computing the cost efficiencies are used in an extensive manner. For most of the organizations spending on IT may be minimal. Only single digits can have spent on IT as a percent of revenue. So let's just call it 2%, which is a number that's relatively well accepted. If I save your company 10% on the technology, I've moved from 2% of revenue to 1.8% of revenue. That's nice, it's important, don't get me wrong. We can create some competitive advantage as a result of it if one takes technology as an investment and find or help to generate new lines of revenue, new products or even more important by moving tens of percent on the top line. When one start looking at that and moving the needle on revenue and profitability just by moving the revenue up, one can naturally move the percent of technology spend down. Understanding business and making the customer happier can be the motto which is very likely to come as use care for big data. All the business persons including an auto manufacturer, a telecom, a retailer often look for opportunities to leverage this data in order to provide better customer, user, shopper or driver experience. Hence big data may be a game changer. But people need not change their business models since it is evolutionary. One can ask them to take advantage of the data in order to provide a much better experience for their consumers, employees, partners and their suppliers. It helps to have right king of creative people who can take what assets and to have look on customers what they are trying to buy and make that transformation. The data can be used as an advantage maybe even to instrument further to get more data. Analysis of the data relied upon at this stage, the results can be used in useful manner since none of the customers are ready to buy the data than insights. They mainly focus on things which are more productive for them or male data easier to use or creates a simpler experience.

Barriers of mid-size companies at the entry

Mindset and developing a Big Data skill set. In terms of mindset, it's sitting down and really thinking through what your customers want is the biggest challenge since the technology is out there. One has to rely on Hadoop as a skill set to do some of this stuff which is true in many cases in terms of skills. This is the major bottle neck for many organizations since they have to re-task the skill sets they have since their people are more familiar with BI and data warehouse. They are well versed with customer thoughts and to interface them. But they don't know to use new tools which allow them to glean new insights out of this data. Hence it is a mindset and retaking the skills challenge to move people from where they are today to give them some new skills. The coding in COBOL has already started. Also they started in FORTRAN, to COBOL, to C++, to Java, and now we're saying you're going to learn how to write Perl and Map Reduce. It's just coding.

Big data is not a technology decision but is a business decision. One can believe the customers and the information w have about them. The customer information exists in every enterprise. The big data helps in ability to mine that and extract and abstract new ways of keeping the customers. IT is not first to initiate the big data but marketing or from sales or from the CFO or the CEO. Hence all companies must understand at an earliest possible point of time and use technology to hold the customers. This is an opportunity to make the money and for the executives to think in

business lines that they have understood the customer in better way. Hence, they got their mindset right and got some bright ideas with their data.

Can big data analytics started in a small way and build from there?

The big data analytics can be a game changer where most of the organizations have started to utilize this. For example, majority of the companies have captured and stored about customer loyalty, transaction data. The organizations have already aggregated the data for reporting purposes, point of sale data from cash registers and sales by store by product category, by product etc. But still they have a detailed data but working off a data warehouse platform that can't handle that detailed data. The organizations are currently using platforms which cannot handle the big data like Green plum. But software's are the means to allow the companies to get in really cheap and grow incrementally without having to take that big giant hurdle of buying some Superdome from HP. They can start small but the thing is to start. Since the competitors are doing something with their data, the most valuable thing by any organization to deal and extract the behavior of the customers from the data they have in the form of transactions and digital data that they have. The big data is goldmine to be mined since it helps the organizations to understand their own customers better and have lot of chance of understanding likeminded customers and prospective customers.

Growing the business by using Big data

The processes adopted by the organizations and their team depends decides the fate of the business. The organizations should conduct a self-assessment before they adopt it. Since the workers in the organization are people who understand the needs and work pattern of the organization. The newly appointed employees may not understand the needs of organizations since they have different business model exposures. Hence there is no need of a new team. The existing team can strengthen their capabilities to adopt the new technology. For example, A major multi-billion dollar manufacturer looked that at their own data in terms of revenue and customer satisfaction, etc. The company wanted a fine deeper metrics and went deeper and found that they actually had more data in field service and field support of their products. A predictive analysis was ordered on failure rates to the point that they could predict with fairly high degree of certainty the possible failures which are likely to happen as a result of wear and tear on thins at their customer sites. It helps them to be preemptive and proactive about maintenance. The organization grew insight from inside with their own people using data they already had.

Starting the new venture

Some very interesting problems by using the stuff can make the organization successful in its implementation. The technology helps the organizations to have an executive sponsor who understands business not only about technology.

It poses a question that the use of big data is a game changer or is it evolution. The companies that can be successful by convincing the IT that its an evolutionary move which helps them to build the assets. But the technology has made to convince the business executives that it is a game changer in reforming the value chain by which they can integrate and service their customers. A competitive business can be built by using the technology.

4. EVALUATION :

DATA STRATEGY Technology has advanced to provide the necessary computing power, memory, storage, software and network capabilities to handle vast amounts of data. Companies are beginning to realize the promise of better analytics, increased accuracy and greater confidence in decision making. Data strategies are focusing on the quality of the data and identifying what information can drive better performance, reduce risk, and predict customer behaviors. The accuracy and trustworthiness of data have assumed greater value for yielding reliability and results. An important component of a company's data strategy is predictive analytics. Raw data in a company's customer relationship management systems and other databases can be combined and analyzed to create useful insights into customer behavior and to predict future behavior and trends. This can help a company to improve its operations and performance through better investments and strategic decisions.

5. CONCLUSION :

In a recent survey of CIO conducted for EMC by CIO Magazine, almost half the respondents said they agreed with this statement: "Big Data analytics is an evolution but not a revolution in the area of data warehousing, databases, and big file systems." The rest were about evenly divided among "game changer," "pipedream," "and not sure." What's up with that? Isn't Big Data changing the world?

Big data is not a new technology. It's actually just an improvement on the old. So the people who answered the survey are not entirely wrong in having that position. Nothing about big data and its processes are anything revolutionary from that perspective. What is revolutionary and game-changing is that for the first time we have the necessary compute power, the necessary memory and network, the storage, and most important, the necessary software to actually consider the entire set of data

It's evolutionary in the technology, but it's game-changing how it's applied to the business. That's what makes this so interesting, that we're not talking game-changing from a technology perspective. That is, leverage your BI and data warehouse assets to get more out of them. We're talking game-changing in how you deploy it at the point of customer engagement.

The challenge is that companies that don't do this will be out of business. Big Data is an evolutionary game-changer, where companies are figuring out, how do I bring data into my product to make it more effective, more productive, more relevant to the user?

Big data sets are so large that traditional processing methods often are inadequate. Big data challenges data analysis, capture, management, search, sharing, storage, transfer, visualization and privacy protection.3 As companies work through these data processing and management issues, the focus is shifting to the areas of data strategy and data governance.

REFERENCES:

1. "The World's Technological Capacity to Store, Communicate, and Compute Information". MartinHilbert.net. Retrieved 13 April 2016.
2. New Horizons for a Data-Driven Economy – Springer. doi:10.1007/978-3-319-21569-3.
3. Crawford, Kate (September 21, 2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.
4. "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.
5. Hilbert, Martin; Lopez, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information". Science. **332** (6025): 60–65. doi:10.1126/science.1200970. PMID 21310967.
6. "IBM what is big data? – Bringing big data to the enterprise". Www.ibm.com. Retrieved 2013-08-26.
7. Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", December 2012
8. Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue.
9. Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". Release 2.0. Sebastopol CA: O'Reilly Media (11).
10. John R. Mashey (25 April 1998). "Big Data ... and the Next Wave of InfraStress"(PDF). Slides from invited talk. Usenix. Retrieved 28 September 2016.