A pathway to Big Data Analytics from Information age to Data age : A Study

U. Padma Mohan , S. Sagar

Abstract : With the widespread use of databases and their exponential growth in their sizes, today organizations are faced with a problem of information overload. Enterprises are increasingly looking to find actionable insights into their data. Business Intelligence uses descriptive statistics with data with high information density to detect trends, uncover patterns and relationships. The techniques of capturing raw data at sources, storing and analyzing the data for corporate decision making from these volumes of data has activated the development of Big Data Technology. This paper presents theoretical overview of the tools , techniques, challenges and applications of Big data.

Index terms – business intelligence, datasets, data warehouse, data analysis, framework , Iot, ,knowledge, machine learning, predictive analysis.

1 Introduction :

Today we are living in a " Data Age " , deluged by terabytes and petabytes of data from diverse sources. Data is generated from Business, Medicine, Engineering , Telecommunication , Health Industry, Banking , Web data, e-commerce, social media, , search engines and so on..

- Google processes 20 PB of data in a day (2008)
- Face book has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN's Large Hydron Collider (LHC) generates 15 PB a year

All enterprises are benefitted by this explosive and tremendous data.

Table -1 Data generation points example

Mobile Devices
Microphones
Readers / Scanners
Science Facilities
Programs / Software
Social Media
Camera
Medical devices
POS machines
Kiosks

Table-2 Big Data: a growing Torrent

5 Billion mobile phones in use -2010
30 Billion pieces of content is shared on Face book every month
40% projected growth in global data generated per year
5 % growth in global IT spending
235 Terabytes of data collected by the US library of congress-2011
15 out of 17 sectors in US have more data stored per company than the US library of congress

1.1 How this data is used by Enterprises?

- Aggregation and Statistics : Data warehouse and OLAP
- Indexing, Crawling, Searching, Page ranking, and Querying : Keyword based search and Pattern matching (XML/RDF)
- Knowledge discovery : Data Mining and Statistical Modeling

This unbridled growth in data has necessitated for new techniques and powerful and versatile automated tools to uncover valuable information and to transform the data into organized " Knowledge " . Consequently important and strategic decisions are made by organizations from these golden " nuggets " of Knowledge.

1.2 Business Intelligence :

Business Intelligence is a framework of concepts, tools, technology, and practices that help a business understand its core capabilities, its customers, competitors, business partners, competitive environment and its own internal operations.BI provides snapshot of the company's situation, and identify key opportunities to create competitive advantages. Data is managed as a corporate asset . The comprehensive and decision support system (DSS) within organizations is BI. BI is about creating Intelligence about a business. Monitoring a business' health is crucial to know where the company stands and what its future is going to be. The organizations use a special type of metrics known as Key Performance Indicators (KPI) that assesses the company's effectiveness or success in reaching its strategic and operational goals.

1.2.1 BI Architecture : From the point of data acquisition to storage, transformation, integration, analysis, monitoring, presentation, and archiving that is the entire data life cycle of data, BI covers a range of technologies. There is no single BI Architecture**.** BI integrates people

and processes using technology to add value to the business. BI tools focus on the strategic and tactical use of information.

1.2.2 Techniques Used : Business Intelligence uses a confluence of techniques like Statistical, Database System, Data warehouse , Information Retrieval, Machine Learning, Pattern recognition, Visualization and High performance computing.

2 Big Data : Data mining is going through a shift in paradigm which eventually resulted in the birth of Big Data. It deals with volumes of data which are usually complex to process with the existing application software.

Models used in Big Data : Inductive statistical methods and non-linear regression models are used in big data to predict outcomes and behaviors from large sets of data with low information density to reveal relationships and trends.

The five Vs : Big Data is associated with five V's : *volume*, *variety*, *velocity* *veracity* and *value.* Big data is defined with these five Keywords as :

Volume : All enterprises capture data from a variety of sources like sensor, mobile devices, social media, software, POS etc . Managing these massive volumes of data thus collected is a herculean task, but latest technologies such as *Hadoop* have eased the burden.

Velocity : The speed at which data is streaming into the database is unprecedented .This data has to be dealt with, in a time frame. RFID tags, sensors and smart metering are necessitating the need to deal with torrents of data in realty.

Variety : Data arrives in a variety of formats – from structured, numeric data in traditional databases to unstructured text documents, images, sequences, email, video, audio, stock ticker data and financial transactions.

Veracity : Veracity of data questions the correctness, accuracy and trustworthiness of the data . For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc. and the reliability and accuracy of all that content.

Value : When we talk about value, it is about the worthiness of the tons and tons of data being extracted. This data is to be transformed into potentially useful information. The costs and benefits of collecting and analyzing the data to ensure that ultimately the data that is reaped can be monetized.

The Big Data environment has the following components:

- Techniques for analyzing data, such as machine learning and natural language processing
- Big data technologies, like Business Intelligence, Cloud computing and Databases.
- Visualization, such as charts, graphs, tables, crosstabs.

3 The Need for Big Data Analysis

Managers are looking for a competitive advantage. They understand that business climate is dynamic. In a complex business environment it is desired to create a support system dedicated to quick decision making. Big Data Analytics is the process of collecting, storing , organizing and analyzing large sets of data to uncover patterns and knowledge. Big Data analytics can help organizations to interpret the information within the data and will also aid in identifying the data that is most useful to the business needs and futuristic decision making. Business Analysts working with Big Data are usually interested in the *knowledge* that comes from analyzing the data. Some specialized Software Tools and Applications are used in Big Data analytics to process extremely large volumes of data for Predictive Analytics, Text Mining, Forecasting and Data Optimization. These processes are independent but highly coherent and integrated functions of high-performance analytics.

4 How Big Data Analytics is used Today

: As the technology that helps an organization to break down data silos and analyze data, it improves business, and can be transformed in all sorts of ways. Today's advances in analyzing big data allow researchers to :

- Decode human DNA in minutes.
- Determine which gene is most likely to be responsible for certain diseases.
- which ads you are most likely to respond to on Facebook.
- To improve public health and safety ,by tracking cell phone data and trace their details in case of emergency .

5 The Benefits of Big Data Analytics

As there are many big data analytics platforms in place there are innumerable benefits .Due to the growing needs of businesses and to answer multiple business queries an enterprise can benefit from big data analysis in many fields :

5.1 Media

- Data journalism: publishers and journalists use big data tools to provide unique and innovative insights and info graphics.
- For marketing and advertising.
- Channel 4, the British public-service television broadcaster, is a leader in the field of big data and data analysis.

5.2 Internet of Things

Big data and the IoT work in conjunction. IoT collects sensory data and this sensory data has been used in medical , industry and manufacturing .

5.3 Retail

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes of data.
- In businesses big data helps a lot in knowing the shopping behaviors and spending appetites of customers.

5.4 Science

- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.
- Google's DNA Stack compiles and organizes DNA samples of genetic data from around the world to identify diseases and other medical defects.

5.5 Sports

- Big data uses sports sensors to enhance sports training and understanding competitors, it is also possible to predict winners in a match .
- In Formula One races, race cars with hundreds of sensors generate terabytes of data. Besides using big data, race teams try to predict the time they will finish the race.

5.6 Technology

- **e-Bay** uses two data warehouses at 7.5 petabytes and 40 PB, as well as a 40 PB Hadoop cluster for search, consumer recommendations, and merchandising.

- **Amazon** handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers.
- **Google** was handling roughly 100 billion searches per month as of August 2012.
- **Face book** handles 50 billion photos from its user base.

6 Industries propelled by Big Data Analytics are –

1. Healthcare Industry
2. Public Sector Services.
3. Learning Services.
4. Insurance Services.
5. Industrialized and Natural Resources.
6. Private Industry
7. Transportation Services.
8. Banking Sectors

6.1 Healthcare

Medical devices Today's Medical devices are Big Data oriented like BP / Temperature / Heartbeat / Pulse monitoring systems. Data is also captured from Scanning / X-Ray , ECG machines and doctors prescribe drugs even for remote patients who are having these devices fitted on to their body.

Hospitals Global hospitals collect patient data which is viewed , shared and accessed by Researchers and Doctors for Data Mining and Predictive Analysis.

Diseases Also helps in disease prevention and treatment.

6.2 Public Sector

Governments generate and capture huge quantities of data through their day-to-day activities, such as power investigation, Infrastructure, Financial data, pension disbursement, payments, tax collection, national health systems, traffic data, and issuing official document. The public sector units are increasingly in demand for technological requirements and gain the potential value from Big Data . Some of the benefits are as follows:

- **Tax Agencies** These agencies have large datasets both structured and unstructured which are integrated from multiple sources like social media and other sources. Few automated algorithms are used to analyze and validate the information. It also helps in fraud detection.
- **Economic Analysis** Correlation of multiple sources of data will help government economists with more accurate financial forecasts and budgetary control.
- **Open Data**: Organizations can build trust and can create transparency between citizens and government by free flow of information in line with Data Initiatives

6.3 Learning Services

Big data has great influence in the education world as almost every courseware is online. Applications for online Tests, Assessments are done through the cameras of the mobile phones.

6.4 Insurance Services

The big data enables for the better purchase of policy from insurance agencies by the extrapolative big data. Business Analytics has been used in claims, administration and to provide rapid service in scam discovery.

6.5 Industrial and Natural Resources

In the natural wealth industry big data enables for analytical modeling to sustain judgment creation that is used to consume and incorporate huge amounts of information from geographical information, graphical information, manuscripts and chronological statistics.

6.6 Private Sector

Private organizations are also benefitted by Big Data :

- **Consumer product** Manufacturers and Retailers use big data in increasing sales and consumer loyalty.
- **Logistics** : Private sector uses the big data in revenue administration, industrial improvements and logistics.

6.7 Transport and Navigation

In transport management, drivers choose efficient delivery routes and traffic management in peak hours.

6.8 Banking Sectors and Fraud Detection

Big data is vastly used in detecting the fraudulent usage of credit and debit cards in the banking and financial sectors.

7 Big Data Analytics : Challenges

Big data challenges include data capturing ,data storage, data analysis, sharing, transfer, visualization, querying, analysis, search, updating, information privacy and data source. Today, data is generated at lightning speed. Every minute sees production of huge amounts of data. Big companies are struggling to find ways and means to find insights into their data. However, this is not an easy task. The amount of data produced makes it very difficult to store, manage, analyze and utilize it. The development of various big data analysis tools has helped with data handling to a great extent. The main challenges faced in Big Data Analysis are as follows :

7.1 Data Storage and Quality

Companies and organizations are growing at a very fast pace. Moreover, the growth of the companies rapidly increases the amount of data produced. It is a challenge to manage the enormous number of sources that are producing the data. The data comes from the company's internal sources like finance, marketing etc. Moreover, external sources like social media produce a huge amount of data therefore, making the data extremely diverse and massive.

The storage of this data is becoming a challenge for everyone. Repositories like **Data Lakes** and **Data Warehouses** are used to collect and store massive quantities of unstructured data in its native format. The problem, however is when a data lake/ warehouse try to

combine inconsistent data from disparate and heterogeneous sources, it encounters problems. Inconsistent, duplicate missing and redundant data causes logic conflicts, and result in 'Dirty Data'.

7.2 People who understand Big Data Analysis

It is very important to transform the huge amount of data being generated every minute into useful information. Therefore, there is a huge need for Big Data Analysts and Data Scientists. This is another challenge faced by companies. The number of data scientists available is very less in comparison to the amount of data being produced.

7.3 Good Quality Analysis

Companies and organizations use big data produced ,to make the best decisions possible. Consequently, the data they are using should be accurate, else it will result in ill-advised decisions that would ultimately be detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue. This requires a lot of resources to ensure the accuracy of the information provided. The process of creating accurate data is very time consuming and requires the use of tools that can be expensive.

7.4 Security and Privacy of the Data

When it comes to the security and the privacy of the data, it also involves big risks . The tools used for analysis stores, manages, analyzes, and utilizes the data from a different variety of sources. This ultimately leads to a risk of exposure of the data, making it highly vulnerable to data leaks and theft of data. Therefore, the production of more and more data increases security and privacy concerns.

8 Technologies Adapted by Big Data

As the world becomes more information-driven than ever before, a major challenge is how to deal with the explosion of data. The mechanical production of data forces us to adapt new strategies and tools for in-depth analysis and present knowledge which is easily human comprehensible. The Technologies used in Big Data environment are as follows:

- Hadoop
- NoSQL Database Systems
- MapReduce
- Massively parallel computing

Hadoop :

Apache Hadoop is a highly scalable storage platform designed to process very large data sets across hundreds to thousands of computing nodes that operate in parallel. It provides a cost-effective storage solution for large data volumes with no specific format requirements. Hadoop is a software ecosystem that allows for massively parallel computing. A distributed parallel architecture distributes data across multiple servers: these parallel execution environments can dramatically improve data processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and NoSQL databases.

Map Reduce:

A computational model that basically takes intensive data processes and spreads the computation across a potentially endless number of servers (generally referred to as a Hadoop cluster). It has been a game-changer in supporting the enormous processing needs of big data. A large data procedure which might take 20 hours of processing time on a centralized relational database system, may only take 3 minutes when distributed across a large Hadoop cluster of commodity servers, all processing in parallel. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.

MapReduce is a framework using which we can write applications to process huge amounts of data in parallel, on large clusters of commodity hardware in a reliable manner. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely :

Map. : Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

Reduce : Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

The Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides.
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

  Map stage : The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

  Reduce stage : This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.

- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

MapReduce, is core of Hadoop .It is the programming paradigm that allows for this massive scalability. The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. Hadoop has two main components - HDFS and YARN

NoSQL (Not Only SQL) :

It is a completely different framework of databases that allows for high-performance, agile processing of information at massive scale. In other words, it is a database infrastructure that has been very well-adapted to the heavy demands of big data. NoSQL distributed databases (such as HBase) can allow for data to be spread across thousands of servers with little reduction in performance

The efficiency of NoSQL can be achieved because unlike relational databases that are highly structured, NoSQL databases are unstructured in nature, trading off stringent consistency requirements for speed and agility. NoSQL centers around the concept of distributed databases, where unstructured data may be stored across multiple processing nodes, and often across multiple servers. This distributed architecture allows NoSQL databases to be horizontally scalable as data continues to explode, just add more hardware to keep up with no slowdown in performance. The NoSQL distributed database infrastructure has been the solution to handling some of the biggest data warehouses on the planet – i.e. the likes of Google, Amazon, and the CIA.

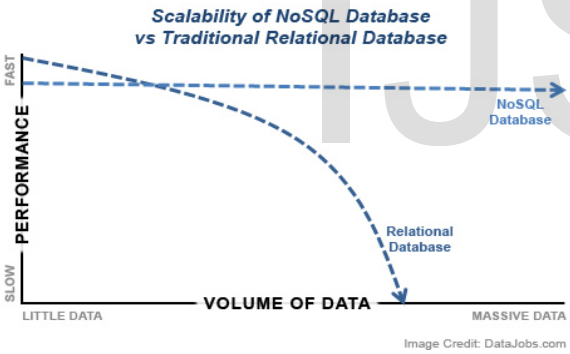

Fig-1 showing the performance of Nosql database and RDBMS

. **Massively Parallel Processing** :

Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data mining, distributed file systems, distributed databases, cloud and HPC-based infrastructure (applications, storage and computing resources) and the Internet .

9.1 Tableau Public

Tableau is an interactive Business Intelligence tool for visually analyzing the data which depict the trends, variations, and density of the data in the form of graphs, charts and dashboards. These dashboards are shareable and interactive. Tableau can connect to files, databases and big data sources to acquire and process data. The software allows data blending and real-time collaboration, which makes it very unique.

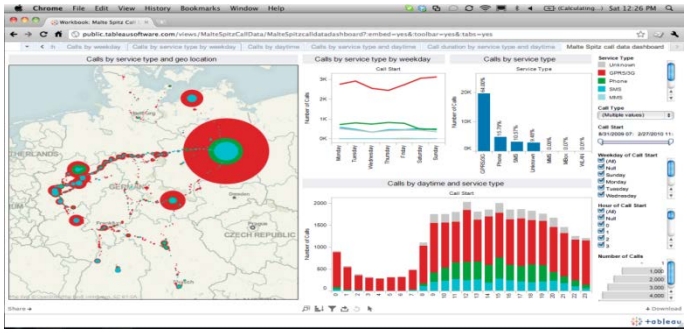

Fig-.2 Tableau Public's data Visualization

9.1.1 Uses of Tableau Public

- Interactive data visualizations can be published to the web and can be embedded into blogs and web pages.
- No programming skills required.
- Shared through email or social media and the content is available for downloads.
- Million row limit, which is easy to use.

9.1.2 Limitations of Tableau Public

- All data is public and offers very little scope for restricted access.
- Data size limitation.
- Cannot be connected to R.
- The only way to read is via OData sources, i.e Excel or txt.

9.2 RapidMiner

RapidMiner provides Machine Learning procedures and data mining including Data Visualization, processing, statistical modeling, deployment, evaluation, and predictive analytics. RapidMiner written in Java is fast gaining acceptance as a Big Data Analytics tool.

9 Data Analytics Tools

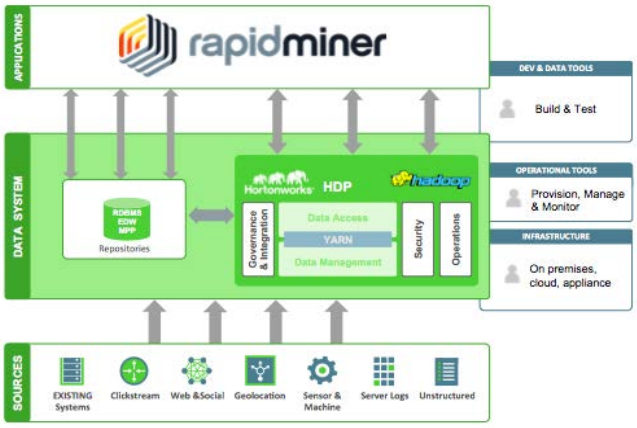

Fig-3 Rapidminer's integrated environment for business analytics, predictive analysis, text mining, Data Mining, and Machine Learning.

9.2.1 Uses of RapidMiner

Along with commercial and business applications, RapidMiner is also used for application development, rapid prototyping, training, education, and research.

9.2.2 Limitations of RapidMiner

- RapidMiner has size constraints with respect to the number of rows.
- For RapidMiner, you need more hardware resources than ODM and SAS.

9.3 Google Fusion Tables

Google spreadsheets are cooler, larger, and nerdier .It is an incredible tool for data analysis, mapping, and large dataset visualization.

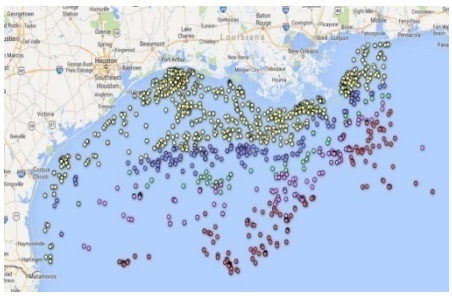

Fig-4 Google fusion tables showing Data analysis, mapping, and large dataset visualization

9.3.1 Uses of Google Fusion Tables

- Visualize bigger table data online.
- Filter and summarize across hundreds of thousands of rows.
- Two or three tables can be merged to generate a single visualization that includes sets of data.
- Maps can be created in minutes

9.3.2 Limitations of Google Fusion Tables

- Only the first 100,000 rows of data in a table are included in query results or mapped.
- The total size of the data sent in one API call cannot be more than 1MB.

9.4 NodeXL

It is a visualization and analysis software of relationships and networks. NodeXL provides exact calculations. It is one of the best statistical tools for data analytics which includes advanced network metrics, access to social media network data importers, and automation.

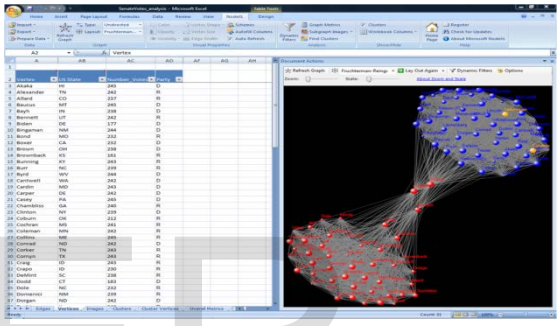

Fig-5 Graph Visualization and graph Analysis

9.4.1 Uses of NodeXL : This data analysis tools in excel that helps in following areas:

1. Data Import
2. Graph Visualization
3. Graph Analysis
4. Data Representation

- This software integrates into Microsoft Excel .It opens as a workbook with a variety of worksheets containing the elements of a graph structure like nodes and edges.
- This software can import various graph formats like adjacency matrices, Pajek .net, UCINet .dl, GraphML, and edge lists.

9.4.2 Limitations of NodeXL

- You need to use multiple seeding terms for a particular problem.
- Running the data extractions at slightly different times.

9.5 WEKA : WEKA is a Data mining machine learning tool. It is a collection of machine learning algorithms .The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression,

clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Data can be imported from a file in various formats : ARF , CSV , C4.5 , binary.

Data can also be read from a SQL database using ODBC / JDBC . Preprocessing tools in WEKA are called filters and these filters are used for Discretization, Normalization , Resampling , Attribute Selection and Transformation .

9.5.1 Main features of WEKA :

- 49 preprocessing tools
- 76 classification algorithms
- 8 clustering algorithms
- 15 attribute / subset evaluation and 10 search algorithms
- 3 algorithms for association

9.5.2 WEKA Application Graphical User Interface :

- **The Explorer** : preprocessing, attribute selection, learning and visualization of data.
- **The Experimental** : Testing and evaluating machine learning algorithms
- **The Knowledge flow** : The visual design of KDD process.
- **Simple command line** : A simple interface for typing commands.

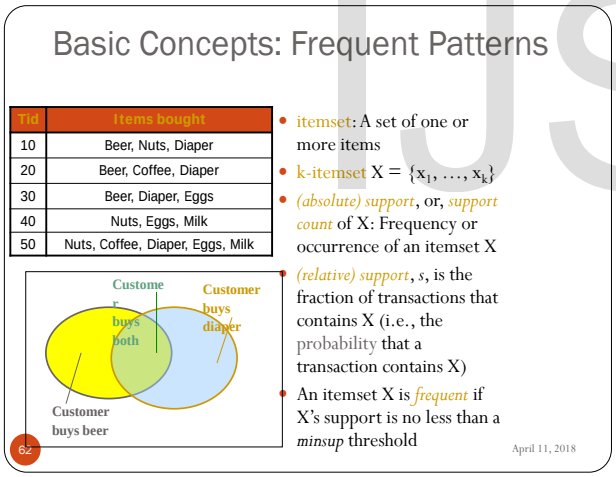

Fig-6 Association analysis example

Conclusion

Big Data is a great boon to various companies and organizations, as it is helping them take better decisions, thus profiting the company. The use of this data to the best of its abilities, however, remains a dream.

The Bottom Line

As big data continues down its path of growth, there is no doubt that these innovative approaches – utilizing NoSQL database architecture and Hadoop software will be central to allowing companies reach full potential with data. Additionally, this rapid advancement of data technology has sparked a rising demand to hire the next generation of technical geniuses who can build up this powerful infrastructure. The cost of the technology and the talent may not be cheap, but for all of the value that big data is capable of bringing to table, companies are finding that it is a very worthy investment.

REFERENCES :

[1] Data Mining concepts and Techniques- Jaiwei Han Micheline Kamber Jian Pei

[2] Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data : Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch",

[3] Hadoop: The Definitive Guide by Tom White

[4] Hadoop in Action by Chuck Lam

[5] Hadoop in Practice by Alex Holmes

[6 Mining of massive datasets, Anand Rajaraman, Jeffrey D Ullman

[7] www.zapmeta.com

[8] www.eduoni.com

[9] www.qubole.com

[10] www.datamation.com

[11] www.researchgate.net

[12] www.ngdata.com