Sub:	**Big Data Analytics**							**Sub Code: 20MCA352**
Date:	**13/11/2021**		**Max Marks:**	**50**		**Sem:**	**III**	**Branch:MCA**
Q.No	**Description**							**Max Marks**

1 **Describe any four characteristics (4 Vs) of big data and discuss the application of big data analytics**. 10

Characteristics of Big Data:
1. Volume
2. Velocity
3. Variety
4. Value

Volume: The main characteristic that makes data “big” is the sheer volume. It makes no sense to focus on minimum storage units because the total amount of information is growing exponentially every year.
Variety: is one the most interesting developments in technology as more and more Information is digitized. Traditional data types (structured data) include things on a bank statement like date, amount, and time.
Velocity is the frequency of incoming data that needs to be processed.
Value: Analysis add value to your business is measured.

Example Applications

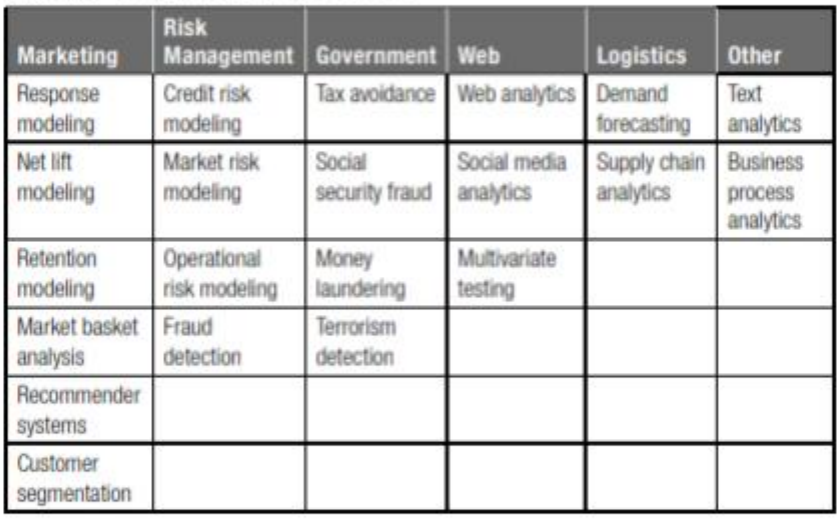
Analytics is everywhere and strongly embedded in our daily lives.

The relevance, importance and impact of analytics are now bigger than ever before and, given that more and more data are being collected and that there is strategic value in knowing what is hidden in data, analytics will continue to grow.

Physical mail box: a catalogue sent to us through mail most probably as a result of a response modeling analytical exercise that indicated, given my characteristics and previous purchase behaviour, we are likely to buy one or more products from it.

Behavioral Scoring Model: Checking account balance of the customer from the past 12 months and credit payments during that period, together with other kinds of information available to the bank, to predict whether a customer will default on the loan during the next year.

Social Media: As we logged on to my Facebook page, the social ads appearing there were based on analyzing all information (posts, pictures, my friends and their behaviour, etc.) available to Facebook. Twitter posts will be analyzed (possibly in real time) by social media analytics to understand both the subject tweets and the sentiment of them.

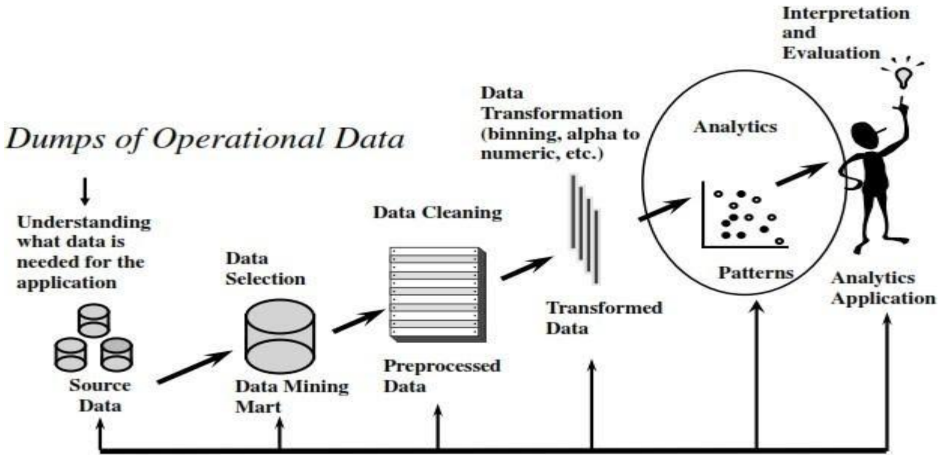


Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

2 **With a neat diagram, explain the working of analytical processing model.** 10

1. Define the business problems to be solved using analytics.
2. All source-data need to be identified that could be of potential interest. (Select of data will have a deterministic impact on the analytical model).
3. All data will be gathered in a staging area which could be data mart/ data warehouse. Basic exploratory analysis will be considered, for example: OLAP (Online Analytical Processing) facilities for multi-dimensional data analysis.
4. Data Cleaning steps to get rid of all inconsistencies – like missing data / values, outliers and duplicate data. Additional transformations may also be considered, such as binning, alphanumeric to numeric coding, geographical aggregation etc.
5. In the analytics steps, an analytical model will be estimated on the pre-processed and transformed data. Once the model is built, it will be interpreted and evaluated by the business experts. Many trivial patterns will be detected by the model.

Knowledge Pattern: Unexpected yet interesting and actionable patterns (referred to as knowledge pattern). Once analytical model has been appropriately validated and approved, it can be put into production as an analytical application.


Dumps of Operational Data
Understanding what data is needed for the application
Source Data
Data Selection
Data Mining Mart
Data Cleaning
Preprocessed Data
Data Transformation (binning, alpha to numeric, etc.)
Transformed Data
Analytics
Patterns
Interpretation and Evaluation
Analytics Application

3 **Mention and elaborate the different types of data sources for big data analytics and its data elements.** 10

Types of Data Sources

Data can originate from a variety of different sources. They are as follows:

- Transactional data
- Unstructured data
- Qualitative/Export based data
- Data poolers
- Publicly available data

Transactional Data: Transactions are the first important source of data. Transactional data consist of structured, low level, detailed information capturing the key characteristics of a customer transaction (e.g., purchase, claim, cash transfer, credit card payment). This type of data is usually stored in massive online transaction processing (OLTP) relational databases. It can also be summarized over longer time horizons by aggregating it into averages, absolute/relative trends, maximum/minimum values, and so on.

Unstructured data: Embedded in text documents (e.g., emails, web pages, claim forms) or multimedia content can also be interesting to analyze. However, these sources typically require extensive pre-processing before they can be successfully included in an analytical exercise.

Qualitative/Expert based data: Another important source of data is qualitative, expert based data. An expert is a person with a substantial amount of subject matter expertise within a particular setting (e.g., credit portfolio manager, brand manager). The expertise stems from both common sense and business experience, and it is important to elicit expertise as much as possible before the analytics is run. This will steer the modelling in the right direction and allow you to interpret the analytical results from the right perspective. A popular example of applying expert based validation is checking the univariate signs of a regression model. For example, one would expect a priori that higher debt has an adverse.

Data poolers: Nowadays, data poolers are becoming more and more important in the industry. Popular examples are Dun & Bradstreet, Bureau Van Dijck, and Thomson Reuters. The core business of these companies is to gather data in a particular setting (e.g., credit risk, marketing), build models with it, and sell the output of these models (e.g., scores), possibly together with the underlying raw data, to interested customers. A popular example of this in the United States is the FICO score, which is a credit score ranging between 300 and 850 that is provided by the three most important credit bureaus: Experian, Equifax, and TransUnion. Many financial institutions use these FICO scores either as their final internal model or as a benchmark against an internally developed credit scorecard.

Publicly available data: Finally, plenty of publicly available data can be included in the analytical exercise. A first important example is macroeconomic data about gross domestic product (GDP), inflation, unemployment, and so on. By including this type of data in an analytical model, it will become possible to see how the model varies with the state of the economy. This is especially relevant in a credit risk setting, where typically all models need to be thoroughly stress tested. In addition, social media data from Facebook, Twitter, and others can be an important source of information. However, one needs to be careful here and make sure that all data gathering respects both local and international privacy regulations.

Types of Data Elements

It is important to appropriately consider the different types of data elements at the start of the analysis. The different types of data elements can be considered:

- Continuous
- Categorical

Continuous: These are data elements that are defined on an interval that can be limited or unlimited. Examples include income, sales, RFM (recency, frequency, monetary).

Categorical: The categorical data elements are differentiated as follows:

Nominal: These are data elements that can only take on a limited let of values with no meaningful ordering in between. Examples: marital status, profession, purpose of loan.

Ordinal: These are data elements that can only take on a limited set of values with a meaningful ordering in between. Examples: credit rating; age coded as young, middle aged, and old.

Binary: These are data elements that can only take on two values. Example: gender, employment status.

Appropriately distinguishing between these different data elements is of key importance to start the analysis when importing the data into an analytics tool. For example, if marital status were to be incorrectly specified as a continuous data element, then the software would calculate its mean, standard deviation, and so on, this is obviously meaningless.

4 **List the various factors required for analytics model and explain its importance.**

A good analytical model should satisfy several requirements, depending on the application area. 10

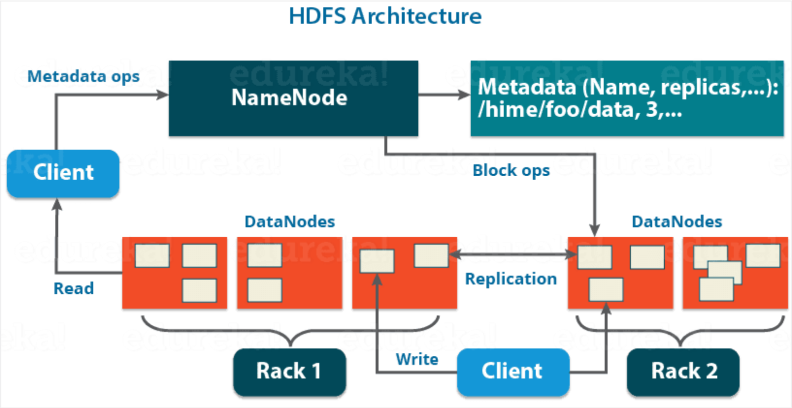
- A first critical success factor is business relevance. The analytical model should actually solve the business problem for which it was developed.
- A second criterion is statistical performance. The model should have statistical significance and predictive power. Example: in a classification setting (churn or fraud) the model should have good discriminative power.
- Depending on the application, the model should also be interpretable and justifiable. Interoperability refers to understanding the pattern and justifiability refers to the degree to which model corresponds to prior business knowledge and institution
- Analytical models should also be operationally efficient. This refers to the efforts needed to collect the data, preprocess it, evaluate the model, and feed its outputs to the business application. Example: campaign management, capital calculation etc…
- Another key attention point is the economic cost needed to set up the analytical model. This includes the cost to gather and process the data, cost to analyze the data and cost to put the resulting analytical models into production
- Finally, analytical models should also comply with both local and international regulation and legislation. Example: Credit risk setting.

5 **Discuss the critical or core components of Hadoop and their working along with a neat diagram.** 10

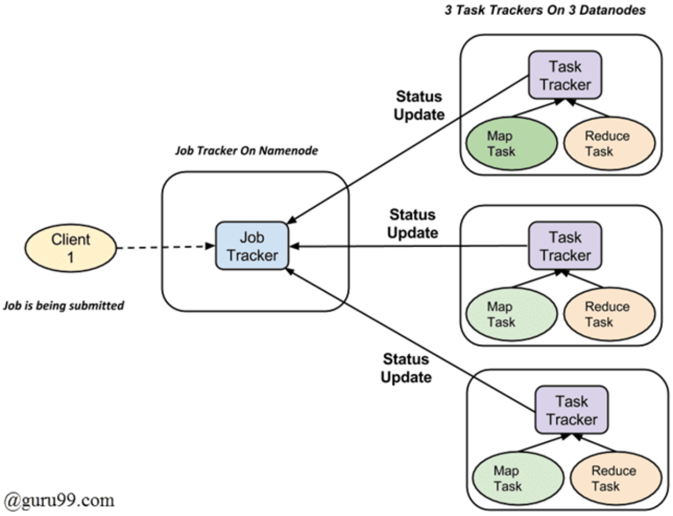
The two critical components of Hadoop are:

- The Hadoop Distributed File System (HDFS)
- MapReduce

The Hadoop Distributed File System (HDFS): HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.



MapReduce: Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the agent that distributes the work and collects the results.



Both HDFS and MapReduce are designed to continue to work in the face of system failure. HDFS continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails, or data is damaged, whether due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster. Likewise, when an analysis job is running, MapReduce monitors progress of each of the servers participating in the job. If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data.

Because of the way that HDFS and MapReduce work, Hadoop provides scalable, reliable, and fault-tolerant services for data storage and analysis at very low cost. The working of DFS and MapReduce is shown in the figure below.

6 **What is predictive analytics? Why are they required? Discuss the leading trends of predictive analytics with examples.** 10

To master analytics, enterprises will move from being in reactive positions (business intelligence) to forward leaning positions (predictive analytics). Using all the data available-traditional internal data sources combined with new rich external data sources-will make the predictions more accurate and meaningful. Because analytics are contextual, enterprises can build confidence in the analytics and the trust will result in using analytics insight to trigger business events.

Leading trends that are making their way to the forefront of businesses today:

- Recommendation engines similar to those used in Netflix and Amazon that use past purchases and buying behavior to recommend new purchases.
- Risk engines for a wide variety of business areas, including market and credit risk, catastrophic risk, and portfolio risk.
- Innovation engines for new product innovation, drug discovery, and consumer and fashion trends to predict potential new product formulations and discoveries.
- Customer insight engines that integrate a wide variety of customer- related info, including sentiment, behavior, and even emotions. Customer insight engines will be the backbone in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetime value, optimizing marketing campaigns for revenue lift, and targeting individuals or companies at the right time to maximize their spend.
- Optimization engines that optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales, such as when, where, and how to seek natural resources to maximize output while reducing operational costs or what potential competitive strategies should be used in a global business that takes into account the various political, economic, and competitive pressures along with both internal and external operational capabilities.

7 **List and explain the technical features of Hadoop.** 10

Hadoop is an open source, Scalable, and Fault tolerant framework written in Java. It efficiently processes large volumes of data on a cluster of commodity hardware.

Hadoop is not only a storage system but is a platform for large data storage as well as processing.
1. Open Source Apache Hadoop is an open source project. It means its code can be modified according to business requirements.
2. Distributed Processing As data is stored in a distributed manner in HDFS across the cluster, data is processed in parallel on a cluster of nodes.
3. Fault Tolerance By default 3 replicas of each block is stored across the cluster in Hadoop and it can be changed also as per the requirement. So if any node goes down, data on that node can be recovered from other nodes easily. Failures of nodes or tasks are recovered automatically by the framework. This is how Hadoop is fault tolerant. Department of Computer Applications- CMR Institute of Technology-Even Sem 2018
4. Reliability Due to replication of data in the cluster, data is reliably stored on the cluster of machine despite machine failures. If your machine goes down, then also your data will be stored reliably.
5. High Availability Data is highly available and accessible despite hardware failure due to multiple copies of data. If a machine or few hardware crashes, then data will be accessed from another path.

8 **Write a brief note on i) crowd sourcing ii) Mobile business intelligence** 10

Crowdsourcing is a great way to capitalize on the resources that can build algorithms and predictive models

Kaggle: Kaggle describes itself as "an innovative solution for statistical/analytics outsourcing." That's a very formal way of saying that Kaggle manages competitions among the world's best data scientists. Here's how it works: Corporations, governments, and research laboratories are confronted with complex statistical challenges. They describe the problems to Kaggle and provide data sets. Kaggle converts the problems and the data into contests that are posted on its web site. The contests feature cash prizes ranging in value from $100 to $3 million. Kaggle's clients range in size from tiny start-ups to multinational corporations such as Ford Motor Company and government agencies such as NASA.

As per Anthony Goldbloom, Kaggle's founder and CEO: The idea is that someone comes to us with a problem, we put it up on our website, and then people from all over the world can compete to see who can produce the best solution."

Kaggle's approach is that it is truly a win-win scenario—contestants get access to real-

world data (that has been carefully "anonymized" to eliminate privacy concerns) and prize

sponsors reap the benefits of the contestants' creativity.

Crowdsourcing is a disruptive business model whose roots are in technology but is extending beyond technology to other areas.

There are various types of crowd sourcing, such as crowd voting, crowd purchasing, wisdom of crowds, crowd funding, and contests.

Take for example:

- 99designs.com/, which does crowdsourcing of graphic design
- agentanything.com/, which posts "missions" where agents vie for to run errands
- 33needs.com/, which allows people to contribute to charitable programs that make a social impact

Mobile Business Intelligence and Big Data

Analytics on mobile devices is what some refer to as putting BI in your pocket. Mobile drives straight to the heart of simplicity and ease of use that has been a major barrier to BI adoption since day one.

Kerzner explains his view on this topic:

We have been working on Mobile BI for a while but the iPad was the inflection point where I think it started to become mainstream. I have seen customers over the past decade who focused on the mobile space generally and mobile applications in particular. One client in particular told me that he felt like he was pushing a boulder up a hill until he introduced mobility to enhance productivity. Once the new smart phones and tablets arrived, his phone was ringing off the hook and he was trying to figure out which project to say yes to, because he couldn't say yes to everyone who suddenly wanted mobile analytics in the enterprise.

Ease of Mobile Application Deployment

Three elements that have impacted the viability of mobile BI:

1. Location—the GPS component and location . . . know where you are in time as well as the movement.
2. It's not just about pushing data; you can transact with your smart phone based on information you get.
3. Multimedia functionality allows the visualization pieces to really come into play.

Three challenges with mobile BI include:

1. Managing standards for rolling out these devices.
2. Managing security (always a big challenge).
3. Managing "bring your own device," where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

9 **Explain the steps to construct a box plot and Construct the box plot for 54,60,65,66 67,69,70,72,73,75,76 and identify lower and higher outliers.** 10

Box plot construction

The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the medianand the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range or IQ.

Box plots with fences
A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. The following quantities (called *fences*) are needed for identifying extreme values in the tails of the distribution:

1. lower inner fence: Q1 - 1.5*IQ
2. upper inner fence: Q3 + 1.5*IQ
3. lower outer fence: Q1 - 3*IQ
4. upper outer fence: Q3 + 3*IQ

10 **Define Data Analytics with its importance. Explain the four types of data analytics with suitable use cases.** 10

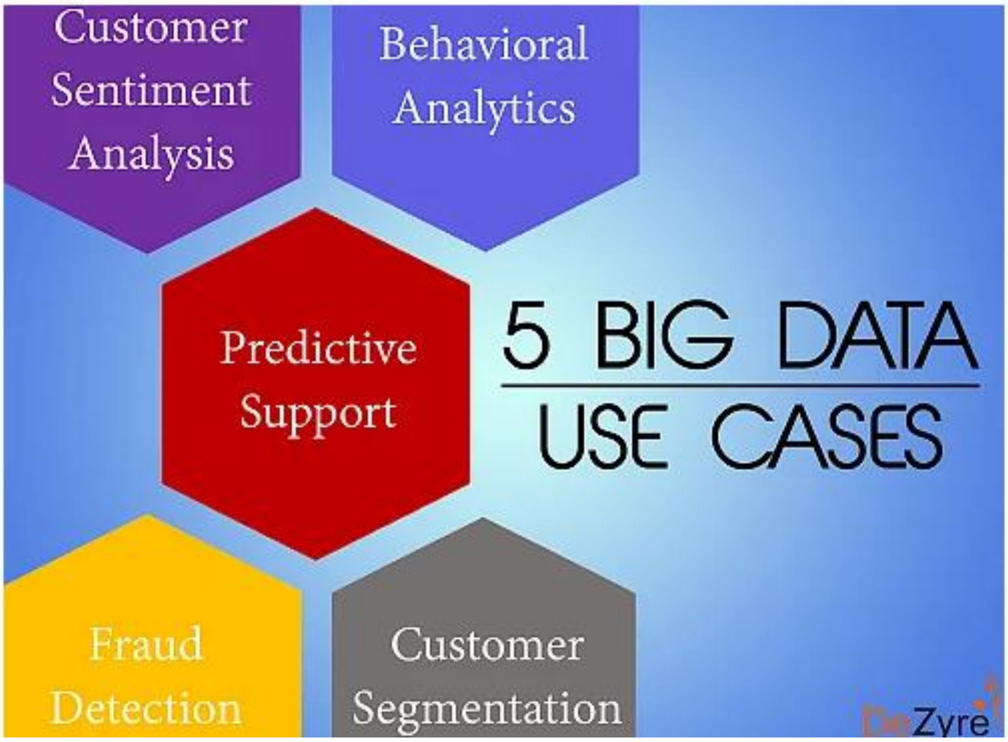
- Data analytics is the science of analyzing raw data to make conclusions about that information.
- The techniques and processes of data analytics have been automated into

mechanical processes and algorithms that work over raw data for human consumption.

- Data analytics help a business optimize its performance.

Data analytics is broken down into four basic types.

1. Descriptive analytics: This describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
2. Diagnostic analytics: This focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
3. Predictive analytics: This moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
4. Prescriptive analytics: This suggests a course of action. If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output.


Customer Sentiment Analysis
Behavioral Analytics
Predictive Support
5 BIG DATA USE CASES
Fraud Detection
Customer Segmentation
DeZyre