Sub:	Big Data Analytics - 22MCA412 – MCA – IAT1 – Solution – July 2025

Note : Answer FIVE FULL Questions, choosing ONE full question from each Module

PART I

1 **Define Big Data. Explain its main characteristics with examples and give example applications.**

Big Data is also data but with a huge size. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

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 behavior, 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 behavior, 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.

Table 1.1 Example Analytics Applications

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 **Describe the analytical process model used in Big Data Analytics.**

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

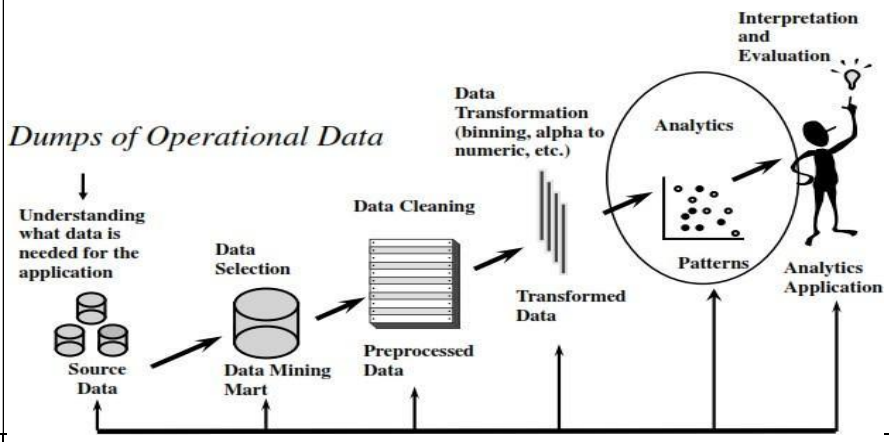
have a deterministic impact on the analytical model).

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.



PART II

3 **A dataset of exam scores (out of 100) for 15 students is: [56, 60, 62, 65, 66, 68, 70, 72, 74, 75, 77, 78, 80, 85, 95].Compute the mean and standard deviation of the data. Compute the Z-score for each score.**

Mean: 72.2

Standard Deviation: 9.80 (approx)

Score	Z-score
56	-1.65
60	-1.24
62	-1.04
65	-0.73
66	-0.63
68	-0.43
70	-0.22
72	-0.02
74	0.18
75	0.29
77	0.49
78	0.59

	80 0.80 85 1.31 95 2.33
4	**Given the dataset representing processing times (in seconds): [14, 18, 22, 26, 28, 32, 33, 34, 36, 39, 41, 44, 46, 100].Draw the Box Plot, marking: min (within fences), Q1, Q2, Q3, Max (within fences) Outliers with dots.** [14, 18, 22, 26, 28, 32, 33, 34, 36, 39, 41, 44, 46, 100] **Step 1: Find Quartiles** • **n = 14** (even) **Median (Q2):** • Middle = average of 7th and 8th values: Q2 = (33 + 34) / 2 = **33.5** **Q1 (lower quartile):** • Lower half: [14, 18, 22, 26, 28, 32, 33] • Q1 = 4th value = **26** **Q3 (upper quartile):** • Upper half: [34, 36, 39, 41, 44, 46, 100] • Q3 = 4th value = **41** **Step 2: Calculate IQR** IQR=Q3−Q1=41−26=15\text{IQR} = Q3 - Q1 = 41 - 26 = 15IQR=Q3−Q1=41−26=15 **Step 3: Determine Fences for Outliers** • **Lower Fence** = Q1 − 1.5 × IQR = 26 − 22.5 = **3.5** • **Upper Fence** = Q3 + 1.5 × IQR = 41 + 22.5 = **63.5** **Step 4: Identify Outliers** • **Any point < 3.5 or > 63.5 is an outlier** • Only **100** > 63.5 → Outlier **Step 5: Determine Min/Max (within fences)** • **Min (within fences)** = 14 • **Max (within fences)** = 46

Final Summary for Box Plot:

Statistic	Value
Min (non-outlier)	14
Q1	26
Q2 (Median)	33.5
Q3	41
Max (non-outlier)	46
Outlier(s)	100

Textual Box Plot Representation

```
14   26    33.5   41    46              100
|-----|========|=======|------|------------------o
      Q1     Q2     Q3     Max          Outlier
```

PART III

5 **Evaluate different approaches for handling missing data and justify your choice in credit scoring applications.**

Missing Values

Missing values can occur because of various reasons.

- Information can be non-applicable
- Information can also be undisclosed

Missing data can also originate an error during merging. Some analytical techniques (e.g., decision trees) can directly deal with missing values. They are:

- Replace (impute)
- Delete
- Keep

Replace: This implies replacing the missing value with a known value. Consider the table 1.2 one could calculate the missing credit bureau scores with the average or median of the known values. For marital status the mode can then be used.

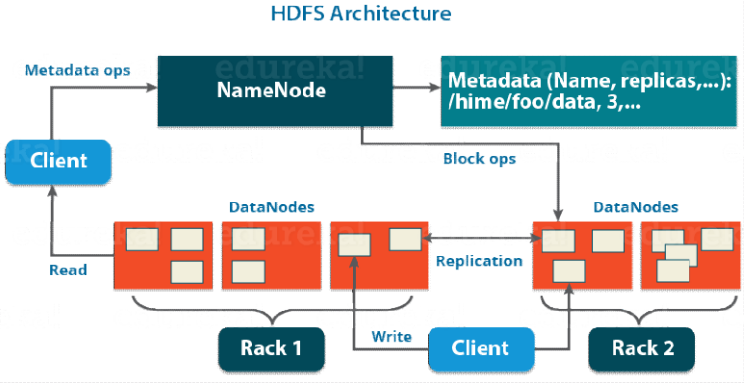
Dealing with missing values

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800	?	620	Churner
2	28	1,200	Single	?	Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married	?	Nonchurner
6	44	?	?	?	Nonchurner

Delete: This is the most straight forward option and consists of deleting observations or variables

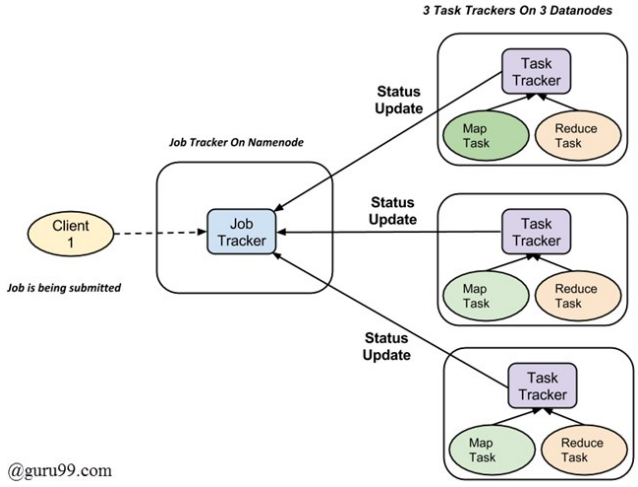
	with lots of missing values. This, of course assumes that information is missing at random and has not meaningful interpretation and/or relationship to the target. *Keep:* Missing values can be meaningful *(e.g a customer didn't disclose his/her income because if he/she is currently unemployed)*. Obviously, this is clearly related to the target *(e.g good/bad risk or churn)* and needs to be considered as a separate category. ***Justification for Credit Scoring Applications:*** *Credit scoring is a **high-stakes domain**, where **bias and inaccurate data handling** can lead to **poor lending decisions or discrimination**.* ***Multiple Imputation** is preferred due to its **statistical rigor** and ability to **retain variability**.* *For operational simplicity, **KNN** or **regression imputation** may be used in real-time scoring environments after validation.* *In critical cases, missing values may themselves be used as a **predictive feature** (e.g., missing income may indicate risk).*
6	**List the various factors required for analytical model and explain.** A good analytical model should satisfy several requirements, depending on the application area. • 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.
	PART IV
7	**Briefly describe the roles of HDFS and MapReduce in Hadoop.** **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.

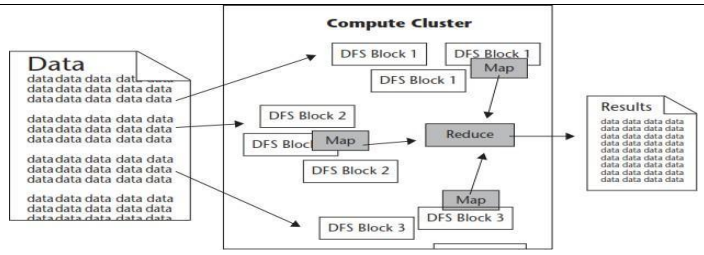


MapReduce: Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back 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.



8 **What is the role of open-source technology in Big Data Analytics?**

Definition: Open source software is computer software that is available in source code form under an open-source license that permits users to study, change and improve and at a times to distribute the software.

Origin: The open-source name came out of a 1998 meeting in Palo Alto in reaction to Netscape's announcement of a source code release for navigation (as Mozilla).Although the source code is released, there are still governing bodies and agreements in place. The most prominent and popular example is the GNU General Public License (GPL), which "allows free distribution under the condition that further developments and applications are put under the same license."

As per David Smith - *vice president of marketing at Revolution Analytics in Palo Alto*

In the past, the pace of software development was moderated by a relatively small set of proprietary software vendors. But there are clear signs that the old software development model is crumbling, and that a new model is replacing it.

The old model's end state was a monolithic stack of proprietary tools and systems that could not be swapped out, modified, or upgraded without the original vendor's support. This model was largely unchallenged for decades.

The status quo rested on several assumptions, including:

1. The amounts of data generated would be manageable
2. Programming resources would remain scarce
3. Faster data processing would require bigger, more expensive hardware

The sudden increase in demand for software capable of handling significantly larger data sets, coupled with the existence of a worldwide community of open-source programmers, has upended the status quo.

The old model was top-down, slow, inflexible and expensive. The new software development model is bottom-up, fast, flexible, and considerably less costly.

A traditional proprietary stack is defined and controlled by a single vendor, or by a small group of vendors. It reflects the old command and control mentality of the traditional corporate world and the old economic order.

An open-source stack is defined by its community of users and contributors. No one "controls" an open-source stack, and no one can predict exactly how it will evolve. The open-source stack reflects the new realities of the networked global economy, which is increasingly dependent on big data.

As per Tasso Argyros: *copresident of Teradata*

This is a significant step forward from what was state-of-the-art until yesterday. This means

	that [in the past] getting data from Hadoop to a database required a Hadoop expert in the middle to do the data cleansing and the data type translation. If the data was not 100% clean (which is the case in most circumstances) a developer was needed to get it to a consistent, proper form. Besides wasting the valuable time of that expert, this process meant that business analysts couldn't directly access and analyze data in Hadoop clusters. SQL-H, an industry-first, solves all those problems.
	PART V
9	**Evaluate the value of crowdsourcing in Big Data using platforms like Kaggle.** **Crowdsourcing:** 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
10	**Imagine you're a data scientist in healthcare. Propose how inter-firewall analytics can enhance patient care.** Over the last 100 years, supply chain has evolved to connect multiple companies and enable them to collaborate to create enormous value to the end-consumer through concepts like CPFR (collaborative planning, forecasting and replenishment—a collection of business practices that leverage the Internet and electronic data interchange to reduce inventories and expenses while improving customer service), VMI (vendor-managed inventory—a technique used by customers in which manufacturers receive sales data to forecast consumer demand more accurately), etc. Decision sciences will witness a similar trend as enterprises begin to collaborate on insights across the value chain. For instance, in the healthcare industry, rich consumer insights can be generated by collaborating on data and insights from the health insurance provider, pharmacy delivering the drugs and the drug manufacturer. In fact, this is not necessarily limited to companies within the traditional demand-supply chain. There are instances where a retailer and a social media company can come together to share insights on

consumer behaviour that will benefit both concerns. Some of the more progressive companies will take this a step further and work on leveraging the large volumes of data outside the firewall such as social data, location data, etc.

In other words, it will not be long before internal data and insights from within the enterprise firewall is no longer a differentiator. We call this trend the move from intra- to inter- and trans-firewall analytics. Yesterday, companies were doing functional silo-based analytics. Today they are doing intra- firewall analytics with data within the firewall. Tomorrow they will be collaborating on insights with other companies to do inter-firewall analytics as well as leveraging the public domain to do trans- firewall analytics.

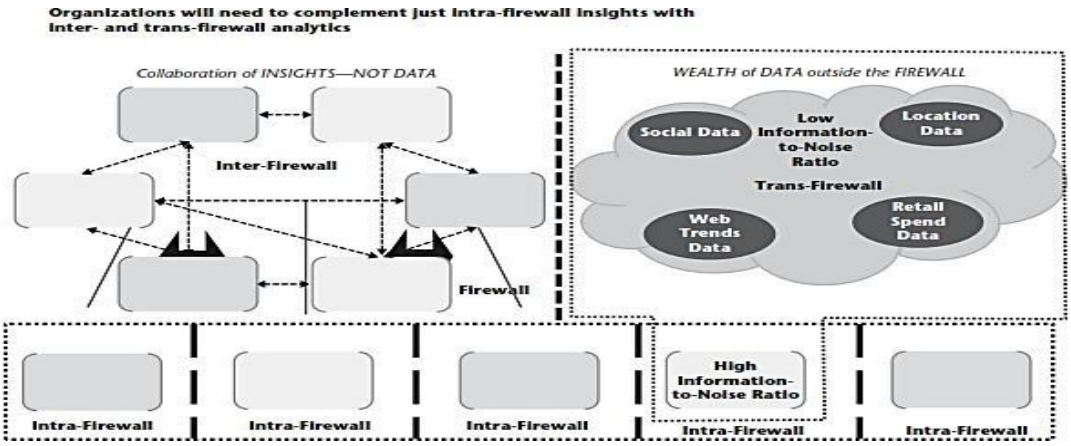


Figure given below depicts, setting up inter- and trans-firewall analytics will add significant value. However, it does present some challenges. They are: As one moves outside the firewall, the information-to-noise ratio increases putting additional requirements on analytical methods and technology requirements Organizations are often limited by a fear of collaboration and overreliance on proprietary information The fear of collaboration is mostly driven by competitive fears, data concerns, and proprietary orientation that limits opportunities for cross-organizational learning and innovation.

While it is clear that the transition to an inter- and trans-firewall paradigm is not easy, we feel it will continue to grow and at some point it will become a key weapon, available for decision scientists to drive value and efficiencies.

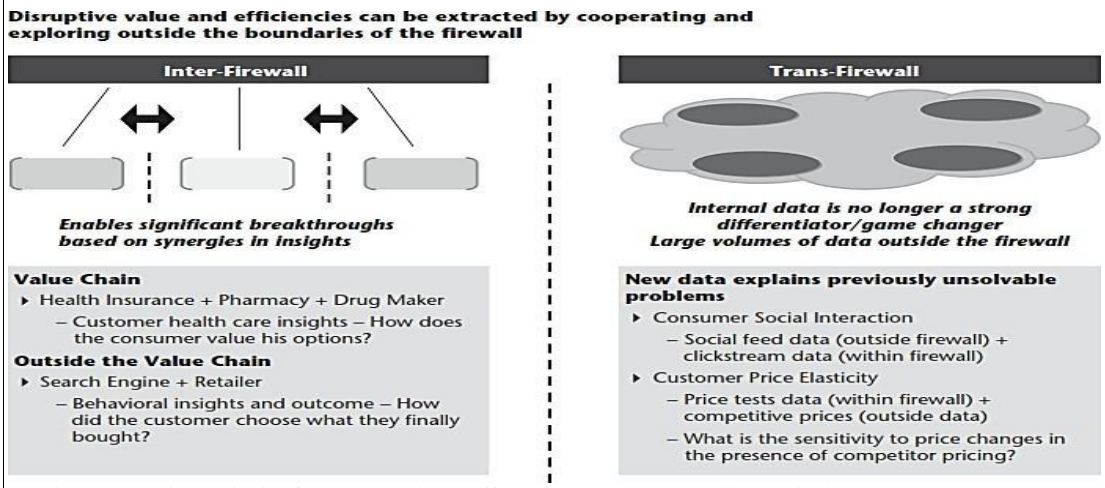


Figure: Value Chain for Inter-Firewall and Trans-Firewall Analytics