USN : □

CMR Institute of Technology, Bangalore

I - INTERNAL ASSESSMENT

Semester: 4-CBCS 2018
Subject: BIG DATA ANALYTICS (18MCA454)
Faculty: Ms Gomathi T

Date: 8 Apr 2021
Time: 02:00 PM - 03:30 PM
Max Marks: 50

PART A

Answer any1 question(s)

Q.No		Marks	CO	PO	BT/CL
1	With a neat diagram, describe the working of analytical processing model	10	CO1	PO1	L2
2	Explain the various factors required for good analytical model.	10	CO1	PO2	L2

PART B

Answer any1 question(s)

Q.No		Marks	CO	PO	BT/CL
3	Discuss the critical components of hadoop with neat diagram.	10	CO2	PO2	L2
4	What is predictive analysis? Why are they required?Discuss the leading trends of predictive analysis.	10	CO2	PO2	L2

PART C

Answer any1 question(s)

Q.No		Marks	CO	PO	BT/CL
5	Classify the difference between Map Reduce and RDBMS.	10	CO3	PO2	L4
6	Explain Volunteer Computing and Grid Computing.	10	CO3	PO2	L5

PART D

Answer any1 question(s)

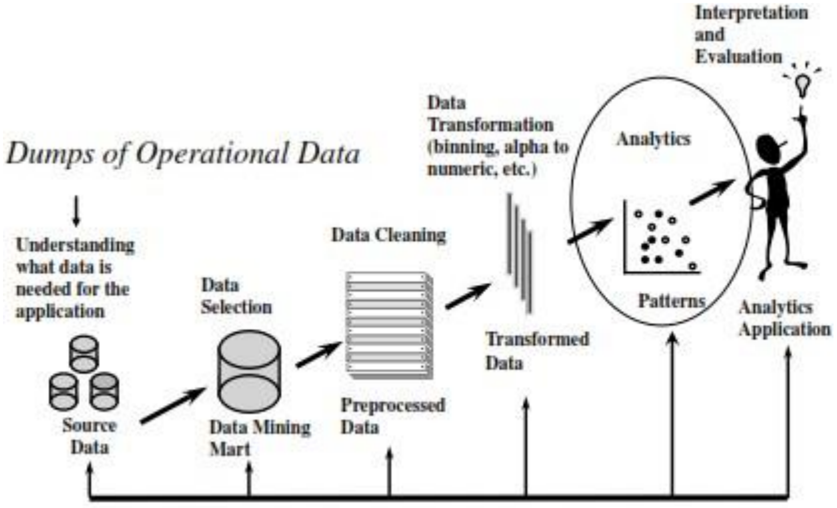
Q.No		Marks	CO	PO	BT/CL
7	Calculate the Z-Score and detect the outlier for the following data. Where mean = 40 Standard deviation = 10 and Data= 30 50 10 40 60 80	10	CO1	PO2	L3
8	Discuss the application of big data analytics.	10	CO1	PO2	L2

PART E

Answer any1 question(s)

Q.No		Marks	CO	PO	BT/CL
9	Classify the different types of data sources and data elements.	10	CO1	PO2	L4
10	Carl works at a computer store. He also recorded the number of sales he made each month. In the past 12 months, he sold the following numbers of computers: 51, 17, 25, 39, 7, 49, 67, 41, 20, 2, 43, 13. Construct the box plot for the above sales.	10	CO1	PO2	L6

1. **With a neat diagram, describe the working of analytical processing model**



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

Figure 1.2 The Analytics Process Model

1. Define the business problems to be solved
2. All source-data need to be identified that could be of potential interest.
3. All data to be gathered in a staging area
4. Basic exploratory analysis will be considered.
5. Data cleaning step to get rid of all inconsistencies
6. In the analytics step, an analytical model will be estimated on the preprocessedand transformed data.

Once the model is built it will interpreted and evaluated by the business experts.

2. Various factors required for analytical model:

1. Business relevance
2. Statistical performance
3. Operational efficient
4. Economic cost
5. Local and International regulations and legislation

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. It makes no sense to have a working analytical model that got sidetracked from the original problem statement. In order to achieve business relevance, it is of key importance that the business problem to be solved is appropriately defined, qualified, and agreed upon by all parties involved at the outset of the analysis.

A second criterion is statistical performance. The model should have statistical significance and predictive power. How this can be measured will depend upon the type of analytics considered. For example, in a classification setting (churn, fraud), the model should have good discrimination power. In a clustering setting, the clusters should be as homogenous as possible. In later chapters, we will extensively discuss various measures to quantify this.

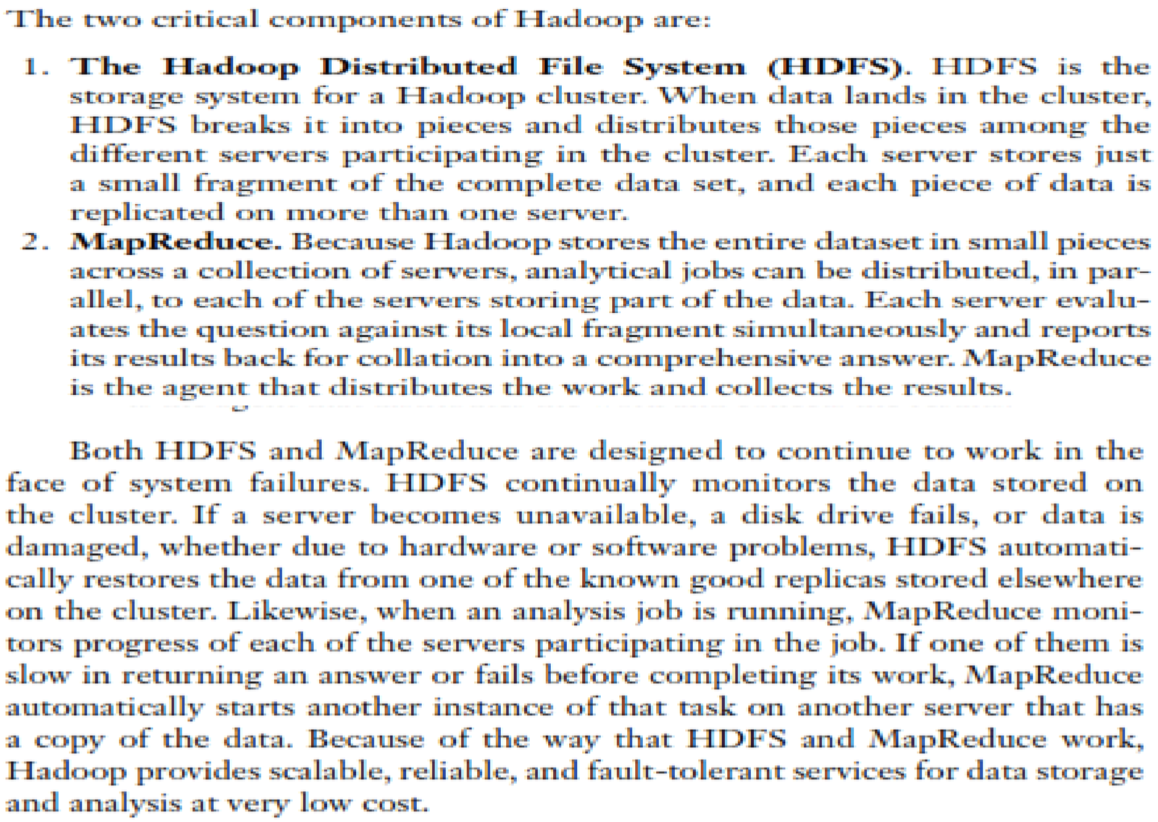
Depending on the application, analytical models should also be interpretable and justifiable. *Interpretability* refers to understanding the patterns that the analytical model captures. This aspect has a certain degree of subjectivism, since interpretability may depend on the business user's knowledge. In many settings, however, it is considered to be a key requirement. For example, in credit risk modeling or medical diagnosis, interpretable models are absolutely needed to get good insight into the underlying data patterns. In other settings, such as response modeling and fraud detection, having interpretable models may be less of an issue. *Justifiability* refers to the degree to which a model corresponds to prior business knowledge and intuition.[6] For example, a model stating that a higher debt ratio results in more creditworthy clients may be interpretable, but is not justifiable because it contradicts basic financial intuition. Note that both interpretability and justifiability often need to be balanced against statistical performance. Often one will observe that high performing analytical models are incomprehensible and black box in nature. A popular example of this is neural networks, which are universal approximators and are high performing, but offer no insight into the underlying patterns in the data. On the contrary, linear regression models are very transparent and comprehensible, but offer only limited modeling power.

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 (e.g., campaign management, capital calculation). Especially in a real-time online scoring environment (e.g., fraud detection) this may be a crucial characteristic. Operational efficiency also entails the efforts needed to monitor and backtest the model, and reestimate it when necessary.

Another key attention point is the *economic cost* needed to set up the analytical model. This includes the costs to gather and preprocess the data, the costs to analyze the data, and the costs to put the resulting analytical models into production. In addition, the software costs and human and computing resources should be taken into account here. It is important to do a thorough cost–benefit analysis at the start of the project.

Finally, analytical models should also comply with both local and international *regulation and legislation*. For example, in a credit risk set-

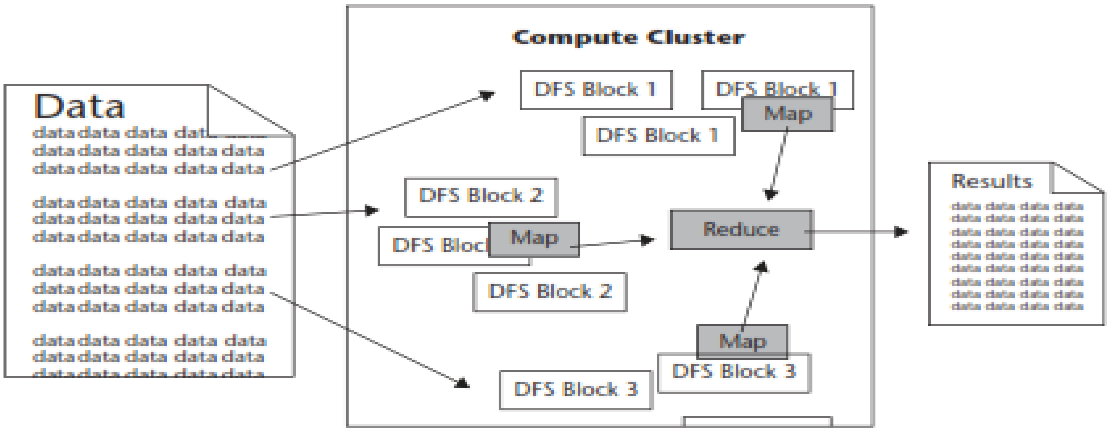
3. **Discuss the critical components of hadoop with neat diagram**



The two critical components of Hadoop are:

1. **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.
2. **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 failures. 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.



4. **What is predictive analysis? Why are they required?Discuss the leading trends of predictive analysis.**

To master analytics, enterprise will move from being in reactive positionsto forward leaning position. Recommendation engines similar to those used in Netflix and Amazon that use past purchases and buying behaviorto 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 relatedinfo, including sentiment, behavior, and even emotions. Customer insightengines will be the backbone in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetimevalue, optimizing marketing campaigns for revenue lift, and targeting individuals or companies at the right time to maximize their spending habit. 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.

5. **Classify the differences between map reduce and RDBMs**

- MapReduce suits in an application where the data is written once and read many times like in your Facebook profile you post your photo once and that picture of your seen by your friends many times, whereas RDBMS good for data sets that are continuously updated.
- The RDBMS is suits for an application where data size is limited like it's in GBs,whereas MapReduce suits for an application where data size is in Petabytes.
- The RDBMS accessed data in interactive and batch mode, whereas MapReduce access the data in batch mode.
- The RDBMS schema structure is static, whereas MapReduce schema is dynamic.
- The RDBMS suits with structure data sets, whereas MapReduce suits with un-structure data sets.
- The RDBMS scaling is nonlinear, whereas MapReduce is linear.

6. **Volunteer Computing and Grid Computing:**

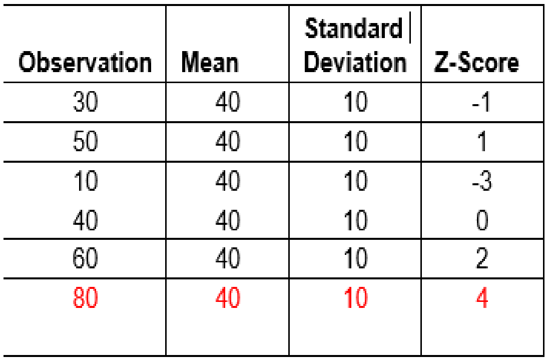
Grid Computing:

The High Performance Computing (HPC) and Grid Computing communities have been doing large-scale data processing for years, using such APIs as Message Passing Interface (MPI). Broadly, the approach in HPC is to distribute the work across a cluster ofmachines, which access a shared filesystem, hosted by a SAN. This works well for predominantly compute-intensive jobs, but becomes a problem when nodes need to access larger data volumes (hundreds of gigabytes, the point at which MapReduce reallystarts to shine), since the network bandwidth is the bottleneck and compute nodes become idle. MapReduce tries to collocate the data with the compute node, so data access is fast since it is local. This feature, known as data locality, is at the heart of MapReduce and isthe reason for its good performance. Recognizing that network bandwidth is the mostprecious resource in a data center environment (it is easy to saturate network links by copying data around), MapReduce implementations go to great lengths to conserve it byexplicitly modelling network topology. Notice that this arrangement does not precludehigh-CPU analyses in MapReduce

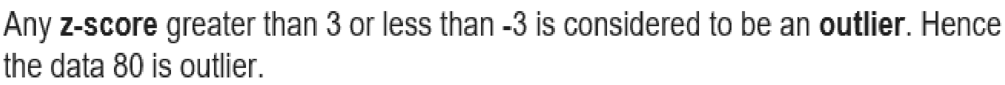
Volunteer Computing:

Volunteer computing projects work by breaking the problem they are trying to solve into chunks called work units, which are sent to computers around the world to be analyzed. For example, a SETI@home work unit is about 0.35 MB of radio telescope data, and takes hours or days to analyze on a typical home computer. When the analysis is completed, the results are sent back to the server, and the client gets another work unit. As a precaution to combat cheating, each work unit is sent to three different machines and needs at least two results to agree to be accepted. Although SETI@home may be superficially similar to MapReduce (breaking a problem into independent pieces to be worked on in parallel), there are some significant differences. The SETI@home problem is very CPU-intensive, which makes it suitable for running on hundreds of thousands of computers across the world,8 since the time to transfer the work unit is dwarfed by the time to run the computation on it. Volunteers are donating CPU cycles, not bandwidth. MapReduce is designed to run jobs that last minutes or hours on trusted, dedicated hardware running in a single data center with very high aggregate bandwidth interconnects.

7. **Calculate the Z-Score and detect the outlier for the following data. Where mean = 40 Standard deviation = 10 and Data= 30 50 10 40 60 80**



Observation	Mean	Standard Deviation	Z-Score
30	40	10	-1
50	40	10	1
10	40	10	-3
40	40	10	0
60	40	10	2
80	40	10	4



Any **z-score** greater than 3 or less than -3 is considered to be an **outlier**. Hence the data 80 is outlier.

8. **Discuss the application of big data analytics**

   Marketing: Response modeling Net-Lift Modeling Retention Modeling Market-based analytics Recommender Systems Customer segmentation

   Risk Management: Credit risk modeling Market risk modeling Operational risk modeling Fraud detection

   Government Tax: avoidance Social Security Fraud Money Laundering Terrorism detection

   Web: Web analytics Social media Multi-variate trusting

   Logistics: Demand forecasting Supply chain analytics

9. **Data Sources and Data Elements**

   Transaction: - Transactional data consists of structured, low-level, detailed information capturing the key characteristics of a customer transaction.
   Un-Structured data: – are stored in form of text documents.
   Qualitative, expert based data:-Subject matter expertise
   Data-Poolers:- Dun & Bradstreet, Thomson Reuters
   Social Media: Data from face book and twitter etc.

   Continuous: - There are data elements that can be defined on an interval that can be limited / unlimited.
   Categorical: -Nominal: Take limited set of values Ordinal: Take limited set of values with a meaningful ordering in-between. Binary: Take on 2 values.

10. **Construct box plot: 51 17 25 39 7 49 67 41 20 2 43 13**

Box - plot Construction

51, 17, 25, 39, 7, 49, 67, 41, 20, 2, 43,

Ascending order 2 7 13 17 20 25 39 41 43, 49 51

2 7 13 17 20 25 39 41 43 49 51 67

15 $\boxed{32}$ 46

Q1 Median Q3

$M = 32$ $Q_1 = 15$ $Q_3 = 46$

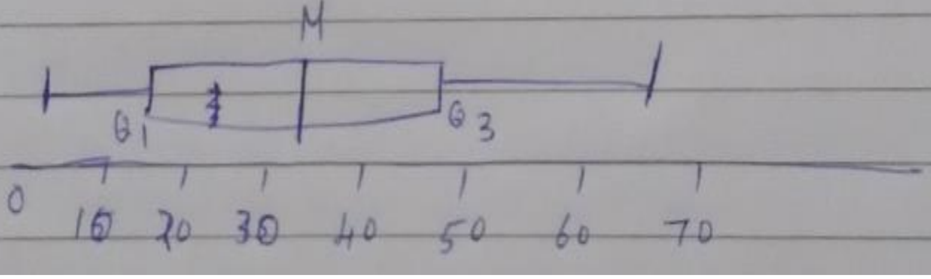
$IQR = (Q_3 - Q_1) \times 1.5$

$= (46 - 15) \times 1.5$

$IQR = 23.5$

$LQ = Q_1 - 23.5$ $UQ = Q_3 + 23.$

$= 15 - 23.5$ $= 46 + 23.$


M
Q1
Q3
0
10
20
30
40
50
60
70