Solution
Internal Assessment Test 1 – March.2019

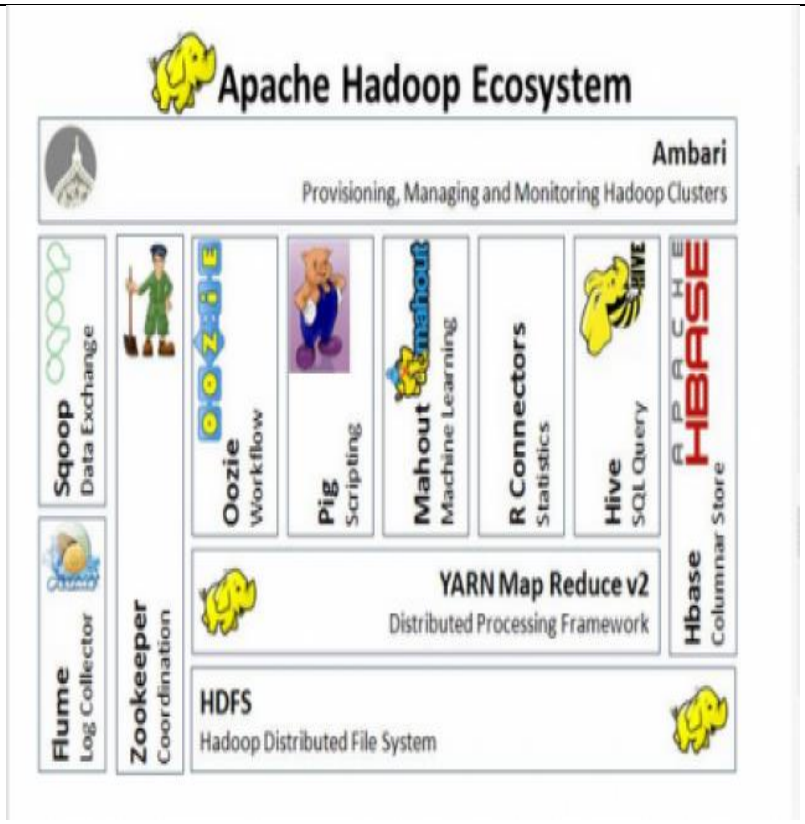
Sub:	Big Data Analytics							Code:	17MCA452
Date:	07/03/2018	Duration:	90mins	Max Marks:	50	Sem:	IV	Branch:	MCA

Note: Answer Any Five Questions

Question #		Description	Marks Distribution		Max Marks
1	a)	**Discuss the Example Applications in analytics.** Analytics is everywhere and it is embedded into our daily lives. 1. Physical mail box 2. Behavioral scoring model 3. Telephone service 4. Social Ads 5. My Twitter post		05 M	10 M
	b)	**Examine and summarize the analytics application in marketing, risk management, government, web and logistics** **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		05 M	
2	a)	**List the various factors required for analytical model and explain.** Business relevance Statistical Performance		10 M	10 M

		Interpretability Justifiability Operational efficient Economic Cost International Regulations and Legislations			
3	a)	**Explain Analytical 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 preprocessed and transformed data. 7. Once the model is built it will be interpreted and evaluated by the business experts.		10M	10 M
4	a)	**List the Basic Nomenclatures discussed in analytics** CLV – Customer lifetime value Can be measured for either individual customer/ at household level. Account behavior – Consider credit score exercise Aim – to find whether customer is defaulter Customer can play different roles parents can buy for kids		5 M	10 M
	b)	**Explain the requirements to satisfy good analytical model** 1. Business relevance 2. Statistical performance 3. Operational efficient 4. Economic cost 5. Local and International regulations and legislation		5 M	
5	a)	**List types of data sources.** 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.		5 M	10 M
	b)	**List and explain the types of data elements.** Continuous: - There are data elements that can be defined on an interval that can be limited / unlimited. Categorical: - 5 Nominal:		5 M	

		Take limited set of values Ordinal: Take limited set of values with a meaningful ordering in-between. Binary: Take on 2 values.			
6	a)	**Describe outlier. How will you detect and treat outliers?** Outliers are extreme observations that are very dissimilar to the rest of the population. Two activities are essential for characterizing a set of data: 1. Examination of the overall shape of the graphed data for important features, including symmetry and departures from assumptions. 2. Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. Two graphical techniques for identifying outliers, scatter plots and box plots. 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 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.		10 M	10 M
7	a)	**Discuss the critical components of Hadoop with neat diagram** Core Hadoop Components 1) Hadoop Common- 2) Hadoop Distributed File System (HDFS) - 3) HDFS 4) MapReduce		5 M	10 M



b) **Why Open Source technology popular?**

Open-Source software is computer software that is available in source code form under an open source license that permits user to study, change and improve.

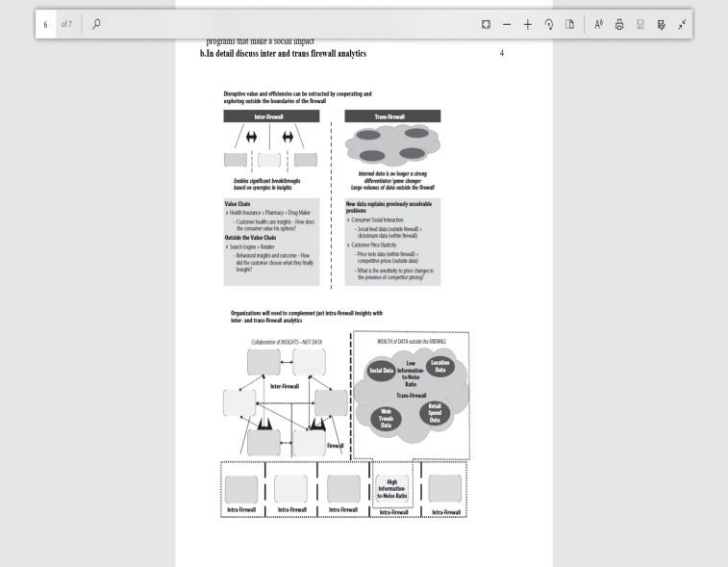
Open source software is software with source code that anyone can inspect, modify, and enhance.
"Source code" is the part of software that most computer users don't ever see; it's the code computer programmers can manipulate to change how a piece of software—a "program" or "application"—works. Programmers who have access to a computer program's source code can improve that program by adding features to it or fixing parts that don't always work correctly.

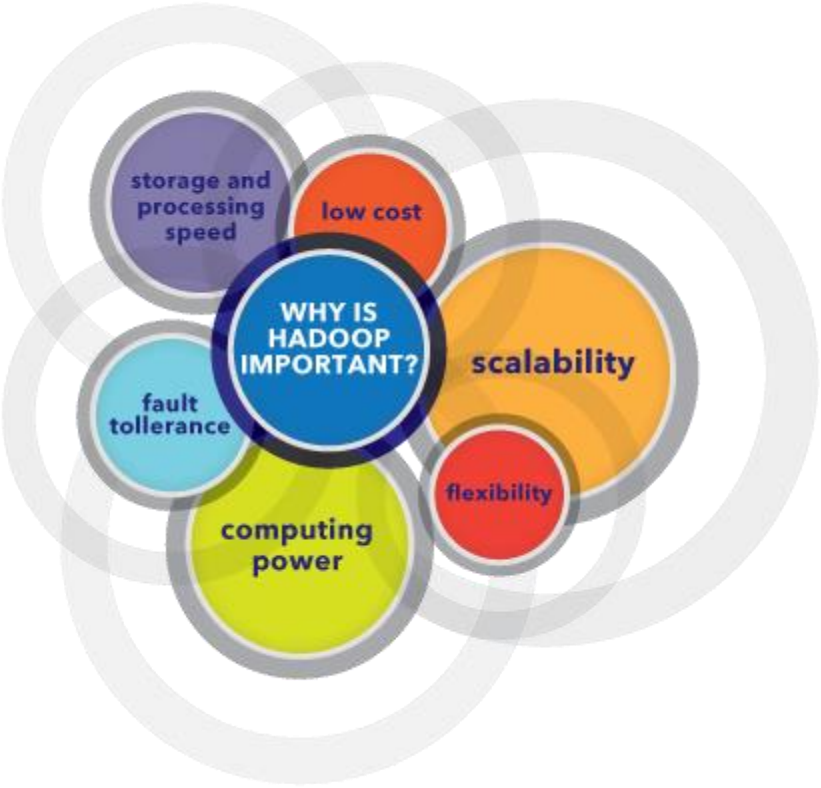
GPL- General Public License: Governing bodies & agreements in place.

5 M

8	a)	**Explain predictive analytics in detail.** To master analytics, enterprise will move from being in reactive positions to forward leaning position. 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 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 M	10 M
	b)	**Explain Mobile Intelligence.** 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. Mobile		5 M	

		devices are a great leveling fi eld where making complicated actions easy is the name of the game. 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	a)	**Summarize the various types of crowd sourcing involved in analytics** Crowdsourcing is a great way to capitalize on the resources that can build algorithms and predictive models. Crowd sourcing is a cost and time effective method for moderating and curating data. It has no over head costs and produces high quality results with little investment. Crowdsourcing is a disruptive business model whose roots are in technology but is extending beyond technology to other areas. There are various types of crowdsourcing, 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		5 M	10 M

	b)	**In detail discuss inter and trans firewall analytics**  		5 M	
10	a)	**List and Explain the technical features of Hadoop** Hadoop is a framework that allows for distributed processing of large data sets across clusters of commodity computers using a simple programming model. Hadoop is a framework that has the ability to store and analyze data present in different machines at different locations very quickly and in a very cost-effective manner. It uses the concept of MapReduce which enables it to divide the query into small parts and process them in parallel.		5 M	10 M

					
	b)	**Construct the box plot for 54, 60, 65, 66, 67, 69, 70, 72, 73, 75, 76 and identify lower and higher outliers**		5 M	

54 60 65 66 67 69 70 72 73
75 76

M

54 60 [65] 66 67 [69] 70 72 [73] 75 76

Q_1 Q_2

$IQR = (Q_2 - Q_1) * 1.5$

$73 - 65$

$IQR = 12$

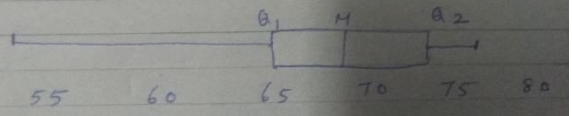
Lower quartile = $Q_1 - IQR$

$65 - 12 = 53$

Higher quartile = $Q_2 + IQR$

$73 + 12 = 85$

In the given data, LQ & HQ are within the limit.



50 55 60 65 70 75 80