CMR
INSTITUTE OF
TECHNOLOGY

USN	1	C								

Internal Assessment Test 1 – March 2018(Answer Key)

Sub:	Big Data Analytics							Code:	16MCA452
Date:	13-03-18	Duration:	90 mins	Max Marks:	50	Sem:	IV	Branch:	MCA

Note: Answer any 5 questions. All questions carry equal marks. Total marks: 50

Marks

1. a. **Discuss the Example Applications in analytics.** 5

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
6. Supermarkets

b. **Examine and summarize the analytics application in marketing, risk management, government, web and logistics.** 5

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 trsting

Logistics:
Demand forecasting
Supply chain analytics

2. a. **Explain Analytical Process Model** 10

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 interpreted and evaluated by the business experts.

3. a. **List the Basic Nomenclatures discussed in analytics.** 5

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.

b. **Explain the requirements to satisfy good analytical model.** 5

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

4. a. **List types of data sources.** 5

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.

b. **List and explain the types of data elements.** 5

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.

a. **Describe outlier. How will you detect and treat outliers?** 10

5

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. 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

Outlier detection criteria:

A point beyond an inner fence on either side is
considered a **mild outlier**. A point beyond an outer fence
is considered an **extreme outlier**.

6 a. **Explain Hadoop Parallel world** 6

Apache Hadoop is an Open-Source platform for storage & processing of diverse data types for storage and processing of diverse data types that enables data driven enterprise to rapidly derive the complete value from all their data.

Cloudera: Leading provider of Apache Hadoop

Creators: Dough Cutting & Mike Cafarella

Nutch: goal of creating a large web index.

b. **Why Open Source technology popular?** 4

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.

7 a. **Explain predictive analytics in detail.** 5

To master analytics, enterprise will move from being in reactive positions to forward leaning position.

Recommendation engines similar to those used in Netflix and A mazon
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 customerrelated 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.

b. Explain Mobile Intelligence. 5

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 devices are a great leveling field 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.

8 a. Summarize the various types of crowd sourcing involved in analytics 6

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:

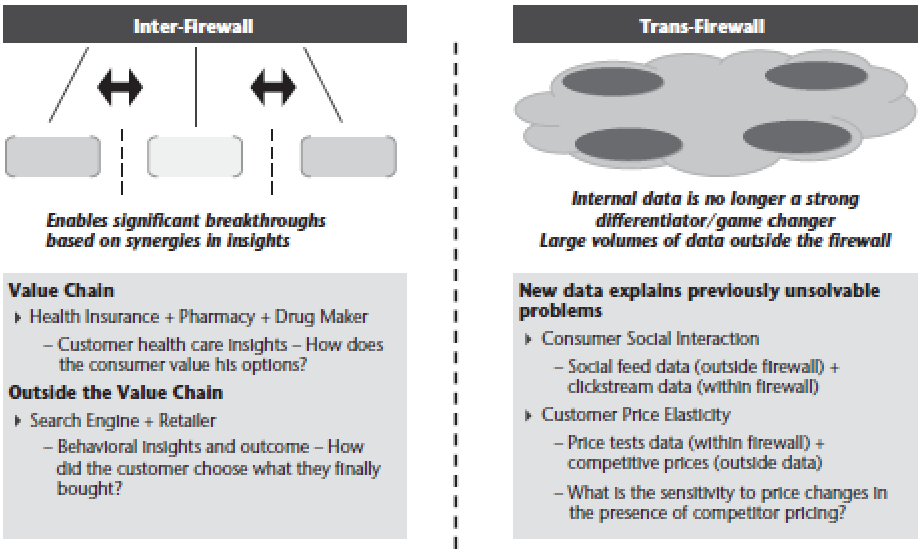
■ 99designs.com/ , which does crowdsourcing of graphic design
■ agentanything.com/ , which posts "missions" where agents vie for to

run errands

■ 33needs.com/ , which allows people to contribute to charitable programs that make a social impact

b.In detail discuss inter and trans firewall analytics 4

Disruptive value and efficiencies can be extracted by cooperating and exploring outside the boundaries of the firewall



Inter-Firewall

Enables significant breakthroughs based on synergies in insights

Value Chain

- Health Insurance + Pharmacy + Drug Maker
  - Customer health care insights – How does the consumer value his options?

Outside the Value Chain

- Search Engine + Retailer
  - Behavioral insights and outcome – How did the customer choose what they finally bought?

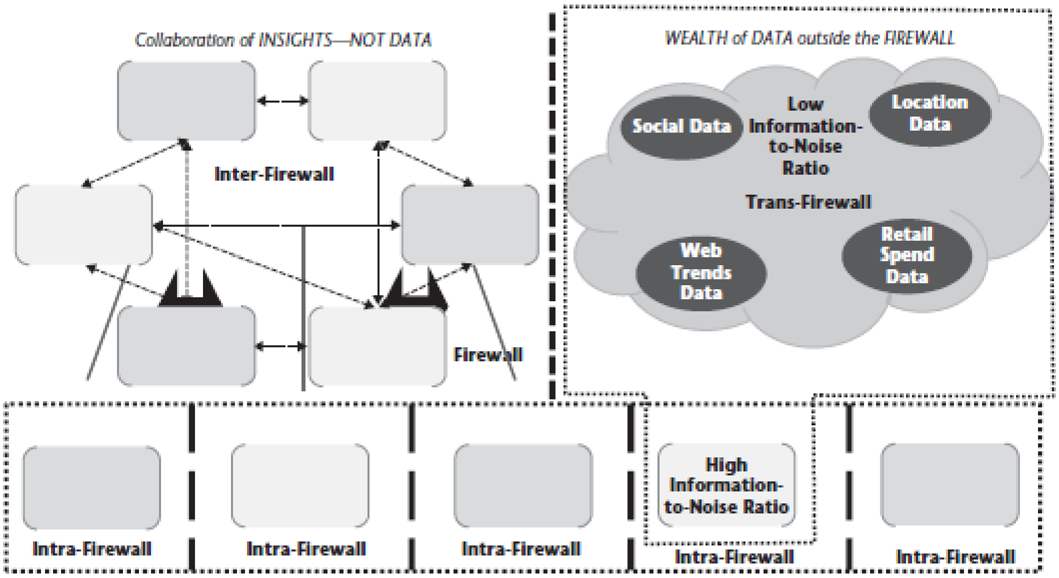
Trans-Firewall

Internal data is no longer a strong differentiator/game changer
Large volumes of data outside the firewall

New data explains previously unsolvable problems

- Consumer Social Interaction
  - Social feed data (outside firewall) + clickstream data (within firewall)
- Customer Price Elasticity
  - Price tests data (within firewall) + competitive prices (outside data)
  - What is the sensitivity to price changes in the presence of competitor pricing?

Organizations will need to complement just intra-firewall insights with inter- and trans-firewall analytics


Collaboration of INSIGHTS—NOT DATA
Inter-Firewall
Firewall
WEALTH of DATA outside the FIREWALL
Social Data
Low Information-to-Noise Ratio
Location Data
Trans-Firewall
Web Trends Data
Retail Spend Data
Intra-Firewall
Intra-Firewall
Intra-Firewall
High Information-to-Noise Ratio
Intra-Firewall
Intra-Firewall