ST.JOSEPH COLLEGE OF ENGINEERING
DEPARTMENT OF INFORMATION TECHNOLOGY
QUESTION BANK

IT6701 INFORMATION MANAGEMENT

OBJECTIVES:

- To expose students with the basics of managing the information
- To explore the various aspects of database design and modelling,
- To examine the basic issues in information governance and information integration
- To understand the overview of information architecture.

UNIT I DATABASE MODELLING, MANAGEMENT AND DEVELOPMENT 9

Database design and modelling - Business Rules and Relationship; Java database Connectivity (JDBC), Database connection Manager, Stored Procedures. Trends in Big Data systems including NoSQL - Hadoop HDFS, MapReduce, Hive, and enhancements.

UNIT II DATA SECURITY AND PRIVACY 9 Program Security, Malicious code and controls against threats; OS level protection; Security – Firewalls, Network Security Intrusion detection systems. Data Privacy principles. Data Privacy Laws and compliance.

UNIT III INFORMATION GOVERNANCE 9 Master Data Management (MDM) – Overview, Need for MDM, Privacy, regulatory requirements and compliance. Data Governance – Synchronization and data quality management.

UNIT IV INFORMATION ARCHITECTURE 9 Principles of Information architecture and framework, Organizing information, Navigation systems and Labelling systems, Conceptual design, Granularity of Content.

UNIT V INFORMATION LIFECYCLE MANAGEMENT 9 Data retention policies; Confidential and Sensitive data handling, lifecycle management costs. Archive data using Hadoop; Testing and delivering big data applications for performance and functionality; Challenges with data administration; **TOTAL: 45 PERIODS**

OUTCOMES: At the end of the course the students will be able to:

- Cover core relational database topics including logcal and physical design and modeling
- Design and implement a complex information system that meets regulatory requirements; define and manage an organization's key master data entities
- Design, Create and maintain data warehouses.
- Learn recent advances in NOSQL , Big Data and related tools.

TEXT BOOKS:

1. Alex Berson, Larry Dubov MASTER DATA MANAGEMENT AND DATA GOVERNANCE, 2/E, Tata McGraw Hill, 2011

2. Security in Computing, 4/E, Charles P. Pfleeger, Shari Lawrence Pfleeger, Prentice Hall; 2006

3. Information Architecture for the World Wide Web; Peter Morville, Louis Rosenfeld ; O'Reilly Media; 1998

REFERENCES:

1. Jeffrey A. Hoffer, Heikki Topi, V Ramesh - MODERN DATABASE MANAGEMENT, 10 Edition, PEARSON, 2012

2. http://nosql-database.org/ Next Gen databases that are distributed, open source and scalable.

3. http://ibm.com/big-data - Four dimensions of big data and other ebooks on Big Data Analytics

4. Inside Cyber Warfare: Mapping the Cyber Underworld- Jeffrey Carr, O'Reilly Media; Second Edition 2011

UNIT I
PART-A (2 MARKS)

1. **Define data modeling**.

A database model contains all the information about the entities.
It can be defined as a collection of logical representation of the data structures.
A model can be constructed in each of the design phase Emphasis is on algorithm.

2**. List the types of data models.**

Conceptual model
Logical model
Physical Data Models

3. Define schemas.

Schemas are generally stored in a data dictionary. Although a schema is defined in text database language,
The term is often used to refer to a graphical depiction of the database structure.

4. Define Normalization

Normalization provides a mechanism for investigating and refining the schema created through ER modeling process
It uses functional dependency to remove the anomalies and get the database into a consistent state.

5. Name any two sources of Business rules.

Business Rules are used every day to define entities, attributes, relationships and constraints.

Usually though they are used for the organization that stores or uses data to be an explanation of a policy, procedure, or principle.

Some examples of business rules:

Departments ------offers---------Course

Course----------generates---------Class

Professor ->->->->->teaches->->->->->->Class

6. Summarize the functionalities of JDBC.

JDBC allows multiple implementations to exist and be used by the same application. The API provides a mechanism for dynamically loading the correct Java packages and registering them with the JDBC Driver Manager.
These may be update statements such as SQL's CREATE, INSERT, UPDATE and DELETE, or they may be query statements such as SELECT. Additionally, stored procedures may be invoked through a JDBC connection.

7. List the types of JDBC drivers

Type 1- JDBC-ODBC Bridge Driver
Type 2-Java Native Driver
Type 3-Java network protocol
driver Type 4-Pure java driver

8. What are the steps are involved to access the database using JDBC

1. Register the JDBC Driver
2. Creating a database connection
3. Executing queries
4. Processing the results
5. Closing the database connection

9. Describe the connection object.

DriverManager.getConnection() method to create a connection object.

getConnection() method with appropriate username and password to get a **Connection object** as follows −

```
String URL = "jdbc:oracle:thin:@amrood:1521:EMP";
String USER = "username";
String PASS = "password"
Connection conn = DriverManager.getConnection(URL, USER, PASS);.
```

10. Define Big data

Big Data is a phrase used to mean a massive volume of *both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.*

11. What are the characteristics of big data?

1. Volume

To store the large amount of data

2. Variety

Different type of data format can be stored.

3. Velocity

Speed of data processing

4. Variability

Inconsistency of the data set can hamper processes to handle and manage it.

5. Veracity

The quality of captured data can vary greatly, affecting accurate analysis.

12. Define HDFS

HDFS is the file system required by Hadoop.It is a typical file system, which does not format the hard drives in the cluster. it can be store and manage the data.

HDFS divides the file into a block of either 64 MB or 128 MB.

13. Explanation about the MapReduce

MapReduce is a programming model for processing large data sets with a parallel distributed algorithm on a cluster.

In the traditional systems,data are brought from the datacenter in the main memory, where the application is running

14. Explain the features of Hive.

1. Fits the low level interface requirement of Hadoop perfectly.

2. Supports external tables which make it possible to process data without actually storing in HDFS.

3. It has a rule based optimizer for optimizing logical plans.

4. Supports partitioning of data at the level of tables to improve performance.

5. Metastore or Metadata store is a big plus in the architecture which makes the lookup easy.

6. Hive support file formats which are textFile, SequenceFile, ORC, RCFile,Avro Files,Parquet,LZO Compression etc

15. List out the application of hive

Log processing
Customer facing business intelligence
Data mining and analysis of data
Document indexing

16. Give the types of NoSQL

Graph database
Key-Value database
Column store database
Document database

17. Define NoSQL

NoSQL database also known as Not Only SQL,is an approach to data management and database design.
That is useful for very large set of distributed data.
It incorporates a wide variety of different technologies

18. Explain about YARN

YARN means Yet another Resource Negotiator was added in Hadoop 2.0.It is a resource manager that enables hadoop to improve its distributed processing capabilities.

.

PART-B (16 MARKS)

1. Explain the following terms briefly :entity,attributes,domain,relationship,entity set, Relationship set, one-to-many relationship, many-to-many relationship and normalization.
2. (a) Draw an ER diagram for the bank application .
   (b) Explain about the JDBC in detail.
3. List the various data models in database design(16)
4. Explain the following SELECT statement syntax with examples in HiveQL. i) Computing with Columns (4) ii) WHERE Clauses (4) iii) GROUP BY Clauses(4) iv) HAVING Clauses (4)
5. Analyse various databases used in NoSQL.(16)
6. Explain the following in MapReduce i) Enterprise Storage (6) ii) Database (6) iii) Event streaming
7. Develop a program to establish Java Database connectivity(16)

UNIT –II
PART-A (2 MARKS)

1. Define malicious code.

Malicious code is generally imbibed into the program, with an intension of either modifying the contents or extracting the contents.

Malicious code is by coding a separate program that gets attached to other program during their execution.

2. List out the some security related terms

1. **Computer security**
   A generic name for the collection of tools designed to protect data
2. **Network Security**
   Measures to protect data during transmission
3. **Internet Security**
   Measures to protect data during transmission over a collection of inter connected networks
4. **Data Security** Preventing
   data from theft

3. Define fault and failure

FAULT

The fault caused side effects in areas that were not directly related to it. it cannot be fixed properly because system performance would be hampered.

4. Define Quick Patch

A patch is usually developed and distributed as a replacement for or an insertion in compiled code (that is, in a *binary file* or object module).

5. What is meant by program security flaw?

Non-malicious flaws. Introduced by the programmer overlooking something:

Buffer overflow
Incomplete mediation
Time-of-check to Time-of-use (TOCTTU) errors

Malicious flaws. Introduced deliberately (possibly by exploiting a non-malicious vulnerability):

Virus, worm, rabbit
Trojan horse, trapdoor
Logic bomb, time bomb

Class objects can be initialized dynamically. The initial values of an object may be provided during run time. The advantage of dynamic initialization is that various initialization formats can be used. It provides flexibility of using different data formats.

6. Define Trojan horse

In computing , **Trojan horse**, or **Trojan** , is any malicious computer program which is used to hack into a computer by misleading users of its true intent. rojans are generally spread by some form of social engineering, for example where a user is duped into executing an e-mail attachment disguised to be unsuspicious, (e.g., a routine form to be filled in), or by drive-by download. Although their payload can be anything, many modern forms act as a backdoor, contacting a controller which can then have unauthorized access to the affected computer

7. Define OS level Protection.

Time-sharing operating systems schedule tasks for efficient use of the system and may also include accounting software for cost allocation of processor time, mass storage, printing, and other resources.

8. Give the administrative controls for security

Administrative controls define the human factors of security. It involves all levels of personnel within an organization and determines which users have access to what resources and information by such means as:

- Training and awareness
- Disaster preparedness and recovery plans
- Personnel recruitment and separation strategies
- Personnel registration and accounting

9. Define Firewall.

Acting as a barrier between a trusted network and other untrusted networks -- such as the Internet -- or less-trusted networks -- such as a retail merchant's network outside of a cardholder data environment -- a firewall controls access to the resources of a network through a positive control model.

10. Give the firewall characteristics

A firewall is a protective system that lies, in essence, between your computer network and the Internet. When used correctly, a firewall prevents unauthorized use and access to your network. The job of a firewall is to carefully analyze data entering and exiting the network based on your configuration

11. Give the advantages of application level gateway

Advantages:

Application inspection firewalls can prevent more kinds of attacks than stateful firewalls can. For example, application inspection firewalls can stop an attacker from trying to set up a virtual private network (VPN) tunnel (triggered from inside the network) through an application firewall by way of tunneled HTTP requests.

12. Define bastion host

A **bastion host** is a special purpose computer on a network specifically designed and configured to withstand attacks. The computer generally hosts a single application, for example a proxy server, and all other services are removed or limited to reduce the threat to the computer. It is hardened in this manner primarily due to its location and purpose, which is either on the outside of a firewall or in a demilitarized zone (DMZ) and usually involves access from untrusted networks or computers.

13. Define Intrusion detection systems.

Intrusion detection (ID) is a type of security management system for computers and networks. An ID system gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization).

14. Define data protection.

The term *data protection* is used to describe both operational backup of data and disaster recovery/business continuity (BC/DR). A data protection strategy should include data lifecycle management (DLM), a process that automates the movement of critical data to online and offline storage

15. Explain about Data Privacy laws and Compliance.

.

The Freedom of Information Act 2000 created a new category of data which extended the definition of "data" in the Data Protection Act to include any information held by a public authority which would not otherwise be caught by the definition. Where information requested under the FOI Act includes information about identifiable individuals, public authorities must consider whether its release would breach the Data Protection Act.

Compliance

Compliance is either a state of being in accordance with established guidelines or specifications, or the process of becoming so. Software, for example, may be developed in compliance with specifications created by a standards body, and then deployed by user organizations in compliance with a vendor's licensing agreement. The definition of *compliance* can also encompass efforts to ensure that organizations are abiding by both industry regulations and government legislation.

PART-B (16 MARKS)

1. Explain about different types of firewalls.
2. Discuss in detail about Network Security Intrusion Detection Systems.
3. Discuss the general technologies involved in Firewall.
4. Discuss in detail about OS level protection
5. Discuss about control against threats. List out the rules for overloading operators with example.
6. Give the types of malicious code and explain
7. Describe in detail about Program security.
8. Explain data privacy principles.
9. Explain about Data Privacy laws and Compliance.

..

UNIT-III
PART –A (2 MARKS)

1. Define MDM

Master data management (MDM) is a comprehensive method of enabling an enterprise to link all of its critical data to one file, called a master file that provides a common point of reference. When properly done, MDM streamlines data sharing among personnel and departments.

2. What are all the implementation styles of MDM? Transaction Style
Registry
Consolidation

3. What is the need for Privacy?

The ability to control the information one reveals about oneself over the Internet, and who can access that information, has become a growing concern. These concerns include whether email can be stored or read by third parties without consent, or whether third parties can continue to track the web sites someone has visited. Another concern is web sites which are visited collect, store, and possibly share personally identifiable information about users.

4. What are all the regulatory requirements?

In general, compliance means conforming to a rule, such as a specification, policy, standard or law. **Regulatory** compliance describes the goal that organizations aspire to achieve in their efforts to ensure that they are aware of and take steps to comply with relevant laws and regulations.

5. Give the manage data control objectives.

Decision making for fisheries policy-making, planning and management relies largely on processed information, not raw data. Data have to be interpreted before they can be utilised. The volume of raw primary data is often very large, and so can only be used effectively if held in a Data Base Management System (DBMS).

6. Give three mission of data governance

Data governance is a control that ensures that the data entry by an operations team member or by an automated process meets precise standards, such as a business rule, a data definition and data integrity constraints in the data model. The data governor uses data quality monitoring against production data to communicate errors in data back to operational team members, or to the technical support team, for corrective action. Data governance is used by organizations to exercise control over processes and methods used by their data stewards and data custodians in order to improve data quality.

7. What are all the goals of data governance?

1. Enable better decision-making
2. Reduce operational friction
3. Protect the needs of data stakeholders
4. Train management and staff to adopt common approaches to data issues
5. Build standard, repeatable processes
6. Reduce costs and increase effectiveness through coordination of efforts
7. Ensure transparency of processes

8. Mention the data related rules

The process of ensuring that a program operates on clean, correct and useful data. It uses routines, often called "validation rules" "validation constraints" or "check routines", that check for correctness, meaningfulness, and security of data that are input to the system. The rules may be implemented through the automated facilities of a data dictionary, or by the inclusion of explicit application program validation logic.

PART-B (16 MARKS)

1. What is need for MDM and Explain?
2. Explain MDM Privacy.
3. Explain Regulatory Requirements and Compliance..
4. Explain the role of data management in regulatory
5. Explain the regulatory compliance through data Management.
6. Give the neat diagram for data governance and explain.
7. What includes in governance in IT? Explain.
8. Mention the data governance program lifecycle and explain.

UNIT-IV

PART –A (2 MARKS)

1. Define Organization systems..

All organizations have a management structure that determines relationships between the different activities and the members, and subdivides and assigns roles, responsibilities, and authority to carry out different tasks. Organizations are open systems--they affect and are affected by their environment.

2. **List Organization schemes.**

1. Alphebetic scheme
2. Chronical scheme
3. Hierarchical scheme
4. Database Oriented Scheme
5. Geographic scheme

4. Summarize types of navigation systems.

A complex web site often includes several types of navigation systems. To design a successful site, it is essential to understand the types of systems and how they work together to provide flexibility and context.

1. Hierarchical Navigation Systems
2. Global Navigation Systems
3. Local Navigation Systems

5. Describe iconic labels.

Labeling or labeling is **describing** someone or something in a word or short phrase. For example, **describing** someone who has broken a law as a criminal. Labeling theory is a theory in sociology which ascribes labeling of people to control and identification of deviant behaviour

6. Discuss Content models.

As the demand for content grows, we need better tools to help us structure it. Content models are an effective way of keeping a multi-disciplined project team aligned in their understanding of structured content.

7. Illustrate Hypertext.

Hypertext is text displayed on a computer display or other electronic devices with references (hyperlinks) to other text which the reader can immediately access, or where text can be revealed progressively at multiple levels of detail (also called StretchText). **8.** Analyse the sources of labeling systems.

9. Explain Social Navigation.

The process of guiding activities aimed at determining our position and planning and following a specific route based on what other people have done or what other people have recommended doing.

PART-B

1. Describe the granularity of content in detail.(16)
2. Discuss the following in detail
   i) Navigation System Choices (6)
   ii) Index Terms (6)
   iii) Iconic Labels (4)
3. i) Describe a Top Down approach in organization structures (8)
   ii) Describe a data base model for bottom-Up Approach (8)
4. Illustrate the following in detail about Embedded Navigation Systems
   i) Global Navigation Systems (6)
   ii) Local Navigation Systems (6)
   iii) Contextual Navigation Systems (4)
5. Briefly explain the following in detail
   i) Personalization and Customization (8)
   ii) Visualization and Social Navigation (8)

UNIT-V

1. Define Data retention policy..

Data retention, also called records retention, is the continued storage of an organization's data for compliance or business reasons. An organization may retain data for several different reasons. One reason is to comply with state and federal regulations.

2.Tell about Confidential/Regulated Data.

confidential data are personal identifiers deemed confidential at Cornell because of their direct link to individuals' financial resources. These identifiers include: Social Security numbers, credit card numbers, drivers license numbers and bank account numbers.A graph in which every edge is directed is called a directed graph.

3.Demonstrate Big data testing strategy.

In Big data testing QA engineers verify the successful processing of terabytes of data using commodity cluster and other supportive components. It demands a high level of testing skills as the processing is very fast. Processing may be of **three types**

1.Batch
2.Real Time
3.Interactive

4.Classify testing of Big Data.

.Volume: big data doesn't sample; it just observes and tracks what happens

- Velocity: big data is often available in real-time
- Variety: big data draws from text, images, audio, video; plus it completes missing pieces through

5.Create an archive in Hadoop.

Creating a Hadoop Archive. Where -archiveName is the name of thearchive you would like to create. The archive name should be given a .har extension. The <parent> argument is used to specify the relative path to the location where the files are to be archived in the HAR.

6.Give the Challenges in Big Data Testing.

Huge Volume and Heterogeneity

Understanding the Data

Dealing with Sentiments and Emotions

Lack of Technical Expertise and Coordination

7. Define Sensitive Data.

Sensitive data encompasses a wide range of information and can include: your ethnic or racial origin; political opinion; religious or other similar beliefs; memberships; physical or mental health details; personal life; or criminal or civil offences. These examples of information are protected by your civil rights.

PART-B

1. i)Briefly describe the requirement for protecting data and data collections based on classification (12) ii)Tell about Canadian Privacy registration (4)
2. Summarize the Information Security and the internet(16)
3. i) Differentiate sensitive information and confidential information?(8) ii) Briefly explain Data protection and human rights Act (8)
4. Explain Challenges in Big Data Testing (16)
5. Prepare a case study for handling confidential information(16)