Module: DATA MINING AND WAREHOUSE

Group D Members:

NIYONSENGA Edithe	**2209000306**
KARANGUZA Anicet	**--**
NSENGIYUMVA Christian	**E/BCS/21/01/14216**
KWIZERA Steven	**E/BIT/20/01/12683**
UMUTESIWASE Marie Claire	**2201000875**
MUGISHA Yves	**2209001122**
NDAYISHIMIYE Jean Claude	**W/BCS/22/01/15611**
NZENGIYUMVA Jean d amour	**2209000565**
RUSHINGABIGWI Emmanuel	**2205001111**

Group D Assignment

Q1. In the process of building Data Warehouse, one need to do Data warehouse modeling.

1.1 Discuss the meaning of Data Warehouse Modeling

Data Warehouse Modeling

Data warehouse modeling is the process of designing the schemas of the detailed and summarized information of the data warehouse. The goal of data warehouse modeling is to develop a schema describing the reality, or at least a part of the fact, which the data warehouse is needed to support.

Data warehouse modeling is an essential stage of building a data warehouse for two main reasons.

Firstly, through the schema, data warehouse clients can visualize the relationships among the warehouse data, to use them with greater ease.

Secondly, a well-designed schema allows an effective data warehouse structure to emerge, to help decrease the cost of implementing the warehouse and improve the efficiency of using it.

Data modeling in data warehouses is different from data modeling in operational database systems. The primary function of data warehouses is to support DSS processes. Thus, the objective of data warehouse modeling is to make the data warehouse efficiently support complex queries on long term information.

In contrast, data modeling in operational database systems targets efficiently supporting simple transactions in the database such as retrieving, inserting, deleting, and changing data. Moreover, data warehouses are designed for the customer with general information knowledge about the enterprise, whereas operational database systems are more oriented toward use by software specialists for creating distinct applications.

1.2 Discuss the need of Data Warehouse

Data Warehouse is needed for the following reasons:

Business User

Store historical data

Make strategic decisions.

For data consistency and quality

High response time

Business User: Business users require a data warehouse to view summarized data from the past. Since these people are non-technical, the data may be presented to them in an elementary form.

Store historical data: Data Warehouse is required to store the time variable data from the past. This input is made to be used for various purposes.

Make strategic decisions: Some strategies may be depending upon the data in the data warehouse. So, data warehouse contributes to making strategic decisions.

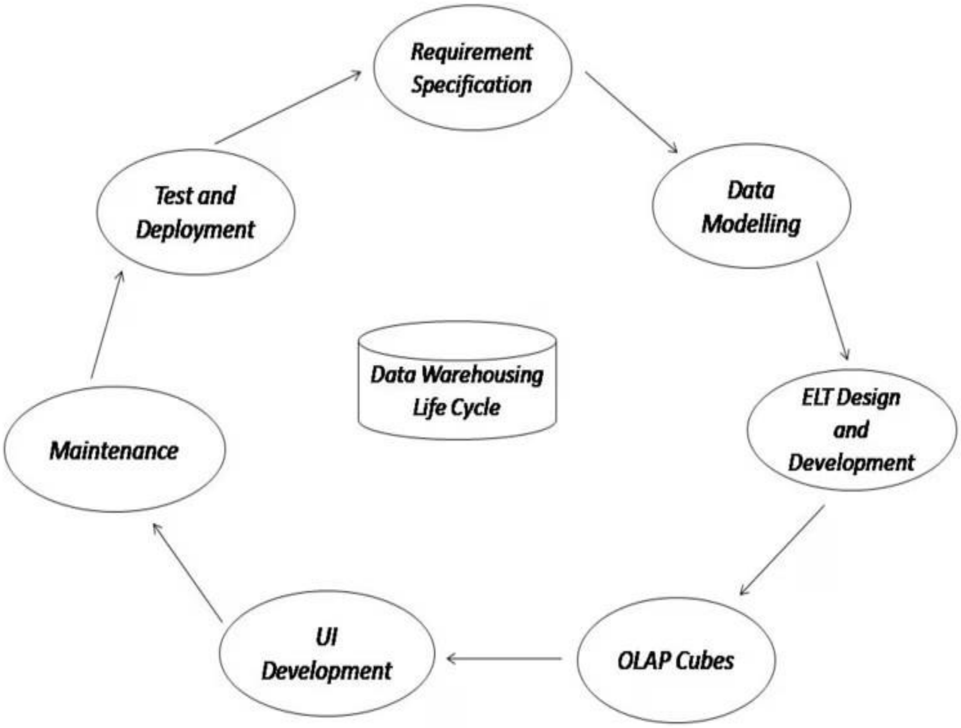
For data consistency and quality: Bringing the data from different sources at a commonplace, the user can effectively undertake to bring the uniformity and consistency in data.

High response time: Data warehouse has to he ready for somewhat unexpected loads and types of queries, which demands a significant degree of flexibility and quick response time.

1.3 What is Data Warehouse Life Cycle

Data warehouse development life cycle

Data Warehousing is a flow process used to gather and handle structured and unstructured data from multiple sources into a centralized repository to operate actionable business decisions. With all of your data in one place, it becomes easier to perform analysis, reporting and discover meaningful insights at completely different combination levels. A data warehouse setting includes extraction, transformation, and loading (ELT) resolution, an online analytical processing (OLAP) engine, consumer analysis tools, and different applications that manage the method of gathering data and delivering it to business. The term data warehouse life cycle is used to indicate the steps a data warehouse system goes through between when it is built. The following is the Lifecycle of Data Warehousing:


Requirement Specification
Data Modelling
ELT Design and Development
OLAP Cubes
UI Development
Maintenance
Test and Deployment
Data Warehousing Life Cycle

Requirement Specification: It is the first step in the development of the Data Warehouse and is done by business analysts. In this step, Business Analysts prepare business requirement specification documents. We can say that this is an overall blueprint of the data warehouse. But, this phase is more about determining business needs and placing them in the data warehouse.

Data Modelling: This is the second step in the development of the Data Warehouse. Data Modelling is the process of visualizing data distribution and designing databases by fulfilling the requirements to transform the data into a format that can be stored in the data warehouse. For example, whenever we start building a house, we put all the things in the correct position as specified in the blueprint. That's what data modeling is for data warehouses. Data modelling helps to organize data, creates connections between data sets, and it's useful for establishing data compliance and its security that line up with data warehousing goals. It is the most complex phase of data warehouse development. And there are many data modelling techniques that businesses use for warehouse design. Data modelling typically takes place at the data mart level and branches out in a data warehouse. It's the logic of how the data is stored concerning other data. **There are three data models for data warehouses:**

- Star Schema
- Snowflake Schema
- Galaxy Schema.

ELT Design and Development: This is the third step in the development of the Data Warehouse. ETL or Extract, Transfer, Load tool may extract data from various source systems and store it in a data lake. An ETL process can extract the data from the lake, after that transform it and load it into a data warehouse for reporting. For optimal speeds, good visualization, and the ability to build easy, replicable, and consistent data pipelines between all of the existing architecture and the new data warehouse, we need ELT tools. This is where ETL tools like SAS Data Management, IBM Information Server, Hive, etc. come into the picture. A good ETL process can be helpful in constructing a simple yet functional data warehouse that's valuable throughout every layer of the organization.

OLAP Cubes: This is the fourth step in the development of the Data Warehouse. An OLAP cube, also known as a multidimensional cube or hypercube, is a data structure that allows fast analysis of data according to the multiple dimensions that define a business problem. A data warehouse would extract information from multiple data sources and formats like text files, excel sheets, multimedia files, etc. The extracted data is cleaned and transformed and is loaded into an OLAP server (or OLAP cube) where information is pre-processed in advance for further analysis. Usually, data operations and analysis are performed using a simple spreadsheet, where data values are arranged in row and column format. This is ideal for two-dimensional data. However, OLAP contains multidimensional data, with data typically obtained from different and unrelated sources. Employing a spreadsheet isn't an optimum choice. The cube will store and analyze multidimensional data in a logical and orderly manner. Now, data warehouses are now offered as a fully built product that is configurable and capable of staging multiple types of data. OLAP cubes are becoming outdated as OLAP cubes can't deliver real-time analysis and reporting, as businesses are now expecting something with high performance.

UI Development: This is the fifth step in the development of the Data Warehouse. So far, the processes discussed have taken place at the backend. There is a need for a user interface for how the user and a computer system interact, in particular the use of input devices and software, to immediately access the data warehouse for analysis and generating reports. The main aim of a UI is to enable a user to effectively manage a device or machine they're interacting with. There are plenty of tools in the market that helps with UI development. BI tools like Tableau or PowerBI for those using BigQuery are great choices.

Maintenance: This is the sixth step in the development of the Data Warehouse. In this phase, we can update or make changes to the schema and data warehouse's application domain or requirements. Data warehouse maintenance systems must provide means to keep track of schema modifications as well, for instance, modifications. At the schema level, we can perform operations for the Insertion, and change dimensions and categories. Changes are, for example, adding or deleting user-defined attributes.

Test and Deployment: This is often the ultimate step in the Data Warehouse development cycle. Businesses and organizations test data warehouses to ensure whether the required business problems are implemented successfully or not. The warehouse testing involves the scrutiny of enormous volumes of data. Data that has to be compared comes from heterogeneous data sources like relational databases, flat files, operational data, etc. The overall data warehouse project testing phases include Data completeness, Data Transformation, Data is loaded by means of ETL tools, Data integrity, etc. After testing the data warehouse, we deployed it so that users could immediately access the data and perform analysis. Basically, in this phase, the data warehouse is turned on and lets the user take the benefit of it. At the time of data warehouse deployment, most of its functions are implemented. The data warehouses can be deployed at their own data center or on the cloud.

Q2. Web mining is an application of data mining techniques to find information patterns from the web data.

2.1. Discuss at least 5 challenges that are associated with web mining.

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

Challenges in Web Mining

The web poses great challenges for resource and knowledge discovery based on the following observations −

- **The web is too huge** − The size of the web is very huge and rapidly increasing. This seems that the web is too huge for data warehousing and data mining.

- **Complexity of Web pages** − The web pages do not have unifying structure. They are very complex as compared to traditional text document. There are huge amount of documents in digital library of web. These libraries are not arranged according to any particular sorted order.

- **Web is dynamic information source** − The information on the web is rapidly updated. The data such as news, stock markets, weather, sports, shopping, etc., are regularly updated.

- **Diversity of user communities** − The user community on the web is rapidly expanding. These users have different backgrounds, interests, and usage purposes. There are more than 100 million workstations that are connected to the Internet and still rapidly increasing.

- **Relevancy of Information** − It is considered that a particular person is generally interested in only small portion of the web, while the rest of the portion of the web contains the information that is not relevant to the user and may swamp desired results.

- **Human activities feedback:** Web page authors provide links to ―authoritative Web pages and also traverse those Web pages they find most interesting or of highest quality. Unfortunately, while human activities and interests change over time, Web links may not be updated to reflect these trends.

- **Effective of deep-Web Extraction:** A research analysts estimated that searchable databases on the Web numbered more than 100,000. These databases provide high-quality, well-maintained information, but are not effectively accessible. Because current Web crawlers cannot query these databases, the data they contain remains invisible to traditional search engines. Conceptually, the deep Web provides an extremely large collection of autonomous and heterogeneous databases, each supporting specific query interfaces with different schema and query constraints. To effectively extract the deep Web, we must integrate these databases and implement efficient web mining approaches.

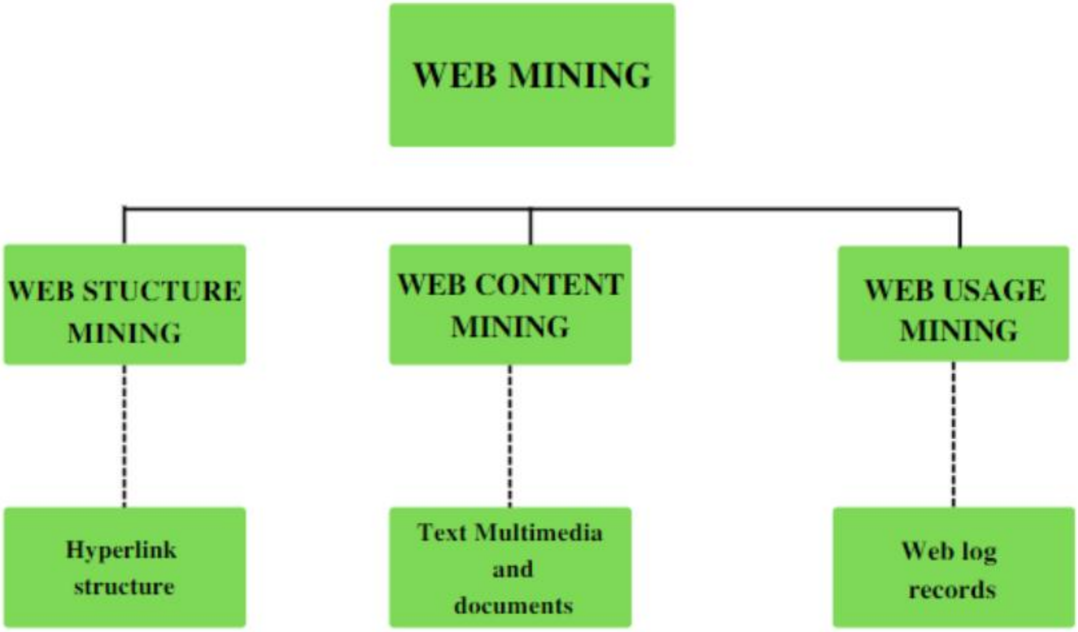
- **Counter Terrorism:** Privacy is a major challenge with respect to data mining for counterterrorism. In this scenario, the challenge is to extract the structure and usage patterns or mine useful information form data mining but at the same time maintain privacy. Different efforts are under way for privacy preserving data.

- **Quality of keyword-based searches:** The quality of keyword-based searches suffers from several inadequacies such as a search often returns many answers, especially if the keywords posed include words from popular categories such as sports, politics, or entertainment. It overloaded keyword semantics and it can return low-quality results. For example, depending on the context, an apple could be a fruit, juice, company or computer and a search can miss many highly related pages

- **Fraud and Threat Analysis:** The main problem issue is that, are they ready for detecting and/or preventing fraud activities and can we completely remove the false positive and false negatives? The challenge is to find how we can gather knowledge directed data mining to eliminate false positives and false negatives. Another challenge of data mining is in real-time. The available tools of data mining have the ability to detect credit card violations and calling card violations. The research community should have a challenge to build a real time model. The challenge is necessary for many companies where they have interactions with up to millions of external parties. Details the subgroups of internal, insurance, credit card, and telecommunications fraud detection which is very concerned for both the researchers and particular organization.

2.2. Discuss are different categories of web mining.

Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.



Categories of Web Mining

Web Content Mining: Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

Web Structure Mining: Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.

Web Usage Mining: Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.

Q3. Security in data mining is essential and vital for data protection.

3.1.Discuss 6 dimension of data security.

There are six dimensions of data security:

Confidentiality: This refers to the protection of data from unauthorized access.

Integrity: This refers to the protection of data from unauthorized modification.

Availability: This refers to the protection of data from unauthorized destruction or loss.

Authenticity: This refers to the protection of data from being tampered with or falsified.

Non-repudiation: This refers to the ability to prove that a particular action was performed by a particular individual or entity.

Traceability: This refers to the ability to track the movement of data within an organization.

These dimensions are interrelated and should be considered together when designing and implementing data security measures. For example, confidentiality and integrity are often implemented together through encryption, which scrambles data so that it cannot be read without a key. Availability can be protected through measures such as backup and disaster recovery, which ensure that data can be restored if it is lost or damaged. Authenticity can be protected through digital signatures, which verify the identity of the sender of a message. Non-repudiation can be protected through logging, which records all actions performed on data. Traceability can be protected through auditing, which analyzes logs to identify suspicious activity.

3.2.Discuss the threats that are likely to affect data mining a web mining

Here are some of the threats that are likely to affect data mining and web mining:

Data breaches: Data breaches are one of the most common threats to data security. When a data breach occurs, sensitive data such as personal information, financial information, or intellectual property can be stolen by unauthorized individuals. This data can then be used for identity theft, fraud, or other malicious purposes.

Malware: Malware is software that is designed to harm a computer system. Malware can be used to steal data, install backdoors, or disrupt operations. Data mining and web mining systems are especially vulnerable to malware attacks because they often collect and store large amounts of sensitive data.

Phishing: Phishing is a type of social engineering attack that is used to trick users into revealing their personal information. Phishing emails often appear to be from legitimate sources, such as banks or credit card companies. When users click on a link in a phishing email, they are redirected to a fake website that looks like the real website. Once the user enters their personal information on the fake website, the attacker can steal it.

Spam: Spam is unsolicited electronic messages that are sent in bulk. Spam can be used to distribute malware, phish for personal information, or simply annoy users. Data mining and web mining systems are especially vulnerable to spam attacks because they often collect and store large amounts of email addresses.

Data poisoning: Data poisoning is a malicious attack that is used to corrupt data. Data poisoning can be used to make data mining and web mining systems produce inaccurate results. For example, an attacker could inject false data into a data mining system in order to skew the results of a prediction model.

Privacy concerns: Data mining and web mining often collect and store large amounts of personal information. This information can be used to track users' online activities, identify their interests, and predict their behavior. This raises privacy concerns for users, who may not want their personal information to be used in this way.

3.4. Discuss the possible solutions for those threats.

Here are some possible solutions to these threats:

Data breaches: Organizations can protect themselves from data breaches by implementing strong security measures, such as encryption, access control, and data validation. They should also monitor their systems for suspicious activity and educate their employees about security risks.

Malware: Organizations can protect themselves from malware attacks by using antivirus software, keeping their software up to date, and being careful about what websites they visit and what emails they open.

Phishing: Organizations can protect themselves from phishing attacks by educating their employees about phishing scams and by using spam filters.

Spam: Organizations can protect themselves from spam by using spam filters and by educating their employees about how to spot spam emails.

Data poisoning: Organizations can protect themselves from data poisoning attacks by using data validation techniques and by monitoring their systems for suspicious activity.

In addition to these specific solutions, organizations can also take general steps to improve their security posture, such as:

Conducting regular security assessments

Implementing a security incident response plan

Having a disaster recovery plan in place