Name: Kunal Sinha Roll no: 2K17/CO/164 Batch: CO-A3

Experiment 1

Aim: Briefly explain various tools used for Data Warehouse and Data Mining.

Theory:

Data Warehouse

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis, and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources. Some of the most popular open-source tools used for Data Warehouse are:

1. **QuerySurge:** QuerySurge is an ETL testing solution developed by RTTS. It is built specifically to automate the testing of Data Warehouses & Big Data. It ensures that the data extracted from data sources remains intact in the target systems as well. Some of its features:
   - Improve data quality, data governance and accelerate your data delivery cycles
   - It helps to automate the manual testing effort and also provides up to 100% data coverage
   - Provide testing across different platforms like Oracle, Teradata, IBM, Amazon etc.
   - It integrates an out-of-the-box DevOps solution for most Build, ETL & QA software
   - Deliver shareable, automated email reports and data health dashboards

2. **Oracle:** Oracle data warehouse software is a collection of data which is treated as a unit. The purpose of this database is to store and retrieve related information. It helps the server to reliably manage huge amounts of data so that multiple users can access the same data.
   - Distributes data in the same way across disks to offer uniform performance
   - Works for single-instance and real application clusters
   - Offers real application testing
   - Common architecture between any Private Cloud and Oracle's public cloud
   - Hi-Speed Connection to move large data
   - Works seamlessly with UNIX/Linux and Windows platforms
   - It provides support for virtualization
   - Allows connecting to the remote database, table, or view

3. **Amazon Redshift:** Amazon Redshift is easy to manage, simple, and cost-effective data warehouse tool. It can analyze almost every type of data using standard SQL.
   - No Up-Front Costs for its installation
   - It allows automating administrative tasks to monitor, manage, and scale your DW
   - Possible to change the number or type of nodes.
   - Helps to enhance the reliability of the data warehouse cluster
   - Every data centre is fully equipped with climate control

- Continuously monitors the health of the cluster. It automatically re-replicates data from failed drives and replaces nodes when needed

4. **Domo:** Domo is a cloud-based Data warehouse management tool that easily integrates various types of data sources, including spreadsheets, databases, social media and almost all cloud-based or on-premise Data warehouse solutions.
   - Help you to build your dream dashboard and Integrates all existing business data
   - Stay connected anywhere you go
   - Helps you to get true insights into your business data
   - Easy Communication & messaging platform
   - It provides support for ad-hoc queries using SQL
   - It can handle most concurrent users for running complex and multiple queries

5. **Teradata Corporation:** The Teradata Database is the only commercially available shared-nothing or Massively Parallel Processing (MPP) data warehousing tool. It is one of the best data warehousing tool for viewing and managing large amounts of data. Features:
   - Simple and Cost Effective solutions with quick and most insightful analytics
   - The tool is best suitable option for an organization of any size
   - Get the same Database on multiple deployment options
   - It allows multiple concurrent users to ask complex questions related to data
   - Offers High performance, diverse queries, and sophisticated workload management

6. **SAP:** SAP is an integrated data management platform, to maps all business processes of an organization. It is an enterprise-level application suite for open client/server systems. It has set new standards for providing the best business information management solutions.
   - It provides highly flexible and most transparent business solutions
   - The application developed using SAP can integrate with any system
   - It follows modular concept for the easy setup and space utilization
   - You can create a Database system that combines analytics and transactions. These next
   - next-generation databases can be deployed on any device
   - Provide support for On-premise or cloud deployment
   - Simplified data warehouse architecture
   - Integration with SAP and non-SAP applications

7. **IBM – DataStage:** IBM data Stage is a business intelligence tool for integrating trusted data across various enterprise systems. It leverages a high-performance parallel framework either in the cloud or on-premise. This data warehousing tool supports extended metadata management and universal business connectivity.
   - Support for Big Data and Hadoop
   - Additional storage/ services can be accessed without installing new software or hardware.
   - Real time data integration

- Provide trusted ETL data anytime, anywhere and solve complex big data challenges
- Optimize hardware utilization and prioritize mission-critical tasks
- Deploy on-premises or in the cloud

8. **Informatica:** Informatica PowerCenter is Data Integration tool developed by Informatica Corporation. The tool offers the capability to connect & fetch data from different sources.
   - It has a centralized error logging system which facilitates logging errors and rejecting data into relational tables
   - Build in Intelligence to improve performance
   - Ability to Scale up Data Integration with enforced best practices on code development.
   - Foundation for Data Architecture Modernization
   - Code integration with external Software Configuration tools
   - Synchronization amongst geographically distributed team members

Data Mining

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Some of the most popular open-source tools used for Data Mining are:

1. **Rapid Miner:** RapidMiner is one of the best predictive analysis systems developed by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It provides an integrated environment for deep learning, text mining, machine learning & predictive analysis. The tool can be used for over a vast range of applications including for business applications, commercial applications, training, education, research, application development, machine learning.
   Rapid Miner offers the server as both on premise & in public/private cloud infrastructures. It has a client/server model as its base. Rapid Miner comes with template based frameworks that enable speedy delivery with reduced number of errors (which are quite commonly expected in manual code writing process). Rapid Miner constitutes of three modules, namely
   - RapidMiner Studio- This module is for workflow design, prototyping, validation etc.
   - RapidMiner Server- To operate predictive data models created in studio
   - RapidMiner Radoop- Executes processes directly in the Hadoop cluster to simplify predictive analysis.

2. **Orange:** Orange is a perfect software suite for machine learning & data mining. It best aids the data visualization and is a component based software. It has been written in Python computing language. As it is a component-based software, the components of orange are called ‘widgets’. These widgets range from data visualization & pre-processing to an evaluation of algorithms and predictive modeling. Widgets offer major functionalities like
   - Showing data table and allowing to select features
   - Reading the data and training predictors and to compare learning algorithms

- Visualizing data elements etc.

Additionally, Orange brings a more interactive and fun vibe to the dull analytic tools. It is quite interesting to operate. Data coming to Orange gets quickly formatted to the desired pattern and it can be easily moved where needed by simply moving/flipping the widgets. Users are quite fascinated by Orange. Orange allows users to make smarter decisions in a short time by quickly comparing & analyzing the data.

3. **Weka:** Also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for data analysis and predictive modeling. It contains algorithms and visualization tools that support machine learning. Weka has a GUI that facilitates easy access to all its features. It is written in JAVA programming language. Weka supports major data mining tasks including data mining, processing, visualization, regression etc. It works on the assumption that data is available in the form of a flat file. Weka can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query.

4. **KNIME:** KNIME is the best integration platform for data analytics and reporting developed by KNIME.com AG. It operates on the concept of the modular data pipeline. KNIME constitutes of various machine learning and data mining components embedded together. KNIME has been used widely for pharmaceutical research. In addition, it performs excellently for customer data analysis, financial data analysis, and business intelligence. KNIME has some brilliant features like quick deployment and scaling efficiency. Users get familiar with KNIME in quite lesser time and it has made predictive analysis accessible to even naive users. KNIME utilizes the assembly of nodes to pre-process the data for analytics and visualization.

5. **Apache Mahout:** Apache Mahout is a project developed by Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering. Mahout is written in JAVA and includes JAVA libraries to perform mathematical operations like linear algebra and statistics. Mahout is growing continuously as the algorithms implemented inside Apache Mahout are continuously growing. The algorithms of Mahout have implemented a level above Hadoop through mapping/reducing templates. To key up, Mahout has following major features
   - Extensible programming environment
   - Pre-made algorithms and math experimentation environment
   - GPU computes for performance improvement.

6. **DataMelt:** DataMelt, also known as DMelt is a computation and visualization environment that provides an interactive framework to do data analysis and visualization. It is designed mainly for engineers, scientists & students. DMelt is written in JAVA and it is a multi-platform utility. It can run on any operating system which is compatible with

JVM(Java Virtual Machine). It contains Scientific & mathematical libraries. Scientific libraries: To draw 2D/3D plots. Mathematical libraries: To generate random numbers, curve fitting, algorithms etc.DataMelt can be used for analysis of large data volumes, data mining, and stat analysis. It is widely used in the analysis of financial markets, natural sciences & engineering.

Findings and Learnings :

We learned about data warehousing and data mining and their respective significance. We also learned about the open source tools which can be used to assist us in the practice of data warehousing and data mining. Finally we learned about the features of the tools used for data warehouse and data mining.