Data Warehousing & Data Mining LAB - G2
EXPERIMENT 1

- ASHISH KUMAR
- 2K18/SE/041

Aim:- List out various open source data warehousing tools and techniques and explain them.

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.

Data Warehousing Tools

Data Warehousing Tools are the software components used to perform various operations on a large volume of data. Data warehousing tools are used to collect, read, write, and migrate large data from different sources. Data warehouse tools also perform various operations on databases, data stores, and data warehouses like sorting, filtering, merging, aggregation, etc.

Some of the most popular open-source tools used for Data Warehousing 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. **Xplenty:** Xplenty is a cloud-based data integration platform to create simple, visualized data pipelines to your data warehouse. It will bring all your data sources together. With Xplenty you will be able to centralize all your metrics and sales tools like your automations, CRM, customer support systems, etc. Xplenty is an elastic and scalable platform for data integration. It can work with structured and unstructured data. It can integrate data with a variety of sources like SQL data stores, NoSQL databases, and cloud storage services.

   - Xplenty can be integrated with a variety of sources like SQL data stores, NoSQL databases, and cloud storage services.
   - It can work with relational databases such as Oracle, Microsoft SQL Server, Amazon RDS, etc.
   - You will be able to connect with online analytical data stores such as AWS Redshift and Google BigQuery.

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 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.

9. Panoply: Panoply is the only smart data warehouse that automates and simplifies all three key aspects of the data lifecycle i.e. data integration, data management, and query performance optimization.

- Panoply allows you to ingest data from any source with just few clicks. This takes minutes not days, which means business users no longer depend on IT/Data Engineering for ETL processes
- Works with popular analytics and business intelligence tools.
- Keeps data stack maintenance to a minimum by handling chores like vacuuming and API updates.

- Table-level data governance ensures you have all the control you need.
- Industry-leading support ranging from robust documentation to expert data architects.
- Panoply learns as you use it. Queries are saved, cached, and continuously optimized, thereby saving your time across all your data analytics reporting tasks. This means lightning-fast queries to fuel any BI tool or statistical package.

With Panoply, you can get a data analytics stack up and running with just a few clicks, thereby saving time, resources, and cost for any size business operating in any industry vertical.

Finding & Learning:- We learned about data warehousing 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. Finally we learned about the features of the tools used for data warehousing.