Program – 1

AIM: List down various open source data warehouse tools.

Introduction and Theory

In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), 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. They store current and historical data in one single place that are used for creating analytical reports for workers throughout the enterprise.
The data stored in the warehouse is uploaded from the operational systems (such as marketing or sales). The data may pass through an operational data store and may require data cleansing for additional operations to ensure data quality before it is used in the DW for reporting.
The typical extract, transform, load (ETL)-based data warehouse uses staging, data integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups, often called dimensions, and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data.

QuerySurge

QuerySurge is 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.
Features:

1. Improve data quality & data governance
2. Accelerate your data delivery cycles
3. Helps to automate manual testing effort
4. Provide testing across the different platform like Oracle, Teradata, IBM, Amazon, Cloudera, etc.
5. It speeds up testing process up to 1,000 x and also providing up to 100% data coverage
6. It integrates an out-of-the-box DevOps solution for most Build, ETL & QA management software
7. Deliver shareable, automated email reports and data health dashboards

MarkLogic

MarkLogic is a data warehousing solution that makes data integration easier and faster using an array of enterprise features. This tool helps to perform very complex search operations. It can query data including documents, relationships, and metadata.
Features:

1. The Optic API can perform joins and aggregates over documents, triples, and rows.
2. It allows specifying more complex security rules for all the elements within documents
3. Writing, reading, patching, and deleting documents in JSON, XML, text, or binary formats
4. Database Replication for Disaster Recovery

5. Specify Output Options on the App Server Configuration
6. Importing and Exporting Configuration Information

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.

Features:

1. Distributes data in the same way across disks to offer uniform performance
2. Works for single-instance and real application clusters
3. Offers real application testing
4. Common architecture between any Private Cloud and Oracle's public cloud
5. Hi-Speed Connection to move large data
6. Works seamlessly with UNIX/Linux and Windows platforms
7. It provides support for virtualization
8. Allows connecting to the remote database, table, or view

Amazon RedShift

Amazon Redshift is an easy to manage, simple, and cost-effective data warehouse tool. It can analyze almost every type of data using standard SQL.

Features:

1. No Up-Front Costs for its installation
2. It allows automating most of the common administrative tasks to monitor, manage, and scale your data warehouse
3. Possible to change the number or type of nodes
4. Helps to enhance the reliability of the data warehouse cluster
5. Every data center is fully equipped with climate control
6. Continuously monitors the health of the cluster. It automatically re-replicates data from failed drives and replaces nodes when needed

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.

Features:

1. Help you to build your dream dashboard
2. Stay connected anywhere you go
3. Integrates all existing business data
4. Helps you to get true insights into your business data
5. Connects all of your existing business data
6. Easy Communication & messaging platform
7. It provides support for ad-hoc queries using SQL
8. It can handle most concurrent users for running complex and multiple queries

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:

Program – 1

1. Simple and Cost Effective solutions
2. The tool is best suitable option for organization of any size
3. Quick and most insightful analytics
4. Get the same Database on multiple deployment options
5. It allows multiple concurrent users to ask complex questions related to data
6. It is entirely built on a parallel architecture
7. Offers High performance, diverse queries, and sophisticated workload management\

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.

Features:

1. It provides highly flexible and most transparent business solutions
2. The application developed using SAP can integrate with any system
3. It follows modular concept for the easy setup and space utilization
4. You can create a Database system that combines analytics and transactions. These next next-generation databases can be deployed on any device
5. Provide support for On-premise or cloud deployment
6. Simplified data warehouse architecture
7. Integration with SAP and non-SAP applications

SAS

SAS is a leading Datawarehousing tool that allows accessing data across multiple sources. It can perform sophisticated analyses and deliver information across the organization.

Features:

1. Activities managed from central locations. Hence, user can access applications remotely via the Internet
2. Application delivery typically closer to a one-to-many model instead of one-to-one model
3. Centralized feature updating, allows the users to download patches and upgrades.
4. Allows viewing raw data files in external databases
5. Manage data using tools for data entry, formatting, and conversion
6. Display data using reports and statistical graphics

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.

Features:

1. Support for Big Data and Hadoop
2. Additional storage or services can be accessed without need to install new software and hardware
3. Real time data integration
4. Provide trusted ETL data anytime, anywhere
5. Solve complex big data challenges
6. Optimize hardware utilization and prioritize mission-critical tasks
7. Deploy on-premises or in the cloud

Program – 1

Informatica:

Informatica PowerCenter is Data Integration tool developed by Informatica Corporation. The tool offers the capability to connect & fetch data from different sources.

Features:

1. It has a centralized error logging system which facilitates logging errors and rejecting data into relational tables
2. Build in Intelligence to improve performance
3. Limit the Session Log
4. Ability to Scale up Data Integration
5. Foundation for Data Architecture Modernization
6. Better designs with enforced best practices on code development
7. Code integration with external Software Configuration tools
8. Synchronization amongst geographically distributed team members

Findings and Learnings:

1. We learned about data warehousing and its significance.
2. We learned which tools can be used to assist us.
3. We learned about the features of the tools.