WIE3007 DATA MINING AND WAREHOUSING - Tutorial 1

1. Discuss the significance of having a layered architecture in Data Warehousing. How does each layer contribute to the overall functionality and efficiency of a Data Warehouse?

Layer	Significance	Contribution
Source Layer	Extract data from various source systems such as databases, operational systems and external data sources.	It ensures data is gathered from different sources efficiently and can be transformed into a consistent format.
Staging Layer	Store and centralize all of your data before it's transformed	To consolidate data to build models to power your business.
Modeling Layer	To transform and merge data into coherent models that answer important questions unique to business.	Models built here can be reused effectively across multiple use cases and business units.
Presentation Layer	Enable users to interact with the data stored in the data warehouse.	Optimize query performance.

2. Your organization is experiencing an exponential increase in data volumes. How would you modify the existing Data Warehouse architecture and infrastructure to ensure scalability and maintain optimal performance? Consider modifications in different layers of the architecture.

Layer	Improvement
Source Layer	● **Data Prioritization** ○ Prioritize critical data sources and cleaning data that can be excluded. ● **Real-time Data Ingestion** ○ To ensure data freshness
Staging Layer	● **Batch Processing** ○ Enhance batch processing capabilities to handle larger volumes of data efficiently. ● **Data Compression**

	○ Reduce storage and processing requirements. ● **Data Validation** ○ Ensure data quality, eliminate unuseful data and reduce cost for data cleaning later.
Modeling Layer	● **Data Partitioning** ○ Enhance query performance especially for large datasets. ● **Data Aggregation** ○ To pre-calculate and store aggregated data for commonly used queries, reducing query response times.
Presentation Layer	● **Query Optimization** ○ Ensure efficient data retrieval. ● **Data Caching** ○ Store frequently accessed data to reduce the load on the data warehouse. ● **Data Indexing** ○ Speed up data retrieval.

3. Explain how you would design a secure Data Warehouse architecture. Discuss the security measures that can be implemented at different layers of the architecture to protect sensitive data and ensure compliance with data protection regulations.

Layer	**Security Measure**
Source Layer	● **Data Encryption** ○ Implement encryption for data in transit to protect data as it moves from source system to data warehouse. ● **Access Control** ○ Restrict access to source systems and ensure that only authorized people can access the data.
Staging Layer	● **Secure Data Transfer** ○ Ensure that the staging environment is encrypted and protected from unauthorized access. ● **Data Validation** ○ Detects and rejects data anomalies or potential security threats by implementing validation rules.
Modeling Layer	● **Data Retention Policy** ○ Enforce data retention policies to remove or archive

	data when it is no longer needed to reduce the risk of data exposure.
Presentation Layer	● **Authentication and Authorization** ○ Implement authentication and authorization mechanisms to control access. ○ Role-based access control to grant varying levels of access to different users. ● **Data Encryption** ○ Encrypt data transmitted from the data warehouse to the presentation layer.

4. Your company is planning to update its Data Warehouse technology stack. Propose a set of technologies and tools for each layer of the Data Warehouse architecture, providing justifications for your choices based on the nature of the business, data volumes, and user requirements.

Layer	**Technologies**
Source Layer	● **Data Extraction: Talend Data Quality** can be used for data profiling, cleansing, and validation. It ensures the data quality before it enters the data warehouse.
Staging Layer	● **Data Storage: Amazon S3** for storing large volumes of data. It is scalable and cost-effective.
Modeling Layer	● **Data Warehouse: Google BigQuery** for scalable, cloud-based data warehousing solutions with built-in security and query optimization features.
Presentation Layer	● **Reporting and Analytics: Tableau** or **Power BI** for interactive, user-friendly dashboards and reports.