Assessment Details for Assessment Item 2: Report - Design business intelligence system and data warehouse.

### 1. BI Architecture Overview

The Business Intelligence (BI) architecture is designed to provide real-time insights from the retail store's data by integrating data collection, processing, and reporting mechanisms. This system is built on a central cloud-based database that gathers real-time data from multiple store branches and enables advanced analytics through tools like Power BI. The architecture includes the following components:

#### Data Sources:

The BI system gathers data from various store branches in real-time. Each store’s operational database feeds into the central system with key data categories such as:

* **Sales Data**: Real-time transactional information, including invoice numbers, product names, quantity, unit price, and total sales amounts. This data will provide insights into store performance, best-selling products, and revenue trends.
* **Customer Information**: Demographic data such as age, gender, and customer loyalty (e.g., membership status). This will help in segmenting customers for personalized marketing and customer behavior analysis.
* **Product Lines**: Categorization of products into specific groups such as "Food & Beverages," "Electronics," "Clothing," etc. This helps to analyze sales performance at the product category level.
* **Customer Ratings**: Feedback provided by customers on a scale of 1 to 10, capturing their satisfaction with the service and product offerings.
* **Payment Methods**: Detailed information on transaction payments, including whether purchases were made through credit cards, eWallets, or cash. This will allow analysis of payment trends and customer preferences.

To enable real-time data collection from these sources, tools such as **Apache Kafka** or **AWS Kinesis** will be utilized. These tools allow for the streaming of data from various stores' operational systems (such as POS terminals or customer databases) to the central cloud database, ensuring data is captured continuously without delays.

#### ETL Process:

The **Extract, Transform, Load (ETL)** process is fundamental to converting raw data from the retail stores into a form suitable for analysis. The real-time ETL process ensures that data is immediately available for reporting and analysis.

* **Extraction**:
  + **Tools**: **Apache NiFi** or **AWS Glue** will be used to automate the data extraction process. These tools are equipped for real-time data extraction, enabling data flow from the operational systems (POS, CRM, etc.) at each store into the central cloud-based database.
  + Data will be extracted continuously from these systems, capturing transactional, customer, and product information as soon as it’s generated.
* **Transformation**:
  + **Tools**: **Talend** or **Azure Data Factory** will handle the transformation step. The transformation process includes:
    - **Data Cleaning**: Handling missing or incomplete data, such as missing customer demographic information or inconsistent product descriptions.
    - **Standardization**: Ensuring consistent data formats, such as converting different date formats into a standard format (e.g., YYYY-MM-DD) and standardizing currency formats for consistency.
    - **Normalization**: Aggregating data where necessary, such as combining sales data from multiple store branches into a unified format for analysis. For example, aligning product names and categories to a common naming convention.
    - **Validation**: Ensuring data integrity by validating that the data conforms to predefined rules, such as valid payment methods or product IDs.
* **Loading**:
  + **Tools**: The cleaned and transformed data is then loaded into a cloud-based data warehouse using platforms like **Amazon Redshift** or **Google BigQuery**. These platforms are optimized for large-scale data storage and are capable of handling real-time data ingestion.
  + The data warehouse is structured to support efficient querying, with data organized into **fact** and **dimension tables**. The use of a **Star Schema** design is recommended, where fact tables store transactional data (e.g., sales data) and dimension tables contain descriptive attributes (e.g., customer details, product categories).

#### Data Storage:

The central cloud-based database serves as the core data repository for the BI system.

* **Cloud-Based Storage**: **Amazon Redshift**, **Google BigQuery**, or **Microsoft Azure SQL Data Warehouse** will be used to store and manage the large volumes of data from multiple stores. These platforms offer scalability, high availability, and fast querying capabilities.
* **Fact and Dimension Tables**: The data warehouse will be structured using a **Star Schema** or **Snowflake Schema**:
  + **Fact Tables**: These tables will store transactional data, such as sales figures, product quantities sold, and revenue generated.
  + **Dimension Tables**: These tables will provide contextual information, such as customer demographics, store locations, product categories, and payment methods.

By using cloud-based storage, the system will benefit from real-time data availability, which is crucial for enabling immediate reporting and analytics.

#### Data Access & Reporting:

The BI system will be connected to tools like **Power BI** and **Tableau**, which will serve as the primary platforms for data access, visualization, and reporting.

* **Visualization Tools**: **Power BI** will be used for creating interactive dashboards and visualizations. It will allow users to explore the data through drag-and-drop interfaces, enabling real-time analysis of sales performance, customer behavior, and product trends. Visualizations may include:
  + Sales trends over time (e.g., sales by day, week, or month).
  + Product performance by category or store.
  + Customer segmentation analysis based on demographic data.
* **Real-Time Reporting**: Since the data is stored in a cloud-based environment and updated in real-time, users will have access to up-to-the-minute reports and dashboards. **Power BI’s DirectQuery** feature will allow the tool to query data from the cloud database directly, ensuring real-time reporting without the need for manual data refreshes.
* **Ad-Hoc Analysis**: Users can perform ad-hoc analysis using **Excel** connected to the cloud database, allowing for flexible reporting and in-depth analysis.

#### Security & Permissions:

Ensuring data security and proper access controls is critical in this architecture.

* **User Access Control**: Tools like **Azure Active Directory** or **AWS Identity and Access Management (IAM)** will be used to manage user roles and permissions. These tools ensure that only authorized users can access sensitive data. Different levels of access will be granted based on the user's role (e.g., store managers, regional managers, or corporate executives).
* **Data Encryption**: All data transmitted between the stores and the cloud, and between the cloud database and Power BI, will be encrypted using **SSL/TLS** protocols.
* **Data Privacy**: To comply with data privacy regulations such as **GDPR** or **CCPA**, customer data will be anonymized or masked where necessary. For example, sensitive fields such as customer names and contact information will be stored in encrypted formats.

#### Conclusion:

This BI architecture is designed to provide a robust, scalable, and secure platform for real-time data analytics. With real-time data sourcing from stores, a well-structured ETL process, and powerful visualization tools like Power BI, the system will enable the retail store to make data-driven decisions. The cloud-based architecture ensures scalability, while advanced security measures safeguard sensitive information, providing a complete and efficient solution for the store’s business intelligence needs.