WIE3007 DATA MINING AND WAREHOUSING 2024/2025 S1

Assignment (15%) – Individual

You are tasked with designing an optimized data model for an e-commerce dataset using Featuretools and modern cloud-based data warehousing technologies. The dataset contains information about customers, products, orders, and order details.

Scenario:

You are provided with an e-commerce dataset comprising several entities:

1. Customers: with attributes like CustomerID, Name, Email, SignupDate.
2. Products: with attributes like ProductID, Name, Category, Price.
3. Orders: with attributes like OrderID, CustomerID, OrderDate, ShipDate.
4. OrderDetails: with attributes like OrderID, ProductID, Quantity, Discount.

Customers:
CustomerID: [101, 102, 103]
Name: ['John Doe', 'Jane Smith', 'Mike Jordan']
Email: ['john.doe@example.com', 'jane.smith@example.com', 'mike.jordan@example.com']
SignupDate: ['2023-01-10', '2023-01-15', '2023-01-20']

Products:
ProductID: [201, 202, 203]
Name: ['Laptop', 'Tablet', 'Smartphone']
Category: ['Electronics', 'Electronics', 'Electronics']
Price: [1000, 500, 800]

Orders:
OrderID: [301, 302, 303]
CustomerID: [101, 102, 103]
OrderDate: ['2023-02-01', '2023-02-05', '2023-02-10']
ShipDate: ['2023-02-03', '2023-02-07', '2023-02-12']

OrderDetails:
OrderID: [301, 302, 303]
ProductID: [201, 202, 203]
Quantity: [1, 2, 1]
Discount: [0, 0.1, 0]

Requirements:

- **Automated Feature Engineering**:
    - Use **Featuretools** or **Google Cloud AutoML** to perform feature engineering on the dataset.

- Compare feature engineering results from **cloud-based solutions** with traditional methods.

- **Star Schema Design**:

  - Design a **Star Schema** for a cloud-based data warehouse (e.g., Snowflake, Google BigQuery).
  - Simulate **real-time data ingestion** using streaming tools like **Apache Kafka**.

- **Data Governance and Security**:

  - Design a data governance and security strategy to comply with modern data privacy regulations (e.g., **GDPR**, **CCPA**).

- **Serverless ETL Pipeline**:

  - Implement a **serverless function** using AWS Lambda or Google Cloud Functions to automate a part of your ETL pipeline.

- **Data Visualization**:

  - Use **Tableau** or **Power BI** to create a simple dashboard that visualizes insights from your data.

Deliverables:

- **Feature Engineering Report**: Summarize the results and insights gained using Featuretools and cloud-based tools.
- **Star Schema Diagram**: Depict the optimized schema for cloud implementation.
- **Data Dictionary**: Include detailed attributes for fact and dimension tables.
- **Data Governance Plan**: Explain the security and governance strategy.
- **Python Code**: Provide the code for feature engineering and serverless ETL.
- **Dashboard**: A simple visualization of key metrics from the e-commerce data.

Incorporating modern cloud-based technologies into this assignment can introduce some potential costs, but there are ways to minimize or eliminate these costs for you. Here's an overview of potential costs and how to manage them:

1. Cloud-based Data Warehousing (e.g., Snowflake, Google BigQuery, AWS Redshift)

- **Cost Involvement**:
  - Most cloud-based data warehousing platforms have free tiers or trial credits. For example:
    - **Snowflake** offers $400 in free credits for 30 days.
    - **Google BigQuery** offers a free tier with 1 TB of query processing per month.
    - **AWS Redshift** offers a free trial with 750 hours of usage for 2 months.
- **Cost Management**:

    - students to make use of **free tiers**. Set assignment limits that can be handled within the free-tier constraints (e.g., limit dataset size and query frequency).
    - Ensure you understand how to monitor usage to avoid exceeding free-tier limits.

2. Featuretools (Open-source Python Library)

- **Cost Involvement**:
    - **Featuretools** is an open-source Python library, so there are no direct costs associated with using it.
- **Cost Management**: No additional steps are needed, as this tool is free.

3. Google Cloud AutoML / AWS Sagemaker Autopilot / Azure Machine Learning

- **Cost Involvement**:
    - **Google Cloud AutoML** has a free tier that includes up to 40 hours of free training time on lightweight models.
    - **AWS Sagemaker Autopilot** offers free-tier usage, including 250 hours of **ml.t2.medium** instances for 2 months.
    - **Azure Machine Learning** offers a free tier with limited training and inference instances.
- **Cost Management**:
    - students to utilize **free credits** and restrict the usage to smaller datasets and limited processing time to stay within the free-tier limits.
    - Set expectations on data volume and complexity to fit within free usage limits.

4. Serverless Computing (AWS Lambda, Google Cloud Functions, Azure Functions)

- **Cost Involvement**:
    - **AWS Lambda** offers 1 million free requests per month, and 400,000 GB-seconds of compute time.
    - **Google Cloud Functions** offers 2 million free invocations per month.
    - **Azure Functions** provides 1 million free executions per month.
- **Cost Management**:
    - For assignments involving serverless functions, ensure that workloads remain lightweight (e.g., simple ETL jobs) to stay well within the free-tier limits.

5. Data Streaming Services (Apache Kafka, AWS Kinesis, Google Cloud Pub/Sub)

- **Cost Involvement**:
    - **Apache Kafka** can be run locally for free, although setting it up on cloud services could incur costs.
    - **AWS Kinesis** offers free-tier usage with 1 million PUT transactions per month.
    - **Google Cloud Pub/Sub** offers 10 GB of messages per month for free.
- **Cost Management**:
    - Encourage students to simulate data locally or use mock streaming datasets. Cloud services should be used within free-tier constraints to avoid incurring extra charges.

6. Visualization Tools (Tableau, Power BI, Looker)

- **Cost Involvement**:
    - **Tableau Public** is free, and **Power BI** offers a free version with basic features.

    - **Google Data Studio** is another free alternative for building dashboards.
- **Cost Management**:
    - Instruct students to use the free versions of these tools (e.g., **Tableau Public** or **Google Data Studio**) to complete the assignment.

7. NoSQL Databases (MongoDB, Amazon DynamoDB, Cassandra)

- **Cost Involvement**:
    - **MongoDB Atlas** provides 512 MB of storage for free.
    - **Amazon DynamoDB** offers 25 GB of storage and 25 write capacity units for free.
    - **Cassandra** can be run locally for free, or students can use managed services with trial credits.
- **Cost Management**:
    - Use the free-tier options for NoSQL databases. Set limitations on data volume to ensure students stay within these constraints.

8. Version Control and CI/CD Pipelines (GitHub Actions, AWS CodePipeline, Azure DevOps)

- **Cost Involvement**:
    - **GitHub Actions** provides 2,000 free minutes per month for private repositories.
    - **AWS CodePipeline** offers a free tier for up to 1,000 pipeline executions per month.
    - **Azure DevOps** provides 1,800 minutes of pipeline usage per month for free.
- **Cost Management**:
    - Encourage students to use **GitHub** for version control and pipeline automation, as it offers more than enough free minutes for the scope of most student projects.

General Cost Management Strategies:

- **Free-Tier and Credits**: Encourage you to sign up for cloud providers' **free-tier** accounts and make use of free trial credits. how to monitor your usage to avoid exceeding free limits.
- **Simulation**: For some tasks like streaming data ingestion or serverless computing, simulate the process locally using smaller datasets, reducing the need for cloud-based execution.
- **Budget Awareness**: how to set **budgets** and **usage alerts** on cloud platforms to ensure you don't accidentally incur costs.

Conclusion:

With careful planning and guidelines, you can complete the assignment using free-tier offerings from major cloud providers and open-source tools. By providing clear instructions on how to manage resources, avoid any unnecessary costs while still gaining experience with modern technologies.

Evaluation Criteria (Total: 15%):

1. **Feature Engineering with Featuretools or AutoML (3%)**:
    - **Entity and EntitySet Creation (1%)**: Evaluate whether entities and relationships were correctly defined.

    - **Feature Synthesis (1%)**: Assess the quality and relevance of new features generated.
    - **Insights Gained (1%)**: Award marks based on the student's ability to extract meaningful insights from the features.
2. **Star Schema Design (5%)**:
    - **Fact and Dimension Tables (2%)**: Accurate and complete identification of tables, correct usage of primary/foreign keys.
    - **Schema Relationships (1%)**: Appropriate relationships between tables that reflect the dataset structure.
    - **New Features Incorporation (1%)**: Successful inclusion of new features discovered during feature engineering.
    - **Scalability (1%)**: Consideration of schema design for scalability in cloud environments.
3. **Data Governance and Security Plan (2%)**:
    - **Security Strategy (1%)**: Clear and accurate explanation of data governance practices and privacy regulations (e.g., GDPR, CCPA).
    - **Data Access Control (1%)**: Design of role-based access control using modern cloud tools (e.g., **AWS IAM**, **Azure Active Directory**).
4. **Python Code Quality and Serverless ETL (2%)**:
    - **Correctness of Code (1%)**: Code should run correctly and meet the assignment requirements.
    - **Code Quality and Documentation (1%)**: Code should be clean, well-structured, and documented.
5. **Data Visualization and Dashboard (2%)**:
    - **Dashboard Clarity (1%)**: Effectiveness of the dashboard in presenting key metrics.
    - **Data Presentation (1%)**: Use of appropriate charts/graphs to communicate insights.
6. **Reflection on Modern Technologies (1%)**:
    - **Use of Modern Tools (1%)**: Students should reflect on how the modern tools (e.g., cloud services, Featuretools) influenced their workflow and design decisions.

Due Date: Week 7, 5pm.