**DEPI Round 1 – Final Project Assessment**

**Part 1: Final Project Plan**

|  |  |
| --- | --- |
| **Instructor Name:** | Shrief Mohamed Abdelaziz |
| **Track Name:** | Microsoft Data Engineer |
| **Group Code:** | CAI1\_AIS4\_S7e |

|  |  |
| --- | --- |
| **Project Title:** | **Customer data management and analysis** |
| **Start and End Date Project:** | **Start Date: 18/9/2024** |
| **End Date: 18/10/2024** |
| **Project Team:** | * Ayaat Ahmed * Shahd Ahmed * Maryam Mohamed * Renad Haytham * Reem Salah * Salma Ashraf |
| **Project Objectives:** | The primary objective of this Data Engineering project is to create an end-to-end pipeline that facilitates the efficient handling, storage, and analysis of customer and transaction data. The specific goals of the project are:  Data Ingestion:  Extract customer and transaction data from multiple sources (such as databases, CSV files, or APIs).  Clean, transform, and format the data to ensure consistency and quality for storage.  Data Warehouse (DWH) Setup:  Design and create a Data Warehouse to store the customer and transaction data using optimal schema design (star or snowflake schema).  Load the cleaned and transformed data into the DWH using ETL (Extract, Transform, Load) processes.  Data Integration with Azure:  Transfer the data from the Data Warehouse to an Azure-based platform (e.g., Azure Data Lake or Azure SQL Database) for further processing and scalability.  Ensure the data pipeline to Azure is efficient, secure, and reliable, supporting real-time or batch data transfers as needed.  Data Analysis:  Conduct exploratory data analysis (EDA) on the customer and transaction data to uncover patterns, trends, and insights.  Create visualizations and reports to help stakeholders make data-driven decisions. |
| **Project Methodology:**  *(Outline the methods, tools, or techniques you plan to use to achieve the project objectives)* | The methodology for this Data Engineering project involves a systematic approach to ingest, store, process, and analyze customer and transaction data using Azure's cloud services, Databricks for machine learning, and various SQL-based tools. The project will follow a hybrid cloud-based and on-premises approach to ensure scalability, efficiency, and data security. The key steps of the methodology include:  1. Data Ingestion & ETL:  Azure Data Factory (ADF):  Azure Data Factory will be used to orchestrate and automate the data ingestion process from various data sources such as on-premises databases, APIs, or cloud storage.  ADF pipelines will manage the Extract, Transform, and Load (ETL) processes by transforming raw customer and transaction data into a usable format.  SQL Server Integration Services (SSIS):  SSIS will be used to perform additional ETL processes within the on-premises environment, loading data into the Data Warehouse for further analysis.  2. Data Warehousing:  SQL Server/SSIS:  A structured Data Warehouse (DWH) will be implemented using SQL Server to store historical customer and transaction data. This provides a single source of truth for analysis.  SSIS will help in managing the data load into the DWH from different data sources.  SQL backups will be implemented to ensure data integrity and provide data recovery mechanisms.  3. Data Storage & Integration:  Azure Data Lake Storage (ADLS):  The processed data will be stored in Azure Data Lake for scalable storage and further data integration.  ADLS will serve as a central repository for raw, transformed, and processed data, making it easily accessible for different downstream processes.  4. Data Transformation & Processing:  Azure Synapse Analytics:  Synapse Analytics will be used to create large-scale analytics queries and to transform data for various business use cases.  It integrates with Azure Data Lake and supports both on-demand and provisioned resources for efficient data processing.  5. Data Analysis and Reporting:  Power BI:  Power BI will be used for real-time data analysis, visualization, and reporting. Reports will be generated to offer insights into customer behavior, transaction trends, and business performance.  Power BI will also provide dashboards for decision-makers to interact with the data.  6. Machine Learning & Advanced Analytics:  Azure Databricks:  Databricks will be utilized to perform advanced analytics and machine learning (ML) tasks. The customer and transaction data will be used to build ML models, such as customer segmentation, predictive analysis, or recommendation systems.  Databricks notebooks will allow collaboration between data scientists and engineers to ensure efficient model deployment.  7. Backup & Recovery:  SQL Backups:  Regular backups of the SQL Data Warehouse will be performed to ensure that the data can be recovered in case of failures or data corruption.  Azure SQL Backup and Restore functionalities will be integrated to provide seamless backup options for cloud data.  8. Automation & Monitoring:  Azure Data Factory Monitoring:  Azure Data Factory will provide automated scheduling of pipelines, allowing for the periodic ingestion of new data. Data quality checks and error handling mechanisms will be embedded in the pipelines.  Monitoring tools within Azure will be used to track pipeline performance, trigger alerts for failures, and ensure the integrity of data processing.  9. Deployment & Maintenance:  The final solution will be deployed in the Azure environment with ongoing monitoring, optimizations, and updates as needed to ensure performance at scale.  Maintenance tasks will include ensuring the continuous flow of data, performance tuning, and enhancing machine learning models based on updated data. |
| **Project Resources and Tools:**  *(List any materials, software, or resources you will need to complete the project)* | To successfully complete this Data Engineering project, the following materials, software, and resources will be required:  1. Software and Platforms:  Azure Services:  Azure Data Factory: For orchestrating and automating data pipelines.  Azure Synapse Analytics: For large-scale data transformation and analytics.  Azure Data Lake Storage (ADLS): For storing raw and processed data.  Azure Databricks: For advanced analytics and machine learning model building.  Azure SQL Database: For storing and managing data in a structured format.  On-Premises Tools:  SQL Server/SQL Server Integration Services (SSIS): For managing data warehousing and running ETL processes.  SQL Backup Tools: For ensuring data security and recovery.  Visualization and Reporting Tools:  Power BI: For creating data visualizations, reports, and dashboards.  2. Development Tools:  Python/PySpark: For scripting ETL processes, data transformations, and machine learning tasks.  SQL Management Studio: For managing databases and writing SQL queries.  Tutorials on setting up Azure Data Factory, Azure Synapse, and Databricks.  Step-by-step guides for using Azure Data Lake and Power BI for reporting.  Videos on deploying machine learning models in Databricks and integrating with Synapse.  Medium Articles:  Articles on best practices for building ETL pipelines in Azure.  Deep dives into using Databricks for machine learning and advanced analytics.  Case studies on data warehouse design and Power BI dashboarding.  Example sources:  Towards Data Science (Medium)  Azure Data Engineering (Medium) |
| **Project Expected Outcomes:**  *(Describe the anticipated results or deliverables of the project)* | The anticipated results and deliverables of this Data Engineering project include a fully operational data pipeline and analysis system that integrates customer and transaction data across various platforms. The key expected outcomes are:  1. End-to-End Data Pipeline:  A seamless and automated data pipeline, managed by Azure Data Factory and SSIS, that regularly ingests, cleans, transforms, and loads customer and transaction data into a Data Warehouse.  The pipeline will support real-time or scheduled batch processing to handle incoming data efficiently.  2. Data Warehouse Implementation:  A well-structured Data Warehouse hosted in SQL Server, storing cleansed and integrated customer and transaction data.  The data will be organized into a star or snowflake schema, enabling fast query execution and analytics.  3. Data Integration in Azure:  The customer and transaction data will be successfully transferred to Azure Data Lake and Azure Synapse Analytics, making it accessible for further analysis and scalable for growing data volumes.  Data security measures and compliance with regulatory standards (e.g., data encryption, access control) will be implemented.  4. Analytics and Reporting Dashboards:  Power BI dashboards and reports will be developed, providing real-time insights into customer behaviors, transaction trends, and business performance.  The reports will offer detailed visualizations (e.g., charts, graphs, heatmaps) to help stakeholders make data-driven decisions.  5. Machine Learning Models:  Machine learning models built using Azure Databricks will provide advanced analytics capabilities, such as customer segmentation, prediction of transaction patterns, and personalized recommendations.  The models will be integrated into the data pipeline, allowing for continuous updates as new data is ingested.  6. Backup and Recovery Mechanisms:  A reliable backup and recovery system for the SQL Data Warehouse will be established, ensuring that all critical data is securely stored and can be recovered in case of system failures or data corruption.  7. Scalability and Automation:  The system will be fully scalable to accommodate future data growth, with automated processes for data ingestion, transformation, and load.  Azure-based automation tools, such as Azure Monitor and ADF Monitoring, will be used to track performance, detect anomalies, and ensure the continuous operation of the data pipeline.  8. Improved Decision-Making:  Business users will be empowered to make faster and more informed decisions based on the comprehensive insights provided by Power BI reports and machine learning outcomes.  Actionable insights into customer trends and transaction patterns will help drive business strategies, improve customer experiences, and optimize operations.  9. Documentation and Knowledge Transfer:  Detailed project documentation will be created, including:  ETL pipeline workflows, data flow diagrams, and architecture.  Technical documentation on how to manage, monitor, and scale the system.  User guides for generating reports and accessing machine learning models.  Knowledge transfer sessions will be conducted with the team to ensure smooth handover and future scalability of the system. |
| **Project Risk Assessment:**  *(Identify any potential challenges that may arise during the project. Propose solutions plans for handling them)* | Data Integration Complexity  Description: Integrating data from multiple sources (on-premises and cloud) into Azure services like Data Factory, Synapse, and Data Lake can be complex. Inconsistent data formats, missing values, or poor data quality can lead to delays and errors in processing.  Proposed Solution:  Implement a robust data validation and cleansing process using Azure Data Factory or Databricks to ensure data is clean before ingestion.  Design standardized data formats for all incoming data, using conversion tools or scripting to ensure consistency.  Perform regular data quality checks and use automated alerts for missing or inconsistent data. |

**Part 2: Final Project Evaluation**

**Evaluation Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criteria | Description | Excellent  (90-100%) | Good (80-89%) | Satisfactory  (70-79%) | Needs Improvement (<70%) |
| Depth of Research | Check if students have provided a thorough **literature review** or background analysis, using relevant sources. Evaluate the **accuracy and relevance** of the content to the course objectives. | In-depth, thorough, and insightful research | Good research, minor gaps | Sufficient research, but lacks depth | Limited or superficial research |
| Methodology and Execution | Examine if the chosen methodology was appropriate and if the project followed a structured approach. Look at the **problem-solving process** and whether students overcame challenges logically. | Flawless, well-organized, and logical execution | Strong, minor errors present | Adequate but disorganized in parts | Poor planning and lack of focus |
| Creativity and Innovation | Determine whether the student approached the topic in a **unique way**, proposed **new ideas**, or **presented solutions** that differ from standard approaches. | Highly original and creative | Good creativity, some original ideas | Some creativity, but mostly standard | Lacks creativity, very basic |
| Practical Application | Examine whether the project has **real-world relevance** and if the students successfully connected theory to practical use. Consider the potential **impact** of the project. | Demonstrates real-world relevance or impact | Has some practical relevance | Limited practical connection | Little to no real-world relevance |
| Presentation Skills | Evaluate the **clarity, organization, and confidence** in the oral presentation. Look at the quality and effectiveness of any **visual aids.** | Clear, confident, and engaging | Good, with minor issues | Adequate, lacks engagement | Unclear, unorganized, hard to follow |
| Written Report | Review the **structure** and **quality** of required reports/documentations. | Well-structured, no errors | Clear, minor errors | Basic structure, some errors | Disorganized, many errors |
| Teamwork | Evaluate the **distribution of roles** and whether each team member contributed equally. Assess the team’s ability to **collaborate** effectively and manage any conflicts. | Excellent collaboration | Good teamwork, minor issues | Some collaboration, uneven | Poor teamwork, unequal contribution |

**First Week**

|  |  |
| --- | --- |
| **Date:** | 28/9/2024 |
| **Deliverables:** | 26/8/2024  o A well-designed SQL database schema and populated database.  o SQL queries for data extraction and basic analysis. |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student ID** | **Student Name** | **Depth of Research (%)** | **Methodology and Execution (%)** | **Creativity and Innovation (%)** | **Practical Application (%)** | **Presentation Skills (%)** | **Written Report (%)** | **Teamwork (%)** |
| 21017428 | Shahd Ahmed |  |  |  |  |  |  |  |
| 21050066 | Ayat Ahmed |  |  |  |  |  |  |  |
| 21042779 | Renad Haytham |  |  |  |  |  |  |  |
| 21045055 | Maryam Mohamed |  |  |  |  |  |  |  |
| 21014734 | Salma Ashraf |  |  |  |  |  |  |  |
| 21015311 | Reem Salah |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**Second Week**

|  |  |
| --- | --- |
| **Date:** | 5/10/2024 |
| **Deliverables:** | 4/10/2024  o A functioning SQL Data Warehouse with integrated data.  o Python scripts for data extraction and preparation. |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student ID** | **Student Name** | **Depth of Research (%)** | **Methodology and Execution (%)** | **Creativity and Innovation (%)** | **Practical Application (%)** | **Presentation Skills (%)** | **Written Report (%)** | **Teamwork (%)** |
| 21017428 | Shahd Ahmed |  |  |  |  |  |  |  |
| 21050066 | Ayat Ahmed |  |  |  |  |  |  |  |
| 21042779 | Renad Haytham |  |  |  |  |  |  |  |
| 21045055 | Maryam Mohamed |  |  |  |  |  |  |  |
| 21014734 | Salma Ashraf |  |  |  |  |  |  |  |
| 21015311 | Reem Salah |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**Third Week**

|  |  |
| --- | --- |
| **Date:** | 12/10/2024 |
| **Deliverables:** | 10/10/2024  o Analysis report with insights and predictive models.  o Integrated Azure Data services setup and documentation. |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student ID** | **Student Name** | **Depth of Research (%)** | **Methodology and Execution (%)** | **Creativity and Innovation (%)** | **Practical Application (%)** | **Presentation Skills (%)** | **Written Report (%)** | **Teamwork (%)** |
| 21017428 | Shahd Ahmed |  |  |  |  |  |  |  |
| 21050066 | Ayat Ahmed |  |  |  |  |  |  |  |
| 21042779 | Renad Haytham |  |  |  |  |  |  |  |
| 21045055 | Maryam Mohamed |  |  |  |  |  |  |  |
| 21014734 | Salma Ashraf |  |  |  |  |  |  |  |
| 21015311 | Reem Salah |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**Fourth Week**

|  |  |
| --- | --- |
| **Date:** | 19/10/2024 |
| **Deliverables:** | 16/10/2024  o Deployed machine learning model.  o Final report and presentation. |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student ID** | **Student Name** | **Depth of Research (%)** | **Methodology and Execution (%)** | **Creativity and Innovation (%)** | **Practical Application (%)** | **Presentation Skills (%)** | **Written Report (%)** | **Teamwork (%)** |
| 21017428 | Shahd Ahmed |  |  |  |  |  |  |  |
| 21050066 | Ayat Ahmed |  |  |  |  |  |  |  |
| 21042779 | Renad Haytham |  |  |  |  |  |  |  |
| 21045055 | Maryam Mohamed |  |  |  |  |  |  |  |
| 21014734 | Salma Ashraf |  |  |  |  |  |  |  |
| 21015311 | Reem Salah |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**Final Project Delivery Evaluation**

|  |  |
| --- | --- |
| **Date:** | 19/10/2024 |
| **Project Final Evaluation and Grade (%):** | *Comment by project’s evaluation according to its Functionality and Performance, Tools integrations, Design and Architecture, Creativity, Alignment with Project Requirements, Real-World Relevance, Scalability and Maintainability* |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Student ID** | **Student Name** | **Training Final Evaluation**  *Comment by student’s evaluation according to Knowledge Acquisition and application, Skills Development, Attendance and Participation, Learning Goal Achievement, Tasks Completion Rate* | **Final Grade (%)** |
| 21017428 | Shahd Ahmed |  |  |
| 21050066 | Ayat Ahmed |  |  |
| 21042779 | Renad Haytham |  |  |
| 21045055 | Maryam Mohamed |  |  |
| 21014734 | Salma Ashraf |  |  |
| 21015311 | Reem Salah |  |  |
|  |  |  |  |
| **Student ID** | **Student Name** | **Project Final Evaluation**  *Comment by student’s evaluation according to Visual Presentation, Functional Output, Oral Presentation,*  *Live Demonstration, Deadline submission, Team Collaboration* | **Final Grade (%)** |
| 21017428 | Shahd Ahmed |  |  |
| 21050066 | Ayat Ahmed |  |  |
| 21042779 | Renad Haytham |  |  |
| 21045055 | Maryam Mohamed |  |  |
| 21014734 | Salma Ashraf |  |  |
| 21015311 | Reem Salah |  |  |
|  |  |  |  |