UBER DATA ENGINEERING ETL PROJECT USING AZURE

BUSINESS PROPOSAL

نزلني يسطا
TRANSPORTATION

CONTENTS

1. Project Overview
2. Introduction
3. Project Objectives
4. Scope of Work
5. Team Responsibilities
6. Azure Services Overview
7. Key Deliverables
8. Project Timeline
9. Conclusion

نزلني يسطا

TRANSPORTATION

1. INTRODUCTION

- We propose the development of a comprehensive Uber-like application that leverages Azure's cloud platform for both real-time and historical data processing. The goal of this project is to design an optimized database that captures trip data for a ride-hailing service, processes this data in real-time, and performs historical analysis. Our system will enable the business to improve decision-making, enhance user experience, and optimize operational efficiency.

2. PROJECT OBJECTIVES

THE OBJECTIVES OF THIS PROJECT ARE TO:

BUILD A DATABASE THAT STORES REAL-TIME TRIP DATA (PICKUP, DROP-OFF, DURATION, FARE, DRIVER INFO, ETC.).

IMPLEMENT A ROBUST ETL PIPELINE THAT PROCESSES REAL-TIME DATA FROM THE APPLICATION AND HISTORICAL DATA FOR ANALYTICAL PURPOSES.

ANALYZE HISTORICAL TRIP DATA TO PROVIDE INSIGHTS INTO TRENDS SUCH AS PEAK TIMES, POPULAR ROUTES, AND CUSTOMER BEHAVIOR.

INTEGRATE MACHINE LEARNING MODELS TO PREDICT DEMAND, OPTIMIZE PRICING STRATEGIES, AND ENHANCE OPERATIONAL EFFICIENCY.

DESIGN DYNAMIC DASHBOARDS IN POWER BI FOR BUSINESS STAKEHOLDERS TO VIEW REAL-TIME KPIS AND HISTORICAL TRENDS.

3. SCOPE OF WORK

THE PROJECT ENCOMPASSES BOTH **REAL-TIME** DATA PROCESSING AND **HISTORICAL** ANALYSIS USING AZURE SERVICES:

نزلني يسطا
TRANSPORTATION

Real-time Data Processing

Streaming data from ride requests, ongoing trips, and driver locations.

Capturing trip events in real-time, performing necessary transformations, and storing the data in a structured format.

Historical Data Processing

Archiving completed trips and analyzing historical data to identify trends, anomalies, and opportunities for improvement.

Database Design

Design a schema that can handle both real-time and historical data efficiently.

Implement best practices for scalability, data consistency, and data durability.

4. TEAM RESPONSIBILITIES

نزلني يسطا
TRANSPORTATION

Mahmoud Ahmed (Team Leader)	**Role**: Oversight and coordination of the team. Lead the data transformation and warehouse design process
Abdelrahman Mohamed	**Role**: Handle **Azure Synapse Analytics** and manage data storage in **Azure Data Lake**.
Marwan Atef	**Role**: Manage data ingestion using batch and stream processing, orchestration with **Azure Data Factory**.
Hisham Mohamed	**Role**: Implement **Machine Learning models** for demand forecasting, pricing optimization, and trip classification.
Mohamed Ahmed	**Role**: Design **Power BI dashboards** for business reporting and insights.

5. AZURE & CLOUD SERVICES AND APACHE SOFTWARE OVERVIEW

نزلني بسطا
TRANSPORTATION

AZURE DATA FACTORY: TO BUILD, ORCHESTRATE, AND MONITOR THE ETL PIPELINES.

AZURE DATABRICKS: FOR DATA TRANSFORMATION, CLEANING, AND INTEGRATION WITH MACHINE LEARNING MODELS.

AZURE DATA LAKE STORAGE GEN2: AS A SCALABLE AND SECURE STORAGE SOLUTION FOR BOTH RAW AND PROCESSED DATA.

AZURE SYNAPSE ANALYTICS: FOR MANAGING AND QUERYING LARGE DATASETS, BOTH REAL-TIME AND HISTORICAL.

AZURE MACHINE LEARNING: TO BUILD, TRAIN, AND DEPLOY MACHINE LEARNING MODELS FOR DEMAND FORECASTING AND PRICING OPTIMIZATION.

POWER BI: TO CREATE REAL-TIME AND HISTORICAL DASHBOARDS FOR BUSINESS USERS TO TRACK PERFORMANCE AND TRENDS.

APACHE KAFKA: FOR BUILDING REAL-TIME DATA PIPELINES AND STREAMING PROCESSING

DOCKER: STREAMLINE THE ENTIRE WORKFLOW, FROM DEVELOPING AND TESTING ETL PROCESSES ON A LOCAL MACHINE TO DEPLOYING THEM IN PRODUCTION ENVIRONMENTS.

6. KEY DELIVERABLES

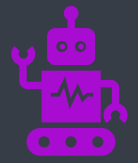
نزلني يسطا
TRANSPORTATION

Database:

Trip Data Storage in a Docker container running SQL Server for scalable, isolated storage.

ELT (Extract, Load, Transform):

Use Azure Synapse Analytics to manage the ELT pipeline, performing transformations and loading the processed data for advanced analytics.



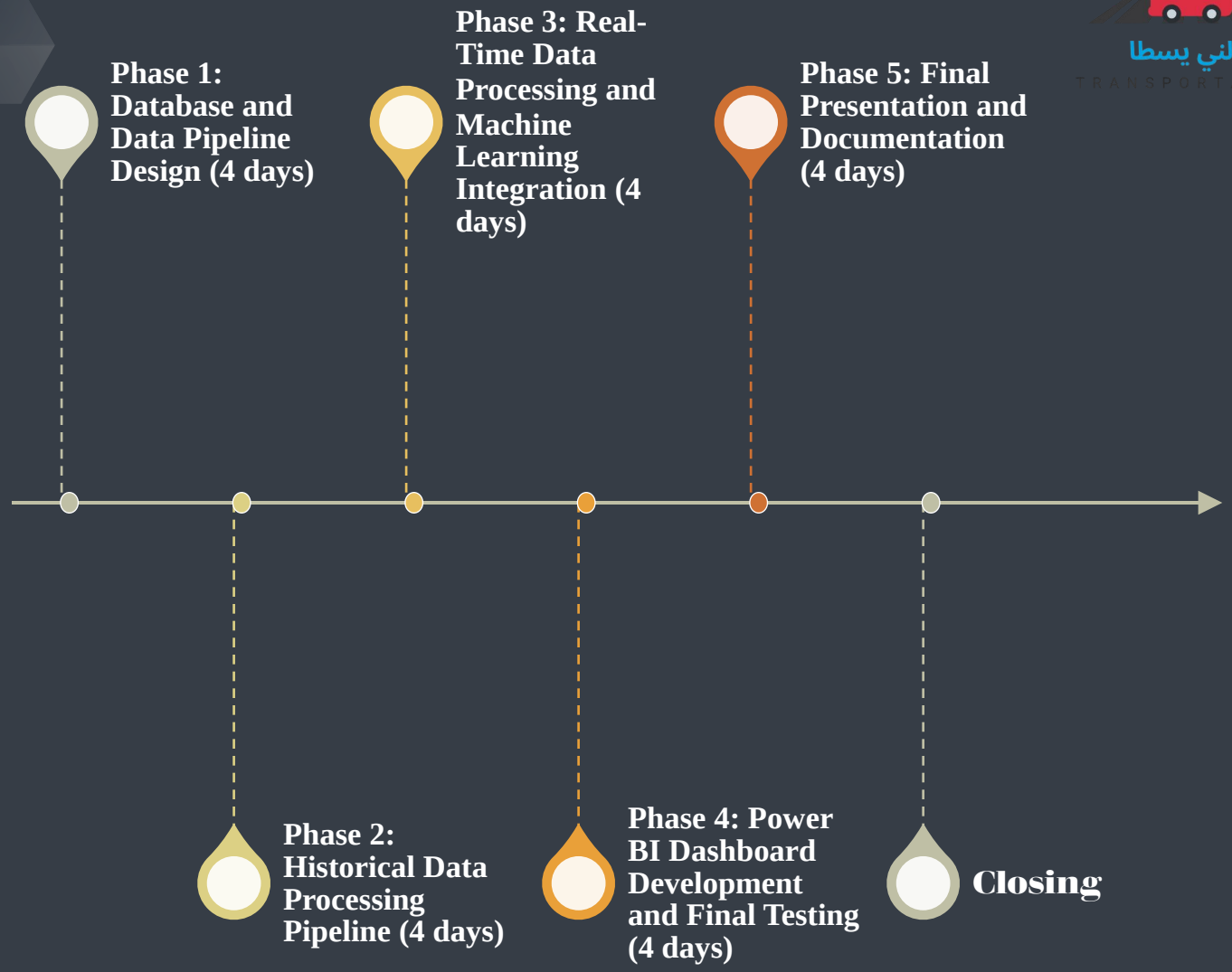
Machine Learning Interactive Model:

Analysis of stored trip data, implementation of ML models.

Power BI Dashboard:

Interactive Visualizations to track KPIs such as total trips, customer satisfaction, driver ratings and revenue trends.

7. TIMELINE


Phase 1: Database and Data Pipeline Design (4 days)
Phase 2: Historical Data Processing Pipeline (4 days)
Phase 3: Real-Time Data Processing and Machine Learning Integration (4 days)
Phase 4: Power BI Dashboard Development and Final Testing (4 days)
Phase 5: Final Presentation and Documentation (4 days)
Closing

8. CONCLUSION

- This Uber data engineering project is designed to provide valuable insights into both real-time and historical ride data, empowering businesses with actionable information. By leveraging the power of Azure services such as Data Factory, Synapse, Databricks, and Power BI, we aim to deliver a comprehensive and scalable solution that can support ride-hailing businesses. Our project will not only enhance operational efficiency but also serve as a real-world example of how cloud-based technologies can drive business growth.