Graduation Project Documentation

Project Title:

Smart Data Integration and Analytics Platform for Scalable Data Warehousing

Team Members:

Ahmed Hany Hafez

Youssef Mohammad Mehanna

Mariam abdelkader

Rokia Mohammad

Instructor:

Eng. Ahmed Elsaid

Project Planning

Objective:
Develop a **smart and scalable data engineering system** that automates data integration, transformation, storage, and visualization using advanced ETL processes, distributed computing, and cloud analytics.
The project aims to provide a robust and efficient platform for organizations to process and analyze large datasets in real-time and gain actionable business insights.

Scope:

Automate the extraction, cleaning, and transformation of structured and unstructured data.

Implement **ETL workflows** using Talend and Python.

Design a **Data Warehouse** using hybrid architecture (SQL Server + Hadoop).

Deploy and orchestrate pipelines in **Azure Cloud** for scalability and performance.

Build **Power BI dashboards** for analytical reporting and visualization.

Milestones:

ETL Pipeline Design & Implementation (Talend + Python).

Database and Data Warehouse Modeling.

Big Data Integration with Hadoop and Hive.

Cloud Deployment via Azure Data Services.

Interactive Dashboard Development using Power BI.

Technologies Used:
Talend, Python, SQL Server, Hadoop, Hive, Apache Airflow, Azure Data Lake, Power BI.

Stakeholder Analysis

Stakeholder	Role	Interest	Responsibility
Ahmed Hany Hafez	Project Leader / Data Engineer	High	Leads project planning, ensures deadlines and quality standards.
Youssef Mohammad Mehanna	Data Engineer	High	Develops ETL workflows, manages data integration and modeling.
Eng. Ahmed Elsaid	Instructor / Supervisor	High	Provides technical guidance and evaluates deliverables.
Cloud Administrator	Manages Azure Cloud resources	Medium	Ensures scalability, security, and cloud optimization.
End Users (Business Analysts)	Data consumers	Medium	Use dashboards for data-driven insights and performance tracking.

Database Design

Overview:
The system employs a **Galaxy Schema** Data Warehouse structure that integrates multiple **Fact** and **Dimension** tables.
This schema supports analytical queries, time-based reporting, and performance optimization for large-scale datasets.

Database Goals:

Provide a unified, consistent, and scalable data storage solution.

Maintain historical data through **Slowly Changing Dimensions (SCD Type 2)**

Enable multi-dimensional analytics (Customer, Driver, Vehicle, Location, Time, etc.).

Support both real-time and batch processing using hybrid architecture (SQL + Hadoop).

Schema Components

Dimension Tables:

Dim_Customer:

`customer_id`, `customer_name`, `phone`, `email`, `signup_date`, `city`, `created_at`.

Stores customer information and registration details.

Dim_Driver:

`driver_id`, `driver_name`, `phone`, `license_number`, `join_date`, `scd_start`, `scd_end`, `city`, `scd_active`.

Maintains driver records with Slowly Changing Dimensions to track history.

Dim_Vehicle:

`vehicle_id`, `model`, `make`, `plate_number`, `capacity`, `year`, `color`, `insurance_expiry`, `scd_start`, `scd_end`, `scd_active`.

Contains all vehicle details and operational data.

Dim_Location:

`location_id`, `city`, `area`, `pickup_location`, `dropoff_location`.

Defines the geographic hierarchy for trip analysis.

Dim_Payment_Method:

`method_id`, `method_name`, `created_at`.

Stores supported payment methods.

Dim_Date:

`date_id`, `full_date`, `day`, `month`, `quarter`, `year`, `weekday`.

Time dimension to support daily, monthly, and quarterly trend analysis.

Fact Tables:

Fact_Rides:

`ride_id`, `driver_id`, `customer_id`, `vehicle_id`, `date_id`, `fare_amount`, `distance_km`, `duration_minutes`, `ride_status`, `pickup_time`, `dropoff_time`, `tip_amount`, `total_amount`, `location_id`, `status`.

Stores transactional data of each ride including financials and duration.

Fact_Payments:

`payment_id`, `ride_id`, `customer_id`, `method_id`, `date_id`, `amount`, `discount`, `payment_status`.

Contains payment transactions and related metrics.

Fact_Ratings:

`rating_id`, `ride_id`, `driver_id`, `customer_id`, `comment`, `rating_score`, `date_id`.

Captures user feedback and service quality scores.

Integration & Optimization

Data flows through the **ETL Pipeline** (Talend + Python) into SQL Server and Hadoop.

Hadoop handles **large-scale data** (logs, big files) via **HDFS** and **Hive**.

Transformed data is loaded into **Azure SQL Database** for Power BI visualization.

Airflow automates workflows, ensuring reliability and scheduling.

Optimization Techniques:

Partitioning and indexing of Fact Tables.

Incremental data loads for better performance.

Data backup and versioning via Azure Blob Storage.

Validation and error handling at each ETL stage.

UI/UX Design

Dashboard Platform: Power BI

Dashboard Features:

Executive Overview: KPIs (Total Rides, Revenue, Active Users, Average Rating).

Sales & Revenue Dashboard: Monthly and regional performance visualization.

Customer Insights: Retention rates, engagement metrics, user demographics.

Driver & Vehicle Analytics: Driver performance and utilization analysis.

Payment Insights: Payment method usage, revenue breakdown.

ETL Monitoring Dashboard: Job status, data freshness, and last load time.

Design Principles:

Consistent color palette (blue, white, gray).

Clear typography and visual hierarchy.

Interactive filters for region, time, driver, and payment method.

Responsive design compatible with desktop and web.

Conclusion

The **Smart Data Integration and Analytics Platform** successfully demonstrates how a modern data engineering ecosystem operates — from ETL automation to big data processing and cloud-based analytics.

By integrating **Talend, Hadoop, Azure, and Power BI**, the project delivers a scalable, reliable, and insightful data solution suitable for real-world enterprise applications.

The advanced **database design** ensures data consistency, historical tracking, and analytics readiness — empowering stakeholders with actionable insights and business intelligence.