Project Designed Document

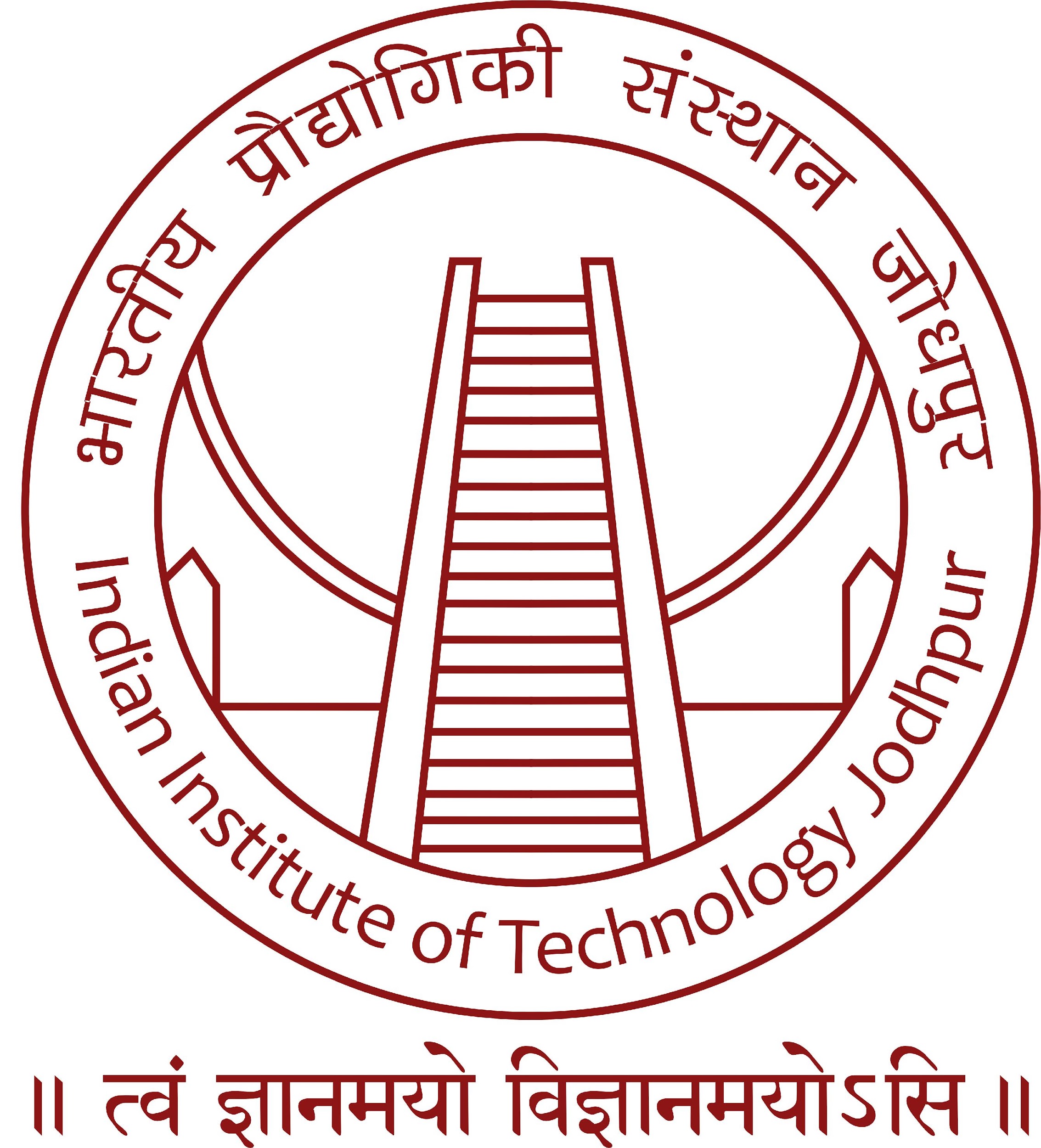
A Pg. Diploma Project Report

Submitted by

**Prince Dubey (G23AI2014)**

Under the Supervision

Of

Prof. Sumit Kalra

Department of Artificial Intelligence

Indian Institute of Technology Jodhpur

### **Project Plan Outline**

### **Define Objectives and KPIs**

**Objective**

XYZ Corp seeks to develop a comprehensive, year-long data engineering project. The primary goal is to transform raw flight data (2015, 2016, and 2017) into actionable business insights using Python, Google Cloud Platform (GCP), and MLOps. The project will involve building a robust data pipeline, including data cleaning, augmentation, ingestion, transformation, warehousing, reporting, and visualization. This project will simulate the journey of a new data engineer and provide business strategies to optimize performance and decision-making

**KPIs**:

* **On-Time Performance (OTP)**: Punctuality.
* **Load Factor**: Optimizing aircraft capacity is essential for profitability.
* **Revenue Kilometers per Passenger**: Measures the total distance traveled by paying passengers.
* **Cost per Available Seat Kilometer (CASK)**: A financial metric.
* **Customer Satisfaction**: Ensures passenger experience.
* **Ancillary Revenue**: Additional income sources.
* **Sustainability**: Focuses on environmental impact.

**Data Source**

* **Internal Data**:
  + There is no internal data available.
* **External Data**:
  + Various online sources.
  + Reddit.
  + Kaggle.

### **Project Phases and Deliverables**

### **1. Data Collection and Preparation**

* **Data Cleaning and Preprocessing**:
  + Remove duplicates, fill in missing data, and ensure data quality for accurate ML models

#### **Phase 2: Data Ingestion and Exploration (todo)**

* **Objective:** Acquire, explore, and understand raw flight data.
* **Key Activities:**
  + Connect to flight data sources (local/cloud databases, APIs, etc.).
  + Load data into GCP Cloud Storage.
  + Initial exploratory data analysis (EDA) using Python to assess data quality, missing values, and outliers.
  + Define the schema and metadata for data warehousing.
* **Deliverables:**
  + Data ingestion pipelines.
  + Data quality report (including insights into completeness, accuracy, and distribution).
  + Exploratory data analysis report.

#### **Phase 3: Data Cleaning and Transformation (todo)**

* **Objective:** Clean and standardize flight data for analysis.
* **Key Activities:**
  + Implement data cleaning processes (handle missing data, remove duplicates, and standardize formats).
  + Apply transformations to enrich data (e.g., time-zone normalization, data augmentation with external sources like weather).
  + Create automated data pipelines using GCP Dataflow and Python scripts.
* **Deliverables:**
  + Data cleaning scripts.
  + Transformation logic and pipeline (including augmentation).
  + Documented data dictionaries and schemas for the cleaned data.

#### **Phase 4: Data Warehousing and Storage (todo)**

* **Objective:** Set up a scalable data warehouse for the cleaned data.
* **Key Activities:**
  + Set up BigQuery for data warehousing.
  + Design efficient data models and partitioning strategies (fact/dimension tables for flights, passengers, routes).
  + Load cleaned and transformed data into BigQuery.
  + Implement security and governance (e.g., access controls, data retention policies).
* **Deliverables:**
  + Data warehouse schema design.
  + Data loading pipelines (ETL scripts).
  + Access control policies and documentation.

#### **Phase 5: Data Analysis and Machine Learning Integration (todo)**

* **Objective:** Use machine learning models to derive business insights.
* **Key Activities:**
  + Identify relevant business problems (e.g., flight delay predictions, customer segmentation).
  + Build machine learning models using Python and TensorFlow.
  + Integrate ML pipelines with MLOps (Kubeflow) for continuous model training and deployment.
* **Deliverables:**
  + Trained ML models (e.g., delay prediction, customer segmentation).
  + MLOps pipeline for automated model retraining and deployment.
  + Performance reports on models (accuracy, precision, recall, etc.).

#### **Phase 6: Reporting and Visualization (todo)**

* **Objective:** Build reports and dashboards to visualize business insights.
* **Key Activities:**
  + Design and create interactive dashboards using Google Data Studio/Looker.
  + Create detailed reports showing key metrics, trends, and business recommendations.
  + Implement real-time dashboards to monitor flight operations.
* **Deliverables:**
  + Interactive business dashboards (flight performance, customer segmentation, etc.).
  + Static reports on key business insights.
  + Recommendations for business strategy optimization.

#### **Phase 7: Final Presentation (todo)**

* **Objective:** Deliver the final solution and ensure smooth handoff to XYZ Corp's internal team.
* **Key Activities:**
  + Deliver final presentations to stakeholders, outlining key insights and recommendations.
  + Provide all documentation and codebase (pipeline scripts, models, etc.).
  + Ensure proper documentation of maintenance and future scalability guidelines.
* **Deliverables:**
  + Final presentation.
  + Complete project documentation (data pipeline, models, visualization).
  + Handoff of data warehouse, ML models, and dashboards.

### **Project Milestones:**

|  |  |  |
| --- | --- | --- |
| **Phase** | **Timeline** | **Key Deliverables** |
| Data Ingestion and Exploration | todo | Data pipelines, EDA report |
| Data Cleaning and Transformation | todo | Cleaned data, transformation logic, pipelines |
| Data Warehousing and Storage | todo | BigQuery setup, ETL scripts, schema design |
| Data Analysis and Machine Learning | todo | ML models, MLOps pipeline, performance reports |
| Reporting and Visualization | todo | Dashboards, reports, business insights |
| Final Presentation and Handoff | todo | Final presentation, documentation, knowledge transfer |