# Background:

You have recently joined AxionRay, a leading technology company revolutionizing engineering quality and safety with AI-driven solutions. AxionRay's innovative tools leverage cutting-edge AI technologies, including Generative AI and Large Language Models (LLMs), to address complex challenges in data processing, automation, and analytics for next-generation products like electric vehicles and airplanes.

This assignment is designed to evaluate your skills in data validation, cleaning, integration, and exploratory data analysis (EDA) using Python, as well as your ability to derive actionable insights.

# Task 1: Data Validation

**Data for task 1 - Link**

**Guidelines:**

1. **Column-Wise Analysis:**
   a. Perform a column-wise analysis of the provided dataset.
   b. Describe each column in terms of its data type, unique values, distribution, and overall significance for stakeholders
2. **Data Cleaning:**
   a. Handle missing or invalid values using appropriate methods (e.g., imputation, deletion).
   b. Address inconsistencies in categorical columns (e.g., typos, inconsistent capitalization).
   c. Ensure numerical columns are in the correct format and free from outliers, where applicable.
3. **Identifying Critical Columns:**
   a. Select the **top 5 critical columns** that might be most insightful for stakeholders according to your data understanding.
   b. Provide reasoning for your selection.
   c. Generate visualizations (e.g., bar plots etc) using Python to represent these insights effectively. **(atleast 3)**
4. **Generating tags/features from free text available :**
   a. Generate meaningful tags from the free text fields to summarize information, example - failure conditions and components etc etc..
5. **Overall Synthesis/key takeaways:** ***(Food for thought and has bonus marks)***
   a. Highlight discrepancies in the dataset (e.g., null values, missing primary keys) and how did you approach.
   b. A summary of the tags generated, including potential insights derived from the dataset.
   c. Provide actionable recommendations for stakeholders based on your analysis.
   d. Discuss any additional observations or key findings.

**Deliverables for Task 1:**

1. Submit the cleaned and tagged records in **CSV or Excel format**.
2. Submit a detailed report covering the above pointers, including (Maximum 2 page):
   a. Column analysis
   b. Data cleaning summary and
   c. Visualizations
   d. Generated tags & Key takeaways
3. Attach Python scripts used for the analysis.

# Task 2: Data Preparation and Integration

**Data for task 2 - Link**

**Guidelines:**

1. **Primary Key Identification**
   a. Analyze both datasets and identify a **Primary Key** for integration.
   b. Justify your selection and highlight potential challenges in identifying a unique key.
2. **Data Cleaning**
   a. Load the datasets using Python.
   b. Inspect the column structure, and clean the data:
      i. Handle missing values, duplicates
      ii. Format Correction - Consistent data types across dataset
      iii. Apply **language translation** if applicable.
   c. Provide a brief summary of your data cleaning process.
3. **Data Integration**
   a. Merge the two datasets on the identified primary key to create a comprehensive view of the datasets.
   b. Choose the appropriate type of join (inner, left, etc.), and justify your choice. Discuss the implications of using other join types in this context.

# Task 3: Exploratory Data Analysis (EDA)

**Data for task 3: Use the merged data from task 2**

**Guidelines -**

1. **Trend Analysis:** Analyze the merged dataset to identify trends or patterns.
   a. Use visualizations (e.g., line charts, heatmaps) to present your findings **(at least 3)** and provide interpretations.

Ex: Correlation between failed component and cost/actual hrs.

2. **Root Cause Identification**: Identify and investigate the trends of failure/fix condition and Failure/fix component and summarise your observation for stakeholders

**Deliverables for Task 2 and 3:**

1. Submit the cleaned and merged dataset in csv/excel format.
2. Attach Python scripts used for the analysis.
3. A short summary document for stakeholders explaining **(Maximum 2 page)** -
    a. cleaning process,
    b. Integration approach, and
    c. Rationale for join type selection.
    d. Identified trends and investigation of 3.1 & 3.2.

# Overall Evaluation Criteria:

- Demonstrates clear understanding of the dataset provided in summary reports.
- Present trends and root causes clearly, with well-supported insights derived from exploratory data analysis.
- Visualizations are relevant and easy to interpret.
- Outputs, including scripts, datasets, and reports, are well-organized, professional, and demonstrate attention to detail.