**LSDA Project**

Seeya Thakre – 11

Snehal Rakas - 12

Tanya Thawani – 13

Yadhnika Wakde – 14

Yashika Jain – 15

**Airline Data Analysis using Dask**

**Objective**

The main goal of this project is to perform large-scale data analysis on an airline dataset using Dask. The project aims to analyze passenger and flight data to extract useful insights such as flight trends, passenger demographics, and country-wise flight distribution.

**Introduction**

In today’s world, airlines generate massive amounts of data every single day. Handling and analyzing such large datasets efficiently requires scalable tools. Traditional tools like Pandas can become slow or crash when data grows too large to fit in memory.

To overcome this, Dask is used in this project. Dask is a parallel computing library in Python that can handle large datasets by breaking them into smaller chunks and processing them in parallel.

**Tools and Technologies**

1. Python – Programming language used for implementation
2. Dask – For distributed data processing and handling large datasets
3. Pandas – For additional data manipulation and conversions
4. Matplotlib/Seaborn – For data visualisation
5. CSV File – Dataset containing airline passenger and flight details

**Dataset Description**

The dataset contains passenger-level details for multiple flights. The major columns include:

Passenger ID, First Name, Last Name, Gender, Age

Nationality, Country Name, Airport Name, Continent

Departure Date, Arrival Airport, Pilot Name, Flight Status

This data helps in studying flight operations, passenger distribution, and flight outcomes.

**Methodology**

* **Data Loading**

The dataset is loaded using Dask’s read\_csv() function which allows reading large files efficiently without loading the entire file into memory.

* **Data Cleaning**

Missing values are removed from key columns like Age, Gender, Country Name, and Flight Status. Date columns are converted to proper datetime format.

* **Data Analysis**

Total number of passengers

Average age of passengers

Count of flight statuses (e.g., On Time, Delayed, Cancelled)

Flight distribution across continents

Top 5 countries with the highest number of passengers

* **Data Export**

The cleaned dataset is exported back to CSV for further use or visualization.

**Results**

The project successfully processes and analyzes the dataset without requiring heavy computing infrastructure. It provides meaningful statistics about the flights and passengers. Dask performs the computations efficiently even when the data is large.

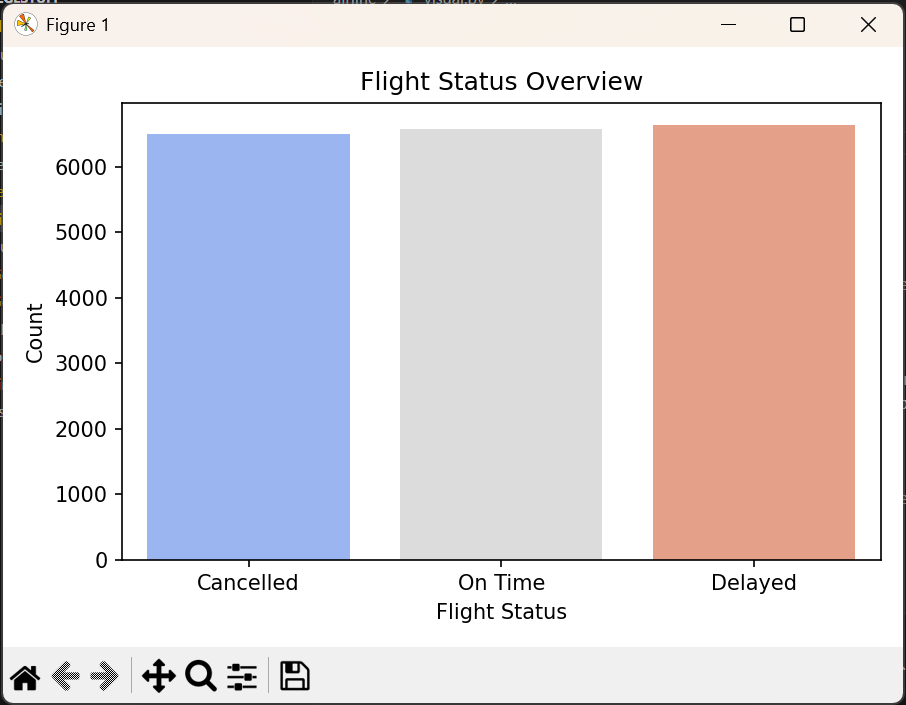
**Conclusion**

This project demonstrates how Dask can be used as an alternative to Spark for large-scale data processing. It provides a simple and efficient approach to perform big data analytics directly in Python without setting up a Java environment. Using Dask, data scientists can scale up their analysis from small datasets to large ones while keeping the same familiar Pandas-like syntax.

A screenshot of a computer

AI-generated content may be incorrect.





A screenshot of a graph

AI-generated content may be incorrect.

