The Impact of Covid-19 on Air Traffic:

Spatiotemporal/Time Series Forecasting and Benchmarking

Bo Yan, Yuan Hu, & Adelle Driker
(Group 6)

Team Introductions

Yuan Hu

- Budget Manager
- Data Engineer/Solution Architect

Bo Yan

- Record Keeper
- Software/ML/DL Engineer

Adelle Driker

- Project Coordinator/Manager
- Data/Business Analyst

Problem Definition

Many global industries have been affected by the COVID 19 pandemic, the airline industry being one of the most heavily hit

- E.g. London's Heathrow Airport reported a 97% decrease in passenger numbers between May 2019 and May 2020

Creates uncertainty for both passengers and airline companies, especially due to the multiple waves of virus mutations

- How should airlines plan future flights? When should passengers schedule their travels?

In other words, given a country's COVID situation, how should an airline/passengers plan ahead?

Definition of Success

Opportunity for Improvement

Individuals who interact with the airline industry should have a trustworthy and accurate forecast to help them make logistical decisions

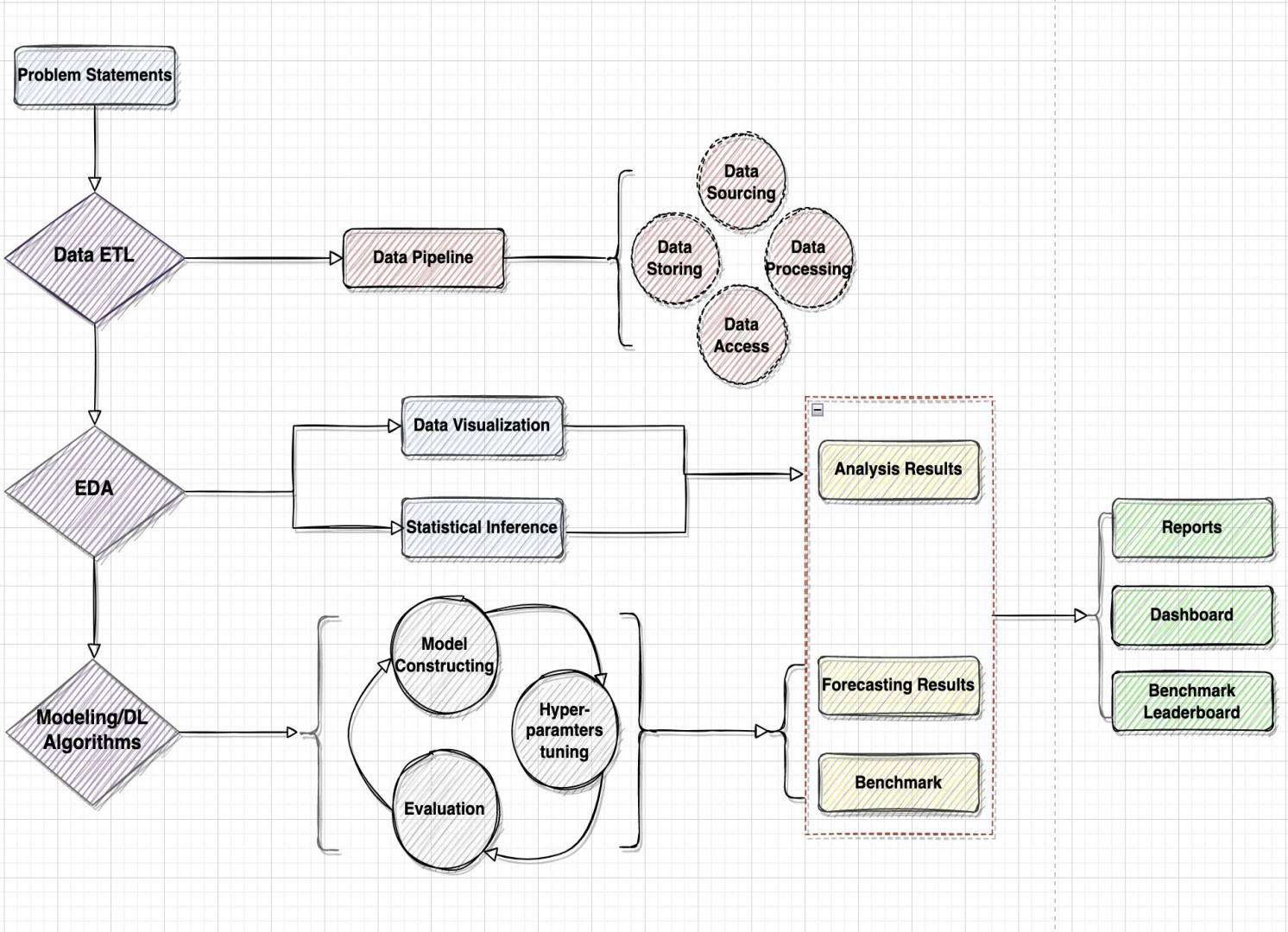
Success From a Stakeholder Perspective:

- Airline companies will have a way to schedule safer and more efficient flight patterns
- Travelers will have more confidence on when the best time to fly and would encounter fewer rebookings/cancellations

Success From a Technical Perspective:

- Create a maintainable and robust system
- Design and deploy a reliable model that employs sensible evaluation metrics

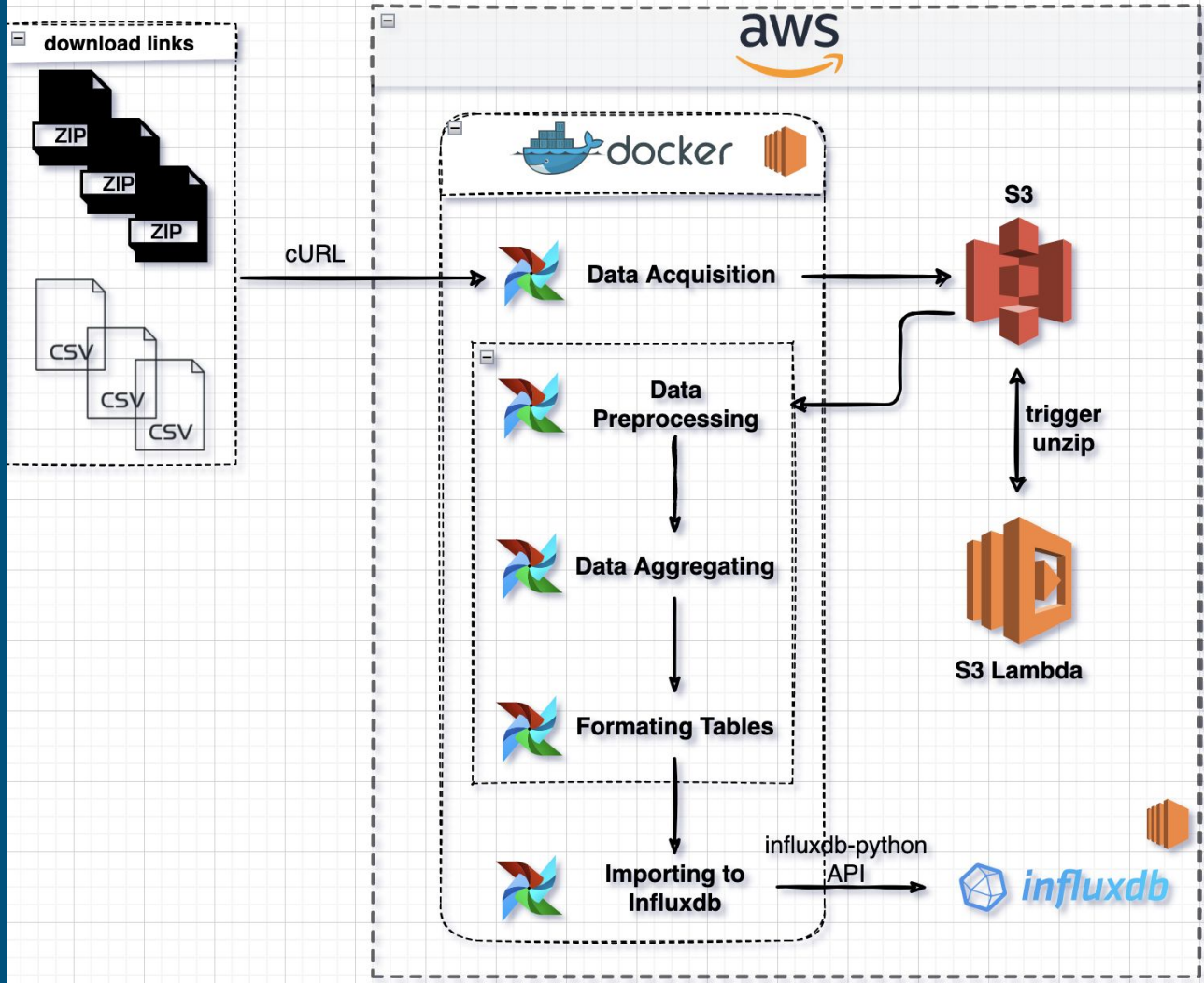
Proposed Solution /Approach - Project Workflow



Proposed Solution /Approach - Data ETL

Features:

- Automated by Airflow
- S3 and influxDB provides data access point
- AWS EC2 environment



Proposed Solution/Approach - EDA

- Data Assessment
- Statistical inference

  E.g. Hypothesis testing for comparing the number of recorded flights during two certain periods of time.

- Data Visualization

  E.g. Flights’ trajectory Networks, the number of flights trend analysis

Proposed Solution/Approach - Modeling

ML/DL Modeling for spatiotemporal/time series forecasting

- Baseline methods
  - Naive or persistence forecasting and averaging methods
  - Autoregressive forecasting methods, such as ARIMA and Seasonal ARIMA (SARIMA)
  - Exponential smoothing forecasting methods
- Deep learning methods
  - MLPs: the classical neural network architecture
  - CNNs: simple CNN models
  - LSTMs: simple LSTM models, Stacked LSTMs, Bidirectional LSTMs
  - Hybrids: hybrids of MLP, CNN and LSTM models

Data Sources

OpenSky Flight Data

- As-is, new files released on a monthly basis, multiple entries with missing data
- CallSign*, Number, ICAO24, Reg, TypeCode, Origin, Dest, First/Last Seen, Lat/Long/Alt of Origin & Dest

Johns Hopkins COVID19 Data

- Updates for historical inaccuracies, new files released daily
- Province/State, Country/Region**, Lat, Long, Dates (01/22/2020 - present)

Airline Code and Country Mapping

- Sourced from IATA and ICAO, mostly complete
- Airline Name, IATA Designator, 3-Digit Code, ICAO Designator*, Country**
- Will be used to link together the OpenSky Flight and COVID19 datasets

Assessment and Strategy

- Access points for retrieving data: Publicly accessible
- Machine Readability: Valid input format
- Data Quality: Dataset completeness/coverage, Amount-of-data, Relevancy, Trust dimensions
- Durability of the data-source: Check variety of reliable sources
- Identify suitable data, verify data update/refresh rates
- Incorporate mapping tables if datasets do not contain direct links to each other

Using the above criteria, acquire all available data

Feasibility and Resources Needed

Data Storage: AWS S3, influxDB

Processing Units: AWS EC2, AWS Sagemaker

Automation: Apache Airflow, Docker

Open Source: Python, Pandas, Pytorch, Scikit-learn, Torch-TS

Visualization: Matplotlib, Plotly, Grafana

Risks and Mitigation

Risk	Mitigation
Data pipeline may be fragile and easy to break if AWS has an outage	Build redundant data storage to ensure continuity during the next outage
Potential performance issues	Train DL models with large amount of valuable data and add more GPUs
Long runtime for advanced models	Plan the tasks ahead, utilize/distribute computational resource reasonably
Security/Stability - open source frameworks may pose unstable risks	Maintain version control, simplify process, choose popular frameworks, seek help from open source community
Data ingestion format	Create additional checks to ensure correct structure and format which will be accepted by pre-processing stage