# Distributed Data Systems - Group Project

MAHESH CHAUDHARI, PH.D

# Group Project

# Your Team Your Project Your Success

Five to Six people per team.

Canvas to create the groups automatically.

Everyone needs to contribute and work!

Peer-feedback is important.

# Group Project Objective

Build an automated data pipeline

Use Airflow to connect different systems like MongoDB, Spark Cluster, GS storage

Understand Map-Reduce or aggregation pipeline in MongoDB

Do data exchange between MongoDB and Spark

*** Use your data science skills to implement a machine learning model of your choice





***

Apache
Airflow

# Group Project Goals

Gather a suitable dataset (preferred if there are more than 1 dataset for more realistic data processing scenario).

More use cases for data aggregation and building ML models.

Consider your final goal:

- Make sure that the goal is reachable within the 7 weeks.
- Detailed project report (template soon available on canvas)

# Dataset selection

Some Data Sources

- Your own data (the best) : e.g. Apple SensorLog, etc.
- data.gov: https://www.data.gov/
- Twitter API: https://developer.twitter.com/en/docs/twitter-api
- Meetup API: https://www.meetup.com/api/guide/
- Football Data API: https://www.football-data.org/
- US Patent and Trademark Office API: https://www.uspto.gov/learning-and-resources/open-data-and-mobility
- Many more!! - https://github.com/public-apis/public-apis

# Things to consider

- Data fusion from multiple data sources
- Conceptualize what data processing/aggregations/new analytics you want to define
- Develop and/or Apply novel data processing / ML algorithms
- Compare results of different ML algorithms
- Compare results from different machine specs- Costs/Speed/DataSize

# Task - 1 Datasets + Project Plan

- Finalize the dataset/s
- Write up the summary of what analytics or what machine learning model to build
- Timeline - As a team, come up with goals for each week and meeting schedule. Try to meet those goals for each week
- Keep that final goal in mind
- Discuss these goals with the team and everyone agrees on the details

## Due: February 7, 2026 Midnight

# Task - 2 Airflow and MongoDB

- Load data into GCS (Google Cloud Storage)
- Import the data into MongoDB in collection
- Create at least 1-2 new datasets from the original datasets based on some analytics, aggregation
- Store the aggregates in a separate collection in MongoDB on MongoDB Atlas
- Query the data in MongoDB (both original and new aggregate data) using MongoDB queries.

## Due: February 21, 2026 Midnight

# Task -3 Airflow MongoDB + SparkSQL

- Create Dataframes from the data from MongoDB Atlas
- Run SparkSQL queries over the data frames
- Build machine learning algorithms to derive analytics on top of the original and aggregated datasets

## Due: March 14, 2026 Midnight

# Submission (Blog Post/short paper style)

- Your dataset/s and goals
- Rational behind the goals and the choice of the dataset/s
- Overview of your data engineering pipeline
- Preprocessing goals, algorithms and time efficiency
- Explain the different analytics and queries you ran over the datasets in MongoDB and on Dataframes in Spark
- Compare and contrast the time of the queries over both the version of the datasets
- Lessons learned
- Conclusion

# Questions?