MSDS 697
Group Project Overview

DIANE WOODBRIDGE, PH.D

Group Project

Your Team

After reviewing the problem proposals, join a team to work for this module.

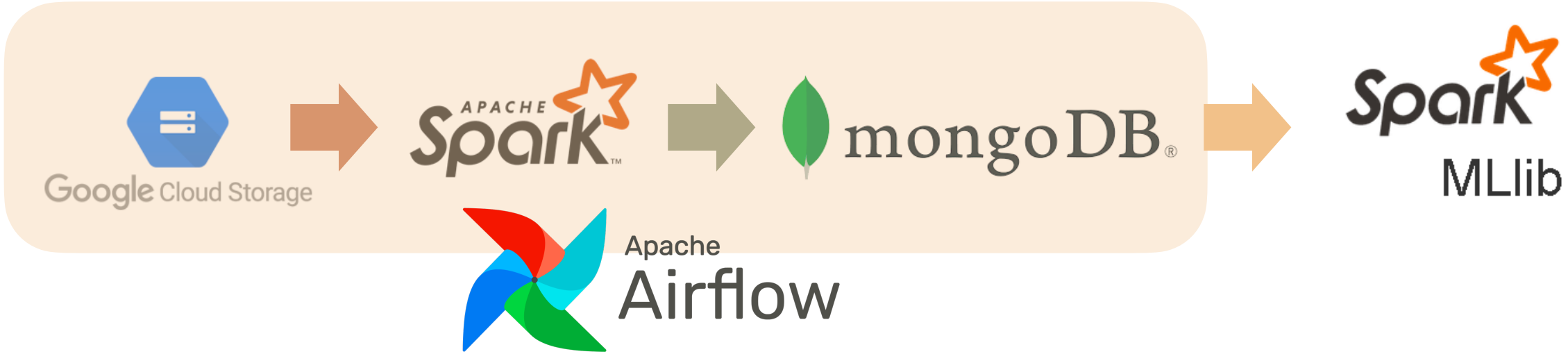
Three to Five people per team.

Everyone needs to work !!

- We are going to survey and the final group project grade will be weighed accordingly.

Overview

Goal : Building an Automated Scalable ML Pipeline



Apache Spark & MongoDB

Import data from GCS into MongoDB Atlas (Should have replicated shards).

- Make sure to automate the following processes using Apache Airflow.
  - Fetching data using API (or crawling) and store to GCS
  - Creating an aggregate using data in GCS and storing to MongoDB

Query data.

Machine Learning

Create Spark RDD/DataFrame on Databricks by importing the data from MongoDB.

Develop a ML model using algorithms in Spark ML on Databricks.

Task Summary

More details will be added to Canvas Assignments

Task 0 - Project Proposal (Optional)

You would need at least two data sources to create data aggregate and build ML models.

Consider what your final goal is.

- Level 1 : Able to pass the class (Minimum work)
- Level 2 : Able to add new techniques and algorithms on your resume with an interesting story
- Level 3 : Publish articles at the MongoDB Student Spotlight, Present at DSCO
- Level 4 : Publish peer-reviewed conference/journal papers

Consider that it is a reachable goal within the 7 weeks.

Note : I will consider topics proposed on Task 0 only to create a group.

Task 0 - Project Proposal (Optional)

Some Data Sources

- Your own data (the best) : ex) 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

Task 0 - Project Proposal (Optional)

Things to Consider

- Data fusion from multiple data sources
- Develop and/or Apply novel data processing / ML algorithms
- Compare results of different ML algorithms
- Compare results from different machine specs- Costs/Speed

Task 1 - Join a Group

Submit a 1-page summary including

- Data sources (URL) - At least one should be an API call.
- ML Objectives - # of objectives should be same as the number of team members.
- Timeline - For each week, what are the goals to achieve and meeting schedule.
- Your final goal for this course is : Level 1 - 4 (See the previous slides).

Task 2 - MongoDB

1. Load data to GCS.
2. Import data from GCS to Apache Spark for data preprocessing to create data aggregates.
    1) Indicate your preprocessing algorithms and time efficiency (seconds to run) - with the cluster specification.
3. Store the aggregates in MongoDB on MongoDB Atlas.
4. Query data (Submit code and screenshot).

Task 3 - Machine Learning

1. Create Spark Dataframe by importing the data from MongoDB Atlas.
2. Apply various machine learning algorithms via Spark ML and other distributed ML frameworks to achieve the proposed goal.
3. Submit codes and reports (Blog post/short paper style).
    1) Your dataset and analytic goals
    2) Overview of data engineering pipeline.
    3) Preprocessing goals, algorithms and time efficiency (seconds to run) - with the cluster specification.
    4) Machine Learning Outcome Comparison on the same cluster specs including the number of instances, same machine types, disk, and memory sizes (One model per person).
    5) Should indicate each member's cluster setting and execution time/model (F1 score, R^2, etc) comparison
    Ex. 1) 1,2,3,4,5 node cluster 2) 3 node clusters with different specs (CPU, Memory, Disk, etc.)
    6) Lesson Learned
    7) Conclusion

Questions?