Flight Price Prediction System

Spring 2022 CSYE 7200

Team 1

Project Description

- Project: Flight Price Prediction System

  A system used to predict the price trend of a flight

- Team Members:
  - Yuhan Yang 1094267
  - Luo Chen 1564677
  - Kang Shentu 1569432

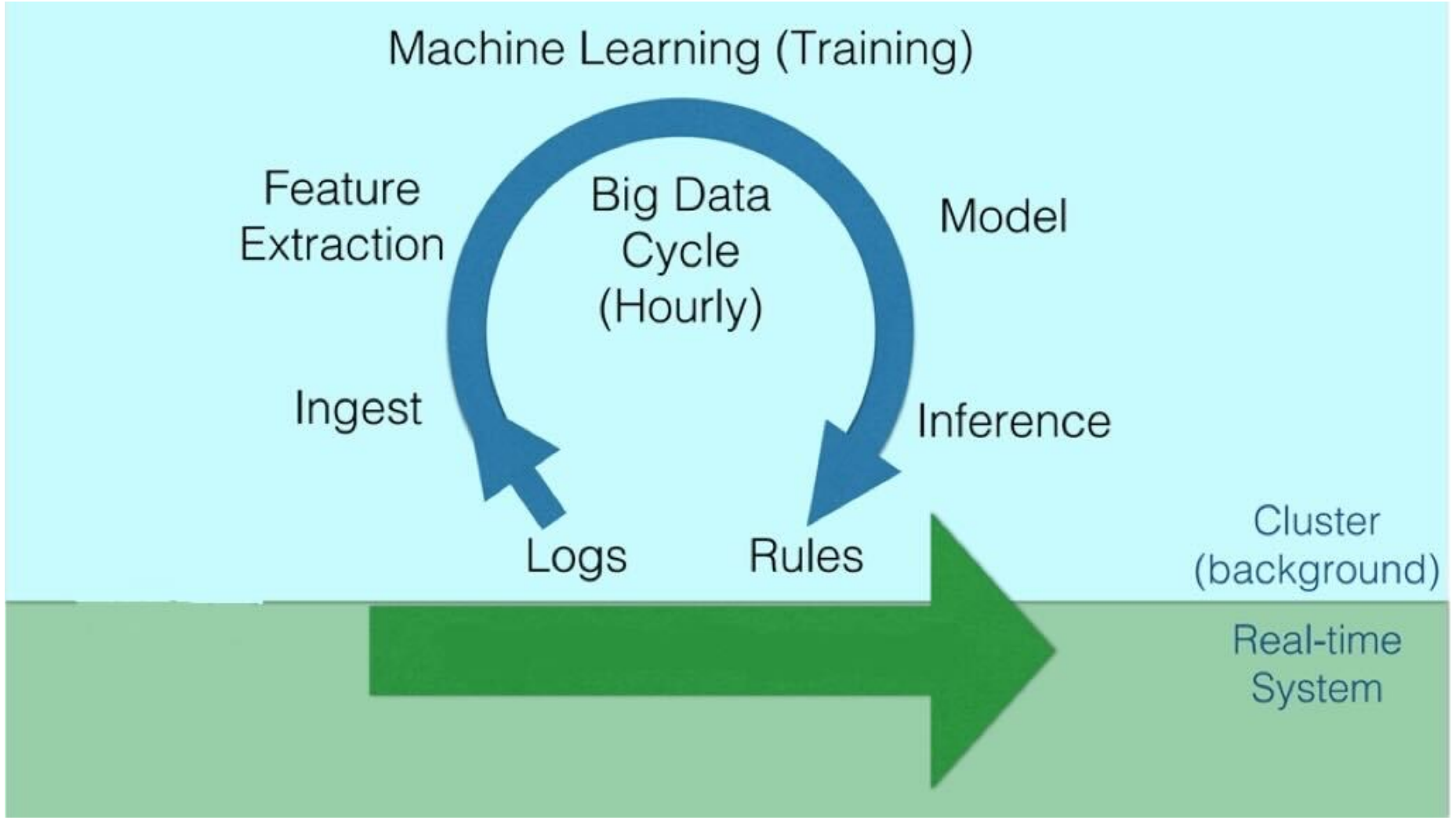
Project Description

- Project: Flight Price Prediction System

  A system used to predict the price trend of a flight

- Input
  - Airline, Days before departure, Departure/Arrival time, Source/Destination city, Economy/Business class
- Output
  - Predicted Prices

Methodology



Methodology

- Ingest & Feature Extraction
  - Web Crawler
    - Data Crawling
  - Spark
    - Stream Processing
    - One-Hot Encoder
  - TableParser / Spark native read

Methodology

- Machine Learning
  - XGBoost Regression Model

Methodology

- Service
  - Play Framework
    - Online prediction API
    - In Batch prediction API
    - Reactive Streaming prediction API

Data Source

Original Dataset:

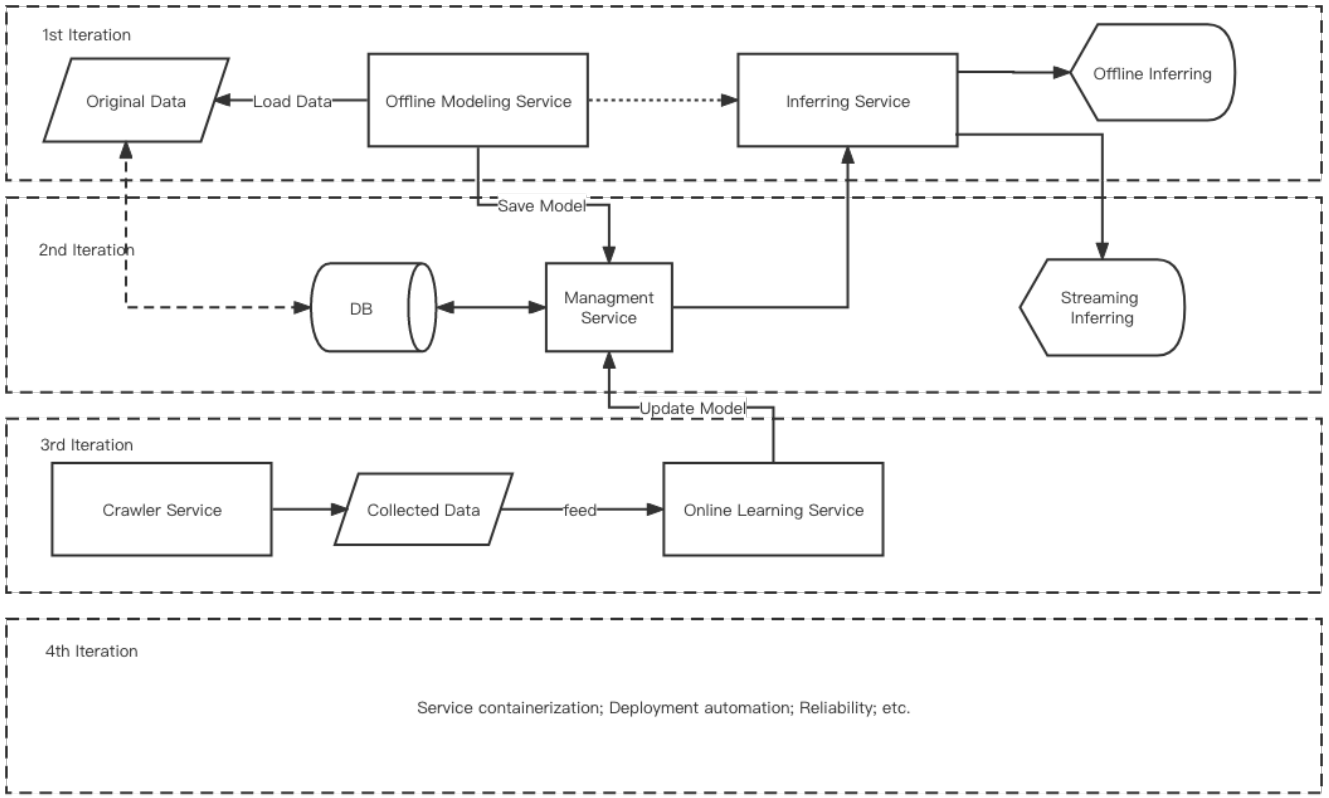
Kaggle: https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

Dataset contains information about flight booking options from the website EaseMyTrip for flight travel between India's top 6 metro cities.
There are 300261 datapoints and 11 features in the cleaned dataset.

Following Data:

Crawl from EaseMyTrip
90k per day

Milestones



1st Week: Implement the basic system to perform offline training and batch predicting.
- **Yes, we did it**

2nd Week: Add the service to manage data, models and predictions. Implement streaming processing for inferring service.
- Finished these functions independently without integration

3rd Week: Complete crawler service. Update offline learning To online learning. Implement the workflow for the whole system.
- Implemented web crawler and streaming API.

4th Week: Optional work. Strengthen reliability of our system.
- Joint Debugging and deployment

Repository

- Web Service repo link:

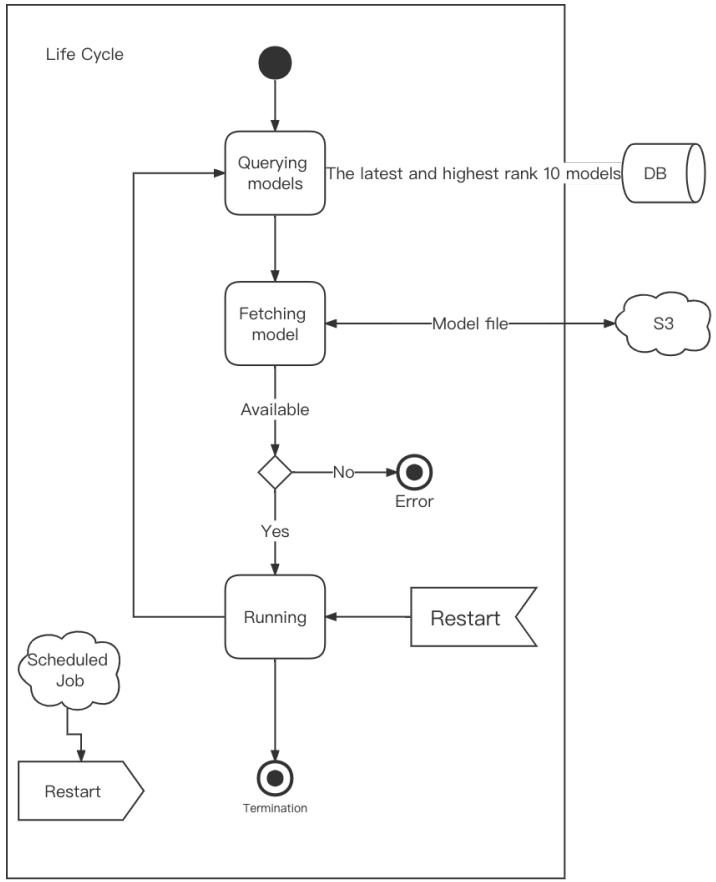
  https://github.com/ScalaTeam1/web

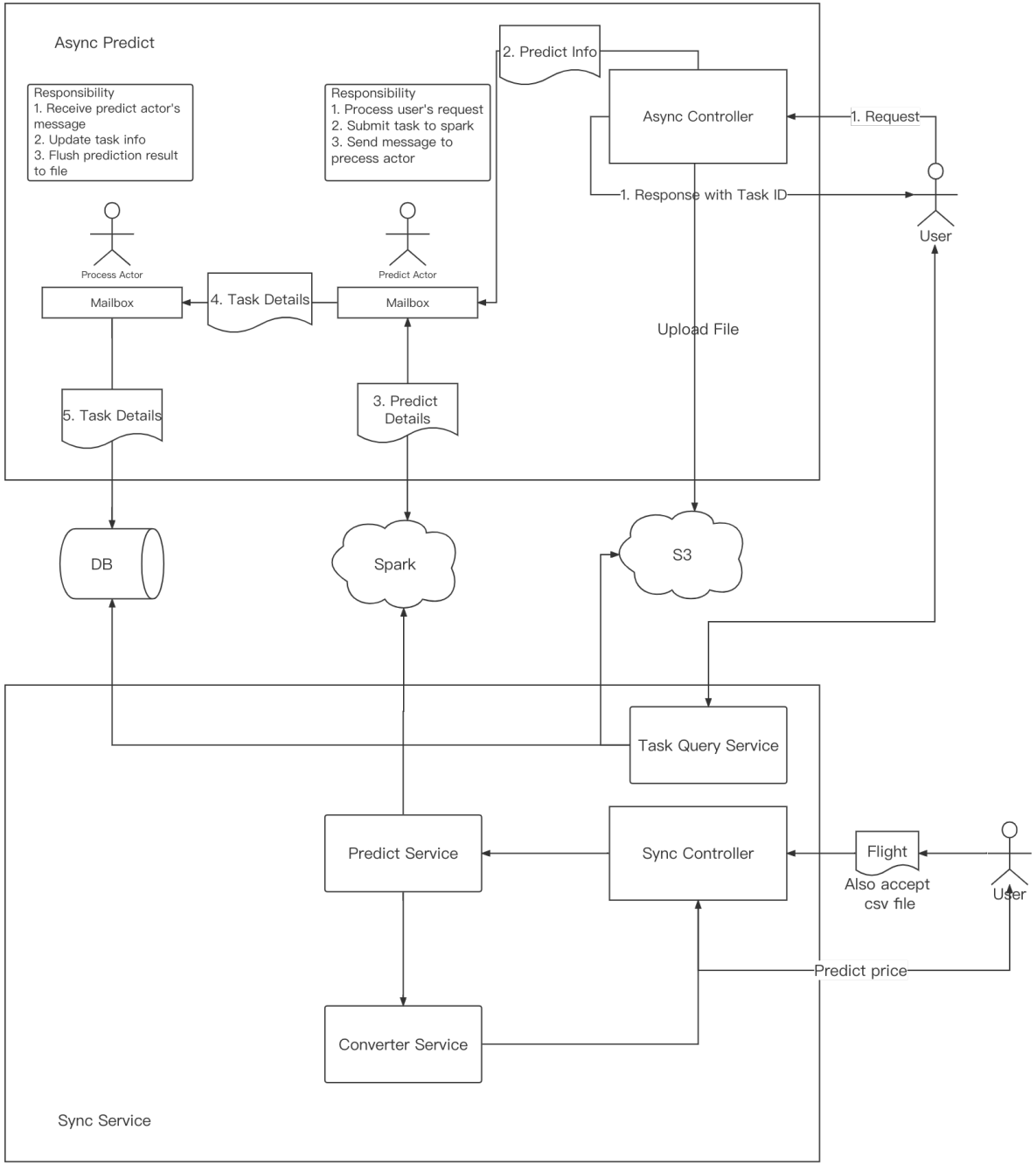
- Web Crawler repo link:
  https://github.com/ScalaTeam1/FlgihtPriceCrawler

- Predictor and Trainer repo link:

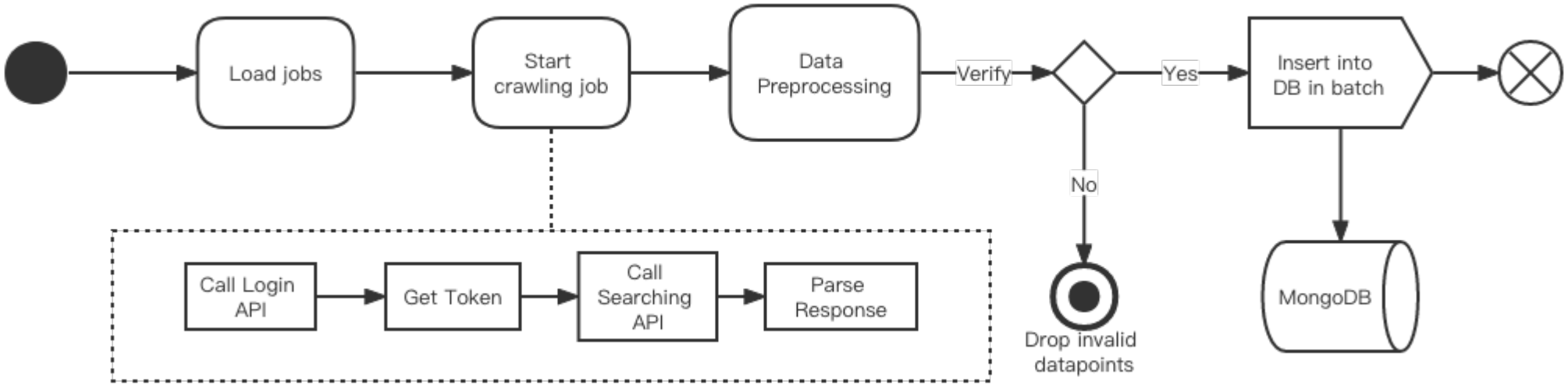
  https://github.com/ScalaTeam1/flight-prices-prediction-xgboost

Web Service


Life Cycle
Querying models
The latest and highest rank 10 models
DB
Fetching model
Model file
S3
Available
No
Error
Yes
Running
Restart
Scheduled Job
Restart
Termination

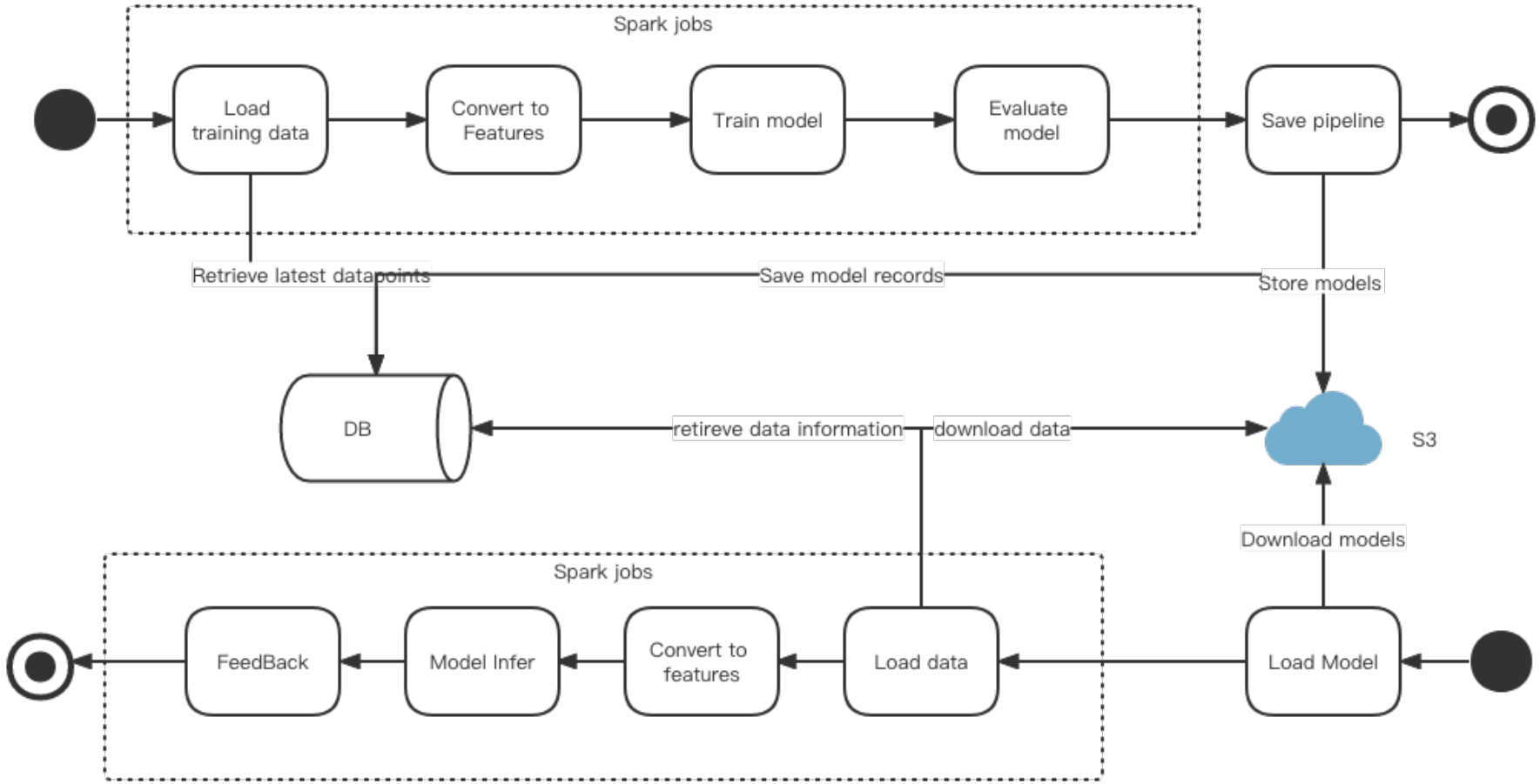

Async Predict
2. Predict Info
Responsibility
1. Receive predict actor's message
2. Update task info
3. Flush prediction result to file
Responsibility
1. Process user's request
2. Submit task to spark
3. Send message to precess actor
Async Controller
1. Request
1. Response with Task ID
User
Process Actor
Predict Actor
Mailbox
4. Task Details
Mailbox
Upload File
5. Task Details
3. Predict Details
DB
Spark
S3
Task Query Service
Predict Service
Sync Controller
Flight
Also accept csv file
User
Predict price
Converter Service
Sync Service

Web Crawler



3k datapoints / min

Trainer & Predictor



Acceptance Criteria

- The response time of the API for prediction for one input is less than 1s
  500ms for single prediction (1.9k per second for in batch prediction, AWS t2.medium)
- Training time of static model (offline training) should less than 1 hour
  About 3 mins give enough memory
- Updating model by new data retrieving from web-crawler every 2 hours
  Cron Job
- The R2 score for the regression model should be higher than 0.7
  About 0.75

Goals of the project

- What we still need to do
- Ensure Web Service availability when restarting the service
- External Spark cluster
- Generalization performance of the model
- Integrate three projects into one multi-module project
- Deploy elegantly

Thank You!

Spring 2022 CSYE 7200

Team 1