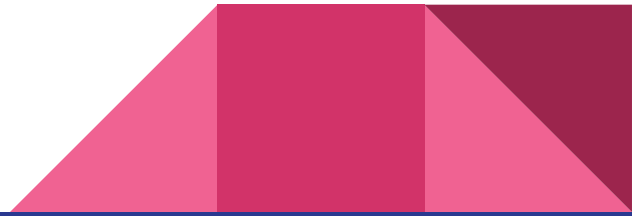
Data Analytics Pipeline for Historical Taxi Data

Anirudha Tambolkar (atambol)
Parth Nagori (pnagori)

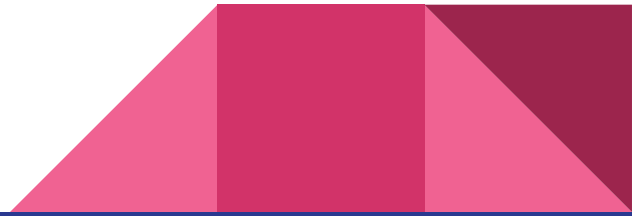
Overview

- Historical data of NYC-TLC
- Big data - volume
- Predicting taxi fare and journey time
- Scalability and fault tolerance



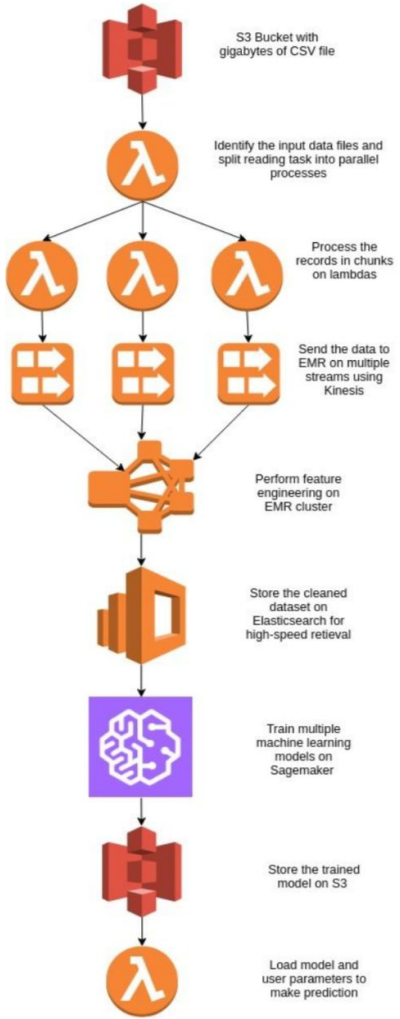
Tools and technologies

- Amazon Web Services
  - Elastic map reduce - EMR
    - Spark
  - Sagemaker
  - S3
- Jupyter Notebooks
- Python Flask

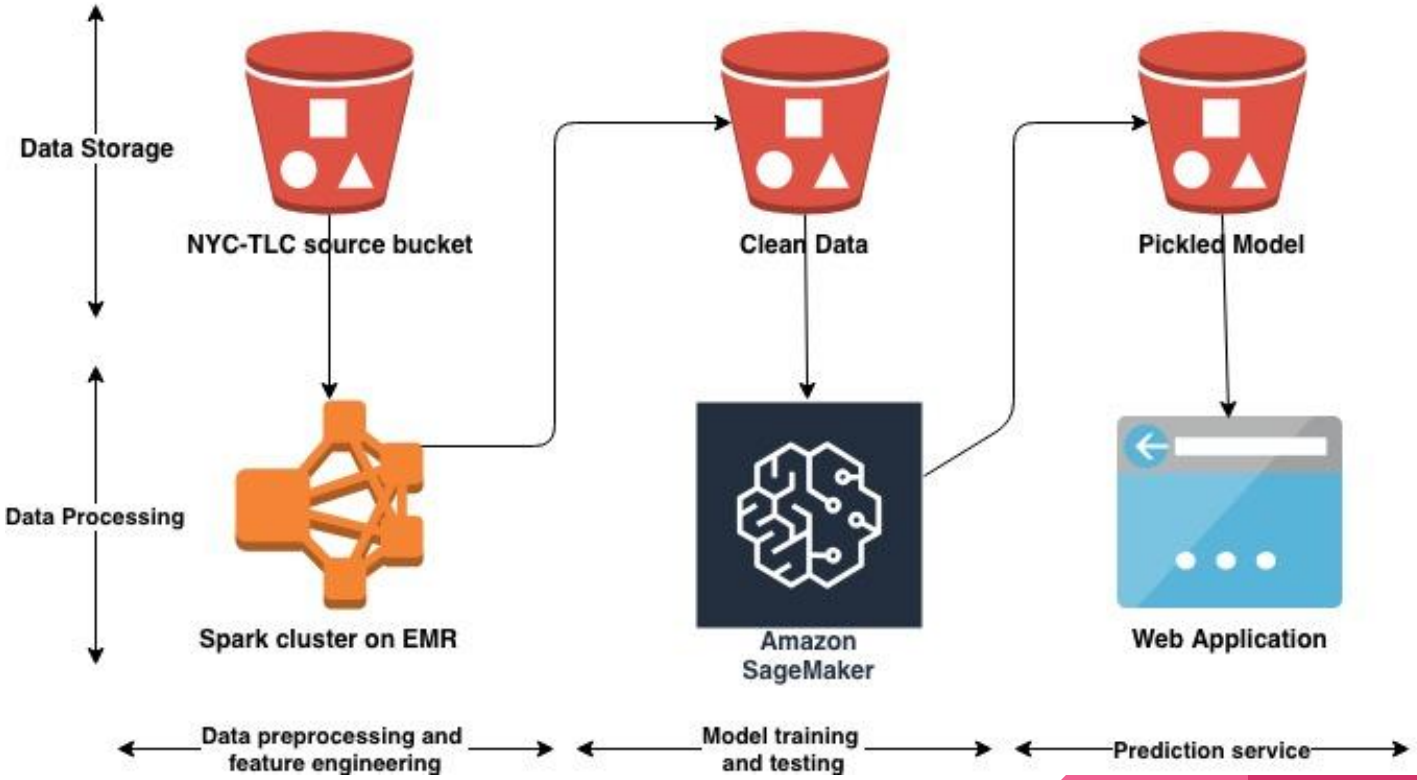


Approach

- Initial approach
  - Lambda architecture
- Why we moved away?
  - Over engineered solution
  - AWS is inherently well integrated
  - Complexity
  - Cost
- Final solution
  - Dropped speed layer
  - Batch processing and serving layer

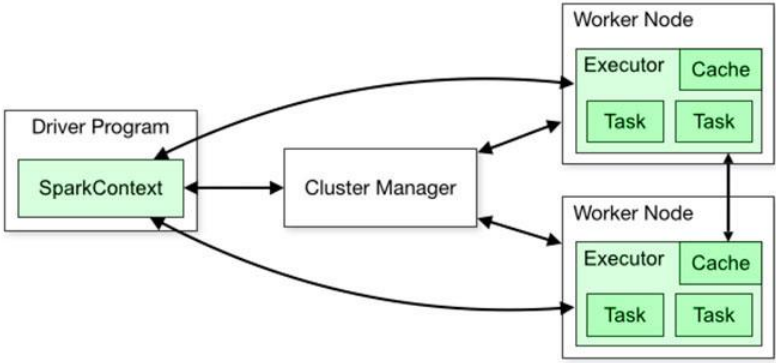


Architecture



Spark

- 3 node cluster
- Two approaches to parallelization
  - Split the file
    - Serial code was a bottleneck
    - Lots of Disk IO
  - Pandas is in-memory
    - Max input file size is 2 GB
    - Each worker instance has 16 GB RAM and 8 cores

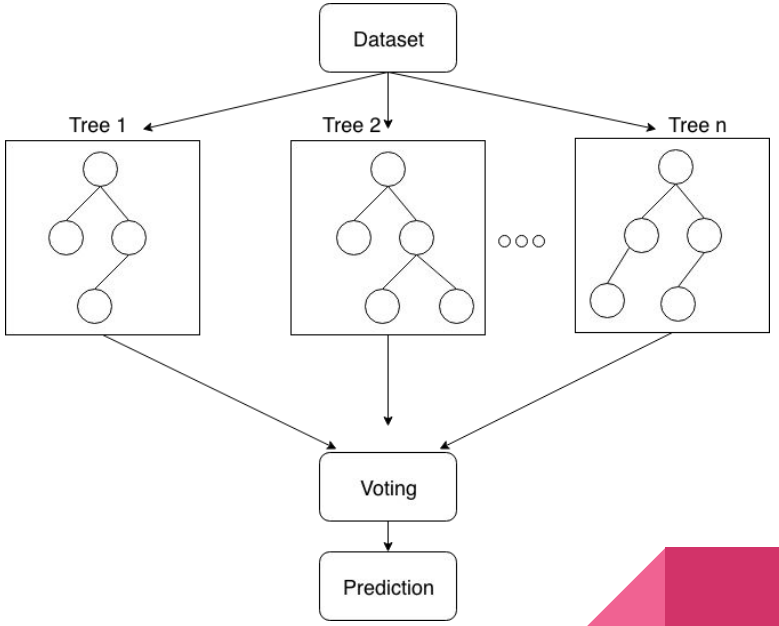


Data Preprocessing

- Cleaning the data
  - Attribute selection
  - Data validation
    - Invalid fare
    - Invalid date
    - Missing values
- Feature Engineering
  - Zipcode
  - Trip Frequency

Data modelling and machine learning

- Sagemaker
  - Random Forest
  - K-mean clustering
  - XgBoost



Performance

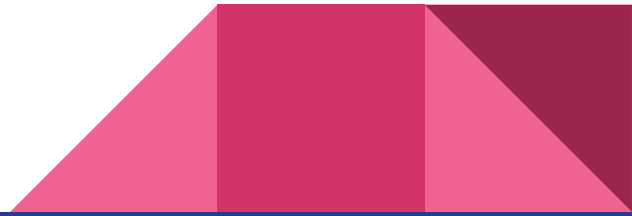
- Data preprocessing
  - Serial processing time ~46 hours
  - Parallel processing
    - 2 * m4.xlarge instance
    - 16 cores
    - duration ~ 3 hours
  - Linear scaleup (16 * 3 ~= 46)
- Data modelling
  - Used inbuilt parallelisation provided by ML algorithms
  - m4.2xlarge instance - 32 GB RAM

Challenges

- Design choices
  - Lambda or batch only processing
  - Chunking file or not for parallelism
- Data preprocessing
  - Selecting attributes
  - Storage cost
- Integrating various components

Future Scope

- Stream processing instead of batch processing
  - Use Kafka
- Automating the manual steps



Conclusion

- Built a data analytics pipeline
- Batch processing of over 100 million records
- Trained machine learning models
- Built a web application to query
- All in 100 bucks!

Questions?