INTERIM DELIVERY

- TIME TURNERS -

KHYATI PAREKH
NIHARIKA SHARMA
RAJIV VEERARAGHAVAN

1.

MOTIVATION

Our goal is to predict the future behavior of customers from a given set of multivariate time series.

Being able to predict a user’s propensity to churn out of a subscription, and the timing of the churn event, makes marketing more efficient.

We also saw this project as an opportunity to learn more about working with time series data and cloud infrastructure as well.

2.

PROPOSED WORK

Our objective is, among other things, to perform a multivariate time series analysis on the data to find the probability of a user churn out.

We aim to give Amplero a fair idea about what the probability of churn of a user is and when they are most likely to churn out.

Using Machine Learning, we would like to predict user churn among prepaid users given a multivariate time series.

3.

WORK WITH THE DATA

DATA

State Recharge Time series

- CarrierReportedSubscriptionStateDeltaTimeSeries
- VoiceCallsPerDayTimeSeries RechargeTimeSeries

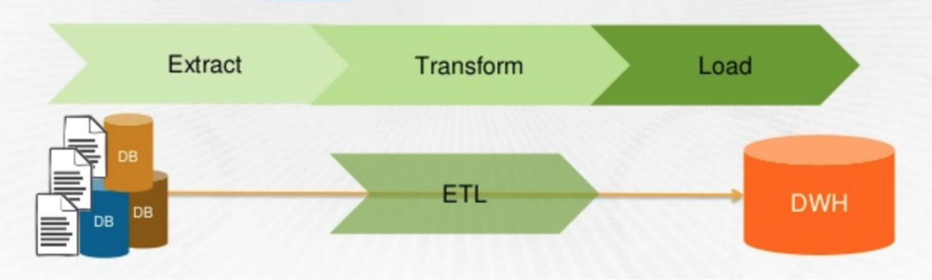
Usage Time series

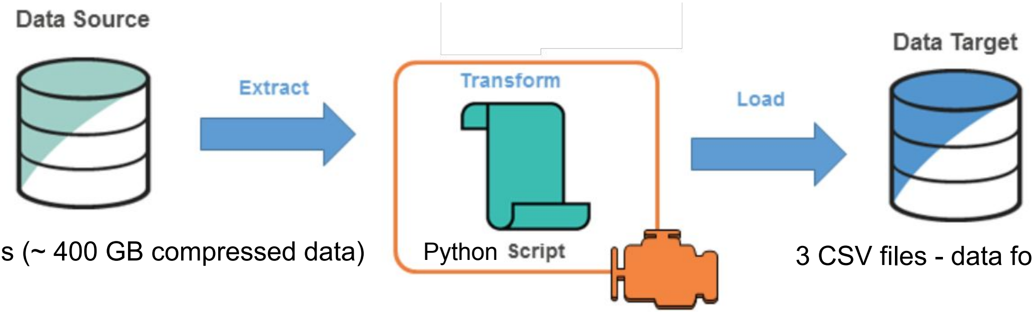
- CompactDataKBPerDayTimeSeries
- CompactSMSPerDayTimeSeries

Social Time series

- OutboundVoiceCountNetworkPageRankLast7DaysTimeSeries
- OutboundVoiceCountNetworkPageRankQuantileLast7DaysTimeSeries
- OutboundSMSNetworkPageRankLast7DaysTimeSeries
- OutboundSMSNetworkPageRankQuantileLast7DaysTimeSeries

ETL PROCESS





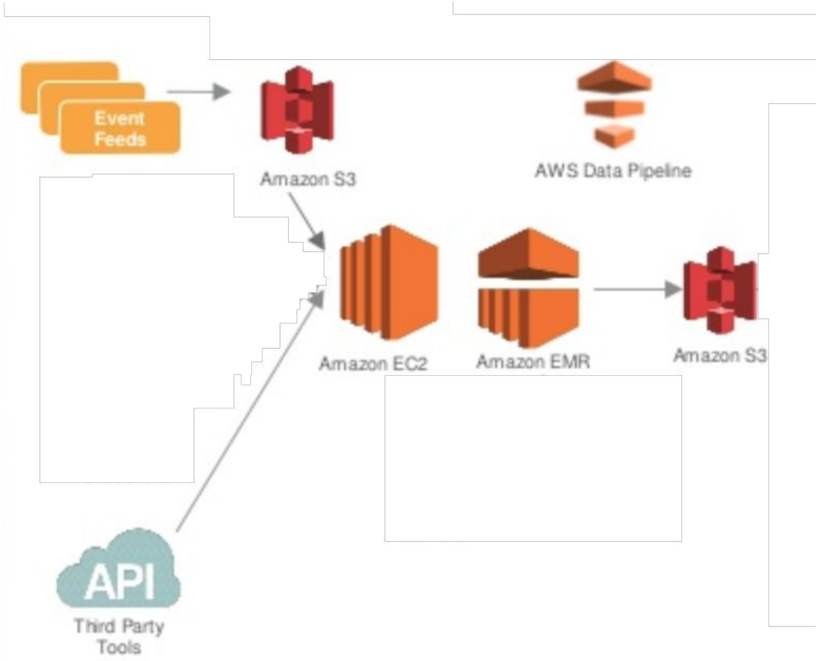
S3 51 JSON files (~ 400 GB compressed data)

3 CSV files - data for ~60000 entity ids

AWS ARCHITECTURE

INFRASTRUCTURE REQUIREMENT

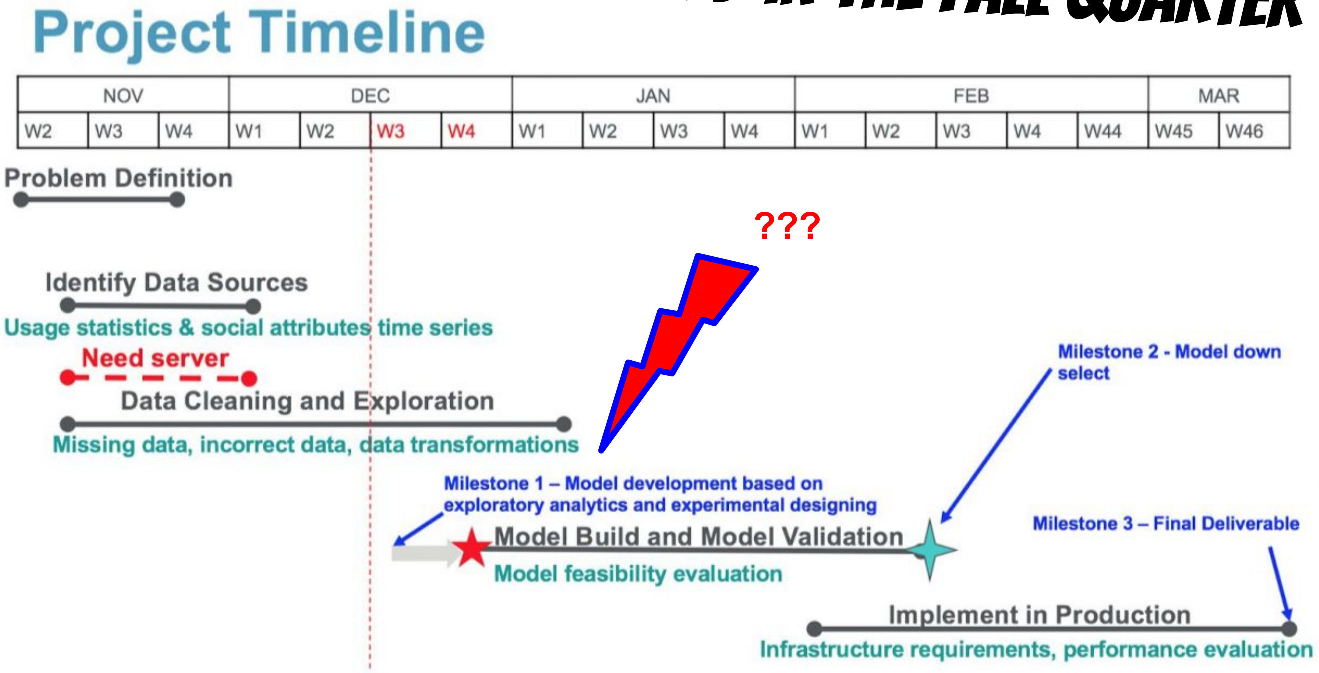
- AWS EC2 Instances– m3.2xlarge
- S3 Storage
- Storage: Amazon EBS Volumes
- Amazon EMR



4.

SCHEDULE REVIEW

ORIGINAL SCHEDULE PROPOSED IN THE FALL QUARTER



Project Timeline

NOV			DEC				JAN				FEB					MAR	
W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W44	W45	W46

Problem Definition

???

Identify Data Sources

Usage statistics & social attributes time series

Need server

Data Cleaning and Exploration

Missing data, incorrect data, data transformations

Milestone 1 – Model development based on exploratory analytics and experimental designing

Model Build and Model Validation

Model feasibility evaluation

Milestone 2 - Model down select

Milestone 3 – Final Deliverable

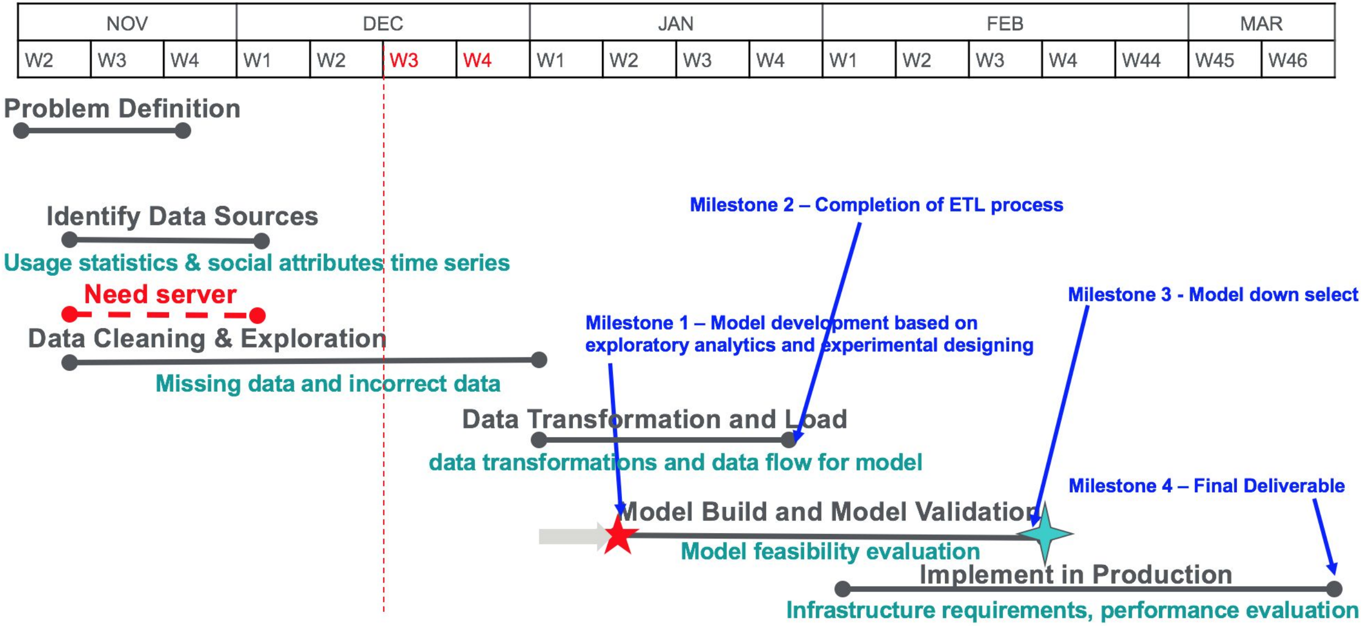
Implement in Production

Infrastructure requirements, performance evaluation

BIG PICTURE

- We are on track and developing the model currently
- Underestimated the data-set
- Didn’t consider the ETL process while devising the project roadmap/ timeline

REVISED SCHEDULE



NOV			DEC				JAN				FEB					MAR	
W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W44	W45	W46

Problem Definition

Identify Data Sources

Usage statistics & social attributes time series

Need server

Data Cleaning & Exploration

Missing data and incorrect data

Data Transformation and Load

data transformations and data flow for model

Model Build and Model Validation

Model feasibility evaluation

Implement in Production

Infrastructure requirements, performance evaluation

Milestone 1 – Model development based on exploratory analytics and experimental designing

Milestone 2 – Completion of ETL process

Milestone 3 - Model down select

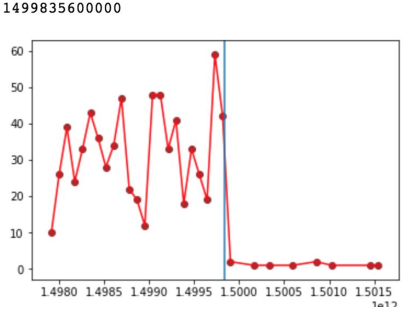
Milestone 4 – Final Deliverable

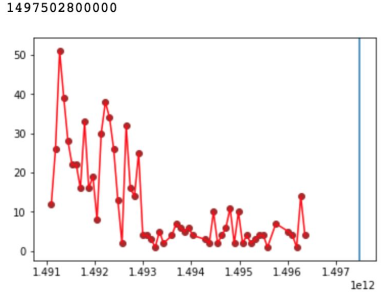
5.

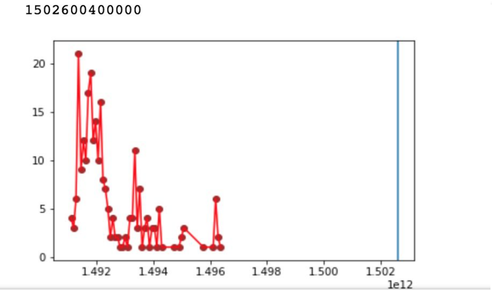
MODEL TRAINING AND DEVELOPMENT

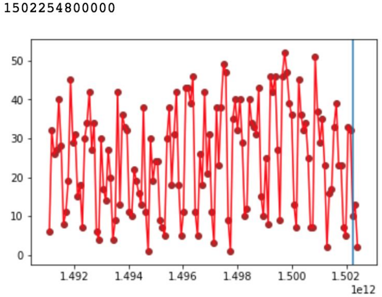
GET A FEEL OF THE DATA, HIGHLIGHT SOME ISSUES

VOICE CALLS TIME SERIES & DELTA TIME SERIES DATA









MODEL DEVELOPMENT

- **Sampling**
  - Data is super biased – 95% non churn to 5% churn
  - Removed the class bias by using 50-50 split
- **Timeseries Considered**
  - VoiceCallTimeseries
  - SMSTimeSeries
  - DataTimeSeries
  - RechargeTimeSereis
- **Features Considered for each timeseries**
  - Mean, Variance
  - Min, Max

MODEL TRAINING

- **Logistic Regression**
  - Poor Performance – 60% accuracy
- **SVM**
  - Relatively better performance ~ 60% accuracy
- **SVM with RBF kernel**
  - Excellent performance ~ 99.92% accuracy

MODEL TRAINING

- **Logistic Regression**
  - Poor Performance – 60% accuracy
- **SVM**
  - Relatively better performance ~ 60% accuracy
- **SVM with RBF kernel**
  - Excellent performance ~ 99.92% accuracy

REALITY CHECK – Overfitting!!

MODEL EVALUATION

- Currently we are considering accuracy – Not the best!
  - Update to Precision and Recall (ROC/AUC) so that we can them weigh them differently
    - It might be okay to have low precision but we definitely want the recall to be high

6.

CHALLENGES

MODEL RELATED CHALLENGES

- Feature extraction seems to be the most complicated part
  - Need to extract the most relevant features from the time series to predict churn
- We are thinking of doing the following to improve model performance
  - Using ensemble techniques
  - Explore ARIMA models (which consider autocorrelation and lag)
  - Consider more time series
  - Extract more advanced features from time series
    - Time window features
    - Longest sequence of consecutive activity

DATA & CLOUD RELATED CHALLENGES

- Large Compressed files
  - Read line by line (parallel process also time consuming),
  - Can't store all the data in one variable,
  - Cannot store Time Series data to a dataframe because of memory issues.
- Not enough/ satisfactory GZip library documentation
- AWS account hacked – charged ~$50,000

THANKS!

Any questions?