Architecture Decisions for Big & Tiny Data.
We are

UnnatiData Labs

www.unnati.xyz

@raghothams @nischalhp
What do we do?

Build Data Science Platforms
What do we do?

Educate
+

= Build Data Science Platforms
What do we do?

Educate
+
Frame the problem
+

= Build Data Science Platforms
What do we do?

Educate
+
Frame the problem
+
Technology
=
Build Data Science Platforms
3 Stories

#Fintech  #UnstructuredHealthcare  #CampaignManagement
3 Criteria

V.V.V Talent Pool Cost
#Fintech
Tiny data | Early Startup | Data Driven
What are we solving?
What are we solving?

Creditworthiness of Students
Challenges?
Challenges?

No data
Challenges?

No data

Credit history of Students is scarce
Challenges?

- No data
- Credit history of Students is scarce
- Using Alternate Data Sources
Architecture

ETL Layer
- mongoDB
- PostgreSQL

Analytics Layer
- pandas
- Python

API Layer
- Flask
- NGINX
Criteria

V.V.V

100K records in the first 6 months

60K Students in the first year
Criteria

Talent Pool
3 member tech team
MEAN stack shop
Criteria

Cost

Early Stage Startup
Tiny data problems are tricky
Tiny data problems are tricky

Get hold of the low hanging fruits
Tiny data problems are tricky
Get hold of the low hanging fruits
Beware of Data Sanity with NoSQL
Tiny data problems are tricky
Get hold of the low hanging fruits
Beware of Data Sanity with NoSQL
Embrace Data Science Early
Campaign Management
Medium data | Established Startup
What are we solving?
What are we solving?

Predict behavior of Users
What are we solving?

- Predict behavior of Users
- Educate the team on Data Science
What are we solving?

Predict behavior of Users
Educate the team on Data Science
Optimize campaign costs
Challenges?
Challenges?

No analytics DW in place
Challenges?

No analytics DW in place

Time to market
Criteria

V.V.V

50+ million rows

Exponential growth
Criteria

Talent Pool
5 member tech team
Python+Django Shop
Criteria

Cost

Bootstrapped
Established Startup
Learnings
Learnings

Know your databases well
Know your databases well
Data pipelines break
Learnings

Know your databases well
Data pipelines break
Exception handling - Logging - Notifications
Learnings

Know your databases well
Data pipelines break
Exception handling - Logging - Notifications
Monitor performance of your libraries
#Unstructured Healthcare

Big data | Generic Platform
What are we solving?
What are we solving?

Analytics on healthcare spend
What are we solving?

Analytics on healthcare spend
Medical claims - Many Data Providers - no standard
Challenges?
Challenges?

Ever changing schema
Challenges?

Ever changing schema

Large number of ETL jobs
Adding more servers is easy, writing good code is hard.
Criteria

V.V.V.V

150+ million rows

Every source has their own schema
Criteria

Talent Pool

SQL + Tableau team
Criteria

Cost
Limited budget
Learnings

Data security is very important
Learnings

Data security is very important
Parallelize ingestion
Learnings

Data security is very important
Parallelize ingestion
Data redundancy is not an issue
Learnings

Data security is very important
Parallelize ingestion
Data redundancy is not an issue
Polyglot of services
Tools

Munin
Server-Monitoring

#!/bin/bash

elasticsearch
logstash
Kibana

git