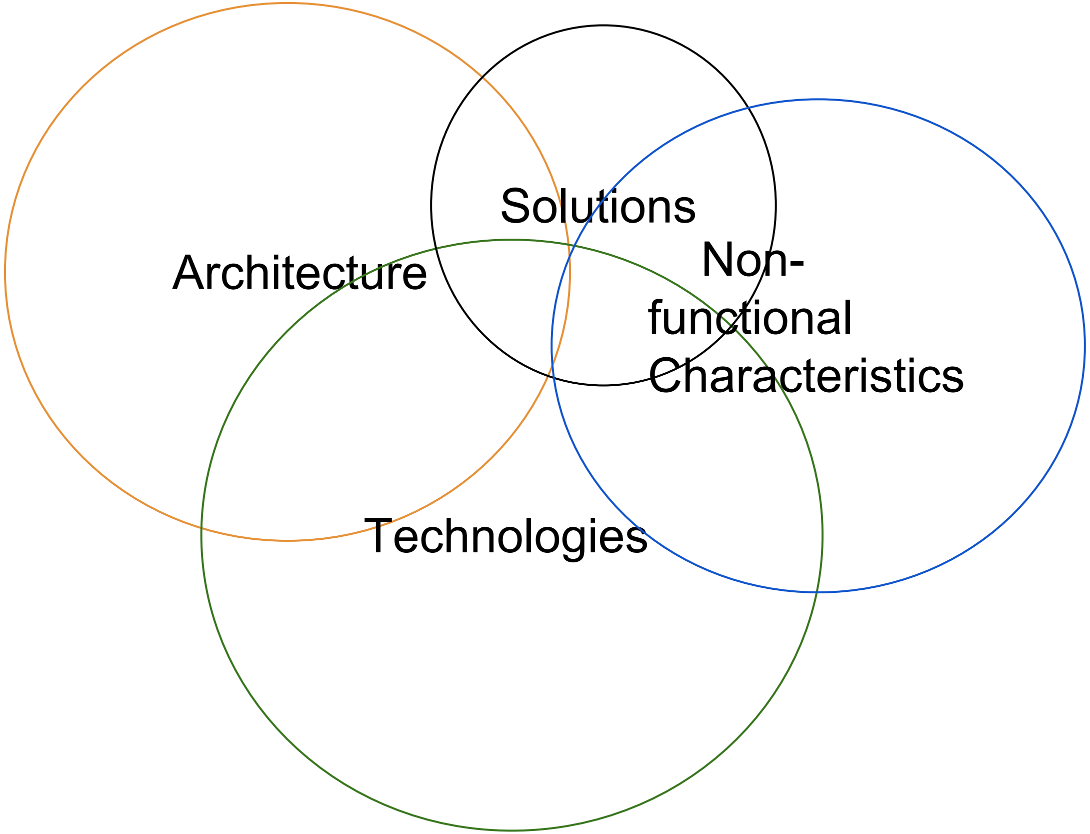
W205 Overview

Jari Koister

Objectives of this presentation

- Provide an overview of learning objectives and changes since last semester.
- Overview of architecture concepts.
- Related each module to the overall picture.

Understanding how to build storage and retrieval



Architecture: How to piece together a solution from parts.

Solution: processes, representation, or models for various analytics problems.

Non-functional: scale, performance, reliability, security aspects.

Technologies: individual technologies that are used in architectures.

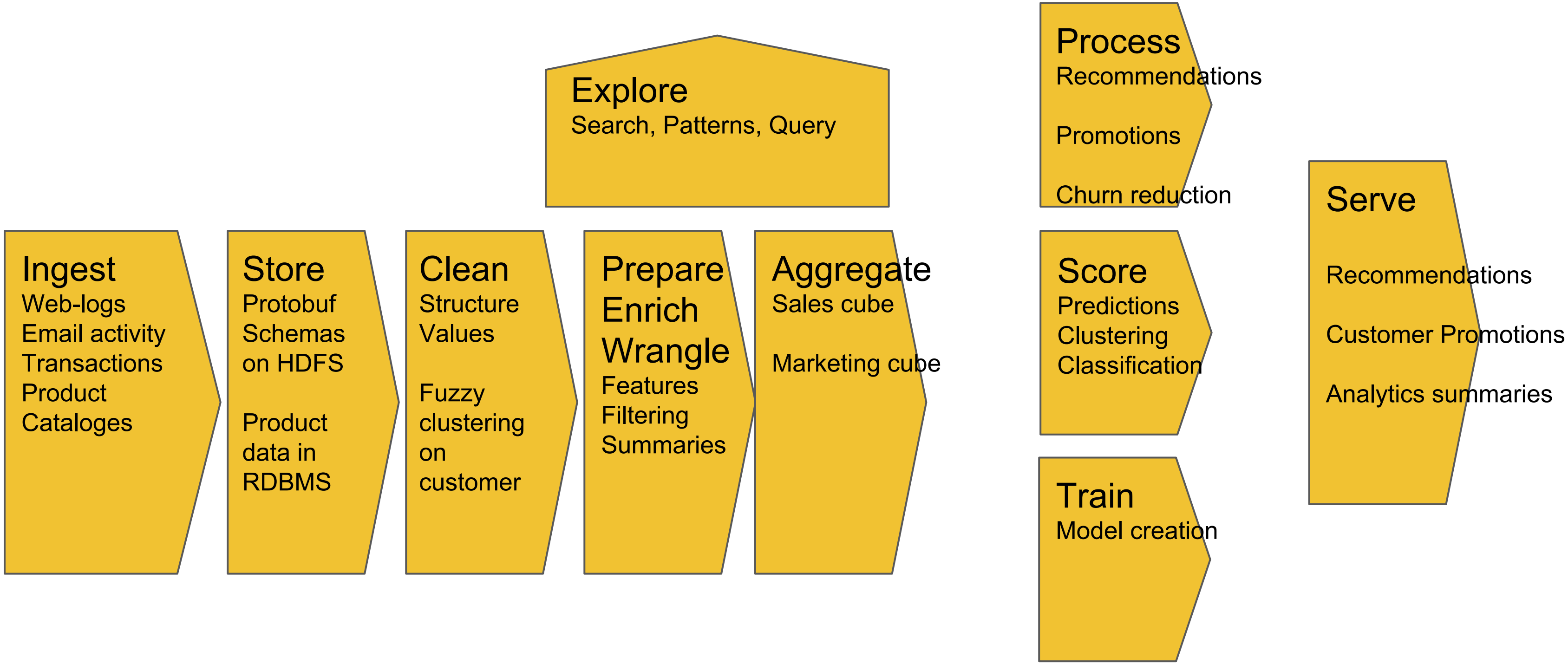
Course Learning Objectives

1. Understand all main **architectural components** involved in **building analytics processes** and applications that results from data science activities. This includes for example data definition and storage, data ingestion, data processing querying, data cleaning, data serving.
2. **Understand fundamental non-functional architectural concepts and characteristics that are considered** when building analytics processes and applications. This includes data scale, processing complexity, network performance.
3. **Understand the nature and needs of processing and storage** for the various processes involved in data analytics. Be able to evaluate any solution according to some fundamental concepts (dimensions)
4. Understand trade-offs between different technology choices. **Conceptual model** differences, **functional** differences, **scale** and **performance** differences.
5. **Hands-on experience** and introductory knowledge of selected technologies.
6. Ability to analyze a problem and **select an appropriate architecture based on functional and non-functional** requirements as well as known characteristics of technical solutions.

Feedback from last semester

1. Difficult to get initial environments up which had downstream impact on later labs and exercises.
   a. We have redone labs 1-4 to facilitate the initial learning and set up process.
   b. We suggest bridge course or self learning of python and Unix/Linux.
2. Hard to see the forest for the tree’s. How to put everything in context.
   a. We will provide more guidance on context.
3. Some readings were hard to digest and to place in the right context.
   a. We will provide reading guidance on challenging or long material.
4. Breadth v.s depth on technology.
   a. More focus on Lab technologies in synchronous sessions.
   b. Depth learning during exercises and projects, pick which tech you like to learn.

Example: Marketing Analytics



Ingest
Web-logs
Email activity
Transactions
Product
Cataloges

Store
Protobuf
Schemas
on HDFS

Product
data in
RDBMS

Clean
Structure
Values

Fuzzy
clustering
on
customer

Prepare
Enrich
Wrangle
Features
Filtering
Summaries

Explore
Search, Patterns, Query

Aggregate
Sales cube

Marketing cube

Process
Recommendations

Promotions

Churn reduction

Score
Predictions
Clustering
Classification

Train
Model creation

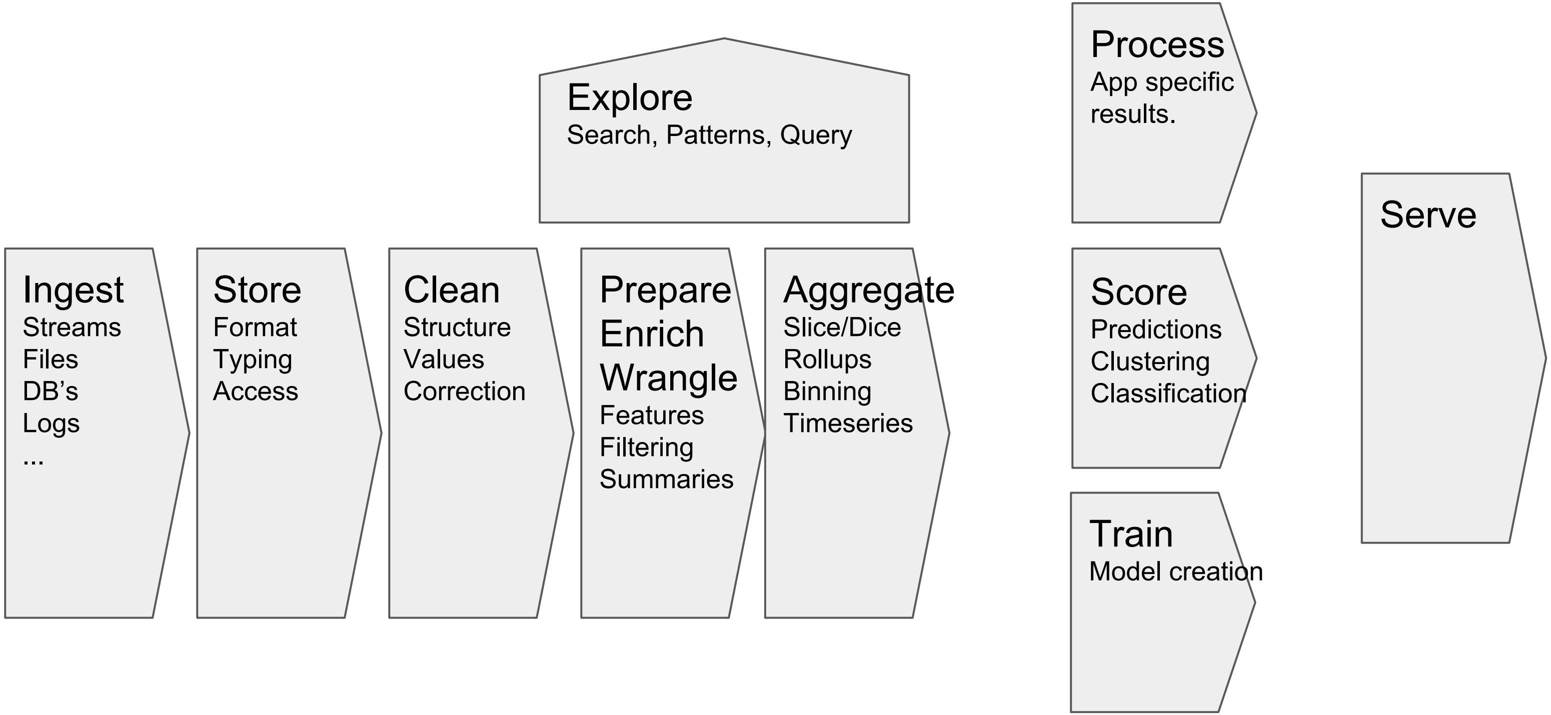
Serve

Recommendations

Customer Promotions

Analytics summaries

Overall Processes



Explore
Search, Patterns, Query

Process
App specific results.

Ingest
Streams
Files
DB's
Logs
...

Store
Format
Typing
Access

Clean
Structure
Values
Correction

Prepare
Enrich
Wrangle
Features
Filtering
Summaries

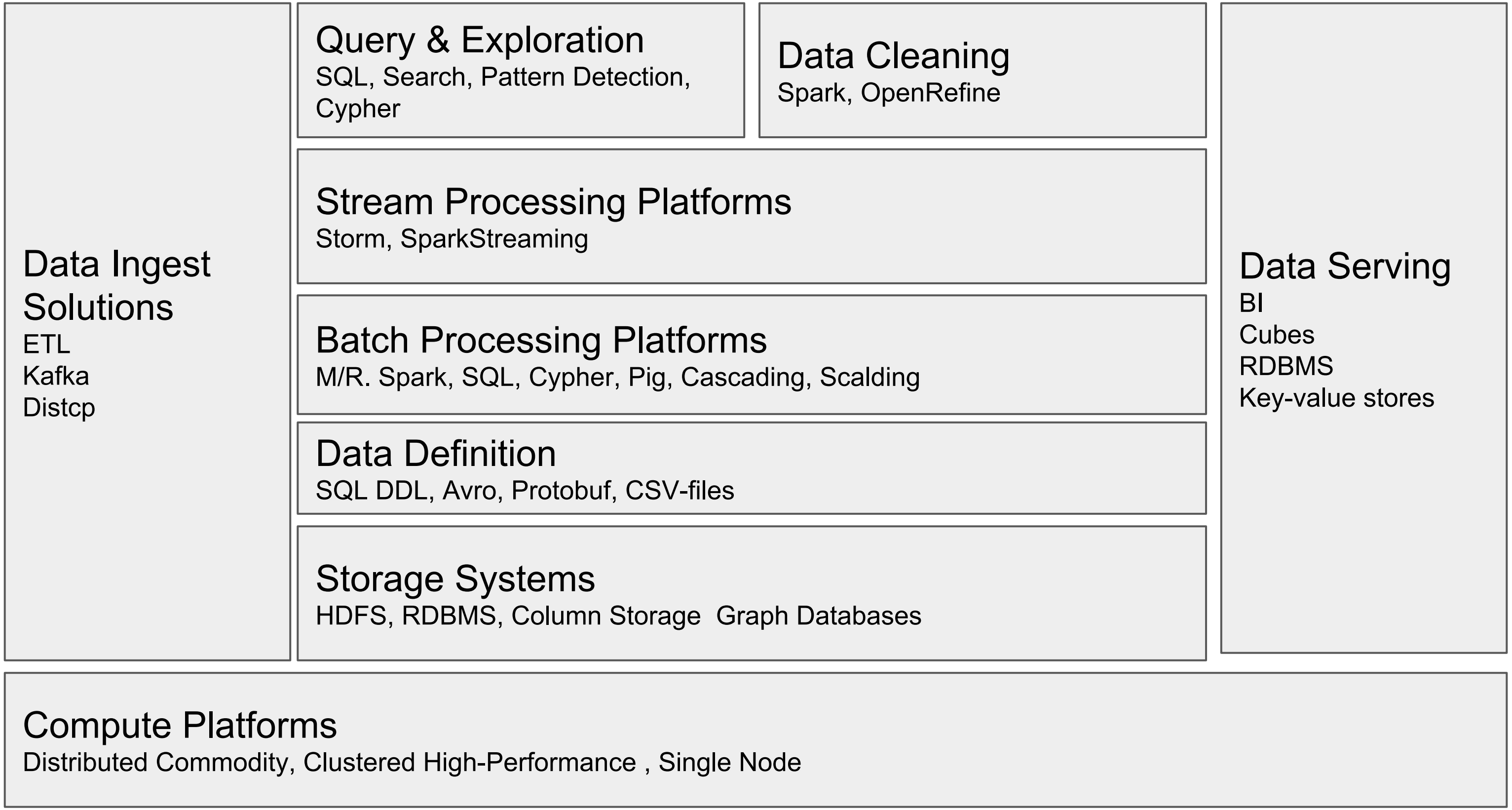
Aggregate
Slice/Dice
Rollups
Binning
Timeseries

Score
Predictions
Clustering
Classification

Train
Model creation

Serve

Overall Architecture: Layer Diagram



Data Ingest Solutions
ETL
Kafka
Distcp

Query & Exploration
SQL, Search, Pattern Detection, Cypher

Data Cleaning
Spark, OpenRefine

Stream Processing Platforms
Storm, SparkStreaming

Batch Processing Platforms
M/R. Spark, SQL, Cypher, Pig, Cascading, Scalding

Data Definition
SQL DDL, Avro, Protobuf, CSV-files

Storage Systems
HDFS, RDBMS, Column Storage Graph Databases

Data Serving
BI
Cubes
RDBMS
Key-value stores

Compute Platforms
Distributed Commodity, Clustered High-Performance , Single Node

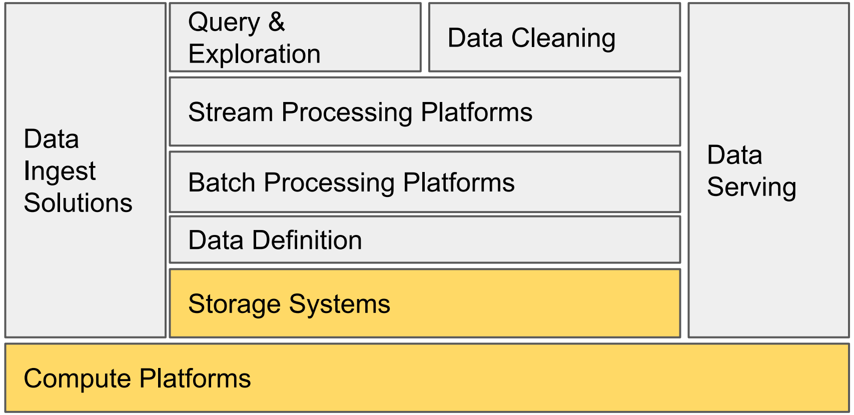
Module 1 and 2: Infrastructure Basics.

What are the considerations when selecting and using technologies?

What are important aspects for scaling: data size, network, processing.

What are common architectural patterns?

What are the considerations between single node and distributed processing?



Data Ingest Solutions

Query & Exploration

Data Cleaning

Stream Processing Platforms

Batch Processing Platforms

Data Definition

Storage Systems

Data Serving

Compute Platforms

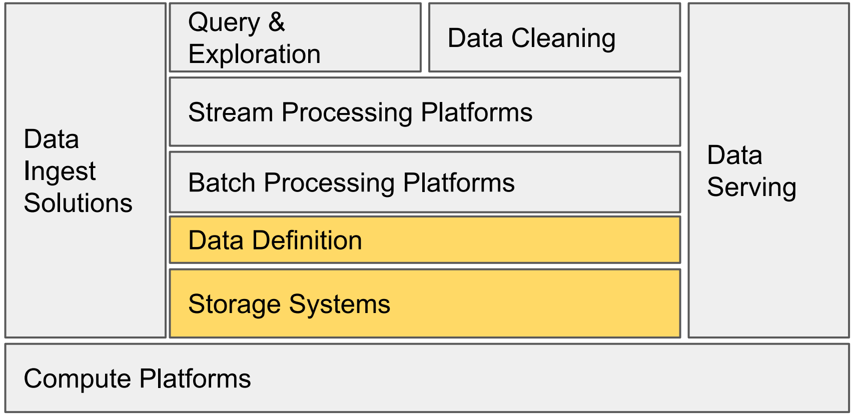
Module 3 and 4: Storing and Defining Data

Understand storage and definition of data.

Schemas at different levels.

Different storage architectures: HDFS, RDBMS, NOSQL, Object Storage

What is the concept of a Data Lake.



Data Ingest Solutions

Query & Exploration

Data Cleaning

Stream Processing Platforms

Batch Processing Platforms

Data Definition

Storage Systems

Data Serving

Compute Platforms

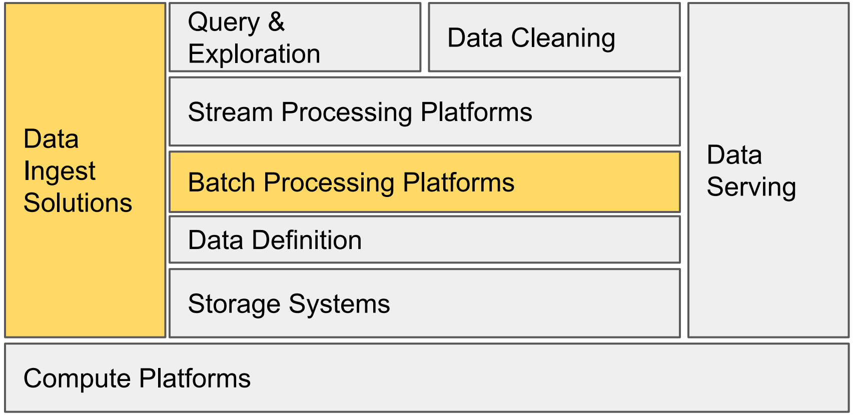
Module 5 and 6: Ingestion and Processing

How do you get data to your analytics processing.

Batch and real-time ingestion.

Schemas at different levels.

What are methods for processing data: grouping, filtering, aggregation.



Module 7, 8 and 9: Querying, Exploring and Cleaning

Using queries: SQL,Joins, Indexes

Approximate SQL

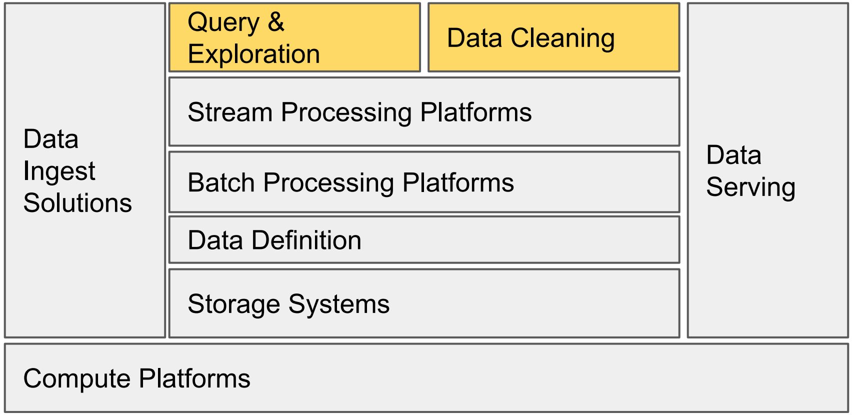
Exploring Data: Enriching, sampling.

Exploratory v.s Confirmatory Discovery

Unsupervised techniques.

What is involved in cleaning data. What are the processing needs.

How do you scale cleaning of data.

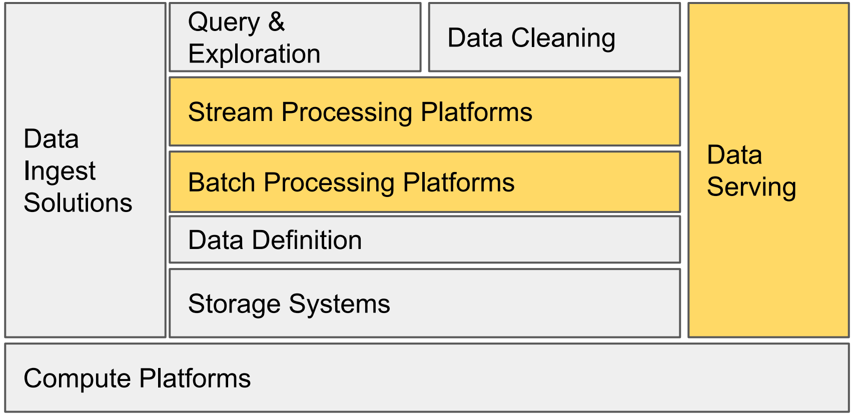


Module 10, 11, and 12: Streaming, Graphs and Serving

An infrastructure of real-time analytics of streams.

An infrastructure for representing and querying data as graphs.

How are analytics results served.



Summary

1. Learning objectives are broad and ambitious; included conceptual and hands-on learnings.

2. We changed labs, provide more overview, will focus more on tech i sync sessions and will provide guidance on readings. You need to make sure you have the basics in python and linux.

3. When you feel lost in the woods, related back to the overall process and architectural context.