Drug Abuse Project

Architecture Design
And
Planning

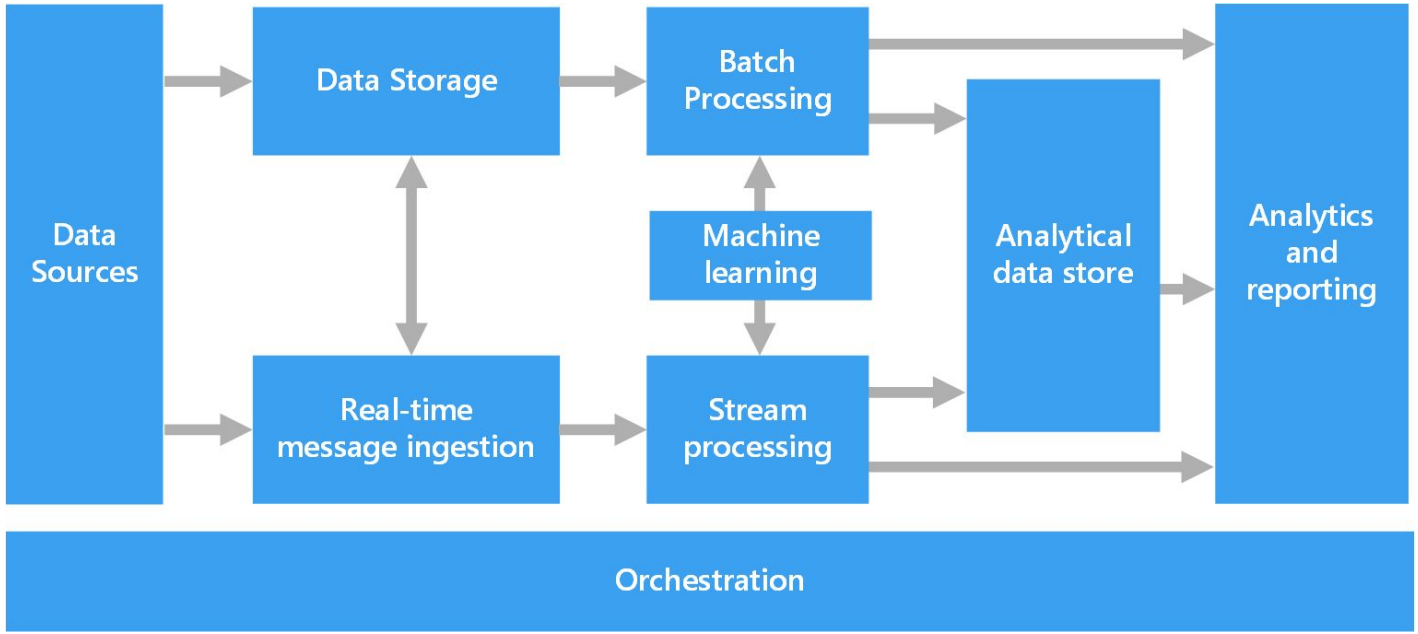
Julina Maharjan

May 20, 2020

Architecture Design

Approach

Big Data Architecture Solution

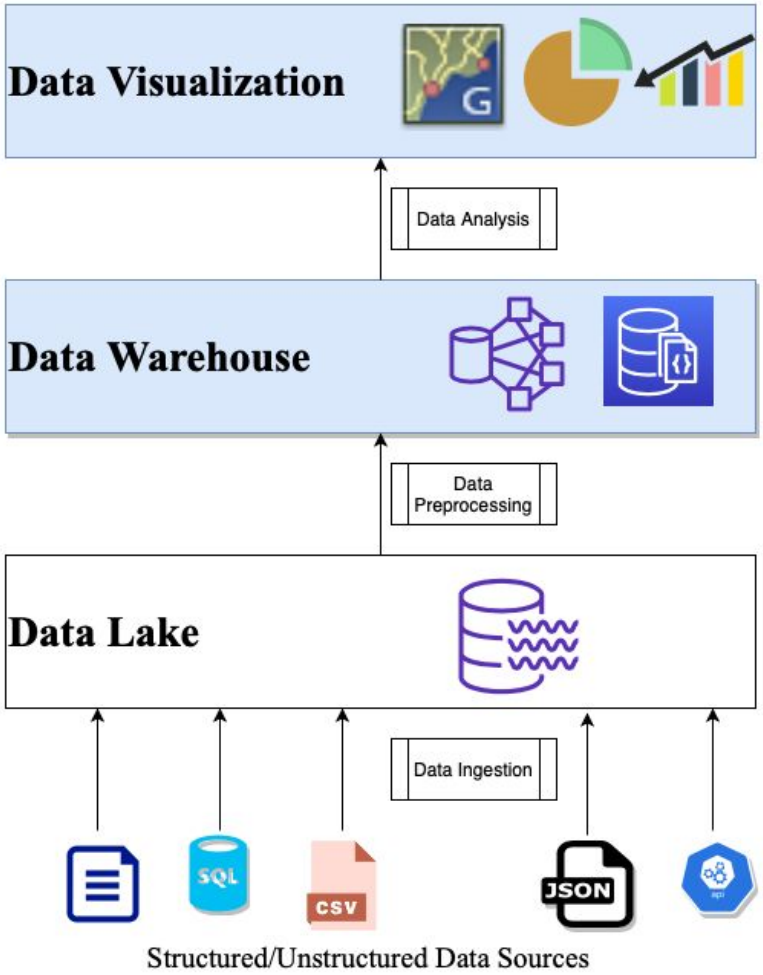


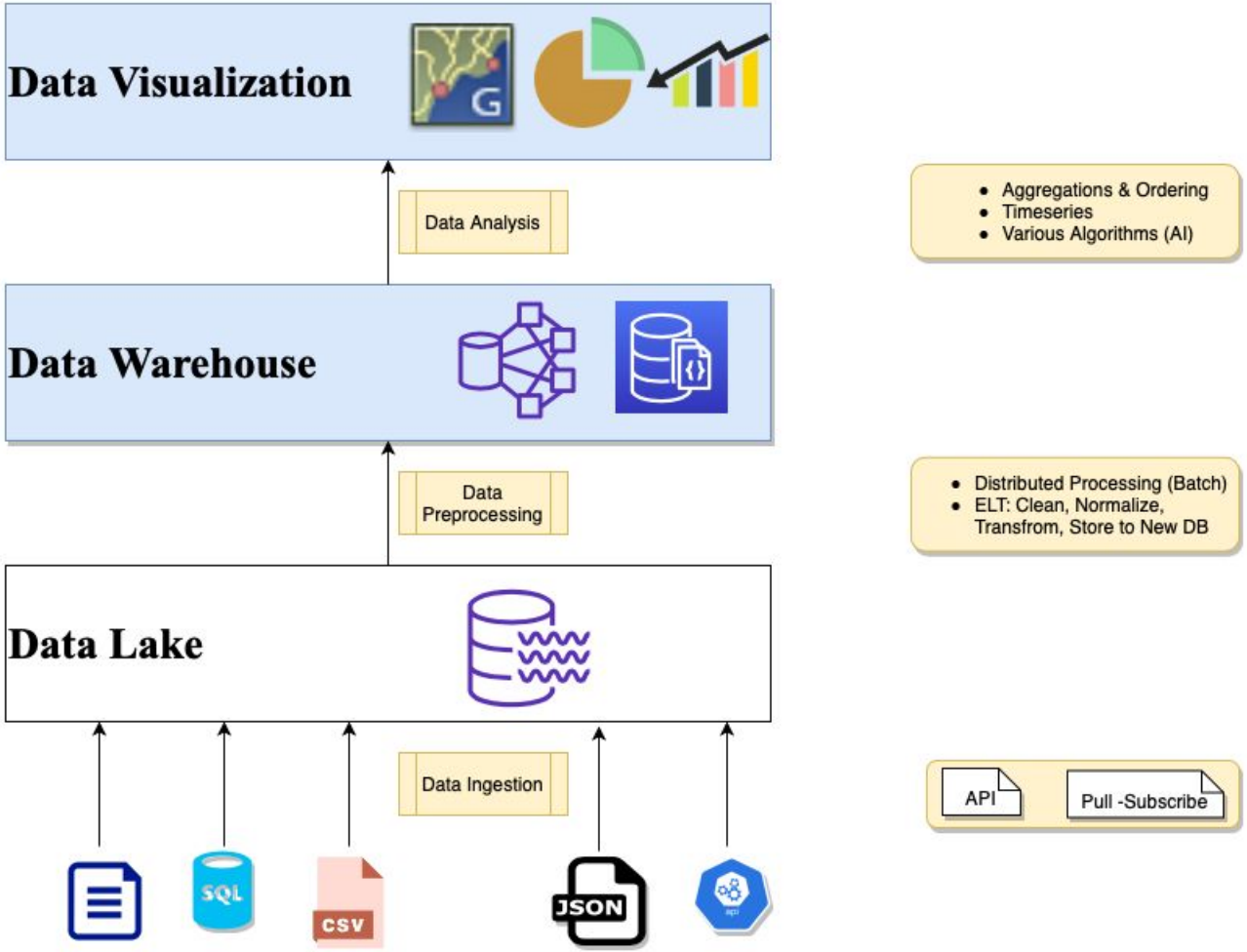
Main Consideration

- Big Data
    - Varieties of Data (Structured/Unstructured)
    - Volume of Data
    - Velocity

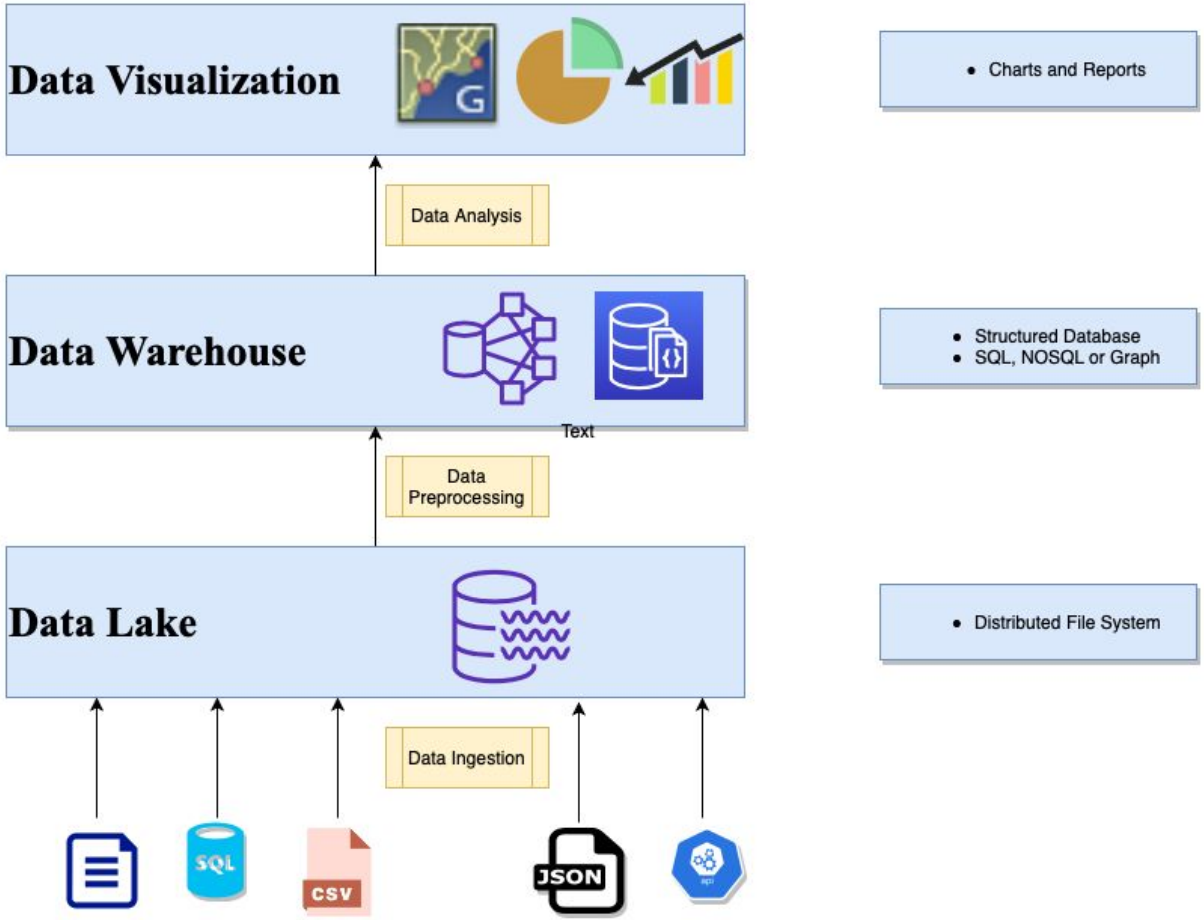
- Scalable Software

Abstract Design

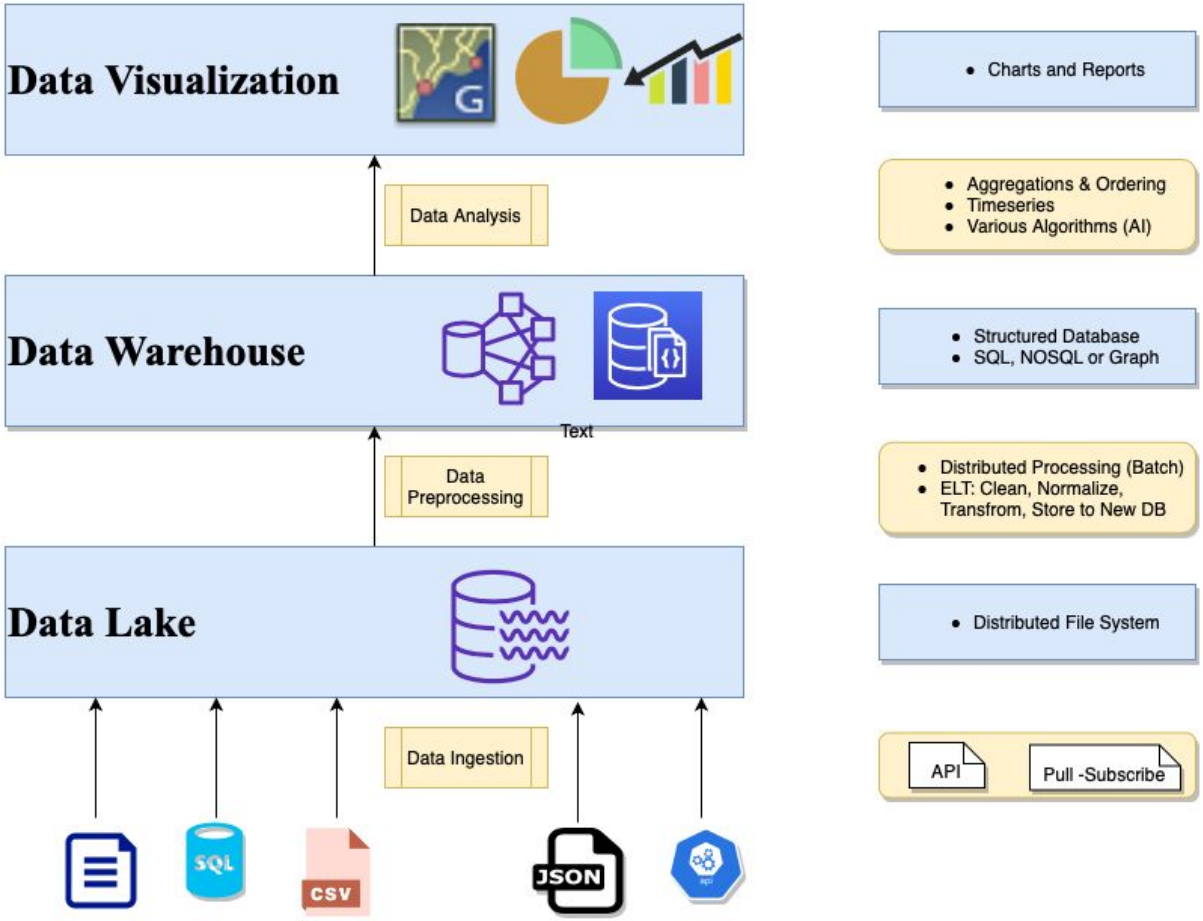




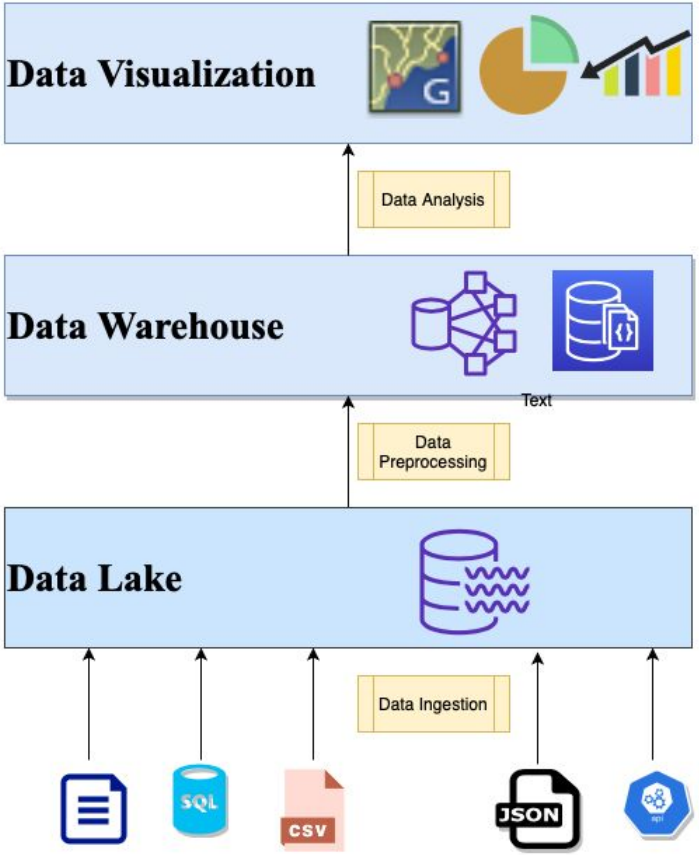
Structured/Unstructured Data Sources



Structured/Unstructured Data Sources



Structured/Unstructured Data Sources



Structured/Unstructured Data Sources

Tools

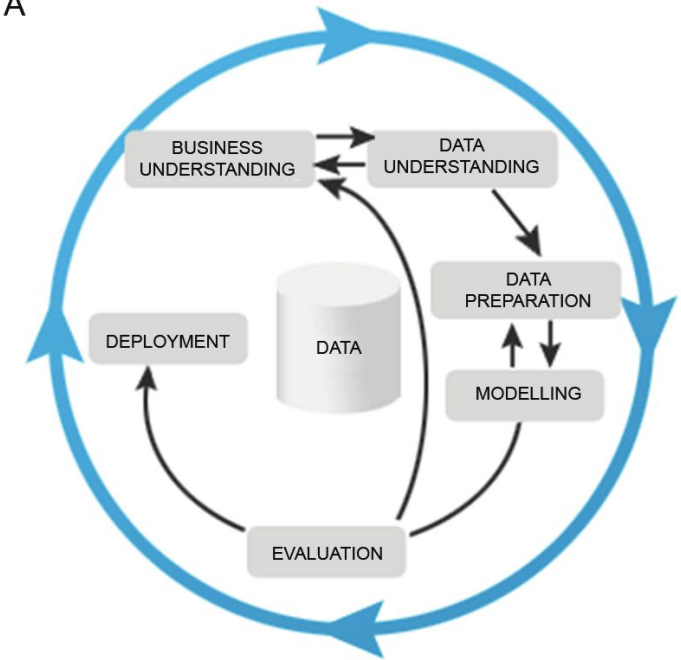
• Charts and Reports	• D3JS • Chart tools
• Aggregations & Ordering • Timeseries • Various Algorithms (AI)	Algorithms - • Lyrics Mining • Sentiment Analysis (NLP-BERT)
• Structured Database • SQL, NOSQL or Graph	Neo4j Graph DB mongoDB
• Distributed Processing (Batch) • ELT: Clean, Normalize, Transfrom, Store to New DB	Spark Data Processing - • Scala/Python
• Distributed File System	HDFS cluster
API Pull -Subscribe	• Tweet API Crawler
	• Tweets • Hospitalization Data • Census Data • Geo-Location Data • Lyrics

Plannings

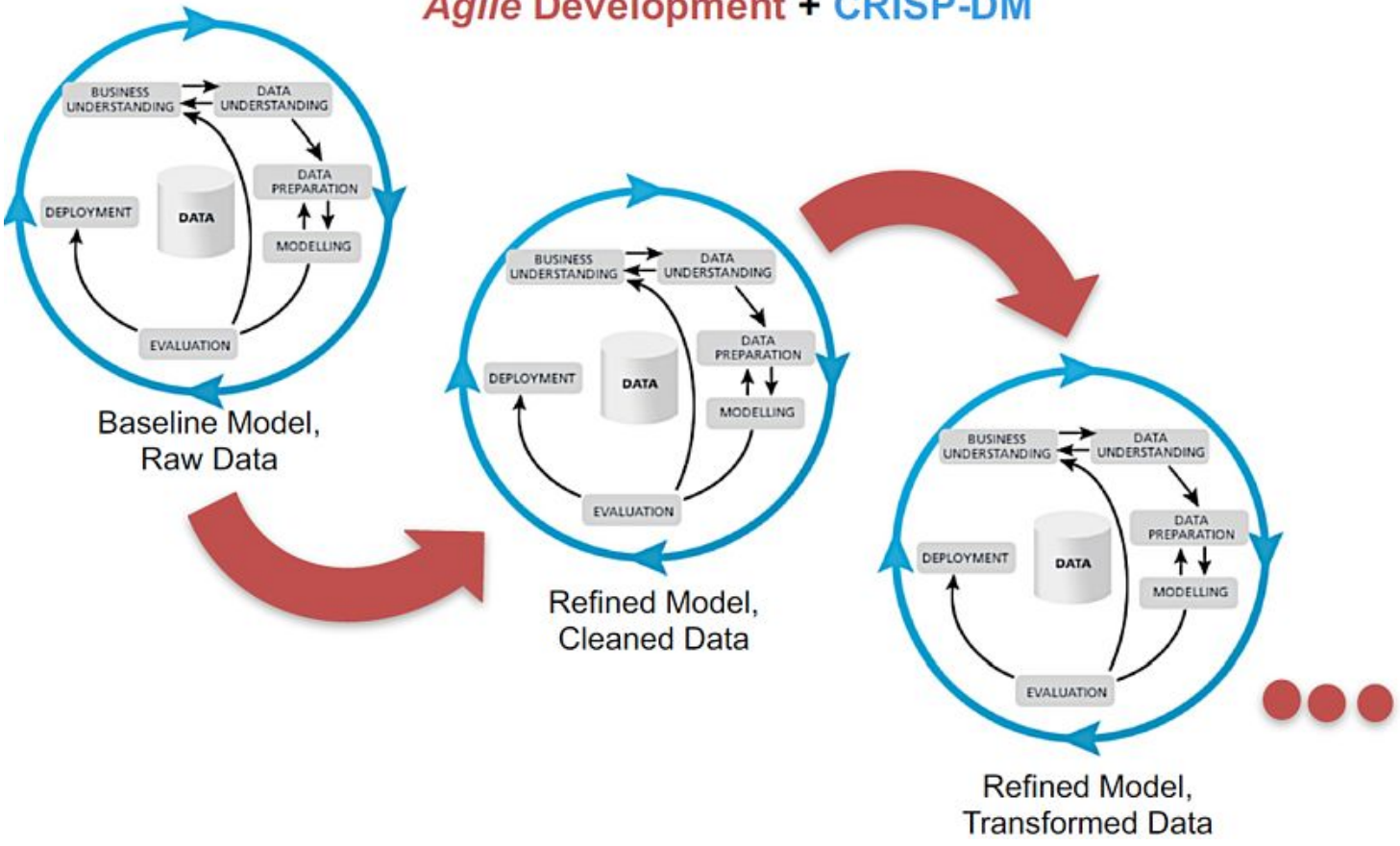
Methodologies : CRISP DM

- Cross-industry standard process for data mining

A



Agile Development + CRISP-DM



Monthly Timeline

May 2020	June 2020	July 2020
Data Warehouse (Tweets, Hospitalization Data, Census Data, Geo-Loc Data, Lyrics) ● Cluster Set-up ● Package Installation <u>For each data</u> ❑ Data Pre-processing Clean, Normalize, Store ❑ Data Modelling ❑ Writing to DB ❑ CRUD Operations	Data Visualization ● Front End Design ● Reports ● Heatmaps, Charts ❑ Framework installation ❑ Reporting	Data Lake ● Operational tasks ● Deployment ❑ API Crawling ❑ Set up HDFS ❑ Read data from HDFS and write to Data Warehouse ❑ Automations

Additional Tools

Code Repository	GitHub
Development Server	scorpion.cs.kent.edu

Thank You