Cycle 1

Cycle Goal

- Load data into Hive
- Be able to pull data from the web source into Hadoop and lay initial groundwork for automation
- Profile Data
-produce a refined, and joined table
-create a visualization on refined table

Setup Phase

-Search for and Aggregate public data sources
- Choose tables with interesting and correlatable information
- Find tables that have foreign keys into each other
-Set Up Hadoop environment with cloudera virtual box instance

ETL Phase

- Develop Naive BASH Extraction script that aggregates data from web source on hadoop edge node
- Develop BASH Transformation Script to prep data for entry into HIVE
- Develop Bash Script that Strips headers from the tables and creates a HQL Query to Load data into a HIVE raw zone

Refined Zone

- Research Data profiling tools: Trifacta, Ataccama, Talend
-Research data prep tools: Trifacta, Talend, datamere
- Profile and Refine data in Trifacta desktop wrangler
- Join refined data sets together in one table with an inner join using Trifacta

Data Discovery and Visualization

- Look for interesting correlations in data
- Develop initial visualizations
- Ensure data completeness and that you have a large enough sample size to work from
- Find flaws in initial refined zone model, and opportunities to make connections with new data sources

Cycle 2

Cycle Goal

- Improve on cycle 1 process
- add new tables to the data lake
- Improve usability of automated ingestion script to support additional tables and beeline
- transfer cleansing and profiling done in Trifacta to Kariba and Hadoop native environments
- Improve on previous visualizations to incorporate new tables and be more polished

Setup Phase

-Find new Data Sources that has possible correlations to your current data pool, and can be joined on a foreign key
- get access to Kariba

ETL Phase

- Extend the extraction script to pull down new data sources
- Extend the Transform scripts to accommodate the new data sets and prepare them for loading into HIVE
- Extend Load scripts to be able to interface with beeline, so it can load data into kariba servers

Refined Zone

- load data into Kariba profiler
> later scrapped due to it not working correctly and done in Trifacta again
- Refine Script transformed to include data quality modifications needed for new data sets, and improvements to old ones
- Attempt to reproduce Refine scripts in Apache PIG, however it was tedious and ineffective. Eventually Used HIVE to do all the Refine jobs.
- Write HIVE query to join tables together based on Process developed in Trifacta

Data Discovery and Visualization

- Used inferences from old visualizations to create new, more statistically sound visualizations
- Found correlations with new data sources that we added to make more interesting visualizations
- Refined the overall apperance and experience of the visualizations created

Cycle 3

Cycle Goal

- Incorporate Financial Data sets
- Extend Ingestion Framework to be more flexible to be able to incorporate the new data sets, and to be more reliable and usable
- Continue polishing and adding to tableau visualizations
- Experiment with outer Joins to see if it is worth keeping null data and if it improves the accuracy of the visualizations

Setup Phase

-Identify points of instability in Ingestion framework ---> bash was unreliable and hard to work with
- Find financial data on schools to be added to data lake

ETL Phase

- Re-write Ingestion framework to do most of the legwork in python, and to take in easy to edit csv files to allow for easy addition of new data sets without hard coding them
> Extraction done in Bash
> Transformation done in python
> Loading done in Bash and Hive QL
- Python Script now writes the HIVE QL Query to load the tables into hadoop

Refined Zone

- Profile new data financial data in Trifacta
- Due to different Granularity of new data create a new refined join of only financial data with both inner and outer join
- Copy old join script to perform an outer join
See if outer join is better --> outer join was better for now, but causes its own problems

Data Discovery and Visualization

- Union the two tables in tableau on the district. Financial data only district level, while other data was school level.
- polish old visualizations
- generated newer more interesting dashboard
> problems with nulls and tableau