Analysis Platform: PySpark + MLLib

Preprocessing

- Parallelized with Spark, scales to larger datasets
- Extracts logically related subsets of data
- Filters out records with missing data
- Filters out irrelevant feature columns

Targeted Subset of Data

Data is now ready to be ingested by an analysis tool, or directly visualized

MLLib Analysis Tools

- Descriptive Statistics
- Classification
- Clustering

Visualization

- Generate plots specific to analysis technique