Big Data Science Workloads

Milind Bhandarkar
(Chief Architect, Greenplum Labs, EMC)

Applications Drive Systems

- Data Science
  - Machine Learning
- Analytics & Reporting
- Visualization

Data Science Workload

- http://www.dataists.com/2010/09/a-taxonomy-of-data-science/
- Obtain, Scrub, Explore, Model, Interpret

Obtain

- Usable & sufficient corpus from multiple independent sources
- Automated for streams
- Efficient ingestion for one-time data

Scrub

- Raw data is always messy
  - Missing data, inconsistent data, charsets
    - NY, New York, NYC, Big Apple etc
- Growing Dictionaries

Explore

- Visualize, Clustering, Dimensionality reduction
  - Feature correlations (scatter plots)
  - Single feature histograms

Model

- Find correlation of past data and outcome
- Find and label good training set
- Derive model parameters
- Apply model, and validate

Interpret

- Models are built for prediction and interpretation
- Check that there are no “surprises”
- Reason about models
- Improve models

Data Science Data Flow

- Raw Data (Timed, Partitioned, Crowdsourced, De-duped etc)
- Derived data (simple aggregates, other statistics)
- Models (Feature weights, decision trees)
- Indexes

Benchmarks

- Need to emulate real data science workloads at various scales
- TeraSort, Grep and Wordcount not enough ☺