OpenChorus: Building a Tool-Chest for Big Data Science

Milind Bhandarkar
Chief Scientist, Machine Learning Platforms
EMC Greenplum
Agenda

• Tools for Data Science
• Data Science Workflow
• Greenplum OpenChorus
• How Chorus Works
Data Science Tools: Abundance of Riches

- Proliferation of tools
- Languages & Libraries
  - R, Matlab, Python – SciPy, NLTK, Madlib, Mahout
- Frameworks
  - Graphlab, Pregel (Giraffe), Mesos, CEP
- Platforms/Data Stores
  - MPP Databases, Hadoop, NoSQL (Hbase, Cassandra, MongoDB), SciDB
Choice of Tool(s)?

• Hammer?
  ‒ Hadoop ought to be sufficient for most tasks
  ‒ “If all you have have a hammer, throw away everything that is not a nail” – Jimmy Lin (http://arxiv.org/abs/1209.2191)
  ‒ Operational complexity / learning curve not worth efficiency

• Tool-Chest?
  ‒ Use the right tool for the right job
  ‒ How to reduce complexity
Hammer or Tool-Chest?

Let the workload decide
Data Science Workload

- Obtain
- Scrub
- Explore
- Model
- Interpret
Obtain

- Corpus needs to be usable & sufficient
- Possibly from multiple independent sources
- Needs to be automated for streams
- Needs to have 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

- Join with Crowdsourcing
  - Mechanical Turk etc
Explore

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

- Challenge: How not to lose these observations
Model

• Find correlation of past data and outcome
  – Find good training set
  – Label the training set
  – Derive model parameters
  – Apply model, and validate

• Ensemble Models: Model of models
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, Deduped etc)
- Derived data (simple aggregates, other statistics)
- Models (Feature weights, decision trees)
- Indexes
Data Diversity

- Natural Language Text, and Annotations
- (Bags of words) : Concept
- Graphs (sparse matrices)
- Dense Matrices
- Location (proximity)
Too Many Tools for One Data Science Project

Data Exploration and Sharing
- Data marts
- Excel spreadsheets
- Flat files

Analytics Tools
- Data mining
- BI/Visualization
- Data integration

Content and File Management
- Share drives
- Wiki
- Content mgmt app

Communications
- Emails
- Meetings
- IMs

Process Documentation
- Some project plans, no up-to-date team collaborative documentation
High Cost of Knowledge Sharing

- Data science process breaks when organization structure changes
- Very difficult knowledge transfer
- No "insurance policy" for the data science intellectual assets
Delayed Time-to-Market

1. Find the data
2. Get access to data
3. Learn about the data
4. Move to sandbox
5. Analysis Finally!
6. Operationalize the model

6-9 Months
Greenplum Chorus

- Collaborative analytics
- Powerful extensibility
- The freedom of open source

*Greenplum’s Social Platform for Collaborative Data Science*
Chorus Enables Collaborative Data Science

- Collaborate within projects, share **data, content, and findings** across teams
- Make projects more transparent
- Iterate faster for accelerated insights with real-time social collaboration
Powerful Extensibility

- Integrated development environment for analytics
- Expand insights with simple access to third-party data
- Fusion with leading analytics and visualization tools
The Freedom of Open Source

- Modify and extend to any environment
- Promotes an ecosystem of applications, startups, and data scientists community

www.openchorus.org
How Chorus Works
How Chorus Works

Chorus Workspace

Chorus View

Data

Sandbox

Greenplum DB

Source Data

GPDB External Table

GPDB

EDW

Non-GPDB

DB

Summary of next few slides, with animation built in
Data Exploration
Search and Data Discovery

- Automatic indexing of metadata, work files, comments, and insights
- Quickly find data across the enterprise regardless of location
**Data Exploration**

**Data Preview and Visualization**

- Data preview for instant understanding
- Quick and easy data visualizations
  - Visualize data for faster insight into datasets
  - No need to export to third-party applications like R
  - Not a replacement for advanced visualization tools
Data Exploration
Living Data Dictionary

• Bring everything about the data to the data
  – Attach documents
  – Ask questions
  – Add comments
• Build a living data dictionary
  – Everything is current
  – No more spreadsheets
Workspace – Streamlines Collaboration

- Chorus includes unlimited workspaces, each representing individual project
- Streamlines complex user-user and user-data interactions
Multi-level Secure Collaboration

- **Authentication**
  - Integrates with LDAP and AD for password management

- **Application access control**
  - User roles: Admin vs. general user
  - Workspace types: Public or private

- **Data access control**
  - Chorus enforces database rules and permissions
Data – Dataset Types

1. **Source Dataset**
   - Pointer to the source data
   - Both internal and external data
   - Support both native connectivity for GPDB and flat files
   - Use GPDB External Tables for Non-GPDB databases and Hadoop

2. **Sandbox Dataset**
   - Copy of the source data to be used for analytics
   - Data generated from analytics
Data – Sandbox

- Container of all the analytics data
- Ease of self-service provisioning of sandboxes
  - Free up IT bandwidth
  - Minimize data proliferation to uncontrolled/unmanaged data marts
Data – Populating Sandbox

Import data easily from anywhere:
- Directly from sources
- Through Chorus View
- Flat file import
Data – Chorus View Utility

- Single-view GUI utility for exploring, filtering, aggregating, and moving the desired data from sources to sandbox
- Data exploration and visualization prior to bringing the data into sandbox
- Derive variation of the basic source datasets without bringing the data into sandbox
Data - Chorus View

Select a.userid, a.customer_name, a.gender, a.customer_state, b.ipaddress, b.device,
From customers AS a
INNER JOIN weblog_2012q1 as b
ON a.userid = b.userid
Data – Automated Data Services

- Subscribe to receive automatic updates
  - Schedule imports from multiple data sources
  - Define and share data sets within the data science team
  - Removes manual data refresh activities
Work Files

- Work files are **non-data assets**
  - SQL query statements with code editor interface
  - Execution of in-database analytics, ex: MADLib, PL/R
  - Third-party tool files
  - PowerPoint, Word doc, etc.

- Analytics asset management with version, compare, and archive work files
Integration with Analytics Tools

- Third-party tools
  - Execute in-database analytics functions (ex: MADLib, R) from Chorus work files
  - Publish and execute Alpine Miner Workflow from Chorus native interface
  - Data preparation for analysis using SAS and other analytics tools

- Code-design UI for SQL
Insight and Data Sharing

- Post comments and ask questions on any analytics artifacts
- Share and publish any activities or insights
- Promote fast iteration on data and ideas
Activities and Insights

- Build a living library of activities and insights
  - Define, publish, and share new insights
  - Discover and learn from existing insights
- Iterate faster, model less