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

- http://www.dataists.com/2010/09/a-taxonomy-of-data-science/
- 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

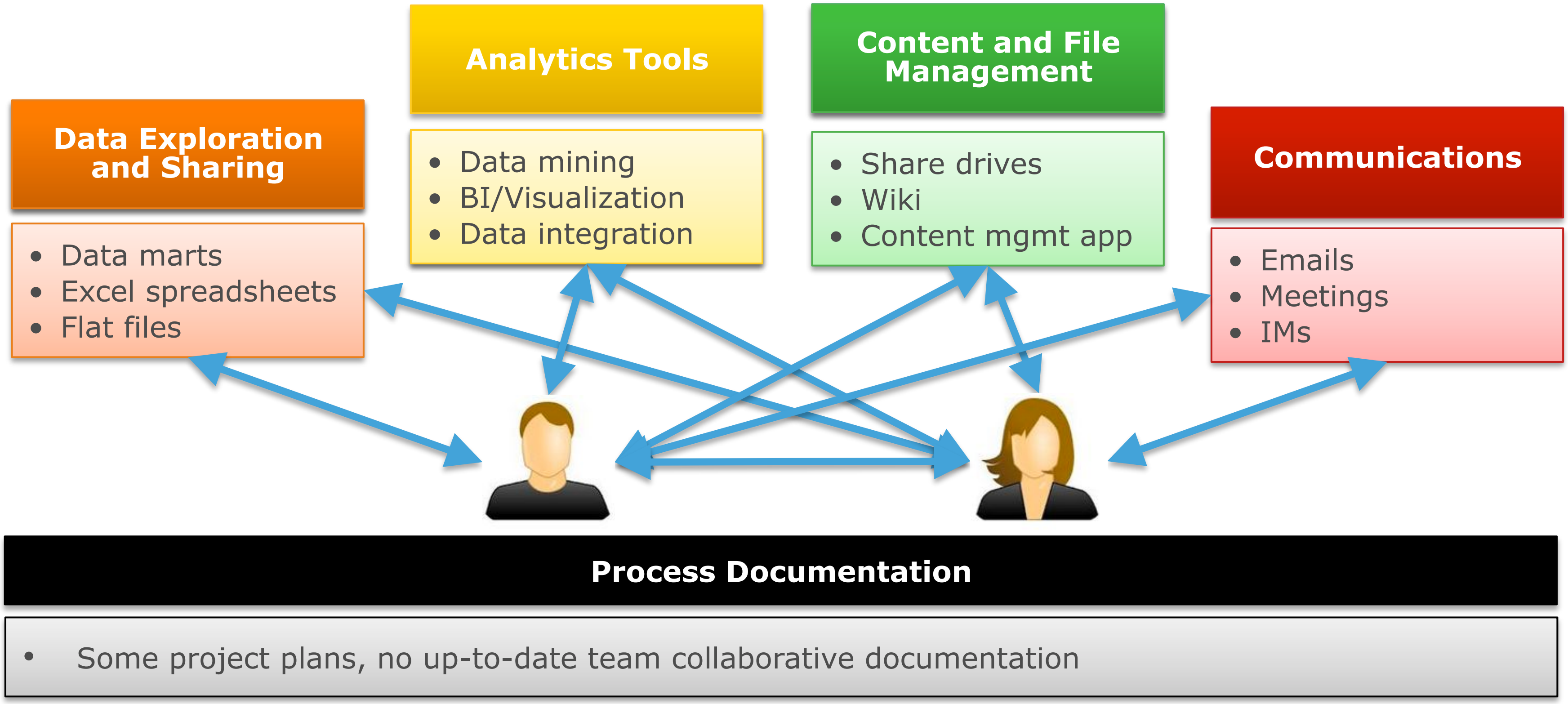
Data Science Data Flow

- Raw Data (Timed, Partitioned, Crowdsourced, De-duped 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

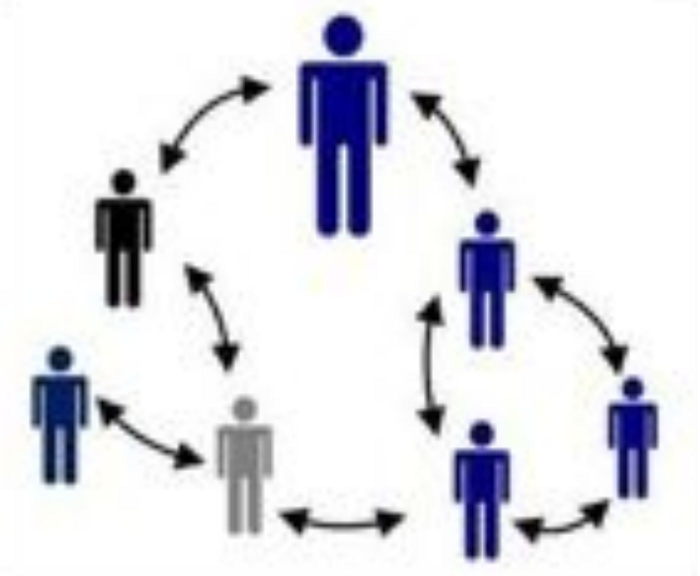
- Emails
- Meetings
- IMs

Process Documentation

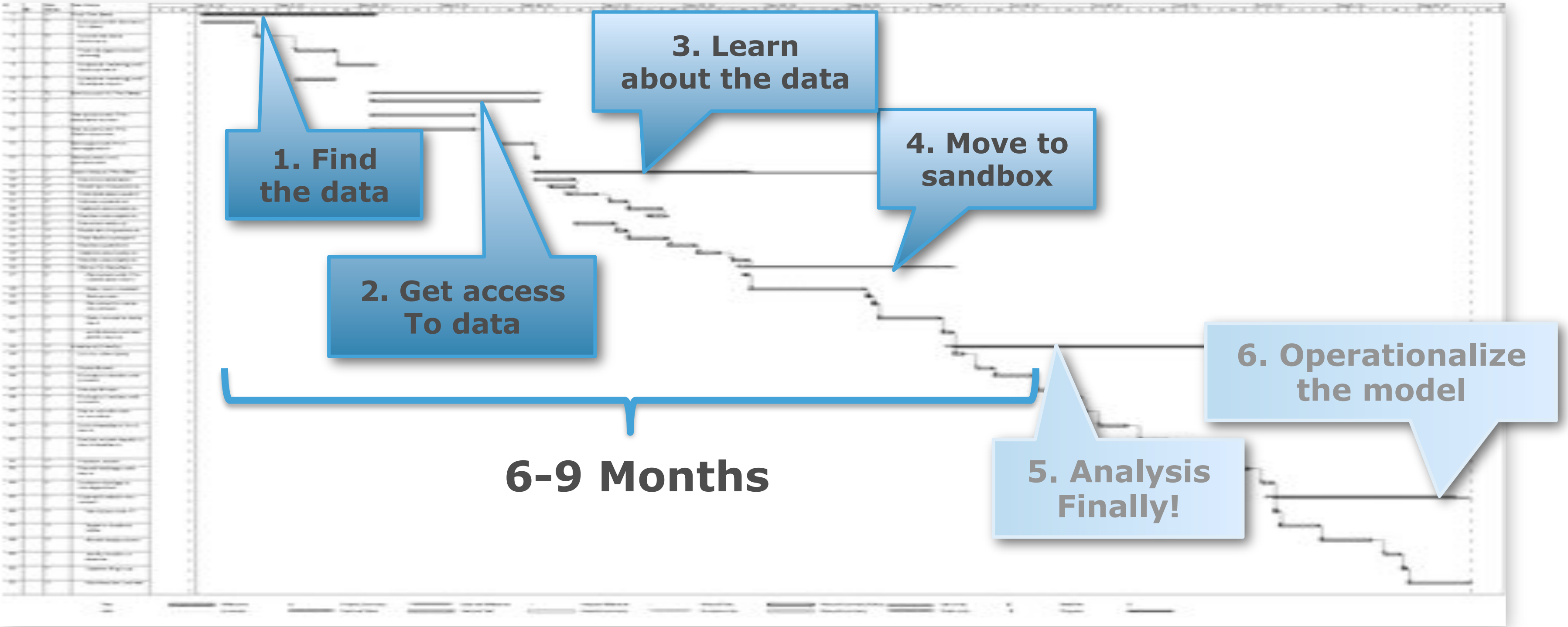
- Some project plans, no up-to-date team collaborative documentation

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



GREENPLUM. EMC²

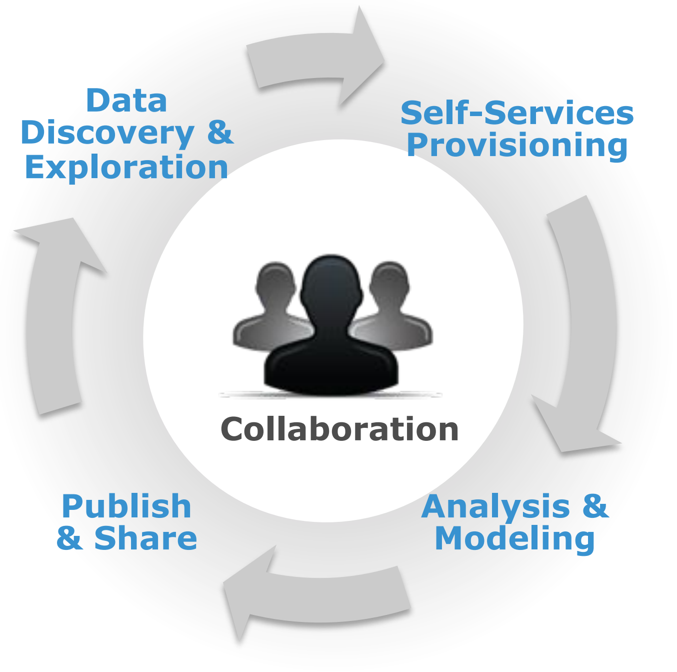
Greenplum Chorus

- 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

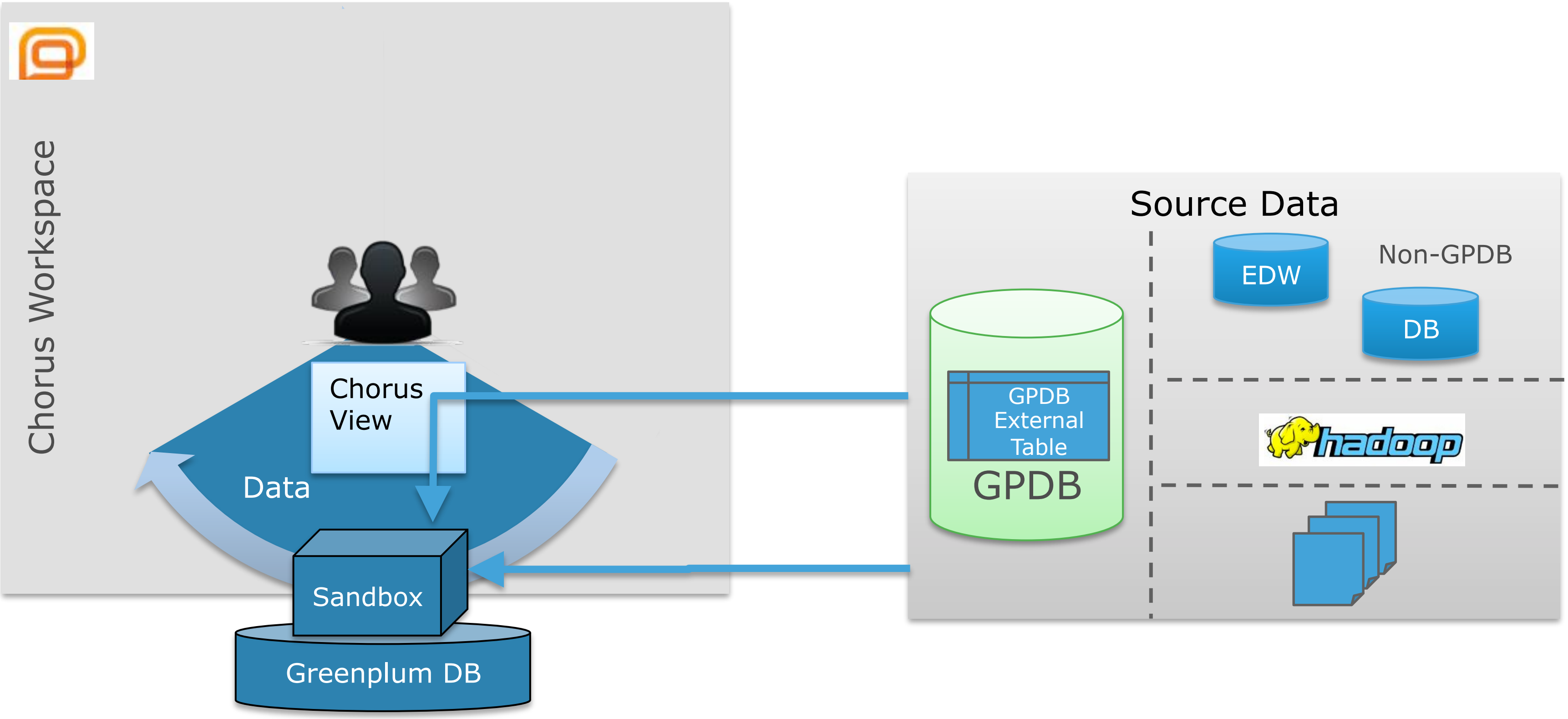
The Freedom of Open Source

www.openchorus.org

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

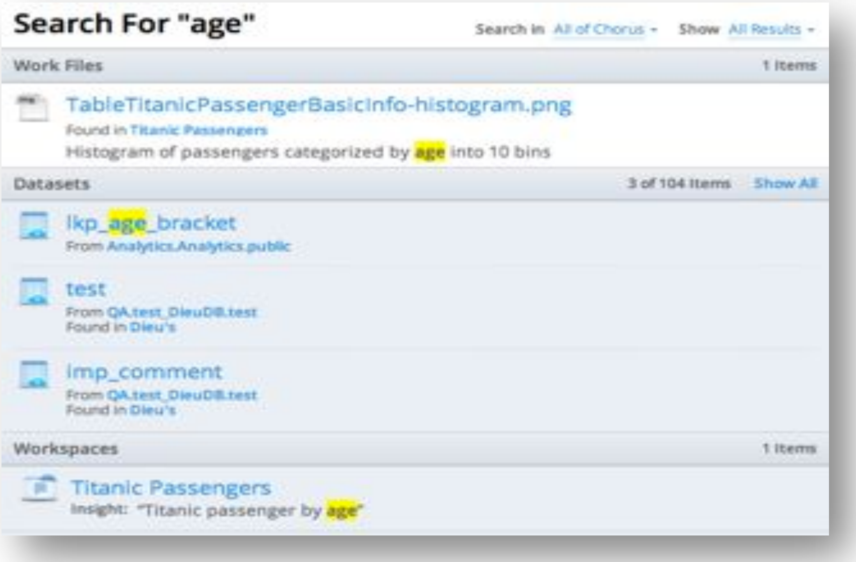
How Chorus Works

How Chorus Works



Data Exploration
Search and Data Discovery

- Automatic indexing of meta-data, work files, comments, and insights
- Quickly find data across the enterprise regardless of location



Search For "age"

Search in All of Chorus ▾   Show All Results ▾

Work Files   1 Items

TableTitanicPassengerBasicInfo-histogram.png
Found in Titanic Passengers
Histogram of passengers categorized by age into 10 bins

Datasets   3 of 104 Items   Show All

lkp_age_bracket
From Analytics.Analytics.public

test
From QA.test_DieuDB.test
Found in Dieu's

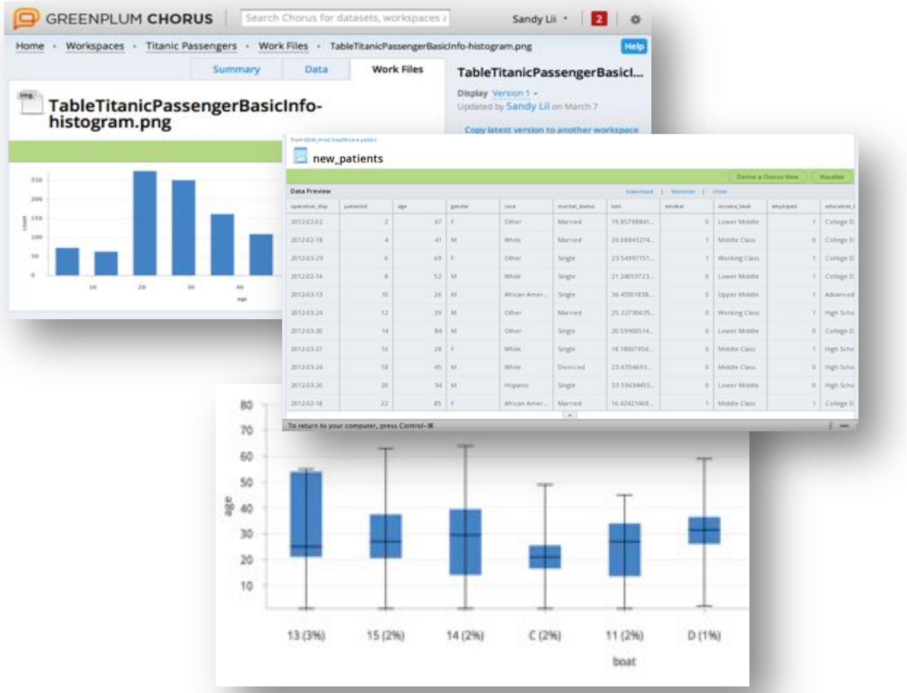
imp_comment
From QA.test_DieuDB.test
Found in Dieu's

Workspaces   1 Items

Titanic Passengers
Insight: "Titanic passenger by age"

Data Exploration
Data Preview and Visualization

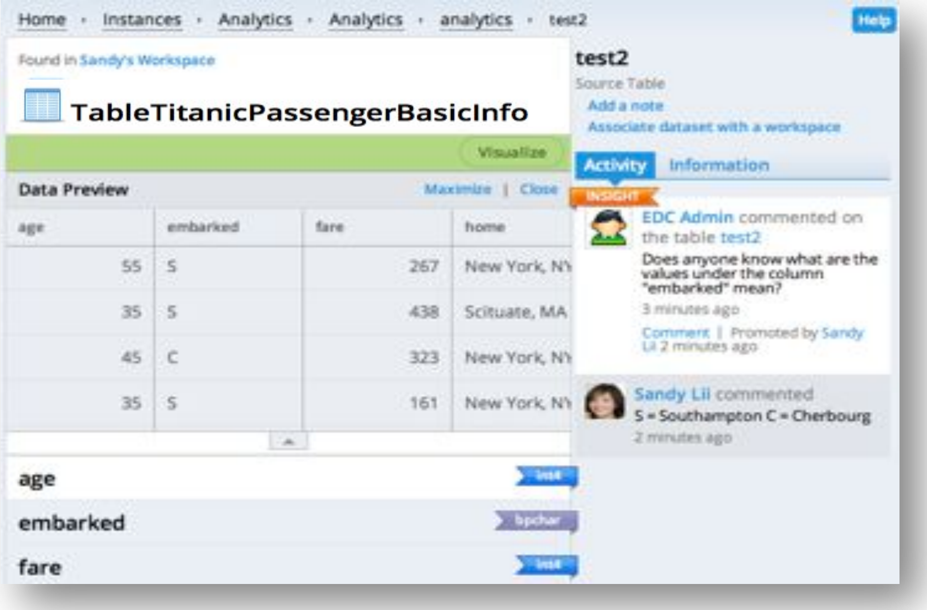
- Data preview for instant understanding
- Quick and easy data visualizations
  - Visualize data for faster insight into datasets
  - No need to export to third-party applications like R
  - Not a replacement for advanced visualization tools


GREENPLUM CHORUS
Search Chorus for datasets, workspaces
Sandy Lii
2
Home
Workspaces
Titanic Passengers
Work Files
TableTitanicPassengerBasicInfo-histogram.png
Help
Summary
Data
Work Files
TableTitanicPassengerBasicI...
Display Version 1
Updated by Sandy Lii on March 7
Copy latest version to another workspace
TableTitanicPassengerBasicInfo-histogram.png
250
200
150
100
50
0
10
20
30
40
age
count
From EDW_Prod.healthcare.public
new_patients
Derive a Chorus View
Visualize
Data Preview
Download
Minimize
Close
To return to your computer, press Control-⌘
80
70
60
50
40
30
20
10
age
13 (3%)
15 (2%)
14 (2%)
C (2%)
11 (2%)
D (1%)
boat

operation_day	patientid	age	gender	race	marital_status	bmi	smoker	income_level	employed	education_l
2012-02-02	2	67	F	Other	Married	19.85798841...	0	Lower Middle	1	College D
2012-02-18	4	41	M	White	Married	26.08843274...	1	Middle Class	0	College D
2012-03-29	6	69	F	Other	Single	23.54997151...	1	Working Class	1	College D
2012-02-14	8	52	M	White	Single	21.24059723...	0	Lower Middle	1	College D
2012-03-13	10	26	M	African Amer...	Single	36.47001838...	0	Upper Middle	1	Advanced
2012-03-26	12	39	M	Other	Married	25.22730635...	0	Working Class	1	High Scho
2012-03-30	14	84	M	Other	Single	20.59900514...	0	Lower Middle	0	College D
2012-03-27	16	28	F	White	Single	18.18007956...	0	Middle Class	1	High Scho
2012-03-26	18	45	M	White	Divorced	23.6354693...	0	Middle Class	0	High Scho
2012-03-20	20	34	M	Hispanic	Single	33.59694455...	0	Lower Middle	0	High Scho
2012-02-18	22	85	F	African Amer...	Married	16.42421468...	1	Middle Class	1	College D

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



Found in Sandy's Workspace

TableTitanicPassengerBasicInfo

Data Preview

age	embarked	fare	home
55	S	267	New York, NY
35	S	438	Scituate, MA
45	C	323	New York, NY
35	S	161	New York, NY

age

embarked

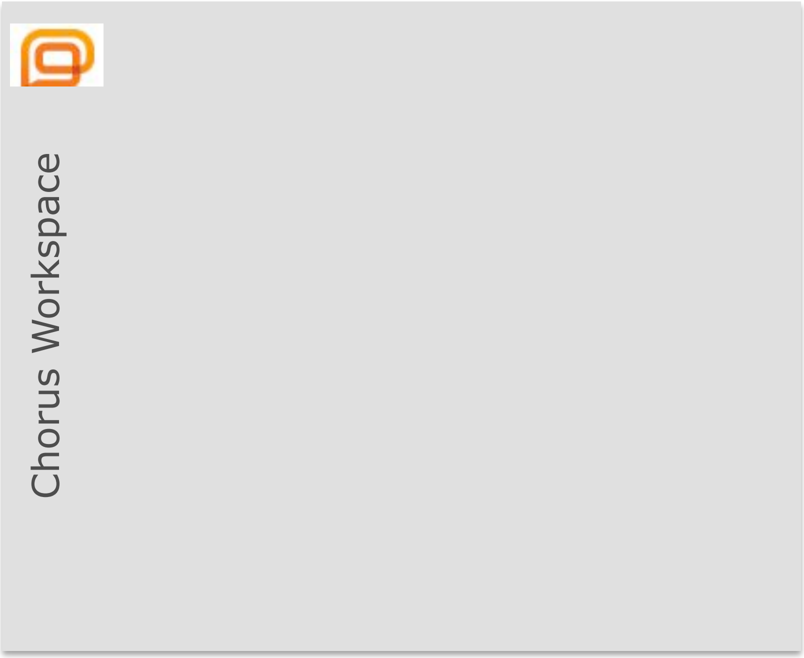
fare

test2
Source Table
Add a note
Associate dataset with a workspace

EDC Admin commented on the table test2
Does anyone know what are the values under the column "embarked" mean?
3 minutes ago
Comment | Promoted by Sandy Lii 2 minutes ago

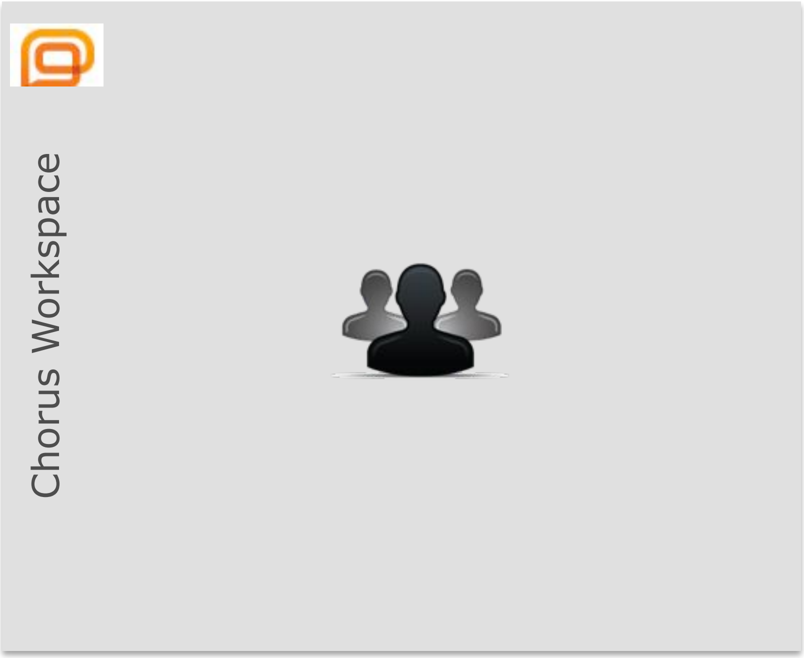
Sandy Lii commented
S = Southampton C = Cherbourg
2 minutes ago

Workspace – Streamlines Collaboration



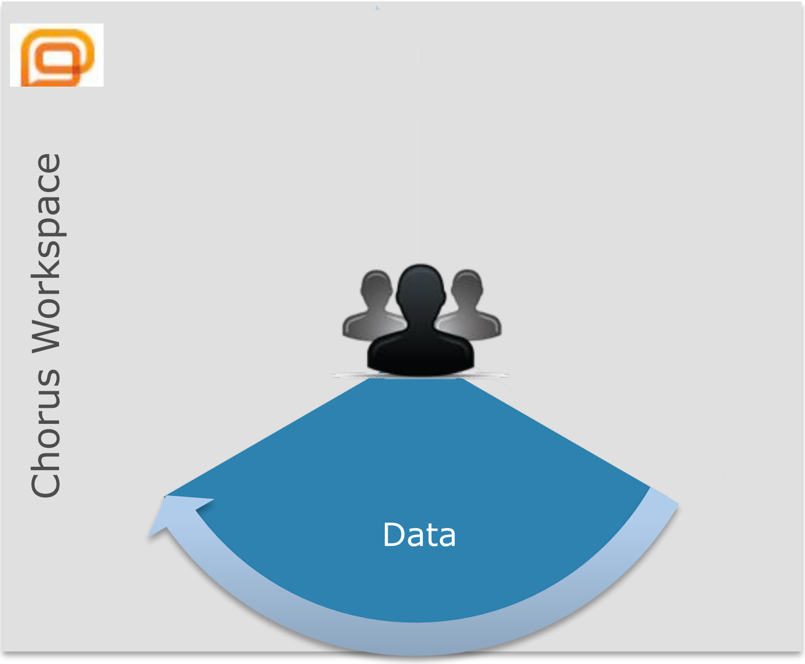
- Chorus includes unlimited workspaces, each representing individual project
- Streamlines complex user-user and user-data interactions

Multi-level Secure Collaboration



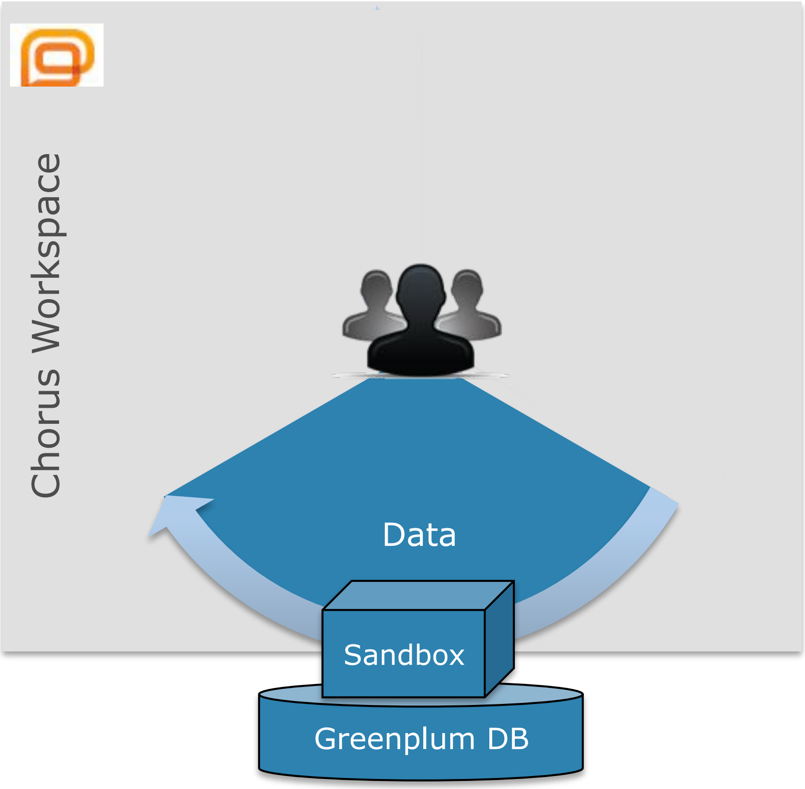
- Authentication
  - Integrates with LDAP and AD for password management
- Application access control
  - User roles: Admin vs. general user
  - Workspace types: Public or private
- Data access control
  - Chorus enforces database rules and permissions

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

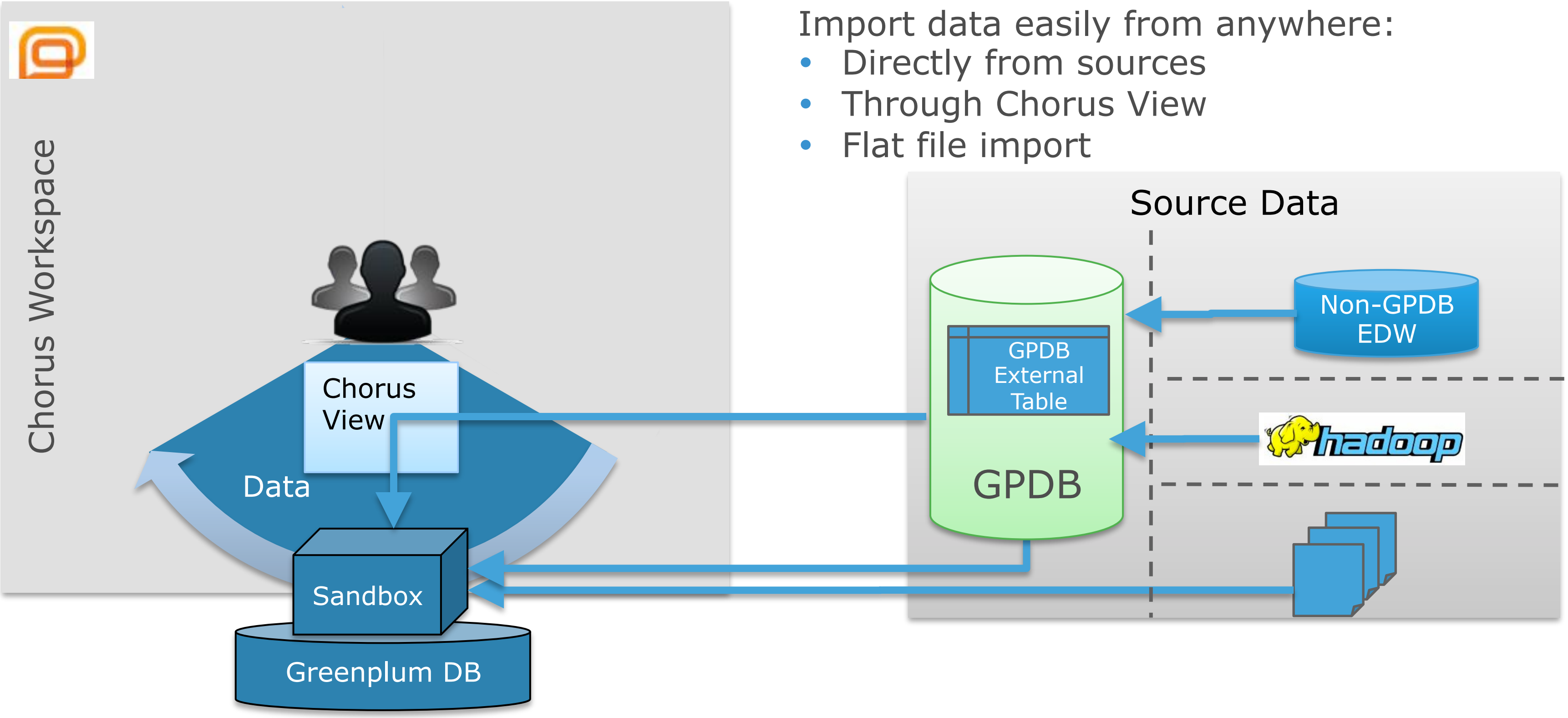


- Container of all the analytics data
- Ease of self-service provisioning of sandboxes
  - Free up IT bandwidth
  - Minimize data proliferation to uncontrolled/unmanaged data marts

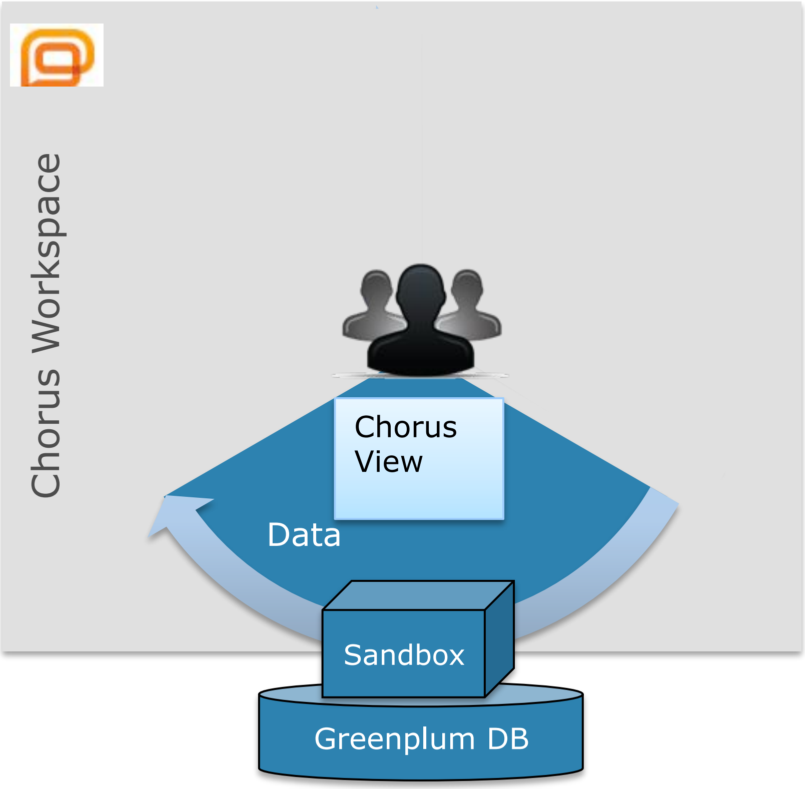
Data – Populating Sandbox

Import data easily from anywhere:

- Directly from sources
- Through Chorus View
- Flat file import

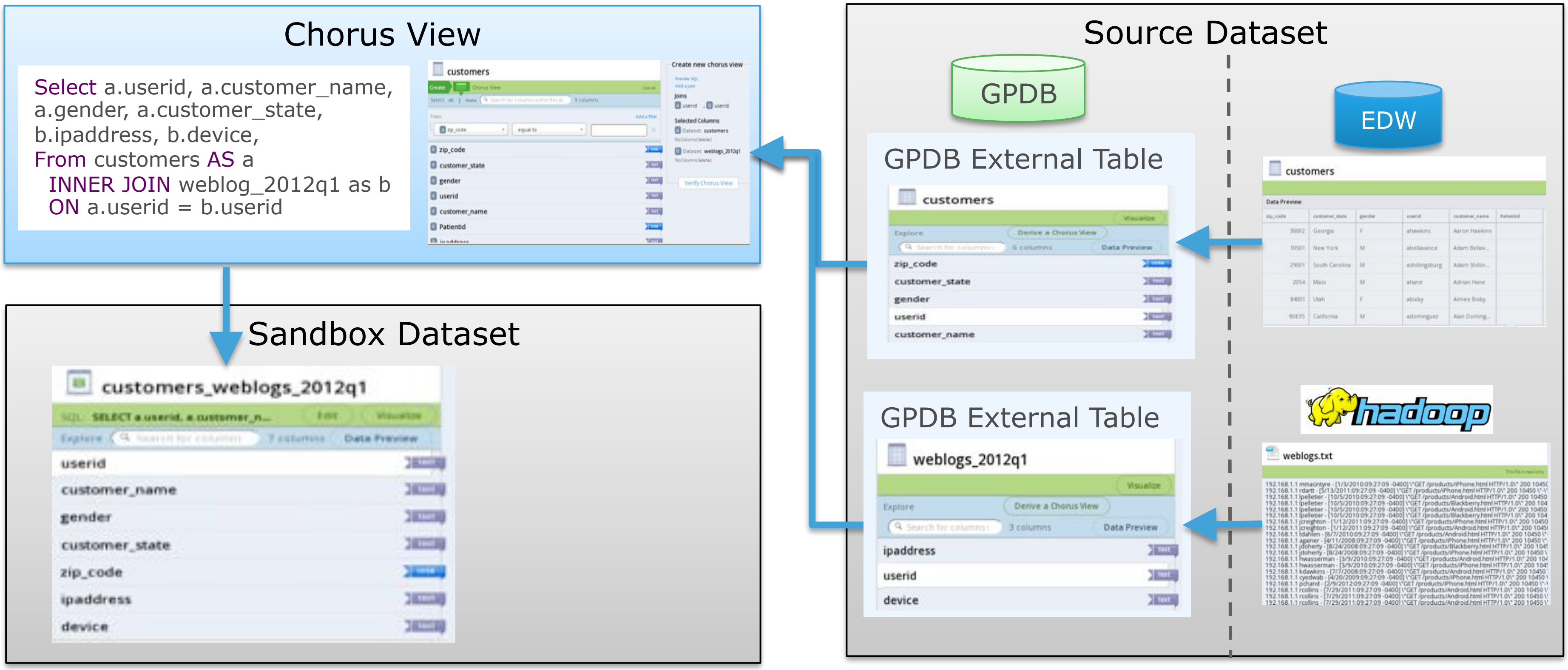


Data – Chorus View Utility



- 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
```

customers

Data Preview

zip_code	customer_state	gender	userid	customer_name	Patientid
30002	Georgia	F	ahawkins	Aaron Hawkins	
10501	New York	M	abellavance	Adam Bellav...	
29001	South Carolina	M	ashillingsburg	Adam Shilin...	
2054	Mass	M	ahane	Adrian Hane	
84001	Utah	F	abixby	Aimee Bixby	
90835	California	M	adominguez	Alan Doming...	

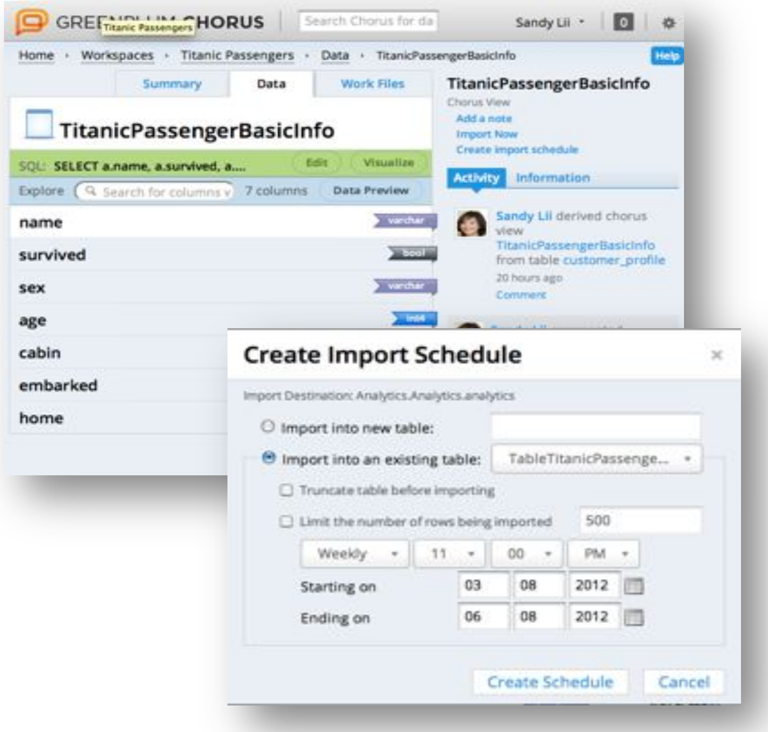
weblogs.txt

This file is read-only

```
192.168.1.1 mmacintyre - [1/3/2010:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 1045(
192.168.1.1 rdartt - [5/13/2011:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 10450 \"-\"
192.168.1.1 lpelletier - [10/5/2010:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 10450
192.168.1.1 lpelletier - [10/5/2010:09:27:09 -0400] \"GET /products/Blackberry.html HTTP/1.0\" 200 104
192.168.1.1 lpelletier - [10/5/2010:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 10450
192.168.1.1 lpelletier - [10/5/2010:09:27:09 -0400] \"GET /products/Blackberry.html HTTP/1.0\" 200 104
192.168.1.1 jcreighton - [1/12/2011:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 10450
192.168.1.1 jcreighton - [1/12/2011:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 1045
192.168.1.1 tdahlen - [6/7/2010:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 10450 \"-
192.168.1.1 againer - [4/11/2008:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 10450 \"-
192.168.1.1 jdoherty - [8/24/2008:09:27:09 -0400] \"GET /products/Blackberry.html HTTP/1.0\" 200 1045
192.168.1.1 jdoherty - [8/24/2008:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 10450 \
192.168.1.1 hwasserman - [3/9/2010:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 104
192.168.1.1 hwasserman - [3/9/2010:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 104
192.168.1.1 kdawkins - [7/7/2008:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 10450 \
192.168.1.1 cyedwab - [4/20/2009:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 10450 \
192.168.1.1 pchand - [2/9/2012:09:27:09 -0400] \"GET /products/iPhone.html HTTP/1.0\" 200 10450 \"-\
192.168.1.1 rcollins - [7/29/2011:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 10450 \
192.168.1.1 rcollins - [7/29/2011:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 10450 \
192.168.1.1 rcollins - [7/29/2011:09:27:09 -0400] \"GET /products/Android.html HTTP/1.0\" 200 10450 \
```

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



TitanicPassengerBasicInfo

SQL: SELECT a.name, a.survived, a....

name
survived
sex
age
cabin
embarked
home

TitanicPassengerBasicInfo
Chorus View

Sandy Lii derived chorus view TitanicPassengerBasicInfo from table customer_profile
20 hours ago

Create Import Schedule

Import Destination: Analytics.Analytics.analytics

○ Import into new table:

● Import into an existing table: TableTitanicPassenge...

□ Truncate table before importing

□ Limit the number of rows being imported 500

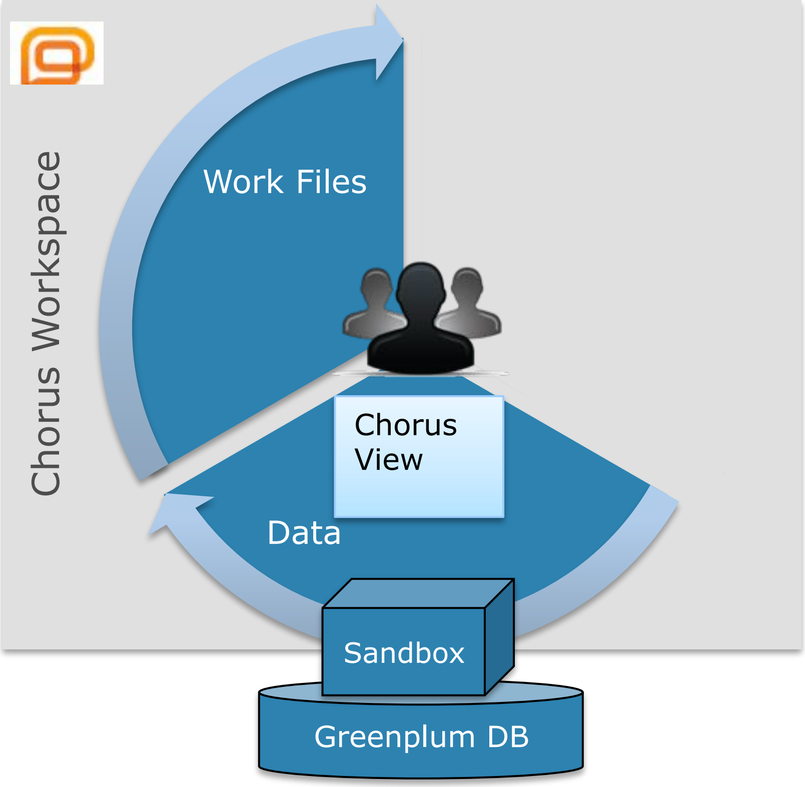
Weekly 11 00 PM

Starting on 03 08 2012

Ending on 06 08 2012

Create Schedule Cancel

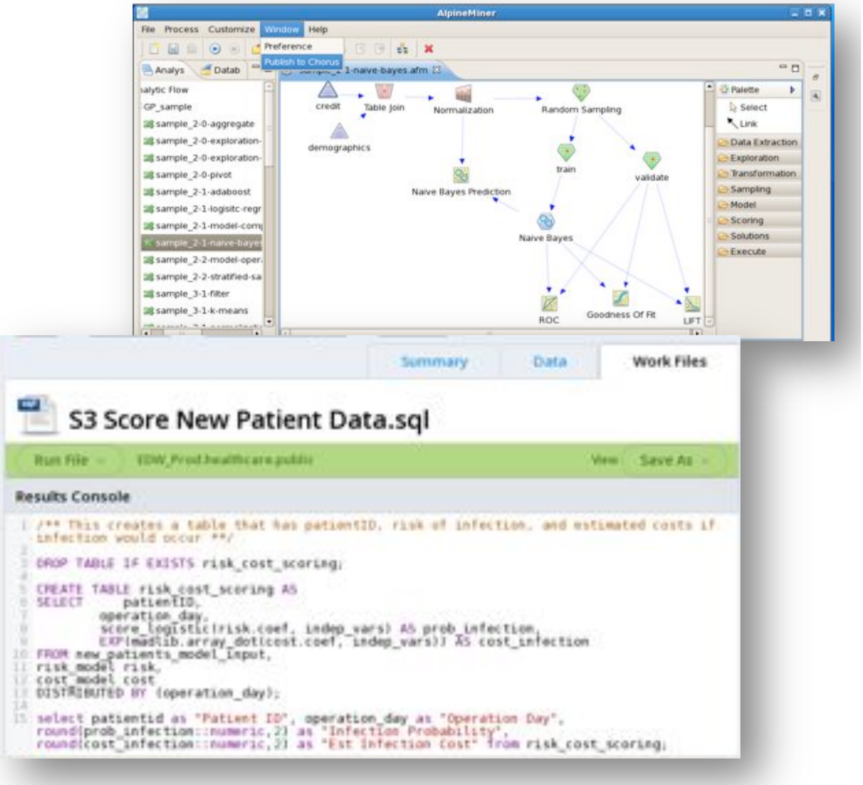
Work Files



- 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

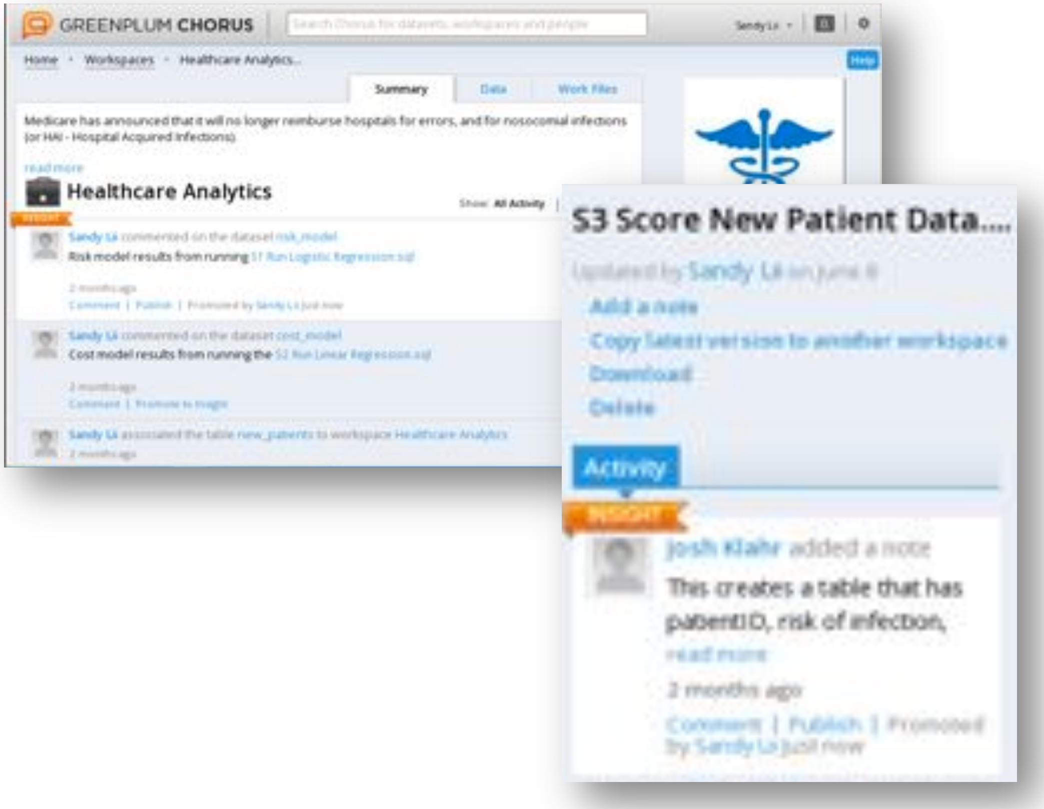


S3 Score New Patient Data.sql

```
/** This creates a table that has patientID, risk of infection, and estimated costs if
infection would occur **/

DROP TABLE IF EXISTS risk_cost_scoring;

CREATE TABLE risk_cost_scoring AS
SELECT      patientID,
          operation_day,
          score_logistic(risk.coef, indep_vars) AS prob_infection,
          EXP(madlib.array_dot(cost.coef, indep_vars)) AS cost_infection
FROM new_patients_model_input,
risk_model risk,
cost_model cost
DISTRIBUTED BY (operation_day);

select patientid as "Patient ID", operation_day as "Operation Day",
round(prob_infection::numeric,2) as "Infection Probability",
round(cost_infection::numeric,2) as "Est Infection Cost" from risk_cost_scoring;
```

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


GREENPLUM CHORUS
Search Chorus for datasets, workspaces and people
Sandy Li
Home · Workspaces · Healthcare Analytics...
Help
Summary
Data
Work Files

Medicare has announced that it will no longer reimburse hospitals for errors, and for nosocomial infections (or HAI - Hospital Acquired Infections)

read more

Healthcare Analytics

Show: All Activity

INSIGHT

Sandy Li commented on the dataset risk_model
Risk model results from running S1 Run Logistic Regression.sql
2 months ago
Comment | Publish | Promoted by Sandy Li just now

Sandy Li commented on the dataset cost_model
Cost model results from running the S2 Run Linear Regression.sql
2 months ago
Comment | Promote to Insight

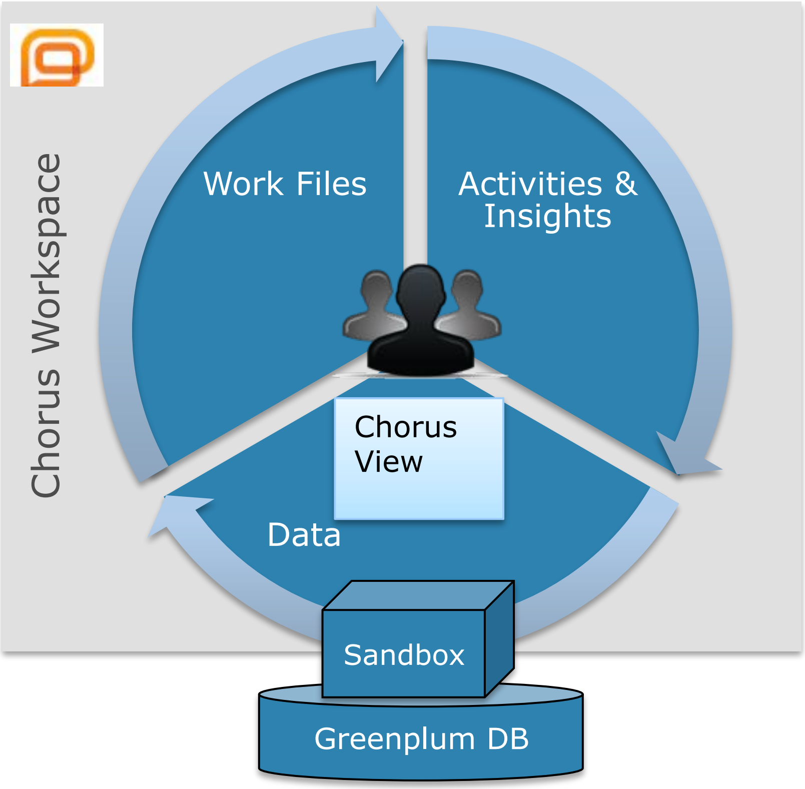
Sandy Li associated the table new_patients to workspace Healthcare Analytics
2 months ago

S3 Score New Patient Data....

Updated by Sandy Li on June 8

Josh Klahr added a note
This creates a table that has patientID, risk of infection,
read more
2 months ago
Comment | Publish | Promoted by Sandy Li just now

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

BIG DATA
THE ANSWER to agile