P/N:
Rev: A01
Updated: June 18, 2012

Greenplum Analytics Workbench

Greenplum Analytics Workbench is a large-scale cluster (1000+ nodes) whose primary purpose is running regular scale validation on Apache Hadoop releases. This facilitates Hadoop innovation. Bringing the best hardware, software, and engineers onto a single platform lets data scientists, scholars, and analysts fuel their research using unstructured data.

This guide explains how to get started on the Greenplum Analytics Workbench in the following sections:

Analytics Workbench Overview

This section outlines the components installed on Greenplum Analytics Workbench and provides a sample user scenario.

Components

By default, Analytics Workbench runs Greenplum Hadoop (GPDHD) running on Red Hat Enterprise Linux 6.1 to process unstructured data. Additional components, such as Zookeeper, Mahout, and Hive are upcoming features .

Data sets

Analytics Workbench includes data sets from a variety of sources while offering users the chance to upload data specific to their projects. Some rules for data running on Analytics Workbench:

- Loading public data sets requires approval by Greenplum.
- Data loaded on the Workbench can be shared amongst the user community and Greenplum. Shared information should never include private information such as social security numbers, credit card numbers, etc. Read the End User License Agreement (EULA) for complete details about data privacy and ownership.
- Data continues to reside in the Analytics Workbench after a project finishes. It is your responsibility to clean up any data you do not want to share after you complete your project. Should you require assistance with data cleanup, please contact the Analytics Workbench project management team.

Always check your End User License Agreement for the latest rules governing data usage on the Analytics Workbench.

Sample data sets on the Analytics Workbench include but are not limited to:

- Twitter feeds
- Wikipedia
- Data.gov sets, such as geo data.
- CNN data

Greenplum maintains a list of data available on the Analytics Workbench as part of the Analytics Workbench Portal, coming soon.

User scenario

Analytics Workbench exists to support research and drive innovation. For example, a public policy data scientist might be researching black market activities in poverty-stricken nations. Traditional structured data, such as transaction records, is difficult to find, but she has accumulated a significant amount of Twitter feeds describing locations, goods sold, prices, and other data that make up a financial transaction. The data scientist can load the Twitter feed data set into Analytics Workbench, apply k-means clustering algorithms, and begin identifying metrics such as most common black market product, top five locations for most lucrative black market activities, etc. Her finished project provides much richer detail than it could possibly have without access to the unstructured data.

As she works through her research, an engineer working on Hadoop notices the new Twitter data feed and becomes interested in the k-means clustering analytics the scientist performs. The way she uses the MapReduce function of Hadoop sparks an idea for handling large data loads made up of very small individual records. He takes his idea to the open source community, who encourages him to create the feature and incorporate it into the next release. The Analytics Workbench picks up this latest release, runs integration testing, and makes it available to the analytics community.

Prerequisites

Before you begin work on Greenplum Analytics Workbench, you must propose a project for addition to the platform. This section explains how to request Greenplum to select your project:

Discovery

Analytics Workbench is an invitation-only platform that drives research and Hadoop innovation. To receive an invitation, your project must go through the Discovery phase. During Discovery you will provide:

- Proposal outlining your project and its purpose.
- Data sets you want to use.
- Applications, other than the platform default, required.
- Estimated number of users.
- Project administrator name and contact information.

- Length of project.

Note: Greenplum grants accounts on Analytics Workbench for a finite time only.

To apply, email the information detailed above to **`analyticsworkbench@emc.com`**.

Selection Review

During Selection Review, the Analytics Workbench team reviews your proposal; some communication/iteration may be necessary. An approved project:

- Does not generate revenue either for you or Greenplum.
- Does not exceed 90 days.
- Subject to available resources/capacity.
- Does not require production-level support. There is no Support Level Agreement (SLA) between Workbench users and Greenplum.
- Displays interesting or new ways of working with unstructured data and Hadoop.

If your project meets the criteria for selection, you move into the provisioning process.

Provisioning

The provisioning process accomplishes the following tasks:

- Creates the Customer Account
- Creates the Project Administrator account
- Enables the Project Administrator to create User accounts

Greenplum creates a Customer Account when a project proposal gets approved. The Customer Account notification goes to the Project Administrator, and includes the Greenplum-specified project name, login and password. With this information, the Project Administrator may request user accounts within the Analytics Workbench portal.

Operations

The Analytics Workbench is a zero-support cluster used at your own risk. You should be familiar with SQL queries and Hadoop; many excellent tutorials exist on the web. These include:

- Linux:
  http://linux-tutorial.info/
- PostgreSQL:
  http://www.cis.temple.edu/~vasilis/Courses/CS33/Documentation/tutorial.pdf
- Apache Hadoop:
  http://hadoop.apache.org/common/docs/r0.20.2/mapred_tutorial.html

Data loading

The Analytics Workbench data loading layer allows you to load data onto HDFS or GPDB. This data layer provides a way to push / pull data from external data sources such as a network or disk drive. (You may want to use disk drives for large data sets to avoid delays due to network traffic.) The layer also serves as a data preparation stage where the Workbench rebalances the data using a proprietary algorithm. The Workbench also offers a bulk loader to pull data from designated locations at a frequency you specify.

Should you request data loads through Greenplum, note that the data structure you provide is the data structure you will see. For example, if you send a single directory with 100,000 files, that is how your data appears in HDFS.

External Data over a network

Transferring data across a network follows this workflow:

- Open a Data Load request in the Analytic Workbench Portal.
- When Greenplum receives and approves your request, we create a directory for your data.
  - If you asked for a bulk loader process to pull data automatically, the Analytic Workbench Team schedules and configures it at this point.
- The Analytic Workbench team updates your ticket; you may upload data onto the cluster.
- After you load your data, validate your load and update your ticket with the validation information.

External data loads by Greenplum

If you have limited bandwidth available and a large data set, you may choose to ship your data to Greenplum so we may load it for you. Use the table and formulae below for guidelines on when shipping data works better than transferring it over a network.

Formula

Calculate the number of days it takes to ship data over a network by:

- Multiply Megabits per second * 125 * 1000 * Network Utilization * 60 seconds * 60 minutes * 24 hours)
- Divide by total data bytes.

Note: This calculation assumes 50% network utilization to account for multiple loads and latency.

Interpreting results

Use the table below to determine if you should ship your data.

Table A.1

Internet Connection Speed	Time to transfer 10GB in days	Time to transfer 100GB in days	Time to transfer 1 TB in days	Time to transfer 2 TB in days	Ship if data set is larger than:
T1 - 1.544 Mbps	1.27954362	12.7954362	131.0252667	262.0505333	70 GB
10 Mbps	0.198841079	1.98841079	20.36132644	40.72265288	400 GB
45 Mbps	0.044186906	0.441869063	4.524739209	9.049478418	2 TB
100 Mbps	0.019884108	0.198841079	2.036132644	4.072265288	4 TB

Shipping data for loading

If you determine your dataset takes more than 7 days to transfer via a network, we require that the data be shipped to our data center. Ship your data using the following mechanism:

- Prepare portable storage device - requirements listed below
- Submit a Drive Load request to the Analytics Workbench team
- Ship your device to the address provided by the Analytics Workbench team.

Note: Customers must acquire drives and cover all shipping costs.

To have your data loaded onto the Analytics Workbench you must use a compatible storage device. Storage device requirements are:

Table A.2 Shipping form requirements

Item	Requirement
Interface type	SATA
Dimensions	3.5”
Format	EXT3 or EXT4

When we receive your package, we connect your device and attach it to our data integration layer to begin loading data. When finished. you must validate the data. When you are satisfied with your data set, we return the device. If you do not validate your data within 30 days after we load it, we return the device. EMC claims no responsibility for any damage that occurs to your device during transit. If your device arrives damaged, we will return it to the Return Address provided in the Drive Load request.

Monitoring

Analytics Workbench provides a portal so you can manage your workbench activity from a single place.

The Analytics Workbench Portal lets you perform the following tasks:

- Request data uploads and application installations.

- Request new projects.
- Report a bug.
- Track the status of the cluster.
- Refresh the view of the cluster.
- View metrics on disk usage, files and directories, jobs in progress and more.
- Track user accounts.
- View job and usage history.

The Analytics Workbench Portal is a work in progress; Greenplum will add more features as more users provide feedback on the platform.

Closing a project

You must close your project within the terms of your original time agreement; the default project lengths is 90 days. Should you require more time, contact the Analytics Workbench team with a request for extension that includes:

- Length of extension
- Extension justification
- Change in project description/scope, if necessary.

The Analytics Workbench team will try to accommodate your request as long as you have been using the system in good faith and adhere to the rules set out in the Approval section of this document.

Closed projects

Once Analytics Workbench closes your project, it suspends your accounts and archives your data. Data remains the property of EMC Greenplum and can only be deleted via special request. Should you need to re-open your project and/or re-enable your account, contact the Analytics Workbench team.