Petabyte Scale Data Warehousing with Open Source Greenplum Database

It's more than just storing and retrieving data. Equally important are loading high volume data in parallel and running analytics in the database. This hands-on session will lead you through the entire process of creating, loading, and analyzing data in the Greenplum MPP database. It's PostgreSQL, but bigger and DWH focused.

Goal:

At the end of this tutorial, attendees will learn modern DWH techniques in a PostgreSQL based Massively Parallel Processing platform.   This includes the basic architecture of the Greenplum Database, the parallel techniques for loading, querying, and analyzing structured and semi-structured data, as well as the tools Greenplum provides for doing analytics in the database.

Prerequisites:

Laptop with a modern browser and SSH client
    Instruction on using SSH on Windows
Basic knowledge of SQL
Users will connect to a cloud based Greenplum Cluster

Instructors:

Marshall Presser and others (tentative)

Audience:

This tutorial will be offered both Mon and Tue of conference week, 8 hours with lunch and other breaks. There will be a maximum of 25 attendees.

Tutorial agenda:

- Introduction to MPP and Greenplum
- Distribution -- a key to good performance in Greenplum
- Parallel loading -- loading multi Terabytes per hour

    - Loading from s3 and external connectivity
    - Polymorphic storage and external partitions - Louis
    - Compare external tables to Foreign Data Wrappers -- Andreas
- Partitioning vs. Distribution -- how they interact
    - Difference between PG and GP partitions
- Query response time -- comparisons between Greenplum and standard PostgreSQL
    - Test I/O perf on single node system and show multi-node system (trainer only)
- Running Analytics in Greenplum: MADlib exercise
    - Run on FAA data, logistic and linear regression, clustering?
    - MADlib on Postgresl -- test this
- Analyzing Free Form Text with SOLR and GPText -- lower priority
- Monitoring and Managing Greenplum with Command Center - show it
- Managing Concurrency with Resource groups and Workload Manager
    - Mention Spark and GF connector
    - Mention not Open Source
- Running PL/Python and PL/R as Trusted Languages with PL/Container -- yes

Suggested Pre work

Videos on YouTube Channel

GP Database basics -
https://www.youtube.com/watch?v=cCuGX_fLNl8&list=PL4duir3J-8GUodk1uS9ONPU_eWvfCeVjT
GP & analytics:
https://www.youtube.com/watch?v=3K1PRZNYHZE&list=PL4duir3J-8GXgVNvHVE8Y86W79Gzu5oEk
GP & MADlib
https://www.youtube.com/watch?v=Nza2F2dU-Q0&list=PL4duir3J-8GUcubGGpudx6KCCxp8onTI8