Pivotal

Petabyte Scale Data Warehousing Greenplum

Introduction

Marshall Presser
Craig Sylvester
Andreas Scherbaum
17 April 2018

Pivotal Greenplum Workshop Overview

What you will take home from this workshop

- **An understanding of the Pivotal Greenplum database**

What we assume you know already

- **SQL and maybe a little bit PostgreSQL**

What you will have in this workshop

- **Some slideware, some discussion, some hands on labs**

Where will you run the lab exercises

- **A small cloud based Greenplum instance**

Agenda

- Introduction to MPP and Greenplum
- Differences Between PostgreSQL and Greenplum
- Distribution -- a key to good performance in Greenplum
- Parallel loading -- loading multiple Terabytes per hour
- Loading from S3 and external connectivity
- Polymorphic storage and external partitions
- Partitioning vs. Distribution -- how they interact
- Query response time exercises
- Running Analytics in Greenplum: MADlib exercise
- Analyzing Free Form Text with Solr and GPText
- Running PL/Python and PL/R as Trusted Languages with PL/Container

Topics without exercises in this Workshop

- Installation
- Updates
- Backup and recovery
- Migration from other RDMS
- Hadoop and other integration
- Workload Management
- Greenplum Command Center
- Writing PXF extensions
- PostGIS
- Encryption
- Security

But first, a little history

But first, a little history

But first, a little history

Massively Parallel Processing (MPP)
Scaling out, not up.

Massively Parallel Processing (MPP)
Scaling out, not up.

Massively Parallel Processing (MPP)
Scaling out, not up.

A brief analogy

A brief analogy

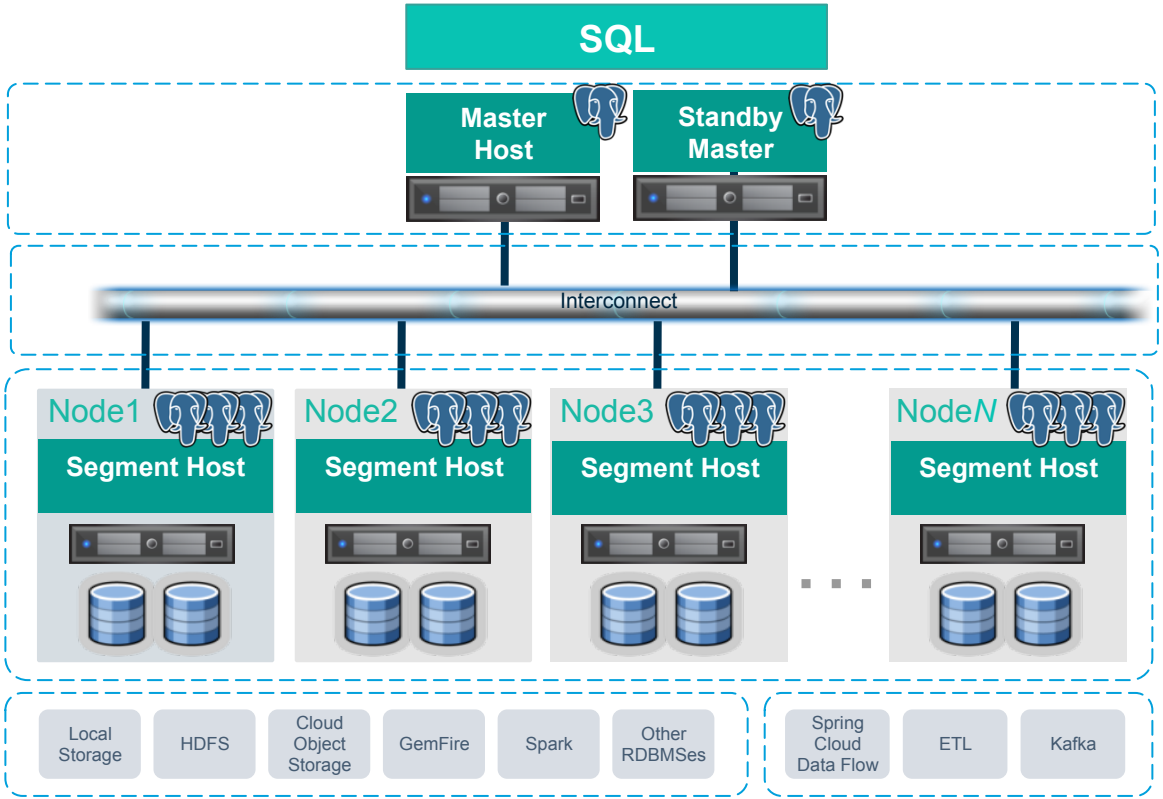
Greenplum = Massively Parallel Postgres for Analytics

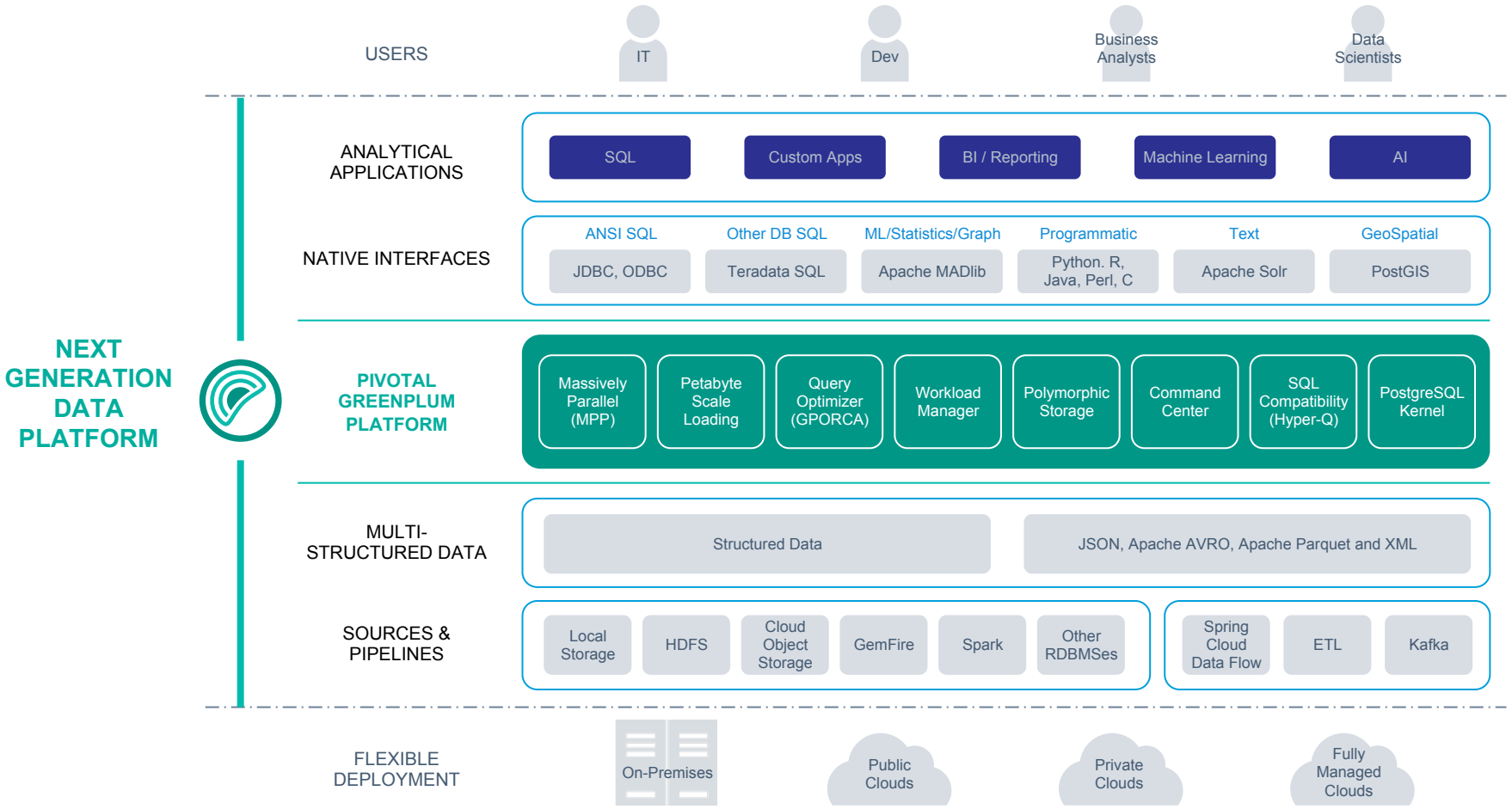
Master Servers
Query planning and dispatch

Interconnect

Segment Servers
Query processing and data storage

External Sources & Pipelines
Parallel loading and streaming




USERS
IT
Dev
Business Analysts
Data Scientists
NEXT GENERATION DATA PLATFORM
ANALYTICAL APPLICATIONS
SQL
Custom Apps
BI / Reporting
Machine Learning
AI
NATIVE INTERFACES
ANSI SQL
JDBC, ODBC
Other DB SQL
Teradata SQL
ML/Statistics/Graph
Apache MADlib
Programmatic
Python. R, Java, Perl, C
Text
Apache Solr
GeoSpatial
PostGIS
PIVOTAL GREENPLUM PLATFORM
Massively Parallel (MPP)
Petabyte Scale Loading
Query Optimizer (GPORCA)
Workload Manager
Polymorphic Storage
Command Center
SQL Compatibility (Hyper-Q)
PostgreSQL Kernel
MULTI-STRUCTURED DATA
Structured Data
JSON, Apache AVRO, Apache Parquet and XML
SOURCES & PIPELINES
Local Storage
HDFS
Cloud Object Storage
GemFire
Spark
Other RDBMSes
Spring Cloud Data Flow
ETL
Kafka
FLEXIBLE DEPLOYMENT
On-Premises
Public Clouds
Private Clouds
Fully Managed Clouds

Run your analytics anywhere you need it

Infrastructure-Agnostic

Bare-Metal

Private Cloud

vmware vSphere

openstack.

Public Cloud

amazon web services

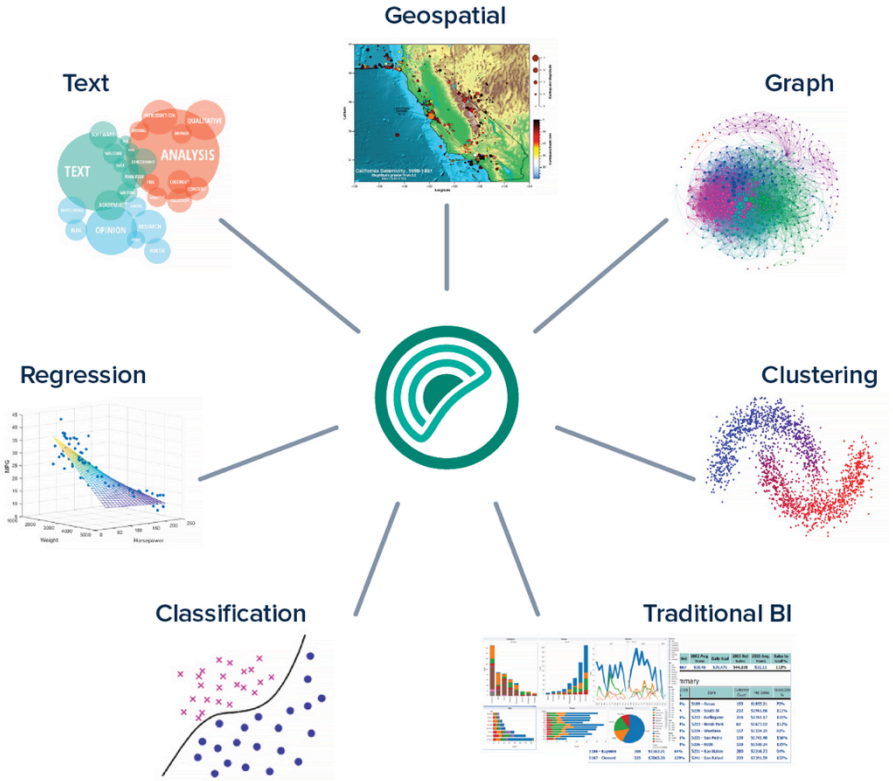
Microsoft Azure

Google Cloud Platform

- Infrastructure Agnostic: A portable, 100% software solution
- Same platform, no switching/migration cost

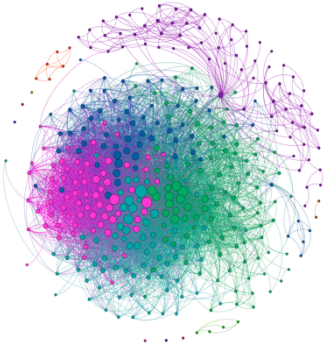
Pivotal Pivotal Greenplum

Greenplum Integrated Analytics
Traditional and Advanced In-Database Analytics



Greenplum Analytics: Graph

- Designed for very large graphs (billions of vertices/edges)
- No need to move data and transform for external graph engine
- Support for most popular graph algorithms
- Familiar SQL interface



```
SELECT madlib.pagerank(
                        'vertex',              -- Vertex table
                        'id',                  -- Vertix id column
                        'edge',                -- Edge table
                        'src=src, dest=dest',  -- Comma delimted string of edge arguments
                        'pagerank_out',        -- Output table of PageRank
                        NULL,                  -- Default damping factor (0.85)
                        NULL,                  -- Default max iters (100)
                        0.00000001,            -- Threshold
                        'user_id');            -- Grouping column name
```

Vertex Table

Vertex	Vertex Params	...
0	...	
1	...	
2	...	
3	...	

Edge Table

Source Vertex	Dest Vertex	Edge Weight	Edge Params	...
0	3	1.0	...	
1	0	5.0	...	
1	2	3.0	...	
2	3	8.0	...	
3	0	3.0	...	
3	1	2.0	...	

Greenplum Analytics: Text

Use Cases

- Communications compliance and monitoring
- Customer Sentiment analysis
- Document Search and Query
- Social Media Processing, etc.

Greenplum Technology: GPText

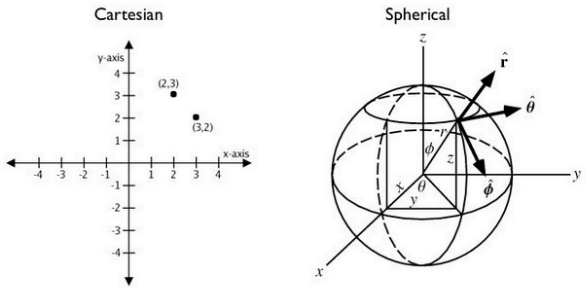
- Leverages Apache Solr and Greenplum
- Python and Java integration for Natural Language Processing
- Apache Madlib integration for machine learning on text data

Greenplum Analytics: GeoSpatial

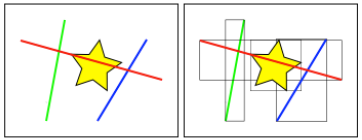
PostGIS
Geospatial Objects for PostgreSQL

- Points, Lines, Polygons, Perimeter, Area
- Intersection, Contains, Distance, Long/Latitude

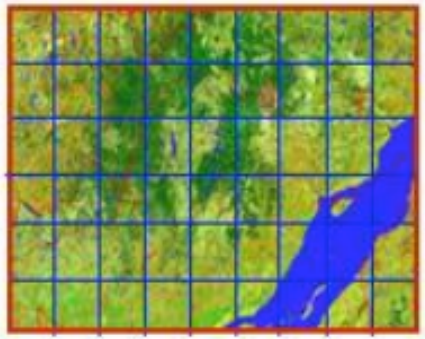
Round earth calculations



Spatial Indexes & Bounding Boxes

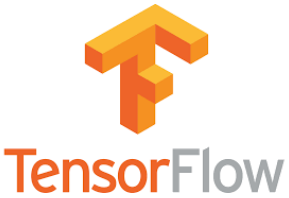


Raster Support

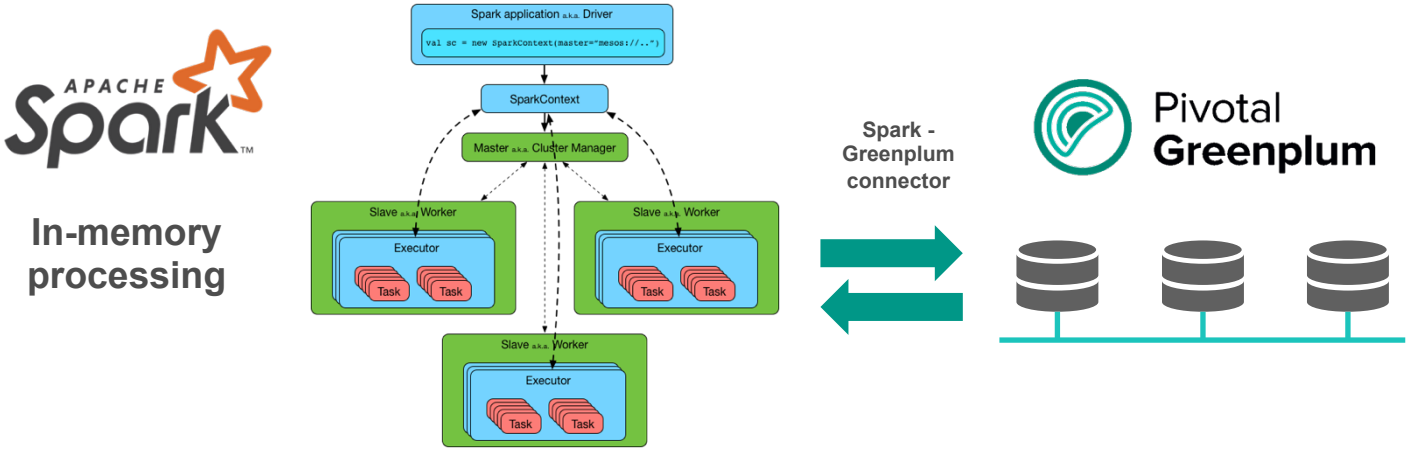


Pivotal Pivotal Greenplum

Greenplum Analytics: R and Python Libraries



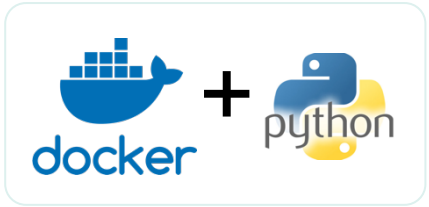
Greenplum Analytics: Spark



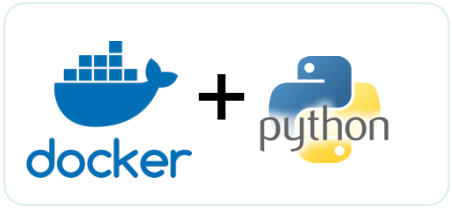
- Provide Data Access to Greenplum Data
- Leverage SPARK Skill Set of Data Scientists
- Leverage off-cluster compute resources to do computations
- Push result sets back into Greenplum for storage

Greenplum Analytics: Language Agnostic

- **Interfaces**
  - User Defined Types
  - User Defined Functions
  - User Defined Aggregates
- **Foundational work for containerized Python and R compute environments**



R/Python Containerization on Greenplum: PL/Container



- Deploy Custom R & Python Developer Environment(s) To Cluster
- Execute Functions in Isolated Secure Containers
- Deploy code and functions as non super-user
- Package any custom Python and R modules in the deployment
- Pre-configured for Data Science or customized images by users
- Multiple developer environments on same cluster

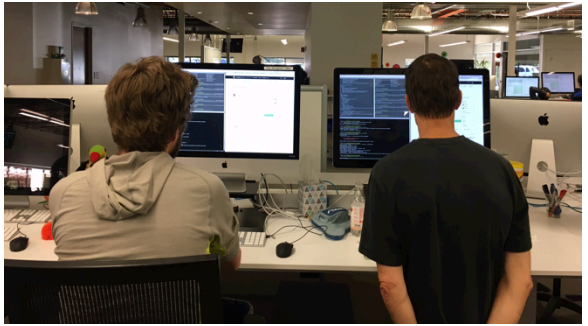
SQL Containerization: Greenplum Resource Groups

- Resource isolation for multi-tenancy and mixed analytical SQL workloads
- Enhances stability and manageability
- Leverages Linux Cgroups

Greenplum Open Source Agility = Analytical Innovation

- 100% based on PostgreSQL innovation, the most advanced analytical database for PostgreSQL users

- Agile development
- A new Pivotal Greenplum release every month



Download at Greenplum.org

Pivotal Greenplum by Pivotal - Support Offering

- Binaries and Installers
- Production Ready Releases
- 24x7 Premium Support
- Pivotal is the top Greenplum sponsor and steward worldwide
- 12 years of engineering experience with Greenplum
- Drives the Roadmap of Pivotal Greenplum
- Largest contributor and committer to open source Greenplum Database

Pivotal Greenplum Modules - Proprietary Pivotal Modules

- Greeplum Command Center: Graphical DBA Console for Centralized Management
- GPText: Apache Solr Integration for Indexing and Search of Text Data
- Gemfire Connector: Import/Export for Low Latency, High Concurrency Access
- Spark Connector: Consume Greenplum data in Spark Analytics Jobs
- Informatica Connector: ETL with industry leading solution integration
- ODBC & JDBC Drivers: Industry standard data provided by Progress Software
- QuickLZ Compression: High speed proprietary data compaction algorithm

Transforming How The World Builds Software