



Greenplum 5.0 and Roadmap

Brian Lu

Pivotal

Safe Harbor

- *“Any information regarding pre-release of Pivotal offerings, future updates or other planned modifications is subject to ongoing evaluation by Pivotal and therefore **subject to change**. This information is provided without warranty of any kind, express or implied. Customers who purchase Pivotal offerings should make their **purchase decision** based upon features that are currently available. Pivotal has no obligation to update forward looking information in this presentation.”*



Greenplum is Growing Steady

- Greenplum is Growing Steady
 - Operating in 34 countries globally
 - Customer count and revenue growing
 - Pivotal engineering investment growing
 - 9 Greenplum Database releases in 2016
 - Open source code contribution growing
 - 1417 commits to the github repo of Greenplum in 2016
 - 111 unique contributors on github repo of Greenplum in 2016
 - Major Greenplum 5.0 release planned early 2017

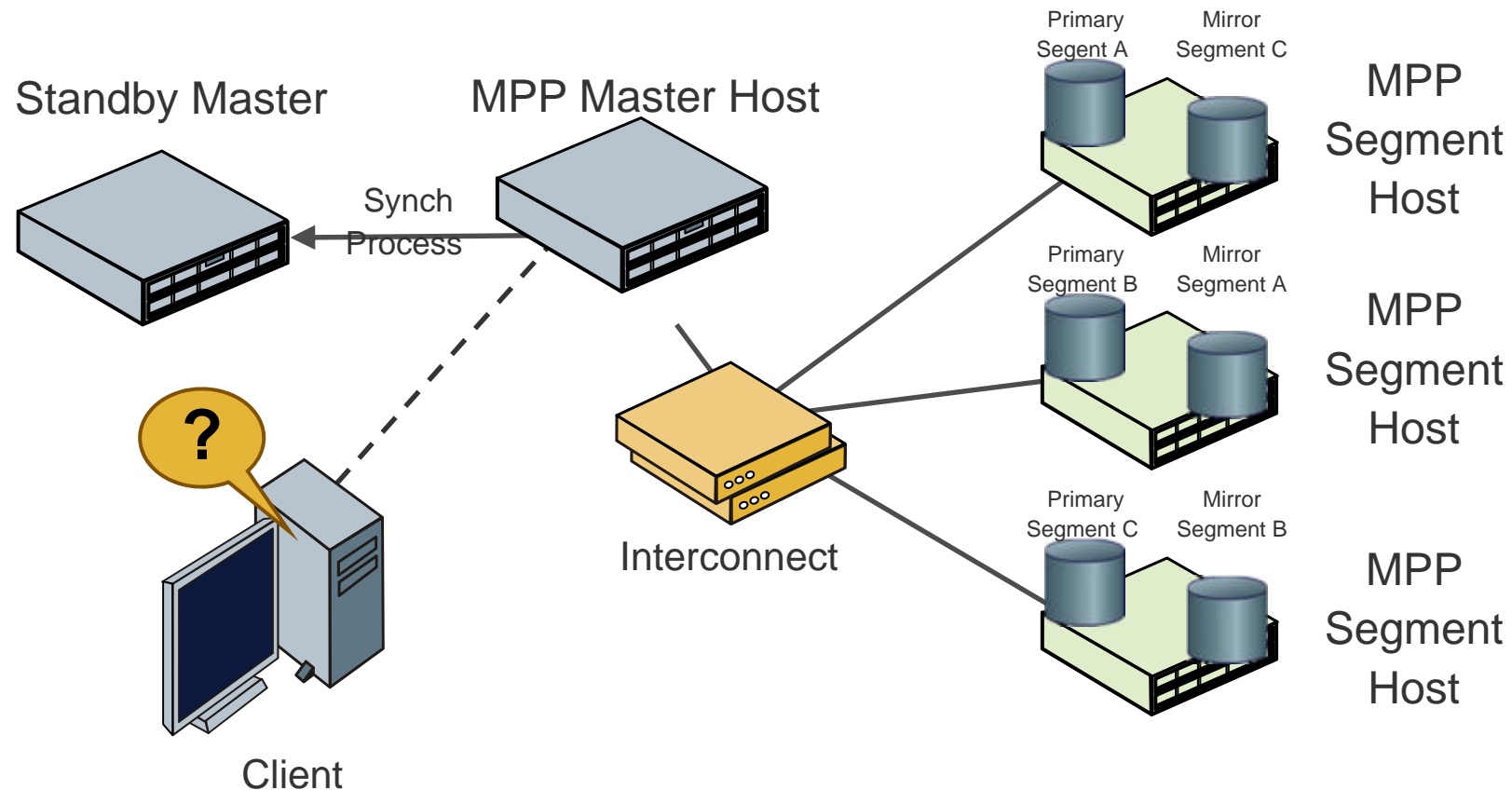


Greenplum Database Overview

- Massively Parallel Processing (MPP) database system
 - Scales out to hundreds(*) of nodes
- Shared nothing architecture
- Comprehensive SQL support with OLAP extensions
- Full ACID support
- Data distributed across nodes
 - Hashed distribution
 - Random distribution



Greenplum Database Architecture



PostgreSQL Heritage



Greenplum
Open Source
Launch



- Widely used
- Open Source
- Enterprise class relational engine



PostgreSQL Base



Vision

Greenplum in the long run will be based on latest PostgreSQL

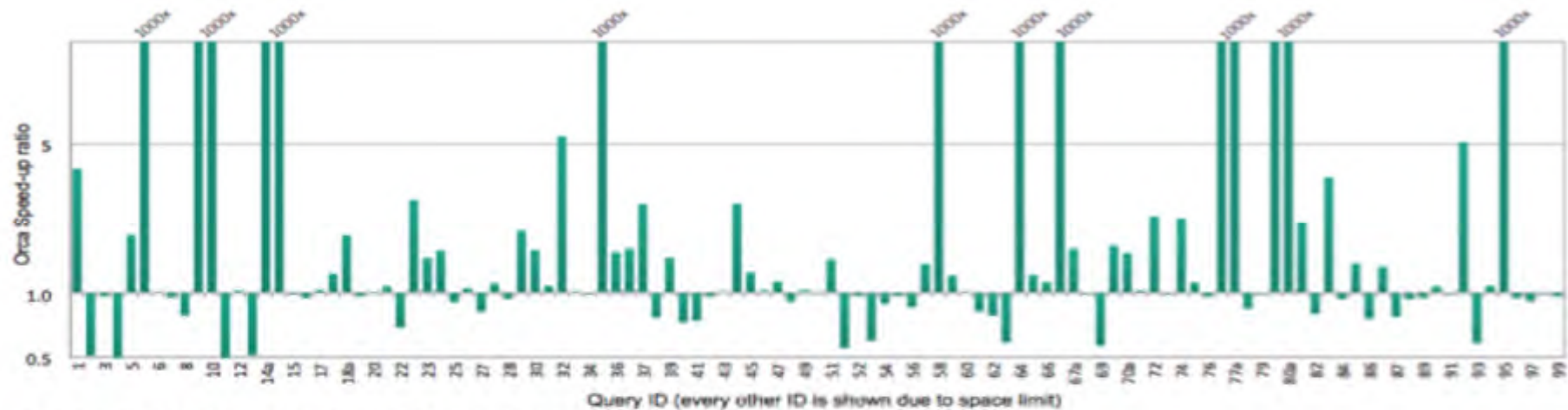
Upcoming Roadmap

- GPDB 5.0 release upgrade from PG 8.2 to PG 8.3 (2017 time frame)
- JSON/JSONB
- Full Text Search
- Improved XML Type/Functions
- UUID Type
- Raster PostGIS
- Anonymous Code Blocks
- PostgreSQL based Analyze (faster)
- Extension Framework
- Foreign Data Wrapper (FDW)



Pivotal Query Optimizer — ORCA

- First Open Source Cost Based Optimizer for BIG data
- Applies broad set of optimization strategies at once
 - Considers many more plan alternatives
 - Optimizes a wider range of queries
 - Optimizes memory usage
- New Extensible Code Base
 - Rapid adoption of emerging technologies



TPC-DS 10TB, 16 nodes, 48 GB/node



Performance: Query Optimization

Vision

Our new cost-based optimizer, Orca, will become the default optimizer in GPDB for all workloads, performing equal or better than legacy optimizer in all cases.

Current Status

Complex workloads for analytics produce large gains with ORCA

Upcoming Roadmap

- Parallelizing Union and Union All Queries
- Expanding ORCA's index support to larger class of predicates
- Reduce optimization time:
 - Auto-disable unnecessary transformations
 - Investigation: Optimization Levels



Performance: Query Execution

Vision

Dynamic Code Generation is a next gen performance enabling technology

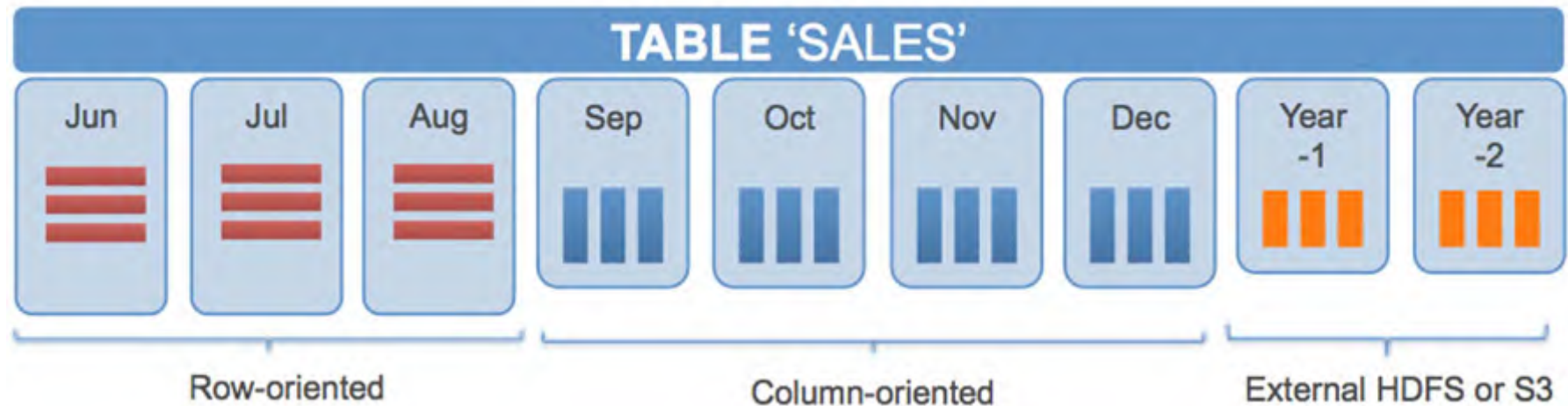
Upcoming Roadmap

- LLVM Dynamic Code Generation for faster query execution
- Dispatcher refactoring for improved performance and scalability
- More accurate query memory accounting internally
 - Optimizer and zlib memory usage accounting can be improved
- Reduce Intra-Transaction Memory
- Reduce Idle-Time Memory Usage
- Catalog data caching in the optimizer to speed short running queries



Polymorphic Storage™

User Definable Storage Layout



- Row oriented faster when returning all columns
- HEAP for many updates and deletes
- Use indexes for drill through queries

- Columnar storage compresses better
- Optimized for retrieving a subset of the columns when querying
- Compression can be set differently per column: gzip (1-9), quicklz, delta, RLE

- Less accessed partitions on external and seamlessly query all data
- All major Hadoop distributions
- Amazon S3 storage
- Others in development



External Tables

Vision

- Wide variety of data source and targets for external data querying
- Leveraging external partitions on cheap and deep storage for online archiving

Upcoming Roadmap

- S3 Writable External Tables
- Certification of GPHDFS with latest Cloudera, MapR, Hortonworks
- Porting PostgreSQL Foreign Data Wrappers to GPDB (longer term)



Storage & Backup

Vision

More '9s', and increased support for mission critical systems

Upcoming Roadmap

- PostgreSQL WAL Replication Segment Mirroring (Longer Term)
- Data Domain and NetBackup Version Upgrades
- Reduce pg_class locking during backups
- Support for all special characters in catalog names
- Discovery investigations on next-gen backup improvements





Scalable, In-Database Machine Learning

Apache MADlib (incubating): Big Data Machine Learning in SQL for Data Scientists

Open source,
commercially friendly
Apache license

Supports PostgreSQL,
Greenplum Database™,
and Apache HAWQ
(incubating)

Powerful analytics for
big data

- Open source <https://github.com/apache/incubator-madlib>
- Downloads and docs <http://madlib.incubator.apache.org/>
- Wiki <https://cwiki.apache.org/confluence/display/MADLIB/>



Madlib



Vision

In database analytics, machine learning and data science tools

Upcoming Roadmap

- Pivotal R Support for SVM and LDA
- Grouping Support and Cross Validation in Elastic Net
- Improved Python Language Support
- Investigation on Graph Support
- Investigation on GPU support
- Performance improvements



GPDB Geospatial



Current Key Features:

- Points, Lines, Polygons, Perimeter, Area, Intersection, Contains, Distance

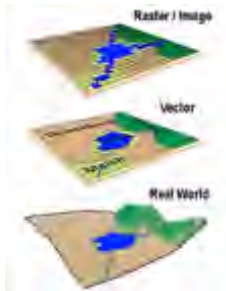
Ability to store geospatial data and query with joins and operators

```
geodemo=# SELECT
nyc_subway_stations.long_name AS subway,
nyc_neighborhoods.name AS neighborhood
FROM nyc_neighborhoods
JOIN nyc_subway_stations
ON ST_Contains(nyc_neighborhoods.geom, nyc_subway_stations.geom)
WHERE nyc_neighborhoods.name = 'Greenwich Village';
```

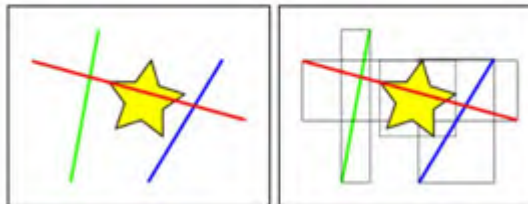
subway	neighborhood
W 4th St (B,D,F,V) Manhattan	Greenwich Village
14th St / Union Sq (4,5,6) Manhattan	Greenwich Village
14th St (1,2,3) Manhattan	Greenwich Village
Bleecker St / Broadway-Lafayette St (6) Manhattan	Greenwich Village
Christopher St / Sheridan Sq (1) Manhattan	Greenwich Village
Union Sq / 14th St (L,N,O,R,W) Manhattan	Greenwich Village
6th Ave / 14th St (F,L,V) Manhattan	Greenwich Village
8th St / New York University (N,R,W) Manhattan	Greenwich Village
Astor Pl (6) Manhattan	Greenwich Village
W 4th St (A,C,E) Manhattan	Greenwich Village

(10 rows)

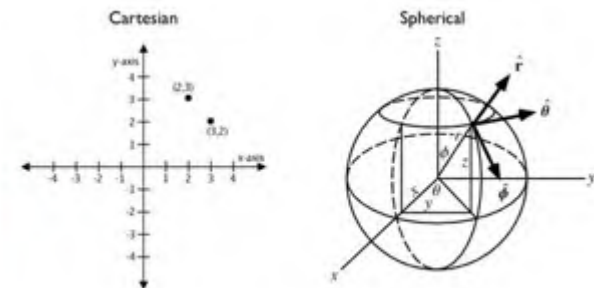
Raster
Image
Processing



Spatial Indexes & Bounding Boxes



Round earth calculations



GP Text: Full Text Search and Text Analysis (proprietary)



Vision

Integrated Data Warehouse for SQL and Text Search in one system leveraging Apache Solr and GPDB

Upcoming Roadmap

- GPText 2.0 GA
- Solr Cloud based high availability and Solr mirroring
- Gptext-recover
- Gptext-expand
- Gptext-backup
- Gptext-restore previous index



Command Center (proprietary)

Vision

Clean and Rich Graphical User Interface for GPDB DBAs

Current Status

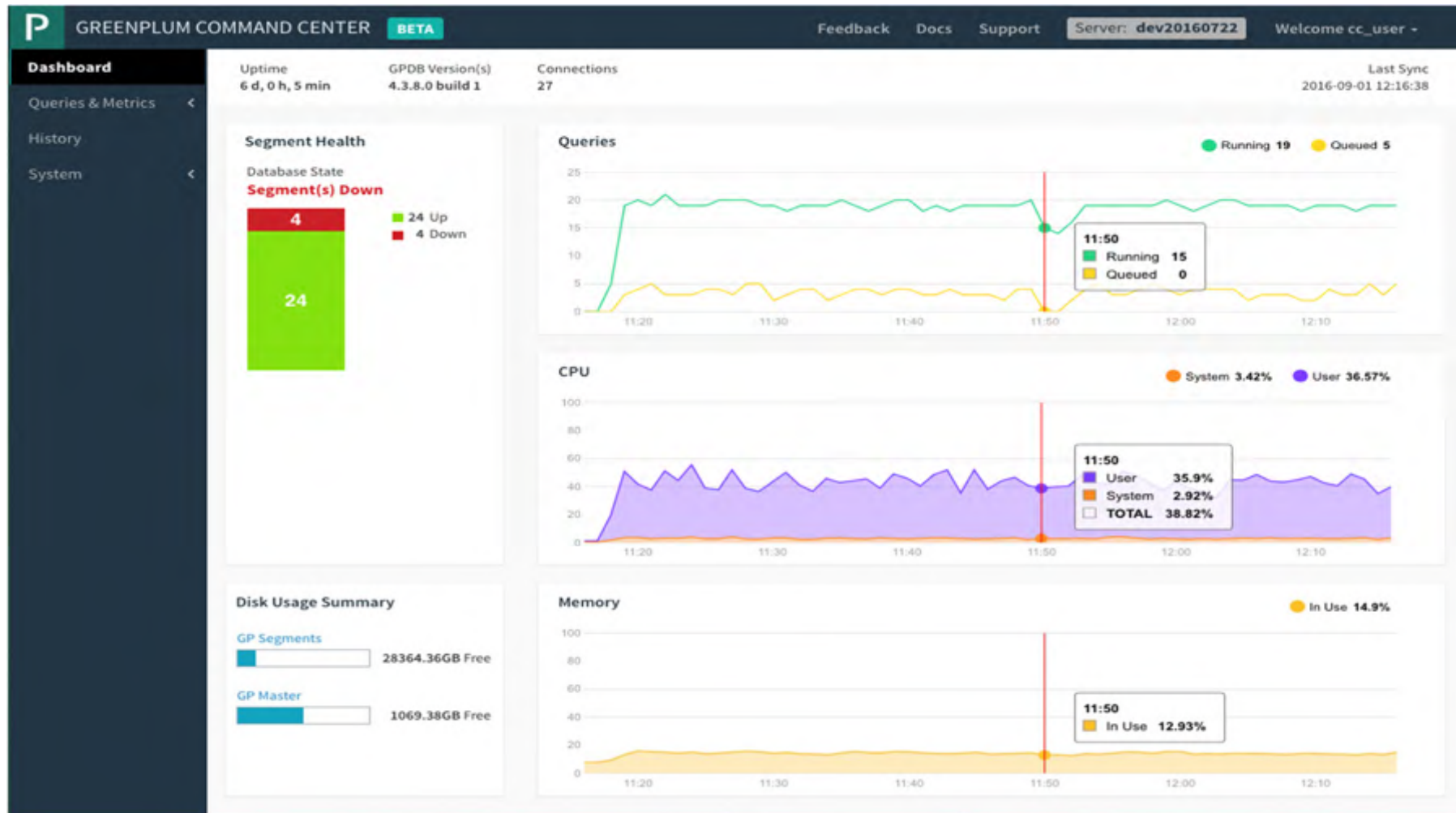
HTML5 rewrite nearly at parity with original Flash-based GUI

Upcoming Roadmap

- Richer history
- Integration with Greenplum Workload Manager for Graphical control of WLM



GPCC New UI Glance



Greenplum Workload Manager

Rule based query management to monitor and manage queries and resource queues

- Monitors Greenplum Database queries and host utilization statistics
- Logs when a query exceeds a threshold
- Throttles the CPU usage of a query when it exceeds a threshold
- Terminates a query
- Detects memory, CPU, or disk I/O skew occurring during the execution of a query
- Creates detailed rules to manage queries
- Adds, modified, or deletes Greenplum Database resource queues



Workload Management (proprietary)

Vision

Hands off DBA policy management of multi-user environment

Upcoming Roadmap

- Fine grain configuration management through centralized console
- IDLE session termination ability
- SuSE support
- Multi-action rules (starting with terminate and log)
- Better event reporting and rollup views
- Deeper resource queue integration (still in design)



G2C (Greenplum Gemfire Connector) (proprietary)



Vision

Bring the real-time and high concurrency feature of Gemfire together with the full SQL analytics and reporting of Greenplum into an “Operational Data Warehouse” solution that combines the benefits of both

Upcoming Roadmap

- GA of Gemfire driven Java class library for IMPORT/EXPORT operations to/from Gemfire and GPDB
- Greenplum based External Tables to provide READ/WRITE to Gemfire



GPDB Pivotal Cloud Foundry (PCF) Tile (proprietary)

Vision

Bring GPDB to the Pivotal Cloud Foundry ecosystem with a smooth deploy and provisioning experience.

Upcoming Roadmap

- Single Node Non-Production Release
- Incrementally improve Day 2 operations
- Incorporate Single Node feedback into multi-node



PL/Container (proprietary)

Vision

Containerized execution of Python and R (PL/Python and PL/R) providing a security model and an isolated environment to install the interpreter and any dependent libraries independent of the Database and DBA environment

Upcoming Roadmap

- Docker based containers
- Features which improve usability



Open vs. Closed

- Open: Core database components, GPDB, ORCA
- Closed
 - 3rd party components, eg: compression
 - PI/Containers
 - G2C (Greenplum Gemfire connector)
 - WLM
 - GPCC
 - GPText



Greenplum Community

- Since open source from 2015/10/27
 - GPDB: 1652 stars, 472 fork, 299 watch
- Contributions
 - Pull Request(PR): 31 open, 987 closed within 12 months
 - External contributions from China: China Mobile, Alibaba, Huawei, ...
- User groups without any advertisement
 - wechat group: 436



Thanks!

Q & A