

Agenda

- What is Greenplum? Who is using it?
- History: Where does it come from!
- How does it work!
- Where to get it!

Time Management – Everything is easy

45 Minutos → 15 Minutos

Sin demonstración

Andreas Scherbaum

- Working with PostgreSQL since 1997
- Co-founded PostgreSQL Europe in 2008, Board of Directors
- Organizing pgconf.de, pgconf.eu
- Joined EMC in 2011 (Big Data Division)
- Moved to Pivotal, Principal Software Engineer
  - working on Greenplum & PostgreSQL
- Twitter: @ascherbaum




What is Greenplum?

What is Greenplum?

- Massively Parallel (MPP)
- Shared Nothing
- Database

Massively Parallel Processing (MPP)

- Scales out from 1 single system to dozens, or possibly 100s of servers
- Everything is one big database
- The software handles the data distribution, and the query planning and execution

Little bit of history …



Little bit of history …





Little bit of history …

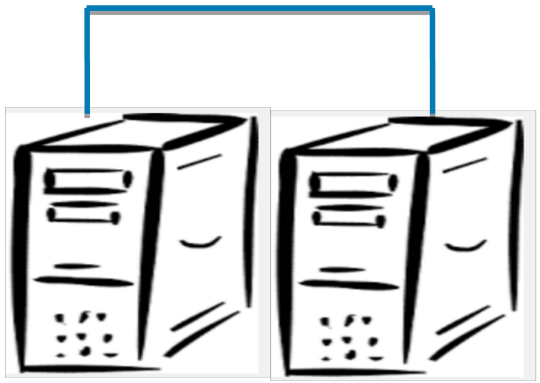






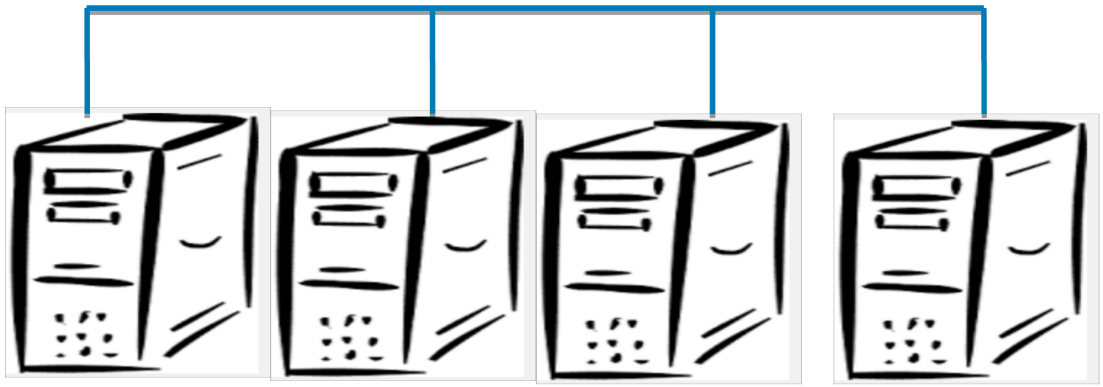
Massive Parallel Processing (MPP)

Scaling out, not up



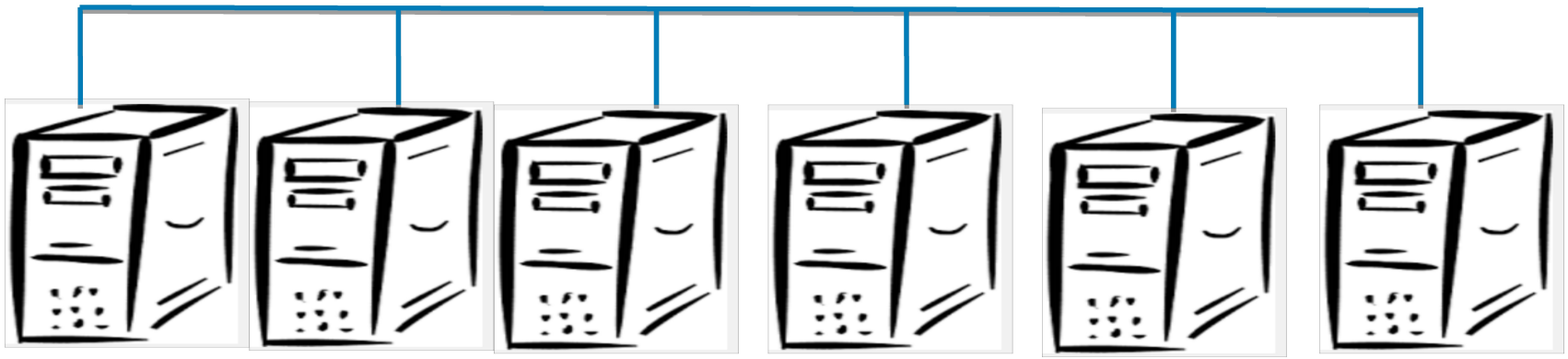
Massive Parallel Processing (MPP)

Scaling out, not up



Massive Parallel Processing (MPP)

Scaling out, not up



A brief analogy





A brief analogy





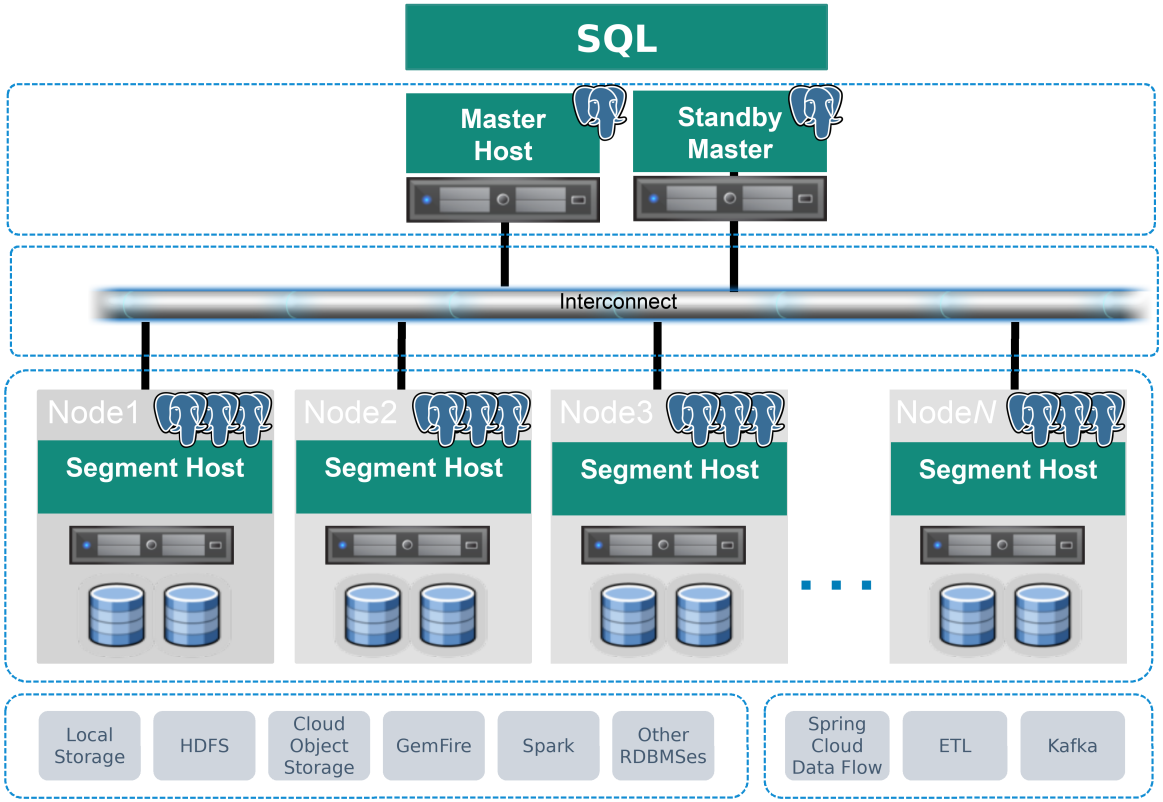
Greenplum = MPP for Analytics

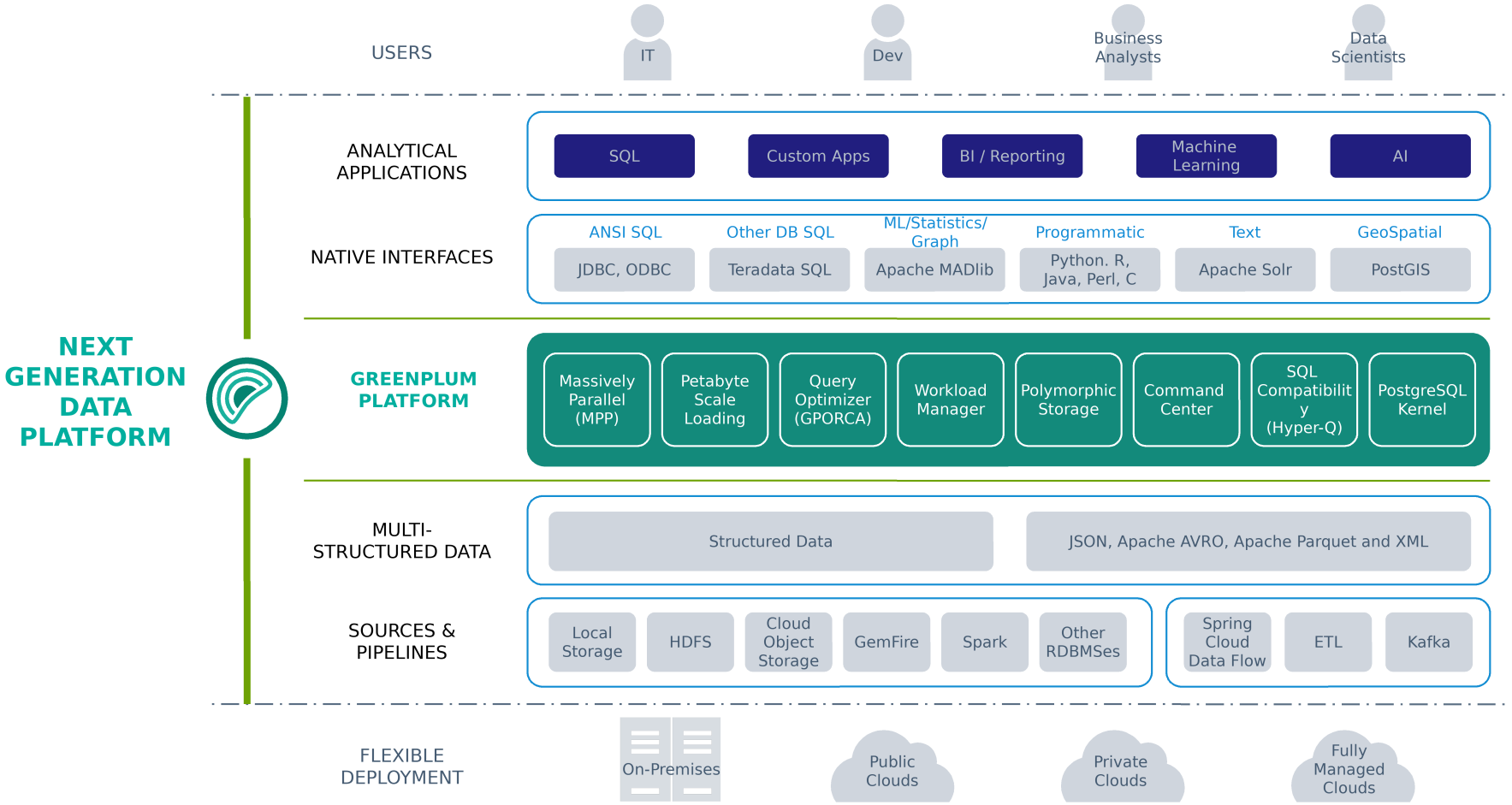
Master Servers
Query planning and dispatch

Interconnect

Segment Servers
Query processing and data storage

External Sources & Pipelines
Parallel loading and streaming




USERS
IT
Dev
Business Analysts
Data Scientists
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
NEXT GENERATION DATA PLATFORM
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

Shared Nothing

- Servers do not share infrastructure (like disks, or CPU)
- Every server is autonomous
- All servers are connected through a high-speed network (Interconnect)
- Data redistribution is handled by the software

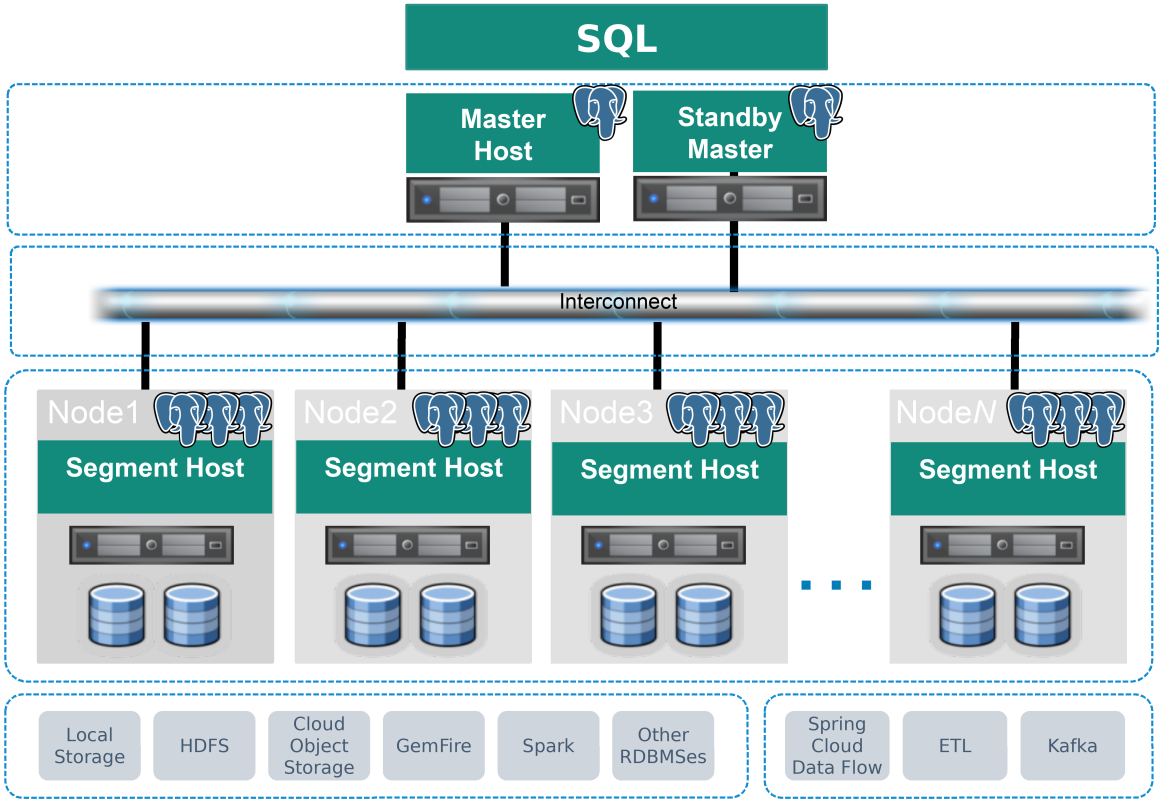
Greenplum = MPP for Analytics

Master Servers
Query planning and dispatch

Interconnect

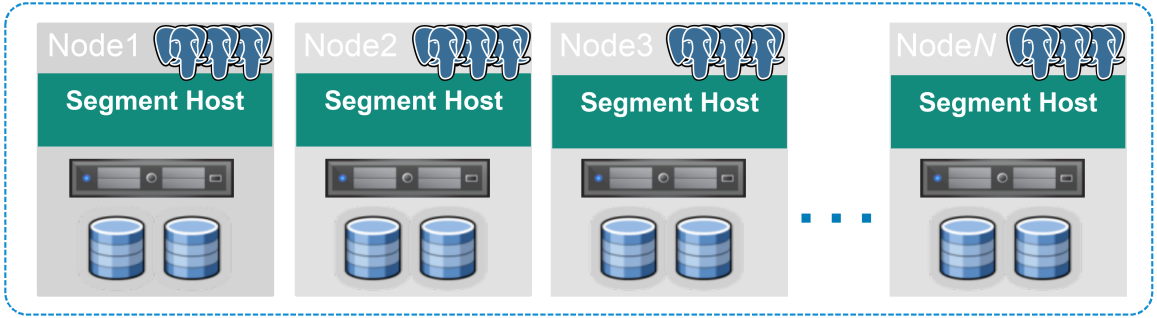
Segment Servers
Query processing and data storage

External Sources & Pipelines
Parallel loading and streaming



Shared Nothing

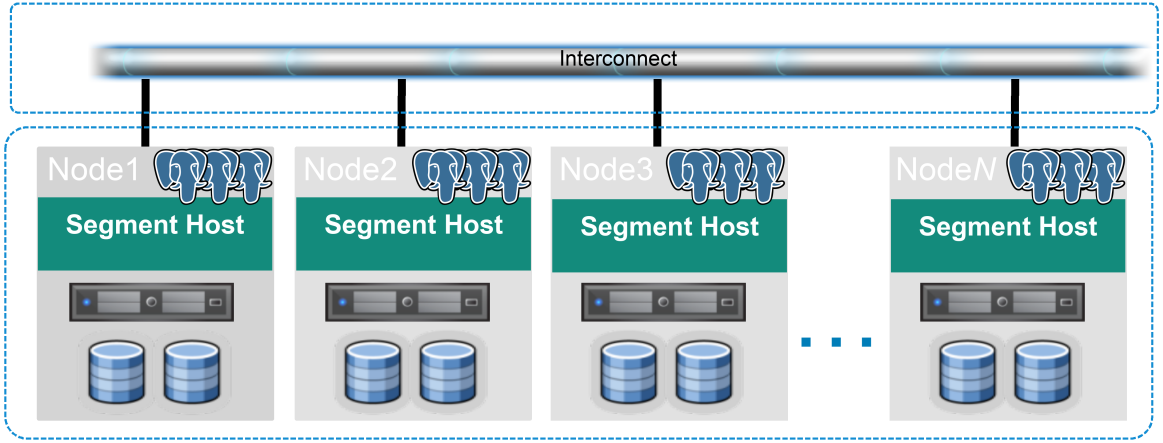
Segment Servers
Query processing and data storage



Shared Nothing

Interconnect

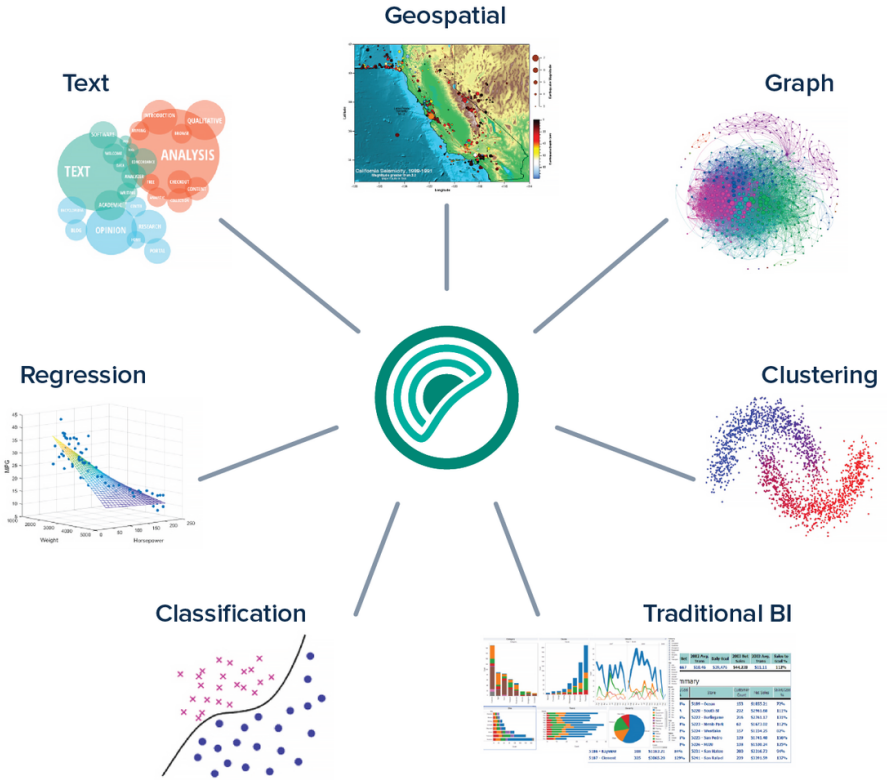
Segment Servers
Query processing and data storage




Who is using it?

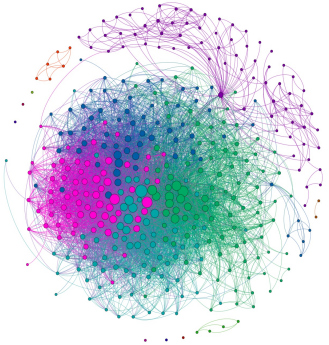
Greenplum Integrated Analytics

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 e
- 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.



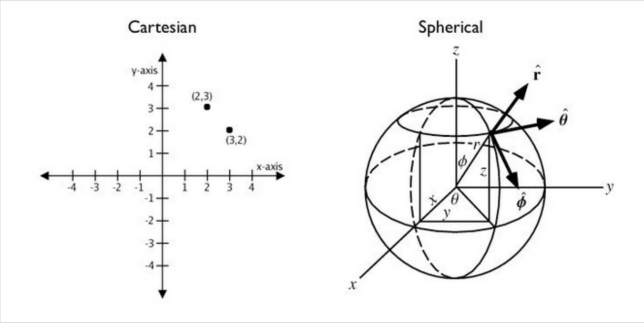
Greenplum Technology: GPText

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

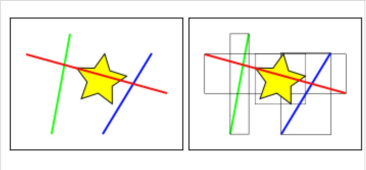
Greenplum Analytics: GeoSpatial

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

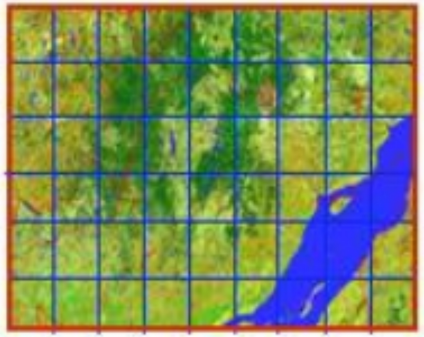
Round earth calculations



Spatial Indexes & Bounding Boxes



Raster Support



Greenplum Analytics: R & Python Libraries



















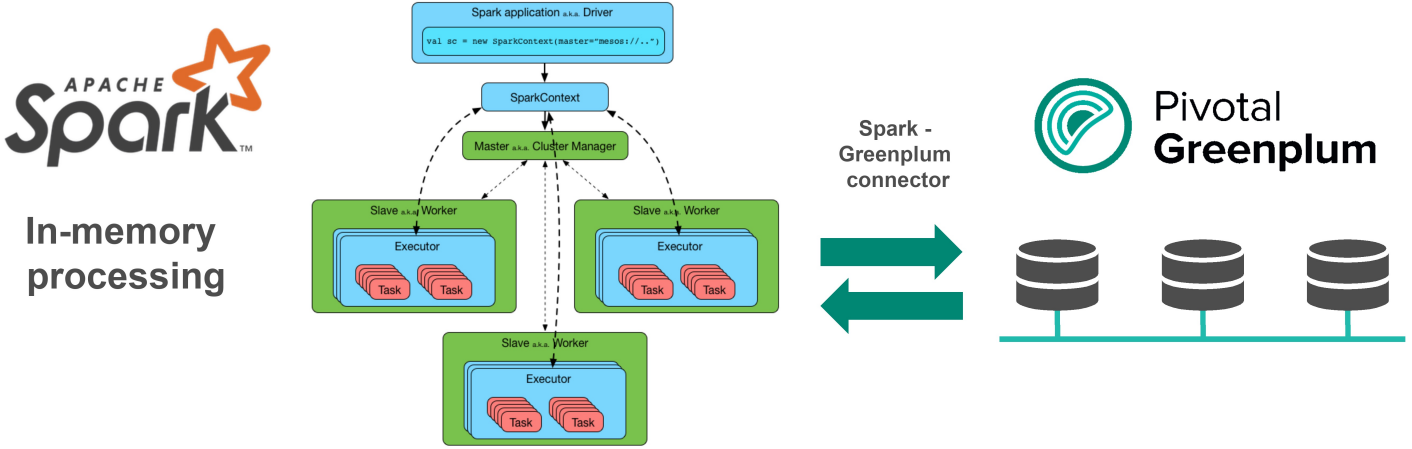








Greenplum Analytics: Spark



- 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







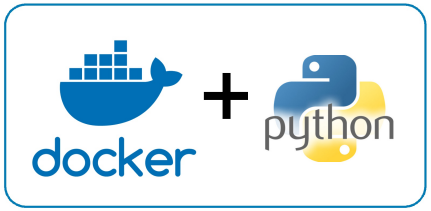


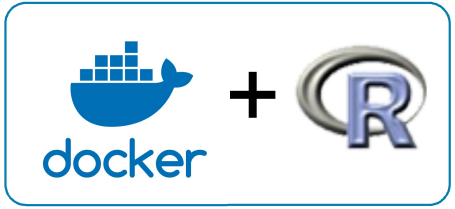




- **Interfaces**
  - User Defined Types
  - User Defined Functions
  - User Defined Aggregates

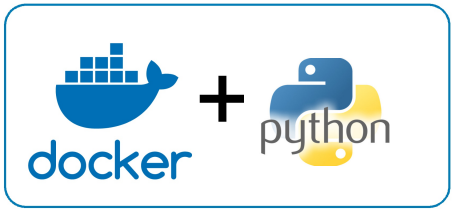
- **Foundational work for containerized Python and R compute environments**

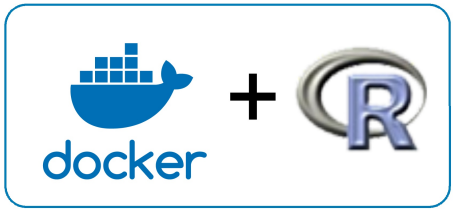




R/Python Containerization on Greenplum

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

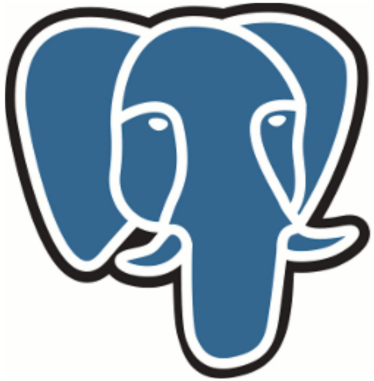



Where does it come from?

Where does in come from?

- Greenplum is based on an Open Source database named “PostgreSQL”





PostgreSQL

the world's most advanced open source database

PostgreSQL starts in Berkeley

- 1985: Postgres in born at Berkeley University
- 1985: First registered domain name
- 1989: Postgres 2.0 (still no SQL)
- 1989: First commercial ISP
- 1990: First website at CERN
- 1994: Netscape browser, Yahoo! born
- 1995: Postgres95
- 1997: PostgreSQL
- 1997: Netflix founded (renting DVD by mail)
- 1998: Google founded, IPv6 introduced
- 2019: PostgreSQL v12 (31st major release)



Greenplum starts in Los Angeles

- 2003: Metapa acquires Didera, founds Greenplum
- 2003: Myspace and Skype founded
- 2005: Greenplum releases “Bizgres”
- 2005: YouTube launches
- 2010: Acquired by EMC
- 2010: Digital Storage reached the “Zetta”
- 2015: Greenplum is Open Source
- 2015: IoT
- 2017: Greenplum Database v5
- 2019: Greenplum Database v6



Where does it go?

- Greenplum is based on an Open Source database named “PostgreSQL”
- For a long time based on PostgreSQL 8.2.x (Fork)
- Merge with newer versions is in progress
  - GP 5.x based on PG 8.3
  - current source is based on PG 9.6
- Code and features flow into both directions


How does it work?

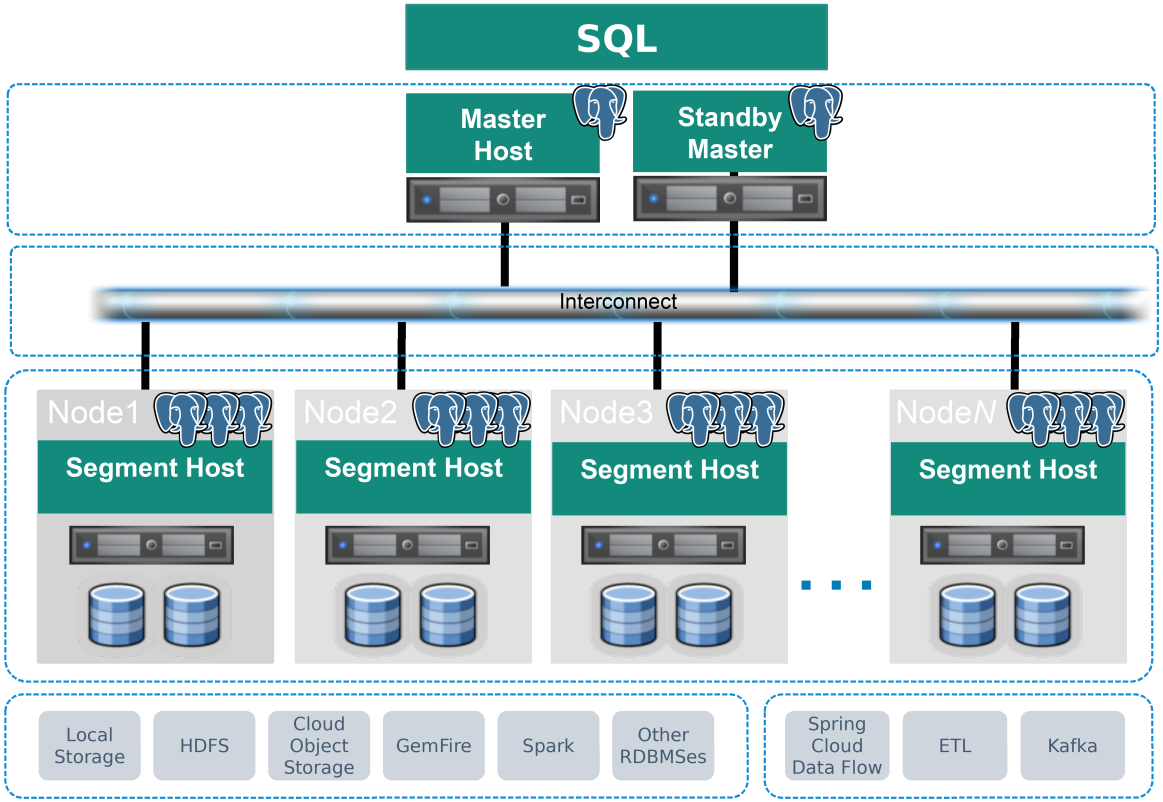
Greenplum = MPP for Analytics

Master Servers
Query planning and dispatch

Interconnect

Segment Servers
Query processing and data storage

External Sources & Pipelines
Parallel loading and streaming



First: software installation

- The Greenplum software is installed once per host

```
[gpadmin@dell-emc-world ~]$ ls -ld /usr/local/greenplum-db/*
drwxr-xr-x  7 gpadmin wheel   4096 14. Aug 2017  /usr/local/greenplum-db/bin
drwxr-xr-x  2 gpadmin wheel     32 14. Aug 2017  /usr/local/greenplum-db/demo
drwxr-xr-x  5 gpadmin wheel     52 14. Aug 2017  /usr/local/greenplum-db/docs
drwxr-xr-x  2 gpadmin wheel     44 14. Aug 2017  /usr/local/greenplum-db/etc
drwxr-xr-x  3 gpadmin wheel     20 14. Aug 2017  /usr/local/greenplum-db/ext
-rwxr-xr-x  1 gpadmin wheel  43025 14. Aug 2017  /usr/local/greenplum-db/GPDB-LICENSE.txt
-rw-r--r--  1 gpadmin wheel    786 28. Apr 01:06 /usr/local/greenplum-db/greenplum_path.sh
drwxr-xr-x  6 gpadmin wheel   4096 14. Aug 2017  /usr/local/greenplum-db/include
drwxr-xr-x 11 gpadmin wheel   8192 14. Aug 2017  /usr/local/greenplum-db/lib
-rwxr-xr-x  1 gpadmin wheel 192912 14. Aug 2017  /usr/local/greenplum-db/LICENSE.thirdparty
drwxr-xr-x  2 gpadmin wheel   4096 14. Aug 2017  /usr/local/greenplum-db/sbin
drwxr-xr-x  4 gpadmin wheel     35 14. Aug 2017  /usr/local/greenplum-db/share
[gpadmin@dell-emc-world ~]$
```

Second: Initialize a number segments + a master

- The data directories are separate PostgreSQL-like installations
- Every master and segment is hosted in it’s own directory
  - Can reside on different file systems

```
[gpadmin@dell-emc-world ~]$ ls -ld /data/*
drwxrwx--- 3 gpadmin gpadmin 21 28. Apr 01:04 /data/master
drwxrwx--- 3 gpadmin gpadmin 20 28. Apr 01:05 /data/seg1
drwxrwx--- 3 gpadmin gpadmin 20 28. Apr 01:05 /data/seg2
drwxrwx--- 3 gpadmin gpadmin 20 28. Apr 01:05 /data/seg3
drwxrwx--- 3 gpadmin gpadmin 20 28. Apr 01:05 /data/seg4
```

Second: Initialize a number segments + a master

```
drwx------ 17 gpadmin wheel 4096 28. Apr 01:06 /data/master/gpsne-1
[gpadmin@dell-emc-world ~]$ ls -ld /data/master/gpsne-1/*
drwx------ 6 gpadmin wheel    54 28. Apr 01:06 /data/master/gpsne-1/base
drwx------ 2 gpadmin wheel  4096 28. Apr 01:06 /data/master/gpsne-1/global
-r-------- 1 gpadmin wheel   109 28. Apr 01:05 /data/master/gpsne-1/gp_dbid
drwxr-xr-x 5 gpadmin wheel    42 28. Apr 01:04 /data/master/gpsne-1/gpperfmon
-rw-r--r-- 1 gpadmin wheel   860 28. Apr 01:04 /data/master/gpsne-1/gpssh.conf
drwx------ 2 gpadmin wheel     6 28. Apr 01:04 /data/master/gpsne-1/pg_changetracking
drwx------ 2 gpadmin wheel    18 28. Apr 01:04 /data/master/gpsne-1/pg_clog
drwx------ 2 gpadmin wheel    18 28. Apr 01:04 /data/master/gpsne-1/pg_distributedlog
drwx------ 2 gpadmin wheel     6 28. Apr 01:04 /data/master/gpsne-1/pg_distributedxidmap
-rw-r--r-- 1 gpadmin wheel  4101 28. Apr 01:04 /data/master/gpsne-1/pg_hba.conf
-rw------- 1 gpadmin wheel  1636 28. Apr 01:04 /data/master/gpsne-1/pg_ident.conf
drwx------ 2 gpadmin wheel   141 28. Apr 01:06 /data/master/gpsne-1/pg_log
drwx------ 4 gpadmin wheel    36 28. Apr 01:04 /data/master/gpsne-1/pg_multixact
drwx------ 2 gpadmin wheel    25 28. Apr 01:06 /data/master/gpsne-1/pg_stat_tmp
drwx------ 2 gpadmin wheel    18 28. Apr 01:04 /data/master/gpsne-1/pg_subtrans
drwx------ 2 gpadmin wheel     6 28. Apr 01:04 /data/master/gpsne-1/pg_tblspc
drwx------ 2 gpadmin wheel     6 28. Apr 01:04 /data/master/gpsne-1/pg_twophase
drwx------ 2 gpadmin wheel     6 28. Apr 01:06 /data/master/gpsne-1/pg_utilitymodedtmredo
-rw------- 1 gpadmin wheel     4 28. Apr 01:04 /data/master/gpsne-1/PG_VERSION
drwx------ 3 gpadmin wheel    60 28. Apr 01:06 /data/master/gpsne-1/pg_xlog
-rw------- 1 gpadmin wheel 19373 28. Apr 01:04 /data/master/gpsne-1/postgresql.conf
-rw------- 1 gpadmin wheel 19327 28. Apr 01:04 /data/master/gpsne-1/postgresql.conf.bak
-rw------- 1 gpadmin wheel   168 28. Apr 01:06 /data/master/gpsne-1/postmaster.opts
-rw------- 1 gpadmin wheel    47 28. Apr 01:06 /data/master/gpsne-1/postmaster.pid
```

Three: Run SQL queries

- Greenplum is using the traditional 5432 TCP port (from PostgreSQL)
- Drivers and many tools are compatible

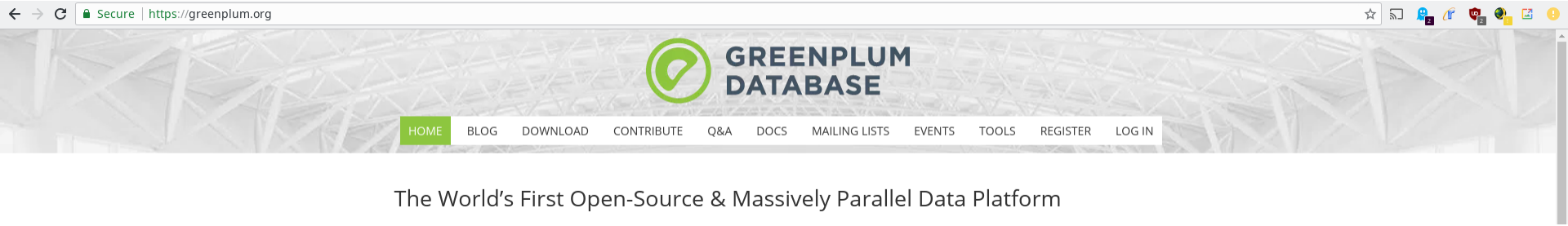
```
[gpadmin@dell-emc-world ~]$ source /usr/local/greenplum-db/greenplum_path.sh
[gpadmin@dell-emc-world ~]$ psql
psql (8.2.15)
Type "help" for help.

gpadmin=# select version();
                                                                        version
------------------------------------------------------------------------------------------------------------------------------------------------------
 PostgreSQL 8.2.15 (Greenplum Database 4.3.16.1 build 1) on x86_64-unknown-linux-gnu, compiled by GCC gcc (GCC) 4.4.2 compiled on Aug 14 2017 22:20:16
(1 row)

gpadmin=# 
```


Where to get it

Where to get it



Download Greenplum Database



SOURCE TARBALLS

Each community release is available as a source code tarball which can be compiled on a number of platforms

Download Source

Instructions for compiling source



PIVOTAL GREENPLUM

Pivotal offers a fully supported commercial distribution of Greenplum Database with some proprietary addons, both for local download as well as cloud based.

Download (Product) Sandbox

AWS Marketplace BYOL and On Demand

Azure Marketplace



APACHE BIGTOP

Bigtop is an Apache project providing packages for data science and big data tools, Greenplum Database is starting with 5.0.0-alpha supported by Bigtop.

Download

More information



UBUNTU

For users of Ubuntu Operating System, Greenplum can be installed via the apt-get command using the Ubuntu Personal Package Archive system.

Personal Package Archive



DOCKER IMAGES

There are also pre-built Docker images with Greenplum Database.

Docker Hub

Where to get it

- greenplum.org / GitHub
- Apache BigTop
- Docker
- Amazon AWS (?)
- Microsoft Azure (?)
- Google Cloud (?)

Collaboration

- gpdb-users & gpdb-dev mailing list
  - https://greenplum.org/mailing-lists/
  - Development happens there
- Greenplum Slack
  - https://greenplum.slack.com/
- Greenplum @ Twitter
  - https://twitter.com/greenplum


The End

Contact

- Andreas Scherbaum
- Email: ads@pgug.de
- Twitter: @ascherbaum
- Blog: https://andreas.scherbaum.la/

