Hello Greenplum

An Introduction to the World's First Open-Source & Massively Parallel Data Platform

Bradford D. Boyle & Andreas Scherbaum

Outline

1. What is Greenplum? Who is using it?
2. History: Where does it come from?
3. How does it work?
4. Demo
5. Where to get it?

Audience

If you

- don’t know what Greenplum is
- heard about it but don’t know the details
- are curious

Then this talk is for you!

Bradford D. Boyle

- Working for Pivotal since 2016
- Work on
  - GemFire-Greenplum Connector
  - Greenplum-Spark Connector
- Twitter: @BradfordDBoyle

Bradford Boyle
bradfordboyle

Add a bio

Menlo Park, CA
http://www.bradfordboyle.com

Organizations

What is Greenplum?

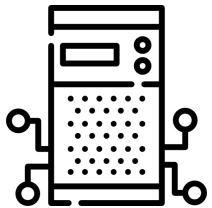
What is Greenplum

- Massively Parallel Processing (MPP)
- Shared Nothing
- Database

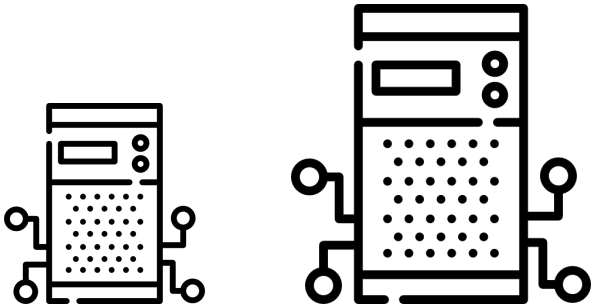
Massively Parallel Processing (MPP)

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

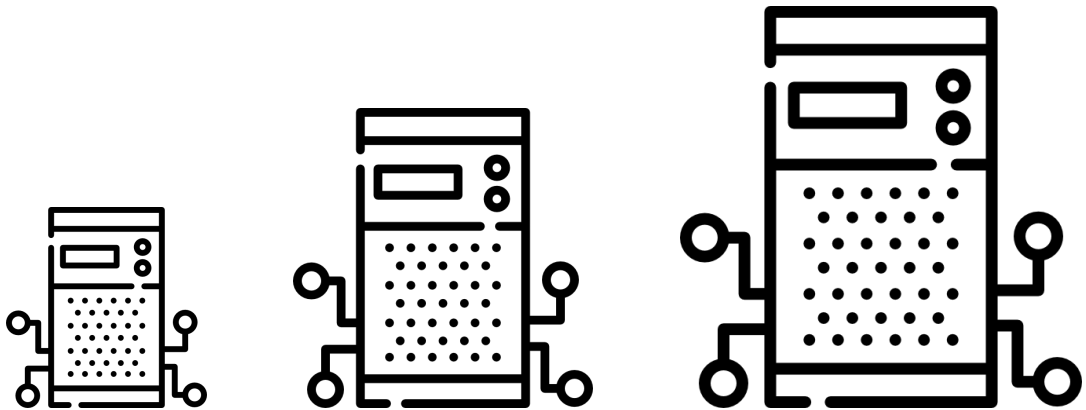
Little Bit of History



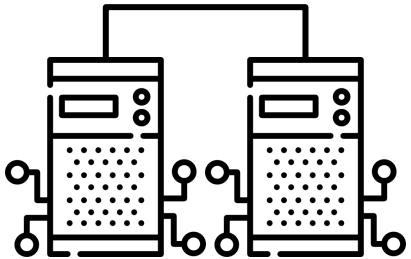
Little Bit of History



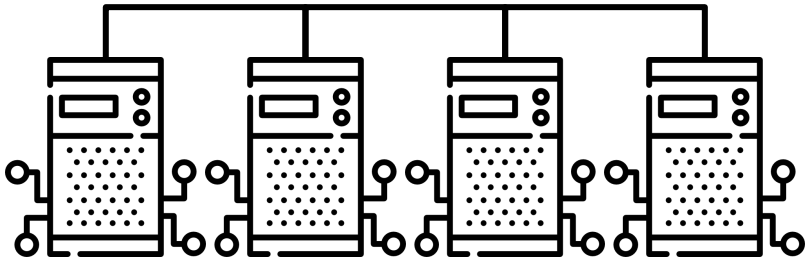
Little Bit of History



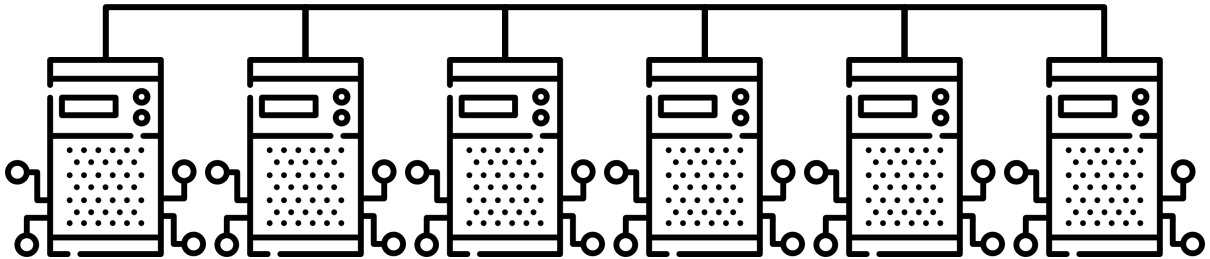
Scale Out, Not Up



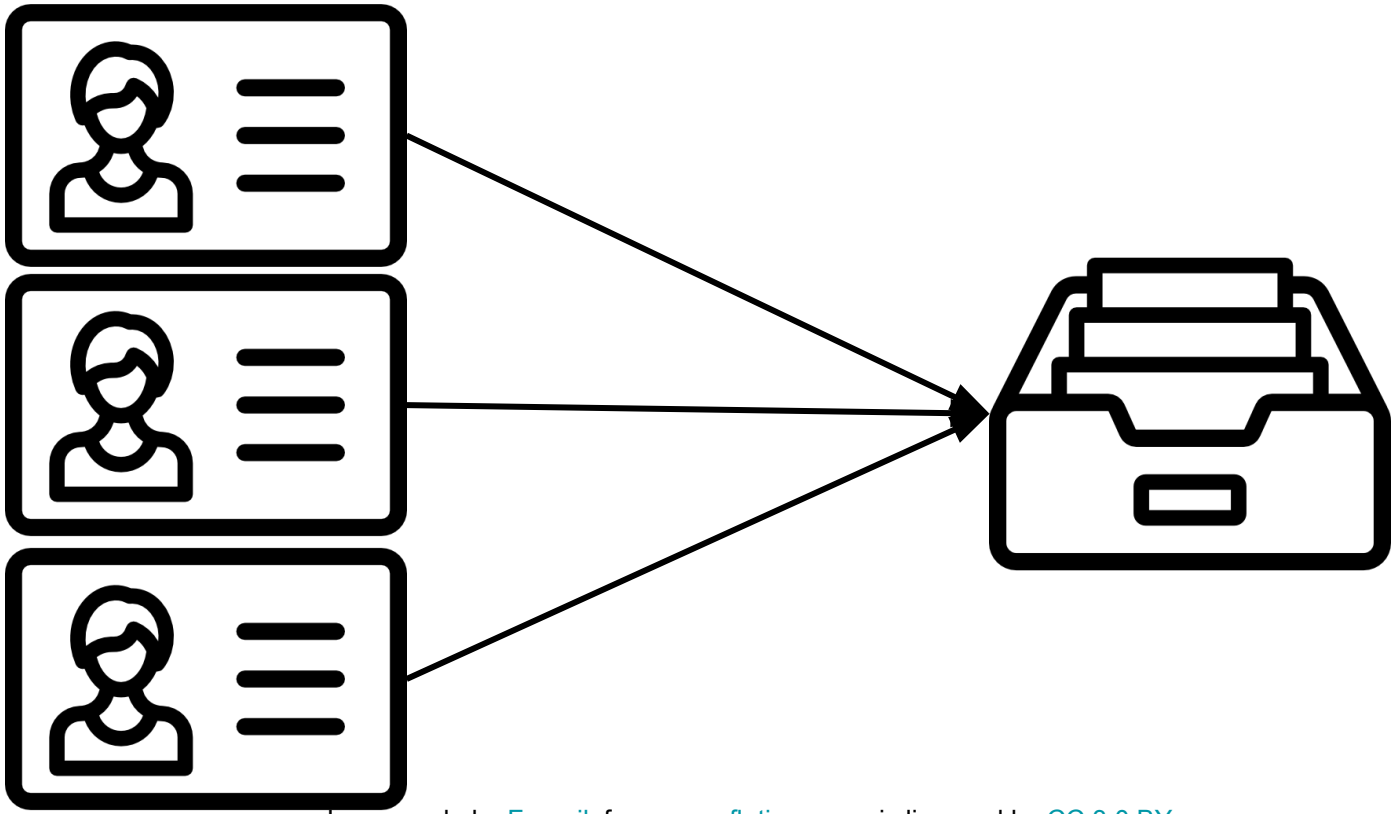
Scale Out, Not Up



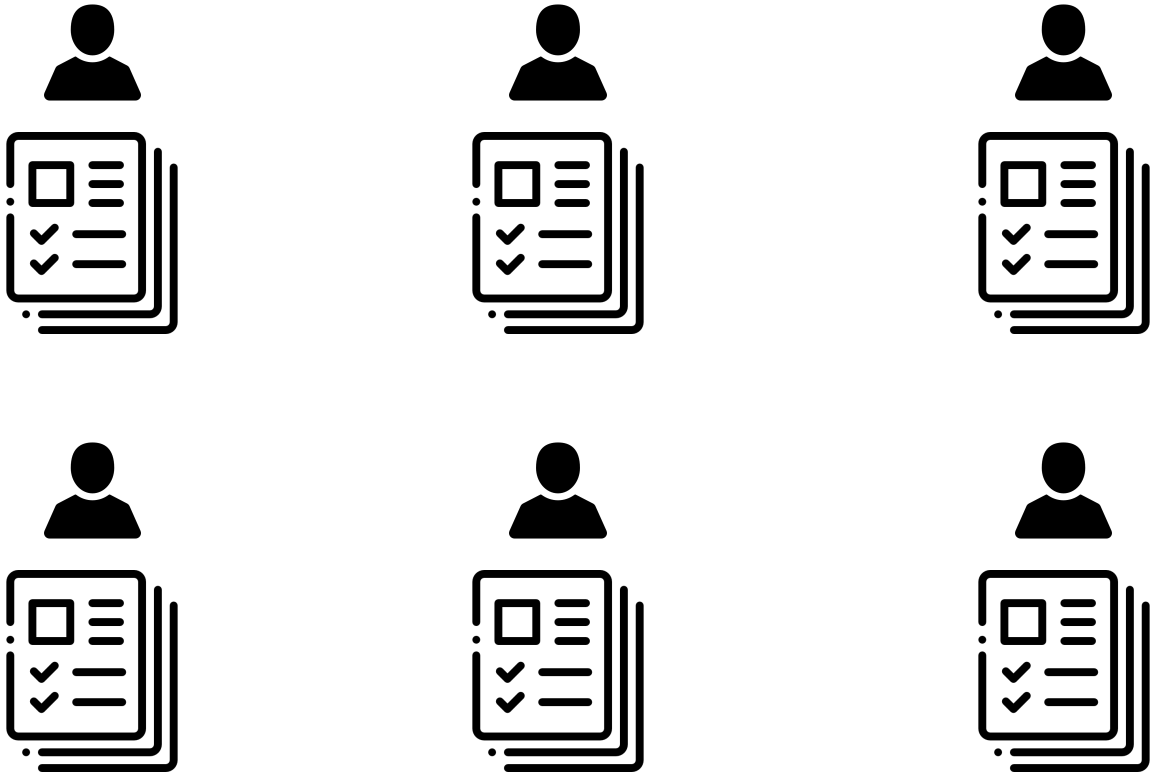
Scale Out, Not Up



A Brief Analogy



A Brief Analogy



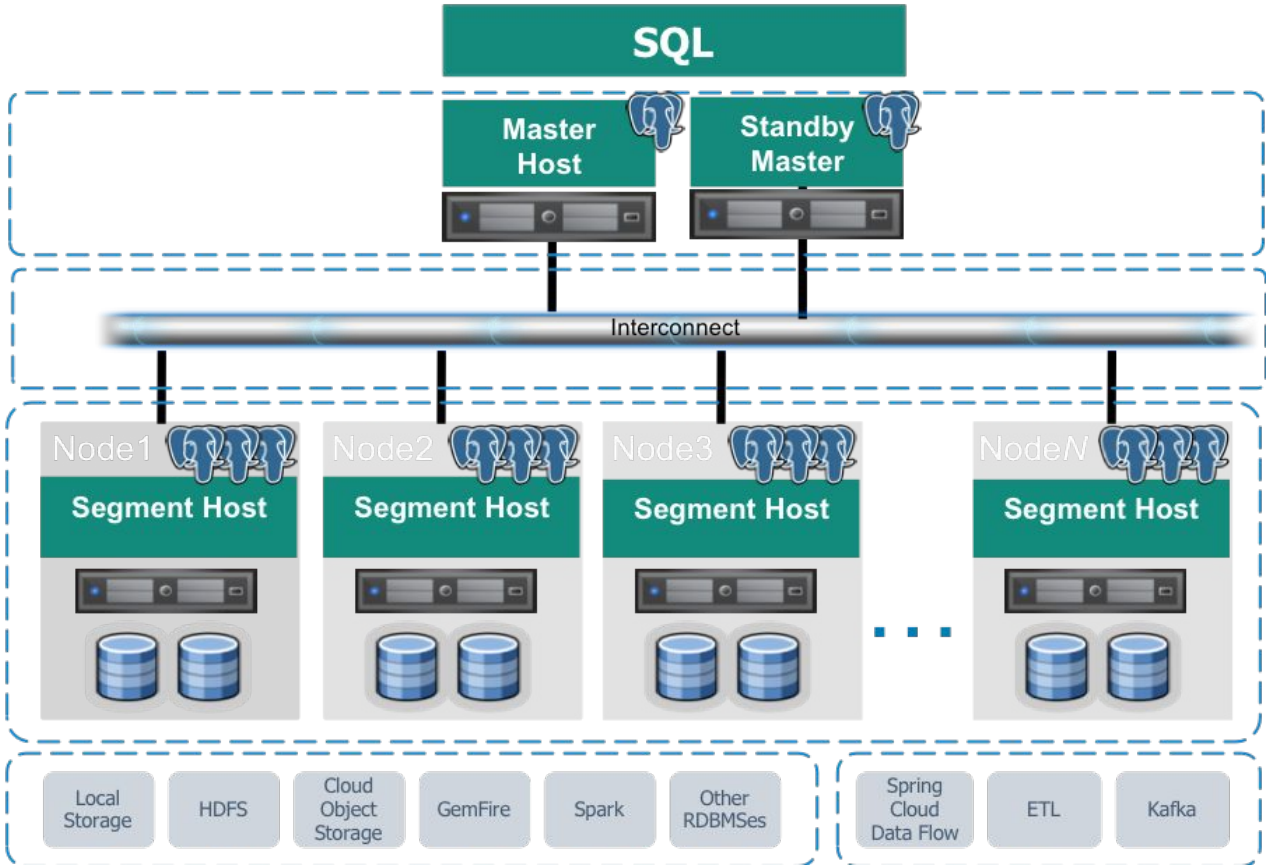
Greenplum is MPP for Analytics

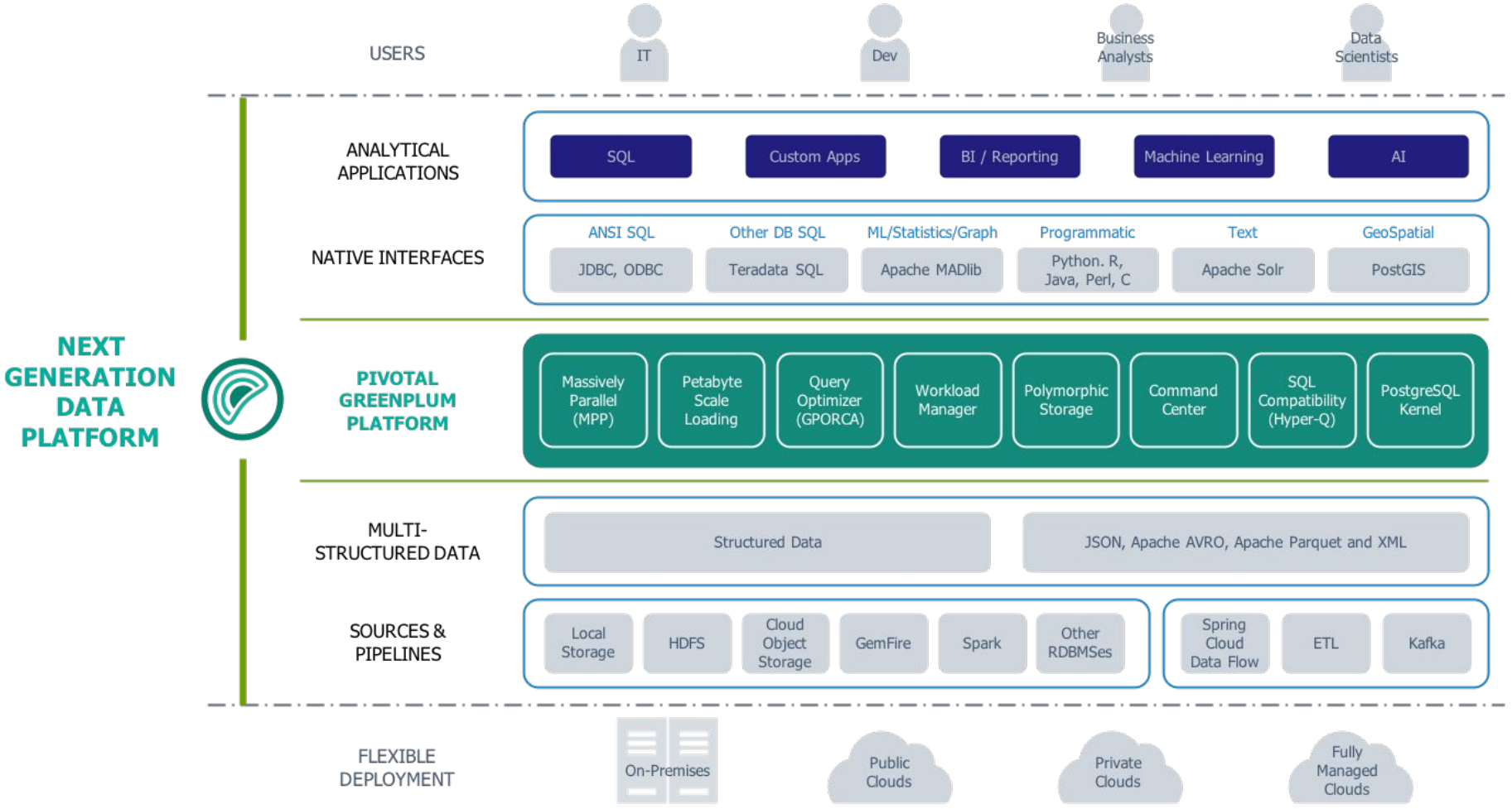
Master Server
Query planning & dispatch

Interconnect

Segment Servers
Query processing & data storage

External Resources
Parallel loading & 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

Shared Nothing

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

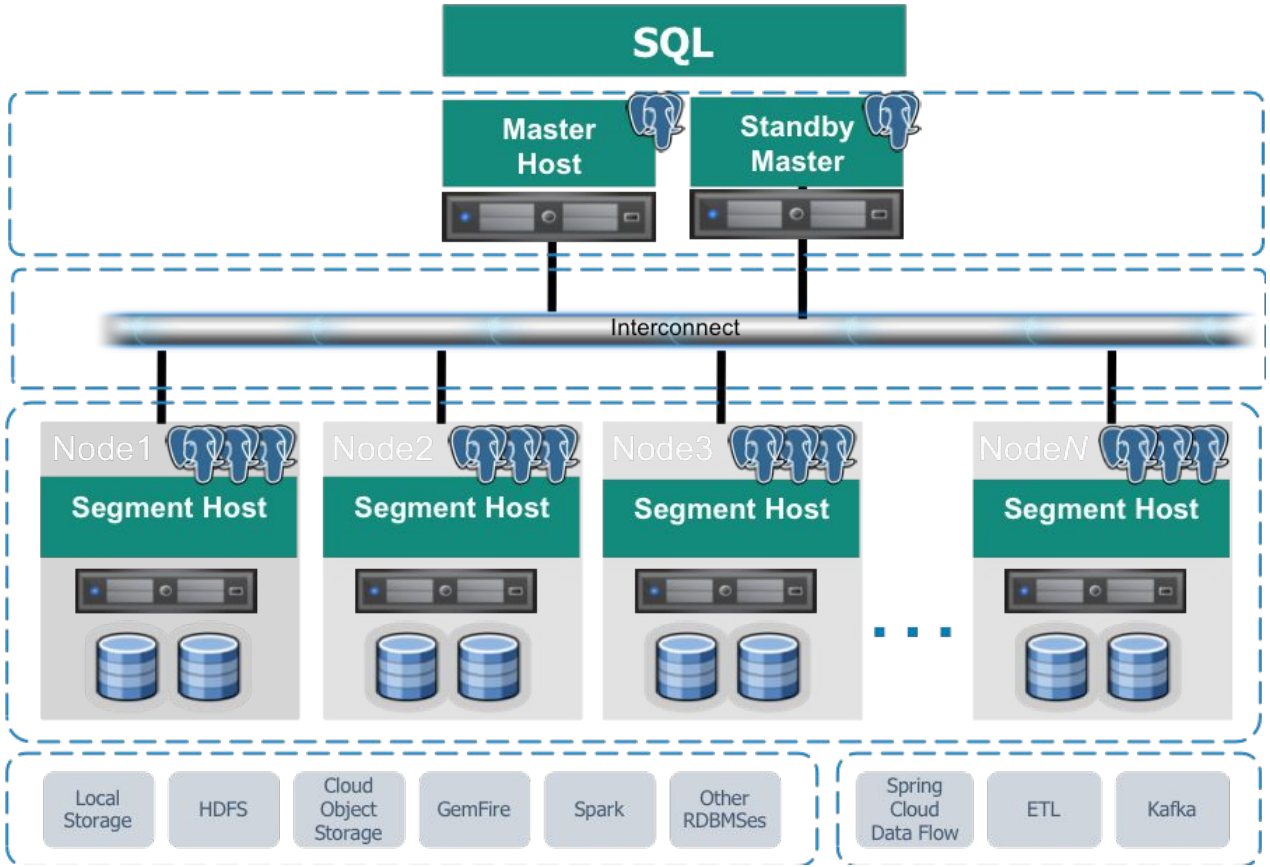
Greenplum is MPP for Analytics

Master Server
Query planning & dispatch

Interconnect

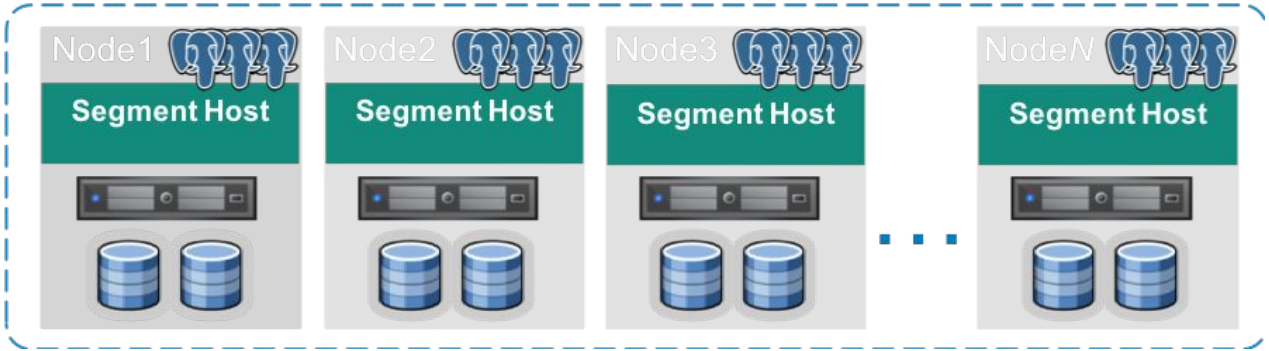
Segment Servers
Query processing & data storage

External Resources
Parallel loading & streaming



Shared Nothing

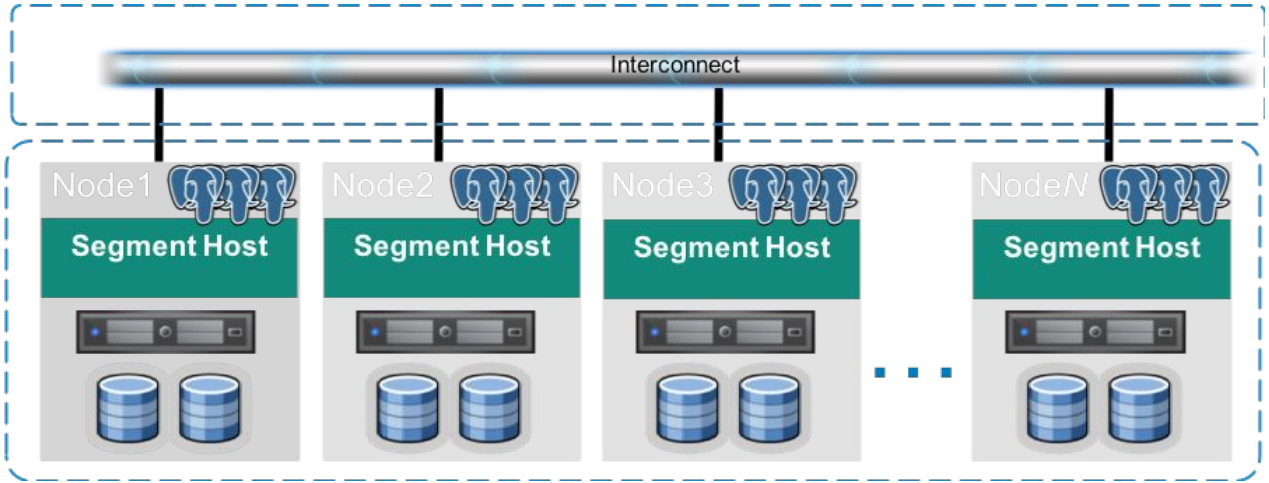
Segment Servers
Query processing & data storage



Shared Nothing

Interconnect

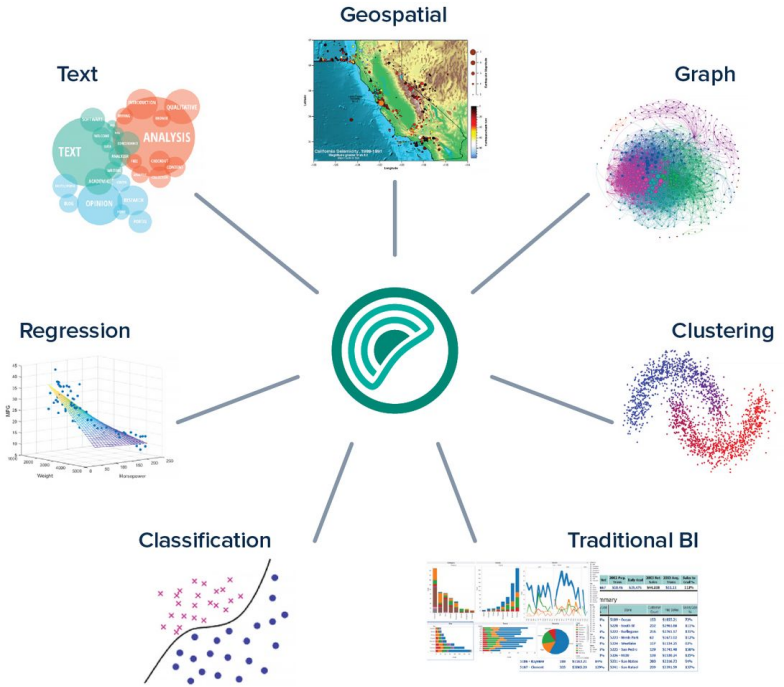
Segment Servers
Query processing & data storage



Who is Using Greenplum?

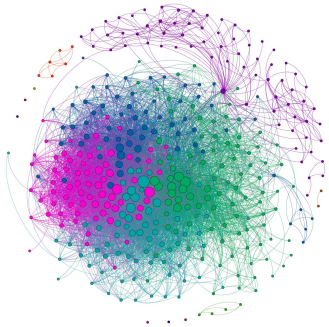
Greenplum Integrated Analytics

- Traditional & Advanced In-Database Analytics



Greenplum Analytics: Graph

- Designed for very large graphs (billions of vertices/edges)
- No need to move data & 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 & monitoring
- Customer sentiment analysis
- Document search & query
- Social media processing, etc.

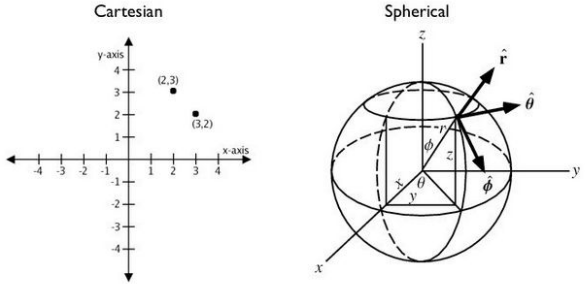
GPText

- Leverages Apache Solr & Greenplum
- Python & Java integration for natural language processing (NLP)
- Apache MADlib for machine learning on text data

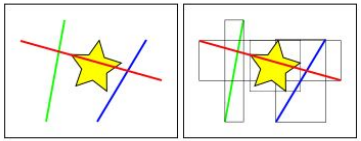
Greenplum Analytics: GeoSpatial

- Points, Lines, Polygons, Perimeter, Area
- Intersections, Contains, Distance, Longitude & Latitude

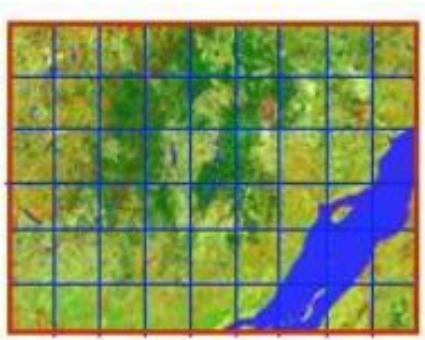
Round earth calculations



Spatial indexes & bounding boxes



Raster support



Greenplum Analytics: R & Python Libraries

gensim

MCMCpack

NumPy

TensorFlow

pyLDAvis

scikit learn

LIFELINES

spaCy

XGBoost

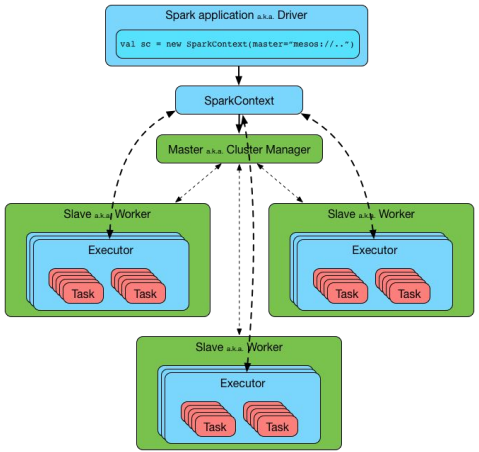
SM

BeautifulSoup

Greenplum Analytics: Spark

- Provide data access to Greenplum data
- Leverage Spark background of data engineers & data scientists
- Utilize off-cluster compute resources for computations
- Persists results to Greenplum

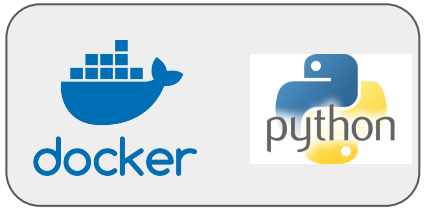
In-memory processing



Greenplum-Spark connector

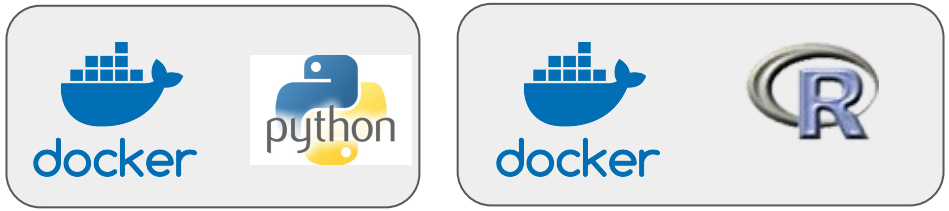
Greenplum Analytics: Language Agnostic

- Interfaces
  - User defined types
  - User defined functions
  - User defined aggregates
- Foundational work containerized Python & R comput environments



R/Python Containerization on Greenplum

- PL/Container
  - Deploy custom R & Python developer environments in the cluster
  - Execute functions in isolated, secure containers
  - Deploy code & functions as non-superuser
  - Package and custom Python or R modules in the deployment
  - Pre-configured for data science or customized images by users
  - Multiple developer environments on same cluster



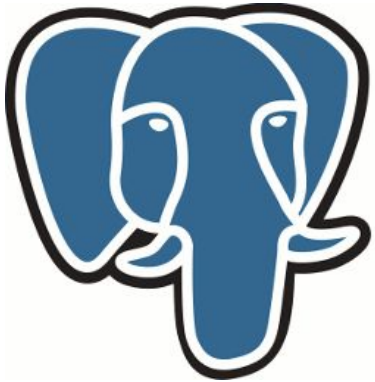
SQL Containerization

- Greenplum Resource Groups
  - Resource isolation for multi-tenancy & mixed analytical SQL workloads
  - Enhance stability & manageability
  - Leverages Linux cgroups

Where Does Greenplum Come From?

Where Does Greenplum Come From?

- Greenplum is based on PostgreSQL



PostgreSQL

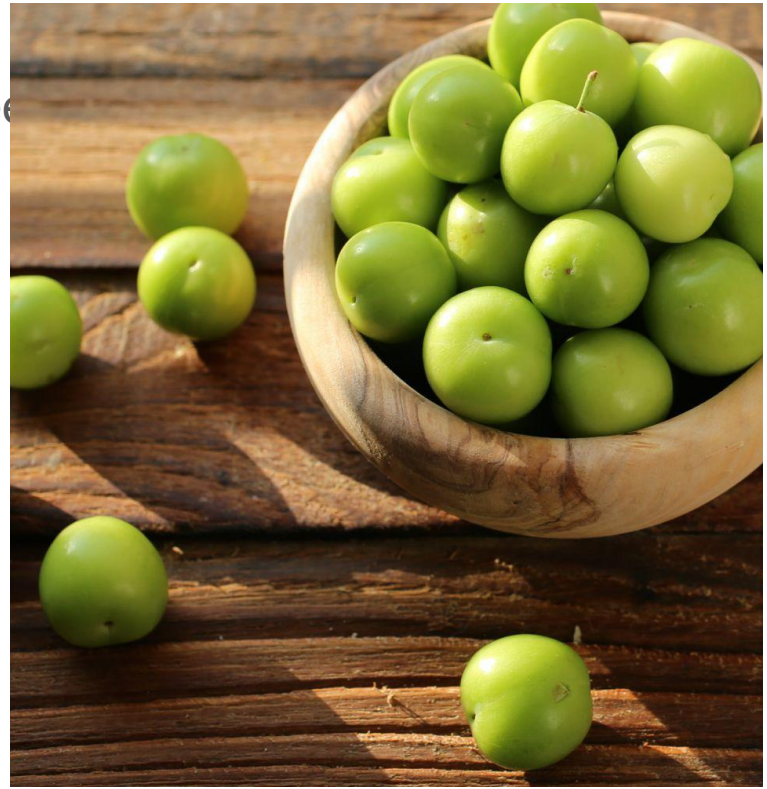
the world's most advanced open source database

PostgreSQL History

- 1985: Postgres in born at Berkeley University
  - First registered domain name
- 1989: Postgres 2.0 (still no SQL)
  - First commercial ISP
- 1990: First website at CERN
- 1994: Netscape browser, Yahoo! born
- 1995: Postgres95
- 1997: PostgreSQL
  - Netflix founded (renting DVD by mail)
- 1998: Google founded, IPv6 introduced
- 2018: PostgreSQL v11 (30th major release)

Greenplum History

- 2003: Metapa acquires Didera, founds Gree
  - MySpace & Skype founded
- 2005: Greenplum releases *Bizgres*
  - YouTube launches
- 2010: Acquired by EMC
  - Digital storage reached the Zetta
- 2015: Greenplum is open-sourced
  - Internet-of-Things (IOT)
- 2017: Greenplum v5 released



Greenplum’s Futures

- Greenplum v4 based on a fork of PostgreSQL 8.2.x
- Greenplum v5 based on 8.3.x
- Current development is based 9.1
- Code & features flow in both directions

How Does Greenplum Work?

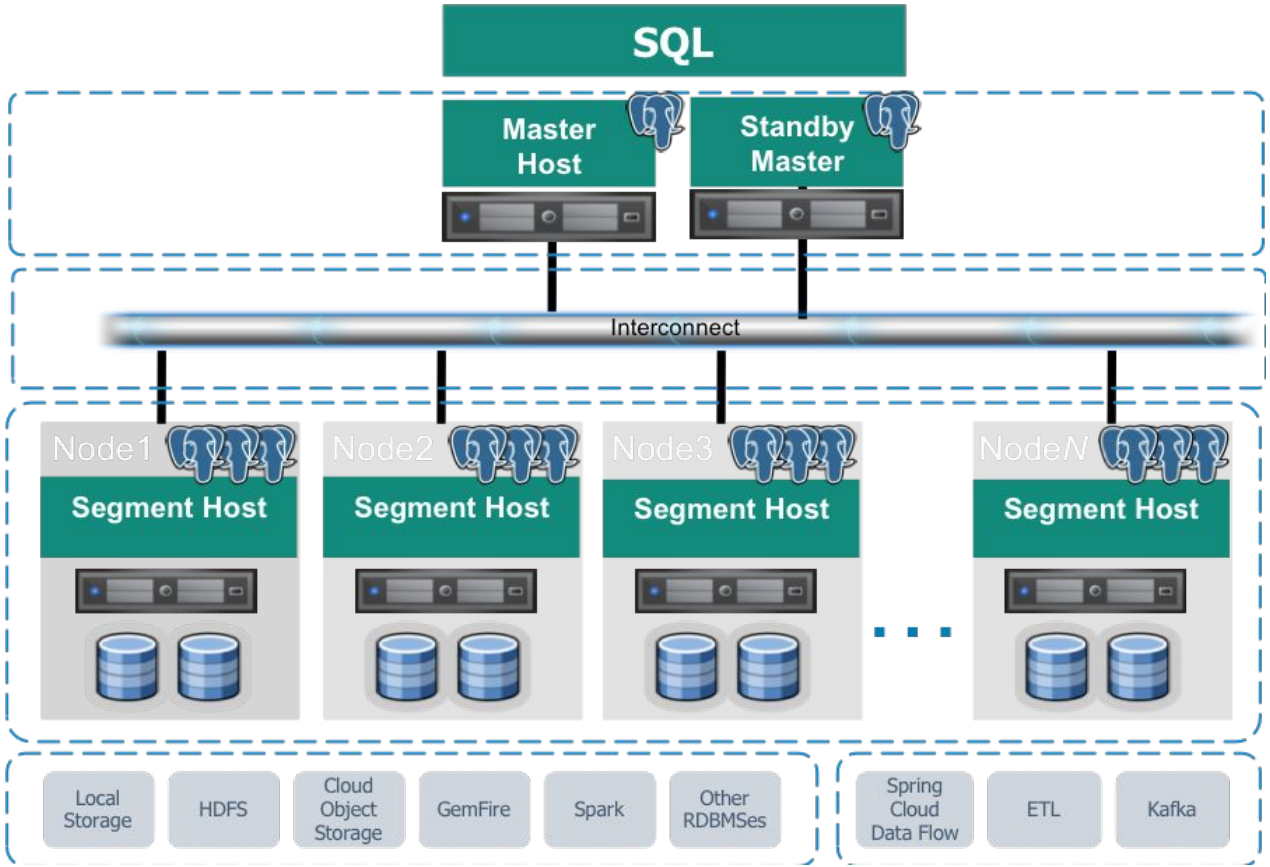
Greenplum is MPP for Analytics

Master Server
Query planning & dispatch

Interconnect

Segment Servers
Query processing & data storage

External Resources
Parallel loading & streaming



Software Installation

- Greenplum is installed once per host
- Install Greenplum OSS on Ubuntu
  - https://greenplum.org/install-greenplum-oss-on-ubuntu/
- Download & build from source
  - https://greenplum.org/download/
- Download & install from Pivotal Network
  - https://network.pivotal.io/products/pivotal-gpdb/
  - https://gpdb.docs.pivotal.io/510/install_guide/install_guide.html

Initialize Segments & Master

- Data directories are separate PostgreSQL-like installations
- Every master & segment is hosted in its own directory
  - Can reside on different file systems

```
gpadmin@greenplum-singlenode:~$ tree -L 2 -pugD data
data
├── [drwxr-xr-x gpadmin  gpadmin  Sep  7  1:39]  master
│   └── [drwx------ gpadmin  users  Sep  7  1:39]  gpsne-1
├── [drwxr-xr-x gpadmin  gpadmin  Sep  7  1:39]  seg0
│   └── [drwx------ gpadmin  users  Sep  7  1:39]  gpsne0
└── [drwxr-xr-x gpadmin  gpadmin  Sep  7  1:39]  seg1
      └── [drwx------ gpadmin  usersSep  7  1:39]  gpsne1

6 directories, 0 files
```

Initialize Segments & Master

```
gpadmin@greenplum-singlenode:~$ tree -L 1 -pugD data/master/gpsne-1
data/master/gpsne-1
├── [drwx------ gpadmin  usersSep  7  1:39]  base
├── [drwx------ gpadmin  usersSep  7  1:39]  global
├── [-r-------- gpadmin  usersSep  7  1:39]  gp_dbid
├── [drwxr-xr-x gpadmin  usersSep  7  1:39]  gpperfmon
├── [-rw-r--r-- gpadmin  usersSep  7  1:39]  gpssh.conf
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_changetracking
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_clog
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_distributedlog
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_distributedxidmap
├── [-rw-r--r-- gpadmin  usersSep  7  1:39]  pg_hba.conf
├── [-rw------- gpadmin  usersSep  7  1:39]  pg_ident.conf
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_log
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_multixact
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_stat_tmp
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_subtrans
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_tblspc
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_twophase
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_utilitymodedtmredo
├── [-rw------- gpadmin  usersSep  7  1:39]  PG_VERSION
├── [drwx------ gpadmin  usersSep  7  1:39]  pg_xlog
├── [-rw------- gpadmin  usersSep  7  1:39]  postgresql.conf
├── [-rw------- gpadmin  usersSep  7  1:39]  postmaster.opts
└── [-rw------- gpadmin  usersSep  7  1:39]  postmaster.pid

15 directories, 8 files
```

Run SQL Queries

- Greenplum uses TCP 5432 (from PostgreSQL)
- Drivers & many tools are compatible

```
gpadmin=# SELECT VERSION();
                                 version
-------------------------------------------------------------------------
 PostgreSQL 8.3.23 (Greenplum Database 5.10.2 build b3c02f3-oss) on x86
 _64-pc-linux-gnu, compiled by GCC gcc (Ubuntu 5.4.0-6ubuntu1~16.04.10)
  5.4.0 20160609, 64-bit compiled on Aug 10 2018 08:54:39
(1 row)
```

Demo

Demo

- Using simple “world” database from the PostgreSQL project
  - Includes cities, countries, & language codes
- Show “psql”, “data import” using “copy”, “file://”, and “gpfdist://”
- Run some queries
- Follow along
  - https://github.com/bradfordboyle/pgsv2018-hello-greenplum

Where to Get Greenplum

Where to Get Greenplum

- greenplum.org
- GitHub
- Pivotal (w/ support & services)
- Apache BigTop
- Docker
- Pivotal Cloud Foundry
- Amazon AWS
- Microsoft Azure

Collaboration

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

The End

Contact

- Bradford D. Boyle
  - bboyle@pivotal.io
  - @BradfordDBoyle
- Pivotal Engineering
  - https://engineering.pivotal.io/
- This talk
  - https://github.com/bradfordboyle/pgsv2018-hello-greenplum