VMware Tanzu Greenplum

An Introduction

May 2020

Future Looking Statements

Disclaimer

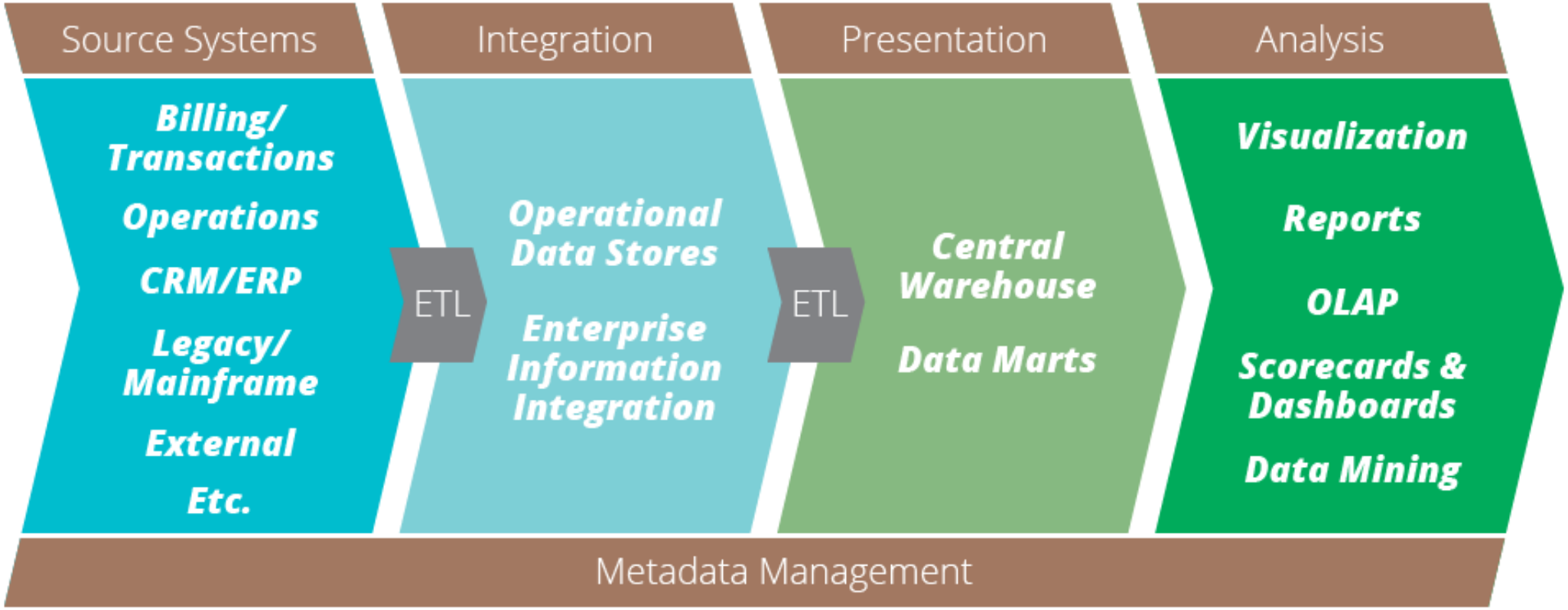
- ***Presentations may contain product features or functionality that are currently under development.***
- ***This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.***
- ***Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.***
- ***Technical feasibility and market demand will affect final delivery.***
- ***Pricing and packaging for any new features/functionality/technology discussed or presented, have not been determined.***
- ***This information is confidential.***

Greenplum Keynote

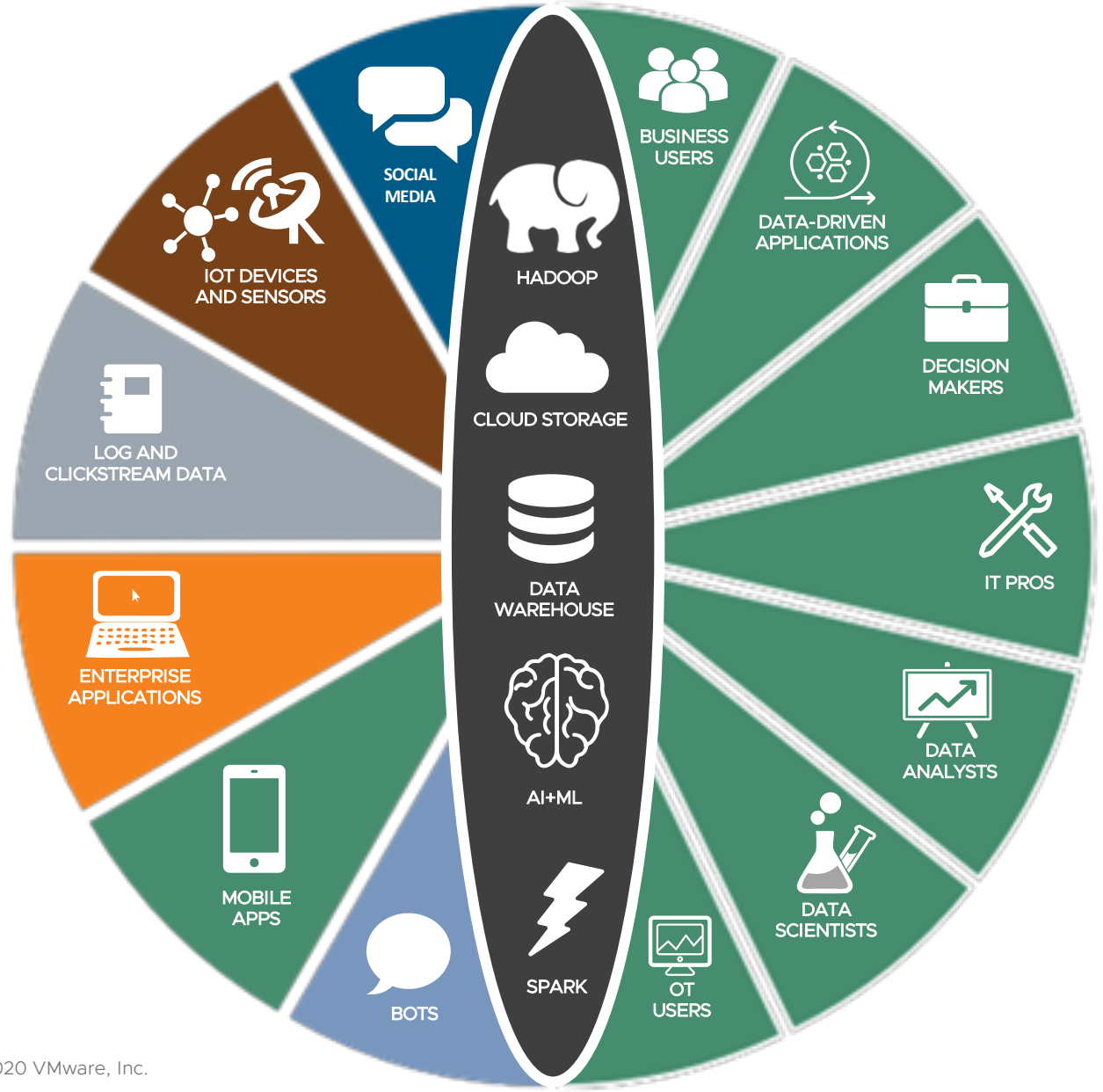
Greenplum is the platform that can power your analytics needs now and, in the future. Let us show you how.

The Greenplum Analytics Platform

Historic Traditional Data Warehouse Process



Use Cases for Analytics Have Expanded Dramatically



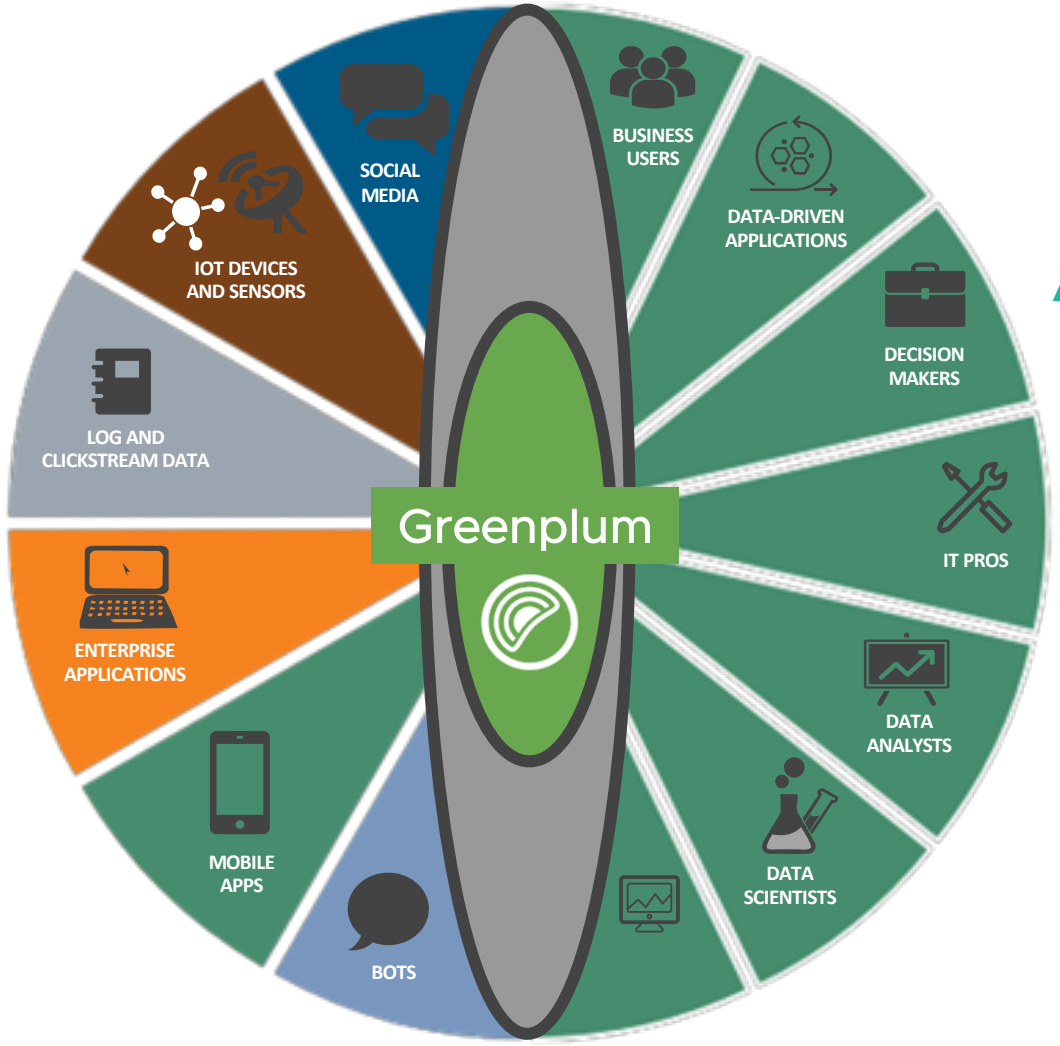
Today's Data Architect can be Easily Overwhelmed

- Enterprise DB Data
- IoT Data
- Logs & Security Data
- Web, Mobile, Clickstream
- Image, Video, Voice Data
- JSON, XML, Geo, Graph


BIG DATA LANDSCAPE
INFRASTRUCTURE
ANALYTICS
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
OPEN SOURCE
DATA SOURCES & APIS
DATA RESOURCES
CROSS-INFRASTRUCTURE ANALYTICS
V2 – Last updated 5/3/2017
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)
mattturck.com/bigdata2017
FIRSTMARK

Redefining The Data Platform

Simplify for Efficiency and Cost Savings



Analyze Across Any New Dimensions

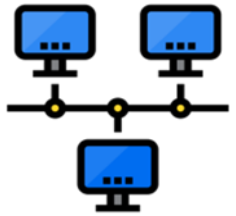
With Greenplum you can have a manageable single solution

GREENPLUM.

- Enterprise DB Data
- IoT Data
- Logs & Security Data
- Web, Mobile, Clickstream
- Image, Video, Voice Data
- JSON, XML, Geo, Graph

Enhanced PostgreSQL

Massively Parallel Database



Massive compute grid

Python and R High Performance Computing

Federated Query Database

What is Greenplum?

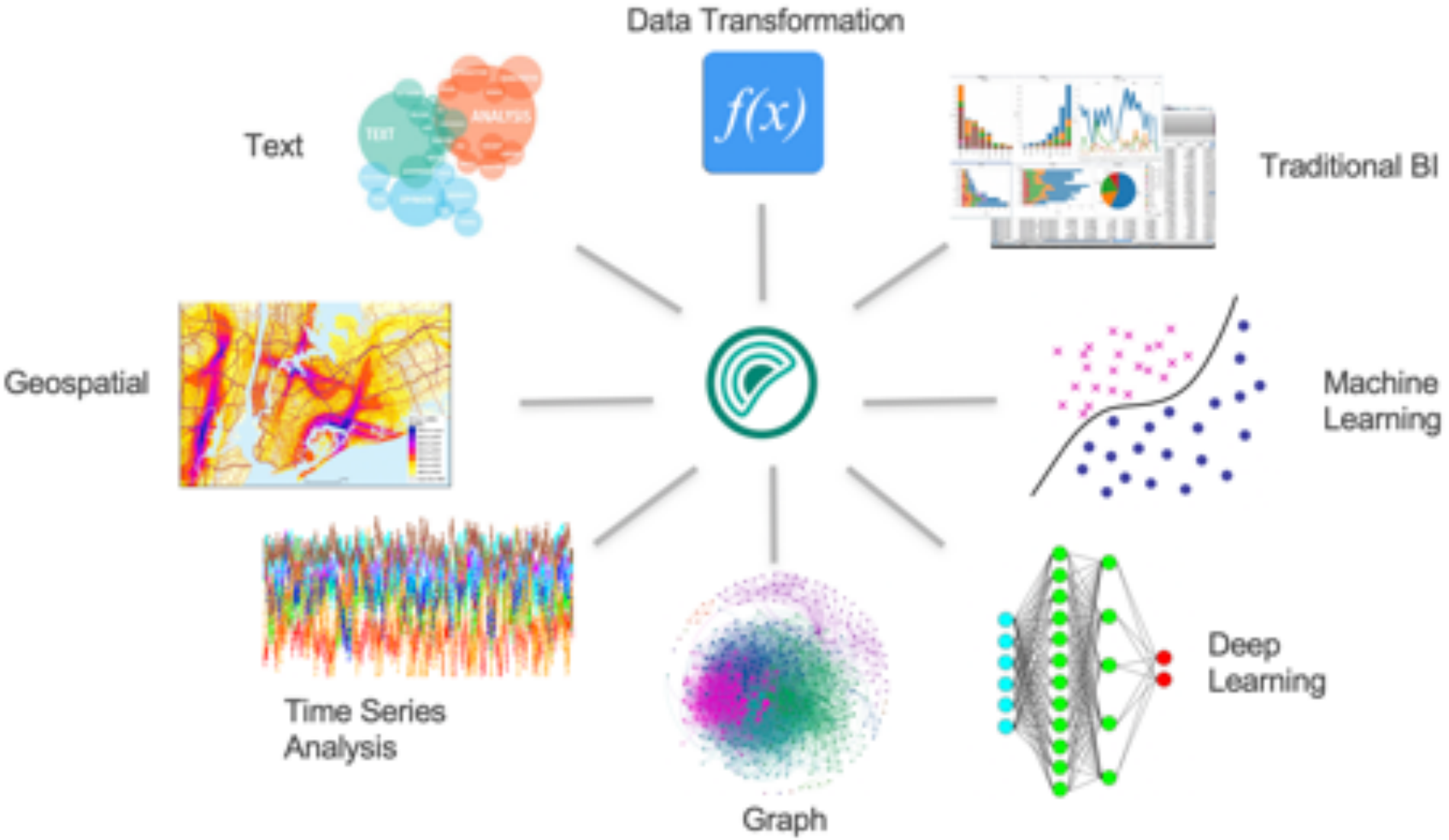
A Structured and Semi-Structured Platform leveraging ANSI-SQL

Enterprise Search Platform

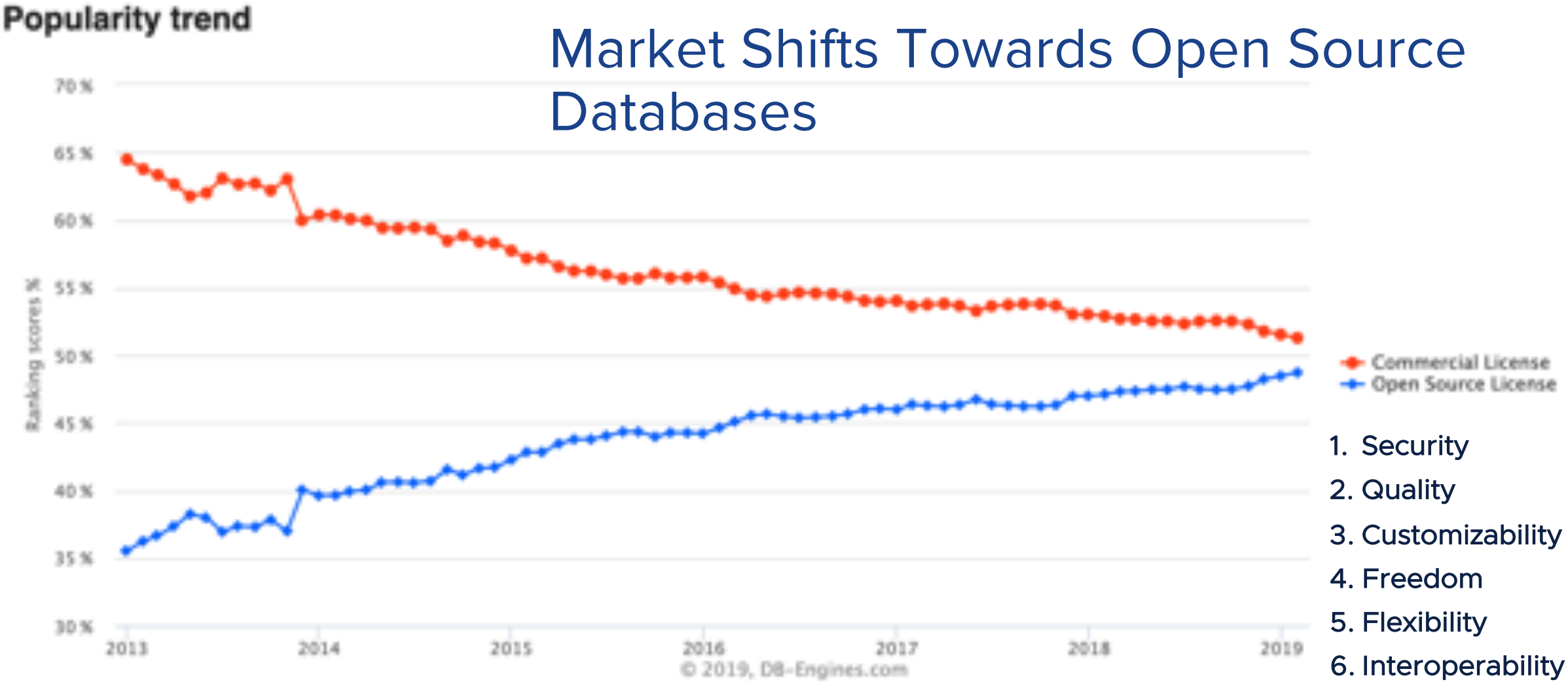
Analytics Database

Graph, Geospatial, Time Series and Image recognition

Greenplum Integrated Analytics



Market Shifts Towards Open Source Databases



1. Security
2. Quality
3. Customizability
4. Freedom
5. Flexibility
6. Interoperability
7. Auditability
8. Support Options
9. Cost
10. Try before you buy

Infrastructure Advances Accelerating

Data & Compute Architectures Becoming More Powerful

Data Center Revolution

Consolidation of data center's provides unprecedented economies of scale

Industry Standards

x86 Arch, Linux OS, Kubernetes, Ethernet, TCP/IP and SQL provide today's fabric for data computing

New Workloads

Large scale deployments of big data

And compute

"Cloud is about how you do computing, not where you do computing."

Paul Maritz, CEO VMWare / Pivotal

Flexibility in Hybrid Cloud

Answer Real Questions With Greenplum

"Find anyone who *works at Pivotal* and knows each other *directly* and whose name *sounds* like 'Peter' or 'Pavan' and have *withdrawn an amount* > $200 *within 24 hours* at an ATM less than *2 KM from a reference* latitude and longitude"

Find anyone who works at 'Pivotal' and know each other 'directly' and whose names sound like 'Peter' or 'Pavan' and have withdrawn an amount > $200 within 24 hours at an ATM less than 2 KM from reference latitude and longitude.

```
CREATE FUNCTION get_people(person1 text, person2 text, amount int, duration int, longit float,latitude float) RETURNS int
AS $$
declare
linkchk integer; v1 record; v2 record;
begin
 execute 'truncate table results;';
 for v1 in select distinct a.id,a.firstname,a.lastname,amount,tran_date,c.lat,c.lng,address,a.description,d.score from people a,transactions b,location c,
    (SELECT w.id, q.score FROM people w, gptext.search('gpadmin.public.people' , 'Pivotal') q
     WHERE (q.id::integer) = w.id order by 2 desc) d
     where soundex(firstname)=soundex($1) and a.id=b.id and amount > $3 and (extract(epoch from tran_date) - extract(epoch from now()))/3600 < $4

and st_distance_sphere(st_makepoint($5, $6),st_makepoint(c.lng, c.lat))/1000.0 <= 2.0 and b.locid=c.locid and a.id=d.id
 loop
   for v2 in select distinct a.id,a.firstname,a.lastname,amount,tran_date,c.lat,c.lng,address,a.description,d.score from people a,transactions b,location c,
    (SELECT w.id, q.score FROM people w, gptext.search(TABLE(SELECT 1 SCATTER BY 1), 'gpadmin.public.people' , 'Pivotal', null) q
     WHERE (q.id::integer) = w.id order by 2 desc) d
     where soundex(firstname)=soundex($2) and a.id=b.id and amount > $3 and (extract(epoch from tran_date) - extract(epoch from now()))/3600 < $4
    and st_distance_sphere(st_makepoint($5, $6),st_makepoint(c.lng, c.lat))/1000.0 <= 2.0 and b.locid=c.locid and a.id=d.id
 loop
     execute 'DROP TABLE IF EXISTS out, out_summary;';
     execute  'SELECT madlib.graph_bfs(''people'',''id'',''links'', v1.id,''out'');';
     select 1 into linkchk from out where dist=1 and id=v2.id;
     if linkchk is not null  then
          insert into results values (v1.id,v1.firstname,v1.lastname,v1.amount,v1.tran_date,v1.lat,v1.lng,v1.address,v1.description,v1.score);
          insert into results values (v2.id,v2.firstname,v2.lastname,v2.amount,v2.tran_date,v2.lat,v2.lng,v2.address,v2.description,v2.score);
      end if;
    end loop;
 end loop;
 return 0;
end
$$ LANGUAGE plpgsql;
-- Call the function now
select get_people('Pavan','Peter',200,24,103.912680, 1.309432) ;
```

Greenplum Fuzzy String Match function **Soundex()** to know if people name sounds like 'Pavan' or 'Peter'

GPText.search() function is used to know if both people work at 'Pivotal'

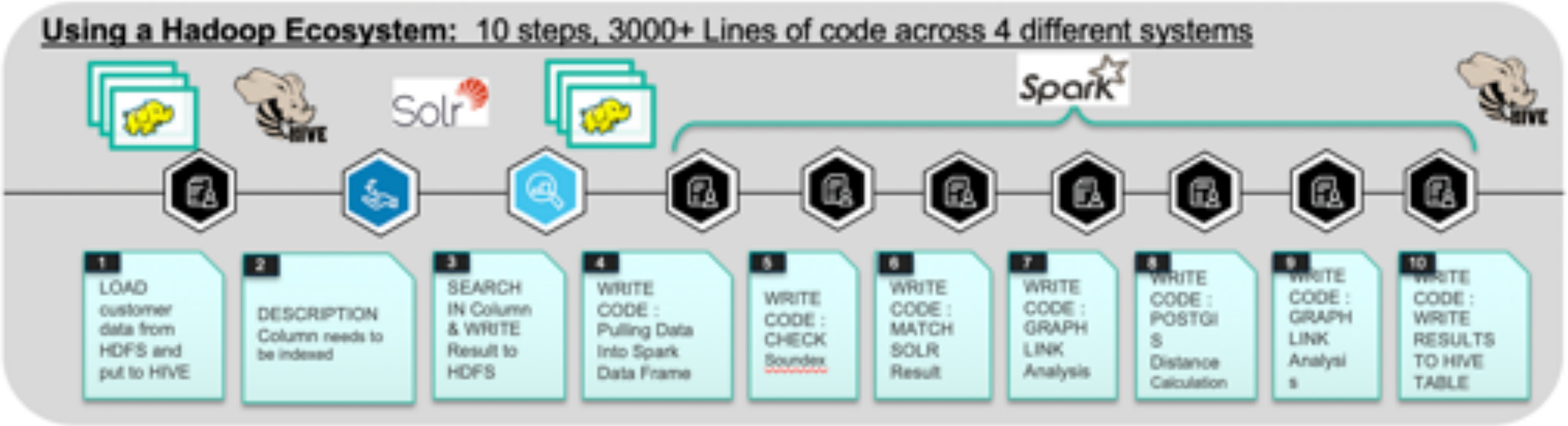
Amount > $200

Greenplum and **Apache MADlib BFS** search to know if there are direct or indirect links between people

Greenplum **Time functions** to calculate difference in amount withdrawn time < 24 hours

Greenplum POSTGIS functions **st_distance_sphere() and st_makepoint()** calculate distance between ATM location and reference latitude, longitude < 2 KM

Coding Productivity Gain 100x vs Competition



Using Greenplum: 1 step, 1 query – 34 Lines of Code

One query – using built-in functions: Soundex (sounds like), NLP (work at same company), Machine Learning MADlib (know directly), Time (yesterday), PostGIS (within 2km)

Greenplum Product Overview

Greenplum Data Sheet

Analytics Data Platform

Enhanced PostgreSQL for Analytical Workloads

ANSI standard SQL, ACID RDBMS, Optimized for Analytics and Big Data Storage & Processing

Massively Parallel Shared Nothing Database

Scale out horizontally from to hundreds of instances or servers

Streaming and Real-time Data Processing

Greenplum Streaming Server connects to Apache Kafka and has an extensible API supporting continual streaming ingestion

Federated Query Processing

S3, HDFS, Hive, Oracle, Parquet, AVRO, JSON and other systems and file formats accessible via Greenplum for query Processing

Hybrid Analytical and Transactions Processing

Simultaneous thousands of Update and Delete transactions per second, Dashboard index lookups and heavy analytics and reporting

Advanced Analytics Beyond OLAP

Machine learning, Deep Learning with GPUs, Time Series, Geospatial and Graph analytics all in the database.

Python, R, Java, Perl, C

User Defined Functions in popular programming languages programmatically modify and transform data. Import libraries in these languages

Enterprise Search Platform

Embedded parallel Apache Solr engines enabled full text indexing and search

Deploy Workloads on Any Infrastructure

One Analytics Data Platform Anywhere

Operating System

Containers

Infrastructure

Data Center

Google Container Engine

Pivotal Container Service

Other Kubernetes (on VMs or not)

Bare-Metal

Private Data Center

Greenplum on Cloud

Elastic Capacity in Public Cloud

Automated Single Click Deploy

Deploy an optimized configuration to the public cloud through a wizard interface

Self Healing Cloud Infrastructure

Machines fail, and when they do Greenplum will automatically reallocate and leverage new cloud instances to replace failed ones

Database Snapshots

Database snapshots can be taken and replicated in the public cloud while users are connected to Greenplum and running workloads

Performance & Horizontal Scale

Bare metal equivalent performance and ability to scale out to larger clusters with cloud provisioned instances

Autonomous Database Operations

Greenplum Building Blocks

Greenplum Reference Configuration Based on Dell

DELL EMC PowerEdge

Optimized Configuration for Greenplum

Greenplum and Dell partner to select the ideal configuration for performance, usability and availability.

Simple & Flexible Blocks Design

Choose between compute, storage, or balanced block types and combine together into a single or multi-rack system that is easily expandable

2020 Dell Tech Components

Leverage NVMe Storage, 40 gig networking, Terabyte RAM, up to 192 cores per host, and multiple dedicated network channels for maximum performance

Extensible Data Types

Run Greenplum on any Data

JSON & XML

Store documents with flexible schemas and introspect document structures during query processing

Text, Image, Video

Store rich ”unstructured” data in tables perform search and deep learning recognition on these types

Network Traffic, IoT, Logs

Ip Addresses, Ranges of Addresses, Packet Captures, System Logs, and IoT sensors stored and analyzed

Geo & Graph

Locations and relationships can be stored and natively analyzed

Extensible User Defined Data Types and Custom Processing Per Type

Data is Stored Everywhere

Greenplum Federated Query

Federated Query Processing

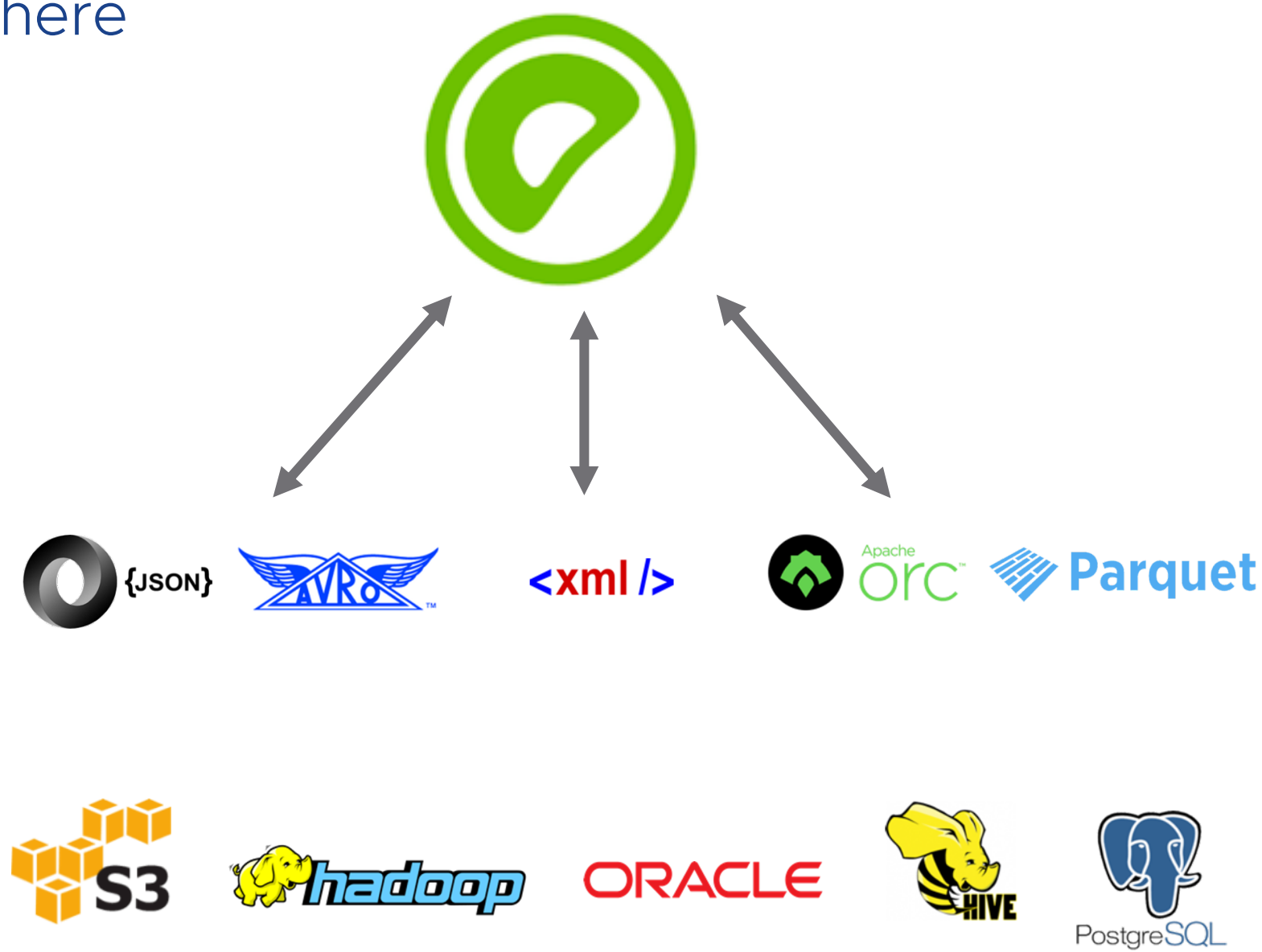
PXF extensible design can query external data in multiple formats and locations

Massively Parallel External Data Access

Each segment scans external data sources in parallel for Terabyte & Petabyte scale external tables

Smart Processing

Optimizer and query processing engine can push down filters and project column selection to remote system for minimized data transfers over the network



Multi-Temperature Data Storage

Greenplum Federated Query

Vertical Partitioning

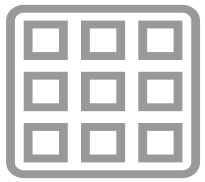
Large fact tables divided into time ranges for efficient data access and retention policies

Polymorphic Partitioning

Different ranges in partitioned table can use different storage parameters and mediums

Optimizer Partition Elimination

Query processing will automatically only scan the storage mediums that contain data needed based on query conditions



Indices
Row Store

Column
Store

External
S3 and
HDFS

Temperature

HOT DATA

WARM DATA

COLD DATA

- Storage based on operational requirements
- Can I work with data created *few second* ago ?
- Can I run a report on data from few days ago ?
- Can I inspect the data archived months or years ago ?

MPP Shared Nothing Architecture

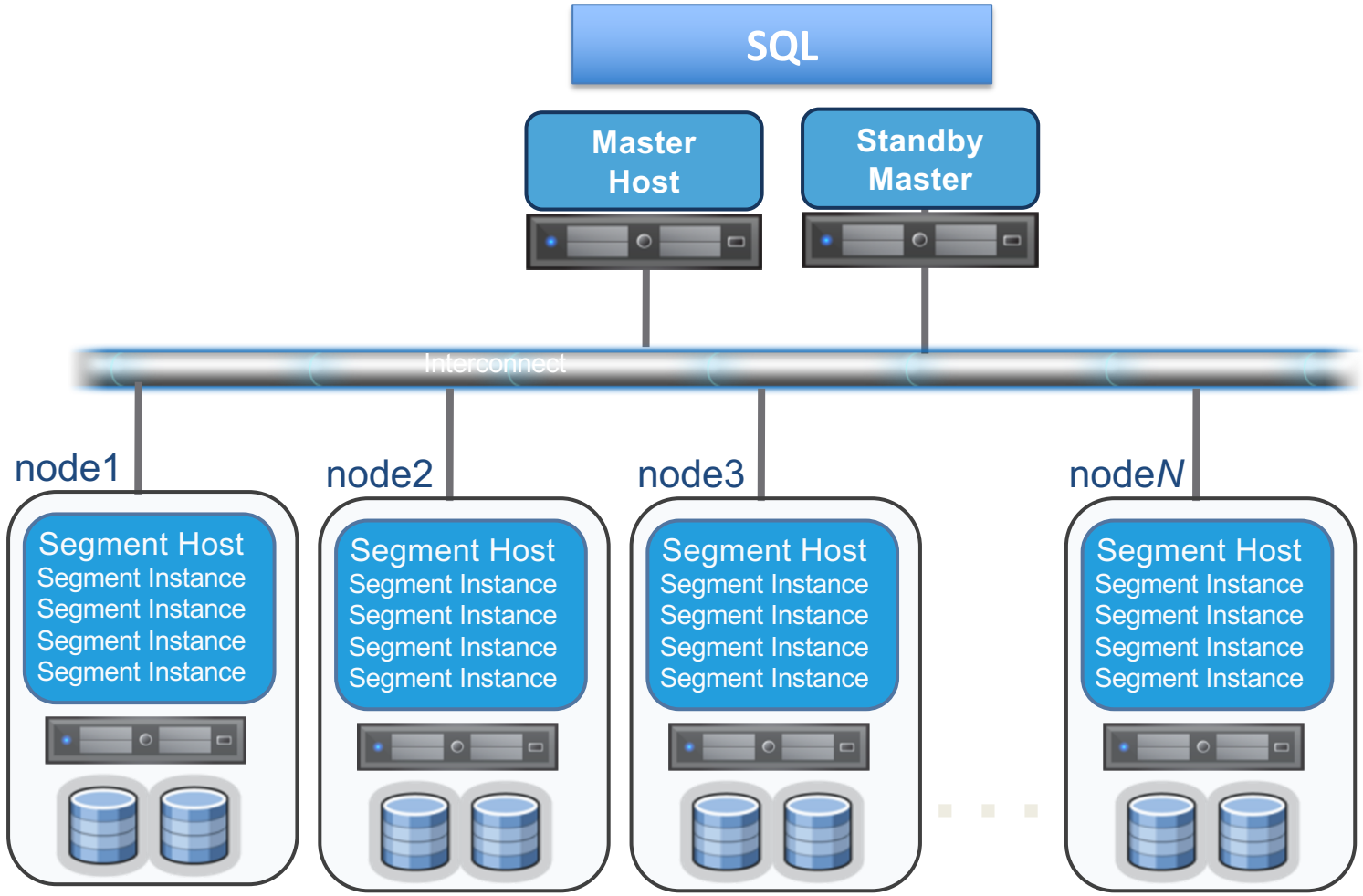
Performance Through Parallelism

Master Host connects with users and coordinates work with Segment Hosts

Segment Host Manages Data and Processes Queries

Segment Hosts have their own CPU, disk and memory (shared nothing)

High speed interconnect for continuous pipelining of data processing



Parallel Query Optimizer

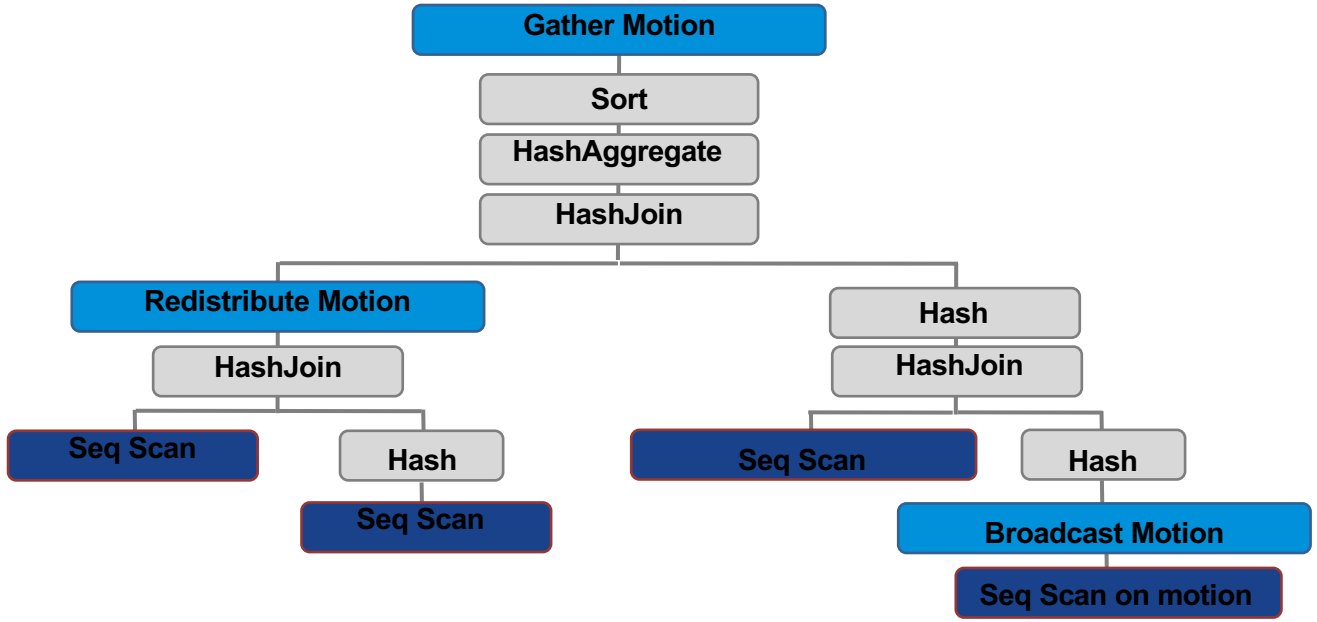
ORCA

Cost-based optimization looks for the most efficient query execution plan

Query execution plan composed of “slices” for scans, joins, sorts, aggregations, etc

Slices are performed in parallel across segment instances

Motion operators for inter-segment communication



Scalable Complex Correlated Queries

Common Table Expression Push Downs

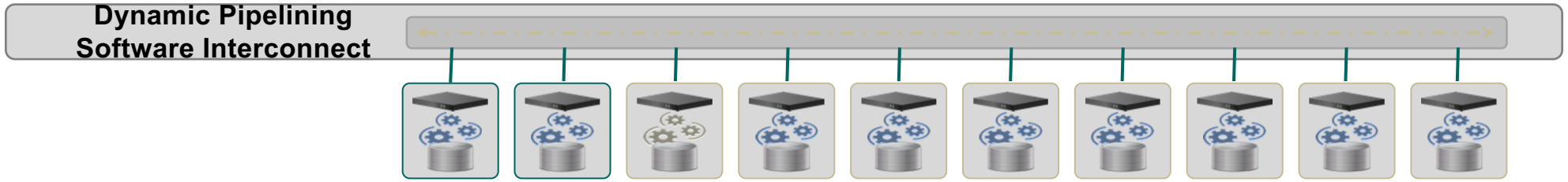
Dynamic Partition Elimination

Dynamic Pipelining

High speed interconnect

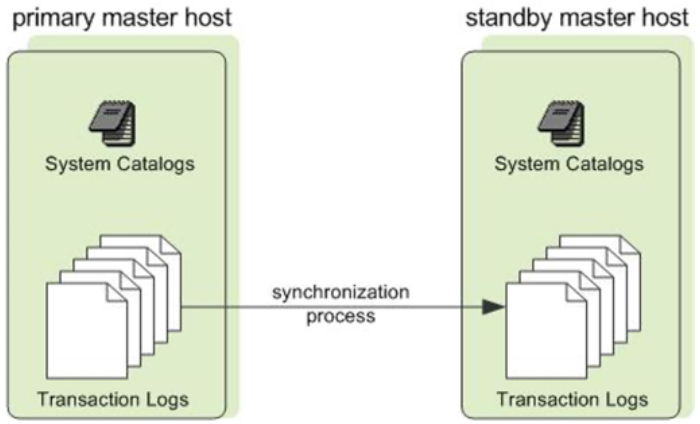
A supercomputing-based “soft-switch” responsible for

- Efficiently pumping streams of data between motion nodes during query-plan execution
- Delivers messages, moves data, collects results, and coordinates work among the segments in the system
- UDP or TCP Intersegment interconnect protocol

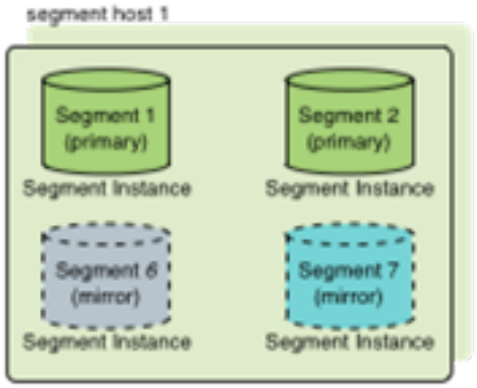


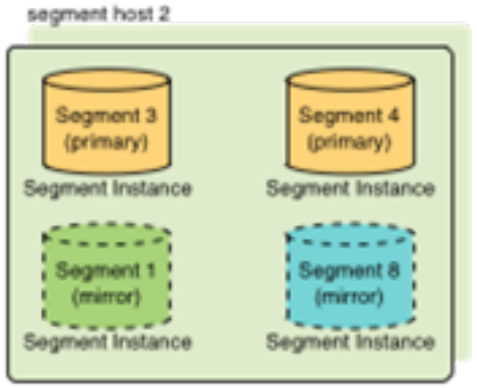
High Availability

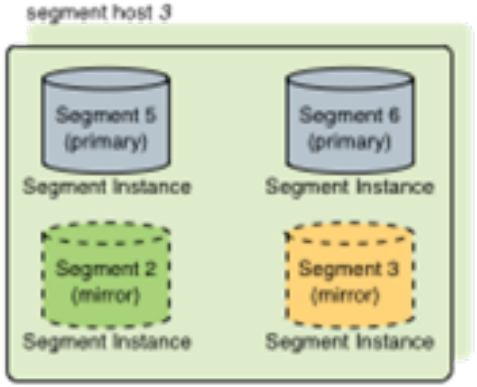
Master View

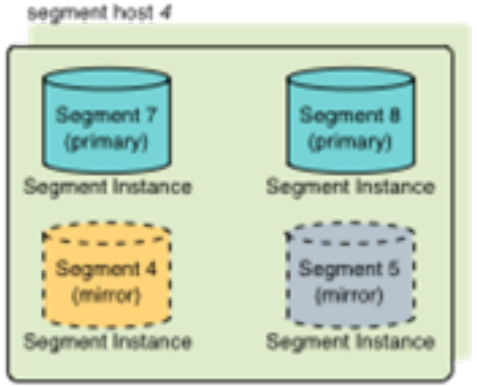


Segment View









- 2 copies of each segment data
- Automatic mirroring
- Automatic failover when hardware fails
- Proven in production over decade of runtime

Hybrid Transactional and Analytical Processing

Mixed Workloads for Analytics

Workload Management

Define resource groups to ensure allocation of allotted resource for each important workload

High Concurrency Analytics

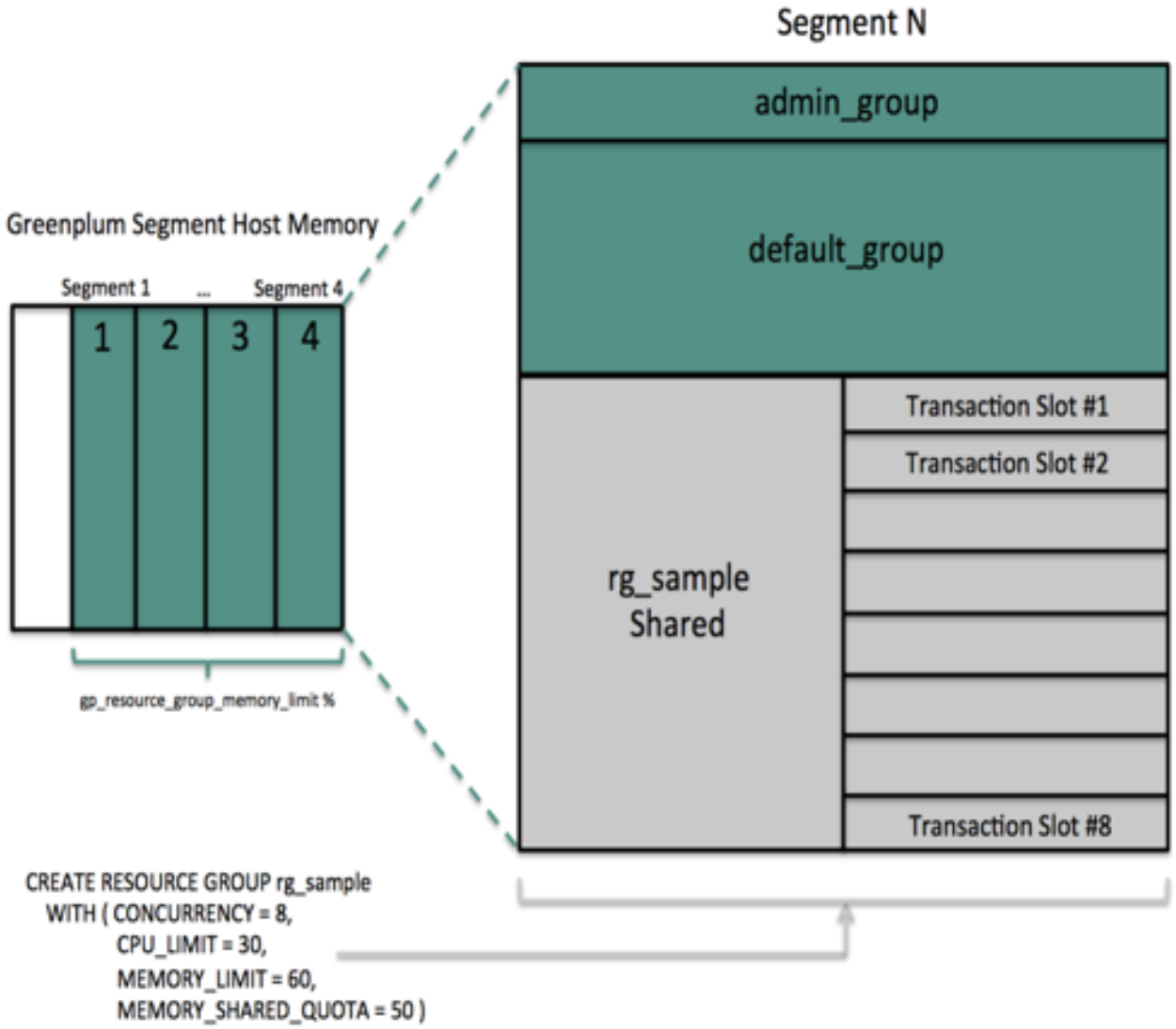
Hundreds of parallel complex queries run in parallel

Index Lookups

100,000 plus index lookups per second for targeted queries

Updates and Deletes

Thousands of concurrent updates and deletes on the same table enabled by row locking and low overhead distributed transactions



Simple Interface to Advanced Functions

Powered by Apache Madlib

Train (build a predictive model)

```
SELECT madlib.linregr_train( 'houses',                      -- Historical prices
                             'houses_linregr_bedroom',      -- Output model table
                             'price',                       -- Variable to predict
                             'ARRAY[1, tax, bath, size]',    -- Features
                             'bedroom'                      -- Diff models by #bedrooms
                           );
```

Predict (use model on new data)

```
SELECT houses_test.*,
       madlib.linregr_predict( model.coef,                  -- Trained model
                               ARRAY[1,tax,bath,size]       -- Features
                             ) as predicted_price
FROM houses_test, houses_linregr_bedroom as models
WHERE houses_test.bedroom = model.bedroom;
```

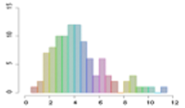
Simple Interface to Advanced Functions

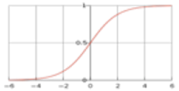
Powered by Apache Madlib

From house pricing model

id	tax	bedroom	bath	size	lot	predicted_price
1	590	2	1	770	22100	43223.5393423991
2	1050	3	2	1410	12000	111527.609949684
3	20	3	1	1060	3500	20187.9052986334
4	870	2	2	1300	17500	99354.9203362624
5	1320	3	2	1500	30000	124508.080626413
6	1350	2	1	820	25700	96640.8258367596
7	2790	3	2.5	2130	25000	224650.799707329
8	680	2	1	1170	22000	138458.174652714
9	1840	3	2	1500	19000	138650.335313723
10	3680	4	2	2790	20000	240000
11	1660	3	1	1030	17500	62911.27521866
12	1620	3	2	1250	20000	117007.693446415
13	3100	3	2	1760	38000	189203.861766405
14	2070	2	3	1550	14000	143322.539831872
15	650	3	1.5	1450	12000	82452.4386727394
etc...						

MADlib Functions




Supervised Learning
Neural Networks
Support Vector Machines (SVM)
Conditional Random Field (CRF)
Regression Models
- Clustered Variance
- Cox-Proportional Hazards Regression
- Elastic Net Regularization
- Generalized Linear Models
- Linear Regression
- Logistic Regression
- Marginal Effects
- Multinomial Regression
- Naïve Bayes
- Ordinal Regression
- Robust Variance

Tree Methods
- Decision Tree and Random Forest

Unsupervised Learning
Association Rules (Apriori)
Clustering (k-Means)
Principal Component Analysis (PCA)
Topic Modelling (Latent Dirichlet Allocation)

Deep Learning
Keras Fit/Evaluate/Predict
Load Model Architectures
Preprocessor for Images
Parallel Image Loading

Graph
All Pairs Shortest Path (APSP)
Breadth-First Search
Hyperlink-Induced Topic Search (HITS)
Average Path Length
Closeness Centrality
Graph Diameter
In-Out Degree
PageRank and Personalized PageRank
Single Source Shortest Path (SSSP)
Weakly Connected Components

Nearest Neighbors
- k-Nearest Neighbors

Time Series Analysis
- ARIMA

Sampling
Balanced
Random
Stratified

Statistics
Descriptive Statistics
- Cardinality Estimators
- Correlation and Covariance
- Summary

Inferential Statistics - Hypothesis Tests
Probability Functions

Data Types and Transformations
Array and Matrix Operations
Matrix Factorization
- Low Rank
- Singular Value Decomposition (SVD)

Norms and Distance Functions
Sparse Vectors
Encoding Categorical Variables
Path Functions
Pivot
Sessionize
Stemming

Utility Functions
Columns to Vector
Conjugate Gradient
Linear Solvers
- Dense Linear Systems
- Sparse Linear Systems

Mini-Batching
PMML Export
Term Frequency for Text
Vector to Columns

Model Selection
Cross Validation
Prediction Metrics
Train-Test Split

Graph Analytics

Terabyte and Petabyte Scale Analysis of Graphs

Graph Algorithms and Measures

Types		Question	Features
Graph-based Features	1 Group	*"What are the sub-graphs, component, communities?"*	weakly-connected component
	2 Structure	*"What is the character of the network structure?"*	Density, Diameter, Average path length, Modularity
	3 Path	*"What is the shortest path (distance) among vertices?"*	Single source shortest path, All pairs shortest path, Breadth-First Search
	4 Centrality	*"What are the most important vertices within a graph?"*	Degree (in/out, weight), Closeness, PageRank, Hub, Authority, Betweenness, Clustering coefficient

Social Network

* Grandjean, M. (2016)

Epidemiology

* http://www.netminer.com/community

Bank Risk

* https://cambridge-intelligence.com

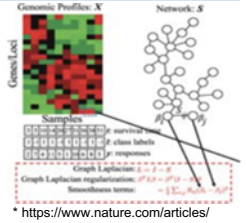
1st Party Fraud

* www.infoglide.com

MMO Role-Playing Game

* www.researchgate.net

Chemistry

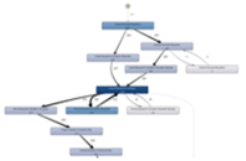


* https://www.nature.com/articles/

Gene

* www.researchgate.net

Manufacturing



* https://blog.trifinance.com

Deep Learning in our Super Computing Grid

GPU Accelerated

Train Neural Networks

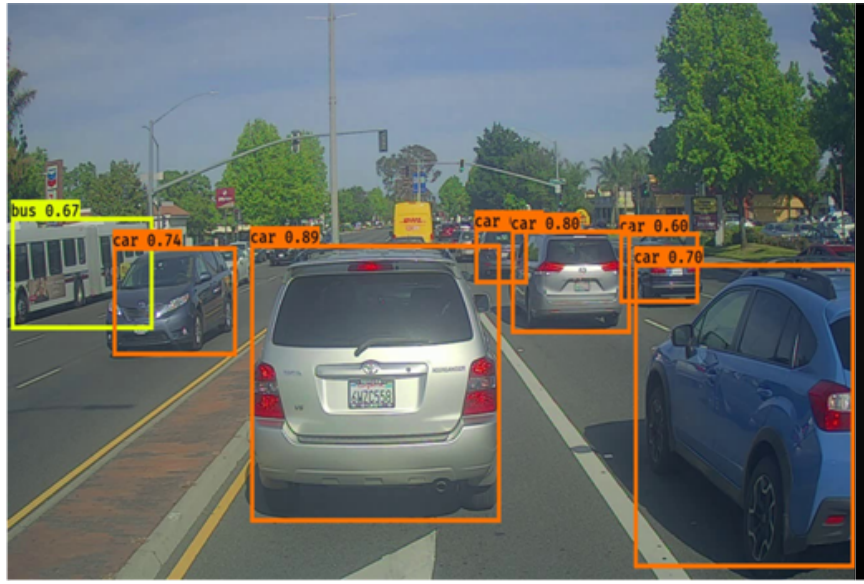
Use unstructured data like images and text and have Greenplum train models to recognize patterns for identification

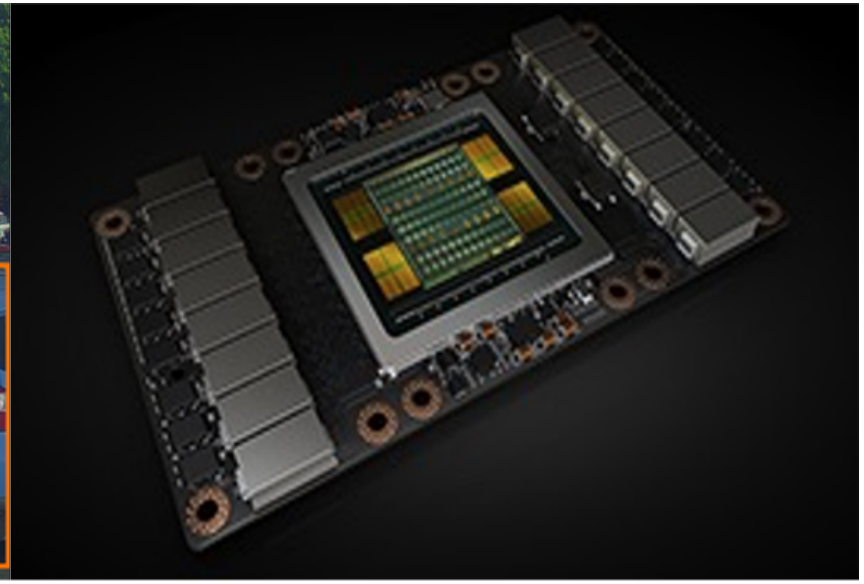
MPP Scale Performance

Train and compare thousands of models using the compute grid of Greenplum

Tensor Flow, Keras, GPUs

Industry standard libraries are used under the hood, complexity is managed for users by Greenplum





```
SELECT madlib_keras_fit('cifar10_train',                    -- training dataset
                        'cifar10_model',                    -- trained model weights
                        'model_arch_library',               -- model architecture table
                         1,                                 -- model architecture id
                         $$ loss='categorical_crossentropy', optimizer='adam',
                            metrics=['accuracy'] $$,        -- compile parameters
                         $$ batch_size=32, epochs=3 $$,     -- fit parameters
                         20,                                -- number iterations
                         4,                                 -- GPUs per host
                        'cifar10_test'                      -- test/validation dataset
                        );
```

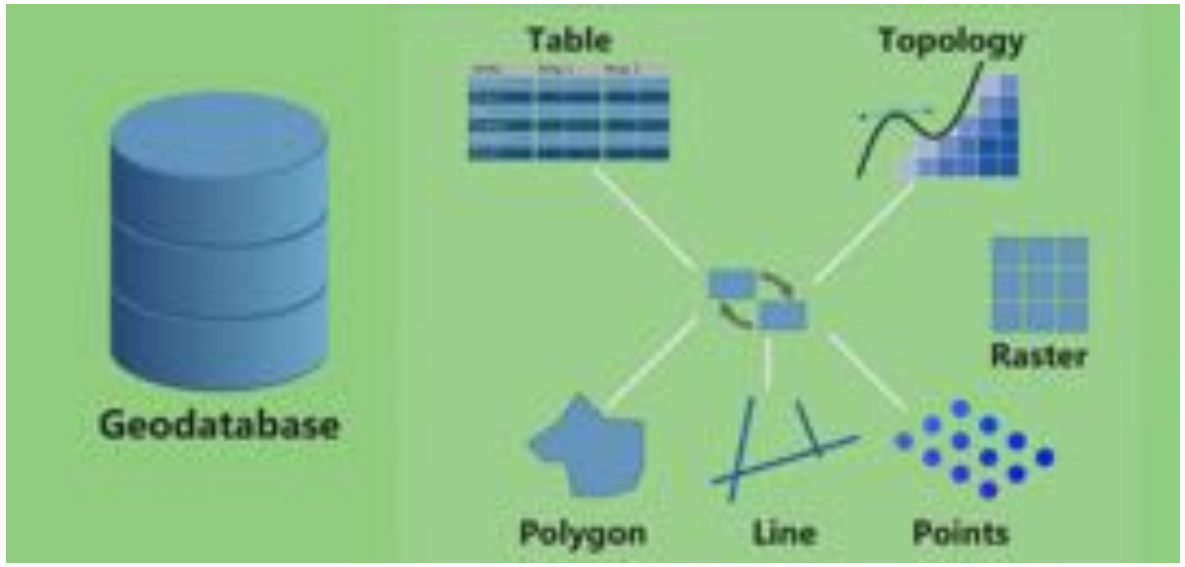
GeoSpatial Analytics

Storage and Query of Geo Data

Turn your big data database in a Geo

database to store, search and analyze

data based on locations




The UTM grid.

Text Search & Analytics

Index and find matching documents

- **Extract** data from binary or human readable formats into data that a machine can understand and operate on.
- **Index** the text data, so we can quickly search for specific text and documents.
- **Search** the text for patterns and keywords.
- **Analyze** what the text actually means.

Greenplum R

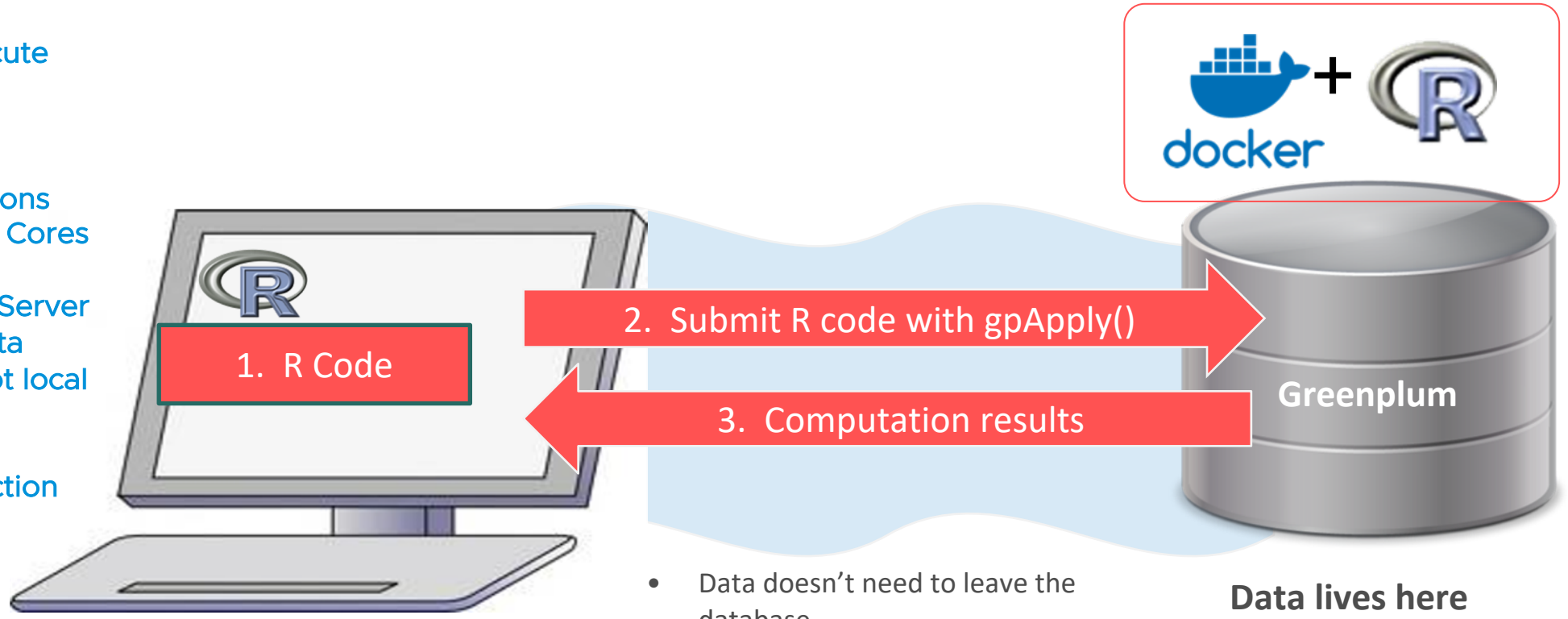
Server Side Compute Grid

Test Locally and Execute Remotely

Massively Parallel Execution of R Functions on thousands of CPU Cores

Secure Execution on Server Side & Security of Data Living Greenplum (not local data)

Dynamic push of function code, no UDF writing needed



- Data doesn't need to leave the database
- Push code to server

No data here

Data lives here

Procedural Programming Languages

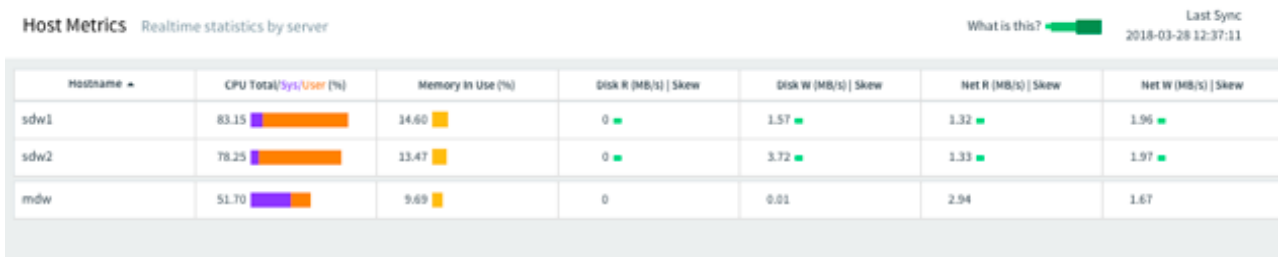
Custom User Defined Functions

Server-Side Functions

- Process data row by row
- Massively parallel execution model
- Transform each row using a procedural language
- Security via containerized execution when needed
- Import OSS libraries for advanced functions (e.g. NLTK)
- Import enterprise libraries for acess to your proprietary code logic
- User defined aggregates for grouping
- Call OSS Machine Learning algorithms

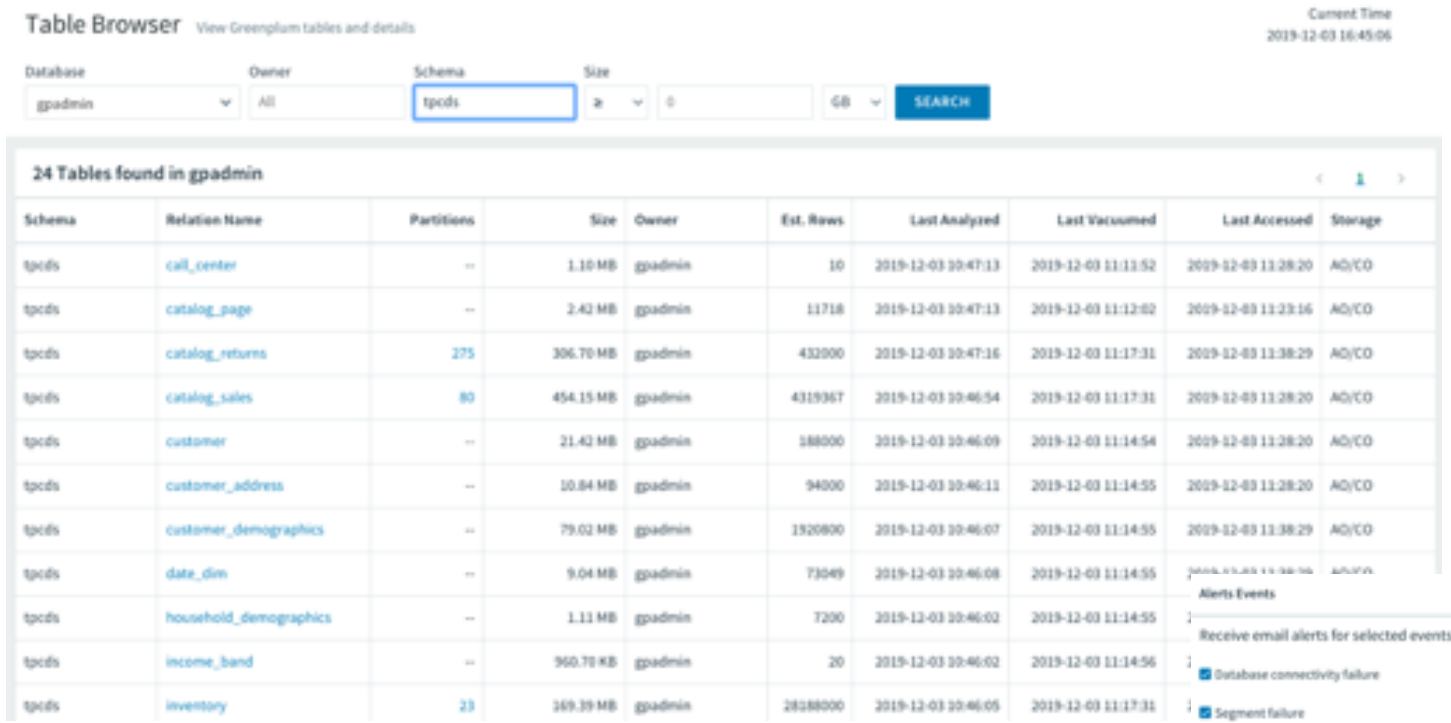
Greenplum Command Center

Single Pane of Glass for Greenplum Database Administrators



Greenplum Command Center

Alerts and Table Browser

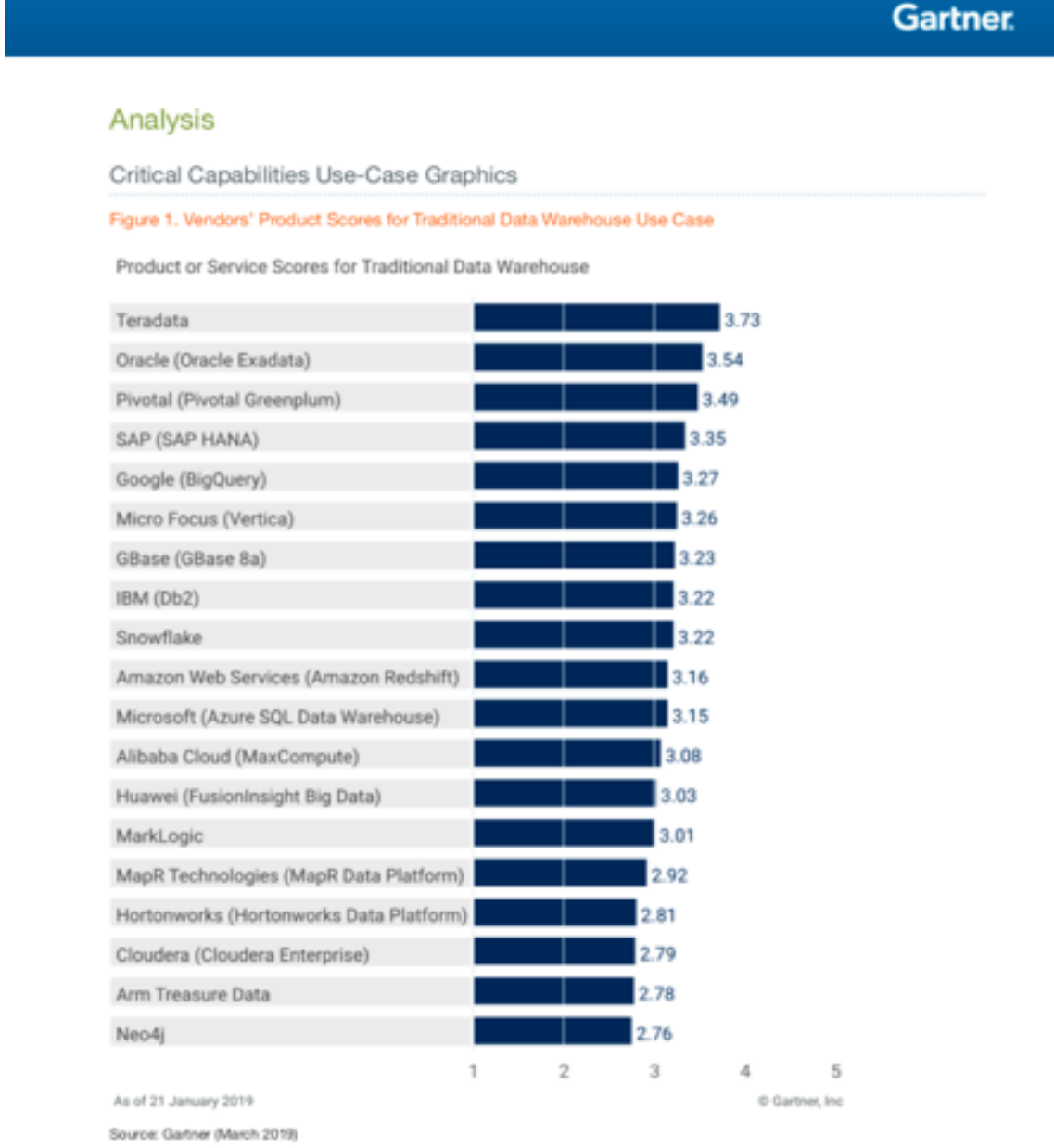


Alerts Management

Alerts Management

Gartner Loves Greenplum

Ranked Number 1 Open Source Data Warehouse in the World!



VMware Tanzu Greenplum Roadmap

April 2020

Future Looking Statements

Disclaimer

- ***Presentations may contain product features or functionality that are currently under development.***
- ***This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.***
- ***Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.***
- ***Technical feasibility and market demand will affect final delivery.***
- ***Pricing and packaging for any new features/functionality/technology discussed or presented, have not been determined.***
- ***This information is confidential.***

VmWare Tanzu Greenplum Roadmap

Contents of Presentation

1. Greenplum Platform Next Generation
2. DBA Operational Improvements
3. ETL and Data Integration Improvements
4. Analytics & Data Science Enhancements
5. Server Feature Release & Support Calendar

Introducing VMware Tanzu Greenplum

Greenplum Platform Roadmap

Expand on Greenplum's traditional strength on bare metal with VMware's virtualization expertise

Off Platform

- Bare Metal
- Public Cloud

Virtualized

- Vsphere
- VxRail

Containerized

- Kubernetes
- VMware Cloud

Foundation

Greenplum on vSphere

Testing and certification of vSphere platform

- Greenplum has been supported on vSphere for over 5 years
- Used heavily in test & dev scenarios
- Used in small to medium production scenarios

We can do more!

- Targeting large production clusters
- VMware Ready Node specification certifications
- Optimal configuration options documented
- Monitoring, troubleshooting and tuning guides
- VM deployment automation

Greenplum For Kubernetes Value Proposition

Advantages of containerization

Speed

Deploy in minutes

Consistently Repeatable

Agile Analytics Workbench

Savings

Operational Efficiency With No Mirror Configurations Leveraging Central Storage

Leverage org's K8s skills

Quick New User Ramp Up

Security

Pre-hardened

Pre-networked

Secured Docker Image

Stability

Automated Recovery

Resource availability for recovery

Faster Upgrade / Patching

No Degraded Mode

Build CI / CD Pipelines

Scalability

Self-Service Deployments

K8s Volume Expansion

Compute-Storage Separated

Available Now!

Continued optimization & tuning ongoing

Full stack component GP component list with Greenplum 6 in development

Greenplum Building Blocks on VxRail

Next gen platform appliance available for Greenplum

- 2010: EMC DCA v1, 2012 EMC DCA, 2016 EMC DCA V3, 2018 Dell Building Blocks
- 2020: Greenplum VxRail Building Blocks
- Software defined architecture consolidates compute, storage, virtualization, and management
- Smart Fabric Services for VxRail automates network setup, simplifying and accelerating deployment
- Provides a single point of support by default for all software and hardware
- Integrated vSan Storage provides mirroring at storage level not DB segment level
- VxRail tech specs per host:

56 Cores, 1.5 TB RAM, 76 TB All Flash Storage, Write Caching 2x25 gb network for interconnect, 2x25 gb network for vSan storage

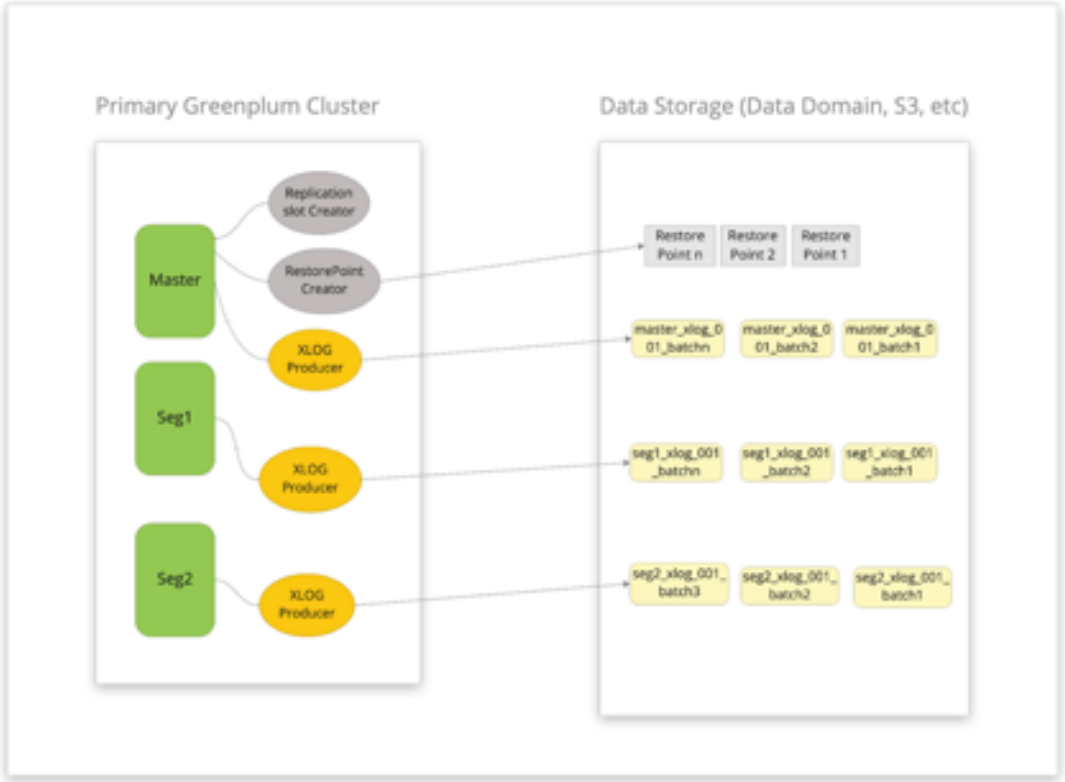
Greenplum DBA Operational Improvements

Multi Site Replication

- WAL **streaming** across two data centers
- Read-only live mirror cluster
- Allow for user defined, consistent Restore Points across all segments
- Supports failover and failback
- Maximum data availability

Point In Time Recovery

- WAL **archiving** on existing storage (Data Domain, S3, etc)
- A secondary Greenplum cluster is initialized, and Recovered up to an existing, defined Restore Point
- May be used to recover specific objects on a specific date
- May be used to seed Dev/QA databases



Greenplum Command Center

Greenplum Roadmap

Autonomous Database Features

Recommendation engine for key tasks like vacuum and analyze

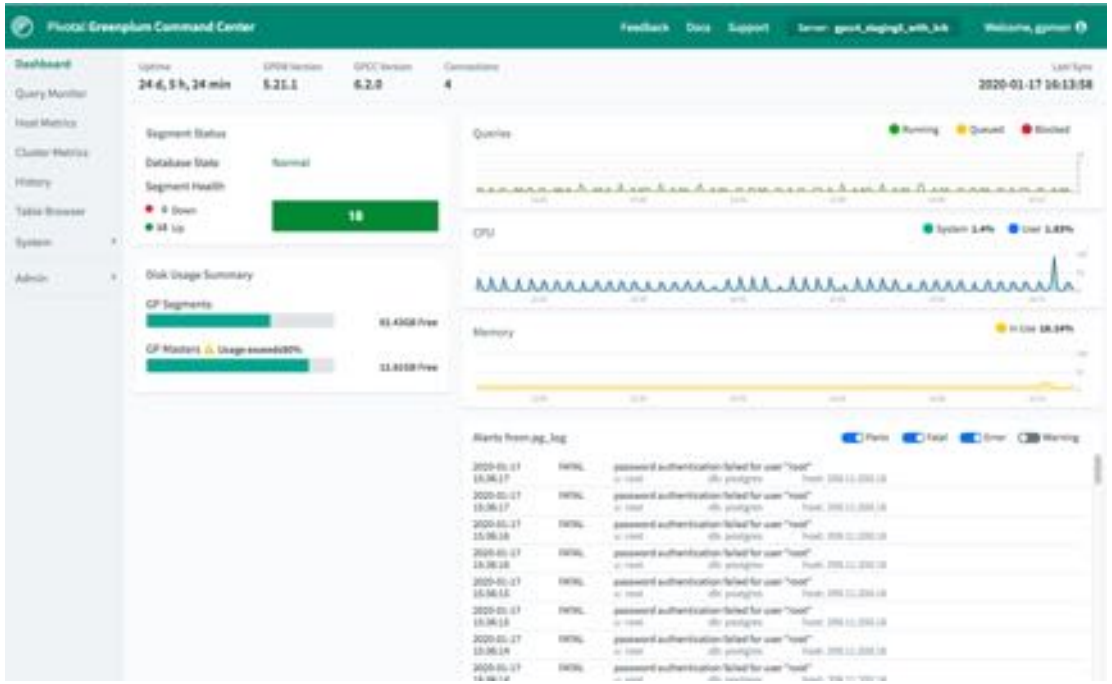
In-App notifications for alerts in the dashboard

Solutions for More Scenarios

Plugin framework to create custom screens in GPCC

New plugins for:

* Greenplum Text
* PXF
* Streaming Server
* Kafka
* GP Backup Manager
* Greenplum For Kubernetes



Greenplum Database Server Release Calendar and Roadmap

Greenplum Releases

Greenplum Roadmap

Greenplum 4.3

Initial release: March 2015

Current Status: Maintenance Mode

End of General Support: Nov 30, 2020

Greenplum 5.X

Initial release: Sep 2017

Current Status: Limited Feature Release

End of General Support: TBD, targeted for 18 months from last minor version

Greenplum 6.X

Initial release: Sep 2019

Current Status: Active Development

End of General Support: TBD

Greenplum 7.X

Target release beta: March-2021

Target release GA: Sep-2021 GA

Current Status: Active Development; Not released yet

End of General Support: TBD

Greenplum Component Releases Ongoing

Greenplum Command Center

Greenplum Backup Manager

Greenplum Streaming Server

Greenplum Data Copy Utility

Greenplum Text

Apache Madlib

Greenplum 7 Server

Greenplum 7 Roadmap

Postgres Merge

Greenplum 7 is targeting Postgres 12

Greenplum 6 is based on Postgres 9.4

Greenplum 5 is based on Postgres 8.3

Query Performance

BRIN indices enable tracking of min and max values per block to bypass IO and speed analytical queries (similar to zone maps) and allow for immediate simple answers from your data

Postgres **parallel query execution** for CPU intensive operators enable elastic scaling up and down to meet available CPU resources

Just in Time (JIT) compilation allows rewriting of machine code execution at run-time speed analytical query execution

Query Federation

Statistics on external data allow complex workloads to run directly on external data leveraging optimized query plans

Improved Resource Utilization for PXF

Parallel Scan Operators for PXF

PXF caching to reduce IO to external data sets

Greenplum to Greenplum (GP2GP) foreign data wrapper allows multi-cluster architectures and cross cluster queries with MPP performance

DBA Operations

WAL Named Restore Points to support DR and PITR

Auto-vacuum on the catalog allows GPDB to run optimally with less attention from DBAs

Thank You

VMWare Tanzu Greenplum

Open Source Software Vs Commercial

Apache License V2.0

"Benevolent Dictator"

- VMware is the Steward

Value Proposition

- Added value to OSS
- Best Support

Naming

- Commercial: VMWare Tanzu Greenplum
- OSS: Greenplum Database

Available in VMWare Tanzu Greenplum Only

- Product packaging and installation scripts
- Greenplum Command Center
- Greenplum Workload Manager
- Text Analytics (GPText)
- Compression: Support for QuickLZ/ZStd
- Connectors:
  - Apache Spark Connector
  - VMware Tanzu Gemfire Connector
  - Apache Kafka Connector (gpkafka)
  - Apache Nifi Connector (Coming soon)
- Progress ODBC/JDBC drivers
- Premium Support