

Pivotal®

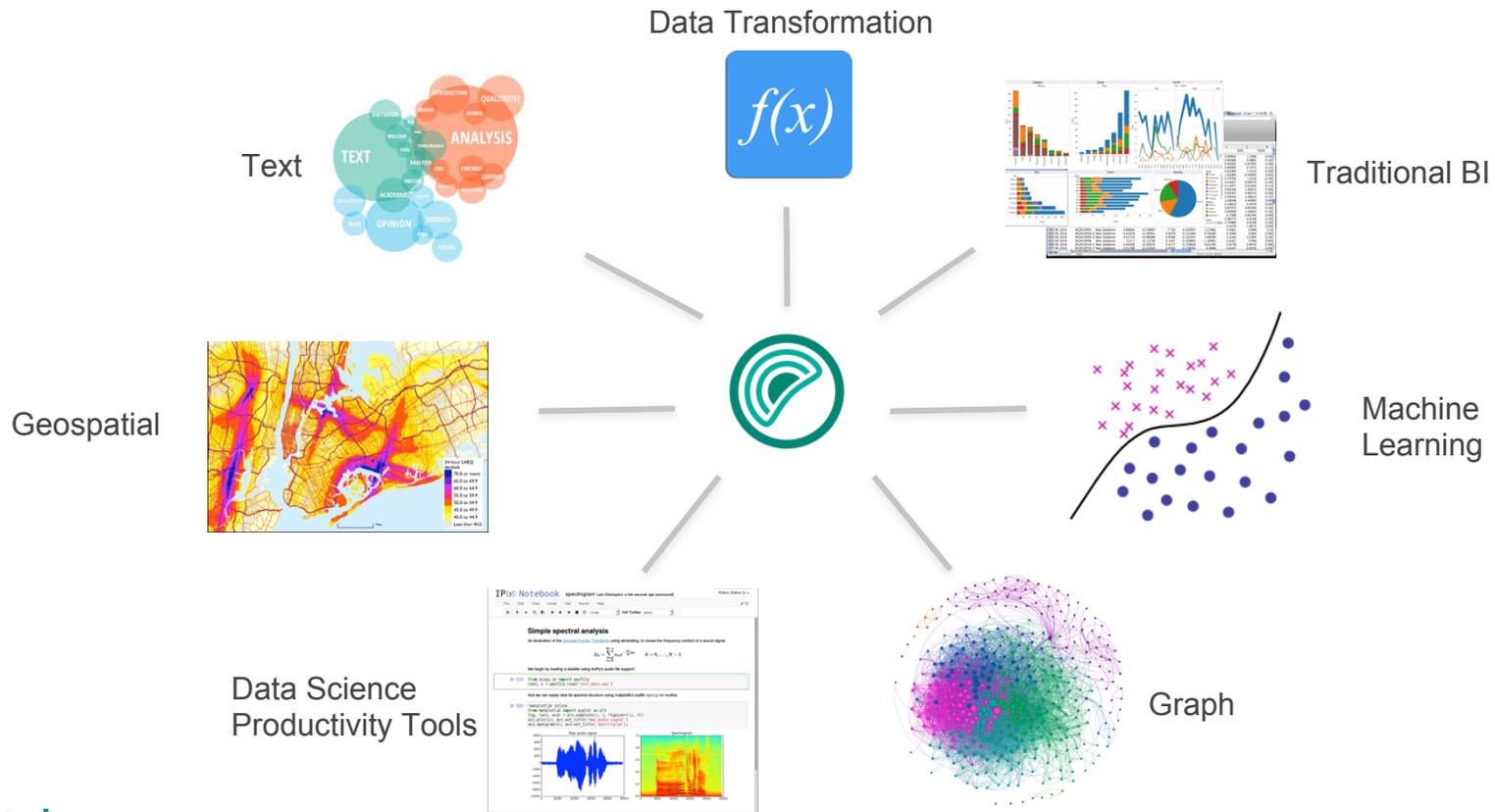
# Machine Learning, Graph, Text and Geospatial on PostgreSQL and Greenplum



---

Frank McQuillan  
Bharath Sitaraman

# Greenplum Integrated Analytics



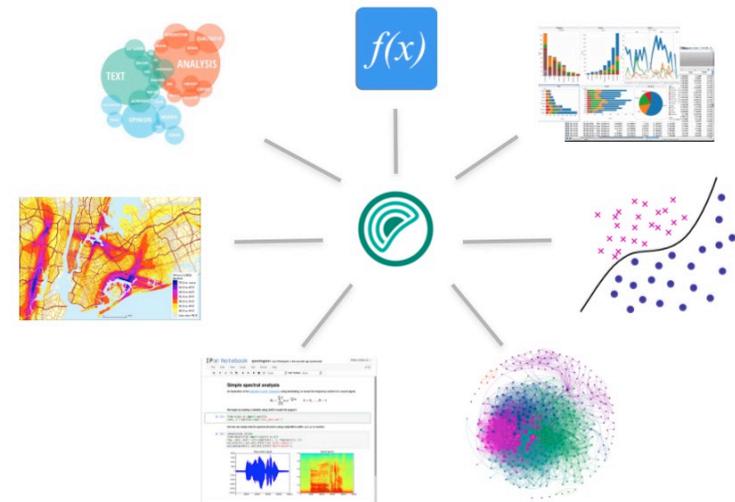
Pivotal

Pivotal

# Agenda

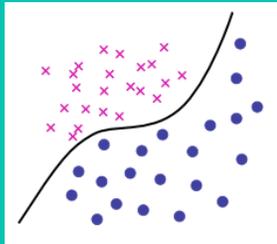


1. Machine learning with Apache MADlib
2. Data transformation
3. Graph
4. Data science productivity tools
5. Geospatial with PostGIS
6. Text analytics with GPText
7. Connectivity
8. Example use cases
9. Looking ahead





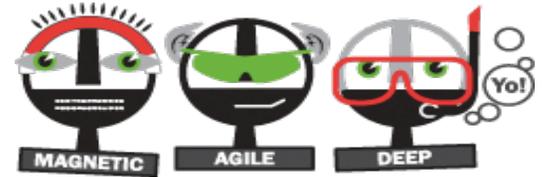
# 1. Machine Learning with Apache MADlib



Pivotal



# Scalable, In-Database Machine Learning



## Apache MADlib: Big Data Machine Learning in SQL



Open source,  
top level  
Apache project

For PostgreSQL  
and Greenplum  
Database



Powerful machine  
learning, graph,  
statistics and analytics  
for data scientists

- Open source
- Downloads and docs
- Wiki

<https://github.com/apache/madlib>

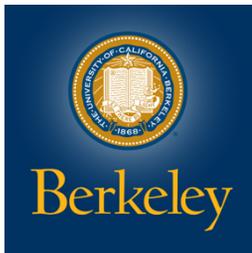
<http://madlib.apache.org/>

<https://cwiki.apache.org/confluence/display/MADLIB/>

# History



MADlib project was initiated in 2011 by EMC/Greenplum architects and Professor Joe Hellerstein from University of California, Berkeley.



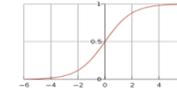
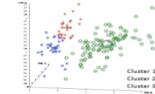
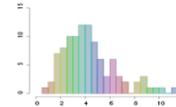
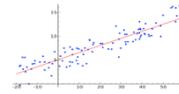
UrbanDictionary.com:

**mad** (adj.): *an adjective used to enhance a noun.*

1- dude, you got skills.

2- dude, you got **mad** skills.

# MADlib Functions



Comprehensive and mature  
data science library

## Supervised Learning

- Neural Networks
- Support Vector Machines (SVM)
- 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
  - Random Forest
- Conditional Random Field (CRF)

## Graph

- All Pairs Shortest Path (APSP)
- Breadth-First Search
- Hyperlink-Induced Topic Search (HITS)
- Average Path Length
- Closeness Centrality
- Graph Diameter
- In-Out Degree

## Data Types and Transformations

- Array and Matrix Operations
- Matrix Factorization
  - Low Rank
  - Singular Value Decomposition (SVD)
- Norms and Distance Functions
- Sparse Vectors
- Principal Component Analysis (PCA)
- Categorical Variables

## Unsupervised Learning

- Association Rules (Apriori)
- Clustering (k-Means)
- Topic Modelling (Latent Dirichlet Allocation)

## PMML Export

- Sampling
  - Balanced
  - Random
  - Stratified
- Sessionize
- Term Frequency for Text Analysis

## g

- s
- ve Statistics
- inality Estimators
- elation and Covariance
- mary
- al Statistics
- Hypothesis Tests
- Probability Functions

## Nearest Neighbors

- k-Nearest Neighbors

## Time Series Analysis

- ARIMA

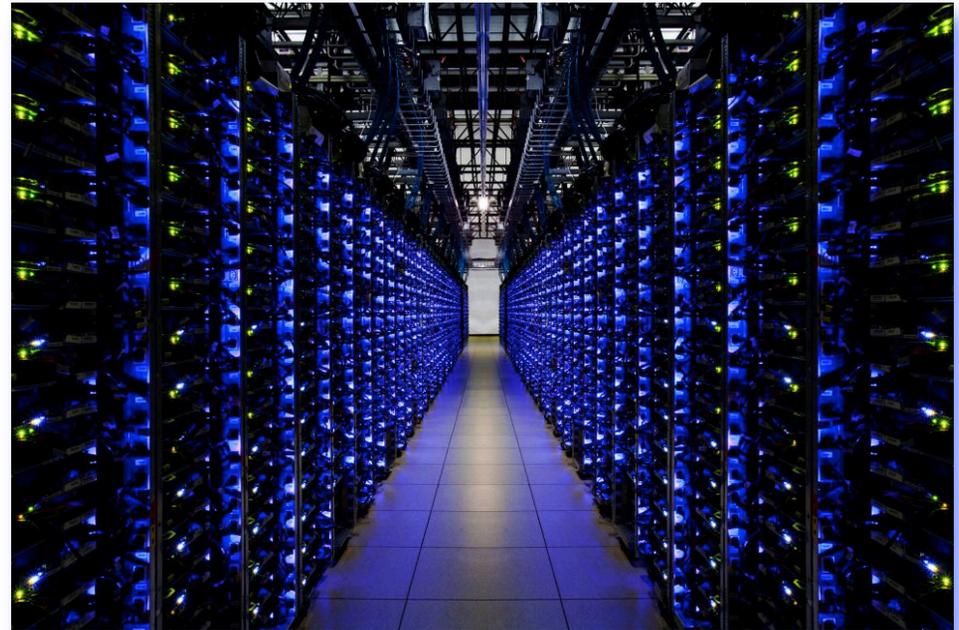
## Model Selection

- Cross Validation
- Prediction Metrics
- Train-Test Split

# Why MADlib on Greenplum?

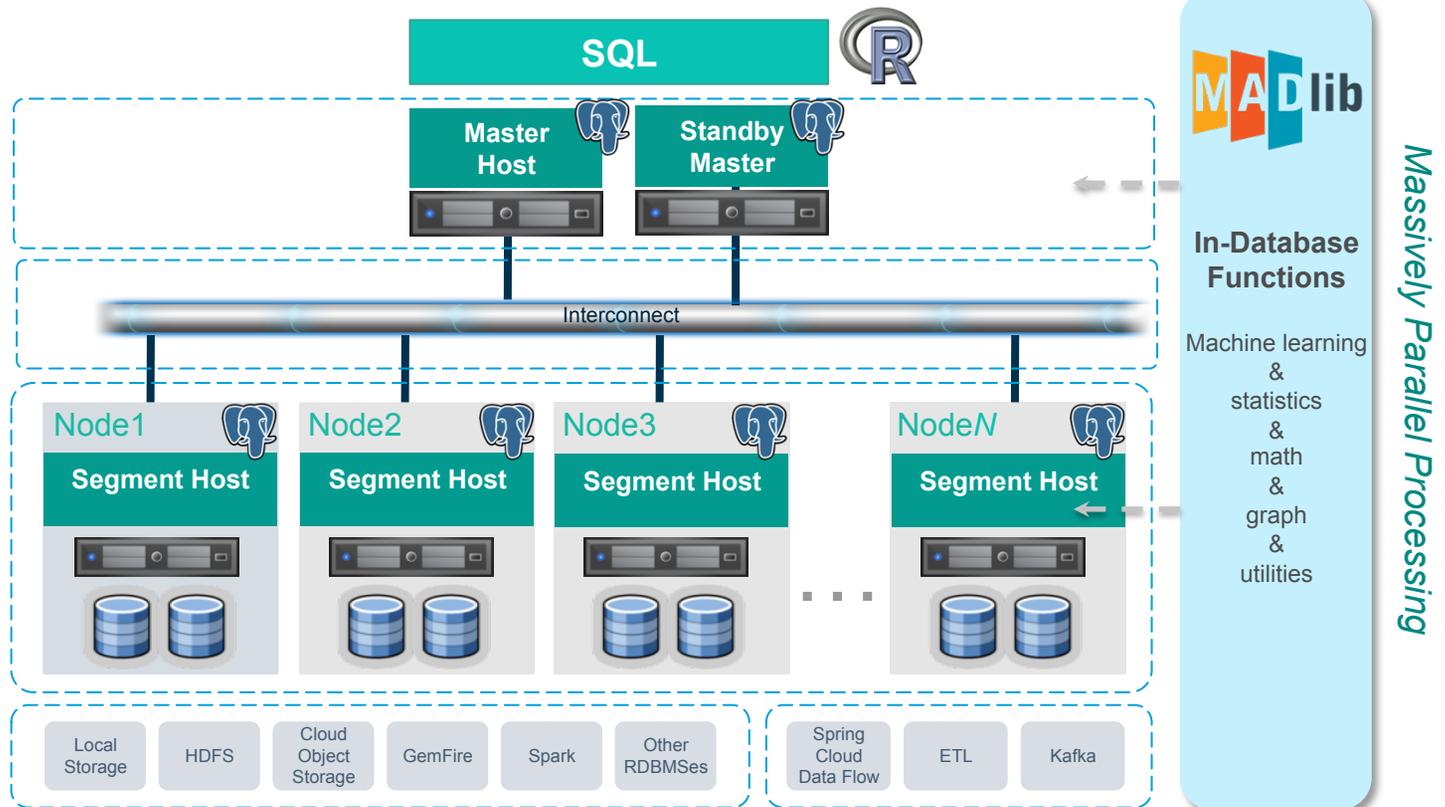


- Better parallelism
- Better scalability
- Higher predictive accuracy
- Top level ASF project



“Apache MADlib Comes of Age”, Frank McQuillan, Oct. 2017,  
<https://content.pivotal.io/blog/apache-madlib-comes-of-age>

# Greenplum Database with MADlib



# Familiar SQL Interface



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;
```

Pivotal

Pivotal

# Familiar SQL Interface

↓ From house pricing model

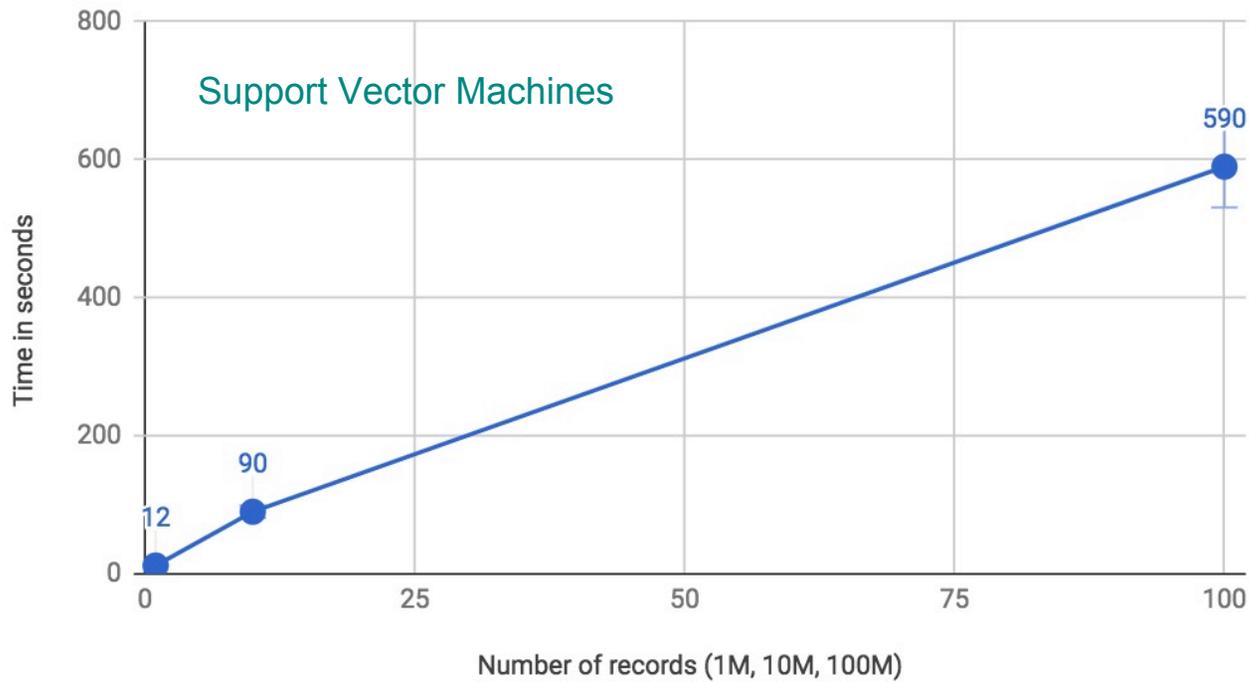
id	tax	bedroom	bath	size	lot	predicted_price
1	590	2	1	770	22100	43223.5393423991
2	1050	3	2	1410	12000	111527.609949684
3	20	3	1	1060	3500	20187.9052986334
4	870	2	2	1300	17500	99354.9203362624
5	1320	3	2	1500	30000	124508.080626413
6	1350	2	1	820	25700	96640.8258367596
7	2790	3	2.5	2130	25000	224650.799707329
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...						



# Built to Scale



Classification, 100 features, no grouping



Greenplum cluster:

- 1 master
- 4 segment hosts with 6 segments per host



## 2. Data Transformation



Pivotal

# Native PostgreSQL Data Transformations

- Rich library of functions and operators
  - Array functions
  - Aggregate functions
  - Window functions



```
SELECT
  product_name,
  price,
  group_name,
  AVG (price) OVER (PARTITION BY group_name)
FROM
  products;
```

product_name	price	group_name	avg
HP Elite	1200.00	Laptop	850.0000000000000000
Lenovo Thinkpad	700.00	Laptop	850.0000000000000000
Sony VAI0	700.00	Laptop	850.0000000000000000
Dell Vostro	800.00	Laptop	850.0000000000000000
Microsoft Lumia	200.00	Smartphone	500.0000000000000000
HTC One	400.00	Smartphone	500.0000000000000000
Nexus	500.00	Smartphone	500.0000000000000000
iPhone	900.00	Smartphone	500.0000000000000000
iPad	700.00	Tablet	350.0000000000000000
Kindle Fire	150.00	Tablet	350.0000000000000000
Samsung Galaxy Tab	200.00	Tablet	350.0000000000000000

(11 rows)

“Comparing Window Function Features by Database Vendors”, Jiri Mauritz, Sonra Intelligence, Sept. 15, 2017



# Data Transformations



Array and Matrix Operations  
Conjugate Gradient  
Encoding Categorical Variables  
Linear Solvers

- Dense Linear Systems
- Sparse Linear Systems

Matrix Factorization

- Low Rank
- Singular Value Decomposition (SVD)

Norms and Distance Functions

Path



Pivot  
PMML Export  
Principal Component Analysis (PCA)  
Sampling

- Balanced
- Random
- Stratified

Sessionize  
Sparse Vectors  
Stemming  
Term Frequency for Text Analysis

# Path Functions in E-commerce



?



?

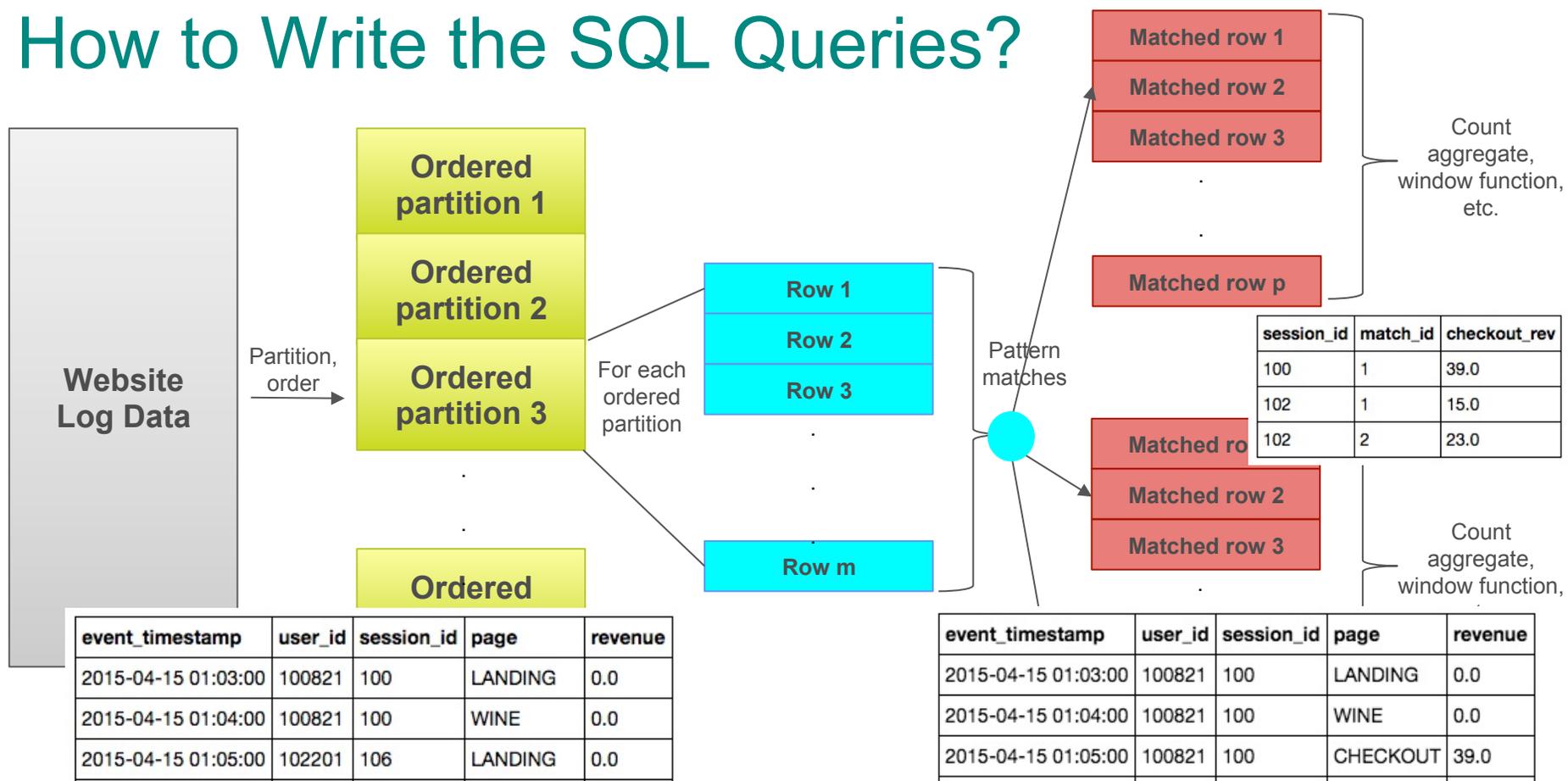




# Raw Data from Website Logs

event_timestamp	user_id	session_id	page	revenue
2015-04-15 01:03:01	100821	100	LANDING	0
2015-04-15 01:03:14	100829	200	LANDING	0
2015-04-15 01:03:19	100839	300	LANDING	0
2015-04-15 01:04:00	100839	300	WINE	0
2015-04-15 01:04:00	100829	200	WINE	0
2015-04-15 01:04:21	100821	100	WINE	0
2015-04-15 01:05:00	100829	200	CHECKOUT	59
2015-04-15 01:05:00	102204	206	LANDING	0
2015-04-15 01:05:00	102224	306	LANDING	0
2015-04-15 01:05:01	100839	300	CHECKOUT	19
2015-04-15 01:05:21	102201	106	LANDING	0
2015-04-15 01:05:44	100821	100	CHECKOUT	39
2015-04-15 01:06:00	102224	306	HELP	0
2015-04-15 01:06:44	102201	106	HELP	0
etc...				

# How to Write the SQL Queries?



# MADlib Path Functions



```
SELECT madlib.path(  
  'eventlog',           -- Name of input table  
  'path_output',       -- Table name to store path results  
  'user_id, session_id', -- Partition input table by user and session  
  'event_timestamp ASC', -- Order partitions in input table by time  
  
  $$ land:=page='LANDING',  
  wine:=page='WINE',  
  beer:=page='BEER',  
  buy:=page='CHECKOUT',  
  other:=page<>'LANDING' AND page<>'WINE' AND page<>'BEER' AND page<>'CHECKOUT'  
  $$,                  -- Symbols for page types  
  
  '(land) [^(land)(buy)]{0,2}(buy)', -- Pattern for purchase within 4 pages  
  
  'sum(revenue) as checkout_rev', -- Sum revenue by checkout  
);
```

Partition and order

Define symbols

Pattern match across rows

Sum revenue

“Path Functions in Apache MADlib”, Frank McQuillan, May 2016,  
<https://content.pivotal.io/blog/path-functions-in-apache-madlib>



# High Value Quick Shoppers



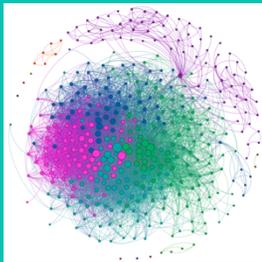
user_id	session_id	checkout_rev
101163	302	75
100829	200	59
101123	202	55
100821	100	39
101163	302	33
101121	102	23
100839	300	19
101121	102	15
101123	202	13
etc...		



Sorted descending by revenue



# 3. Graph

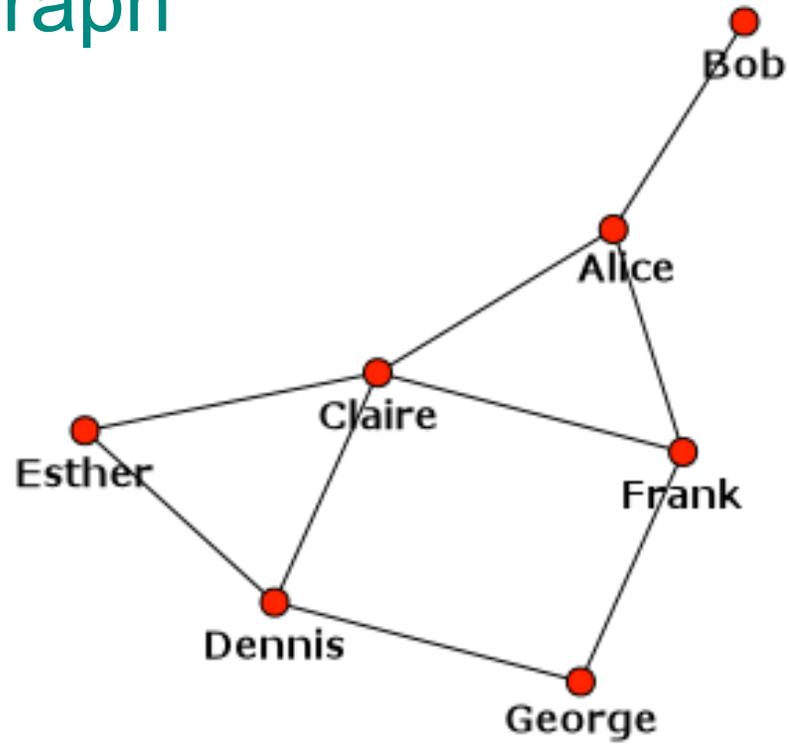


Pivotal

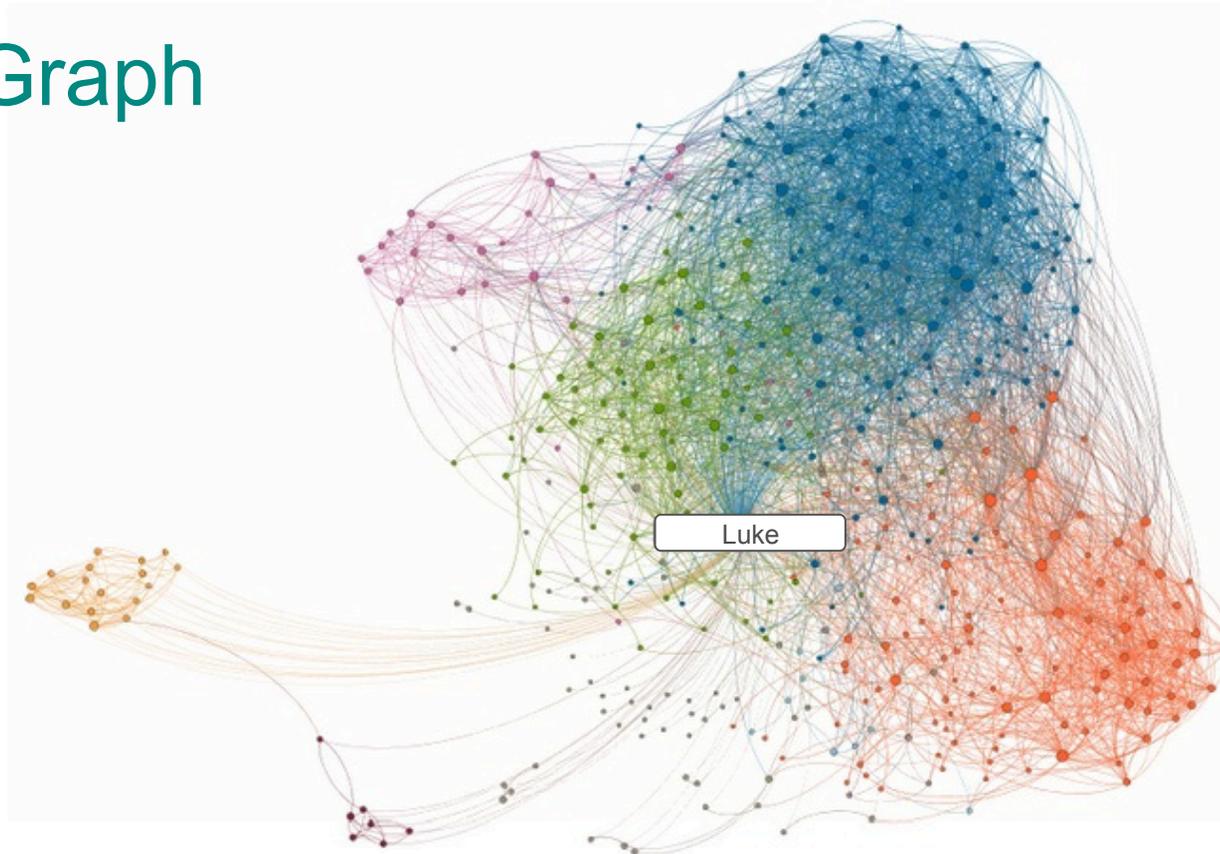


**MPP databases are an effective tool for  
*graph analytics at scale in enterprise***

# A Small Graph

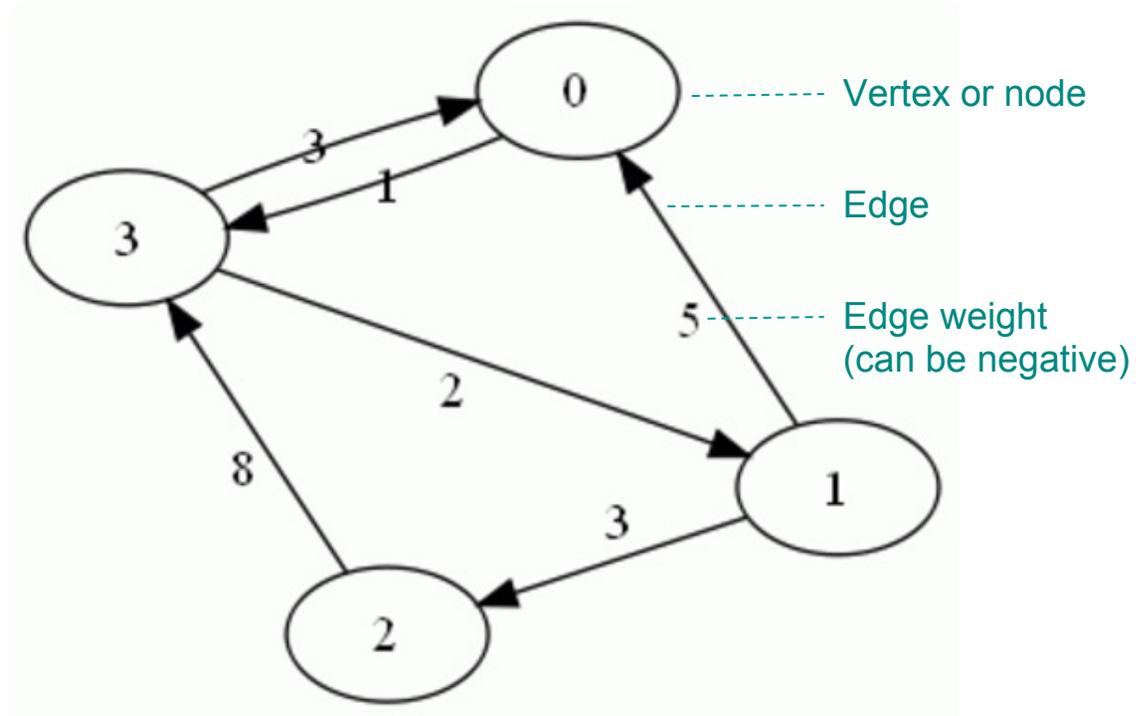


# A Big Graph



Sample LinkedIn social graph

# Directed Graph





# Graph Representation in MADlib

*Vertex Table*

<b>Vertex</b>	<b>Vertex Params</b>	...
0	...	
1	...	
2	...	
3	...	

▪  
▪  
▪

*Edge Table*

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

▪  
▪  
▪

# PageRank



- Web search
- Scientific impact of researchers
- Street and space usage
- Neuroscience

Google

Google Search

I'm Feeling Lucky

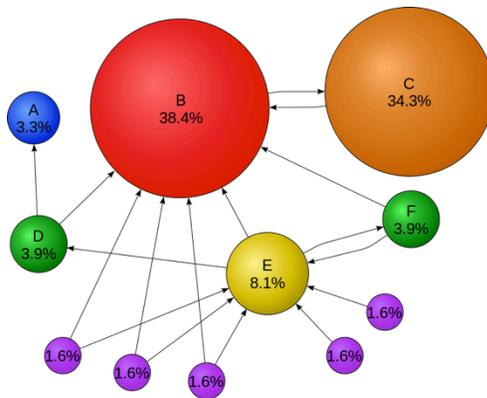


Image from <https://en.wikipedia.org/wiki/PageRank>

Pivotal

# PageRank in MADlib



```
SELECT pagerank(  
    'vertex',           -- Vertex table name  
    'id',              -- Vertex id column  
    'edge',            -- Edge table name  
    'src=start_id, dest=end_id', -- Edge source and dest columns  
    'pagerank_out'    -- Output table with PageRank  
);
```

id	pagerank
0	0.287518161212111
3	0.210171199451415
2	0.146637377532288
4	0.102910437211324
1	0.102910437211324
6	0.097274664434317
5	0.0525777229481976
etc...	



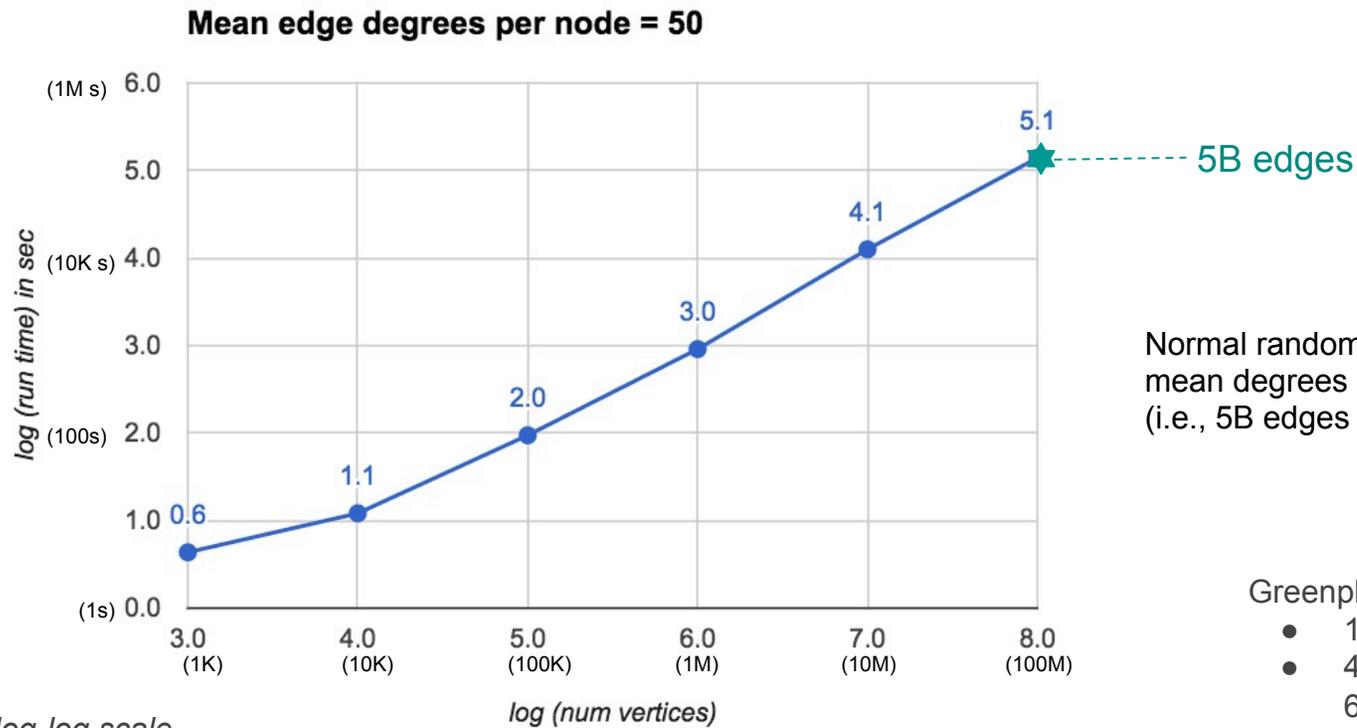
# PageRank in MADlib

```
SELECT pagerank(  
    'vertex_table',           -- vertex table  
    'vertex_id',            -- col in vertex table containing vertex IDs  
    'edge_table',           -- edge table  
    'edge_args',            -- source, dest and edge weights col in the edge table  
    'out_table',            -- output table with PageRank distribution  
    'damping_factor',       -- damping  
    'max_iter',             -- maximum iterations  
    'threshold',           -- stopping criterion  
    'grouping_cols',        -- grouping columns for multiple PageRank distributions  
    'personalization_vertices' -- for personalized PageRank  
);
```

Optional parameters

“Graph Processing on Greenplum Database using Apache MADlib”, Frank McQuillan, Jan 2018,  
<https://content.pivotal.io/blog/graph-processing-on-greenplum-database-using-apache-madlib>

# PageRank Performance on Greenplum



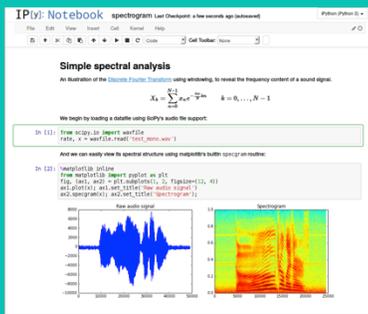
Note: log-log scale

Pivotal

Pivotal



# 4. Data Science Productivity Tools





# PivotalR

- Familiar R interface + performance/scalability of in-database analytics

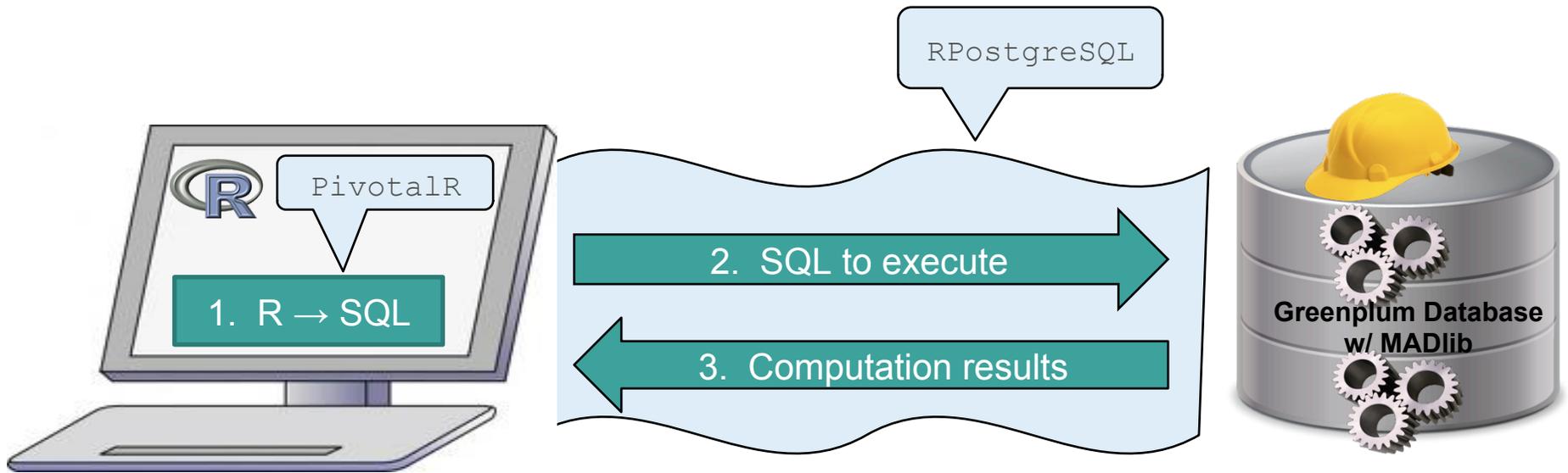
## PivotalR

```
d <- db.data.frame("houses")
houses_linregr <-
  madlib.lm(price ~ tax
            + bath
            + size
            , data=d)
```

## SQL Code

```
SELECT madlib.linregr_train( 'houses',
                             'houses_linregr',
                             'price',
                             'ARRAY[1, tax, bath, size]');
```

# PivotalR Workflow



No data here

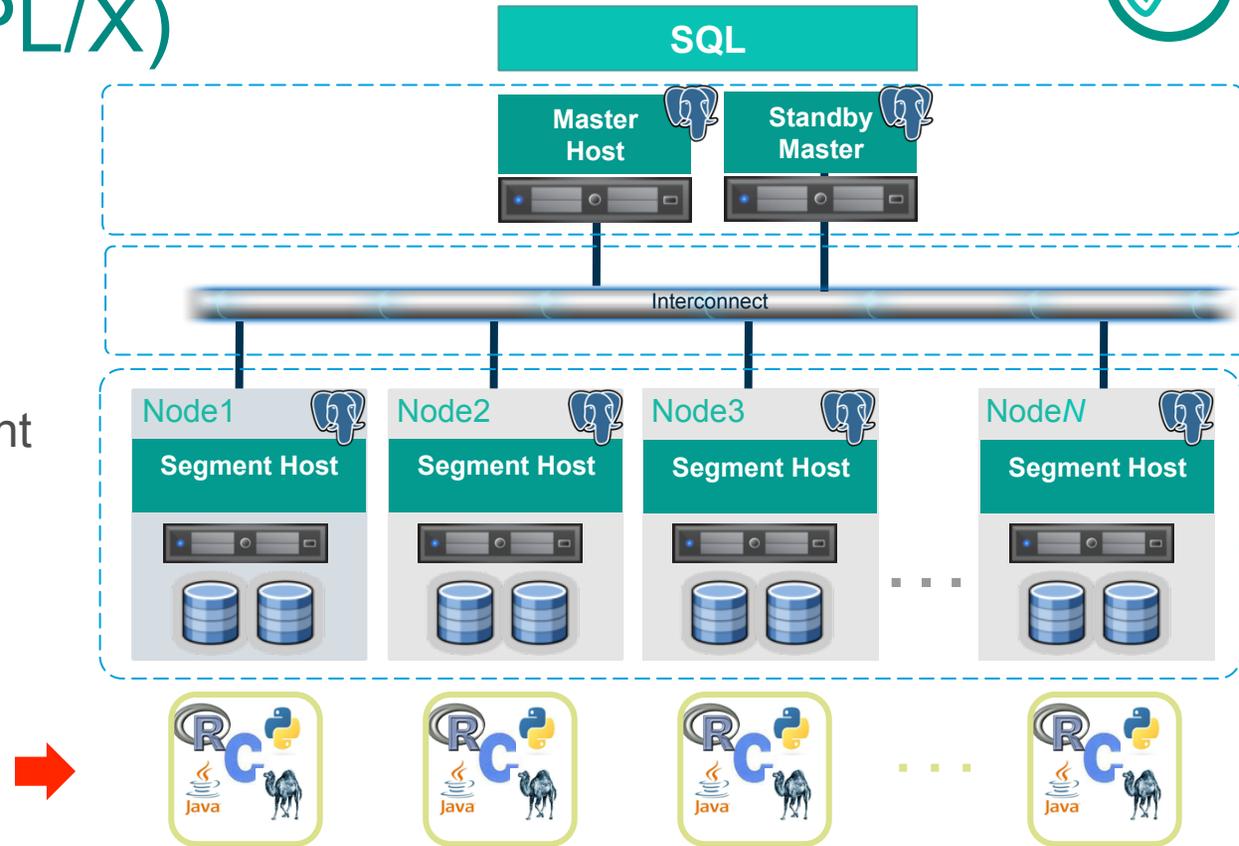
***Data Never Leaves DB***

Data lives here

# Procedural Language Extensions (PL/X)



- Write functions in Python, R, C, Java, pgsql, Perl
- Run on each segment in data parallel manner



Pivotal

Pivotal

# Data Science Bundle



100+ Libraries  
in Python & R

gensim

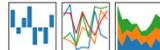
NumPy

MCMCpack



TensorFlow

pandas



pyLDAvis

spaCy

LIFELINES

XGBoost



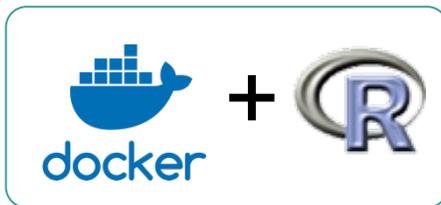
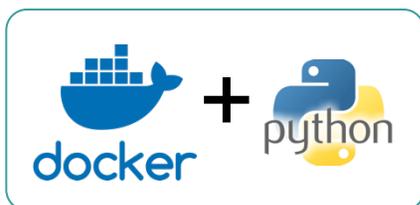
BeautifulSoup

Pivotal

Pivotal



# PL/Container



- Execute functions in isolated secure containers
- Deploy code and functions as non super-user

“Customize and Secure the Runtime and Dependencies of Your Procedural Languages Using PL/Container”

Hubert Zhang



Jack Wu



**Date:**  
**Duration:**  
**Room:**  
**Conference:**  
**Language:**  
**Track:**

2018 April 20 08:50  
50 min  
Liberty II-III  
PostgresConf US 2018  
English  
**Greenplum Summit**

# Data Science Notebooks



Jupyter mlp-mnist Last Checkpoint: 03/14/2018 (unsaved changes)

File Edit View Insert Cell Kernel Help Trusted Python 2

### Train MLP model

In [7]: `# From https://db-blog.web.cern.ch/blog/luca-casali/2016-07-neural-network-scoring-engine-plsql-reco...  
Image("../images/mlp-2.png")`

Out[7]:

Input image: 28x28 pixels

Output: predicted digit value

Input layer: 784 neurons, one per pixel

Hidden layer: 20 neurons

Output layer: 10 neurons

In [24]: `%%sql  
  
DROP TABLE IF EXISTS mnist_result_summary, mnist_result, mnist_result_standardization;  
  
SELECT madlib.mlp_classification(  
 'mnist_train', -- Source table  
 'mnist_result', -- Destination table  
 'x', -- Independent  
 'y', -- Dependent  
 ARRAY[5], -- Hidden layer sizes  
 'learning_rate_init=0.001,  
 n_iterations=1,  
 learning_rate_policy=constant,  
 lambda=0.0001, -- Regularization  
 tolerance=0',  
 'tanh', -- Activation function  
 '+', -- No weights  
 FALSE, -- No warmstart  
 TRUE); -- Verbose  
  
SELECT * FROM mnist_result;`

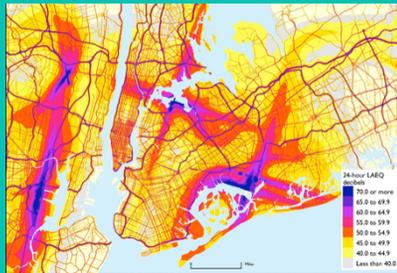
Zeppelin Notebook - Job

MADlib Zeppelin demo 1

2.02 2.12020002024040 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6 4.8 5.0 5.2 5.4 5.6 5.8 6.0 6.2 6.4 6.6 6.8 7.0 7.2 7.4 7.6 7.8 8.0 8.2 8.4 8.6 8.8 9.0 9.2 9.4 9.6 9.8 10.0 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 14.4 14.6 14.8 15.0 15.2 15.4 15.6 15.8 16.0 16.2 16.4 16.6 16.8 17.0 17.2 17.4 17.6 17.8 18.0 18.2 18.4 18.6 18.8 19.0 19.2 19.4 19.6 19.8 20.0 20.2 20.4 20.6 20.8 21.0 21.2 21.4 21.6 21.8 22.0 22.2 22.4 22.6 22.8 23.0 23.2 23.4 23.6 23.8 24.0 24.2 24.4 24.6 24.8 25.0 25.2 25.4 25.6 25.8 26.0 26.2 26.4 26.6 26.8 27.0 27.2 27.4 27.6 27.8 28.0 28.2 28.4 28.6 28.8 29.0 29.2 29.4 29.6 29.8 30.0 30.2 30.4 30.6 30.8 31.0 31.2 31.4 31.6 31.8 32.0 32.2 32.4 32.6 32.8 33.0 33.2 33.4 33.6 33.8 34.0 34.2 34.4 34.6 34.8 35.0 35.2 35.4 35.6 35.8 36.0 36.2 36.4 36.6 36.8 37.0 37.2 37.4 37.6 37.8 38.0 38.2 38.4 38.6 38.8 39.0 39.2 39.4 39.6 39.8 40.0 40.2 40.4 40.6 40.8 41.0 41.2 41.4 41.6 41.8 42.0 42.2 42.4 42.6 42.8 43.0 43.2 43.4 43.6 43.8 44.0 44.2 44.4 44.6 44.8 45.0 45.2 45.4 45.6 45.8 46.0 46.2 46.4 46.6 46.8 47.0 47.2 47.4 47.6 47.8 48.0 48.2 48.4 48.6 48.8 49.0 49.2 49.4 49.6 49.8 50.0 50.2 50.4 50.6 50.8 51.0 51.2 51.4 51.6 51.8 52.0 52.2 52.4 52.6 52.8 53.0 53.2 53.4 53.6 53.8 54.0 54.2 54.4 54.6 54.8 55.0 55.2 55.4 55.6 55.8 56.0 56.2 56.4 56.6 56.8 57.0 57.2 57.4 57.6 57.8 58.0 58.2 58.4 58.6 58.8 59.0 59.2 59.4 59.6 59.8 60.0 60.2 60.4 60.6 60.8 61.0 61.2 61.4 61.6 61.8 62.0 62.2 62.4 62.6 62.8 63.0 63.2 63.4 63.6 63.8 64.0 64.2 64.4 64.6 64.8 65.0 65.2 65.4 65.6 65.8 66.0 66.2 66.4 66.6 66.8 67.0 67.2 67.4 67.6 67.8 68.0 68.2 68.4 68.6 68.8 69.0 69.2 69.4 69.6 69.8 70.0 70.2 70.4 70.6 70.8 71.0 71.2 71.4 71.6 71.8 72.0 72.2 72.4 72.6 72.8 73.0 73.2 73.4 73.6 73.8 74.0 74.2 74.4 74.6 74.8 75.0 75.2 75.4 75.6 75.8 76.0 76.2 76.4 76.6 76.8 77.0 77.2 77.4 77.6 77.8 78.0 78.2 78.4 78.6 78.8 79.0 79.2 79.4 79.6 79.8 80.0 80.2 80.4 80.6 80.8 81.0 81.2 81.4 81.6 81.8 82.0 82.2 82.4 82.6 82.8 83.0 83.2 83.4 83.6 83.8 84.0 84.2 84.4 84.6 84.8 85.0 85.2 85.4 85.6 85.8 86.0 86.2 86.4 86.6 86.8 87.0 87.2 87.4 87.6 87.8 88.0 88.2 88.4 88.6 88.8 89.0 89.2 89.4 89.6 89.8 90.0 90.2 90.4 90.6 90.8 91.0 91.2 91.4 91.6 91.8 92.0 92.2 92.4 92.6 92.8 93.0 93.2 93.4 93.6 93.8 94.0 94.2 94.4 94.6 94.8 95.0 95.2 95.4 95.6 95.8 96.0 96.2 96.4 96.6 96.8 97.0 97.2 97.4 97.6 97.8 98.0 98.2 98.4 98.6 98.8 99.0 99.2 99.4 99.6 99.8 100.0 100.2 100.4 100.6 100.8 101.0 101.2 101.4 101.6 101.8 102.0 102.2 102.4 102.6 102.8 103.0 103.2 103.4 103.6 103.8 104.0 104.2 104.4 104.6 104.8 105.0 105.2 105.4 105.6 105.8 106.0 106.2 106.4 106.6 106.8 107.0 107.2 107.4 107.6 107.8 108.0 108.2 108.4 108.6 108.8 109.0 109.2 109.4 109.6 109.8 110.0 110.2 110.4 110.6 110.8 111.0 111.2 111.4 111.6 111.8 112.0 112.2 112.4 112.6 112.8 113.0 113.2 113.4 113.6 113.8 114.0 114.2 114.4 114.6 114.8 115.0 115.2 115.4 115.6 115.8 116.0 116.2 116.4 116.6 116.8 117.0 117.2 117.4 117.6 117.8 118.0 118.2 118.4 118.6 118.8 119.0 119.2 119.4 119.6 119.8 120.0 120.2 120.4 120.6 120.8 121.0 121.2 121.4 121.6 121.8 122.0 122.2 122.4 122.6 122.8 123.0 123.2 123.4 123.6 123.8 124.0 124.2 124.4 124.6 124.8 125.0 125.2 125.4 125.6 125.8 126.0 126.2 126.4 126.6 126.8 127.0 127.2 127.4 127.6 127.8 128.0 128.2 128.4 128.6 128.8 129.0 129.2 129.4 129.6 129.8 130.0 130.2 130.4 130.6 130.8 131.0 131.2 131.4 131.6 131.8 132.0 132.2 132.4 132.6 132.8 133.0 133.2 133.4 133.6 133.8 134.0 134.2 134.4 134.6 134.8 135.0 135.2 135.4 135.6 135.8 136.0 136.2 136.4 136.6 136.8 137.0 137.2 137.4 137.6 137.8 138.0 138.2 138.4 138.6 138.8 139.0 139.2 139.4 139.6 139.8 140.0 140.2 140.4 140.6 140.8 141.0 141.2 141.4 141.6 141.8 142.0 142.2 142.4 142.6 142.8 143.0 143.2 143.4 143.6 143.8 144.0 144.2 144.4 144.6 144.8 145.0 145.2 145.4 145.6 145.8 146.0 146.2 146.4 146.6 146.8 147.0 147.2 147.4 147.6 147.8 148.0 148.2 148.4 148.6 148.8 149.0 149.2 149.4 149.6 149.8 150.0 150.2 150.4 150.6 150.8 151.0 151.2 151.4 151.6 151.8 152.0 152.2 152.4 152.6 152.8 153.0 153.2 153.4 153.6 153.8 154.0 154.2 154.4 154.6 154.8 155.0 155.2 155.4 155.6 155.8 156.0 156.2 156.4 156.6 156.8 157.0 157.2 157.4 157.6 157.8 158.0 158.2 158.4 158.6 158.8 159.0 159.2 159.4 159.6 159.8 160.0 160.2 160.4 160.6 160.8 161.0 161.2 161.4 161.6 161.8 162.0 162.2 162.4 162.6 162.8 163.0 163.2 163.4 163.6 163.8 164.0 164.2 164.4 164.6 164.8 165.0 165.2 165.4 165.6 165.8 166.0 166.2 166.4 166.6 166.8 167.0 167.2 167.4 167.6 167.8 168.0 168.2 168.4 168.6 168.8 169.0 169.2 169.4 169.6 169.8 170.0 170.2 170.4 170.6 170.8 171.0 171.2 171.4 171.6 171.8 172.0 172.2 172.4 172.6 172.8 173.0 173.2 173.4 173.6 173.8 174.0 174.2 174.4 174.6 174.8 175.0 175.2 175.4 175.6 175.8 176.0 176.2 176.4 176.6 176.8 177.0 177.2 177.4 177.6 177.8 178.0 178.2 178.4 178.6 178.8 179.0 179.2 179.4 179.6 179.8 180.0 180.2 180.4 180.6 180.8 181.0 181.2 181.4 181.6 181.8 182.0 182.2 182.4 182.6 182.8 183.0 183.2 183.4 183.6 183.8 184.0 184.2 184.4 184.6 184.8 185.0 185.2 185.4 185.6 185.8 186.0 186.2 186.4 186.6 186.8 187.0 187.2 187.4 187.6 187.8 188.0 188.2 188.4 188.6 188.8 189.0 189.2 189.4 189.6 189.8 190.0 190.2 190.4 190.6 190.8 191.0 191.2 191.4 191.6 191.8 192.0 192.2 192.4 192.6 192.8 193.0 193.2 193.4 193.6 193.8 194.0 194.2 194.4 194.6 194.8 195.0 195.2 195.4 195.6 195.8 196.0 196.2 196.4 196.6 196.8 197.0 197.2 197.4 197.6 197.8 198.0 198.2 198.4 198.6 198.8 199.0 199.2 199.4 199.6 199.8 200.0 200.2 200.4 200.6 200.8 201.0 201.2 201.4 201.6 201.8 202.0 202.2 202.4 202.6 202.8 203.0 203.2 203.4 203.6 203.8 204.0 204.2 204.4 204.6 204.8 205.0 205.2 205.4 205.6 205.8 206.0 206.2 206.4 206.6 206.8 207.0 207.2 207.4 207.6 207.8 208.0 208.2 208.4 208.6 208.8 209.0 209.2 209.4 209.6 209.8 210.0 210.2 210.4 210.6 210.8 211.0 211.2 211.4 211.6 211.8 212.0 212.2 212.4 212.6 212.8 213.0 213.2 213.4 213.6 213.8 214.0 214.2 214.4 214.6 214.8 215.0 215.2 215.4 215.6 215.8 216.0 216.2 216.4 216.6 216.8 217.0 217.2 217.4 217.6 217.8 218.0 218.2 218.4 218.6 218.8 219.0 219.2 219.4 219.6 219.8 220.0 220.2 220.4 220.6 220.8 221.0 221.2 221.4 221.6 221.8 222.0 222.2 222.4 222.6 222.8 223.0 223.2 223.4 223.6 223.8 224.0 224.2 224.4 224.6 224.8 225.0 225.2 225.4 225.6 225.8 226.0 226.2 226.4 226.6 226.8 227.0 227.2 227.4 227.6 227.8 228.0 228.2 228.4 228.6 228.8 229.0 229.2 229.4 229.6 229.8 230.0 230.2 230.4 230.6 230.8 231.0 231.2 231.4 231.6 231.8 232.0 232.2 232.4 232.6 232.8 233.0 233.2 233.4 233.6 233.8 234.0 234.2 234.4 234.6 234.8 235.0 235.2 235.4 235.6 235.8 236.0 236.2 236.4 236.6 236.8 237.0 237.2 237.4 237.6 237.8 238.0 238.2 238.4 238.6 238.8 239.0 239.2 239.4 239.6 239.8 240.0 240.2 240.4 240.6 240.8 241.0 241.2 241.4 241.6 241.8 242.0 242.2 242.4 242.6 242.8 243.0 243.2 243.4 243.6 243.8 244.0 244.2 244.4 244.6 244.8 245.0 245.2 245.4 245.6 245.8 246.0 246.2 246.4 246.6 246.8 247.0 247.2 247.4 247.6 247.8 248.0 248.2 248.4 248.6 248.8 249.0 249.2 249.4 249.6 249.8 250.0 250.2 250.4 250.6 250.8 251.0 251.2 251.4 251.6 251.8 252.0 252.2 252.4 252.6 252.8 253.0 253.2 253.4 253.6 253.8 254.0 254.2 254.4 254.6 254.8 255.0 255.2 255.4 255.6 255.8 256.0 256.2 256.4 256.6 256.8 257.0 257.2 257.4 257.6 257.8 258.0 258.2 258.4 258.6 258.8 259.0 259.2 259.4 259.6 259.8 260.0 260.2 260.4 260.6 260.8 261.0 261.2 261.4 261.6 261.8 262.0 262.2 262.4 262.6 262.8 263.0 263.2 263.4 263.6 263.8 264.0 264.2 264.4 264.6 264.8 265.0 265.2 265.4 265.6 265.8 266.0 266.2 266.4 266.6 266.8 267.0 267.2 267.4 267.6 267.8 268.0 268.2 268.4 268.6 268.8 269.0 269.2 269.4 269.6 269.8 270.0 270.2 270.4 270.6 270.8 271.0 271.2 271.4 271.6 271.8 272.0 272.2 272.4 272.6 272.8 273.0 273.2 273.4 273.6 273.8 274.0 274.2 274.4 274.6 274.8 275.0 275.2 275.4 275.6 275.8 276.0 276.2 276.4 276.6 276.8 277.0 277.2 277.4 277.6 277.8 278.0 278.2 278.4 278.6 278.8 279.0 279.2 279.4 279.6 279.8 280.0 280.2 280.4 280.6 280.8 281.0 281.2 281.4 281.6 281.8 282.0 282.2 282.4 282.6 282.8 283.0 283.2 283.4 283.6 283.8 284.0 284.2 284.4 284.6 284.8 285.0 285.2 285.4 285.6 285.8 286.0 286.2 286.4 286.6 286.8 287.0 287.2 287.4 287.6 287.8 288.0 288.2 288.4 288.6 288.8 289.0 289.2 289.4 289.6 289.8 290.0 290.2 290.4 290.6 290.8 291.0 291.2 291.4 291.6 291.8 292.0 292.2 292.4 292.6 292.8 293.0 293.2 293.4 293.6 293.8 294.0 294.2 294.4 294.6 294.8 295.0 295.2 295.4 295.6 295.8 296.0 296.2 296.4 296.6 296.8 297.0 297.2 297.4 297.6 297.8 298.0 298.2 298.4 298.6 298.8 299.0 299.2 299.4 299.6 299.8 300.0 300.2 300.4 300.6 300.8 301.0 301.2 301.4 301.6 301.8 302.0 302.2 302.4 302.6 302.8 303.0 303.2 303.4 303.6 303.8 304.0 304.2 304.4 304.6 304.8 305.0 305.2 305.4 305.6 305.8 306.0 306.2 306.4 306.6 306.8 307.0 307.2 307.4 307.6 307.8 308.0 308.2 308.4 308.6 308.8 309.0 309.2 309.4 309.6 309.8 310.0 310.2 310.4 310.6 310.8 311.0 311.2 311.4 311.6 311.8 312.0 312.2 312.4 312.6 312.8 313.0 313.2 313.4 313.6 313.8 314.0 314.2 314.4 314.6 314.8 315.0 315.2 315.4 315.6 315.8 316.0 316.2 316.4 316.6 316.8 317.0 317.2 317.4 317.6 317.8 318.0 318.2 318.4 318.6 318.8 319.0 319.2 319.4 319.6 319.8 320.0 320.2 320.4 320.6 320.8 321.0 321.2 321.4 321.6 321.8 322.0 322.2 322.4 322.6 322.8 323.0 323.2 323.4 323.6 323.8 324.0 324.2 324.4 324.6 324.8 325.0 325.2 325.4 325.6 325.8 326.0 326.2 326.4 326.6 326.8 327.0 327.2 327.4 327.6 327.8 328.0 328.2 328.4 328.6 328.8 329.0 329.2 329.4 329.6 329.8 330.0 330.2 330.4 330.6 330.8 331.0 331.2 331.4 331.6 331.8 332.0 332.2 332.4 332.6 332.8 333.0 333.2 333.4 333.6 333.8 334.0 334.2 334.4 334.6 334.8 335.0 335.2 335.4 335.6 335.8 336.0 336.2 336.4 336.6 336.8 337.0 337.2 337.4 337.6 337.8 338.0 338.2 338.4 338.6 338.8 339.0 339.2 339.4 339.6 339.8 340.0 340.2 340.4 340.6 340.8 341.0 341.2 341.4 341.6 341.8 342.0 342.2 342.4 342.6 342.8 343.0 343.2 343.4 343.6 343.8 344.0 344.2 344.4 344.6 344.8 345.0 345.2 345.4 345.6 345.8 346.0 346.2 346.4 346.6 346.8 347.0 347.2 347.4 347.6 347.8 348.0 348.2 348.4 348.6 348.8 349.0 349.2 349.4 349.6 349.8 350.0 350.2 350.4 350.6 350.8 351.0 351.2 351.4 351.6 351.8 352.0 352.2 352.4 352.6 352.8 353.0 353.2 353.4 353.6 353.8 354.0 354.2 354.4 354.6 354.8 355.0 355.2 355.4 355.6 355.8 356.0 356.2 356.4 356.6 356.8 357.0 357.2 357.4 357.6 357.8 358.0 358.2 358.4 358.6 358.8 359.0 359.2 359.4 359.6 359.8 360.0 360.2 360.4 360.6 360.8 361.0 361.2 361.4 361.6 361.8 362.0 362.2 362.4 362.6 362.8 363.0 363.2 363.4 363.6 363.8 364.0 364.2 364.4 364.6 364.8 365.0 365.2 365.4 365.6 365.8 366.0 366.2 366.4 366.6 366.8 367.0 367.2 367.4 367.6 367.8 368.0 368.2 368.4 368.6 368.8 369.0 369.2 369.4 369.6 369.8 370.0 370.2 370.4 370.6 370.8 371.0 371.2 371.4 371.6 371.8 372.0 372.2 372.4 372.6 372.8 373.0 373.2 373.4 373.6 373.8 374.0 374.2 374.4 374.6 374.8 375.0 375.2 375.4 375.6 375.8 376.0 376.2 376.4 376.6 376.8 377.0 377.2 377.4 377.6 377.8 378.0 378.2 378.4 378.6 378.8 379.0 379.2 379.4 379.6 379.8 380.0 380.2 380.4 380.6 380.8 381.0 381.2 381.4 381.6 381.8 382.0 382.2 382.4 382.6 382.8 383.0 383.2 383.4 383.6 383.8 384.0 384.2 384.4 384.6 384.8 385.0 385.2 385.4 385.6 385.8 386.0 386.2 386.4 386.6 386.8 387.0 387.2 387.4 387.6 387.8 388.0 388.2 388.4 388.6 388.8 389.0 389.2 389.4 389.6 389.8 390.0 390.2 390.4 390.6 390.8 391.0 391.2 391.4 391.6 391.8 392.0 392.2 392.4 392.6 392.8 393.0 393.2 393.4 393.6 393.8 394.0 394.2 394.4 394.6 394.8 395.0 395.2 395.4 395.6 395.8 396.0 396.2 396.4 396.6 396.8 397.0 397.2 397.4 397.6 397.8 398.0 398.2 398.4 398.6 398.8 399.0 399.2 399.4 399.6 399.8 400.0 400.2 400.4 400.6 400.8 401.0 401.2 401.4 401.6 401.8 402.0 402.2 402.4 402.6 402.8 403.0 403.2 403.4 403.6 403.8 404.0 404.2 404.4 404.6 404.8 405.0 405.2 405.4 405.6 405.8 406.0 406.2 406.4 406.6 406.8 407.0 407.2 407.4 407.6 407.8 408.0 408.2 408.4 408.6 408.8 409.0 409.2 409.4 409.6 409.8 410.0



# 5. Geospatial

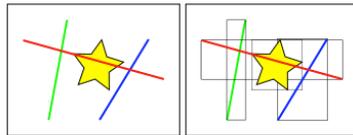


# Geospatial Analytics with

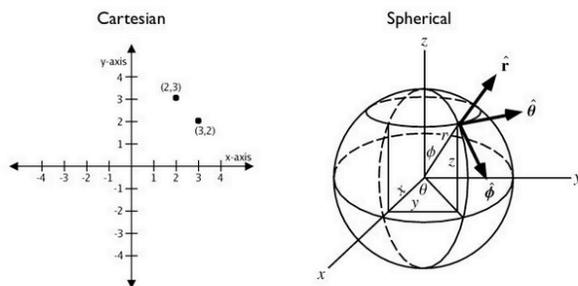


- PostGIS is a spatial database extension which allows for analysis and processing of GIS objects

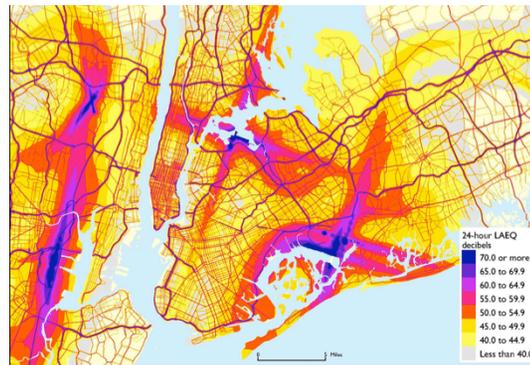
## Spatial Indexes & Bounding Boxes



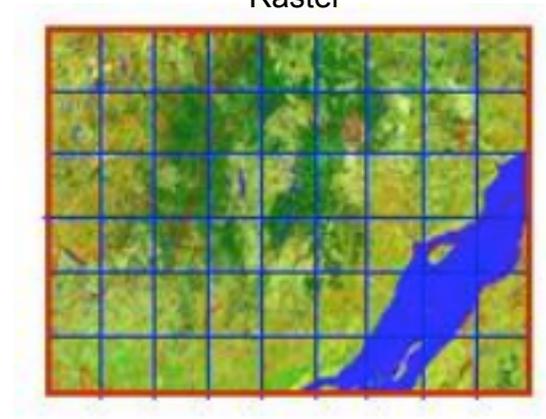
## Round earth calculations



Vector



Raster





# Spatial Relationships & Joins

- Relationships
  - ST\_Equals
  - ST\_Intersects
  - ST\_Crosses
  - ST\_Disjoint
  - ST\_Overlaps
  - ST\_Touches
  - ST\_Within
  - ST\_Contains

Spatial joins use spatial relationships as the join key

Example:

Subway stations: POINT



Neighborhoods: MULTIPOLYGON

```
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,Q,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)

From Introduction to PostGIS, <http://workshops.boundlessgeo.com/postgis-intro/>



# The State of Unstructured Data

*“...most industry experts agree that **80% to 90% of the world’s data is unstructured**. Yet, only **0.5%** is effectively analyzed and used today.*

*In the business world, most unstructured data lies in **customer-related text**...Done right, extracting valuable predictive insights from huge quantities of text takes just **seconds**.”*

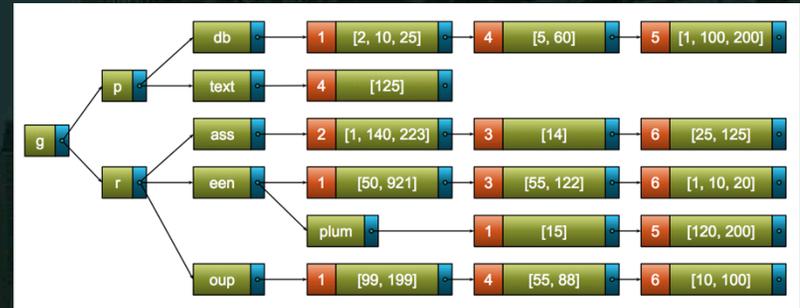
- Osvaldo Driollet (PhD), Sr. Data Scientist, FICO

# GPText Overview

- GPDB + Apache Solr (+ MADLib!)
  - Only DB that integrates text at scale
- Combination of semi-structured and structured data
- Process mass quantities of raw text for large-scale analytics
- Exposed as SQL UDFs

# GPText Index

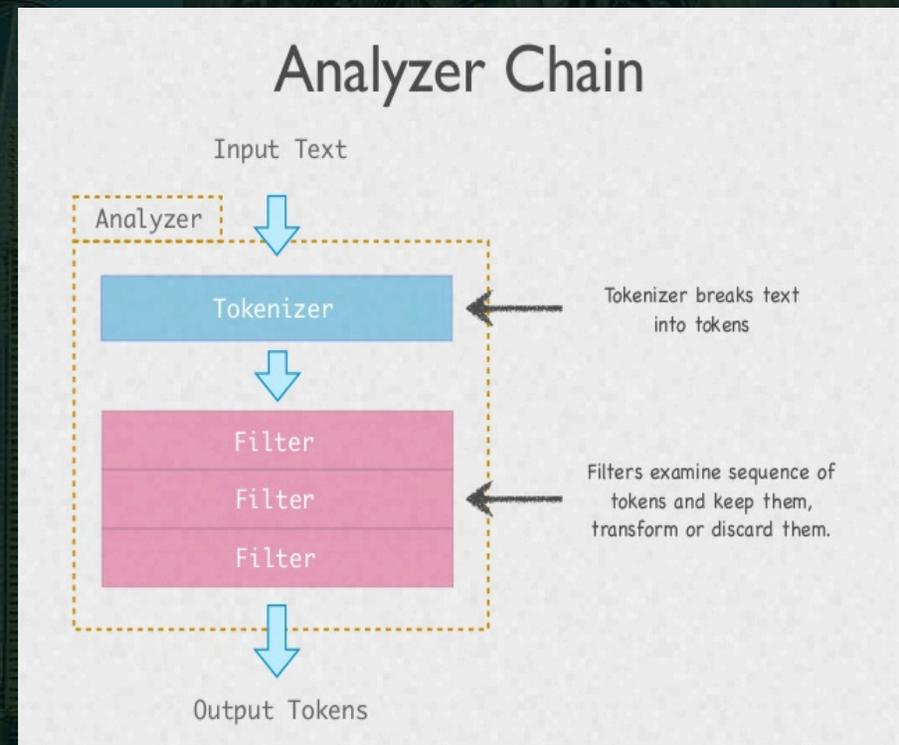
- Efficient Storage
  - Word, Position, Synonyms, Stem, Relevancy, Emoticons
- Fast Search
  - Indexed, not Scanning
- Relevant Results



$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

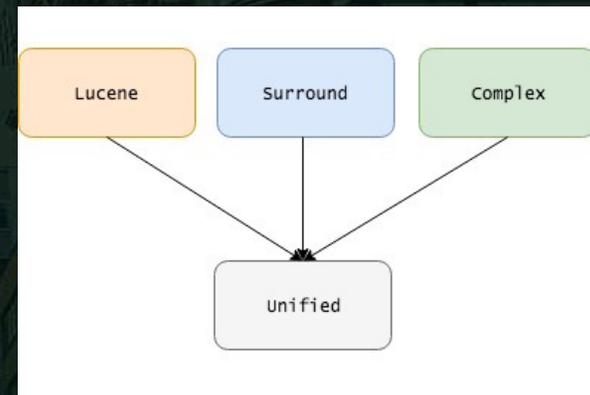
# Analyzer Chains

- Document formats are NOT standard
  - International, Social Media, Logs, etc.
- Parse and Extract without losing meaning!



# Unified Query Parser

- Designed to support multi-faceted queries
  - Boolean
  - Proximity
  - Wildcard
- No need to write multiple individual queries with joins

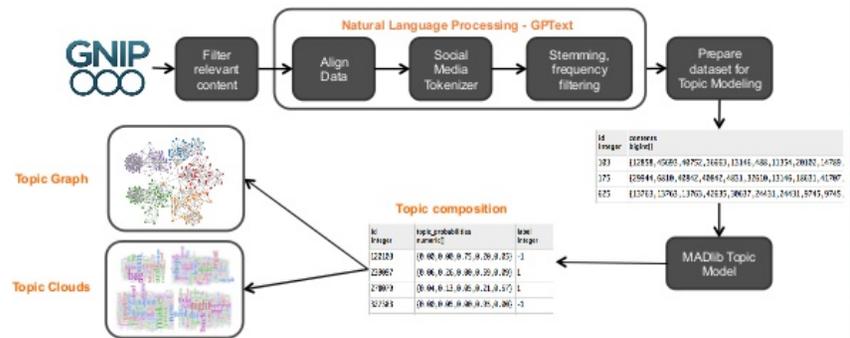


```
SELECT l.id, l.score, r."TO"  
FROM gptext.search(  
  TABLE(SELECT 1 SCATTER BY 1),  
  'demo.public.enron',  
  '{!gptextqp} content:2w(Phillips Petroleum)  
  AND to:"Christine Stokes"  
  AND date:["2000-01-01T00:00:00Z" TO "2001-01-01T00:00:00Z"]',  
  NULL) l,  
  enron r WHERE l.id=r.id;
```

# GPTtext + MADLib

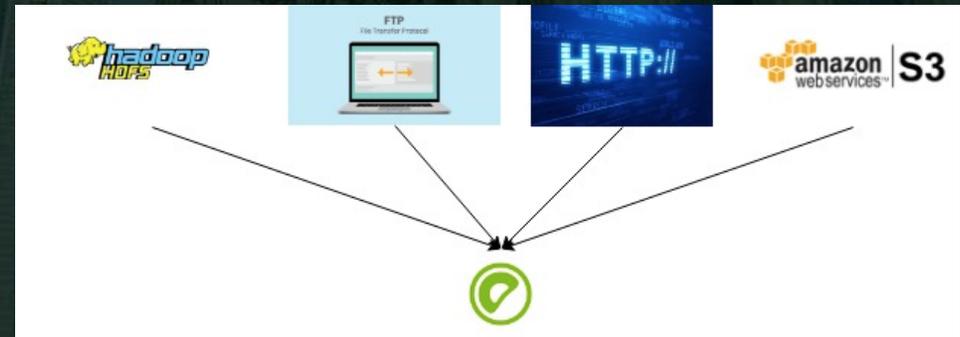
- Integrated with MADLib
  - Topic Modeling
  - Clustering
  - Sentiment Analysis
  - Sequence Pattern Mining

## Topic Analysis – MADlib pLDA



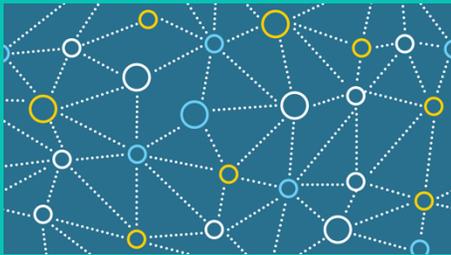
# External Indexing

- Ability to connect to external data sources
  - Currently: HTTP, HDFS
  - Planned: FTP, S3
- Index and Store raw files (PDF, Word, Mail, etc.)
- Access and search your data, no matter where, no matter what.





# 7. Connectivity

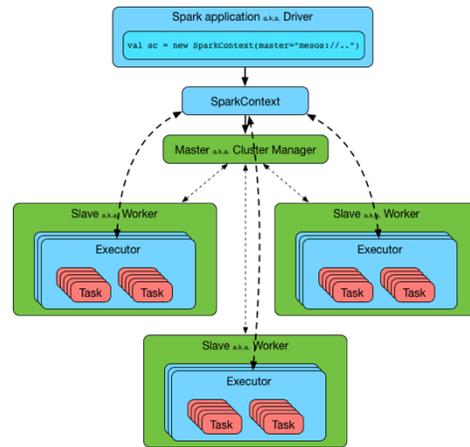


Pivotal

# Greenplum - Spark Connector



In-memory processing



Spark -  
Greenplum  
connector



Pivotal  
Greenplum

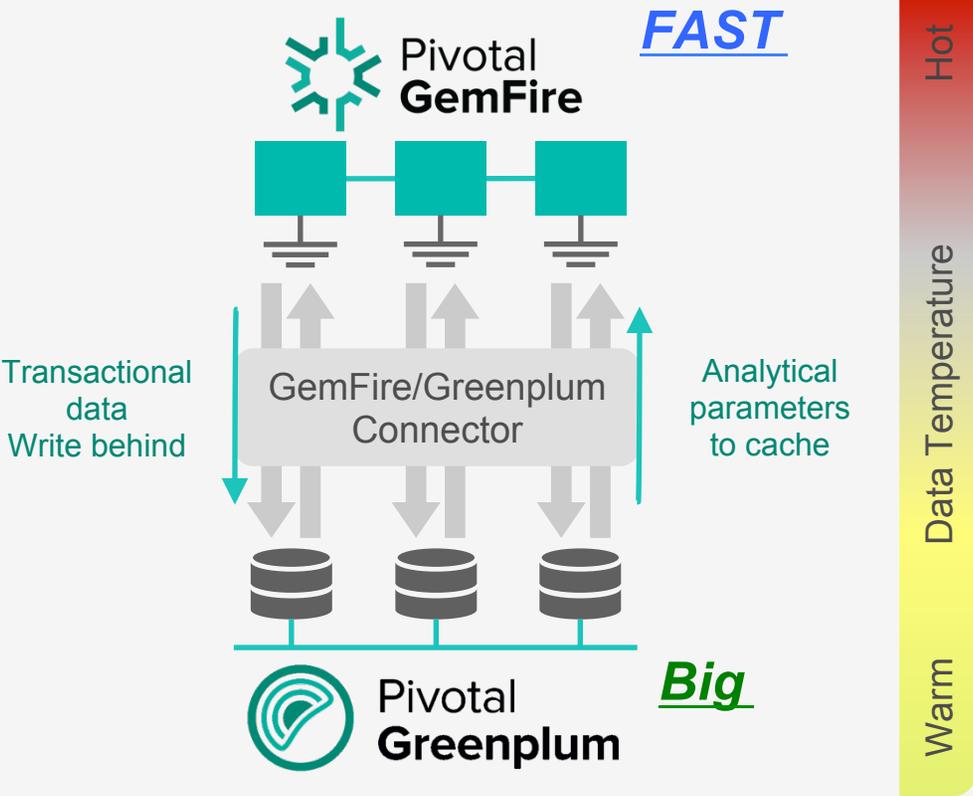


- Provide Data Access to Greenplum Data
- Leverage SPARK Skill Set of Data Scientists
- Use off-cluster resources to do computations
- Push result sets back into Greenplum for storage

# Greenplum - Gemfire Connector

Seamlessly share data between GemFire and Greenplum

**Bi-Directional  
Direct Connection**  
--  
**GemFire and  
Greenplum Segment  
Servers**





# 8. Example Use Cases



Pivotal



## Event Data Warehouse (EvDW) System Architecture

### Applications

#### GIS tools



QGIS

#### Navigation



#### Mobile Apps



### Machine Learning/Data Processing Infrastructure



Risk map creation

Route analysis

Event Prediction

Event Data Archives

Event Association analysis

Spatial/temporal analysis

### Sources



#### Raster data

- Weather radar
- Air pollution
- Floating population



#### Vector data

- Road network
- Traffic data
- Atmospheric observation
- Mobile/automotive sensing



#### Social Network Service data

- Geo-tagged Twitter

#### Other

- Origin-destination (people) flows
- Epidemiological surveillance data

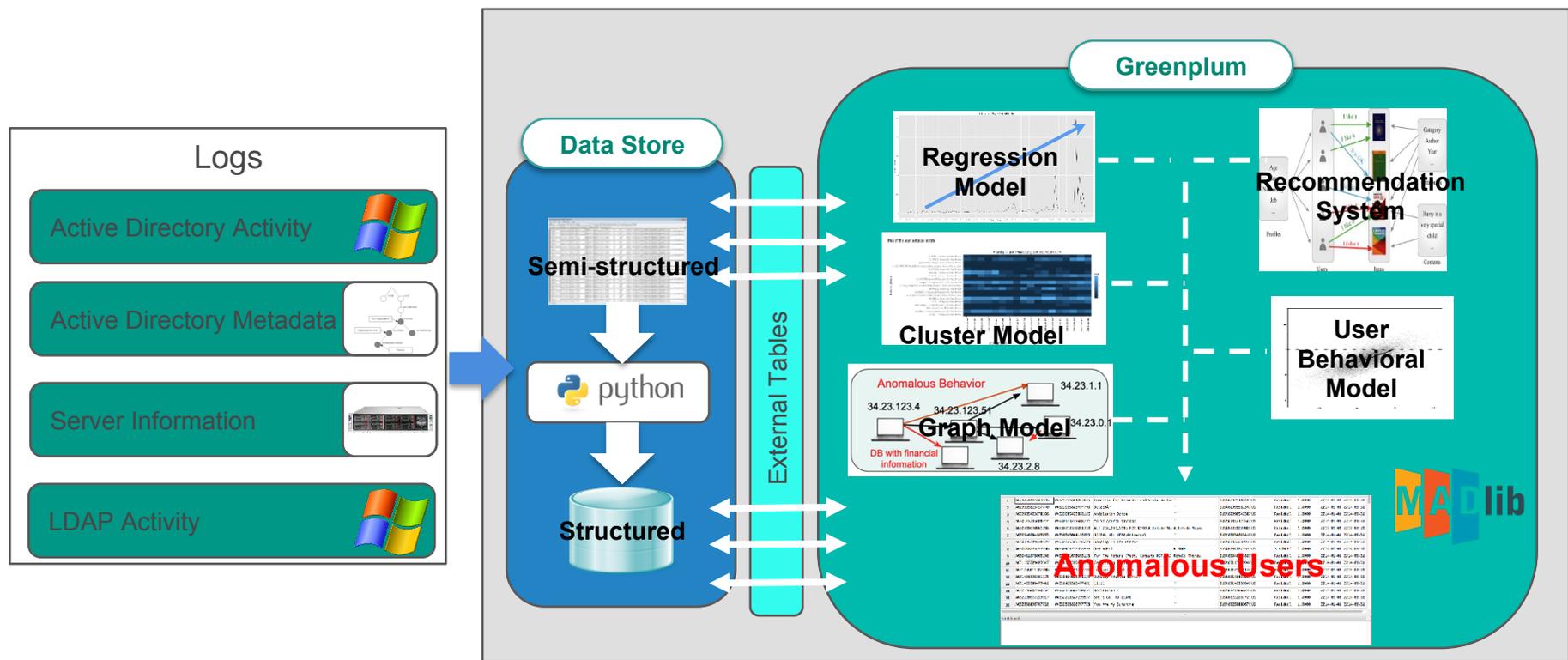
The QGIS logo is a registered trademark of the QGIS project (qgis.org).

# Operations - Parts Monitoring

- Monitoring 100s of different models + parts
- Structured Data + Operator Notes
- Minimized Recall Risk and Improved Reliability



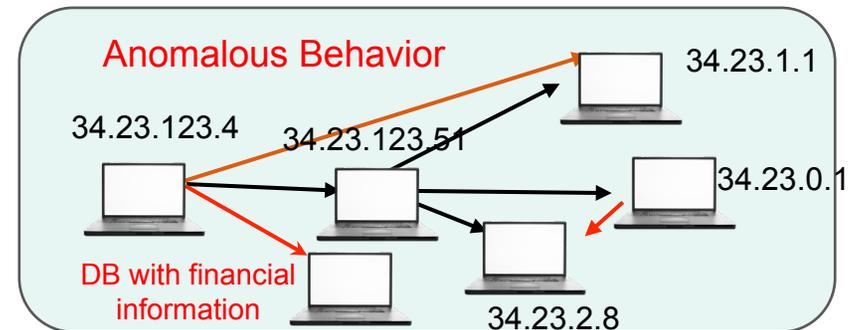
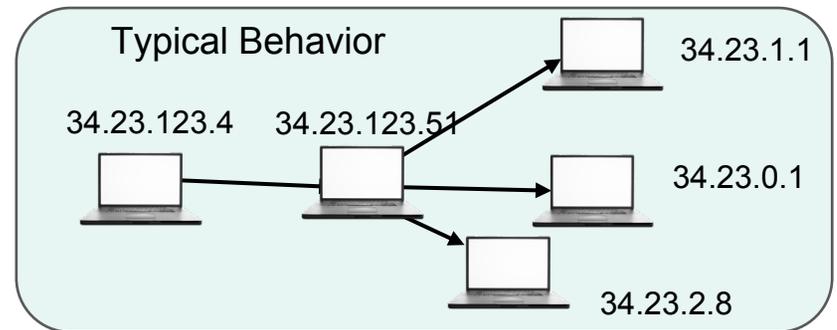
# Cyber Security - Lateral Movement Detection



<https://content.pivotal.io/blog/better-threat-detection-and-response-with-analytics-for-lateral-movement>

## Cyber Security (continued)

- Using historical window events data to build graphs of typical user behavior\*
- Is this behavior typical?
- Graph models are sensitive to direction, order, and frequency.



\*Reference: Alexander D. Kenta, Lorie M. Liebrock, Joshua C. Neila.  
*Authentication graphs: Analyzing user behavior within an enterprise network.*

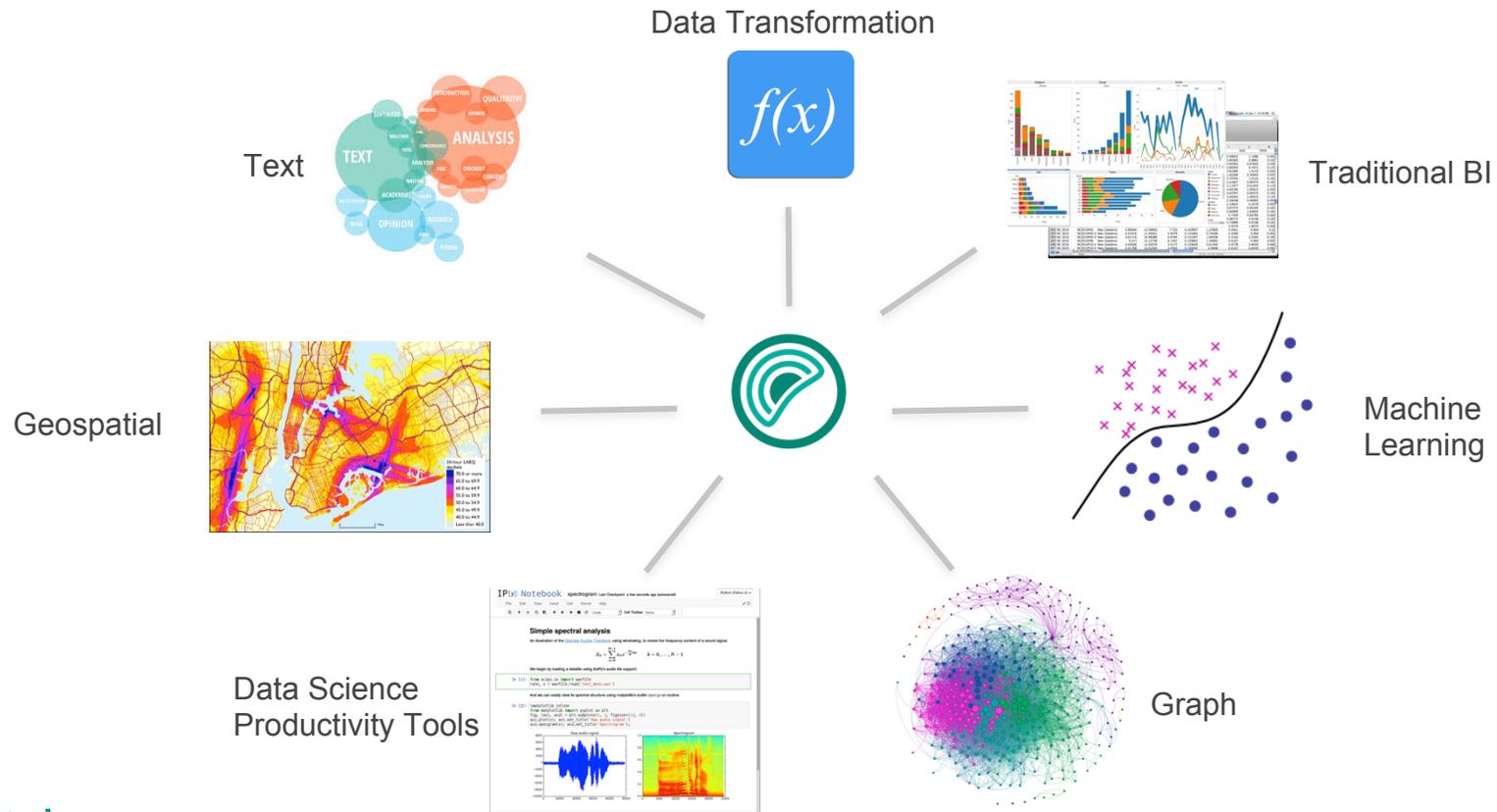


# 9. Looking Ahead



Pivotal

# Greenplum Integrated Analytics



Pivotal

Pivotal



**Thank you!**