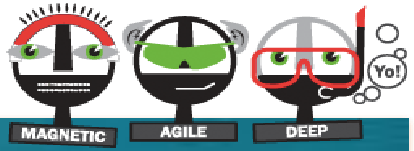
[Data Science Webinar]

Greenplum人工智能与机器学习工具集
—Apache MADLib引论

Pivotal资深产品经理 吴疆(jwu@pivotal.io)

MADlib: 可扩展的, In-Database 机器学习工具集



Apache MADlib: Big Data Machine Learning in SQL

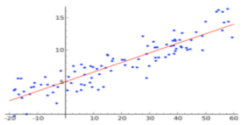
Open source, **top level Apache project**

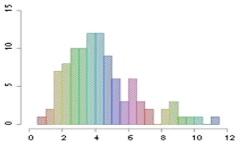
For PostgreSQL and Greenplum Database

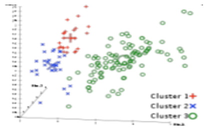
Powerful machine learning, graph, statistics and analytics for data scientists

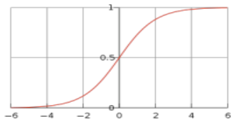
- Open source https://github.com/apache/madlib
- Downloads and docs http://madlib.apache.org/
- Wiki https://cwiki.apache.org/confluence/display/MADLIB/

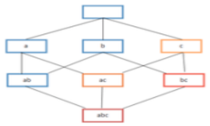
MADlib Functions











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)

Nearest Neighbors

- k-Nearest Neighbors

Time Series Analysis

- ARIMA

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

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

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

Statistics

Descriptive Statistics

- Cardinality Estimators
- Correlation and Covariance
- Summary

Inferential Statistics - Hypothesis Tests
Probability Functions

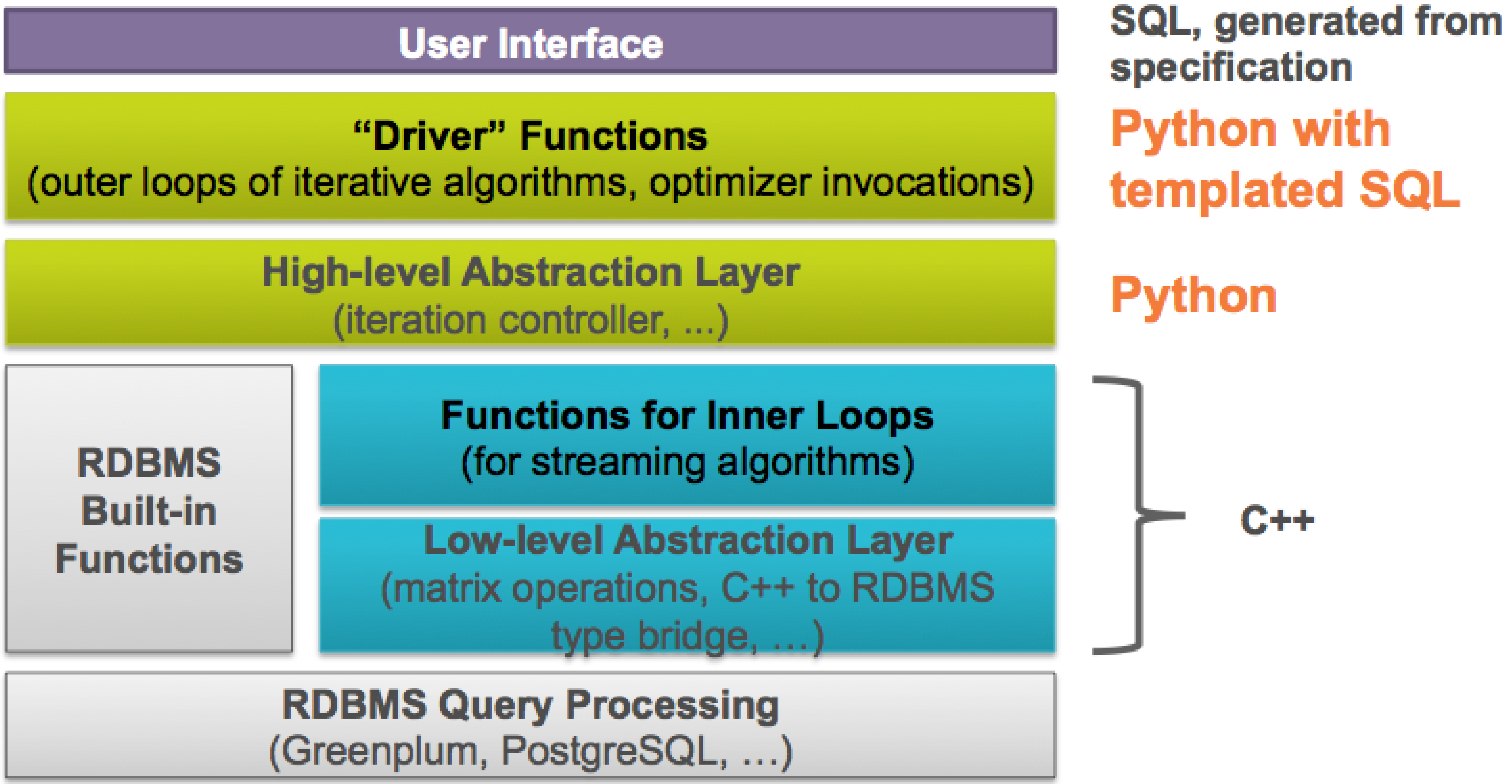
Model Selection

Cross Validation
Prediction Metrics
Train-Test Split

Sampling

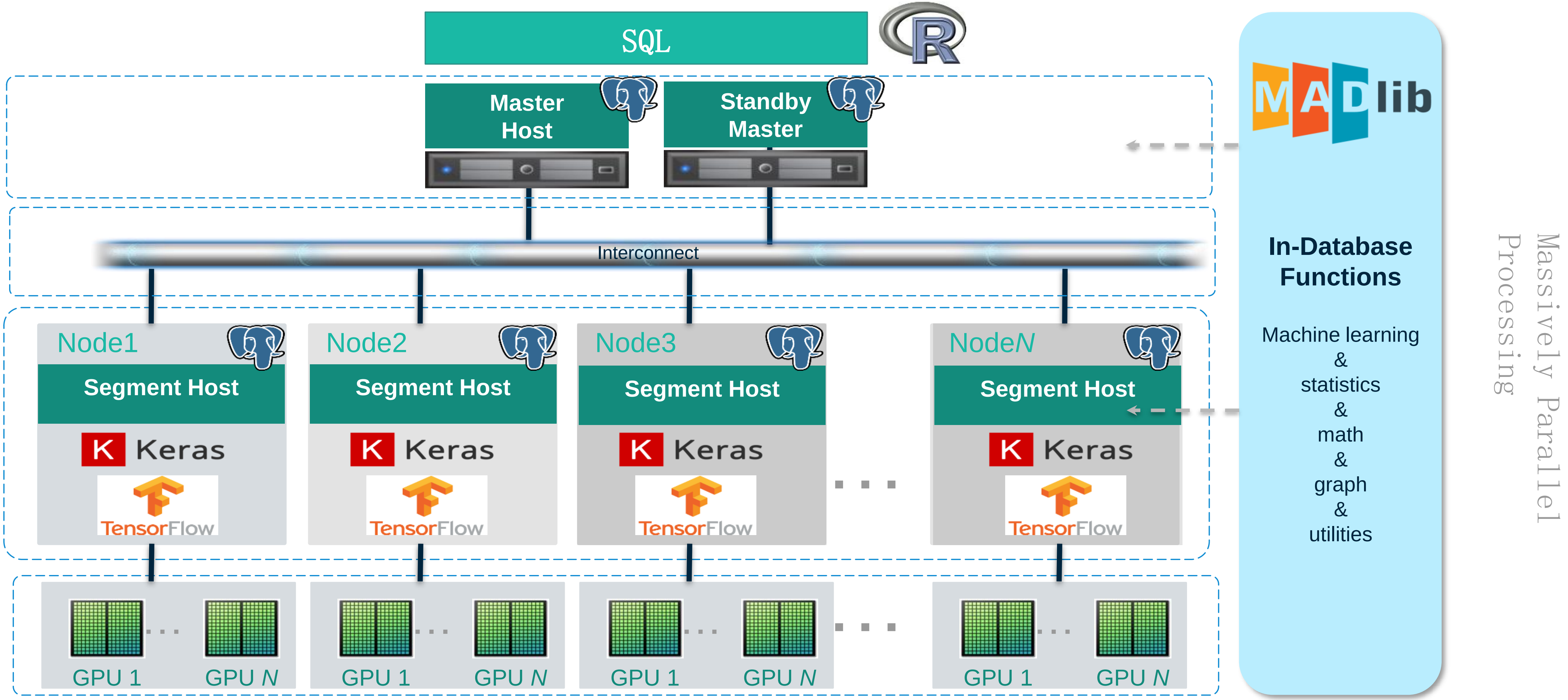
Balanced
Random
Stratified

Apache MADlib总体架构图



User Interface
"Driver" Functions
(outer loops of iterative algorithms, optimizer invocations)
High-level Abstraction Layer
(iteration controller, ...)
RDBMS Built-in Functions
Functions for Inner Loops
(for streaming algorithms)
Low-level Abstraction Layer
(matrix operations, C++ to RDBMS type bridge, ...)
RDBMS Query Processing
(Greenplum, PostgreSQL, ...)
SQL, generated from specification
Python with templated SQL
Python
C++

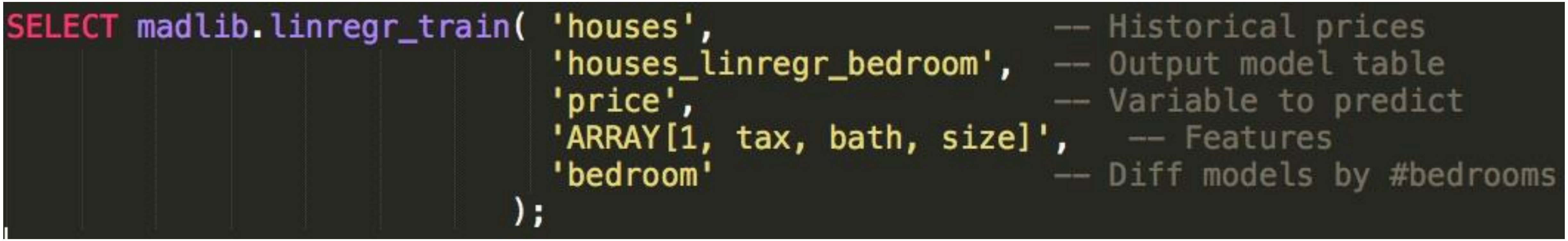
Greenplum+MADlib部署架构图



Supporting the full spectrum of data science workloads:
Data preparation, feature generation, machine learning, geospatial, deep learning, etc

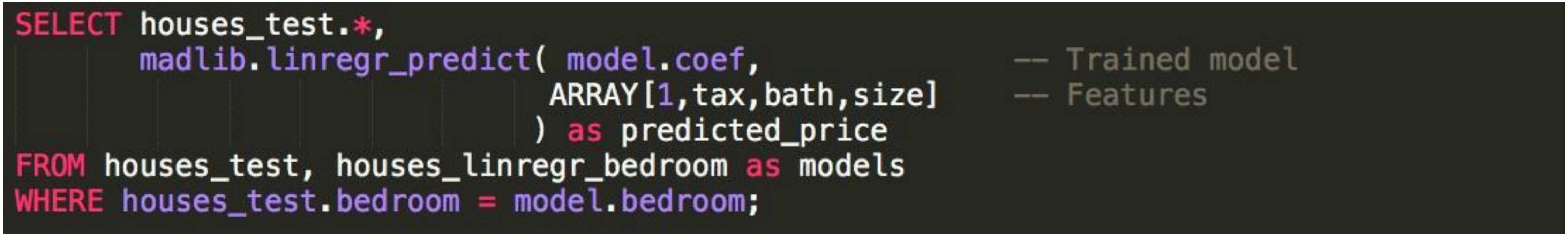
Greenplum+MADlib进行in-database Analysis

模型训练

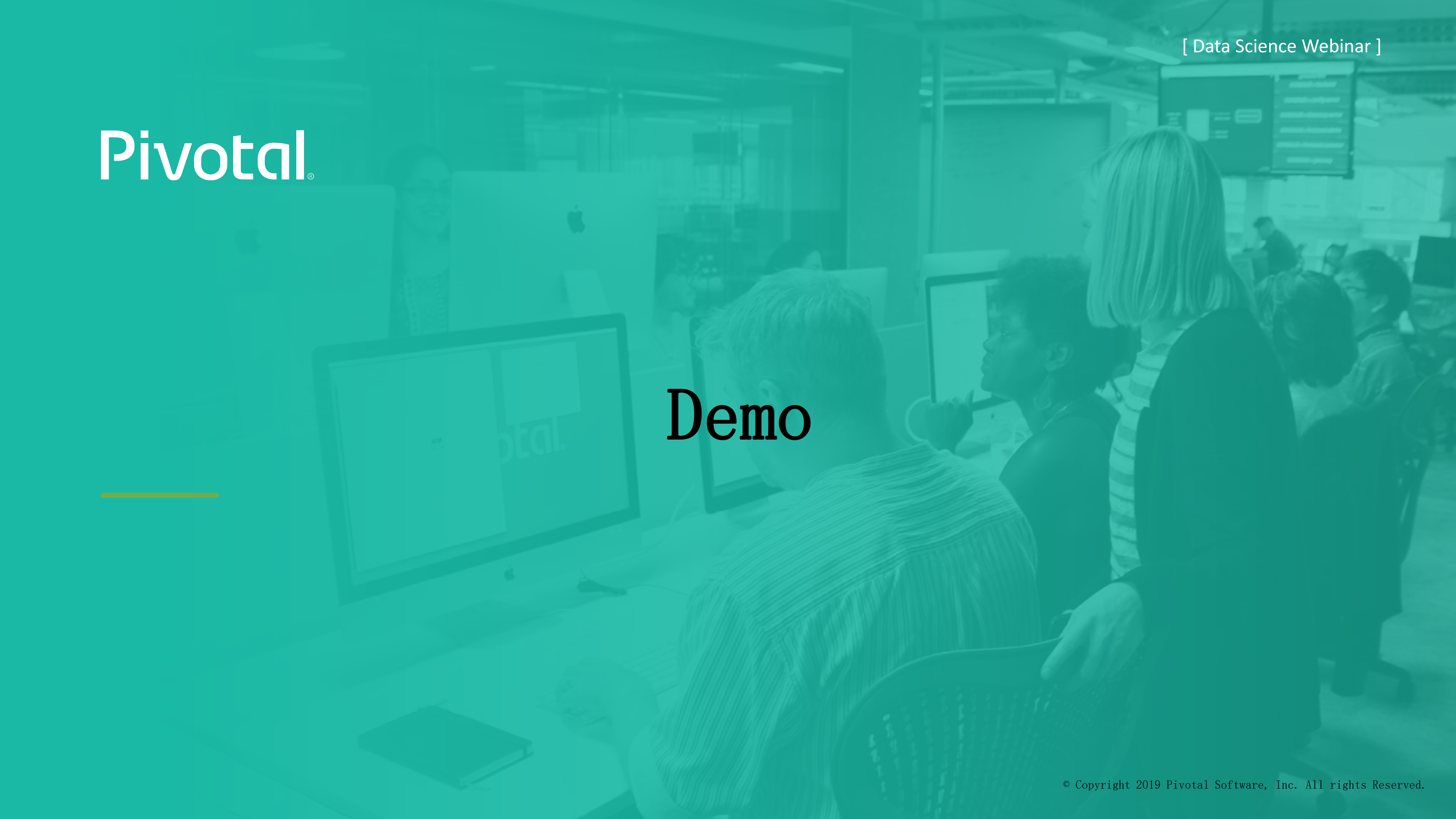


```
SELECT madlib.linregr_train( 'houses',                       -- Historical prices
                             'houses_linregr_bedroom',       -- Output model table
                             'price',                        -- Variable to predict
                             'ARRAY[1, tax, bath,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

预测



```
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;
```



Demo