THE MADLIB ANALYTICS LIBRARY

Presented by : Meghna Garg

CMPT 843

“In God we trust.
All others must bring data.”

- *Dr. W. Edwards Deming*

- Data is cheap

Friendly Fact : World's largest data warehouse of ~15 years ago can be stored on disks for less than about $2000

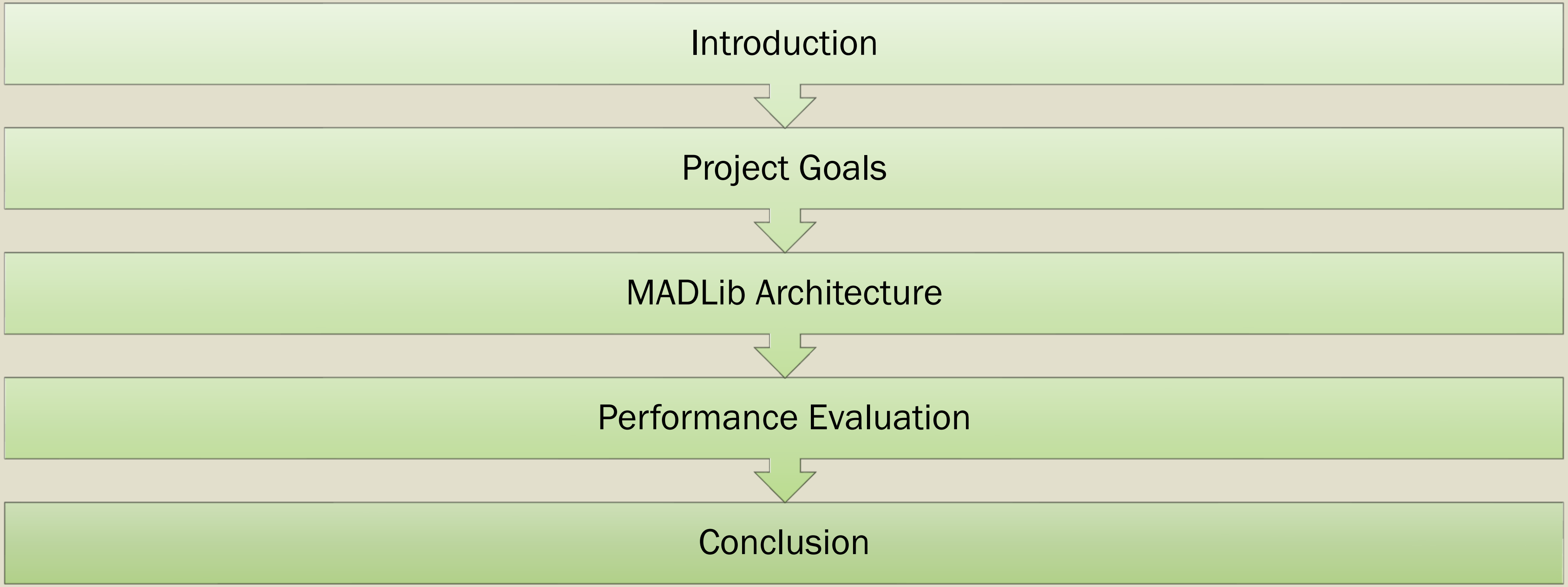
- Data is cheap

Friendly Fact : World’s largest data warehouse of ~15 years ago can be stored on disks for less than about $2000

- Makes “analysis” a common culture

“ THE GOOD NEWS IS, PROFITS ARE UP 74%, THE BAD NEWS IS, WE DON'T KNOW WHY. ”

Roadmap



Introduction

Project Goals

MADLib Architecture

Performance Evaluation

Conclusion

Introduction

M A D lib

Library for :

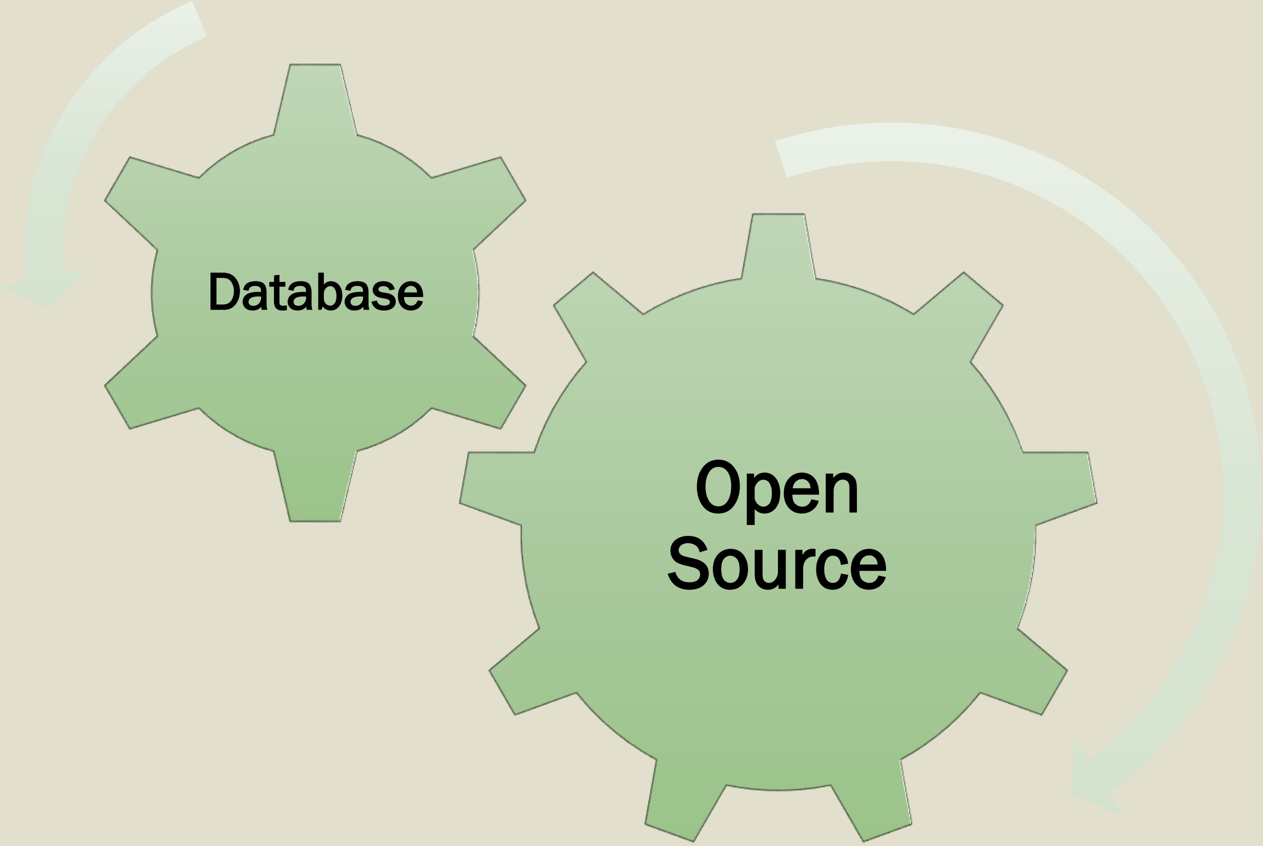
- Advanced (mathematical, statistical, machine learning)
- Parallel and scalable
- In-database functions

Attracts all kinds of data

Fast, Progressive queries and code.

Sophisticated ML methods for large data sets.

Project Goals



Project Goals

Databases

- Develop scalable and full-dataset analytics.
- Growing SQL-based analytics ecosystem.

Open source

- The benefits of customization
- Valuable data vs. valuable software
- Closing the research-to-adoption loop.
- Leveling the playing field, encouraging innovation.

Query Examples:

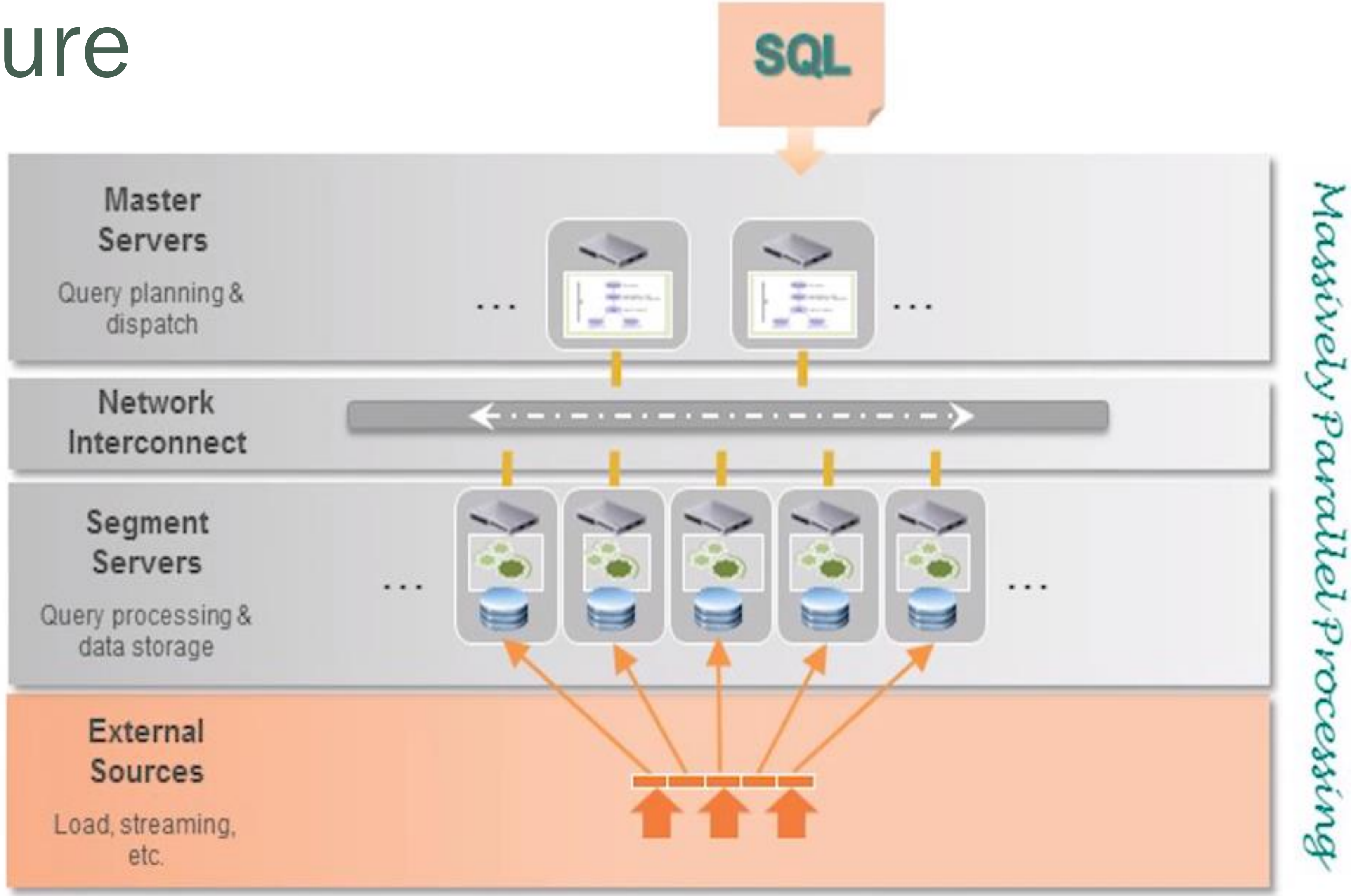
“How many people under the age of 30 visited the Toyota community in the past four days?”

- Simple data retrieval query in SQL.

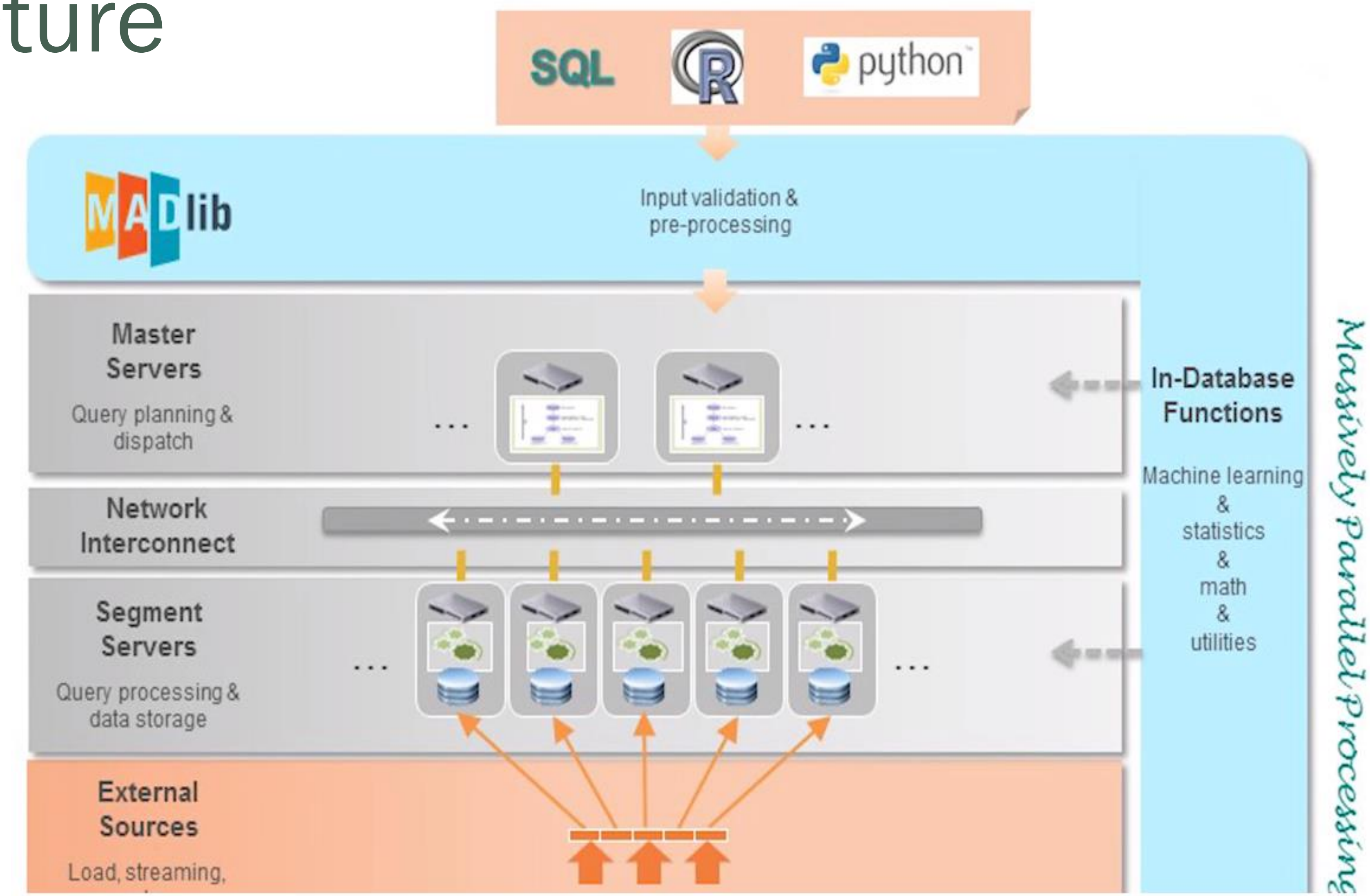
“How are these people similar to those that visited Nissan?”

- Open-ended, requires some statistics and the analyst to be in the loop.

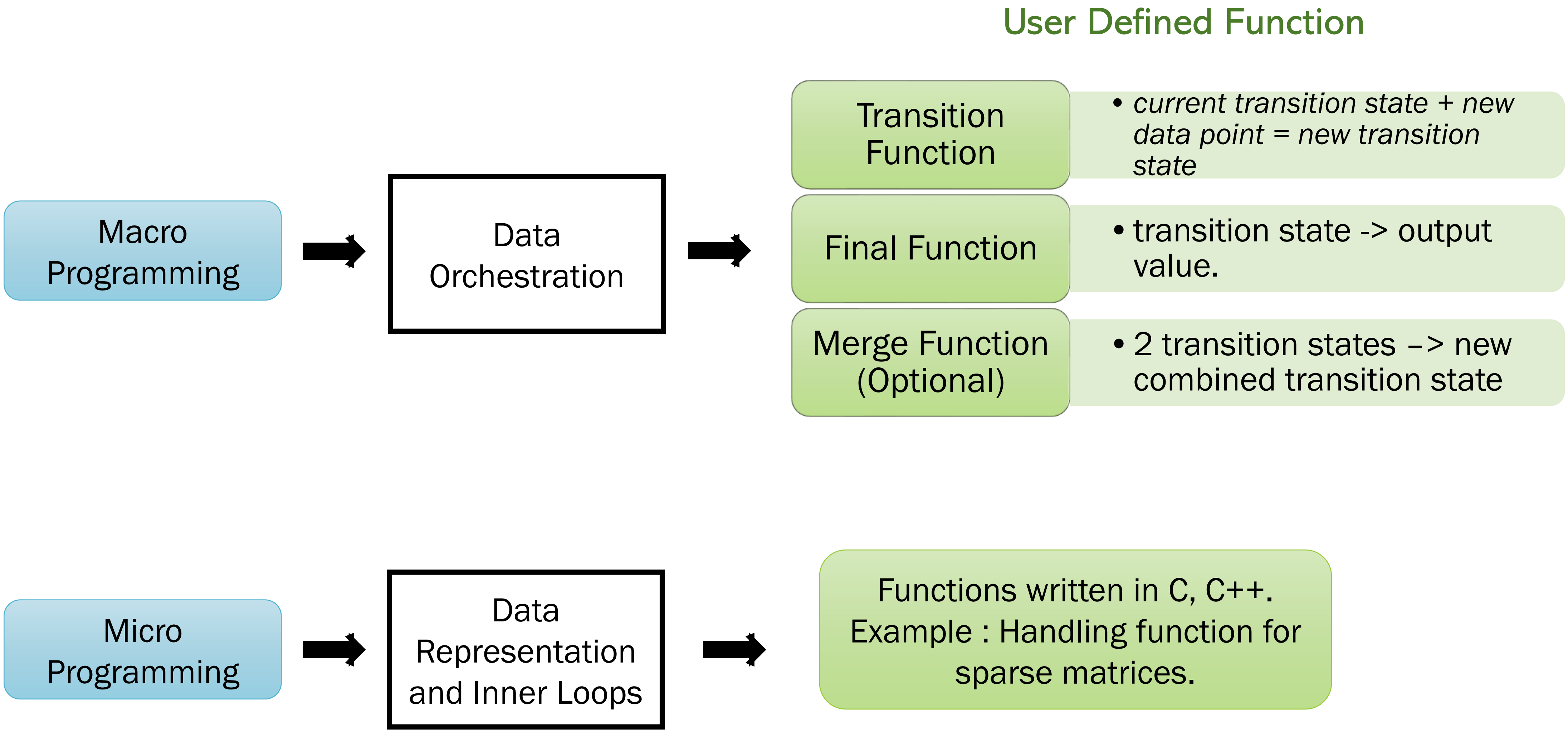
Architecture



Architecture



Architecture



User Defined Function

Macro Programming

Data Orchestration

Transition Function

- *current transition state + new data point = new transition state*

Final Function

- transition state -> output value.

Merge Function (Optional)

- 2 transition states –> new combined transition state

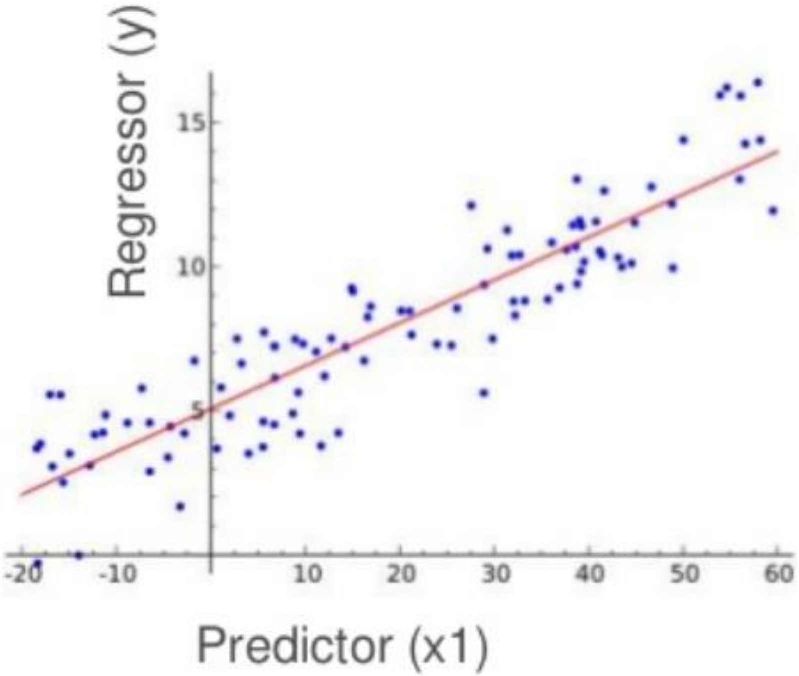
Micro Programming

Data Representation and Inner Loops

Functions written in C, C++. Example : Handling function for sparse matrices.

Linear Regression : Single Pass Iteration

GOAL : Find vector b that minimizes $\sum_{i=1}^{n}(y_i - \langle \widehat{\boldsymbol{b}}, \boldsymbol{x}_i \rangle)^2$



B can be calculated as:

$$\widehat{\boldsymbol{b}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

Where,

$$X^T X = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$$

$$X^T \boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{x}_i y_i$$

Summation is associative,
parallelization can be achieved.

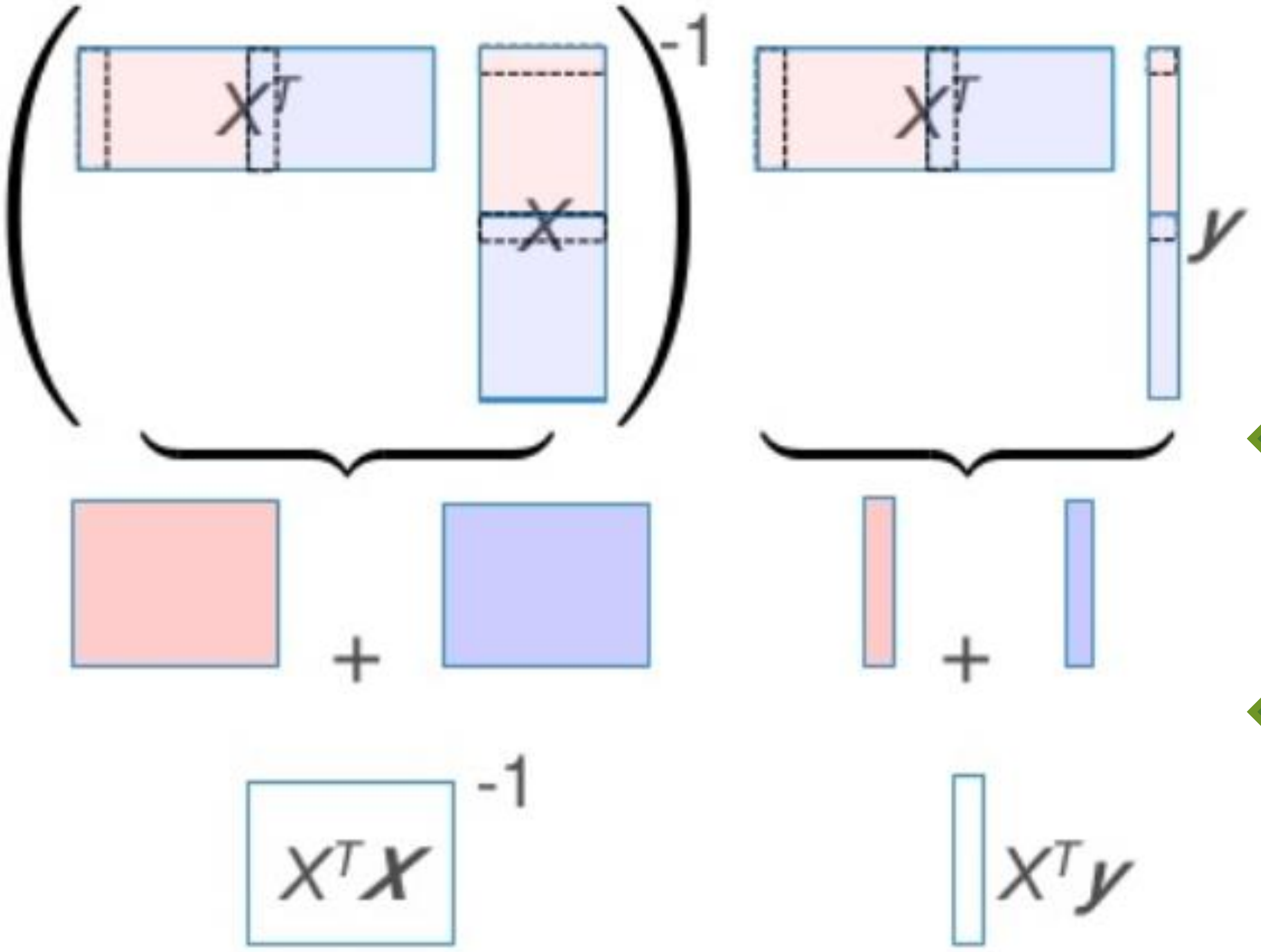
$$\widehat{\boldsymbol{b}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

$$\overset{X^T}{\begin{bmatrix} a & c & e & g \\ b & d & f & h \end{bmatrix}} \overset{X}{\begin{bmatrix} a & b \\ c & d \\ e & f \\ g & h \end{bmatrix}}$$

$$= \begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix} \begin{bmatrix} c & d \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} \begin{bmatrix} e & f \end{bmatrix} + \begin{bmatrix} g \\ h \end{bmatrix} \begin{bmatrix} g & h \end{bmatrix}$$

$$= \begin{bmatrix} a^2+c^2+e^2+g^2 & ab+cd+ef+gh \\ ab+cd+ef+gh & b^2+d^2+f^2+h^2 \end{bmatrix}$$

$$\widehat{\boldsymbol{b}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

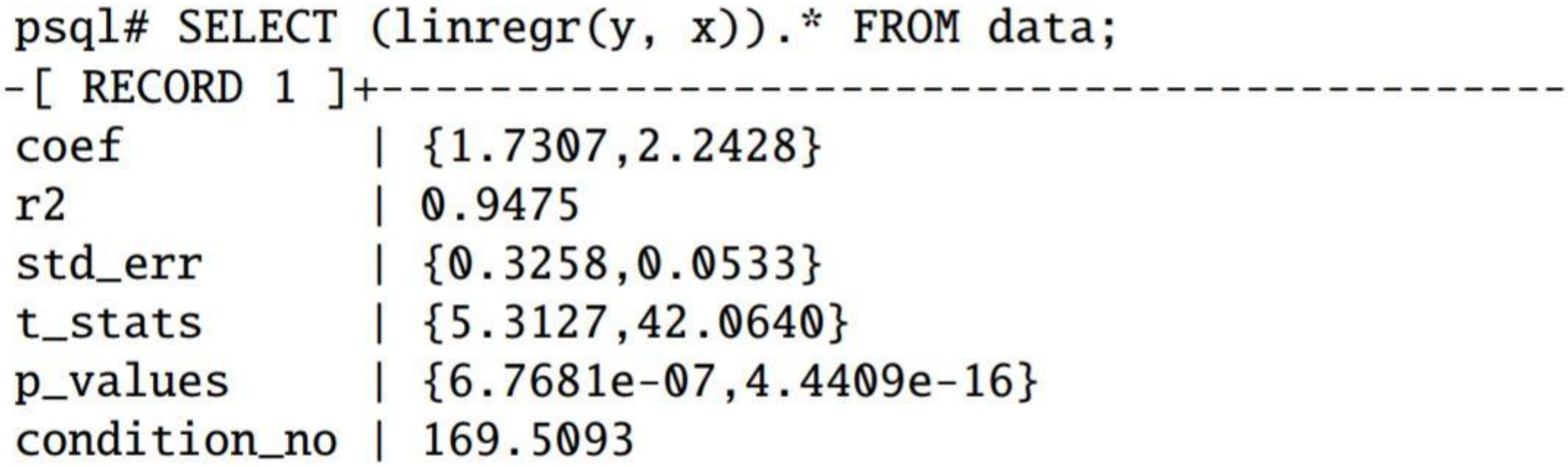


1st Phase

2nd Phase

Linear Regression : Single Pass Iteration

MADlib Implementation

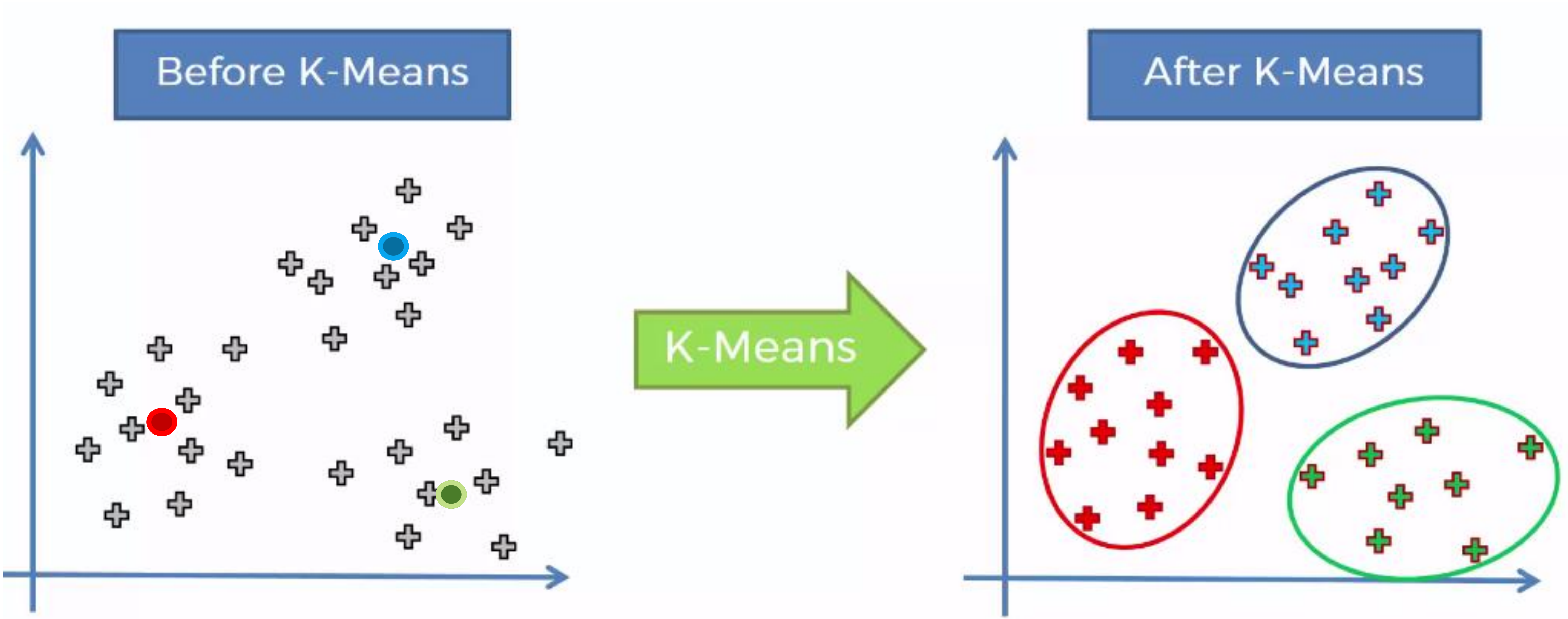


```
psql# SELECT (linregr(y, x)).* FROM data;
-[ RECORD 1 ]+--------------------------------------------
coef         | {1.7307,2.2428}
r2           | 0.9475
std_err      | {0.3258,0.0533}
t_stats      | {5.3127,42.0640}
p_values     | {6.7681e-07,4.4409e-16}
condition_no | 169.5093
```

Large State Iteration: k-means

Problem Statement: $x_1, \ldots, x_n \in \mathbb{R}^d$ $c_1, \ldots, c_k \in \mathbb{R}^d$

Goal: Minimize $\sum_{i=1}^{n} \min_{j=1}^{k} ||x_i - c_j||^2$



Large State Iteration: k-means

MADLib solution:

centroids

K

Centroid_ id	x	y
1	1	2
2	6	9

points

coords	Centroid_ id
(3,4)	1
(5,8)	2
(6,7)	2

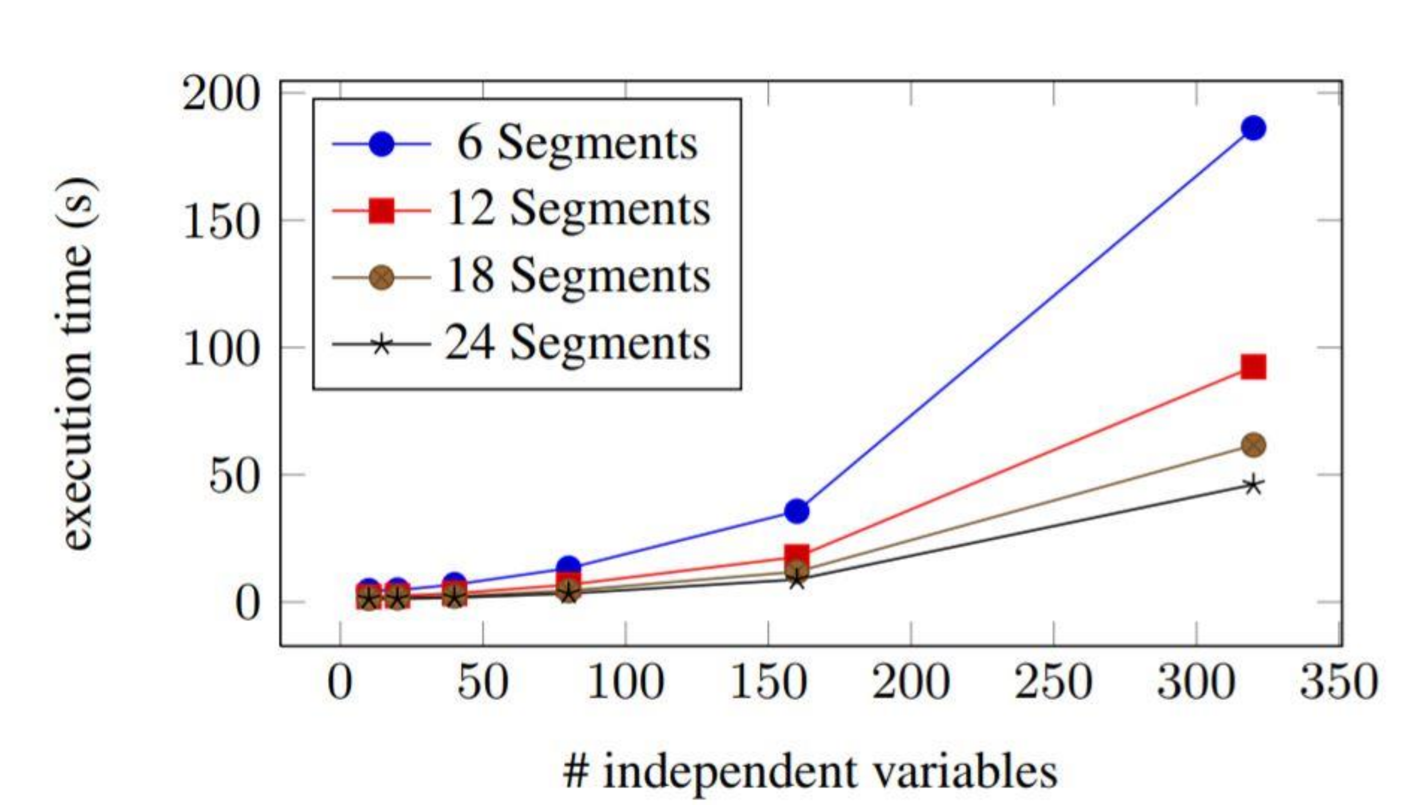
```
UPDATE points
SET centroid_id = closest_column(centroids, coords)
```

MADlib UDF

Matrix of centroids

Coordinate attribute of points table

Performance Graph



- Linear regression execution times using MADlib, 10 million rows
- As the number of segments increase, the execution time reduces.

Conclusion

- Designed to **fill a vacuum** for scalability analytics in SQL DBMS, and connect database research to market needs.
- Popular alternative to a DBMS infrastructure today is **Hadoop MapReduce**, which provides much lower-level programming APIs than SQL.
- **Room for enhancements** in its core treatment of mathematical kernels (e.g., linear algebra over both sparse and dense matrices) especially in out-of-core settings.
- It is still in its **early stages of development**, but is already in use both at research universities and at customer sites.

Further Reading..

- Joseph M. Hellerstein , Christoper Ré , Florian Schoppmann , Daisy Zhe Wang , Eugene Fratkin , Aleksander Gorajek , Kee Siong Ng , Caleb Welton , Xixuan Feng , Kun Li , Arun Kumar, The MADlib analytics library: or MAD skills, the SQL, Proceedings of the VLDB Endowment, v.5 n.12, August 2012
- Documentation: https://madlib.apache.org/docs/latest/index.html
- Website: http://madlib.net
- Online resources : https://www.youtube.com/watch?v=DGPZwpB92Aw

Questions