Session: 3973 – Integrating SAS, Apache Hadoop, and an Enterprise Data Warehouse in a Single Solution

Bob Matsey – Teradata Senior Advanced Analytic Consultant

Agenda

- SAS & Teradata Partnership
- Benefits of In Database
  - Coding Example
- Customer Improvement Examples
- VIYA Integration with Teradata
- Teradata's UDA
- Agile Analytics with Data Labs
- In-DB Decision management with Decision Manager
- IoT Example – Wearables
- Questions?

The SAS & Teradata Partnership Overview

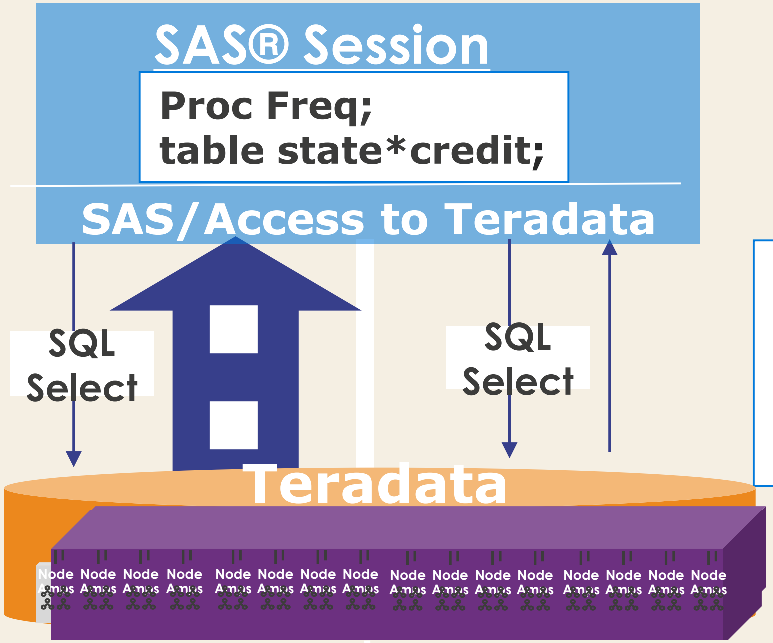
- Teradata is an Authorized Global Reseller of SAS Solutions
- Partnership began in 2007 to improving analytic performance
- Focus on joint product collaboration and customer success
- More than 450 sales to over 240 customers already
- Teradata has dedicated R&D teams onsite at SAS
- Regular collaboration on Joint Product Roadmap to ensure seamless product integration

TERADATA

Example of In Database with Proc FREQ

Traditional Technique

- Request all rows
- Select state, credit from credit data;
- Calculate frequency count



SAS® Session

```
Proc Freq;
table state*credit;
```

SAS/Access to Teradata

SQL Pushdown

- Select count(*), state, credit from . . . group by state, credit;
- Return only count

	Traditional	SQL Pushdown
Rows Returned	9,000,000	51
Time to Process	55 seconds	2 seconds

In Database Coding Example

Testing In-database Functionality

Not Running In Database Example: (SQLGENERATION=NONE;) will tell the code to NOT run In database.

Example 1 – Shows running a simple Proc Freq in a SAS program against a larger dataset (at least 1- 2 million rows) without in-database capabilities turned on & with SAS log turned on. Then review the SAS log for duration and database performance

Code Example:

```
12   libname tdXXXX teradata server="XXXserver" database=XXXXP user=&user password=&password;
13
17   options sastrace=(,,,ds) sastraceloc=saslog nostsuffix;
20   OPTIONS SQLGENERATION=NONE;
21        PROC FREQ DATA=tdxxxx.xxxxx;
22          TABLES XXXX_XXXX;
23        RUN;
```

Running In Database Example: (SQLGENRATION=DBMS;) Will tell the code to run In database

2nd Example is: Running the same Proc Freq code in a SAS program with the following options: options SQLGENERATION=DBMS . This option says to run the code In database whenever it can, so I highly recommend putting this on ALL your SAS code.

```
12   libname tdXXXX teradata server="XXXserver" database=XXXXP user=&user password=&password;
13
17   options sastrace=(,,,ds) sastraceloc=saslog nostsuffix;
20   OPTIONS SQLGENERATION=DBMS DBIDIRECTEXEC set=truncate_bigint 'yes' MSGLEVEL=1;
21        PROC FREQ DATA=tdxxxx.xxxxx;
22          TABLES XXXX_XXXX;
23        RUN;
```

Running these two test will show,
Example 1 – this will NOT run In database.
Example 2 – will run IN database.

In-Database Functionality

SAS/Access to Teradata Base Procedures:

- PROC APPEND
- PROC CONTENTS
- PROC COPY
- PROC DATASETS
- PROC DELETE
- PROC FORMAT
- PROC FREQ
- PROC MEANS
- PROC PRINT
- PROC RANK
- PROC REPORT
- PROC SORT
- PROC SQL
- PROC SUMMARY
- PROC TABULATE

DQ Accelerator for Teradata

- Match code
- Parsing/Casing
- Gender/Pattern/Identification analysis
- Standardization

SAS Code Accelerator for Teradata

- PROC DS2

SAS Scoring Accelerator for Teradata

- EM/STAT* Models

SAS Analytics Accelerator for Teradata

Statistical Analysis Procedures:

- PROC CANCORR
- PROC CORR
- PROC FACTOR
- PROC PRINCOMP
- PROC REG
- PROC SCORE
- PROC TIMESERIES
- PROC VARCLUS

SAS Enterprise Miner

- PROC DMDB
- PROC DMINE
- PROC DMREG (Logistic Regression)
- Also nodes for Input, Sample, Partition, Filter, Merge, Expand

- PROC SCORE works with coefficients from:

- PROC ACECLUS
- PROC CALIS
- PROC CANDISC
- PROC DISCRIM
- PROC FACTOR
- PROC PRINCOMP
- PROC TCALIS
- PROC VARCLUS
- PROC ORTHOREG
- PROC QUANTREG
- PROC REG
- PROC ROBUSTREG

Itaú

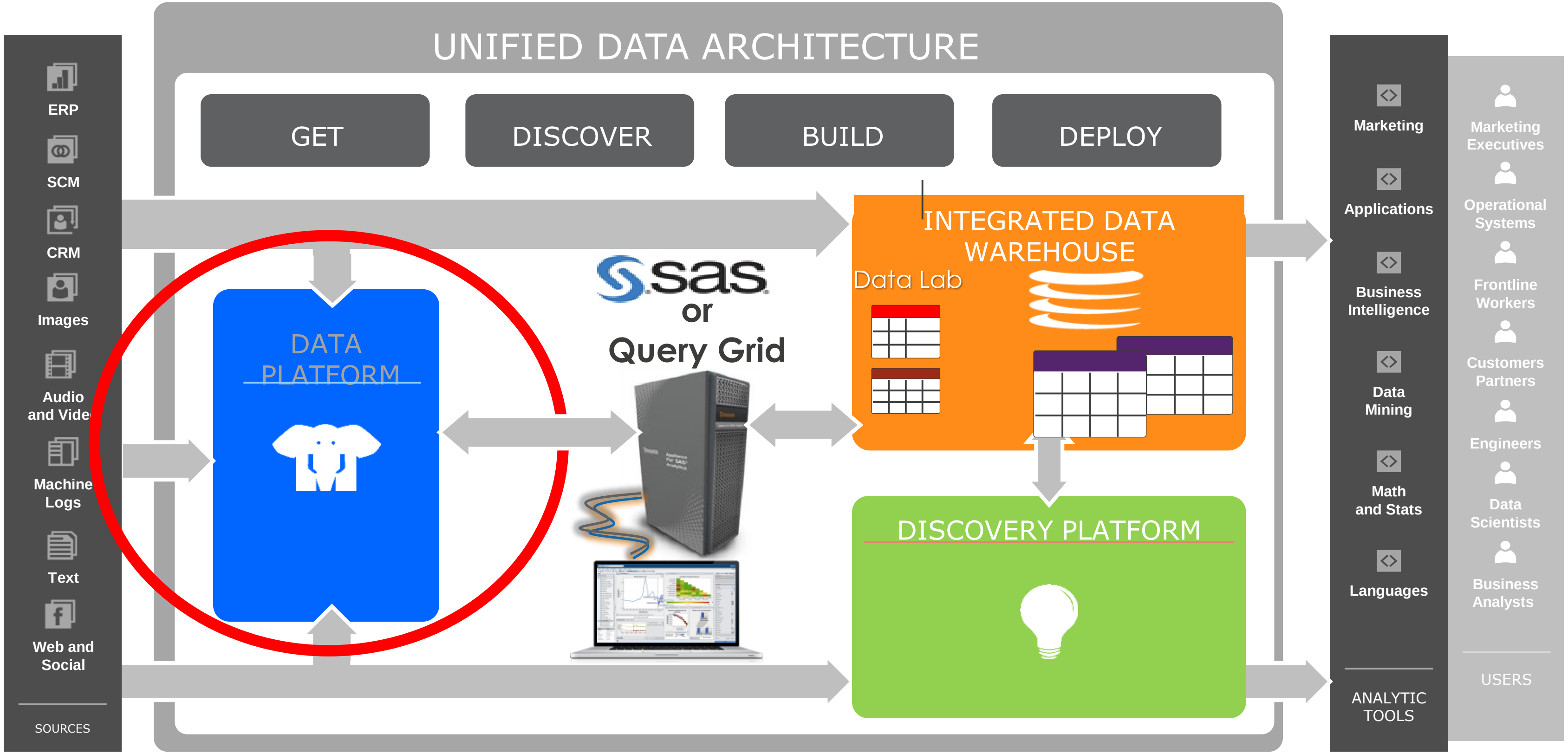
#	Process Name	SAS + Oracle	SAS + 2 Node Teradata	X Faster
1	Horizontalization	18 hrs 7 mins	32 mins	**34 X**
2	Horizontalization	15 hrs 3 mins	33 mins	**27 X**
3	Variable Calculation	6 hrs 57 mins	4 mins	**104 X**
4	Scoring	10 hrs 56 mins	11 mins	**60 X**
5	Data Mart Generation	27 hrs 50 mins	1 hour 28 mins	**19 X**

SAS Programs Results

- **Highlights**
  - **GE – long running queries with sort**
    - **Execution in Teradata only took 3.75 minutes – 1600X – Old way 103 hours!**
  - **OSCAR – running against Commercial Market Scan data**
    - **Execution in Teradata was 1 hour 50 minutes against 3 times larger data set – Old way 231 hours**

				SAS Only			SAS + Teradata				
#	Business Line	SAS Log Name	# of Steps	Days	Hours	Minutes	Days	Hours	Minutes	% of SAS Only	X Times Faster
1	oscar	oscar_mdcd_v3.log	945	9.6	231.6	13,894.1		1.83	110.0	1%	126.3
2	GE	mk_text_observation_f_sort.log	3	4.3	103.0	6,178.0			3.8	0%	1,625.8
3	ingenix	dcf ~ i3_qc.log	3,401		15.1	908.2			45.8	5%	19.8
4	humana	humana_dups.log	28		5.6	333.3			18.8	6%	17.7
5	ingenix	analysis ~ 100_indentifying_initial_patients.log	12		1.7	99.4			1.5	2%	66.3
6	ingenix	analysis ~ 200_extracting_mx_claims.log	11		1.1	68.1			1.0	1%	68.1
7	ingenix	analysis ~ 210_extracting_rx_claims.log	12			28.5			0.4	1%	71.3
8	ingenix	dcf ~ mk_s2009_r12q2.log	20		1.6	98.2			3.8	4%	25.8
9	ingenix	dcf ~ mk_s2010_r12q2.log	20		1.5	87.8			3.6	4%	24.4
10	ingenix	dcf ~ mk_s2011_r12q2.log	20		1.0	61.8			3.4	6%	18.2
11	ingenix	dcf ~ mk_m2011_r12q2.log	20			56.8			2.3	4%	24.7
12	ingenix	dcf ~ mk_r2011_r12q2.log	20			41.9			3.3	8%	12.7
13	pharmetrics	130_af_all_claims.log	12		1.7	101.2			4.7	5%	21.5
14	pharmetrics	110_af_claims.log	6			52.0			2.7	5%	19.3
15	pharmetrics	183_table8d.log	43			30.8			3.4	11%	9.1
16	pharmetrics	183_table8b.log	39			30.4			1.5	5%	20.3
17	pharmetrics	162_table2b.log	30			20.6			2.8	13%	7.4
18	pharmetrics	182_table8d.log	43			23.8			1.8	8%	13.2

Agile Analytics – Integrating Data into a Single Solution



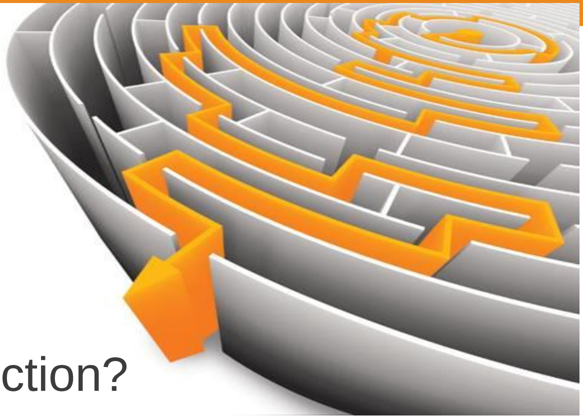
TERADATA

Dealing with All Types of Data

Enabling Self Service Data Loading & Analytics with a Teradata's Data Labs

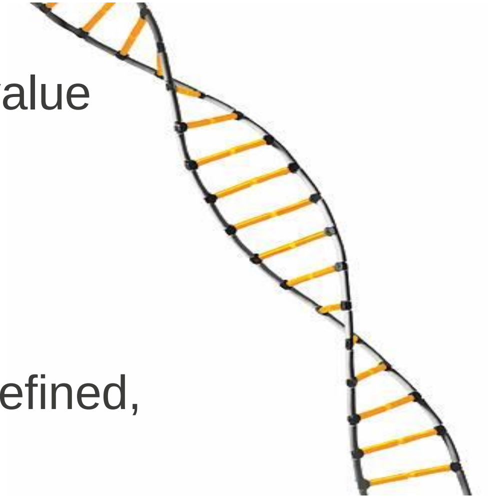
Business Need for Agile Analytics

Flexibility vs. IT Process



- Analyze quickly
  - Test New Theories
  - New Data
- Does the new data provide additional insight?
- Does the new insight cause a change in thinking or direction?
- Test Fast
  - Was the theory right? (**<u>Success or Failure</u>**)
- Productionize what works; discard what doesn’t!
  - Add the new application
  - Add the new data
  - Or delete and move on!

Don’t Just Use Production Data – Evolve It



3rd Party Data

- Often rented, supplier data and/or format needs to change, value needs validation, only applies to some projects

Temporary & Research Data

- Exploratory metrics and aggregates, requirements not fully defined, short lived, early stage work

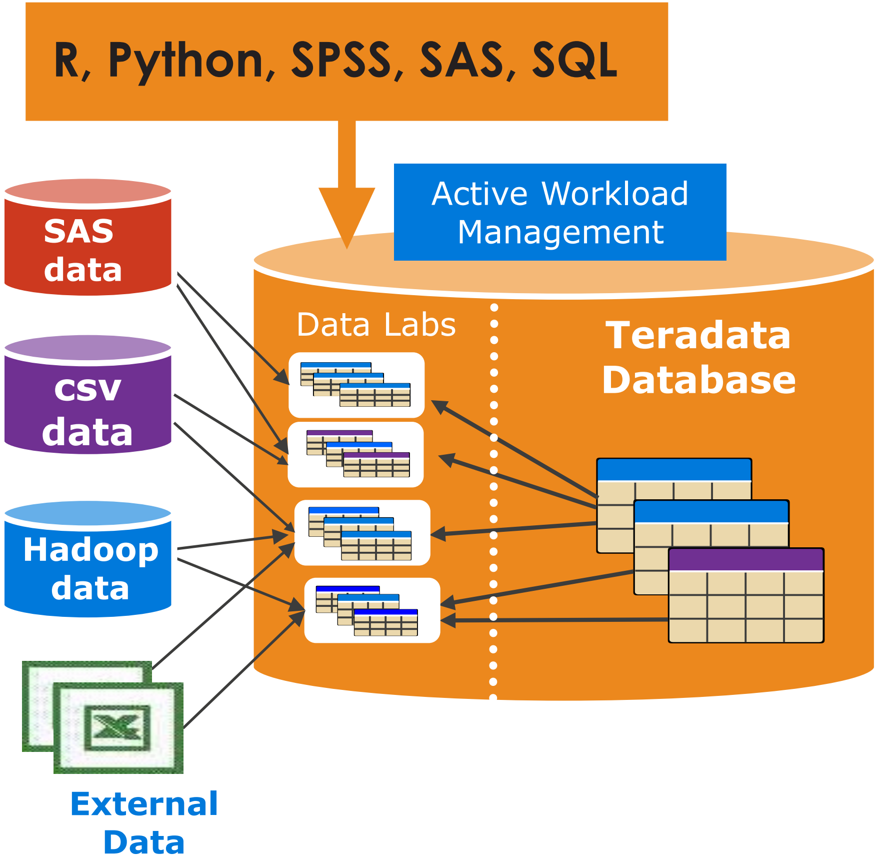
Pre-Production Data & Prototypes

- Excel Spreadsheets
- Oracle, SQL Server, SAS datasets, Access DB, others can be loaded
- Comma delimited, space delimited, other data types

Teradata Data Labs Architecture

Analytic Sandboxes with Governance

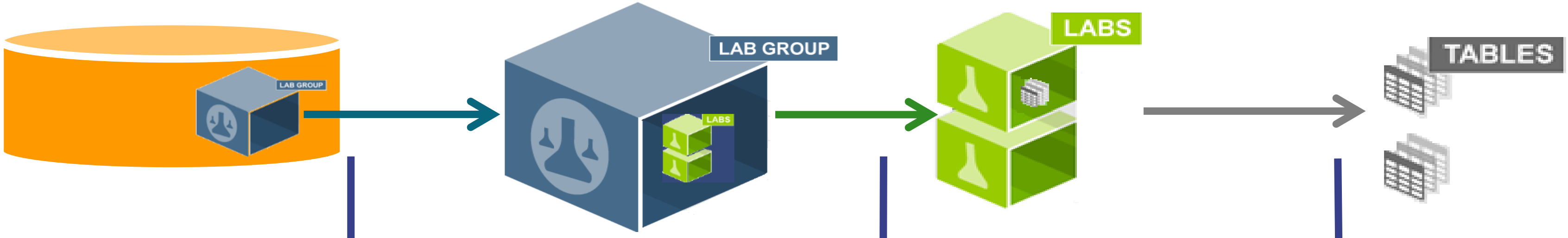
- Data Lab(s) inside the EDW or DW Appliance to easily join to production data via Views
- Load experimental, untested data from external sources
- Rapid prototyping, exploratory and experimentation analysis
- Beyond a Sandbox
  - An architecture that enables governance
    - ✓ Works within your current data warehouse environment
  - Data lab portlets for IT and Business analyst
    - ✓ Self-provisioning system that simplifies implementation, management and use



Teradata Data Lab Hierarchy

Data Lab Objects

Data Lab hierarchy to manage user groups, space, and workload



Database

Database where the lab group resides

- Normal Teradata user-database

Lab Group

Workspace allocated for a group of users to create their own data labs.

Groups can be arranged by department or project

Groups can be made private

Lab Group is a fixed size that's shared by users.

Data Labs

Workspaces allocated for analysis

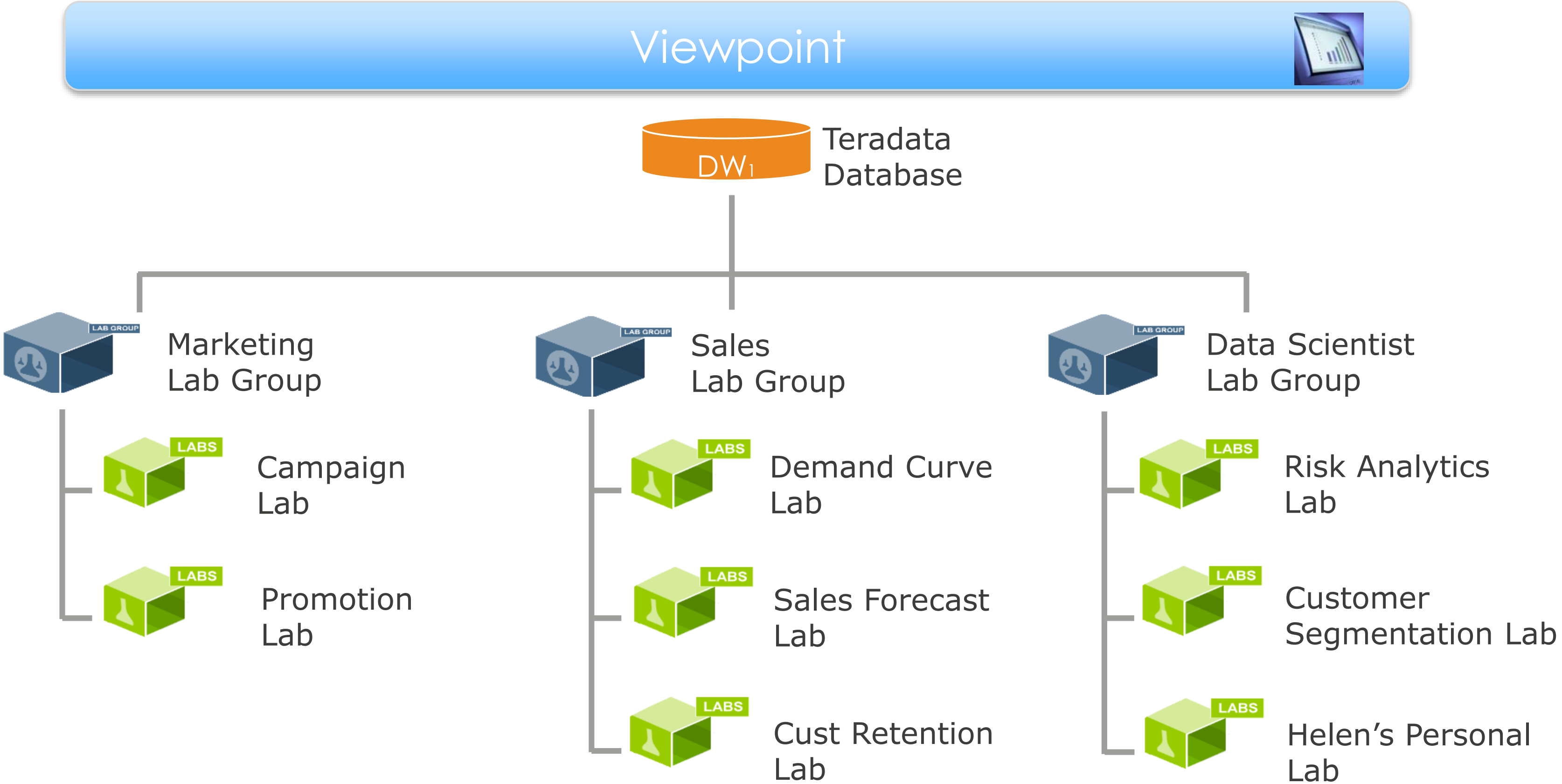
- Can be for a single user or X number of users
- Data Labs expire
- Data Labs are allocated with a fix size, but are elastic

Table

Database table to store the data

- User can create table and load data

Example: Lab Group Hierarchy



Viewpoint

DW1

Teradata Database

Marketing Lab Group

Campaign Lab

Promotion Lab

Sales Lab Group

Demand Curve Lab

Sales Forecast Lab

Cust Retention Lab

Data Scientist Lab Group

Risk Analytics Lab

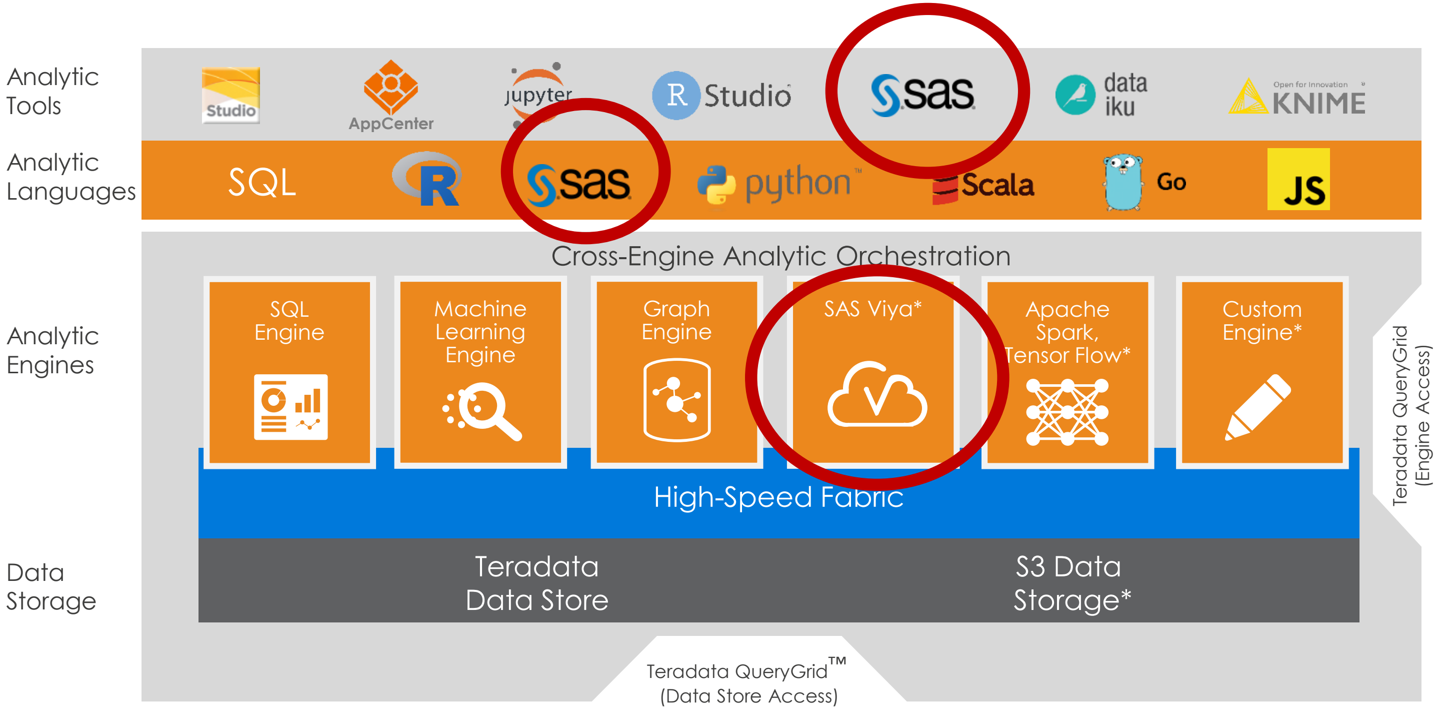
Customer Segmentation Lab

Helen’s Personal Lab

TERADATA.

SAS is Built into the Teradata Analytics Platform

Teradata's strategy is to allow the customer to choose the tools they want



* future capabilities

QUESTIONS ???

TERADATA®