Bob Matsey – Senior Analytic Consulant

- Bob Matsey is a Senior Analytic Consultant at Teradata Corporation. With more than 26 years of experience in the information technology industry, Bob consults with customers on how to push Analytic process in database and implementing Agile Analytics in Data Labs to deliver value-added business solutions in analytics, data warehousing and data management. His combined technical and business background includes extensive experience as a customer for 16 years, the last 10 years with SAS & Teradata (5 –SAS, 5 – TD)
- Bob holds a Bachelor of Science in Mgmt & Computer Science from Idaho State University and an MBA degree from the McColl School of Business at Queens University, Charlotte, NC.

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

12660 - Integrating SAS, Hadoop, and the Data Warehouse in a Single Solution!

What is Hadoop?

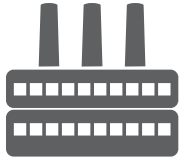
- Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware.
- Hadoop offers massive data storage
- Hadoop handles structured and unstructured data (including audio, visual and free text).

Why is Hadoop Important?

A Massive Influx of Data from New Sources

Retail

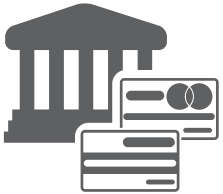
Social Media



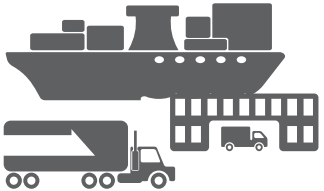
Manufacturing

Geolocation

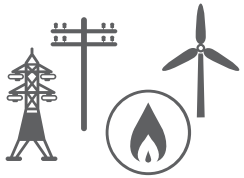
Medicare



Banking / Credit



Shipping & Distribution



Utilities

More Data Requires Better Data Management

More data from more sources comes with specific challenges:

- Format (voice, text, images, logs, etc.)
- Quality
- Data Value
- Storage
- Accessibility
- Security
- **Cost**

What are Companies Doing with Hadoop?

- 46% - Data warehouse extensions
- 46% - Data exploration and discovery
- 39% - Data staging area for data warehousing and data integration
- 36% - Data lakes
- 36% - Queryable archive for nontraditional data (web logs, sensor, social, etc)
- 33% - Computational platform and sandbox for advanced analytics

Source: The TDWI Best Practices Report, *Hadoop for the Enterprise* by Philip Russom

The Common Denominator with Hadoop

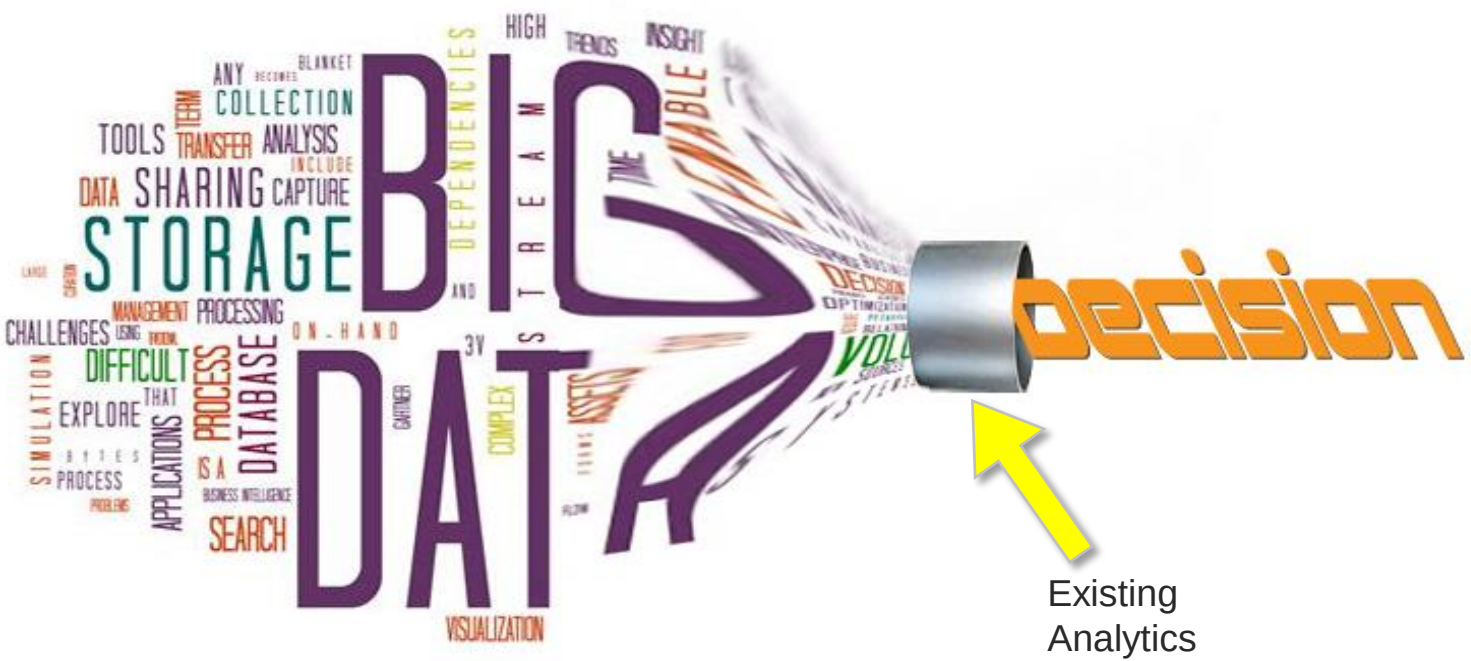
Hadoop is a ~~*replacement*~~ ***complement*** to:

- Business Intelligence;
- Data Warehousing;
- Data Integration;
- Analytics

The Keys to Unleashing Value From Data with Hadoop

- **Efficiently** and **economically** manage the larger volume of data
- **Operationalizing** the data to solve real problems
- Generating **new insights** from the data
- **Demonstrating the value** of the added insight to your business
- To make the **insights easily available** across the organization

The Big Data Analytics Conundrum

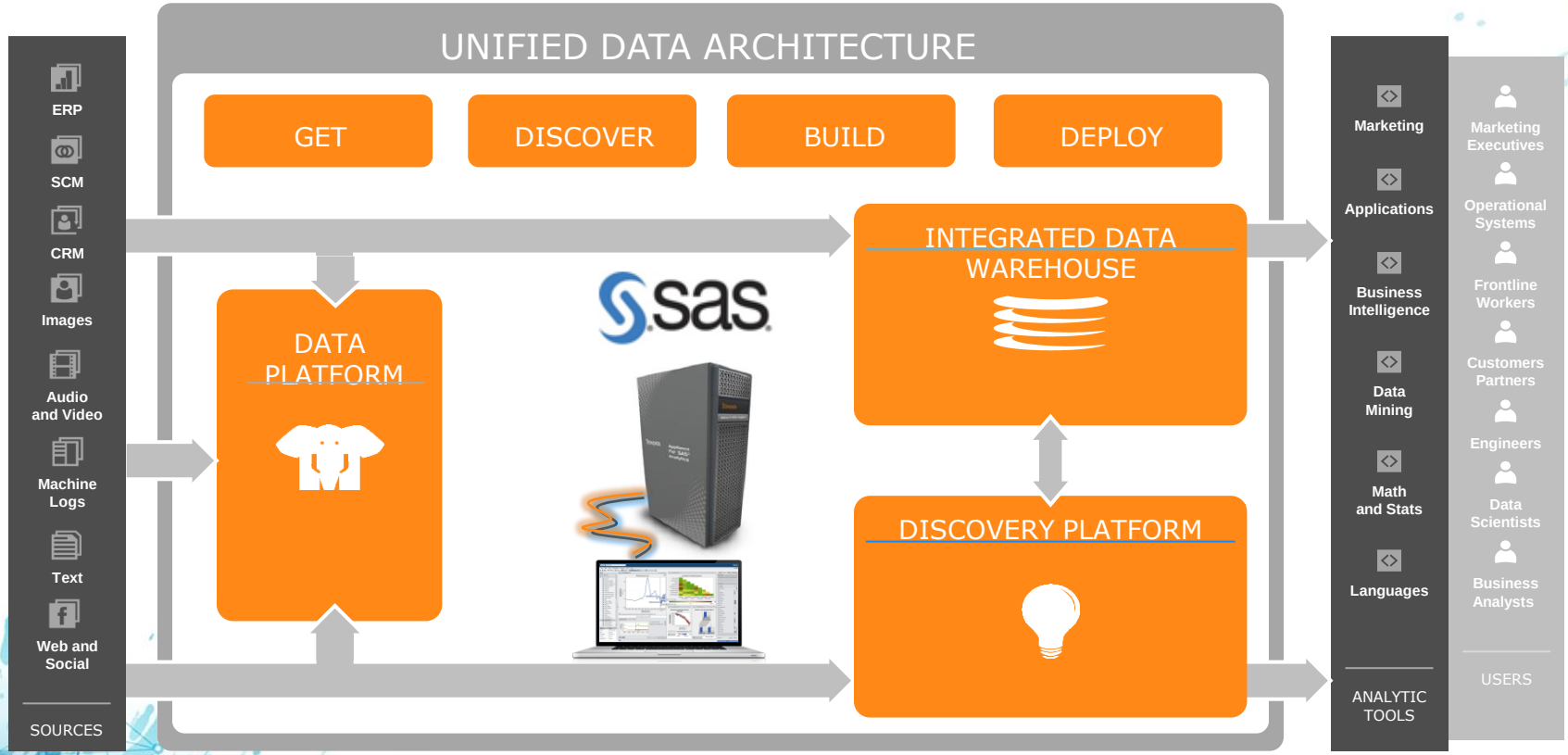


Existing
Analytics
Processes

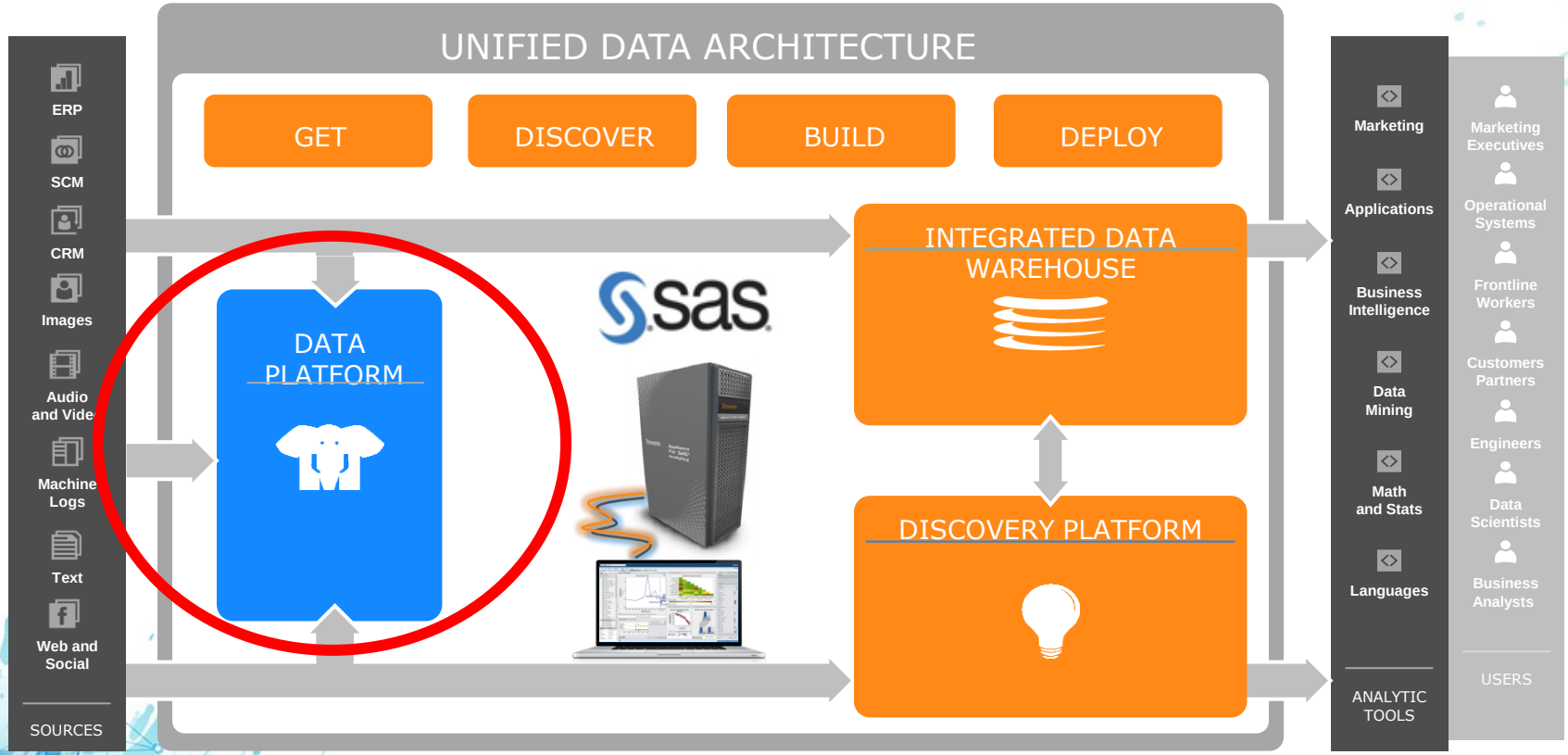
Gartner's Logical Data Warehouse

- Workload-specific engines/platforms for storage and analysis
- Type-specific analytic functions and engines
- Integration technologies to copy and move data
- Easy-to-use environments for business and tech analysts

Teradata's UDA Creates a Single Analytics Platform



Hadoop in a Key Component of the Teradata UDA



Teradata's Appliance for Hadoop

- Pre-configured and optimized specifically for big data
- Flexible configuration with Hortonworks or Cloudera
- Avoid the high costs and hassle of a do-it-yourself system

Integrated, Enterprise-ready Hadoop - Delivered Ready to Run

Teradata's Appliance for Hadoop **with SAS**

Add a SAS server directly to the cabinet to connect your analytics directly to your data.

Creating a Complete Enterprise Analytics Ecosystem

Teradata Active
Data Warehouse

Teradata
Appliance for SAS
In-Memory Analytics

Teradata Appliance
for Hadoop

Enterprise Analytics
Ecosystem

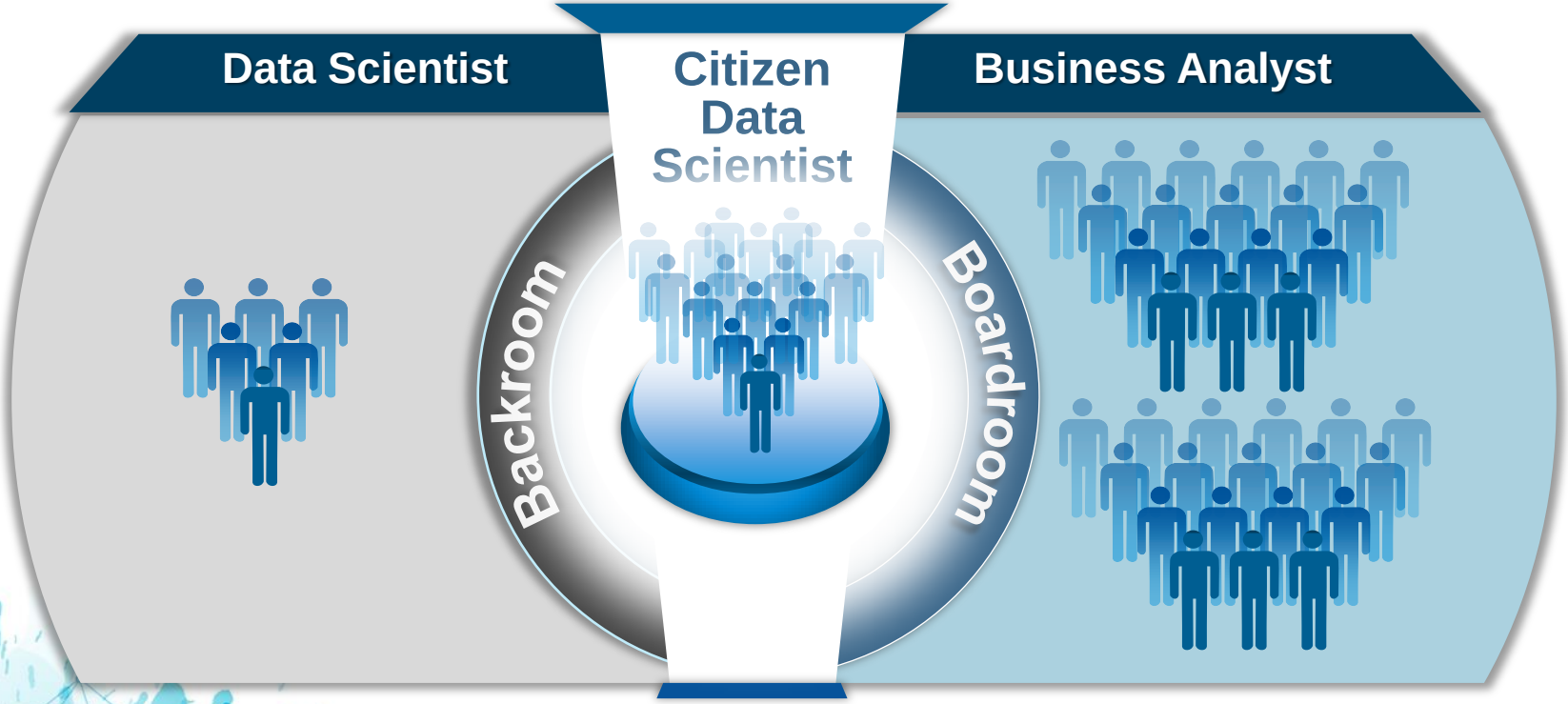
SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

A Use Case for Getting Value from Hadoop within the Analytics Enterprise

#SASGF

Analytics Shaping the Future of the Organization

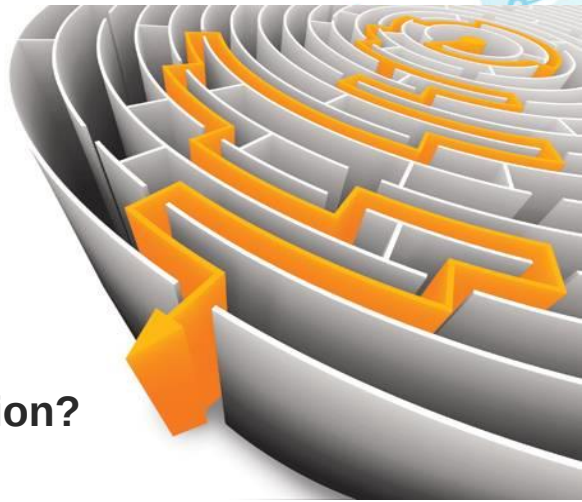


The Citizen Data Scientist

> A person who creates models that **use predictive or prescriptive analytics**, but whose **primary job function is outside of the field of statistics** and advanced analytics. They are "power users" who will be able to perform simple and moderately sophisticated analytic applications that would previously have required more expertise. They often reside in the lines of business and have deep domain expertise
>
> *- Gartner Inc.*

Enabling Self-Service Data for Citizen Data Scientists

Flexibility vs. IT Process



- **Analyze quickly**
  - New Theory
  - 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? (Success or Failure)
- **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

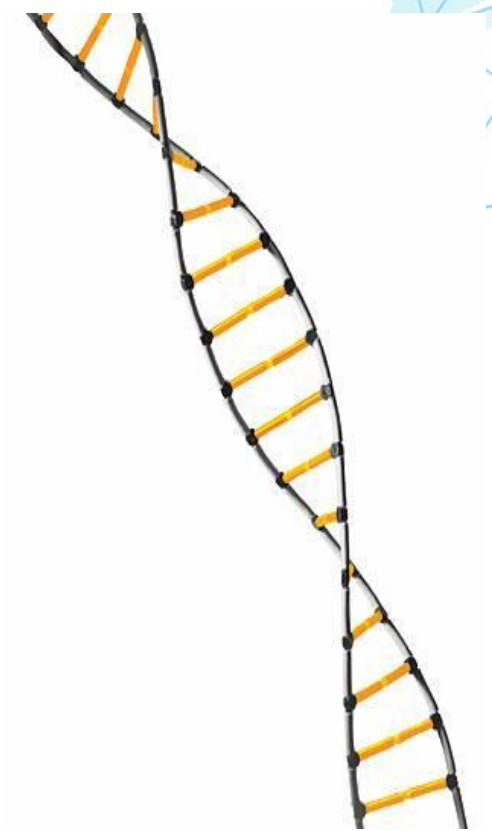
- Often rented, supplier and/or format can change, value needs validation, only applies to some projects

Temporary & Research Data

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

Pre-Production Data & Prototypes

- Either of the above can transform into this
- Process is defined and proven, there is interest in formalizing it, but it only exists in the Data Lab



What is an Analytic Data Lab?

Collection of data on which in-depth analysis can be done to answer critical business questions

- Ideal for data exploration, data transformation, analytic development, POC and prototyping
- User allowed to drop data in for brief time periods without meeting production warehouse criteria
- Data is segregated from the production database
- Data has a limited shelf life (Duration)
- Accessed by a set of known users making ad hoc request or process intensive analytic tasks

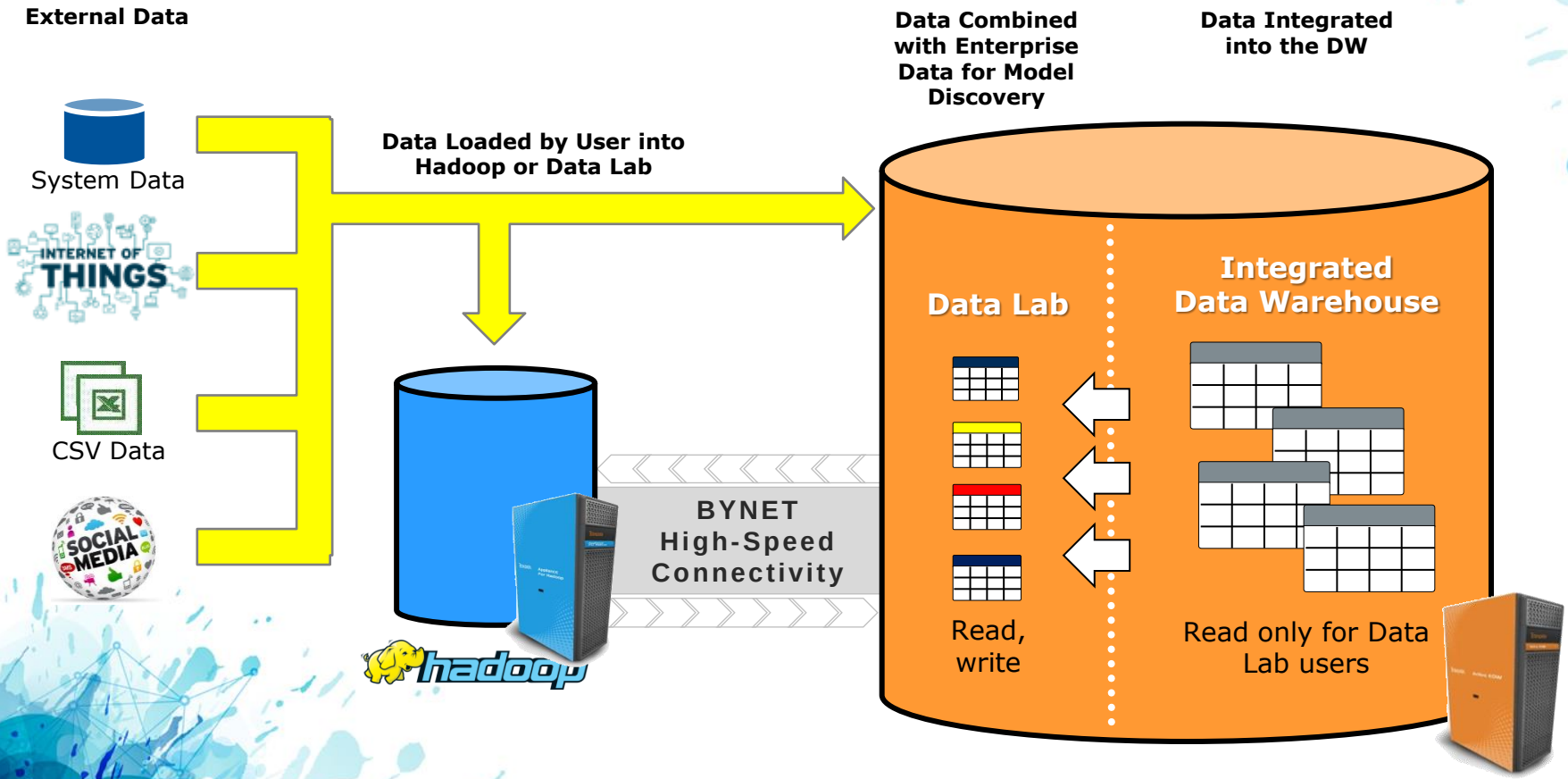
What a Data Lab is NOT!!

- It is not a ‘Production environment’
- It is not a place you can get access to data that you don’t have access in Production (no cheating..)
- It is not a place you can stay in ‘forever’
  - There is a defined & agreed amount of time
  - Examples of Customer ‘Best Practice’ Provisioned timeframes are:
    - » 7 days
    - » 1 month
    - » 3 months
    - » 6 months (with Business Justification)

Difference Between a Sandbox & Data Labs

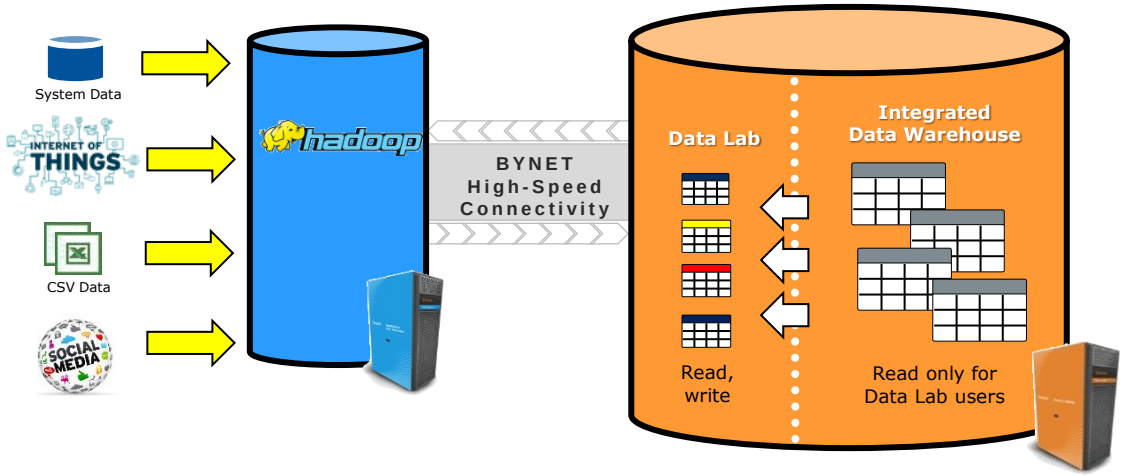
Function	Sandbox	Data Labs
Runs Unsupported Production Apps	Yes	No
Environment Backup & Recoverable	No	Yes
Speed of Processing & Priority	No	Yes
DBA Support (agreement)	No	Yes
Users can impact & impact other users	Yes	No
Space is **never** cleaned up or reclaimed	Yes	No
Work load management set up	No	Yes
Users Trained on Optimal use	No	Yes

Enabling Self-Service Data for Citizen Data Scientists



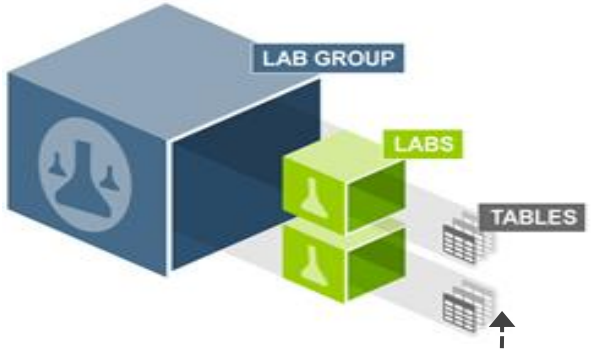
Data Lab Within Your Teradata Production DW

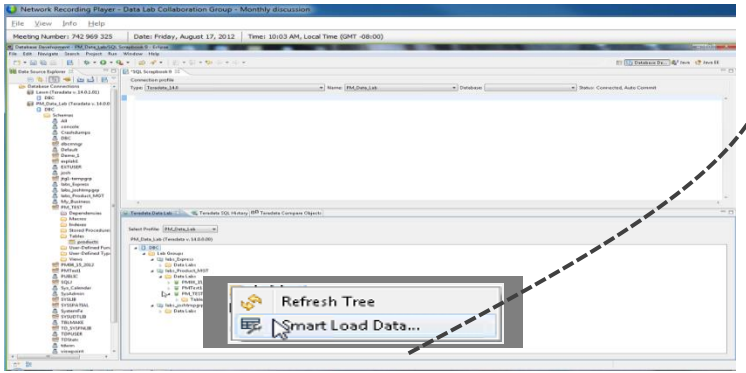
- Join with IDW data (No data exports!)
- New or experimental data quickly loaded into your data lab
- Used for rapid prototyping, experimentation, and exploratory analysis
- Easy to use self-provisioning and management
  - Extend analytics to more users
- Minimal IT support required after initial setup



Teradata Studio Express – Data Lab

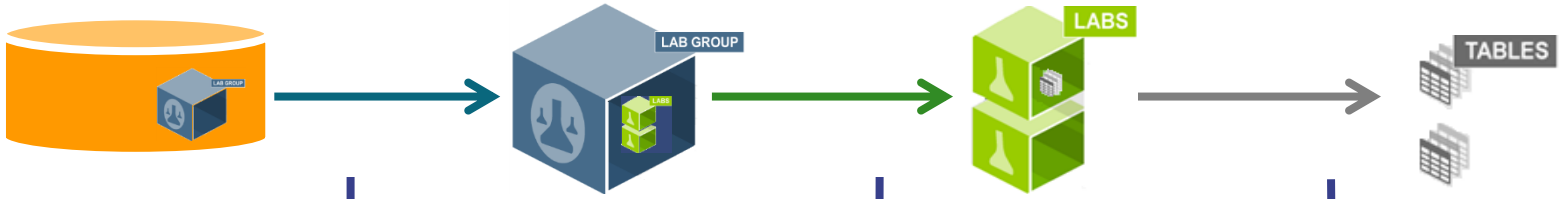
- Self-Service smart loader
  - Automatically determines data types
  - Automatic table creation
  - Loads, appends or replaces data
  - Excel or CSV files
- Smart loader from Hadoop
  - Browser for Hadoop files
  - Automatically maps data types
  - Creates new dbs tables





Teradata Data Lab Hierarchy

Data Lab hierarchy to manage user groups, space and workload



Database

Database where the Lab Group resides

- A Teradata user-database

Ownership: DBA

Lab Group

Allocated space for a group of users to create data labs

- Distributed ownership
- Privacy option
- Multiples based on need

Ownership: DBA

Data Labs

Workspaces for analysis

- For a single user or a number of users
- Automated options for management

Ownership: Business User

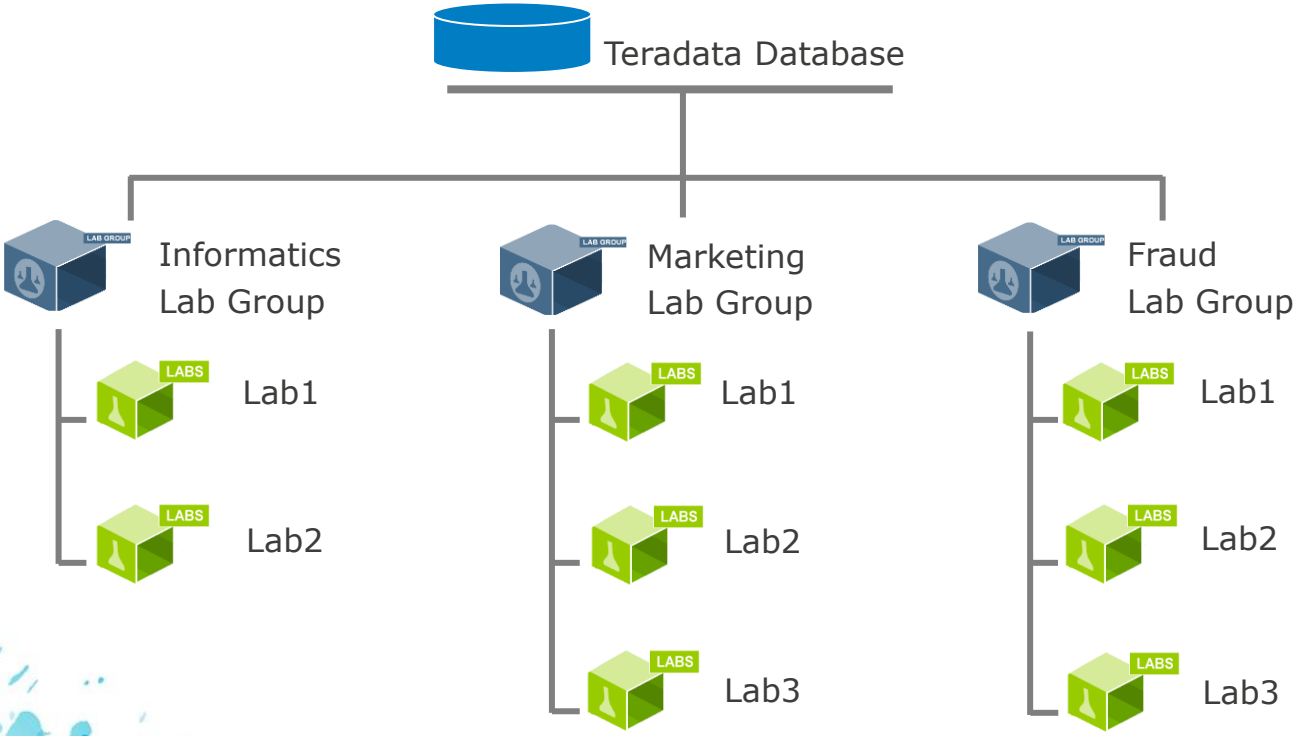
Table

DB table for data storage

- Users can create and load data

Ownership: Development/ Modelers

Sample Lab Group Hierarchy



Data Lab Customer Examples

Reasons For & Benefits of a Data Lab (From–To)

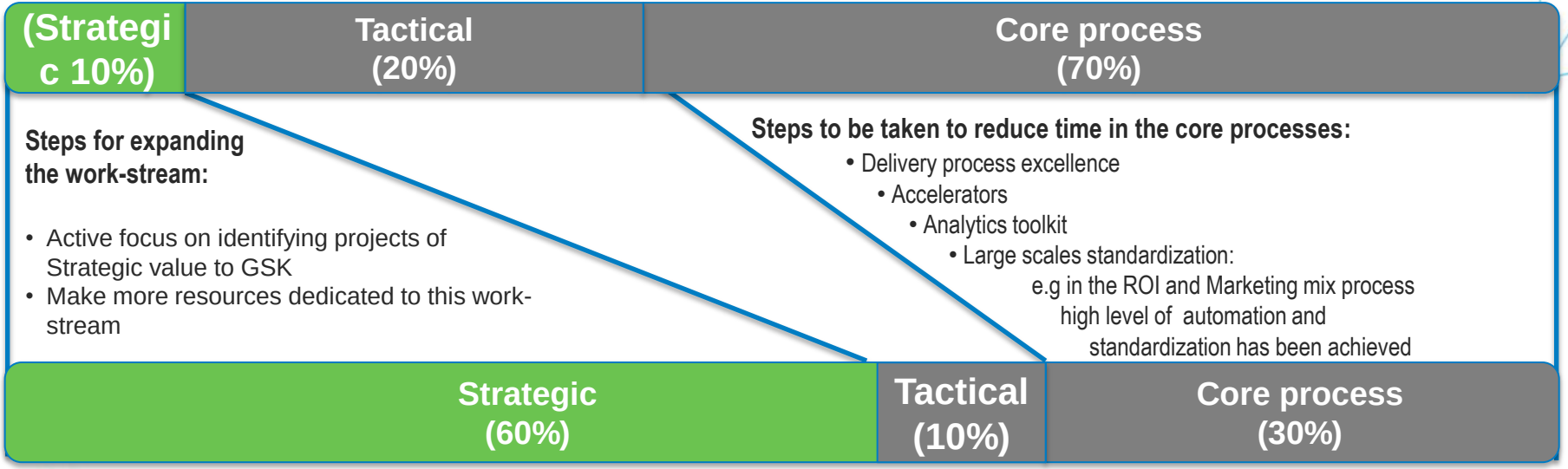
FROM:	TO:
Cannot ….. • create add, modify or delete Data easily! • Existing data models (SPDS PRx, CRx…) did not have data needed. • Static environment leads to long extensive data preparation work	Ability to….add short/long term data aggregations. Add new content to CBI specific data models for collaboration, understanding and reduced prep time & allow quick loading of external data
Suffer from … • computing and space bottlenecks in SAS SPDS & Oracle analytics environment	Ability to….Use Teradata for high volume data exploration, processing, and supported analytical work. Use SAS for specialized or highly-iterative analytical work
Cannot….. • easily use SAS data and other types of Data easily in an agile analytic environment • Need quickly analyze data with next generation analytic tools.	Ability to…. Allow data access with any tool (Tableau, SPSS, SAS, JMP, Qlikview, Spotfire, MicroStrategy, Business Objects)

Data Lab PoC / ROI Metrics Before & After

	Before		After		Gains	
Core Process /ROI Modeling	***Tools***	***Measure***	***Tools***	***Measure***	***Difference***	***Improvement***
Data Aggregation	Base SAS / SPDS	1200 Minutes	SQL / SAS DI / In-DB	2 Minutes	-1198	59900%
Model Execution	Base SAS / SPDS	1800 Minutes	SQL / SAS / In-DB	30 Minutes	-1770	5900%
Model Fit/QC	Base SAS / SPDS	1200 Minutes	SQL / SAS / In-DB	240 Minutes	-960	400%
Manual QC	Excel/SAS	3600 Minutes	Data Lab / SAS / Excel	15 Minutes		
Total Time		***130 Hours***		***5 Hours***	***-125***	***2768%***
FTE's		3		1	-2	200%
Brands		5 (18 Possible combos)		5 (18 Possible combos)		

Reallocate Resources to More Strategic Projects

Efficiencies gained in core process and tactical projects would be funneled into doing more strategic projects



- Reduce the time spent in Core and Tactical projects through delivery process excellence and Accelerators
- Actively invest resources in Strategic projects

Data Labs Runtime Comparisons : Before / After

ETL of Patient History Tables

Table	Volume	Clarity to Oracle	Clarity to Data Lab	Data Lab to Oracle
hx_surgical	129.9G	6 Hours	50 Seconds	1 Hour 14 Minutes
hx_social_alc_use_detail	5.9G	11 Minutes	36 Seconds	~ 5 Minutes
hx_family	197.4G	8 Hours 28 Minutes	1.45 Minutes	~ 2 Hours
hx_social	81.4G	5 Hours 14 Minutes	6.46 Minutes	~ 1 Hours
hx_medical	240.8G	14 Hours 37 Minutes	9.04 Minutes	~ 4 Hours

Clarity to Data Lab: Extract data from Clarity and manage/store data in data lab.

Data Lab Proof of Value – Sample 1

Proc name	SAS Code	TD	SPDS
Proc Freq	proc freq data=td.pyr_prsc_rx_unalgn_unfct_stg; tables mo_id*sls_chnl_cd; run;	**14 s**	**17m 00s**
Proc Means	proc means data=TD.PYR_PRSC_RX_UNALGN_UNFCT_STG ; class cid; var RTLRX_EQU_NRX_CNT RTLRX_EQU_TRX_CNT RTLRX_NRX_CNT; quit;	**30s**	**22m 33s**
Proc Report	proc report data=td.pyr_prsc_rx_unalgn_unfct_stg; column mo_id (rtlrx_nrx_cnt) ; define mo_id/group; define rtlrx_nrx_cnt/analysis; title 'report count of cids in each month'; run;	**15s**	**8m 30s**
Proc Tabulate	proc tabulate data=TD.PYR_PRSC_RX_UNALGN_UNFCT_STG ; class sls_chnl_cd mo_id ; var RTLRX_NRX_CNT; table sls_chnl_cd, MO_ID*RTLRX_NRX_CNT ; title 'By SLS_CHNL_CD and YRMO'; run;	**23s**	**6m 31s**

Data Lab Proof of Value – Sample 2

Proc name	SAS Code	TD	SPDS
Proc Sql	PROC SQL; CREATE TABLE agg_indb_rx AS SELECT t1.prsc_CID, (SUM(t1.RTLRX_EQU_NRX_CNT)) AS SUM_of_RTLRX_EQU_NRX_CNT, (SUM(t1.RTLRX_EQU_TRX_CNT)) AS SUM_of_RTLRX_EQU_TRX_CNT, (SUM(t1.RTLRX_NRX_CNT)) AS SUM_of_RTLRX_NRX_CNT FROM TD.PYR_PRSC_RX_UNALGN_UNFCT_STG t1 GROUP BY t1.prsc_CID; QUIT;	**51sec**	**16m 39s**
Proc Rank	proc rank data=td.pyr_prsc_rx_unalgn_unfct_stg out=yrmo_rank(keep=MO_ID PRSC_CID RTLRX_NRX_CNT RX_RANK) descending ties=high; var RTLRX_NRX_CNT; ranks rx_rank; where MO_ID in (201002) and prsc_cid in (273816); run;	**14s**	**In sufficient memory**
Data step	data mon_rx; set td.pyr_prsc_rx_unalgn_unfct_stg; where cid in (61998); run;	**0.33sec**	**2sec**

Summary: Real Value For Your Business

Rapid Time to Business Value

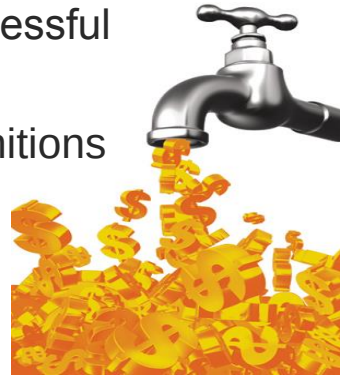
- Self service of individual or collaborative analytical data lab environment
- Exploration within the warehouse results in improved accuracy, consistency, and precision of results

Cost Effective – “Sharing of Resources & Data”

- Faster implementation and less effort than physical servers

Promotes Structured Corporate Analytics

- Proactively controls proliferation of data marts
- Enables rapid analysis for new projects and allows for “promotion” of successful projects
- Helps in development of more precise requirements with clearer ROI definitions



Questions?