TERADATA | THE BEST DECISION POSSIBLE™

Behind the Big Data Hype: New Tools and Approaches available to Researchers

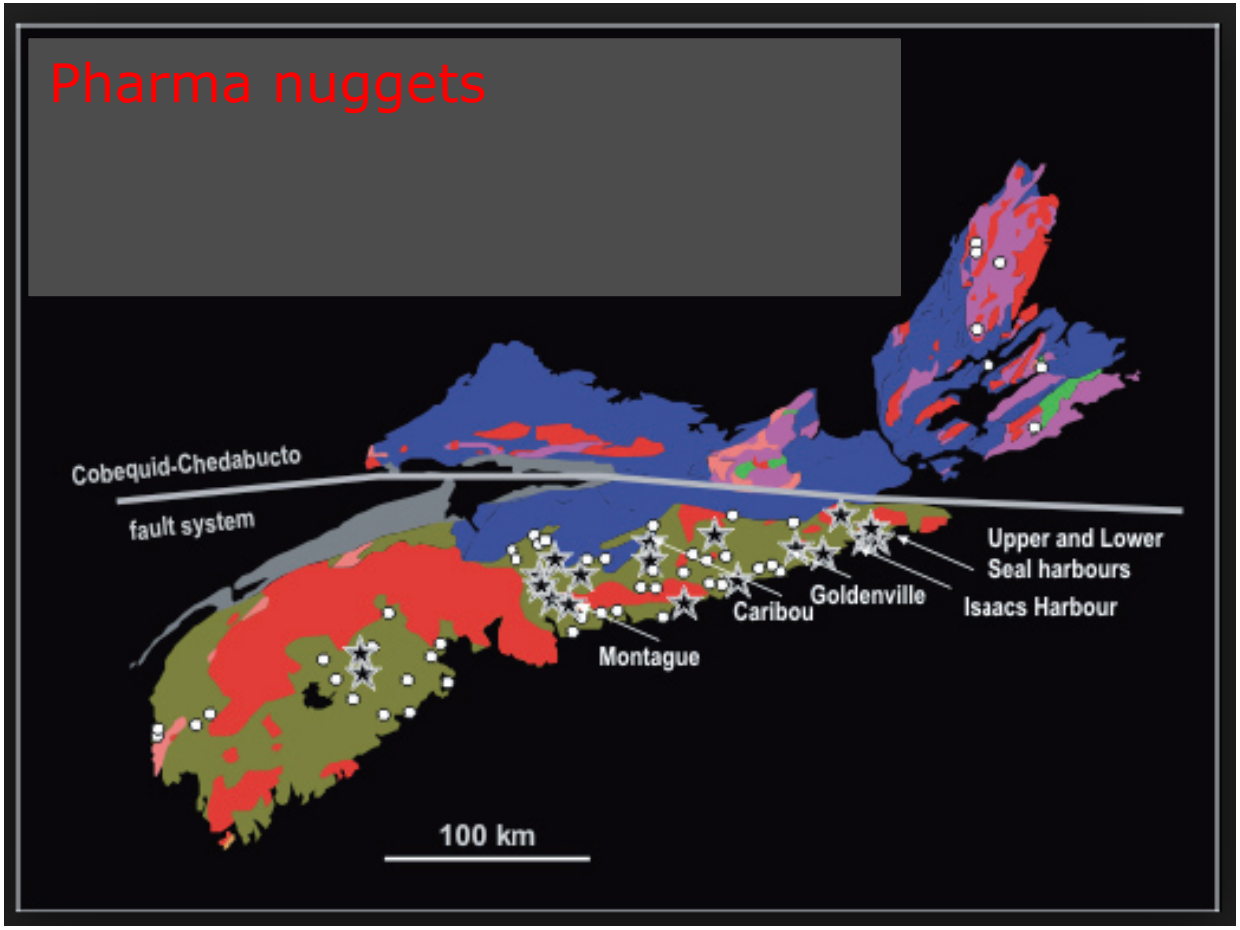
28 June 2013
PhUse conference

Peter Grolimund, Senior Industry consultant Life Sciences

Agenda

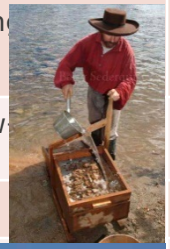
- The mining history
- Assumptions
- Big Data and Pharmaceutical industry
- The process and engines
- Agility
- The soft part
- Integration and Hadoop
- The UDA
- The SAS integration
- The application library
- Some visionary approaches

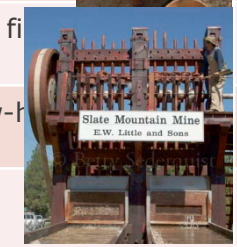
Where are the nuggets?

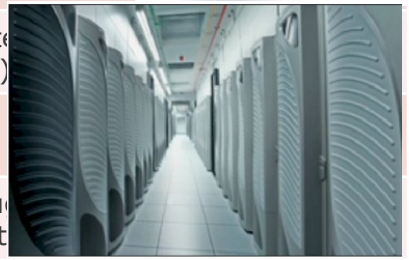


The mining history

		Gold mining	Data «mining»
Generation 1	**Tools:**	Simple tools (showel, pick, pannin	Statiscial functions and mathe Paper and listings
	Skills:	Physical strength, geological know patience	Mathematics, statistics, graph
	Quantity:	Couple of kg up to a few tonnes	A few tables, p-values, or gra
Generation 2	**Tools:**	Drilling machine, Explosives, fi mecahnisation of crunching	Programming language, dig
	Skills:	Engineering, geological know-h & logistics	Programming, computer su archiving
	Quantity:	Several tonnes	Several studies a year, with patients (a few Mbytes)
Generation 3	**Tools:**		Larger computes, big memory, c several statiscial tools and graph software, storage infrastructure.
	Skills:		Several programming languages, data-know-how beyond, busines mathematics, multidisciplinary kn
	Quantity		More and differnt data (studies, E several Terabytes and
Generation 4	**Tools**	Specialized machin waste from resourc	New Software not relate (e.g. Map-Reduce, etc.)
	Skills:	Logistics, flow of re engineering, envirc	Data mining, sciences, team
	Quantity	Waste quantity is h	A lot of 'noise' data, hu Terabyies up to Petabyt







Assumption

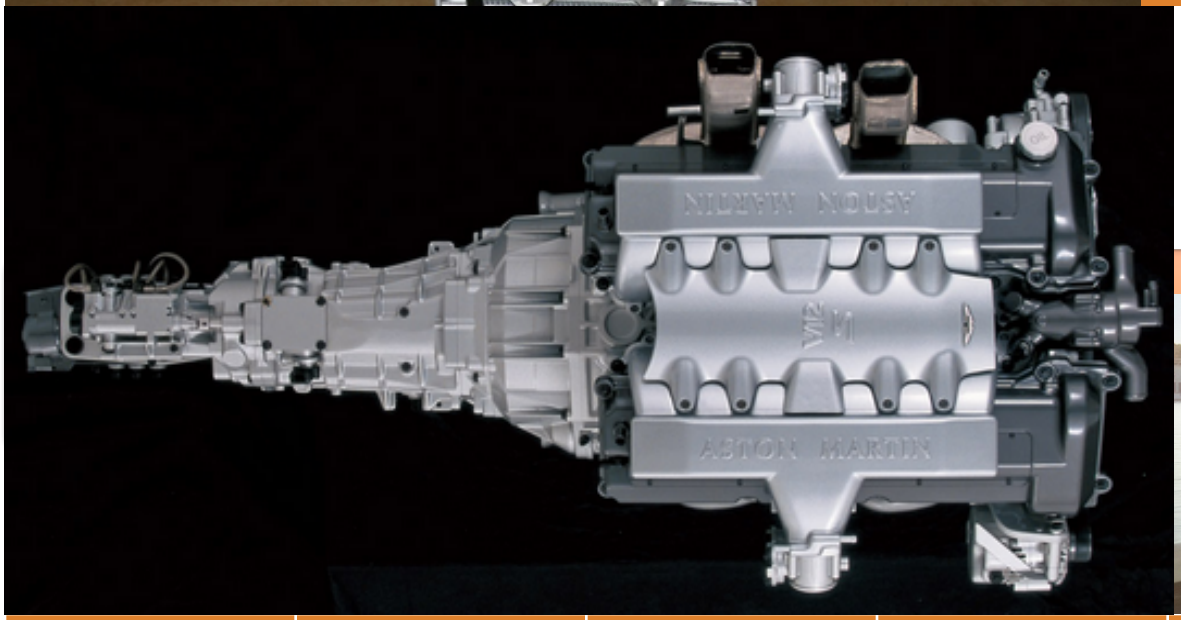
- Big data and new approaches (like non-SQL) etc. will play a role in the future (Remark: We do have evidence)
- The skillset owned by statistical programmers and statisticians is a key asset for the future
- Reproducibility and traceability of analytics will stay (validation) in place
- The workbench extension for analytical tools will continue to grow rapidly
- New data sources will become available or will be created
- Historic data will be used more
- Increased efficiency of what we do today is a must in order to extend into new fields (Remark: less maintenance, higher re-usability)

Why big data in big Pharma?

- There are new needs which are going beyond the traditional analytics
  - Translational sciences
  - Increased Safety
  - More available public data (signals, competition)
  - New in-license activity (assessments, research results)
  - Real world data (claims, health-economics)
  - Social network (patient feedback, off label use indications)
  - Pharmaceutical companies get closer to the end-consumer
  - Conditional approvals and adaptive trial design
  - Repurposing of existing drugs by using secondary analytics and old dr
  - Devices (Apps, Gym, Epidemiologic data)
  - Pesonalised medicine (Genome, Epidemiology, ..)
  - Temporal (changes over time...)

CONCLUSION:

It is not about does it happen; it is about does it happen with or without me

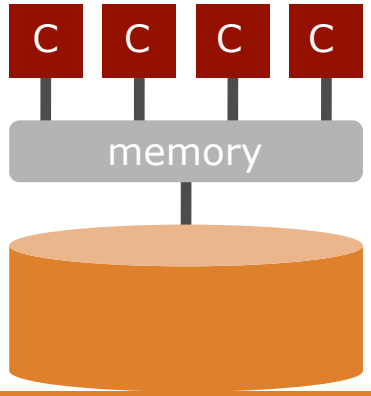


Process	Hardware	Skills	Data	Analysis	Visualization
Agility **Validation** **Culture** **External** Efficiency Reproducability Transparency Diversity Legal …	**Scalability** **Performance** Availability ToC Reliability Investment Maintenance …	**Teamwork** **Data-Science** Mathematics Statistics Data Insight …	**Integration** **Privacy** **Temporal** **Geospatial** Availability Diversity Scalability Transparency Reprocuability	**Diversity** **Reproducability** Dynamic Peformance …	**Dynamic** **Diversity** Attractive Meaning …

The Parallel Computing Spectrum

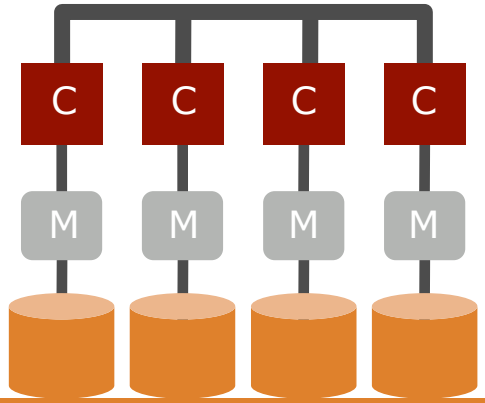
Multiprocessor
Shared disks and memory

SQL/SERVER
DB2
ORACLE



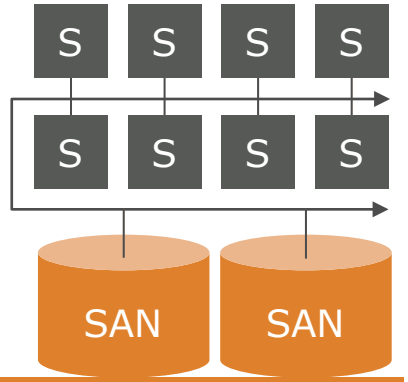
Multicomputer
Shared disk or shared nothing via interconnect

TERADATA
DB2 10
ORACLE RAC



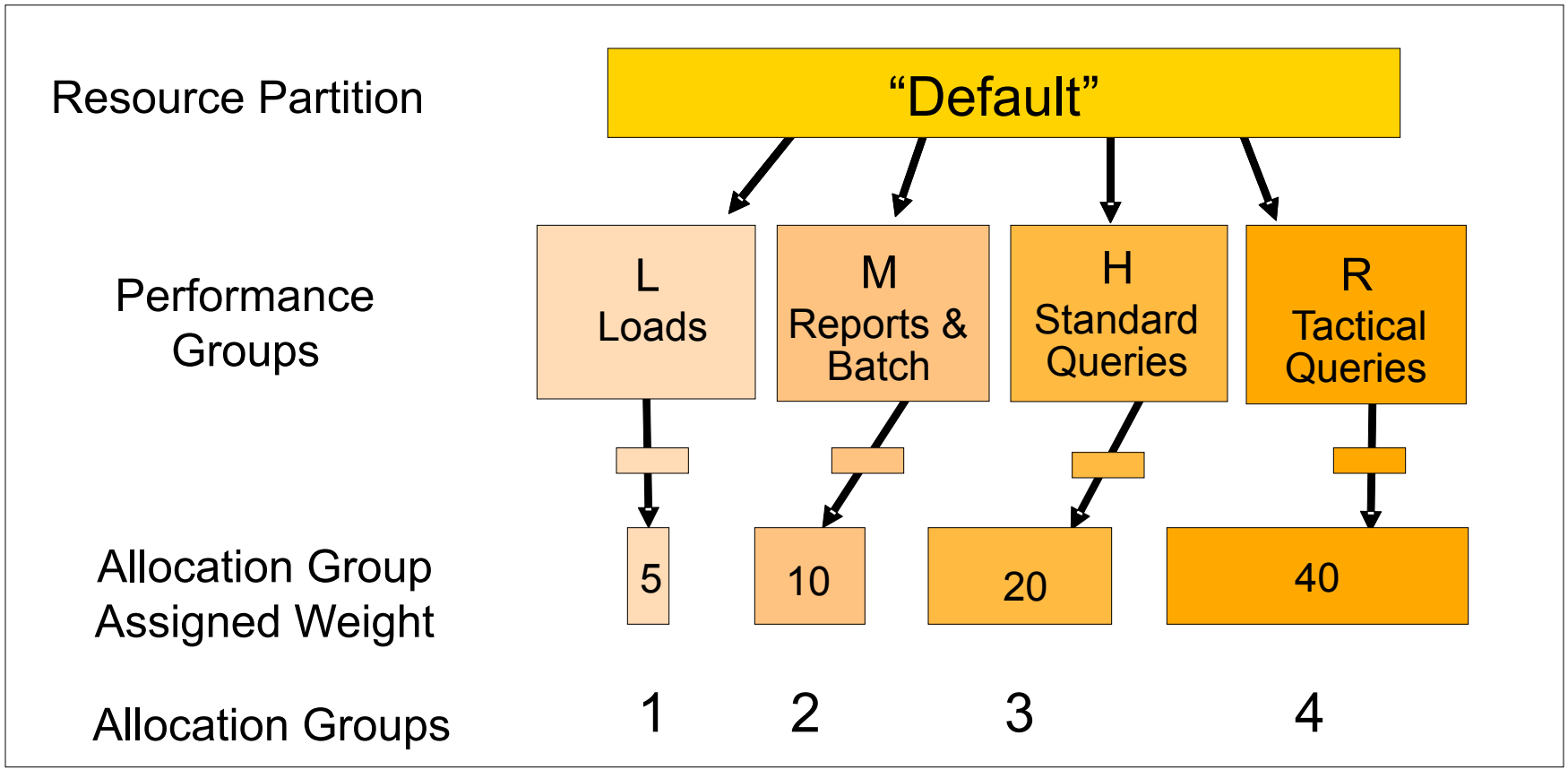
Grid
Federated servers via LAN

SAS Grid
Hadoop
Clouds



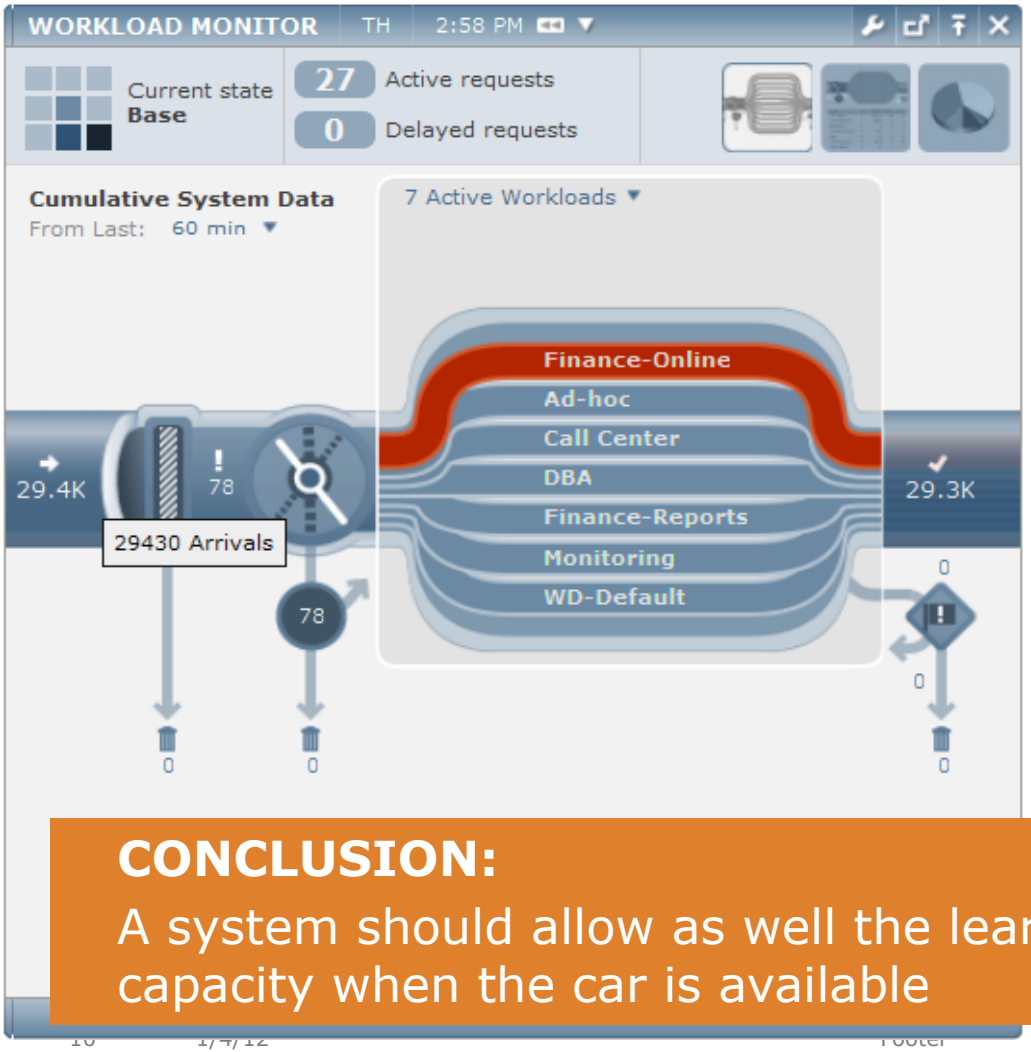
CONCLUSION:
To use all cylinders is the most economic way to drive into big data

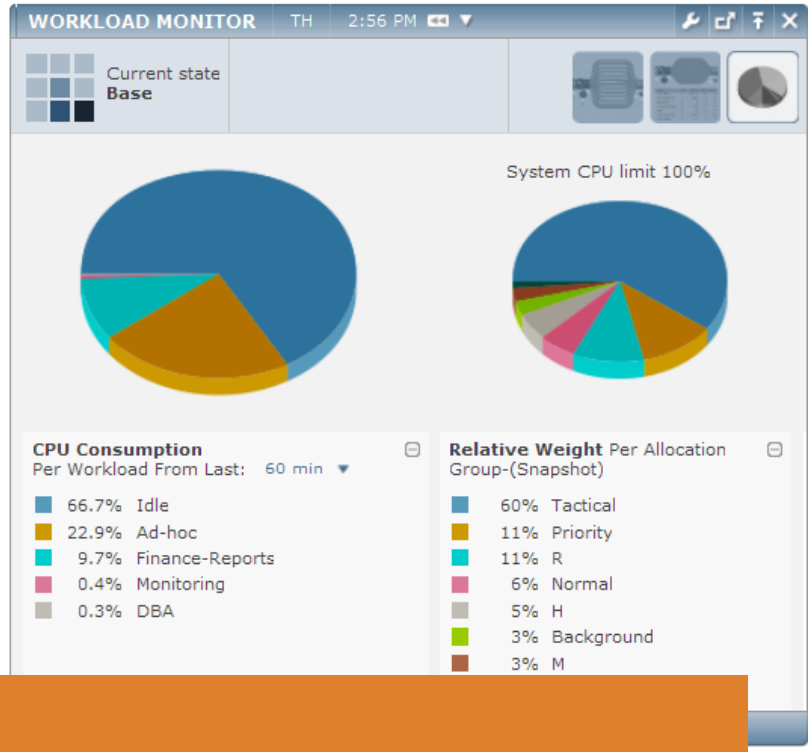
Default setup of Priority Scheduler



Workload Monitor

- TASM workload management of queries in a Teradata system



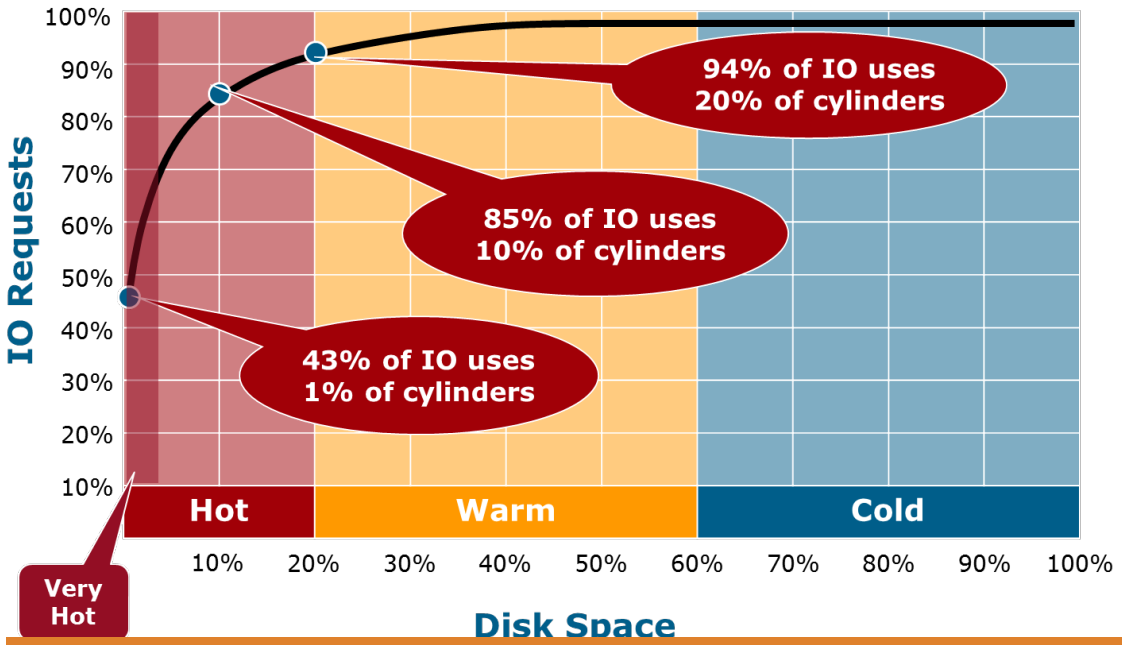


CONCLUSION:

A system should allow as well the learning drivers to use full capacity when the car is available

Teradata Intelligent Memory

Performance of in-memory databases without their cost



- 1% of data satisfies 43% of queries
- Hottest data in memory/not all the data
- Integrated into Teradata system
- No need for

CONCLUSION:

Economic investments is still an element to be considered even memory seems to be available at no cost

When to Use Which?

The best approach by workload and data type

Processing as a Function of Schema Requirements, by Analytic Type

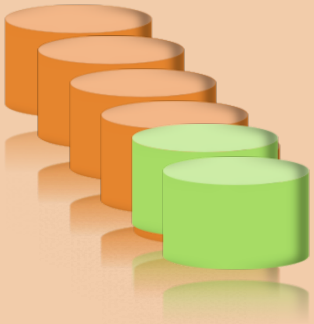
	Low Cost Storage & Fast Loading	Data Pre-Processing, Refining, Cleansing	"Simple math at scale" (Score, filter, sort, avg., count…)	Joins, Unions, Aggregates	Reporting	Analytics (Iterative and data mining)
Stable Schema	Trial performance, Safety reporting, Clinical data, Trial performance, Electronic health records					
Evolving Schema	Devices, Genome, Claims, Web clickstream, Mobile Internet, public data					

CONCLUSION:

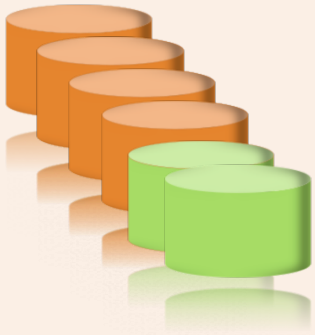
The right choice for the right thing paired with data insights is a corner –stone for big data

The regulation trap or the “hard wired” approach

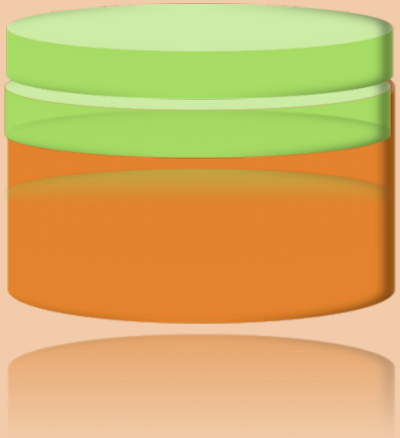
Internal data sources



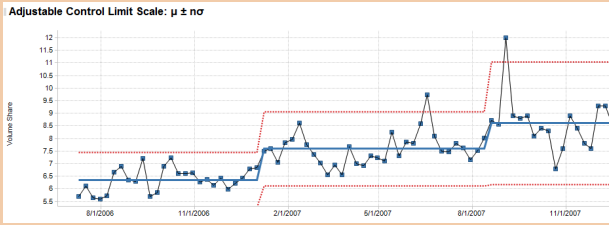
External data sources



Data integration layer
Datawarehouse



The analytical layer



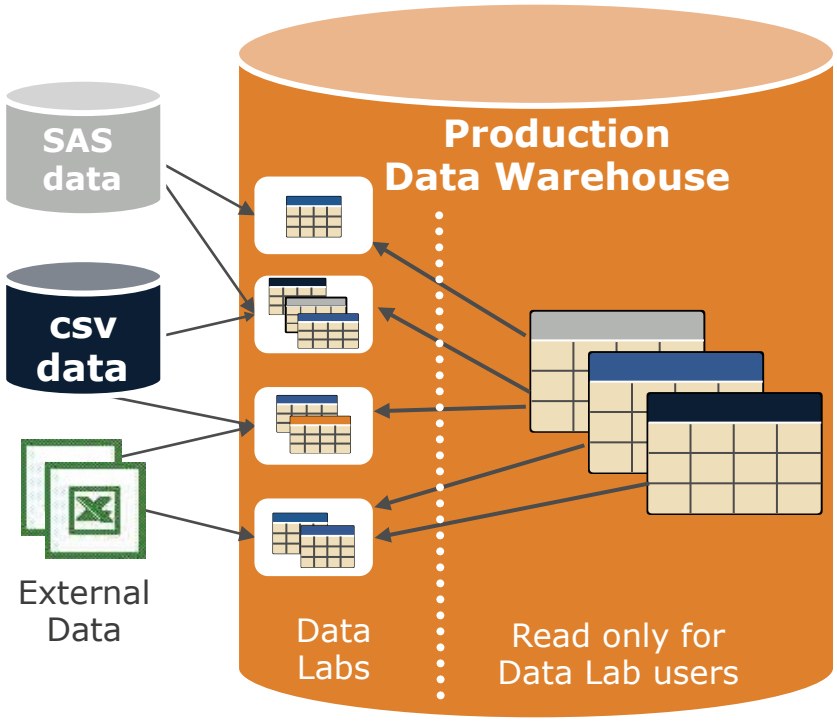
URS FS DS

URS FS DS

URS FS DS

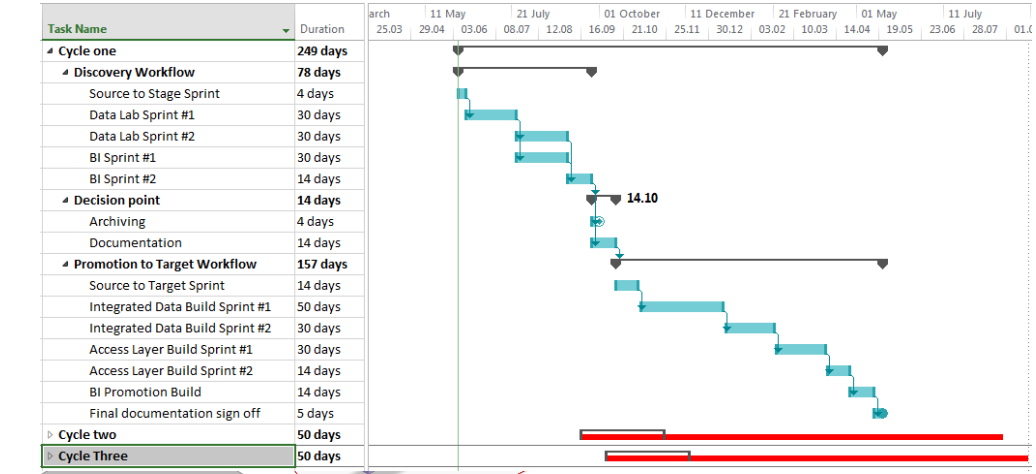
The agile approach, a data lab for the scientist

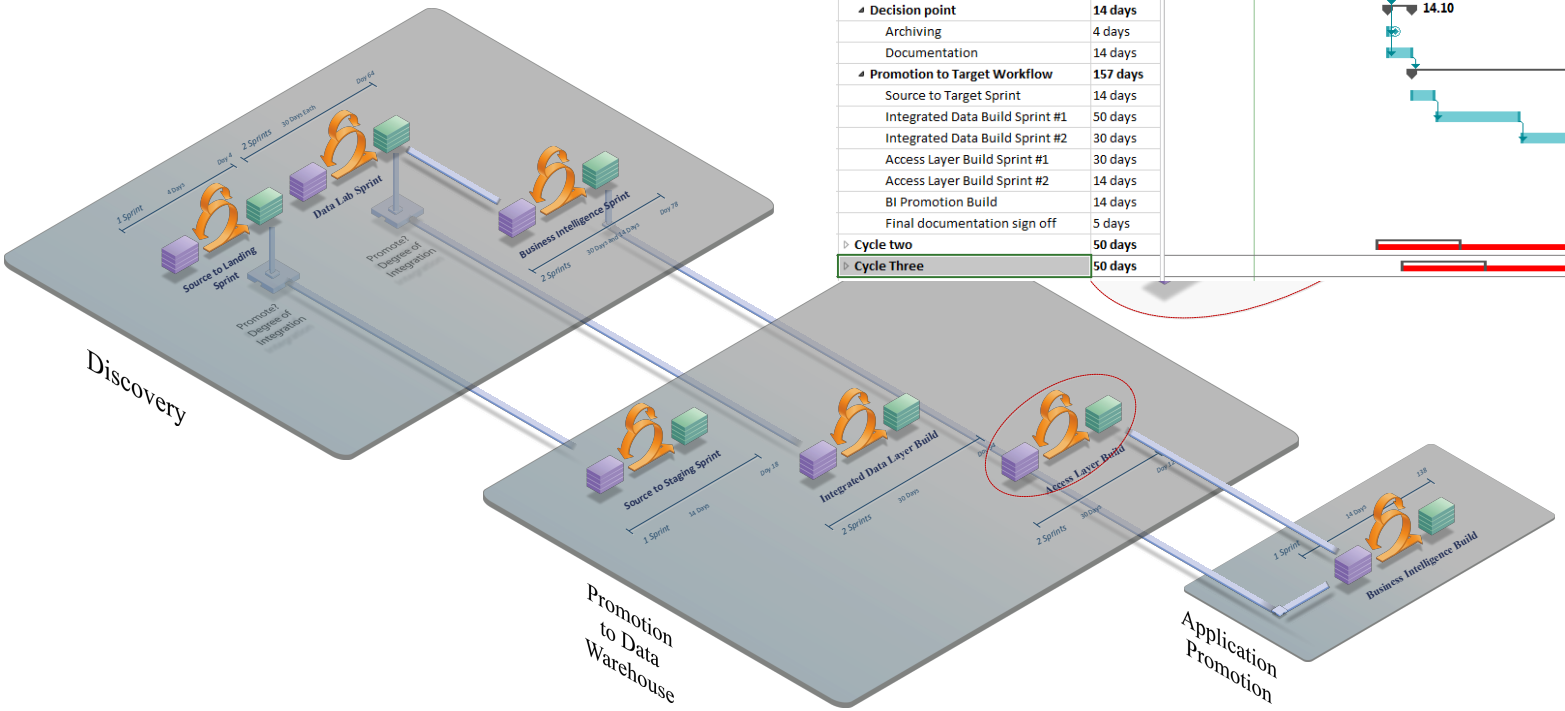
- Data Lab(s) inside system
- Join with production data
- Load experimental data from external sources
- Temporary data
- Rapid prototyping, experimentation, exploratory analysis
- Self-service provisioning and management
- Minimal IT support after initial setup
- Easy to use tool
  - DBA – Governance and management
  - User – Self-service & Collaboration



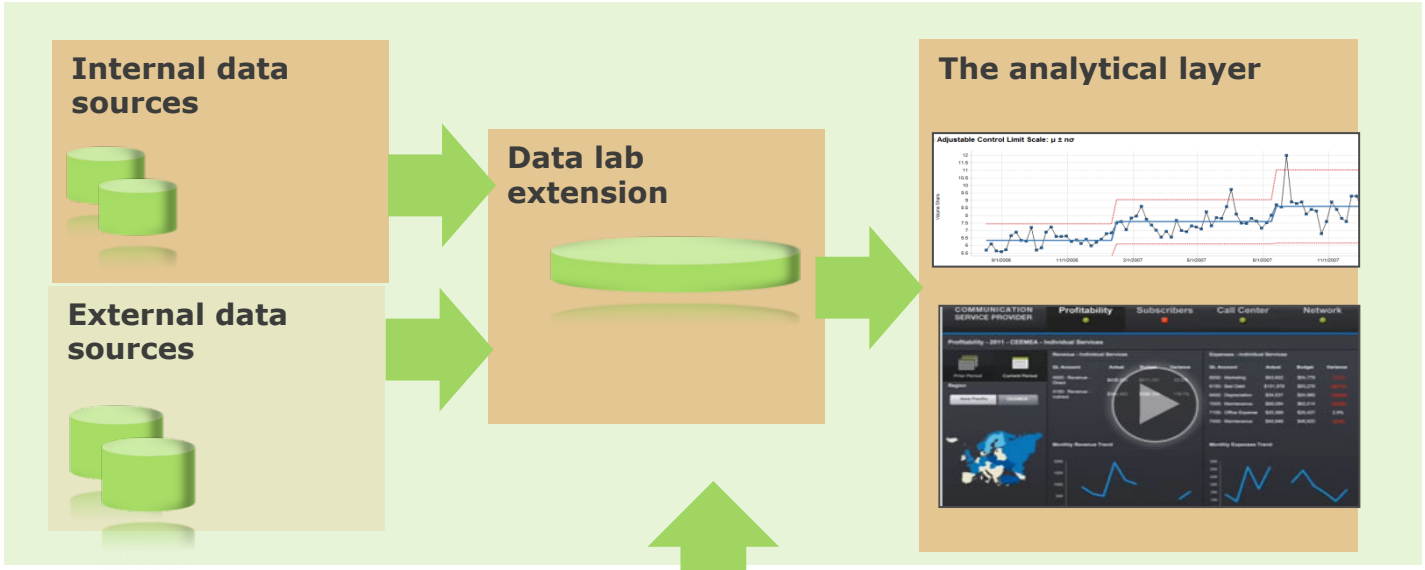
The agile approach, a data lab for the scientist

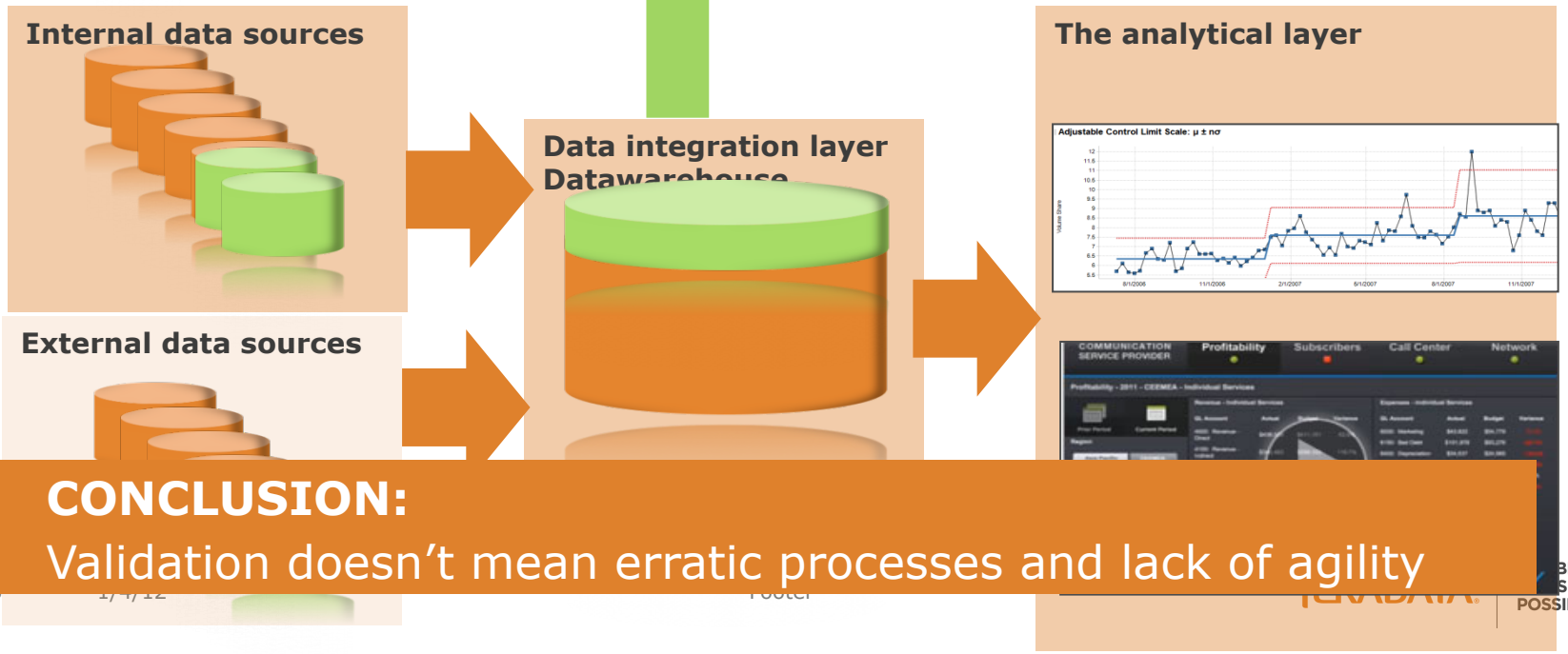
Task Name	Duration
Cycle one	249 days
Discovery Workflow	78 days
Source to Stage Sprint	4 days
Data Lab Sprint #1	30 days
Data Lab Sprint #2	30 days
BI Sprint #1	30 days
BI Sprint #2	14 days
Decision point	14 days
Archiving	4 days
Documentation	14 days
Promotion to Target Workflow	157 days
Source to Target Sprint	14 days
Integrated Data Build Sprint #1	50 days
Integrated Data Build Sprint #2	30 days
Access Layer Build Sprint #1	30 days
Access Layer Build Sprint #2	14 days
BI Promotion Build	14 days
Final documentation sign off	5 days
Cycle two	50 days
Cycle Three	50 days





The agile approach, a data lab for the scientist





CONCLUSION:

Validation doesn't mean erratic processes and lack of agility

Labs can be dangerous

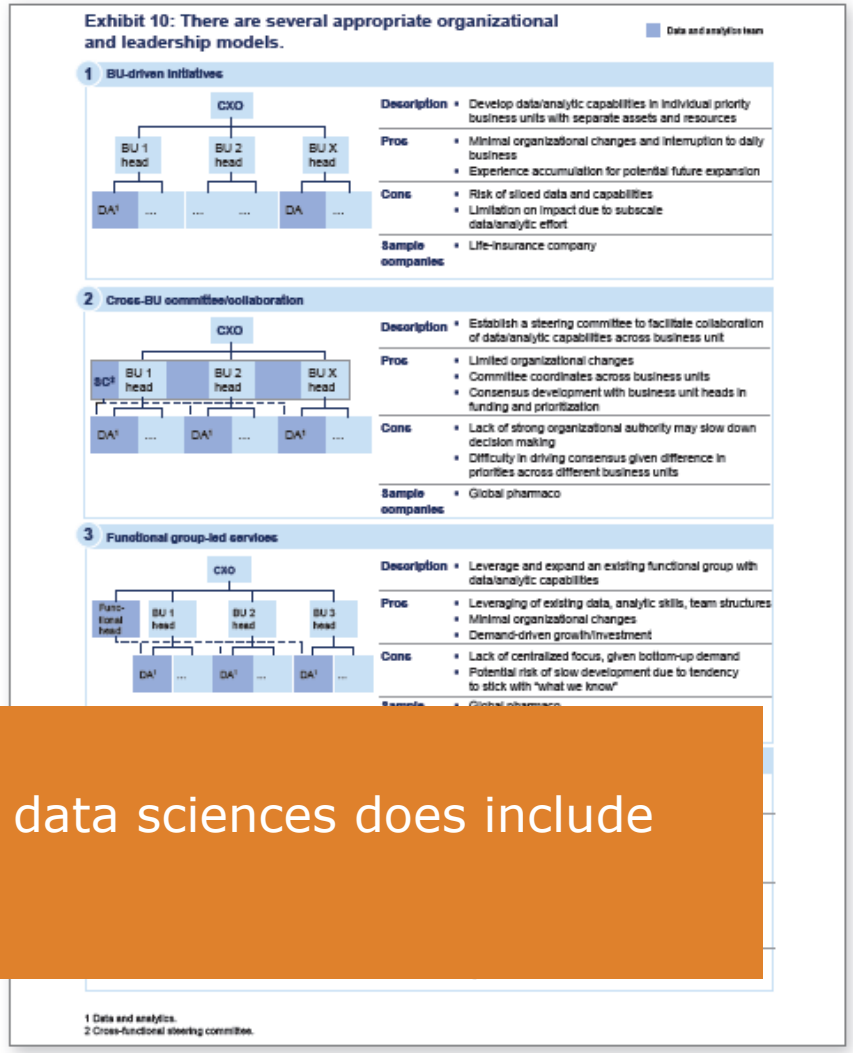
- The explosive parts of the data-lab
  - Findings might
    - speak against objectives
    - Show KPI's do not measure correctly
    - Question existing assumptions/findings
    - Might ask for major process changes
    - ….

but as well big fun

- The fun parts of the data-lab
  - Findings might
    - Releave new opportunities of the company
    - Get more with less effort
    - Improve patients outcom and customer service
    - Get you closer to the patient
    - Drive innovation
    - New repurposed drugs …..
    - ….

The soft part

- Organization set-up
- Senior management support
- Skill development
- Development plan
- The readiness to take
- Data-sharing culture
- Data science as science
- …..



CONCLUSION:

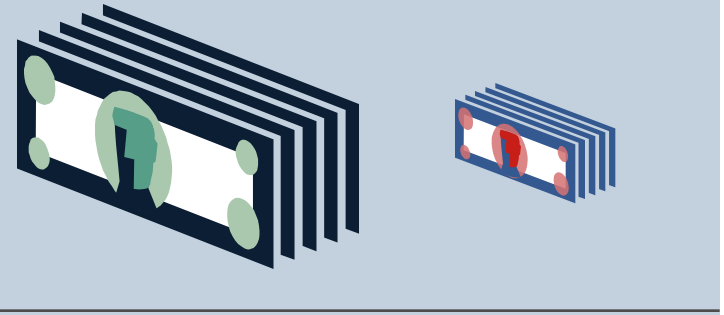
Make it a culture and make sure, data sciences does include the peer review

Options for Enterprise Analytics

Business value Cost to deliver

Integrated, Enterprise Data Warehouse

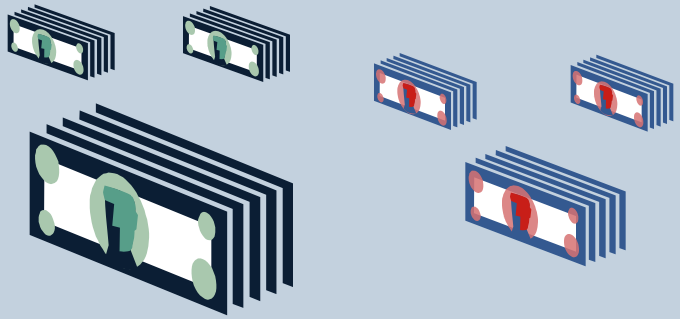
Customer Name, Customer #, Address, List Price
Country, Region, Item #, Transaction #
Revenue, Margin, Vendor, Cost



Some Data Integration

Customer Name, Customer #, Address
Country, Region, Item #
Revenue, Margin

Vendor, Revenue
Address, Margin
Item #

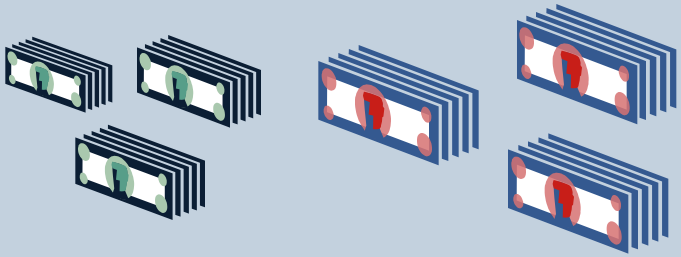


No Data Integration (Data Mart Strategy)

Region, Vendor
Country, Revenue
Customer Name

Country
Revenue, Item #
Customer #, Margin
Customer Name

Vendor, Revenue
Address, Margin
Item #



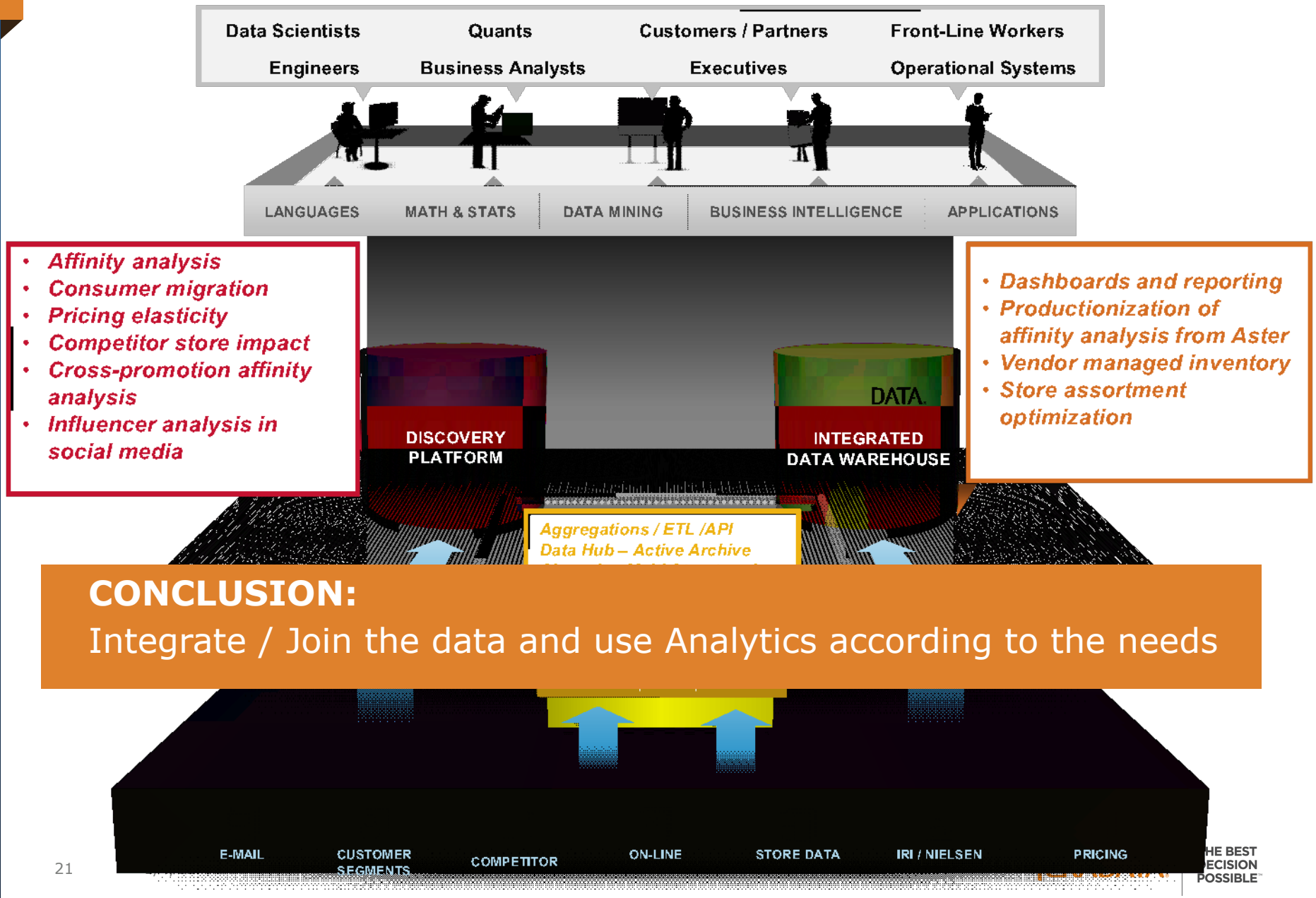
Does Hadoop solve the integration challenge?

- It can help to accelerate the integration (mapping and identify common elements)
- It allows to load without integration and postpone the integration (drop-box)
- It can give new integrations by Map-reduce (aggregated additional information from unstructured data, e.g. social media, web-clicks)

CONCLUSION:

It is not about Hadoop does replace traditional Datawarehouse approaches, the effort and pay back of properly integrated data is proven and the persistence needed to do it doesn't move away because of Hadoop

TERADATA UNIFIED DATA ARCHITECTURE



CONCLUSION:

Integrate / Join the data and use Analytics according to the needs

Hortonworks + Teradata = UDA Delivered

TERADATA.

Teradata Unified Big Analytics Appliance

+hadoop

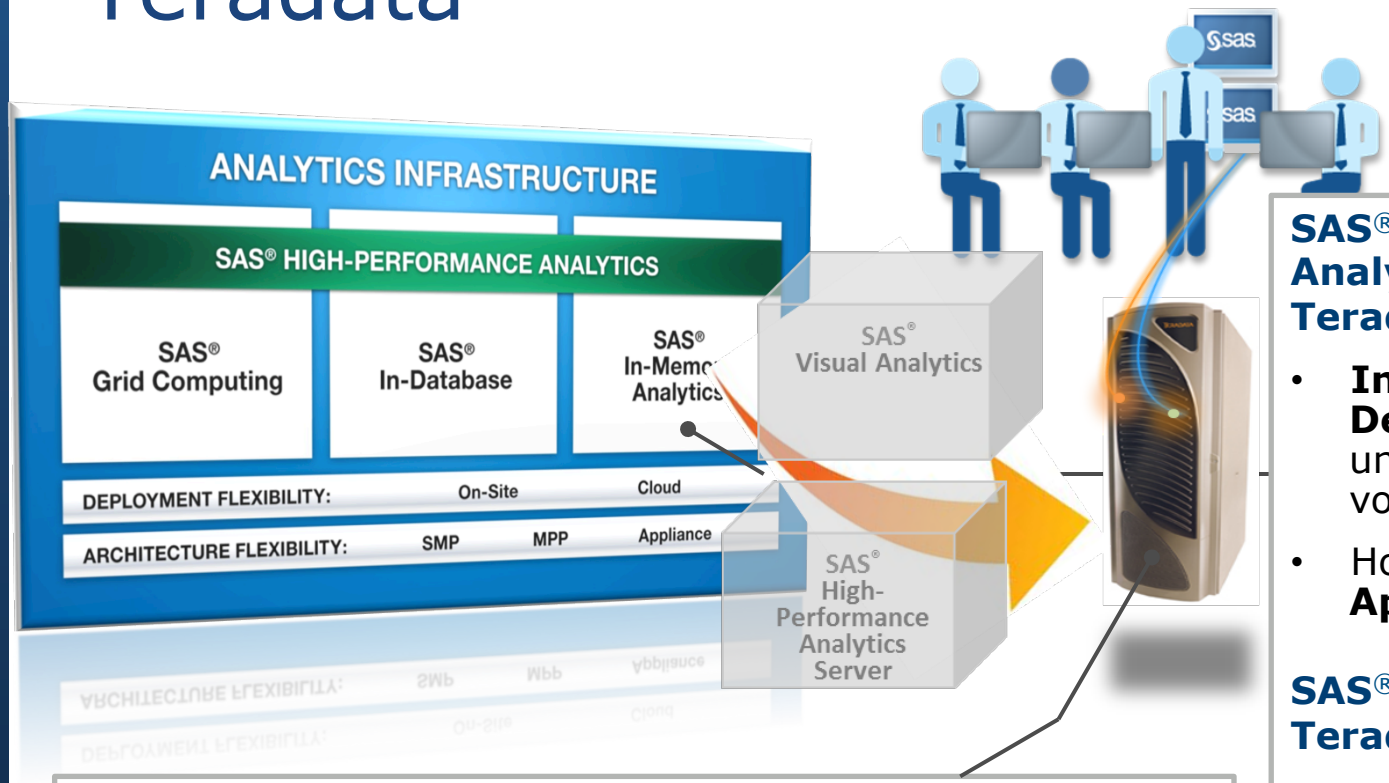
Integration with

- **Teradata**
- **Teradata Aster**
- **Viewpoint**
- **Teradata Vital Infrastructure**

- **Power of Hadoop provided to analysts through familiar Teradata toolset**
- **Manage & Analyze Structured, Semi-structured & Unstructured Data**
- **Hortonworks Data Platform provides the Hadoop capabilities**

- **An engineered, open source solution: only HDP delivers**
  - SQL-H, powered by HCatalog provides SQL-Like interface to Teradata Aster and Teradata Warehouse
  - Deep integration w/Teradata via parallel SQOOP connector
  - Manage/Monitor from a single pane of glass
  - Appliance hardware optimized for big data

SAS® High-Performance Analytics for Teradata



Teradata Appliance for SAS®

- Teradata model 720 built and optimized specifically for SAS® In-Memory Analytics
- MPP infrastructure that executes SAS applications, co-resident with Teradata database

SAS® High-Performance Analytics Server for Teradata

- **In-Memory Model Development** for unsampled and large volumes of data
- Hosted on **Teradata Appliance for SAS®**

SAS® Visual Analytics for Teradata

- **In-Memory Data Exploration** for large and very granular data
- Hosted on **Teradata Appliance for SAS®**

3 YARN

4 HBase

5 HCatalog

6 Serengeti

7 Oozie

8 Ambari

9 Sqoop

10 Chillax7

11 Mahout

Related Projects

Geospatial

Using Teradata to Refine Analysis

Zipcodes
- 92024
- 92009

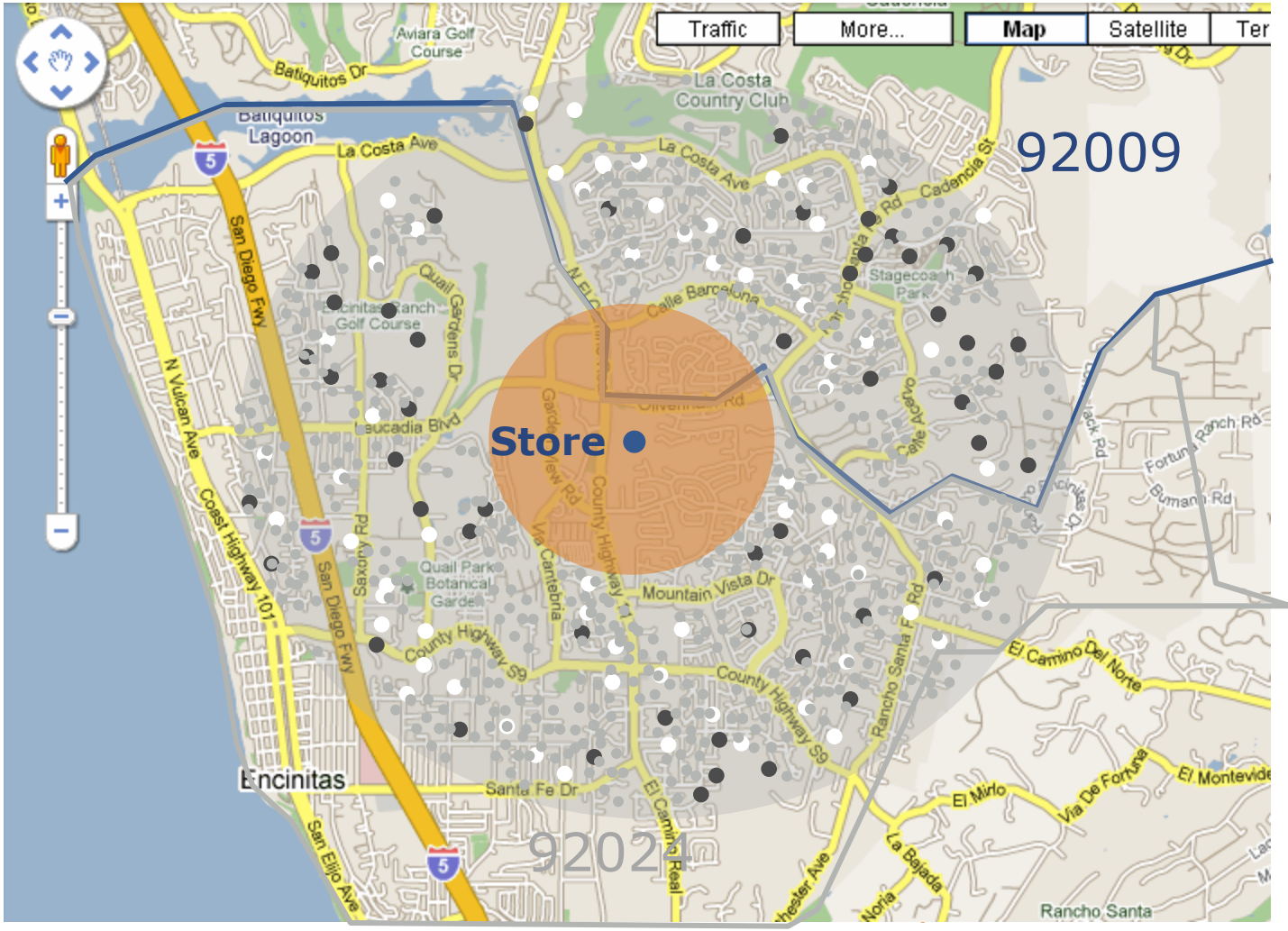
Identify driving radius

Regions
- 5-mile radius
- 15-mile radius
- *Deliver Zone

Display Customers

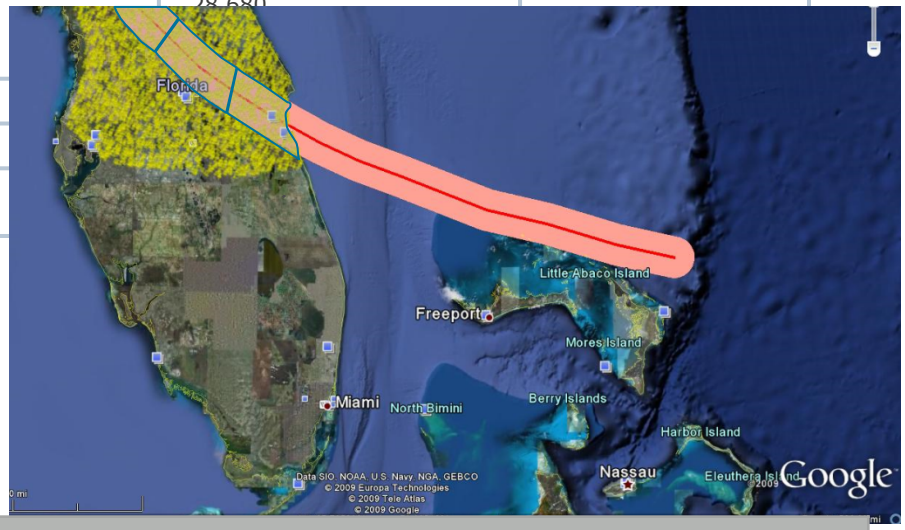
Customer segments
- All customers

Filter customers:
- >= $50 basket Profitability < 75
- >= $50 basket Profitability >= 75



Geospatial

Agents						
Agent ID	Last Name	First Name	Home Office	Contact #	Support region	# of customers
3293	Taylor	John	#253	407-555-2324	Polygon (27.997 -81.103, 28.680	40
2873						
1281						
3889						



Getting Answers with Teradata Database

- Visualization is nice, but you need answers
- With Teradata Geospatial, you can calculate the following based on each path
  - Which policyholders are in the path of the hurricane?
  - How many agents do we deploy and where?
  - How do I contact the policyholders with evacuation sites and agents contacts?
  - What is the hurricane exposure and cost?

Customer List						
Cust ID	Last Name	First Name	Policy #	Contact #	Location	Impact Zone
48737	Riedel	Chris	FA3239-2314	407-555-3243	Point (27.997, -81.103)	Zone 2
29837	Fritz	Lisa	SE2812-2342	312-555-9958	Point (28.680, -82.614)	Zone 1
23421	Monda	Alan	FL9937-020	312-555-27	Point (28.261	Zone 1
39234	Jones					

```
SELECT (C.id, C.name, C.policy, C.Contact)
FROM Insurance_Customers C, Regions R
WHERE R.id = 5
AND C.location.ST_Within(R.geometry) = 1;
```

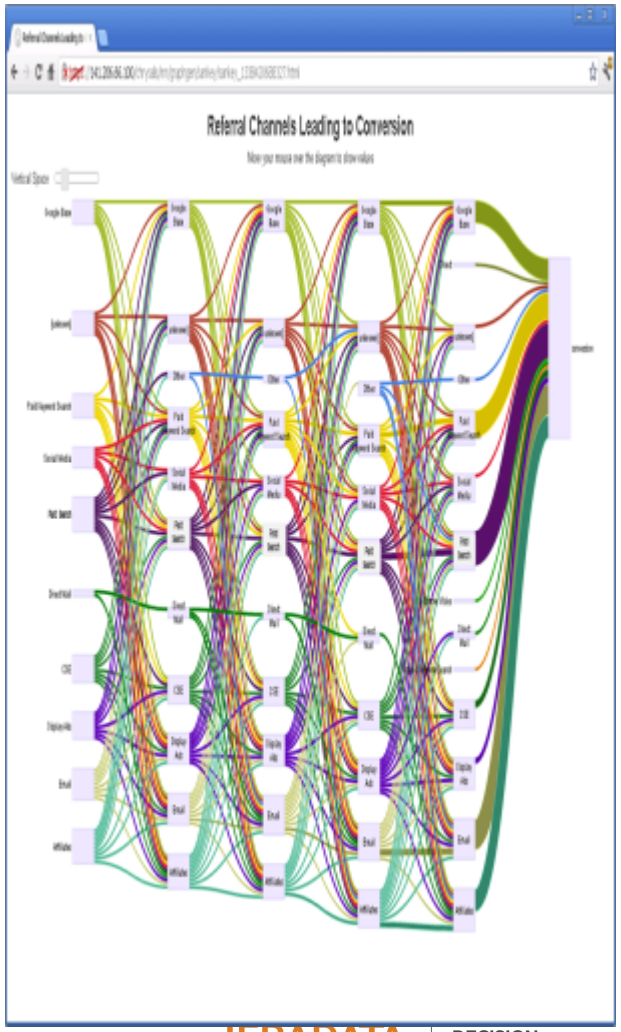
Example - A New Approach

1 SQL Statement to Acquire, Prepare, Analyze & Visualize

Id	Web_log_data
1	75.36.209.106 - - [20/May/2008:15:43:57 -0400] "GET / HTTP/1.1" 200 15251 "http://www.google.com/search?hl=en&q=%22Aster+Data+Systems%22" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; YPC 3.2.0; .NET CLR 1.1.4322; .NET CLR 2.0.50727; MS-RTC LM 8)"
2	98.210.132.218 - - [27/Mar/2011:11:45:47 -0700] "GET /about/management.php HTTP/1.1" 200 10048 "http://www.bing.com/search?q=aster+data&form=QBLH&qs=n&sk=&sc=8-10" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; . NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"
3	llf520029.crawl.yahoo.net - - [29/May/2008:23:15:15 -0400] "GET /resources/images/support HTTP/1.0" 301 187 "" "Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"
4	159.41.1.23 - - [06/Jul/2010:07:19:45 -0400] "GET /public/js/common.js HTTP/1.1" 200 16711 "http://www.wooloo.org/wonjoo" "Mozilla/5.0 (Windows; U; Windows NT 5.1; it; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3"

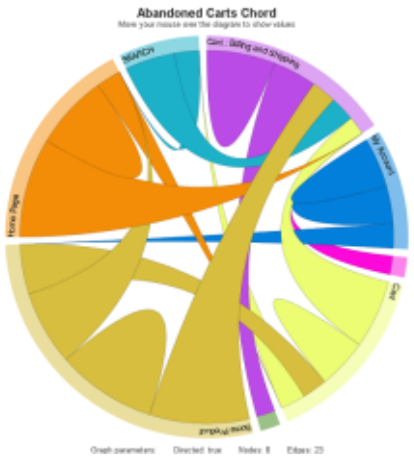
Single SQL statement:

```
SELECT * FROM nPathViz( on
SELECT *
FROM nPath (
  ON (SELECT * FROM
SESSIONIZE ( ON SELECT *
FROM
LOAD_FROM_TD_HADOOP)
  PARTITION BY sba_id
SYMBOLS (
    event LIKE '%REFERRAL%' AS
START_EVENT,
)
  RESULT (...)
) n;
```

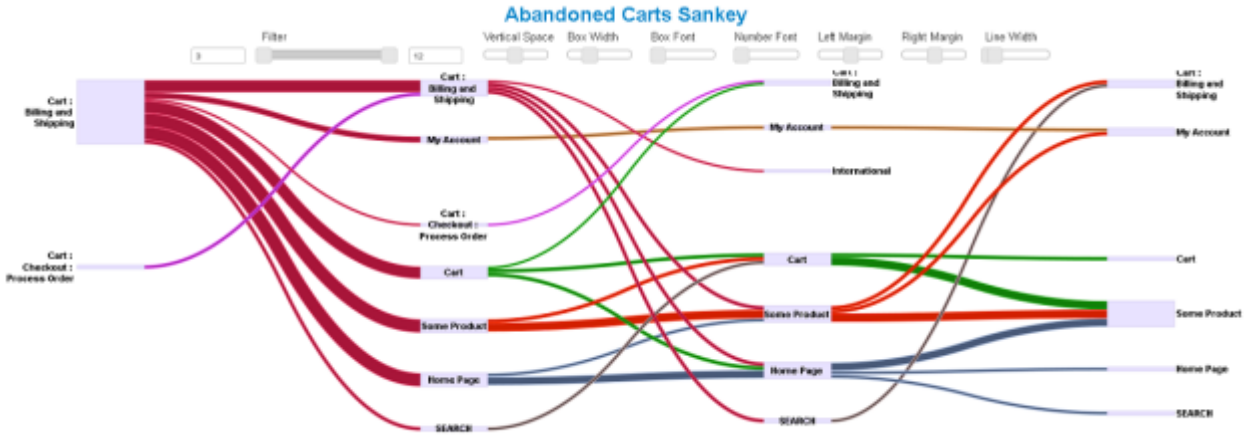


Visual SQL-MapReduce® Functions

Flow, Affinity, Hierarchy Visualizers



Chord

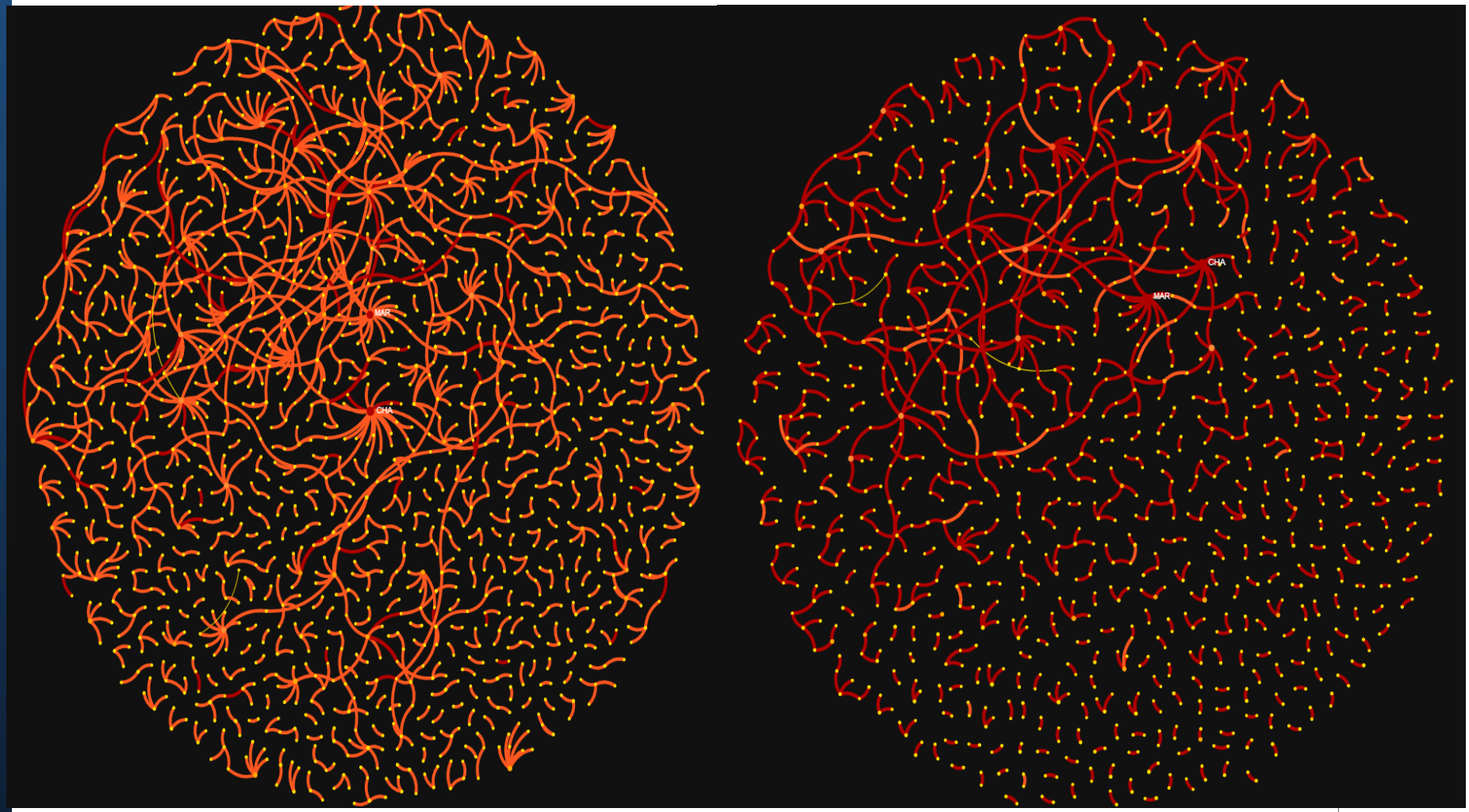


Flow

Hierarchy

Affinity

Synch Issues by Hub Type



Vision: The automated traceability (Lab)

Extend existing approach to any analytics

Programmes (version control and process)

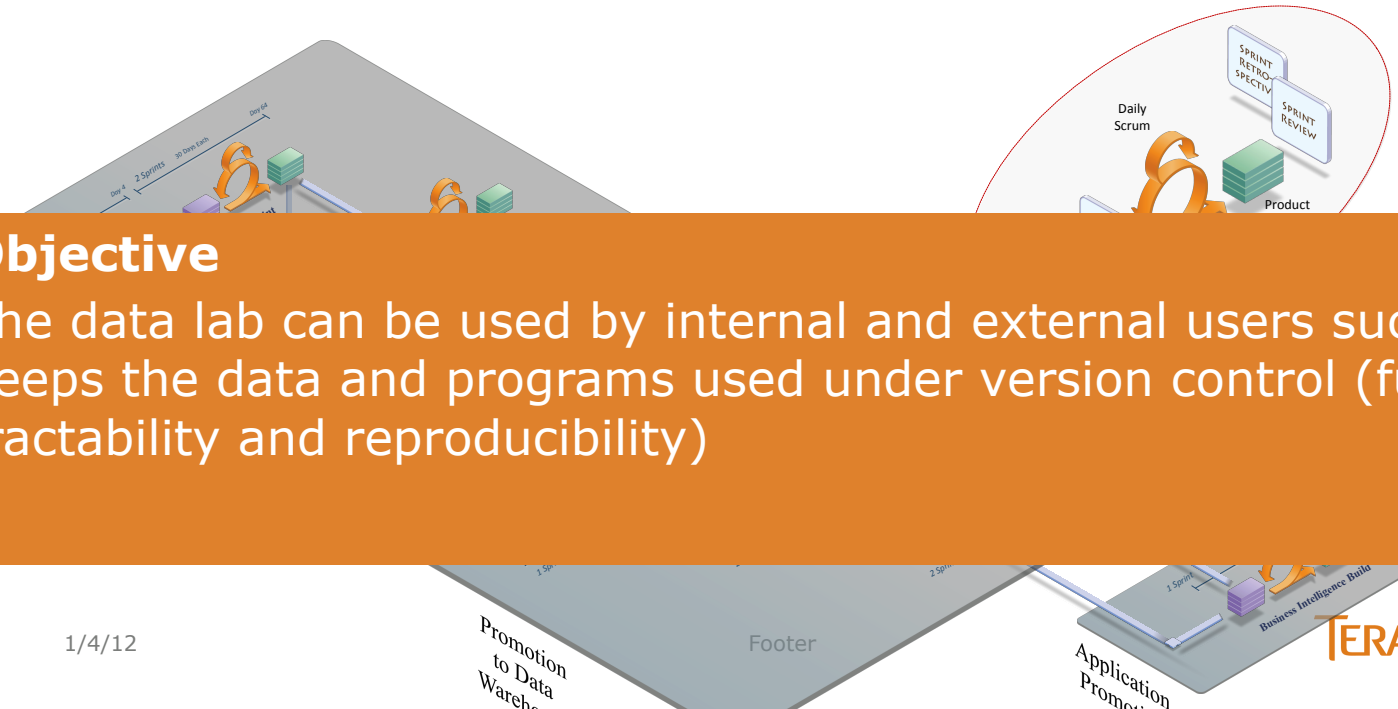
All program versions are kept and available

Third party

Data & query traceability

All queries and data versions are kept and available

TERADATA. THE BEST DECISION POSSIBLE™



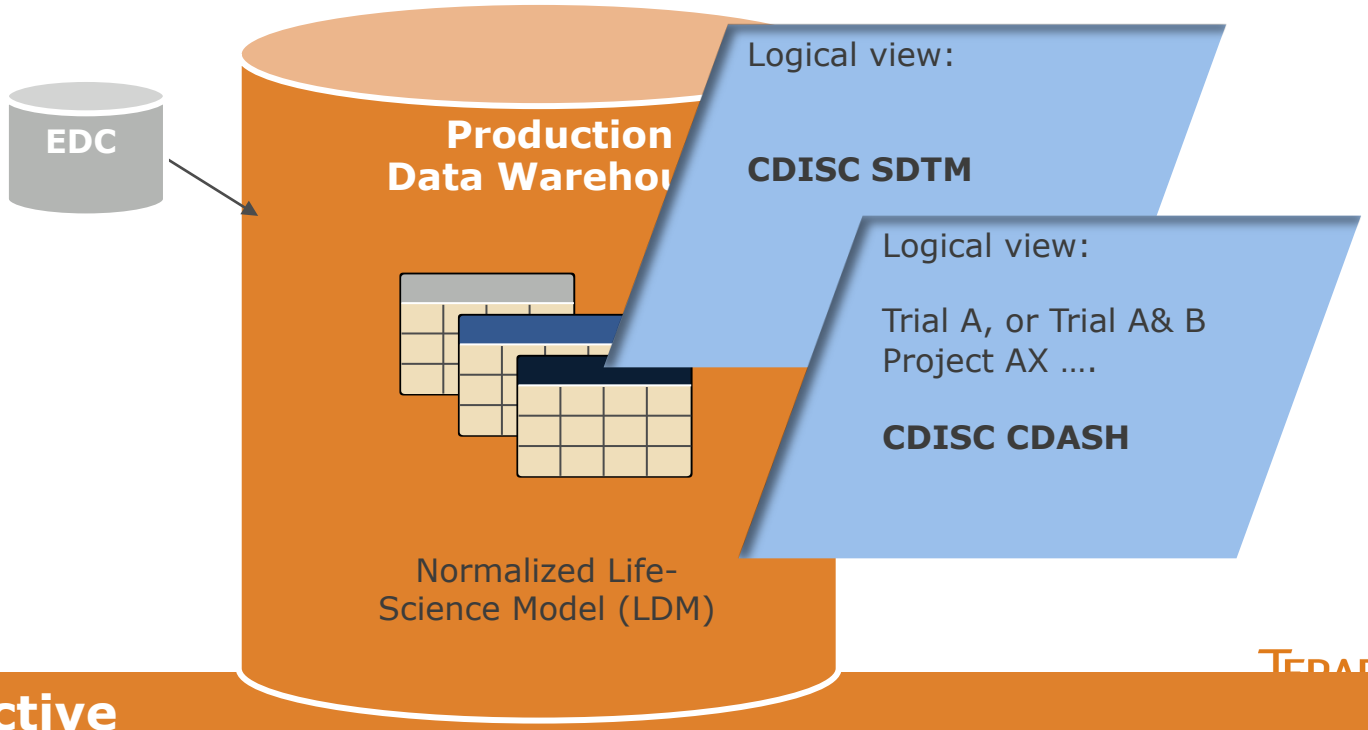
Objective

The data lab can be used by internal and external users such, that it keeps the data and programs used under version control (full tractability and reproducibility)

TERADATA. THE BEST DECISION POSSIBLE™

Vision: The real clinical data warehouse

Pooled by default



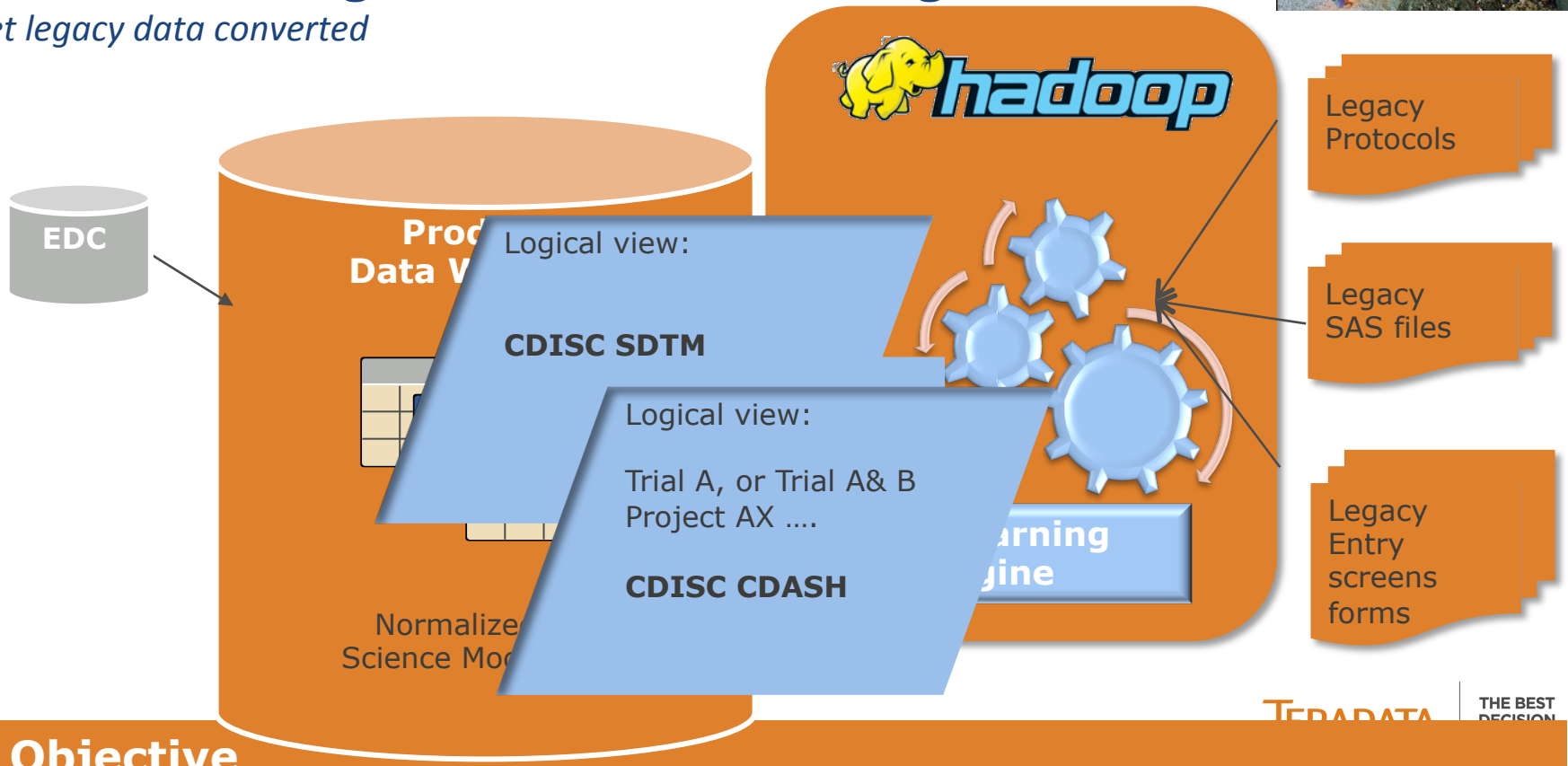
Objective

Keep clinical data once, expose it in the format needed by logical views

Instead of data files, expose logical views and let re-run authorities the scripts in the database

Vision: Leverage the “waste” mining

Get legacy data converted



Objective

Lower the entry cost for data mapping/pooling of legacy data

Use new approaches to create, improve the mappings

Use existing data to validate the result

Use the data lab to do so

Some messages for your bosses

Thank you