Big Data and Your Data Warehouse

Philip Russom

TDWI Research Director for Data Management

May 7, 2013

Sponsor

TERADATA®

THE BEST
DECISION
POSSIBLE™

Speakers

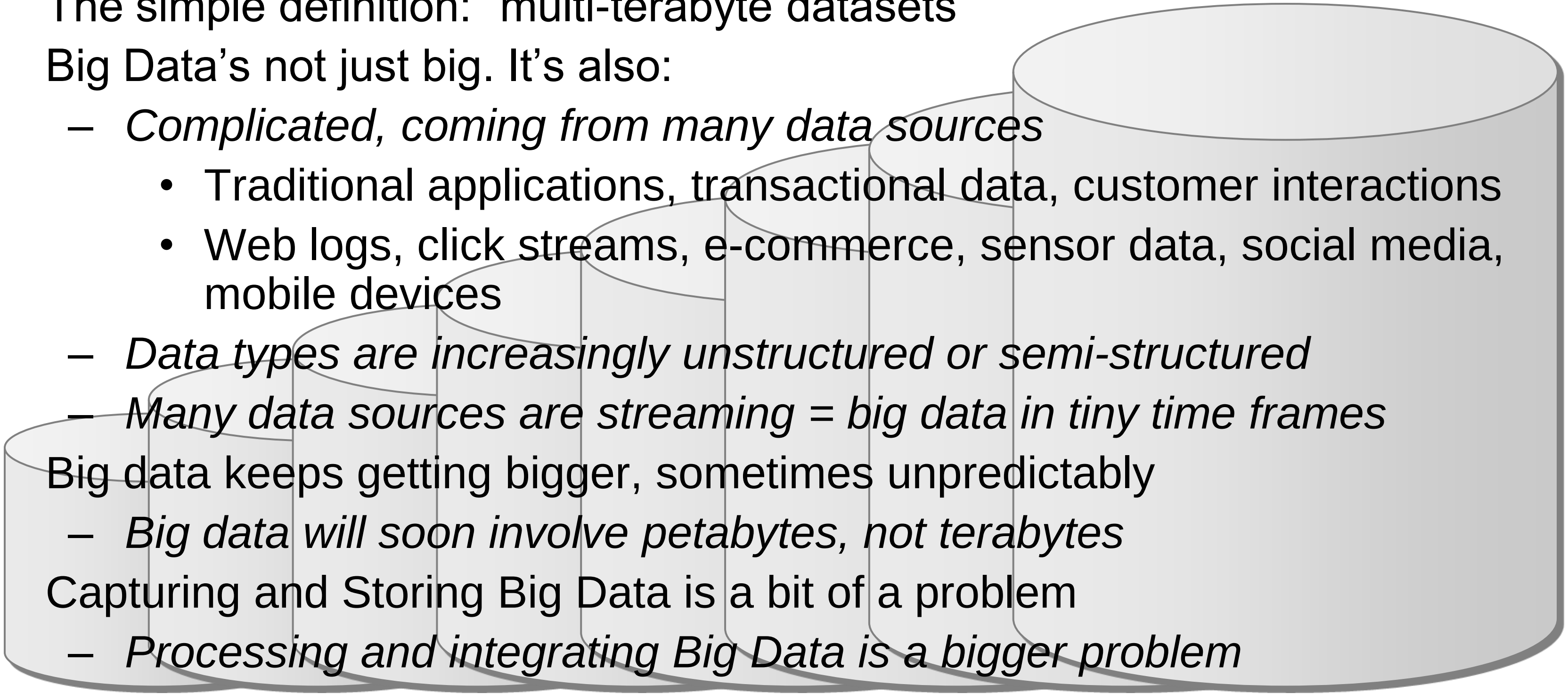
Philip Russom
TDWI Research Director,
Data Management

Chris Twogood
VP, Product and Services
Marketing, Teradata

Today’s Agenda

- Big Data
  - *Was a problem; now it’s an opportunity*
  - *Analysis is primary path to Big Data Value*
- Success with Big Data Analytics may require changes to your Data Warehouse Architecture
  - *DW Architecture, integration architecture*
  - *Trend toward “DW environment”*
  - *Multiple, diverse data platforms for big data’s multiple data types and the multiple forms of analytics associated with big data*
  - *Diversity of DW workloads on the rise*
  - *Logical design vs physical deployment*
  - *Adjustments to data integration, quality, governance, metadata*
  - *Data staging, real time, file-based data, Hadoop*
- Future Trends and Recommendations

Defining Big Data

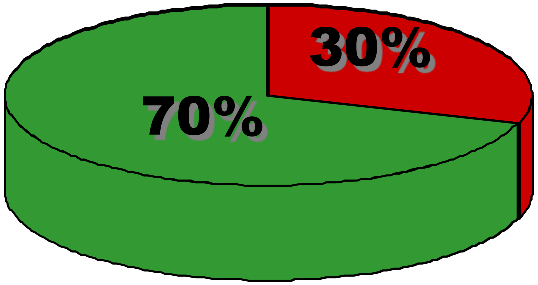


- The simple definition: “multi-terabyte datasets”
- Big Data’s not just big. It’s also:
  - *Complicated, coming from many data sources*
    - Traditional applications, transactional data, customer interactions
    - Web logs, click streams, e-commerce, sensor data, social media, mobile devices
  - *Data types are increasingly unstructured or semi-structured*
  - *Many data sources are streaming = big data in tiny time frames*
- Big data keeps getting bigger, sometimes unpredictably
  - *Big data will soon involve petabytes, not terabytes*
- Capturing and Storing Big Data is a bit of a problem
  - *Processing and integrating Big Data is a bigger problem*
- Big data has its challenges
  - *But it also presents useful advantages you can leverage.*

Big Data as Opportunity

- **Big Data** used to be a scalability crisis.
  - *But today it's not the problem it used to be.*

In your organization is big data considered mostly a problem or mostly an opportunity?



🟩 **Opportunity** – because it yields detailed analytics for business advantage

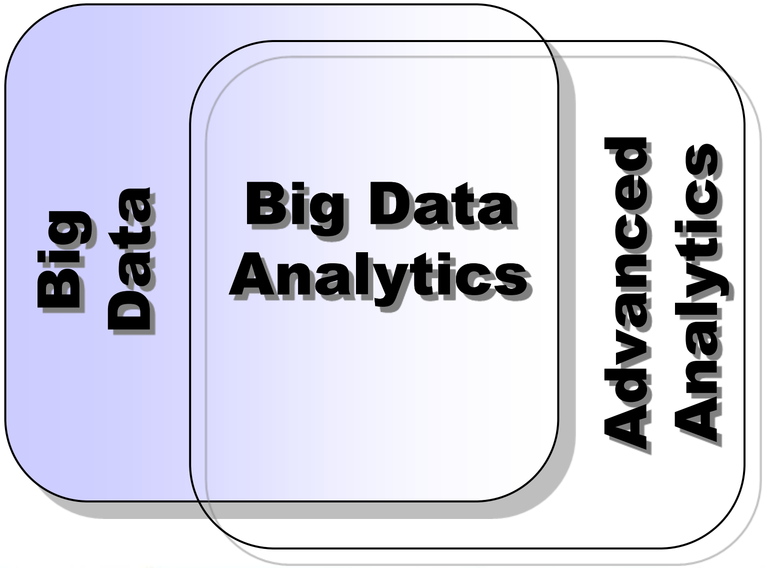
🟥 **Problem** – because it's hard to manage from a technical viewpoint

Source TDWI. Survey of 325 respondents, June 2011

- Oddly enough, the challenge of Big Data today is to get business value out of it.
  - ***Advanced Analytics** yields valuable Business Insights from Big Data*
  - *But there are other paths to Business Value via Big Data, as well.*

Definition of Big Data Analytics

- It's where advanced analytic techniques operate on big data sets
- It's about two things: big data AND advanced analytics
  - *The two have teamed up to leverage big data*
  - *The combo turns big data into an opportunity*
- Big Data isn't new.
  Advanced Analytics isn't new.
  - *Their successful combination is new*
  - *Hundreds of terabytes of data just for analytics is new*



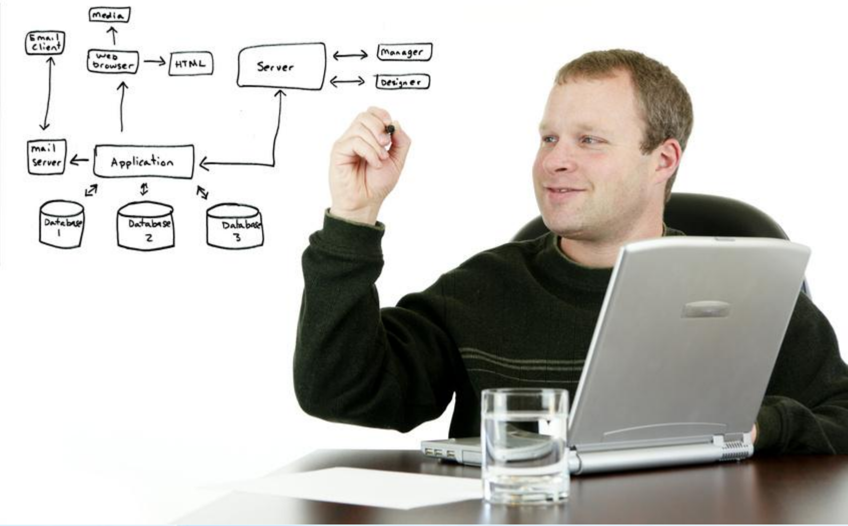
Use Cases for Your DW and Big Data Analytics

- Big Data enables exploratory analytics. Discover new:
  - *Customer base segments*
  - *Customer behaviors and their meaning*
  - *Forms of churn and their root causes*
  - *Relationships among customers and products*
- Analyze big data you've hoarded. Finally understand:
  - *Web site visitor behavior*
  - *Product quality based on robotic data from manufacturing*
  - *Product movement via RFID in retail*
- Use tools that handle human language for visibility into:
  - *Claims process in insurance*
  - *Medical records in healthcare*
  - *Sentiment analysis in customer-oriented industries*
  - *Call center applications in any industry*
- Big data improves data samples for older analytic apps:
  - *Fraud detection*
  - *Risk management*
  - *Actuarial calculations*
  - *Anything involving statistics or data mining*
- Big data adds more granular detail to analytic datasets:
  - *Broaden 360-degree views of customers and other entities, from hundreds of attributes to thousands*

SOUNDS TOO GOOD TO BE TRUE, RIGHT?

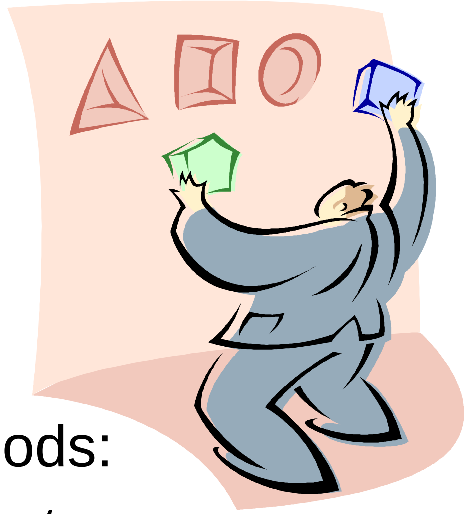
Well, there is a catch.

- To enter the brave new world of Big Data Analytics, you'll probably need to extend your DW environment and redesign your DW architecture for:
  - *Capturing and Managing Big Data*
  - *Processing and Analyzing Big Data*
  - *Discovery Analytics*
    - Extreme SQL, Data Mining, Statistical Analyses, Natural Language Processing, etc.
- Accommodate other workloads:
  - *Streaming (Big) Data and other Real-Time Data*
  - *Un-Semi-Multi-Structured Data*



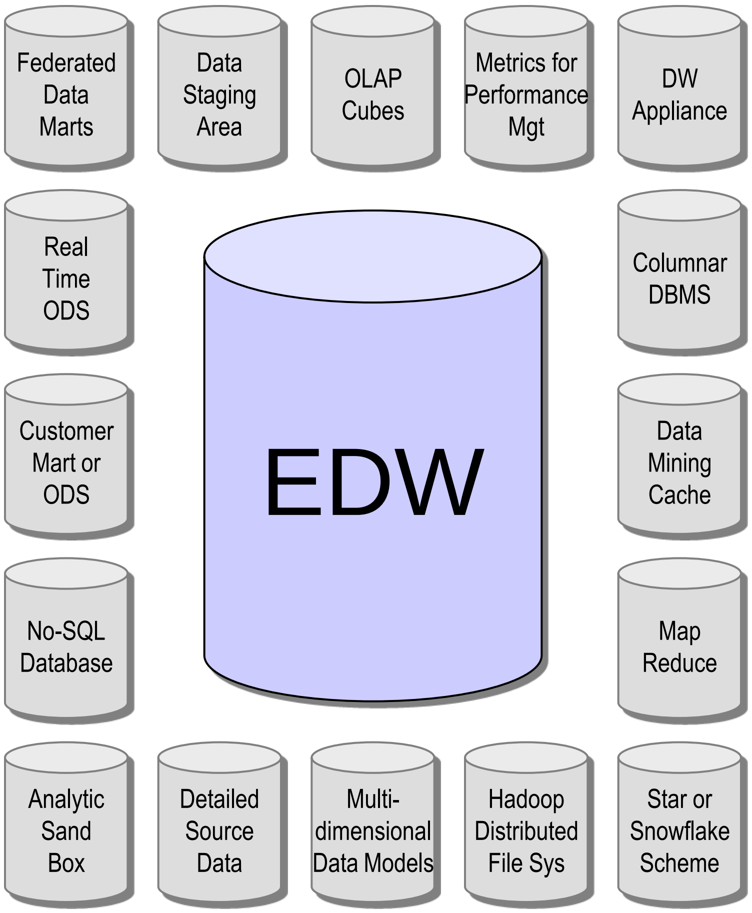
THE CATCH

“Square Peg” Workloads may not fit “Round Hole” DW Architectures



- Most data warehouses were designed and optimized for common deliverables and methods:
  - *Standard reports, dashboards, performance mgt, online analytic processing (OLAP)*
  - *This is a design and architectural decision made by users, not a failing of vendor platforms*
- Can/should all DW & analytic workloads run on your EDW?
  - *If your EDW can handle multiple mixed concurrent workloads with performance and without impeding other workloads, then run all workloads (including analytics) on the EDW, for simplicity’s sake*
  - *If not, you may need additional data platforms for some workloads, including an ADBMS for analytic workloads*

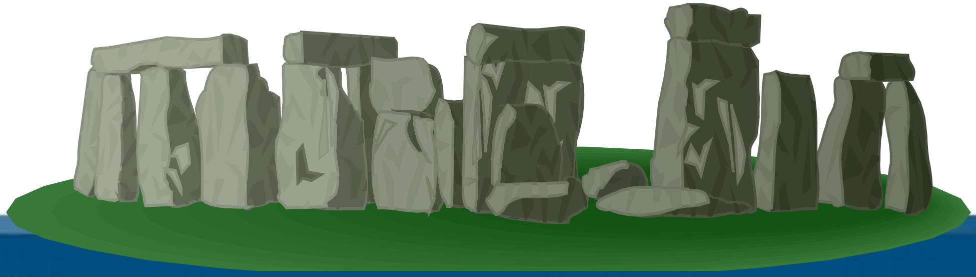
New Big Data Workloads affect Your DW Logical Designs & Physical Deployments



- System on the Side (SOS) or Edge System
  - *A workload and its data that's deployed on a platform separate from the EDW*
  - *Usually integrates with EDW via shared data or data models*
- Long-standing tradition of SOSs w/EDWs
  - *Data marts, operational data stores (ODSs), data staging areas, file systems (for flat files, documents, logs)*
  - *Workload types: analytics, real-time, detailed source data, unstructured data*
- Trend: Workload proliferation driving up SOSs
  - *Big data management vs big data processing*
  - *Each analytic method (or even each analytic application) may need its own SOS*
  - *Streaming, real-time data; multi-structured data*
- Outcome
  - *To provide performance and optimization for workloads users are deploying more standalone data platforms to on edge of distributed DW architecture*

Monolithic EDWs vs Distributed Architectures

- Monolithic DW Architecture – EDW
  - *All or most BI/DW workloads via a single DBMS instance for the EDW*
    - Usually involves mart/ODS consolidation; sometimes a change of DBMS platform for the EDW; "Green field" EDWs may start with a single DBMS
    - Requires a hefty DBMS platform and a great user design to handle so-called "mixed workloads" = multiple, diverse, concurrent DW workloads
- Distributed DW Architecture – EDW*E*
  - *Users deploy separate DBMS instances and standalone data platforms outside and alongside the EDW for nonstandard workloads*
  - *Warning: If not controlled, data marts, ODSs, analytic databases may proliferate. Complexity increases, which deters standards, tuning, sys mgt, etc.*
- Hybrid DW Architecture
  - *Monolith managing core reporting/OLAP data, plus most workloads*
  - *Only a few workloads are deployed on separate systems*
    - Offload invasive or unpredictable analytic workloads, like extreme SQL



Real-Time Data Affects DI & DW Architecture

"Time is money."

- Most real-time DW functionality is provided by Data Integration (DI) techniques
  - *Many DI techniques are conducive to real-time*
    - Micro batches, federation, change data capture (CDC), complex event processing (CEP), RT data quality, interoperability with message buses, Web services, service-oriented arch. (SOA)
- Real-time data demands adjustments to DW Architecture
  - *Real-time data in*
    - Data landing and/or staging area specifically for real-time data
    - Whatever the selected real-time data integration techniques demand
  - *Real-time data out*
    - Very fast queries
    - DW DBMS may need native support for events, alerts, triggers, services, message buses
    - In-memory database (cached in DW's server memory)

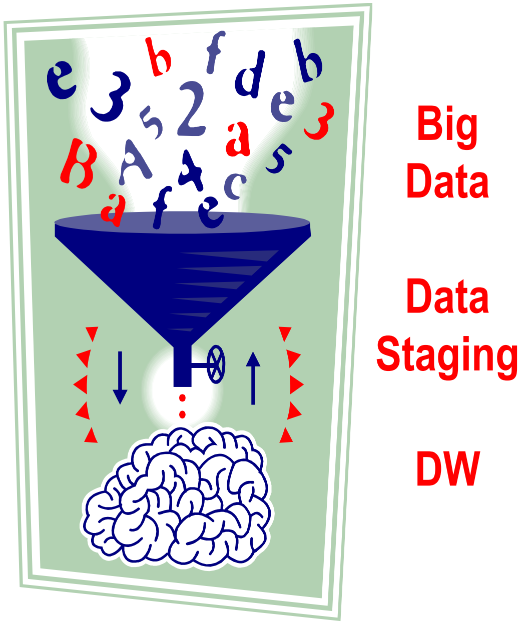
Analytic Tools for Big Data affect Your DW

- There's a cross-road where you choose an analytic method – or multiple methods!
  1. *Online Analytic Processing (OLAP)*
  2. *Extreme SQL*
  3. *Statistical Analysis*
  4. *Data Mining*
  5. *Other: Natural Language Processing (NLP), Artificial Intelligence (AI)*
- Each analytic method has requirements for data and analytic tool types.
  - *Multiple analytic methods can lead to multiple data stores, DBMSs, DW arch. components – and multiple analytic tools*

Big Data Analytics affects Data Management for Your DW

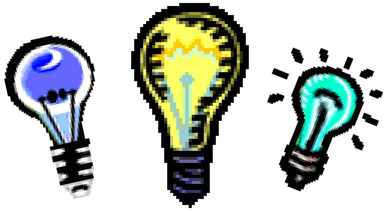
- Analyze data first
  - *Later, improve it for a more polished analysis*
- Analytic discovery depends on data nuggets
  - *Both query-based and predictive analytics need:*
    - Big data, raw data
- Data quality for analytic databases
  - *Do discovery work before addressing data anomalies and standardization*
    - E.g., fraud is often revealed via non-standard or outlier data
- Data modeling for analytic databases
  - *Modeling data can speed up queries and enable multidimensional views*
    - But it loses details & limits queries
    - Do only what's required, like flattening and binning
- Data for post-analysis use in BI
  - *Apply best practices of DI, DQ, modeling*

Big Data even affects your DW's Data Staging Area



- Data staging areas evolved to do more than stage data
  - *Now they must evolve again to accommodate big data*
- Originally data staging areas were temporary holding bins
  - *In that spirit, some are good for "analytic sandboxes"*
- Most data staging areas are optimized for detailed source data
  - *Can manage detailed source data as found in transient big data*
- Data is regularly processed while managed in the staging area
  - *E.g., sort prior to a DW load. SQL temp tables held in staging for later merging*
  - *Some analytic workloads (especially columnar) may run well in a staging area*
- Data staging must scale to big data's volume, which comes and goes
  - *A cloud could be an elastic platform for unpredictable data staging volumes*

Hadoop is a Useful Addition to DW Arch, Because it enables new, compelling apps.

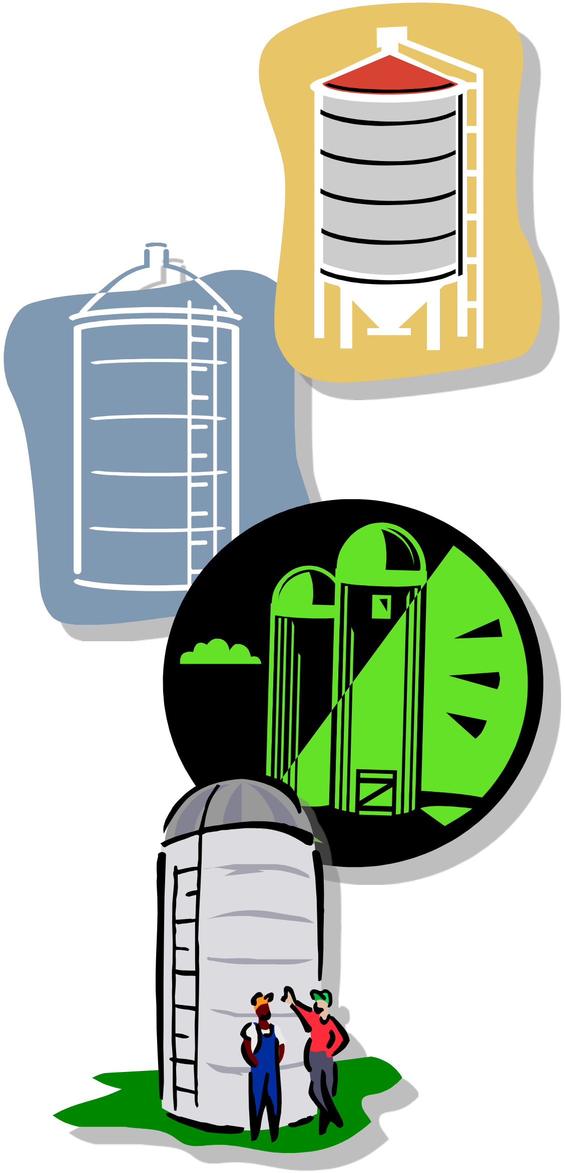


- Hadoop scales with file-based big data
  - *Imagine HDFS as shared infrastructure, similar to SAN & NAS*
  - *Imagine a huge, live archive*
  - *Imagine content mgt on steroids*
  - *Imagine low price per terabyte*
- HDFS extends BI, DW, analytics…
  - *Managing multi-structured data*
  - *Repository for detail source data*
  - *Processing big data for analytics*
  - *Advanced forms of analytics*
  - *Data staging on steroids*

AN AGE-OLD PROBLEM

Analytic Silos

- As advanced analytics becomes more common, it becomes more silo'd
- Why the silos? Most analytic apps are departmental, by nature
  - *Marketing owns customer segmentation, customer profitability*
  - *Procurement owns supplier analytics*
- Big data analytics can be very specialized (but doesn't have to be)
  - *Can have its own platform, tools, team*
  - *Be careful not to silo Hadoop & its data*
- POINT – Integrate new analytic apps into distributed DW architecture:
  - *To avoid analytic silos; share analytic insights; make the analytics better; get better data governance*



A Look Into the Future of Big Data Analytics

- Big data analytics is here to stay
  - *No.1 adoption trend w/TDWI members*
- Big data will be petabytes, not terabytes
  - *Half petabyte will be common in 3 yrs*
- Big data is less & less a mgt problem
  - *Due to advances in DBMSs & hardware*
- Analytics will draw biz value from big data
  - *That's why the two have come together*
- New types of analytic apps will appear
  - *Old ones will be revamped*
  - *OLAP & reports won't go away*
- Big Data Analytics is mostly batch today
  - *Will go real time as users/techs mature*
- Big data & analytics are new competencies for many BI/DW & IT shops
  - *They will hire & train, plus acquire tools*

Recommendations

- Foster your DW as a killer platform for reporting, OLAP, performance mgt
  - *Be open to additional data platforms in extended DW environment for other workloads*
- Support multiple data workloads on single DW DBMS instance when you can
  - *E.g., consider offloading analytic workloads to a DW Appliance or analytic DBMS*
- Be open to alternative architectures
  - *Systems on the Side (SOSs) have a place, but you must control them*
  - *Both DW and DI architectures need adjustments to accommodate analytics*
- Be open to new or alternative DW platforms, not just traditional ones.
  - *New DBMS types and brands provide more options, so at least consider them:*
    - Analytic DBMSs, Data Warehouse Appliances, Columnar databases, No-SQL, etc.
    - Also: Hadoop, MapReduce, Clouds for DW/BI & analytics, SaaS
- Incorporate new data types and new data sources
  - *Semi- and un-structured data. Web, machine, and social data*
- Adjust best practices in data management
  - *These still apply to big data analytics, but different order & priority*
- Embrace real-time operation, maybe streaming big data
  - *Real-time is a biz requirement. Many new sources stream big data.*
  - *This requires very special tools, probably outside purview of DW*
- Take command of your architecture(s).
  - *Big Data and Analytics are driving up DW architecture complexity*
  - *Know the biz/tech requirements per analytic app & design arch accordingly*

TERADATA | THE BEST DECISION POSSIBLE™

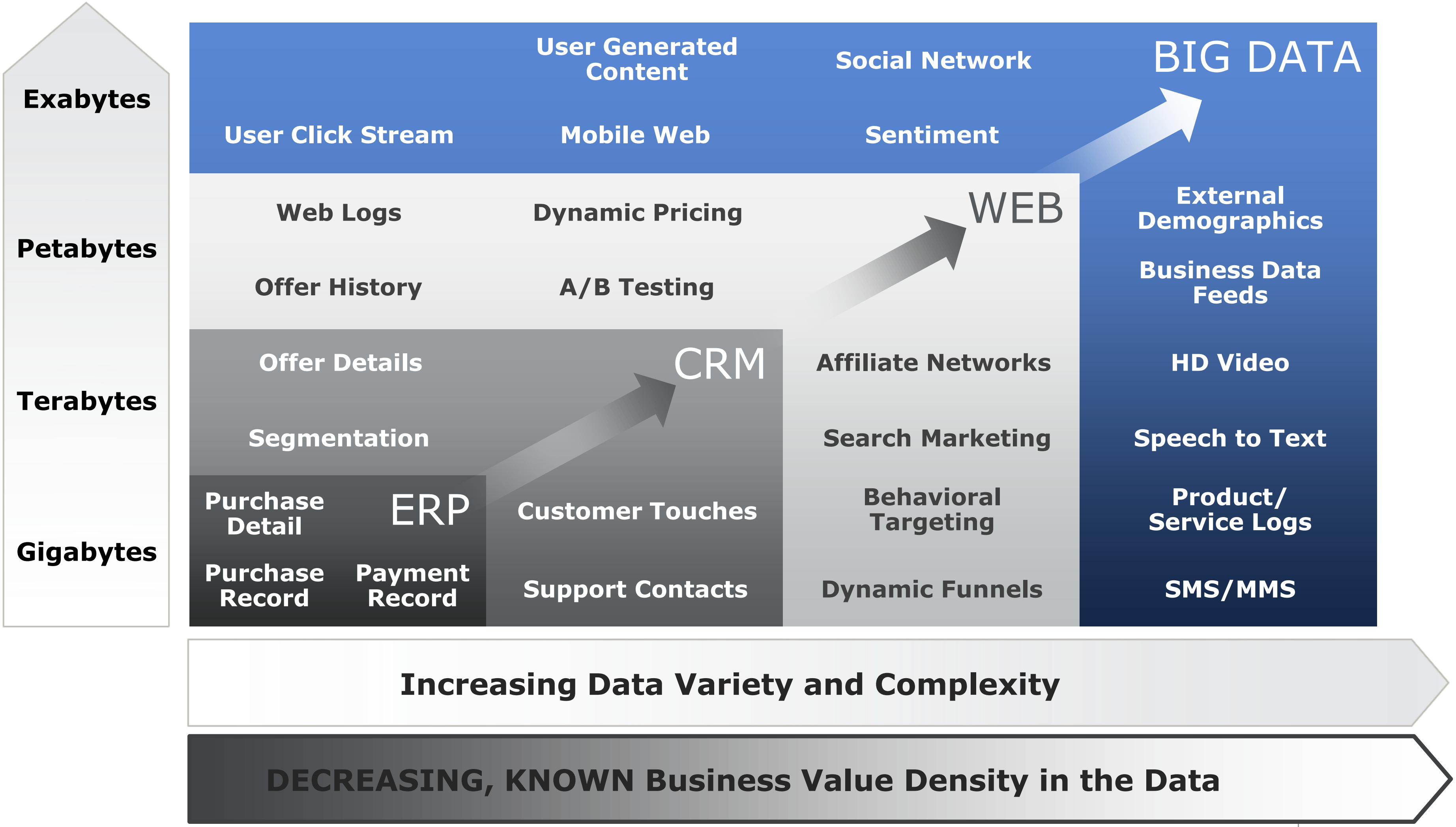
UNLOCKING BIG DATA: UNIFIED DATA ARCHITECTURE

Chris Twogood, VP, Product Marketing

March 14, 2013

Big Data: From Transactions to Interactions

Business Goal – Unlock Unknown Value from "Sparse" Data



Exabytes

Petabytes

Terabytes

Gigabytes

BIG DATA

User Generated Content

Social Network

User Click Stream

Mobile Web

Sentiment

WEB

Web Logs

Dynamic Pricing

Offer History

A/B Testing

External Demographics

Business Data Feeds

CRM

Offer Details

Segmentation

Affiliate Networks

Search Marketing

HD Video

Speech to Text

ERP

Purchase Detail

Purchase Record

Payment Record

Customer Touches

Support Contacts

Behavioral Targeting

Dynamic Funnels

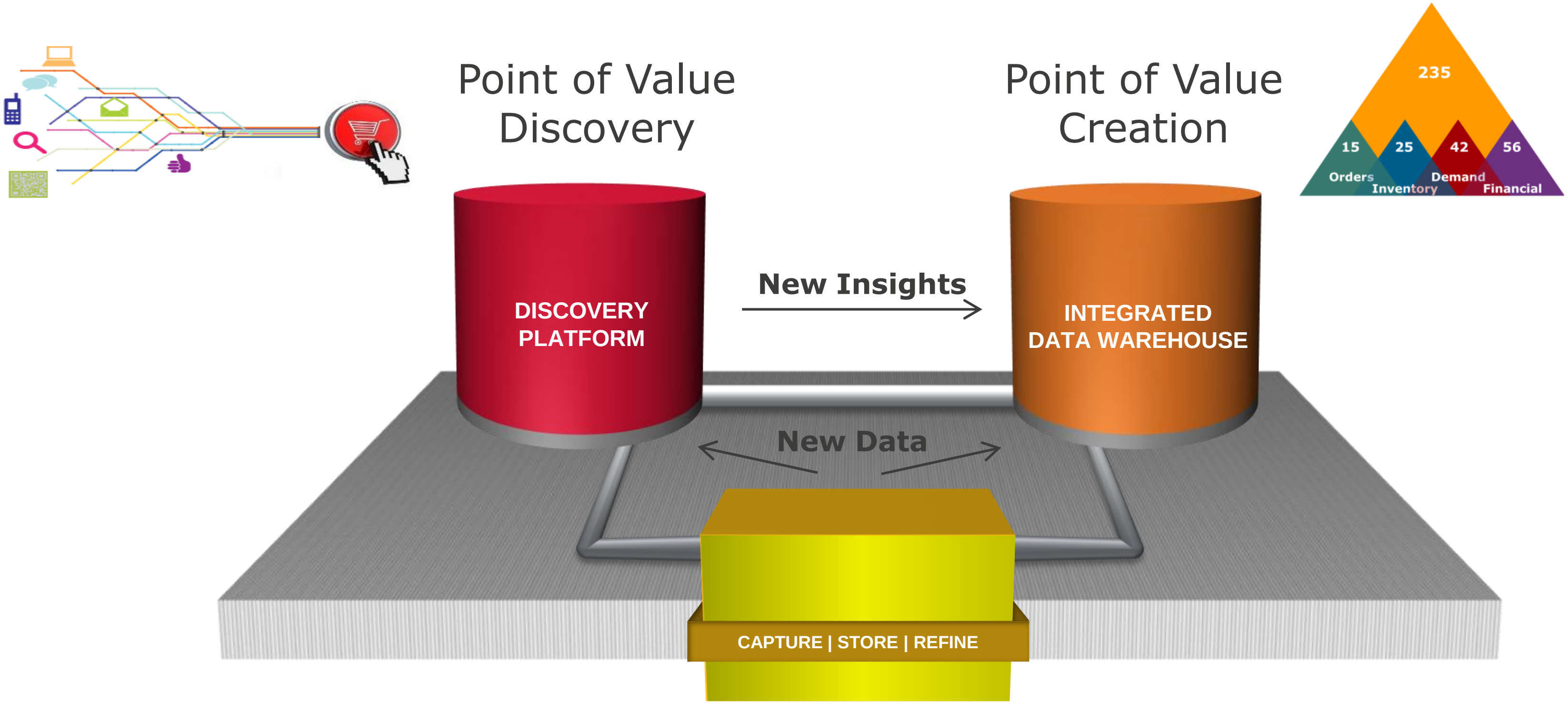
Product/ Service Logs

SMS/MMS

Increasing Data Variety and Complexity

DECREASING, KNOWN Business Value Density in the Data

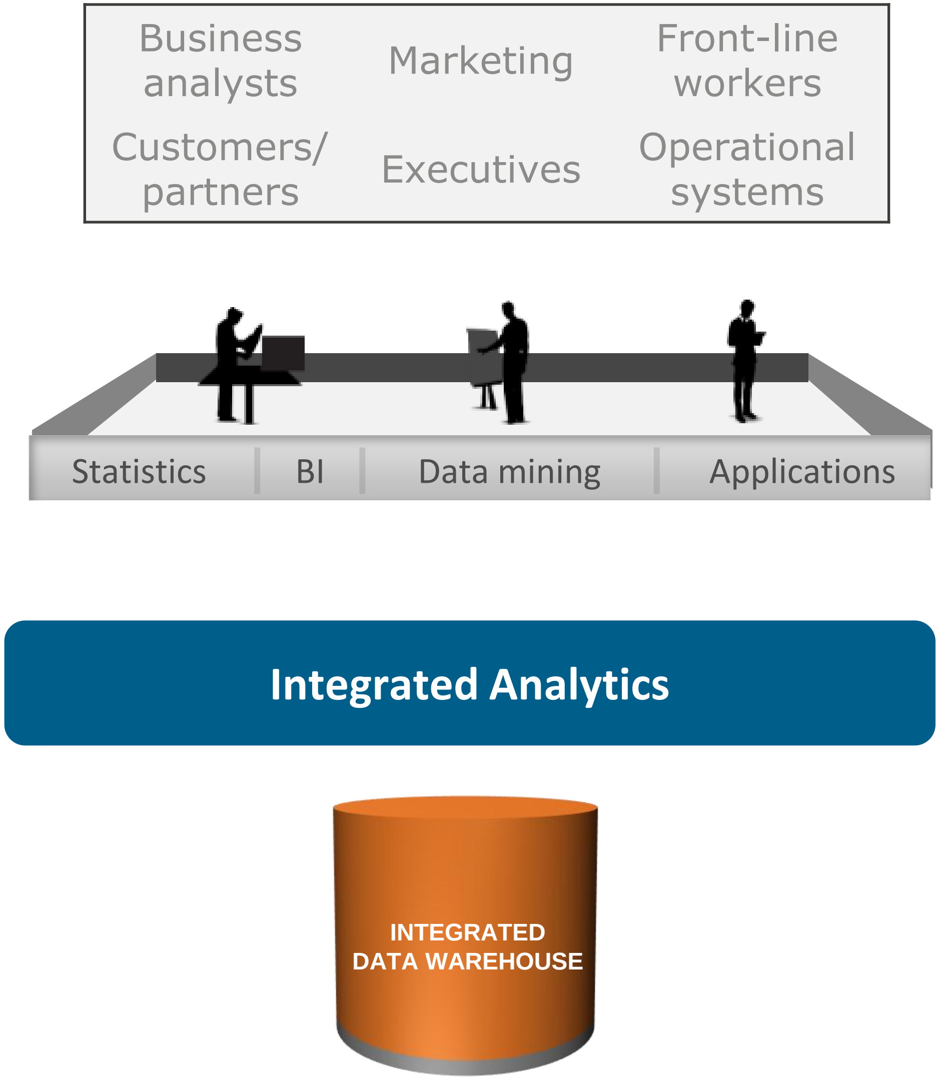
Need for a Unified Data Architecture for New Insights



Point of Data Capture

Operationalizing Insights in the Enterprise

- Single view of your business
- Cross-functional analysis
- Shared source of relevant, consistent, integrated data
- Load once, use many times
- Lowest cost of ownership
- Fast new applications time-to-market



Business analysts

Marketing

Front-line workers

Customers/ partners

Executives

Operational systems

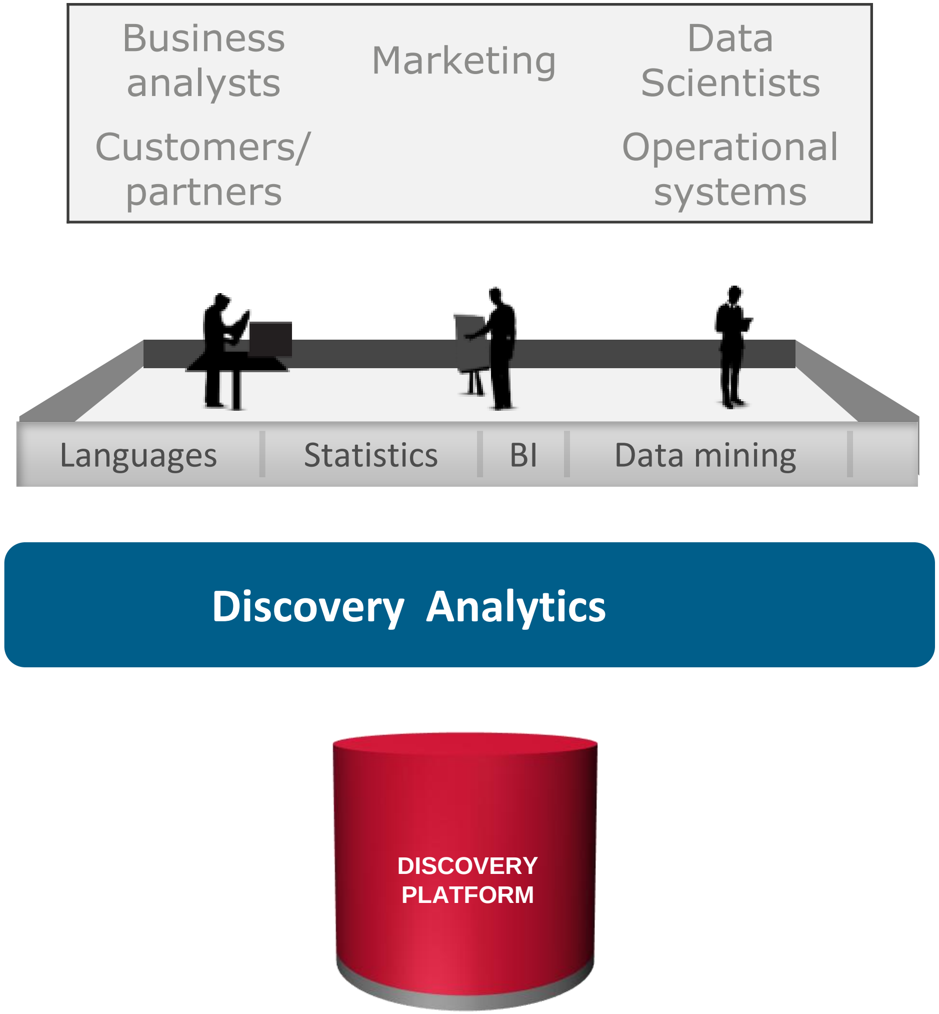
Statistics | BI | Data mining | Applications

Integrated Analytics

INTEGRATED DATA WAREHOUSE

Unlocking Hidden Value in (Any) Data

- Interactive data discovery
  - Web clickstream, social
  - Set-top box analysis
  - CDRs, sensor logs, JSON
- Flexible evolving schema
- MapReduce, SQL, statistics, text, …
- Structured and multi-structured data



Business analysts

Marketing

Data Scientists

Customers/ partners

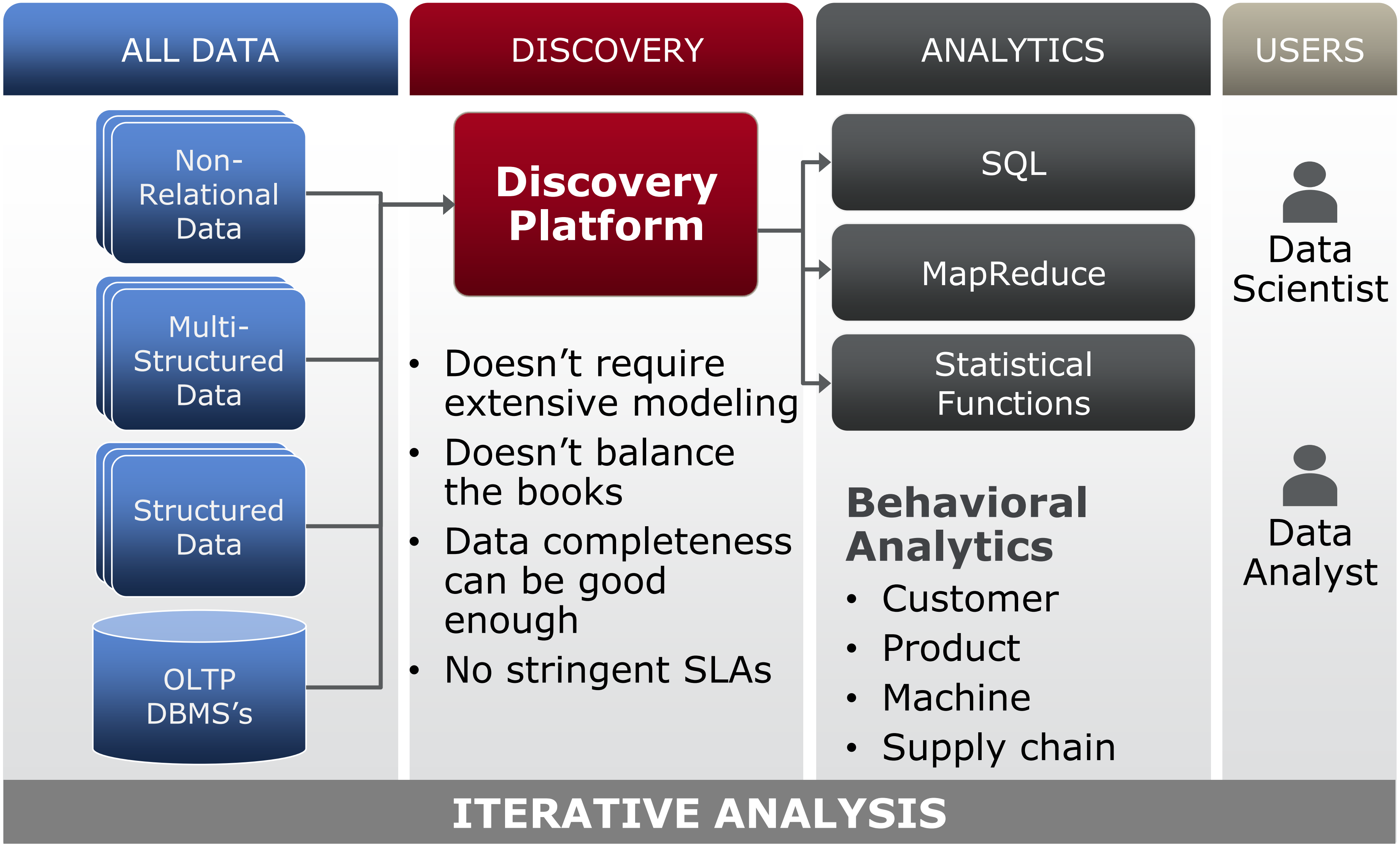
Operational systems

Languages | Statistics | BI | Data mining

Discovery Analytics

DISCOVERY PLATFORM

Discovery Platform *Requirements*



ALL DATA

Non-Relational Data

Multi-Structured Data

Structured Data

OLTP DBMS’s

DISCOVERY

Discovery Platform

- Doesn’t require extensive modeling
- Doesn’t balance the books
- Data completeness can be good enough
- No stringent SLAs

ANALYTICS

SQL

MapReduce

Statistical Functions

Behavioral Analytics

- Customer
- Product
- Machine
- Supply chain

USERS

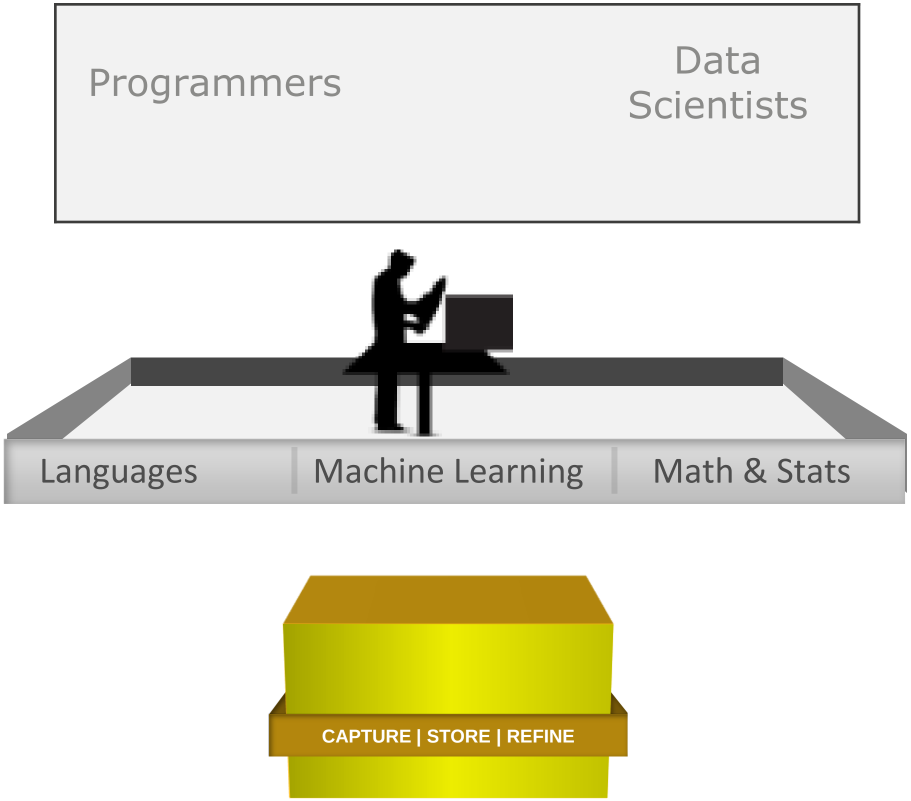
Data Scientist

Data Analyst

ITERATIVE ANALYSIS

Capturing Data for Storage and Refining

- Raw data capture
- History or long term storage
  - Low cost archival
- Transformations
  - Structured, semi-structured
  - Sessionize, remove XML tags, extract key words
- Simple math at scale
- Batch processing



Unified Data Architecture

TECHNICAL REQUIREMENTS

TERADATA SOLUTION

Data Warehousing

- Integrated and shared data environment
- Manages the business
- Strategic & operational analytics
- Extended throughout the organization

Teradata Active IDW

- Market-leading platform for delivering strategic and operational analytics
- Single source of centralized data for reuse

Data Discovery

- Unlock insights from big data
- Rapid exploration capabilities
- Variety of analytic techniques
- Accessible by business analysts

Teradata Aster

- Patented SQL-MapReduce capability for discovery analytics
- Pre-packaged analytics for data-driven discovery

Data Staging

- Loading, storing, and refining data in preparation for analytics

Hadoop

- Effective, low-cost technology for loading, storing, and refining data
- 1xxx and 2xxx recommended for stable schema data

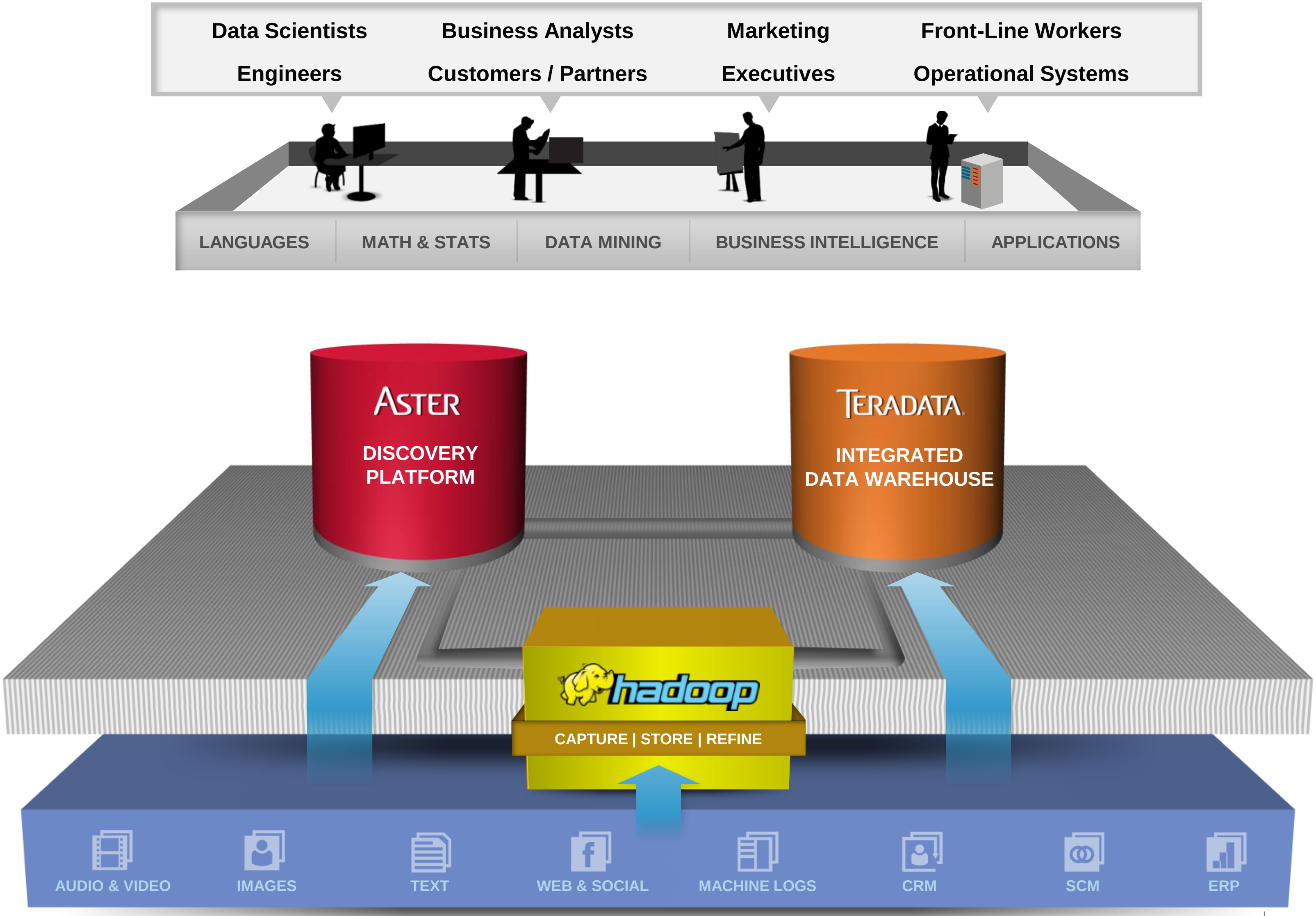
Benefits of Teradata Unified Data Architecture

The only truly integrated analytics solution that unifies multiple technologies into a cohesive and transparent architecture

Best-of-breed and values of Teradata, Teradata Aster, and Hadoop

- Valuable Insights From All Your Data
- Fast, Flexible Deployment
- Sophisticated Analytics For Business and Technical Users

Teradata Unified Data Architecture



Teradata UDA Advanced Analytics

Continue to deliver leading established and emerging advance analytics that enable any type of analytics on any type of data at any time.

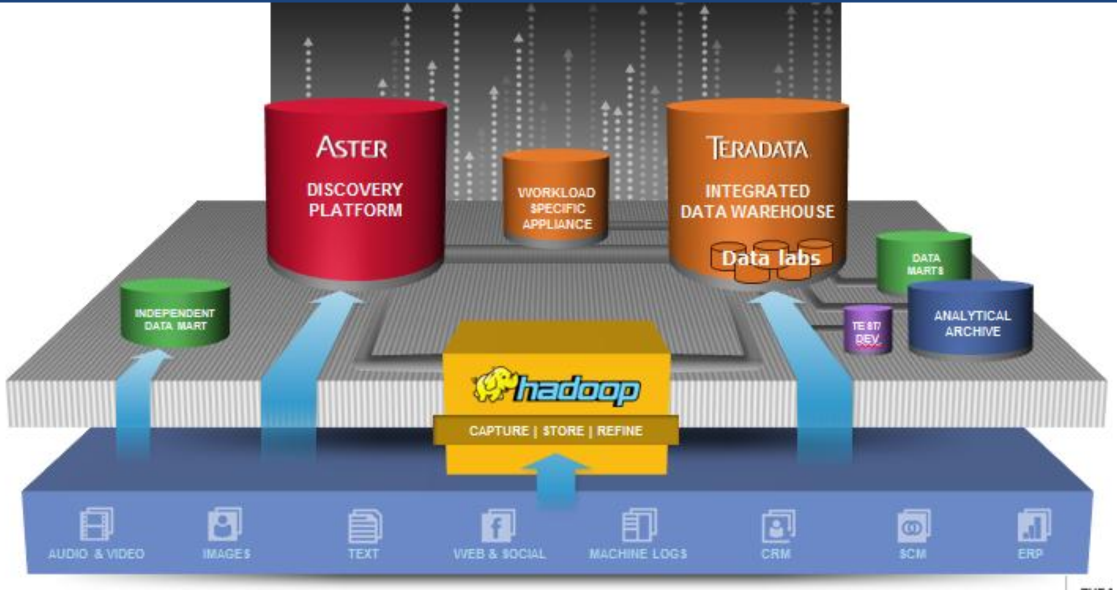
1. Expand in-database analytics
2. Extend the UDA with workload specific platforms
3. Enhance and simplify the user's experience within the UDA

Languages
Big Analytics
Text Analytics

Custom Analytics
Mathematics
Statistics

Model Scoring
Matrix Operations
Modeling Algorithms

Visualization
Transformations
Data Preparation



Teradata Aster Discovery Platform

PATH ANALYSIS
Discover Patterns in Rows of Sequential Data

TEXT ANALYSIS
Derive Patterns and Extract Features in Textual Data

STATISTICAL ANALYSIS
High-Performance Processing of Common Statistical Calculations

SEGMENTATION
Discover Natural Groupings of Data Points

MARKETING ANALYTICS
Analyze Customer Interactions to Optimize Marketing Decisions

DATA TRANSFORMATION
Transform Data for More Advanced Analysis

GRAPH ANALYTICS
Native graph analytics processing engine to simplify storage and processing

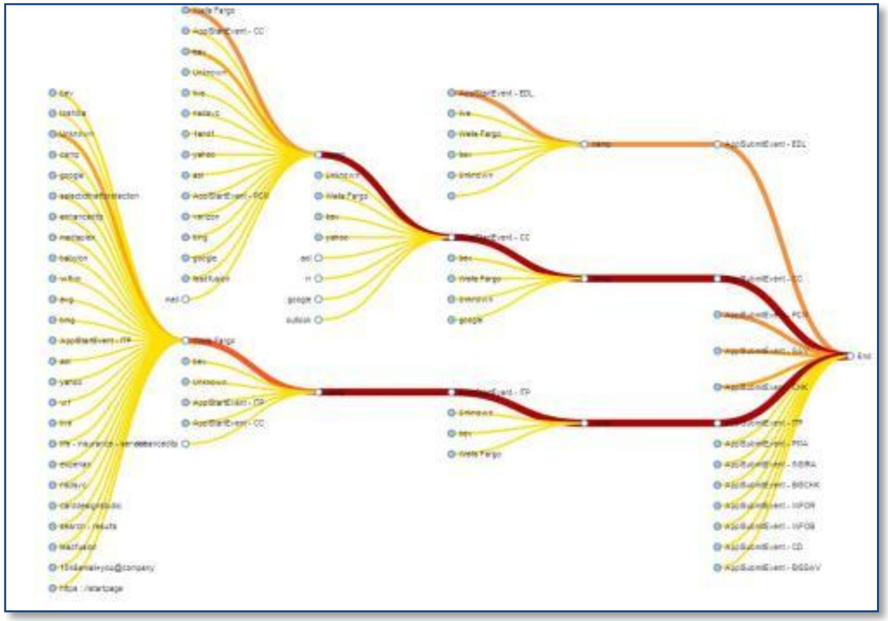
Coming in 2013

SQL-MR VISUALIZATION
Graphing and visualization tools linked to key functions of the MapReduce analytics library

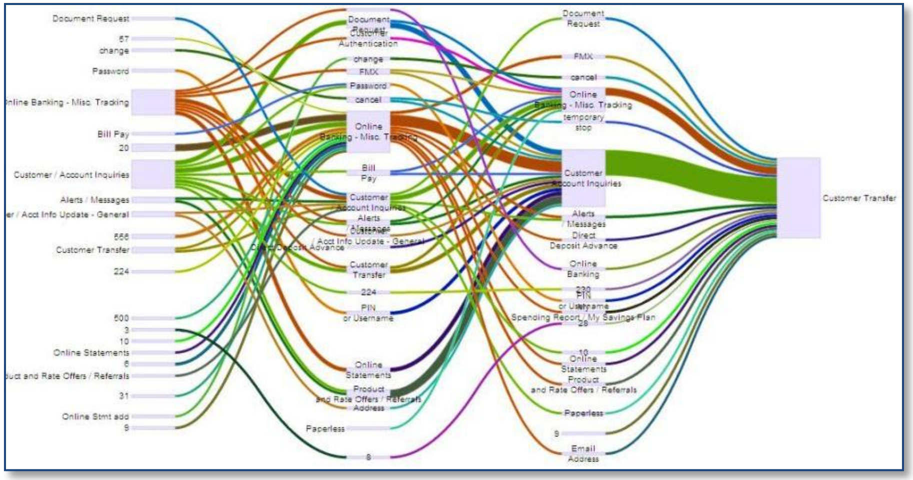
Uncover New Insights & Make Actionable

- Deliver valuable insight to lines of business resulting from deep analysis of all of your data, all of the time

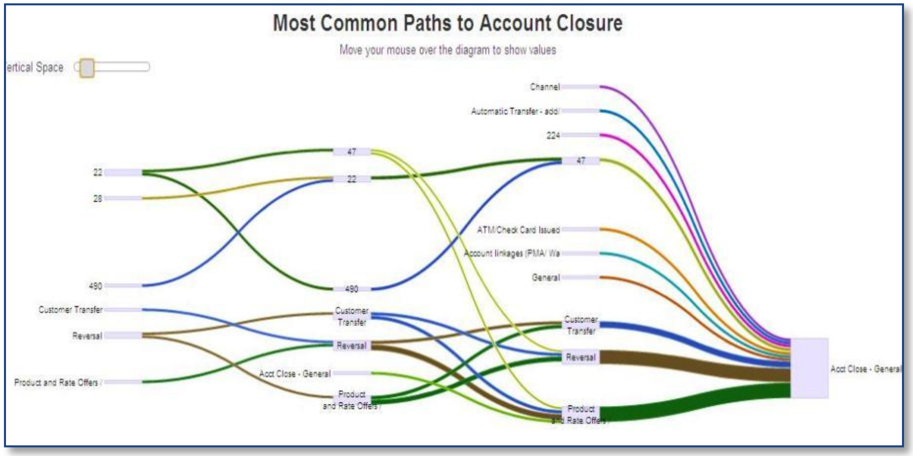
Golden Path to Application Submit



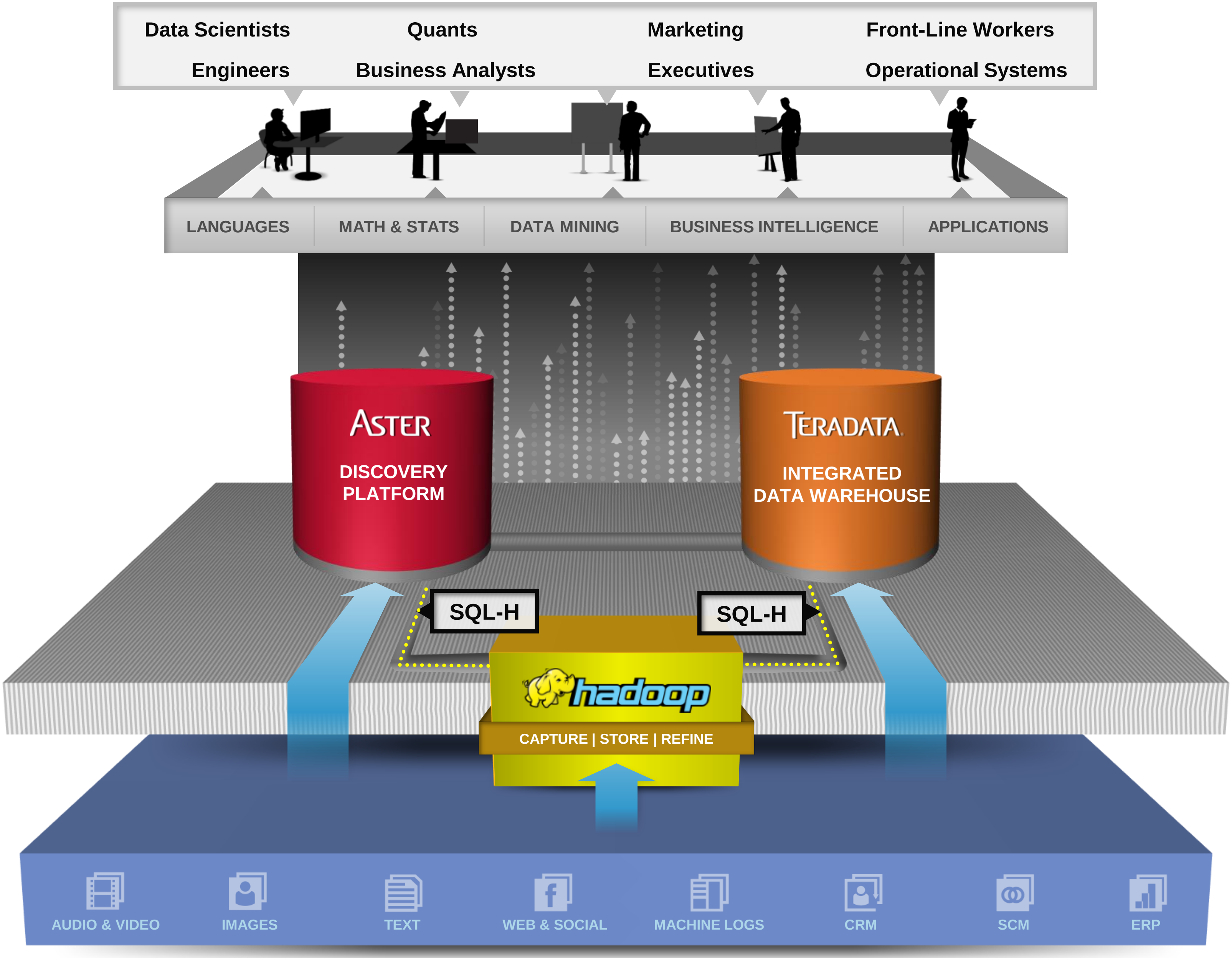
Fraudulent Paths



Paths To Attrition


Most Common Paths to Account Closure
Move your mouse over the diagram to show values
ertical Space
22
28
490
Customer Transfer
Reversal
Product and Rate Offers /
47
22
490
Customer Transfer
Reversal
Acct Close - General
Product and Rate Offers /
Channel
Automatic Transfer - add/
224
47
ATM/Check Card Issued
Account linkages (PMA/ Wa
General
Customer Transfer
Reversal
Product and Rate Offers /
Acct Close - General

TERADATA UNIFIED DATA ARCHITECTURE

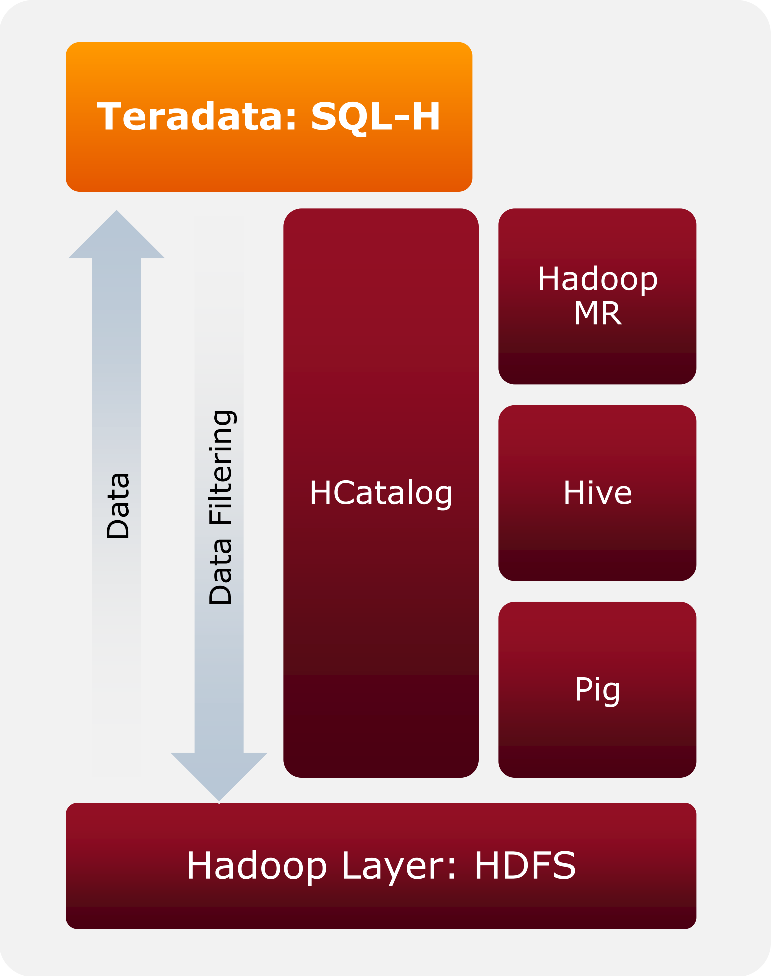


Teradata SQL-H™

Gives business users on-the-fly access to data in Hadoop

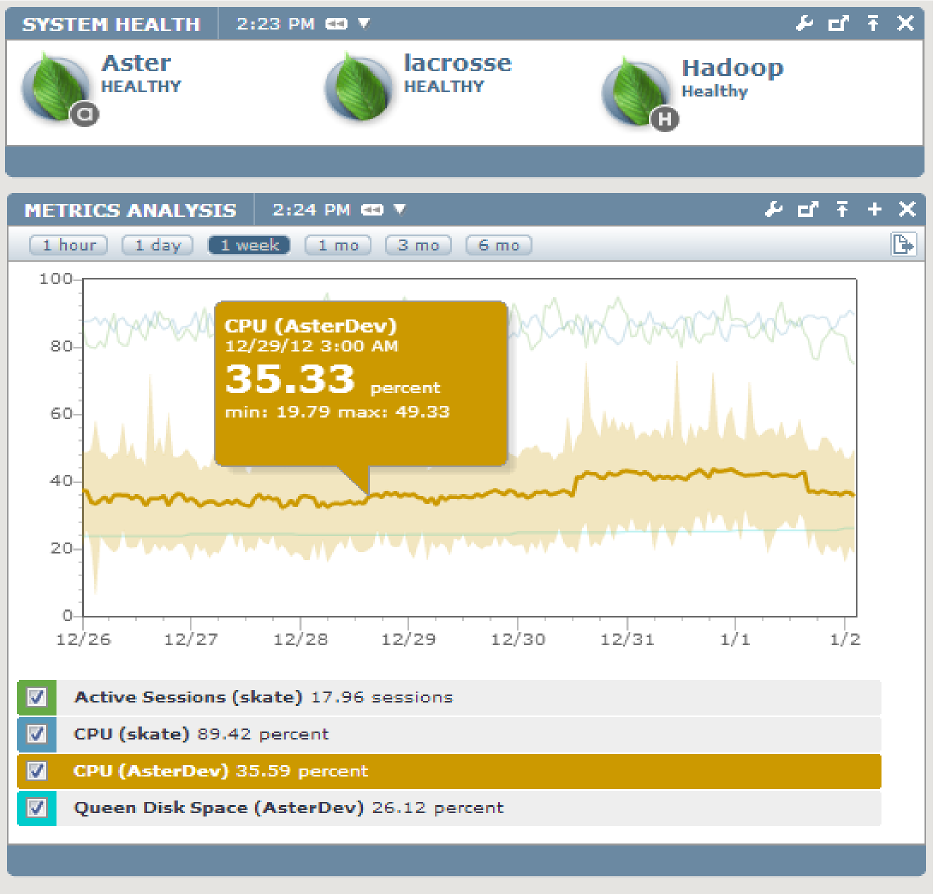
SQL-H Gives Business Users a Better Way to Access Data Stored in Hadoop

- **Trusted**: Use existing tools/skills and enable self-service BI with granular security
- Allow standard ANSI SQL access to Hadoop data
- **Fast:** Queries run on Teradata, data accessed from Hadoop
- **Efficient**: Intelligent data access leveraging the Hadoop HCatalog



Unified System Monitoring and Management

- **Viewpoint shared functionality:**
  - > System Health Monitoring
  - > Alerting
  - > Operational Metrics
  - > Node Monitoring
  - > Metric Trends
  - > Space Management
  - > Administrative Setup
  - > Advanced Analysis with Rewind



Aster HEALTHY

lacrosse HEALTHY

Hadoop Healthy

CPU (AsterDev)
12/29/12 3:00 AM
35.33 percent
min: 19.79 max: 49.33

☑ Active Sessions (skate) 17.96 sessions

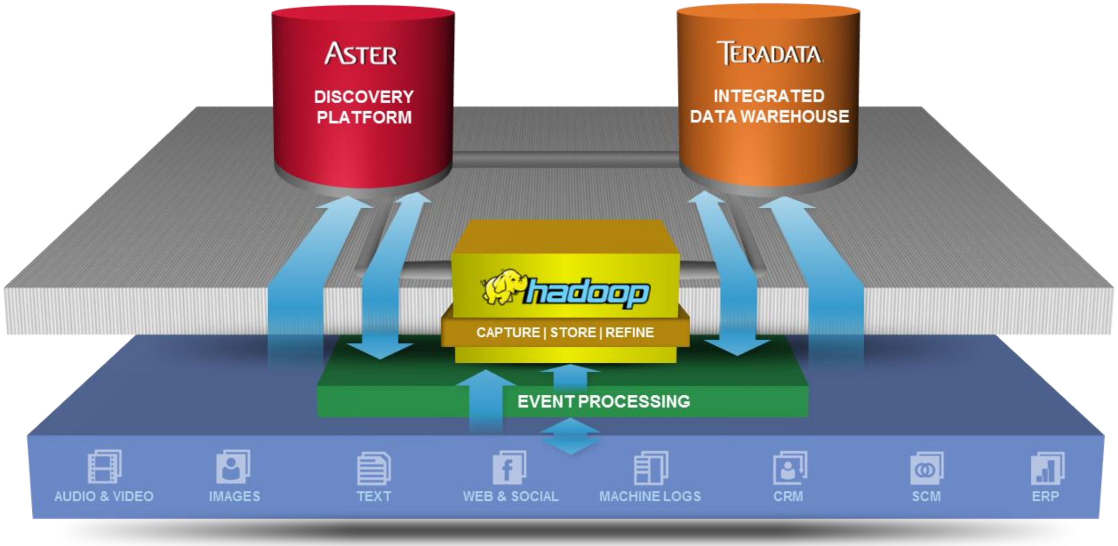
☑ CPU (skate) 89.42 percent

☑ CPU (AsterDev) 35.59 percent

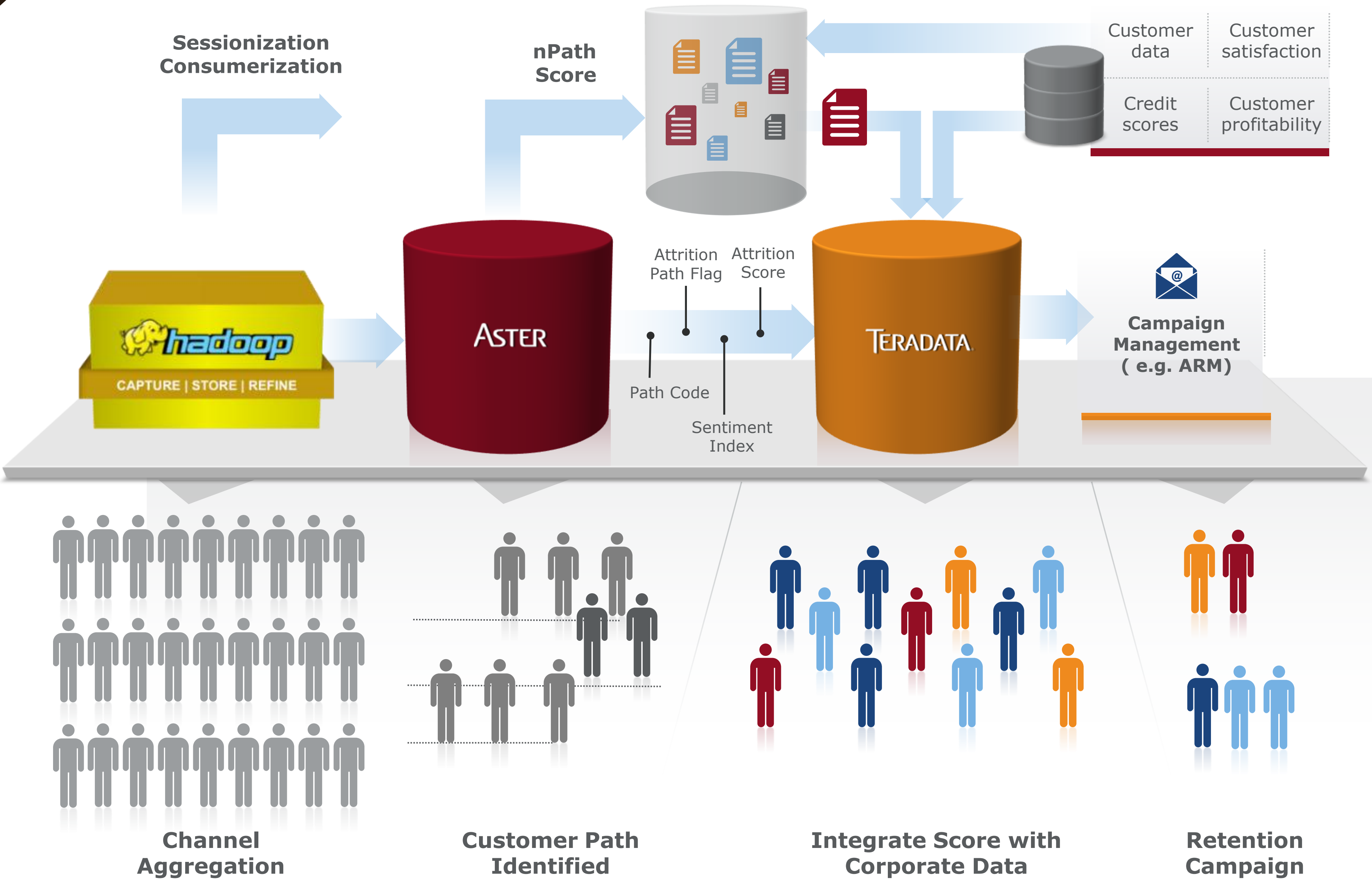
☑ Queen Disk Space (AsterDev) 26.12 percent

Complex Event and Data Streams Processing

- Analysis of “Data in Motion” - while it still has value
  - Streaming data is analyzed against pre-defined queries/models
  - Results are incrementally updated
  - Data in motion typically has a lot of noise
- Use Cases:
  - Update dashboards
  - Feed event driven applications
  - Generate alerts
  - Stored in an EDW for reporting
- Technology
  - Tibco BusinessEvents
  - IBM InfoSphere Streams



Improving Customer Retention



Sessionization Consumerization

nPath Score

Customer data

Customer satisfaction

Credit scores

Customer profitability

hadoop

CAPTURE | STORE | REFINE

ASTER

Attrition Path Flag

Attrition Score

Path Code

Sentiment Index

TERADATA

Campaign Management (e.g. ARM)

Channel Aggregation

Customer Path Identified

Integrate Score with Corporate Data

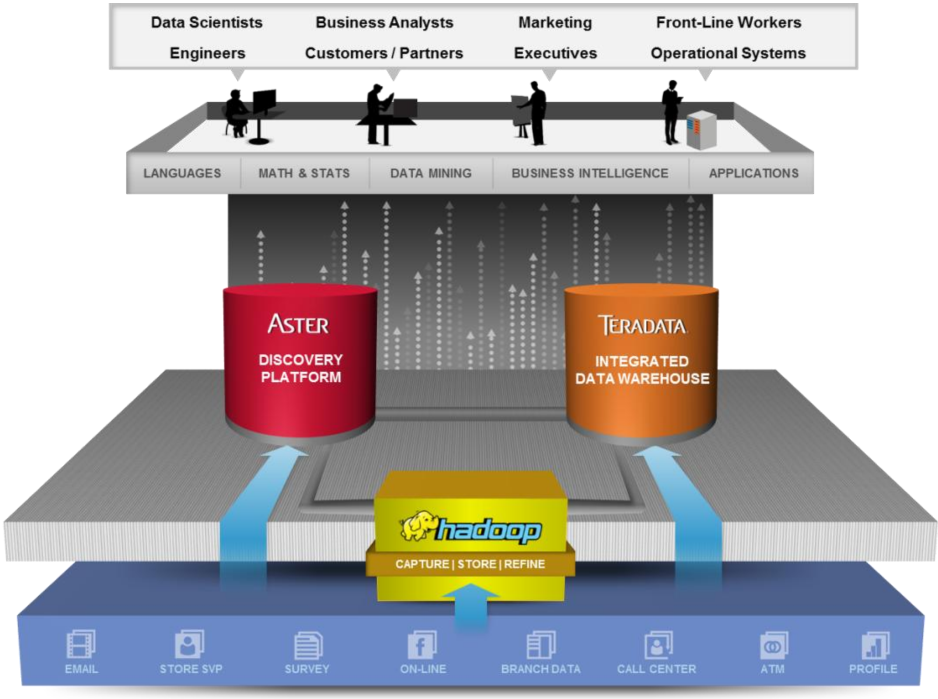
Retention Campaign

Teradata Unified Data Architecture™

Competitive advantage through deeper, comprehensive insights

Truly integrated analytic solution

- Provides best-of-breed value of Teradata, Aster, and Hadoop
- Unifies into comprehensive & transparent architecture
- Supported by data experts with deep industry experience



Questions and Answers

Contact Information

If you have further questions or comments:

Philip Russom, TDWI
prussom@tdwi.org

Chris Twogood, Teradata
chris.twogood@teradata.com