Challenges of Handling Big Data

Ramesh Bhashyam
Teradata Fellow
Teradata Corporation

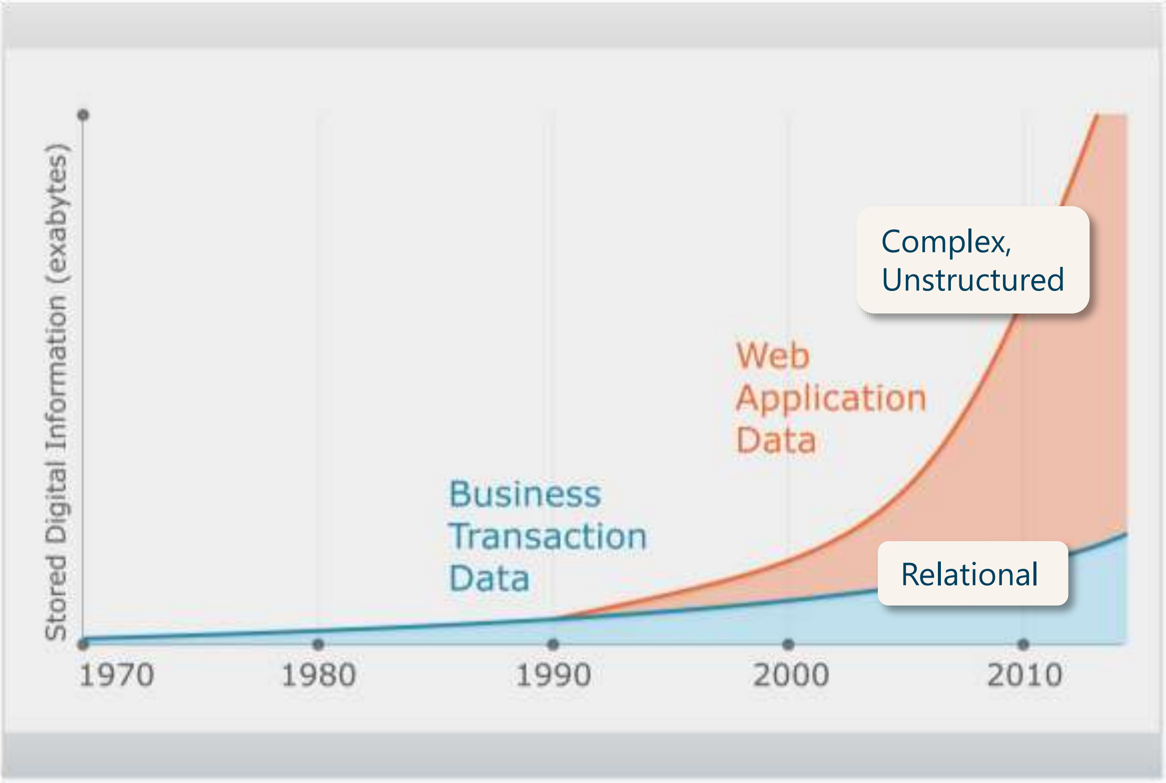
bhashyam.ramesh@teradata.com

Trend

“Too much information is a storage issue, certainly, but too much information is also a massive analysis issue.” Source: Gartner’s Report

- Volume of Data
- Complexity Of Analysis
- Velocity of Data - Real-Time Analytics
- Variety of Data - Cross-Analytics

Reality: Massive Data Growth



2,500 exabytes of new information in 2012 with digital content as the primary driver

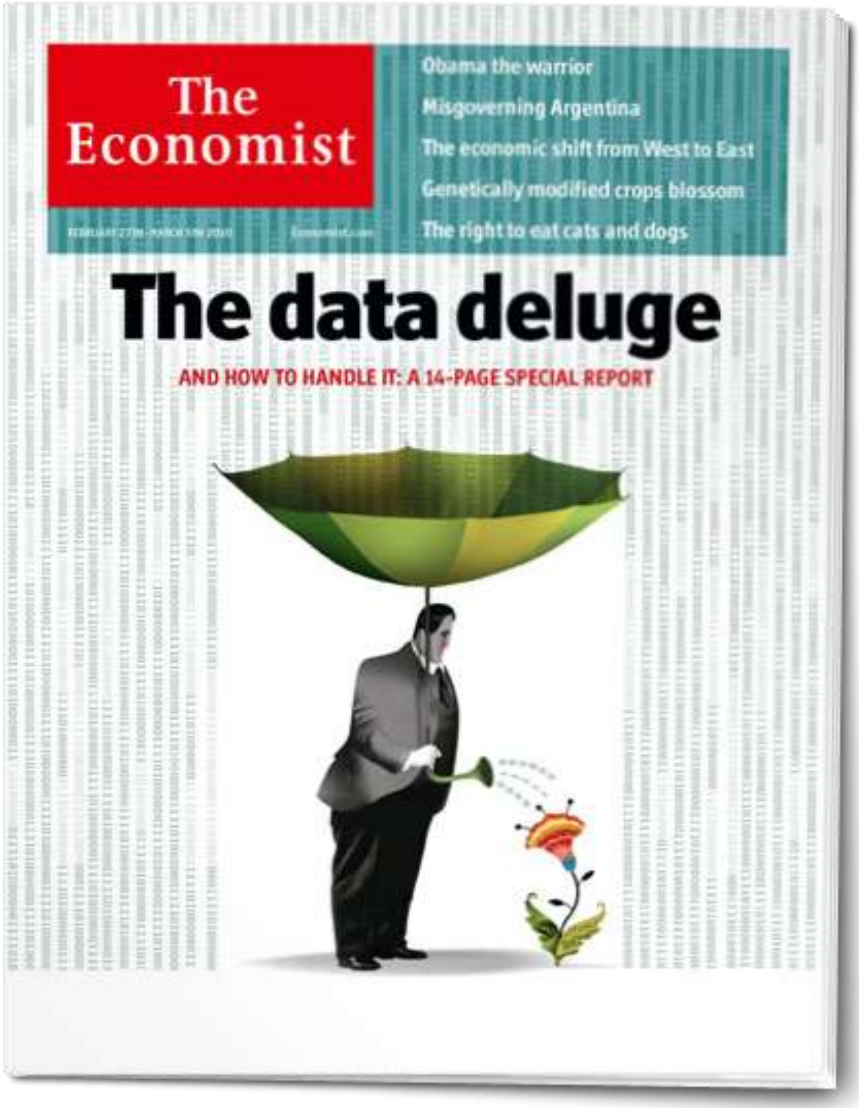
Digital universe grew by 62% last year to 800K petabytes and will grow to 1.2 “zettabytes” this year

Source: An IDC White Paper. As the Economy Contracts, the Digital Universe Expands. May 2009.

“Big Data”

- Structured and Unstructured
  - Data with structure and data with application imposed structure
  - SQL and static ERD; non SQL and dynamic ERD
- 10x – 100x of today’s data warehousing
- Projects to 5-50 EB (10^{19}) by 2015
- Need to separate the useful from the useless
- Shared nothing parallel analysis

Strategic Opportunity


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

"Data is widely available; what is scarce is the ability to extract wisdom from it."

Hal Varian, Chief Economist, Google

The Unmet Need!

Complexity

- Model Complexity
- Query Complexity
- Concurrency

Data Silos Limit Business Value

Limited Business Value

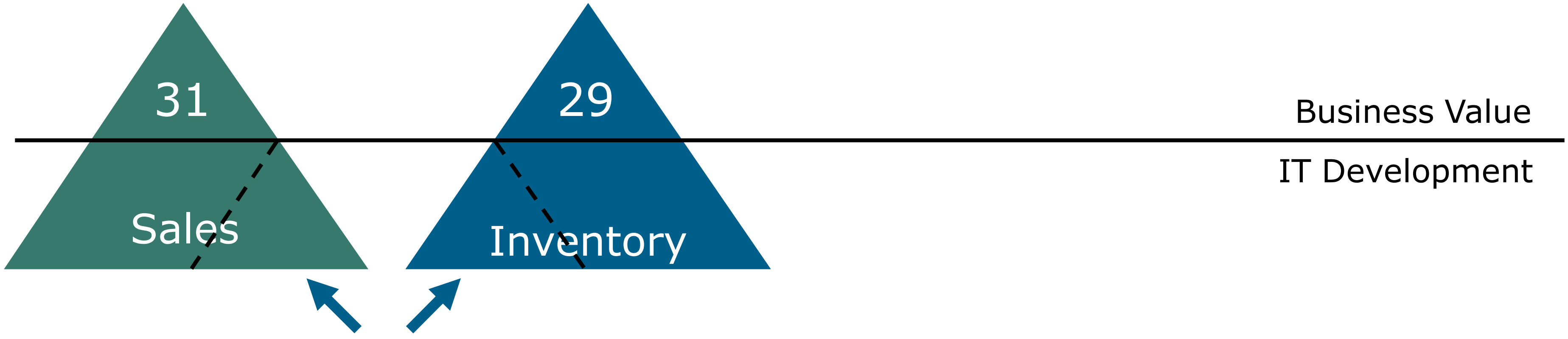
- Subject-specific questions
- Simple data model. Star schema, OLAP
- Many common tables necessary in each mart such as products, store, transactions

Sales

- What is the sales by product

Marketing

- What is the inventory for product X?



Duplicated Data

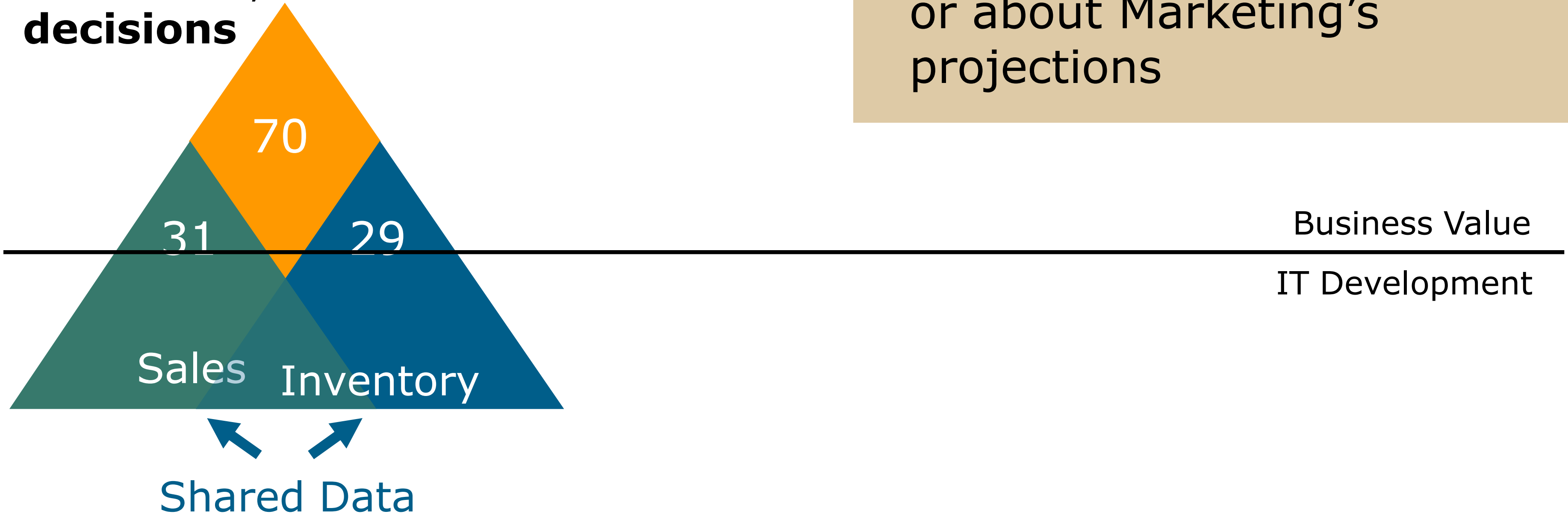
Integrated Data Enables Superior Value

Differentiated Business Value

- Combining the environments requires only incremental work for each new subject area
- Complex data model. 100s of entities and relations - snowflake
- Enables new cross-functional insights that can't be achieved with separate data marts; **new differentiated decisions**

Combined Sales and Inventory

- Which product sales can be increased by 20% in what stores?
- Cannot answer questions about supply chain capability or about Marketing's projections



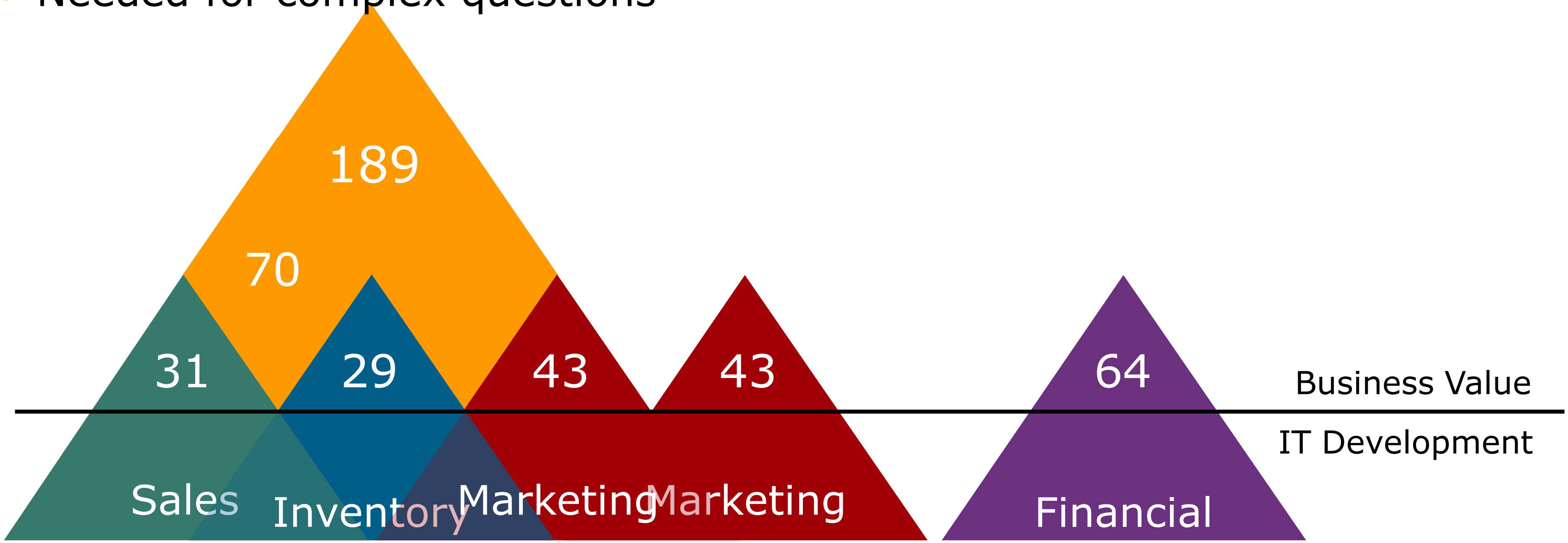
Integrated Data Enables Superior Value

Differentiated Business Value

- Very complex data model. Thousands of entities and relations. Spans across all subject areas. Very large tables.
- Needed for complex questions

Combined Sales, Inventory, Manufacturing, Supply chain

- Can manufacturing support sales projections

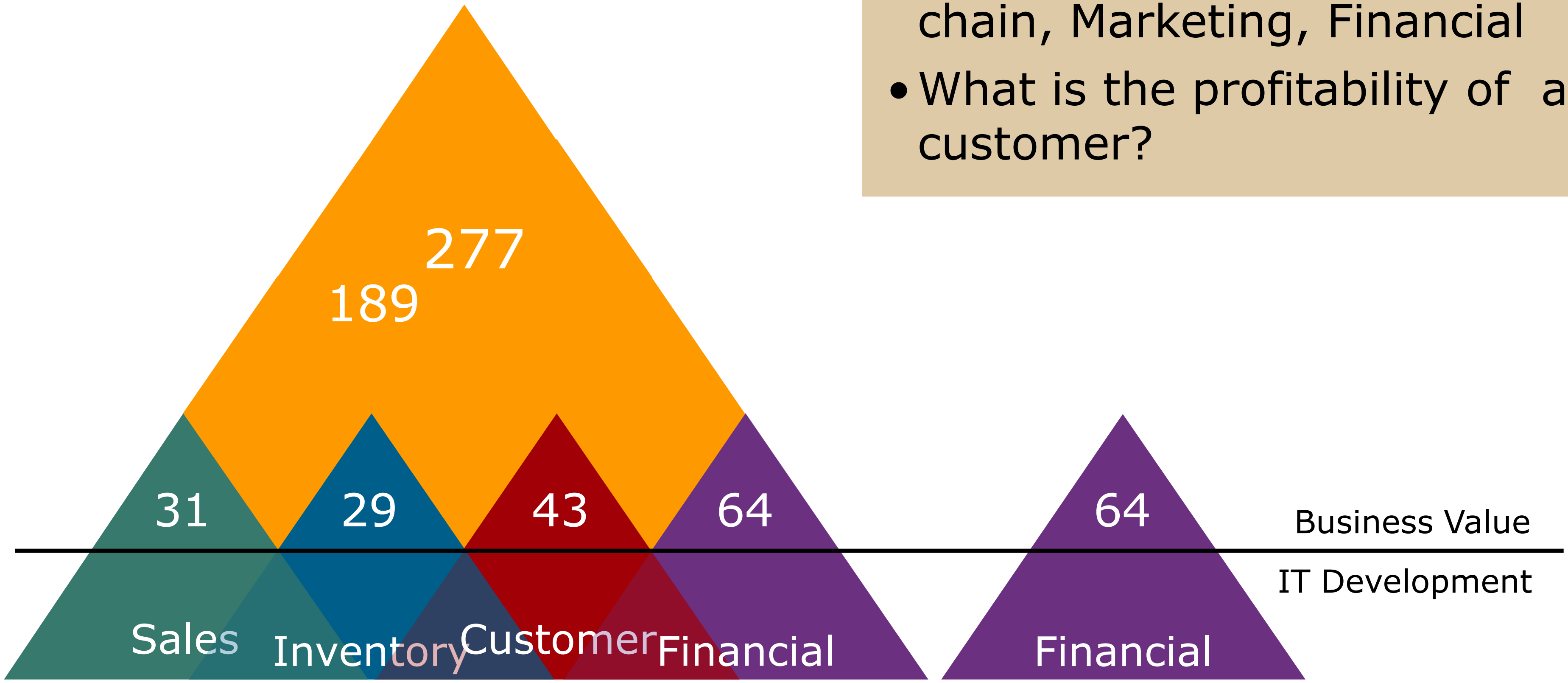


Integrated Data Enables Superior Value

Differentiated Business Value

Combined Sales, Inventory, Manufacturing, Supply chain, Marketing, Financial

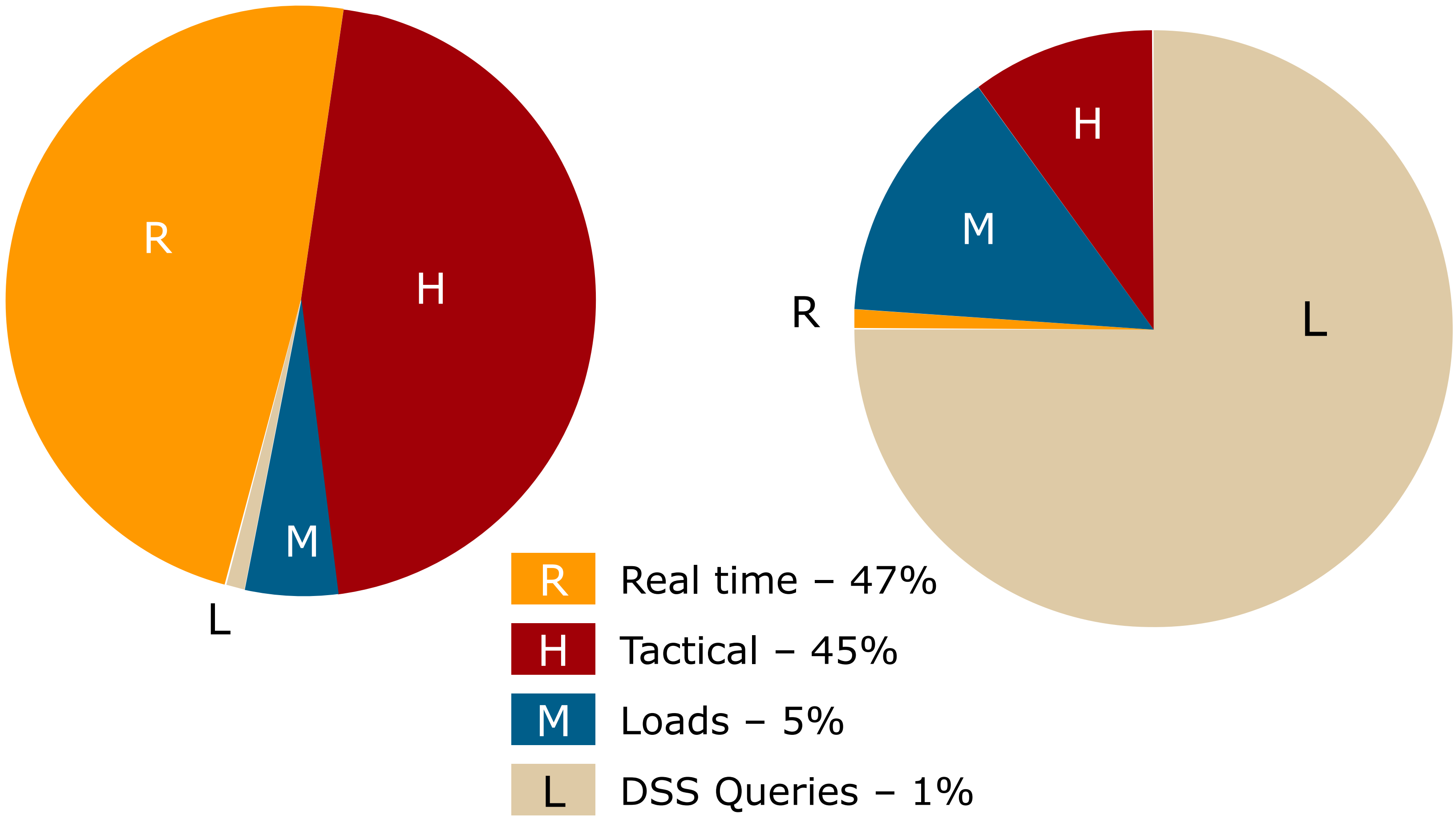
- What is the profitability of a customer?



Complexity - Query and Analytics

- Complex query plans
  - 128 way joins
  - Enormously large set of solution possibilities
- Comprehensive query operations
  - Joins
  - Aggregations
  - OLAP (rank, window analytics)
  - Time Series Analysis
- Scalable – No Fat-Node bottleneck
- Millions of queries
- Data loading throughput and latency

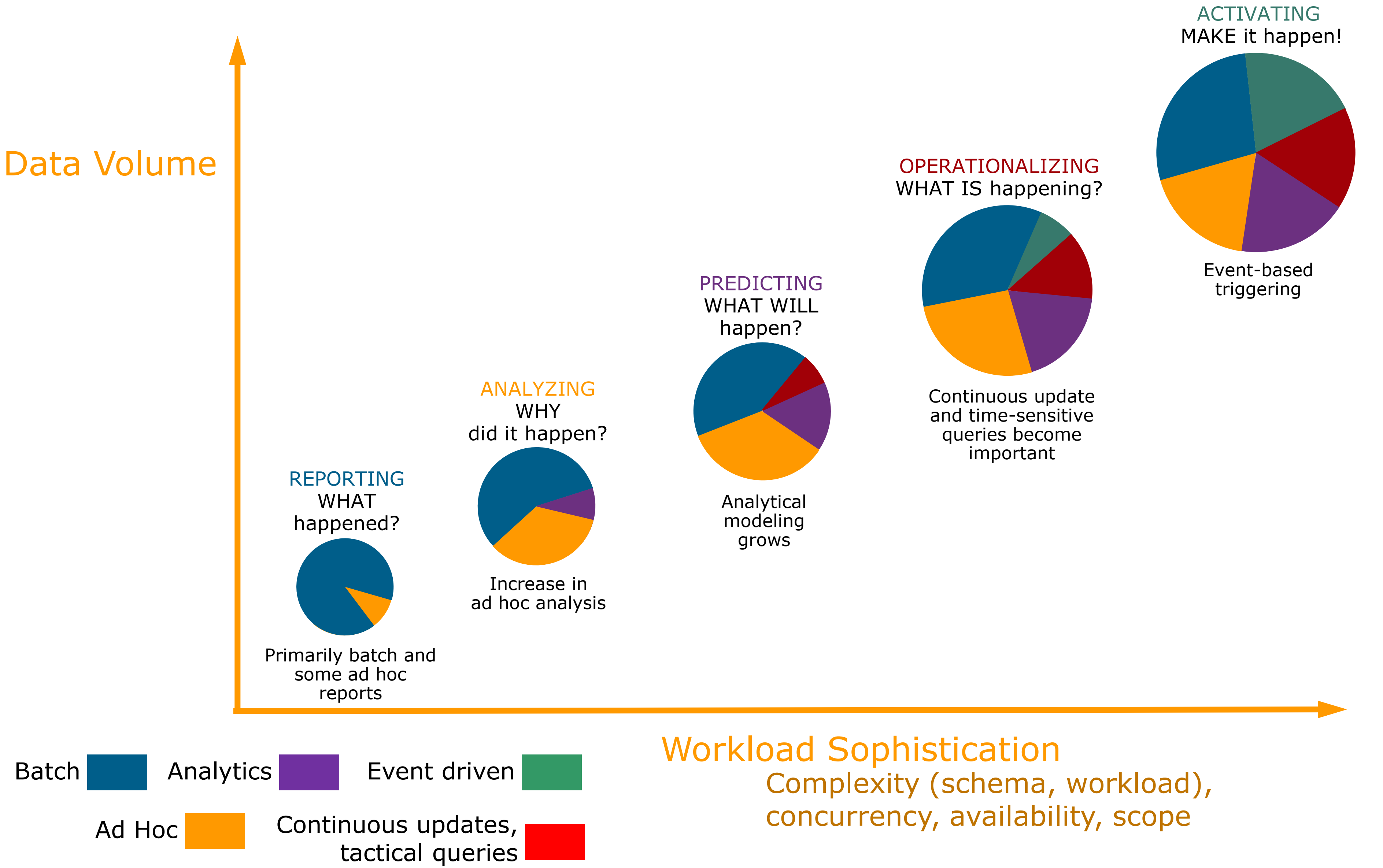
Workload Management



Workload Priorities

Actual Server Utilization

Five Stages of Analytic Evolution



Data Volume

REPORTING
WHAT
happened?

Primarily batch and
some ad hoc
reports

ANALYZING
WHY
did it happen?

Increase in
ad hoc analysis

PREDICTING
WHAT WILL
happen?

Analytical
modeling
grows

OPERATIONALIZING
WHAT IS happening?

Continuous update
and time-sensitive
queries become
important

ACTIVATING
MAKE it happen!

Event-based
triggering

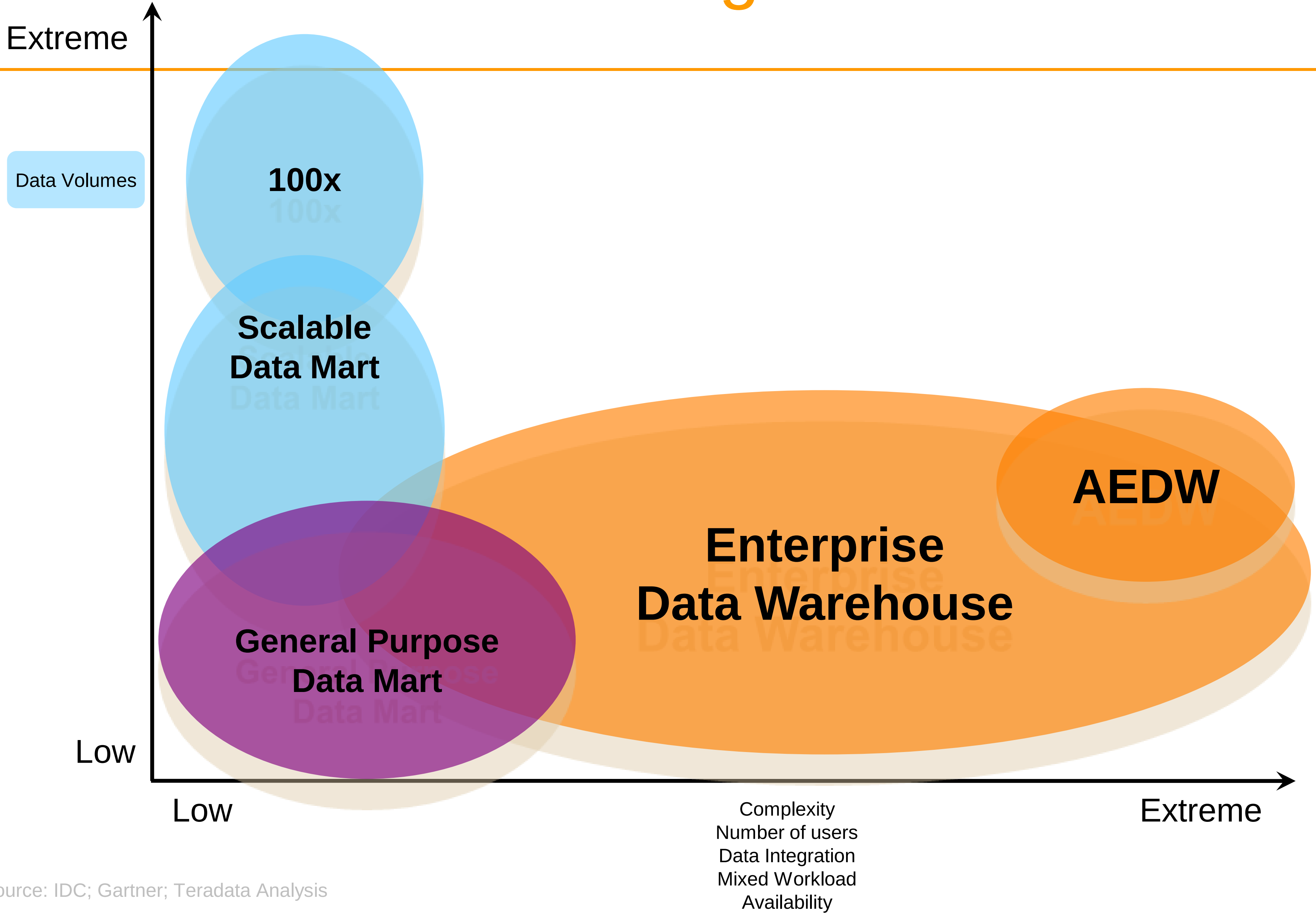
Workload Sophistication

Complexity (schema, workload),
concurrency, availability, scope

Batch 🟦 Analytics 🟪 Event driven 🟩

Ad Hoc 🟧 Continuous updates, tactical queries 🟥

Problem Segmentation

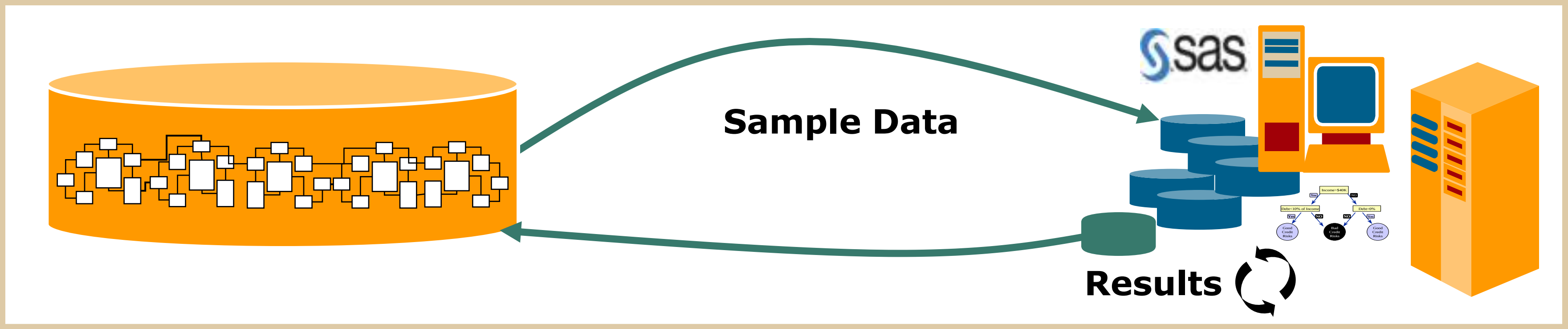


Source: IDC; Gartner; Teradata Analysis

In-Database Data Mining Optimization

More Models and More Business Value

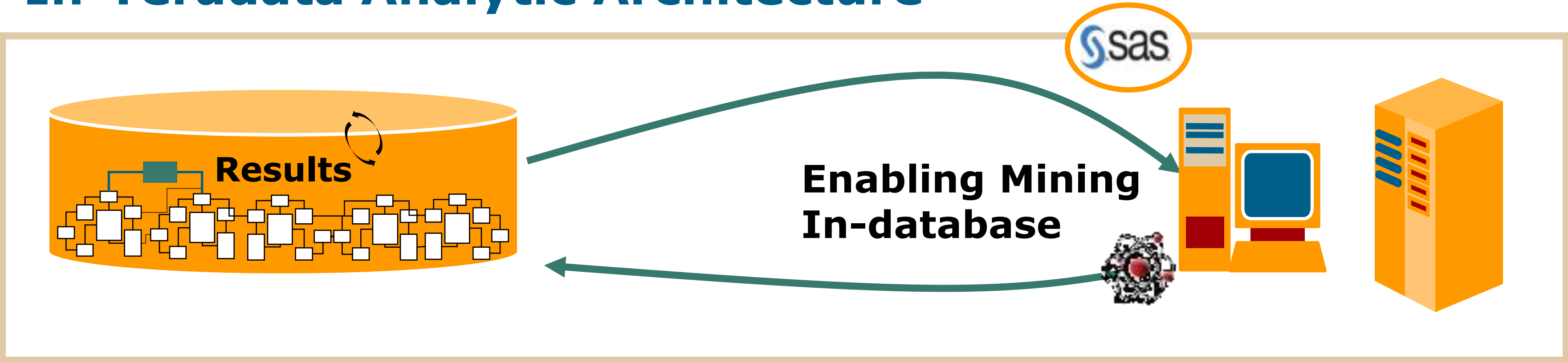
Desktop and Server Analytic Architecture



Sample Data

Results

In-Teradata Analytic Architecture



Results

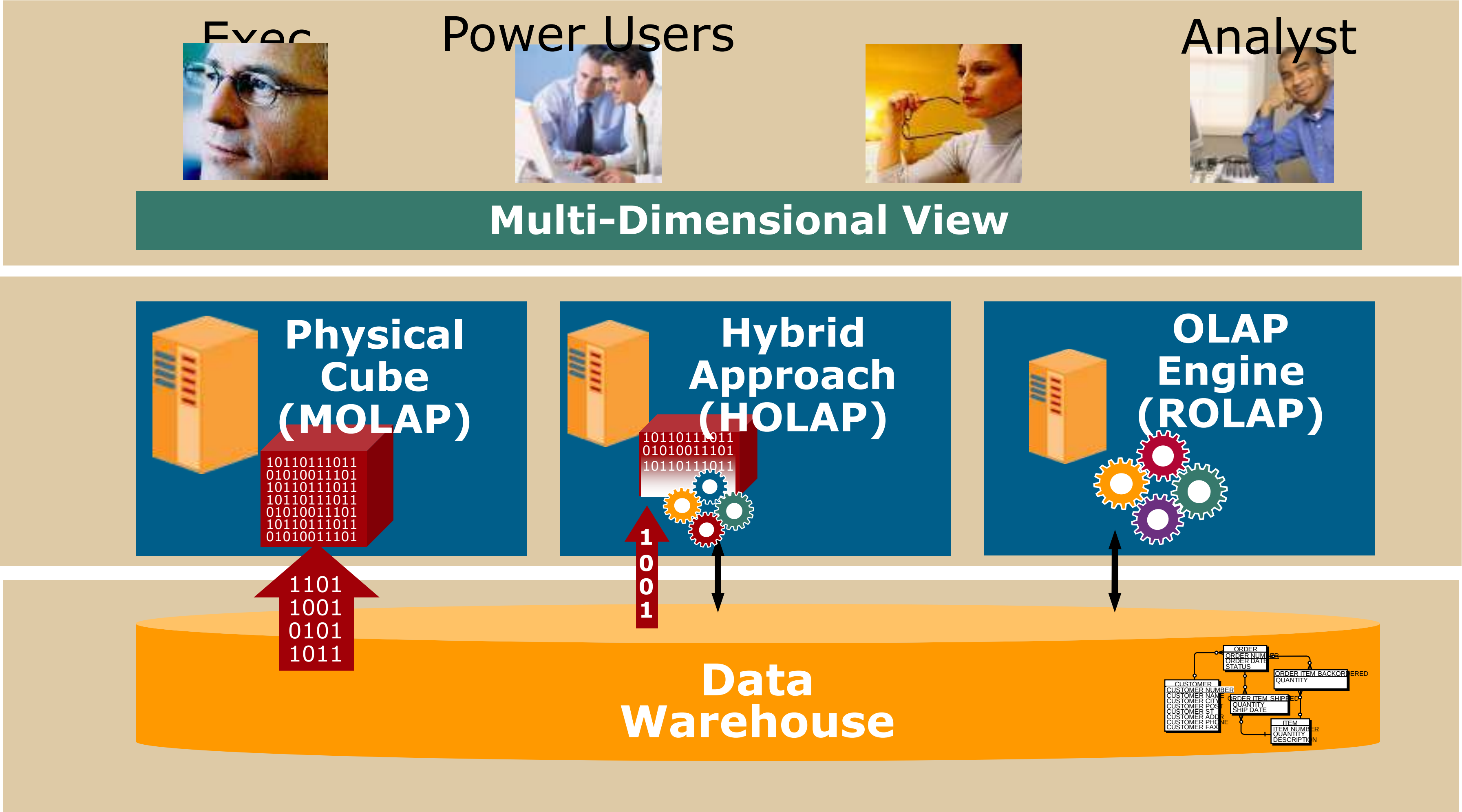
Enabling Mining In-database

Database Processing from Hours to Minutes

Data Mining Process from Days to Hours

OLAP Architectures

Various Amounts of Data Being Moved



Exec

Power Users

Analyst

Multi-Dimensional View

Physical Cube (MOLAP)

Hybrid Approach (HOLAP)

OLAP Engine (ROLAP)

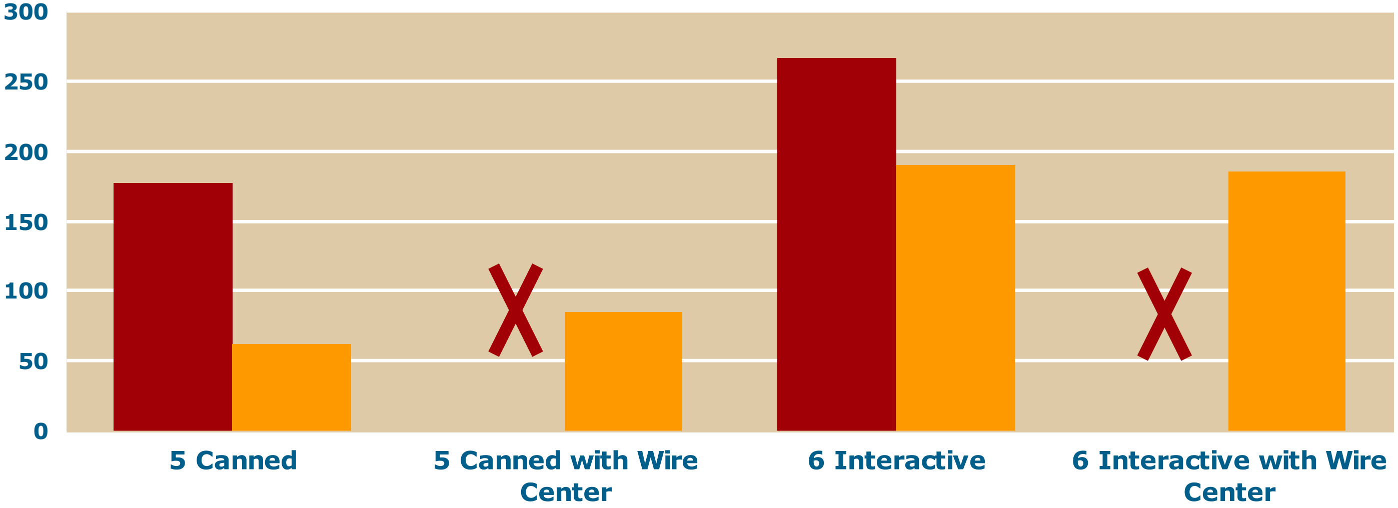
Data Warehouse

OLAP Optimization Results
Landline Communications Provider

- 38 dimensions/24 measures with 5 years of history
  - Add 39th dimension: Wire Center

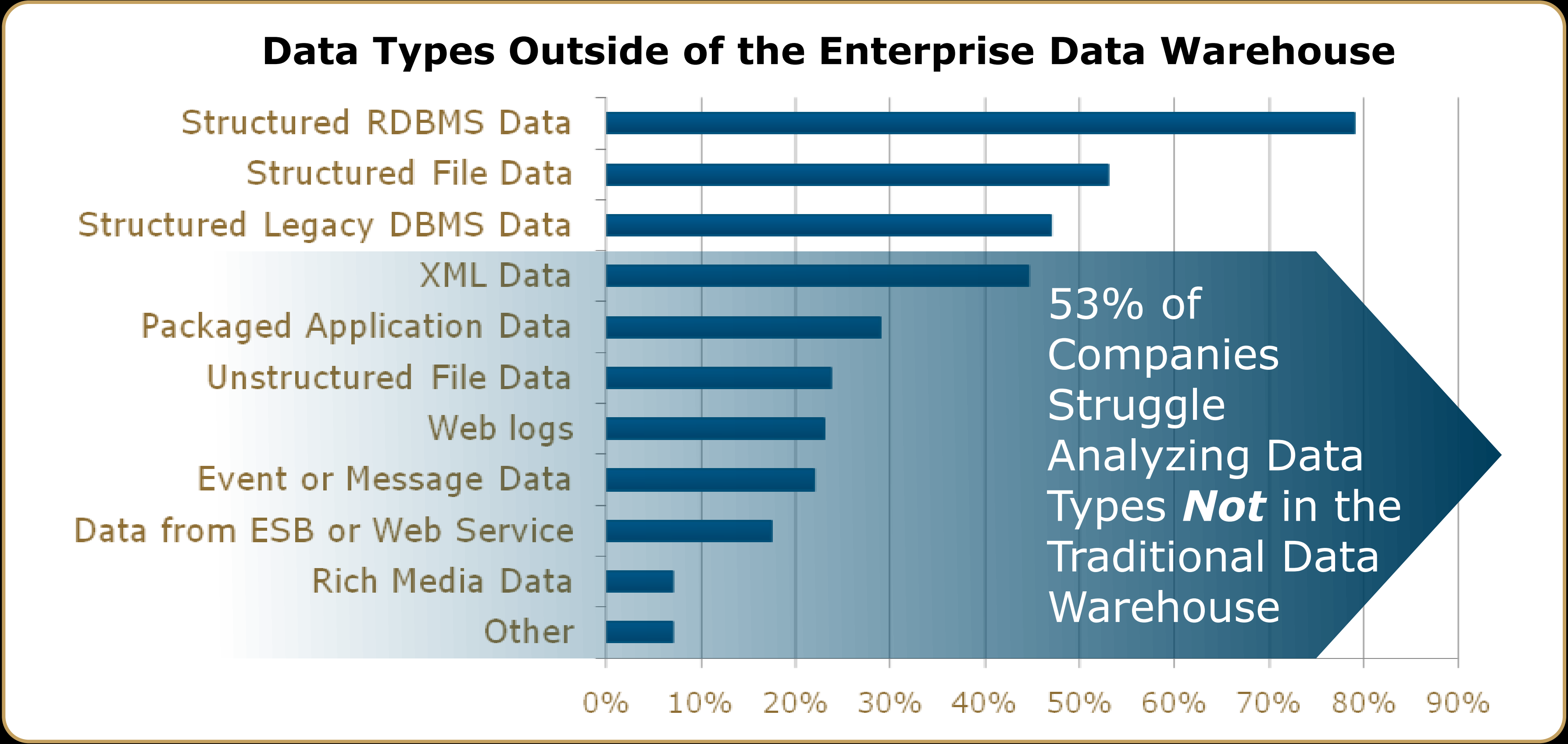
- Maintenance: **13 hours to 3 minutes**
- Cube size: **22.4 GB to <10GB**
- Detail: **Month to Daily**

Response Comparison:



Reality: New Machine-Generated Data

Non-relational and relational data outside of the EDW



53% of Companies Struggle Analyzing Data Types ***Not*** in the Traditional Data Warehouse

[†]Source: Analytics Platforms – Beyond the Traditional Data Warehouse, Survey of 223 companies. BeyeNetwork 2010

Reality: Advanced Iterative Analytics

New investigations require both standard SQL and MapReduce

- Analytics on non-relational, multi-structured, machine-generated data
- Analytics that need to scale to big data sizes
- Analytics that require reorganization of data into new data structures – graph, time & path analysis
- Analytics that require fast, adaptive iteration
- A new generation of data scientists require support for new analytic processes including Python, R, C, C++, Java & SQL.

> "In our survey 53% of respondents said they perform business analysis on data not contained within an RDBMS.
>
> **Nearly two-thirds of them were using hand coded programs."**
>
> *- Colin White & Merv Adrian, Analytics Platforms: Beyond the Traditional Data Warehouse BeyeNetwork 2010*

LinkedIn –World’s Largest Professional Networking Website

- 100+ million members across 200 countries
- A new member joins LinkedIn *every second* and 50% of members are outside the U.S.
- Executives from *all* Fortune 500 companies are LinkedIn members.
- LinkedIn’s products critically dependent on analytic-intensive algorithms for traversing the social graph, user-profile analysis

Unstructured Data

- Unstructured Schema – Multi-Structured Data
- Store as objects of any kind:
  - Key-value pair (hash table)
  - Serialized objects
  - Graph databases
  - Document files
  - Blobs
  - BigTable (GFS)

Need for SQL-MapReduce Combination

Business Question

- Determine the product, user and amount of time in which individuals…
  1. View an advertisement
  2. Possibly view other pages or advertisements before buying the product advertised
  3. Purchase that original product for which they saw the advertisement

Analytics Question

- Events exist in multiple rows in the database, for each user
- How do we attribute a purchase back to a specific ad within a 30 day period?

Manage & Analyze Multi-structured Data

Bind the structure to the data at runtime

Examples:

Raw Click Stream

- Long strings of encoded page clicks, sessions, and actions

Online Search Strings

- Entry points to a website tracked by cookie strings

Twitter , Facebook, & Other Social Network Feeds

- Social connections and influencers indicated by communication flow

New Big Data
Raw formats: Lengthy text strings, binary, blobs, social graphs
Rapid updates, data refreshes: Online click stream, stock orders, social connections/friends
High volume: Embedded processing to eliminate data movement

Examples:

Hierarchical Transactions

- e.g. One stock order split into 100s of transactions over days/weeks

Text Strings/Fields

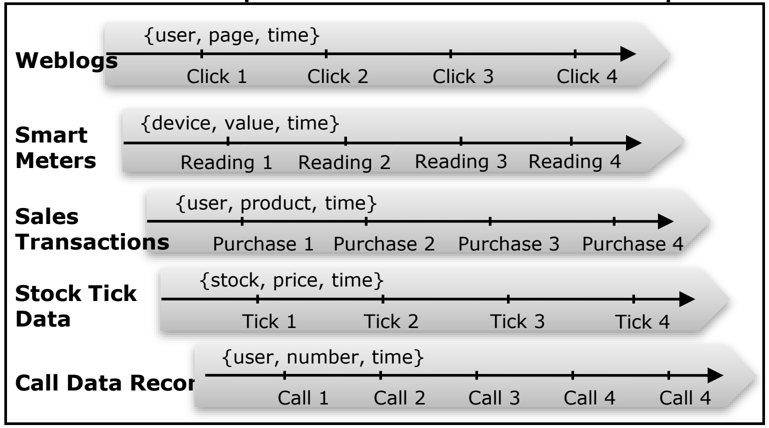
- Wide tables with *highly descriptive textual* strings

  e.g. ACH transactions, Service/Customer Support records, insurance claims

Multi-structured, Big Data Analytics

Example: Pattern Matching Analysis

Discover patterns in rows of sequential data



Weblogs {user, page, time}
Click 1 Click 2 Click 3 Click 4

Smart Meters {device, value, time}
Reading 1 Reading 2 Reading 3 Reading 4

Sales Transactions {user, product, time}
Purchase 1 Purchase 2 Purchase 3 Purchase 4

Stock Tick Data {stock, price, time}
Tick 1 Tick 2 Tick 3 Tick 4

Call Data Recor {user, number, time}
Call 1 Call 2 Call 3 Call 4 Call 5

SQL and MapReduce Approach

- Single-pass of data
- Linked list sequential analysis
- Gap recognition

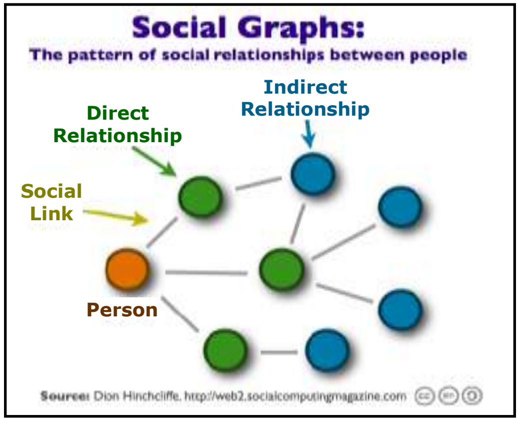
Traditional SQL Approach

- Full Table Scans
- Self-Joins for sequencing
- Limited operators for ordered data

eBiz, Media/Ent	Telecomm	Financial	Government
>Click stream Analysis >Lifecycle Marketing >Revenue Attribution	>Calling Patterns >Signal Processing >Forecasting	>Trade Sequences >Pairs Trading >Fraud Detection	>Pattern Detection >Fuzzy Matching >Inference Analysis

Example: Graph Analysis

Discover links and degree of influence between objects



Source: Dion Hinchcliffe, http://web2.socialcomputingmagazine.com

SQL and MapReduce Approach

- Single-pass of data
- Looping through all nodes

Traditional SQL Approach

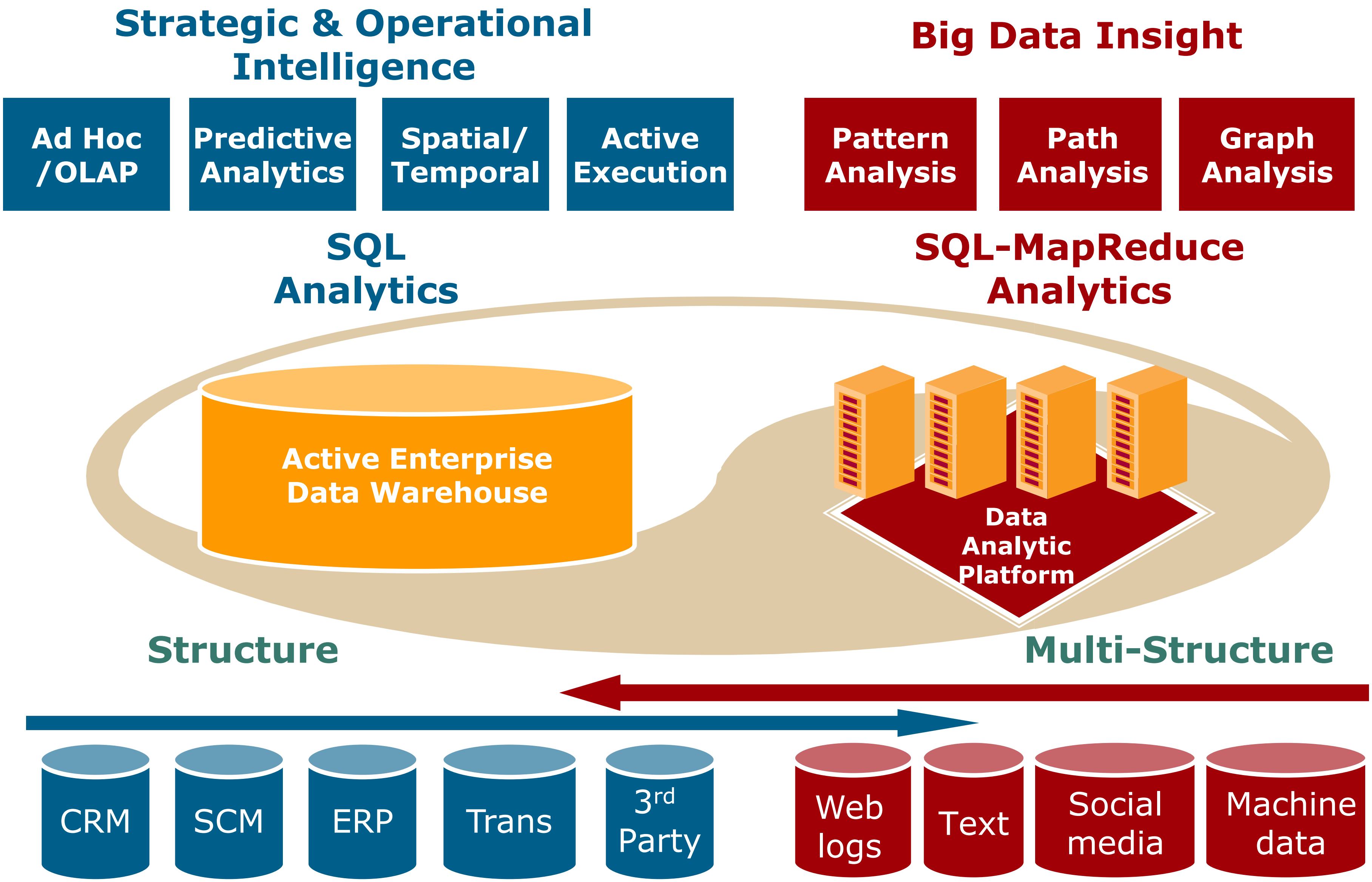
- Full Table Scans
- Self-Joins for all possible paths

eBiz, Media/Ent	Telecomm	Financial	Government
>Social Media >Crowd Sourcing >Viral Delivery >Ad optimization	>Influencers >Calling Groups >Churn Detection >Predictive Modeling	>Social Pairing >Fund Movement >Stress Triggers >Churn Detection	>Follow the Money >Collusion Detection >Pattern Matching >Network Analysis

Other Applications

- Dependency Analysis
- Traffic Analysis and Optimization
- Task Optimization
- Clustering
- Graph Mining
- Scheduling
- Routing
- Logistics
- Shortest Path
- Location Based Services
- Semantic Web ………….

Different Analytics For Different Types of Data



Strategic & Operational Intelligence

Ad Hoc /OLAP

Predictive Analytics

Spatial/ Temporal

Active Execution

SQL Analytics

Big Data Insight

Pattern Analysis

Path Analysis

Graph Analysis

SQL-MapReduce Analytics

Active Enterprise Data Warehouse

Data Analytic Platform

Structure

Multi-Structure

CRM

SCM

ERP

Trans

3rd Party

Web logs

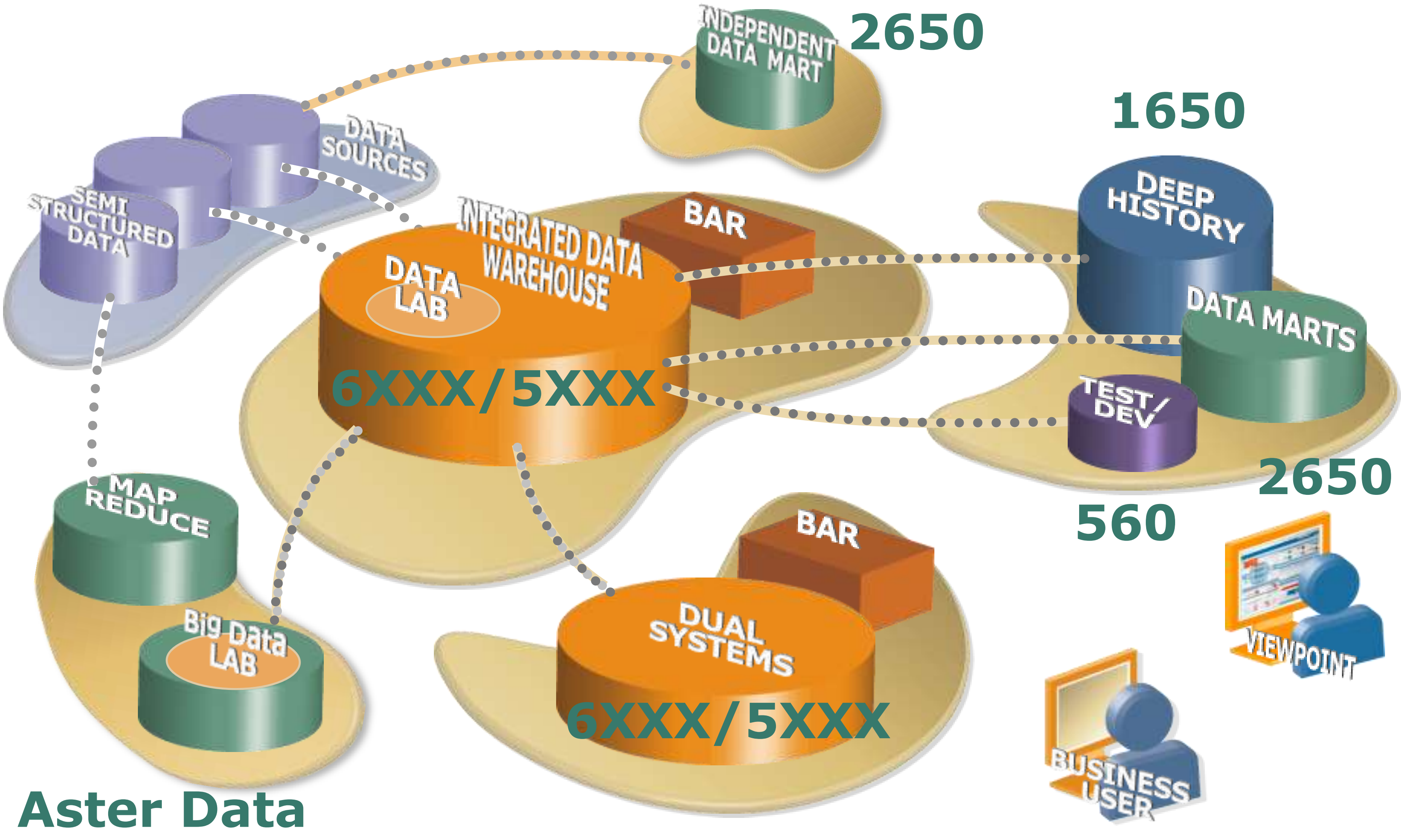
Text

Social media

Machine data

Teradata Analytical Ecosystem Overview

Flexible, Integrated Analytics



Summary

- Analytics is a competitive differentiator
- Big Issue
- Systems must be able to manage different workloads
- Mine relationship that exist in multi-structured data