eBay Analytics Platform Metadata and its Applications

September 2011

SACC2011

Agenda

- **The Born of eBay**

- **eBay Analytics Platforms**

- **Analytics Platform Metadata and Its Applications**

- **Metadata Repository**

- **Other Applications**

- **Q & A**

SACC2011

The Born of eBay . . .

Started with a <u>Broken</u> Laser Pointer . . .

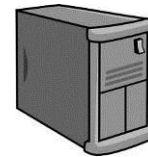**AuctionWeb was born on the Labor Day weekend in September 1995**

$30

eBay Founder

. . . sold for $14.83 USD

Pierre Omidyar

SACC2011

FREE Service Running Off from a Home Server . . .

$240 USD/month

Pierre Omidyar

SACC2011

The Born of eBay . . .
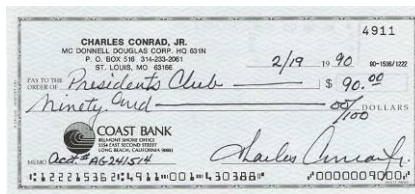
Requesting for donations . . .

Movie Tickets

Coins

Money Order

Personal Check

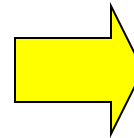**Bills**

Coupons

SACC2011
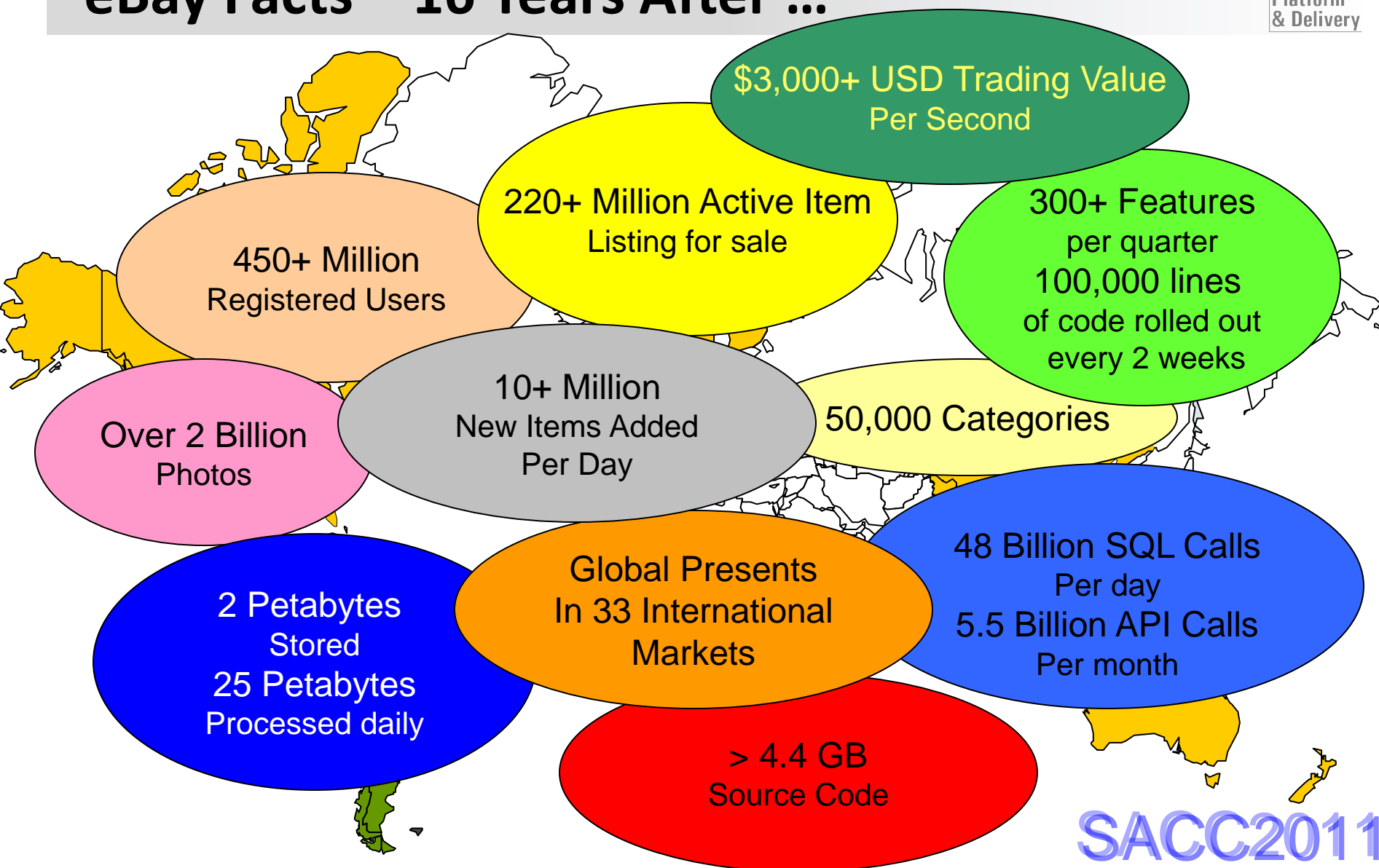
Start Profitable . . .

The Born of eBay . . .

Initial Business Model and Target Users . . .

Build equitable electronic marketplace for Americans to buy and sell their stuff

eBay Facts – 16 Years After …

$3,000+ USD Trading Value
Per Second

220+ Million Active Item
Listing for sale

300+ Features
per quarter
100,000 lines
of code rolled out
every 2 weeks

450+ Million
Registered Users

10+ Million
New Items Added
Per Day

50,000 Categories

Over 2 Billion
Photos

2 Petabytes
Stored
25 Petabytes
Processed daily

Global Presents
In 33 International
Markets

48 Billion SQL Calls
Per day
5.5 Billion API Calls
Per month

> 4.4 GB
Source Code

SACC2011

eBay Analytics Platforms

Born in Year 2000 …

From Oracle to Teradata in 2003 …

Largest Teradata Customer Infrastructure Today …

55,000 daily batch processes …

6,000+ internal users relying on our platforms …

SACC2011

eBay Analytics Platforms

500+ concurrent users

20-50 concurrent users

>5 concurrent users

Analytics & Reporting

Discover & Explore

Operational Analytics
Transactional Analytics
High volume ad hoc queries

"Compare User Activity against last year"
Trending and Forecast Analysis (large history)

Image Fingerprinting
Image Classification
Pattern Recognition
Detect Counterfeits & SNADs

Production Analytics Platform
Large Concurrent User-base

Contextual-Complex Analytics
Deep, Seasonal, Consumable Data Sets

Structure the Unstructured
Detect Patterns

Enterprise-Class System

**Deep Analytics
Enterprise-Class System**

Research System

EDW/ODW

Singularity

Primary & Secondary

Closed Loop, Active Analytics Platform

Knowledge: Integrated, aggregated, augmented

Analytical Reporting

Marketing

Trust & Safety

Customer Support

Analytics Platforms

Site Databases

www.eBay.com

- **Traffic Tracking**
- **Finding**
- **Rules engines**
- **Real time creative**
- **Advertising**
- **Fraud prevention**
- **Fake item detection**
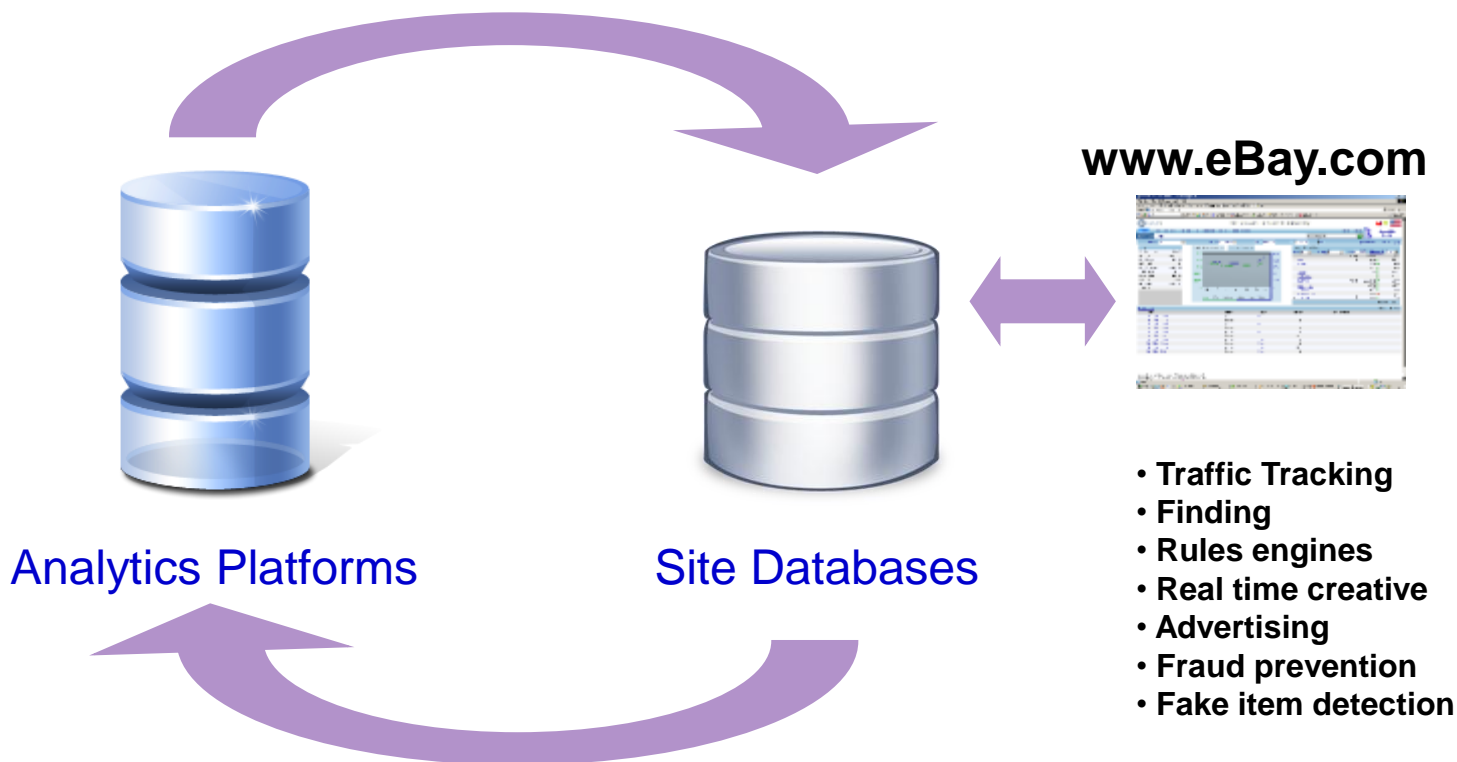
Raw data: daily, hourly feeds

SACC2011

Wisdom: informed, fact based actions

Analytics Platform Metadata

B Data Dictionary

B Logical Data Map (Source to Target Mapping)

T System Inventory

T Physical Source to Target Mapping

T Logical and Physical Data Model

T Data Definitions

T Batch Process Information

T Process Execution Information

Analytics Platform Metadata

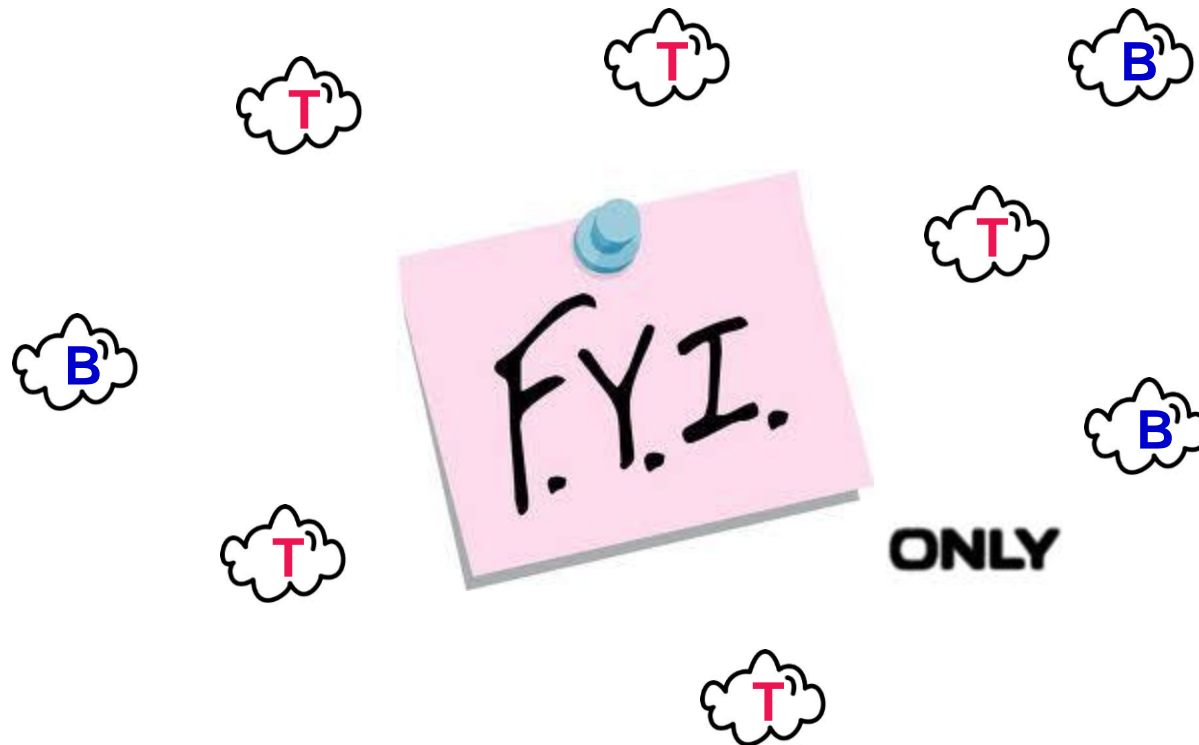**What else do we get?**

Physical Data Flow

Data Utilization

Object Dependency

System/Batch Performance

etc . . .

Typically, metadata is . . .

How does Metadata help us?

Physical Data Flow Visualization

Data Rationalization

Data Quality Monitoring

SACC2011

Application of Metadata

Physical Data Flow Visualization

Problem Statements:

Data Flow Diagram

Manual Drawing of Data Flow is Time Consuming

No Complete Set of Data Flow Diagrams

Easily be Out Dated

Manual Drawing can only provide Limited Information

Accuracy not Guaranteed

SACC2011

Application of Metadata

Physical Data Flow Visualization

Data Flow Diagram

The Data Flow Visualization tool is an automated solution to generate Data Flow Diagrams (DFD) for all Analytics Platform tables.

The DFD shows how data is being flowed through from within the Analytics Platform production environment.

SACC2011

Application of Metadata

Physical Data Flow Visualization

Round Corner Rectangle: The upstream tables from other subject area

Blue line: Stands for the process critical path

The output table of step1, also, it is the input table of step2

Step2: the step number is ordered by the job start time

Set Background as gray to highlight the target table of the diagram

The script(job) name to populate the table in the step

Job Start/End Time(HH:MM:SS)

```
14:05:22--14:06:09
B_COMM
WORKING.STG_COMM_FORUM_ACTV_W

02:41:21--02:41:27
B_USER
GDW_TABLES.DW_USERS_INFO

STEP 1
14:13:03--14:19:17
dw_comm.dw_comm_forum_actv_w.ins.sql
WORKING.DW_COMM_FORUM_ACTV_W

12:52:04--12:52:21
B_COMM
GDW_TABLES.DW_COMM_FORUM_TYPE

STEP 2
15:24:36--15:28:12
dw_comm.dw_comm_forum_vstr_actv.ins.sql
GDW_TABLES.DW_COMM_FORUM_VSTR_ACTV
```

SACC2011

Data Flow Diagram

Application of Metadata

Physical Data Flow Visualization

What questions can the Data Flow Diagram answer:

Where is the source?

How data gets flow into target?

Which SQL statements?

What are the start time and the end time?

When does a target table be ready?

What is the critical path?

Application of Metadata

Physical Data Flow Visualization

What do we use the Data Flow Information for?

Unusual delay of table readiness.

Unusual run time of SQL execution

Data Flow critical path change.

Failure down stream impact analysis.

Better view on business data analysis.

Etc . . .

SACC2011

Data Flow Diagram

Application of Metadata

Data Flow Visualization Tool User Interface
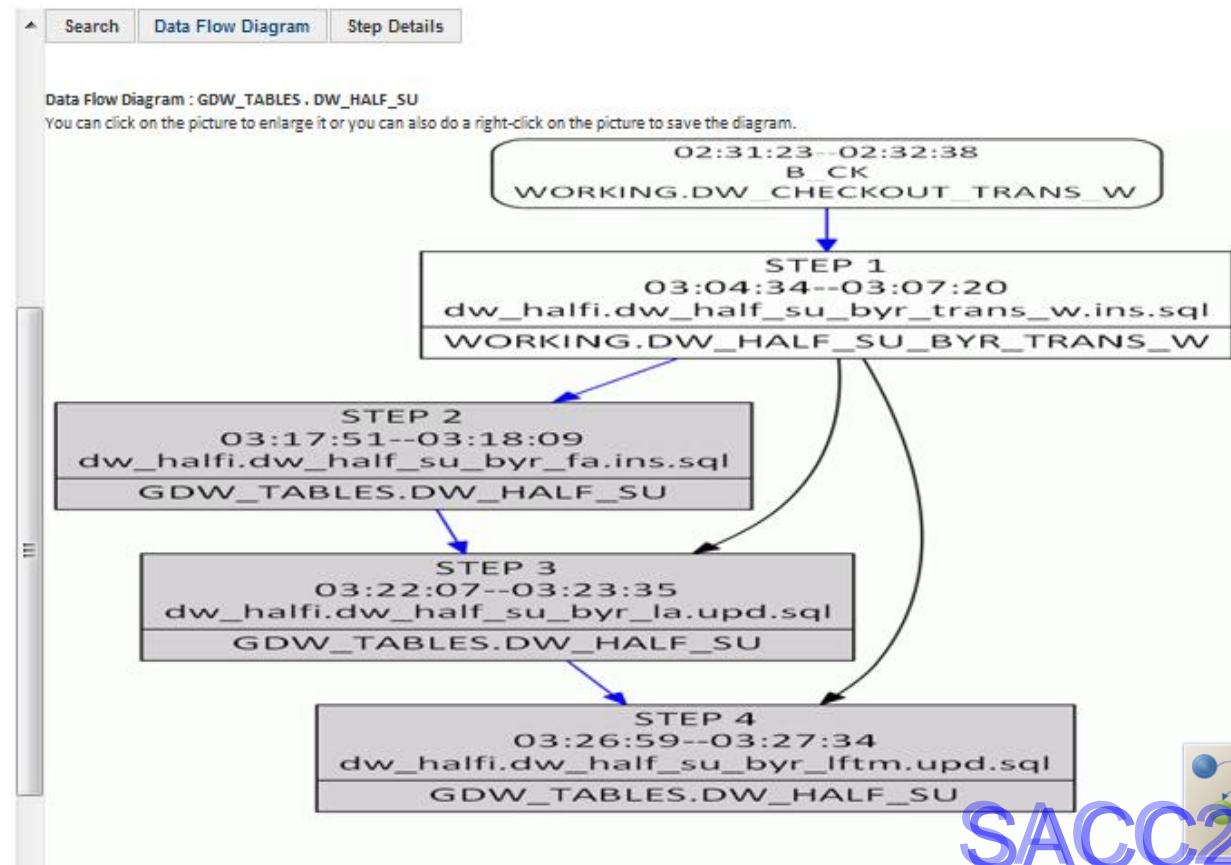
Application of Metadata

Data Rationalization

Problem Statements:

System becomes running out of space

Batches running slower and slower

Risk of missing business SLA

Takes longer on accessing data on the system

Lose end user satisfaction
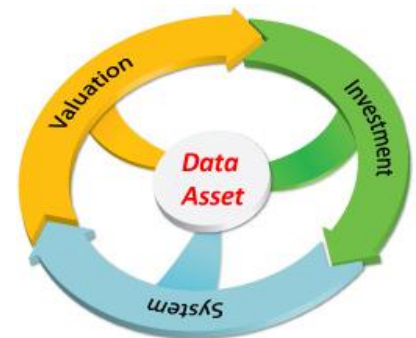
SACC2011

Application of Metadata

Data Rationalization

Data rationalization leverages the technical metadata to determine the optimum collection of data to provide the greatest business benefit to the business user.
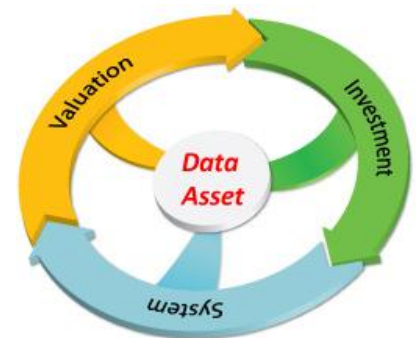
Application of Metadata

Data Rationalization

Benefits

- ✓ Reduce in overall IT expenditures
- ✓ Freeing up valuable resources
- ✓ Extend the life of value-generating system
- ✓ Enhance User Experience
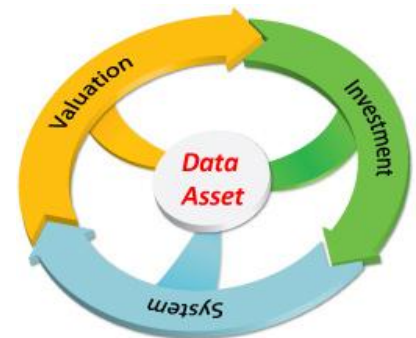
SACC2011

Application of Metadata

Data Rationalization – Table Usage Metadata

The Table Usage Metadata (TUM) collects the table usage and consumption information for analysis and rationalization purposes.
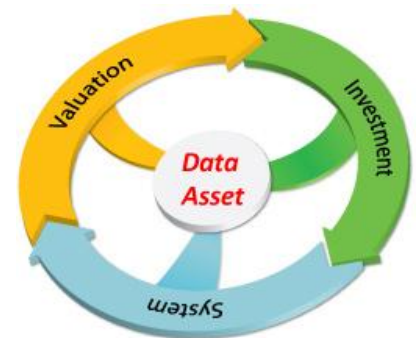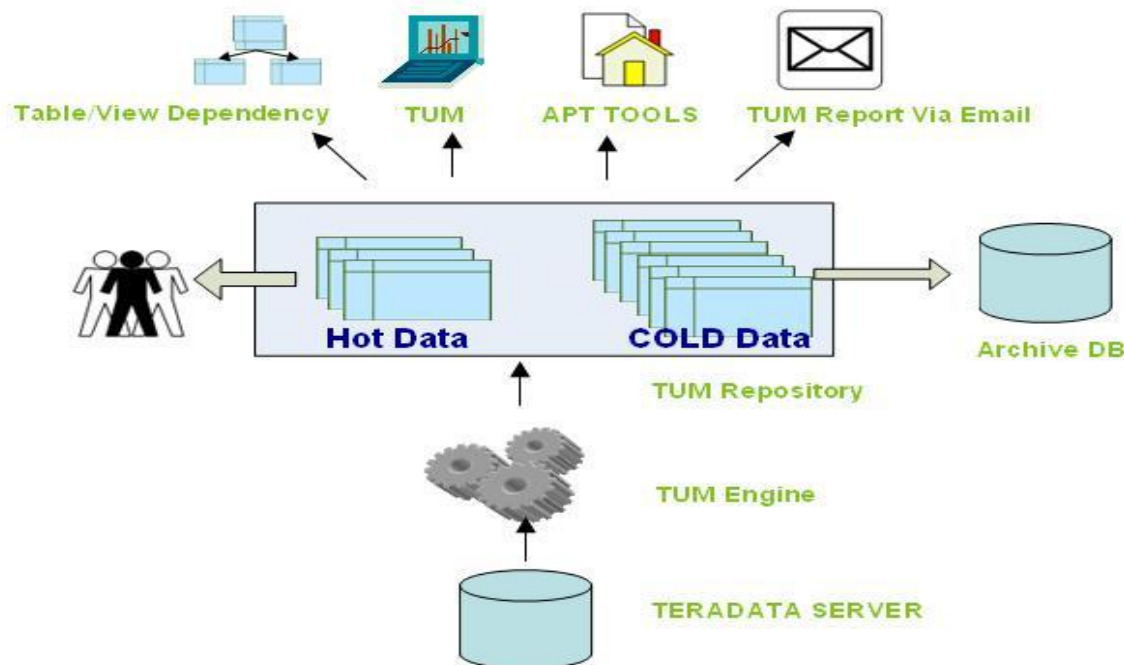
SACC2011

Application of Metadata

Data Rationalization – Table Usage Metadata

Table Basic Info

Table Size
Table Skew Factor
Table PI Info

Table Access Info

Refresh Frequency
Average Completion Time
End User Access Frequency
Column Access Info

Table Usage Info

CPU Cost (both population and consumption)
Downstream Batch Hit Counts
End User Hit Counts

Table Utilization Index

Valuation
Investment
System
Data Asset
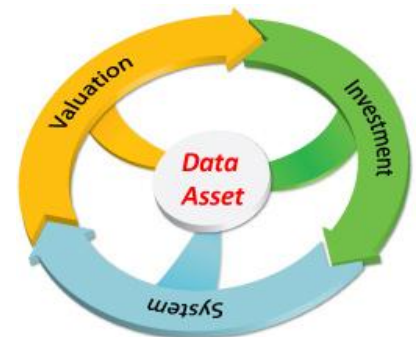
SACC2011

Application of Metadata

Data Rationalization

Typical Approaches

✓ Table Retirement

✓ Loading Frequency Reduction

✓ Table Retention by Removing Old Data Records

✓ Table Retention by removing End of Life Data Elements

SACC2011

Application of Metadata
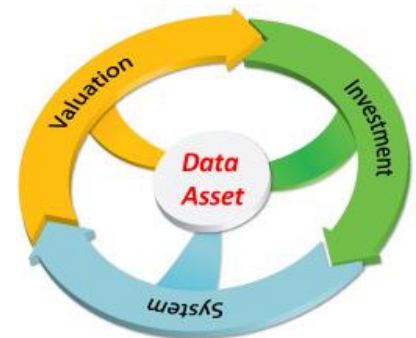
Data Quality Monitoring

Problem Statements:

Data Quality from Source cannot be guarantee

Data Quality issues usually discovered by end user

Data Quality issues cannot be discovered immediately

Data Quality issues cost a large amount of effort to fix

Lose customer satisfaction on data quality

SACC2011

Application of Metadata

Data Quality Monitoring

Data Quality Monitoring tool (DQ Firewall) is designed to monitor data quality on the system close to real time. Issues can be monitored and prevented by leveraging business metadata or customized data quality rules.

SACC2011

Application of Metadata

Data Quality Monitoring

Benefits

✓ Discover Data Quality issues immediately when occur

✓ Prevent downstream impact when issues are discovered

✓ Enhance user experience on Data Quality

✓ Improve data accuracy on the system

SACC2011

Application of Metadata

Data Quality Monitoring

How does it Work?

Data Element Metadata

Data Type

Data Allowable Domain

Referential Integrity

Constraints

Defaults

Monitoring Engine (DQ Firewall)

Alerts and Automated Actions

SACC2011

Data Model Management

Data Model Management by Metadata Repository (MDR) Tool

Subject Areas and Tables

Subject Area List

SA Code	SA Name
ABTM	About Me Pages
ACCT	Accounts
ACCTB	Account Balance
ACCTB_SG	—
ACCT_LINK	Account Linking
ACCT_SG	—
ADJ	Adju

Subject Areas Information Web User Interface

There are **580** tables and views Export entire list

22 Database	86 Tables
ACCESS_VIEWS	ACCT_QLFD_SLR_BILLED_DSCNT_SM
BATCH_VIEWS	ACCT_QLFD_SLR_DSCNT_CNTRCT
CATY1	ACCT_QLFD_SLR_UNBLD_DSCNT_SM
CATY2	ADJ_TRANS_DESCR
CATY3	BMF_TRANS_DESCR
CATY4	CNTRCT_TYPE
CATY5	CNTRY_LSTG_SITE_ELGBL_DSCNT
CATY6	DSCNT_PLAN
CATY7	DSCNT_RATE

SACC2011

Data Model Compliance

Data Model Metadata Repository and Compliance Report
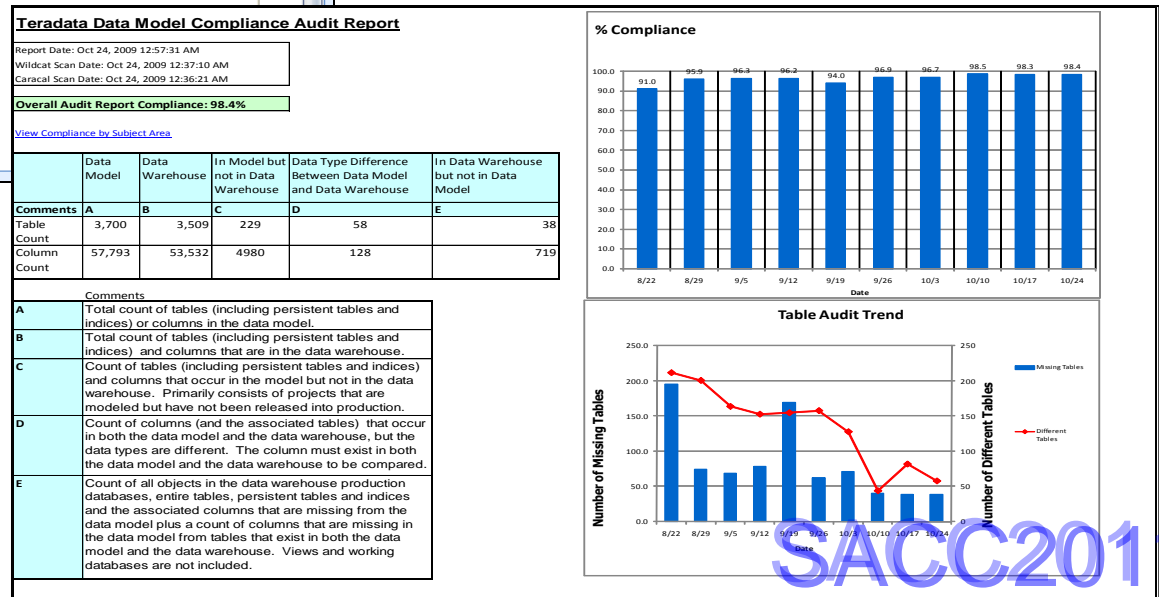
Data Model Domain & Subject Areas
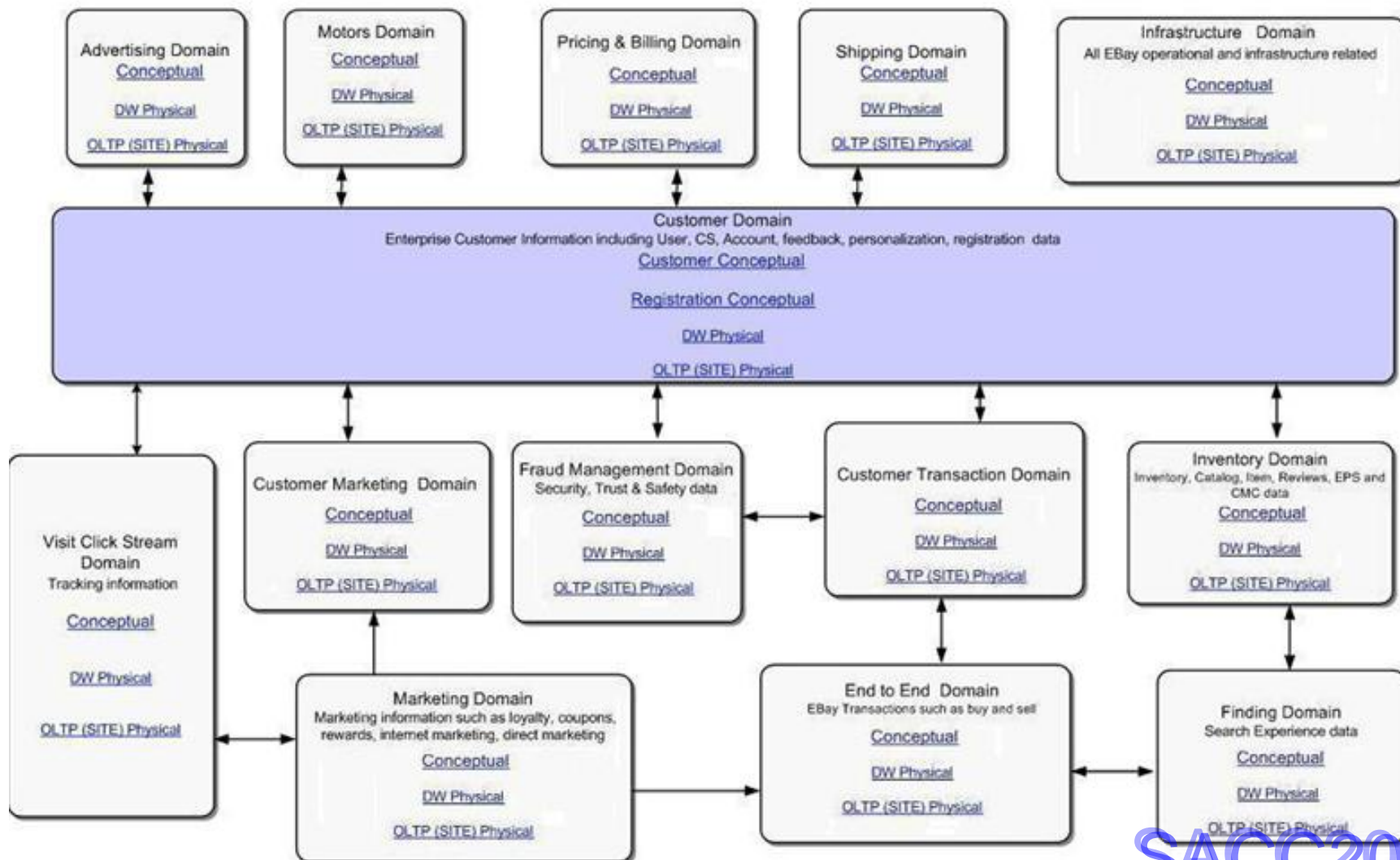
Other Applications

- ETL Performance Monitoring

- Product Return on Investment Management

- Query Pattern Recognition

- Analytics Platform User Behavior Analysis

- ETL Problematic Path Analysis

- Others

SACC2011

Metadata is NOT . . .

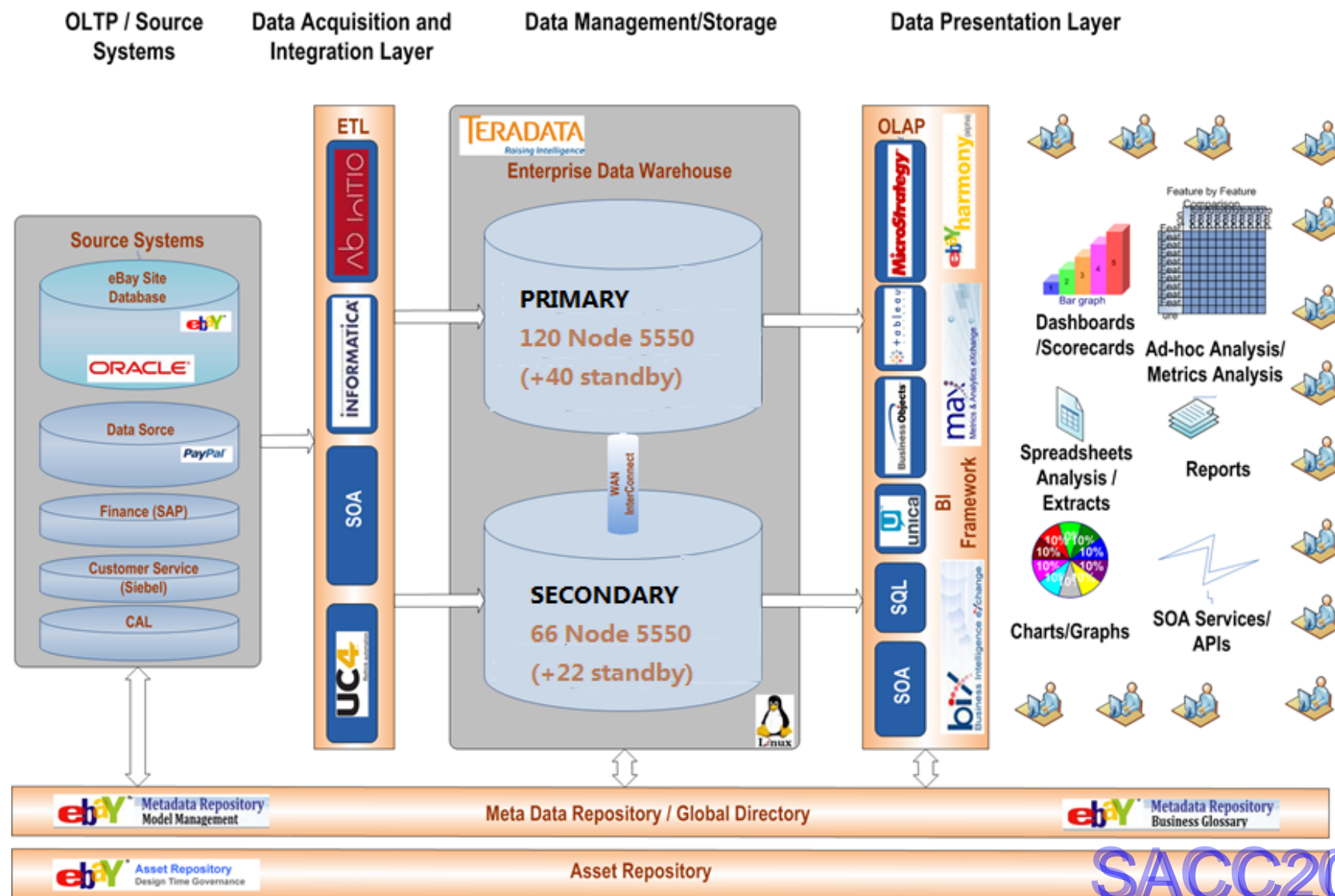**Operational Applications can be designed to rely heavily on Metadata to us help managing our Analytics Platforms . . .**
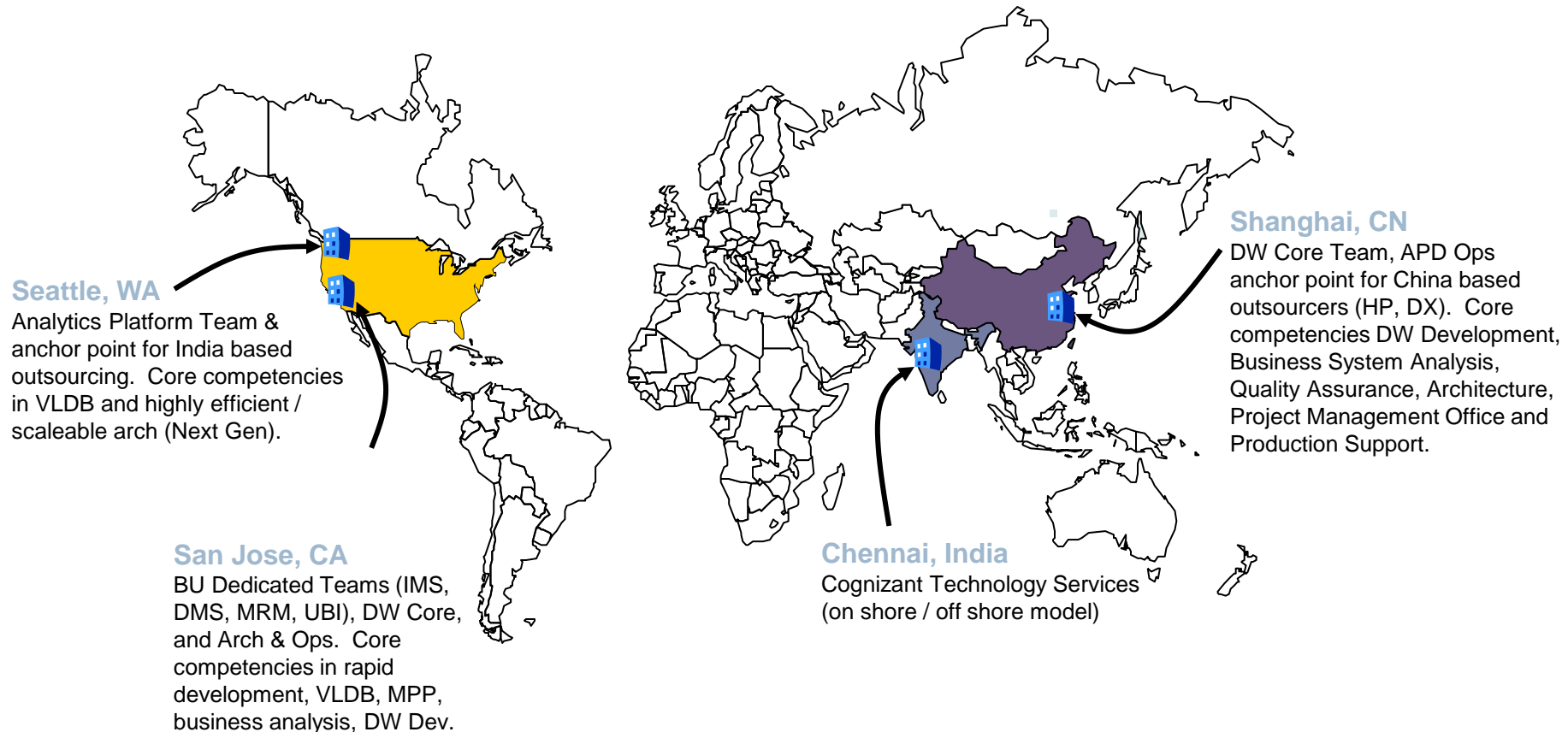
SACC2011

Q & A

SACC2011

Analytics Platform Architecture

APD – Global Team Locations

Seattle, WA
Analytics Platform Team & anchor point for India based outsourcing. Core competencies in VLDB and highly efficient / scaleable arch (Next Gen).

San Jose, CA
BU Dedicated Teams (IMS, DMS, MRM, UBI), DW Core, and Arch & Ops. Core competencies in rapid development, VLDB, MPP, business analysis, DW Dev.

Shanghai, CN
DW Core Team, APD Ops anchor point for China based outsourcers (HP, DX). Core competencies DW Development, Business System Analysis, Quality Assurance, Architecture, Project Management Office and Production Support.

Chennai, India
Cognizant Technology Services (on shore / off shore model)

SACC2011