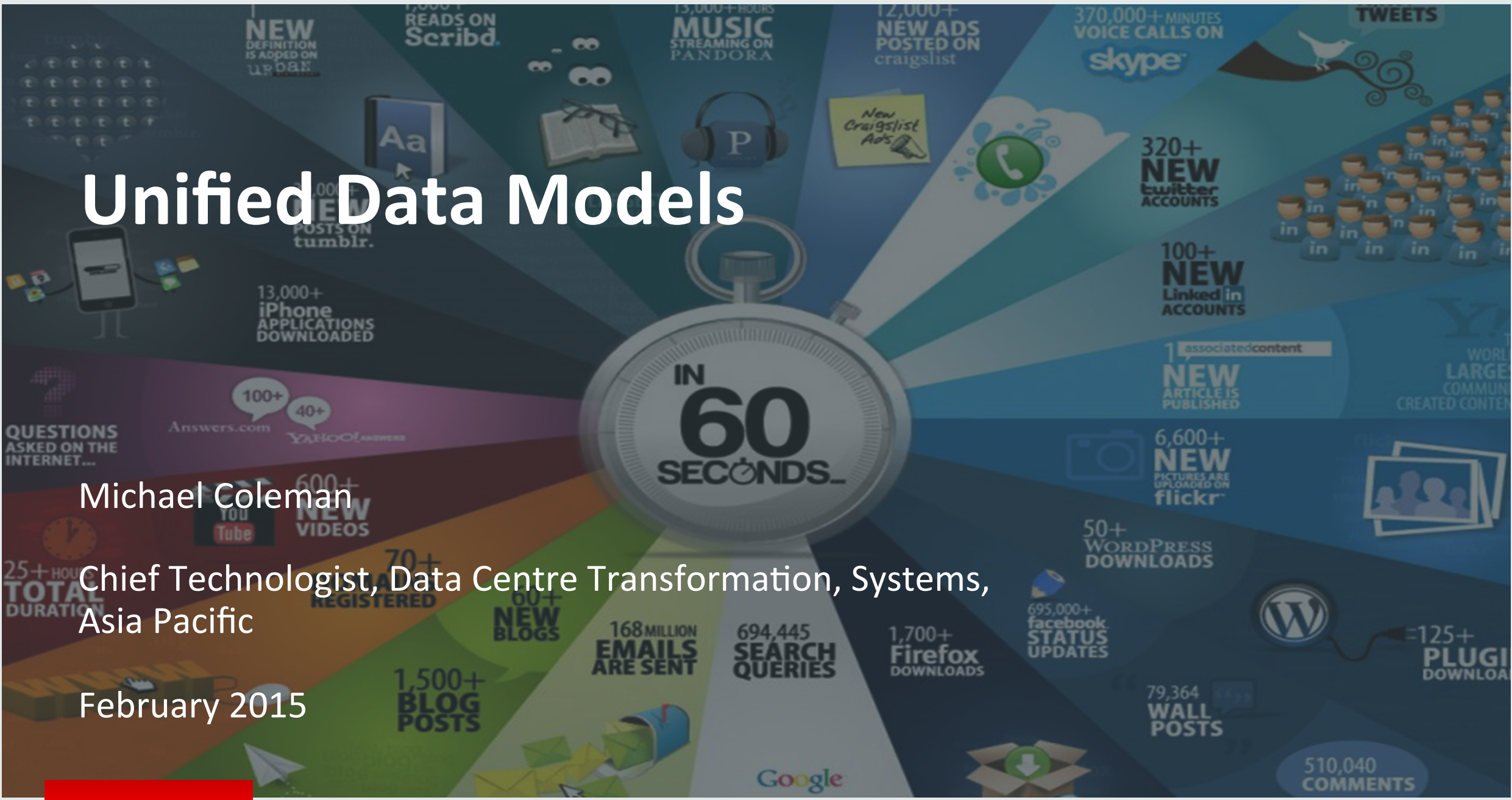# Unified Data Models

Michael Coleman

Chief Technologist, Data Centre Transformation, Systems, Asia Pacific

February 2015

# Predictions 2015 in a Snapshot

US$15 billion of government funding in 2015 will turn ICT plans into battlefield innovators.

60% of enterprises in 2015 will structure IT into core vs. LOB IT.

The software-defined battle lines will get defined in 2015.

The agile development team will be in high demand in 2015.

Consumers will begin to buy IoT from (retail) brands in 2015.

3rd Platform mash-ups will go on steroids in 2015.

The digital experience in 2015 will separate market leaders from followers.

Personalized businesses will thrive on disruptive platforms in 2015.

Data whisperers will emerge in 2015.

Service providers will aggressively compete by leveraging Big Data and analytics in 2015 to challenge for world dominance.

# Prediction #1: The Government Hi-Tech Play

**US$15 billion** of government funding in 2015 will turn ICT plans into battlefield innovators.

*Open innovation will accelerate tech+ partnerships.*

# The Battlefield Innovators

• School of the future
• Target parents
• College fundraising

Analytics and smart education

• Smart healthcare
• Productivity-led
• Holistic patient care and journey

IoT and smart healthcare

• BDA investment program
• Customer privacy act
• Data sovereignty

Data-as-a-strategic asset

• Go-to-market (GTM) accelerator program
• ISV and device "craftsmanship"
• Innovative partnerships/ GTM channels

Start-up accelerators

• Surveillance and border security
• Emergency response
• Disaster management

Public safety and order

• 3rd Platform mash-ups will accelerate cross-functional initiatives
• VC funding, SMB internationalization, open innovation

Inter-government innovation

Source: *IDC Asia/Pacific Government and Healthcare Insights, 2014*

ORACLE®

# Community Data - Governance

## ⬆ CONFIDENCE

### Ensure Information is Trusted

- How do the business users ensure the report is sourcing the correct data?
- How do the business users confirm the data is transformed correctly within each integration layer.
- How can the business and /or technical user see a full report to source data lineage diagram?

## ⬇ RISK

### Impact of Change

- How does the technical user assess the impact of modifying a source table?
- How does the technical user assess the impact of modifying an ETL Mapping?
- What is the current data governance approach for data integration?

## ⬆ COLLAB

### Searchable, Shared Metadata

- How assessable is the metadata?
- Can business and technical users work from one definition of metadata?
- Can business and technical users easily share metadata as well as views of metadata?

# Enterprise Metadata Management

## Solves a significant pain point for a wide variety of business consumers and technical staff
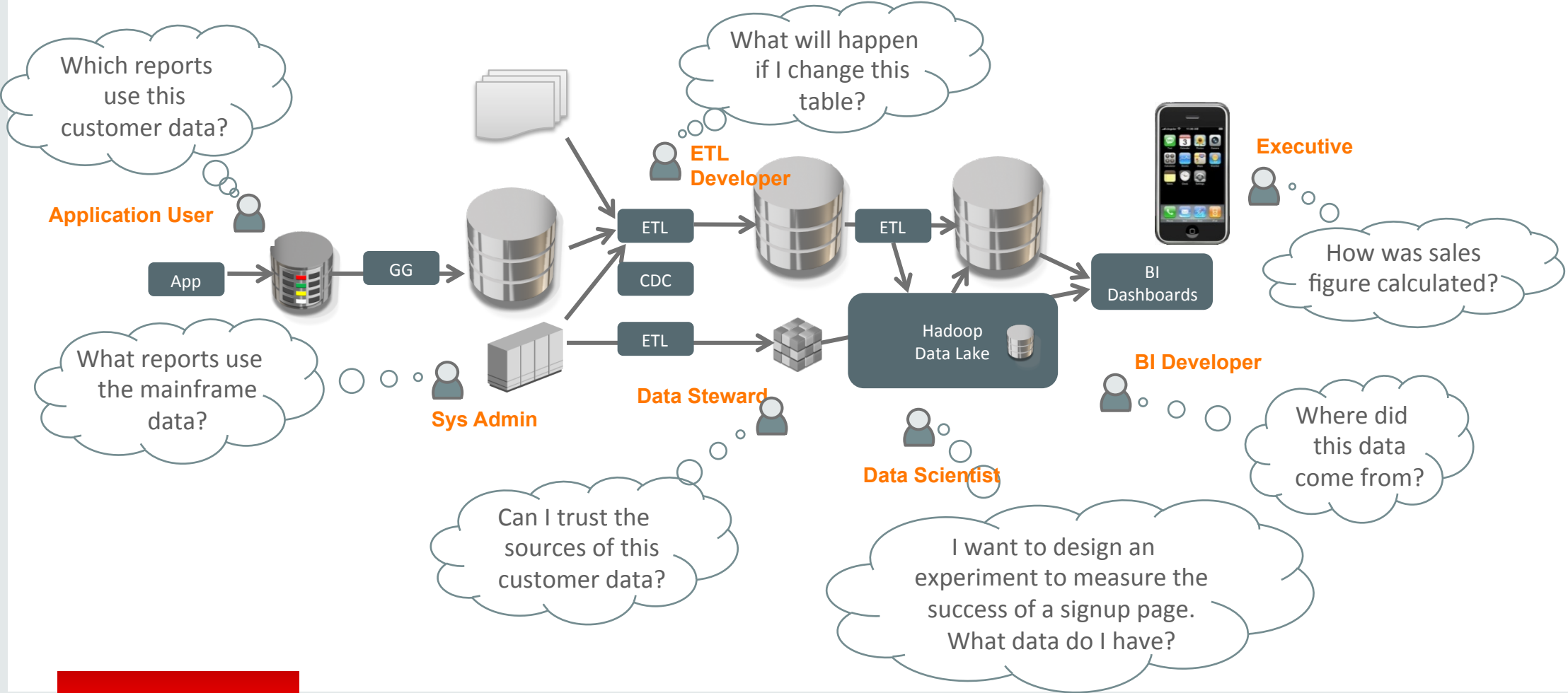
# Introduction
## Big Data sample data

- **Community Data**

(*unstructured*)

```
<tweet><oracletag>DL</oracletag><searchType>stream</searchType><text>@delta we've been sitting
in an aircraft in Atlanta for almost an hour. No one is telling us anything other than an annou
ncement 20 mins ago</text><time>12-27-2011 11:55:15</time><id>151707729151537152</id><userID>30
9762192</userID><userName>Heard at Lunch</userName><userScreenName>heardatlunch</userScreenName
><userFollowersCount>1</userFollowersCount><userFriendsCount>1</userFriendsCount><replyToScreen
Name>Delta</replyToScreenName><replyToStatusID>-1</replyToStatusID><replyToUserID>5920532</repl
yToUserID><isRetweet>false</isRetweet><retweetCount>0</retweetCount><retweetSourceID>0</retweet
SourceID<retweetSourceTime>12-27-2011 11:53:36</retweetSourceTime><retweetUserName></retweetUse
rName><retweetUserScreenName></retweetUserScreenName><retweetUserFollowersCount>0</retweetUserF
ollowersCount><retweetUserFriendsCount>0</retweetUserFriendsCount></tweet>
```

- **Community Data (JSON)**

(*semi-structured*)

```
{ "firstName": "John", "lastName": "Smith", "age": 25, "address": { "streetAddress":
 "21 2nd Street", "city": "New York", "state": "NY", "postalCode": "10021" }, "phone
Number": [ { "type": "home", "number": "212 555-1234" }, { "type": "fax", "number":
"646 555-4567" } ] }
```
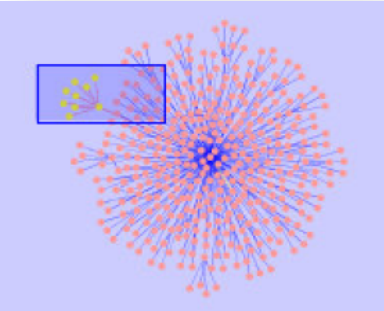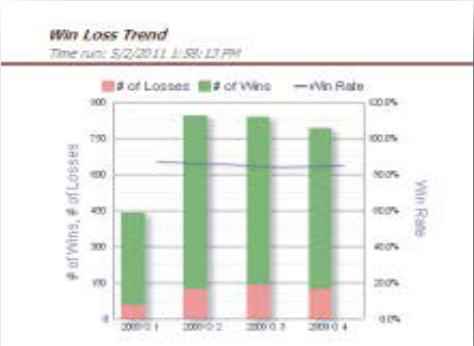
- **Yahoo stock tickers**

(*structured*)

```
UAL,20111220,20.67,20.90,20.47,20.67,2820600
UAL,20111221,20.55,20.98,20.19,20.30,4022800
UAL,20111222,20.33,20.48,20.01,20.26,2761100
UAL,20111223,19.98,20.00,18.40,19.85,10873800
UAL,20111227,19.87,20.03,18.77,19.20,5438800
```
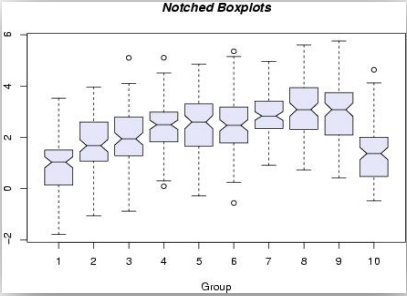
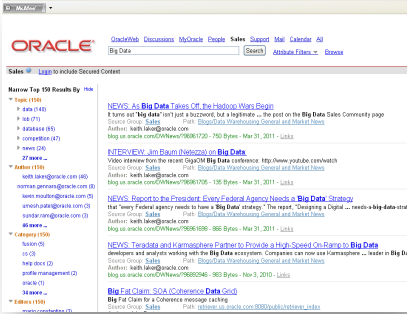# Introduction
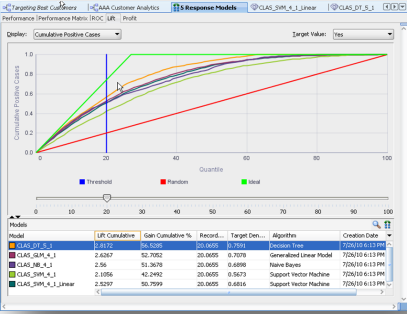## Big Data: New Analytics Capability

**Structured data**



**+**

**Unstructured data**





Graph Analytics



Advanced Statistics



Data Mining



Spatial Analytics



Text Analytics

# IoT Challenge
## Big values: previously impossible can do it now…

**AUTOMOTIVE**
Auto sensors reporting location, problems

**COMMUNICATIONS**
Location-based advertising

**CONSUMER PACKAGED GOODS**
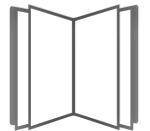Sentiment analysis of what's hot, problems
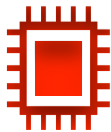
**FINANCIAL SERVICES**
Risk & portfolio analysis

**EDUCATION & RESEARCH**
Experiment sensor analysis

**HIGH TECHNOLOGY / INDUSTRIAL MFG.**
Mfg quality
Warranty analysis

**LIFE SCIENCES**
Clinical trials
Genomics

**MEDIA/ ENTERTAINMENT**
Viewers / advertising effectiveness

**ON-LINE SERVICES / SOCIAL MEDIA**
People & career matching
Web-site optimization

**HEALTH CARE**
Patient sensors, monitoring, EHRs
Quality of care

**OIL & GAS**
Drilling exploration sensor analysis

**RETAIL**
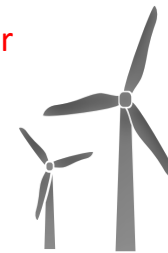Consumer sentiment
Optimized sales & marketing

**TRAVEL & TRANSPORTATION**
Sensor analysis for optimal traffic flows
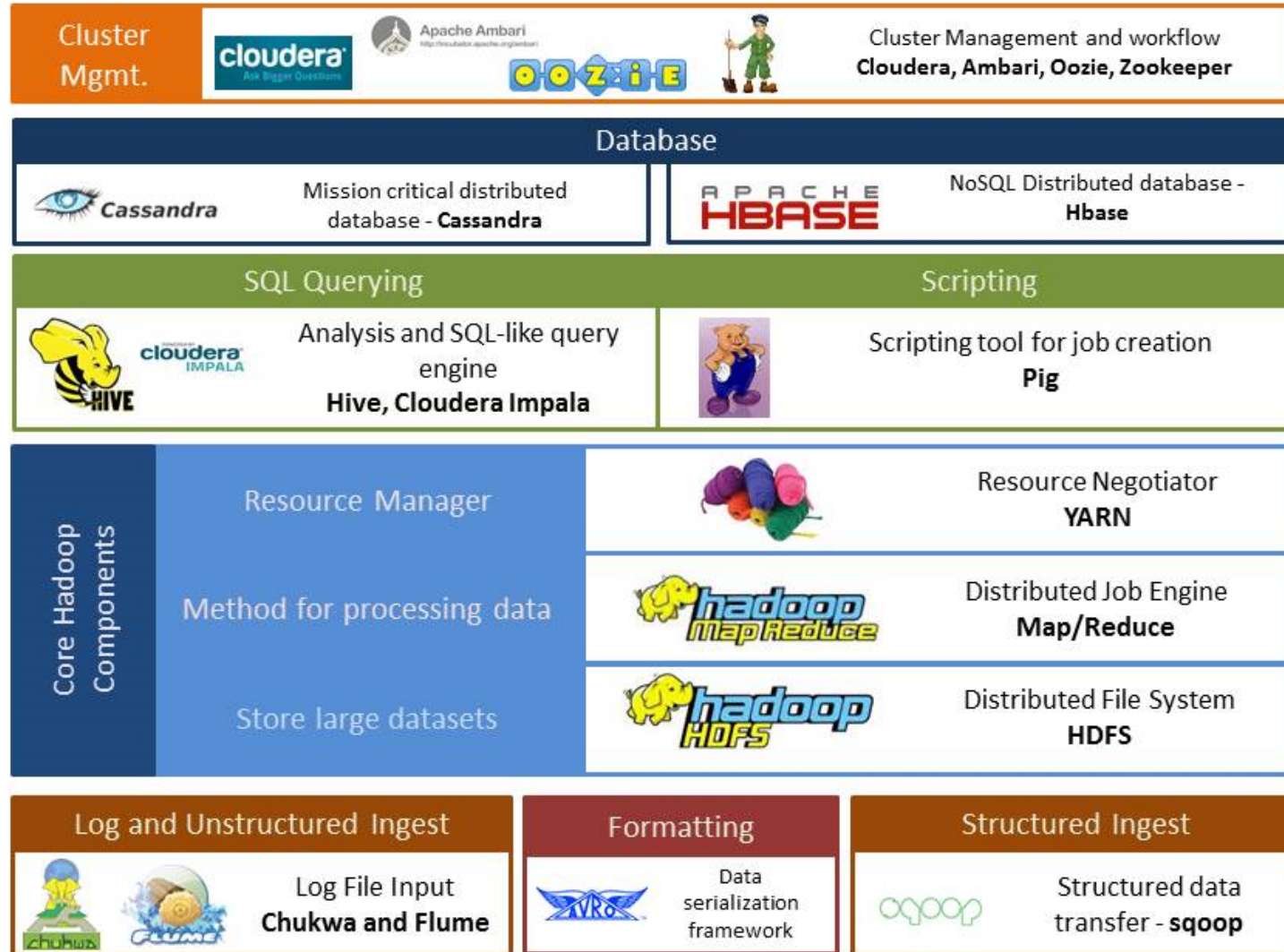Customer sentiment

**UTILITIES**
Smart Meter analysis

**LAW ENFORCEMENT & DEFENSE**
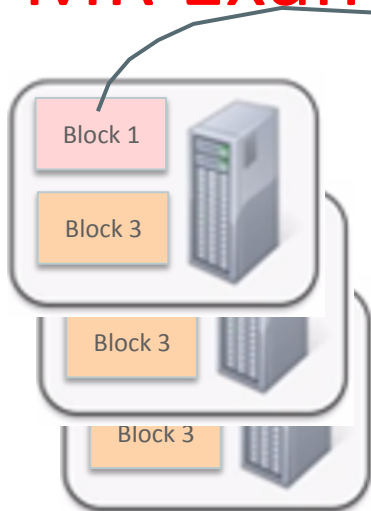Threat analysis - social media monitoring, photo analysis

# Architecture
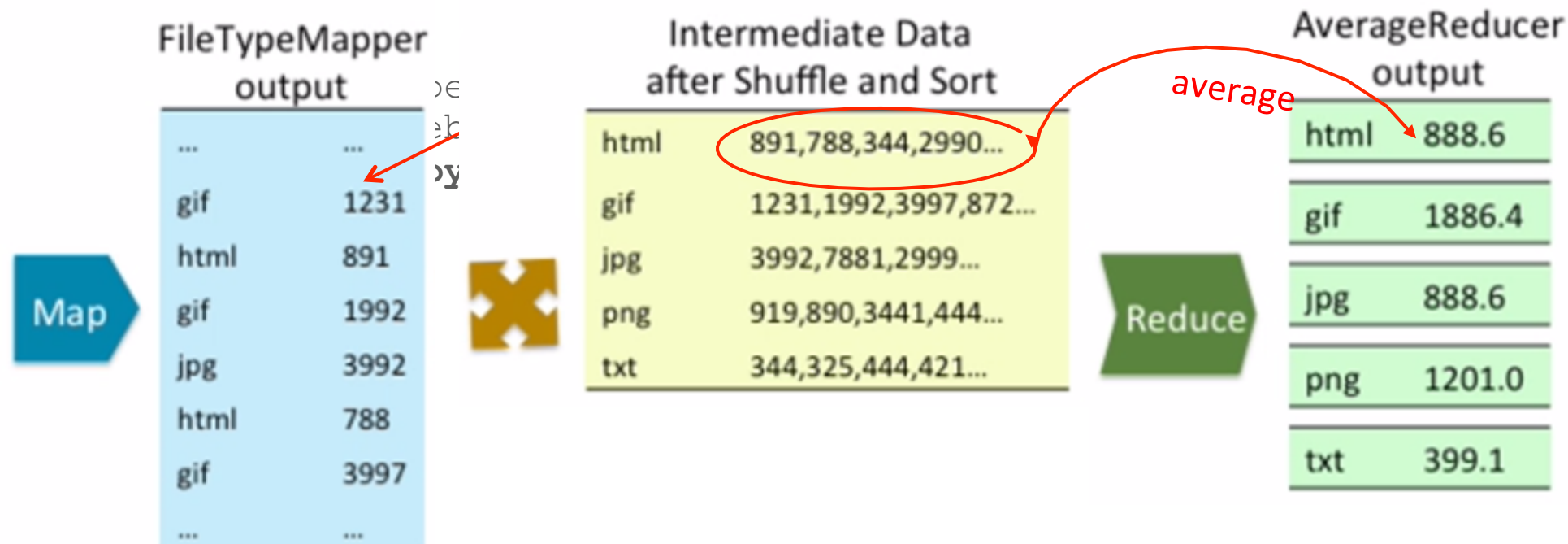## Hadoop=Apache Hadoop=Hadoop Ecosystem

# Architecture
## MR Example: average the "time-taken" from weblog



```
...
2013-03-15 12:39 - 74.125.226.230 /common/logo.gif 1231ms - 2326
2013-03-15 12:39 - 157.166.255.18 /catalog/cat1.html 891ms - 1211
2013-03-15 12:40 - 65.50.196.141 /common/logo.gif 1992ms - 1198
2013-03-15 12:41 - 64.69.4.150 /common/promoex.jpg 3992ms - 2326
...
```
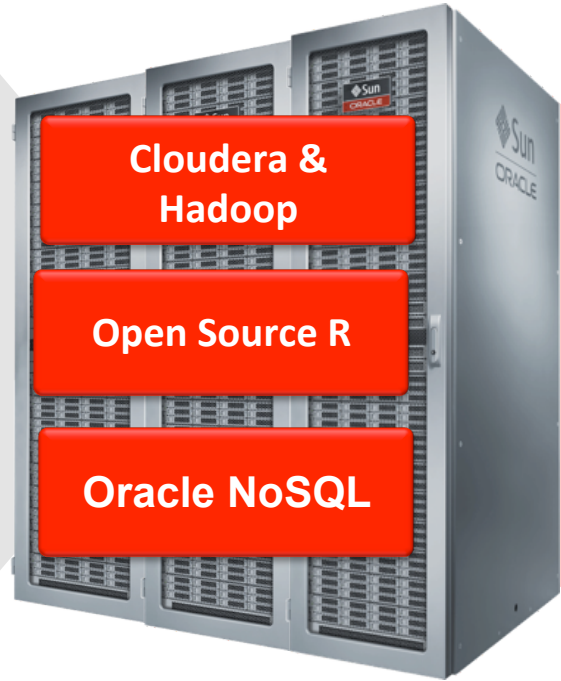
Block 1
Block 3
Block 3
Block 3

**FileTypeMapper output**

| ... | ... |
|-----|-----|
| gif | 1231 |
| html | 891 |
| gif | 1992 |
| jpg | 3992 |
| html | 788 |
| gif | 3997 |
| ... | ... |

Map

**Intermediate Data after Shuffle and Sort**

| html | 891,788,344,2990... |
|------|---------------------|
| gif | 1231,1992,3997,872... |
| jpg | 3992,7881,2999... |
| png | 919,890,3441,444... |
| txt | 344,325,444,421... |

average

Reduce

**AverageReducer output**

| html | 888.6 |
|------|-------|
| gif | 1886.4 |
| jpg | 888.6 |
| png | 1201.0 |
| txt | 399.1 |

ORACLE

# Oracle Big Data Strategy

- Big Data Integrates to Oracle Information Architecture
- SQL across all relational and unstructured big data: RDBMS+Hadoop



**Integrate**

- Scalable
- Flexible
- Cost Effective

SQL — DW & BI

SQL — Analytics

NoSQL — Web
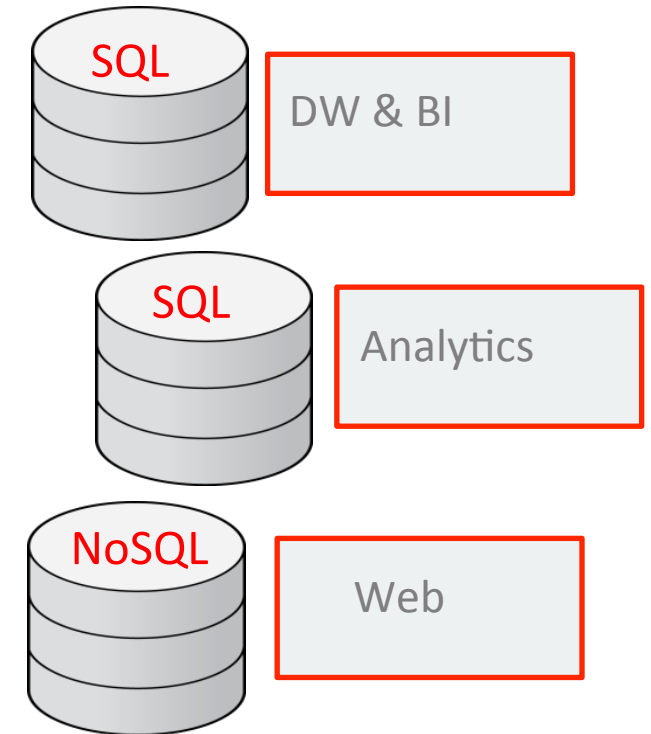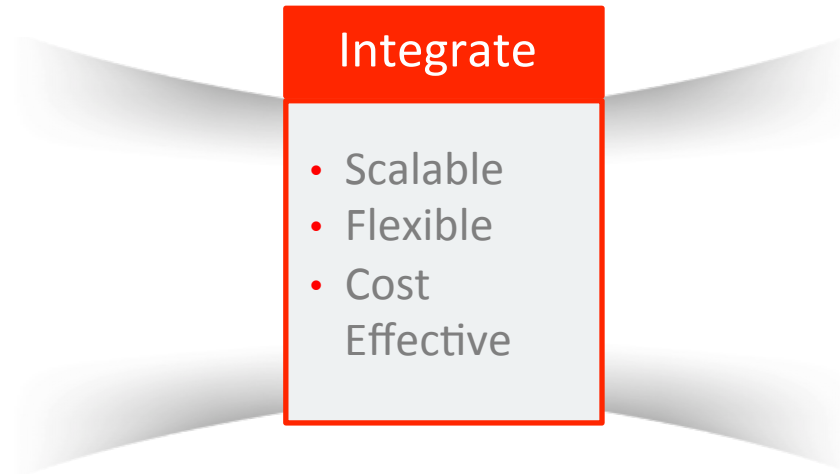
# Oracle Big Data Platform

**Oracle Big Data Appliance**
Optimized for Hadoop, R, and NoSQL Processing

**Oracle Big Data Connectors**

**Oracle Exadata**
"System of Record" Optimized for DW/OLTP

**Oracle Exalytics**
Optimized for Analytics & In-Memory Workloads

Cloudera & Hadoop

Open Source R

Oracle NoSQL

Oracle Big Data Connectors

Oracle Data Integrator

Oracle Advanced Analytics

Data Warehouse

Oracle Database

In-Database Analytics

Oracle Enterprise Performance Management

Oracle Business Intelligence Applications

Oracle Business Intelligence Tools

Oracle Endeca Information Discovery

**Stream** | **Acquire** | **Organize** | **Analyze** | **Visualize**

# Oracle Big Data SQL

**Query All Data from Oracle without Application Change or Data Conversion**

Hive metadata

HDFS Name Node

HDFS Data Node

Oracle NoSQL

**Big Data SQL**
Query all data with Oracle SQL
Smart scan in Hadoop to optimize data requests

Oracle Catalog

Hive metadata

External Table

External Table

```
create table customer_address
( ca_customer_id        number(10,0)
, ca_street_number   char(10)
, ca_state              char(2)
, ca_zip                char(10))
organization external (
TYPE ORACLE_HIVE
DEFAULT DIRECTORY DEFAULT_DIR
ACCESS PARAMETERS
    (com.oracle.bigdata.cluster hadoop_cl_1)
LOCATION ('hive://customer_address')
)
```

**Big Data Appliance**
+
**Cloudera Hadoop**

**Exadata**
+
**Oracle Database**
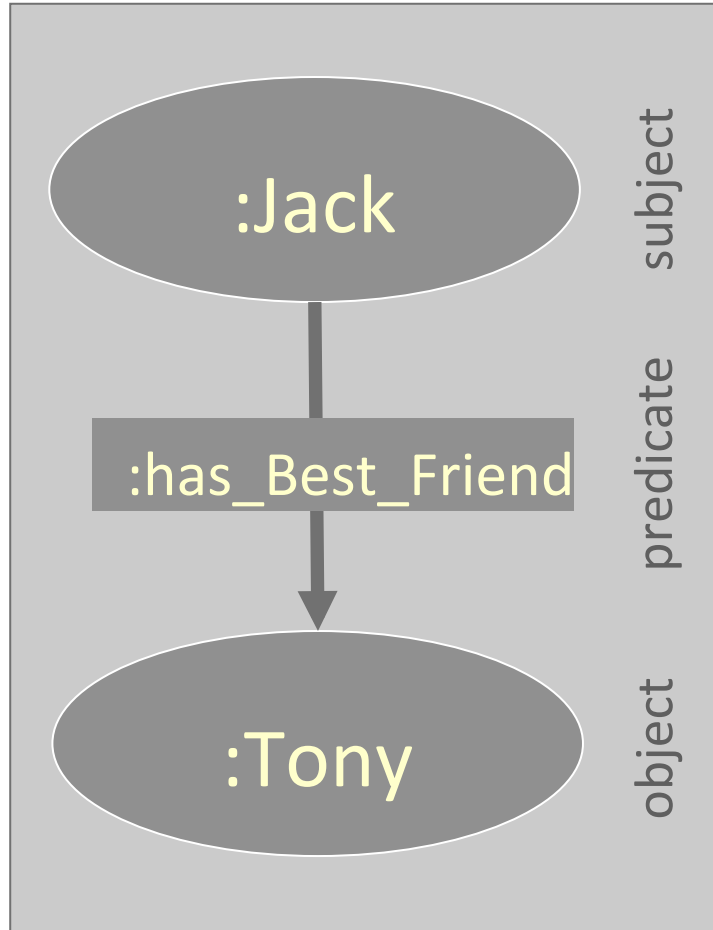
**ORACLE**

# Big Data Requirements
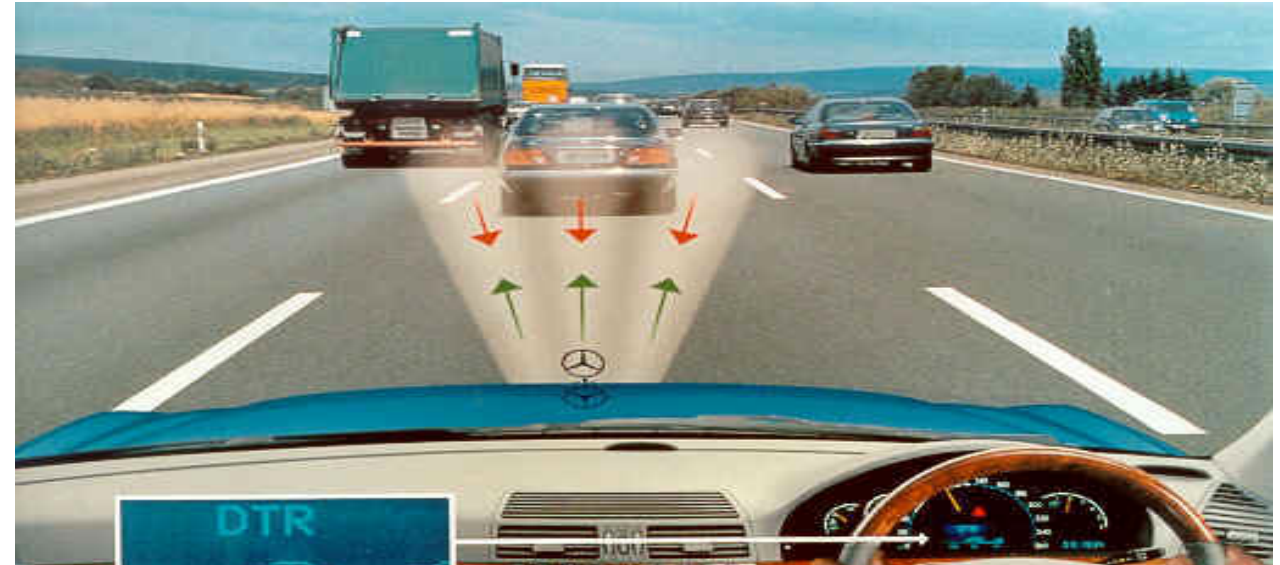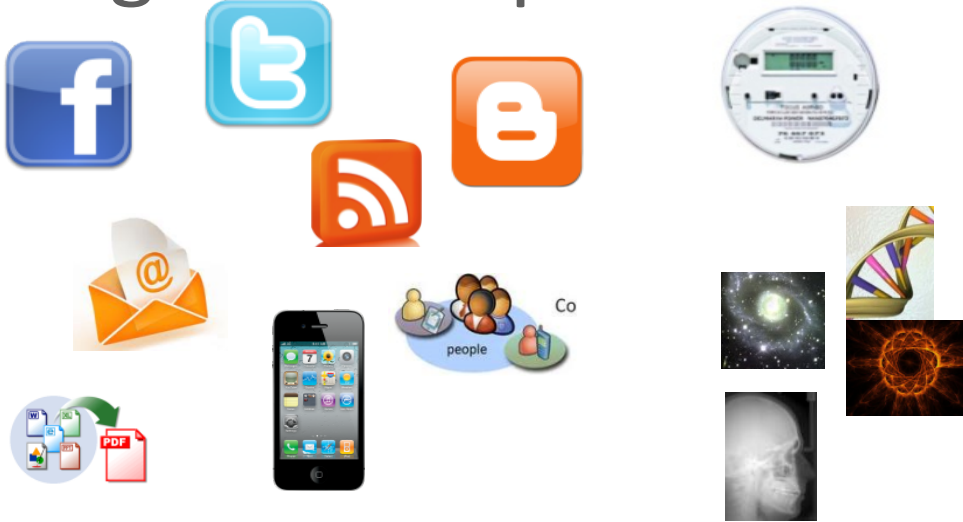


- Device-generated data
- Documents
- Location data
- Audio, Video, Image
- Social Network and Interaction Models

# What is RDF?
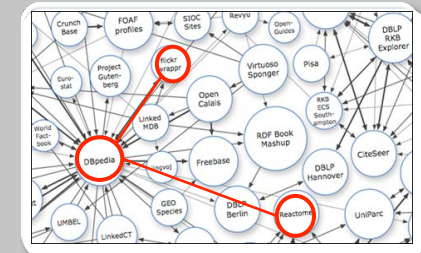## Resource Description Framework



- Basic structure is a "triple"
- RDF can be serialized into XML
- Schemas need not be specified in advance
- RDF data is fully expressible as RDBMS data And
- RDBMS data is also expressible as RDF
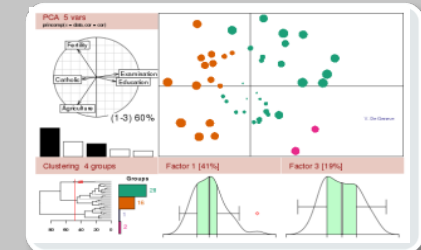
# RDF Semantic Graph
## Use Cases

**Linked Data & Public Clouds**

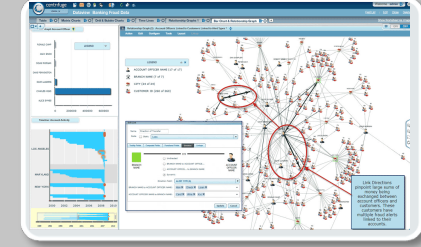- Unified content metadata model for public clouds
- Validate consistency

**Text Mining & Entity Analytics**

- Find related content & relations by navigating connected entities

**Social Media Analysis**

- Analyze content using integrated **metadata**
  - Blogs, wikis, video
  - Calendars, IM, voice

# Network Data Model Graph

## Temporal Modeling/Analysis

- Traffic Patterns
  - Record historical travel
  - Based on time of day and day of the week
- NDM can use traffic patterns to compute shortest paths
- Support NAVTEQ Traffic Patterns format out of the box

# Network Data Model Graph

## Multi-Modal Routing

- Each mode (car, bus, rail, bike, etc) modeled as a separate network

- Single logical network represents all modes of transportation

- Transition nodes where networks meet

- NDM APIs can specify the modes

- Out of the box support for transit data published by transit authorities

Time to compute geometries: 2.419s.

Analysis Result:
From: 575456205
To:   575481535

Drive/Walk to
'CONNECTICUT AV and WYOMING AV'
 (31 meters).

[1]
Board Route 227 (Inbound)
At 'CONNECTICUT AV and WYOMING AV'
Dep. Time : 10:10:42

Get down at 'NW CONNECTICUT AV and NW 20TH ST';

[2]
Transfer to Route 86
Board Route 86 (Outbound)
At 'NW CONNECTICUT AV and NW 20TH ST'
Dep. Time : 10:21:00

Get down at 'NW H ST and NW JACKSON PL';

[3]
Transfer to Route 75
Board Route 75 (Inbound)
At 'NW H ST and NW JACKSON PL'
Dep. Time : 10:32:42

Get down at 'SE INDEPENDENCE AV and SE 1ST ST';

[4]
Transfer to Route 131
Board Route 131 (Outbound)
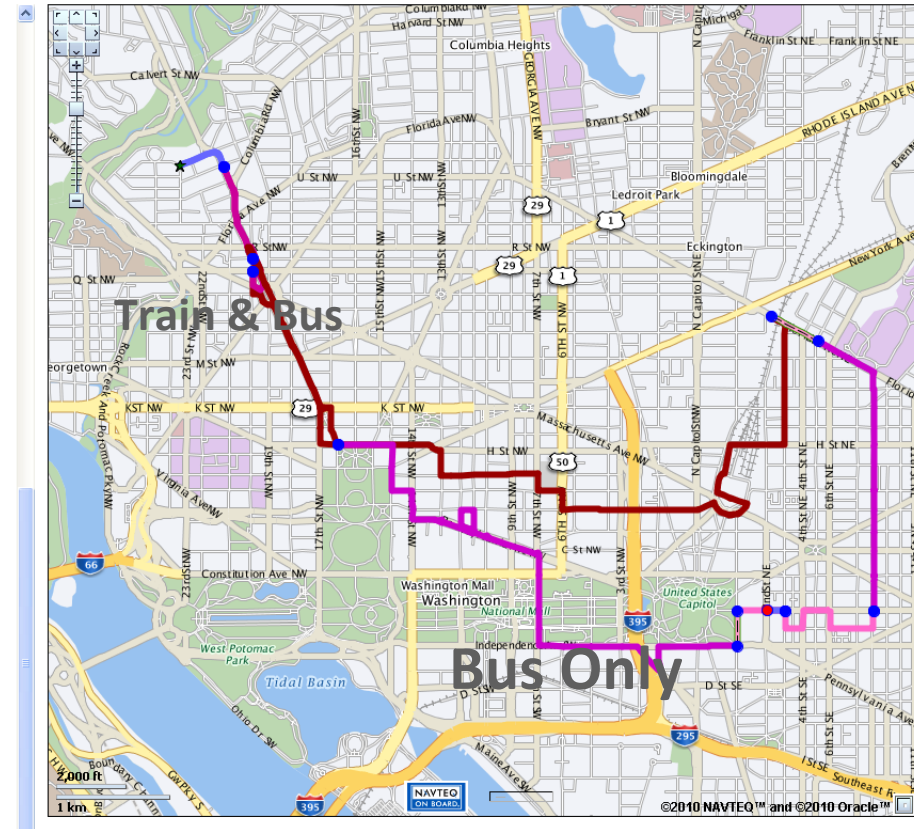At 'E CAPITOL ST and SE 1ST ST'
Dep. Time : 11:01:06

Get down at 'E CAPITOL ST and SE 3RD ST'
At 11:02:00

Drive/Walk from
'E CAPITOL ST and SE 3RD ST'
 (0 meters) to destination.

Trip Travel Time:  51 minutes.

Number of Bus Routes=4
Number of Train Routes=0

Time to analyze the network: 0.914s.

# Network Data Model Graph
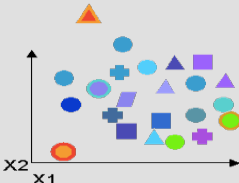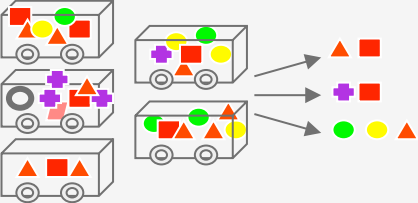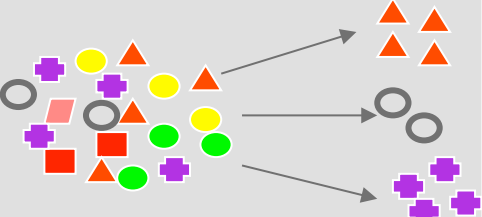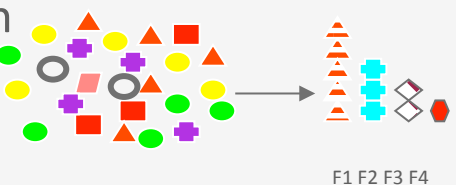## Large Scale Drive Time/Distance Analysis

## Big Data Analysis

- Millions of customers, find closest store within a specified drive time

- Single database query to find closest store and drive time/distance for each customer

- Customers geocode as based on graph segment
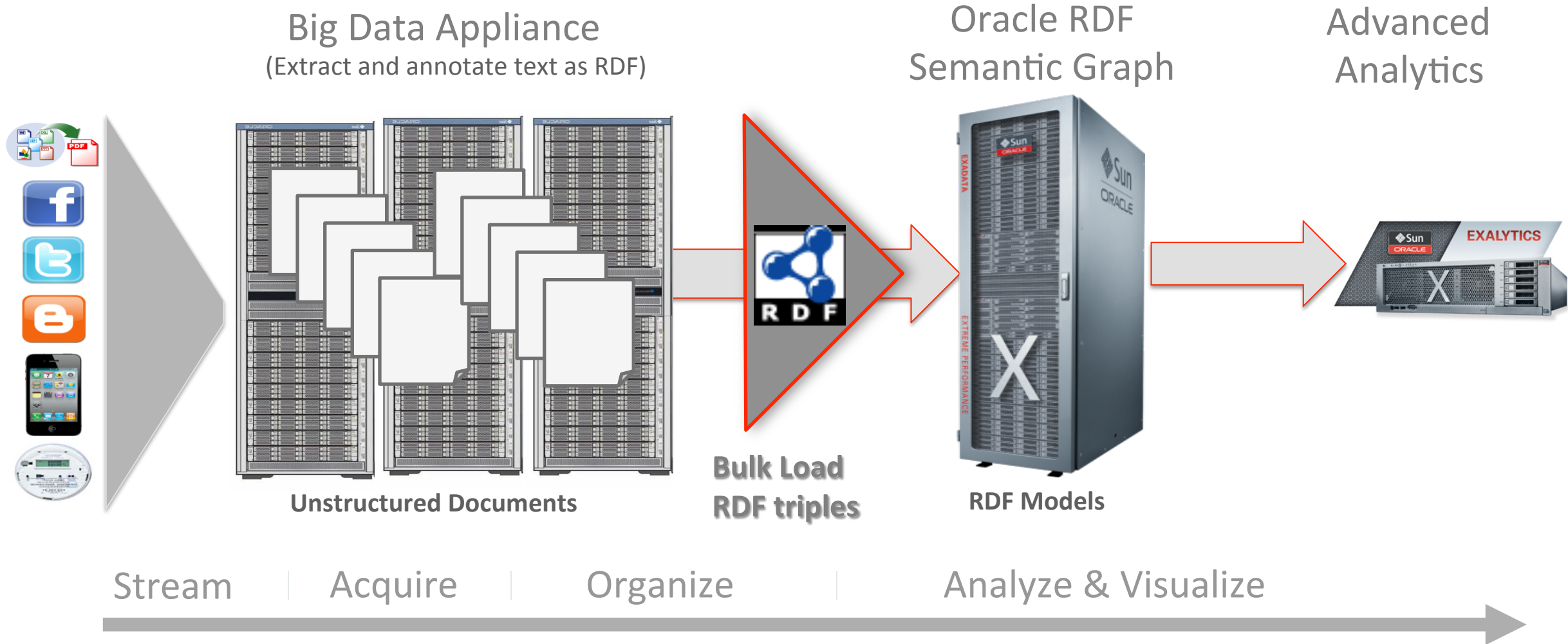
- Network Buffer generates all possible paths



★ Store Location
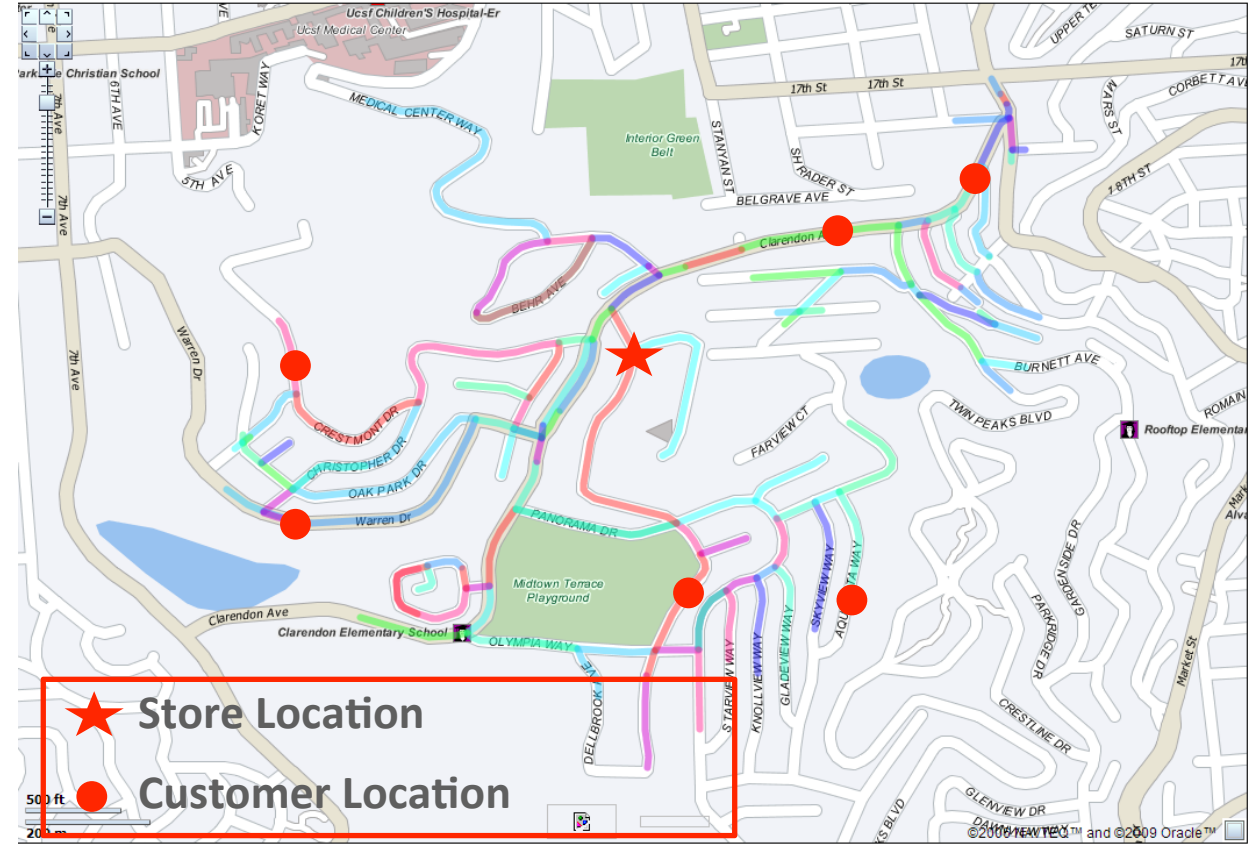
● Customer Location

# Discovery & Predictive Analysis
## DMaaS (Data Mining as a Service)

| Problem Classification | Sample Problem |
|---|---|
| Anomaly Detection | Given demographic data about a set of customers, identify customer purchasing behavior that is significantly different from the norm |
| Association Rules | Find the items that tend to be purchased together and specify their relationship – market basket analysis |
| Clustering | Segment demographic data into clusters and rank the probability that an individual will belong to a given cluster |
| Feature Extraction | Given demographic data about a set of customers, group the attributes into general characteristics of the customers |

# Use Case: Aligning Unstructured Content

**Big Data Appliance**
(Extract and annotate text as RDF)

**Oracle RDF
Semantic Graph**

**Advanced
Analytics**

**Unstructured Documents**

**Bulk Load
RDF triples**

**RDF Models**

Stream | Acquire | Organize | Analyze & Visualize

ORACLE®

# Community Data Model - Enablers

- School of the future
- Target parents
- College fundraising

**Analytics and smart education**

- Smart healthcare
- Productivity-led
- Holistic patient care and journey

**IoT and smart healthcare**

- BDA investment program
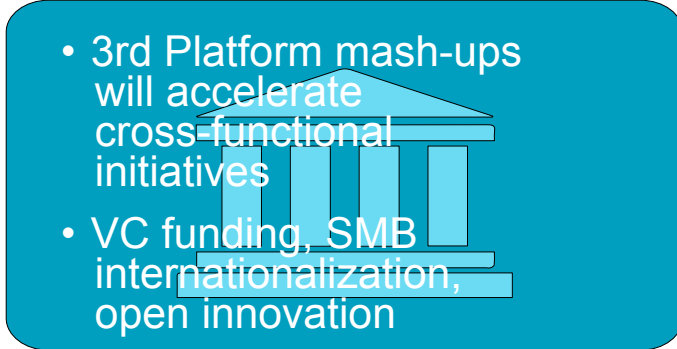- Customer privacy act
- Data sovereignty

**Data-as-a-strategic asset**

- Go-to-market (GTM) accelerator program
- ISV and device "craftsmanship"
- Innovative partnerships/ GTM channels

**Start-up accelerators**

- Surveillance and border security
- Emergency response
- Disaster management

**Public safety and order**

- 3rd Platform mash-ups will accelerate cross-functional initiatives
- VC funding, SMB internationalization, open innovation

**Inter-government innovation**

Source: *IDC Asia/Pacific Government and Healthcare Insights, 2014*

# Hardware and Software
## Engineered to Work Together

ORACLE®