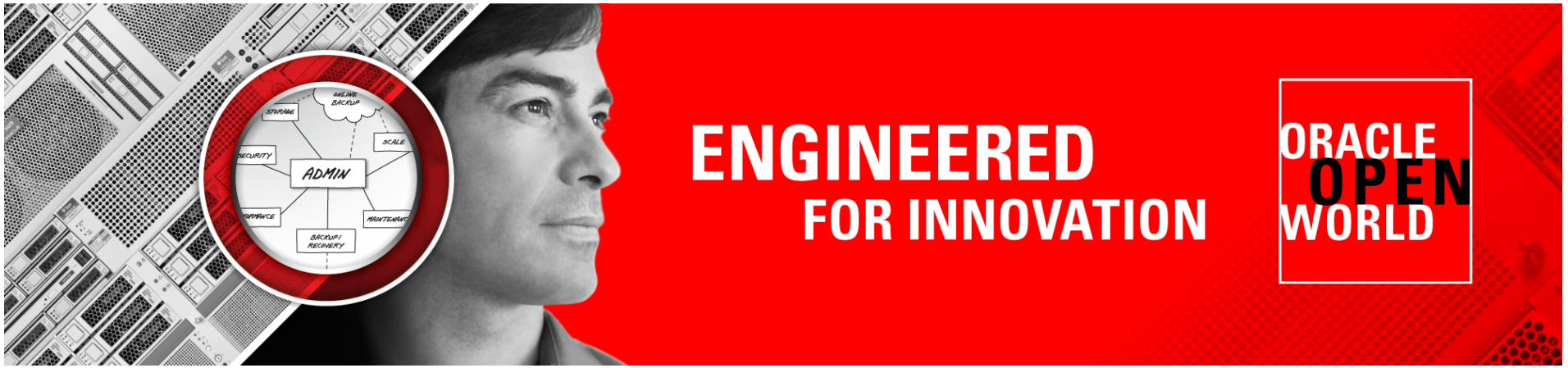

ORACLE®



ORACLE®

Managing Big Data by Using Hadoop and Oracle Exadata

Jim Steiner, Vice President, Server Technologies

ORACLE OPEN WORLD

ENGINEERED FOR INNOVATION

ORACLE

Latin America 2011

December 6–8, 2011

Tokyo 2012

April 4–6, 2012

Oracle OpenWorld Bookstore

- Visit the Oracle OpenWorld Bookstore for a fabulous selection of books on many of the conference topics and more!
- Bookstore located at Moscone West, Level 2
- All Books at 20% Discount

DigitalGuru
Technical Bookshop

ORACLE

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.

The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Big Data Buzz

"Why big data is a big deal"

InfoWorld – 9/1/11

"The challenge–and opportunity–of big data"

McKinsey Quarterly—5/11

"Ten reasons why Big Data will change the travel industry"

Tnooz -8/15/11

"Keeping Afloat in a Sea of 'Big Data"

ITBusinessEdge – 9/6/11

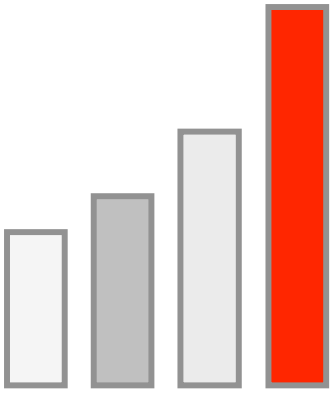
"Getting a Handle on Big Data with Hadoop"

Businessweek-9/7/11

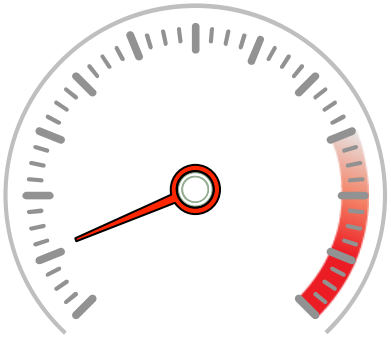
"The promise of Big Data"

Intelligent Utility-8/28/11

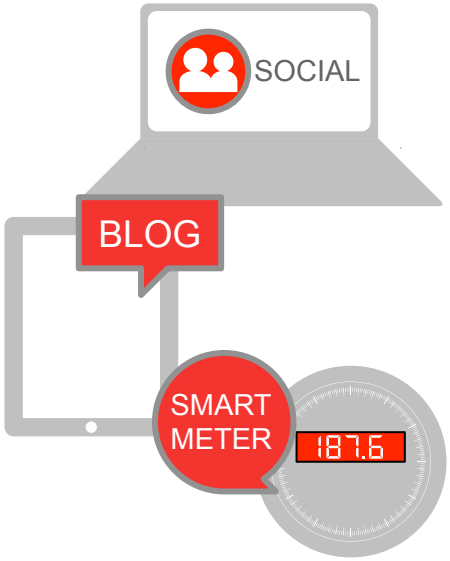
What is Big Data



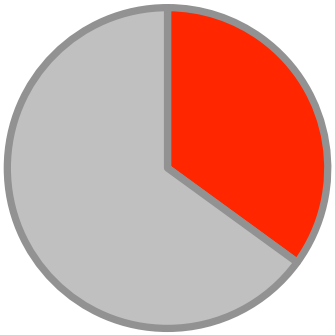
VOLUME



VELOCITY

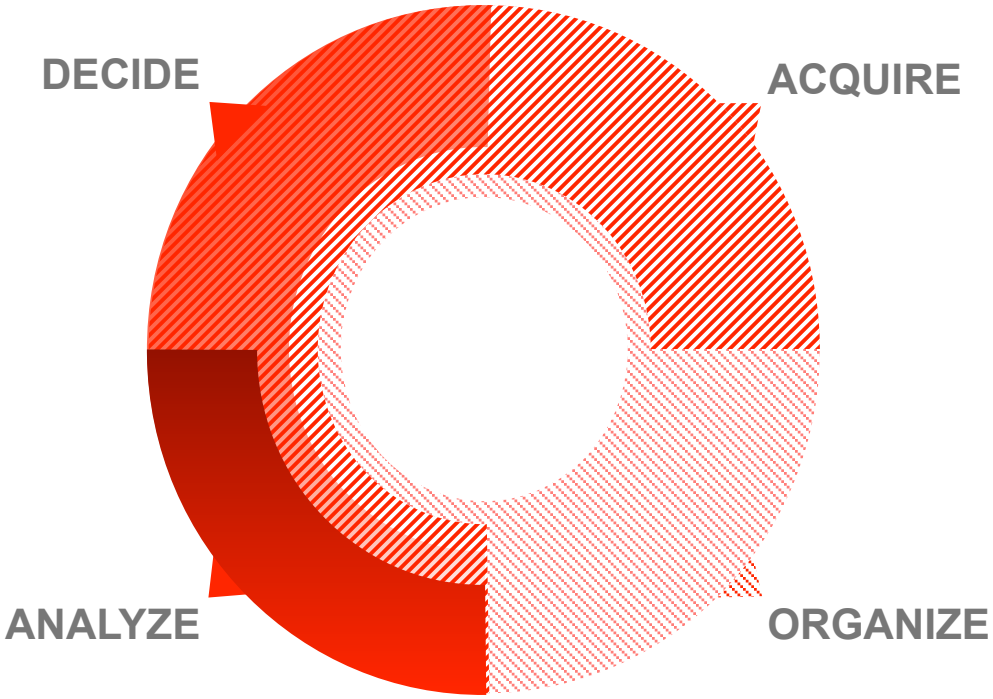


VARIETY



VALUE

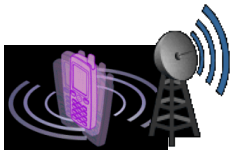
Big Data in Action

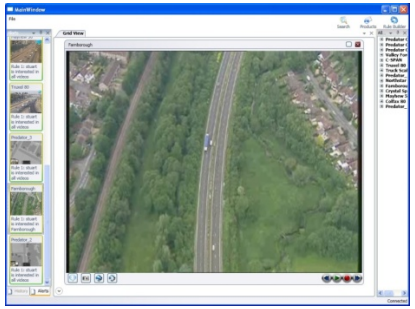


Make Better Decisions Using Big Data

Big Data

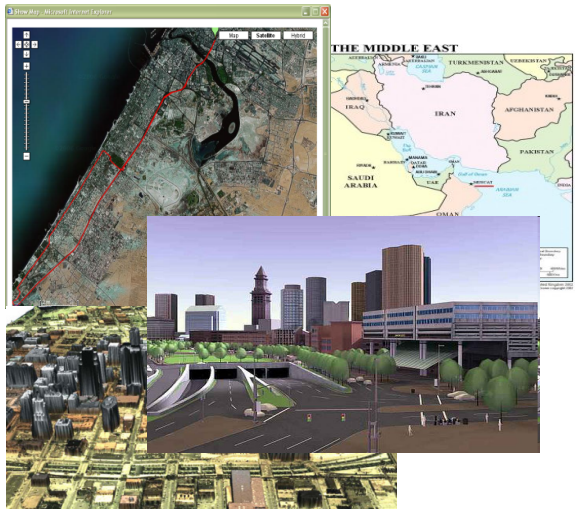
- Sensor data
- Clickstream data (weblogs)
- Social network data and logs
- Imagery
- Video surveillance feeds
- …





Big Data

Diverse types of data

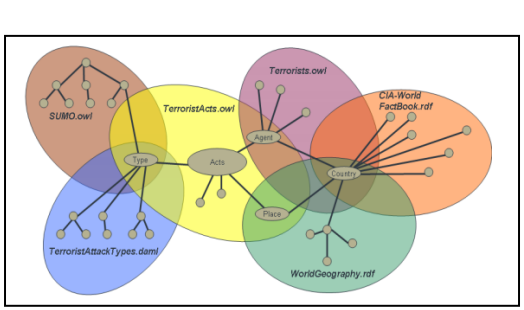


Geospatial, 3-D, Maps

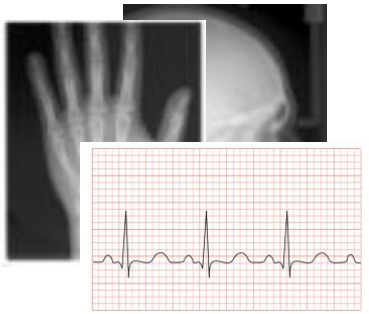
```
<?xml version="1.0"?>
<quiz>
 <question>
  Who was the forty-second
  president of the U.S.A.?
 </question>
 <answer>
  William Jefferson Clinton
 </answer>
 <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

Video



Graphs and networks



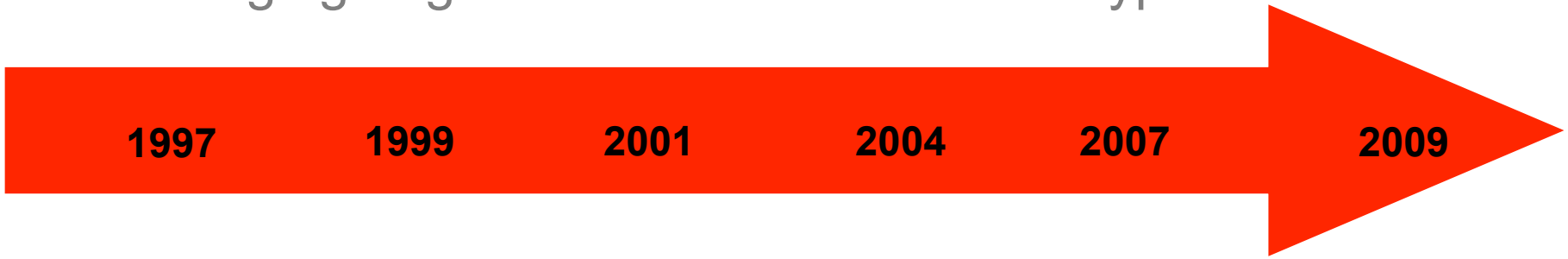
Medical images, ECGs

Oracle has always stored both structured and unstructured data. This is really nothing new. We are constantly adding features to our database to support the storage and searching of unstructured as well as structured data. … Oracle hasn't been just an RDBMS for about 20 years. … Oracle's strategy has always been to integrate additional types of data into the Oracle Database… whether it's video or audio or images … we think the transition continues. We started with relational then objects then text then XML. Now [there's] a lot of different types of unstructured data types all going into the Oracle Database. Finally, big data or the searching of large amounts of data using Hadoop. After Hadoop finishes filtering the data, the place you want to put that data is an Oracle Database, and that's what a lot of our customers are doing.

Larry Ellison, Q1 2012 Earnings Call September, 2011

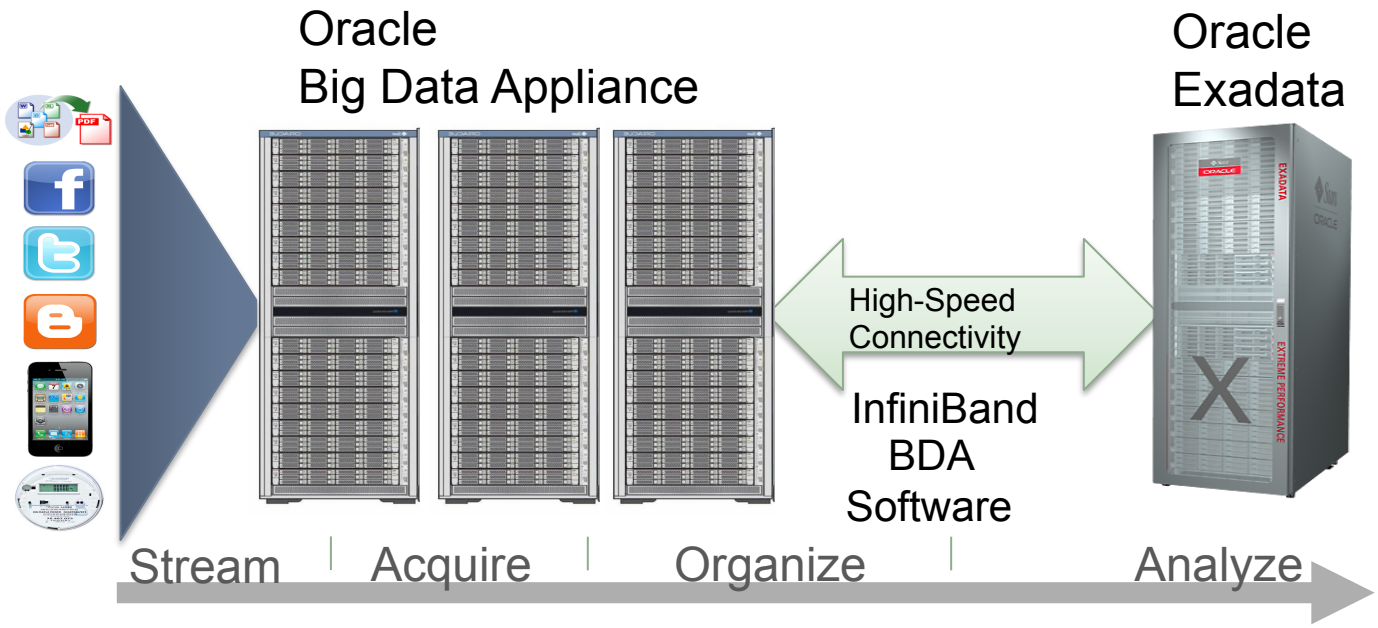
History

Managing large volumes of diverse data types



- Oracle8
  - VLDB
  - LOB's
  - Object-relational
  - Extensibility
- Oracle8*i*
  - Text
  - Spatial
  - Media
- Oracle9*i*
  - XML DB
  - Repository
  - SQL/XML
- Oracle 10g
  - ULDB
  - Location Services
  - XQuery
- Oracle 11g
  - Secure Files
  - Semantics
  - 3D & Spatial
  - Web Services
  - Binary XML
  - DICOM
- Oracle 11gR2
  - DBFS
  - Jena and SPARQL
  - Point Geocoding

Big Data



Drive Value from Big Data

Big Data Appliance

Oracle Big Data Appliance Software

Software	Description
Open Source Distribution of Apache Hadoop	Oracle distributed and supported version of Apache Hadoop open source software
Oracle NoSQL Database EE	A distributed key-value store with enterprise manageability, availability, scalability, and performance
Oracle Data Integrator Application Adapter for Hadoop	Easy to use Visual Mapping, creation, deployment and provisioning of Hadoop jobs
Oracle Loader for Hadoop	Optimized data loading from Hadoop into Oracle Database
	Infrastructure tools that provide for better interoperability with Hadoop and faster file movement

Oracle Big Data Appliance Hardware Engineered Systems

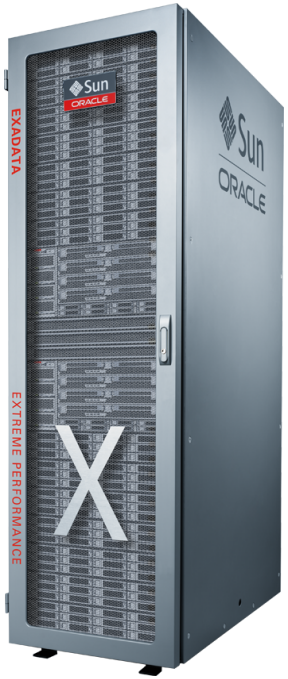
- 18 Sun X4270 M2 Servers
  - –48 GB memory per node = 864 GB memory
  - –12 Intel cores per node = 216 cores
  - –24 TB storage per node = 432 TB storage
- 40 Gb /sec InfiniBand
- 10 Gb /sec Ethernet

Maximizing the Value of Enterprise Big Data

- Hardware and software for Big Data
- Integrates all enterprise data
  - Structured and unstructured
  - SQL and NoSQL
- Fastest time-to-market
- Single vendor support

Exadata Database Machine

Best Platform to Run the Oracle Database



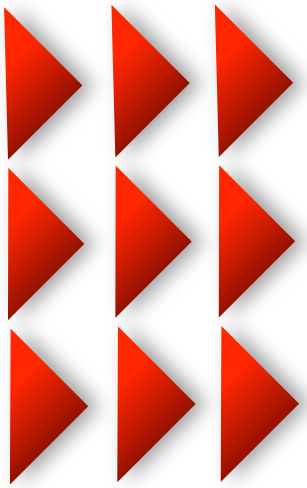
Scaleable Grid of industry standard servers for Compute and Storage

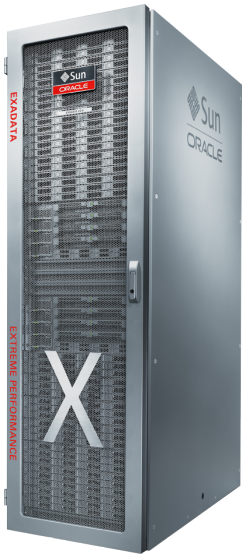
- Database Grid
- Intelligent Storage Grid
- InfiniBand Network
- 5.3 TB PCI Flash
- Unique Architecture Makes it
  - Fastest, Lowest Cost

Oracle Big Data Appliance and Oracle Exadata

Usage Model

- Diverse types of data
- High velocity
- Massive volume
- Unstructured and semi-structured





- High speed connectivity
- Infiniband
- **Oracle Loader for Hadoop**

ORACLE BIG DATA APPLIANCE

ORACLE EXADATA

Transfer Data from Hadoop to Oracle Database

Overview

- Oracle Loader for Hadoop
  - A map/reduce utility for optimized load of data into Oracle Database or Oracle Exadata
- Direct HDFS
  - Make HDFS files accessible to Oracle Database through external table definitions
  - BDA only

Hadoop

- Framework for massively parallel processing
- Management of parallel processing is transparent to the developer
- Storage is on the Hadoop distributed file system (HDFS), highly distributed and available
- Use case
  - Extract relevant data, transform and load

Oracle Loader for Hadoop Features

- Load data into a single partitioned or non-partitioned table
  - Support for scalar datatypes of Oracle Database
- Runs as a Hadoop Map-Reduce job
- Online and offline load modes
- Available on
  - Oracle Big Data Appliance
  - As a software product that can be deployed on Hadoop distributions based on Apache Hadoop 0.20.2

Oracle Loader for Hadoop Advantages

- Offload database server processing to Hadoop:
  - Converts input data to final database format
  - Computes table partition for row
  - Sorts rows by primary key within a table partition
- Generate binary datapump files
- Balance partition groups across reducers
- Works with complete table metadata knowledge

Oracle Loader for Hadoop Input Formats

- Delimited text
- Hive tables
- Write your own input format

Oracle Loader for Hadoop Output Options

Online Load

- Load directly from Hadoop nodes to Oracle database
  - Parallel JDBC
  - Parallel direct path
- No need to write to disk after Hadoop job
- Supports Oracle Wallet – secure external password store

Oracle Loader for Hadoop Load Options

Offline Load

- Datapump format
  - Create binary files for external tables
  - SQL for external table definition
  - Supports parallel direct path load
  - Fastest option for external tables
- CSV, delimited text
  - Load with SQL*Loader or external table

Pick the Output Option for the Use Case

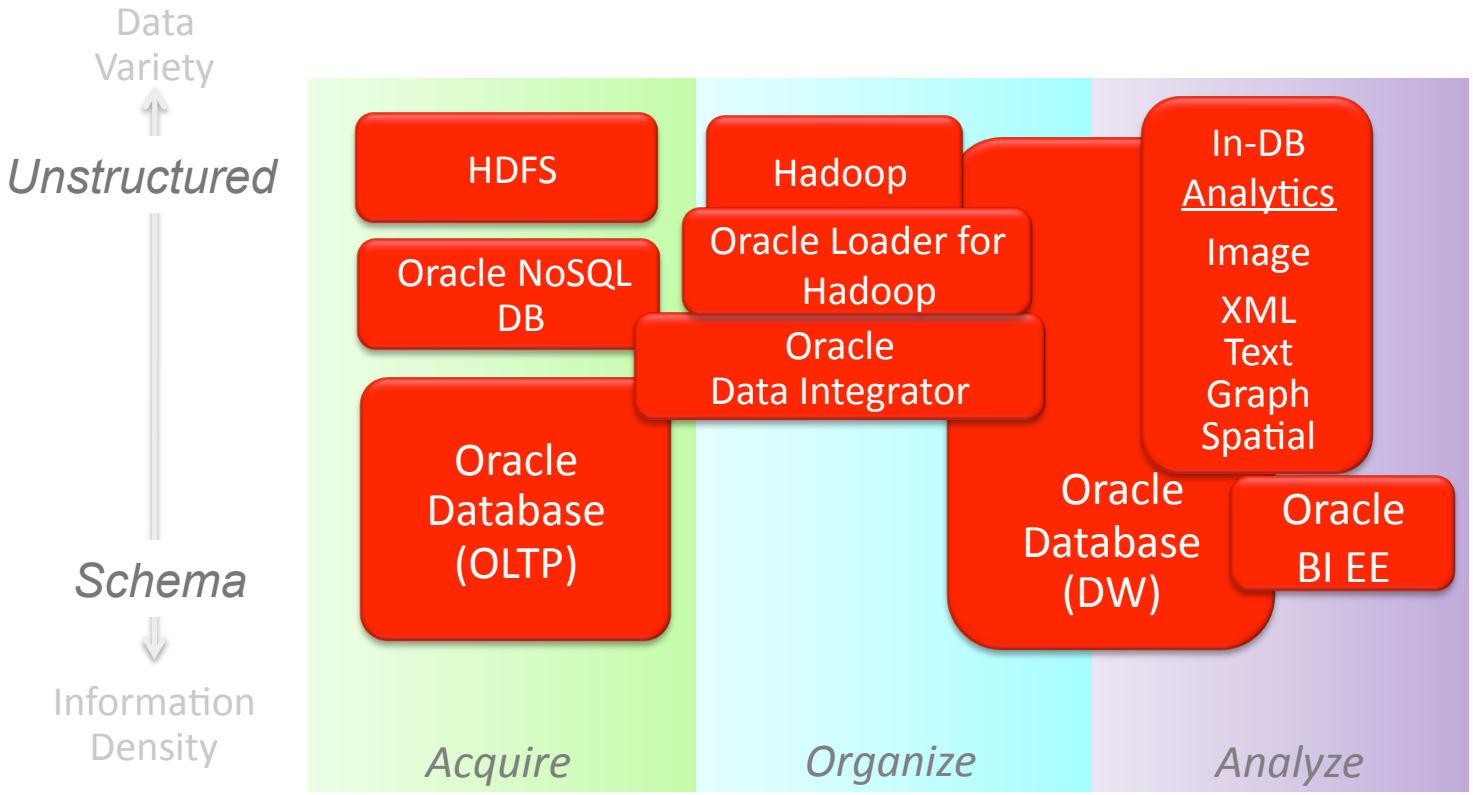
Oracle Loader for Hadoop Option	Use Case Characteristics
Online load with JDBC	The simplest use case for non partitioned tables
Online load with Direct Path	Fast online load for partitioned tables
Offline load with datapump files	Fastest load method for external tables Less load on the database server compared to online load options
On Oracle Big Data Appliance Direct HDFS	Leave data on HDFS Parallel access from database Import into database when needed

Automate Usage of Oracle Loader for Hadoop

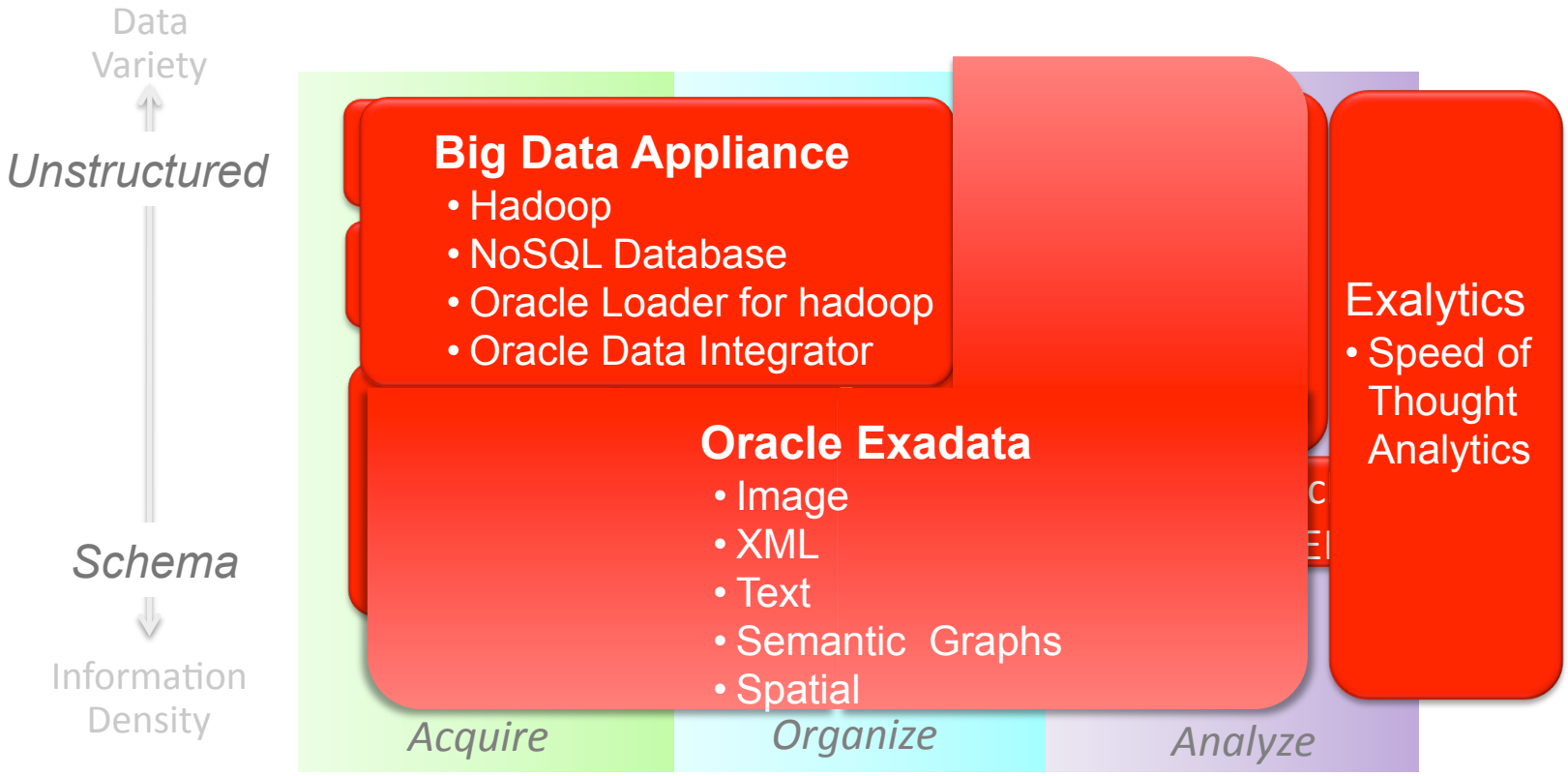
Oracle Data Integrator (ODI)

- ODI has knowledge modules to
  - Generate data transformation code to run on Hadoop
  - Invoke Oracle Loader for Hadoop
- Use the drag-and-drop interface in ODI to
  - Include invocation of Oracle Loader for Hadoop in any ODI packaged flow

Oracle Integrated Software Solution Stack



Oracle Engineered Solutions



Use Cases From Beta Customer

Healthcare

- Non-standard formats
- Data in silo-ed systems
- Diverse data types

- Transform data into standard formats
- Move to a centralized repository for analysis
- Analyze multiple data types together

ORACLE
BIG DATA APPLIANCE

Data analytics hub

- Radiation dosage and patient information, dates from medical image
- Symptoms and progress from physician's notes
- Expert knowledge from ontologies

Analyze patient outcomes to evaluate when a sonogram was not necessary before a CT Scan

ORACLE
EXADATA

Banking

- Data coming in from multiple input streams
- Unstructured text
- Schema-based data

- Consolidate data from multiple streams
- Extract entities
- Generate RDF/OWL capturing relationships between entities

ORACLE
BIG DATA APPLIANCE

Analyze complex semantic graphs

- Identify relationships between data items
- Translate bits and bytes into higher level facts
- Rationalize data across silos
- Query large graphs

Identify fraud: Users with multiple identities

ORACLE
EXADATA

Utilities

- Data feeds coming in from multiple types of smart meters
- Data in multiple vendor-specific formats
- Semi-structured feeds

- Extract relevant location, time and consumption values from raw data feeds
- Transform data into a standard spatial format

ORACLE
BIG DATA APPLIANCE

Spatial analysis

- Aggregate results based on geographic service areas
- Perform what-if analysis on the distribution network

Analyze and display service areas with high or low utilization based on smart meter data

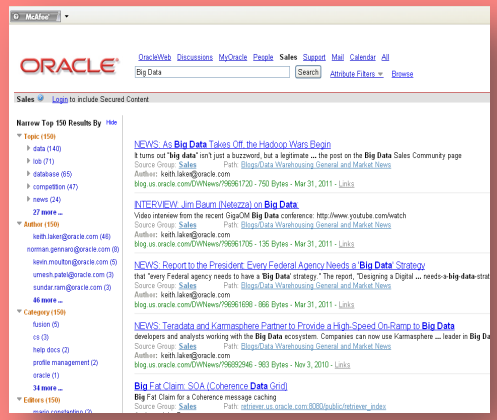
ORACLE
EXADATA

In Database Analytics

In-Database Text Search

Find Relevant Text and Comments:

```
select score(1),notes_URL
from patient_data
where appt_date AFTER '1-Jan-2011'
and Contains(notes,
'appendicitis', 1) > 0;
```



=> Efficient evaluation of structured and unstructured filters

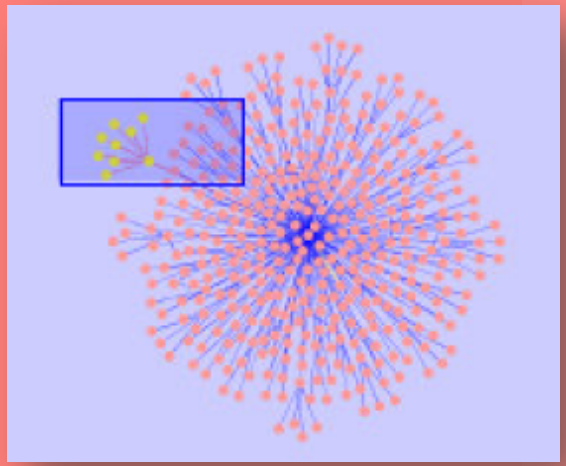
=> Search within document structures

=> Customizable scoring and relevancy

In-Database Graph Analytics

Uncover Social Relationships:

```
SELECT t.c_id, t.name
FROM Customers t
WHERE SEM_RELATED
(t.name,
'rdfs:subClassOf',
'current_customer',
'customer_behavior_graph' = 1)
AND SEM_DISTANCE() <= 2;
```

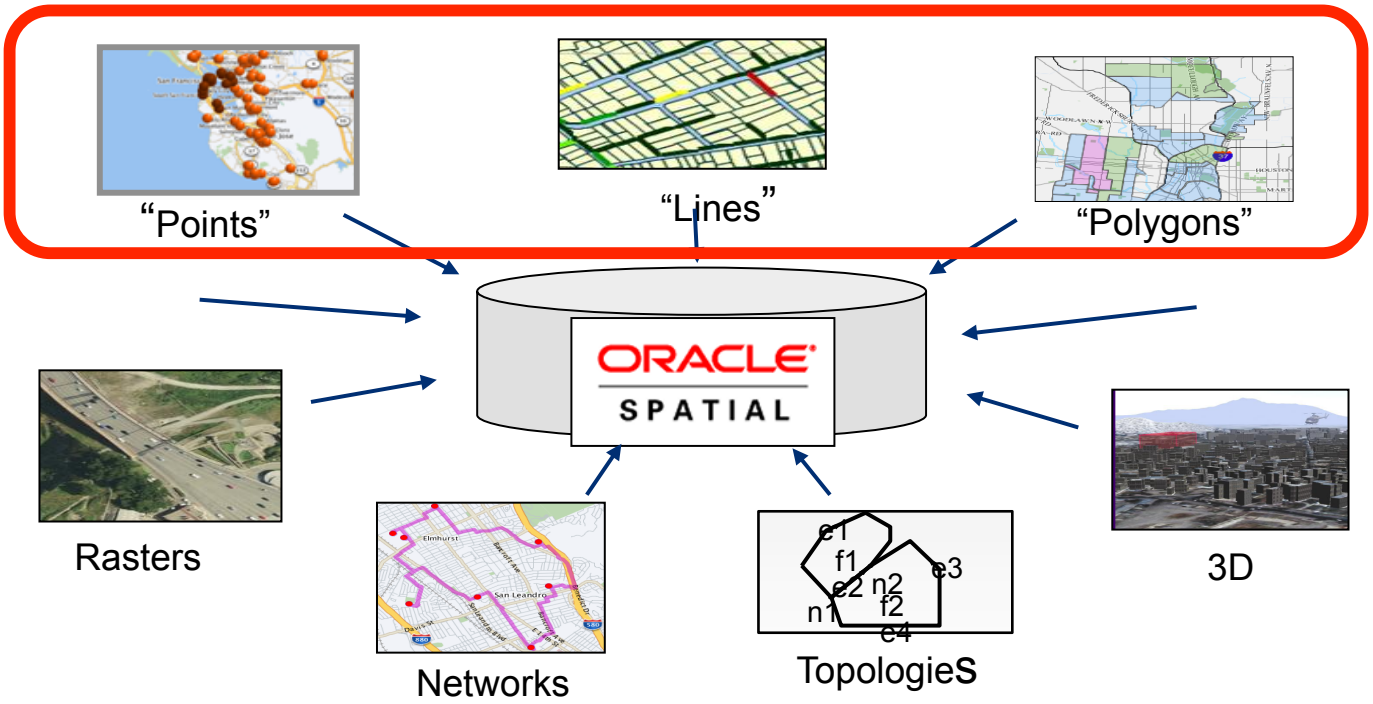


=> Broad user community and all BI tools can leverage relationships

=> Parallelism dramatically and transparently improves performance

In-Database Spatial Analysis

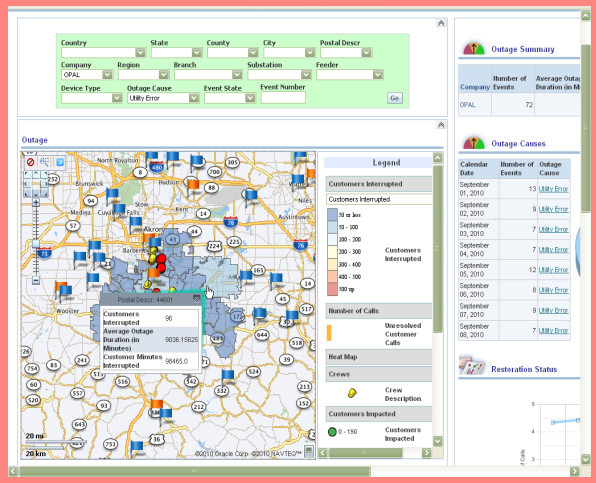
Include multiple GIS data types in analysis and reporting



In-Database Spatial Analytics

Analyze Regional Differences:

```
SELECT b.name
  FROM meters c,
       neighborhood b
 WHERE c.consumption > X
   AND SDO_ANYINTERACT(c.location,
                       b.location)= 'TRUE';
```



⇒OBI EE and MapViewer delivers Spatial Data to any user

⇒Spatial data co-located with all other data

⇒Exceptional performance on Exadata

In-Database XML Storage and Query

Directly Analyze Purchases:

```
let $USER := "SKING"
for $doc in fn:collection("oradb:/OE/PURCHASEORDER")
    where $doc/PurchaseOrder[User = $USER]
    order by $doc/PurchaseOrder/Reference
    return $doc/PurchaseOrder/Reference
```

```
<PurchaseOrder  DateCreated="2011-01-31">
  …
 <LineItems>
   <LineItem ItemNumber="1">
     <Part Description="Octopus">31398750123</Part>
     <Quantity>3.0</Quantity>
   </LineItem>
…..
   <LineItem ItemNumber="5">
     <Part Description="King Ralph">18713810168</Part>
     <Quantity>7.0</Quantity>
   </LineItem>
 </LineItems>
</PurchaseOrder>
```

XQuery operations on XML and Relational data. SQL operations on XML Content

In-Database Image Management

Analyze medical image metadata:

```
select m.id, t.PATIENT_NAME, t.MODALITY
from medical_image_table m,
  xmltable
    (xmlnamespaces
      (default 'http://xmlns.oracle.com/ord/dicom/metadata_1_0'),
             '/DICOM_OBJECT'
      passing m.dicom.metadata
      columns
        patient_name varchar2(100)
          path './*[@name="Patient""s Name"]/VALUE',
        modality varchar2(100)
          path './*[@name="Modality"]‘
```

⇒ Use standard and private metadata tags in analys

⇒Extreme scalability for multiple users

⇒ Store data in a unified repository and share with multiple applications

Exadata Performance

- Spatial warehouse-style queries up to 100x faster; box and distance queries up to 25x faster
- Medical Image operations 29x faster
- XML queries up to 100x faster
- Semantic operations up to 100x faster

Analyze Unstructured Data for New Insights

- Better decisions and lower costs from analytics on unstructured, diverse types of data
- Analyze many different types of data together
  - Use APIs designed for specific data type
- Useful across sectors
  - Healthcare
  - Banking
  - Utilities
  - Many more…

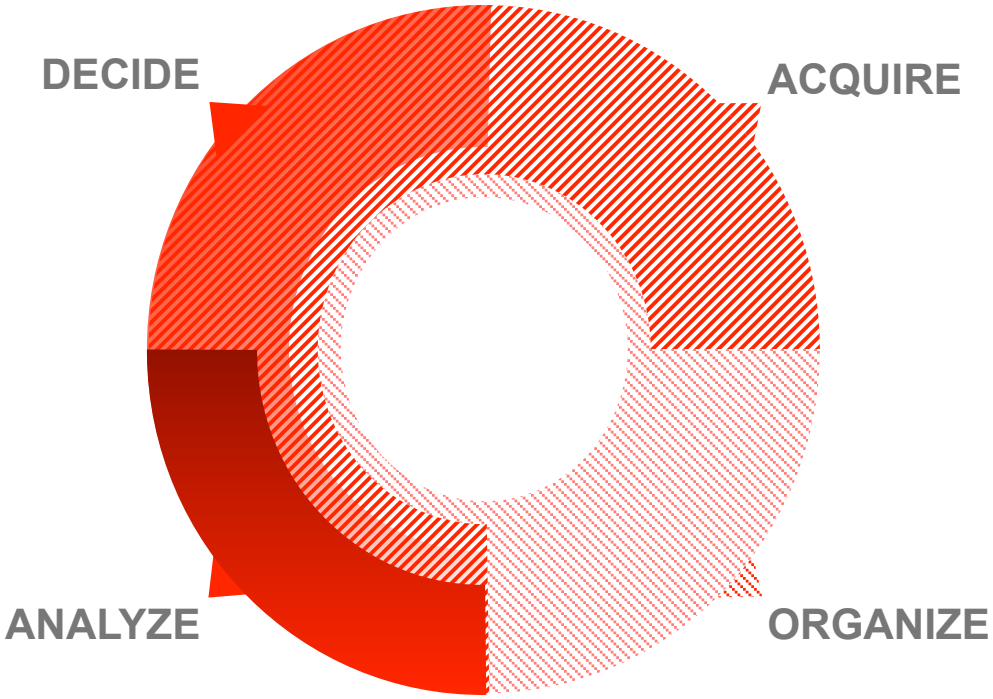
Big Data Business Value

Industry	New Data	What's Possible	Why?
Healthcare Improve Quality and Efficiency	Practitioner's notes; machine statistics	Best practices, reduced hospitalization	Increase industry value by $300 B per year
Retail One size fits all marketing	Weblog, click streams	Micro-segmentation, recommendations	Increase net margin by 60%
Banking Fraud detection; risk analysis	Weblogs, transaction systems, fraud reports	Semantic discovery; pattern detection	Billions of Dollars lost in bank fraud annually
Location-Based Services Based on home zip code	Personal location data	Geo-advertising, traffic, local search, more.	Increase revenue for service providers by $100 B+
Utilities Resilient and adaptible grid	Smart meter reading, call center data	Realtime and predictive utilization analysis	Energy use expected to grow by 22 percent by 2030

ORACLE

Source: * McKinsey Global Institute: Big Data – The next frontier for innovation, competition and productivity (May 2011)

Big Data in Action



Make Better Decisions Using Big Data

Q&A

Hardware and Software

ORACLE

Engineered to Work Together


ORACLE®