Big Data & Predictive Analytics: The Next Big Leap

Rich Niemiec, Rolta 2013

ROLTA

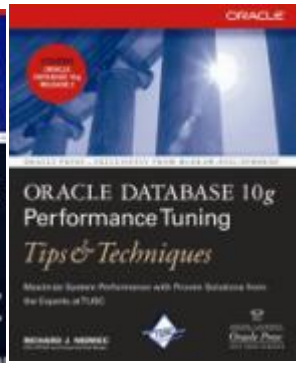
Innovative Technology for Insightful Impact

Rich's Overview

ORACLE ACE Director

ROLTA

- Advisor to Rolta International Board
- Former President of TUSC
  - Inc. 500 Company (Fastest Growing 500 Private Companies)
  - 10 Offices in the United States (U.S.); Based in Chicago
  - Oracle Advantage Partner in Tech & Applications
- Former President Rolta TUSC & Former President Rolta EICT International
- Author (3 Oracle Best Sellers – #1 Oracle Tuning Book for over a Decade):
  - Oracle Performing Tips & Techniques (Covers Oracle7 & 8i)
  - Oracle9i Performance Tips & Techniques
  - Oracle Database 10g Performance Tips & Techniques
  - Oracle Database 11g Performance Tips & Techniques
- Former President of the International Oracle Users Group
- Current President of the Midwest Oracle Users Group
- Chicago Entrepreneur Hall of Fame - 1998
- E&Y Entrepreneur of Year & National Hall of Fame - 2001
- IOUG Top Speaker in 1991, 1994, 1997, 2001, 2006, 2007
- MOUG Top Speaker Twelve Times
- National Trio Achiever award - 2006
- Oracle Certified Master & Oracle Ace Director
- Purdue Outstanding Electrical & Computer and Engineer - 2007



Rolta Leaders in Database Technologies and Worldwide Oracle Platinum Partner

- ✓ Oracle Partner-of-the-Year: multiple times
- ✓ Oracle user group leadership
  - Past president of International Oracle User Group
  - Member of Applications & Technology Advisory Councils
  - Current president of Midwest Oracle User Group
    - Service Oriented Architecture (SOA)
    - Fusion
    - Oracle 11g
- ✓ Oracle Magazine Consultants-of-the-Year

Industry Recognitions

- ✓ Nine Times Oracle Titan Award Winners
- ✓ 6 “Oracle Masters” on staff

ORACLE® Platinum Partner

One of first few companies worldwide with highest level of partner certification for Apps & Technology

Rolta – *Your* Partner in Oracle Excellence

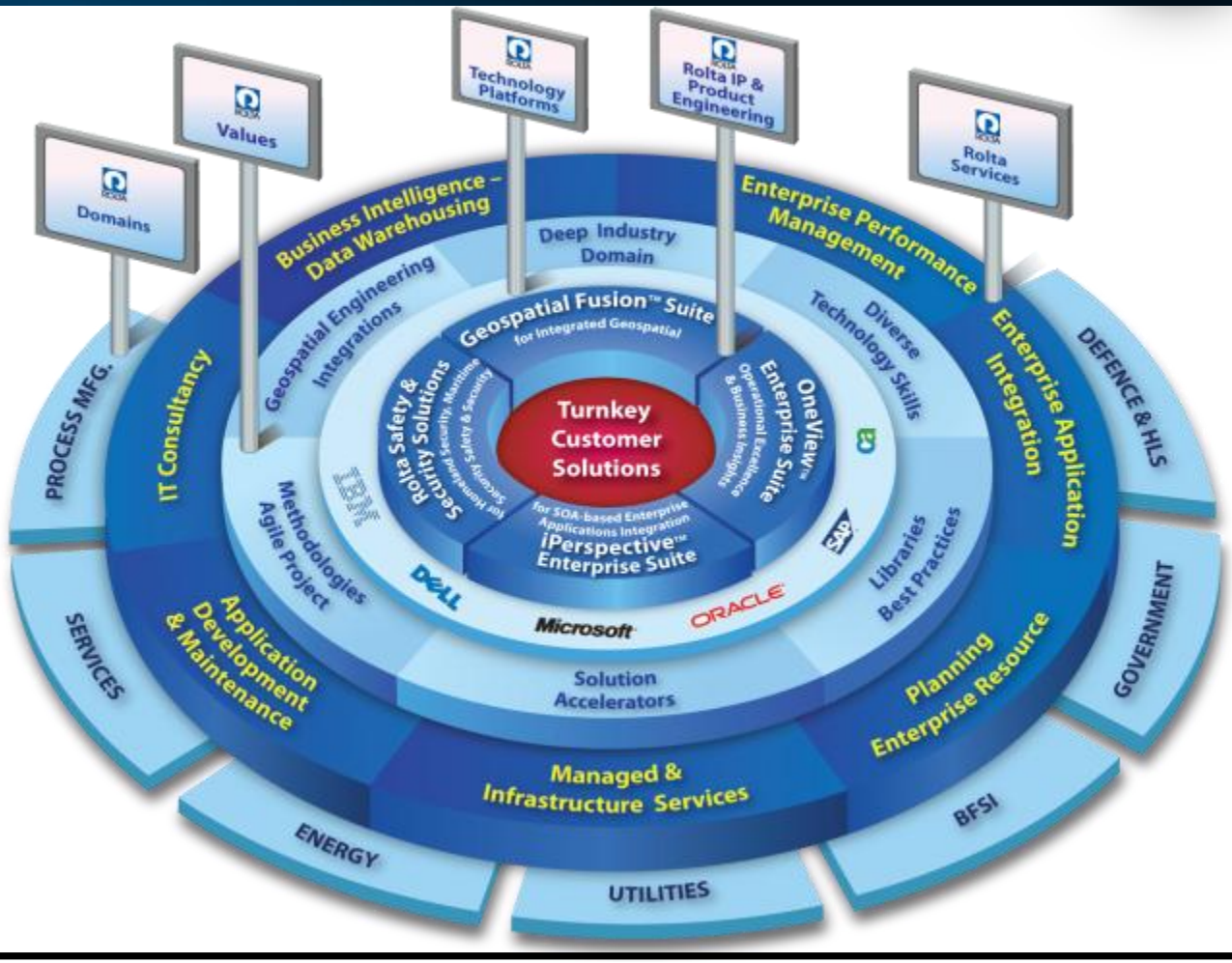
2012 Oracle Excellence Award (BI Application)

(9 Partner of the Year / Titans / Excellence Awards)

Prior Years Winner 2002, 2004*, 2007*,2008,2010, 2011

***Won 2 Awards**

How to Make a Difference in the World!



ROLTA

CMM LEVEL 5

ISO 9001 : 2008 CERTIFIED

ISO/IEC 20000-1:2005

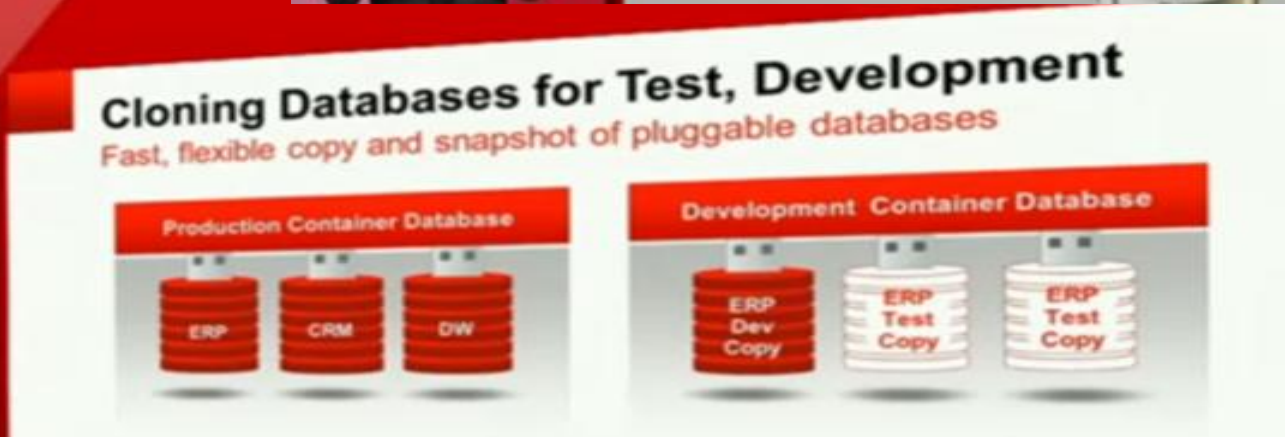
ISO/IEC 27001:2005

ISO 14001:2004
OHS 18001:227

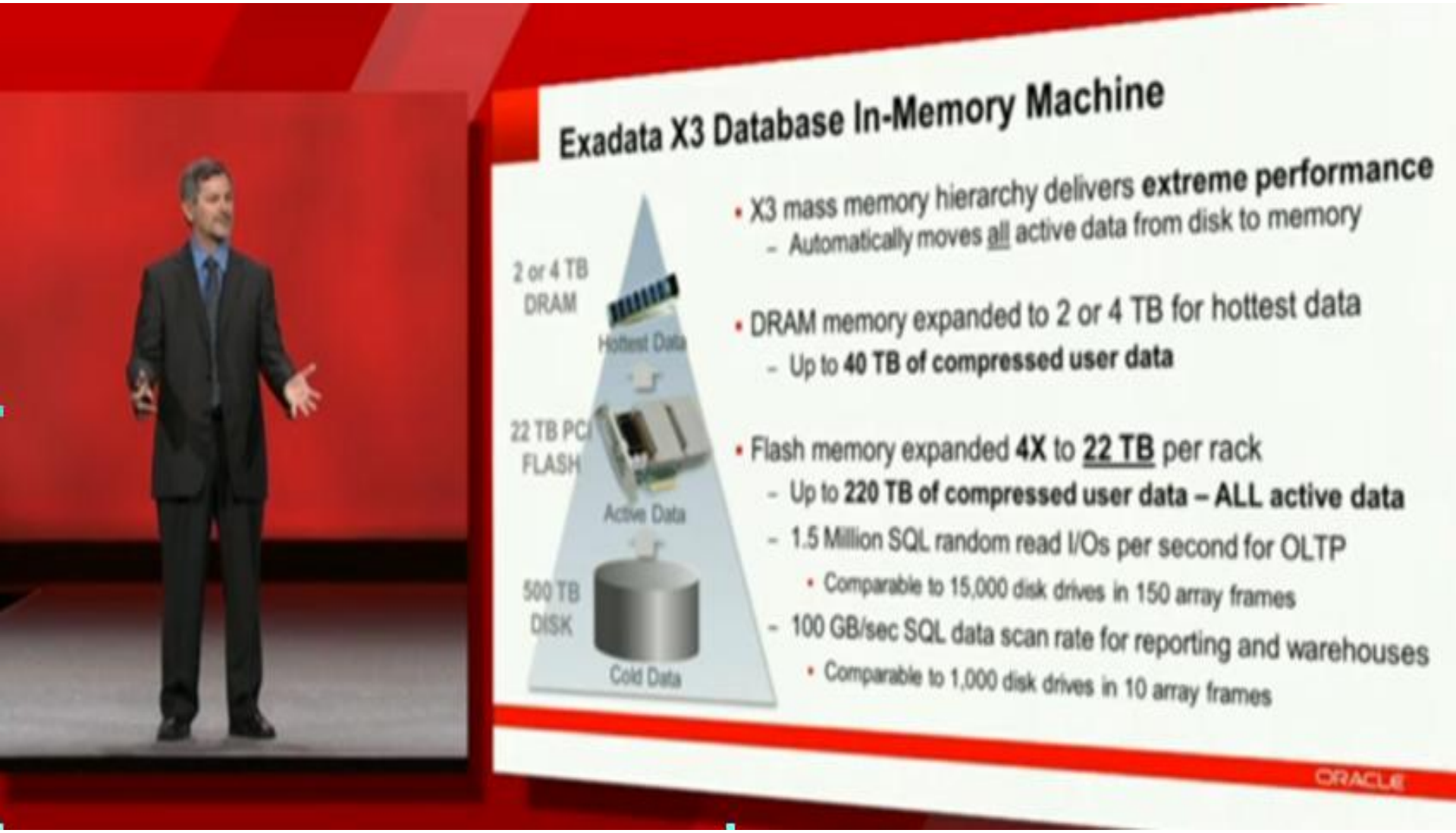
Agenda

- Oracle Trends
- Current Scenario
- Technology Evolution
- The way forward

Oracle Trends
Know the Oracle



Exadata X-3: In-Memory Database
4 T DRAM / 22 T Flash Cache



Oracle Firsts – *Innovation!*

1979 First commercial SQL relational database management system
1983 First 32-bit mode RDBMS
1984 First database with read consistency
1987 First client-server database
1994 First commercial and multilevel secure database evaluations
1995 First 64-bit mode RDBMS
1996 First to break the 30,000 TPC-C barrier
1997 First Web database
1998 First Database - Native **Java** Support; Breaks 100,000 TPC-C
1998 First Commercial RDBMS ported to **Linux**
2000 First database with **XML**
2001 First middle-tier database cache
2001 First RDBMS with **Real Application Clusters**
2004 First **True Grid Database**
2005 First **FREE Oracle Database** (10g Express Edition)
2006 First **Oracle Support for LINUX Offering**
2007 **Oracle 11g Released!**
2008 Oracle Exadata Server Announced (Oracle buys BEA)
2009 Oracle buys Sun – Java; MySQL; Solaris; Hardware; OpenOffice
2010 Oracle announces MySQL Cluster 7.1, Exadata, Exalogic
2011 Oracle X2-2, ODA, Exalytics, SuperCluster, Big Data, Cloud, Social Network
2012 Oracle X3-2, Oracle 12c OEM, Pluggable Databases & X3-8 announced
2013 Oracle12c Released! Oracle X3-8 Exadata, Acquisitions (Acme Packet)!

Business Scenario

- Recession Over, but Outlook is still challenging?
  - Great Recession of 2008-09, biggest blow to economy since 1930
  - For some, 2012 was worse than 2008-09: Europe's Debt crisis, General slowing of Global Economy, Domestic Political Troubles in various nations
- Many country ratings are reduced, *Most* banking unscathed, Reforms will drive some growth, Big Data will drive more!
- World is awash with excess capacities that will find its way into demand centric markets

Worldwide capacity utilization
- Oil Refineries – 84%
- Steel – 76%
- Cement – 75%
- Aluminium – 73%
- Poly-crystalline silicon – 20%

China's production capacity (as % of global capacity)
- Oil Refining ~20% *(excluding OPEC)*
- Steel ~ 50%
- Cement ~ 50%
- Aluminium ~ 40%

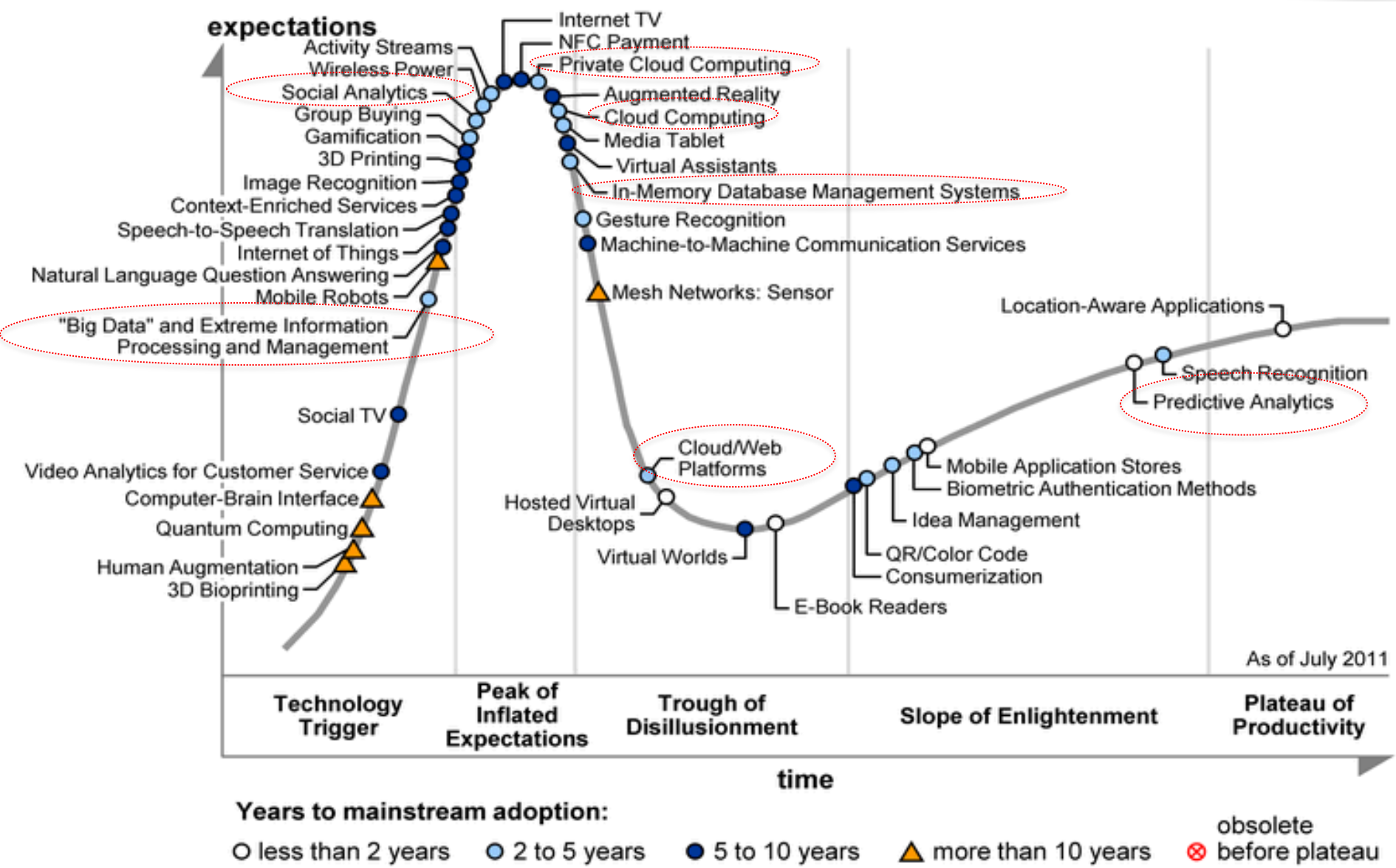
Data collated from various Industry Sources

Deep Analytics – a must, to ensure sustainability, drive growth & profitability

Some of the typical Organizational Challenges and need for Analytics

- Cloud Computing, Mobile Computing, Social Media and Big Data Analytics are driving the New Computing Paradigm.
- This paradigm in-turn sparks-off Business Transformations to improve Efficiency, Compliance with Regulation and overall Business Sustainability based on Customer Centricity.

Technology Trends: Gartner Hype Cycle 2012



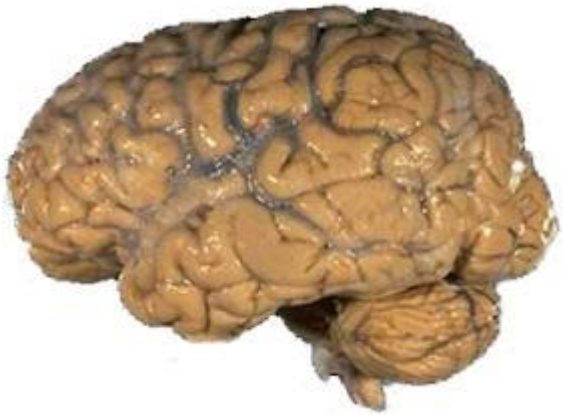
Gartner Trends for 2012

Bigger Data - Data Size Matters…

- Worldwide, data is growing rapidly over the years….
  - 2000: 800 Terabytes (10^{12})
  - 2006: 160 Exabytes (10^{18})
  - 2009: 500 Exabytes (just Internet)
  - 2012: 2.7 Zettabytes (10^{21})
  - 2020: 35 Zettabytes …?
- Data generated in ONE day….?
  - Twitter: 7 TB
  - Facebook: > 10 TB

Big data: The next frontier for innovation, competition, and productivity

McKinsey Global Institute 2011



2.8 x 10^{20} bits of Memory Space – John von Neumann ("Computer and the Brain", Harvard Lecture Notes, Half Century ago)

Data collated from various online sources

We are drowning in data, but thirsting for Information

How Much Data …

- 2004 monthly internet traffic >1E; 2010 it was 21E/month.
- In 2012, **2.5E data created every day** (about 1Z=1000E per year)
- June 2012 – Facebook has **100P Hadoop cluster**
- Facebook: **500T** processed daily – (210T/hr Hive scanned)
- A Single Jet Engine **– 20T/hour** –same rate as Facebook!
- Gmail has **450 million users**
- Wal-Mart – 1 million customer transactions/hour (2.5P DB)
- Large Hadron Colider produced **13P in one year**
- Business data **doubles every 1.2 years**
- 19% of $1B **companies have >1P of data** (31% in 2013)
- 2011 – First **Exabyte tape library** from Oracle
- Decoding Human Genome took 10 yrs; Now takes a week!

Big Data & Basics

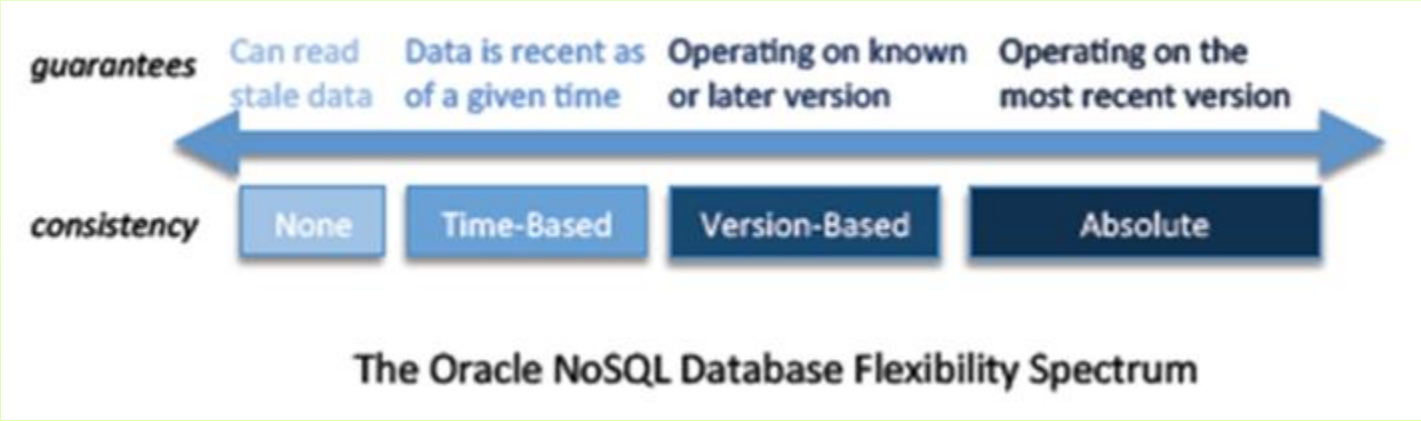
- Big Data includes: Social Media, Sensor Data, Biological, Traffic, RFID Data, Environmental, Aerial, Wireless, Security & Video Data, Retail, Medical, Engineering Systems, Search Data, Photographs, Call Records, CRM/ERP data, etc.

- Goal was to **Organize Data without moving it!** – Hadoop HDFS & MapReduce (Cheaper way to access Petabytes)
- **Acquire & Store data** – NoSQL (simple key value storage) – Amazon DynamoDB (hosted), Apache Cassandra, HBase, BigTable, MongoDB, Oracle NoSQL (distributed key value) or just use the original HDFS / GFS & MapReduce (many are **EVENTUALLY** consistent!)
- **Analyze Data** – Google Dremel, Apache Hive Data Warehouse, Oracle Data Warehouse
- 54% of companies doing Big Data say **projects are critical**

Atomicity Consistency Isolation Durability Test

NoSQL supports BASE:

- Basically Available
- Soft state
- Eventually consistent



The Oracle NoSQL Database Flexibility Spectrum

In the Beginning… How did we get here?

- Larry Page & Sergey Brin wrote BigFiles; **GFS** (Google File System) grew out of that & then **MapReduce** which maps problems across cluster a of worker nodes & then collects results & aggregates/reduces result (**used to generate Google's index of WWW**)
- Apache came out with **Hadoop** (*used by Facebook, Yahoo, Amazon EC2 & S3*) which was an Open Source version with **HDFS & MapReduce** – Batch Processing Jobs going after distributed data & processing it near the data (same node) – not super fast (**seconds vs. ms**) & not good for interactive/analytic
- **Google then came out with BigTable** (compressed, high performance data storage) used by Google Maps, Google Reader, Google Earth, YouTube, and Gmail
- Apache adds **NoSQL DB's: Cassandra & HBase**
- The NoSQL onslaught of systems started (over 100 of them) including **Oracle's NoSQL (BerkeleyDB)**.

NoSQL Trends for 2012

- Hadoop goes Enterprise.
  - Microsoft joins the party (partnership with Yahoo! Spin-off Hortonworks ➔ Hadoop implementation for Windows Server & Azure with connectors to MSSQL)
- NoSQL based solutions
  - Security Issues hamper NoSQL
  - Oracle gets in the NoSQL game in a BIGGER way (Big Data Appliance)

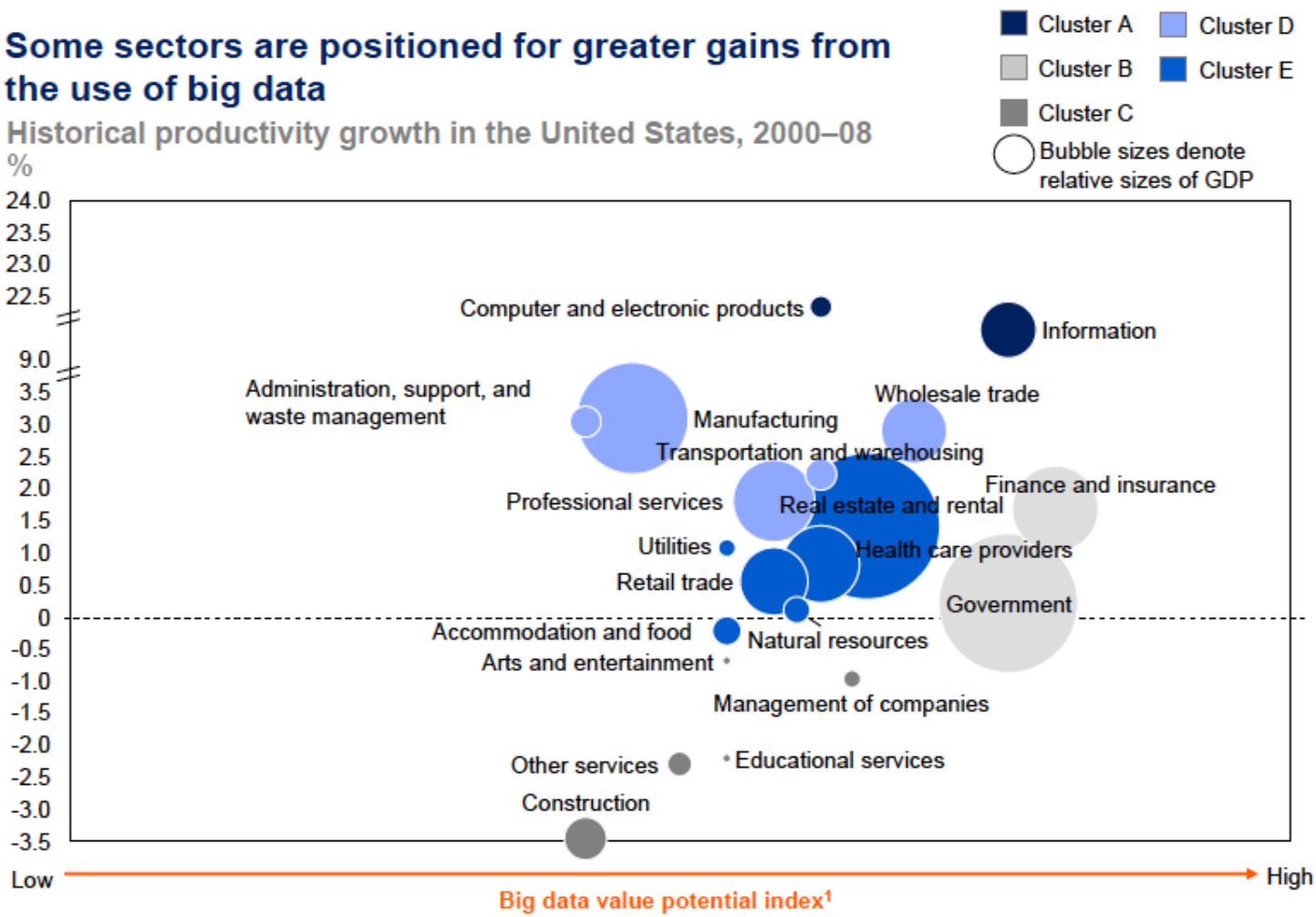
"As customers look to manage the huge explosion in data from new and evolving sources, such as the Web, sensors, social networks and mobile applications, Oracle is helping them unlock the value of this data by providing a highly available, reliable and scalable NoSQL database environment."

– *Oracle SVP, Andrew Mendelsohn*

- Integration of In-Memory Data Grids and NoSQL, leveraging success stories of Facebook & Twitter

DataVersity post on Jan 26, 2012

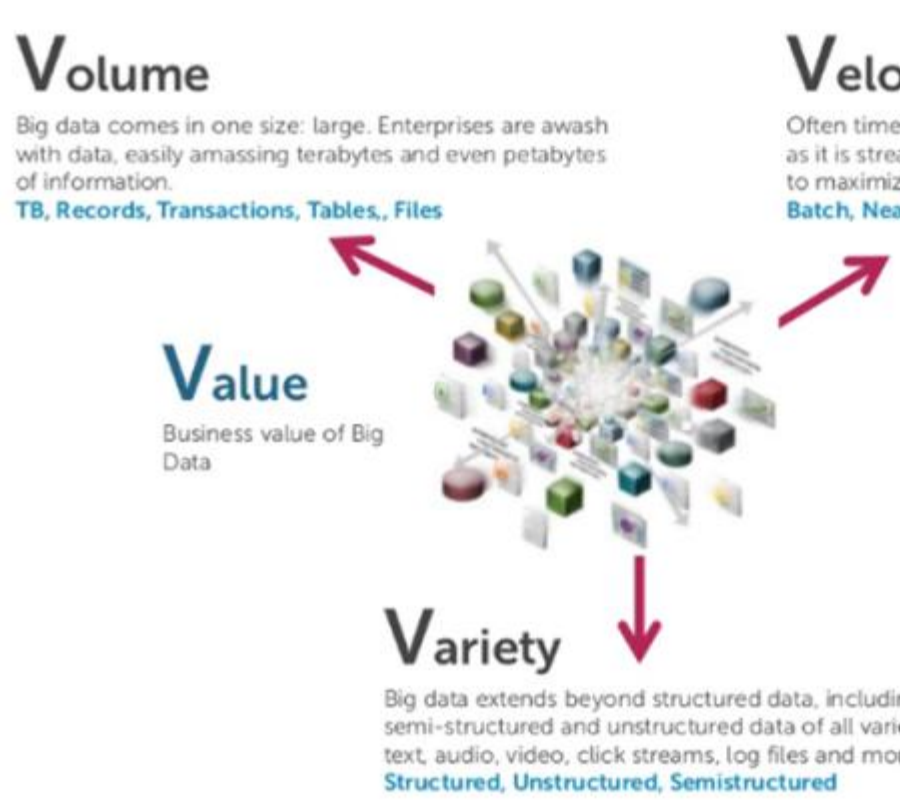
Why Is Big Data Important?



SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Jiawan Zhang
School of Computer Software,
Tianjin University

Characteristics of Big Data




Finance


Telecom


Retail


Life Sciences

Media


Government

Big Data Themes

- HW & SW technologies for large data volumes
- Focus on Web 2.0 technologies
- Database Scale-out
- Relational & Distributed Data Analytics
- Distributed File Systems
- Real Time Analytics

Big Data Domains

- Digital Marketing Optimization
- Data Exploration & Discovery
- Fraud Detection & Prevention
- Social Network & Relationship Analysis
- Machine-generated Data Analytics
- Data Retention

All Data is not similar!

Data Realms	Security	Storage & Retrieval	Modeling	Processing & Integration	Consumption
Master data Transactions Analytical data Metadata	Database, app, & user access	RDBMS / SQL	Pre-defined relational or dimensional modeling	ETL/ELT, CDC, Replication, Message	BI & Statistical Tools, Operational Applications
Reference data	Platform security	XML / xQuery	Flexible & Extensible	ETL/ELT, Message	System-based data consumption
Documents and Content	File system based	File System / Search	Free Form	OS-level file movement	Content Mgmt
Big Data - Weblogs - Sensors - Social Media	File system & database	Distributed FS / noSQL	Flexible (Key Value)	Hadoop, MapReduce, ETL/ELT, Message	BI & Statistical Tools

Data Realm Characteristics (Oracle Information Architecture Framework)

Analytics

Executive Management	Strategy Management	Design Right Strategy, Communicate, Collaborate, Scorecard, Drive Actions
Functional Specialists / Strategist	Performance Analytics	Multi-dimensional analysis of performance, Predictive Analysis, Forecasting
Line Manager, Functional Manager	Performance Management	Functional pre-defined KPI's , Knowledge Data-model, Targets, Alerts,
Engineers, Supervisors, Operators	Real Time Intelligence	Intelligence on Real-time operational and business data, Site Schematics
	Real Time Data	ERP, Historian, CMMS

Pre-packaged Analytics are also available…

Oracle BI Applications

Comprehensive, Prebuilt, Best Practice Analytics

Sales	Service & Contact Center	Marketing	Procurement & Spend	Supply Chain & Order Management	Financials	Human Resources
Pipeline Analysis	Service Effectiveness	Campaign Effectiveness	Direct / Indirect Spend	Revenue and Backlog	General Ledger	Employee Productivity
Forecast Accuracy	Customer Satisfaction	Customer Insight	Buyer Productivity	Inventory	Accounts Receivable	Compensation
Sales Team Effectiveness	Resolution Rates	Product Propensity	Off Contract Purchases	Fulfillment Status	Accounts Payable	Compliance Reporting
Up-sell/ Cross-sell	Service Rep Efficiency	Loyalty & Attrition	Supplier Performance	Customer Status	Cash Flow	Workforce Profile
Cycle Times	Service Cost	Market Basket Analysis	Purchase Cycle Time	Order Cycle Time	Profitability	Retention Analysis
Lead Conversion	Churn & Service Trends	Campaign ROI	Employee Expenses	BOM Analysis	Expense Management	Return on Human Capital

Source adapters: ORACLE SIEBEL PeopleSoft JDEDWARDS SAP and Other Operational & Analytic Sources

Oracle BI Suite Enterprise Edition Plus

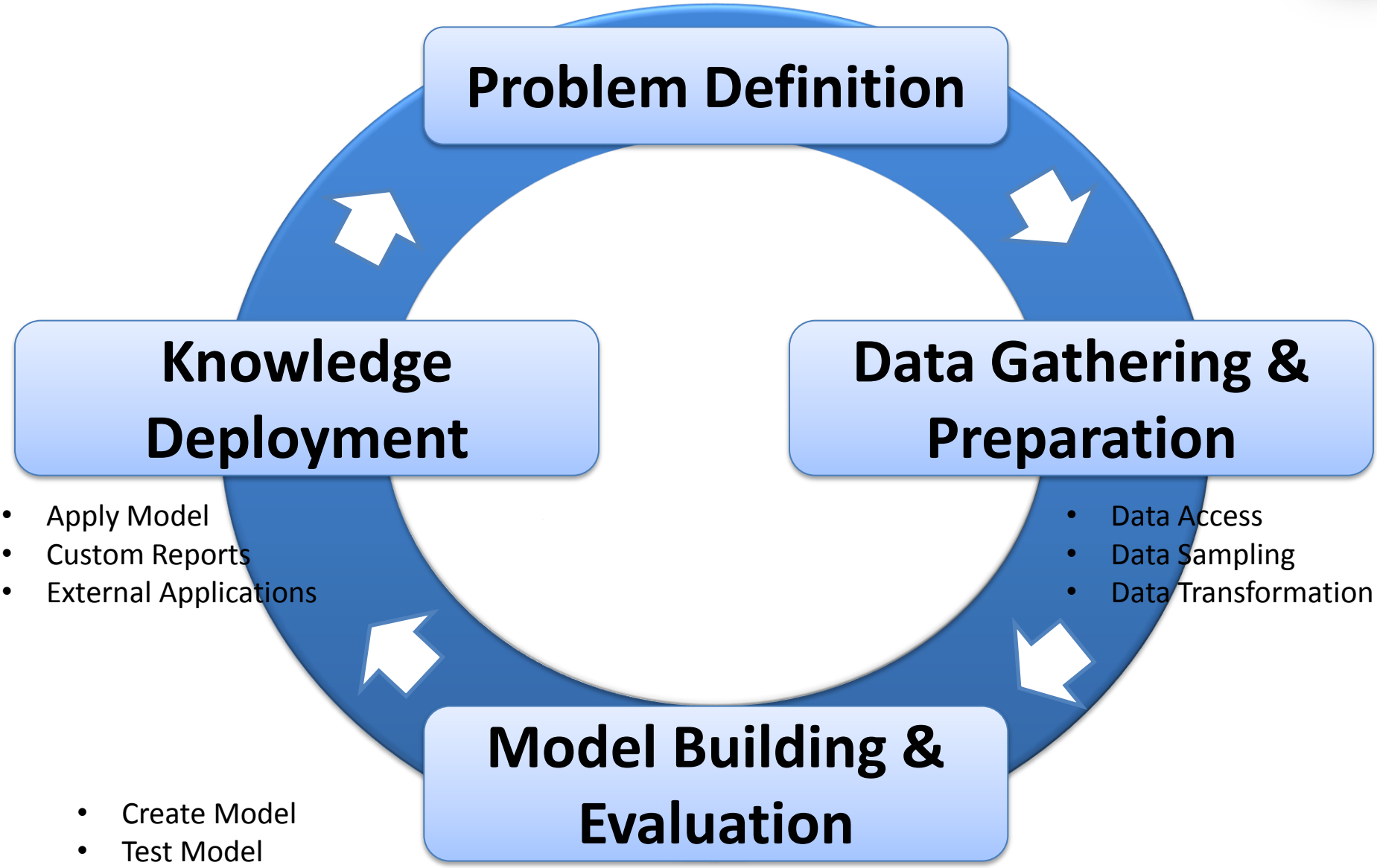
Oracle Database is loaded with Analytics !!

Analytical Feature	Description
Data Mining	Oracle Data Mining implements complex algorithms to discover patterns, predict probable outcomes, identify key predictors, etc.
Complex data transformations	ETL capabilities and SQL expressions or DBMS_DATA_MINING_TRANSFORM package. for missing values, outlier treatments, binning and normalization.
Statistical functions	SQL statistical functions: hypothesis testing (t-test, F-test), pearson correlation, cross-tab/descriptive statistics (median, mode, etc). DBMS_STAT_FUNCS package adds distribution fitting procedures.
Window / Analytic SQL functions	Computing cumulative, moving, and centered aggregates.
Frequent Itemsets	DBMS_FREQUENT_ITEMSET used as a building block for the Association algorithm used by Oracle Data Mining.
Image feature extraction	Oracle Intermedia supports extraction of color histogram, texture, and positional color.
Linear algebra	UTL_NLA package exposes a subset of the popular BLAS and LAPACK libraries for operations on vectors and matrices.
OLAP	Multidimensional analysis beyond drill-downs and roll-ups, Oracle OLAP also supports time-series analysis, modeling, and forecasting
Spatial analytics	Oracle Spatial's analysis and mining capabilities include binning, pattern detection, spatial correlation, colocation mining, and spatial clustering, topology & NW data model analytics - shortest path, minimum cost spanning tree, nearest-neighbors analysis, traveling salesman problem, etc
Text Mining	Std SQL to index, search, analyze text / documents stored in DB, files, and web with automatic classification and clustering

Predictive in Nature?

Hindsight	Insight	Foresight
What is happening?	Why is it happening?	What will / should happen?
❑ Historic orientation ❑ Typical MIS Reporting or BI ❑ Oracle Reports, Hyperion, IBM Cognos, SAP BO, etc	❑ Business / Behaviour Analysis, Trends ❑ What is currently happening / Why?	❑ Forecasting ❑ Optimization ❑ Past behaviour to predict future outcomes

Data Mining Process



- Data Access
- Data Sampling
- Data Transformation

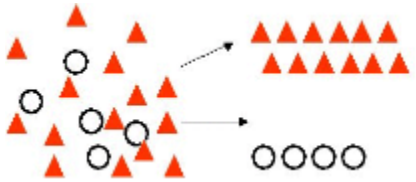
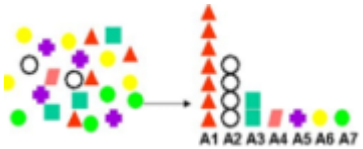
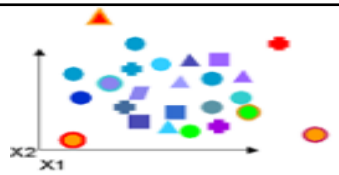
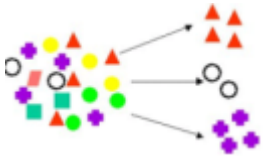
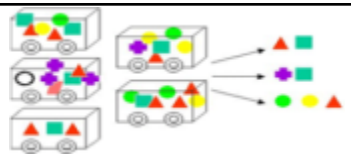
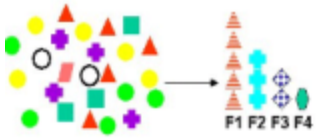
- Create Model
- Test Model
- Evaluate & Interpret Model

- Apply Model
- Custom Reports
- External Applications

Closely Inter-related Concepts

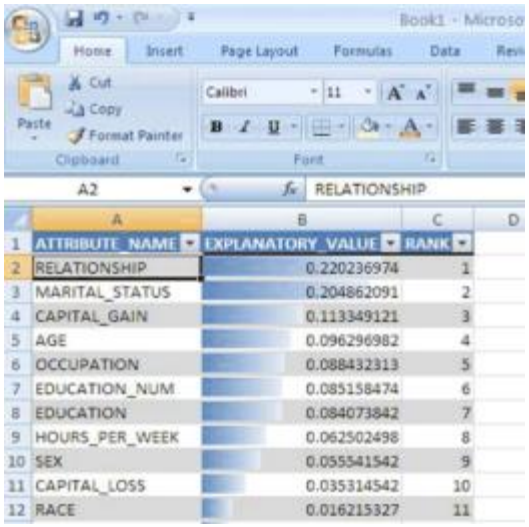
- **Artificial Intelligence (AI)** refers to systems that exhibit autonomous intelligence or behaviour
- **Machine Learning (ML)** are AI techniques that enable devices to learn from their performance and modify their functioning
- **Data Mining (DM)** applies ML concepts to data
  - Each **Data Mining Function (DMF)** specifies a class of problems that can be modelled and solved.
    - Data mining functions fall generally into two categories: **Supervised and Unsupervised**
  - Each **Data Mining Algorithm (DMA)** is a Mathematical Procedure for solving a class of problems

Popular DMFs & DMAs - supported by Oracle

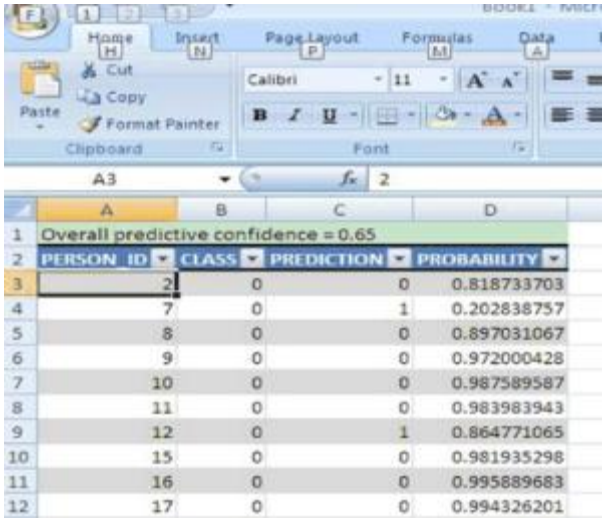
Functions		Some Examples	Algorithms
Classification		▪ High, Medium or Low Value customer ▪ Likely Buy / No-Buy	❑ Logical Regression ❑ Naïve Bayes ❑ Support Vector M/c ❑ Decision Tree
Regression		▪ Customer Lifetime Value ▪ Process Yield Rates	❑ Multiple Regression ❑ Support Vector M/c
Attribute Importance		▪ Medical diagnosis factors ▪ Buyer priorities	❑ Minimum Description Length
Anomaly Detection		▪ Insurance Frauds ▪ Tax compliance	❑ One-class Support Vector Machine
Clustering		▪ Customer segmentation ▪ Life Sciences Discoveries	❑ Enhanced K-Means ❑ Orthogonal Partitioning Clustering
Association		▪ Product Bundling ▪ Defect Analysis	❑ Apriori
Feature Extraction		▪ Pattern Recognition ▪ Data Projection	❑ Non-Negative Matrix Factorization

Oracle Predictive Analytics

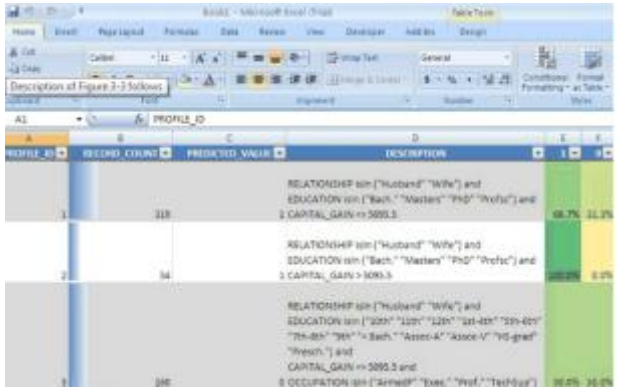
EXPLAIN	PREDICT	PROFILE
Explains how the individual attributes affect the variation of values in a target column	For each case, predicts the values in a target column	Creates a set of rules for cases that imply the same target value



	A	B	C
1	ATTRIBUTE_NAME	EXPLANATORY_VALUE	RANK
2	RELATIONSHIP	0.220236974	1
3	MARITAL_STATUS	0.204862091	2
4	CAPITAL_GAIN	0.113349121	3
5	AGE	0.096296982	4
6	OCCUPATION	0.088432313	5
7	EDUCATION_NUM	0.085158474	6
8	EDUCATION	0.084073842	7
9	HOURS_PER_WEEK	0.062502498	8
10	SEX	0.055541542	9
11	CAPITAL_LOSS	0.035314542	10
12	RACE	0.016215327	11



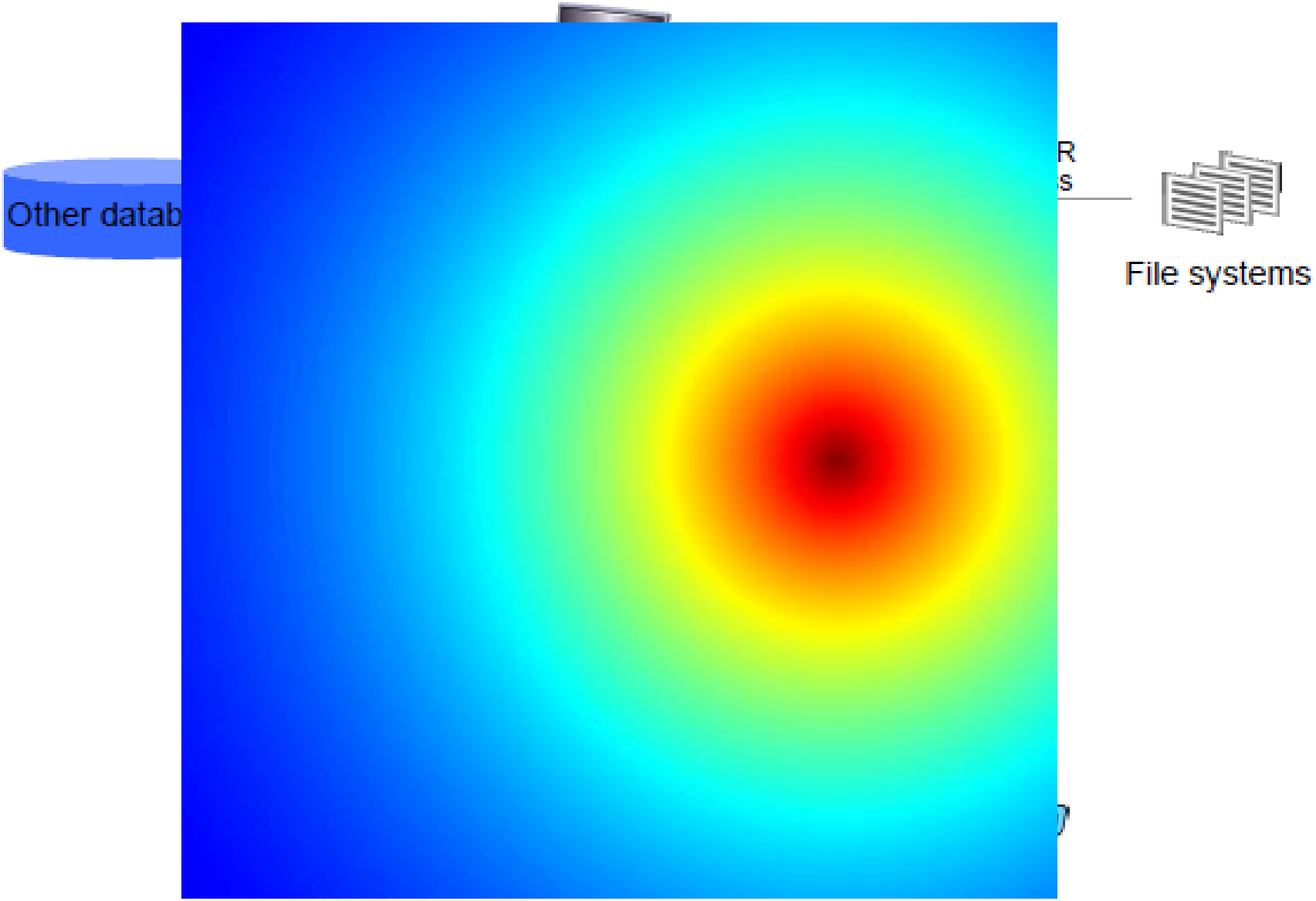
	A	B	C	D
1	Overall predictive confidence = 0.65			
2	PERSON_ID	CLASS	PREDICTION	PROBABILITY
3	2	0	0	0.818733703
4	7	0	1	0.202838757
5	8	0	0	0.897031067
6	9	0	0	0.972000428
7	10	0	0	0.987589587
8	11	0	0	0.983983943
9	12	0	1	0.864771065
10	15	0	0	0.981935298
11	16	0	0	0.995889683
12	17	0	0	0.994326201



Oracle Add-in Spread sheets for Predictive Analytics

Oracle Data Mining implements Predictive Analytics in PL/SQL & Java APIs

Oracle's "Open" Secret Sauce for Predictive Analytics of Big Data

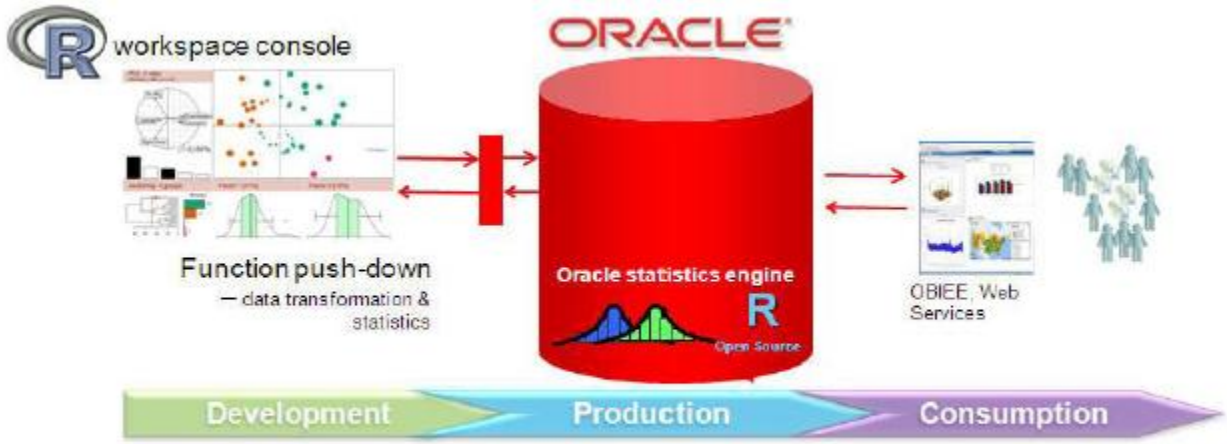


Source: Wikipedia

How does it work?

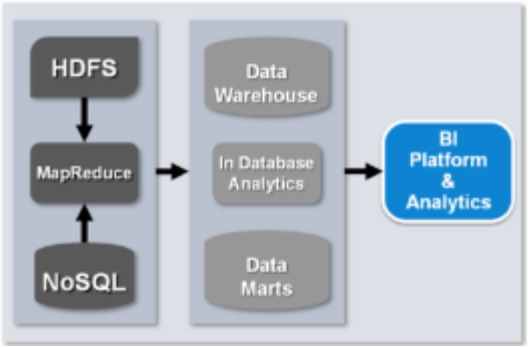
- Coupled with the power of SQL, Oracle Data Mining eliminates data movement and duplication, maintains security and minimizes latency time from raw data to valuable information

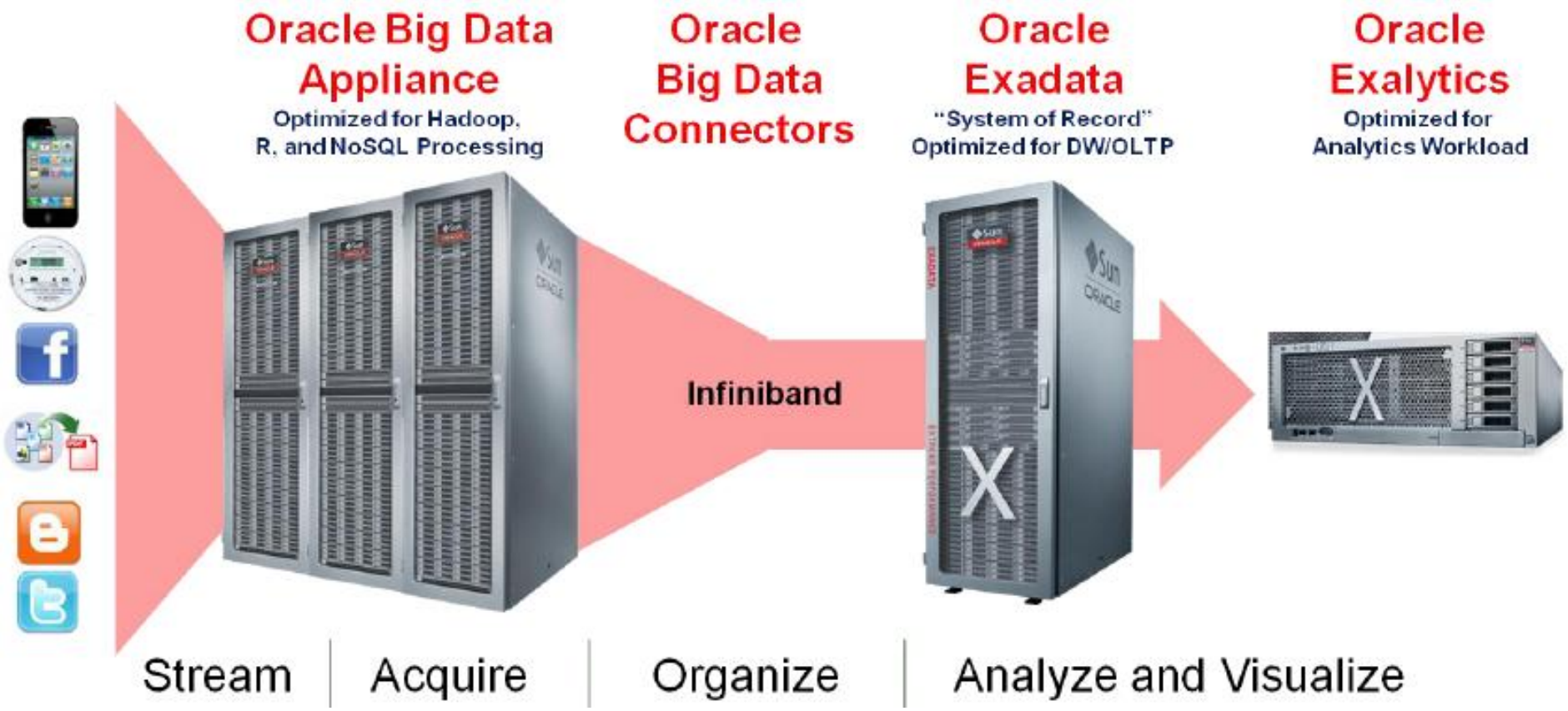
- Oracle Advanced Analytics Option, makes Oracle tables/views accessible to the R environment as if they are native R objects and transparently leverages the massive scalability of the database for big data analysis problems.



Business Intelligence and Advanced Analytics integrated

Oracle Technologies for –
Big Data Predictive Analytics





My Oracle Big Data Benefits

- **It's actually done** and **complete** unlike others
- **Full Hadoop integration and loader**
- **Exadata and Exalytics** BI integration & solution
- **Big Data hardware** which includes Hadoop HDFS, MapReduce, R programming language (statistics and regressions…etc.), Oracle NoSQL, ACID compliant, Simple key-value pair data model (hashes keys over many servers - major/minor keys & byte arrays)
- **Based on Oracle's BerkeleyDB** (commercial 8 years!) which integrates with HDFS (Hadoop File System) using external tables if you want,
- Oracle Loader for Hadoop (OLH) takes the analyzed data from MapReduce & **puts into 11g Database as last step** (easier to do)
- **Concurrency is flexible** at any level & it's horizontally scalable
- Oracle knows clustering & HA well (**no single point of failure!**)
- Oracle **Admin tools** are great as are Oracle professionals
- **BerkeleyDB is the worlds most widely used DB toolkit** >200M deployed copies

Rolta Leaders in Big Data Analytics and Worldwide Oracle Platinum Partner

- ✓ Oracle Partner-of-the-Year: multiple times
- ✓ Oracle user group leadership
  - Past president of International Oracle User Group
  - Member of Applications & Technology Advisory Councils
  - Current president of Midwest Oracle User Group
    - Service Oriented Architecture (SOA)
    - Fusion
    - Oracle 11g
- ✓ Oracle Magazine Consultants-of-the-Year

Industry Recognitions

- ✓ Nine Times Oracle Titan Award Winners
- ✓ 6 “Oracle Masters” on staff

ORACLE® Platinum Partner

One of first few companies worldwide
with highest level of partner certification for Apps & Technology

Final Thoughts… Catching your Wave!

"Things may come to those who wait, but only the things left by those who hustle."

— Abraham Lincoln

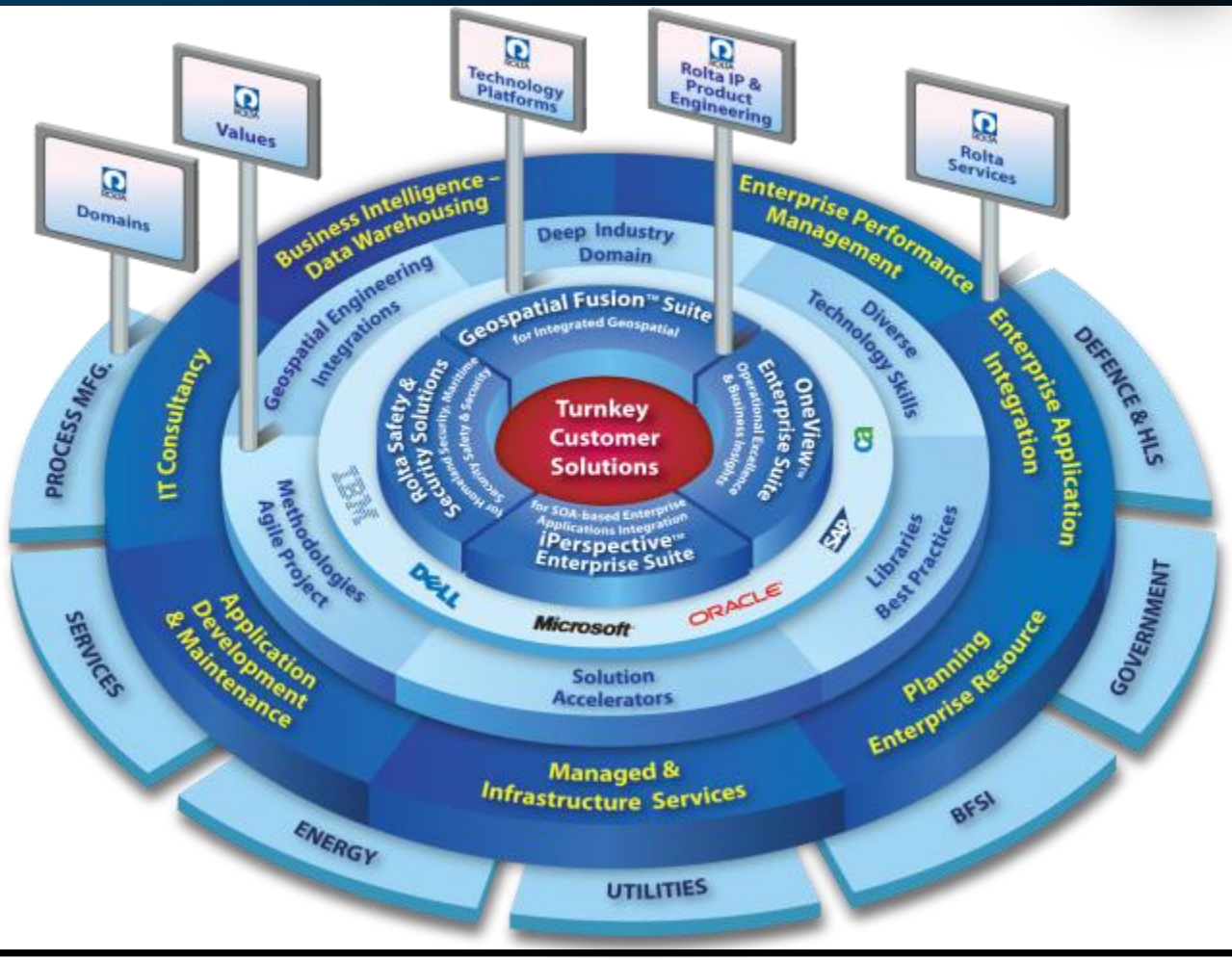
Rolta– *Your* Partner in Success….

2012 Oracle Partner of the Year
(9 Titans/Excellence Awards)

Prior Years: 2002, 2004*, 2007*, 2008, 2010, 2011

***Won 2 Awards**

Thank You & Make a Difference in the World!



ROLTA

CMM LEVEL 5

ISO 9001 : 2008 CERTIFIED

ISO/IEC 20000-1:2005

ISO/IEC 27001:2005

ISO 14001:2004
OHS 18001:227

Copyright

- Neither Rolta nor the author guarantee this document to be error-free. Please provide comments/questions to rich.niemiec@roltasolutions.com. I am always looking to improve!
- Rich Niemiec/Rolta ©2013. This document cannot be reproduced without expressed written consent from Rich Niemiec or an officer of Rolta, but may be reproduced or copied for presentation/conference use.
- References include Rich Niemiec's Exadata Presentation & Oracle11g Database Performance Tuning Tips & Techniques book, www.oracle.com, en.wikipedia.org, slashgear.com, gifsoup.com, www.amazon.com, Tech Crunch, www.rolta.com, The Matrix movie, Information Week, Gartner, Computerworld, & Oracle OpenWorld

Contact Information

Rich Niemiec: rich.niemiec@roltasolutions.com

www.rolta.com