

# Ground System Architectures Workshop



"Innovation on the Ground"

Manhattan Beach Marriott, Manhattan Beach, Calif.

**March 1–4, 2010**

## Oracle Exadata as a Research Platform

John C. Hax – Oracle Corporation, Member IEEE

ORACLE

# Science – A product of data analysis

“Science does not result from the launch of a mission or the collection of data. Rather, science only occurs through the analysis and understanding of that data.”

- Philosophy of the NASA Science Mission Directorate (SMD)

# Oracle's R&D Presence

- National Ignition Facility – Fusion and Laser Research
  - Database, SecureFiles, Orchestration and Middleware, Virtualization, Dataguard, Grid Control, Storage Management, Partitioning
- CERN/Large Hadron Collider
  - Database, Streams, Dataguard, Grid Control, Storage Management, Partitioning
- Max Planck Institute
  - Database, SecureFiles, Dataguard, Grid Control, Storage Management, Partitioning
- NBII.gov – National Biological Information Infrastructure
  - Middleware, Portal, Spatial
  - <http://www.nbio.gov/portal/server.p>
- Jet Propulsion Lab
  - Database, Grid Control, Partitioning, Storage Management

# Future of Scientific Computing and Analysis

Data Intensive

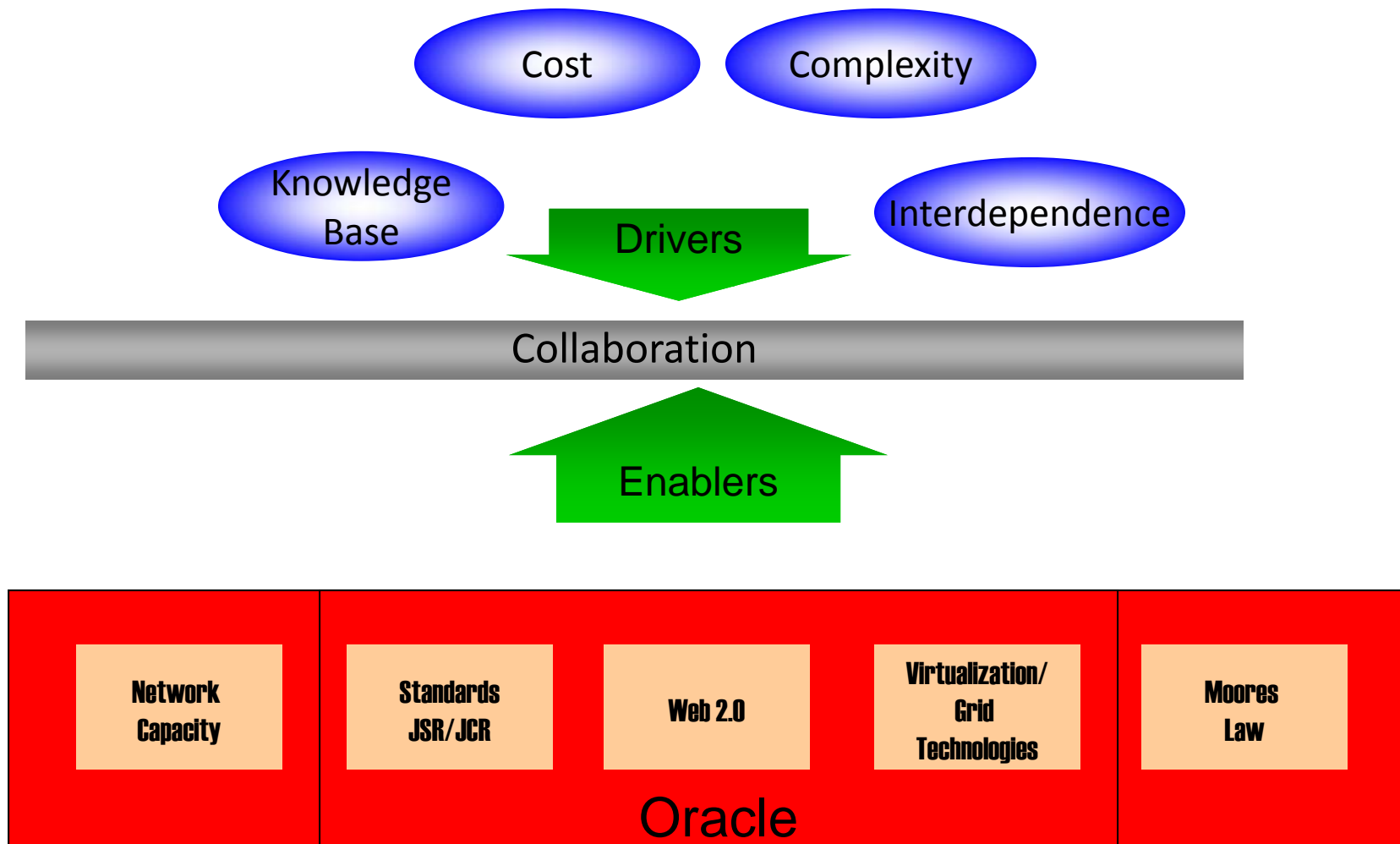
+

Collaborative



Data Intensive Collaborative Science

# Data Intensive Collaborative Science



# Data Challenges for Science

- Stewardship - the long term preservation of data so as to ensure its continued value for both anticipated and unanticipated uses
- Integrity/Provenance - data is complete, accurate, verifiable, if possible reproducible
- Accessibility - availability of research data to researchers other than those who generated the data when the data is needed
- Privacy- ensuring data is accessed in an appropriate manner in a verifiable manner by the appropriate people or resources

# Use Cases for Data Sharing

- Re-analysis
  - New or existing data for same problem
- Secondary Analysis
  - Re-use of same data for different problem
- Replication
  - Different data to study same problem
- Verification
  - 3rd party re-analysis using existing initial data.

# Collaborators

- Initial Investigators
- Subsequent Analysts
- Scientific Community
- Funding Agencies and Foundations

# Obstacles to Data Sharing

## Human

- Lack of Foresight
- Fear of Conflicting Conclusions
- Breach of Confidentiality
- Greater Influence
- Compromising of Potential Profits

## Systematic

- Project Level Funding
- Origination Rules
- Lack of Guidelines
- Lack of Standards
  - Classifying
  - Archiving
  - Documenting
  - Metadata

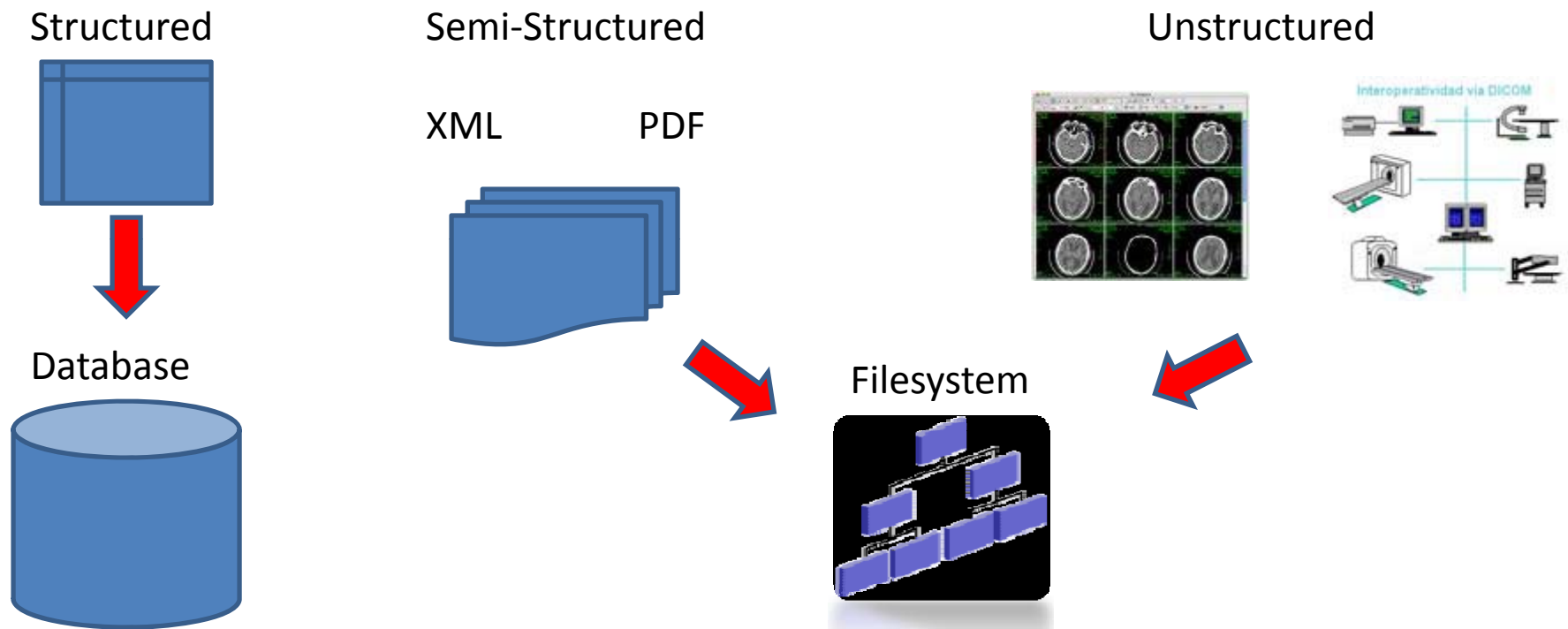
# Technical Obstacles to Collaboration

- Stovepiped/Desktop Systems
- Lack of Institutional IT Support
- Informal Data Sharing Mechanisms
- Lack of Expertise

# Data Challenges to Collaboration

- Physical Limitations
  - I/O Intensive - limitations on max IOPS
  - Network speeds/cost - time/cost to ship data to compute nodes
- Multiple Data Silos
  - Governance issues
    - Pedigree of the data
    - Multiple access policies to get to the data
    - Duplicate data stored in each silo
  - Need to scale disparate systems as data grows
- Increased effort required for Scientists, Developers, Administrators
  - Correlating the data across data silos
  - Coordinated backup and recovery plan
  - Multiple Data Aggregation Efforts

# Research Organizations need to efficiently store, analyze and manage all data



Simplicity and performance of file systems makes it attractive to store file data in file systems, while keeping relational data in DB

# Problem with File Systems (bfiles)

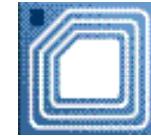
## The Split Architecture – a step in the wrong direction

- Many applications manipulate both files and relational data
  - Rich user experience, compliance, business integration
- This split compromises the **value of the data**.
  - Difficulty merging data
  - Inability to perform Federated Searches
  - Legacy of Stove Piped Data
  - Disjoint security and auditing models
  - Changes cannot be made atomically
  - Backup and recovery are fragmented
  - Search across relational data and files is difficult
  - Space management is complicated
  - Separate interfaces and protocols
  - Application architecture more complex

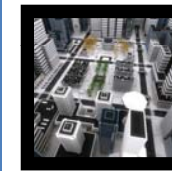
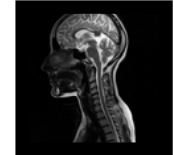
# Integrating Unstructured Data



## New in Oracle Database 11g



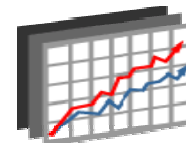
RFID



3D



Binary XML



SecureFiles

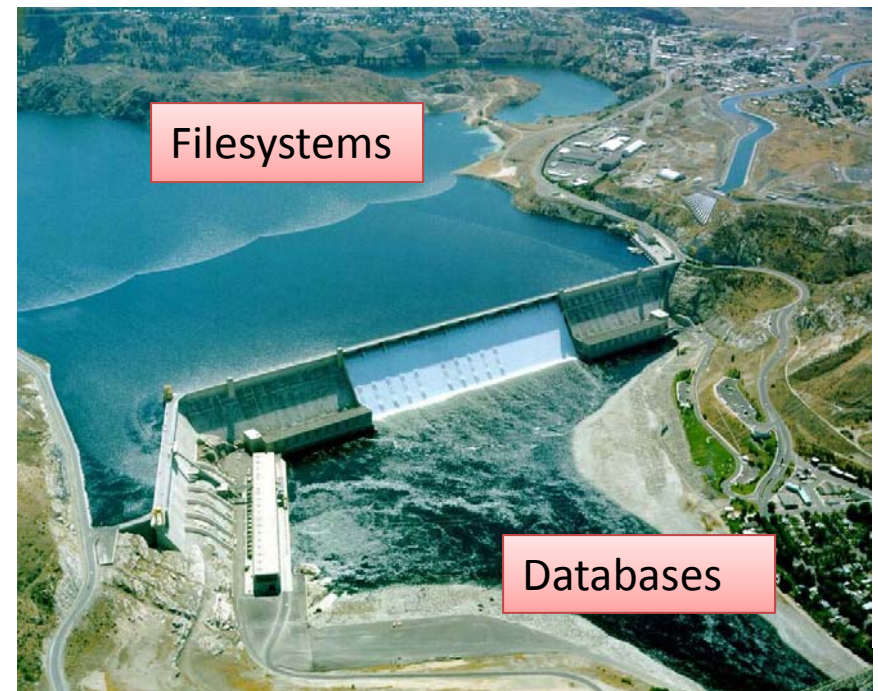
DBFS

# Disparate Data Types

<u>Dataset Category</u>	<u>Examples</u>	<u>Data Type</u>
Optics Metrology	Optics Measurements	XML, Other
Production checklists	LRU manufacturing checklist	XLS
Calibration	Eng Node Sensitivity, Cal ATP	XML, Other
OI Inspection	DMS, IMS, CIM, VIDAR labs	Images(jpeg, GIF)
OI Inspection – Online	FODI, PODI, LOIS	Images(jpeg, GIF)
Auto Alignment	AA Samples	Images
Target Diagnostic Raw	SXI, Dante, FABS	HDF5, Other
Laser Diagnostics Raw	Energy Node, ISP Cal	HDF5, Other
Shot Analysis Results	Analyzed data	HDF5, Other
Operations	Environmental	Scalar

# Database Filesystems

- Bridge the Gap between File systems and Relational Database Systems
  - Maintain Filesystem Performance
  - Leverage multiple access methods
  - Single Security Mechanism
  - Unified Administrative Tools
  - Data Pedigree
  - Unified Architecture and Skill sets
  - Leverage Institutional Resources for IT
  - Enabling Collaboration around Data
  - Optimized for Data Access



# Database Filesystems

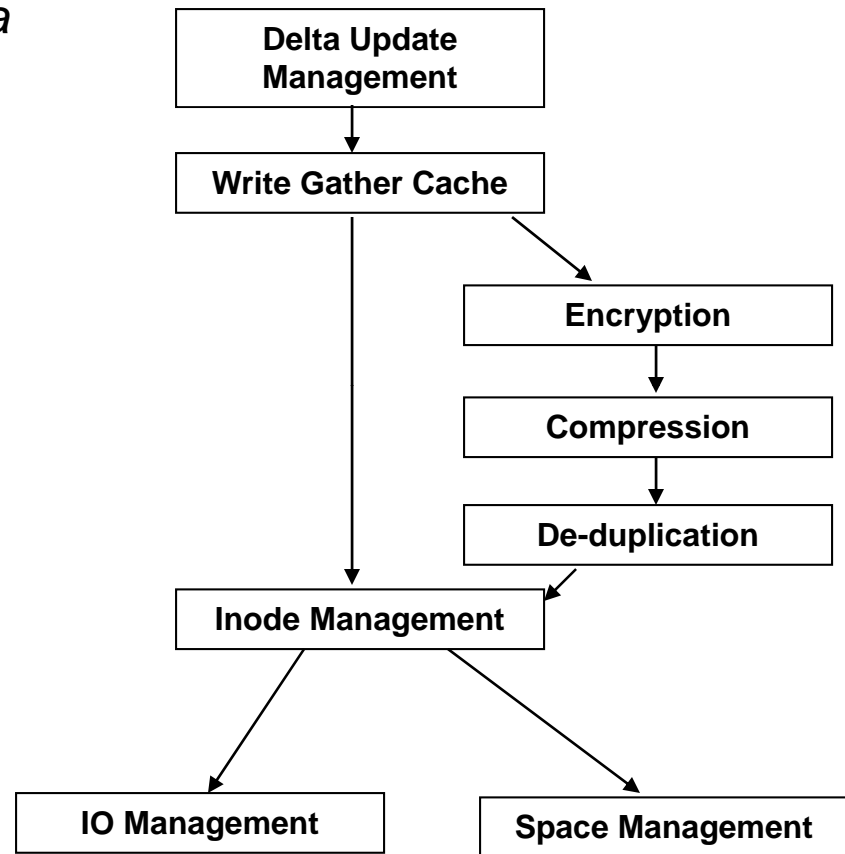
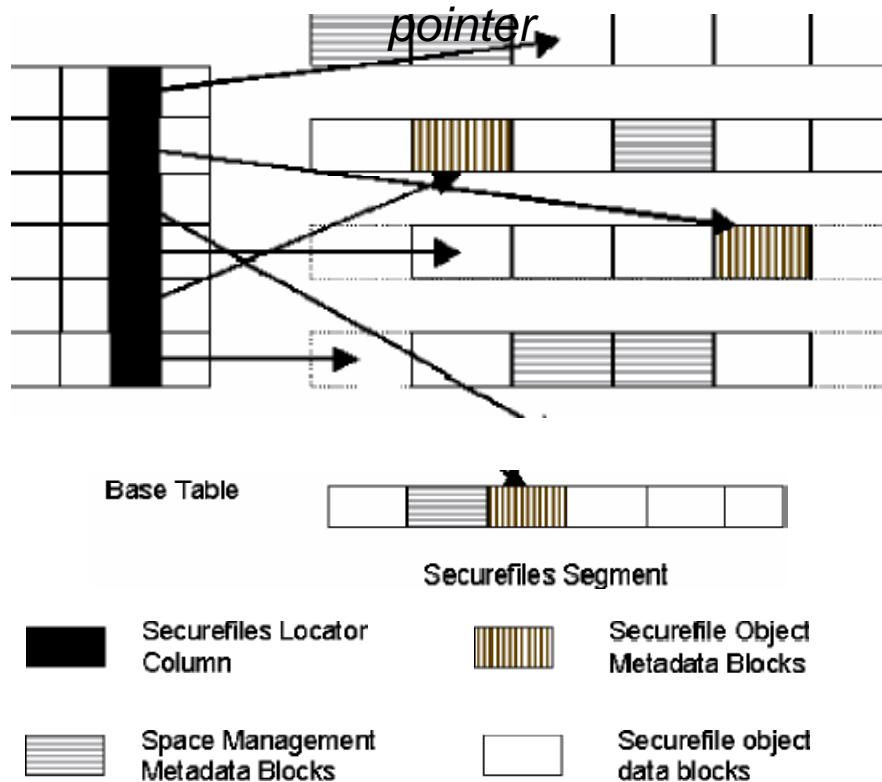
- DBFS is a file system in the database, uses database for storage and brings all of database technology to file systems
- Fuse Client
- DBFS implements the file system interfaces:
  - 2 methods (getpath, list) for a read only file system
  - 5 methods for a file system with read and write support
  - 15 methods for fully functional POSIX file system
- DBFS interface is extensible for easily defining special purpose implementations (providers)
  - DBFS can surface one or more DB tables as a filesystem or a single table through multiple file systems
  - Example, a CheckImages table can have 2 filesystems on it:
    - /CheckImages\_by\_customer/CustomerName/check.jpg
    - /CheckImages\_by\_date/2008/September/check.jpg

# Database Filesystems built on SecureFiles Technology

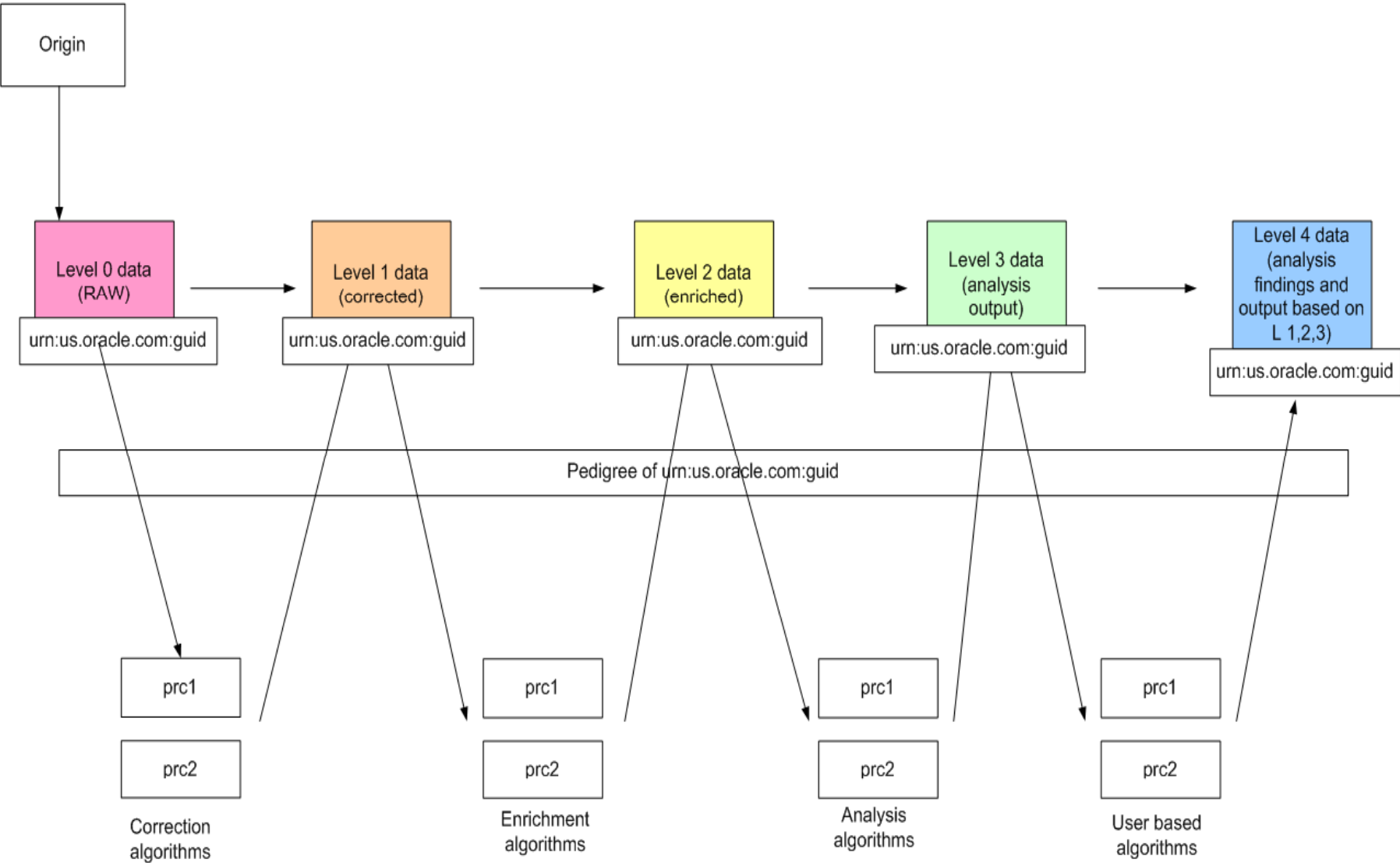
- A new database feature designed to break the performance barrier keeping file data out of databases
- Similar to LOBs but much faster, and with more capabilities
  - Transparent encryption (with Advanced Security Option)
  - Compression, deduplication (with Advanced Compression Option)
  - Preserves the security, reliability, and scalability of database
  - Superset of LOB interfaces allows easy migration from LOBs
  - Enables consolidation of file data with associated relational data
    - Single security model
    - Single view of data
    - Single management of data

# SecureFiles Detail

*Base Table – Oracle table holding metadata plus locator columns similar to a b-file*



# Pedigree with a database filesystem



# Goals of Research Platform

- Optimized for Collaboration
- Optimize for Active Archive
- Minimize Costs
- Extensible Compute Framework
  - Institutional Cloud and External Cloud
- Implements Best Practices
  - Metadata
  - Standards
  - Institutional

# Oracle Exadata

Oracle Exadata provides a mid range capacity computing platform that can meet the needs of many data intensive scientific programs at a cost much lower than traditional scientific platforms. When combined with additional compute nodes, Exadata can scale to meet both compute intensive and IO intensive scientific program requirements.

# Definitions

- **Capacity Computing:** Using smaller and less expensive clusters of systems to run parallel problems requiring modest computational power
- **Capability Computing:**  
Using the most powerful supercomputers to solve the largest and most demanding problems with the intent to minimize time-to-solution

# Modern databases have much to offer in the realm of data analysis

- RDF/OWL can allow semantic searching of data
- Predictive Analytics
- Spatial Data Analysis
- Text Mining of Unstructured Content

# Some of the native data mining techniques and algorithms available

## Technique

## Algorithms

Classification

Logistic Regression

Naive Bayes

Support Vector Machine

Decision Tree

Regression

Multiple Regression

Attribute Importance

Minimum Description Length

Anomaly Detection

One-Class Support Vector Machine

Clustering

Enhanced K-Means

Orthogonal Partitioning Clustering

Association

Apriori

Feature Extraction

Non-negative Matrix Factorization

# Sun Oracle Database Machine Hardware

- Complete, Pre-configured, Tested for Performance
  - Database Servers
  - Exadata Storage Servers
  - InfiniBand Switches
  - Ethernet Switch
  - Pre-cabled
  - Keyboard, Video, Mouse (KVM) hardware
  - Power Distribution Units (PDUs)
- Ready to Deploy
  - Plug in power
  - Connect to Network
  - Ready to Run Database



# Sun Oracle Database Machine Full Rack

- 8 Sun Fire™ X4170 Oracle Database servers
- 14 Exadata Storage Servers (All SAS or all SATA)
- 3 Sun Datacenter InfiniBand Switch 36
  - 36-port Managed QDR (40Gb/s) switch
- 1 “Admin” Cisco Ethernet switch
- Keyboard, Video, Mouse (KVM) hardware
- Redundant Power Distributions Units (PDUs)
- Single Point of Support from Oracle



# Sun Fire™ X4170 – Database Reference Server

Processors	2 Quad-Core Intel® Xeon® E5540 Processors (2.53 GHz)
Memory	72GB
Local Disks	4 x 146GB 10K RPM SAS Disks
Disk Controller	Disk Controller HBA with 512MB Battery Backed Cache
Network	2 InfiniBand 4X QDR (40Gb/s) Ports (Dual-port HCA) 4 Embedded Gigabit Ethernet Ports
Remote Management	1 Ethernet port (ILOM)
Power supplies	Redundant

# Sun Oracle Exadata Storage Servers

Processors	2 Quad-Core Intel® Xeon® E5540 Processors (2.53 GHz)
Memory	24 GB
Disks	12 x 600 GB 15K RPM SAS OR 12 x 2 TB 7.2K RPM SATA
Flash	4 x 96 GB Sun Flash Accelerator F20 PCIe Cards
Disk Controller	Disk Controller HBA with 512MB Battery Backed Cache
Network	2 InfiniBand 4X QDR (40Gb/s) Ports (Dual-port HCA) 4 Embedded Gigabit Ethernet Ports
Remote Management	1 Ethernet port (ILOM)
Power Supplies	Redundant

# InfiniBand Network

- Unified InfiniBand Network
  - Storage Network
  - RAC Interconnect
  - External Connectivity (optional)
- High Performance, Low Latency Network
  - 80 Gb/s bandwidth per link (40 Gb/s each direction)
  - SAN-like Efficiency (Zero copy, buffer reservation)
  - Simple manageability like IP network
- Protocols
  - Zero-copy Zero-loss Datagram Protocol (ZDP RDSv3)
    - Linux Open Source, Low CPU overhead (Transfer 3 GB/s with 2% CPU usage)
  - Internet Protocol over InfiniBand (IPoIB)
    - Looks like normal Ethernet to host software (tcp/ip, udp, http, ssh,...)

# InfiniBand Network

- Uses Sun Datacenter 36-port Managed QDR (40Gb/s) InfiniBand switches
  - Runs subnet manager and automatically discovers network topology
  - Only one subnet manager active at a time
  - 2 “leaf” switches to connect individual server IB ports
  - 1 “spine” switch in Full Rack for scaling out to additional Racks
- Database Server and Exadata Servers
  - Each server has Dual-port QDR (40Gb/s) IB HCA
  - Active-Passive Bonding – Assign Single IP address
    - Performance is limited by PCIe bus, so active-active not needed
  - Connect one port from the HCA to one leaf switch and the other port to the second leaf switch for redundancy
  - Connections pre-wired in the Factory

# Scaling Out to Multiple Full Racks

- Single InfiniBand Network
- Switch to a “Fat Tree” Topology
  - Valid up to 8 Racks
  - Every “leaf” node inter-connected with every “spine” switch
  - “Leaf” switches not connected with other “leaf” switches
  - “Spine” switches not connected with other “spine” switches
  - Database and Exadata Server cabling unchanged.
  - Inter-rack cabling done at installation time
- Up to 3 Racks
  - Extra cables already included with each DB Machine
- Greater than 3 Racks
  - Longer cables need to be purchased

# InfiniBand Network – External Connectivity

- External connectivity ports for
  - Connect to more Exadata servers for on disk backup
  - Connect to media servers for Tape backup
  - Data Loading
  - Client / Application Access
- Validated InfiniBand cable lengths
  - Up to 5m Passive Copper 4X QDR QSFP cables
  - Up to 50m Fiber Optic 4X QDR QSFP cables (more expensive)
- Use available ports on the two “Leaf” switches
  - 12 in the Full Rack (6 per leaf switch)
  - 36 in the Half Rack (18 per leaf switch)
  - 48 in the Quarter Rack (24 per leaf switch)
  - 32 in the Single Server Configuration

# External Connectivity – Ethernet

- Per Database Machine
- Admin Access
  - 1 port from “Admin” Ethernet switch
  - 1 port from KVM Switch
  - Note – For Database Machine Basic System, there is no KVM or Ethernet switch provided and the ILOM and management ports are connected to data center network directly
- Database / Client / Application Access
  - Minimum 1 port per X4170
  - 2 more Ethernet ports per X4170 available
    - Can use them for bonded client / application access or for additional connectivity

# Conclusion

- The ultimate goal of science is to create new knowledge and new discoveries.
- Oracle has a number of features which can benefit the scientific community and ease the burden of pedigree, data management, and analysis
- Using a database filesystem will enable data intensive collaborative science.
- As new discoveries are made and data volumes increase, it is imperative to have a robust database system that is not only capable of managing the pedigree of that data, but also serve as a knowledge repository for the future.
- Exadata provides and ideal platform for program consolidation and scientific collaboration



# For More Information

<http://search.oracle.com>

or

<http://www.oracle.com/>