

**ORACLE  
BUSINESS ANALYTICS**

**See more. Act faster.**



**ORACLE®**

## **Endeca Information Discovery Technical Overview**

David de Santiago

Principal Sales Consultant Business Analytics – Endeca Information Discovery

# Oracle Endeca Information Discovery

Enterprise Platform for Business Self-Service and Discovery

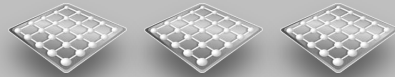
## Studio

Interactive Business Discovery  
Create and Share Apps



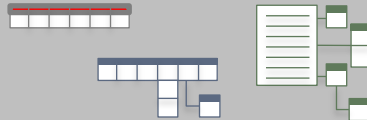
## Endeca Server

Hybrid Search/Analytical Database  
Flexible Data Model



## Integration Suite

Data Integration and Enrichment  
Structured and Unstructured



Helps organizations **quickly explore all relevant data.**

- Combines diverse and changing information from disparate systems.
- Automatically organizes information for exploration and analysis.
- Enables rapid assembly of easy-to-use discovery applications.



# Oracle Endeca Information Discovery

## Endeca Server

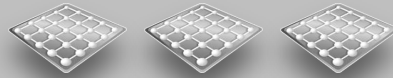
### Studio

Interactive Business Discovery  
Create and Share Apps



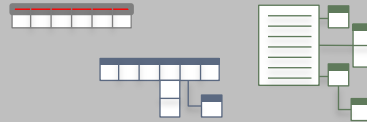
### Endeca Server

Hybrid Search/Analytical Database  
Flexible Data Model



### Integration Suite

Data Integration and Enrichment  
Structured and Unstructured



## Oracle Endeca Server

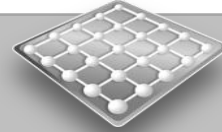
- Hybrid search/analytical database
- Search, navigation, and analytics on diverse, changing information
- Designed for Discovery
  - Flexible data model
  - Columnar storage
  - In-memory analytics



# Endeca Server

Designed for Discovery

Endeca Server



*Flexible **natively semi-structured data model**  
as the foundation for the system architecture*

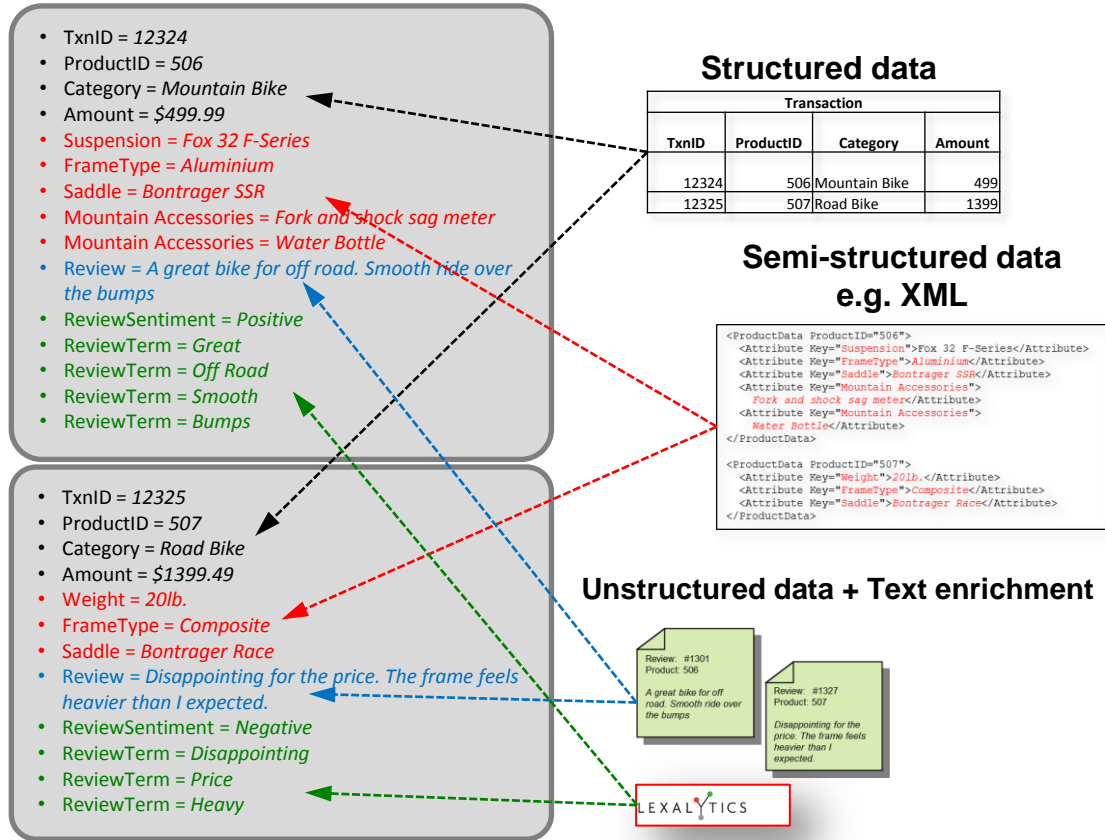
***Unified query capability**  
spanning structured search and navigation,  
analytics, and unstructured search*

***Modern database implementation techniques**  
such as in-memory analytics, columnar physical storage,  
and data parallelism optimized for multicore architectures*

# Endeca Server

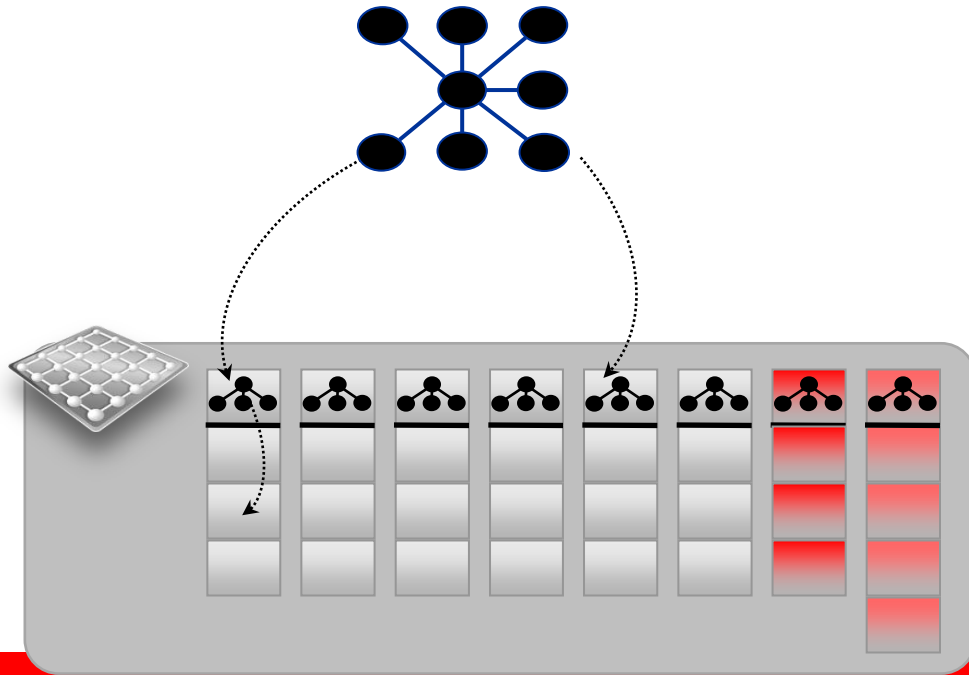
## Flexible Data Model

- Key-Value Store
  - No segmentation into tables
  - No overarching schema
- Simple concepts:
  - Attributes – like columns, except may be sparse, multi-valued, hierarchical
  - Records – each record is a collection of attribute/value pairs
- Accommodates:
  - Idiosyncratic structure... each record is self describing, has its own possibly unique schema
  - Multi-valued fields
  - Large fields of unstructured text



# Endeca Server

Optimized for Exploration and Analysis



## Storage architecture

**High performance column store**, enhanced with:

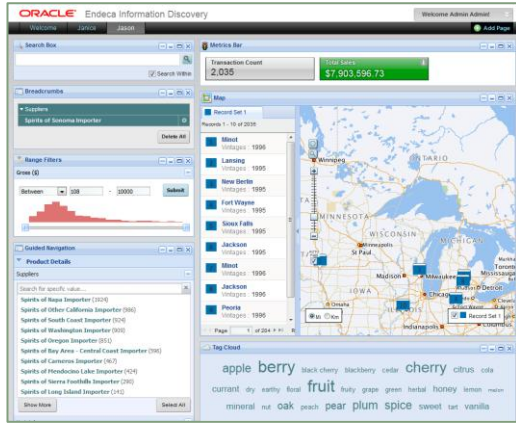
- Integrated index trees
- Membership column
- Full text search indices

## Enhanced in-memory analytics

- Data mapped and cached in memory

# Endeca Server

## Optimized for Exploration and Analysis



## Discovery Capabilities

- **Contextual Navigation**
  - Data filtering and metadata analysis
- **Search**
  - Full text and metadata search
- **Dynamic Data Summarizations / Visualizations**

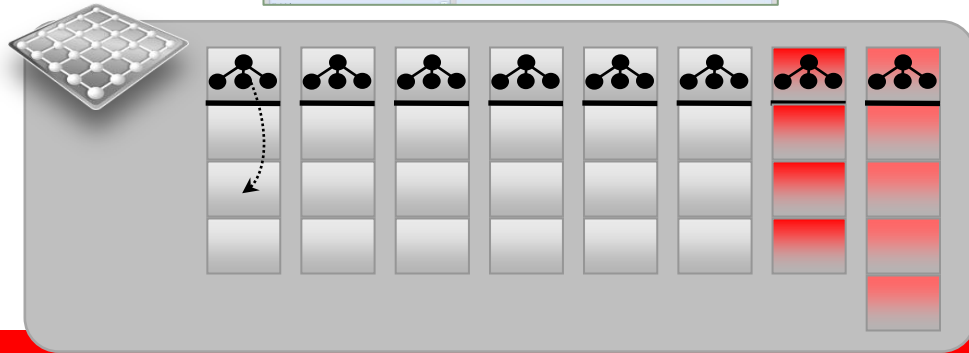
## Storage architecture

High performance column store, enhanced with:

- Integrated index trees
- Membership column
- Full text search indices

## Enhanced in-memory analytics

- Data mapped and cached in memory

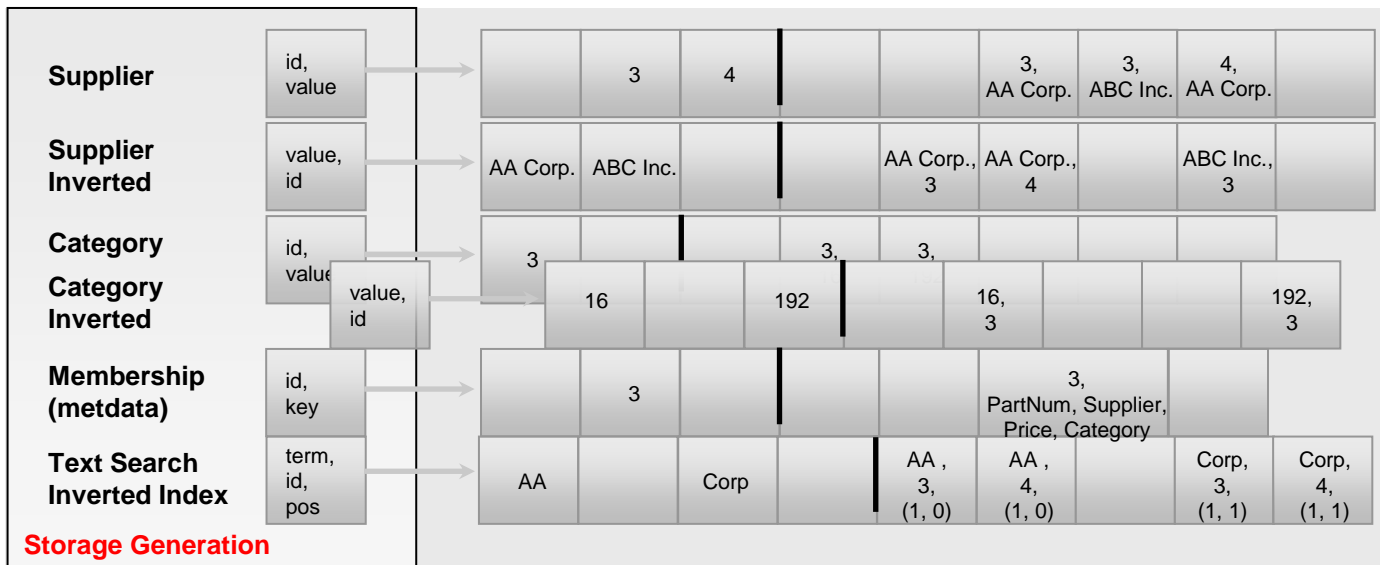


# Endeca Server | Data and Index Structure

```
<record id="3">
  <prop name="PartNum">123-466</prop>
  <prop name="Supplier">AA Corp.</prop>
  <prop name="Supplier">ABC Inc.</prop>
  <prop name="Price">1.75</prop>
  <dval id="192"/>
</record>
```

```
<dimension name="Category">
  <dval id="16" name="Passives">
    <dval id="192" name="Resistor"/>
  </dval>
</dimension>
```

- All physical data storage and indexes based on a columnar storage
- Enables high data compression
- Enables efficient use of I/O, memory, and cache through locality of data access
- Column structures automatically indexed to enable fast seeking and filtering in addition to rapid scanning





# Endeca Server

## In-Memory, But Not Memory-Bound

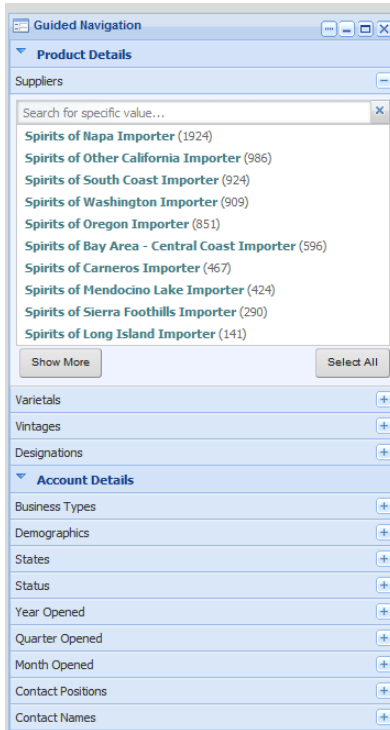


- Storage columns are “mapped” into virtual memory
  - Data resides on disk
  - Fetched into RAM when referenced
- Embedded index trees mean only needed data is scanned
- Data that is referenced will reside in RAM
- Data that is not frequently needed remains on disk
  - Especially common with search indexes and wide records

**Provides key benefits of in-memory analytics,  
without being bound by the size of RAM**

# Endeca Server

## Advanced Analytics Algorithms

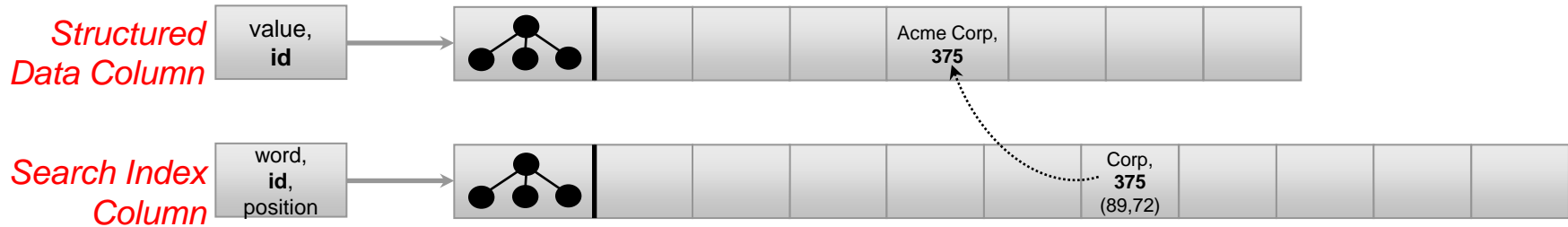


- User experience built to scale with data volume, but also importantly information complexity: wide, sparse dimensionality, with complex jagged hierarchy.
- Requires advanced navigation behaviors:
  - **Automatic Navigation Attribute Detection, Precedence Rules**  
*Detect and present relevant dimensions in current result set*
  - **Attribute Contraction**  
*Hide attributes that do not provide useful navigation options*
  - **Automatic “Implicit Refinement” Detection**  
*Identify refinements that are already implied by in-view records*
  - **Automatic hierarchy drill-down / Refinement Least Common Ancestor (LCA)**  
*Automatically detect maximally specific hierarchy level for navigation*
  - **Dynamic / approximate “top K” refinement ranking**  
*Efficiently and ergonomically handle high-cardinality attributes*

# Endeca Server

## Full Text Search

Search index storage and analysis build on same column storage core as structured data store / indexes



Supports a breadth of structured and unstructured search capabilities:

- Guided Navigation
- Keyword search
- Boolean search
- Parametric search
- Wildcard search
- Dimension search
- Dimension filters
- Dimension precedence rules
- Numeric range filters
- Geospatial filters
- Date/Time filters
- Security filters
- Spell correction/suggestion, DYM
- Find similar
- 1- and 2-way synonyms
- Stemming and lemmatization
- Keyword-in-context snipping
- Results clustering
- Relevance ranking
- Sorting and paging
- Language support

# Endeca Server

## Endeca Query Language (EQL)

- SQL-based syntax for specifying dynamic aggregation queries
- Operates on the intersection results of security constraints and all user-applied search, navigation, and range (e.g. numeric, geospatial) filters

```
DEFINE Customers AS SELECT
  CustomerAge AS CustomerAge,
  CustomerState AS CustomerState
GROUP BY CustomerKey ;

RETURN AvgAges AS SELECT
  AVG(CustomerAge) AS AvgAge
FROM Customers
GROUP BY CustomerState
```

A simple, structured example

```
DEFINE Terms AS SELECT
  COUNT(1) AS Assignments
GROUP BY Term, Category ;

RETURN TermCounts AS SELECT
  COUNTDISTINCT(Term) as NumTerms,
  SUM(Assignments) AS NumAssign
FROM Terms
GROUP BY Category
```

Operating on semi-structured (multi-valued) data extracted from unstructured content

# Oracle Endeca Information Discovery Integration Suite

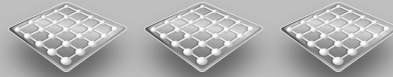
## Studio

Interactive Business Discovery  
Create and Share Apps



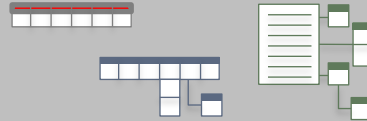
## Endeca Server

Hybrid Search/Analytical Database  
Flexible Data Model



## Integration Suite

Data Integration and Enrichment  
Structured and Unstructured

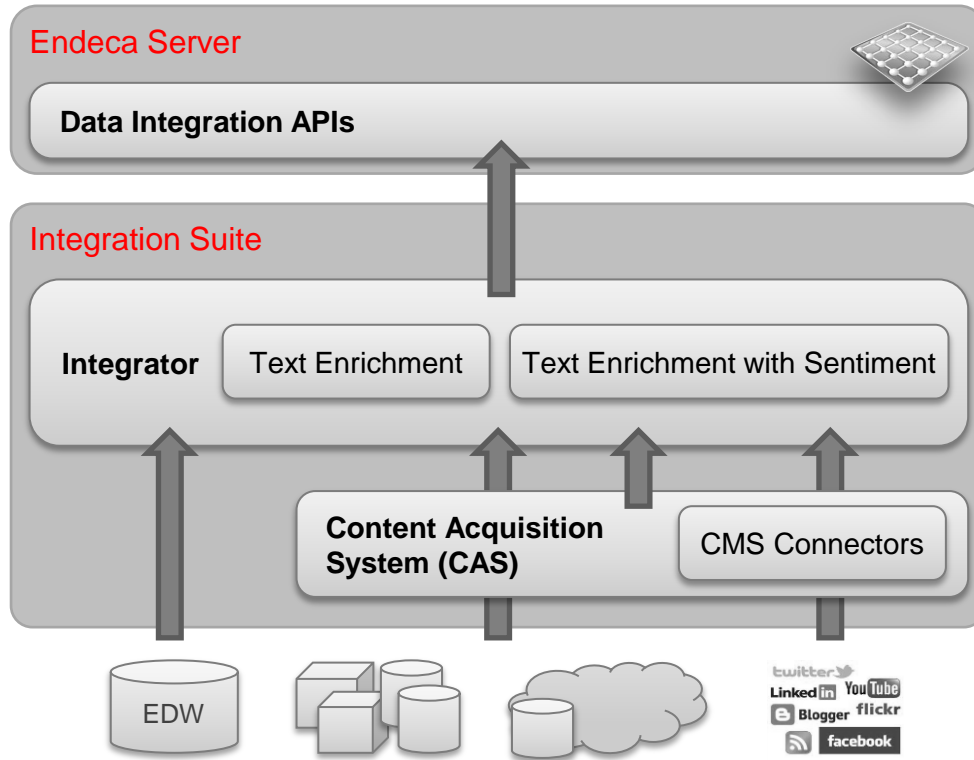


## Integration Suite

- Lightweight ETL with native Endeca Server connectivity
- Java SDK for extensibility – easily connect to additional data feeds and APIs
- Modules for unstructured data:
  - Text Enrichment
  - Sentiment Analysis
  - Content Management System Connectors



# Integration Suite



Quickly load and enrich any data.  
Structured or unstructured.  
Internal or external.

- **Endeca Server**
  - Provides load-and-go web services APIs for incremental, bulk ingest
  - Enables integration with other data tools
- **Integrator**
  - Included agile enterprise ETL tool
  - Adapters for JDBC and common file types (e.g. XML, delimited, fixed-width), library of structured and unstructured data manipulators
  - Production management and monitoring interfaces
  - *Add-on modules for Text Enrichment, Sentiment Analysis*
- **Content Acquisition System (CAS)**
  - Crawls and extracts text from documents in CMS, web, file systems
  - *Add-on modules for common CMS*

# Oracle Endeca Information Discovery Studio

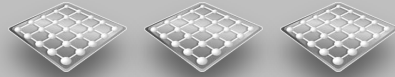
## Studio

Interactive Business Discovery  
Create and Share Apps



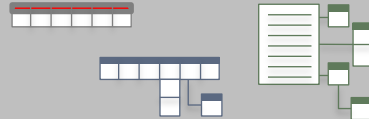
## Endeca Server

Hybrid Search/Analytical Database  
Flexible Data Model



## Integration Suite

Data Integration and Enrichment  
Structured and Unstructured



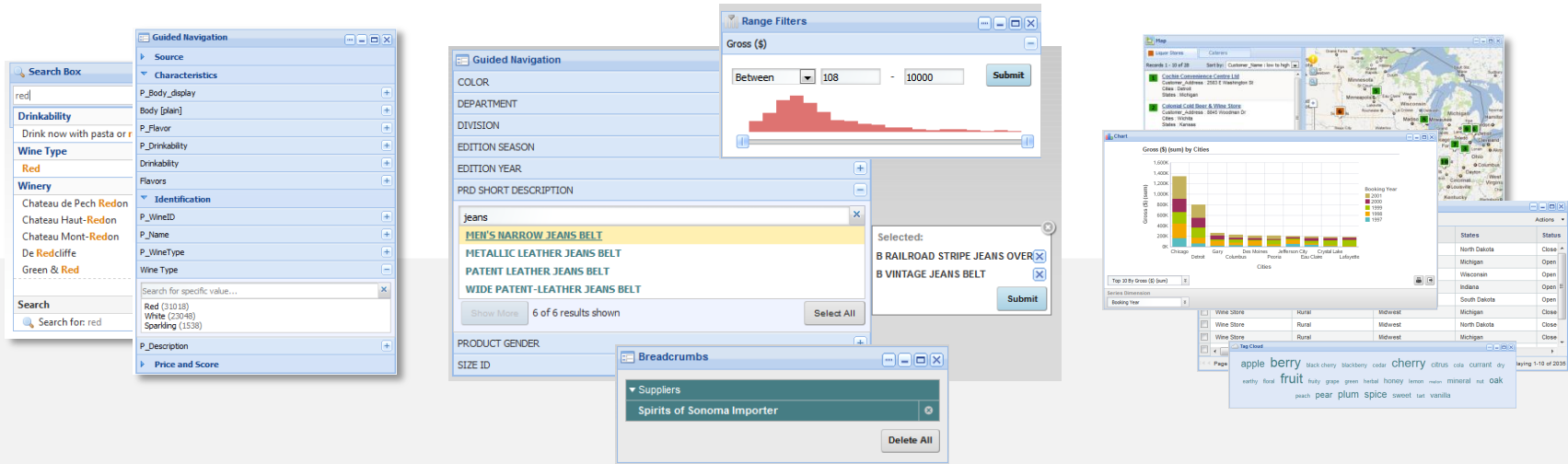
## Studio

- Interactive, visual discovery environment
- Drag-and-drop configuration
- Full-featured component library
- Based on best practice UI design patterns
- Built on Java portal technology standards



# Endeca's Unique User Experience

## Interactive Data Exploration and Analysis



### Advanced Search

- Search across *all* data
- Dynamic typeahead
- Automatic spell correction
- Unlocks unstructured data



### Contextual Navigation

- **Data-Driven.** Freely browse data without predefined paths or writing queries
- **Interactive.** Shows only valid next steps
- **Easy to Use.** Familiar online experience



### Visual Analysis

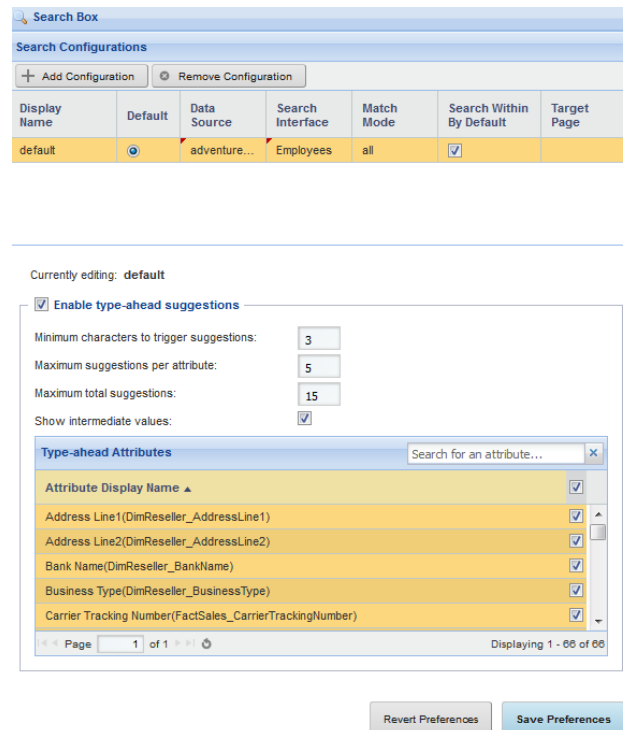
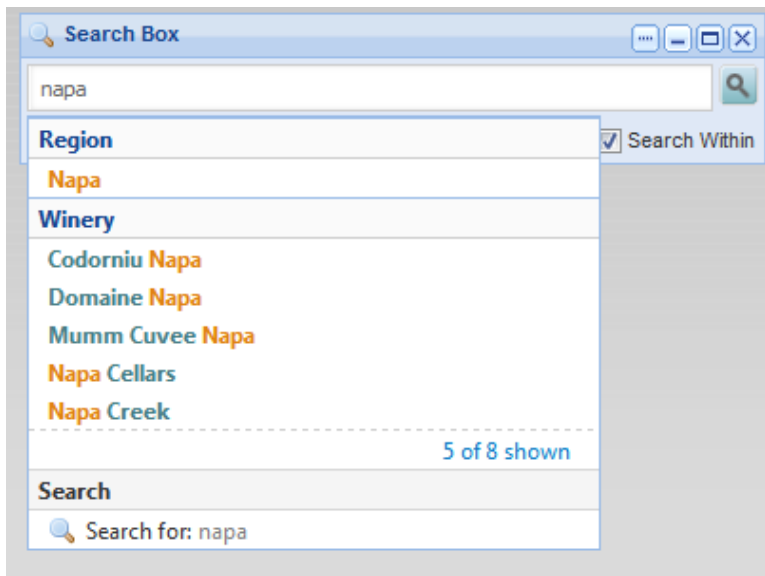
- Charts, crosstabs, key metrics
- Geospatial visualization
- Tag clouds



# Studio

## Component-Based Application Composition

Each component includes a **user control** and an **editor**



# Studio

## Component Library

### Configurable components:

- Alerts
- Bookmarks
- Breadcrumbs
- Chart
- Compare
- Crosstab
- Data Explorer
- Guided Navigation
- Map
- Metrics Bar
- Range Filters
- Record Details
- Results Table
- Results List
- Search Box
- Tag Cloud

The screenshot displays the Studio application interface with several components overlaid. The main window shows a 'Guided Navigation' sidebar on the left with filters for COLOR, DEPARTMENT, DIVISION, EDITION SEASON, EDITION YEAR, PRD SHORT DESCRIPTION, PRODUCT GENDER, and SIZE ID. The central area features a 'Chart' titled 'Gross (\$) (sum) by Cities' showing a stacked bar chart for Chicago, Detroit, Gary, Des Moines, Columbus, and Peoria. Below the chart is a 'Results Table' with columns for States, 1997, and 1998. A 'Map' component shows a map of the United States with blue markers for various cities. A 'Range Filters' dialog is open, showing a histogram for 'Gross (\$)' with a range from 108 to 10000. Other components include 'Alerts' (You have 10), 'Bookmarks' (My Bookmarks table), 'Tag Cloud' (apple, berry, fruit, etc.), and 'Metrics Bar' (Transaction Count: 2,035; \$7,903,596.73). The bottom of the interface shows a 'Page 1 of 204' and 'Records per page' dropdown.

States	1997	1998
Illinois	\$245,956.47	\$398,453.40
Indiana	\$185,869.65	\$222,412.41
Iowa	\$54,575.20	\$127,844.32
Kansas	\$18,885.09	\$37,178.19
Michigan	\$128,790.30	\$282,787.91
Minnesota	\$27,911.96	\$9,417.49
Missouri	\$68,380.89	\$86,346.00
North Dakota	\$32,616.00	\$44,746.20
Ohio	\$56,701.68	\$188,795.80
South Dakota	\$21,936.59	\$1,357.98

Name	Description	Date Created	Actions
apple	black cherry blackberry cedar cherry citrus cola currant dry		
berry	earthy floral fruit fruity grape green herbal honey lemon melon mineral nut oak		
cherry	peach pear plum spice sweet tart vanilla		

Transaction Count	Value
2,035	\$7,903,596.73



# DEPLOYMENT AND ADMINISTRATION

# Discovery Application Lifecycle

Building applications in days, not months

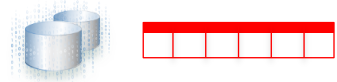
Diverse and changing information integrated and enriched via ETL

Automatically unified in Oracle Endeca Server – no predefined model required

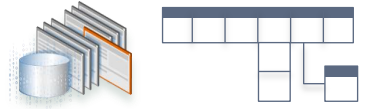
Drag-and-drop application composition in Studio

Interactive search, navigation and visualization for exploration and analysis

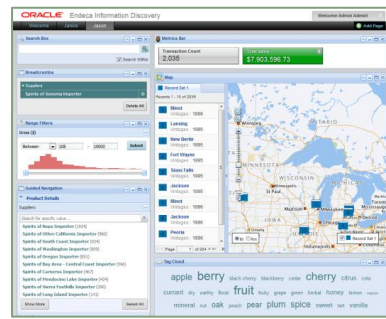
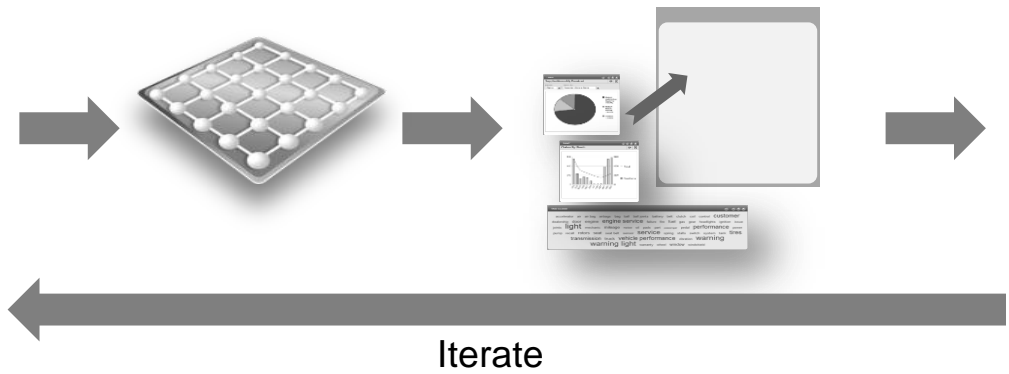
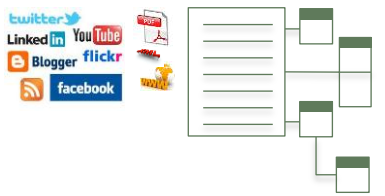
Structured



Semi-Structured





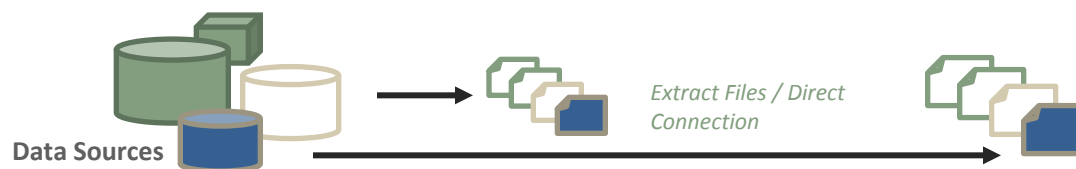
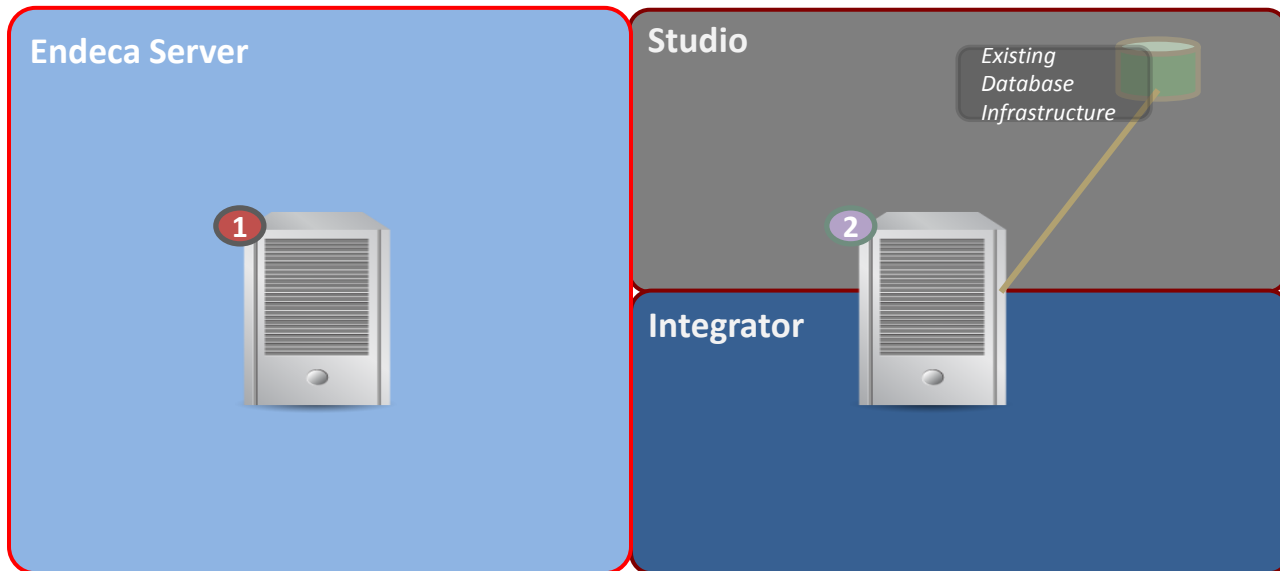
Unstructured



# Example First Release Hardware Topology



- 1 Endeca Server, containing**
  - OS: OEL5 or RHEL 5 Linux 64-bit or Windows 2008 Server Enterprise 64-bit
  - CPU: 4 x (6 / 8 / 10) core Intel Xeon E7 processors
  - RAM: 64 GB
  - Disk: 250GB+ local and/or SAN (preferred)
  - Network: Gigabit Ethernet
- 2 Studio/Integrator Server, containing**
  - OS: OEL5 or RHEL 5 Linux 64-bit or Windows 2008 Server Enterprise 64-bit
  - CPU: 2 x (6 / 8 / 10) core Intel Xeon E7 processors
  - RAM: 32 GB+
  - Disk: 250GB+ local and/or SAN (preferred)
  - Network: Gigabit Ethernet

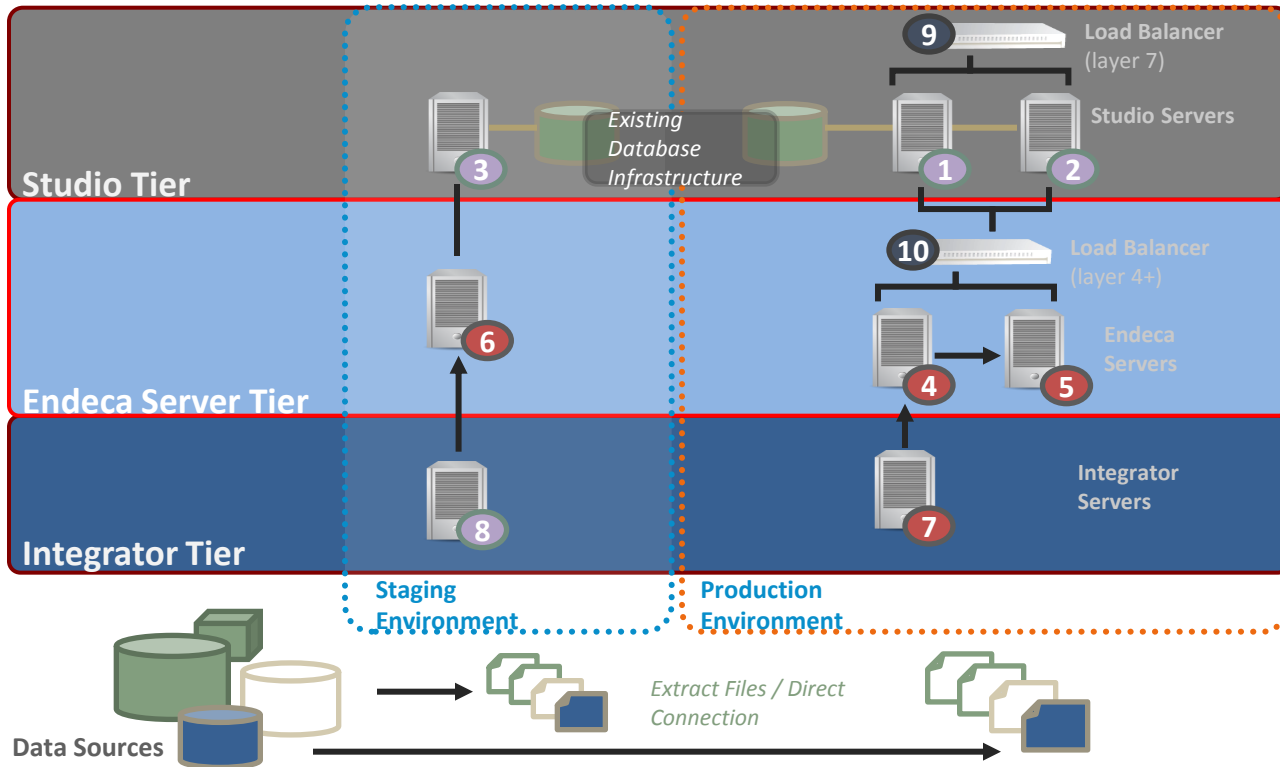
-  Physical server preferred
-  Physical or virtual server



# Example Fully Redundant Hardware Topology

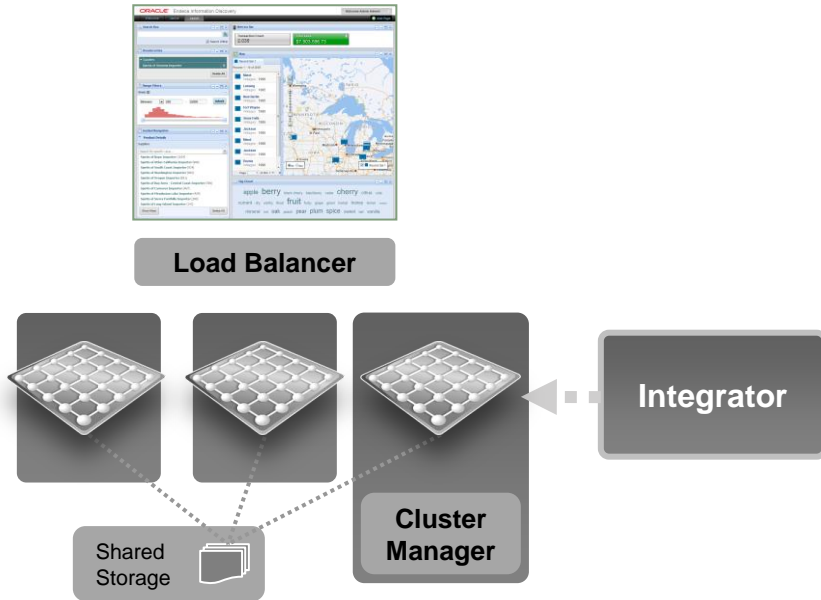
- 1** **3 Studio servers, containing**
- 2** > OS: OEL5 or RHEL 5 Linux 64-bit or Windows 2008 Server Enterprise 64-bit
- 3** > CPU: 2 x (6 / 8 / 10) core Intel Xeon E7 processors  
RAM: 24 GB  
> Disk: 250GB+ local and/or SAN (preferred)  
> Network: Gigabit Ethernet
- 4** **3 Endeca Servers, containing**
- 5** > OS: OEL 5 or RHEL 5 64-bit
- 6** > CPU: 4 x (6 / 8 / 10) core Intel Xeon E7 processors, fastest available clock speed  
RAM: 64GB+  
> Disk: 500GB+ SAN (preferred)  
> Network: Gigabit Ethernet
- 7** **2 Integrator servers, containing**
- 8** > OS: OEL 5 or RHEL 5 64-bit or Windows 2008 Server Enterprise 64-bit
- 8** > CPU: 2 x (6 / 8 / 10) core Intel Xeon E7 processors, fastest available clock speed  
RAM: 32GB+  
> Disk: 1TB+ SAN (preferred)  
> Network: Gigabit Ethernet
- 9** **2 Load Balancers, supporting**
- 10** > OSI Layer: 4+ for MDEX tier and 7 for application tier

-  Physical server preferred
-  Physical or virtual server



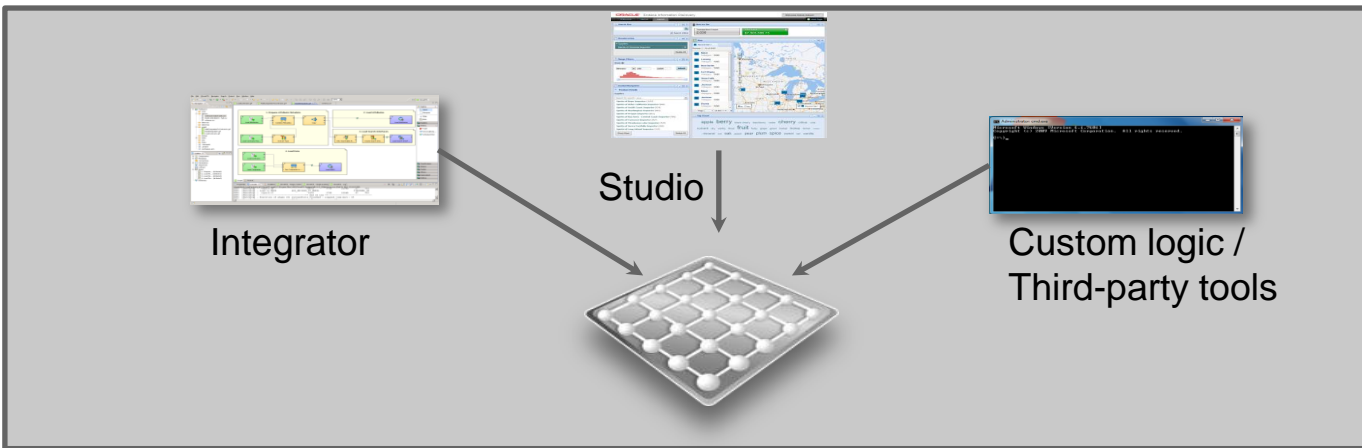
# Endeca Server Clustering

Simple, Fast Redundancy and Throughput Scaling



- Endeca Servers share indices; no local storage
- Data updates and configuration web service calls target a single endpoint
- Cluster Coordinator handles coordination for atomic operations

# Administration and Configuration APIs



Web services for administrative and configuration, including:

Administration:

- Create a snapshot
- List active jobs
- Roll logs
- Ping (load balancer support)
- Get server statistics / performance information

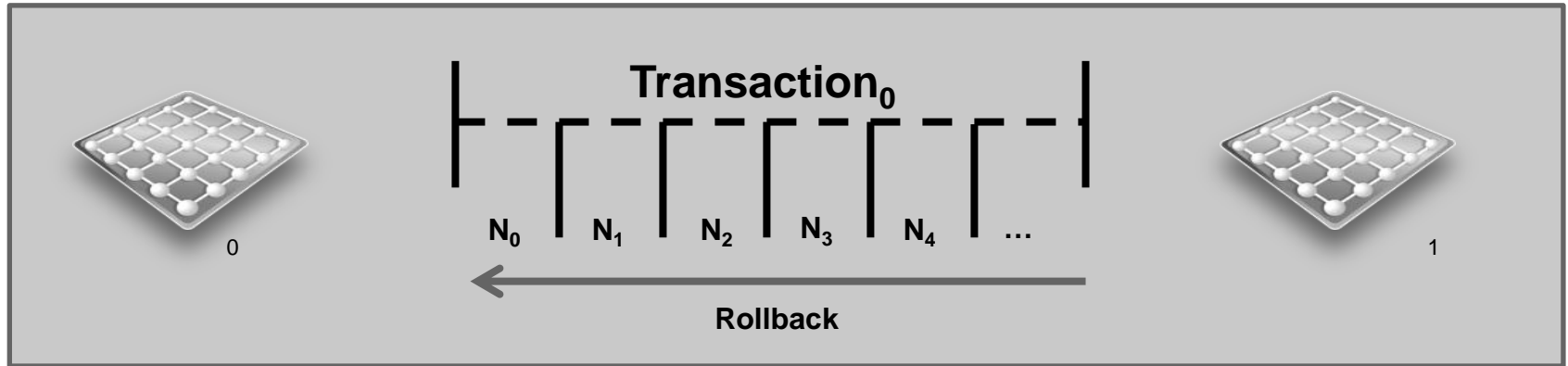
Configuration:

- Manage metadata, such as attribute types
- Manage view creation, backup, and migration
- Indexing options
- Feature configuration, such as precedence rules, thesaurus, etc.
- Export/Import configuration for environment migration

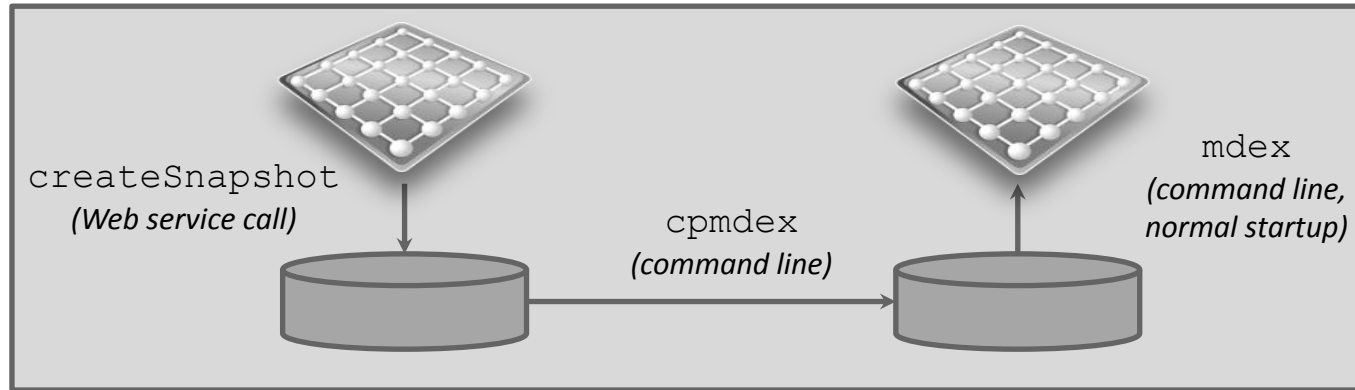


# Transaction Control

- Combine multiple data/configuration web service requests as a single transaction
- Web service for begin, commit, rollback



# Backup and Recovery

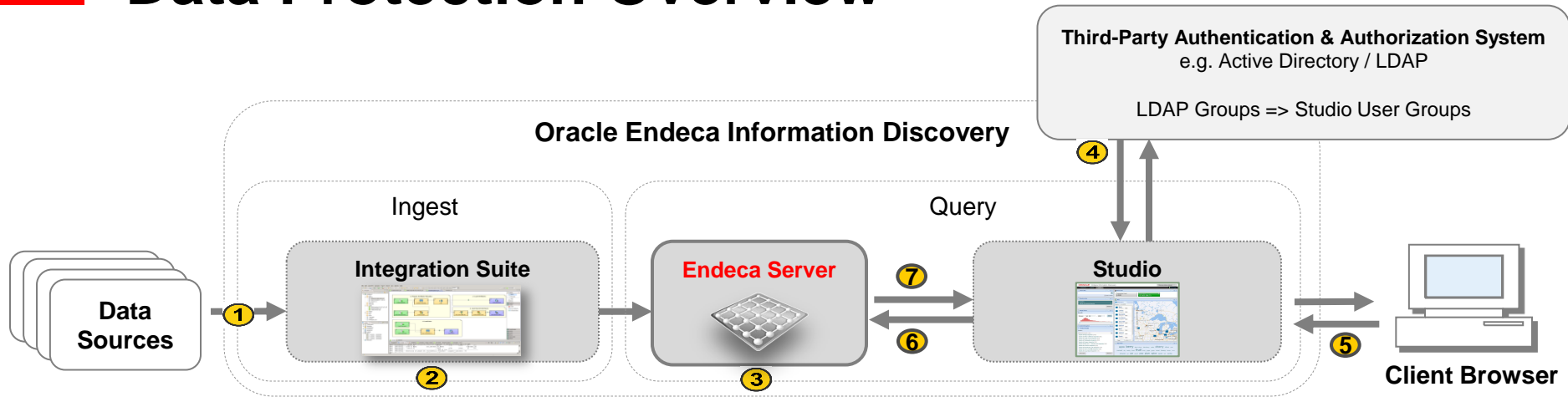


- `createSnapshot` does not disrupt continuous query operations, but is mutually exclusive with data and configuration updates
- Result of snapshot is a simple directory of binary files that can be backed up using any standard file backup approach



# SECURITY

# Data Protection Overview



- 1 Data source(s) may contain not only records, but also metadata describing access rights to those records, such as ACLs.
- 2 On ingest, Integrator tags each record with ACLs and other security metadata.

<b>Record 1</b> Allow: LDAPGroup1 Region: South	<b>Record 3</b> Allow: LDAPGroup2 Region: North
<b>Record 2</b> Allow: LDAPGroup2 Region: South	<b>Record 4</b> Allow: LDAPGroup3 Region: South

- 3 Endeca Server presents security-annotated data for querying.
- 4 End user logs in and receives credentials:
  - **User:** Joe Brown
  - **LDAP:** LDAPGroup1
  - **Region:** South
- 5 Joe does a text-based search

- 6 Studio appends custom security filters, derived from Joe's credentials, to all Endeca Server queries.

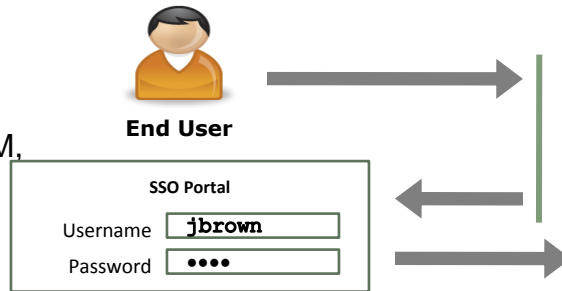
```
Security Filter
OR(Allow:LDAPGroup1,
Region:South)
```

- 7 Query results contain only data that Joe is authorized to see. Additionally, Studio only offers Joe those pages/components that his LDAP group allows.

# Authentication

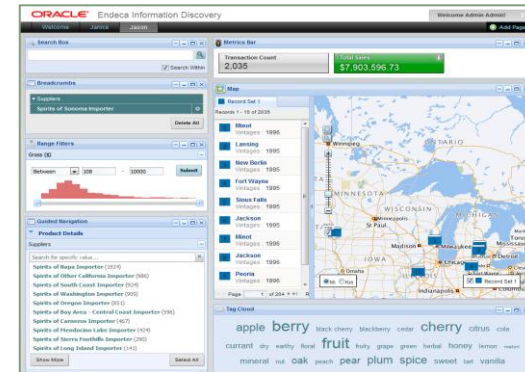
- Oracle Endeca Information Discovery owns authentication: User and role
- User credentials may be validated against external systems

- Configurable plugins for:
  - LDAP/Active Directory, CAS, NTLM, OpenID, OpenSSO, SiteMinder



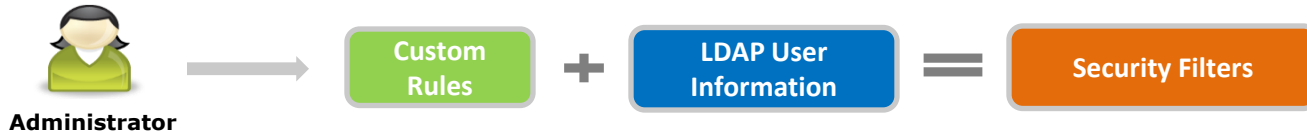
<http://intranet.company.com/secureapp>

- Security provider API for other systems
- Hook for SSO, bypasses login portlet
- Some plugins support one or more:
  - Local cache copy for performance
  - Scheduled import for performance
  - Writeback of changes in management UI



# Authorization / Data Protection

- Data is tagged with access control information (ACLs)
  - Either directly or in a related system
- Users have various individual/group memberships or specific access rights
  - May be retrieved and cached at login from 3<sup>rd</sup>-party system
  - May be combined with administrator-specified rules



# Authorization Levels

1

## Application

- Do users have access to overall application?

2

## Tab/Page

- Do users have access to specific areas?

3

## Record (Row)

- Do users have access to only a subset of Endeca records?

The screenshot displays the Oracle Endeca Information Discovery interface. It features a top navigation bar with the Oracle logo and the text 'Endeca Information Discovery'. Below this, there are several tabs: 'Welcome', 'Janice', 'Jason', and 'kristen'. A 'Welcome Admin Admin!' message is visible in the top right corner. The main content area is divided into several sections:

- Search Box:** Located at the top left, it contains a search input field and a 'Search Within' checkbox.
- Metrics Bar:** Located at the top right, it displays two metrics: 'Transaction Count' (2,035) and 'Total Sales' (\$7,903,596.73).
- Results Table:** The central part of the interface, showing a table with columns for 'Business Types', 'Demographics', 'Regions', and 'States'. A red box highlights a row in this table, indicating a specific record.
- Suppliers:** A section on the left side, listing various suppliers such as 'Spirits of Sonoma Importer'.
- Product Details:** A section on the left side, providing details for selected products.
- Map:** A map at the bottom right, showing a geographical view of the data with markers for 'Minot', 'Lansing', and 'New Berlin'.

Three numbered callouts (1, 2, and 3) are overlaid on the screenshot to indicate authorization levels:

- 1:** Points to the top navigation bar, representing application-level access.
- 2:** Points to the search box, representing access to specific areas or tabs.
- 3:** Points to a row in the results table, representing access to individual records.

# Other Notes

- Endeca Server security filtering includes refinements and spelling corrections
  - Avoids a spell correction inadvertently revealing a code name, for example
  - User sees no evidence of inaccessible data
- Endeca builds on the security of the underlying OS and servers
  - Set up separate user account for Endeca processes and data
  - Isolate Endeca servers from other network segments
  - Mutually-authenticated SSL to secure network traffic (if necessary)
- The Endeca Server data stores constitute a complete copy of the corpus
  - Not encrypted
  - Rely on OS/network to protect sensitive data from unauthorized user access

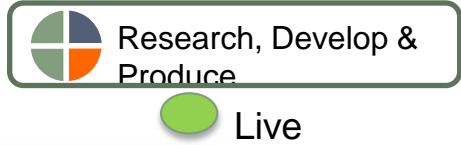




ORACLE®

ORACLE®

# Manufacturing Quality Analysis and Early Issue Detection



## Industry

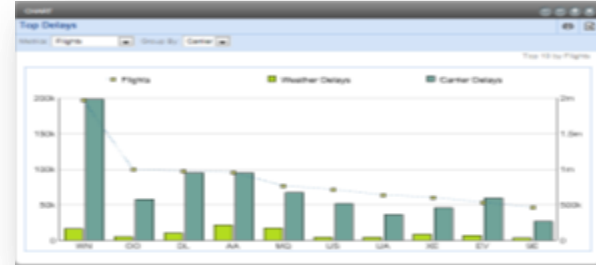
- High Tech

## Business Challenges

- The OEM needed visibility to their products as they went down the assembly line of the contract manufacturers, tracking module test results, product quality and failures.
- Existing approaches took weeks to compile and correlate test data and allowed visibility into only a small set of parameters.

## Overview of Solution

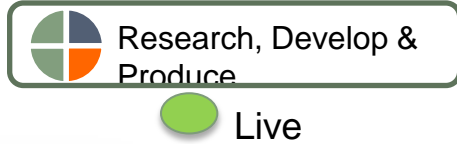
- The Endeca solution allows tracking and analysis of thousands of product test results and quality parameters for each device.
- Quality engineers can detect issues and trends early in the process and make corrections during the manufacturing process.



## Results & Benefits

- Early detection of and response to quality issues
- Consistent high quality end product
- Cost avoidance through consistent monitoring of production process
- Maintenance of world-class quality and customer satisfaction

# Reducing Product Complexity and Improving Quality



## Company

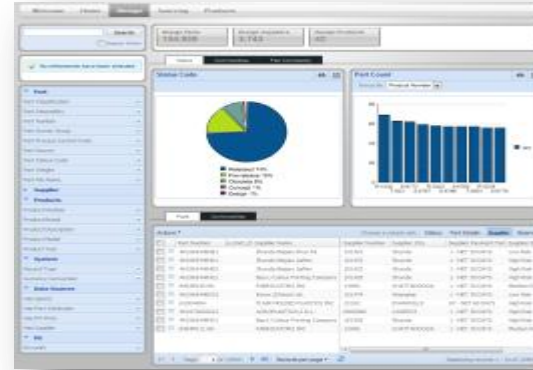
- Whirlpool Corporation engages in the manufacture and marketing of home appliances worldwide.
- \$19.5B in revenue and 69,000 employees

## Business Challenges

- Limited visibility to vendor activity between regions lead to disjointed negotiations and sub optimized supply agreements
- Multiple PLM, ERP and supply chain system due to large acquisitions.
- Multiple organizations maintaining vendor and item master data differently lead to inconsistencies in data maintenance and control

## Overview of Solution

- Unifies 4 regional SAP BW, 3 SAP R/3's with item master, part catalog, engineering change notices, engineering projects and should cost data.
- Enables engineers to find the best part based on all salient technical, commercial and quality attributes
- Provides commodity teams the ability to identify material cost reduction opportunities at a global level



\* Screenshot is representative only and not the actual customer application.

## Results & Benefits

- Improved component reuse
- Reduced direct material and supply chain related costs
- Reduced warranty and cost of quality
- Improved product innovation
- Reduced time-to-market
- Faster insight into product/supply chain redundancies in acquired businesses

# Global Product Information Management



## Industry

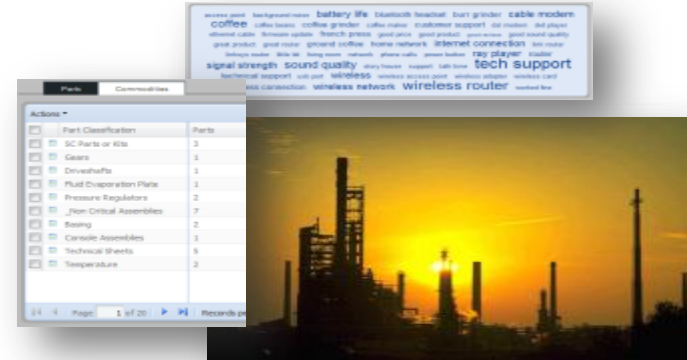
- Chemical Manufacturing, Industrial Manufacturing, Consumer Products

## Business Challenges

- Aging internal systems to support the product information needs of company web sites and applications
- Complex and difficult to maintain interfaces limit extensibility and are costly to maintain.

## Overview of Solution

- hybrid Product Content Management (PCM) and Digital Asset Management (DAM) with integrated search and navigation
- A global repository that stores product information for use by all sites and web applications including PoW, Endeca and EOC
- Manages text, images, documents and videos.



### Results & Benefits

- Global access to product information across business units and web sites
- Low support and maintenance cost
- Adaptable and flexible PIM system
- Supports innovative and continuous product improvement

# Improving product quality and reliability through information visibility



Live

## Industry

- Automotive

## Business Challenges

Reduce warranty and quality related costs:

- Detection to correction times were too long
- Difficulties setting priorities to fix problems in manufacturing
- Engineers not incorporating product failure/ performance data into their creation and improvement processes
- Inadequate supplier reclamation

## Overview of Solution

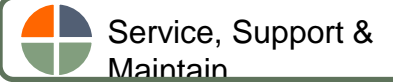
- Correlates structured and unstructured data
- Aggregates 14 independent systems including warranty claims, manufacturing, engineering part history, supplier and program data



## Results & Benefits

- Increased supplier recovery
- Reduce detection to correction time
- Reduced product launch risks
- Reduced warranty related costs

# Improving Quality Leadership and Customer Loyalty Through Early Detection



## Industry

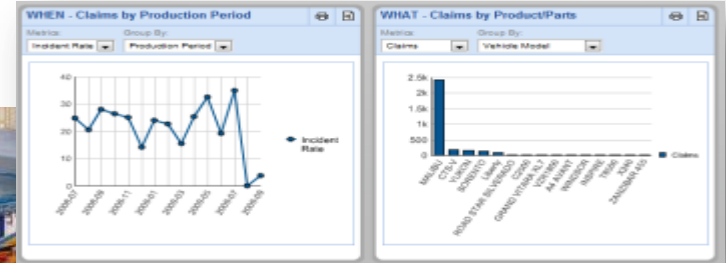
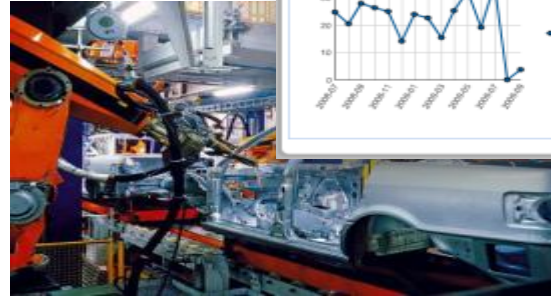
- Automotive

## Business Challenges

- At risk in achieving Targeted Incident Rate objective for 2012.
- Lack of visibility into Early Warning Systems (EWS)
- Time to obtain / analyze warranty data is about 30 Days, 21 Days over target
- Bottleneck getting information to appropriate parties
- Multiple data sources tied to complex reports

## Overview of Solution

- Correlates structured & unstructured data including 6 years worth of history
- Aggregates 14 independent systems including warranty claims, In-House repairs, manufacturing, engineering part history, supplier and counter measure data



## Results & Benefits

- Reduce data gathering process by 21 days
- Increase data and report consistency
- Increase adoption rate
- Reduce analysis time by 3%
- Reduce Incident Rates by 50%