

An Oracle White Paper
June 2012

A Technical Overview Oracle Endeca Information Discovery

Executive Overview	2
Introduction	2
What is Oracle Endeca Information Discovery?	4
The Architecture for Information Discovery	5
Oracle Endeca Server: A Hybrid Search/Analytic Database	6
Flexible Data Model	7
Best-of-Breed Search	9
Support for Queries and Filters	10
Built-In Analytics Language	11
Capabilities and Benefits	12
Studio: Creating Visually Rich Discovery Applications	13
Component Library	15
Administrative Control Panel	15
User Experience	16
Integration Suite: Handling Diverse and Rapidly Changing Data	18
Integrator	19
Content Acquisition System	20
Open Interfaces and Connectors	20
Conclusion	21

Executive Overview

The intuitive mind is a sacred gift. The rational mind is a faithful servant.¹

Applying Albert Einstein’s wisdom to the world of business analytics, traditional business intelligence and performance management solutions have been the faithful servant to organizations for years and will continue to be so. However, to stay ahead of their competition, organizations need to act on their intuition. They need to use new, diversified information regardless of where it resides to explore new paths, ask questions and validate ideas. Information Discovery enables decision makers at all levels in an organization to derive value from diverse data for business advantage—to explore and discover new insights, turn intuition into meaningful conclusions and to make better decisions - *faster*. Information Discovery is the missing link that needs to become an integral part of your organization’s business analytics strategy.

This whitepaper will introduce the newest member of Oracle’s Business Analytics portfolio, Oracle Endeca Information Discovery, provide a technical overview of the exploratory analytics platform and discuss the benefits of its iterative “model-as-you-go” approach. Information Discovery simultaneously supports the broad exploration and analysis needs of business users across new and varied information while it frees IT from the burden of prescriptive data modeling and pre-defined reporting.

Introduction

Our World Today

The mission of a business analytics strategy is to help organizations of all sizes thrive by enabling business users to explore and analyze any data, gain insight into every aspect of their business, plan ahead, make decisions based on all the information they need and act with confidence-anytime, anywhere, on any device. For more than 20 years, business intelligence (BI) and performance management solutions have delivered reliable, operational reporting and analytics, dashboards and scorecards— excellent at monitoring, forecasting and planning activities, with architectures based on pre-defined data models supporting well-known questions and proven evidence to answer those questions.

New Complications Emerge

Increasingly, the volume, variety and velocity of information generated inside the enterprise and now outside, on the web, are complicating the way organizations can leverage this data for competitive advantage. Data for decision-making does not only reside within the data warehouse but also within multiple disjointed legacy, ERP and CRM systems beyond the warehouse. Valuable information is also stored in content management systems, the intranet or on file systems. Data can be externally purchased 3rd party data, crawled output from a competitor's web site, a library of articles published on blogs, or a conversation from a post or tweet. The type of data is varied; ranging from structured database tables, to semi-structured forms containing a mix of structured values and free-form text, to unstructured documents. Couple this with an internet-savvy business culture and the impact of consumer mobile applications on enterprise software, there is an urgent need for business users to have the freedom to easily and quickly explore the relevant information to discover new insight.

The problem is this type of exploratory, undefined analysis is not what typical BI solutions are designed to support. Traditional BI solutions revolve around the premise that you must first know most of the questions you need to answer, then define the data models to support the answers to these questions. You may only then create your reports or dashboards from these pre-defined, static models. The end result is a specific set of reports and dashboards that are published to the user community; answering exactly and only the questions for which they were designed.

When analyzing diverse data that can be structured, semi-structured or unstructured text, it is extremely costly and time intensive to make these radically different data types fit within a rigid data model. Additionally the very nature of performing exploratory analysis is that users do not know what questions to ask next. So how can you pre-define a data model and build reports if you don't know exactly what you're looking for?

Complementing Business Intelligence with Information Discovery

Enter Information Discovery. Information Discovery solutions are designed to handle multiple data sources, diverse information types, and build an intuitive and interactive analytic experience without the all steps and processes required by traditional BI solutions. Information Discovery will index data the same way regardless of whether it's an xml stream exposed over

a web service, a fact table in the data warehouse, a word document on a file system, product reviews on a web site or a metadata repository. It will enrich and mash this information together to offer a single, consumer friendly presentation layer to explore, navigate and discover new insights – a dynamic experience that’s as easy as buying a product on your favorite store’s website. Information Discovery also conforms to IT governance and security standards; allowing administrators to enable business analytics while preserving strict security controls. Information Discovery fosters an environment where business users of any skill level can quickly see more and easily get fast answers to new questions against any data.

Coupling BI solutions, which provide proven answers to known questions, with Information Discovery allows organizations to get more value out of their existing business analytics solutions. For example, when the business intelligence report says that warranty claims on the top-selling product increased 15% last month, the new questions are, “Why? What changed? What’s the root cause? What are customers saying about this?” That exploration happens in an Information Discovery application.

The relationship goes both ways. Information Discovery creates new KPIs for the BI solution to deliver. For example, a consumer packaged goods company learned that preference for seemingly unrelated brands was highly correlated in certain customer segments. This insight came from a discovery application combining relevant brand data with social media and suggested new KPIs they quickly pulled into their operational BI system.

What is Oracle Endeca Information Discovery?

With Oracle Endeca Information Discovery, IT professionals gain the power to bring together diverse data from a variety of sources inside and outside of the organization, with remarkable ease and speed, resulting in better alignment with the business and a more agile IT organization.

Business professionals gain access to all relevant data they need in a powerful, yet simple-to-use, analytics application providing them with the freedom to explore the information in an unconstrained yet intuitive manner using advanced search and highly interactive visualizations. The result is delivering unprecedented visibility, analytic power, and insight to business users of all skill levels.

The Architecture for Information Discovery

Oracle Endeca Information Discovery includes the following components (refer to Figure 1):

- **Oracle Endeca Server.** This hybrid, search/analytical database provides unprecedented flexibility in combining diverse and changing data as well as extreme performance in analyzing that data. Oracle Endeca Server has the performance characteristics of in-memory architecture however is not memory bound. Oracle Endeca Server is also used extensively as an interactive search engine on many major e-commerce and media websites.
- **Studio.** Discovery application composition environment for the Oracle Endeca Server. Studio provides drag and drop authoring to create highly interactive, visually-rich, enterprise-class information discovery applications.
- **Integration Suite.** Oracle Endeca Information Discovery includes a set of tools used for loading and optionally enriching diverse information—including structured, unstructured, and semi-structured content—into Oracle Endeca Server.

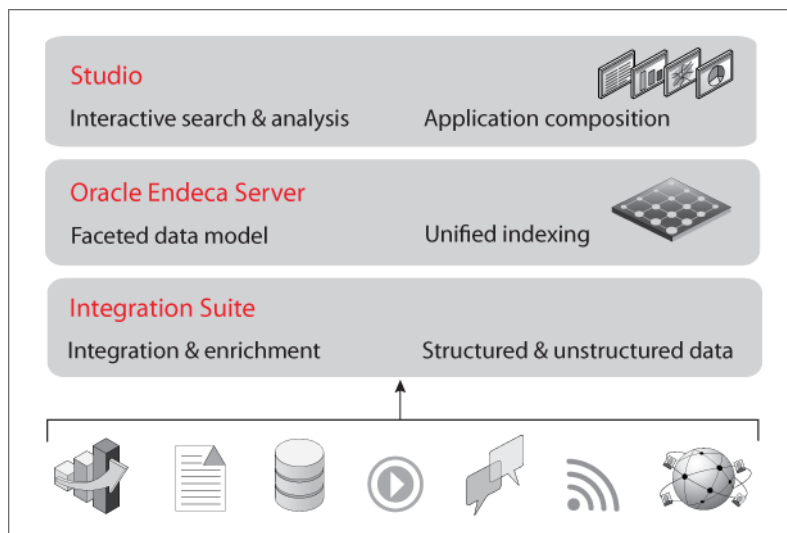


Figure 1. Oracle Endeca Information Discovery is an integrated Information Discovery platform.

Oracle Endeca Server: A Hybrid Search/Analytic Database

The core data management solution of Oracle Endeca Information Discovery is Oracle Endeca Server, a hybrid search/analytical database that is specifically optimized for discovery, not reporting or online transaction processing. Highly scalable, column-oriented, and in-memory - without being memory bound - Oracle Endeca Server supports navigation, search, and analysis of any kind of information including structured, semi-structured, and unstructured content.

Oracle Endeca Server shares many of the characteristics of a state-of-the-art database, yet it is designed for the unique constraints of unstructured, structured, and semi-structured data. Oracle Endeca Server is designed around a vertical record store to operate at high speeds while handling large data volume. Information is stored in columnar form to achieve maximum compression. Oracle Endeca Server is highly memory-efficient and stores each column of information on a combination of disk and memory in two sort orders—by value and by a universal record key identifier. Each of the Oracle Endeca Server columns contains a tree-structured index cached in memory to improve the speed of finding and delivering data in Oracle Endeca Server.

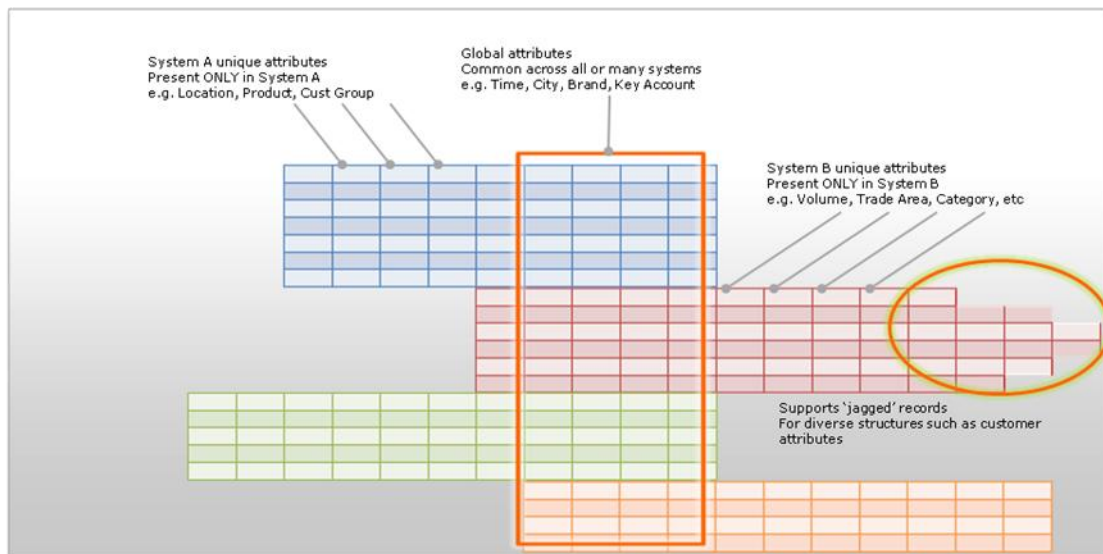


Figure 2. With Endeca Information Discovery, data does not have to conform to a target schema. Records from a variety of sources are simply stacked in their native format based on shared attributes. This allows for the data to be jagged (i.e. differing sets of attributes from one record to the next), semi-structured or completed unstructured (full-text indexed).

Lightning-Fast Query Response

To support an interactive user experience, the database must be able to process and respond to rapid-fire queries in sub-second intervals. It should also be able to accommodate new data and changes to existing data with minimal rework. Oracle Endeca Server does this by providing:

- Multicore query parallelism – increase scale and reduce latency with additional cores.
- In-memory caching of frequently accessed working column segments and query results.
- Per-attribute column storage: record order for value lookup, value order for range filters.
- Full range of type-specific compression techniques including: Prefix, String Coding, Run Length, Delta, Variable Length Numeric Coding, etc.

Figure 3 demonstrates how the Oracle Endeca Server achieves these performance characteristics through an intelligent use of CPU, memory and storage resources.

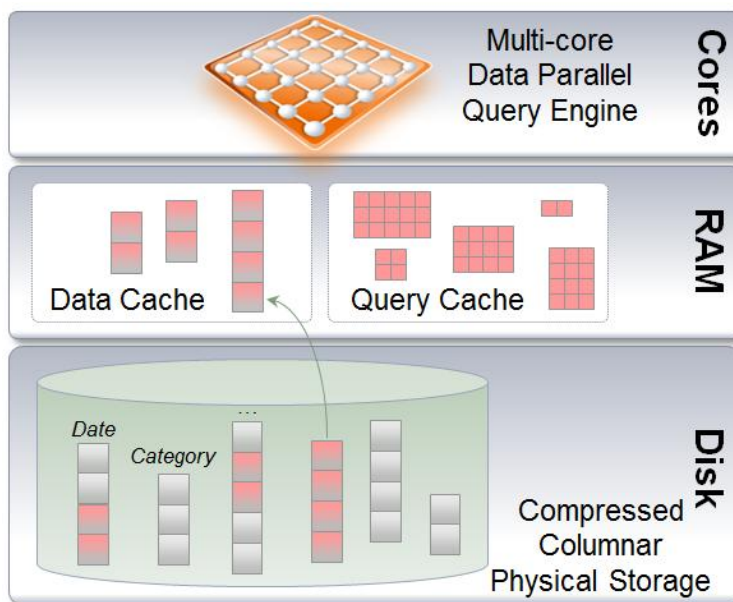


Figure 3. Scalability and performance characteristics of Oracle Endeca Server

Flexible Data Model

Oracle Endeca Server employs a unique flexible data model that reduces the need for up-front modeling, enables the integration of diverse and changing data, and supports the broad exploration and analysis needs of business users.

Oracle Endeca Server's data model is based on the concept of "records". Roughly analogous to data warehouse facts, each record consists of an arbitrary collection of attributes made up of key/value pairs. These attributes can have a single value, multiple values, or a hierarchy of values based on an external taxonomy definition. Values may be text of any length, numeric, date/time, geospatial, or Boolean.

Importantly, Oracle Endeca Server does not employ an overarching schema for data. Instead, every record has its own schema. This enables the rapid integration of any type of data, whether structured or unstructured, from inside or beyond the confines of the enterprise data warehouse, without the efforts associated with traditional relational data modeling.

Iterative, “model-as-you-go” schema of Structured and Unstructured

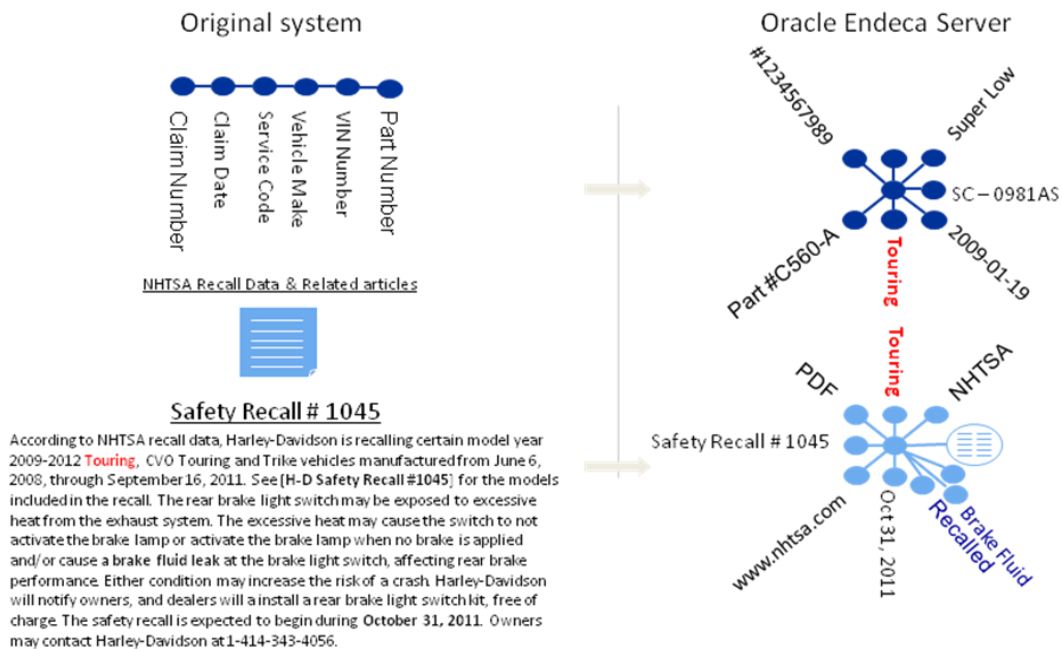


Figure 4. Relationships between structured data (such as a warranty database) and unstructured data (such as a press release) can be created through natural language processing based term-extraction. Oracle Endeca Server supports rapid inclusion of new data sources and changing data.

The process of loading data into Oracle Endeca Server is as simple as submitting a set of rows to its data-ingest web services. These web services – both incremental and bulk load are provided – can either create new records or append new attributes and values to existing records. By providing a flexible load target, Oracle Endeca Server supports an iterative “model-

as-you-go” approach to designing discovery applications. IT can quickly load new data via ETL, observe the results in Studio, and decide whether to keep, enrich, and organize what has been loaded, or discard it. Once data is loaded, Oracle Endeca Server derives common attributes from the available metadata, instantly exposing each for high-performance refinement and analysis in Studio.

As an example of this iterative approach, in the example illustrated in Figure 4, IT began by ingesting warranty claims data from the enterprise data warehouse. Structured tabular data becomes a set of Endeca records, with each column becoming an attribute and every table cell becoming a value. When the business needs to explore additional context around part failures, IT can simply load a set of internal or 3rd party press releases. By applying advanced text enrichment techniques, additional structured data, such as people, places, subjects, and sentiment, are extracted from the raw text and tagged on to the existing Endeca records. Each record now contains a combination of structured warranty claims data and textual attribution from the press releases, enabling business users to explore and analyze the data in ways never before possible.

Best-of-Breed Search

Oracle Endeca Server provides best-of-breed search features that are designed to bridge structured, semi-structured, and unstructured content. Features include type-ahead, automatic alpha-numeric spell correction, positional search, Boolean and wildcard search, natural language processing, category search, and query clarification dialogs. Search is critically important in allowing users to explore diverse data sets contained in Oracle Endeca Server.

The search capabilities in Oracle Endeca Server differ from other BI products (which employ simple “search boxes”) in the following ways:

- **Support for exploratory search.** To explore enterprise content, users typically must engage in ad hoc exploration assisted by guidance that helps them navigate and understand data while avoiding dead ends. However, traditional search does not provide a dialogue—instead it uses relevance ranking to ensure the best guess lands at the top of the results list. This type of search calculates many relevance factors for each query—including word frequency, distance, meaning, and more—and then reduces those factors to a single score based on a proprietary “black box” formula. But how can a business user act on the information that a document is, say, 38.1 percent relevant? In contrast, exploratory search gives business users the opportunity to clarify what information is relevant to them through refinements and summaries.

- **Support for cascading relevance.** Traditional search approaches reduce many relevance weights to a single score. This means that if a result with a good title match gets a similar score to one with an exact phrase match, those results will appear next to each other in a list. Unfortunately, users can't deduce the reasoning behind such rankings—even though that information could be valuable. Oracle Endeca Server takes a different approach, stratifying results by a primary relevance strategy and then breaking ties within a strata by ordering them with a secondary strategy, and so on. Application managers get the explicit means to compose these strategies based on their knowledge of their own domain. This approach gives both users and managers a deterministic way to set and understand relevance.

Support for Queries and Filters

Search may be the most common query type, but it is not the only type. Users need to express a wide range of queries. Oracle Endeca Server addresses this need by including navigation, interactive visualizations, analytics, range filters, geospatial filters, and other query types more commonly associated with BI tools. Unlike other technologies, these queries operate across structured, semi-structured, and unstructured content stored in Oracle Endeca Server. Furthermore, this set is easily extensible because the core engine allows the addition of pluggable features.

Relevancy ranking in traditional search engines is limited to a one-size-fits-all approach where queries are answered with a results list in which the most likely matches appear first. Rather than generalize one strategy for everyone as “black box” relevance solutions do, Oracle Endeca Server uses relevance strategies that vary across domains. As a result, line-of-business owners get a set of relevance modules that let them tune the best results based on their content.

Oracle Endeca Server query result sets are also summarized providing users with guidance on how to refine and explore further. Summaries include navigation, maps, charts, graphs, tag clouds, concept clusters, and clarification dialogs. Users don't explicitly ask for these summaries; Oracle Endeca Information Discovery applications provide the right ones, based on configurable controls and rules. For example, the discovery application might guide a procurement agent filtering for in-stock parts by visualizing the results on a map and calculating their average fulfillment time. Furthermore, the user can interact with summaries and filters without resorting to writing complex SQL queries—he simply clicks to add filters.

Within Oracle Endeca Information Discovery, all parts of the summaries are clickable and searchable.

Built-In Analytics Language

Endeca Query Language (EQL) is a powerful integrated analytics language inside Oracle Endeca Server. EQL is used by power users who want to create new metrics to compose their own discovery applications. It builds on the core capabilities of Oracle Endeca Server to enable analytic applications that examine aggregate information such as trends, statistics, relationships between two dimensions, and comparisons. EQL extends the capabilities of Oracle Endeca Server with a rich analytic language that allows users to explore aggregate and pivoted views of large volumes of data. EQL also supports working with a variety of data types, including numerical, geospatial and date/time values that allow applications to work with temporal data, performing time-based sorting, filtering, and analysis. Technical professionals have full access to the language for the purpose of building special formulas, metrics, and more that can be made available in discovery applications.

The following are some important EQL features:

- Tight integration with search and navigation. With EQL, which users control via the Studio application, analytical visualizations are updated dynamically as the user refines the current search and navigation query. Users can click through analytics results to reveal underlying record details, allowing them to refine their navigation directly from visualization components. Users can employ the Studio application to explore the details behind any aggregates.
- Rich analytical functionality. EQL supports computation of a rich set of analytics on records in Oracle Endeca Server—particularly the results of navigation, search, and other analytics operations. The language supports a wide variety of capabilities, including the following:
 - Aggregation functions including basic (count, sum, average) and advanced (standard deviations, variance)
 - Numeric functions including basic math and trigonometry functions
 - Composite expressions to construct complex derived functions
 - Grouped aggregations such as cross-tabulated totals over one or more dimensions
 - Top-k and percentiles according to an arbitrary function
 - Cross-grouping comparisons such as time period comparisons
 - Intra-aggregate comparisons such as computation of the percentage contribution of one region of the data to a broader subtotal

- Rich compositions of these features
- **Efficiency.** Although EQL allows the expression of a rich set of analytics, its functionality is constrained to allow efficient internal implementation, avoiding multiple table scans, complex joins, and so on. This ensures satisfactory performance for analytics operations—essential for enabling the interactive response time associated with the Studio application.
- **Familiarity.** EQL uses concepts, structure, and terminology familiar to developers experienced with SQL and relational database systems. The competing desires of familiarity and efficiency are balanced by using a subset of SQL with additional enhancements that can be efficiently implemented by the developer.

Capabilities and Benefits

Oracle Endeca Server provides the following enterprise-class capabilities to help IT organizations deploy and manage large-scale applications as well as applications scattered across the enterprise:

- **Real-time query response.** Oracle Endeca Server uses proprietary data structures and algorithms that allow it to provide real-time responses to client requests. Oracle Endeca Server stores the indices created after source data is ingested. After the indices are stored, Oracle Endeca Server receives client requests via the application tier, queries the indices, and returns the results.
- **Stateless.** Oracle Endeca Server is also designed to be stateless—necessitating that session information be maintained between query requests. This stateless design makes it easy to add servers for load balancing and redundancy.
- **Scalability, support for backup and redundancy.** Due to Oracle Endeca Server's stateless design, any Oracle Endeca Server replica on one server can reply to queries independently of a replica on other Oracle Endeca Servers. Consequently, adding replicas of Oracle Endeca Servers provides redundancy and improves query response time. If a server goes down, an Oracle Endeca Server replica provides redundancy by allowing other servers in the environment to continue to respond to queries. In addition, total response time is improved by using load balancers to distribute queries to replica Oracle Endeca Servers on any of the additional servers.
- **Support for parallelism.** Oracle Endeca Server abstracts query evaluation from physical processors, maximizing performance across multicore processors. Queries can be automatically split across processors and processed in parallel, leading to higher performance and lower latency. This allows larger data sets to be processed on a single engine, particularly for the analytic queries common with a set-based approach to search

and guidance. Oracle Endeca Server future-proofs investments for the arrival of the next chipsets, guaranteeing increasingly high-scale deployments while preserving operational simplicity and the lowest hardware costs.

- **Support for Web services.** Oracle Endeca Server offers best-in-class SOA capabilities, supporting an out-of-the-box WS-I-compliant SOAP API for data handling and easy integration into the enterprise stack. In addition, because the HTTP stack in Oracle Endeca Server is exposed, these services aren't limited to SOAP; any HTTP-based Web service, including REST and XML-RPC, can be published.
- **Support for 64-bit Windows and Linux.** Oracle Endeca Server runs on Windows and Linux 64-bit platforms and supports a distributed model for large-scale applications. It also allows queries to be threaded to take advantage of multicore hardware architectures.
- **Data governance and security.** Architected to meet the security demands of leading financial services institutions and U.S. government agencies, Oracle Endeca Server is reliable and secure in high-scale, high-traffic deployments. It readily extends existing IT policies (especially around data governance and data security) without requiring substantial additional IT overhead. Adherence to IT standards simplifies maintenance and allows for rapid integration of disparate data systems.
- **Performance at scale.** Hundreds of millions of source system records, tens of thousands of users, thousands of dimensions, and dozens of diverse source systems and interactive latencies—these are the requirements of today's BI applications. Not surprisingly, traditional solutions are unable to meet these requirements. Oracle Endeca Server combines leading-edge query parallelism technology with innovative hybrid in-memory architecture to deliver interactive performance at the scale business demands. Oracle Endeca Server delivers these capabilities without requiring proprietary hardware.

Studio: Creating Visually Rich Discovery Applications

Studio is a highly interactive, component-based environment for building enterprise-class analytic applications powered by Oracle Endeca Server. Studio is built on robust Web-based infrastructure technologies that facilitate construction of analytic applications delivered through popular Web browsers.

With Studio and its component-based approach to the construction of highly interactive analytic applications, IT professionals gain the power to rapidly prototype applications, expose them to business users, and then refine them to ensure that they identify core business requirements and achieve better alignment with business needs. This approach provides the increased agility required to rapidly deliver analytic applications. Through these applications,

business professionals gain access to all the information they need in a powerful yet easy-to-use analytic application and the freedom to explore the information in an unconstrained and intuitive manner using search and interactive visualizations. As a result, users gain unprecedented visibility, analytic power, and insight.

This new model for information access and analytics has made even the world's most complex enterprises more responsive—in the process helping them decrease costs, increase revenues, and improve productivity.

With its innovative approach to BI, Oracle Endeca Information Discovery brings new analytic power to every organization—even those with mature BI infrastructures. It does so by employing a unique method of unifying structured data and unstructured content. Oracle Endeca Information Discovery's ability to integrate information from virtually any source (including business documents and the Web) yields unprecedented analysis. Oracle Endeca Information Discovery gives users the information to decide and the confidence to act.

Oracle Endeca Information Discovery's breakthrough analytic capabilities are described below:

- **Exploration and discovery.** With Oracle Endeca Information Discovery, users can explore all relevant data in an impromptu manner—without the constraints of preset hierarchies. Providing answers to unanticipated questions and giving users the power to ask “why,” Oracle Endeca Information Discovery allows organizations to uncover the root cause of current conditions.
- **Side-to-side BI.** Drilling up and down in reports and dashboards is good, but with Oracle Endeca Information Discovery, users can walk “sideways” across data sources to discover how different parts of the business or industry inter-relate.
- **High-dimensional analysis.** Oracle Endeca Information Discovery affords superior insight by allowing organizations to unify diverse data from inside and outside the enterprise—including “incompatible,” highly dimensioned and dirty data that would have been too costly to combine using traditional methods.
- **Text Analytics.** For unprecedented insight into customer sentiment, competitive trends, current news trends, and other critical business information, Oracle Endeca Information Discovery explores and analyzes structured data with unstructured content. Unstructured content is free-form text that can come in a variety of forms from various sources including chatter on social media channels, product reviews from the web, business documents in content management systems, and text fields in your data warehouse. Oracle Endeca Information Discovery leverages text analytics and natural language processing to extract

new facts and entities like people, location, and sentiment from text that can be used as part of the analytic experience.

- **Specialized analytics.** Analytic applications from Oracle Endeca Information Discovery are customized to the decision-maker's role, the decisions they make, and the information they want to consider.

Component Library

Rather than requiring specialized application developers to code analytic applications from scratch, Studio provides a wide variety of analytic components and visualizations that can be assembled and configured by BI technology professionals. Major Studio capabilities include:

- Filtering components such as navigation, search box, and range filters (which allow users to explore and refine the data in view)
- Data visualization components—including charts, tables, cross-tabs, metrics bars, tag clouds, and maps—to assist in analyzing complex information
- Personalization components such as bookmarks that allow view customization and collaboration
- Layout and style components, including tabbed containers, a default application skin, and multiple page layout options

If a business' needs fall outside of the functionality delivered by these standard components, Java developers can use the Studio SDK to quickly and easily create custom components.

Business intelligence professionals will find Studio easy to learn and use. As a result, they'll be able to quickly develop and deploy scalable analytic solutions driven by an agile development process. The modular nature of the Studio means that this new class of analytic applications is simple to control, adapt, and extend. Granular layout and configuration control enable users to manage and personalize their experiences.

Administrative Control Panel

Studio enables administrators to manage many operational aspects of the application from an intuitive control panel. They can use this panel to modify data source connections, monitor component and Oracle Endeca Server query performance, configure state and security settings, and control how data attributes are grouped and displayed throughout the application. Studio supports role-based authorization and integrates easily with LDAP/Active directory, NTLM, OpenSSO and SiteMinder.

Administrators can choose to set up multiple user communities, each with a different view. They can even allow users to create their own personalized pages.



Figure 5. This sample analytic application built with Oracle Endeca Information Discovery illustrates how advanced search, BI, and text analytics come together to easily show new insights using interactive exploration.

User Experience

Oracle Endeca Information Discovery analytic applications provide a guided experience that helps a specific set of users complete a specific set of business tasks. In addition, Oracle Endeca Information Discovery analytic applications enable BI developers to continuously improve applications as their requirements and data evolve—an opportunity that does not exist when application coding is required.

The benefits of this approach are described as follows:

- **Increased insight and visibility and decreased costs.** The search and navigation experience provided by Oracle Endeca Information Discovery’s analytic applications

increases task completion rates, helping users find the data they want to analyze. This, in turn, enables users to make optimal decisions as they look to gain deeper insight into their business.

- **Better optimized solutions.** Because analytic applications designed using Studio can be configured instead of coded, BI professionals can iteratively update their Oracle Endeca Information Discovery analytic applications without the need for lengthy development cycles.
- **Access to fresher information.** With Oracle Endeca Information Discovery, data and content can be delivered in near real time, helping people make decisions based on the most current information.
- **Increased reuse of assets.** With search and faceted navigation built into analytic applications, users are better able to find and reuse information assets, eliminating the costs of re-creating these assets. In addition, applications built with Oracle Endeca Information Discovery can be used as the building blocks for new applications for different audiences. For example, an organization that integrates product and sales data into a sales analytics application could deploy a warranty and quality application simply by adding warranty claims information into Oracle Endeca Server and creating some additional analytic views.
- **Lower total cost of ownership.** Oracle Endeca Information Discovery allows IT to launch (and maintain) highly interactive analytic applications in less time and with a smaller financial investment than comparable applications developed using traditional coding methodologies. This is because Oracle Endeca Information Discovery offers easier application development through a highly interactive visual design environment; support for displaying and interacting with all kinds of structured, semi-structured, and unstructured data; reduced data modeling costs through a flexible schema; and easy application administration.
- **Guidance in daily decisions.** Analytic applications created with search and navigation components inform users about the data before they interact with it, helping them direct their attention to the most rewarding areas. navigation is a data-driven user interface that shows the user all of the possible, but only the valid next steps based on the user's interactions thus far, the facets in the data, and any business rules (such as recommendations or security restrictions). Oracle Endeca's navigation differs from other methods of data navigation in that it assists users in navigating the data however does not ask them to relinquish control.
- **Consumer ease of use.** With Oracle Endeca Information Discovery, BI professionals can develop and deliver analytic applications that business professionals will actually want to use—leading to higher adoption rates, lower training costs, and faster time to value. While some BI solutions strive to deliver consumer ease of use, Oracle Endeca Information Discovery is the only platform proven to be successful in consumer environments (where

user training isn't possible). In fact, 45 of the top 100 online retailers use Oracle Endeca Server to power their websites in order to help consumers make better buying decisions in their e-commerce stores.

- **Agile Delivery.** Studio facilitates an iterative approach to deployment that uncovers the true requirements of business users, minimizes risks, and speeds time to benefit. Oracle Endeca Information Discovery reduces the data modeling, integration effort, and application development inherent in traditional software deployments, making it possible to load data as is (that is, without costly cleansing), expose it to users for feedback, and refine the approach—all in a matter of days. This makes it cost-effective for IT departments to load diverse and changing data, configure applications, and iteratively expand them in a fraction of the time required by alternative technologies.

Integration Suite: Handling Diverse and Rapidly Changing Data

Integration Suite is a powerful solution for loading diverse information—including, unstructured, and semi-structured content—into Oracle Endeca Server. Supporting a variety of data integration products, system connectors, and content enrichment libraries, the suite makes it possible for users to efficiently combine information from any source into a single integrated view. Furthermore, it also allows enrichment of the content using natural language processing and regular expression techniques which can be used to drive the navigation experience.

Integration Suite includes:

- **Integrator.** This ETL environment is used for integrating and enriching enterprise data and content.
- **Content Acquisition System (CAS).** This collection of connectors is used for extracting, enriching, and integrating unstructured content from network file systems, Websites, and content management system repositories. Currently supported systems include: EMC Documentum, EMC Documentum eRoom, FileNet P8, FileNet Document & Image Services, Interwoven TeamSite, Lotus Notes/Domino, Microsoft SharePoint, and OpenText LiveLink.
- **Open Interfaces and Connectors.** Using the Oracle Endeca Server's open Web services API, the suite allows direct data integration from industry-standard ETL tools, such as Informatica PowerCenter and Apache Hadoop, and SOA-based services.

Integrator

Integrator is used for data extraction, transformation, and loading when an enterprise ETL solution is not already in place or is not desired. It allows business professionals to easily create data integration processes that connect to a wide variety of source systems, including relational databases, file systems, and more. In addition, Integrator supports the ability to implement business rules that extract information from source systems and transform it into business knowledge in the Oracle Endeca Server in an easy-to-use environment. Additional features include:

- Rich visual environment for creating data integration processes
- Wide variety of source connectors to relational and file sources using open connectors including ODBC and JDBC
- Support for moving data directly into Oracle Endeca Server
- Support for batch-based and real-time data feeds
- Library of transformers for modifying and reformatting data
- Join components for merging related data
- Platform and database independence
- Efficient execution with small footprint
- Scheduling and on-demand execution capabilities
- High performance and scalability

Key benefits of Integrator include:

- Reduced manual workload and time
- Communication among incompatible systems
- Optimized process for data interpretation
- Single, consistent process for business-critical data
- Increased development efficiency

Content Acquisition System

Oracle Endeca Information Discovery also includes the Content Acquisition System, which gathers content from file systems, content management systems, Websites, and other unstructured and semi-structured sources. Key capabilities include:

- Packaged connectors to common enterprise applications
- Documents, metadata, and security information all collected from sources
- CAS Extension API for adding custom functionality
- Administration through GUI or command line interactions

CAS provides file system and Web crawling capabilities along with a set of connectors that enable crawling of content management systems. Example connectors include:

- Microsoft SharePoint
- EMC Documentum
- Lotus Notes/Domino
- EMC Documentum eRoom
- FileNet P8
- FileNet Document & Image Services
- OpenText Livelink
- Interwoven TeamSite
- Alfresco (new in CAS 2.2)
- Any JSR-170-compliant repository

Open Interfaces and Connectors

Oracle Endeca Server is also accessible to other enterprise applications as a SOAP (Simple Object Access Protocol) based web service. This web services interface can be used by commercial ETL tools or from custom code to load data and to query the engine.

Conclusion

Oracle Endeca Information Discovery allows business users to explore, analyze and discover information via interactive and configurable applications. The heart of Oracle Endeca Information Discovery, the Oracle Endeca Server, is a breakthrough technology that allows an iterative "model-as-you-go" approach that simultaneously frees IT from the burdens of traditional data modeling and supports the broad exploration and analysis needs of business users.

While traditional business intelligence and performance management solutions have been the faithful servant to organizations for years and will continue to be so, the addition of Information Discovery solutions will help organizations to act on their intuition and stay ahead of their competition. Oracle Endeca Information Discovery allows organizations to use new, diversified information regardless of where it resides to explore new paths, ask questions and validate ideas. Endeca Information Discovery enables decision makers at all levels in an organization to derive value from diverse data for business advantage—to explore and discover new insights, turn intuition into meaningful conclusions and to make better decisions - *faster*.



A Technical Overview of
Oracle Endeca Information Discovery

May 2012

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com



Oracle is committed to developing practices and products that help protect the environment

Copyright © 2012, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark licensed through X/Open Company, Ltd. 0112

Hardware and Software, Engineered to Work Together

ⁱ Einstein, Albert. In his own words

ⁱⁱ NHTSA data: www.nhtsa.gov