Oracle Endeca Information Discovery: A Technical Overview
Contents

Introduction ......................................................................................................................... 4

Dynamic Questions ............................................................................................................. 4

Diverse Data ....................................................................................................................... 4

Composable Applications, Purposeful Views ................................................................. 5

A Complete Solution ........................................................................................................ 5

Oracle Endeca Information Discovery Architecture ................................................... 6

Oracle Endeca Server: Revolutionary Hybrid Search/Analytic Database .................... 7

  Flexible, Adaptive Data Model ......................................................................................... 7

  Fast Query Processing at Scale ....................................................................................... 11

  Industry-Leading Search and Navigation ..................................................................... 12

  Data Enrichment ............................................................................................................ 13

  Built-In Analytics Language ........................................................................................ 14

  Other Endeca Server Capabilities and Benefits .......................................................... 15

Oracle Endeca Information Discovery Integrator: Easily Manage Diverse Data .......... 16

  Integrator ETL .............................................................................................................. 16

  Text Enrichment and Sentiment Analysis .................................................................. 17

  Web Acquisition Toolkit .............................................................................................. 17

  Integrator Acquisition System ...................................................................................... 17

  Open Interfaces and Connectors .................................................................................. 17

Oracle Endeca Information Discovery Studio: The Art of Visual Discovery ............. 18

  Self-Service Data Management .................................................................................... 18

  Smart Applications ........................................................................................................ 18

  Self-Service Mashups ................................................................................................... 19

  Summary of Studio Data Management Features and Benefits .................................. 19

  Building Visually Rich Discovery Applications ......................................................... 20

  Composability .............................................................................................................. 20

  Integrated Discovery ..................................................................................................... 21

  Enterprise-Class Administrative Control .................................................................... 24

  Summary of EID Studio’s Capabilities and Benefits .................................................. 25

Conclusion ....................................................................................................................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A: EID Success Stories</td>
<td>27</td>
</tr>
<tr>
<td>Automotive Manufacturing</td>
<td>27</td>
</tr>
<tr>
<td>Consumer Beverages</td>
<td>27</td>
</tr>
<tr>
<td>Commercial Food Production</td>
<td>28</td>
</tr>
</tbody>
</table>
Introduction

The last decade has seen an exponential increase in data volume and complexity, and technologies to help business make sense of this data have proliferated accordingly. In addition to enterprise data management and business intelligence products, data discovery solutions have now become “a mainstream BI analytic architecture.” *(5 Feb 2013, Gartner MQ for BI and Analytics)*

Organizations have been managing metrics and structured data for half a century, but are now operating in an environment where the world’s data has doubled in the last two years. Today’s challenge is how to find critical insights hidden in the wealth of unstructured information—from the human dialogues in enterprise text fields, to relevant information from the outside world, in websites, blogs, social media, government reports, consumer reviews—and keep up with the pace of change without drowning in data in the attempt. Traditional methods are too labor- and cost-intensive, meaning many organizations simply cannot include the information they need in business analytics. And the requirement for fast, effective data exploration only grows more pressing as analytics budgets shift from IT to the business, driven in part by business user demand for more control over their analytic destiny.

Dynamic Questions
Traditional business intelligence solutions optimize for operational metrics: this month’s sales in Region A or B; Region A’s sales in this or that month. BI focuses on fast answers to predicted questions and to the same types of questions—Region A’s sales, broken down by territory and rep (hierarchical drilldown). When you want a smooth paved road from a recurring question to a clear answer, a mature BI system is tough to beat.

When the road gets bumpy, or starts to disappear in the brush—in other words, in the face of unpredictable change—the operational strength of traditional BI is less helpful. Managing change is where data discovery shines, because discovery solutions are optimized for the unpredictable, with a specific charter to reveal the why. In pursuit of that why, analysts need to have every tool available to them. Faceted navigation, charts, interactive heatmaps, tables, tag clouds, spotlights—these are the implements that will help an analyst follow the road through the brush to uncover a deep solution to an urgent problem.

Diverse Data
The business intelligence world has gotten used to talking about “the data”, as if it were a permanent and static object, periodically renewing in content, perhaps, but constant in structure. In reporting and performance monitoring, permanence is exactly what we want: when we’re comparing current metrics with past ones, we should be sure to use the same metrics and the same data.

Data discovery is another world, with its own set of highly desirable traits. This is the world of variability, where the constant is change. What we want for this environment is the flexibility and the freedom to shift between different views of different data, to combine and even enrich data as we go, as our analysis requires. This agility is the hallmark of data discovery. Integration happens in the moment, at the hands of the analyst, in an ongoing dialogue with the data.
Composable Applications, Purposeful Views

Data discovery is a cycle of adding new data, asking new questions, and seeing new patterns. Thus, in data discovery, stunning charts and pixel-perfect maps aren’t the end, they’re the beginning. Discovery applications aren’t reports, infographics, or interactive PowerPoint slides; their lifecycle isn’t a gradual progression toward some predetermined point of completion, but rather an organic evolution—radical, if need be—in response to new insights, new questions, and new data.

To support this charter, discovery applications must have certain core characteristics: they must be easy to compose, configure, and change for both business users and IT; they inherently integrate search, navigation, and analysis into a single experience that is interactive but unscripted; and they are fundamentally data-driven, using intelligence from the data itself to determine what to show and how to show it, driving meaningful exploration that improves understanding and decision-making.

Having laid out the essentials of data discovery, let’s look at how Oracle fulfills them for the enterprise.

A Complete Solution

Oracle Endeca Information Discovery delivers a complete solution for agile data discovery across the enterprise, empowering business user innovation in balance with IT governance. Founded on a revolutionary hybrid search-analytical database, EID offers fast, intuitive exploration across both traditional analytic data, leveraging existing enterprise investments, as well as to more exotic, external, and typically unstructured data. This allows organizations to achieve unprecedented visibility into all relevant information, to drive growth while saving time and cutting costs.

This whitepaper introduces Oracle Endeca Information Discovery to a technical audience by describing its unique architecture and explaining how that architecture supports fluid, secure, and scalable data discovery for the enterprise.

With its innovative approach, Oracle Endeca Information Discovery brings new analytic power to every organization—including those with mature BI infrastructures. It does so by employing a unique method of unifying structured data and unstructured content, yielding profitable new insights from the combination. Oracle Endeca Information Discovery’s ability to integrate information from virtually any source (including business documents and the Web) enables unprecedented visibility in 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 interrelate.
• 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 from many sources, including customer complaints, product reviews from the web, call center transcripts, medical records, and text fields in a 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 to enrich the analytic experience. Moreover, by allowing self-service users to enrich data from within their apps, Endeca Information Discovery opens a whole new world for discovery.

• 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.

Oracle Endeca Information Discovery Architecture
Oracle Endeca Information Discovery has three tiers:

• Oracle Endeca Server. This hybrid search/analytical database is at the heart of Oracle Endeca Information Discovery, providing unprecedented flexibility in combining diverse and changing data as well as strong performance in analyzing that data. Oracle Endeca Server has the performance characteristics of in-memory architecture coupled with a highly intelligent approach to using disk, optimizing available resources and avoiding being memory-bound. Oracle Endeca Server is also used extensively as an interactive search engine on many major e-commerce and media websites.

• Oracle Endeca Information Discovery Integrator. Integrator is a suite of industrial strength data management tools that makes it easy for business users and IT to acquire, ingest, and enrich information. In addition to self-service data loading, OEID Integrator is a powerful visual environment for data integration that includes the Information Acquisition System (IAS) for gathering content from file systems, content management systems, and websites; and out-of-the-box ETL purpose-built for incorporating data from a wide array of sources, including Oracle BI Server. Oracle Endeca Web Acquisition Toolkit is a web-based graphical ETL tool that allows IT to enter a URL, collect content, and add structure to it as part of the data acquisition process. Connectivity to data is also available through Oracle Data Integrator (ODI).

• Oracle Endeca Information Discovery Studio. The front end to Endeca Server, Studio is a rich visual application composition environment that provides drag-and-drop authoring to create highly interactive, personal and enterprise-class information discovery applications. Studio also includes self-service data provisioning, which gives business users the ability to add their own data, connect to existing gold-standard enterprise sources, and combine them. Studio enables allows IT to create application templates for self-service and ensure that data security is maintained.
Figure 1. Oracle Endeca Information Discovery, an integrated information discovery platform.

These components combine to provide a powerful discovery platform that empower business users and IT equally. From IT-provisioned applications with myriad discovery components exposing data from several sources, to the personal, incrementally-evolving application developed by a business user, EID enables the discovery of critical insights, whatever the data, and whatever the question.

The magic starts with Endeca Server, the revolutionary database that drove Endeca’s success across e-commerce, enterprise search, and data discovery.

**Oracle Endeca Server: Revolutionary Hybrid Search/Analytic Database**

The engine behind Oracle Endeca Information Discovery is Oracle Endeca Server, the industry's first hybrid search/analytical database, specifically optimized for data discovery. Flexible, scalable, column-oriented, and in-memory without being memory-bound, Oracle Endeca Server enables fluid navigation, search, and analysis of any type of data—structured or unstructured, internal or external.

As an engine optimized for data discovery, Oracle Endeca Server’s sweet spot is precisely at the point where users need to have maximum flexibility in how they query any data, structured or unstructured, numbers or text. Endeca Server provides first-class, fully-integrated support for both keyword searches and analytical queries. Through its innovative, purpose-built architecture, it enables users to ask any question, of any type, of any data and get instant answers that both prompt new questions and fuel decisions. That’s the meaning of data discovery.

**Flexible, Adaptive Data Model**

Oracle Endeca Server employs a unique, flexible data model that reduces the need for up-front modeling, enabling the integration of diverse and changing data while supporting the broad, unpredictable search, exploration, and analysis needs of business users.
Endeca Server organizes data into records. Each record is a sequence of attribute-value pairs. For example, a record with three attribute-value pairs might be:

```
[{ID, 1} {FirstName, Thomas} {Company, Oracle}]
```

This data model means that every record can be different: they don't need to have the same attributes or the same number of attribute-value pairs, and they can even have multiple values for the same attribute. So in the same collection of records, there might also be the records:

```
[{ID, 2} {Company, SAP} {Title, Sales Consultant} {Age, 45} {Comment, “Ich bin ein…”}]
```

```
[{ID, 3} {Hobby, Bowling} {Hobby, Tennis} {Company, Oracle}]
```

It's clear, then, that Endeca Server records offer several technical advantages over rows in a relational database. For example, Endeca Server naturally compresses sparse data: if a record doesn’t have a value for an attribute, it’s simply never associated with that attribute. If, conversely, a record has several values for an attribute, Endeca Server simply stores all of them, without having to duplicate the rest of the record.

Native support for jagged, idiosyncratic records means that Endeca Server can ingest data with no up-front modeling. This lowers the barriers to discovery, both for IT and especially for business users: take some interesting data, dump it into Endeca Server where it’s organized for integrated search, analysis, and navigation, and start discovering in minutes. If later a user wants to ingest data from a different source, that’s no problem at all—just load it in, leaving the old records as they are. Or, if a user wants to enrich data in place—say by running a salient term extractor on customer complaints or patient records—they can do so without concern for the schema. Endeca Server’s pioneering of faceted navigation is the user-facing complement to this adaptive architecture: rather than forcing the user (or IT) to specify or know about a
schema before they can see the data, Endeca Server builds up a schema as it ingests data, then surfaces that schema with the data for the user to refine upon. One of the great virtues of Hadoop is that it lets organizations safely and cheaply store data without having to know much about it first. Endeca Server provides a similar benefit, with the distinction that in its case, it optimizes data for immediate, responsive discovery rather than either batch analytics, schema-driven querying, or complicated statistical data mining.

One of the great virtues of Hadoop is that it lets organizations safely and cheaply store data without having to know much about it first. Endeca Server provides a similar benefit, with the distinction that in its case, it optimizes data for immediate, responsive discovery rather than either batch analytics, schema-driven querying, or complicated statistical data mining.

The one attribute value every record must have is a unique record ID. Here’s why.

For each attribute in the data, Endeca Server keeps two indices that store every value-record pair on that attribute. The forward index is sorted by record ID; this enables quick lookups of the values associated with certain records—useful when users have drilled down and want to see detailed information on certain records, for example in a results table. The reverse index is sorted by attribute value; this optimizes for cases in which the user wants to analyze the distribution of values in the data, like aggregations, range filters, and navigation. Each record, rather than storing its attribute values itself, points to the appropriate position(s) in the appropriate attribute indices. A universal membership index tracks the set of attributes that each record has values for; when a record is updated to have a new attribute, the membership column is updated along with the relevant attribute models.

---

A universal membership index tracks the set of attributes that each record has values for; when a record is updated to have a new attribute, the membership column is updated along with the relevant attribute models.
Attribute models are mapped into virtual memory. To take advantage of the different sort orders, each attribute index is prefixed with a B-tree-like data structure that greatly accelerates the lookup of records and values. Frequently-accessed column segments are cached in physical memory to speed query processing. In this respect, Endeca Server’s storage strategy is designed to exploit a common data discovery usage pattern: users often have some idea of what they’re looking for and so apply early filters such as a keyword search or a spatial/temporal selection that greatly restrict the eligible result set, then make varied forward and backward steps within that subset of data. Maintaining all attribute models in virtual memory allows Endeca Server to supply the breadth needed for those initial starting-point filters, while its caching strategy enables interactive speeds during the back-and-forth ad-hoc exploration phase. Strictly in-memory solutions necessarily restrict the scope of data available for that initial starting point. Also, this strategy enables scalable, iterative expansion both of the analysis and the data. Adding new attributes via text enrichment or mashups is no problem at all because Endeca Server can scale to as much disk as you allow. In contrast, pure in-memory solutions face a hard stop when they exhaust available memory—which means many users (say, more than a single department) cannot freely experiment with enriching and mashing up data. Because Endeca Server’s cache size is easily configurable and controllable per data domain, it’s easy for administrators to tune performance by raising the cache size.

Each attribute model is type-specific, allowing Endeca Server to reap the full benefit of data compression techniques. Endeca Server supports numerics, Booleans, date-times, geocodes, hierarchical values (e.g. Wine -> Red -> Bordeaux), and—crucially—strings of any length. And here “support” means more than just “allow”: Endeca Server builds in optimizations for each data type. For example:

- **Geocodes** have two reverse indices: one sorted by the value’s latitude, one sorted by the value’s longitude. Quick geographical searches are the result of this special optimization.
- **Hierarchical values** point to a position in a tree data structure that captures the structure of the hierarchy. In other words, Endeca Server embeds hierarchies at the most fundamental level of its data storage. This means that when a parent value is requested (e.g. Red), its descendants (e.g. Bordeaux, Claret) are also included in the request—even though they were not stored on a particular record.
- **Strings and text values** are stored only once per distinct value, in a universal index that all attribute models can access. Instead of holding instances of string values, attribute models hold references to their positions in the universal index. This practice of string interning speeds up many queries by 50% or more and cuts down total index size by a third in typical cases.

These examples show how support for diverse query types over diverse data is rooted in the most fundamental layers of Endeca Server. Already, this adaptive data model and type-specific support bespeak a commitment to solving the challenges of data discovery that few other tools can claim—certainly not those that depend on off-the-shelf databases. But if the attribute models suggest this fact, Endeca Server’s integrated search index confirms it.

Endeca Server’s core text search functionality is fueled by an inverted index that directly incorporates the records and attribute model. Search tokens are associated with the record, model, and search interface they appear within. A position column also keeps track of where a term appeared within an attribute value. This intricate architecture allows Endeca Server to do much more than just efficiently retrieve the records that
contain a certain word or phrase—it allows it to return results with all the context that makes them intelligible to users, including matched term highlighting, identification of the facet in which the match occurred, relevance ranking, and, in the case of text fields, snippets that show keywords in context.

Spell-correction, synonym expansion, and any-position wildcard search are made possible by several indices that supplement the core postings index. IT can fine-tune these indices for applications where web-caliber search plays a central role, or trim them for more navigation- or visualization-centric applications. In either case, the fundamental structure of Endeca Server integrates text search with navigation and analysis to deliver an equally-integrated user experience.

The two key points here are schema flexibility and query flexibility. No matter what the data is, Endeca Server will organize it for fast exploration by any query type.

**Fast Query Processing at Scale**

Providing an interactive user experience for many concurrent users is a challenge for any database. Add to that the demands of discovery—complex, changing data; varied query methods; unexpected, ad hoc queries—and building a performant platform is no small task. But Oracle Endeca Server’s innovative architecture, plus the optimizations accrued over a decade of supporting applications with exacting performance requirements, allow it to respond to rapid-fire queries in sub-second intervals.

Oracle Endeca Server achieves high performance through:

- **Dual-sorted type-specific columnar storage.** As described above, maintaining two columns—one sorted by record ID, one sorted by attribute value—ensures fast, scalable performance for any type of query.

- **Query parallelism.** Search, analytic, and navigation queries are split to leverage all available cores to increase throughput and lower latency.

- **Code generation.** Parallel processing can incur several types of overhead that eat into the performance gain it offers. To dodge this overhead and maximize efficiency, Endeca Server continues a long history of technology leadership by converting a parallelized query plan into parameterized machine code that executes on the several cores. The representations used in code generation may themselves be cached to accelerate subsequent processing.

- **Pervasive caching.** Endeca Server’s caching algorithms exploit EID’s navigation-oriented user experience, caching intermediate queries and result sets to accelerate a user’s next query, no matter which direction it goes. The cache is shared among all users.

- **Cache warming.** In many products, updates to a data source flush the cache. This has the direct effect of slowing down queries and the indirect effect of making IT hesitant to perform updates. Endeca Server skirts these perils by quickly restoring the cache after updates.

- **Cluster orientation.** Endeca Server was built to run on clusters, and it shows. Endeca Server is stateless, meaning each query request must carry its full state. This design implies that any Endeca Server instance can reply to any query, and thus adding Endeca Server instances provides redundancy and improved performance. In addition to offering enterprise-grade cluster administration controls, Endeca Server can free resources by automatically idling indexes that are not being used.
A forthcoming Oracle Endeca Information Discovery Performance Whitepaper describes EID’s performance as it scales up to 300M Endeca records on a single machine, while providing interactive speeds for realistic query loads.

**Industry-Leading Search and Navigation**

Oracle Endeca Server provides best-of-breed search and navigation features that help users discover insights hidden in unfamiliar data.

With built-in stemming and spell-correction, along with configurable thesaurus expansion and relevance ranking, Endeca Server’s advanced keyword search optimizes for recall, ensuring that arbitrary choices (such as choosing a singular instead of a plural, or wreck instead of accident) don’t prevent users from making game-changing discoveries. Meanwhile, faceted navigation organizes the data and guides the user through it without requiring advance knowledge of questions or drill paths, cleanly presenting all and only the data that can lead to a useful refinement from the present state. This integration of exploratory search and navigation gives business users the opportunity to clarify what information is relevant to them through refinements and summaries.

Both core components have their roots in Endeca’s e-commerce history, where they have proved so successful at helping consumers navigate through unfamiliar products that 45 of the top 100 online retailers use a version of Endeca Server to power their online stores. The same core technology delivers an intuitive and powerful discovery experience to business analysts.

Endeca Server’s search features include:

- **Attribute-sensitive typeahead.** Because of how Endeca Server stores data, in the web application layer Studio can break out typeahead suggestions by attribute. This context helps users refine their question as much as search helps them answer it. Typeahead only shows values that meet the current filter state.

- **Data-driven spell correction.** During ingest, Endeca Server builds a dictionary using the values in the actual data. Proper names, part numbers, chemical compounds, technical terms—in each of these cases Endeca Server’s data-driven dictionary helps guides users toward what they’re looking for. Endeca Server uses this dictionary to provide spelling correction and did-you-mean suggestions.

- **Did-you-mean suggestions.** If a search term would return very few results while a lexically-close term would return many, Endeca Server can substitute the more popular term. This helps users avoid dead ends.

- **Stemming.** Endeca Server can return all terms that match the roots of a search term (e.g., walks, walked, and walking for the keyword walk). Stemming avoids the arbitrary exclusion of results, based on tense or number, that plagues typical discovery tools.

- **Thesaurus expansion.** If provided with a thesaurus, Endeca Server will expand search terms to include synonyms. Doing so widens the breadth of a user’s query, making it more likely that they’ll be able to use navigation to find what they’re looking for.

- **Many search modes.** From Boolean to wildcard to exact to partial (and more), Endeca Server provides full support for a variety of search use cases.

- **Configurable relevance ranking.** In contrast to the black-box approach favored by many search tools (and in particular ones glommed onto data visualization products), Endeca Server allows IT to build customized
relevance strategies based on factors like proximity, position, number of terms matched, number of matched terms, and number of attributes containing a match (among several others).

- **Inter- and intra-dataset search.** Endeca Server’s support for data mashups extends to search. Users can specify whether they’d like to search all data sets in an application, or just a particular one. Typeahead also breaks out suggestions by source.
- **Robust internationalization.** All the above features are officially supported in 35 languages.

Endeca Server’s faceted navigation includes the following features:

- **Context awareness.** Not only does Endeca Server only show values that pass the current filters, it hides attributes that cannot lead to a useful refinement. For example, if all the records that meet the current filter criteria have Color=Blue, Studio will not show the Color attribute in the available refinements bar, because selecting Color=Blue would not limit the result set.
- **Native hierarchy support.** Because Endeca Server natively stores hierarchical values (e.g. USA -> Massachusetts -> Cambridge) just as it does strings and numbers, users can seamlessly navigate through hierarchies without the performance penalty of on-the-fly hierarchy construction or the bother of a separate hierarchy component.
- **Typeahead for values.** The same search technology that fuels typeahead in the search bar allows users to reduce a list of hundreds or thousands of attribute values to a desired few, just by typing in a few characters.
- **Multiple selection types.** A wide array of selection types, including single-select, multi-or, multi-and, negative, multi-OR, multi-AND, offer users a contextual, dynamic approach to including—and excluding—data from analysis.
- **Precedence rules.** Attributes often become meaningful only in certain contexts. Precedence rules allow for specific attributes to be hidden until the context is created by user refinements on other attributes.
- **Integrated range filters.** Range filters appear alongside value lists and, as with every other component, always match the current filter state, giving feedback to users while guiding them toward answers.

Whether it’s on web search engines or e-commerce sites, most people use search and faceted navigation several times a day, and they do so instinctively. These are the dominant forms of exploration with unfamiliar information today, and they are the core pillars of Endeca Server—so much so that to this day earlier incarnations of Endeca Server power hundreds of the leading e-commerce and enterprise search applications. The result is that Endeca Information Discovery delivers a user experience that’s second nature to any Internet user.

**Data Enrichment**

Endeca Server takes data as it is, but it doesn’t have to leave it that way. Native data enrichment capabilities put advanced natural language processing techniques into the hands of business users, making possible discoveries that couldn’t have been anticipated beforehand. A whitelist component lets business users leverage domain knowledge to turn acronyms, model names, and other industry knowledge into attributes that appear in the application. Meanwhile, salient term extraction exposes key concepts lying hidden in text data.
Data enrichment is a natural fit for Endeca Server, dovetailing with its strengths in managing jagged and unpredictable data, efficient updates, and iterative development. Once kicked off, enrichment processes run in the background while the user continues exploring the app. Behind the scenes, Endeca Server creates a new attribute for the output of the enrichment (e.g. ExtractedTerms, NormalizedProductNames) and establishes values for that attribute for the records that have generated enrichments. When this process completes, the user is alerted, the page refreshes, and the new attribute is immediately available for use in navigation, charts, tag clouds, and any other facet of a discovery application.

Business users can explore hunches and alter their data without having to declare this in advance and hand it off to IT for processing. The data is held in the index, so one user’s changes don’t interfere with anyone else’s.

Endeca Server’s current data enrichment functionality includes the following features:

- **Salient term extraction.** Builds a model of terms that appear in text data, then picks the most important terms in each record, up to a user-specified number of terms. This means that different types of text (e.g. a sales pipeline update and a customer complaint) have distinct models, making mashups more insightful.
- **Whitelist.** Accepts user-entered or uploaded mappings of input terms to output terms.
- **Language support.** Salient term extraction works in seven language, while whitelists are supported in all 35 languages supported by EID.
- **Built by and for Endeca Information Discovery.** These enrichment capabilities are developed in-house and tailored for the discovery use case.

**Built-In Analytics Language**

Endeca Query Language (EQL) is an expressive, SQL-like analytics language that allows IT and power users to define new metrics and views. EQL boosts Endeca Information Discovery’s analytical power by providing an entry point for more complex analytics, including regressions, running averages, part-whole comparisons, and top k analyses. At the same time, its position on top of the index furthers Endeca Information Discovery’s modeling-optional strategy—users load data, play with it in a discovery application, and then use EQL to define customized metrics and views as desired. Different users with different interests can define their own views on top of the same data, then publish their views for others to leverage. Once created, views can be used as the basis for search, navigation, and visualization in Studio.

To understand EQL’s expressiveness, it helps to know that when a user interacts with any Studio component (a chart or a map, for example), that component sends an EQL query back to Endeca Server. EQL supports all the data types of Endeca Server, including geospatial, temporal, and hierarchical data, giving advanced users fine-grained control over their applications. Common use cases include manually joining different data sets to create customized aggregates and metrics. EQL also helps users make the most of multi-assigned attributes, which are treated as sets.

The following are some important EQL features:

- **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. EQL is parallelized and takes full advantage of multiple cores.

- **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.

**Other Endeca Server 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 provide interactive 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.

- **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. This stands in contrast to the many desktop discovery tools that support only Windows and/or only 32-bit 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.
Oracle Endeca Information Discovery Integrator: Easily Manage Diverse Data

EID provides numerous options for loading diverse and rapidly changing data, including structured, unstructured, and semi-structured content, into Endeca Server.

Platforms

- Integrator ETL provides a drag-and-drop interface for building pipelines that integrate data from a variety of sources, including flat files, JSON, XML, databases, HDFS, and Hive. By dragging text enrichment components into their pipelines, IT can extract concepts and entities (companies, people, places, and products) from unstructured text to bring a new dimension to discovery.
- Oracle Data Integrator (ODI) provides native support for Endeca Server, meaning that organizations can seamlessly and securely transfer their data from enterprise data sources through an enterprise data integration platform to an enterprise data discovery platform.

Tools

- Integrator Acquisition System (IAS). Crawl file systems and extract content from binary files (e.g. PDFs, Office files).
- Oracle Endeca Web Acquisition Toolkit. Use a simple visual interface to extract content from a wide variety of web-based unstructured sources—even ones without APIs.
- Advanced Text Enrichment and Sentiment Analysis. Configurable NLP engine that integrates text enrichment and sentiment analysis into data pipelines.

Integrator ETL

Integrator ETL 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 like 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 ETL 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

**Text Enrichment and Sentiment Analysis**
The Text Enrichment component provides information extraction and summarization capabilities. Extracted information includes entities (such as people, places, and organizations), quotations, and themes. It utilizes the Salience Engine from Lexalytics. Text Enrichment with Sentiment Analysis provides the ability to extract sentiment from documents at the document, entity, and theme levels.

The supported text enrichment features include:

- Sentiment Analysis
- Named Entities
- Themes
- Quotations
- Document Summary

**Web Acquisition Toolkit**
The Oracle Endeca Web Acquisition Toolkit offers easy access to myriad web sources—whether they have APIs or not—and integrates readily into any IT environment by supporting a wide variety of enterprise standards. An intuitive point-and-click Integrated Development Environment (IDE) lets users build data integration pipelines that bring together unstructured data from web sources like consumer sites, industry forums, and Big Data systems.

**Integrator Acquisition System**
Oracle Endeca Information Discovery also includes the Integrator Acquisition System (IAS), which gathers content from file systems and other unstructured and semi-structured sources. Key capabilities include:

- IAS Extension API for adding custom functionality
- Administration through GUI or command line interactions
- Documents, metadata, and security information all collected from sources

**Open Interfaces and Connectors**
Oracle Endeca Server is also accessible to other enterprise applications as a Simple Object Access Protocol (SOAP) based web service. This web services interface can be used by commercial ETL tools or with custom code to load data and to query the engine.
**Oracle Endeca Information Discovery Studio: The Art of Visual Discovery**

**Self-Service Data Management**

Studio builds on the robust data integration options described above with easy and elegant data management for self-service discovery.

Spreadsheet sprawl has plagued more than one IT department. Analysts all have their own spreadsheets and their own stories. At the very least this means duplicated effort and wasted resources; more often, the consequences are more dramatic, since no one can tell if data is reliable or whether they can trust the discoveries they make.

Things are different with Endeca. Users can quickly upload their spreadsheets or JSON files via the provisioning service, which will profile the data, present an opportunity to adjust metadata, then load the data into Endeca Server. This in itself is an improvement: users are now leveraging a single, centralized, IT-governed environment instead of siloed on their laptops.

Users can also connect to existing IT-provisioned enterprise data sources to ensure that their discoveries are founded on gold-standard data. Supported enterprise sources include Oracle BI Server and anything with a JDBC interface, including Hive and other SQL-on-Hadoop products. Once IT has established a connection, users can browse the information in the Data Source Library. To use a data set, they simply enter their security credentials to the underlying enterprise source, then are guided through a wizard that helps them select portions of the enterprise data they’d like to include. When they’re satisfied, the chosen data (up to the IT-specified maximum number of records) is loaded into Endeca Server and the user is brought to their new application.

**Smart Applications**

During ingest, the provisioning service profiles the data. Based on that profile it pre-populates a discovery application and drops the user into it. Charts choose metrics and dimensions from the data, and immediately present them for analysis. Other components make smart presentation choices: for example, if the number of values for a numeric attributes exceeds a certain threshold, it displays in faceted navigation as a range filter instead of a list of values. This intelligent auto-configuration lets users start exploring data immediately, without either them or IT having to stop to build a page first. When faced with unfamiliar data and uncertain goals, getting hands-on with the data right away is a huge advantage.
Figure 4. A pre-populated app with search box, faceted navigation, chart, and results table. There has been no manual configuration.

Figure 4 shows Studio’s default template. IT can stick with this or build their own featuring other auto-configuring components like tag clouds, results lists, and maps. Components not only show up ready for interaction but also provide options for on-the-glass configuration, for example changing the metric, dimension, and/or series on a chart.

**Self-Service Mashups**

Users can access a data source library from within any discovery application. From the library, they can add their own data or select any IT-provisioned source. It’s easy to modify data or metadata when selecting a source. After selecting the source they’d like to add, data is ingested in the background and users are brought to a new page in their application that displays the new information as it’s loaded.

Refinement rules link equivalent attributes across data sets, so that filtering on one page of an app filters on the other. For example, a “Product” attribute in a sales enterprise database might correspond to a “Mentioned Product” attribute that’s been derived from online customer reviews; filtering by “camera” in one attribute would filter by “camera” in the other.

The provisioning service automatically creates refinement rules between data sets for attributes that meet the following criteria:

- Same attribute name
- Same data type
- Same assignment type
- Same selection type.

This enables users to seamlessly continue their exploration across datasets.

**Summary of Studio Data Management Features and Benefits**

- **Fast, interactive ingest.** Users can be in a discovery application finding insights in the time it would take them to open a large Excel file on their laptops. The Studio provisioning service previews the data and offers several opportunities for the user to adjust metadata, clean up data, and even split or merge fields.
• **No modeling required.** The provisioning service ingests both spreadsheets and irregular JSON files with nested structures with no demands on the user.

• **Secure connection to IT-curated enterprise sources.** Simple wizards let IT establish a connection to enterprise sources, including databases, data warehouses, OBI subject areas and big data sources. Business users can see all these sources in the Data Source Library. After submitting their credentials for the underlying data source and optionally applying filters or adjusting metadata, they tell Endeca Server to index the data and immediately start exploring.

• **Easy mashup of data with refinement rules.** Shrinks the gap between wanting to explore multiple data sets together and doing it. Choose a source, and the provisioning source automatically maps equivalent fields to each other, so that refining on an attribute in one data set refines on its counterpart in the other data set. A menu provides an opportunity to manually adjust these refinement rules as desired.

• **Jump-start discovery apps.** The provisioning service’s analysis of the data helps Studio create a basic application that gets the user exploring right away. The more unfamiliar the data, the more this intelligence launches the user down a productive path.

**Building Visually Rich Discovery Applications**

OEID Studio is an easy-to-use, visually-rich environment for building and using enterprise-class discovery applications. Blending a core interface pioneered in online commerce with a library of best-practice interactive visualizations, Studio leverages the full power of Endeca Server to let users experience free-form contextual navigation and sophisticated interactive analytics, enabling an ongoing dialogue with the data. With drag-and-drop composition, pre-populated application templates, and smart auto-configuration, any user can start discovering the moment the data loads, then iteratively enhance their application as they learn more.

**Composability**

Studio implements the vision of naturally-evolving, effortlessly-composable discovery apps by making all parts of the discovery experience intuitive, clear, and elegant. Whether it’s searching through existing applications, ingesting data, adjusting metadata, configuring a component, mashing up sources, sharing insights with others—Studio treats every aspect of discovery as essential.

For data discovery to work, anyone who can consume a discovery application should be able to create one. This is why Studio’s charts, tag clouds, and maps not only configure themselves as soon as they’re dragged onto the page but also provide elegant point-and-click configuration menus. Composability might seem a strange thing to tout—vendors will more typically brag about their Pareto charts—but experience has shown that ease-of-use is essential to scaling self-service discovery in the enterprise. Business users wants to add data, ask questions, see patterns. When they need to make a decision and can choose between submitting a request for IT or building it themselves, differences in usability often prove to be decisive. Dragging an auto-configuring component with a sleek, clear menu onto an intuitive discovery dashboard and seeing the data immediately frees analysts to do what they do best: use their domain knowledge and curiosity to make crucial discoveries. Their thirst for information should be the limiting factor in discovery—not their dexterity at navigating complex analytics software.
Integrated Discovery

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.

Typical Studio discovery applications combine some or all of the following components:

- **Search box.** Industry-leading search with contextual typeahead suggestions.
- **Faceted navigation.** Organizes available data at a glance in a familiar e-commerce-style interface. Native support for range filters and hierarchies.
- **Charts.** From simple bar charts to conformed-dimension and multi-dataset scatter-bubble charts, Studio’s dynamic charts capture patterns and trends in an attractive, instantly-digestible form.
- **Tag clouds.** Perfect for exploring terms extracted by Endeca Server’s data enrichment framework. On the fly, users can swap both dimensions and the metrics used to calculate the size of tags in the cloud. Also offers a list view to show terms in descending order.
• **Maps.** Automatically plots data by geocodes and allows visualization of several layers, including aggregate and heat layers.

• **Summarization bars.** Tracks key metrics, spotlights important dimension values, and flags records that meet user-specified criteria.

• **Pivot and result tables.** Splits and summarizes data by a number of dimensions, and provide color highlighting.

• **Results list and record details.** Shows everything you want to know about a certain record.

Each of these components serves a dual purpose: displaying a visual summary of the available data and presenting a way to refine the available data by certain values.

Consider a heatmap.

![Heatmap Image](image)

It instantly draws the user’s eye to areas with heightened activity. By updating automatically in response to filter changing—not only in the value it displays, but in where it pans on the map—the map keeps the user in context. At the same time, it provides three avenues for refinement.
First, a geographical lasso filter lets users select an area on the map.

Second, a search bar lets a user who wants to focus on a certain area zoom directly to that area by typing in a city name.
Third, each dot on the map presents a list of record details when clicked on; values within this popup can be chosen to refine upon.

Every component offers this blend of visualization, summarization, and filtering.

All Studio components respect and obey the filter state. In ways both obvious (charts cascading to a new dimension; tag clouds only showing terms in the available records) and subtle (available refinements showing only attributes that could lead to a further refinement; typeahead only suggesting values that pass the current filter), a Studio discovery application is a coherent, unified whole. A refinement from any one component propagates to all the others—a text search filters a heatmap; a click in a chart narrows a range filter; a range filter limits a text search. Refinements can be as easily removed as they are added, meaning users can move back through their navigation intuitively, and change it as they go. Additionally, Studio offers a unique capability to exclude data (negative refinements), presenting users an elegant, easy way to filter out noise and hone in on critical information. At every step, a Studio discovery application shows the data from several directions and provides multiple avenues for exploration.

**Enterprise-Class Administrative Control**

As befits a data discovery platform built for the enterprise, Endeca Information Discovery comes with a host of essential security and administration features.

- **Integration with existing credentials.** EID integrates with LDAP/Active Directory, NTLM, OpenSSO, and SiteMinder.
- **Role-based access control.** Administrators can establish distinct user communities and assign groups of users with different levels of access to certain applications.
• **Secure self-service.** IT-provisioned data sources like enterprise data warehouses and Oracle BI Server subject areas retain their underlying security; users are prompted for credentials when they try to load data from these sources. EID balances end user innovation with IT governance and control.

• **Attribute-level application filters.** User groups can be limited to viewing only certain values for an attribute, or can be prevented from seeing an attribute at all. All user-facing aspects of EID respect these filters; for example, excluded attributes or values won’t show up in search suggestions or typeahead.

• **Easy access to performance and security settings.** Studio exposes panels for IT administrators to use to adjust performance and other desired settings.

• **Auditing.** Studio visualizations show how and when applications are being used, and who’s using them. These auditing capabilities help administrators spot performance problems or determine which apps should be retired or enhanced.

• **Application templates for self-service.** IT can choose what components will be included in self-service apps by default.

**Summary of EID Studio’s Capabilities and Benefits**

• **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, Oracle Endeca Information Discovery analytic applications can be iteratively updated 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 easy application configuration through a highly interactive visual design environment; support for displaying and interacting with all kinds of structured, semi-structured, and unstructured data; and 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 as 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 possible, 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 without requiring predefinition.

- **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 high-volume consumer environments (where user training isn’t possible).

- **Agile delivery.** Studio facilitates an iterative approach to deployment that uncovers the true requirements of business users, minimizes risks, and speeds time-to-value. 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 hours or 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 alternate technologies.

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.

**Conclusion**

Today, data is widely recognized as a company’s greatest competitive asset, exceeding even the competitive value of its products or services. However, data acquisition alone isn’t enough. The businesses that win are analytics-savvy organizations that can make sense of the vast array of information by tapping insights from diverse sources—inside the enterprise or outside it, structured or unstructured, Big Data or small. These organizations already recognize the importance of unfettered data exploration and know that empowering their business users will yield unprecedented new insights. They also understand the value of their existing enterprise models and definitions, and are looking for a way to extend analytics without compromising security and governance. Their goal is to benefit the entire enterprise through an agile environment for data-driven analysis that inspires confidence and drives innovation.

The combination of ground-breaking enterprise architecture, data-driven orientation, and ease-of-use born of high-volume e-commerce make Oracle Endeca Information Discovery uniquely able to meet the industry’s data discovery needs. By delivering powerful self-service as part of a complete enterprise platform, EID frees business users to do what they do best within a framework of governance and standards, enabling faster and more confident decisions, reducing the IT backlog, increasing innovation, and reducing cost.
Appendix A: EID Success Stories

Many Oracle customers have successfully complemented their existing business analytics investments with Oracle Endeca Information Discovery. Here are three examples:

Automotive Manufacturing

Several years ago a large automotive manufacturer issued a massive vehicle recall related to reports of unintended acceleration leading to several deaths. While the CEO was called before Congress to explain the situation, they faced fundamental questions: “Is this a real quality problem, or something else? How exposed are we if it is a quality issue? What are our customers saying about it and how is it affecting our sales?”

The company is a very happy Oracle Business Intelligence customer, but there were no reports to answer these questions. Using Oracle Endeca Information Discovery, they were able to combine a variety of data from their warehouse and beyond – vehicle data, quality reports, internal warranty claims, sales transactions, service records, supply chain data, and more. When new questions required data from outside the company they were able to readily incorporate claims from the National Highway Transportation Safety Authority and competitor sales data from JD Power. Only by combining all of this data – replete with misspellings and bad grammar – did they have the right infrastructure in place to enable line of business workers to understand what was happening.

The quality engineers, the marketing organization, and the team managing the supplier relationships had the expertise to ask questions about vehicles, suppliers, manufacturing processes and facilities, but they didn't have the expertise to write advanced queries or build reports. Oracle Endeca Information Discovery enabled these business users to easily explore, analyze, and understand this diverse data.

After a thorough investigation, the company was vindicated. The Transportation Secretary concluded there was no electronic-based cause for unintended high-speed acceleration in their cars. Proving a negative – that the cars didn’t have an electronic problem – was tough. Oracle Endeca Information Discovery played a prominent role in exonerating the company.

The company estimated that it would have taken over a year to solve this problem with their traditional BI tools. EID reduced time to market by 80%. The company also estimated that the engineers’ ability to ask and answer their own questions as they unfolded through the investigation saved hundreds of thousands of hours they would have had to spend waiting for reports to answer their new questions.

Consumer Beverages

A major consumer beverages company needed to understand variances between demand forecasts and actuals. While this is typically a problem well served by business intelligence tools, their demand planners still had additional questions based on the need to understand why inaccuracy existed in the demand plan. They wondered: “Could variations be due to unanticipated trade promotions with customers? Does pricing impact the accuracy of the demand plan? What about unanticipated shipments of products between distribution centers?”
They built a discovery application for the demand planners that combined the forecasts out of SAS with the actuals from the distribution transaction system, and then connected a separate marketing database with the other two sources. When they saw that some of the variances were still unexplained the planners had more questions: "Do promotions offered by our sales team lead to unanticipated bulk buying?", To address this they loaded Trade Promotion data from outside the data warehouse. Then the planners asked: "Did our customers affect demand by changing their prices? Did competitor pricing impact demand?" They then combined sales and pricing data acquired from 3rd party sources. All of this happened over the course of 8 weeks.

Finally, planners discovered something they didn't expect. When they asked the question, "How do out-of-lane shipments between distribution centers impact forecast accuracy?", they actually found that unauthorized overrides to the demand plan being performed by individuals in the field had helped to improve accuracy of the forecast. This was due to tribal knowledge of business conditions, impossible to predict in the planning process. These tribal business practices have now been captured and replicated across the business leading to accuracy improvements of between 2-5%.

**Commercial Food Production**

The world population hit 7 billion last year. A large processed food producer realized that corn yields needed to increase from 150 bushels an acre to about 200 to feed a growing world. One division sells and distributes new strains of seed to increase farmers’ crop yields. Because farmers often can’t weather even a single season of poor yields, they were unlikely to use a new strain of seed without a concrete reason for the change. The food producer had to make the case with data.

Fortunately, there is lots of data available, but the challenge lay in combining it cost-effectively and making it usable and useful. Oracle helped this food producer combine data from many sources including a transactional warehouse that indicated which farmer had bought what, a marketing database that indicated which farmer had been pitched what seed, and a separate transactional warehouse with data from "answer plots" that the company had planted all around the US at different latitudes in different soils with different seeds to demonstrate the actual yields. Finally, data from all of these sources were combined with government data on how many acres are planted with which crops. Data from these multiple sources, some of which were outside the company’s control and could change at any time, were combined to derive insights.

This application is now used by thousand of salespeople, many of them former farmers. The company expects higher profit margins as a result. They have estimated they saved 1.5 years and $4M by solving this problem with Oracle Endeca Information Discovery.