CHAPTER 1

Oracle Endeca Information Discovery Architecture
This quotation, from John Kenneth Galbraith's classic title chronicling the stock market crash of 1929, is a warning for all generations. The quote comes from a chapter titled “The Twilight of Illusion,” which discusses the common mental modes in the period leading up to the crash. Galbraith describes these mental modes as a “failure to know what isn’t known.”

The quote has particular relevance for the information age we now inhabit. Information technology has become part of every aspect of our professional lives. Corporate operations, manufacturing, healthcare, communications, and commerce all depend on enterprise systems. The information driving these systems primarily resides in relational databases, and business intelligence systems have been used for more than two decades to provide operational reporting and analytics against this information. However, there are also new sources of information, including social media, e-mail, and blogs, that can be critical to decision making. The traditional approach to business intelligence systems was not designed to support the integration and analysis of data originating in these new sources. The inability to access this information is, in the words of Galbraith, a “failure to know what isn’t known.” Oracle Endeca Information Discovery is complementary to business intelligence systems and facilitates using information from these new data sources. This chapter introduces Oracle Endeca Information Discovery. We start this introduction by comparing and contrasting aspects of Oracle Endeca Information Discovery to business intelligence systems.

Oracle Endeca Information Discovery vs. Business intelligence

To develop an initial understanding of Endeca, you need to appreciate the difference between Endeca and business intelligence systems. Business intelligence systems, like Oracle Business Intelligence Enterprise Edition, provide answers to known questions in reports and dashboards, and they rely on structures known as data models to define the data. Data models are inherent to relational database management systems; they document the relationship between tables in a database. Information is retrieved from the tables in the data model using Structured Query Language (SQL). The tasks of designing and maintaining data models and writing queries can be arduous and require an information technology professional who is skilled with data models and queries as well as knowledgeable with the information the tables contain. This familiarity with the information in the tables is often referred to as domain knowledge: It is the knowledge of subject-matter practices and procedures and how they are implemented in the data model and table data. A further complication is that the information in the tables change, requiring data models and queries to be modified.

Endeca differentiates itself from traditional business intelligence products in many ways. First, Endeca does not require users to formulate questions in advance of looking at the data. This is great for any circumstance where the user has no idea where to start or has only a hunch. Endeca is designed to automatically summarize every attribute in the data and allow users to further refine the scope of the data. Users can explore the data based on summarizations presented “in the moment.” This facilitates an ongoing dialogue with the data, allowing information to be
evaluated immediately and its importance determined, before deciding that the data is relevant to
decision making. Another difference between Endeca and business intelligence tools is related to
data models. Business intelligence systems require data models be designed and implemented
before any meaningful use of the information can occur, whereas Endeca derives its models from
the information. This eliminates the time and labor required when creating data models and
eliminates the complications that can occur when the information structure or content changes.
(This is a key feature of Endeca and will be covered at length in this book.) As was stated
previously, Endeca allows users to explore and evaluate information. By comparison, for business
intelligence systems, this type of activity is typically not performed by users, but requires an
information technology professional with domain knowledge who is skilled in writing queries and
who understands underlying data models. Figure 1-1 shows Endeca’s web-based user interface
that facilitates data exploration.

Perhaps the most compelling way that Oracle Endeca Information Discovery differentiates
itself from business intelligence systems is its ability to handle different types of data. Data
residing in rows and columns in relational databases belongs to a type of data known as
structured data. Structured data has strongly defined data types primarily for storing text or
numeric data, along with a well-defined data model. By contrast, unstructured data lacks the
fixed fields and organizational features of structured data. Information from raw text found in
documents, social media, e-mail, and blogs are all examples of unstructured data. Often, the
information originating in unstructured data is called dirty because it is not subject to data
standards and may contain ambiguous meanings, spelling errors, and abbreviations. Between
structured and unstructured data is semistructured data, which has internal organization but lacks a data model. Two common examples of semistructured data are Extensible Markup Language (XML) with its tags and attributes and JavaScript Object Notation (JSON), an open format used for web services. Data in comma-separated value (CSV) files can also be classified as semistructured.

Business intelligence systems are designed to work with structured data, but have challenges with unstructured data.

Oracle Endeca Information Discovery is designed to process any type of data, regardless of structure, and make it available by inferring flexible data models, accommodating dirty data, and even determining relationships to other data sources that have already been processed, regardless of their type. Table 1-1 summarizes the data types we’ve covered.

### Table 1-1. Data Type Summary

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Characteristics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>Data model that has defined rectangular tables containing rows and columns and has referential integrity between tables</td>
<td>Relational databases (RDBMSs)</td>
</tr>
<tr>
<td>Semistructured</td>
<td>Loosely defined data model, internal structure</td>
<td>XML, CSV, spreadsheet files (Microsoft Excel), JSON</td>
</tr>
<tr>
<td>Unstructured</td>
<td>No data model</td>
<td>Documents, e-mail, social media, blogs, customer review comments, media files including image files, and video</td>
</tr>
</tbody>
</table>

Oracle Endeca Information Discovery is designed to process any type of data, regardless of structure, and make it available by inferring flexible data models, accommodating dirty data, and even determining relationships to other data sources that have already been processed, regardless of their type. Table 1-1 summarizes the data types we’ve covered.

### Introducing Oracle Endeca Information Discovery

Endeca was founded in 1999 in Cambridge, Massachusetts, with a company name derived from the German verb *entdecken*, meaning “to discover.” Endeca was founded at the height of the “dot com bubble” and initially produced a database to be used to expedite web searches related to e-commerce. After the bubble burst, Endeca transformed itself from a product company to a platform company, and Endeca’s core in-memory database engine and associated products diverged, one continuing to support e-commerce solutions and the other emerging as an alternative business intelligence technology. E-commerce solutions powered by Endeca Server are ubiquitous and familiar to anyone who has shopped for products on the Web. Many of America’s “big-box” retailers use Endeca to drive product search capability. Consumers wanting to purchase a washing machine, for example, can refine their search by manufacturer, price, capacity, and many other options. Figure 1-2 shows an example retail web site for washing machines, with refinements on the left.

As the concept of data discovery became more commonplace, Endeca became known as a leader in this new genre of information technology. *Data discovery* is a term used to describe a set of application features, including the following:

- A data server with the capability to store and model data from disparate sources that does not require predefined metadata.
A data server with in-memory processing and indexing to maximize performance and eliminate the need for aggregations, summaries, or precalculations to occur prior to application usage.

Users can develop and refine views and analyses of structured, semistructured, or unstructured data.

The user interface is optimized for ease of use, allowing users to create and modify applications.

On October 18, 2011, Oracle Corporation announced its acquisition of Endeca. At the time of this acquisition, Oracle already had one of the most competitive business intelligence offerings with Oracle Business Intelligence Enterprise Edition (OBIEE). OBIEE offers best-in-class structured data management and analytics, and Oracle’s decision to acquire Endeca enabled Oracle to add best-in-class data discovery and unstructured data management to its portfolio of products. Oracle Corporation has a remarkable track record of acquiring companies and their offerings and quickly integrating these acquisitions with existing offerings. Oracle Endeca Information Discovery is no exception to this; Endeca can be deployed on Oracle’s WebLogic application server running on Oracle Exalytics. Endeca is certified to run on Exalytics, taking the guesswork out of hardware selection and ensuring an optimum configuration to enable superior performance and scalability.
Master Competitive Analytics with Oracle Endeca Information Discovery

One of the latest features of Endeca is tight integration between it and OBIEE. Endeca has the ability to source data from an OBIEE metadata repository into Endeca Server, allowing users to query and load data from OBIEE. Guided navigation and search can be used to analyze data and join data already in Endeca. This opens the door to enriching unstructured data with structured data ingested from OBIEE. Another compelling use of this tight integration would be to allow data discovery to occur on data from OBIEE and find answers to new questions without needing to make changes to the OBIEE repository.

Oracle also has one of the most compelling big data offerings with its Big Data Appliance, and Endeca’s ability to connect to a Hadoop cluster on Big Data Appliance means Oracle has offerings for every possible data storage and analytic scenario.

Oracle Endeca Information Discovery Overview
Endeca is not a single product, but a number of products and product features that are separately installed and licensed. To understand the capabilities of Endeca, it is best to start by contextualizing it against a business intelligence system. Like a typical business intelligence system, Endeca includes the following major components:

- A data server
- A data integration tool
- An analytics toolset

The Oracle Endeca Information Discovery features that differentiate it from a traditional business intelligence system are

- No need to predetermine questions to be answered by the information before ingesting the data
- End-user provisioning and exploration of data via self-service interfaces
- Ease of integration and enrichment of semistructured and unstructured data
- Unification of full-featured, advanced keyword search; data-driven guided navigation; and in-memory analytics

Let’s look at the major components, compare them to a business intelligence system, and discuss how they deliver these differentiating features.

Data Server
A typical business intelligence deployment involves a database server designed specifically to store data for rapid retrieval, using indexed and denormalized tables. Likewise, Endeca features a data server designed for rapid retrieval of information acquired through information discovery activities. This design of the data server is a departure from that of a typical RDBMS, however: Data is stored as key-value pairs, and for every attribute, a full inverted search index and a membership index are created to quickly determine the association between attributes and records. The design is optimized for in-memory performance, allowing the retrieval of data to drive analytics at interactive speeds. The server also features its own query language known as
Endeca Query Language (EQL), which is a rich set of SQL-like features providing both basic and complex aggregations of search results or the current navigation state.

Data Integration Tool
A typical business intelligence deployment features an extract-transform-load (ETL) tool to acquire information from other databases and data stores, perform transformations on the information, and insert or update records in the reporting database with the transformed data. Likewise, Endeca features an agile ETL tool for IT with all the typical features of a commercial-strength ETL product, along with unique unstructured data discovery features. Let’s examine three significant features that provide capabilities above and beyond nominal ETL functionality: Text Enrichment, Text Enrichment with Sentiment Analysis, and the Integrator Acquisition System.

Text Enrichment is the “silver bullet” that allows Endeca to discover significance in unstructured data. It is an add-on feature that provides the ability to find terms and phrases in text and then rank and organize the findings. Text Enrichment includes text analysis capabilities for extracting topics and themes in the data to determine subject matter, as well as for extracting entities to expose people, places, organizations, quotes, products, and custom entities. Text Enrichment also contains summarization capabilities for automatically creating abstracts and topical summaries.

Text Enrichment with Sentiment Analysis delivers all the Text Enrichment capabilities as well as advanced text analysis for extracting opinion or feeling related to each extracted concept. Sentiment is extracted with a score indicating the positive and negative nature of a document, a phrase, or an entity. These scores are used to show varying ranges of positivity and negativity across the data at any point during the search or navigation state.

The Integrator Acquisition System is a set of components that crawl source data stored in a variety of formats, including file systems, JDBC databases, delimited files, web content, and custom data sources.

Analytics Toolset
A typical business intelligence deployment features an analytics toolset. Likewise, Endeca features a web-based analytics suite designed to require little or no training to build or consume new discovery applications. The in-memory performance of the Endeca data server enables analysis to be performed at interactive speeds using an intuitive UI pioneered in online commerce, where ease of use is critical to mass consumer adoption. The Endeca UI combines the best of search, guided navigation, and in-memory analytics to provide the end user with all the tools necessary to discover new insights in structured and unstructured content. The Endeca analytics toolset also allows end users to drag and drop from a library of discovery components to easily create their own applications for personal or small-scale use. It also allows IT to create advanced discovery applications with security and failover that may be published to and consumed by the end-user community. Endeca fully supports both IT provisioned discovery applications and self-service discovery. In self-service mode, end users may upload their own data sets from personal files on their desktop (such as Excel or JSON) and then “mash up” this data with data enterprise sources, including the data warehouse or the Oracle Common Enterprise Information Model.

We’ve just compared the major components of Endeca to typical business intelligence systems, so you should now have a general understanding of Endeca. Next let’s look deeper at these components to gain a better understanding of the Endeca architecture.
Oracle Endeca Information Discovery Core Components

Oracle Endeca Information Discovery comprises three core components. The sections that follow provide an overview of these components and how they relate to the major Endeca features that have been discussed. Chapters 2, 3, and 4 cover these components in greater detail. Figure 1-3 shows an overview of the Endeca core components. These core components are:

- **Endeca Server**  A hybrid search and analytic database with capabilities to incorporate wide, diverse, and changing data; architected for discovery
- **Endeca Information Discovery Integrator**  A set of tools for loading and optionally enriching diverse information, including structured, semistructured, and structured data
- **Endeca Information Discovery Studio**  A discovery application composition environment providing drag-and-drop authoring to create interactive, visually rich information discovery applications

![FIGURE 1-3. Endeca core components](image-url)
Chapter 1: Oracle Endeca Information Discovery Architecture

Endeca Server

Endeca Server is at the heart of Endeca and is the core search-analytical database. In Endeca Server, data is organized using a highly flexible data model known as a faceted data model. With this data model, it is not necessary to define a unified schema before loading and analyzing data; data models are derived from the data that is stored in the database, and every record has its own schema based on its own generated attributes. This is irrespective of the data source or whether the source is structured or unstructured.

Data Model

Structured data can be directly loaded into a faceted model using standard ETL tools. Each row becomes a record, and each column becomes an attribute. An example of such data is a sales transaction. Each transaction row becomes a record, and every element of the transaction record becomes an attribute.

Semistructured data from enterprise applications, various feeds, and XML sources can also be loaded as attribute and value pairs. This is a common cause of heterogeneous record structure. For the sales transaction record, it is possible to extend it with more information about specific products. The attributes that go with each product could be very different depending on the product. For example, a road bike has different components than a mountain bike. With Endeca, attributes become attribute and value pairs, and jagged records begin to emerge that look dissimilar, meaning the data sets do not have the same data model but have some commonality between them. With relational database technologies, this is difficult to implement, but the key-value pair data structure in Endeca makes it possible to implement and extend.

With Endeca Server, unstructured data can be linked to records by any available key. In addition, unstructured elements can be stored as their own records for side-by-side analysis. Some examples are documents, RSS feeds, Twitter and Facebook data, and data feeds from discussion forums. To continue with the sales transaction data mentioned previously, you could now integrate into the record online customer reviews and the customers’ Facebook or Twitter comments on the product and transaction.

The way to accomplish this through Oracle Endeca Information Discovery is to take the textual fields into the records as new attributes. These attributes can be combined in the same record with the sales transactions, product details, and customer review information. Here is a summary of the data sources in the examples we mentioned:

<table>
<thead>
<tr>
<th>Data Source Name</th>
<th>Data Source Type</th>
<th>Data Source Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>Relational data</td>
<td>Oracle Database</td>
</tr>
<tr>
<td>Product details</td>
<td>Semistructured</td>
<td>XML</td>
</tr>
<tr>
<td>Employee information</td>
<td>Relational data</td>
<td>Oracle Database</td>
</tr>
<tr>
<td>Online reviews</td>
<td>Unstructured</td>
<td>Text files</td>
</tr>
</tbody>
</table>

Endeca allows the mapping of sales transactional records with product details and employee information through a product ID and a sales rep ID. The online reviews can also be mapped to the transaction records through transaction IDs that are captured. In addition, users can create a whitelist text tagger to tag the employee first name or last name mentioned in the review text to an employee by full name. Now unstructured, semistructured, and structured data is aligned and loaded and could be analyzed side by side.
Data Domains
Data domains are the largest unit of data over which Endeca Server allows queries to be expressed. No data domains exist on an Endeca Server instance immediately after installation, so a server administrator must create each data domain. Users can specify meaningful names such as staffing, sales, or marketing to identify data domains. The Data Ingest Web Service (DIWS) facilitates loading data into a data domain.

Records and Attributes
Records are the fundamental unit of data in a data domain on Endeca Server. As data is loaded, or ingested, it is stored in data records. Data records generally correspond to traditional records in a source database, but differ in that they are standardized for consistency and classified with attributes. Attributes are the “facets” of the data; they are the storage for metadata. As an example, consider opinion data ingested from unstructured data regarding consumer opinions on automobiles. The data records would have attributes indicating automobile make, model, year, a sentiment such as “recommend” or “lemon,” and a feature of the automobile such as “handling” or “transmission.”

Dgraph
For each data domain, Endeca Server creates a run time process called Dgraph that manages data domain operations. The Dgraph process of Endeca Server is the main computational module that provides the features of Endeca Server, such as search, refinement computation, and guided navigation. Dgraph maintains indexes of records that are searchable documents of domain data associated with attributes.

The Dgraph is stateless, which facilitates the addition of Dgraph processes for load balancing and redundancy. When more processing capability is required than can be achieved on one Endeca Server instance, an Endeca Server cluster can be deployed.

Endeca Server Clustering
Endeca Server clustering can be used in production deployments to handle heavy workloads and is deployed across multiple servers. Nodes can be added to an Endeca Server cluster as additional processing needs emerge, ensuring the scalability of the Endeca Server cluster. The central feature of the Endeca cluster architecture is a data domain cluster, which is a set of Dgraph processes that handles requests across multiple nodes. One of the Dgraph processes is designated as the leader node, and all other Dgraph processes are referred to as follower nodes. The leader node handles all write and update requests, while the follower nodes allow only read operations. A shared file system is used for disk-based versions of indexes, and only the leader node has write access to the file system. The node hosting the leader node Dgraph process must also host the Cluster Coordinator service, which is responsible for intercluster communications between Dgraph processes and for notifying follower nodes when indexes or data has changed.

Endeca Query Language
Endeca Query Language is a powerful integrated analytics language built within Oracle Endeca Server. EQL enables power users to define and create new metrics to compose their own discovery applications. Built on the core capabilities of Oracle Endeca Server, EQL extends the capabilities of Oracle Endeca Server with a rich analytic language that allows users to explore aggregated and pivoted views of large volumes of data. EQL also supports a variety of data types,
including numerical, geospatial, and date/time values that enable applications to work with temporal data, performing time-based sorting, filtering, and analysis. IT 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. Some of the most important EQL features include tight integration with search and navigation, rich analytical functionality, and processing efficiency, as well as a familiar development experience.

This is an example of a simple EQL statement:

```
RETURN Sales AS SELECT
    AVG(FactSales_Amount) AS AverageSalesVolume,
    SUM(FactSales_Amount) AS TotalSaleVolume,
GROUP BY DimDate_CalendarYear
```

An EQL statement starts with either DEFINE or RETURN. DEFINE doesn’t return the result set. Rather, it creates the data set as a temp table. A DEFINE statement is typically followed by a RETURN statement that consumes the result set. Endeca uses the DEFINE statement to create views that can then be used to generate charts and other advanced visualization.

The statement then needs to have one or many SELECT elements, separated by commas. The SELECT clause is composed of an expression followed by an alias. Expressions are usually one or more attributes, operators, or functions such as summation or average, as you see in the previous example.

The GROUP BY clause specifies the method of aggregation. Other EQL capabilities include joining, sorting, paging, and filtering. Oracle’s documentation for Endeca Server includes product documentation titled “Oracle Endeca Server: EQL Guide.” This is an extensive reference for EQL, and will help developers and users learn more about EQL’s capabilities.

**Services**

Most of the Endeca Server APIs are exposed as SOAP web services. These services are used by other Endeca components to interact with Endeca Server. The major services are briefly discussed in the sections that follow.

**Data Ingest Web Service**

Data Ingest Web Service (DIWS) provides an interface to ETL tools to load data into the data domains hosted in Oracle Endeca Server.

**Conversation Web Service**

This web service provides the primary means of querying data in the data domain hosted in Oracle Endeca Server. This service is used by Endeca Information Discovery Studio to query Oracle Endeca Server.

**Endeca Server Version Information**

Be aware that Endeca Server version numbers do not coincide with the version numbers of the other core components. The current version of Endeca Server is 7.6, whereas the current versions of Endeca Information Discovery Integrator and Endeca Information Discovery Studio are 3.1.
Master Competitive Analytics with Oracle Endeca Information Discovery

**Entity and Collection Configuration Web Service**
This web service, also known as sConfig, allows you to create and update collections and views for collections or for data sets.

**Configuration Web Service**
This web service is for updating the schema and configuring the records in a data domain.

**Endeca Information Discovery Integrator**
This section provides more detail about the major components of Endeca Information Discovery Integrator. The Endeca Information Discovery Integrator consists of the following five components:

- Integrator ETL
- Integration Server
- Integrator Acquisition System
- Web Acquisition Toolkit
- IKM SQL to Endeca Server

**Integrator ETL**
Integrator ETL is part of Endeca Information Discovery Integrator and is an integration platform that enables source records to be extracted from many types of sources. Integrator is a powerful graphical-based development environment; its primary purpose is to develop, debug, and execute ETL processes. The graphic diagrams of these processes are referred to as graphs. A graph is a series of sequential components that process data. The user interface for Integrator ETL is based on the popular open-source Eclipse integrated development environment (IDE). Integrator ETL can run graphs without Integrator Server; when the integration server is not used, the graphs are run on the machine running the Integrator ETL IDE.

Integrator ETL is very capable and supports connectivity to a wide variety of other systems and software. JDBC is used to provide connectivity to the following databases:

- Oracle RDBMS
- MySQL
- Microsoft SQL Server
- Informix
- IBM DB2
- PostgreSQL

Bulk loader capability is available for Oracle, Microsoft SQL, Informix, and IBM DB2.

Integration with systems supporting Java Message Service (JMS) is available for the following:

- IBM WebSphere MQ
- Apache Active MQ
- JBoss Messaging
Chapter 1: Oracle Endeca Information Discovery Architecture

The following protocols are also supported:

- HTTP
- HTTPS
- FTP
- FTPS
- SMTP

Some proprietary formats are also supported, which include

- Microsoft Excel
- FoxPro (dBase)

**Integration Server**

The integration server allows graphs to be treated as a production process and run in an enterprisewide team environment. As discussed in the previous section, graphs can run in Integrator ETL. The integration server provides an alternative to this and should be considered for any large-scale deployment of Endeca. As with any production process tool, the integration server allows graphs to be scheduled and monitors the execution status of graphs.

The integration server comes with a web-based user interface for configuration and administration. In addition, the integration server features an API for remote operations and interoperability with other systems.

**Integrator Acquisition System**

The Integrator Acquisition System (IAS) is a set of components that crawl source data stored in a variety of formats, including file systems, JDBC databases, flat files, web sources, and custom data sources. IAS transforms the data, if necessary, and outputs it to an XML file or a record store that can be accessed by Integrator ETL. Within IAS there are two major processing entities: the IAS Server and the Web Crawler.

The IAS Server crawls JDBC sources, file systems, or custom-created sources. The number of types of sources and file formats supported is extensive. Table 1-2 lists the types of sources and some examples to illustrate the power and versatility that is available with this product. For a complete list of supported formats, please refer to the appendix.

The IAS Web Crawler is installed by default as part of the IAS. The IAS Web Crawler gathers data by crawling HTTP and HTTPS web sites. Once a crawl is completed, the Integrator ETL can access the data acquired during the crawl. A crawl usually writes data directly to Endeca Server. However, data can also be written to an XML file for debugging and development. The Web Crawler is for large-scale crawling and has an architecture that allows developers to create custom plug-ins. Plug-ins provide a means to extract additional content, such as HTML meta tags, from web pages.

**Web Acquisition Toolkit**

The Endeca Web Acquisition Toolkit (WAT) provides an intuitive, simple-to-use graphical interface for collecting content from the Web, allowing users to rapidly access and integrate any information
Master Competitive Analytics with Oracle Endeca Information Discovery

<table>
<thead>
<tr>
<th>File Format Type</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive</td>
<td>File compression formats</td>
<td>ZIP, UNIX tar, UNIX Gzip, self-extracting EXE, 7z, RAR, LZH</td>
</tr>
<tr>
<td>Database</td>
<td>Database formats generally used for stand-alone databases</td>
<td>Microsoft Access, Microsoft Works DB, Paradox, Dbase</td>
</tr>
<tr>
<td>E-mail</td>
<td>E-mail formats, generally used by e-mail clients</td>
<td>Outlook formats, Apple Mail Message (EMLX), many encoded mail message formats, Lotus Notes</td>
</tr>
<tr>
<td>Multimedia</td>
<td>Formats used for audio and video</td>
<td>AVI, MPEG, MP3, Real Media, QuickTime, WMA, WAV</td>
</tr>
<tr>
<td>Other formats</td>
<td>A variety of formats, used for many different applications</td>
<td>vCard, vCalendar, Yahoo Messenger, MS Live Messenger, AOL Messenger, Microsoft Windows files including DLL, executables, shortcuts, help</td>
</tr>
<tr>
<td>Presentation</td>
<td>Used by products such as Microsoft PowerPoint and similar</td>
<td>PowerPoint, OpenOffice Impress, Lotus Freelance, Harvard Graphics</td>
</tr>
<tr>
<td>Raster image</td>
<td>Files used for image storage, including all major formats and many lesser-known formats</td>
<td>JPEG, GIF, TIFF, PNG, .BMP</td>
</tr>
<tr>
<td>Spreadsheet</td>
<td>All major formats supported and many older, lesser-known formats</td>
<td>Microsoft Excel, Lotus 1-2-3, OpenOffice Calc, Quattro Pro</td>
</tr>
<tr>
<td>Text and markup</td>
<td>Includes character sets and markup languages</td>
<td>ASCII text, EBCDIC, MS-DOS Character Set, RTF, HTML, XHTML, XML (text only), Unicode text, UTF-8</td>
</tr>
<tr>
<td>Vector image</td>
<td>Formats used by document and drawing programs</td>
<td>Adobe PDF, Adobe Illustrator, AutoCAD, Visio, CorelDRAW, OpenOffice Draw</td>
</tr>
<tr>
<td>Word processing</td>
<td>Supports all of the best-known word processing programs, along with legacy types</td>
<td>Microsoft Word, Microsoft Works, OpenOffice Writer, WordStar, WordPerfect</td>
</tr>
</tbody>
</table>

TABLE 1-2. Example IAS Server Crawl File Types

Exposed through a web front end. Endeca WAT allows for information to be collected from many sources, including content from consumer sites, industry forums, government or supplier portals, cloud applications, and other big data sources.

Endeca Web Acquisition Toolkit Design Studio is an IDE for building data integration workflows. Endeca Web Acquisition Toolkit Design Studio combines the best aspects of a web browser and a visual flow editor and eliminates the need to write code by enabling developers to visually navigate applications and data sources and by using a powerful XML editor to generate workflows in minutes.
IKM SQL to Endeca Server
IKM SQL to Endeca Server provides an integration module that allows users of the Oracle Data Integrator (ODI) to write directly to Endeca Server.

Endeca Information Discovery Studio
Oracle Endeca Information Discovery Studio is a web-based application that allows business analysts to rapidly assemble dashboard applications. These applications enable analysts and other end users to explore a full range of structured and unstructured enterprise data from Endeca Server. Each application consists of one or more pages, with each page containing a set of graphical components. Endeca Information Discovery Studio components include functions to

- Navigate to or search for specific data
- Display detailed information about data
- Display graphical representations of the data
- Manipulate and analyze the data
- Highlight specific data values

Endeca Information Discovery Studio is easy to deploy and is ideal for the agile development of enterprise-quality applications. Endeca Information Discovery Studio provides a library of UI components that embody best practices in information discovery applications. Because Endeca Information Discovery Studio is component based, its applications are simple to control, adapt, and extend.

In Endeca Information Discovery Studio, application locale determines the language used for an application. In addition, application locale determines the default data formatting for currency, numbers, and dates. Endeca Information Discovery Studio supports the following languages:

- French
- German
- Italian
- Spanish
- Japanese
- Korean
- Simplified Chinese
- Traditional Chinese
- Portuguese-European

Note that this is a subset of the languages supported by Oracle Endeca Server.

Endeca Information Discovery Studio has administrative features that are accessible to Endeca Studio administrators and developers. Administrative features are part of the Control Panel, where the following operations are available:

- Monitoring Endeca Studio usage
- Creating and managing data sources
- Creating and managing Endeca Server connections
- Configuring Endeca Studio settings
Monitoring Endeca Studio performance
Managing and configuring applications
Managing Endeca Information Discovery Studio users and user access

Endeca Information Discovery Studio Navigation Overview
When logging in to Endeca Information Discovery Studio, the user is presented with three application subsets on the left sidebar, as shown in Figure 1-4.

Clicking the link for any of the subsets reveals a page with icons for accessing applications.
Table 1-3 describes the groups.
Users who are not administrators for any applications will be able to access only Certified Applications and Community Applications.

Application Information Icon
The application information icon allows users to view summary information about an application. This information includes data sets used, application creator, and application access (public or private).

<table>
<thead>
<tr>
<th>Application Subset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Applications</td>
<td>Displays applications for which the user is an application administrator. Incomplete applications are not displayed in this list.</td>
</tr>
<tr>
<td>Certified Applications</td>
<td>Displays applications that an Endeca Studio administrator has certified.</td>
</tr>
<tr>
<td>Community Applications</td>
<td>Displays applications that have not been certified.</td>
</tr>
</tbody>
</table>

TABLE 1-3. Endeca Information Discovery Studio Application Subsets
Chapter 1: Oracle Endeca Information Discovery Architecture

Application Icon
An application is accessed when a user clicks the application icon.

Anatomy of an Endeca Information Discovery Studio Application
Endeca Information Discovery Studio applications allow users to search and analyze data from Oracle Endeca Server data domains. An Endeca Studio application consists of two or more pages, and each page features components that enable filtering, navigating, or viewing data. Endeca Information Discovery Studio applications promote information discovery, allowing Studio users to uncover previously unknown relationships and trends as they investigate business issues. Endeca Information Discovery Studio applications can be quickly built using the data display and visualization tools provided. The following section summarizes the major features available within an Endeca Studio application.

Chart Components
Endeca applications provide insight into data using chart components. These tools will be familiar to users of spreadsheet software. Many chart components are available, including line charts, bar charts, stacked bar charts, area charts, scatter diagrams, bubble charts, and pie charts.

Tabular Data Elements
Tabular data elements allow detailed viewing of data. These elements include result tables and pivot tables. Tables support extensive formatting options, allowing users to change row and column colors, shades, fonts, and borders.

Summarization Bar
The summarization bar displays key performance indicators and can provide alerts when a critical value or condition for a metric occurs.

Search Box
A search box allows users to search for specific values in attributes. When attributes are located matching the search criteria, display components are updated immediately, including charts, tables, and maps.

Refinement Component
Refinement components allow users to improve the displayed data using search terms or selected attribute values. As users refine the data, the other components may be updated to include only the data for the current refinement.

Tag Cloud
The tag cloud is a visualization tool that may not be familiar to many users, but it provides a unique visualization experience. Tag clouds display all the attribute values for a data set, with more prevalent attributes being displayed in larger text. For example, for attributes whose values are for a brand of car, the text might display the words Ford, Chevrolet, Chrysler, Volkswagen, and so on. A value for each can be displayed with the text. Figure 1-5 shows an example of a tag cloud.
Map Components
The map component uses Oracle's map viewer to display and analyze geographic information contained within data. With the map component, three types of layers are possible:

- Numbered point layers display numbered points on a map and corresponding detail for each numbered point.
- Point layers display points on a map and can have multiple layers active, using colors to differentiate between layers.
- A heat map displays a point for each location on a map and a shaded cloud with gradient colors to show relative density of the points on the map or the change in associated value between locations. Figure 1-6 shows an example of a heat map.

User Convenience Features
In addition to the data display components just covered, an Endeca Information Discovery Studio application has features that are convenient and useful:

- Bookmarks allow users to save a given navigation and component state so that they can return to it at a later time or e-mail it to other users.
- Data can be exported from components to CSV files for use outside of Endeca Information Discovery Studio.

Data Sets
Data used in an Endeca Studio application from one source is called a data set. It is possible for an Endeca Studio application to have only one data set, but most applications will have multiple data sets. Data sets can be from an Endeca Server data domain, and they can be from sources other than an Endeca server, including user-uploaded Excel spreadsheets, user-uploaded JSON files, JDBC data sources, and Oracle BI servers. Application administrators have many options for
managing data sets, including using filters and specifying columns to be used from tables. As data sets are created from imported or uploaded data, their attributes are determined and made available.

Text Enrichment can be used to enhance data by creating new attributes. When Text Enrichment terms are specified, data sources are searched for these terms. When matches for the terms occur, attributes are created. Enrichment can be based on a single term or a list of terms. For example, the terms nice, pleasant, and fun can be used to create an attribute called positive. This is particularly useful on unstructured data. To use Text Enrichment, the data enrichment plug-in needs to be installed on the Endeca Server being used.

**Views and Attribute Groups**

Views and attribute groups are useful organization tools within Endeca Information Discovery Studio because they provide a useful mechanism for adding order to the information from data sets. Views in Endeca Information Discovery Studio are similar in some regard to views in an Oracle RDBMS.
A view is a logical collection of information that is derived from the records in application data sets. Views are composed of attributes, which can be attributes of a data set or of a derived value. For example, a view could be composed of a list of products and a calculated total sales attribute for each product. Views are useful because they allow a subset of data to be defined and encapsulated in intuitive containers for later use.

Within views, attribute groups can organize attributes within a view. For example, an attribute group within a view containing human resources information could be created and named address info to contain all the mailing address information such as street address, city, and ZIP code.

Refinement Rules
For Endeca applications that contain more than one data set, refinement rules allow users to specify relationships between data sets. Refinement rules work best with attributes that contain the same or similar values. When a data set is added to an existing application, Endeca Studio can automatically create refinement rules for attributes that have the same attribute name, data type, multivalue setting, and refinement behavior. Refinement rules are useful because they limit attributes to those that are relevant to both data sets. For example, a data set containing human resources data and a data set containing members of an engineering group can have a refinement rule created that effectively selects only members of the engineering group.

This section summarized the major components of an Endeca Information Discovery Studio application. The components are covered in detail in later chapters.

Provisioning Service
Oracle Endeca Information Discovery Studio includes a provisioning service that facilitates dynamic application creation from data uploaded from a user's desktop. The provisioning service profiles and transforms data into appropriate formats before sending it to Endeca Server for ingestion. Endeca Studio users can upload data from the following sources:

- Microsoft Excel spreadsheets
- JSON files, such as Twitter data files
- Relational databases supported by a JDBC driver
- Oracle Business Intelligence Server

Figure 1-7 shows an example of an application created with the provisioning service using an Excel file containing publicly available city employee data. This application was created with minimal effort.

Application Administration
The Application Settings page in Oracle Endeca Information Discovery Studio is available to administrators and developers and is used for configuring and administering an application and for managing the views and groups for the application data. One of the most important administrative tasks is managing application access and application membership. This is discussed in the next two sections.

Application Access Management
Endeca administrators can set an application to be one of two application types:
Chapter 1: Oracle Endeca Information Discovery Architecture

Public applications are available to all Endeca Information Discovery Studio users.

Private applications are available to application members only. An application member is an Endeca Information Discovery Studio user who has been granted access to an application by the application administrator.

Application Membership
Applications can have two different roles:

- **Application members** These people have view access to all the content in an application but cannot edit or configure the application or its membership. This is the default application role for a new application member.

- **Application administrators** These people have complete control over an application. They can edit the application content, configure the application, and control application membership, including whether a member is an administrator. The user who creates the application is automatically an application administrator. An application administrator can also assign other users as application administrators.

Major Component Summary
In the previous three sections, we covered the three core components of Endeca. You are certainly aware that Endeca has its own architecture and terminology unique from other products. If some of the terminology or concepts presented are somewhat unclear, don’t worry. All of these components will be covered at length in the next three chapters, with each component given a full chapter.
Endeca Licensing
Oracle licenses each of the core components separately. There are three plug-ins that are also licensed separately, listed here:

- **Oracle Endeca Web Acquisition Toolkit**  This is an add-on module for the Endeca Information Discovery Integrator.
- **Oracle Endeca Text Enrichment**  This includes text analysis capabilities for extracting people, places, organizations, quotes, and themes as well as summarization capabilities for automatically creating abstracts and topical summaries.
- **Oracle Endeca Text Enrichment with Sentiment Analysis**  This delivers all the Text Enrichment capabilities as well as advanced text analysis for extracting aggregate sentiment related to each extracted concept. Sentiment is extracted with a score indicating the positive and negative nature of a document, a phrase, or an entity. These scores are used to show varying ranges of positivity and negativity in search, guided navigation, and analytics.

Oracle sales consultants and Oracle license resellers can provide assistance and answer questions regarding Endeca licensing.

Summary
This chapter introduced Oracle Endeca Information Discovery and provided introductory information on each of the Endeca core components. The next three chapters will deal with these core components in detail:

- Chapter 2 covers installing, configuring, and managing Endeca Server.
- Chapter 3 covers getting Endeca Information Discovery Studio up and running and using it to design and run information discovery applications.
- Chapter 4 guides you through Endeca ETL installation and usage, Endeca ETL Server installation and configuration, Integrator Acquisition System installation and configuration, and other topics associated with the Endeca Information Discovery Integrator.