Oracle Endeca Information Discovery

Платформа для исследования данных

Ольга Горчинская
Oracle EE&CIS

Еще раз о Больших Данных

- **Сверхбольшие объемы структурированных и неструктурированных данных, с которыми трудно работать с помощью традиционных средств**
- Впервые термин появился 3 сентября 2008 года, Клиффорд Линч, редактор научного журнала Nature, в связи с проблемой накопления научных данных
- Источники больших данных– интернет-документы измерительные устройства, радиочастотная идентификация, социальные сети, устройства ауди и видеорегистрации
- 3V—определение Больших Данных
- **Новая парадигма анализа данных**

VVV
Velocity, Volume, Variety
(скорость, объем, многообразие)

Новые требования в анализу данных

Аналитика Больших Данных

Большое разнообразие данных		**Не все можно описать традиционными реляционным и многомерными моделями данных**
Большая скорость изменения		**Нельзя ориентироваться на заранее установленную модель данных**
Больше непредвиденных запросов		**Неизвестно заранее, какую информацию можно запрашивать**

Что такое Data Discovery?

Gartner

- Инструменты обеспечивают высокую степень удобства (usability), гибкости управления процессом моделирования и создания контента
- Высокая степень интерактивности и расширенные возможности визуализациии интерфейса, основанного на in-memory архтектурах и быстрым развертыванием
- Возможность пользователям исследовать данные самостоятельно без предварительного обучения
- Примеры: QlikTech QlikView, Tableau, Tibco Spotfire, Microsoft PowerPivot, Oracle Endeca Information Discovery (previously Endeca Latitude)

**Source: Gartner, June 17, 2011, “Emerging Technology Analysis: Visualization-Based Data Discovery Tools”

FORRESTER®

- Быстрая разработка и ускоренная реакция на быстро меняющиеся бизнес-требования
- Инкрементальное внедрение, быстрое прототипирование
- Данные, метаданные и приложения (отчеты, дэшборды) не разделяются («one and the same»)
- Легкость и простота создания новых приложений и изменения существующих

**Source: Forrester, “April 22, 2010, “Agile BI Out of the Box,” Boris Evelson

Когда возникает необходимость нового подхода к анализу данных

«Исследование» вместо «бизнес-анализа»

- Пользователи не знают какие вопросы они будут формулировать:
  - *Огромное число различных параметров их состав заранее неизвестен*
  - *Быстрые изменения информационных источников*

- ИТ-специалисты не знают, какая модель данных будет использоваться:
  - *Разнообразие данных и наличие большого числа внешних источников*
  - *Неструктурированные данные*
  - *Схемы исходных данных часто меняются*

Oracle Endeca Information Discovery – платформа для исследования структурированных, слабо-структурированных и неструктурированных данных с меняющейся структурой в условиях нечетких критериях поиска

Endeca

Историческая справка

- Основана в Кембридже, MA в 1999
- Более 600 клиентов
- 33% of the Fortune 100
- Анилиз неструктурированной информации, Большие данные
- От систем поиска для электронной коммерции к «BI beyond the data warehouse»
- *Entdecken* (немецкий) = to discover, открытие

Направления деятельности

eCommerce

- Интуитивный пользовательский интерфейс
- Быстрота предоставления информации

Public Sector

- Высокая производитльность, масштабируемость,
- Безопасность

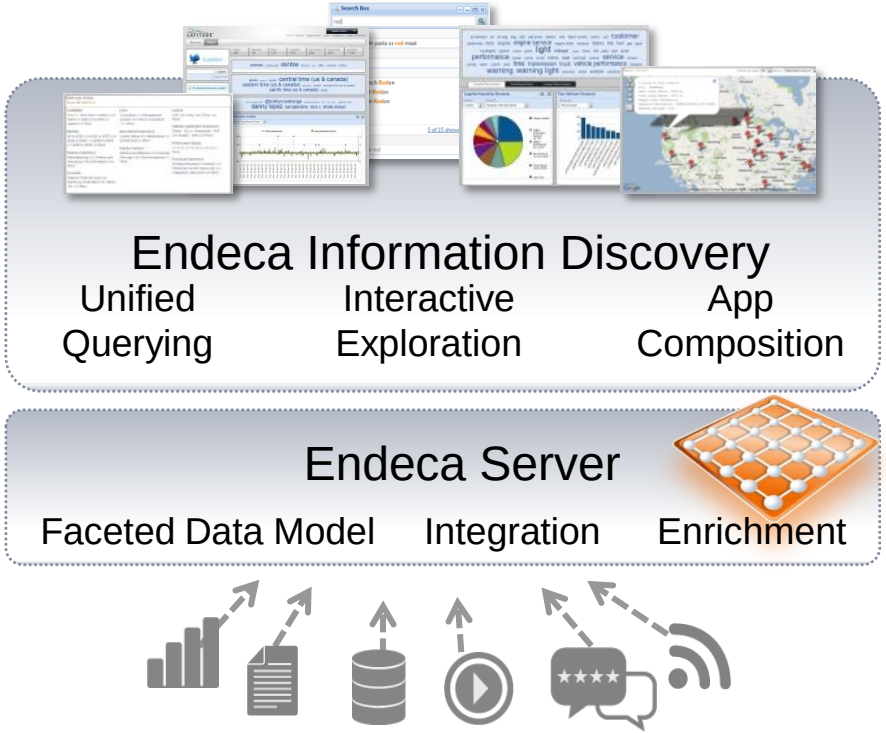
Enterprise

- Распределенные, сложные, изменяющиеся источники данных

Oracle Endeca Information Discovery

Платформа для исследования информации

- **Endeca Server**
  - Поисково-аналитическая база данных
- **Information Integration Suite**
  - Загрузка данных в Endeca Server
- **Endeca Studio**
  - Быстрая компонентная разработка приложений для исследования данных



Особенности интерфейса пользователей

- Удобство и простота использования
  - На основе 10-летнего опыта работы в области разработки поисковых систем для электронной коммерции
- Поиск + Фасетная навигация + Визуальный анализ
  - Поиск и выбор атрибутов в стиле вэб сайтов
- Интерактивные исследования
  - Без заранее определенного сценария

Разработка приложений

Источники данных

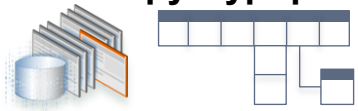
Загрузка в Endeca Server

Drag-and-drop инструменты создания приложения

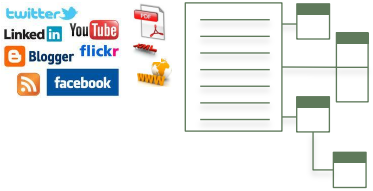
Интерактивный поиск, навигация и анализ

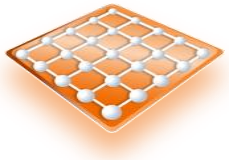
Структурированные

Слабоструктурированные



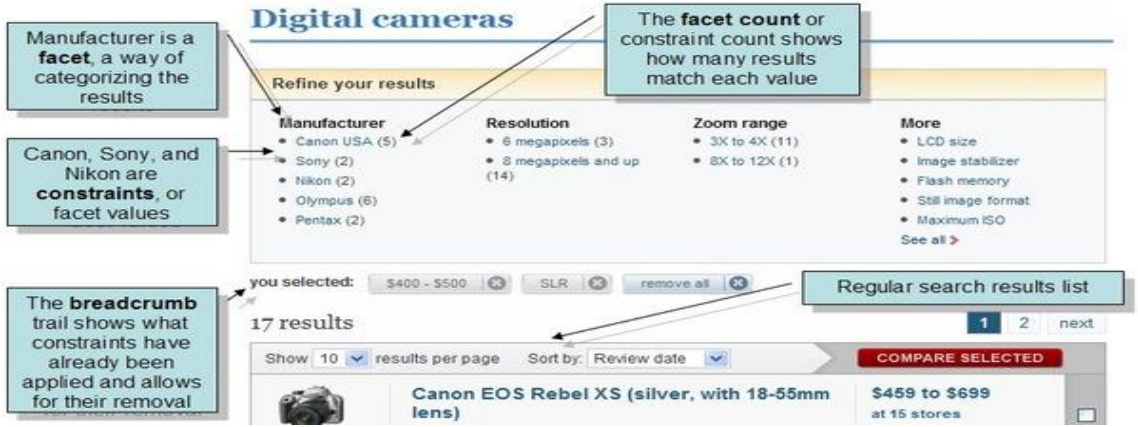
Неструктурированныеd





Фасетный поиск, фасетная навигация

- Технологии доступа к информации, организованной на основе системы фасетной классификации
- Поиск путем уточнений, навигация по независимым параметрам
- Модель информационного поиска – набор характеристик
- Теоретические основы фасетной классификации -- Ш. Р. Ранганатан («Классификация двоеточием», 1933)



Фасетная модель данных

Быстрая загрузка и изменение данные из любого источника

- Набор записей, каждая из которых имеет собственную «структуру»
  - Многозначные поля
  - Неструктурированные поля (тексты)
- Один из видов «Key Value» модели
- Модель состоит из
  - Записей
  - Атрибутов
- Каждая запись – это набор пар (атрибут, значение)
- Нет никакого разбиения на таблицы
- Нет понятия схемы данных

- TxnID = 12324
- ProductID = 506
- Category = Mountain Bike
- Amount = $499.99
- Suspension = Fox 32 F-Series
- FrameType = Aluminium
- Saddle = Bontrager SSR
- Mountain Accessories = Fork and shock sag meter
- Mountain Accessories = Water Bottle
- Review = A great bike for off road. Smooth ride over the bumps
- ReviewSentiment = Positive
- ReviewTerm = Great
- ReviewTerm = Off Road
- ReviewTerm = Smooth
- ReviewTerm = Bumps

- TxnID = 12325
- ProductID = 507
- Category = Road Bike
- Amount = $1399.49
- Weight = 20lb.
- FrameType = Composite
- Saddle = Bontrager Race
- Review = Disappointing for the price. The frame feels heavier than I expected.
- ReviewSentiment = Negative
- ReviewTerm = Disappointing
- ReviewTerm = Price
- ReviewTerm = Heavier

Загрузка данных : *Структурированные источники*

- Структурированные данные загружаются в фасетную модель естественным образом
  - Каждый кортеж становится записью
  - Каждый столбец становится атрибутом

Transaction			
TxnID	ProductID	Category	Amount
12324	506	Mountain Bike	499
12325	507	Road Bike	1399

Реляционная таблица

- **TxnID = 12324**
- **ProductID = 506**
- **Category = Mountain Bike**
- **Amount = $499.99**
- Suspension = Fox 32 F-Series
- FrameType = Aluminium
- Saddle = Bontrager SSR
- Mountain Accessories = Fork and shock sag meter
- Mountain Accessories = Water Bottle
- Review = A great bike for off road. Smooth ride over the bumps
- ReviewSentiment = Positive
- ReviewTerm = Great
- ReviewTerm = Off Road
- ReviewTerm = Smooth
- ReviewTerm = Bumps

- **TxnID = 12325**
- **ProductID = 507**
- **Category = Road Bike**
- **Amount = $1399.49**
- Weight = 20lb.
- FrameType = Composite
- Saddle = Bontrager Race
- Review = Disappointing for the price. The frame feels heavier than I expected.
- ReviewSentiment = Negative
- ReviewTerm = Disappointing
- ReviewTerm = Price
- ReviewTerm = Heavier

Загрузка данных : *Слабо-структурированные источники*

- Слабоструктурированные данные из различных источников -- промышленные системы, XML источники, внешние приложения. Загружаются в виде пар атрибут-значение
- Гетерогенные структуры записей

```
<ProductData ProductID="506">
  <Attribute Key="Suspension">Fox 32 F-Series</Attribute>
  <Attribute Key="FrameType">Aluminium</Attribute>
  <Attribute Key="Saddle">Bontrager SSR</Attribute>
  <Attribute Key="Mountain Accessories">
    Fork and shock sag meter</Attribute>
  <Attribute Key="Mountain Accessories">
    Water Bottle</Attribute>
</ProductData>

<ProductData ProductID="507">
  <Attribute Key="Weight">20lb.</Attribute>
  <Attribute Key="FrameType">Composite</Attribute>
  <Attribute Key="Saddle">Bontrager Race</Attribute>
</ProductData>
```

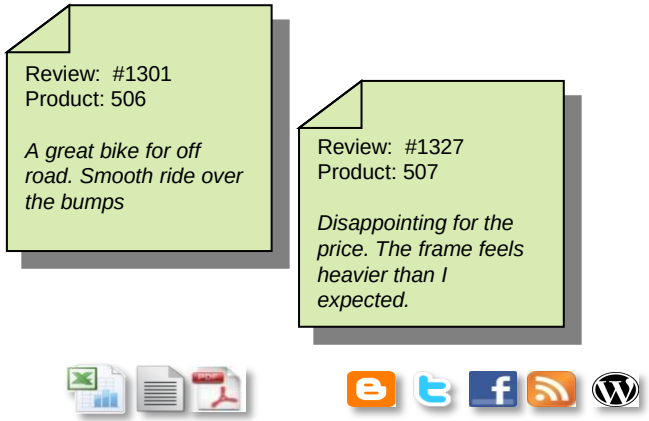
ETL

- TxnID = 12324
- ProductID = 506
- Category = Mountain Bike
- Amount = $499.99
- **Suspension = Fox 32 F-Series**
- **FrameType = Aluminium**
- **Saddle = Bontrager SSR**
- **Mountain Accessories = Fork and shock sag meter**
- **Mountain Accessories = Water Bottle**
- Review = A great bike for off road. Smooth ride over the bumps
- ReviewSentiment = Positive
- ReviewTerm = Great
- ReviewTerm = Off Road
- ReviewTerm = Smooth
- ReviewTerm = Bumps

- TxnID = 12325
- ProductID = 507
- Category = Road Bike
- Amount = $1399.49
- **Weight = 20lb.**
- **FrameType = Composite**
- **Saddle = Bontrager Race**
- Review = Disappointing for the price. The frame feels heavier than I expected.
- ReviewSentiment = Negative
- ReviewTerm = Disappointing
- ReviewTerm = Price
- ReviewTerm = Heavier

Загрузка данных : *Неструктурированные источники*

- Неструктурированные данные «присоединяются» к остальным данным на основе значения ключа
- В определеных случаях неструктурированные данные могут храниться в виде независимых записей
- Текстовые документы, RSS, Twitter, Facebook, форумы и др..



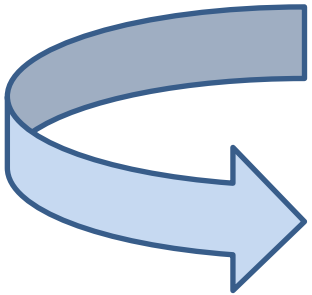
- TxnID = 12324
- ProductID = 506
- Category = Mountain Bike
- Amount = $499.99
- Suspension = Fox 32 F-Series
- FrameType = Aluminium
- Saddle = Bontrager SSR
- Mountain Accessories = Fork and shock sag meter
- Mountain Accessories = Water Bottle
- **Review = A great bike for off road. Smooth ride over the bumps**
- ReviewSentiment = Positive
- ReviewTerm = Great
- ReviewTerm = Off Road
- ReviewTerm = Smooth
- ReviewTerm = Bumps

- TxnID = 12325
- ProductID = 507
- Category = Road Bike
- Amount = $1399.49
- Weight = 20lb.
- FrameType = Composite
- Saddle = Bontrager Race
- Review = Disappointing for the price. The frame feels heavier than I expected.
- ReviewSentiment = Negative
- ReviewTerm = Disappointing
- ReviewTerm = Price
- ReviewTerm = Heavier

Обогащение (Enrichment)

Добавление в модель дополнительной информации

- Использование дополнительных пакетов работы с текстами, пространственными данными и др.
- Примеры:
  - Извлечение объектов по именам
  - Анализ эмоциональной окраски текстов
  - Совмещение с геопространственными
  - Term extraction
  - Geospatial matching



- TxnID = 12324
- ProductID = 506
- Category = Mountain Bike
- Amount = $499.99
- Suspension = Fox 32 F-Series
- FrameType = Aluminium
- Saddle = Bontrager SSR
- Mountain Accessories = Fork and shock sag meter
- Mountain Accessories = Water Bottle
- Review = A great bike for off road. Smooth ride over the bumps
- **ReviewSentiment = Positive**
- **ReviewTerm = Great**
- **ReviewTerm = Off Road**
- **ReviewTerm = Smooth**
- **ReviewTerm = Bumps**

- TxnID = 12325
- ProductID = 507
- Category = Road Bike
- Amount = $1399.49
- Weight = 20lb.
- FrameType = Composite
- Saddle = Bontrager Race
- Review = Disappointing for the price. The frame feels heavier than I expected.
- **ReviewSentiment = Negative**
- **ReviewTerm = Disappointing**
- **ReviewTerm = Price**
- **ReviewTerm = Heavier**

efox

Staffing - endeca.com

ENDECA LATITUDE®

Welcome! Demo User | Toggle Controls | Logout | Demo Home | Create new bookmark

Welcome | Staffing | Utilization

Refine your results:

Availability
More than 4 weeks (1500) Within 2 weeks (816) Now (739) More

Bill Rate
$0 to $100 (627) $100 to $200 (789) $200 to $300 (401) $300 to $400 (423) $400 to $500 (195) More

Industry Experience
Manufacturing (2038) Finance and Insurance (1567) Information (1338) More

Accounts
Cisco Systems, Inc. (151) Intel Corporation (127) Xerox Corporation (111) More

Level
Consultant (716) Senior Consultant (517) Management Analyst (502) More

International Experience
United States (2376) Netherlands (188) Switzerland (117) More

Practice Vertical
Electronics Software (754) Financial Services (366) Communications (263) More

Country
USA (1466) China (358) Germany (87) More

Software Application Experience
Siebel - BI (862) Peoplesoft - CRM (818) Siebel - CRM (786) More

Performance Rating
1 (461) 2 (560) 3 (853) 4 (363) 5 (198) More

Functional Experience
Enterprise Human Resources (710) Integration IT Strategy (705) Infrastructure Services Management (514) More

Search

Search Within

Geography Search:

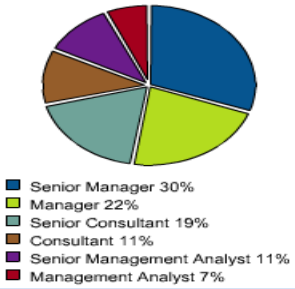
miles from

Submit

Averages

Metrics: Billing Rate

Group By: Level



Consultants	Average Performance Rating	Avg Billing Rate
2,435	2.70	$199.40

Actions

Choose a column set: Position | Location | Resume

Employee Name	Billing Rate	Level	Performance	Practice Area	Practice Vertical
Chesick, Harry	437	Manager	1	North America	Electronics Software
Fleitas, Cleo	361	Senior Consultant	1	North America	Electronics Software
Gantzler, Harvey	111	Management Analyst	5	North America	Electronics Software
Dibenedict, Ruthe	110	Management Analyst	5	North America	Financial Services
Cote, Lissette	373	Senior Consultant	1	North America	Aerospace & Defense
Mabrey, Ilana	280	Senior Consultant	5	EMEA	Content
Zwahlen, Lee	265	Consultant	3	North America	Electronics Software
Adelmann, Gus	350	Senior Consultant	2	North America	Financial Services
Baldiviez, Becki	148	Senior Management Ar	4	North America	Electronics Software
Suryanarayana, Mia	134	Management Analyst	2	North America	Automotive

Более 600 клиентов

Полная аналитическая платформа Oracle