Oracle для анализа и исследования Больших Данных

© Ольга Горчинская, главный консультант по аналитическим технологиям
Корпорация Oracle
Москва
olga.gorchinskaya@oracle.com

Аннотация

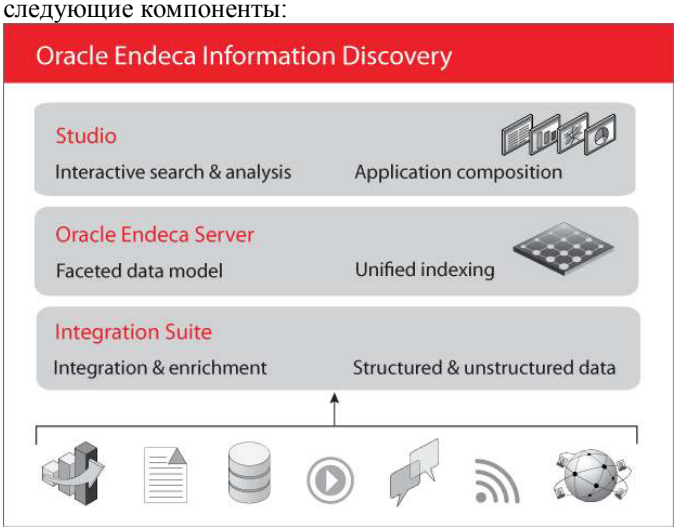
Доклад посвящается инструментальной среде Oracle для совместного анализа и исследования структурированной, слабоструктурированной и неструктурированной информации.

Стремительно развивающееся направление Больших Данных меняет концепцию аналитики – огромные объемы разнообразной, чаще всего неструктурированной информации требуют новых подходов и технологий для эффективного ее использования.

Традиционный подход к бизнес-анализу данных, основанный на идеях хранилища данных как «единого источника истины», предполагает работу в рамках четко фиксированной модели данных, гарантирует качество и непротиворечивость информации. В этом случае заранее известно, какие структуры данных содержатся в хранилище, и анализ в основном сводится к выполнению вычислительных процедур по агрегированию, детализации, фильтрации данных и визуализации результатов. Специфика Больших Данных связана не только с огромными объемами, но и с огромным разнообразием и изменчивостью, и это не позволяет ориентироваться на какую-либо заранее разработанную модель данных. В этом случае анализ данных должен сопровождаться многочисленными поисковыми операциями, в результате которых уточняется запрос на необходимую информацию. Такой подход к аналитике, при котором собственно аналитические операции интенсивно интегрируются с поисковыми, и нет модели данных в традиционном понимании, лежит в основе нового направления – Information Discovery или Исследование данных. Компания Oracle предлагает технологическую платформу для исследования данных – Endeca Information Discovery.

Продукт Endeca Information Discovery представляет собой платформу для исследования структурированных, слабоструктурированных и неструктурированных данных из различных меняющихся источников в условиях нечетких критериев поиска. Средства семантического анализа позволяют выявлять в неструктурированном тексте понятия, связи, факты и другие релевантные данные. Система может быть дополнена как собственными дополнительными модулями, расширяющими функционал решения, так и модулями сторонних производителей, обеспечивающих, например, улучшенную морфологическую поддержку и/или поддержку различных дополнительных иностранных языков. В результате система предоставляет средства расширенного агрегирования и поиска информации, в сочетании с мощной аналитикой.

В состав Endeca Information Discovery входят следующие компоненты:



- **Oracle Endeca Server.** Основой платформы является гибридная поисково-аналитическая база данных. В этой базе данных собирается информация из различных структурированных и неструктурированных источников и хранится в виде универсальной фасетной модели, которая обеспечивает

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

максимальную гибкость при работе с изменяющимися источниками, не требует предварительной разработки семантической модели и поддерживает эффективные средства поиска информации. Для обеспечения высокой производительности на аналитических запросах, в отличие от традиционных подходов реляционных баз данных, используется колоночное хранение и высокоэффективная при таком способе хранения колоночная компрессия. MDEX хранит каждую колонку информации на диске и в оперативной памяти с использованием двух индексов -- по значению и по ключу. Кроме того, каждая колонка имеет B-Tree индекс, который кэшируются в оперативной памяти. Такой подход обеспечивает высокую производительность, необходимую при работе в условиях изменяющихся нечетких критериев поиска в сочетании с аналитическими вычислениями.

- **Studio.** Интерактивная, компонентно-ориентированная среда для быстрой итеративной разработки и разворачивания приложений для исследования данных. В рамках таких приложений пользователи получают удобные средства поиска и исследования информации, ориентированные на решение конкретных прикладных задач.

- **Integration Suite.** Инструментальный комплекс для загрузки структурированных, слабоструктурированных и неструктурированных данных в базу данных Endeca Server. Инструмент содержит (1) Content Acquisition System для сбора информации из файловых систем, систем управления контентом и веб-сайтов, (2) Integrator, содержащий готовые ETL инструменты для интеграции и обогащения данных и (3) открытый Web Services API для задач прямой интеграции с другими средствами такими, как Oracle Data Integrator, Informatica PowerCenter и Hadoop.

- **Oracle Endeca Content Management System Connectors.** Этот add-on модуль поддерживает интеграцию данных из различных систем управления контентом. Среди поддерживаемых хранилищ документов -- EMC Documentum, EMC Documentum eRoom, FileNet P8, FileNet Document & Image Services, Interwoven TeamSite, LotusNotes/Domino, Microsoft SharePoint, OpenText LiveLink.

- **Oracle Endeca Text Enrichment.** Модуль поддерживает возможности полнотекстового поиска и анализа, включая выявление сущностей – физических лиц, организаций, адресной информации, автоматическое формирование аннотаций и др.

- **Oracle Endeca Text Enrichment with Sentiment Analysis.** Этот add-on module включает средства обогащения текстовых данных, а также предоставляет методы углубленого анализа текста для извлечения эмоциональной окраски или оттенков. Оттенки представляются в виде числовых значений и могут относиться как ко всему тексту в целом, так и к конкретным сущностям. Впоследствии эти значения используются в рамках фасетного поиска, объединяясь с другими данными.

В докладе обсуждаются возможности платформы Endeca Information Discovery, рассказывается об особенностях разработки на ее основе прикладных систем исследования данных, а также обсуждаются примеры использования этого продукта для решения практических задач.

Oracle Tools for Big data Analysis

© Olga Gorchinskaya, Master Principal Sales Consultant, Business Analytics
Oracle Corporation