



Работа с большими объемами данных и оптимизация моделей

Андрей Медведев

Архитектор решений

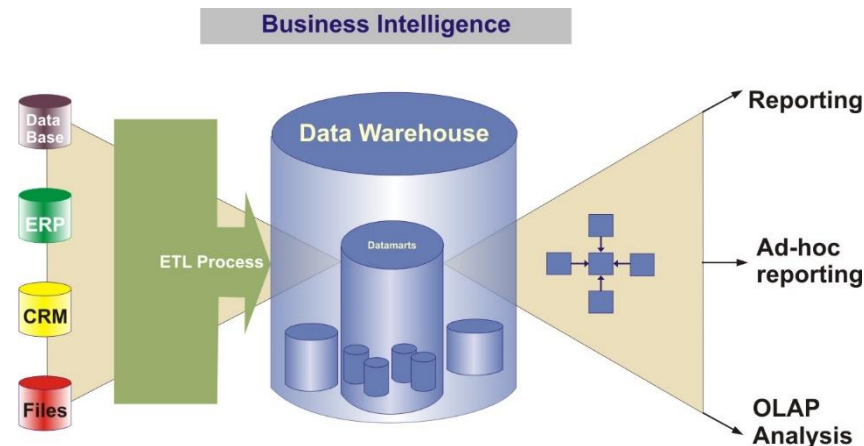
medvedev@iterbi.ru

www.iterbi.ru

02.06.2016

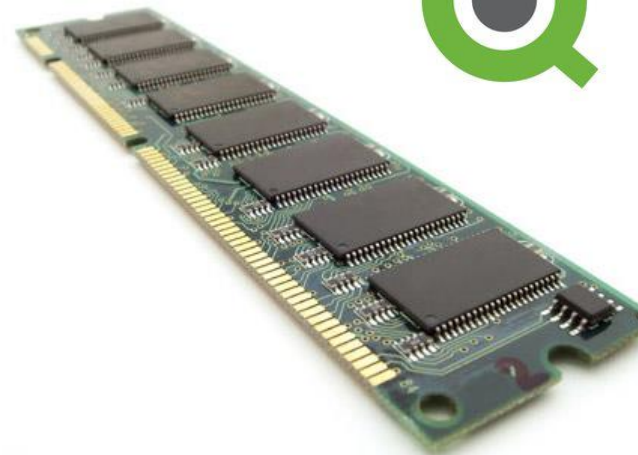
Обработка операционных баз данных и многомерных хранилищ

- Транзакционные тяжело нагруженные базы данных
- Реляционные и многомерные хранилища
- Прямые обращения к API систем
- Работа с коннекторами
- Обработка плоских файлов (CSV, Excel, XML и т.п.)



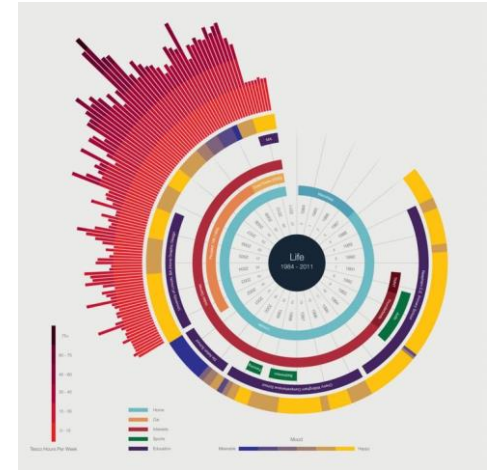
Построение аналитических моделей в памяти

- Классические BI-решения
- Гибридный подход (Exalytics, HANA, Teradata и т.п.)
- In-memory системы
 - Принципиально другой подход
 - Нереляционные базы данных
 - Компрессия
 - Скорость доступа
 - Индексация и поиск
 - Средства визуализации



Типичные проблемы визуализации и расчеты «на лету»

- Попытка сформировать перегруженный элементами интерфейс
- Избыточность данных (миллионы строк в табличных и графических элементах)
- Динамический расчет показателей
- Сложные вычисления
- Условное форматирование
- Неоптимальные выражения, приводящие к полному сканированию таблиц
- Расчеты по большим объемам первичных данных



Инструментарий, комбинирующий ETL-инструменты и средства визуализации



- Задачи
 - SQL запросы
 - Работа с хранимыми процедурами
 - Работа с плоскими файлами как с таблицами базы
 - Скрипты преобразования, обогащения и очистки данных
 - Работа с функциями
 - Формирование новых измерений и значений
 - Подготовка реляционных связей между сущностями



- Обработка данных, формирование локальной прогнозной модели, реализация математических алгоритмов
- Передача данных и использование внешних инструментов
- Работа с СУБД и использование функций обработки больших массивов данных
- Многопоточная обработка
- Сравнительная аналитика
- What-If анализ и параметризация отчетов





Андрей Медведев

Архитектор решений

Medvedev@iterbi.ru

www.iterbi.ru