Анализ, интеграция и качество данных: вчера, сегодня, завтра

Ежегодная BI-конференция TEPH

Валерий Артемьев (Банк России) © 2019

Апрель 2019 года

Москва

Содержание

1. Анализ данных
2. Интеграция данных
3. Качество данных (в широком смысле)

Корпоративное управление данными (Data Management) – ведение и регулирование информационных ресурсов как корпоративных активов с позиций бизнеса

Вчера: Бизнес-аналитика
Business Intelligence

- Традиционная бизнес-аналитика
  - Многомерный анализ, запросы, отчёты, интеграция с Office
  - Ведущая роль у ИТ в подготовке данных и семантического слоя
  - Самообслуживание бизнес-пользователя
  - Описательный анализ
  - Ограниченная интерактивность
  - Интеграция данных: хранилище и витрины данных
- Инфографика
  - Деловая графика, географические карты, dashboard, scorecard, тренды, sparkline
  - Для пользователей и разработчиков
- Исследование данных Data Discovery
  - Подготовка и визуальное исследование данных, storytelling
  - Самообслуживание бизнес-пользователя
  - Описательный и диагностический анализ
  - Свободная интерактивность
  - Анализ структурированных данных без модели

Web BI
Desktop BI
Mobile BI
In-memory BI
BI портал
BI приложения
Аналитические сервисы
MOLAP?
ROLAP
Big Data
NewSQL
BI appliance
Облака

Сегодня: Продвинутая аналитика Advanced Analytics

- Продвинутая аналитика
  - Ведущая роль у expert data scientist.
  - Многообразие источников и видов данных
  - Статистический, описательный, предсказательный
  - Риск- и предписывающий анализ
  - Машинное обучение, глубинное обучение нейронных сетей
  - Data Mining
  - Анализ временных рядов
  - Анализ поведения и рекомендации
  - Аналитическое моделирование
  - Продвинутая визуализация
- Анализ текста Text Mining
  - Нечёткий поиск/ сравнение текста
  - Выделение объектов и свойств из текста
  - Определение тональности текста
- Искусственный интеллект
  - Распознавание текста и голоса
  - Генерация текста и голоса
  - Распознавание образов

Open Source
Экосистемы Big Data
Языки и библиотеки Python, R, Java, Scale
Платформы Data Science
Распределённая обработка Map/Reduce, in-memory, потоки
NoSQL/NewSQL как SQL
Лаборатория данных Облака

Завтра: Дополненная аналитика Augmented Analytics

- Самообслуживание на распутье
  - Продвинутые сложные средства анализа для expert data scientist
  - Новые более простые средства аналитики для бизнес–пользователя
- Демократизация данных и аналитики
  - Устранение барьеров, упрощение доступа, но необходимая защита
  - Упрощение структуры и подготовки данных
  - Автоматизация/облегчение исследования, анализа и машинного обучения
- Дополненная аналитика
  - Ведущая роль у бизнеса (citizen data scientist)
  - Разнообразные виды данных
  - Всеобъемлющий автоописательный, диагностический, предсказательный, предписывающий анализ
  - Автоматизация машинного обучения
  - Автовизуализация релевантных шаблонов
  - Автообнаружение и дополненная подготовка данных
  - Советы в контексте пользователя
  - Запросы и ответы на естественном языке (текст и голос).

Платформы демократизации
Big Data
Облака

Вчера: Хранилище данных на основе Data Vault

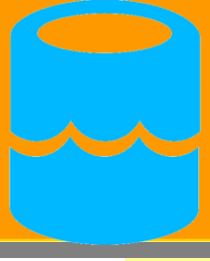
- Многоуровневое хранилище данных
  - Накопительная область или оперативны склад данных
  - Ядро хранилища данных
  - Витрины данных и аналитические наборы данных

- Единая версии правды/ факта
  - Структурированные данные
  - Преобразование, контроль, очистка: сосредоточены при загрузке в хранилище или
  - Распределены при загрузке в хранилище и витрины данных

- Адаптивность за счёт схемы Data Vault
  - Таблицы концентраторы HUB (бизнес-сущности, постоянная часть)
  - Таблицы связей LNK (могут добавляться без изменения других таблиц)
  - Таблицы-спутники STL (могут изменяться атрибуты, добавляться новые таблицы)

ETL
Data Quality
MDM
SQL
DW appliance
Облака

Сегодня: Озеро данных
Data Lake



- Архитектура озера данных
  - Только для накопительной области или оперативного склада данных
  - «Три в одном»: 1) Staging, 2) ядро хранилища, 3) витрины и аналитические наборы данных
- Единая версия факта
  - Плюс полуструктурированные и неструктурированные данные, события/потоковые данные
  - Преобразование, контроль и очистка распределены при загрузке в хранилище, витрины данных и аналитические наборы
- Адаптивность за счёт гибких схем данных
  - Поколоночные расширяемые БД плюс Data Vault или 6NF
  - Документарные и графовые базы данных

Open Source
ETL/ ELT
Data Quality
MDM
NoSQL/ NewSQL
Big Data
Облака

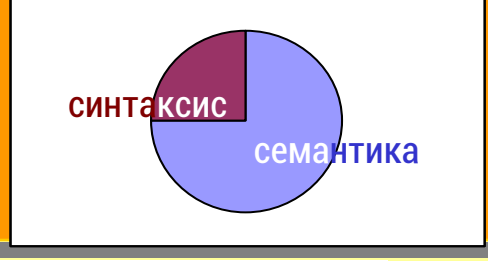
Завтра: Интеграция данных без их перемещения

- Проблемы-катализаторы
  - Перемещение и дублирование данных от 3 до 6 раз
  - Высокая трудоёмкость интеграции с перемещением
  - Потребность в операционной аналитике
- Логическое хранилище данных (Logical Data Warehouse, LDW)
  - Федеративное виртуальное связывание существующих ХД, витрин, БД, Big Data
  - Разные виды и структуры данных
  - Преобразования на лету
- Гибридная обработка данных (Hybrid Transact Analytical Processing)
  - Не разделены операционные и аналитические данные
  - Единая версия факта
  - События, потоки и транзакции, аналитическая обработка

Data Quality
Платформы интеграции данных
SQL/ NoSQL/ NewSQL
Big Data

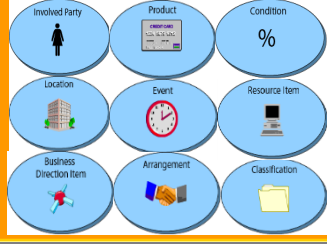
Data Quality
Акселераторы
In-memory DB
In-memory Computing

Вчера: Качество данных
Data Quality



- Характеристики и метрики качества данных
  - В основном синтаксические проверки
  - Полнота, допустимость, согласованность, ссылочная целостность, уникальность, достоверность, своевременность и другие
  - Распределение проверок по характеристикам
  - Бизнес-правила добавляют некоторую семантику
  - Метрики: доля проверок без ошибок/ с ошибками, доля брака, интегральные оценки
- Мониторинг качества данных
  - Уровни качества данных в соглашениях о качестве данных DQA
  - Управление инцидентами/проблемами
- Улучшение качества данных информирование
  - Исправления или игнорирование дефектов
  - Заглушки данных по умолчанию
  - Информирование об ошибках и исправлениях

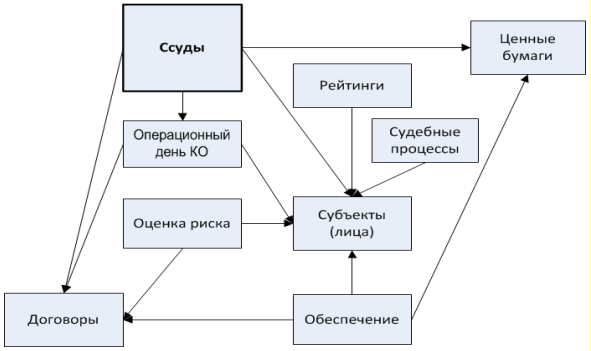
Сегодня: Управление мастер-данными MDM



- Классификационные схемы
  - Определяют аналитическую ценность данных
  - Методы классификации и кодирования
- Ключевые бизнес-сущности
  - Вовлечённые стороны
  - Местонахождение и контакты
  - Продукты/ услуги
  - Контракты
  - События и т.п.
- Качество мастер-данных
  - Формирование «золотых» записей (история изменений, кластеризация, оценка достоверности, приоритеты заполнения)
  - Стандартизация (парсинг, правила/ шаблоны/ словари, matching, определения вида субъекта)

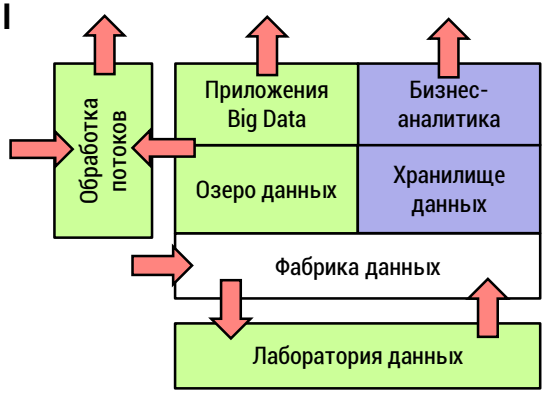
Завтра: Архитектура данных
Data Architecture

- **Корпоративная модель данных** – статическая часть, семантическая основа качества данных
  - Верхнеуровневая модель данных
  - Модели предметных областей
  - Бизнес-глоссарий
  - Прикладные модели данных
  - ER-диаграммы, RDF, таксономии, онтологии



- **Потоки данных** – динамическая часть
  - ЖЦ данных, связь потоков с бизнес-процессами и ролями
  - Диаграммы и матрицы потоков
  - Источники, потребители, внутренние инфоресурсы
  - Учёт активов, каталог данных

- **Архитектура интеграции данных**
  - New: обработка потоков, озеро данных, приложения Big Data, лаборатория данных



Лозунги к Первому Мая

BI ЖИЛ, BI – ЖИВ,
BI БУДЕТ ЖИТЬ!
АНАЛИТИКИ, ПРИМЕНЯЙТЕ
DATA DISCOVERY!
ДЕМОКРАТИЗАЦИЮ АНАЛИТИКИ –
БИЗНЕСУ!

ВЫ ОСВОИЛИ ПОДХОД
DATA VAULT?!
НЕ ДОПУСТИМ ПРЕВРАЩЕНИЯ
ОЗЁР В БОЛОТА!
БУДЬ ГОТОВ К ИНТЕГРАЦИИ ДАННЫХ
БЕЗ ПЕРЕМЕЩЕНИЯ!

КАЧЕСТВО ИЗМЕРЯЙ,
НАБЛЮДАЙ,
ИСПРАВЛЯЙ, ИНФОРМИРУЙ!
БОЛЬШЕ МАСТЕР-ДАННЫХ
ХОРОШИХ И РАЗНЫХ!
МОДЕЛИ ДАННЫХ И ГЛОССАРИЙ –
ЗАЛОГ КАЧЕСТВА ДАННЫХ!

ЗАДАЧИ БИЗНЕСА – НЕ ТЕХНОЛОГИИ
РАДИ ТЕХНОЛОГИЙ!
УПРАВЛЯЕТЕ ДАННЫМИ
В МАСШТАБЕ КОРПОРАЦИИ!

Спасибо за внимание!

Валерий Иванович Артемьев

Департамент статистики
и управления данными
Банк России

+7(495) 753-96-25
avi@cbr.ru