Проектирование хранилищ данных

ФИО преподавателя: Смирнов Михаил Вячеславович

e-mail: smirnovmgupi@gmail.com

Лекция 4

Аспекты реализации проекта хранилища данных

Вопросы лекции

- Типовые фазы разработки хранилища данных.
- Вопросы организации проектной команды для проекта хранилища данных.
- Ключевые точки успеха реализации проекта хранилища данных.
- Тупиковые сценарии проекта хранилища данных и краткосрочные показатели успеха.
- Практическая модель успешного проекта хранилища данных
- Проблема формирования требований к хранилищу
- Форма бизнес-измерения, «информационный пакет» объекта (+примеры)

Типовые фазы проекта хранилища данных

Ключевой принцип организации команды проекта хранилища данных

Полагается, что команда, обслуживавшая ранее проекты баз данных OLTP *не имеет представления о хранилищах OLAP типа*.

В этом случае будет разумно заранее определять «белые» пятна проекта и стартовые ожидаемые требования по навыкам внутри команды, после чего – ассоциировать их по ролям с имеющимися специалистами (определение требований к данным и типов запросов, выбор модели хранения и инструментария, контроль правильности данных и их очистка и т.д.).

Стандартный набор ролей для проекта хранилища данных

- Executive sponsor
  - Поддержка, разрешение споров
- Project manager
  - Распоряжения, контроль выполнения
- User liaison manager
  - Координация с пользователями
- Lead architect
  - Архитектурный проект
- Infrastructure specialist
  - Проектирование инфраструктуры
- Business analyst
  - Определение требований
- Data modeler
  - Моделирование измерений, где это необходимо
- Data warehouse administrator
  - Функции DBA
- Data transformation specialist
  - Проведение ETL процедур
- Quality assurance analyst
  - Процедуры контроля качества данных
- Testing coordinator
  - Тестирование программ, систем, инструментов для хранилища
- End-user applications specialist
  - Проверка и подтверждение data usability
- Development programmer
  - Программы и скрипты, используемые в процессе проектирования
- Lead trainer
  - Координация программ обучения команды и пользователей

Типовые навыки для ролей проекта хранилища данных

Executive sponsor. Знание предметной области, энтузиазм, психология.

Project manager. Тайм-менеджмент, знание навыков группы, опыт управления проектами (знание методик).

User liaison manager. Уважение в сообществе пользователей, организационные навыки, точка зрения пользователя.

Lead architect. Способности масштабирования, аналитические навыки, разумный «евангелизм».

Infrastructure specialist. Знание «железа», операционных систем, вычислительных платформ, опыт в качестве пользователя.

Business analyst. Навыки взаимодействия с пользователями, опыт аналитики в предметной области проекта.

Data modeler. Опыт аналитика данных, опыт работы с OLTP и OLAP базами данных. Опыт работы с распределенными базами данных.

Data warehouse administrator. Опыт администрирования баз данных, знания физической модели данных хранилищ.

Data transformation specialist. Знание типовых структур данных, аналитический опыт, понимание принципов работы источников информации.

Quality assurance analyst. Навыки повышения качества данных, аналитический опыт, понимание принципов работы источников информации.

Testing coordinator. Знание стандартов и методов тестирования, ПО для тестирования, понимание работы «входа и выхода» хранилища, опыт программирования.

End-user applications specialist. Глубокие знания приложений, которые будут использоваться на выходе хранилища.

Development programmer. Опыт программирования как приложений, так и баз данных.

Lead trainer. Навыки обучения. Опыт в IT-образовании, организаторские навыки, психология.

Ключевые точки успеха проекта хранилища данных

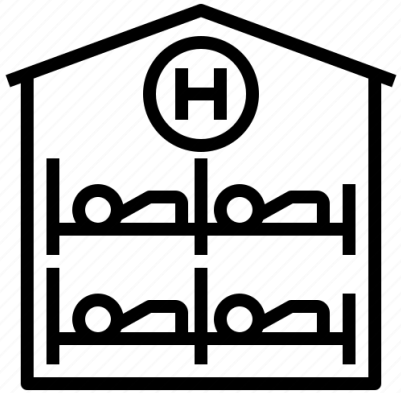
- Финансирование. Проект как правило долгий и кропотливый, без моментальных результатов. Финансовая поддержка должна быть безоговорочной.
- Менеджмент. Должен быть ориентирован не столько на технологии, сколько на конечного пользователя и бизнес.
- Новая парадигма. Переход на хранилище данных для компании, как правило, - радикальная смена технологии работы с данными.
- Роли команды проекта. Роли проекта должны определяться требованиями проекта, а не методологией.
- Качество данных. Критический показатель качества хранилища данных – качество данных внутри.
- Пользовательские требования. Все задачи проекта разработки должны формироваться из требований пользователя (очевидно для Agile).

Ключевые точки успеха проекта хранилища данных

- Потенциал роста. После запуска хранилища количество пользователей и запросов будет постоянно расти. В некоторые моменты экспоненциально.
- Реалистичные ожидания.
- Моделирование измерений. Требует предварительной проработки и создания моделей и эскизов.
- Внешние данные. Обязательно необходимо учитывать для успешного функционирования хранилища.
- Обучение. Конечные пользовательские приложения могут быть совершенно незнакомы даже опытным пользователям. Как и языки взаимодействия с новым хранилищем.

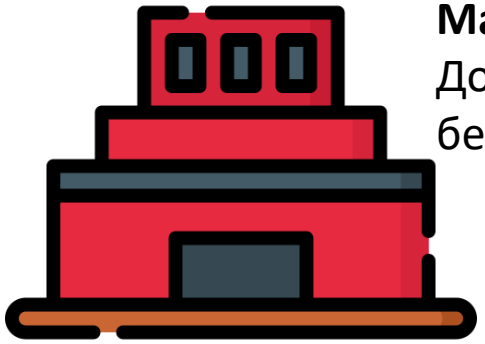
Тупиковые сценарии проектов хранилища

Подвал данных. Плохое качество данных, плохой доступ



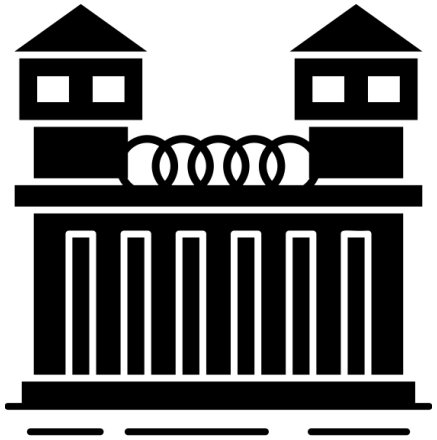
Общежитие данных. Построено на базе Legacy без сбора требований.

Коттедж данных. Отдельные, фрагментированные витрины данных.



Мавзолей данных. Дорогое, но бессмысленное

Шалаш данных. Забагованная свалка данных.



Тюрьма данных. Хранилище с затрудненным и неудобным доступом к данным.

Практическая модель успешного проекта хранилища данных

- Согласованная бизнес-оценка полученного стека технологий.
- Высокая степень вовлеченности конечного пользователя в результаты проекта.
- Наибольшее время и внимание в ходе реализации проекта уделено … (вспомним, предположим, победим).
- Сперва – архитектура, затем – технология, наконец – инструменты.
- Используются самые простые в использовании конечные инструменты.

Практическая модель успешного проекта хранилища данных

- Разработан план эксплуатации и оптимизации результатов проекта.
- Проект управлялся менеджером, ориентированным в первую очередь на пользователей (очевидно для Agile).
- Фокус на проектировании запросов, а не транзакций.
- Правильная подборка источников данных. Загружаются только нужные данные.

Элементы требований при реализации проекта базы данных

- Бизнес-процессы (дополненные DFD диаграммами с описаниями потоков).
- Описания конечных приложений, которые будут «подцеплены» к базе данных.
- Описание отчетов и документации, установленной в организации формы, как результата работы приложения базы данных.

Ничего из этого не работает при сборе требований к хранилищу данных (комплексность)

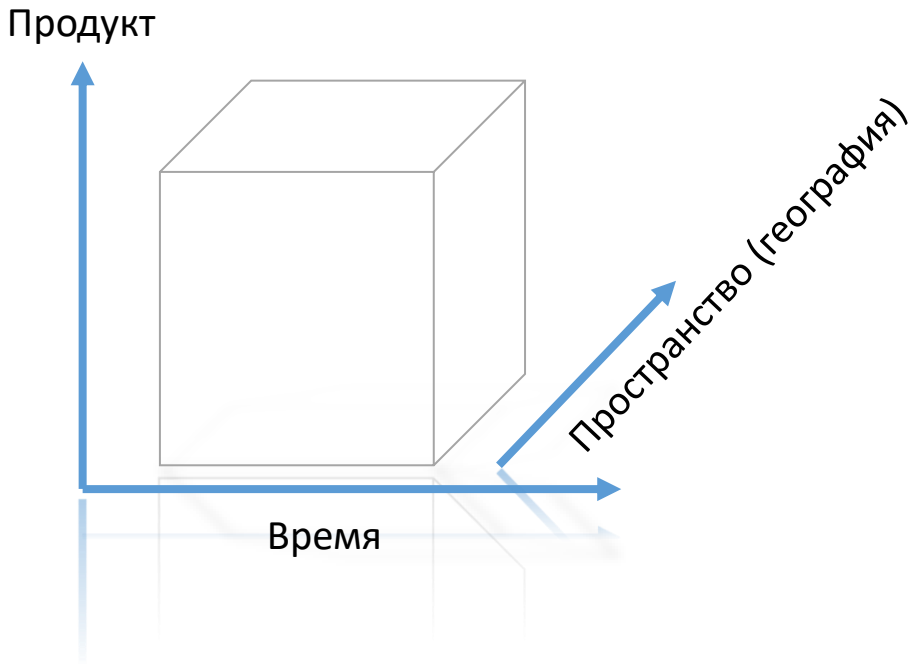
Проблема требований к хранилищу данных

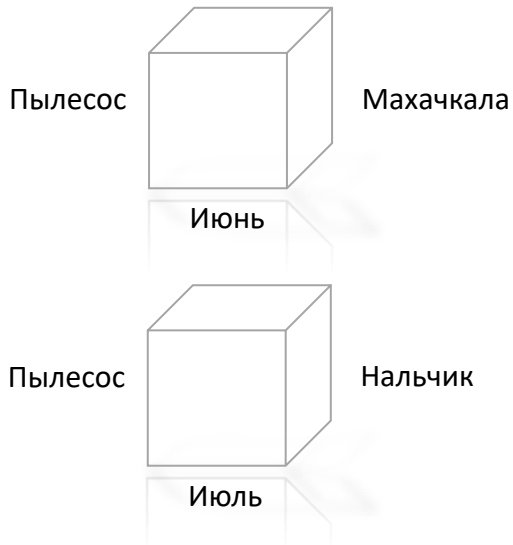
Пользователи хранилища «смотрят» на процессы организации как минимум на уровне «тактического развития» и оперативная информация, как основа традиционных требований, до них не доходит. Это называется «мышление в бизнес-измерении» (принятие решения на основании аналитических данных).

Вице-президент по маркетингу. *Данные по новому продукту: помесячно, в европейском подразделении компании, по группам покупателей, по точкам продаж, по сравнению с прошлой версией, по сравнению с планом…*
Менеджер по маркетингу. *Статистика продаж: по продукту, агрегатный по категориям продукта, за день, за неделю, за месяц, по локации продаж, по каналам распределения…*
Финансовый аудитор. *Затраты: соответствие бюджету, за месяц, за квартал, по географии, агрегатные для всей компании…*

Трехмерное бизнес-измерение (кубы)

Самое простой тип бизнес-измерения. Данные помещены в куб.

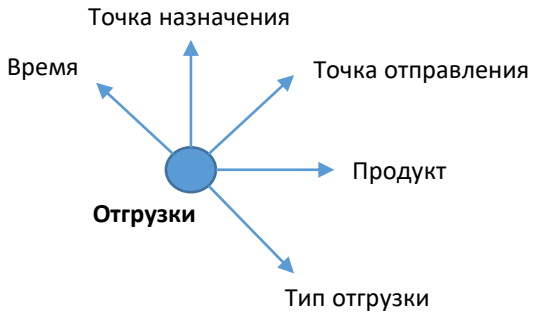




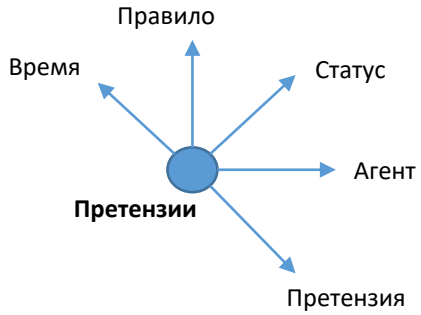
Гиперкуб

Чаще на практике используются более сложные, многовекторные измерения. Они называются гиперкубы.

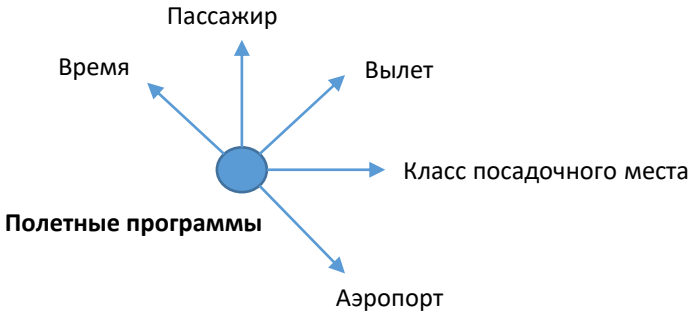
Завод



Страховая компания



Авиакомпания



Неполностью детерминированные требования. Informational package.

Временные периоды	**Локации**	**Продукты**	**Возрастные группы**	**…**	**…**
Год	Страна	Класс	Группа 1		
…	…	…	…		
…	…	…	…		
…	…	…	…		
Измеряемые факты: прогноз продаж, оперативная информация о продажах					

Предметная область информации: Анализ продаж

Informational package. Продажи автопроизводителя.

Время	Продукт	Тип оплаты	Демография покупателя	Дилер	…
Год	Название модели	Тип транзакции	Возраст	Название	
Квартал	Год модели	Длительность	Пол	Город	
Месяц	Дизайн	Процент	Уровень дохода	Страна	
Дата	Линейка продукта	Агент	Семейный статус	Флаг «одного бренда»	
День недели	Категория продукта		Количество автомобилей	Дата открытия	
День месяца	Цвет снаружи		Собственность		
Сезон	Цвет внутри				
Флаг праздников					
Измеряемые факты: текущая цена, полная цена, надбавка дилера, кредиты дилера, первый платеж, текущие платежи, финансы…					

Informational package. Заселенность номеров отеля.

Время	**Отель**	**Тип номера**	**…**	**…**	**…**
Год	Сеть отелей	Тип номер			
Квартал	Название филиала	Размер			
Месяц	Код филиала	Количество кроватей			
Дата	Регион	Тип кровати			
День недели	Адрес	Количество посетителей			
День месяца	Код города/региона	Холодильник			
Сезон	Дата постройки	Кухня			
Флаг праздников	Дата обновления				
Измеряемые факты: свободные номера, занятые номера, недоступные номера, количество посетителей, доход…					

Самостоятельное изучение и кейсы

- Ponniah – Data warehousing. Fundamentals for IT Professionals, стр. 73-120.
- Kimball, Ross – Datawarehouse & BI, стр. 113-132.

Самостоятельное изучение и кейсы.

Вы – аналитик проектной команды, задачей которой является создание хранилища данных для некоторой компании (предметная область исследования диссертации). Составьте один или несколько Informational package для вашей темы (слайды 17-19).

Отчет оформить в шаблоне документа (форма документа для кейса лекции 4.docx)

Спасибо за внимание!