Проектирование хранилищ данных

ФИО преподавателя: Смирнов Михаил Вячеславович

e-mail: smirnovmgupi@gmail.com

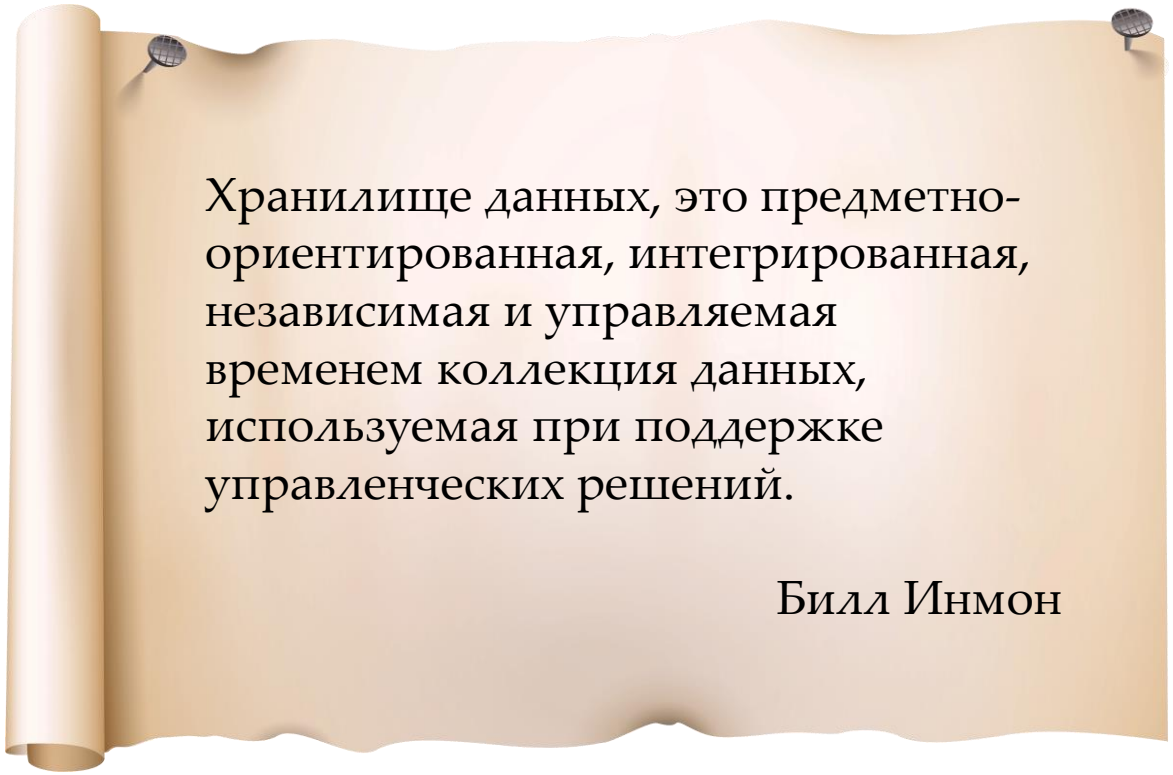
Лекция 2

Хранилище данных. Определение и компоненты.

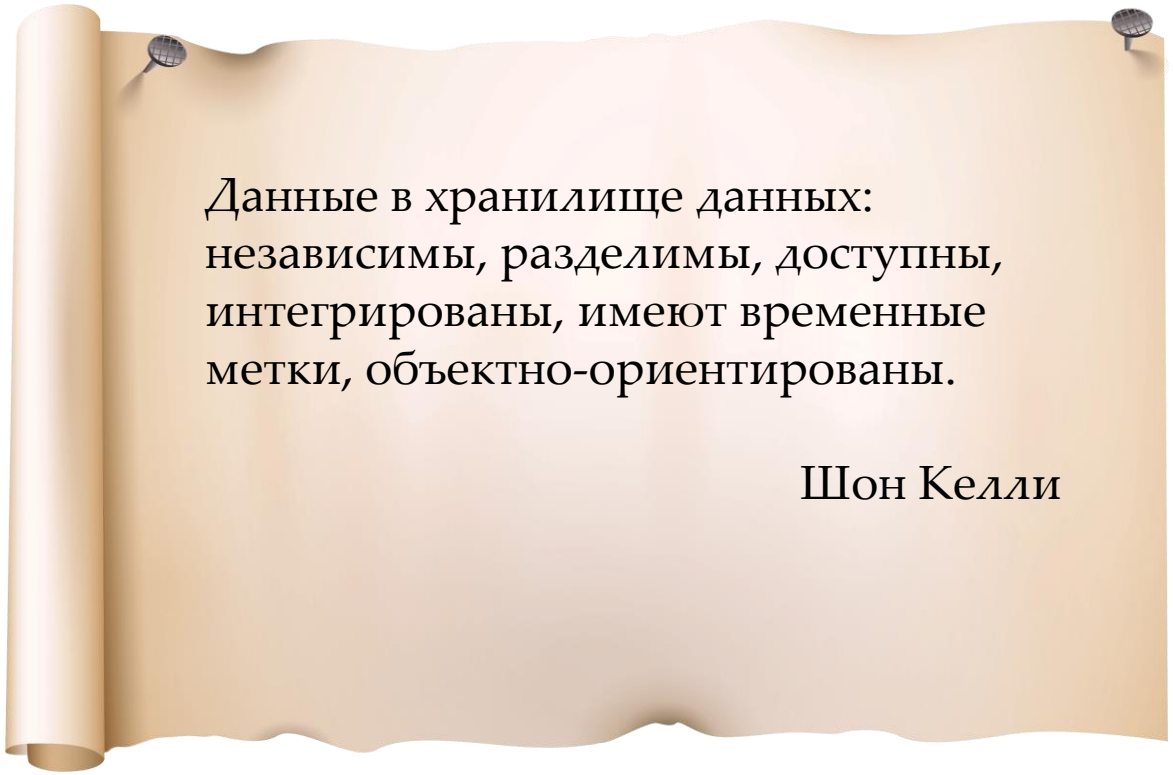
Вопросы лекции

- Определение хранилища данных
- Свойства хранилищ данных
- Классическая трехуровневая архитектура хранилища данных
- Модель Кимбалла и Data-Marts
- Модель Инмона и компоненты Хранилища Данных
- Критерии выбора Kmb-Inm
- Типы метаданных хранилища
- Кейсы для самостоятельного изучения

Определение хранилища данных



Определение хранилища данных



Разберем свойства хранилищ данных

- Предметно-ориентированные данные
  - В отличие от данных в операционной системе, приходящих со стороны приложений, источником данных для хранилища являются реальные бизнес-объекты или события.
- Интегрированные данные
  - В хранилищах данных массив данных представляет собой комплексный агрегат из разнородных, синхронизированных данных, получаемых из разных источников (так, агрегат Банковский счет клиента может состоять из трех потоков данных: вклады, займы, баланс счета). Каждый из потоков данных синхронизируется с остальными, входящими в агрегат (названия массивов данных, коды, типы данных, единицы измерения…)

Разберем свойства хранилищ данных

- Изменяющиеся во времени данные
  - Поскольку хранилища данных в первую очередь используются в аналитике и прогнозировании (текущие паттерны поведения покупателя, прогнозы поведения покупателя…), они оптимизированы для хранения не только текущих, но и исторических данных.
- Независимые данные
  - Данные из операционных систем переносятся в хранилища в определенные интервалы времени. В зависимости от требований бизнеса, они могут переноситься раз в день, в неделю, в две недели. Иногда интервал может быть и больше. При этом для разных наборов данных в хранилище могут использоваться разные интервалы данных (описание продуктов можно обновлять раз в две недели, тогда как данные об их продажах обновляются каждый день)

Разберем свойства хранилищ данных

- Гранулярность данных
  - В отличие от операционных систем (где данные лучше хранить на максимально возможном уровне детализации, - данные с конкретной кассы, конкретного магазина), хранилище подразумевает возможность гранулярности (использование укрупненных, агрегированных данных).

Три уровня грануляции в банковском хранилище данных

Дневная детализация
- Аккаунт
- Даты транзакций
- Количество
- Внесение/снятие средств

Месячная детализация
- Аккаунт
- Месяц
- Количество транзакций
- Снятия средств
- Внесения средств
- Баланс на начало
- Баланс на конец

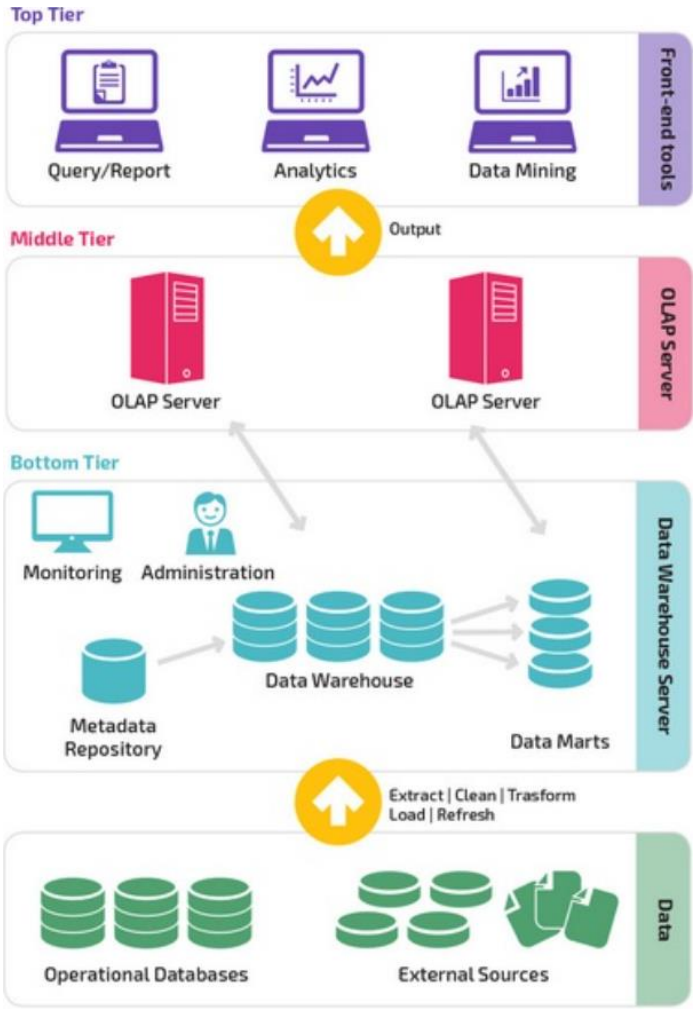
Квартальная детализация
- Аккаунт
- Месяц
- Количество транзакций
- Снятия средств
- Внесения средств
- Баланс на начало
- Баланс на конец

Классическая трехуровневая архитектура хранилища данных

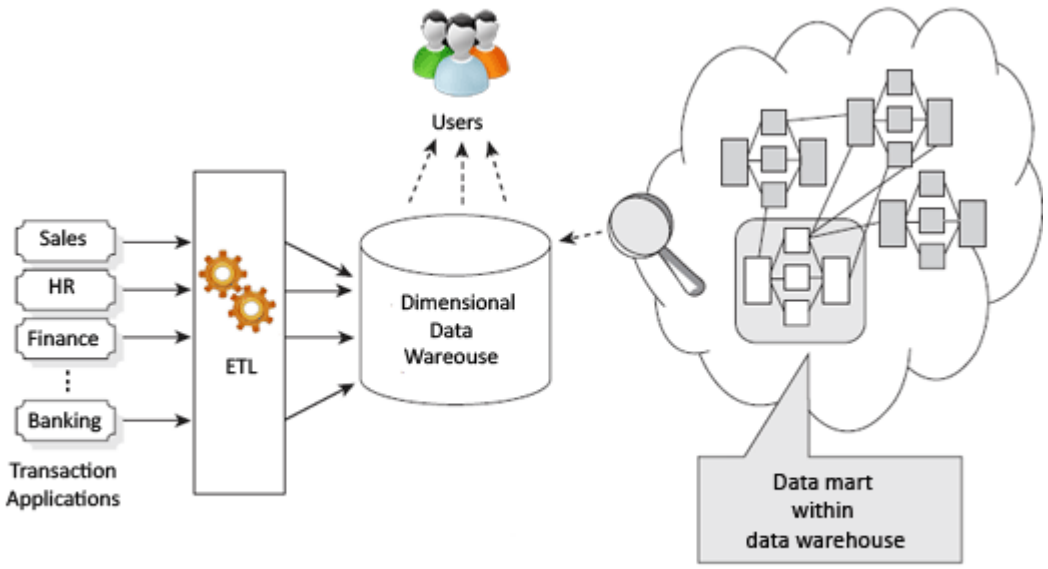
Нижний уровень: сервер базы данных, используемый для извлечения данных из множества различных источников.

Средний уровень: средний уровень содержит сервер OLAP, который преобразует данные в структуру, лучше подходящую для анализа и сложных запросов.

Верхний уровень: уровень клиента. Инструменты, используемые для высокоуровневого анализа данных, создания отчетов и анализа данных.

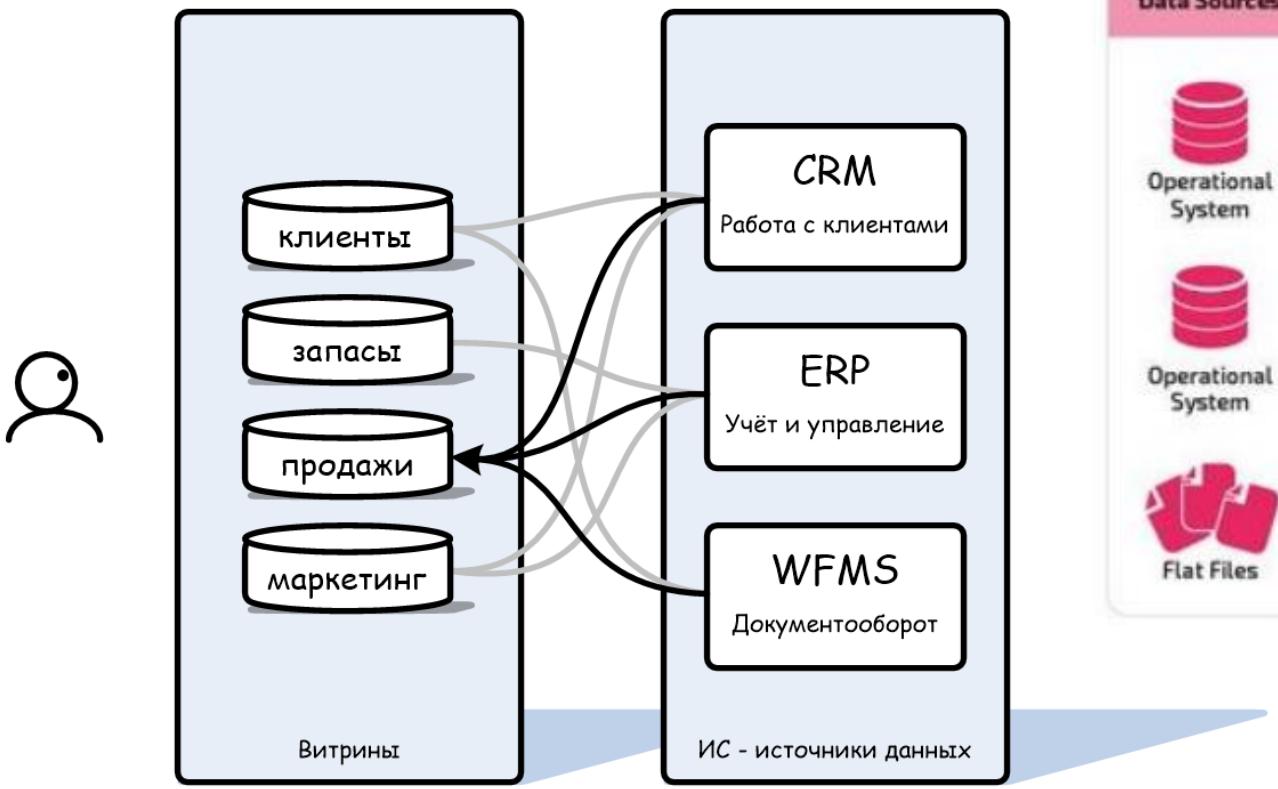


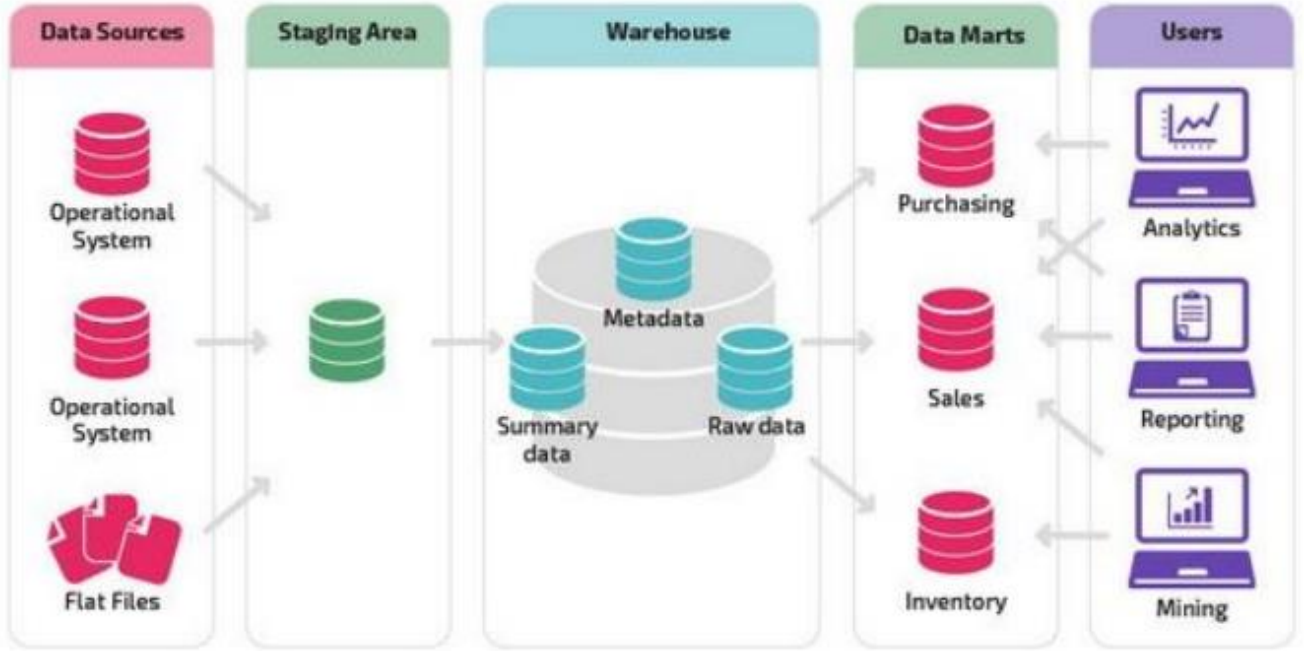
Модель хранилища мистера Кимбалла (Kimball)



- Transaction applications – системы или сервисы, созданные для обработки бизнес транзакций. Результатом обработки являются данные, сохраненные в РМД или в файлах.
- ETL – сервисы приведения данных к стандартному виду. Расшифровывается как достать, изменить, загрузить.
- A dimensional data warehouse. Содержит корпоративные данные в виде, наиболее удобном для гранулирования. Используется модель измерений. Как правило имеет форму звездочки. Аналитические системы способны извлекать данные напрямую из модели.
- Datamart. Витрина данных. Существует за пределами хранилища и представляет собой логический концепт.

Место витрин данных (data mart) в процессе аналитики данных

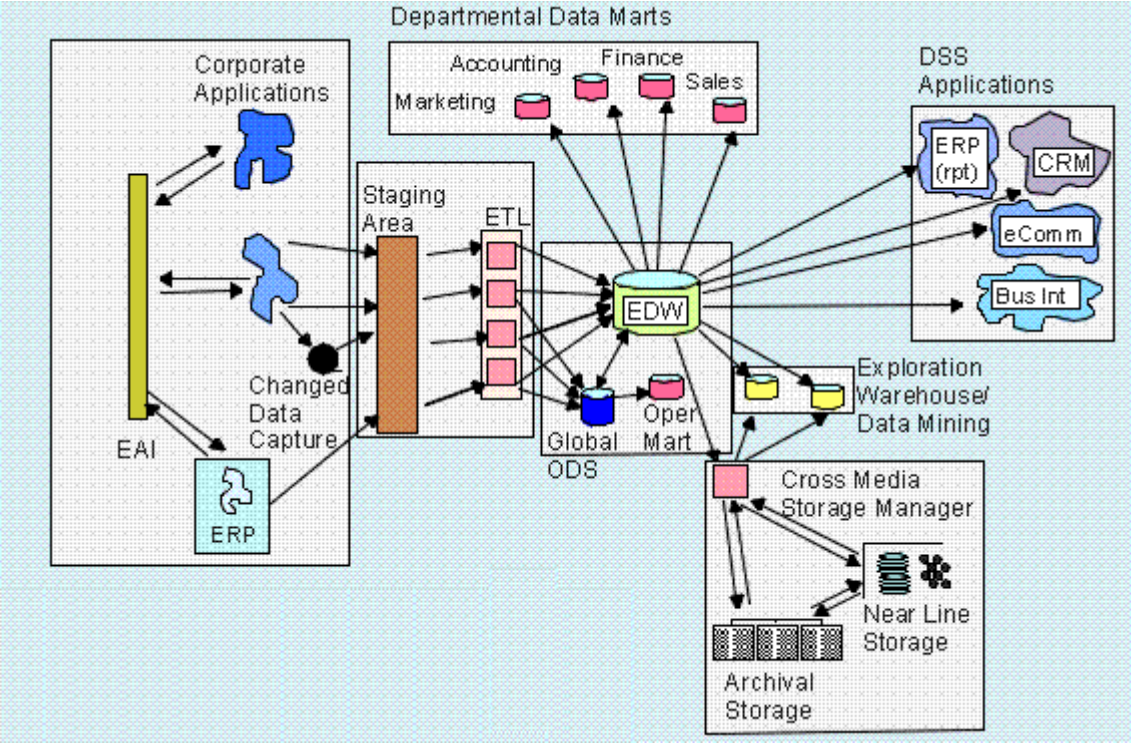




Почему витрины, а не простые запросы и представления?

- Изоляция данных. Обновление данных, например загрузка справочника, может быть растянуто во времени, при этом до окончания загрузки текущая версия справочника должна быть полностью доступна с исключением «грязного чтения» загружаемой версии.
- Гарантии атомарности операций при обновлении данных. В случае сбоев и ошибок загрузки данных витрина остаётся в состоянии, которое предшествовало сбойному процессу. Другими словами, или данные обновляются полностью, или не обновляются вовсе, не оставляя следов сбойных операций.
- Системная темпоральность. Мало какая реляционная СУБД имеет функцию системной темпоральности «из коробки». Ведение системного времени и версионирование записей по системному времени позволяет сравнивать состояние данных витрины между двумя разными моментами времени или проводить «расследование», основываясь на данных, которые были в витрине в определенный момент в прошлом.
- Эффективное выполнение различных видов запросов: сравнительно редких и тяжелых аналитических запросов, затрагивающих большой объем данных (OLAP-нагрузка), и множества одновременных простых запросов (OLTP-нагрузка). Как правило, СУБД заточены на какой-то один вариант нагрузки: OLAP или OLTP.

Модель хранилища мистера Инмона (Inmon)



- Corporate Applications. В данной модели называются также системами источника, и предоставляют данные в хранилище.
- ETL Processes. Также называются data services. После обработки данных в рамках установленного формата, они загружаются в корпоративное хранилище. Процессы ETL могут выполняться в режиме batch или в режиме real-time.
- Enterprise data warehouse. Центральный элемент архитектуры, объединенный репозиторий атомарных данных, находящихся в третьей нормальной форме (см. реляционную модель).
- Data marts. Выбирают данные из хранилища с помощью инструментов агрегации. Аналитические приложения обращаются не к хранилищу, а к заранее приготовленным витринам данных.

Компоненты хранилища данных

- Источники данных
  - Production – данные, поступающие из финансовых систем, систем управления производством, CRM (взаимоотношения с клиентами) и SCM (управление поставками) систем.
  - Internal – внутренние данные: заметки, профили, данные из департаментов (весьма проблемные при очистке).
  - External – внешние данные: статистические, справочные данные из внешних источников (компания сдающая в аренду автомобили интересуется мировой номенклатурой автомобилей и свободными данными производителей авто, также проблемные при очистке).
  - Archived – архивные данные разной степени архивации (отдельная база данных онлайн, файлы на носителе, пленка).

Компоненты хранилища данных

- Зона посадки (Data staging). Состоит из трех последовательных процедур Extraction, Transformation, Loading (сокращенно - ETL).
  - Extraction – функция получения данных из многочисленных источников. Несколько сложнее для традиционных хранилищ, существенно проще для Hadoop (Big Data хранилищ). В итоге получается группа файлов с данными, промежуточная реляционная база данных или комбинация этих вариантов.
  - Transformation – функция приведения данных в состояние «согласованности». Включает в себя этапы: очистка данных, стандартизация типов данных и ограничений, семантическую стандартизацию, при необходимости – агрегацию.
  - Loading – функция загрузки данных согласно установленным бизнес-требованиям.

ETL и хранилище данных

1. Извлечение данных из пула источников данных. Данные хранятся во временной промежуточной базе данных.
2. Выполняются операции преобразования, чтобы структурировать и преобразовать данные в подходящую форму для целевой системы хранилища данных.
3. Структурированные данные загружаются в хранилище и готовы к анализу.

ELT и HDFS

1. Данные сразу же загружаются после извлечения из исходных пулов данных. Промежуточная база данных отсутствует, что означает, что данные немедленно загружаются в единый централизованный репозиторий.
2. Данные преобразуются в системе хранилища данных для использования с инструментами бизнес-аналитики и аналитики.

Компоненты хранилища данных

- Хранилище данных.
  - Реализуется в рамках одного из доступных паттернов.
  - Данные в классическом хранилище работают в режиме «read-only».
  - Loading – функция загрузки данных согласно установленным бизнес-требованиям.
- Компонент доставки информации.
- Метаданные.
  - По сути – содержимое физического словаря данных. Информация о модели хранения, файлах и адресах, индексах, «данные о данных» и т.д..
- Управление и мониторинг.

Критерии выбора Kmb-Inm.

Характеристика	Kmb	Inm
Уровень бизнес-решений	Тактический	Стратегический
Требования по интеграции данных	Отдельные бизнес-требования	Корпоративные бизнес-требования
Структура данных	Простые количественные данные, KPI	Многослойные данные, качественные данные
Стабильность данных	Источники редко меняются	Источники часто меняются
Сложность исполнения	Маленькая команда проекта	Большая команда проекта
Время на развертывание	Срочно требуется хранилище	Есть существенное время на проектирование и релиз
Затраты	Сравнительно низкие затраты	Высокие затраты со старта

Категории метаданных Хранилища данных

- Operational – информация о том, как «понимать» хранящиеся данные. Типы данных, наложенные на данные ограничения.
- Extraction and Transformation – запрограммированный алгоритм сбора данных, бизнес-правила сбора данных, частота сбора данных.
- End-User – навигация по хранилищу данных. Позволяет конечному пользователю получить доступ к данным хранилища.

Самостоятельное изучение и кейсы

- Ponniah – Data warehousing. Fundamentals for IT Professionals, стр. 23-44.
- Архитектура хранилищ данных: традиционная и облачная (статья) https://habr.com/ru/post/441538/
- Концепции хранилищ данных: Kimball vs. Inmon (статья) https://observsp.blogspot.com/2021/07/kimball-vs-inmon.html

Кейс для тренировки. Страховая компания.

Вы – аналитик проектной команды, задачей которой является создание хранилища данных для страховой компании. Перечислите все возможные источники данных, из которых вы можете брать данные для вашего хранилища. Рассмотрите все возможные типы источников (слайд 14).

Дайте пояснение каждому выбранному источнику и потоку данных.

Отчет оформить в шаблоне документа (форма документа для кейса лекции 2.docx)

Спасибо за внимание!