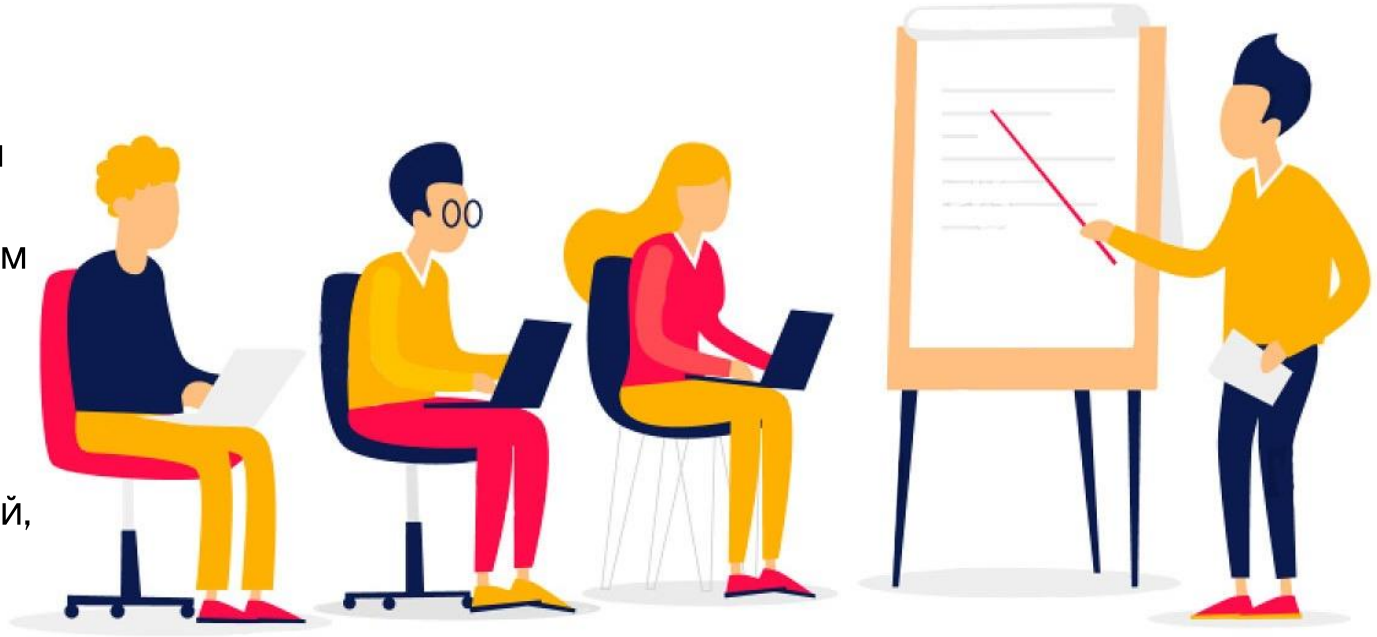
Ростелеком

Эволюция аналитических хранилищ данных

04.03.2019

Когда начинают задумываться о хранилище данных?
Запросы от бизнес-пользователей…

- «Мы имеем горы информации, но не имеем доступ к ней…»
- «Необходимо смотреть на информацию в требуемых разрезах и с достаточной детализацией…»
- «Предоставить бизнес-пользователям прямой доступ к данным…»
- «Требуется выделять только важную информацию из потока данных…»
- «Избегать противоречивых отчетов, приходящих из разных подразделений, при отражении общего бизнес-показателя…»
- «Принимать решения на основе подтвержденных фактов, нежели интуитивным способом…»



Общие предпосылки создания хранилищ данных

развитие технологий → рост внешних и внутренних источников информации → рост видов и объемов данных → рост срезов для анализа

Тенденции в появлении новых видов данных

- Данные финансовых операций и учетных систем предприятия
- Данные систем ресурсного, финансового и производственного планирования
- Данные внешних поставщиков (гос. органов, аналитических агентств, профессиональных сообществ)

1960-2000

Данные социальных сетей и интернет ресурсов

2000-2012

Данные информационных потоков и IoT (Internet of Things)

> 2012

Организационные предпосылки

- Автоматизация процессов формирования отчетных и аналитических данных.
- Исключение влияния централизованных аналитических систем на операционные бизнес процессы.
- Прозрачность (единообразие) методологии подготовки и обработки данных рассчитываемых отчетных и аналитических показателей.
- Обеспечение возможности проведения аудита рассчитываемых отчетных и аналитических показателей.

Технические предпосылки

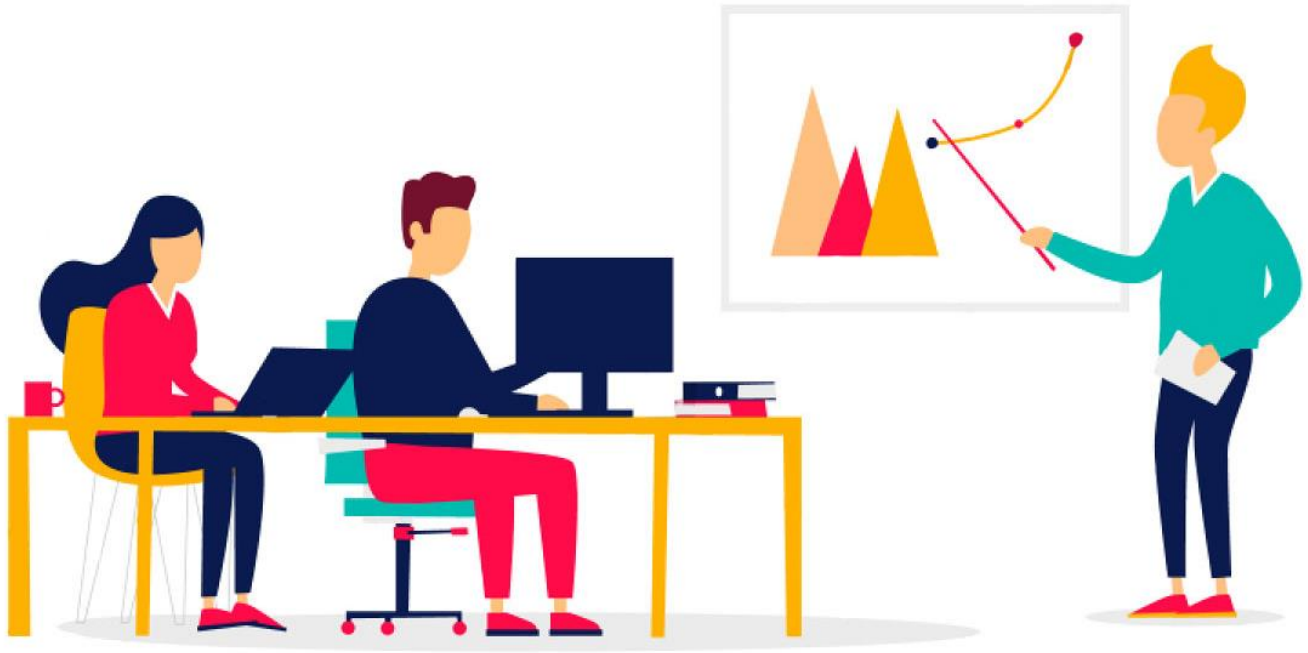
- Необходимость длительного хранения архивных данных для ретроспективного анализа
- Снижение производительности и дополнительная нагрузка на операционные учетные системы предприятия
- Сниженные требования к времени отклика
- Гибкость и сложность аналитических запросов
- Увеличение мощности аппаратной платформы (снижение стоимости хранения данных)
- Появление дружественного интерфейса и возможность работы с ИТ системами бизнес-пользователям

- Разделение решения задач обработки транзакций и задач анализа данных
- Проектирование отдельной структуры хранения архивов для анализа данных

Выделение отдельного класса аналитических приложений

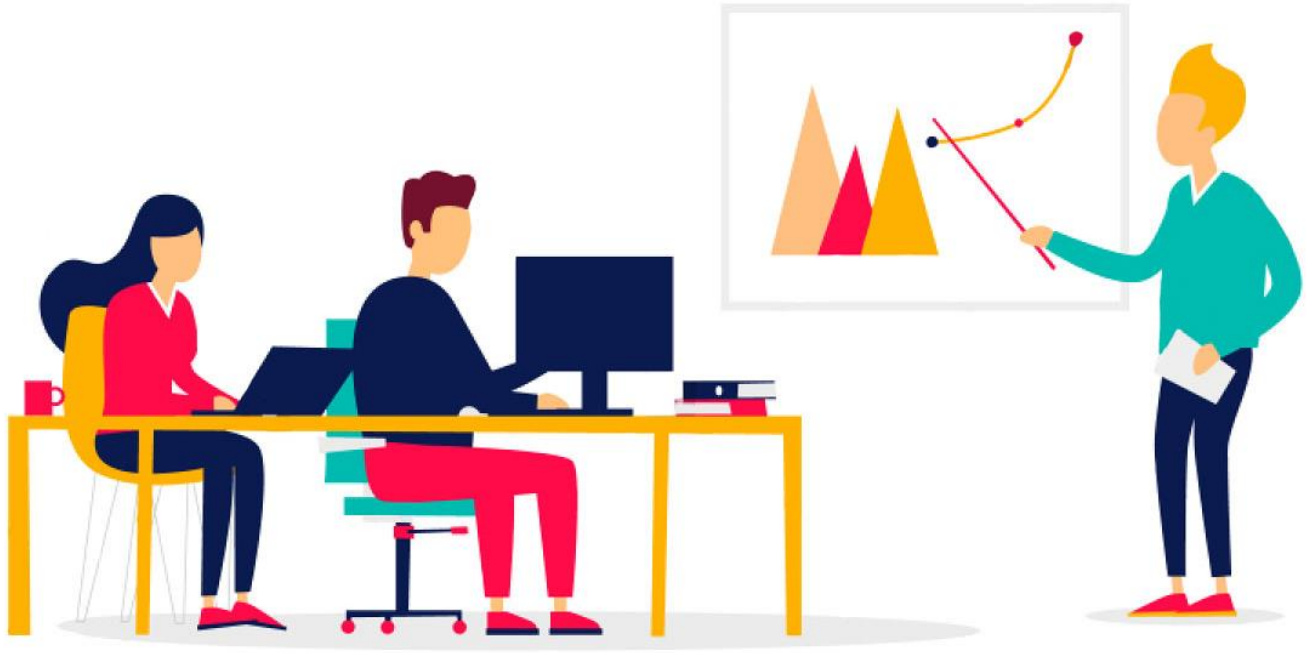
Требования к корпоративному хранилищу данных

- Обеспечение легкого доступа к информации организации
- Обеспечение согласованной информации организации
- Адаптация и устойчивость к изменениям
- Безопасность информационных активов организации
- Основа для совершенствования процессов организации



Требования к корпоративному хранилищу данных

- Обеспечение легкого доступа к информации организации
- Обеспечение согласованной информации организации
- Адаптация и устойчивость к изменениям
- Безопасность информационных активов организации
- Основа для совершенствования процессов организации



Бизнес-сообщество должно принять хранилище данных, чтобы оно считалось успешным

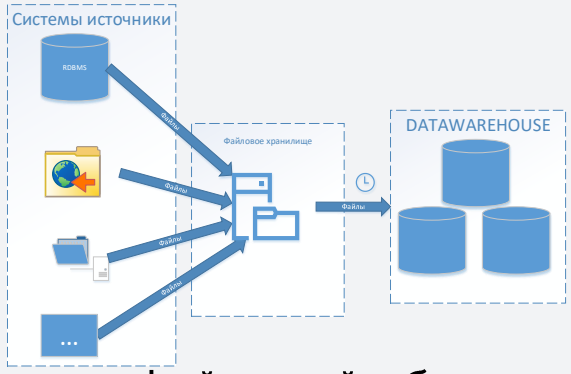
Отличия аналитических хранилищ данных от систем оперативного учета

Оперативные данные	Аналитические данные
Детальные	Сводные (приведены к корпоративному формату)
Актуальные в момент доступа	Представляют версию записи (или снимок) в момент времени
Модель данных ориентирована на приложение (для оперативного учета)	Модель данных ориентирована предметную область (для анализа)
Высокие требования к времени отклика	Сниженные требования к времени отклика
Доступ к чтению и обработке отдельной записи	Доступ к чтению и обработке отдельной набора записей
Управление обновлением записи серьезная проблема процесса обработки данных	Управление обновлением записи не используется
Применяется цикл разработки SDLC	Применяется цикл разработки CDLS

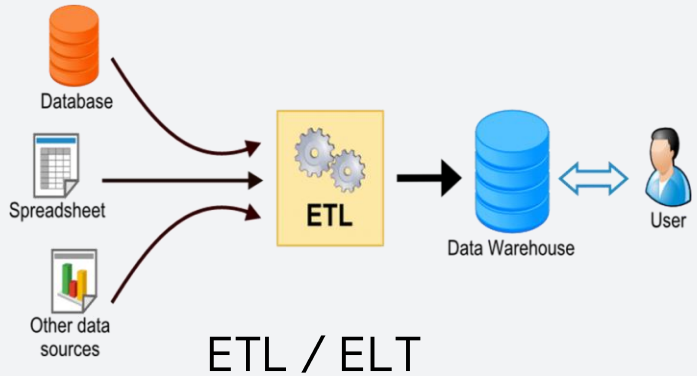
Способы интеграции данных в хранилище

Ростелеком

batch

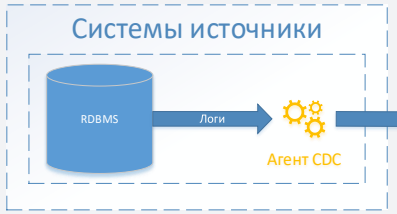


файловый обмен

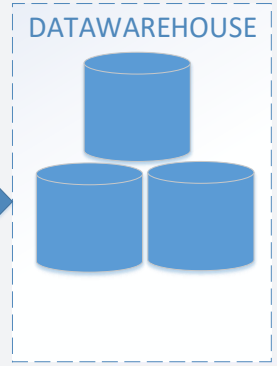


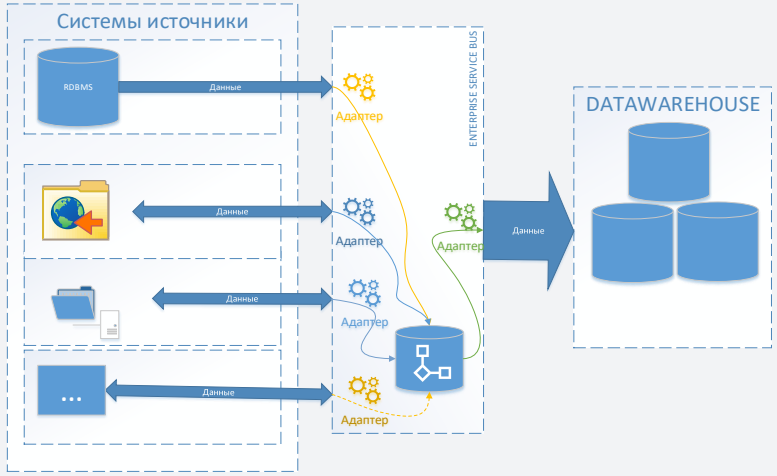
ETL / ELT

online

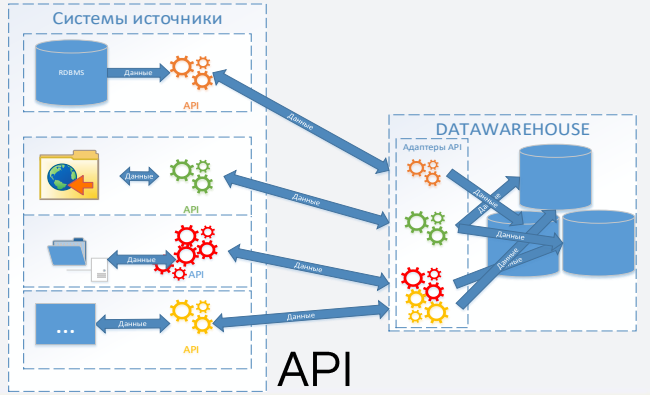


CDC





ESB



API

Подходы ведения историчности данных в OLTP системах

- Данные не версионируются / история изменений не ведется

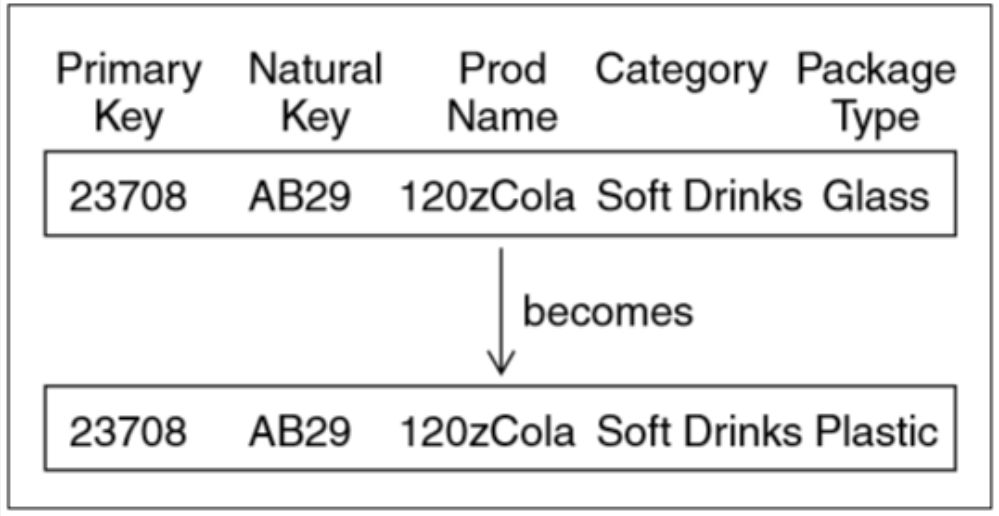
- Отслеживание изменений :

UPDATE_TIME	LAST_STATE	CURRENT_STATE

1. Добавление полей аудита :
2. Сохранение полной истории изменений :
   - Создание отдельных исторических таблиц по изменению атрибутов с определением номера версии, даты изменения и аудита изменяемых атрибутов
   - Сохранение старых значений строки в отдельной исторической таблице (триггер)
   - Версионирование строк таблицы. Данный подход редко применим в OLTP системах, но часто используется в хранилищах данных

Slowly Changing Dimensions.
Тип 1

Тип 1 : SCD – это простая перезапись одного или нескольких атрибутов в существующей записи измерения. Изменяемые значения SCD1 не представляют интереса для сохранения истории.



** Хотя значение атрибутов типа 1 не влияет на ключи таблиц фактов и других измерений, оно может оказывать влияние на совокупность фактических таблиц, если агрегат создается непосредственно на атрибуте, который был изменен.*

*В базах данных хранилища используют собственные первичные ключи

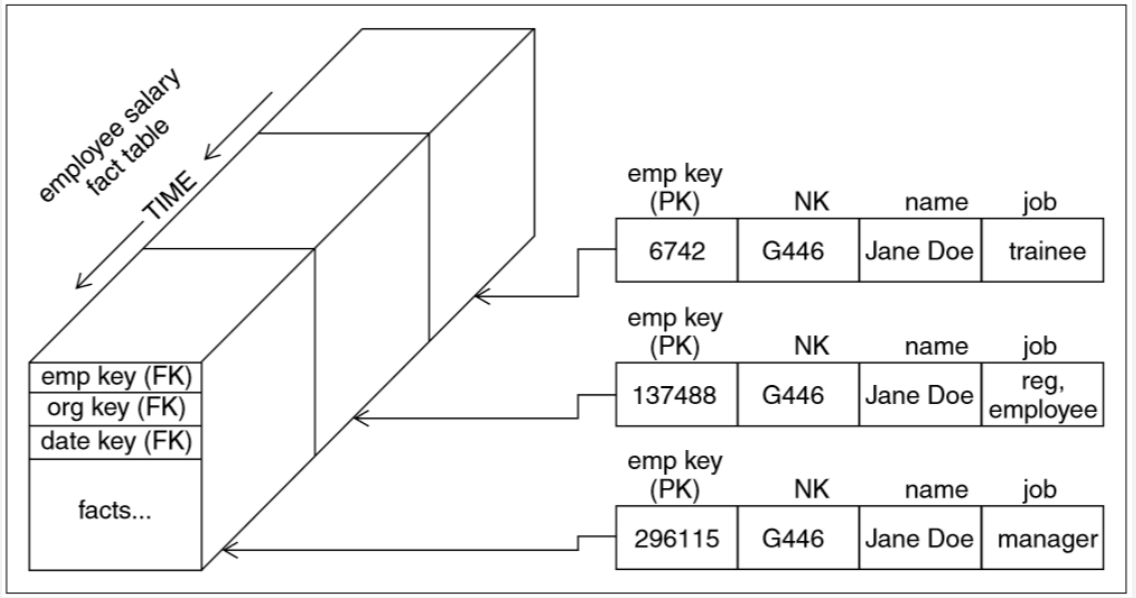
Slowly Changing Dimensions.

Тип 2

- Для новой и измененной записи присваивается новый суррогатный первичный ключ
- Измененная запись не оказывает влияния на уже рассчитанные таблицы фактов и агрегаты

Дополнительные поля для ведения историчности :

- Change DateTime (дата:время изменения записи)
- Effective DateTime (дата:время начала действия версии записи)
- Expiration DateTime (дата:время завершения действия версии записи)
- Active Flag (признак активной записи)
- Deleted Flag (признак логического удаления записи)

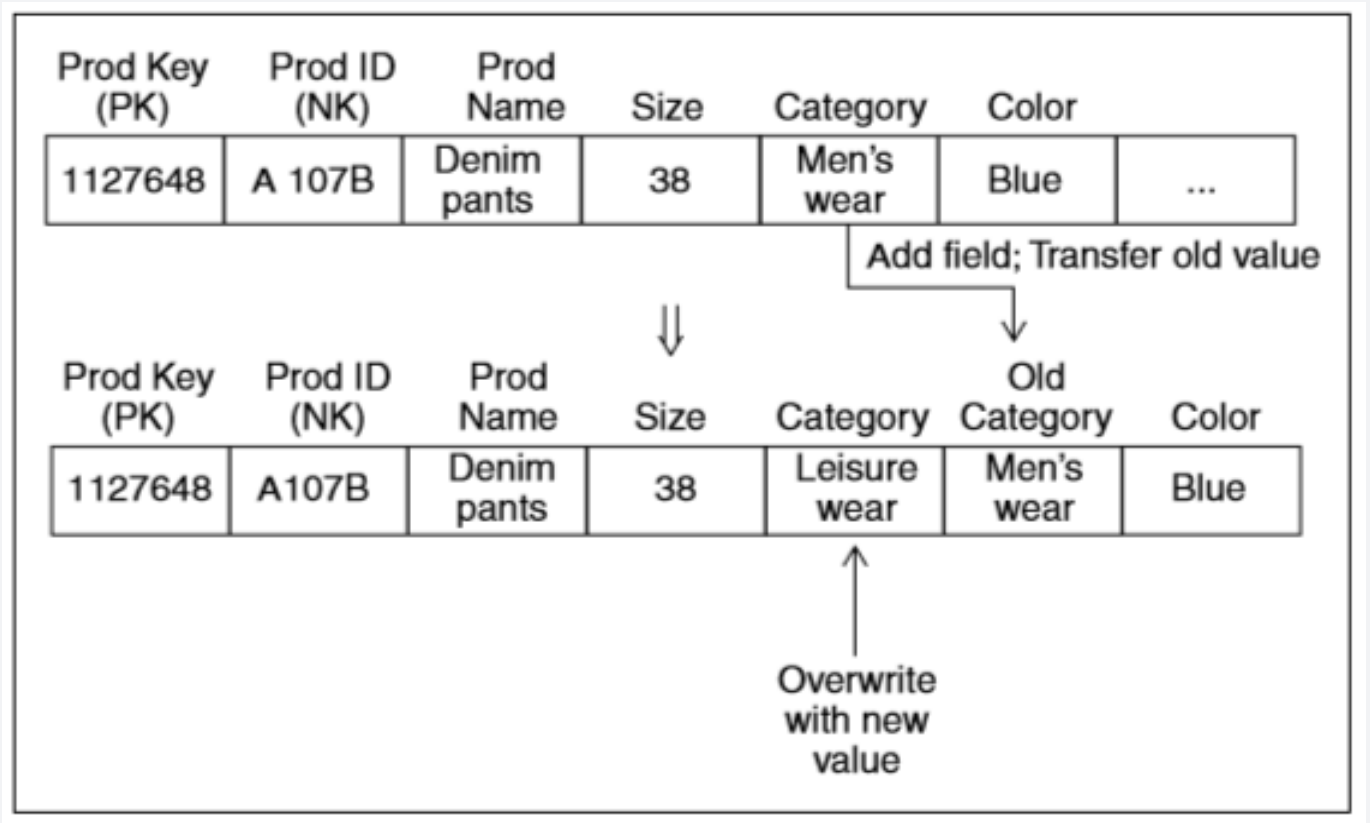


** В случае, если натуральный ключ на источнике был изменен – возникают проблемы с потерей истории в хранилище данных.*

Slowly Changing Dimensions

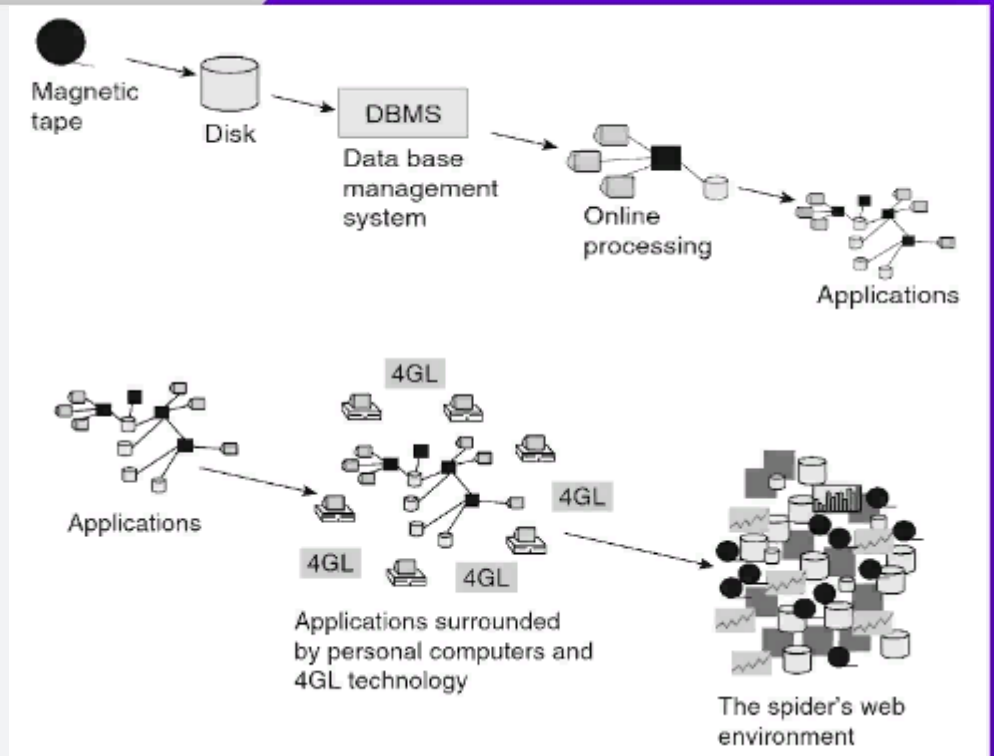
Тип 3: Slowly Changing Dimension (альтернативная реальность)

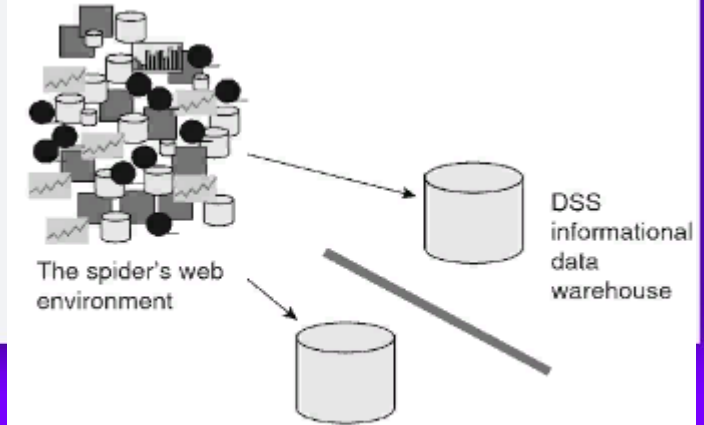
Используется когда пользователи хотят иметь возможность выбора использования старого или нового значения атрибута, таким образом старое значение становится вторым действительным (альтернативным) вариантом



Хронология развития информационных систем и хранилищ данных

- 1960: Данные хранятся на магнитных лентах и перфокартах, последовательное чтение данных. Сложность создания и сопровождения программ на Fortran и COBOL.
- 1970: Дисковое хранение данных, появление DBMS
- 1975: Появление OLTP систем (bank teller systems, Manufacturing control systems, …)
- 1980: Появление PC и языков программирования 4 поколения
- 1985: Появление extract programs и «естественно-развивающейся архитектуры» привело к явлению spider web (неконтролируемый рост обработки и извлечения данных в организации).





Хронология развития хранилищ данных

- 1988: Появились первые статьи, посвященные именно ХД, их авторами были Б. Девлин и П. Мэрфи
- 1992: Вышла монография Билла Инмона (Bill Inmon) «Building the Data Warehouse», где описана концепция «Enterprise Data warehouses» и четко сформулировано понятие хранилищ данных.
- 1996-2013: Вышла книга и множество редакций+дополнений Ральфа Кимбалла «The Data Warehouse Toolkit» (ISBN 9780471153375) с альтернативной концепцией хранилищ данных и понятием «Dimensional data warehouse».
- 1998: Вышла книга Билла Инмона «Corporate Information Factory» с развитием предыдущей концепции EDW в виде концепции «Корпоративная информационная фабрика (Corporate Information Factory, CIF) (ISBN 9780471399612)

Bill Inmon

Ralph Kimball

Хронология развития хранилищ данных

- 2008: Вышла книга Билла Инмона с соавторами «DW 2.0: The Architecture for the Next Generation of Data Warehousing» с развитием предыдущей концепции EDW и CIF. (Bill Inmon, Derek Strauss, Genia Neushloss; ISBN 9780080558332)

Bill Inmon

- 2010-е: Становится распространенным комбинирование подходов в архитектуре и жизненном цикле ХД.

- 2015: Подход «λ архитектура» (книга Натана Марза/Nathan Marz и Джемса Варрена/James Warren «Big Data:Principles and best practices of scalable realtime data systems» ISBN 9781617290343)

Nathan Marz

Хронология развития hadoop

2003-2008 Google и Yahoo разрабатывают framework распределенных вычислений и хранения данных, который лег в основу Hadoop.

2005—2006 годов Hadoop развивался усилиями двух разработчиков — Дуга Каттинга (Doug Cutting) и Майка Кафареллы (Mike Cafarella) в режиме частичной занятости, сначала в рамках проекта Nutch, затем — проекта Lucene.

2008 группа энтузиастов создает компанию «Cloudera». В марте 2009 года была представлена первая версия Hadoop от Cloudera.

2009 - 2011 появляются первые последователи «Cloudera», проверившие эффективность Hadoop на практике. Речь идет о MapR и Hortonworks.

2012 - 2014 направление BigData становится main stream в области хранения и обработки данных.

2014 - 2015 Big Data достигает пика популярности.

Появляются новые технологии и инструменты, например: Apache Spark, Apache Flink, Apache Kafka и т.д.

Появляются концепции - «Data Lake»,"Data Hub" и «Lambda Architecture».

Mike Cafarella

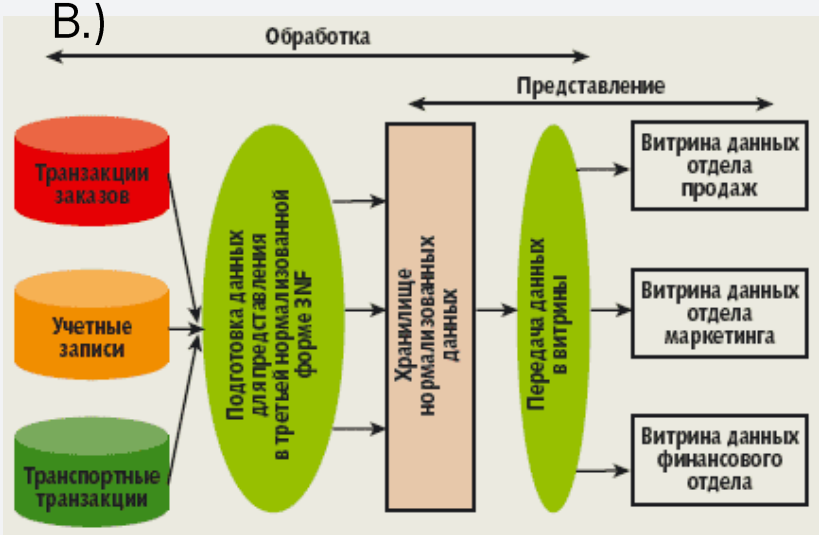
Doug Cutting

Концепция Enterprise Data Warehouse [публикация - 1992 год]

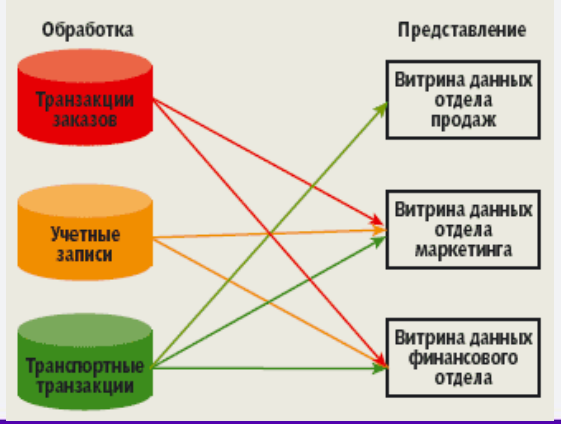
«Хранилище данных - предметно-ориентированный, неизменяемый, поддерживающий хронологию и интегрированный источник данных»

- Стратегический подход
- В хранилище все данные предприятия хранятся в 3-й нормальной форме
- Предметно-ориентированность
- Неизменность
- Поддержка историчности
- Интегрированность
- Витрины данных ориентированы на субъекты или подразделения

Enterprise data warehouse (Inmon B.)



Подход без хранилища



Преимущества концепции Enterprise Data Warehouse

Преимущества:

- Все корпоративные данные полностью документированы
- Данные эффективно хранятся в 3-й нормальной форме в одном хранилище (в единой модели)
- Данные легко доступны для извлечения в витрины данных для бизнес-пользователей

Недостатки:

- Высокая стоимость
- Длительность внедрения первого релиза

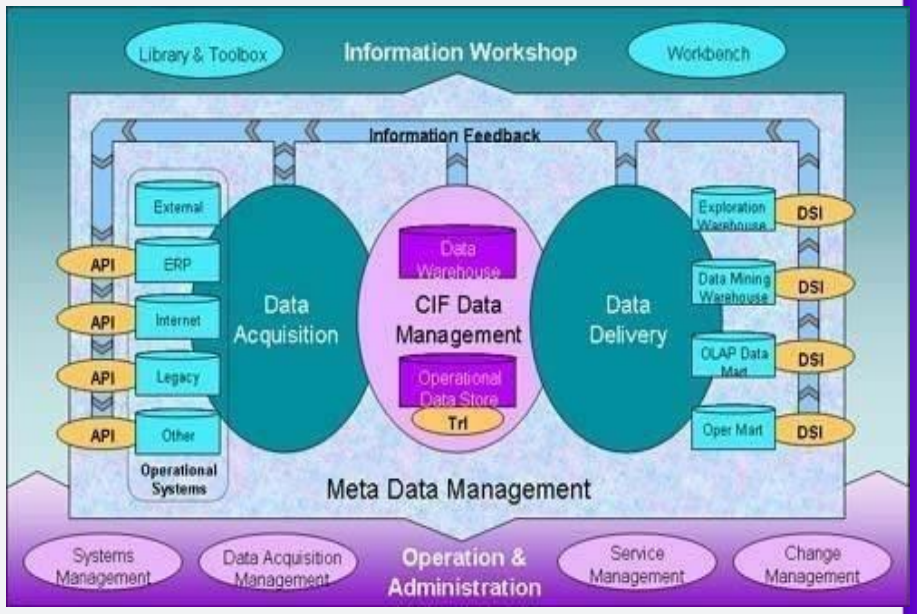
«Корпоративная информационная фабрика» (CIF) [развитие концепции EDW в 1998 году]

Ростелеком

«Корпоративная информационная фабрика» (CIF)

- **Сбор данных(Data Acquisition)**: инструменты и системы управления для извлечения, преобразования и загрузки данных из различных систем-источников (внешних систем, ERP, внутренние системы, файлы и т.д.) в хранилище данных.
- **Data Management(Оперативный склад данных):** хранилище актуальных детальных данных, обновляемых из источника в режиме близкому к реальному времени. Предназначено для принятия бизнесом тактических решений по оперативным данным.
- **Data Management(Хранилище данных)**: хранилище данных, в соответствии с изначальным определением EDW. Является центральной точкой для интеграции данных, централизует всю информацию. Доступ данных извне осуществляется только через витринный слой.
- **Доставка данных (Data Delivery)**: операции агрегирования, фильтрации, переформатирования информации по размеру или бизнес-требованиям для использования конкретных инструментов BI. Передача информации в рамках всей организации (для наполнения конкретных Datamarts или Datawarehouse).

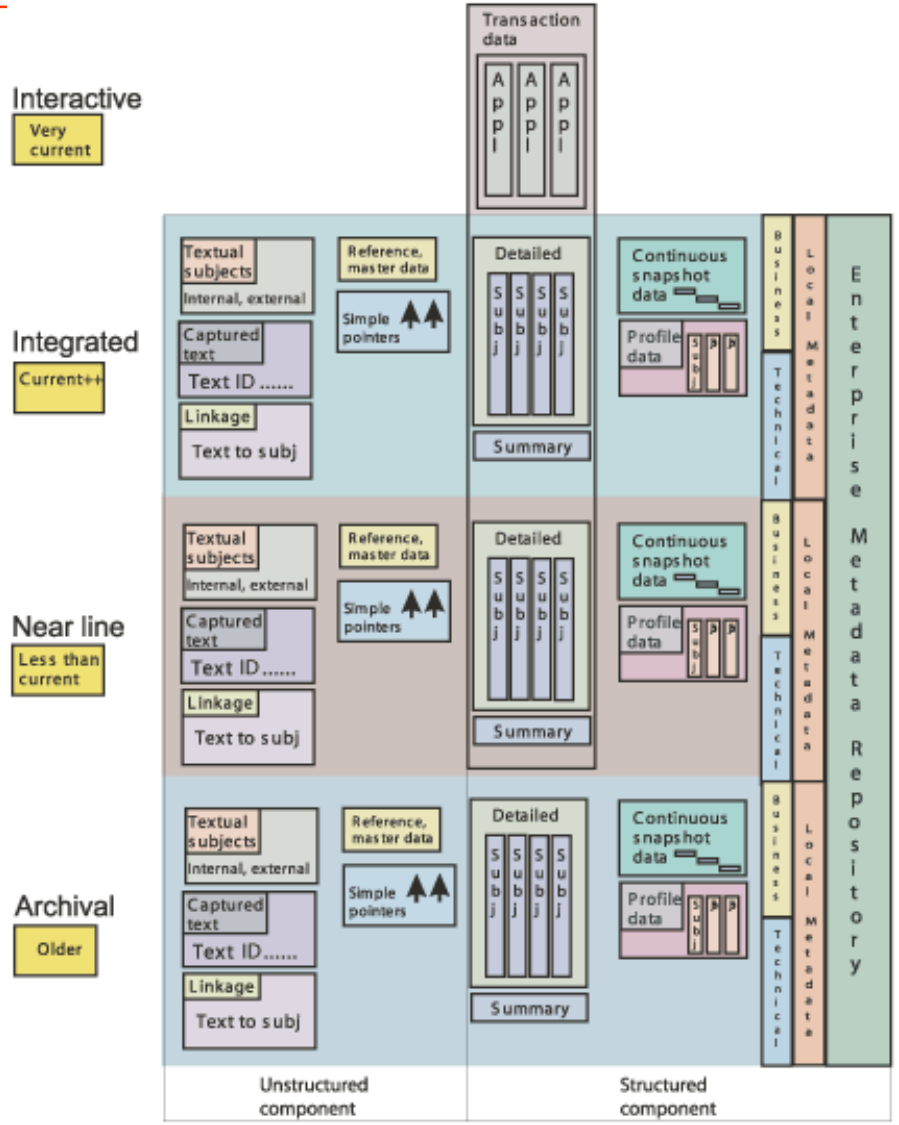
Corporate Information Factory (Inmon B.)



Концепция DW 2.0
[развитие концепции EDW+CIF в 2008 году]

В концепции DW 2.0 описываются следующие аспекты корпоративного хранилища данных:

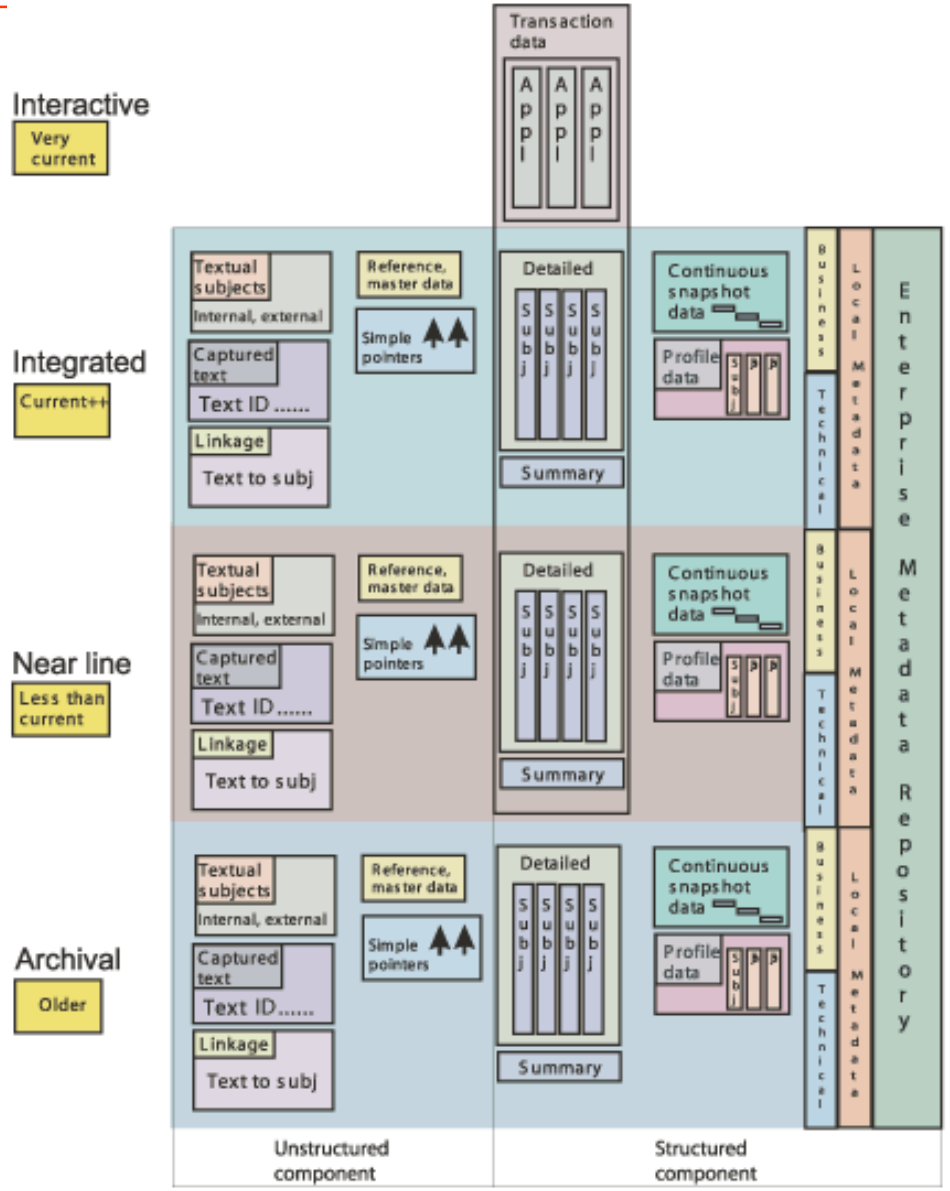
- классификация и редактирование неструктурированных данных, которые могут быть представлены в различных формах;
- управление метаданными, в том числе бизнес-информации и технических метаданных;
- высокоскоростной доступ к данным в интерактивном режиме с возможностью их обновления;
- spiral development methodology



Концепция DW 2.0
[секторы данных]

Сектора/зоны данных

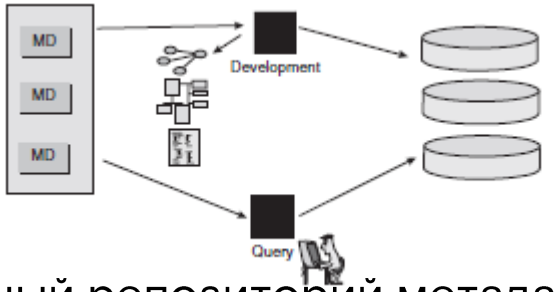
- Интерактивный (< 1 месяца)
- Интегральный (1 день – 2-3 года)
- Ближайший (6 мес – 10 лет)
- Архивный (длительное хранение,)



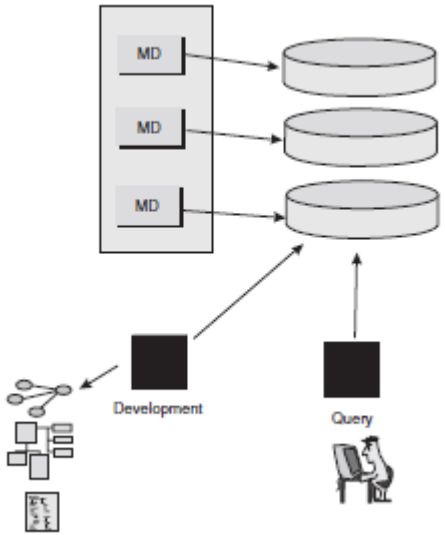
Концепция DW 2.0
[метаданные в корпоративном хранилище]

Уровни метаданных в хранилище:

- Технические метаданные
- Бизнес метаданные
- Локальные
- Корпоративные



активный репозиторий метаданных



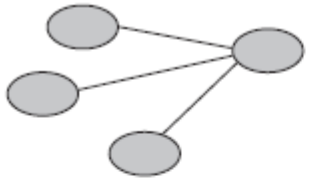
пассивный репозиторий метаданных

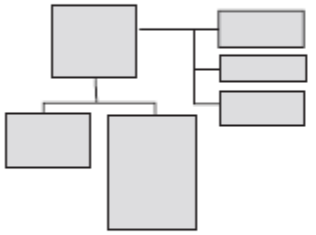
** If by some miracle the passive metadata repository does get built, it is soon out of date as changes to the system are not reflected in the passive repository.*

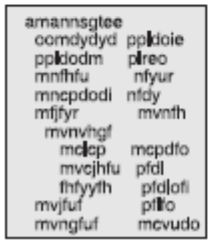
Концепция DW 2.0
[модель данных]

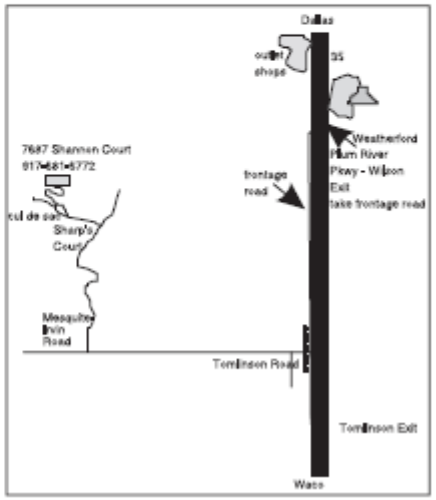
Уровни модели данных в хранилище:

- ERD (Entity Relationship Level)
- DIS (Data Set Item)
- Low-level (Physical model)

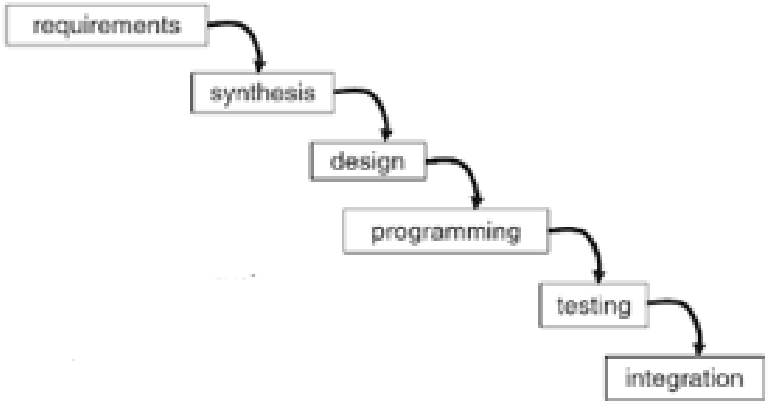


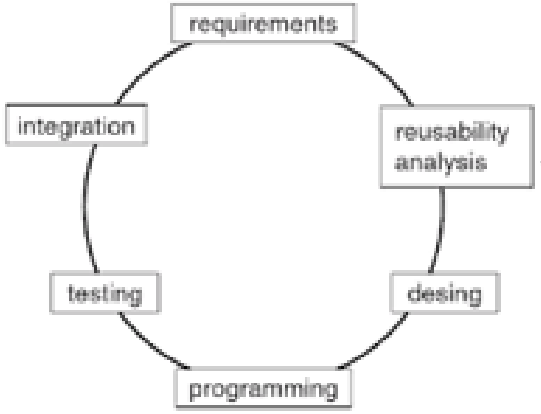






Концепция DW 2.0 [Development Cycle]



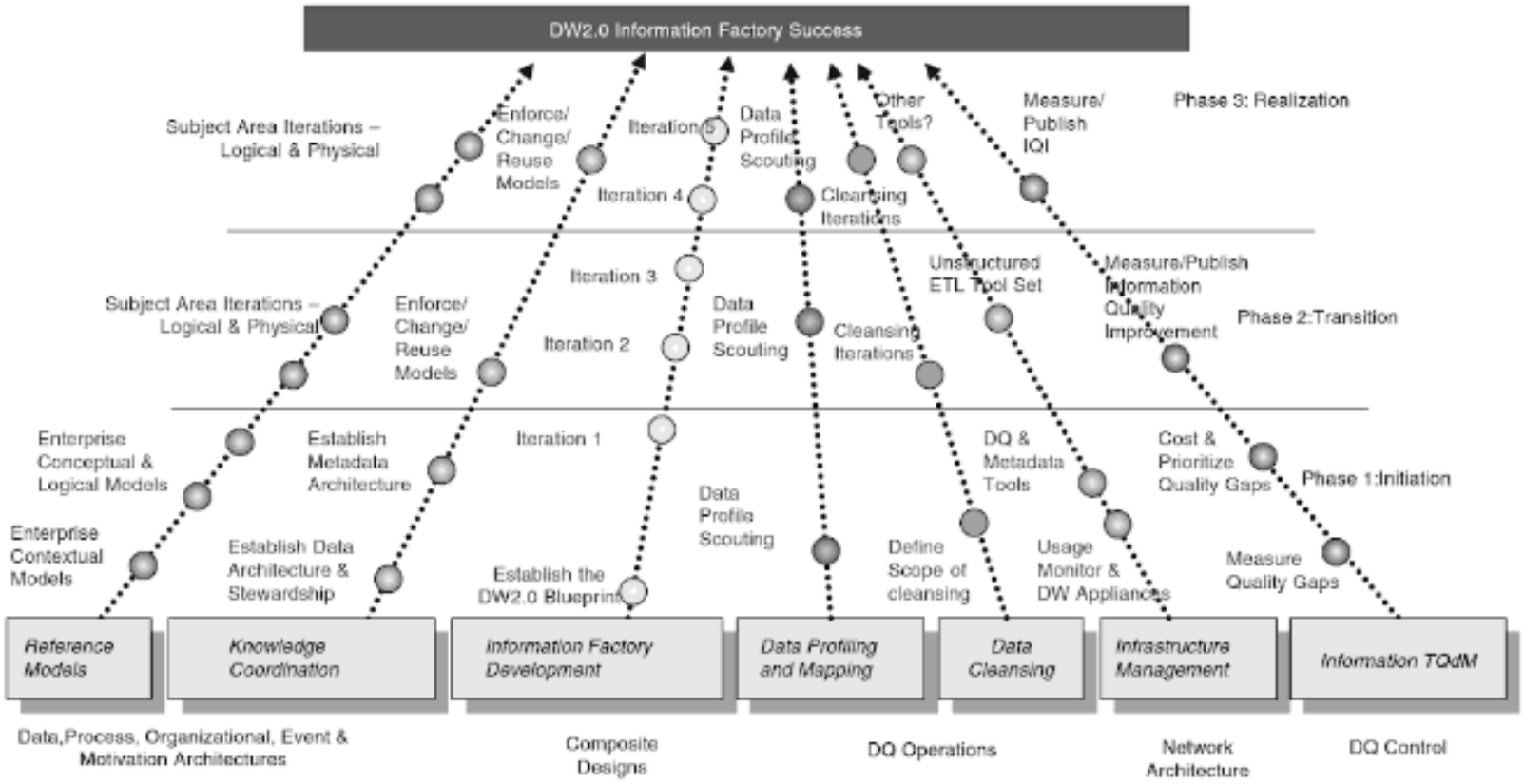


System Development Life Cycle (SDLS)	Cycle Development Life System (CDLS)
Оперативные системы	Аналитические системы и хранилища
Длительные этапы проекта	
Требования полностью понятны перед началом работ	Заранее не известны все требования. Развитие системы происходит в режиме «открытий».
Изменение системы после deploy – ошибка системы	Изменение системы после deploy – новая итерация развития

CASE world is dominated by requirements-driven analysis. Trying to apply CASE tools and techniques to the world of the data warehouse is not advisable, and vice versa. (B. Inmon Building the Data Warehouse Third Edition)

Концепция DW 2.0
[7 streams DW\BI projects]

#	Название активности	Единица измерения
1	Reference Data Models	Subject
2	Knowledge Coordination	Artifact
3	Information Factory Development	Topic
4	Data Profiling	Source
5	Data Correction	Attribute
6	Infrastructure management	Component
7	Information quality management	Process



Концепция DW 2.0
[checklist проектирования корпоративного хранилища]

1. Определить сущности для добавления\изменения модели
2. Определить регламент частоты выгрузок данных из систем источников и частота актуализации сущностей хранилища
3. Определить глубину хранения данных сущностей хранилища
4. Определить способ масштабирования системы (индексирование, хеширование, партиционирование и т.д.)
5. Определить будут ли агрегированные данные выноситься в отдельные области хранилища
6. Определить требования к безопасности данных и аудиту использования данных
7. Определить возможно ли обновление и удаление данных (каким образом и в каких случаях)
8. Определить будут ли выгружаться данные в другие оперативные или аналитические системы (каким образом и в каких случаях)
9. Определить прирост данных
10. Определить как пользователи будут информироваться об изменениях

Концепция Dimensional data warehouse
[1996-2013 книга + несколько дополнений Ральфа Кимбалла]

«Хранилище данных – это копия транзакционных данных, специально структурированных для запроса и анализа»

- Тактический характер концепции.
- Фиксация на решении конкретной бизнес задачи.
- Подход сосредоточен на витринах данных в виде многомерной модели.

Преимущества:

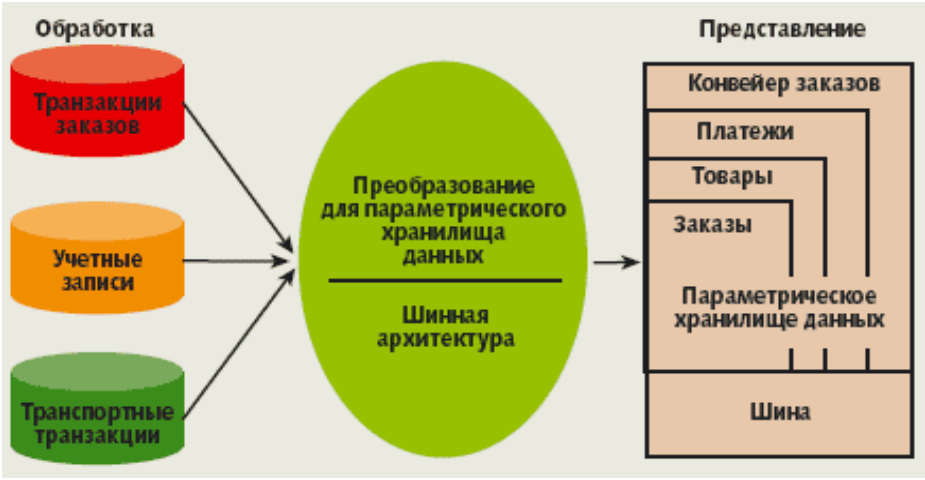
- скорость разработки
- простота и прозрачность модели данных

Недостатки:

- отсутствие корпоративной направленности хранилища данных
- сложность поддержки изменений модели данных

Enterprise data warehouse (Inmon B.)

Dimensional data warehouse (Kimball R.)

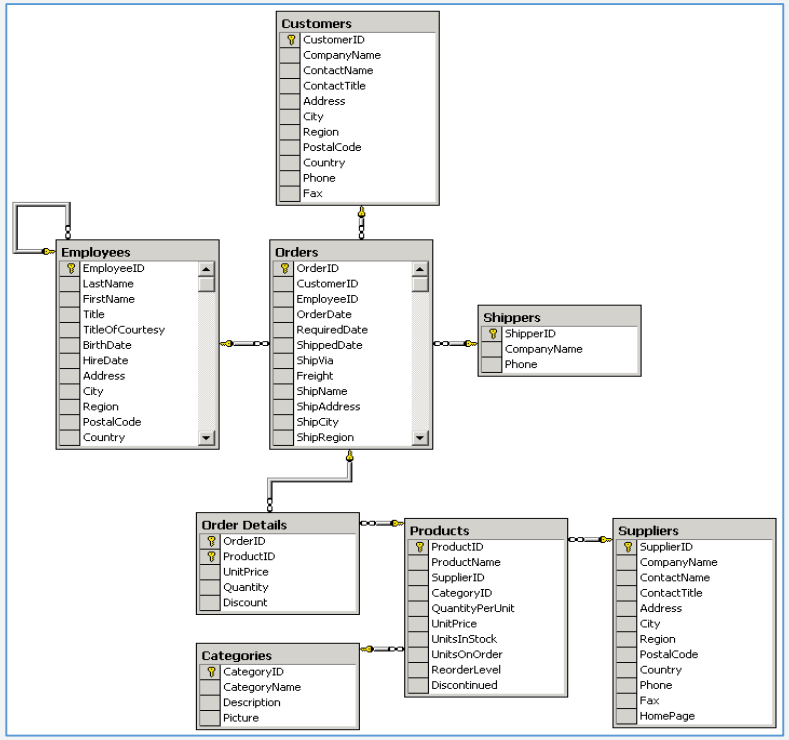


Отличие хранилищ данных от оперативных систем учета организации

Назначение - оперативные системы учета предназначены для повседневной работы пользователей организации с объектами хозяйственной деятельности.

Режим обновления данных - данные изменяются по закрытию транзакции в операции хоз. деятельности

Модель - нормализованная



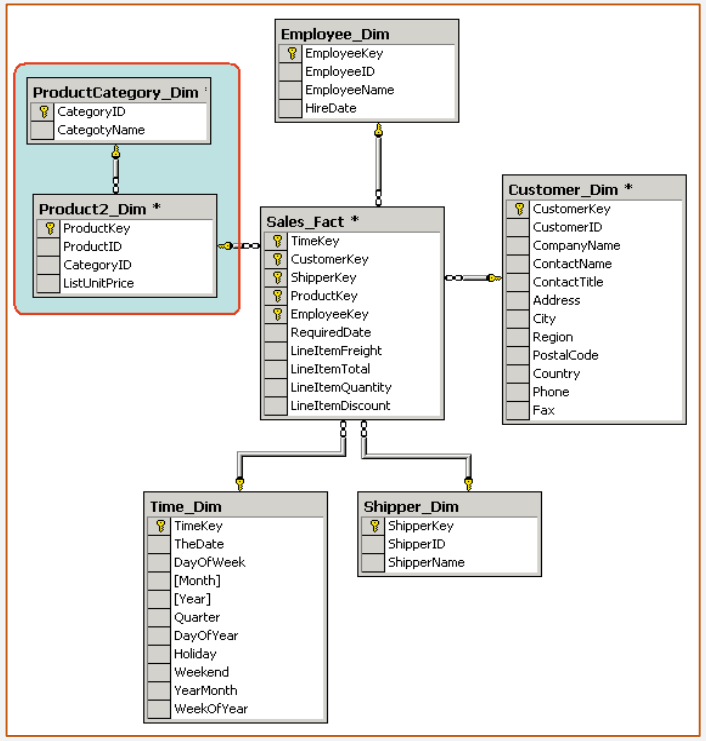
Нормализованная модель данных OLTP систем

Отличие хранилищ данных от оперативных систем учета организации

Назначение – хранилища данных предназначены для принятия решений

Режим обновления данных – данные изменяются по расписанию

Модель – денормализованная



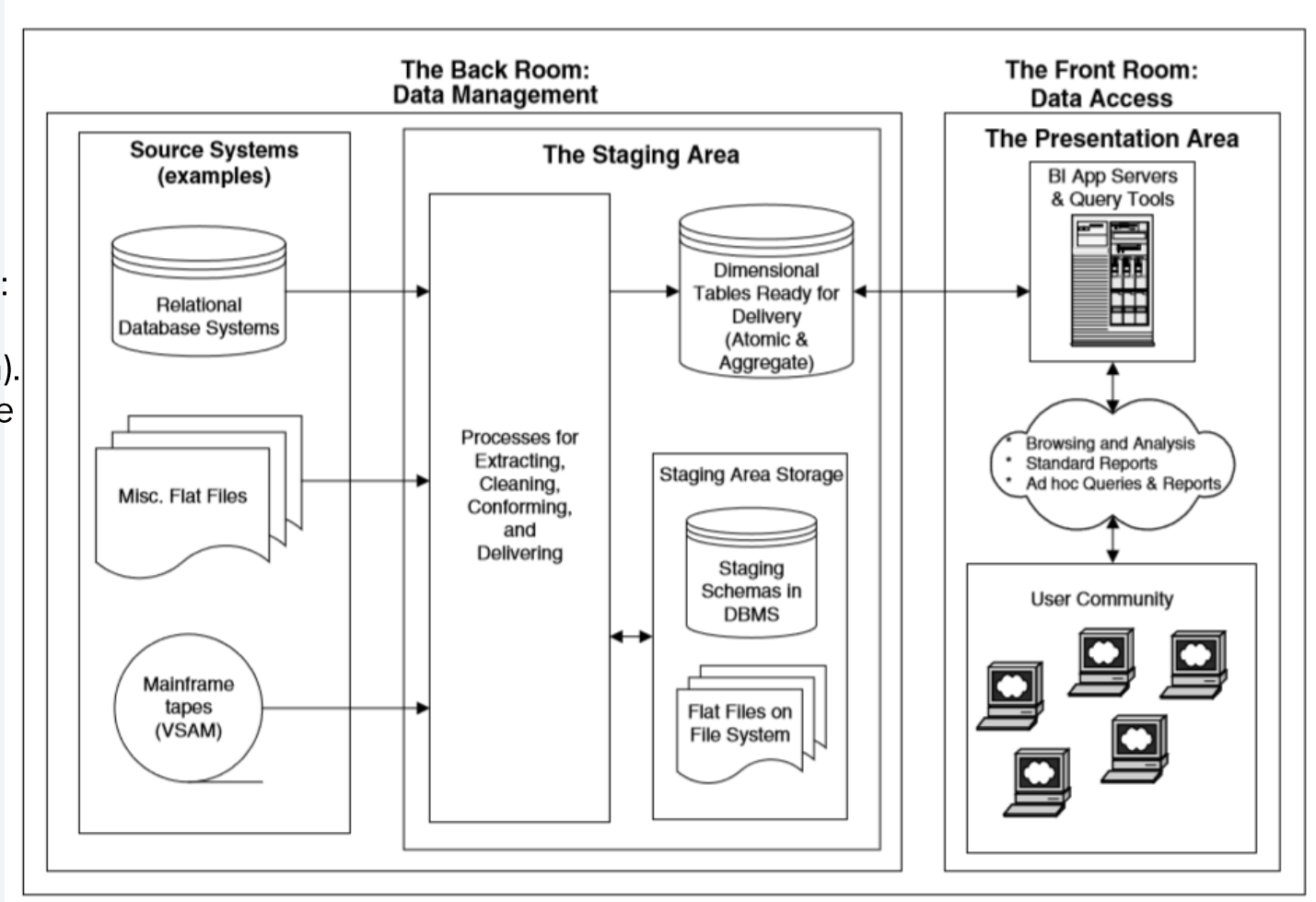
Денормализованная модель DWH систем

Области данных хранилища

Хранилища данных имеют сложную внутреннюю структуру процессов управления данными, которую можно разбить на 2-е основные составляющие :

- Внутренняя область управления данными хранилища (**The Back Room**). Данные предоставлены в сыром виде во множестве промежуточных слоев хранения : Operative data of legacy systems(ODS)/Staging Area (ETL)/Master data model of business entries/Metadata model

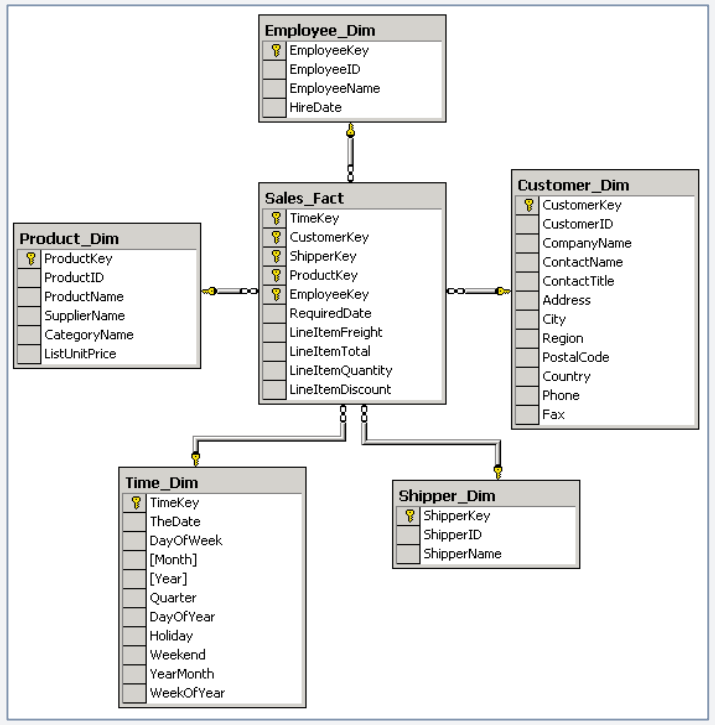
- Область доступных данных или презентационная область (**The Front Room**). Витрины и гиперкубы OLAP



Проектирование модели данных (Dimensional Modeling)
Таблицы измерений. Иерархия измерений

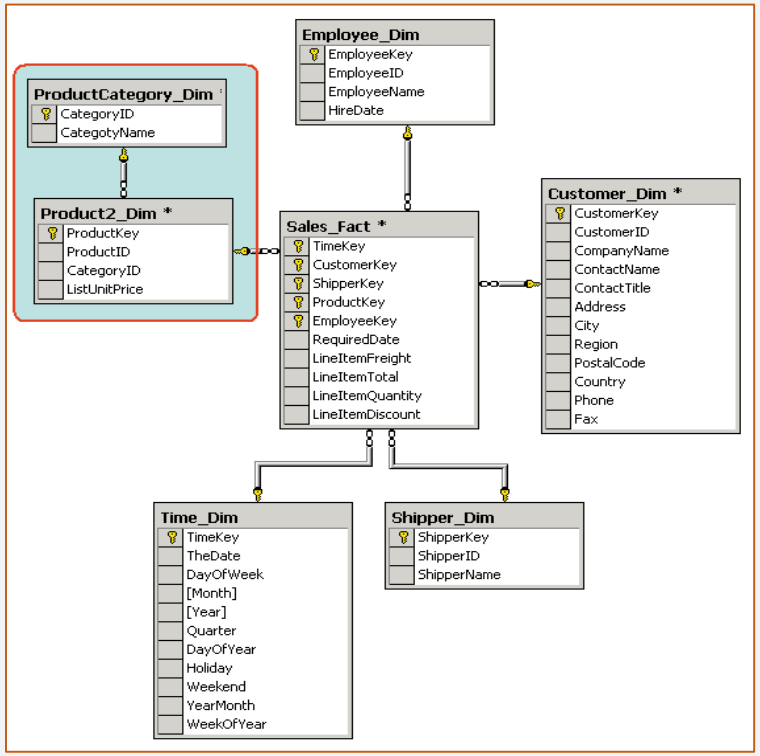
Dimensional Modeling – одна диаграмма сущности-отношения (ER-model) разбивается на несколько моделей связей таблиц фактов и измерений.

- **Таблицы измерений** – контекст для таблиц фактов использующийся для выбора и агрегирования данных на требуемом уровне детализации.
- **Измерения могут быть организованы в иерархию**, состоящую из нескольких уровней, каждый из которых представляет уровень детализации, требуемый для соответствующего анализа.



Модель данных (звезда)

Если же хотя бы одно измерение содержится в нескольких связанных таблицах, такая схема хранилища данных носит название **«снежинка»** (snowflake schema). Дополнительные таблицы измерений в такой схеме, обычно соответствующие верхним уровням иерархии измерения и находящиеся в соотношении «один ко многим» в главной таблице измерений, соответствующей нижнему уровню иерархии, иногда называют консольными таблицами (outrigger table).

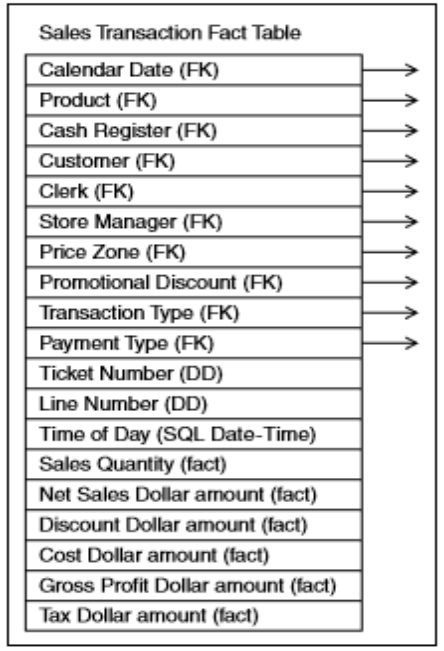


Модель данных (снежинка)

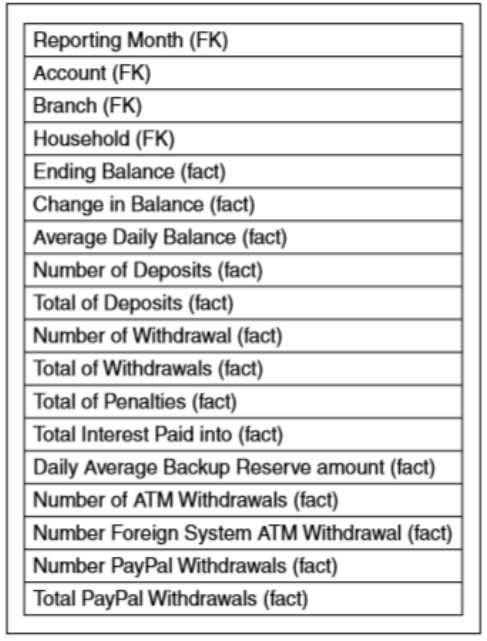
Проектирование модели данных (Dimensional Modeling)
Таблицы фактов

Таблица фактов является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться.

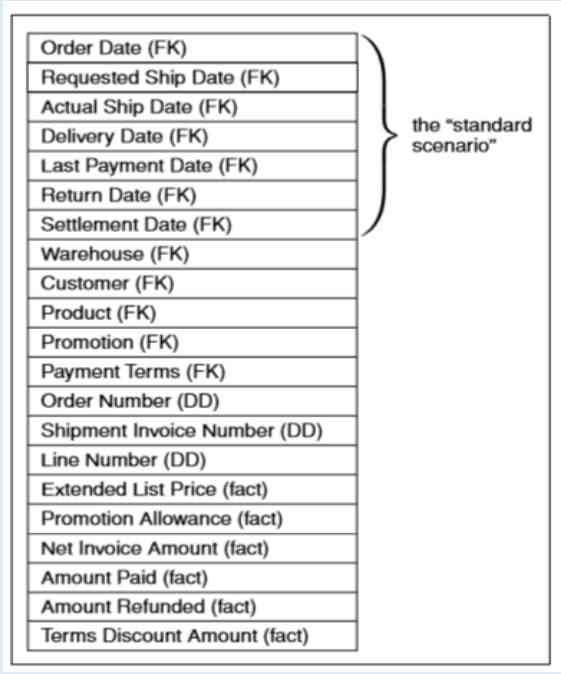
Факты событий (Transaction facts). Пример: телефонный звонок, снятие денег со счета, продажа товара



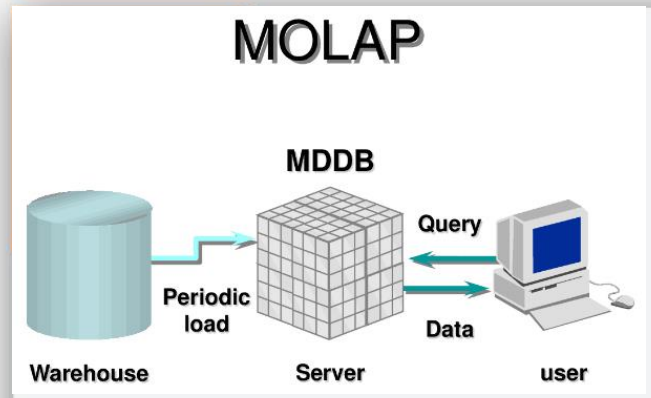
Факты снимков состояний объектов за регулярно повторяемый период времени (Periodic Snapshot facts). Подходит для отслеживания долгосрочных процессов, таких как банковские счета и другие формы финансовой отчетности. Обычно периодом таких снимков является – месяц.



Факты итоговых снимков состояний (Accumulating Snapshot Fact Tables) - используется для описания процессов, которые имеют определенное начало и конец, таких как выполнение заказа, обработка заявок и большинство рабочих процессов. Накопленный снимок не подходит для длительных непрерывных процессов, таких как отслеживание банковских счетов или описание непрерывных производственных процессов.



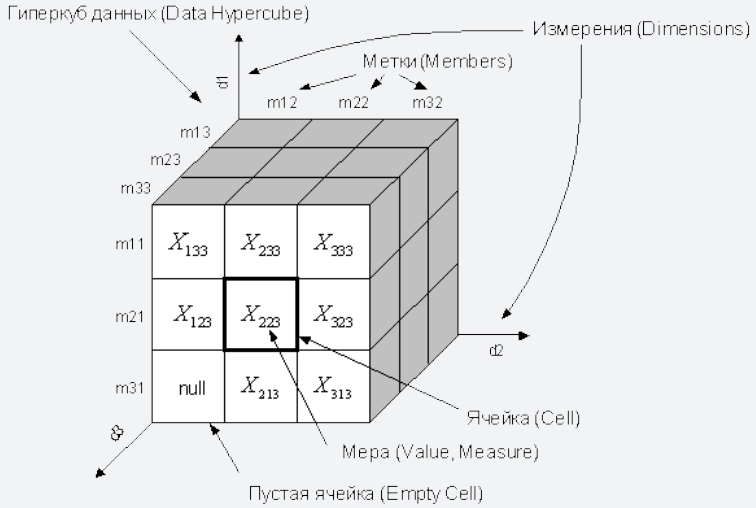
Многомерные базы данных (MOLAP). Модель гиперкуба.



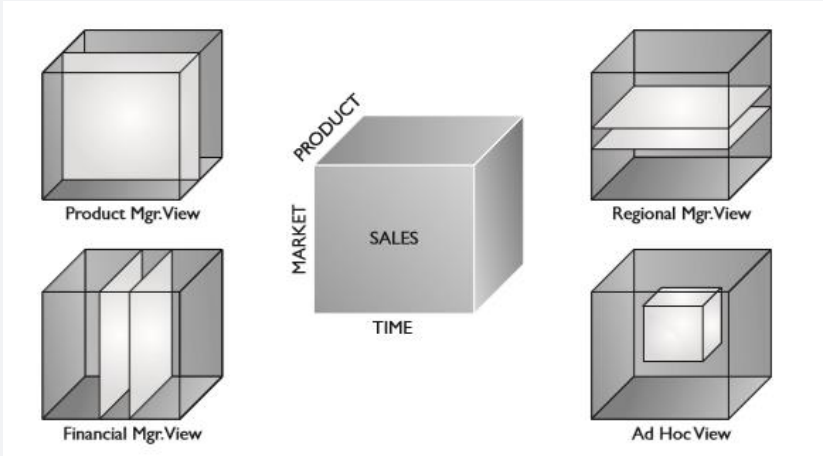
Многомерные базы данных (MDB) рассматривают данные как кубы, которые являются обобщением электронных таблиц на любое число измерений. Кроме того, кубы поддерживают иерархию измерений и формул без дублирования их определений. Набор соответствующих кубов составляет многомерную базу данных (или хранилище данных). Кубами легко управлять, добавляя новые значения измерений. Кроме того такие системы баз данных позволяют производить анализ по сложным иерархиям измерений.

Основными понятиями многомерной модели данных являются:

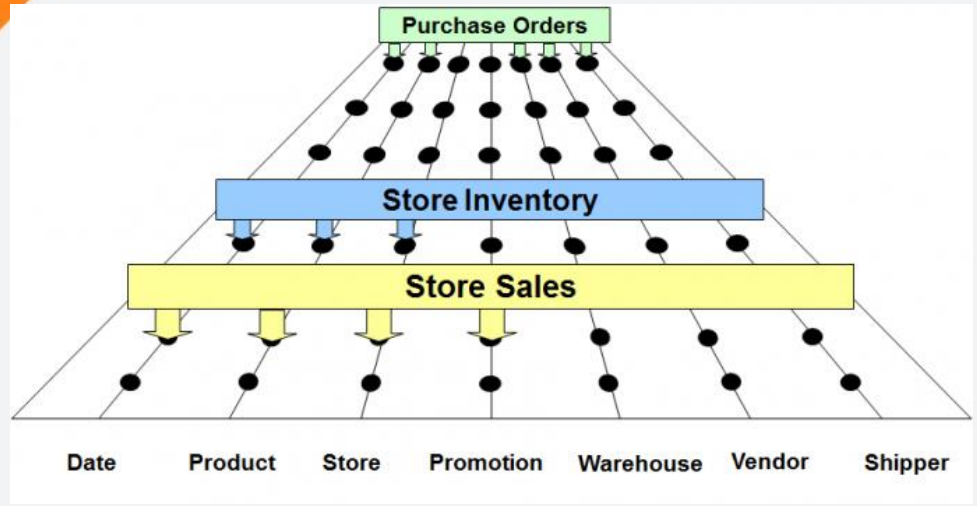
- Гиперкуб данных (Data Hypercube)
- Измерение (Dimension)
- Метки (Memders)
- Ячейка (Cell)
- Мера (Measure)



Многомерные базы данных позволяют производить быстрые срезы, фильтры и агрегации над данными гиперкуба.



Проектирование модели данных (Data Warehouse Bus Architecture). Шина данных хранилища.



BUSINESS PROCESSES	COMMON DIMENSIONS: Date	Product	Warehouse	Store	Promotion	Customer	Employee
Issue Purchase Orders	X	X	X				
Receive Warehouse Deliveries	X	X	X				X
Warehouse Inventory	X	X	X				
Receive Store Deliveries	X	X	X	X			X
Store Inventory	X	X		X			
Retail Sales	X	X		X	X	X	X
Retail Sales Forecast	X	X		X			
Retail Promotion Tracking	X	X		X	X		
Customer Returns	X	X		X	X	X	X
Returns to Vendor	X	X		X			X
Frequent Shopper Sign-Ups	X			X		X	X

Шина данных – архитектура совместного использования измерений витринами данных.

Матрица шины помогает архитектору визуализировать, какие измерения совместно используются или соответствуют различным витринам данных в хранилище данных.

Правила многомерного моделирования R. Kimball.

- Загружать в многомерные структуры детальные данные.
- Строить многомерную модель вокруг производственных бизнес-процессов.
- Таблица фактов должна иметь связь с таблицей размерностей времени.
- Размещать в таблицах фактов данные одинаковой гранулярности или уровня детализации.
- Избавляться в таблицах фактов от отношений многие-ко-многим.
- Избавляться от отношений многие-к-одному между таблицами размерностей.
- Хранить в таблицах размерностей заголовки для отчётов и области значений для фильтров.
- Использовать в таблицах размерностей суррогатный ключ.
- Создавать согласованные размерности для интеграции данных в масштабах предприятия.
- Непрерывно соотносить требования с действительностью, чтобы предоставить бизнес-пользователями подходящий DW/BI-инструмент, помогающий им принимать решения.

Сравнение концепций по характеристикам

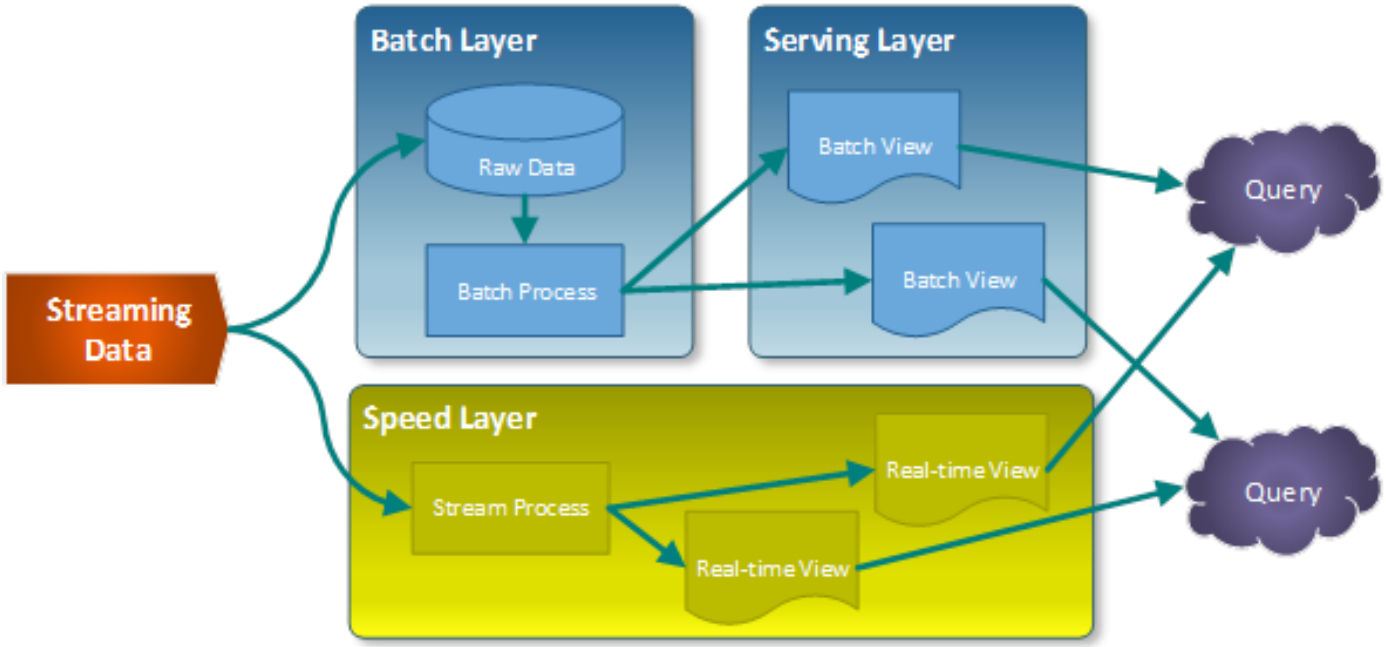
Характеристики	**Dimensional data warehouse** *(*Ральф Кимболл / Ralph Kimball)	**Enterprise data warehouse (**+CIF +DW 2.0) (Билл Инмон / Bill Inmon)
Поддержка принятия решения	Тактическая	Стратегическая
Требования к объединению данных	Индивидуальные бизнес требования	Объединение на базе единого широкого набора измерений и их элементов
Структура данных	KPI, измерения эффективности бизнеса, scorecards	Различные подходы к организации данных в зависимости от бизнес потребностей. Как в виде объектов описывающих предметную область и имеющие сложную систему взаимосвязей, так и объекты выстроенные в виде системы показателей и измерений.
Применимость к изменчивости данных источников	Достаточно стабильные источники	Большое количество изменений на источниках
Требования к размеру компетентной команды	Небольшая команда	Большая команда
Время получения результата	Нужен быстрый результат, пара месяцев	Можно потратить пол года-год для получения результата
Стоимость построения	Низкие первоначальные затраты	Высокие первоначальные затраты

λ архитектура [04/2014]

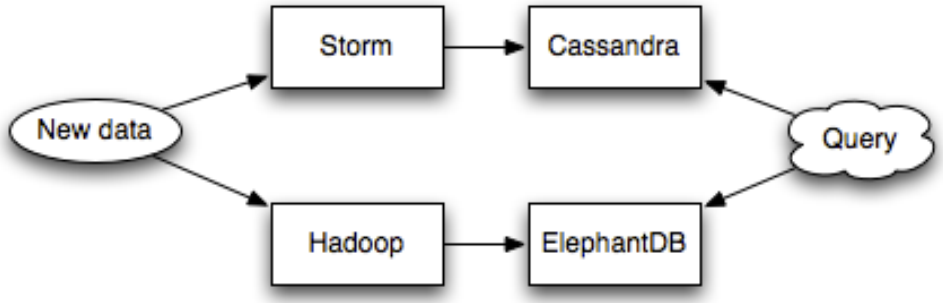
Лямбда (λ) архитектура – концепция сочетания Big Data и Real-time аналитики, которая может быть применена к хранилищам данных

Имеет структуру, состоящую из трех уровней:

- Пакетный уровень – архив сырых исторических данных
- Сервисный уровень - индексирует пакеты и обрабатывает результаты вычислений
- Уровень ускорения - отвечает за обработку данных, поступающих в систему в реальном времени



Пример

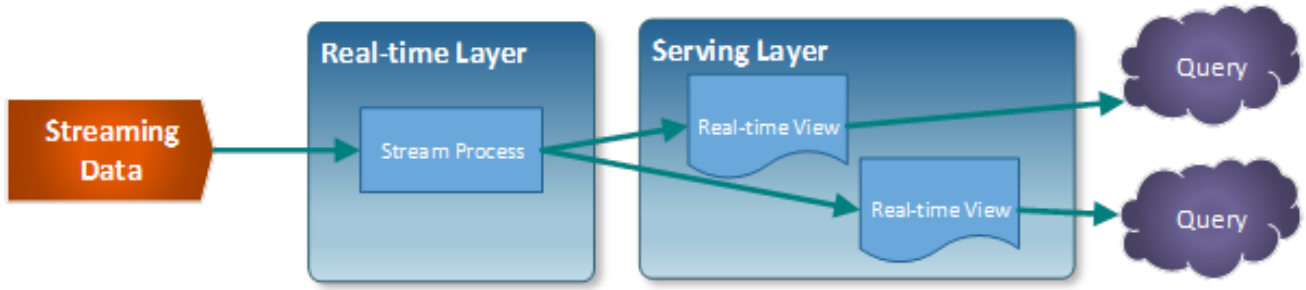


Kappa архитектура [07/2014]

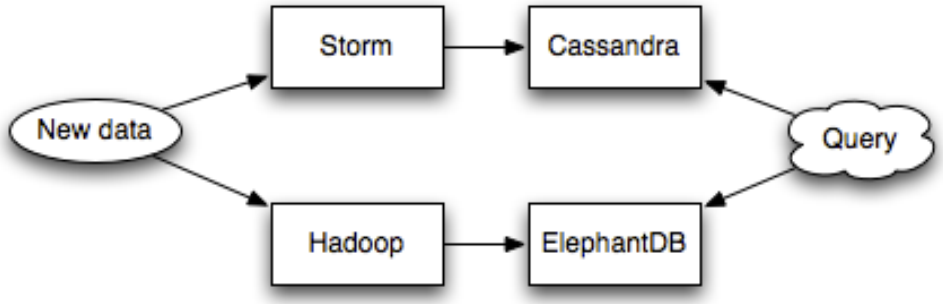
Kappa архитектура – концепция поточной обработки данных в реальном времени и постоянная повторная обработка на одном движке.

Имеет структуру, состоящую из двух уровней:

- Сервисный уровень
- Уровень обработки в реальном времени



Пример



Ростелеком

Проблематика построения хранилищ данных.

- Своевременный переход от QuickWin решения к Enterprise
- Требования к ускорению частоты обновления данных (от batch по регламенту до real time)
- Проектирование масштабируемой модели
- Обследование систем источников (сбор требований при подключении системы и профилирование)
- Изменяющиеся бизнес-процессы
- Изменяющиеся методики расчета показателей отчетности
- Некачественное документирование модели данных и алгоритмов преобразований
- Физическая модель данных должна учитывать специфику платформы хранения и обработки данных

E-mail:

datatalks@rt.ru

Адрес:

Москва, Сущевский вал 26

Ростелеком