Data Warehouse Analyst

Принципы построения DWH

otus.ru

Проверить, идет ли запись

Меня хорошо видно && слышно?

Ставим “+”, если все хорошо
“-”, если есть проблемы

Тема вебинара

Принципы построения DWH

Водовозова Татьяна

Руководитель направления аналитики DWH

Об опыте:

- участвовала в проектах развития хранилищ данных,
- построение хранилищ с нуля,
- миграция данных
- ментор

Правила вебинара

Активно
участвуем

Off-topic обсуждаем
в Telegram @**OTUS DWH-2025-04**

Задаем вопрос
в чат или голосом

Вопросы вижу в чате,
могу ответить не сразу

Условные обозначения

Индивидуально

Время, необходимое
на активность

Пишем в чат

Говорим голосом

Документ

Ответьте себе или
задайте вопрос

Карта курса


1. Введение в DWH Analytics
2. Хранение и трансформация данных
3. Загрузка данных
4. Аналитика и Business Intelligence
5. Оптимизация и продвинутые темы
6. Проектная работа

Цели вебинара

К концу занятия вы сможете

1. Изучить основные концепции в построении Хранилища Данных
2. Разобрать основные понятия, относящиеся к DWH
3. Выявить best practices в организации хранилища данных
4. Моделировать хранилища на примерах из реальной жизни

Смысл

Зачем вам это уметь

1. Узнать основные концепции построения DWH
2. Узнать лучшие практики по построению хранилищ данных
3. Узнать о методах используемых при построении хранилищ данных

Строительные блоки DWH

Шаги проектирования DWH

- Определить какую задачу надо решить
- Проанализировать наличие и качество данных
- Спроектировать модель данных для каждого слоя

Требования при проектировании DWH

Требования к хранилищу:

- Гетерогенность источников
- Поддержка историчности
- Гибкость модели данных
- Частота обновления
- Устойчивость к объёму

Типовые кейсы:

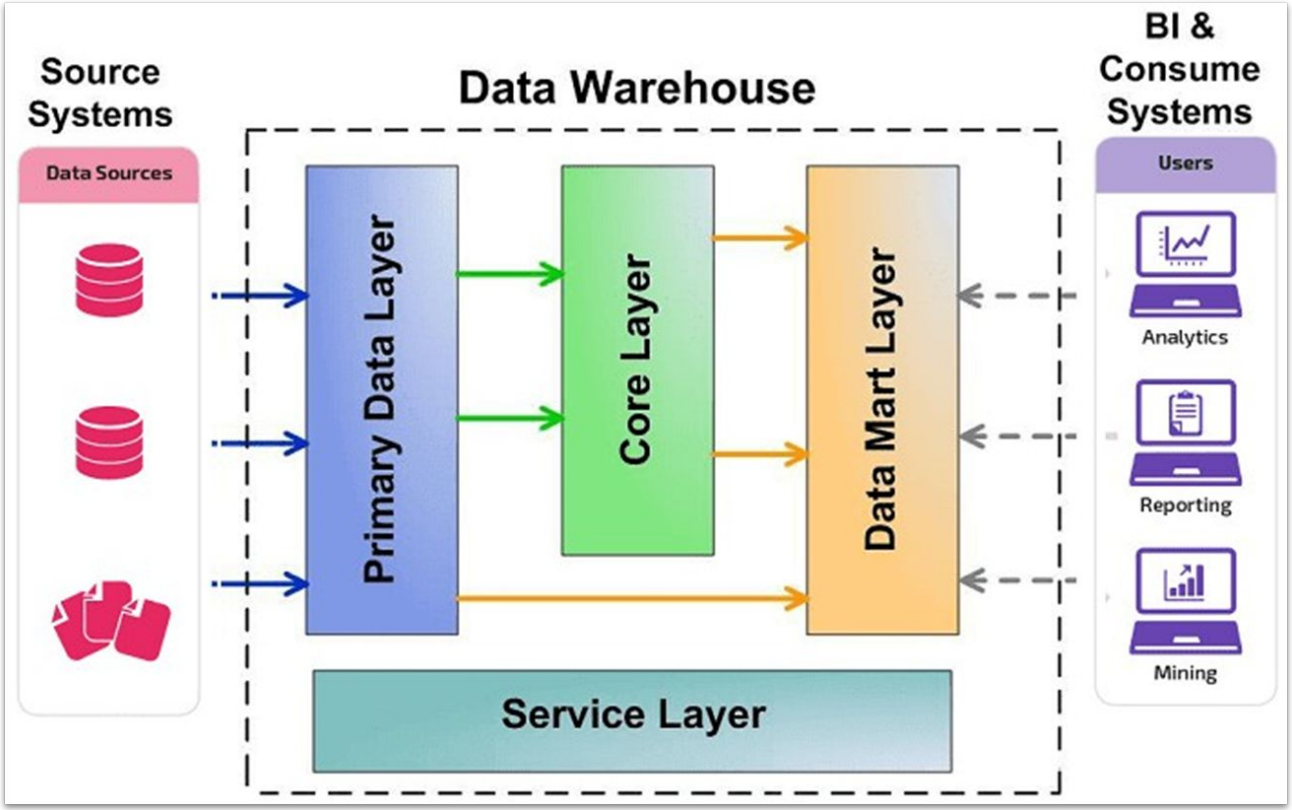
- Добавление новых данных (источников) в существующую модель
- Отработка изменений на источниках
- Создание пользовательских представлений и их сопровождение

Абстрактно. DWH

Stage + CORE + DataMarts

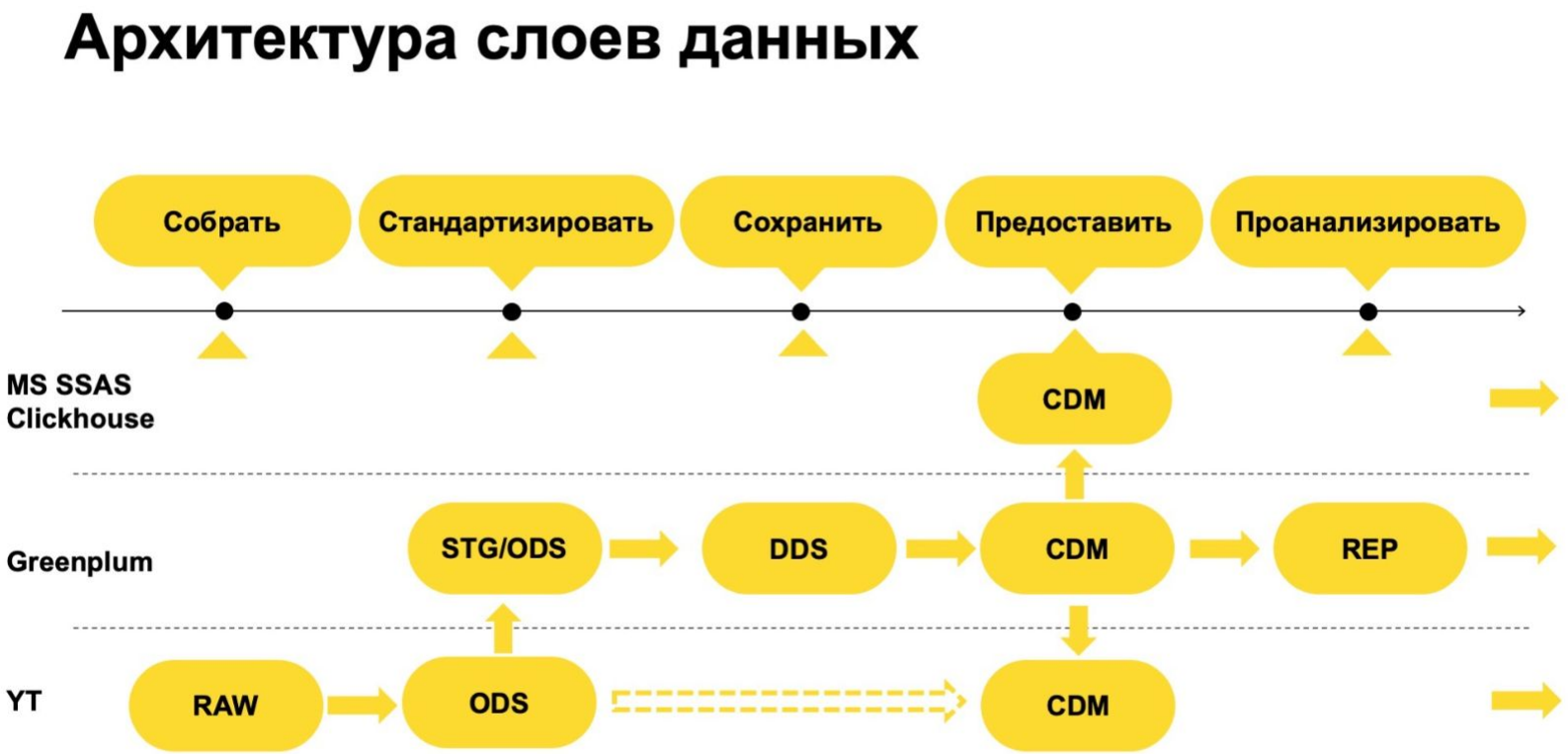
- **Stage –** приземляем данные as is. Нужен для разгрузки источников без трансформаций, если их нельзя делать на ходу
- **CORE –** это то, что мы будем разбирать. Это способ моделирования данных в хранилище, который отвечает нашим целям
- **DataMarts –** слой представления данных для пользователей

Логические слои DWH



Пример организации слоев

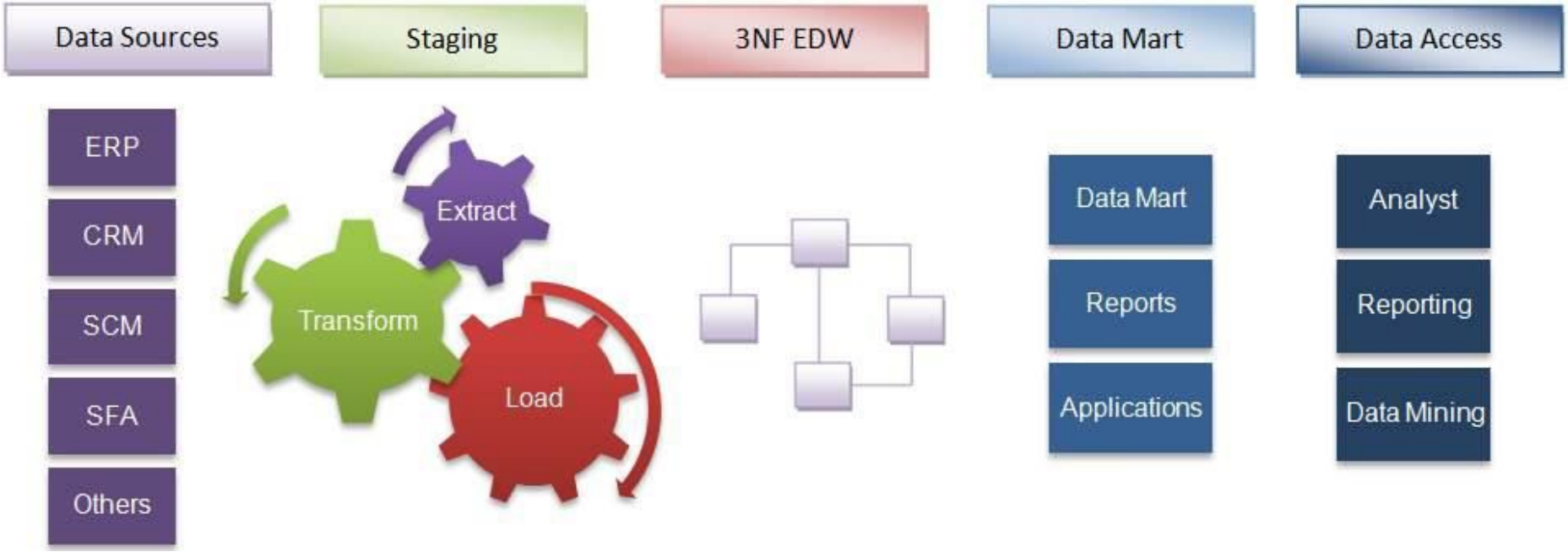
Пример организации слоев



Популярные подходы

- **Dimensional Modeling**
    - **Bill Inmon:** сверху вниз (top-down), нормализация, единое целое, для всего предприятия
    - **Ralph Kimball: снизу вверх** (bottom-up), денормализация, процессно-ориентированный
- **Data Vault**
- **Anchor Modelling**

Подход Bill Inmon



Подход Ralph Kimball

Data Sources

Staging

Star Schema

Data Access

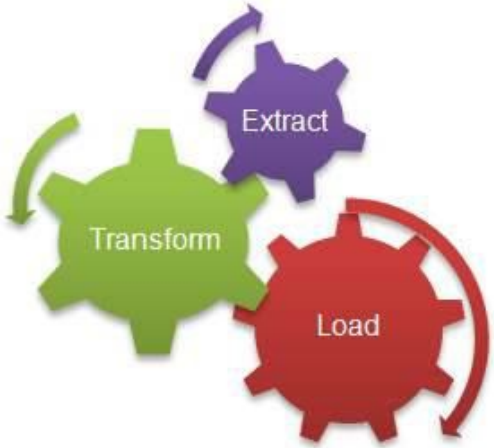
ERP

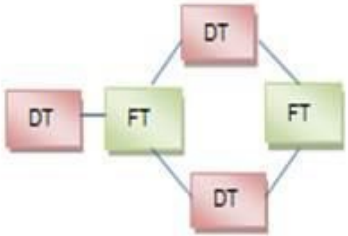
CRM

SCM

SFA

Others





Analyst

Reporting

Data Mining

DT – Dimension Table
FT – Fact Table

Inmon vs Kimball

Характеристики	Inmon	Kimball
Требования к области данных	Уровень предприятия.	Определенная бизнес область.
Структура данных	Данные никак не ранжированы, их можно применять под разные нужды.	Бизнес метрики, показатели эффективности, рейтинг.
Масштабируемость	Рост скоупа и запросы на изменения являются критичными.	Нужно адаптировать под изменения в ограниченном скоупе.
Неизменность данных	Высокий процент изменений в системах источниках.	Системы источники относительно неизменны.
Требования к команде	Большая команда или несколько команд сильных специалистов.	Небольшая команда средней квалификации.
Время на поставку	Требования компании позволяют долгое время на разработку.	Есть срочная необходимость в первой версии хранилища.
Стоимость разворачивания	Высокая стоимость первого развертывания, с более низкой стоимостью следующих итераций.	Невысокая стоимость первого развертывания, каждая следующая итерация стоит примерно столько же.

Практика

Для каждого случая выбрать поход Инмон или Кимбал

- Страхование. Бизнесу необходимо видеть общую картину по компании в отношении, клиентов, продуктов, как с течением времени менялись продукты клиентов, историю страховых случаев, что на них влияет, выполняемость планов агентов
- Отдел маркетинга. Для нужд маркетинга надо разработать быстро возможность получения отчетов о результатах маркетинговых акций.
- CRM в банке. Фокус на продажах продуктов, сопутствующих продажах, продаже дополнительного товара

5 - 7 минут

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Модели данных

3 нормальная форма

В третьей нормальной форме каждый неключевой атрибут зависит от ключа, причем от всего ключа целиком и ни от чего другого, кроме как от ключа.

id (PK)	last_ name	first_ name	second_name	position_id	salary
1	Иванов	Иван	Иванович	1001	15000
2	Петров	Юрий	Николаевич	1002	25000
3	Никитина	Мария	Ивановна	1003	30000
4	Сидорова	Анна	Дмитриевна	1002	25000

3 нормальная форма

id (PK)	last_ name	first_ name	second_name	position_id	salary
1	Иванов	Иван	Иванович	1001	15000
2	Петров	Юрий	Николаевич	1002	25000
3	Никитина	Мария	Ивановна	1003	30000
4	Сидорова	Анна	Дмитриевна	1002	25000

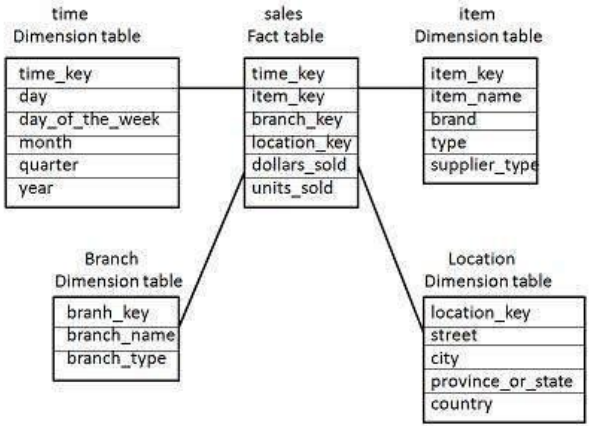
id (PK)	last_ name	first_ name	second_na me	position _id
1	Иванов	Иван	Иванович	1001
2	Петров	Юрий	Николаевич	1002
3	Никитина	Мария	Ивановна	1003
4	Сидорова	Анна	Дмитриевна	1002

position_id	salary
1001	15000
1002	25000
1003	30000

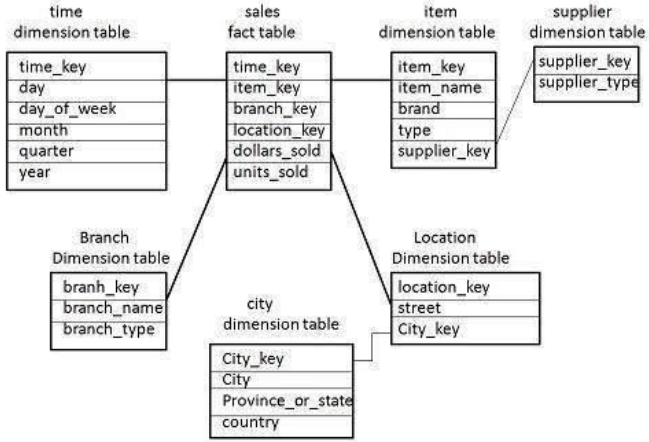
Dimensional Model

- Выделить бизнес процесс
- Определить гранулярность данных в фактовой таблице
- Определить измерения - в каких разрезах надо считать
- Что считаем - сумма, количество, факт события

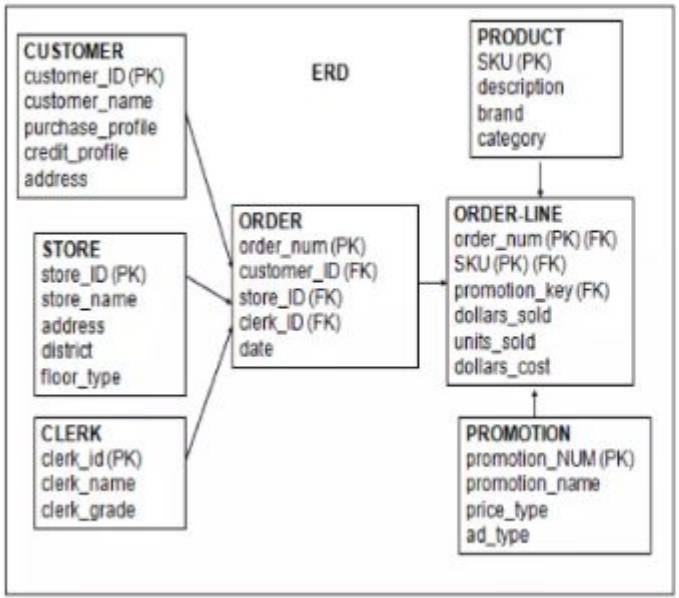
Звезда

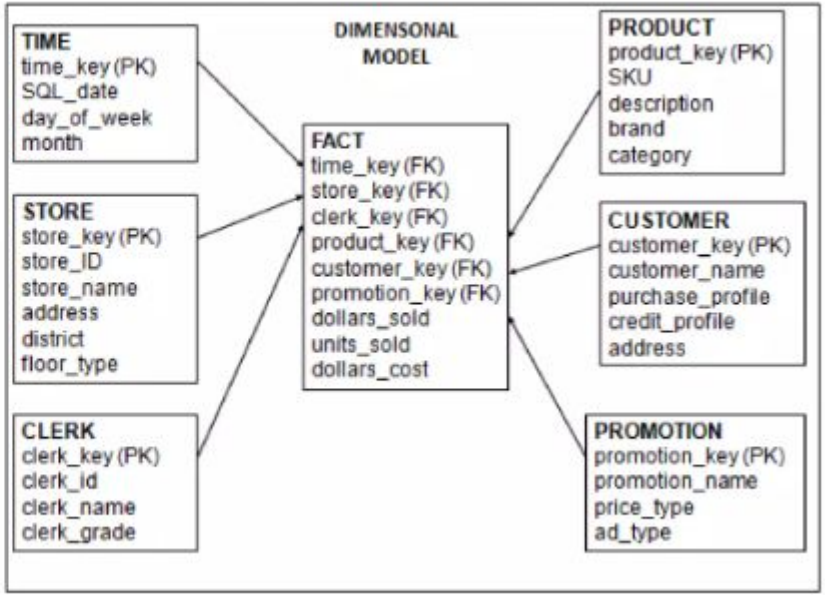


Снежинка



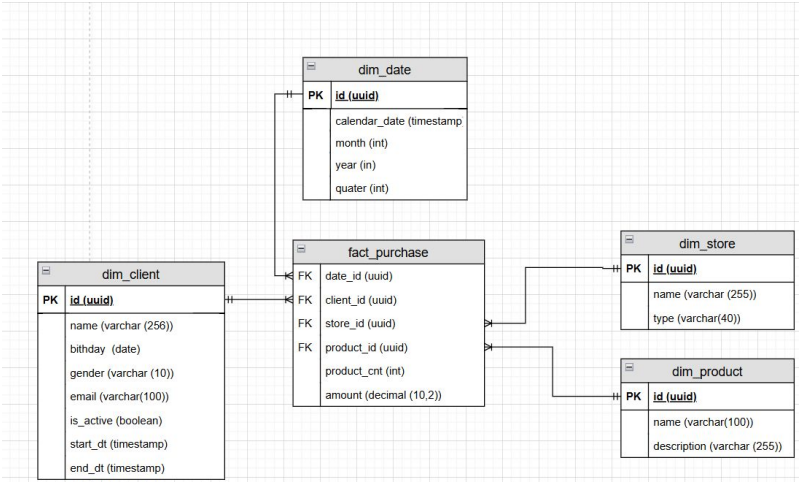
3 NF & Dimensional Model





Факт

- Действие, транзакция
- Имеет ссылку на данные в измерениях (внешний ключ)
- Имеет дату



Типы фактов

- **Транзакции / transaction fact table** - одна строка представляет собой измеряемое событие, произошедшее в конкретный момент времени (например, факт покупки)

Transactional Fact Table

Transaction ID	Customer ID	Product ID	Quantity Sold	Sales Amount
T101	C2001	P51231	2	250
T102	C2002	P51731	5	430
T103	C2002	P51231	6	750
T104	C2003	P59823	3	210
T105	C2006	P54213	1	75

Типы фактов

- **Периодическая фиксация фактов / Periodic snapshot fact tables** - одна строка представляет собой группу событий, измеренных в конкретный момент времени. Если в этот момент событий нет, строка все равно будет занесена в таблицу со значениями null (например, количество продукции на складе по конкретному товару, месяцу и магазину, выручка по клиенту за месяц)

Periodic Snapshot Fact Table

Month ID	Total Sales	Revenue	no of customers
M202301	2340	1420035	450
M202302	3160	3540020	390
M202303	2530	2104240	456
M202304	4302	4570126	523
M202305	51021	5120356	621

Типы фактов

- **Накапливающая фиксация фактов / Accumulating snapshot fact tables** - одна строка представляет цикл жизни короткого процесса. Если процесс предполагает три этапа, то в таблице будут три столбца с датой прохождения соответствующего этапа (например, логистический процесс)

Accumulating Snapshot Fact Table

Order ID	Order Date	Received Date	Stocked Date	Sold Date
O202301021001	2023-01-05	2023-01-08	2023-01-10	2023-01-15
O202301021002	2023-02-10	2023-02-12	2023-02-15	2023-02-20
O202301021003	2023-03-20	2023-03-22	2023-03-25	2023-03-30
O202301021004	2023-04-15	2023-04-18	2023-04-20	2023-04-25
O202301021005	2023-05-02	2023-05-05	2023-05-08	2023-05-15

Типы фактов

- **Factless fact tables** - неизмеряемое событие (например, совещание с информацией о дате, месте проведения и участниках)

Factless Fact Tables

Opportunity ID	Stage	Start Date	End Date
1001	Prospecting	01-01-2023	10-01-2023
1002	Qualification	05-01-2023	15-01-2023
1003	Negotiation	10-01-2023	20-01-2023
1004	Prospecting	15-01-2023	25-01-2023
1005	Qualification	20-01-2023	05-02-2023
1006	Negotiation	25-01-2023	10-02-2023
1007	Prospecting	01-02-2023	15-02-2023
1008	Qualification	05-02-2023	20-02-2023
1009	Negotiation	10-02-2023	28-02-2023

Типы фактов

- **Aggregated fact tables or cubes** - одна строка представляет агрегированную информацию о нескольких событиях. Это позволяет ускорить обработку аналитических запросов.

Детальный транзакционный факт

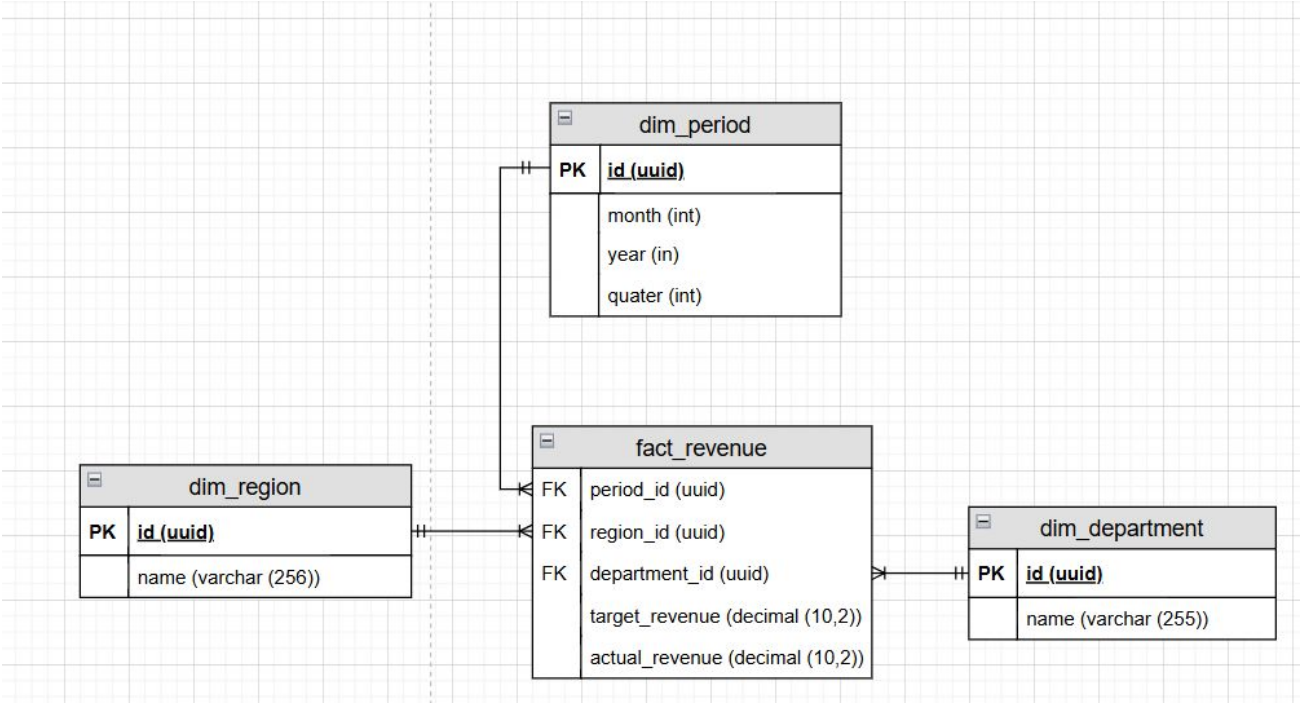
- **Date**
- Product ID
- Client ID
- Category ID
- Quantity Sold
- Sales Amount

Агрегированный факт

- **Month**
- **Product Category**
- Total Quantity Sold
- Total Sales Amount
- Average Sales Amount

Типы фактов

- **Consolidated fact tables -** Одна строка представляет несколько событий с одинаковой степенью гранулярности (например, планы и факты продаж).

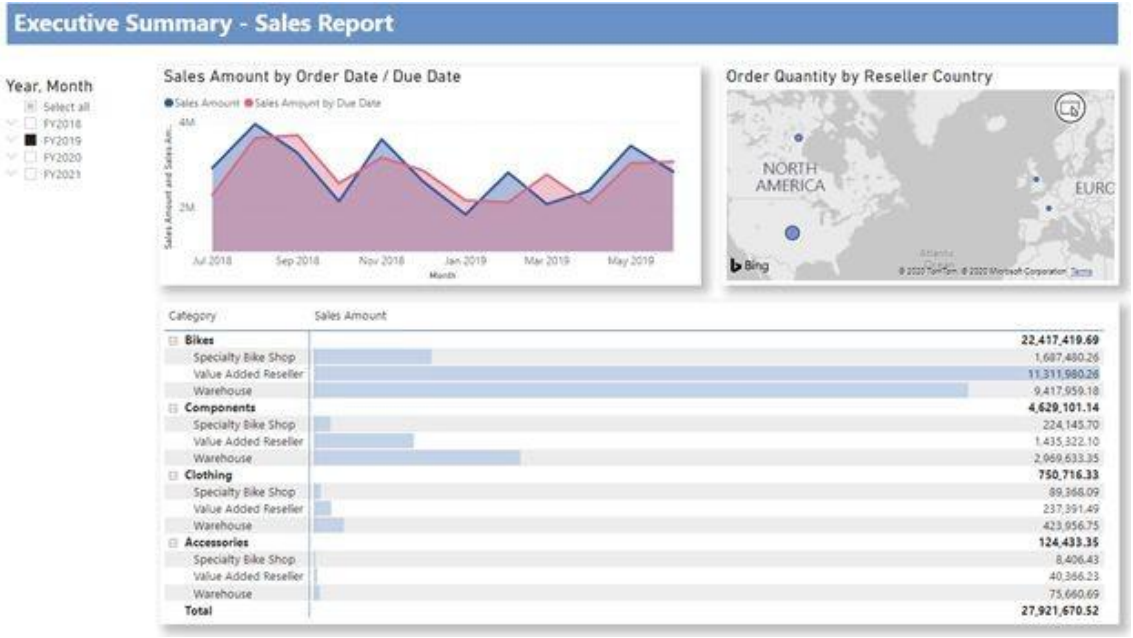


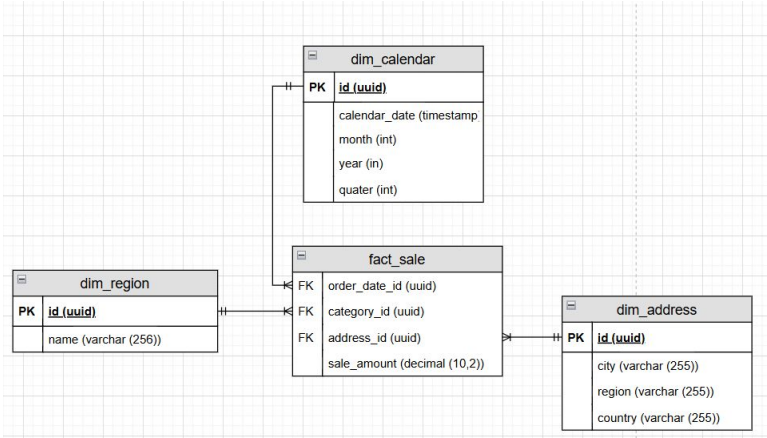
Измерения

- Отвечают на вопросы **Что? Где? Когда? Как? Почему?**
- Содержат атрибуты фактов, которые позволят их сгруппировать в инструментах визуализации данных

Измерения

- Отвечают на вопросы **Что? Где? Когда? Как? Почему?**
- Содержат атрибуты фактов, которые позволят их сгруппировать в инструментах визуализации данных





Типы измерений

- Calendar dimension - отсылает ко всем таблицам фактов, отражает временный аспект данных

Типы измерений

- Calendar dimension - Отсылает ко всем таблицам фактов, отражает временный аспект данных
- Role-playing dimension - Одна и та же таблица измерения может играть разную роль для одного и того же факта. Каждая роль - это внешний ключ в таблице фактов. Например, в банковской транзакции есть роль отправителя и получателя. Оба будут находиться в одной таблице измерений.

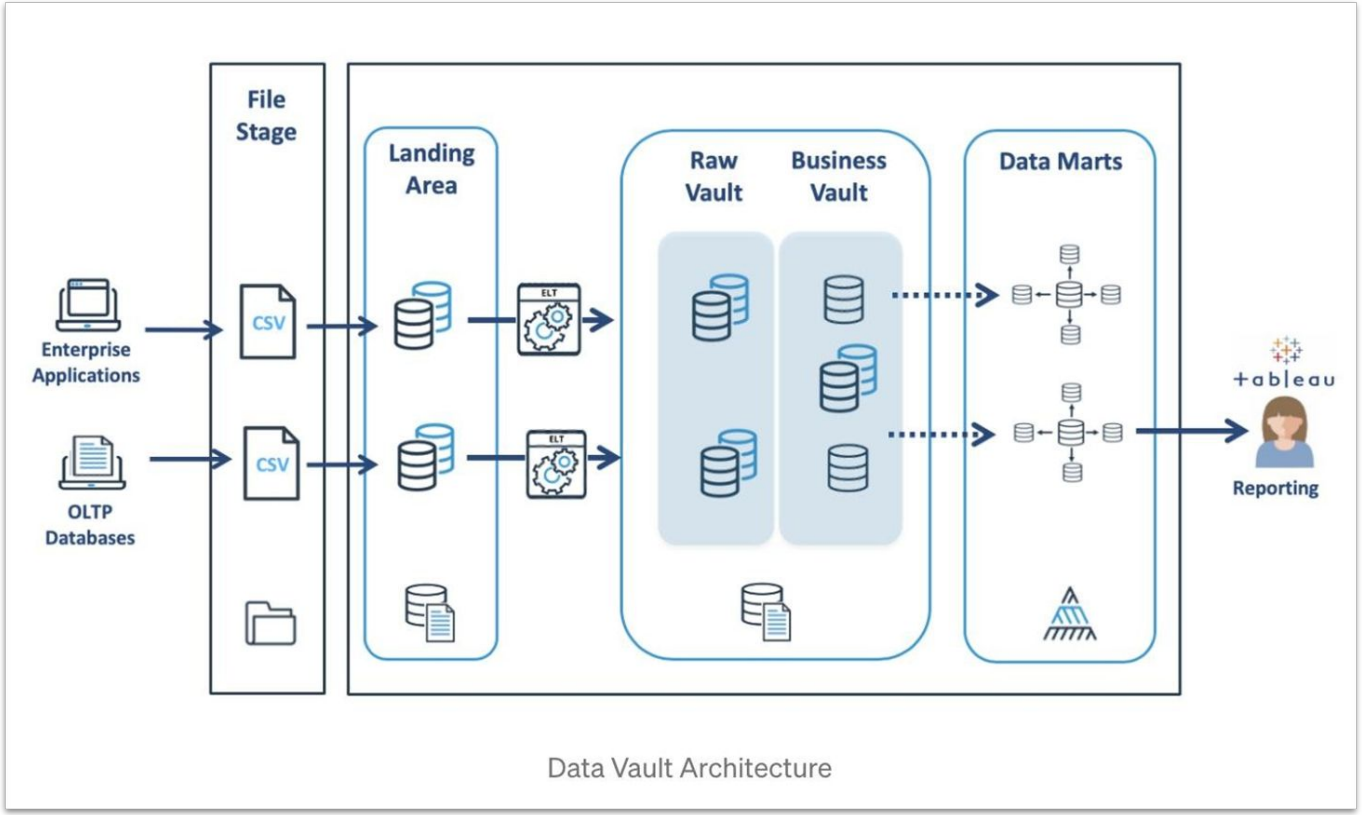
Типы измерений

- Calendar dimension - Отсылает ко всем таблицам фактов, отражает временный аспект данных
- Role-playing dimension - Одна и та же таблица измерения может играть разную роль для одного и того же факта. Каждая роль - это внешний ключ в таблице фактов. Например, в банковской транзакции есть роль отправителя и получателя. Оба будут находиться в одной таблице измерений.
- Junk dimension - Таблица для индикаторов, флагов, текстовых данных, которые не относятся ни к таблице фактов, ни к другим измерениям.

Типы измерений

- Calendar dimension - Отсылает ко всем таблицам фактов, отражает временный аспект данных
- Role-playing dimension - Одна и та же таблица измерения может играть разную роль для одного и того же факта. Каждая роль - это внешний ключ в таблице фактов. Например, в банковской транзакции есть роль отправителя и получателя. Оба будут находиться в одной таблице измерений.
- Junk dimension - Таблица для индикаторов, флагов, текстовых данных, которые не относятся ни к таблице фактов, ни к другим измерениям.
- Conformed dimension - Таблица измерения, которая может использоваться для нескольких фактов

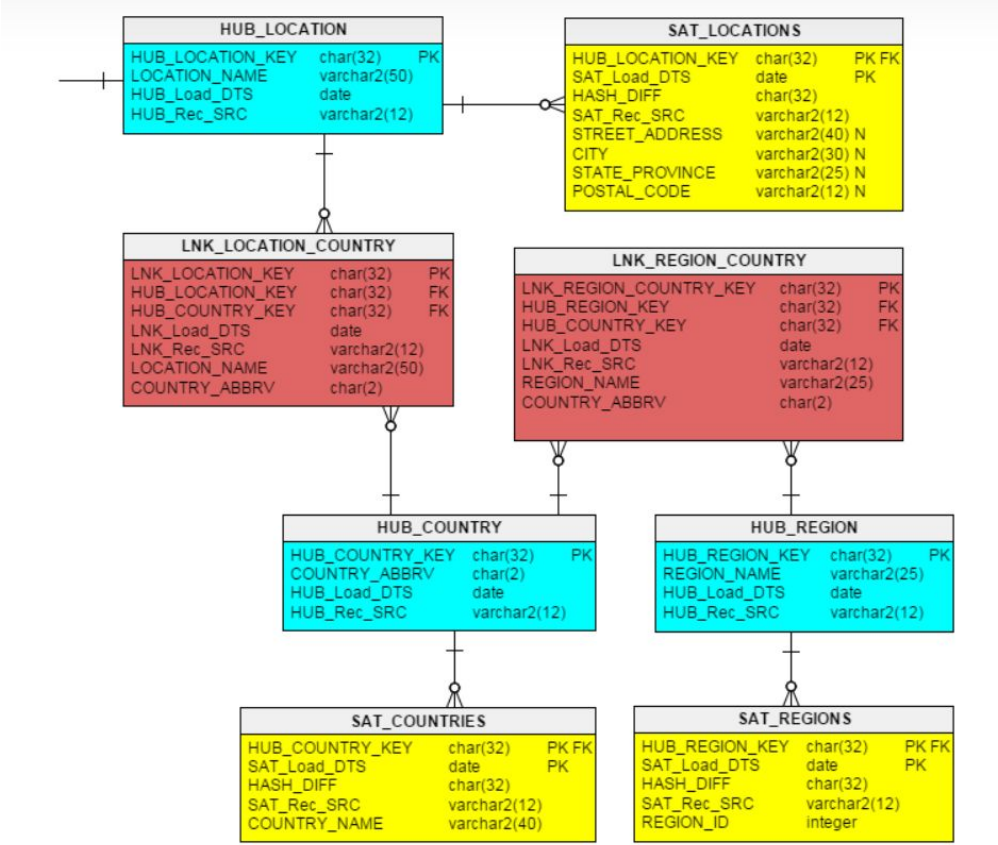
Data Vault 2.0



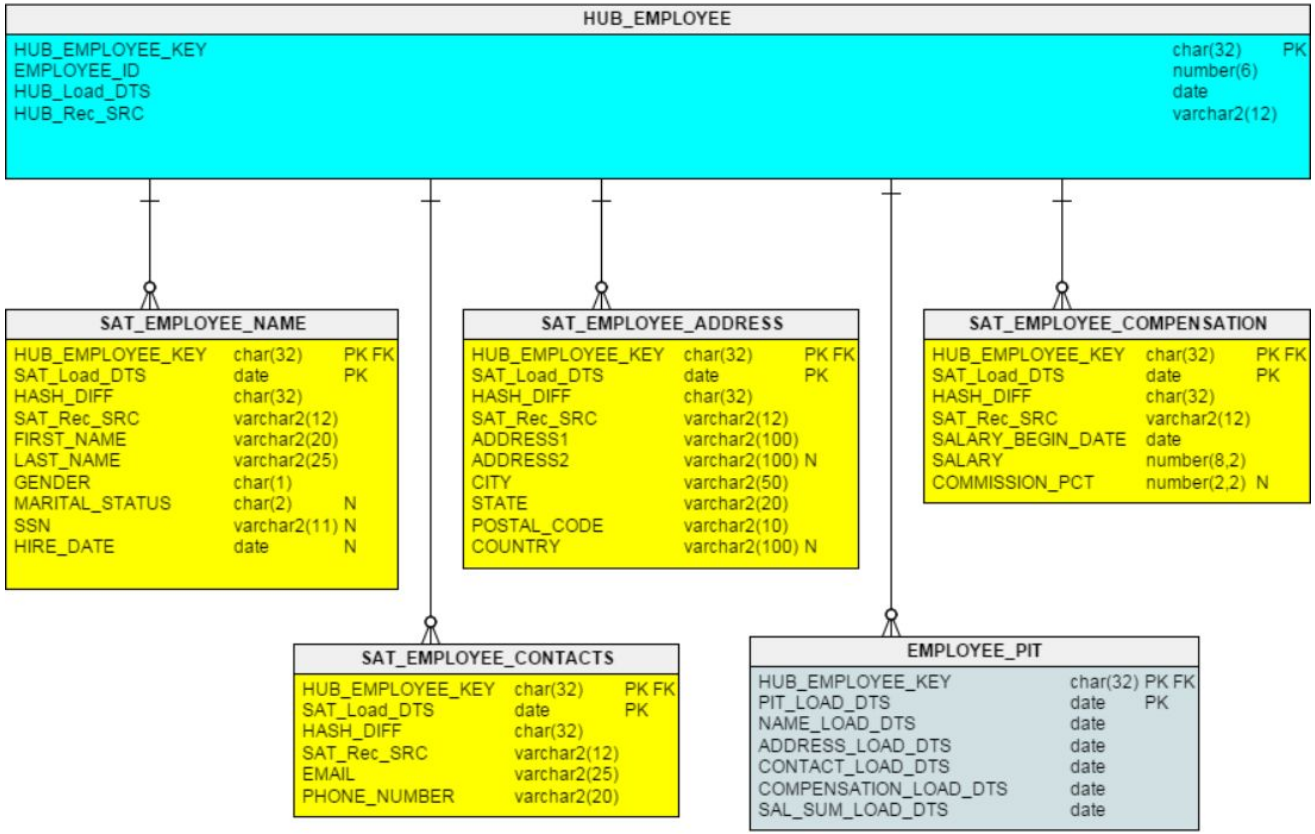
Data Vault Architecture

Raw Vault

- hub - уникальный бизнес ключ+суррогат+дата-время+источник
- link - связь между хабами
- satellite - описательные свойства хабов

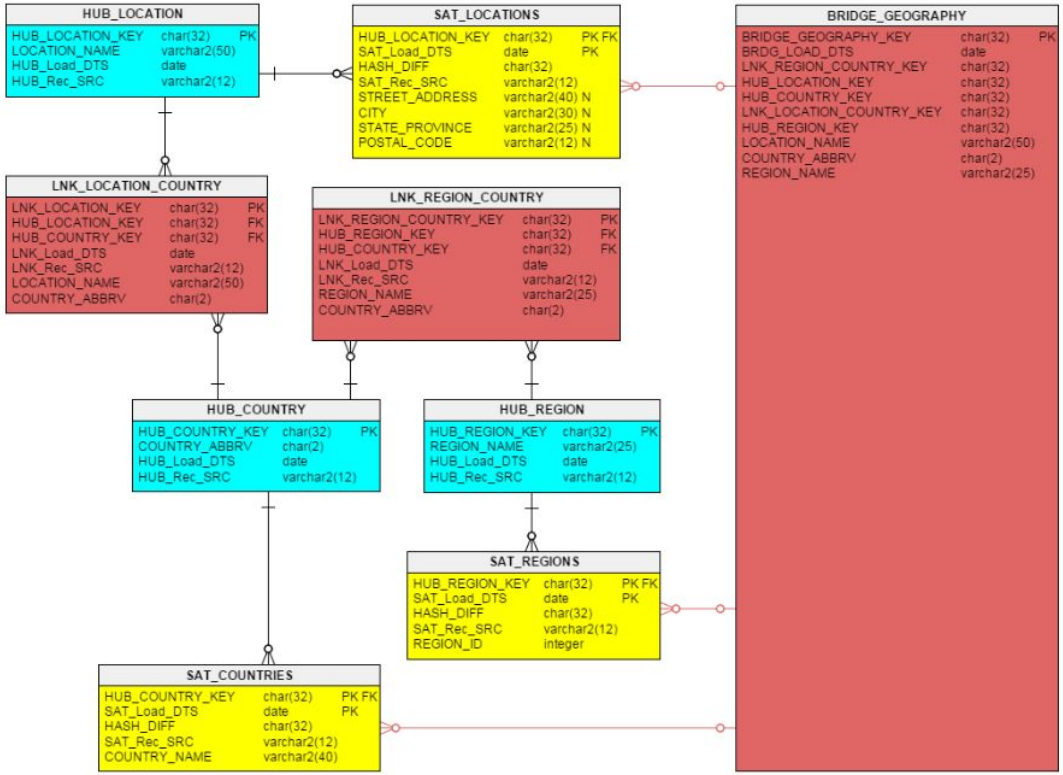


Business Vault. Point-In-Time



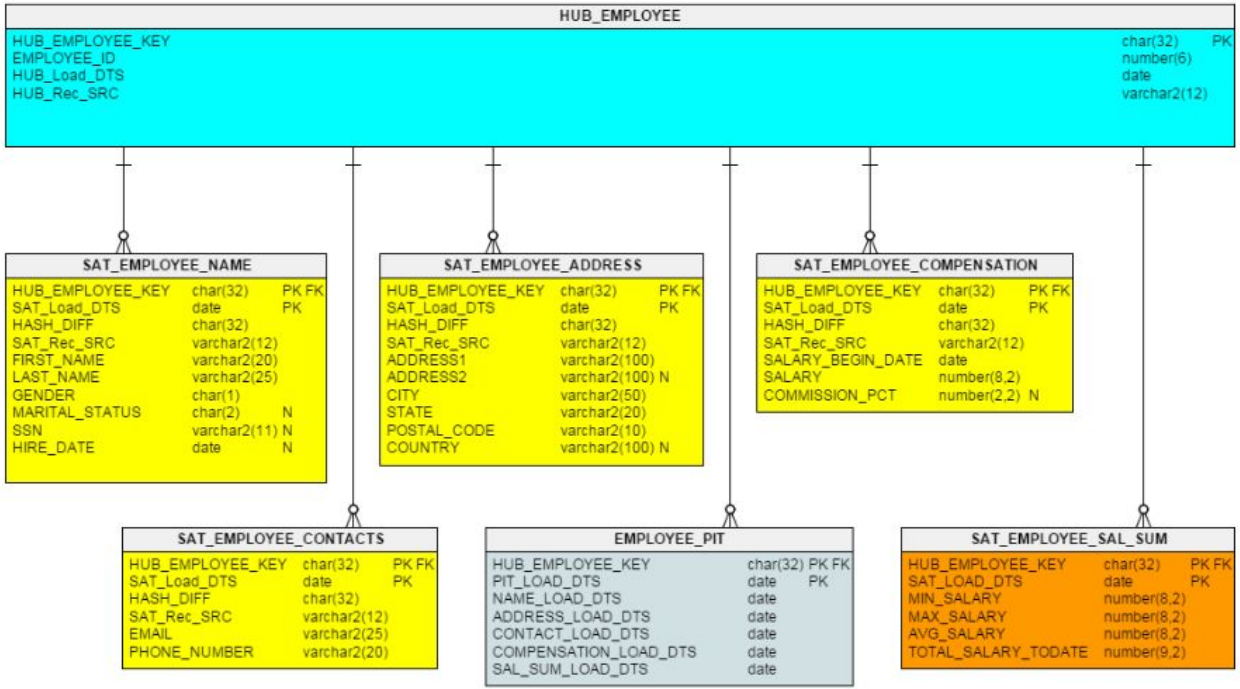
Содержит даты загрузки всех сателлитов

Business Vault. Bridge



Объединяет несколько хабов и линков вместе

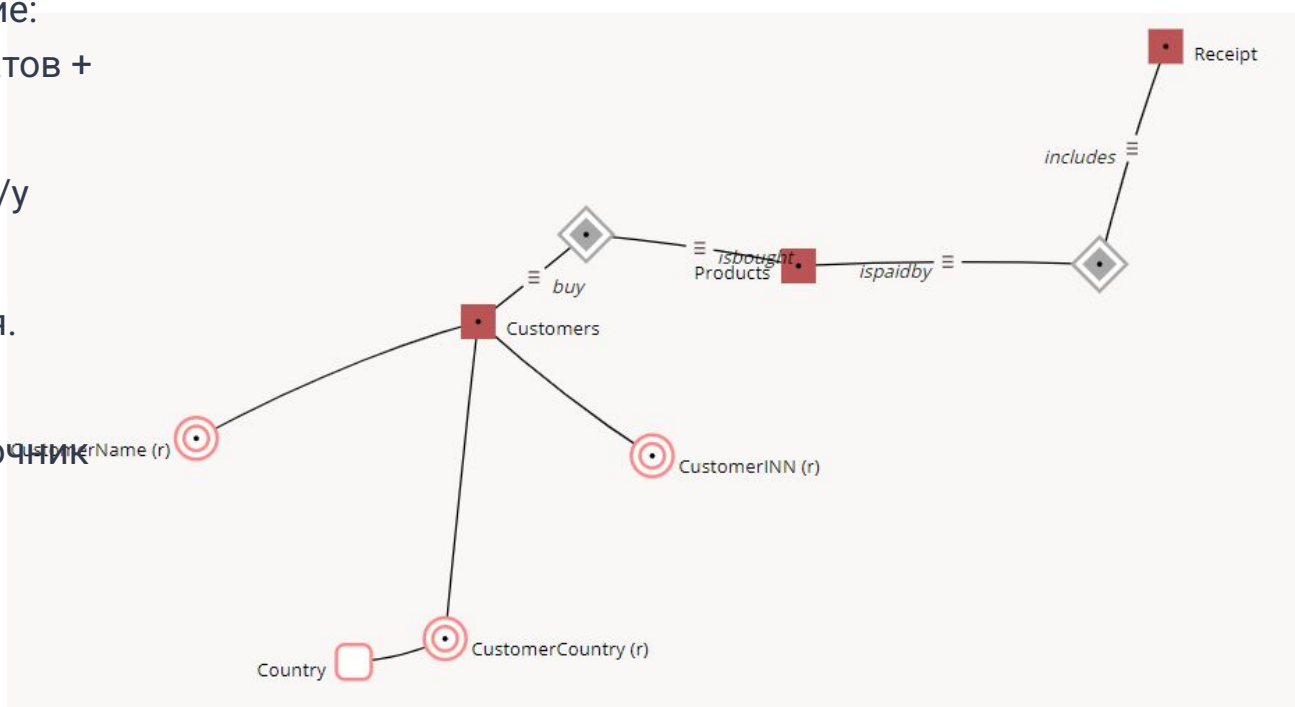
Business Vault. Predefined Calculation



Содержит подсчитанные значения

Anchor Modeling (6NF) + Code Generation

- anchors - сущность, событие: набор уникальных суррогатов + дата-время+источник
- ties - связь (отношение) м/у якорями
- attributes - описание якоря. Имеет конкретный тип.
- knots - статический справочник



Вопросы?

Ставим “+”,
если вопросы есть

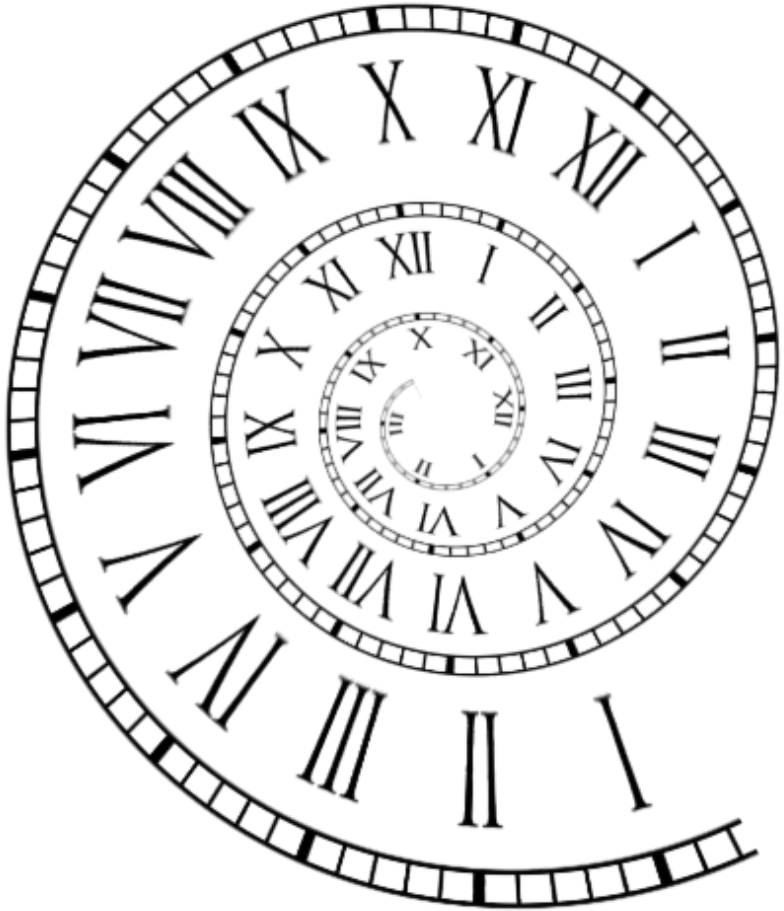
Ставим “–”,
если вопросов нет

Хранение истории. SCD.

Хранение истории

Назначение:

- отчеты по историческим данным
- логирование



Типы историчности

- SCD 0: оставляем оригинал
- SCD 1: перезаписываем последним актуальным значением
- SCD 2: добавляем новую строку - храним всю историю
- SCD 3: добавляем новый атрибут - храним последнее и предыдущее значение
- SCD 4: выносим историю в отдельную таблицу
- SCD 6: комбинация 1, 2 и 3 типа

SCD 1. Перезапись.

Текущее состояние

name	project
Aliaksei	Health monitor
Yuliya	Bank24
Tatsiana	SBL
Margarita	CRM

Новые данные

Margarita	Marketing

Новое состояние

name	project
Aliaksei	Health monitor
Yuliya	Bank24
Tatsiana	SBL
~~Margarita~~	~~CRM~~
Margarita	Marketing

SCD 2. Новая строка. Вся история.

Новое состояние

sk	name	project	date_begin	date_end	is_actual
1	Aliaksei	Health monitor	01.04.2023	31.12.9999	1
2	Yuliya	Bank24	11.05.2023	31.12.9999	1
3	Tatsiana	SBL	11.05.2023	31.12.9999	1
4	Margarita	CRM	11.01.2023	13.06.2023	0
5	Margarita	Marketing	14.06.2023	31.12.9999	1

Старое->

Новое->

SCD 3. Новый столбец. Предыдущее значение.

Новое состояние

name	project	prev_project
Aliaksei	Health monitor	
Yuliya	Bank24	
Tatsiana	SBL	
Margarita	Marketing	CRM

SCD 4. Отдельная таблица

Актуальный срез

sk	name	project
1	Aliaksei	Health monitor
2	Yuliya	Bank24
3	Tatsiana	SBL
5	Margarita	Marketing

Историческая таблица

sk	name	project	create_date
1	Aliaksei	Health monitor	01.04.2023
2	Yuliya	Bank24	11.05.2023
3	Tatsiana	SBL	11.05.2023
4	Margarita	CRM	11.01.2023
5	Margarita	Marketing	14.06.2023

SCD 6. 1+2+3

Новое состояние

sk	name	project	date_begin	date_end	current_project
1	Aliaksei	Health monitor	01.04.2023	31.12.9999	Health monitor
2	Yuliya	Bank24	11.05.2023	31.12.9999	Bank24
3	Tatsiana	SBL	11.05.2023	31.12.9999	SBL
4	Margarita	CRM	11.01.2023	13.06.2023	Marketing
5	Margarita	Marketing	14.06.2023	31.12.9999	Marketing

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Extending: UDF, Macro, Packages

Macro usage magic

- Column values comparison (сравнение значений столбцов)
- Creating UDF (создание функции)
- Handling nested JSON structures (обработка JSON)
- Wrapping reusable pieces of code (обертывание кода)
- Compressing tables (сжатие таблиц)
- Granting database object permissions (предоставление прав доступа)
- Vacuum & analyze tables (обслуживание баз данных)

Extending DWH with UDF

```
{% macro create_udf() -%}

  {% set sql %}

        CREATE OR REPLACE FUNCTION {{ target.schema }}.f_email_hash(mes "varchar")
            RETURNS varchar
            LANGUAGE plpythonu
            STABLE
        AS $$
            import hashlib
            prep = mes.strip().lower()
            ## return prep
            return hashlib.sha256(prep).hexdigest()
        $$
        ;


        CREATE OR REPLACE FUNCTION {{ target.schema }}.f_email_domain(mes "varchar")
            RETURNS varchar
            LANGUAGE plpythonu
            STABLE
        AS $$
            prep = mes.split('@')[1]
            return prep
        $$
        ;

  {% endset %}

  {% set table = run_query(sql) %}

{%- endmacro %}
```

Generating calendar in one line

models > marts > dim > dim_calendar.sql

You, a year ago | 1 author (You)

```
{{
    config(
        materialized='table',
        dist="all",
        sort='date_day'
    )
}}

{{ dbt_date.get_date_dimension('2012-01-01', '2025-12-31') }}
```

	Value
date_day	2021-03-29
prior_date_day	2021-03-28
next_date_day	2021-03-30
prior_year_date_day	2020-03-29
prior_year_over_year_date_day	2020-03-30
day_of_week	1
day_of_week_name	Monday
day_of_week_name_short	Mon
day_of_month	29
day_of_year	88
week_start_date	2021-03-29
week_end_date	2021-04-04
prior_year_week_start_date	2020-03-30
prior_year_week_end_date	2020-04-05
week_of_year	13
iso_week_start_date	2021-03-29
iso_week_end_date	2021-04-04
prior_year_iso_week_start_date	2020-03-30
prior_year_iso_week_end_date	2020-04-05
iso_week_of_year	13
prior_year_week_of_year	14
month_of_year	3
month_name	MARCH
month_name_short	MAR
month_start_date	2021-03-01
month_end_date	2021-03-31
prior_year_month_start_date	2020-03-01
prior_year_month_end_date	2020-03-31
quarter_of_year	1
quarter_start_date	2021-01-01
quarter_end_date	2021-03-31
year_number	2,021
year_start_date	2021-01-01
year_end_date	2021-12-31
fiscal_week_of_year	9

Importing packages allows reusing code

packages.yml

Artemiy Kzr, 2 months ago | 3 authors (Artemiy Kzr and others)

```yaml
packages:
  - package: fishtown-analytics/dbt_utils
    version: 0.6.4
  - package: fishtown-analytics/redshift
    version: 0.4.1
  - package: fishtown-analytics/logging
    version: 0.4.1
  - package: fishtown-analytics/dbt_external_tables
    version: 0.6.2
  - git: "https://github.com/wheely/dbt-date.git"
    revision: 0.2.4
```

Artemiy Kzr, 2 months ago via PR #846 • Ak (#846)

dbt hub - подборка модулей

Logging every run metrics

```
models:
    +pre-hook: "{{ logging.log_model_start_event() }}"
    +post-hook: "{{ logging.log_model_end_event() }}"
```

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Maintenance & Security

DWH Maintenance

- Отслеживать новые показатели
- Определять нужны ли старые метрики
- Настраивать группы привилегий пользователей
- Следить за оптимизацией хранилища, при необходимости изменить подход к модели.

DWH Security

- Ограничить доступ пользователей
- Обеспечить защиту сетей, в которых хранятся данные
- Настройка передачи данных с учетом соблюдения требований безопасности
- Обезличивание данных

DWH Security Best practices

- Шифровать данные
- Классифицировать данные
- Контроль ролей, прав
- Безопасное перемещение данных
- Разделение данных
- Защита КХД

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Список материалов для изучения

1. What is Normalization in DBMS (SQL)?DWH maintenance
2. https://habr.com/ru/company/wheely/blog/549614/
3. https://www.dataprix.com/en/blog-it/guest/data-warehouse-security-best-practices
4. https://docs.getdbt.com/terms/dimensional-modeling
5. https://habr.com/ru/companies/yandex/articles/557140/
6. https://habr.com/ru/articles/101544/
7. https://github.com/Data-Engineer-Camp/dbt-dimensional-modelling/tree/main/docs
8. https://www.sciencedirect.com/topics/computer-science/dimensional-model

Цели вебинара

Проверка достижения целей

1. Изучить основные концепции в построении Хранилища Данных
2. Разобрать основные понятия, относящиеся к DWH
3. Выявить best practices в организации хранилища данных
4. Моделировать хранилища на примерах из реальной жизни

Домашние задание

Спроектировать DWH для вымышленной компании, которая занимается продажей электроники. DWH должен содержать данные о продажах, клиентах, продуктах, складах.

Шаги:

1. Сформулировать бизнес-требования и цели создания data warehouse.
2. Придумать как минимум три системы источника, в каждой из которых данные передаются в своем формате.
3. Разработать dimensional модель данных для data warehouse (dds слой), включая измерения, факты и связи между ними (минимум 5 сущностей).
4. Описать принципы разложения данных по слоям stage - ods - dds - datamart

Рефлексия

Заполните, пожалуйста, опрос о занятии по ссылке в ЛК

Спасибо за внимание!