Data warehouse business concept

for a company importing ordered goods

Written by

Andrey Neveykov, 2022

**Overview**

 Пример решения задачи по проектированию хранилища данных для белорусского бизнеса по импорту товаров из других государств под заказ.

**Business Background**

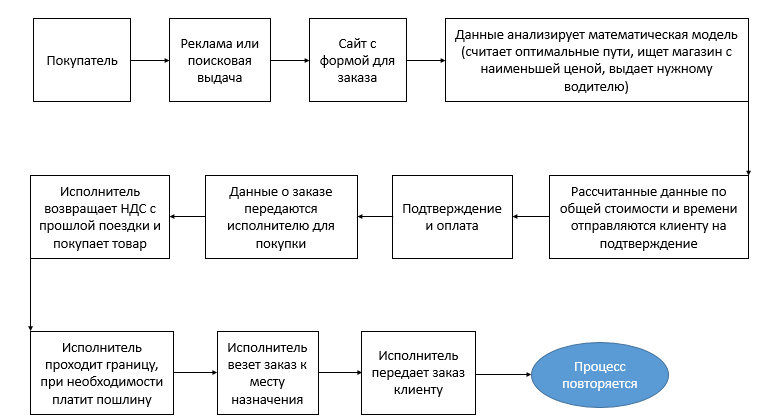
Бизнес по импорту иностранной продукции под заказ. В странах Европейского союза, при вывозе товаров за рубеж, при повторном посещении страны можно вернуть НДС в виде налогового вычета и цены на многие типы товаров ниже. 50 кг продукции можно вывезти без пошлины.

Клиент получает товар по более низкой цене.

Бизнес зарабатывает на:

1. Разнице в цене в РБ и за рубежом.
2. Налоговом вычете при повторном посещении страны
3. Курсовой разнице валют (компания может покупать валюту на бирже, где курс выгоднее)

Схема бизнес-процесса представлена ниже.



Для международного бизнеса с большим оборотом требуется хранилище данных, для быстрого доступа ко всем операциям, способное быстро обновляться и работать под нагрузкой в диапазоне 1000-3000 заказов в день.

**Benefit**

Предлагаемое решение должно быть полезно компании для:

1. Хранения больших объемов данных по всем транзакциям для их анализа и составления выборок для обучение математических моделей на основе нейронных сетей.
2. Упрощения бухгалтерского учета за счет хранения данных в единой системе и сохранения истории изменения данных.
3. Ускорения взаимодействия сайта с базой данных.
4. Уменьшения затрат на физические носители информации за счет возможности использовать более мощные носители для свежей информации и медленные, старые носители для исторической информации.
5. Повышения безопасности данных от утечек и несанкционированного изменения, за счет возможности разделения доступа между сотрудника и отделами.
6. Возможности создавать представления (view) с агрегированной информацией руководству, для принятия стратегических решений.

**Requirements**

**Business Requirements**

|  |  |
| --- | --- |
| **ID** | **Description** |
| B01 | Возможность расчета состояния финансовых ресурсов компании на текущий момент и любой день в исторической перспективе до 5 лет. |
| B02 | Возможность анализировать изменения ключевых метрик компании (доход общих, средний доход с клиента, структура дохода по его источникам, общая окупаемость, средняя окупаемость одного заказа, количество уникальных клиентов, количество клиентов, сделавших заказ за последние 30 дней, самые прибыльные страны для импорта, количество сотрудников)на промежутках месяц, квартал, год и период, задаваемый вручную. |
| B03 | Возможность связать информацию о клиенте с его местом нахождения (с точностью до адреса) и заказом. |
| B04 | Присылать уведомление начальнику подразделения, в котором сотрудник отработал определенное количество времени для рекомендации к повышению по должности или увеличение оклада. |
| B05 | Возможность анализировать, какой из исполнителей принес наибольший доход. |
| B06 | Возможность делать сравнение выборки по любой из хранимых метрик в сравнении с аналогичными периодами прошлых лет. |

**Technical Requirements**

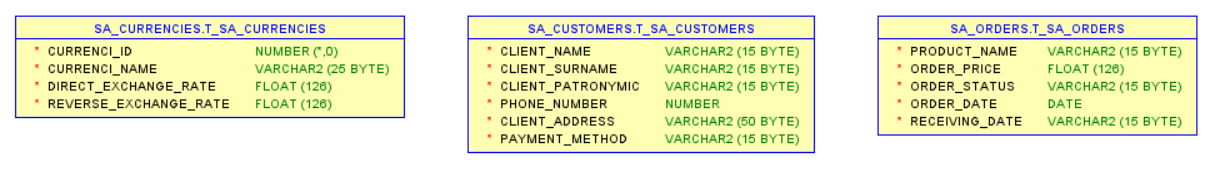
|  |  |
| --- | --- |
| **ID** | **Description** |
| T01 | Обрабатывать данные о примерно 1000 заказов в день, с возможностью увеличения вблизи праздников и мировых распродаж (черных пятниц). |
| T02 | Чтение данных из csv-файлов, генерируемых на back-end части сайта. |
| T03 | Должны осуществляться логические проверки на вносимые данные (невозможность заказать на вчера или на 3022 год). |
| T04 | Разделение прав доступа к таблицам согласно занимаемой должности. В частности, сохранять анонимность клиентов, невозможность изменять финансовую историческую информацию и удалять любую информацию (без получения разрешения от директора). |
| T05 | Возможность быстрого доступа к новой информации: хранить новые данные на быстрых носителях, старые данные – на более медленных. |
| T06 | Доступ к базе данных из любой точки мира, в любое время суток  (В частности возможность замены физических комплектующих без необходимости отключать сервера). |
| T07 | Хранение истории операций с базой данных, возможность восстановления информации. |
| T08 | Хранение информации в нормализованном виде. |

**Solution Sketch**

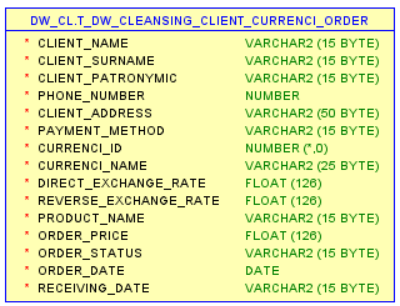
**Source Tables structure**

Данные, которые необходимо загружать в хранилище, генерируются back-end частью сайта. Они хранятся в csv-документах (данные о клиентах и о заказах). Также необходимо обновлять данные получаемые из API биржи, где компания производит валютно-обменные операции. Данные обновляются на момент оплаты заказа.

Таким образом, для решения поставленной задачи предлагается организовать SA-уровень, состоящий из трёх таблиц (схема представлена ниже).

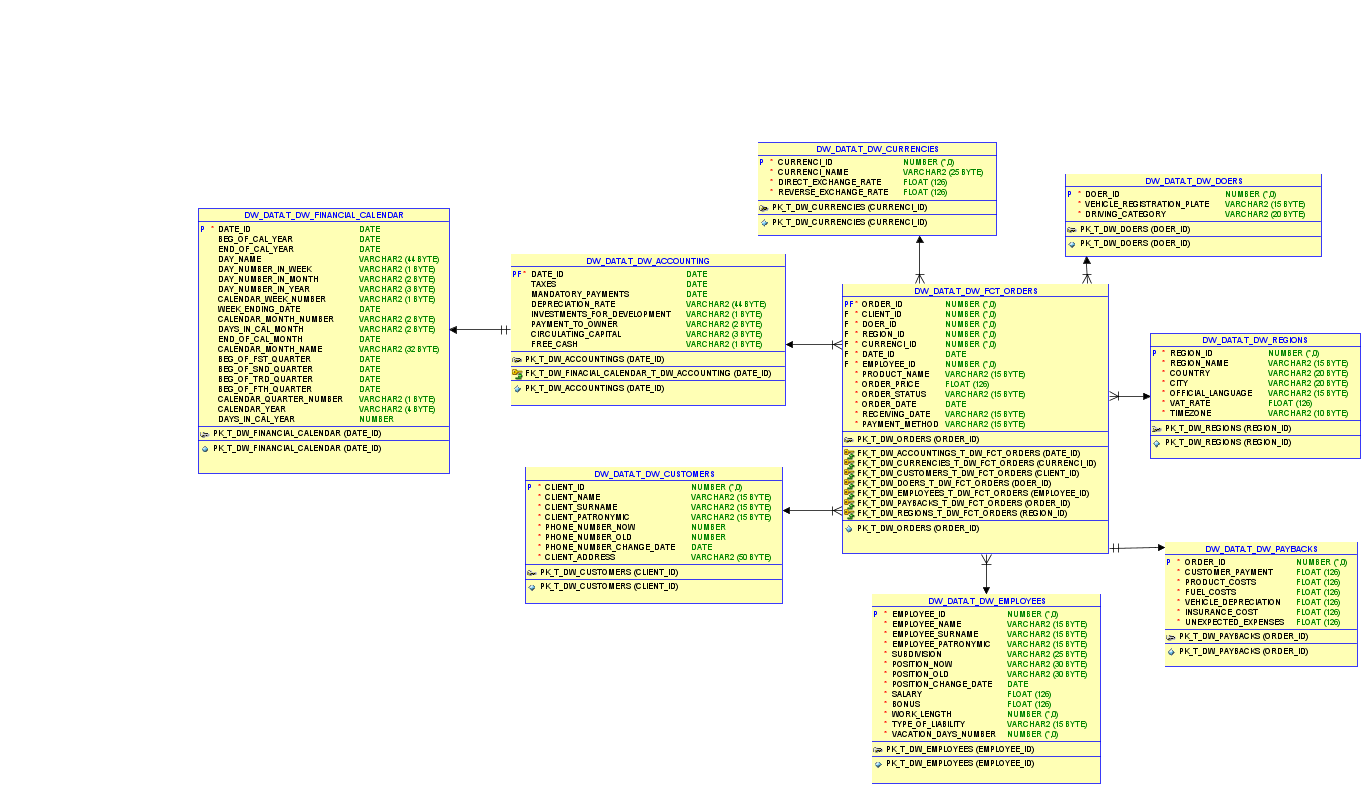


Логическое разделение таблиц, хранящей данные о клиентах и о заказах обусловлено тем, что один клиент может сделать сразу несколько заказов. Далее предполагается устанавливать логическое соответствие каждого заказа конкретному клиенту. Далее данные при помощи соответствующей процедуры попадают на cleansing-уровень, в единую таблицу, где происходят логические проверки.



**Summarize Data Plan**

Для решения поставленной бизнесом задачи предполагается использовать STAR-схему хранилища, представленную ниже.



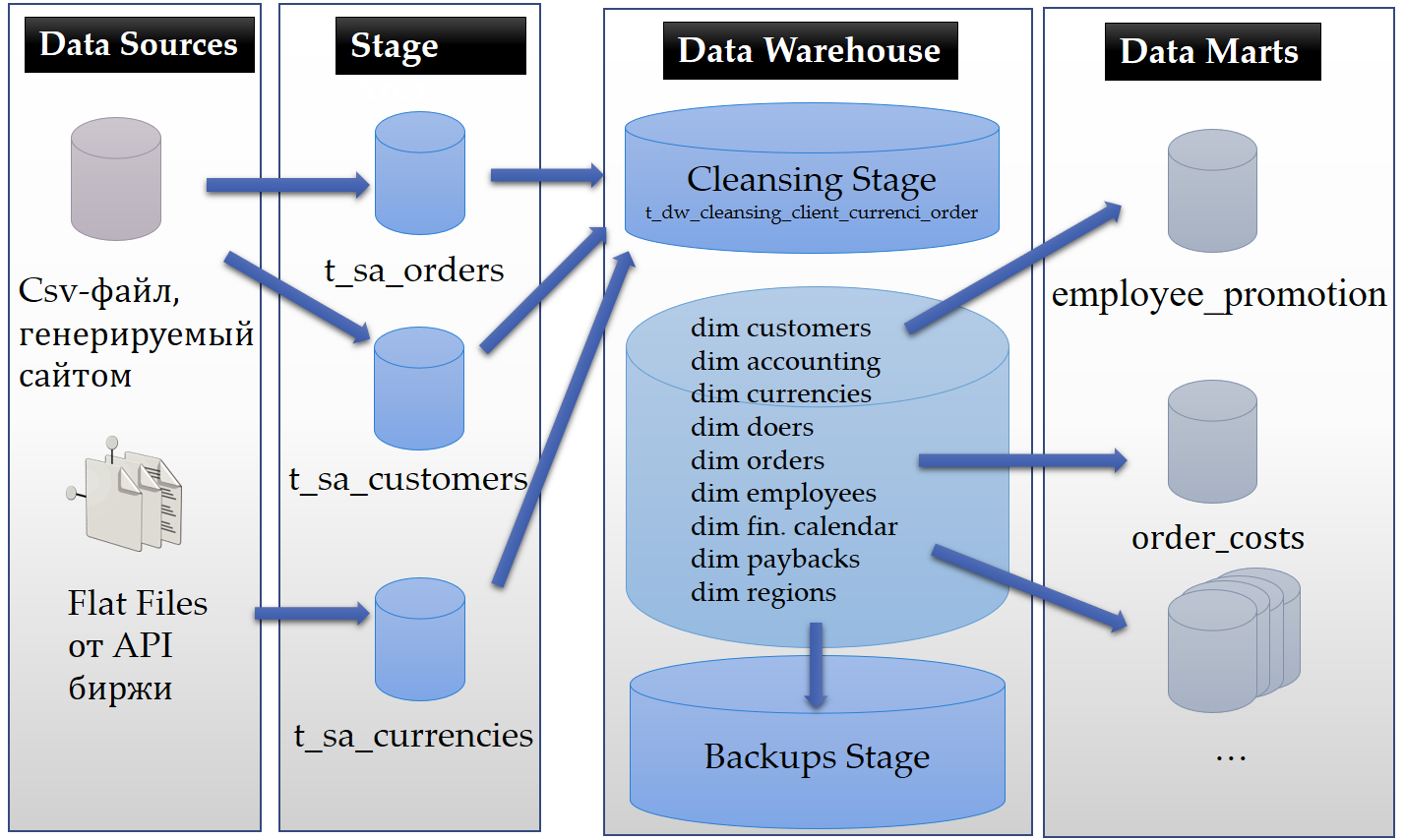
Характеристики представленной схемы:

* Таблицы измерений присоединяется к таблице фактов с помощью внешнего ключа
* Таблицы измерений не соединена друг с другом, кроме таблицы с финансовым календарем компании и таблицы с основными финансовыми показателями, для создания удобных выборок по периоду.
* STAR-схема проста для понимания и обеспечивает оптимальное использование памяти.
* Схема широко поддерживается BI Tools.

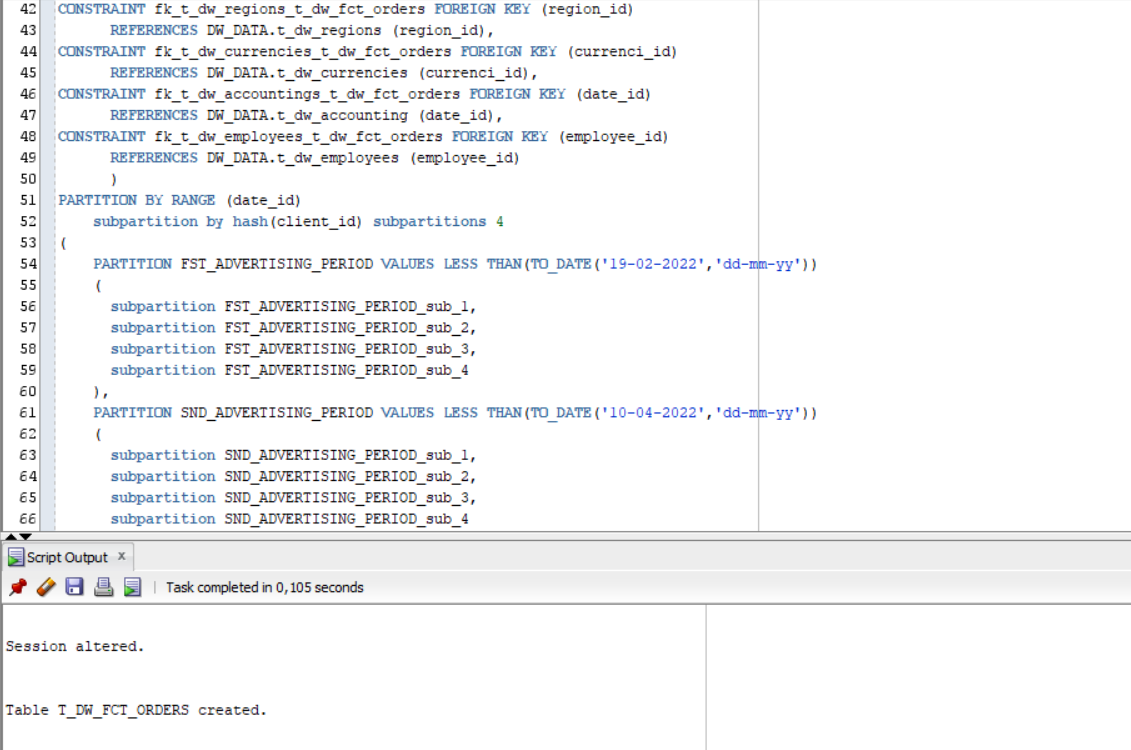
Для повышения безопасности и разделения доступа предлагается использовать в хранилище несколько tablespace­-ов, с обособленным пользователем для каждого. Структура уровней хранилища представлена ниже:

|  |  |  |  |
| --- | --- | --- | --- |
| Level Type | Object Name | Tablespace | Desctiption |
| Storage level  SA\_\* | SA\_CUSTOMERS | ts\_sa\_customers\_data\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Загружается из csv-файла, содержит имя, фамилию, отчество, номер телефона, статус оплаты (оплатил/ нет) |
| SA\_ORDERS | ts\_sa\_orders\_data\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Загружается из csv-файла, содержит наименование товара, сумму заказа, статус заказа (принят/закуплен/передан клиенту) |
| SA\_CURRENCIES | ts\_sa\_currencies\_data\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 50M,  Autoextend clause ON next 10M) | Загружается из файла, содержит наименование валюты, прямой курс конвертации и обратный курс конвертации |
| DW - Cleansing Level | DW\_CL | ts\_dw\_cl\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 100M) | Загружает информацию из Storage level, готовит ее к дальнейшей очистке |
| DW – Level | DW\_DATA | ts\_dw\_data\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Загружает информацию из Cleansing level, нормализует данные. |
| DW– Prepare Star Cleansing Level | SAL\_DW\_CL | ts\_dw\_star\_cls\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Загружает информацию из уровня DW. Содержит представления (view), объединяющие объекты с уровня DW. |
| STAR - Cleansing | SAL\_CL | ts\_sal\_cl\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Загружает информацию из уровня DW\_CL. Содержит представления (view) предыдущего уровня, но без избыточностей. |
| STAR – Level | DM\_EMPLOYEES | ts\_dm\_employees\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 50M,  Autoextend clause ON next 10M) | Хранит информацию о dim employees |
| DM\_CUSTOMERS | ts\_dm\_customers\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Хранит информацию о dim customers |
| DM\_ORDERS | ts\_dm\_orders\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Хранит информацию о фактах (в таблице t\_dw\_fct\_orders) |
| DM\_CURRENCIES | ts\_dm\_currencies\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 100M,  Autoextend clause ON next 10M) | Хранит информацию о dim currencies |
| DM\_DOERS | ts\_dm\_doers\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 50M,  Autoextend clause ON next 5M) | Хранит информацию о dim doers |
| DM\_PAYBACKS | ts\_dm\_paybacks\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Хранит информацию о dim paybacks |
| DM\_ACCOUNTINGS | ts\_dm\_accounting\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 100M,  Autoextend clause ON next 10M) | Хранит информацию о dim accounting и dim financial calendar |
| DM\_REGIONS | ts\_dm\_regions\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 70M,  Autoextend clause ON next 15M) | Хранит информацию о dim regions |

DataFlow хранилища представлен ниже:

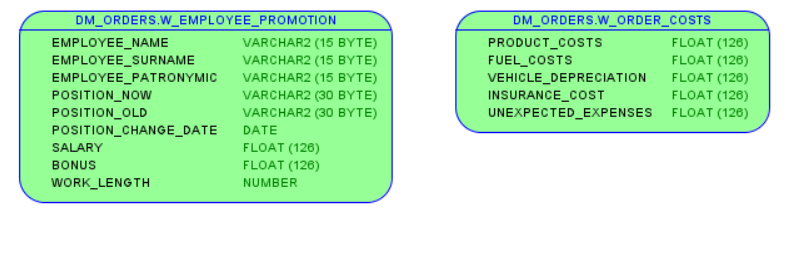


В хранилище рекомендуется для оптимизации запросов и места на диске рекомендуется использовать партиционирование по временным периодам (например, проведения рекламных компаний) и субпартиционирование по хэшу (например, client\_id).

****

Так как чаще всего в системе изменяются dimention-ы: client, order и employee(т.к. компания большая). Поэтому логичнее всего использовать параллелизм в DW, CL и SA уровнях для обновления данных о них. Однако применять параллельные вычисления можно и в DM – data mart уровнях, при проектировании представлений (view), связанных с финансами. Т.к. они часто обновляются из-за большого числа транзакций, а бухгалтерам, маркетологам и т.д. необходима достоверная актуальная информация.

Самые часто просматриваемые views в бизнесе представлены ниже.



**Conclusion**

Таким образом представленное техническое решение поможет бизнесу получить дополнительную прибыль на:

1. Упрощения бухгалтерского учета за счет хранения данных в единой системе и сохранения истории изменения данных.
2. Повысит конверсию за счет ускорения взаимодействия клиента с сайтом (за счет ускорения ответа от базы данных).
3. Уменьшении затрат на физические носители информации за счет возможности использовать более мощные носители для свежей информации и медленные, старые носители для исторической информации.
4. Повышения безопасности данных от утечек и несанкционированного изменения, а значит несения репутационных и материальных издержек.
5. Повышения эффективности маркетингового отдела за счет возможности создавать представления (view) с агрегированной информацией для принятия решений.