Data warehouse business concept

for a company importing ordered goods

Written by

Andrey Neveykov, 2022

**Overview**

 An example of solving the problem of designing a data warehouse for a Belarusian business importing goods from other countries on order.

**Business Background**

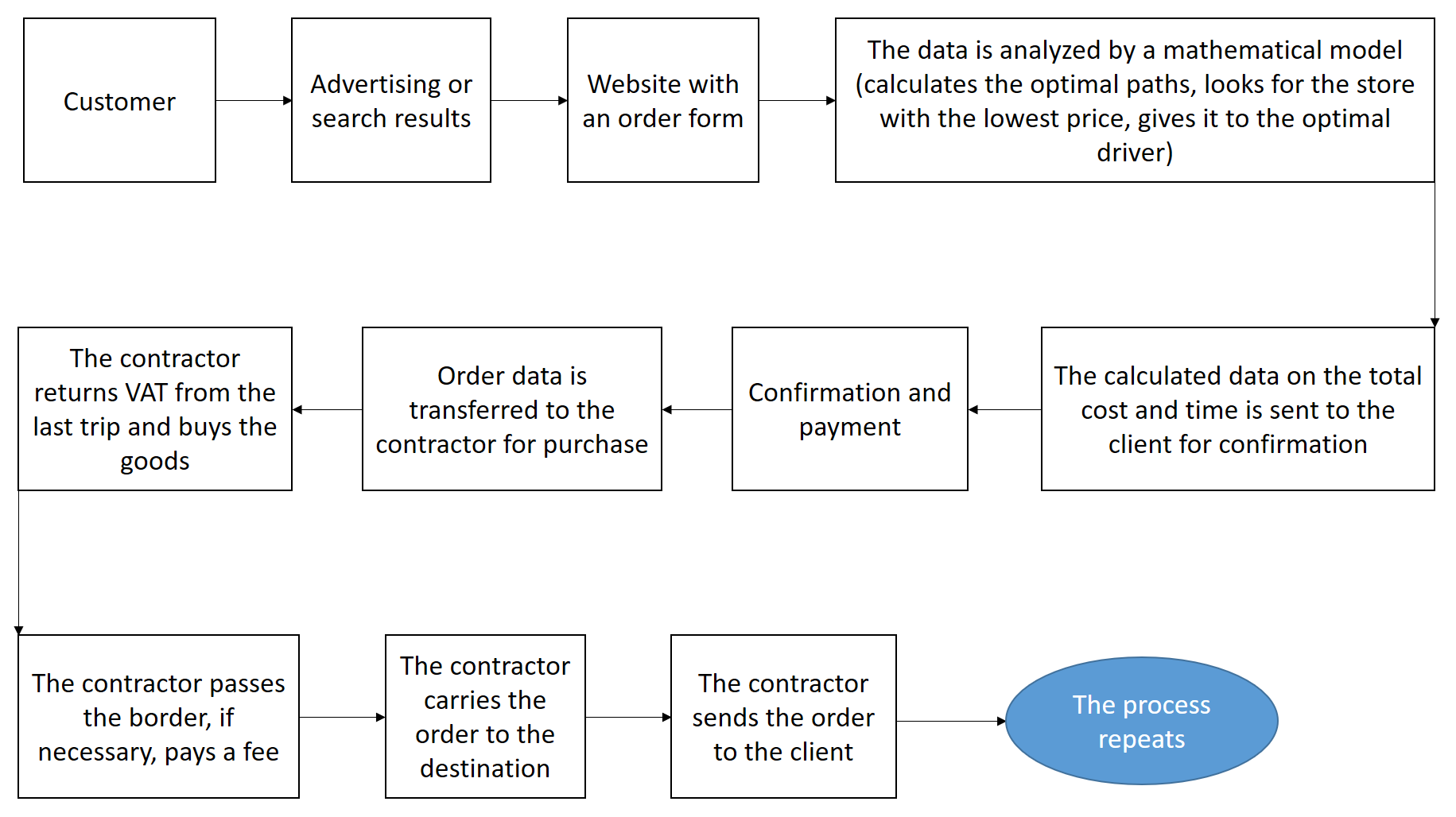
Business for the import of foreign products under the order. In the countries of the European Union, when exporting goods abroad, when revisiting the country, you can return VAT in the form of a tax deduction and prices for many types of goods are lower. 50 kg of products can be exported without duty.

The customer receives the product at a lower price.

Business earns from:

1. The difference in price in Belarus and abroad.
2. Tax deduction when revisiting the country.
3. Exchange rate difference of currencies (the company can buy currency on the exchange, where the exchange rate is more profitable).

The business process diagram shown below.



International business with high turnover requires a data warehouse to quickly access all operations, be able to update quickly and work under load in the range of 1000-3000 orders per day.

**Benefit**

The proposed solution should be useful for the company to:

1. Storing large amounts of data on all transactions for their analysis and sampling for training mathematical models based on neural networks.
2. Simplification of accounting by storing data in a single system and saving the history of data changes.
3. Accelerating the interaction of the site with the database.
4. Reducing the cost of physical storage due to the ability to use more powerful storage for fresh information and slow, old storage for historical information.
5. Increasing data security from leaks and unauthorized changes, due to the possibility of sharing access between an employee and departments.
6. Possibility to create representations (view) with the aggregated information to a management, for acceptance of strategic decisions.

**Requirements**

**Business Requirements**

|  |  |
| --- | --- |
| **ID** | **Description** |
| B01 | The ability to calculate the state of the company's financial resources at the current moment and any day in a historical perspective up to 5 years. |
| B02 | The ability to analyze changes in key company metrics (general income, average income per client, income structure by its sources, total payback, average payback per order, number of unique customers, number of customers who placed an order in the last 30 days, the most profitable countries for import, quantity employees) at intervals of a month, a quarter, a year, and a period set manually. |
| B03 | The ability to link information about the client with his location (accurate to the address) and order. |
| B04 | Send a notification to the head of the department in which the employee has worked a certain amount of time for a recommendation for a promotion or salary increase. |
| B05 | The ability to analyze which of the contractors brought the most income. |
| B06 | Possibility to compare the selection by any of the stored metrics in comparison with the same periods of previous years. |

**Technical Requirements**

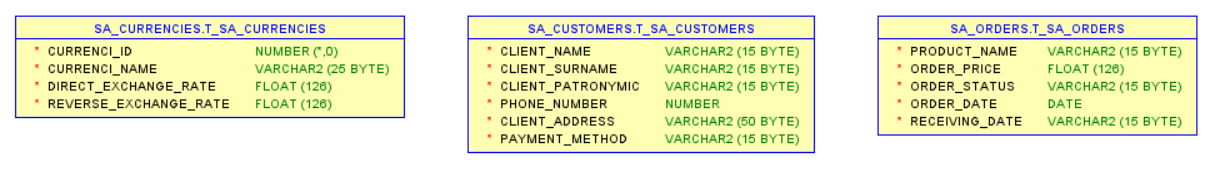
|  |  |
| --- | --- |
| **ID** | **Description** |
| T01 | Process data on approximately 1000 orders per day, with the possibility of increasing around holidays and global sales (Black Fridays). |
| T02 | Reading data from csv files generated on the back-end of the site. |
| T03 | Logical checks on the entered data should carried out (impossibility to order for yesterday or for 3022). |
| T04 | Separation of access rights to tables according to the position held. In particular, to maintain the anonymity of clients, the inability to change financial historical information and delete any information (without obtaining permission from the director). |
| T05 | Ability to quick access new information: store new data on fast media, old data on slower media. |
| T06 | Access to the database from anywhere in the world, at any time of the day  (In particular, the ability to replace physical components without having to turn off the server). |
| T07 | Storing the history of operations with the database, the ability to restore information. |
| T08 | Storing information in a normalized form. |

**Solution Sketch**

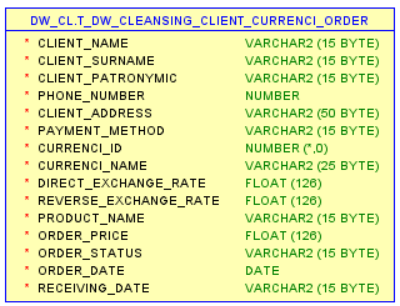
**Source Tables structure**

The data that needs to be loaded into the storage generated by the back-end part of the site. They are stored in csv documents (customer and order data). It is also necessary to update the data received from the API of the exchange where the company performs currency exchange operations. Data is updated at the time of order payment.

As a result, to solve the problem, it is proposed to organize an SA-level consisting of three tables (the diagram is presented below).

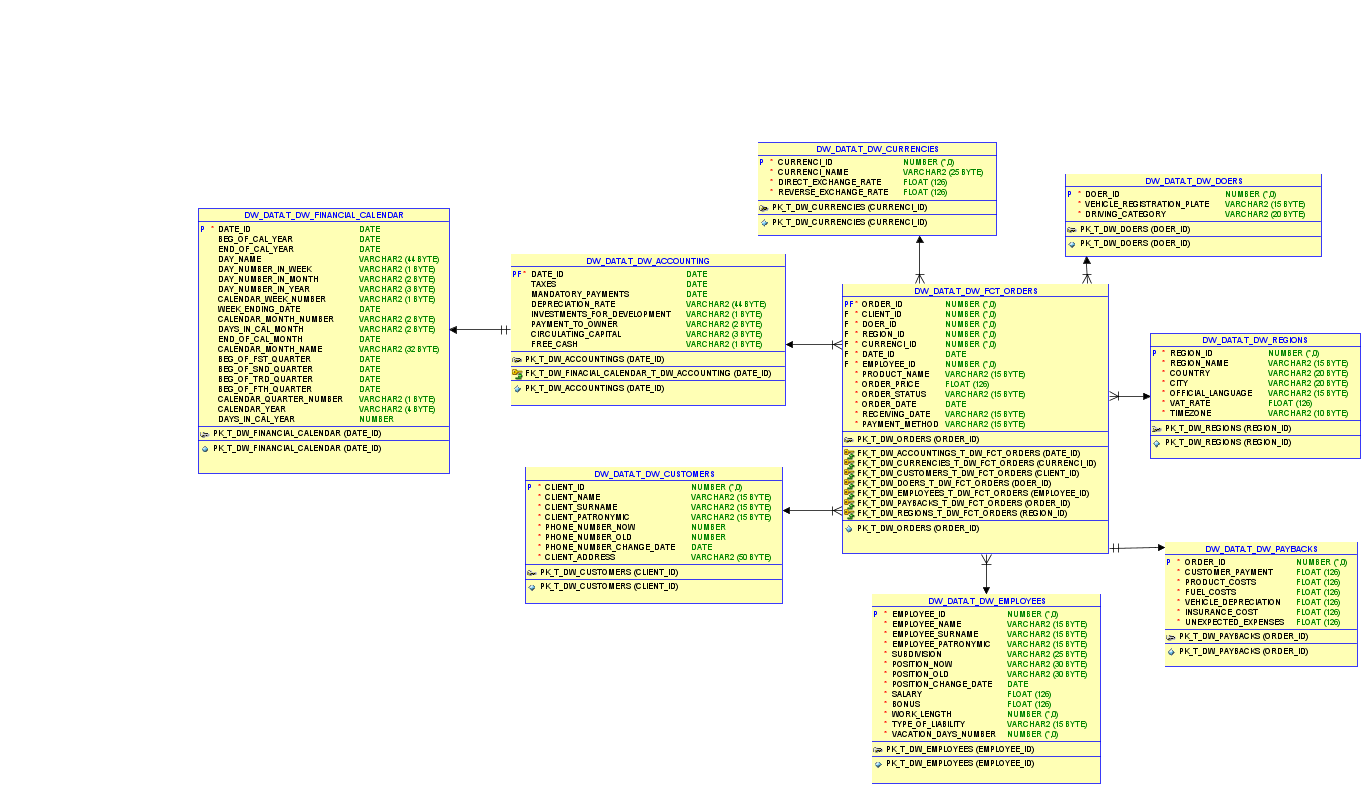


The logical separation of tables storing data about customers and orders is due to the fact that one customer can place several orders at once. Further, it is supposed to establish a logical correspondence of each order to a specific client. Further, the data, using the appropriate procedure, gets to the cleansing level, into a single table, where logical checks take place.



**Summarize Data Plan**

To solve the problem set by the business, it is supposed to use the storage STAR scheme presented below.



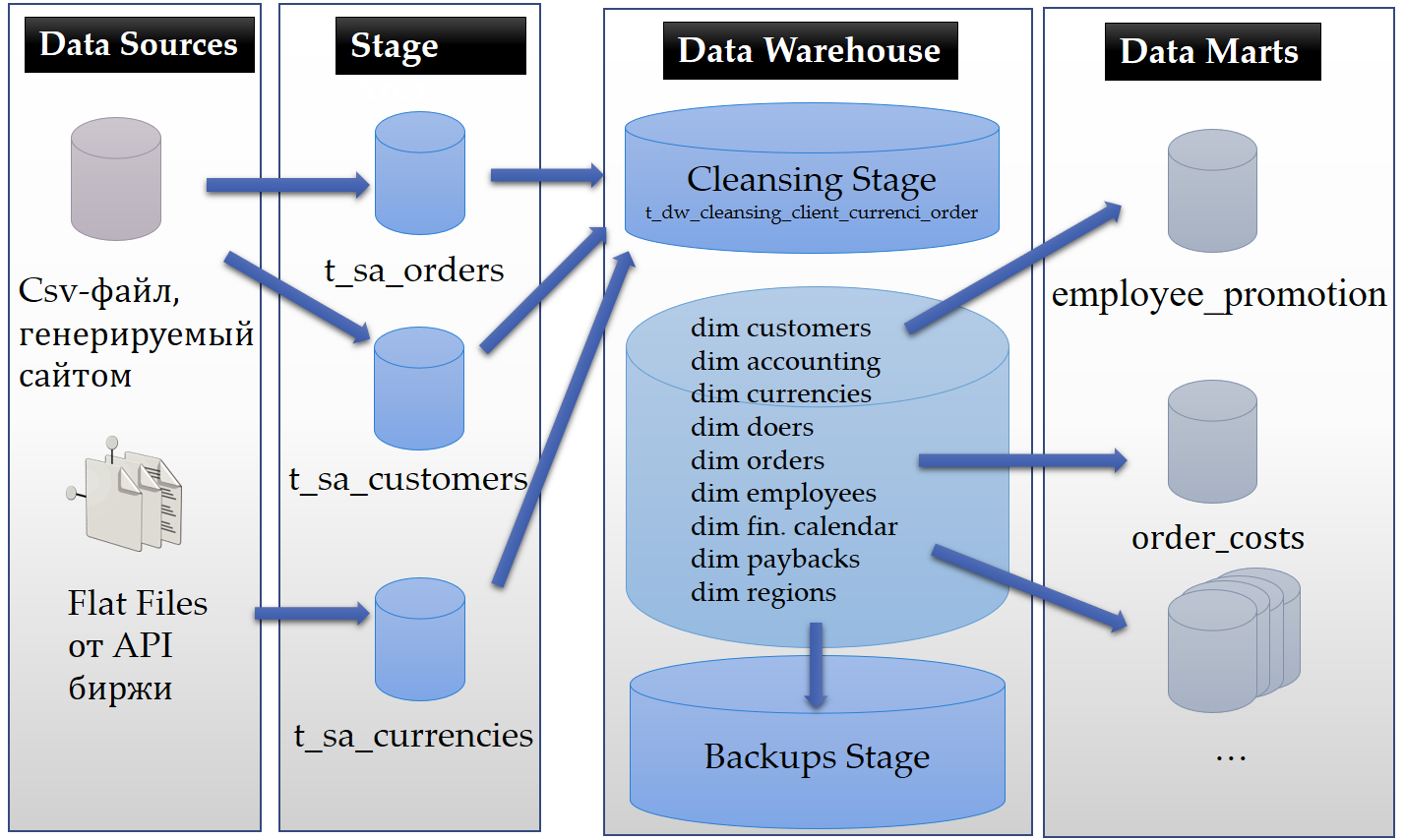
Characteristics of the presented scheme:

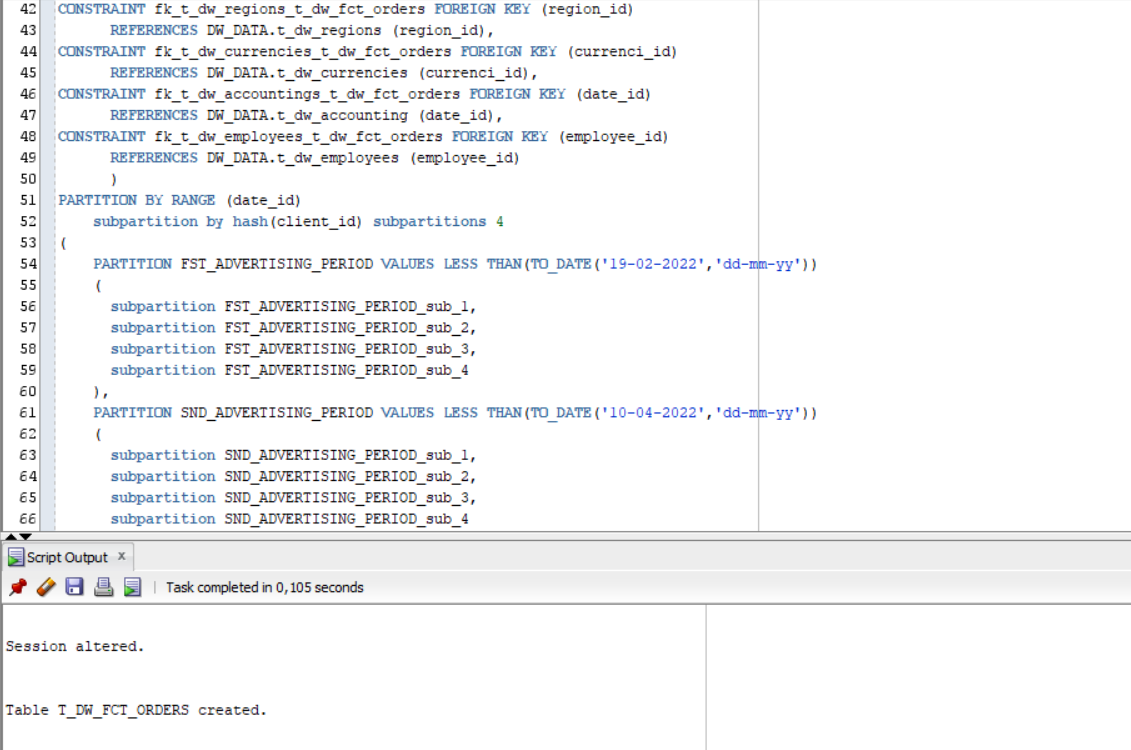
* Dimension tables joined to a fact table using a foreign key.
* Dimension tables not connected to each other, except for the table with the company's financial calendar and the table with the main financial indicators, to create convenient selections by period.
* STAR scheme is easy to understand and provides optimal memory usage.
* Scheme is widely supported by BI Tools.

To improve security and access separation, it proposed to use several tablespaces in the storage, with a separate user for each. The storage tier structure shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| 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) | Loaded from csv file, contains first name, last name, patronymic, phone number, payment status (paid / not) |
| SA\_ORDERS | ts\_sa\_orders\_data\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Loaded from a csv file, contains the product name, order amount, order status (accepted/purchased/delivered to the client) |
| SA\_CURRENCIES | ts\_sa\_currencies\_data\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 50M,  Autoextend clause ON next 10M) | Loaded from a file, contains the name of the currency, direct conversion rate and reverse conversion rate |
| DW - Cleansing Level | DW\_CL | ts\_dw\_cl\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 100M) | Loads information from the Storage level, prepares it for further cleaning |
| DW – Level | DW\_DATA | ts\_dw\_data\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Loads information from Cleansing level, normalizes data. |
| 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) | Loads information from the DW level. Contains representations (view), combining objects from the DW level. |
| STAR - Cleansing | SAL\_CL | ts\_sal\_cl\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Loads information from the DW\_CL level. Contains views of the previous level, but without redundancy. |
| STAR – Level | DM\_EMPLOYEES | ts\_dm\_employees\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 50M,  Autoextend clause ON next 10M) | Stores information about dim employees |
| DM\_CUSTOMERS | ts\_dm\_customers\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Stores information about dim customers |
| DM\_ORDERS | ts\_dm\_orders\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Stores fact information (in t\_dw\_fct\_orders table) |
| DM\_CURRENCIES | ts\_dm\_currencies\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 100M,  Autoextend clause ON next 10M) | Stores information about dim currencies |
| DM\_DOERS | ts\_dm\_doers\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 50M,  Autoextend clause ON next 5M) | Stores information about dim doers |
| DM\_PAYBACKS | ts\_dm\_paybacks\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 150M,  Autoextend clause ON next 50M) | Stores information about dim paybacks |
| DM\_ACCOUNTINGS | ts\_dm\_accounting\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 100M,  Autoextend clause ON next 10M) | Stores information about dim accounting and dim financial calendar |
| DM\_REGIONS | ts\_dm\_regions\_01  (AUTOALLOCATE,  SEGMENT SPACE MANAGEMENT AUTO,  LOGGING,  Size 70M,  Autoextend clause ON next 15M) | Stores information about dim regions |

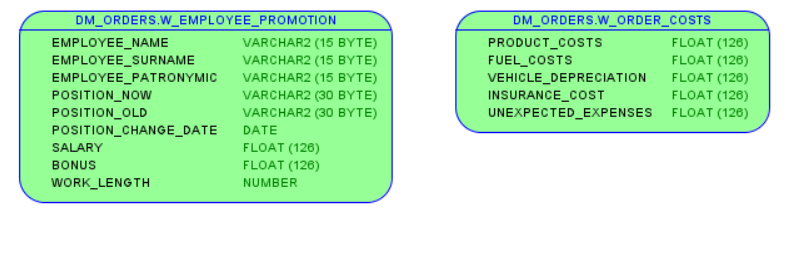
DataFlow of the storage is presented below:



In storage, it is recommended to use partitioning by time periods (for example, advertising campaigns) and hash subpartitioning (for example, client\_id) to optimize requests and disk space.****

Since dimentions change most often in the system: client, order and employee (because the company is large). Therefore, it is most logical to use parallel computing in the DW, CL and SA levels to update data about them. However, parallel computing can be used also in DM - data mart levels, when designing views related to finance. Because they are frequently updates due to the large number of transactions, and accountants, marketers, etc. reliable and up-to-date information is required.

The most used views in business are listed below.



**Conclusion**

As a conclusion, the presented technical solution will help the business to get additional profit on:

1. Simplification of accounting by storing data in a single system and saving the history of data changes.
2. Increase conversion by accelerating the interaction of the client with the site (by accelerating the response from the database).
3. Reducing the cost of physical media due to the ability to use more powerful media for fresh information and slow, old media for historical information.
4. Increasing the security of data from leaks and unauthorized changes, which means incurring reputational and material costs.
5. 5. Improving the efficiency of the marketing department due to the ability to create views with aggregated information for decision making.