**Solution concept.**

1. **Data of interest.**

**1.1 Business background.**

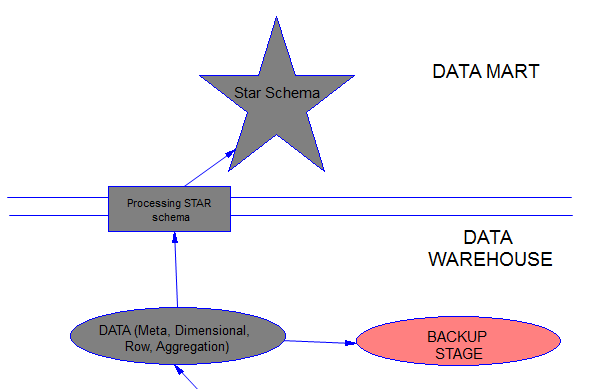
Assume that there is a company «Pear», which manufactures and sells equipment and software (CD players, DVD players, Cameras, Computers, Printers, Peripherals and Accessories, Software and etc.). The company «Pear» recently made ​​a breakthrough and was able to penetrate world markets. As a result, sales volume significantly increased, and the flow of information grew up to 1 million rows per day. In this connection it became necessary to introduce a tool for analyzing historical data to determine the positive / negative trends in sales worldwide acording to the company’s perposes.

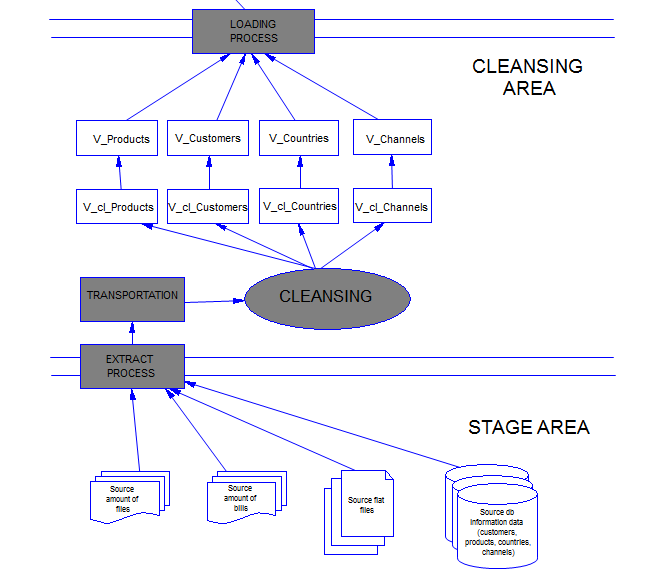
**1.2 Benefits.**

Having spent some time in the world markets in the normal mode, the top management of the company came to the conclusion that for the further successful promotion of the brand «Pear» it is necessary to conduct a thorough analysis of the accumulated statistics on sales of the company's products in world markets, and to use the tools of BI. The main task for the analyst will be the statistical analysis of sales of the company (by category of goods), by countries and regions, so that regional offices of the company where able to adjust their actions in the market and apply the possible economic measures to increase company profits. The primary task of leaders recommended that a further analysis of sales statistics to determine which age groups of customers buy a particular product category (optional).

**1.3 Data sources and sketch diagrams.**

Data sources include all sources (documents, bills, db. elements, reports), that may contain important information for our purposes. Basing on these sources cleaning tables will be produced to prepare the whole amount of all possible information. After this step all the required information will be available for further usage in DWH and, later, while producing star and snowflake models. For our project it is advisable to prepare following sketch diagram, describing processes of data collecting and star diagram producing.





# Requirements.

# Business Requirements

* colculate information about sales (amount, quantity) monthly for each country;
* colculate information about customers monthly;
* colculate information about products monthly;
* control information about total sales for each product category for each country raising alarm, if it’s less then 5% from region sales for this product;

# Technical Requirements

* High performance;
* High availability;
* High reliability;
* System must process all required source data from all mentioned below sources;
* System must perform information according to the company’s standards;
* All the information must be available according to the company’s security Policy.

3.2. **Add time to key.**

Our business data model is a “point-in-time” model. That is, the model portrays the business at the present. The data warehouse data model, on the other hand, is an “over-time” model. An “over-time” model portrays an enterprise with a historical perspective. Since the data warehouse is time variant, this is the type of model that is appropriate for the data warehouse.

Adding a time component to the key enables a distinct record to be created each time a new snapshot is taken and each time data for a new period is captured. Inclusion of this time component has a significant design implication. Since the entity may be the parent in some one-to-many relationships, the generated foreign key would now include the time component. It generates a new occurrence for every child, even if no data in the child entity changed.

For our project the Times table will be used to set time key for each sale all over the world. And it will collect time divisions from years up to days (may be even fiscal time divisions according to the proposal). And uniqueness of these sales will be provided with usage identifier of product and time id of each sale.

3.3. **Add Derived Data.**

The reason that this step is to perform the business impact—to ensure consistency; performance benefits are secondary. One of the common objectives of a data warehouse is to provide data in a way so that everyone has the same facts—and the same understanding of those facts. That’s why in the fact table the calculated columns must be entered: **the quantity of sold products and the amount of sold products.**

3.4. **Determine Granularity Level.**

**Current business need.** The primary determining factor according to the requirements should be the business need. At a minimum, the level of granularity must be sufficient to provide answers to each and every business question being addressed within the scope of the data warehouse iteration. For our project the minimum analyzing time period is set **up to the month aggregating to quarters and then to years**. Providing a greater level of granularity adds to the cost of the warehouse and the development project and, if the business does not need the details, the increased costs add no business value.

**Anticipated business need.** The future business needs should also be considered. A common scenario is for the initial data warehouse implementation to focus on monthly data, with an intention to eventually obtain daily data. If only monthly data is captured, the company may never be able to obtain the daily granularity that is subsequently requested. Therefore, if the interview process reveals a need for daily data at some point in the future, it should be considered in the data warehouse design.

It is also possible to provide so more granularity levels: product subcategories up to categories and (optional) customers according to

3.5. **Summarize Data.**

Summaries for Period of Time Data.

Simple accumulations and rolling summaries apply to data that pertains to a period of time. Simple accumulations represent the summation of data over one of its attributes, such as time. For example, a monthly sales summary provides a summary of all sales for the month across the common ways that people access it. Managers often need to have sales quantity and amounts by month, quarter, year, sales person, store, and product subcategory and category, the summary table.

3.6. **Merge Entities.**

The sixth step in developing the data warehouse model is to merge entities by combining two or more entities into one. According to the proposal merging will be introduced for products->subcategories->categories also birth data up to year groups (0-20;21-30;31-50;51-rest). Merging the entities improves the data delivery process performance by reducing the number of joins, and also enhances consistency. Merging entities is a form of denormalizing data and, in its ultimate form, it entails the creation of conformed dimensions for subsequent use in the data mart.

3.7. **Create Arrays.**

In our case it is useful to create an array to represent sales according to different year groups aiming to provide deep analyses of possible influence factors on sales for these age groups. The same idea is about subcategories and categories of products. It is important to make sure that our arrays are useful. The arrays are useful if all of the following conditions exist:

* The number of occurrences is relatively small. In the example cited above, there are five occurrences. Creating an array for sales at each of 50 regions would be inappropriate.
* The occurrences are frequently used together. In the example, when accounts receivable analysis is performed, people often look at the amount in each of the five categories together.
* The number of occurrences is predictable. In the example, there are always exactly five occurrences.
* The pattern of insertion and deletion is stable. In the example, all of the data is updated at the same time. Having an array of quarterly sales data would be inappropriate since the data for each of the quarters is inserted at a different time. In keeping with the data warehouse philosophy of inserting rows for data changes, there would actually be four rows by the end of the year, with null values in several of the rows for data that did not exist when the row was created.

Our arrays satisfy all this conditions and may be used for creating data warehouse model.

3.8. **Segregate Data.**

This last transformation step recognizes that data in the operational environment changes at different times, and therefore groups’ data into sets based on insertion patterns. If taken to the extreme, a separate entity would be created for each piece of data. That approach will maximize the efficiency of the data acquisition process and result in some disk space savings. For our purposes it will be useful to introduce segregation levels for time (month, quarter, year) and for products (subcategories and categories).