**Legal Notice:** This document contains privileged and/or confidential information and may not be disclosed, distributed or reproduced without the prior written permission of Datamola.

U1.Exit Task.Oracle DB.Introduction to Data Warehousing

Alina Sadovskaya

Contents

[2. Overview 4](#_Toc45633664)

[2.1. Business Background 4](#_Toc45633665)

[2.2. Benefits 4](#_Toc45633666)

[3. Requirements 4](#_Toc45633667)

[3.1. Business Requirements 4](#_Toc45633668)

[3.2. Technical Requirements 5](#_Toc45633669)

[3.3. Functional Requirements 5](#_Toc45633670)

[4. Solution Sketch 5](#_Toc45633671)

[4.1. Reporting Mockups 6](#_Toc45633672)

[4.2. DWH Solution Sketch 6](#_Toc45633673)

[5. DWH Solution Concept 6](#_Toc45633674)

[5.1. Logical Diagram 6](#_Toc45633675)

[5.2. Physical diagram 6](#_Toc45633676)

[5.3. Dimensions 6](#_Toc45633677)

[5.3.1. Dimension Types 6](#_Toc45633678)

[5.3.2. Dimension Hierarchies 7](#_Toc45633679)

[5.4. Facts 7](#_Toc45633680)

[5.4.1. Facts Aggregations 7](#_Toc45633681)

[5.5. Dataflow Diagram 7](#_Toc45633682)

[6. Open Questions and Risks 7](#_Toc45633683)

[7. Estimates 7](#_Toc45633684)

# 2. Overview

As an example of a business idea, consider the work of the Belarusian clothing brand Mark Formelle.

## 2.1. Business Background

For 10 years, MARK FORMELLE has been producing clothing for home, life and sports, helping customers lift their spirits and giving them ease and self-confidence. The company's products are successfully sold not only on the territory of Belarus, but also far beyond its borders. The production process involves both traditional cotton, viscose, and synthetic materials, but also uses the latest developments of fabrics with an admixture of polyester, as well as bamboo and modal. The company's professionals participate in the development and design. They also come up with their own prints and drawings. All masters have specialized education and have been trained in the best institutions. The main vectors of activity are: clothing for everyday life, work and leisure called Life, options for active athletes line Sport, insulated underwear line Zima, as well as underwear, socks, stockings, tights.

Let this clothing brand face certain difficulties related to incorrect data handling. Some of the possible problems:

1. lack of business intelligence from multiple sources;

2. reduced query and system performance;

3. lack of timely access to data;

4. the lack of historical intelligence.

## 2.2. Benefits

We offer the customer to use DWH for the following reasons:

1. It is better to make decisions. Corporate decision makers will no longer have to make important business decisions based on limited data and guesswork. The data warehouse will store reliable facts and statistics, and decision makers will be able to extract this information from the data warehouse based on their personal needs.

2. Quick and easy access to data. Speed is an important factor that puts the company above its competitors. Business users can quickly access data from multiple sources from the data warehouse, which means that precious time will not be wasted extracting data from multiple sources. This allows the company to make fast and accurate decisions, with little or no support from the it Department.

3. The quality and consistency of data. As data warehouses collect information from various sources and convert it into a single and widely used format, departments will produce results that are consistent and consistent with each other. When data is standardized, a company can be confident in its accuracy, and accurate data is what makes it possible to make strong business decisions.

# 3. Requirements

## 3.1. Business Requirements

| **ID** | **Description** |
| --- | --- |
| B01 | Сalculation of sales information (Volume, Quantity) on a monthly basis(you can collect information for large time periods due to the time hierarchy) for each product |
| B02 | Сalculation of information about the benefits of used shares for products on a monthly basis |
| B03 | Control information about total sales for each product category for each country of concern if it is less than 3% of the region's sales for that product |
| B04 | Our brand has several stores in different countries. Each office has its own geographical location. The warehouse must take this fact into account |
| B05 | Storing information about products in the warehouse, for quick search of the selected product |
| B06 | Each retail sale must be associated with a valid customer |
| B07 | Retail sales are always linked to the payment method and associated with a specific employee |

## 3.2. Technical Requirements

| **ID** | **Description** |
| --- | --- |
| T01 | Keep info from the beginning of business |
| T02 | Clean products that have been discontinued |
| Т03 | Тhe access time to the storage– 24/7/365 |
| Т04 | High performance and high availability |
| T05 | All the information must be protected according to the company’s security Policy |
| T06 | Ability to process large amounts of information per day (1 million rows) |

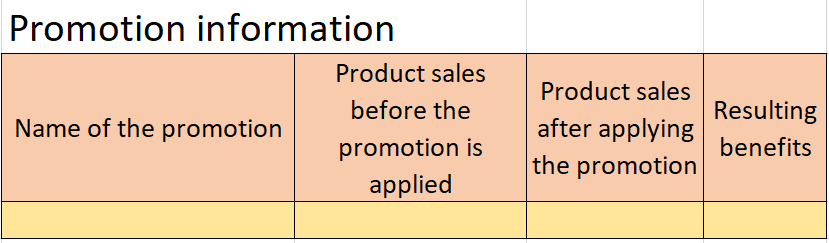
## 3.3. Functional Requirements

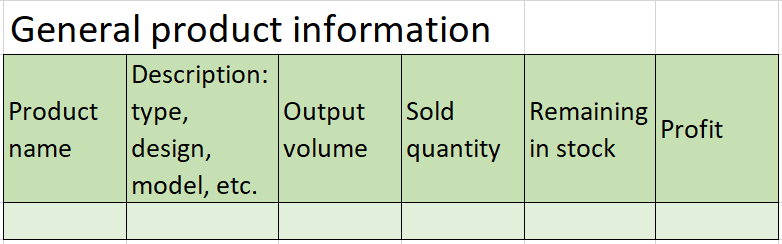
| **ID** | **Description** |
| --- | --- |
| F01 | Ability to load data from third-party resources must be implemented |
| F02 | The software automatically checks that clients meet all necessary requirements |
| F03 | The Sales system should allow users to record customers sales |
| F04 | The software automatically checks the availability of products in stock |
| F05 | Only Managerial level employees have the right to view revenue data |
| F06 | The software system should be integrated with banking API |
| F07 | Sales reports should be generated every day (24 hours) |

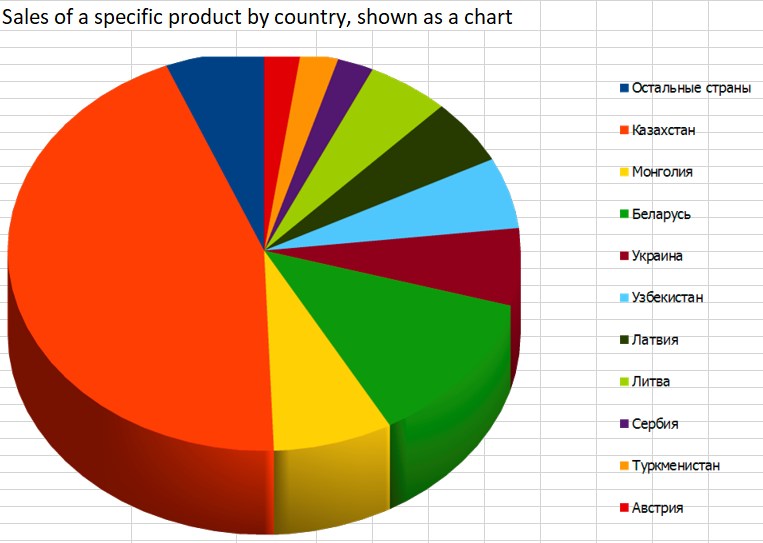
# 4. Solution Sketch

## 4.1. Reporting Mockups

At the output, the customer will receive the following data:

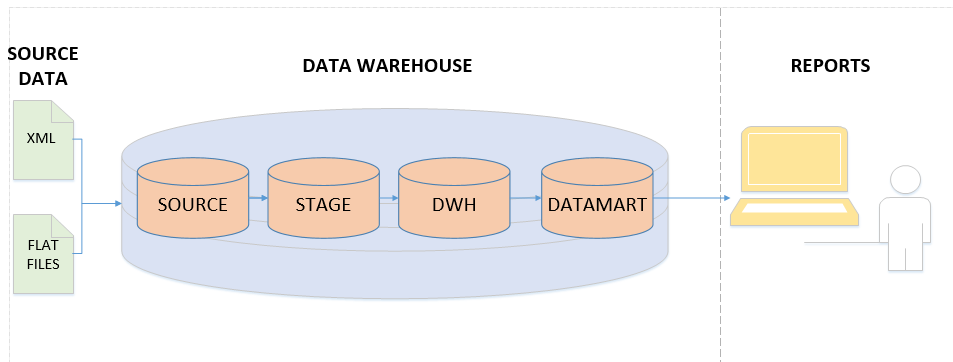






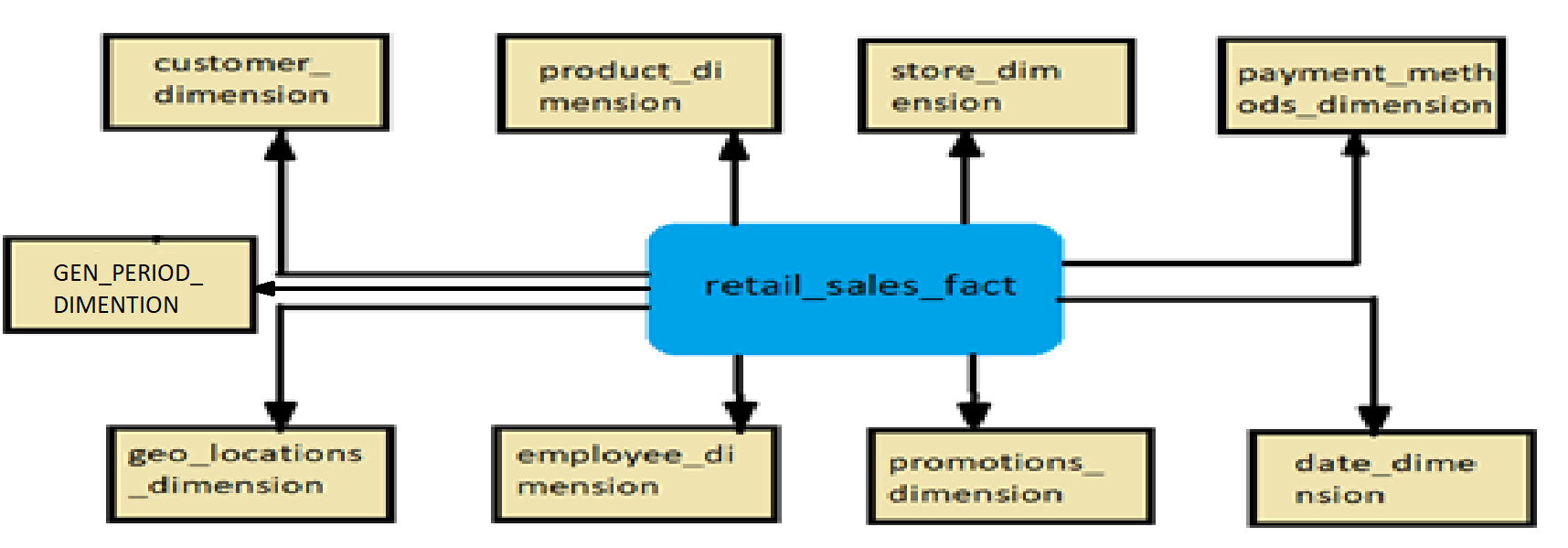
And a complete table with all product sales data.

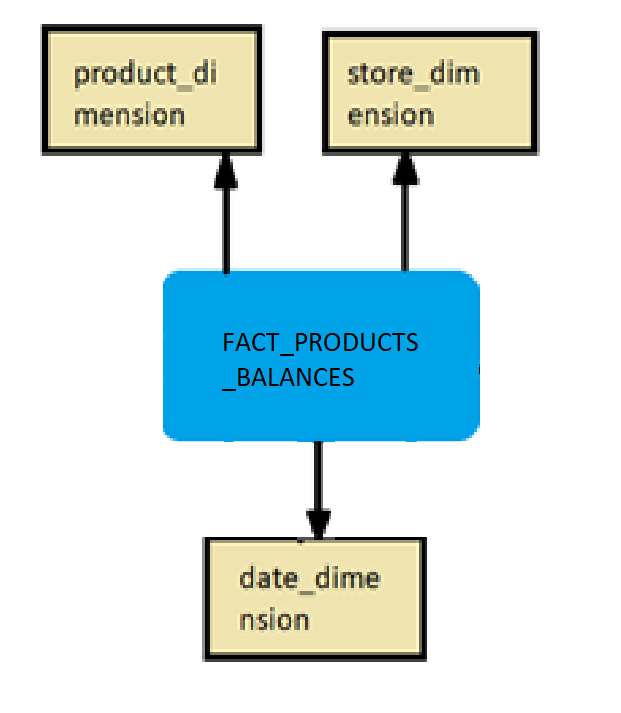
## 4.2. DWH Solution Sketch



# 5. DWH Solution Concept

## 5.1. Logical Diagram

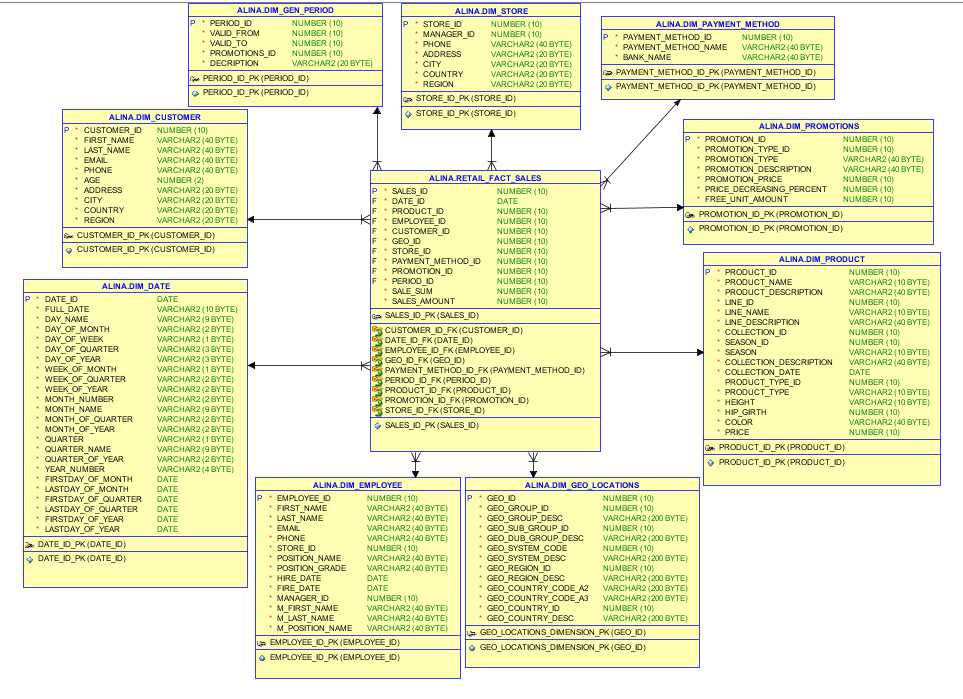




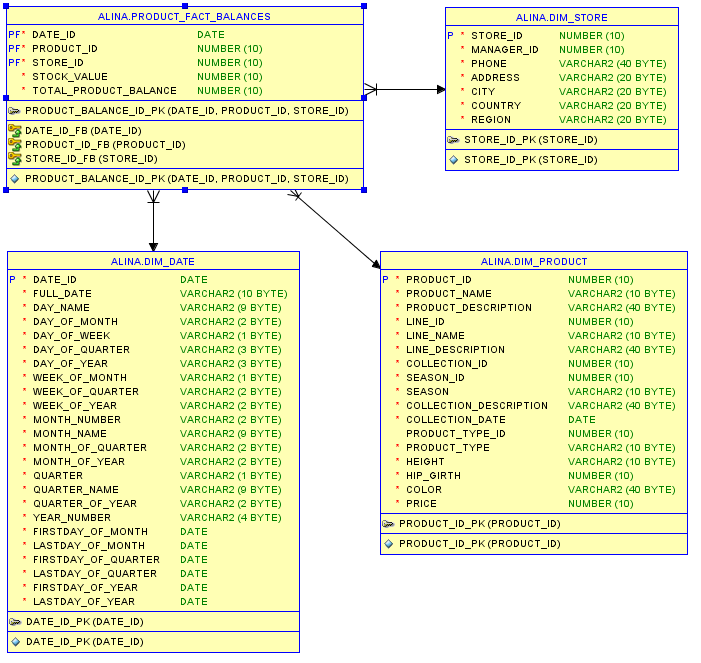
## 5.2. Physical diagram

As a result we have the following **STAR SCHEME**:

1)fact table – RETAIL\_SAILES\_FACT



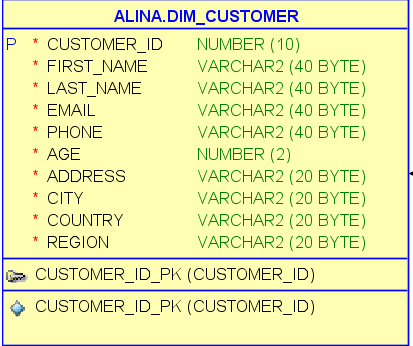
2)fact table –FACT\_PRODUCT\_BALANCE



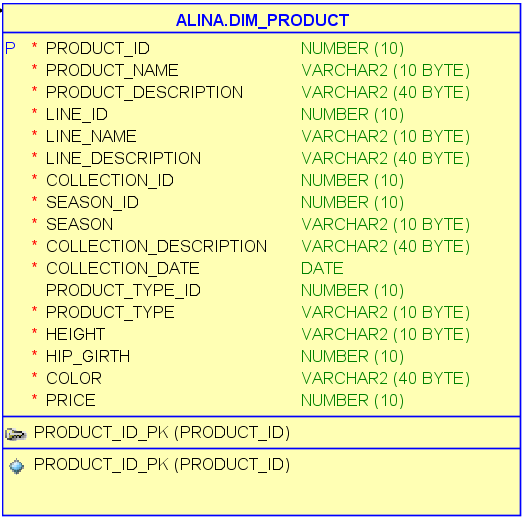
## 5.3. Dimensions

Schema should contains next dimensions:

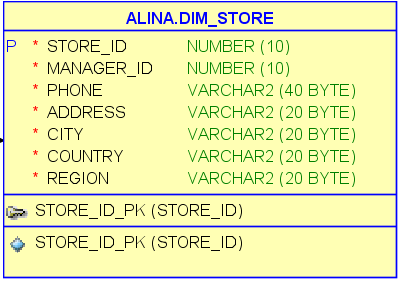
1. Customer dimension (Primary Key is Customer\_ID)



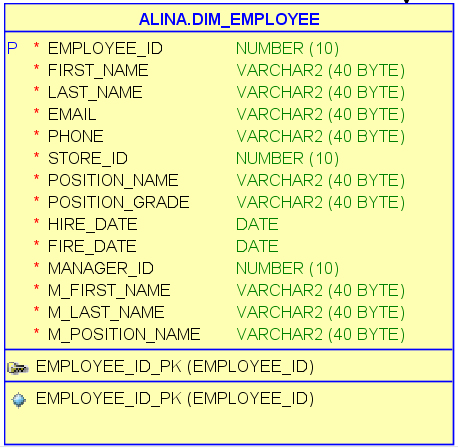
2. Product dimension (Primary Key is Product\_ID)



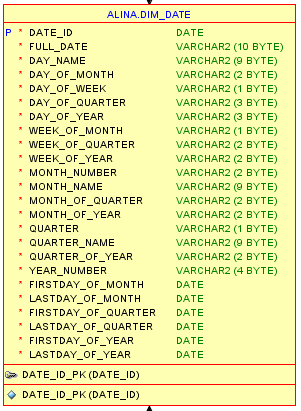
3. Store dimension (Primary Key is Store\_ID)



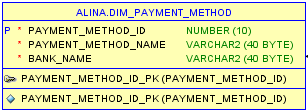
4. Employee dimension (Primary Key is Employee\_ID)



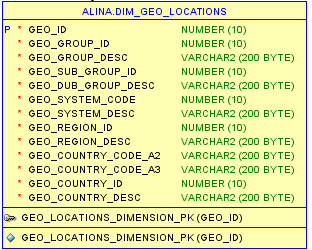
5. Date dimension (Primary Key is Date\_ID)



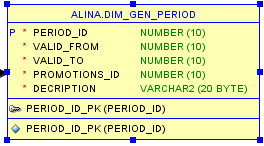
6. PAYMENT METHOD DIMENTION



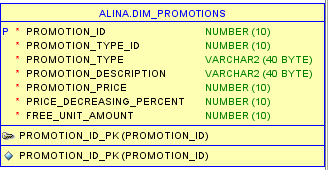
7. GEO LOCATION DIMENTION



8. GEN PERIOD DIMENTION



9. PROMOTION DIMENTION



5.3.1. Dimension Types

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Type | Size | DW – Merged Dimensions | Descriptions |
| DIM\_GEN\_TIMES | SCD1 | BIG | DW.T\_DAYS, DW.T\_WEEKS, DW.T\_MONTHS, DW.T\_QUARTERS,  DW.T\_YEARS | TBD – Example row |
| DIM\_GEO\_LOCATIONS | SCD1 | SMALL | Geo\_id  Geo\_group\_id  Geo\_group\_desc  Geo\_sub\_group\_id  Geo\_dub\_group\_desc  Geo\_system\_code  Geo\_system\_desc  Geo\_region\_id  Geo\_region\_desc  Geo\_country\_code\_a2  Geo\_country\_code\_a3  Geo\_country\_id  Geo\_country\_desc | This kind of dimension contains information about all countries, subregions, regions of the world where the company's stores are located. And also enters information on the types of economic development and unions according to the international classification. |
| DIM\_CUSTOMER | SCD1 | BIG | customer\_id  first\_name  last\_name  email  phone  age  address  city  country  region | This kind of dimension contains detailed information about clients (including age and address of the residence for informative presentation). |
| DIM\_PRODUCT | SCD1 | BIG | PRODUCT\_ID  PRODUCT\_NAME  PRODUCT\_DESCRIPTION  LINE\_ID  LINE\_NAME  LINE\_DESCRIPTION  COLLECTION\_ID  SEASON\_ID  SEASON  COLLECTION\_DESCRIPTION  COLLECTION\_DATE  PRODUCT\_TYPE\_ID  PRODUCT\_TYPE  HEIGHT  HIP\_GIRTH  COLOR  PRICE | This kind of dimension contains detailed information about the company's products, including the name of an individual product, category and subcategory. To do so, provided the opportunity for dimension Type SCD 2 perfectly partitions history because each detailed version of a dimensional entity is correctly connected to the span of fact table records for which that version is exactly correct |
| DIM\_STORE | SCD1 | SMALL | store\_id  manager\_id  phone  address  city  country  region | Provides information about store including the store's address, it’s phone and information about the Manager |
| DIM\_GEN\_PERIOD | SCD2 | BIG | period\_id  valid\_from  valid\_to  promotions\_id  decription | A specific type of dimension that allows grouping facts based on logic (the duration of product discounts). |
| DIM\_PAYMENT\_METHODS | SCD1 | SMALL | PAYMENT\_METHOD\_ID  PAYMENT\_METHOD\_NAME  BANK\_NAME | Provides information about the payment method used |
| DIM\_EMPLOYEE | SCD1 | SMALL | employee\_id  first\_name  last\_name  email  phone  store\_id  POSITION\_NAME  POSITION\_GRADE  HIRE\_DATE  FIRE\_DATE  MANAGER\_ID M\_FIRST\_NAME  M\_LAST\_NAME  M\_POSITION\_NAME | Provides information about the company's employees(including theres Manager) |
| DIM\_PROMOTIONS | SCD1 | SMALL | PROMOTION\_ID  PROMOTION\_TYPE\_ID  PROMOTION\_TYPE  PROMOTION\_DESCRIPTION  PROMOTION\_ PERCENT PRICE\_DECREASING\_PERCENT  FREE\_UNIT\_AMOUNT | provides information about discounts, their description, amount, and so on |
| DIM\_DATE | SCD1 | BIG | date\_id  full\_date  day\_name  day\_of\_month  day\_of\_week  day\_of\_quarter  day\_of\_year  week\_of\_month  week\_of\_quarter  week\_of\_year  month\_number  month\_name month\_of\_quarter  month\_of\_year  quarter\_name  quarter\_of\_year  year\_number  firstday\_of\_month  lastday\_of\_month  firstday\_of\_quarter  lastday\_of\_quarter  firstday\_of\_year lastday\_of\_year | These dimention consists days, weeks, months, quarters and so on. |

### 

### 5.3.2. Dimension Hierarchies

**DATE\_DIMENSION:**

**Hierarchy DAY-WEEK-MONTH-YEAR**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| **DAY** | DAY | Store all day at the week | DAY\_OF\_WEEK |
| **WEEK** | WEEK | Store all weeks at the month | WEEK\_OF\_MONTH |
| **MONTH** | MONTH | Store all months at the year | MONTH\_OF\_YEAR |
| **YEAR** | YEAR | Store all years | YEAR\_NUMBER |

**Hierarchy DAY-MONTH- QUARTER -YEAR**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| **DAY** | DAY | Store all day at the month | DAY\_ OF\_MONTH |
| **MONTHS** | MONTH | Store all months at the quarter | MONTH\_ OF\_ QUARTER |
| **QUARTER** | QUARTER | Store all quarters at the year | QUARTER\_ OF\_YEAR |
| **YEAR** | YEAR | Store all years | YEAR\_NUMBER |

**Hierarchy DAY - QUARTER -YEAR**

|  |  |  |  |
| --- | --- | --- | --- |
| **DAY** | DAY | Store all day at the quarter | DAY\_OF\_ QUARTER |
| **QUARTER** | QUARTER | Store all quarters at the calendar year | QUARTER\_ OF\_YEAR |
| **YEAR** | YEAR | Store all years | YEAR\_NUMBER |

**CUSTOMER\_DIMENSION:**

**Hierarchy PRODUCT–TYPE**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| **PRODUCT** | PRODUCT\_NAME | Store all possible products for each type. | PRODUCT\_ID |
| **TYPE** | PRODUCT\_TYPE | Store all product types of our company. | PRODUCT\_TYPE \_ID |

**Hierarchy PRODUCT–LINE-COLLECTION**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| **PRODUCT** | PRODUCT\_NAME | Store all possible products for each LINE. | PRODUCT\_ID |
| **LINE** | LINE\_NAME | Store all possible LINES for each COLLECTIONS. | LINE\_ID |
| **COLLECTION** | COLLECTION\_NAME | Store all COLLECTIONS of our company. | COLLECTION\_ID |

**GEO\_LOCATIONS\_DIMENSION:**

**Hierarchy COUNTRY – REGION - GEO\_GROUP - GEO\_SUB\_GROUP**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| COUNTRIES | GEO\_COUNTRY | Store all countries for each region. | GEO\_COUNTRY\_ID |
| REGIONS | GEO\_REGION | Store all regions for each GEO\_GROUP . | GEO\_REGION\_ID |
| GEO\_GROUP | GEO\_GROUP | Store all regions for each GEO\_SUB\_GROUP. | GEO\_GROUP\_ID |
| GEO\_SUB\_GROUP | GEO\_SUB\_GROUP | Store all Geo\_sub\_group of the world. | GEO\_SUB\_GROUP\_ID |

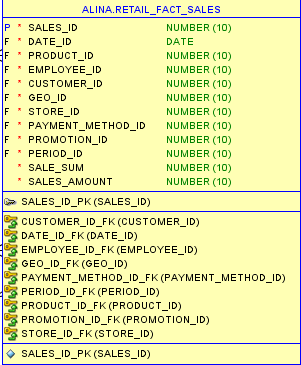
**EMPLOYEE\_DIMENSION**

**Hierarchy EMPLOYEE -** **MANAGER**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| **EMPLOYEE** | EMPLOYEE | Store all employees for each Manager | EMPLOYEE\_ID |
| **MANAGER** | MANAGER | Store all Manager of the company. | MANAGER\_ID |

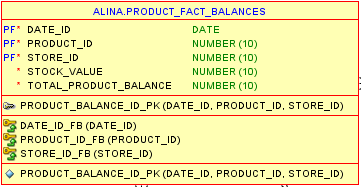
## 5.4. Facts

## 1) Retail\_fact\_ sales



I think it is necessary to add another fact table that stores information about the remaining products in stock.

2) FACT\_PRODUCT\_BALANCES



### 5.4.1. Facts Aggregations

* + Facts Aggregations for table RETAIL\_SALES\_FACT

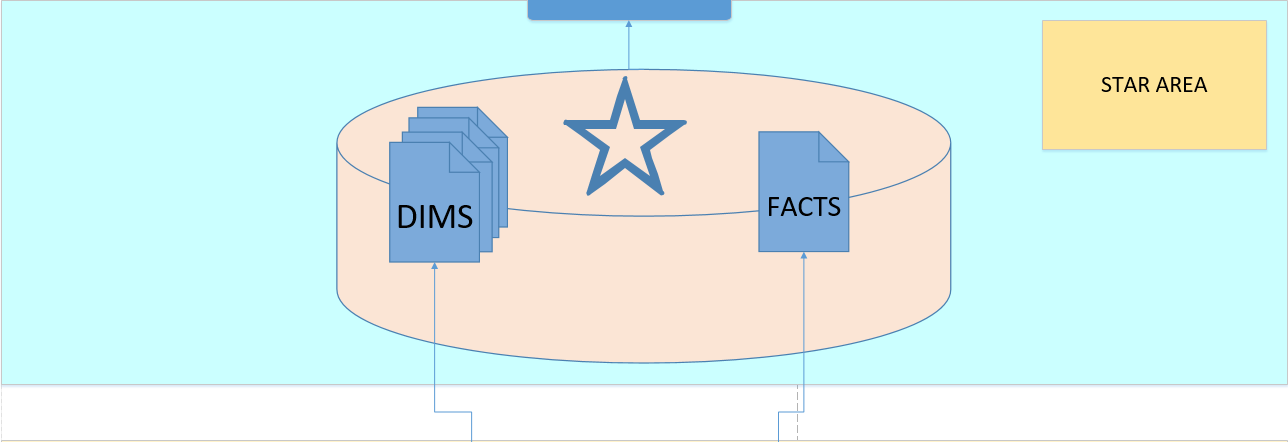
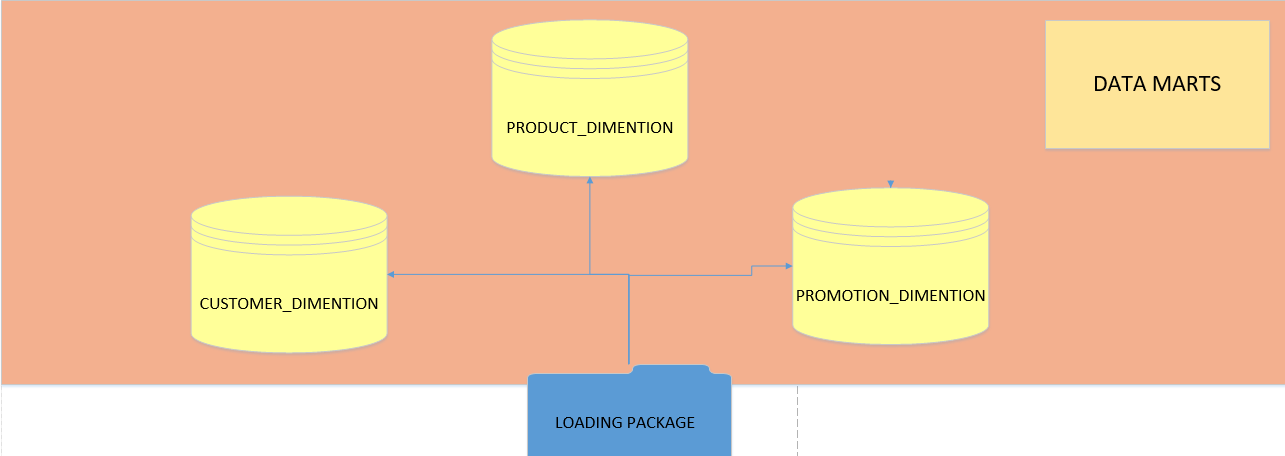
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Code | Table Name | Additive | Descriptions |
| value of the total sum of products sold | SALE\_SUM | RETAIL\_FACT \_SALES | + | Calculate the total amount of sales in the selected period. |
| value of the total number of products sold | SALES\_AMOUNT | RETAIL\_FACT \_SALES | + | Сalculates the number of sales for a specific period |

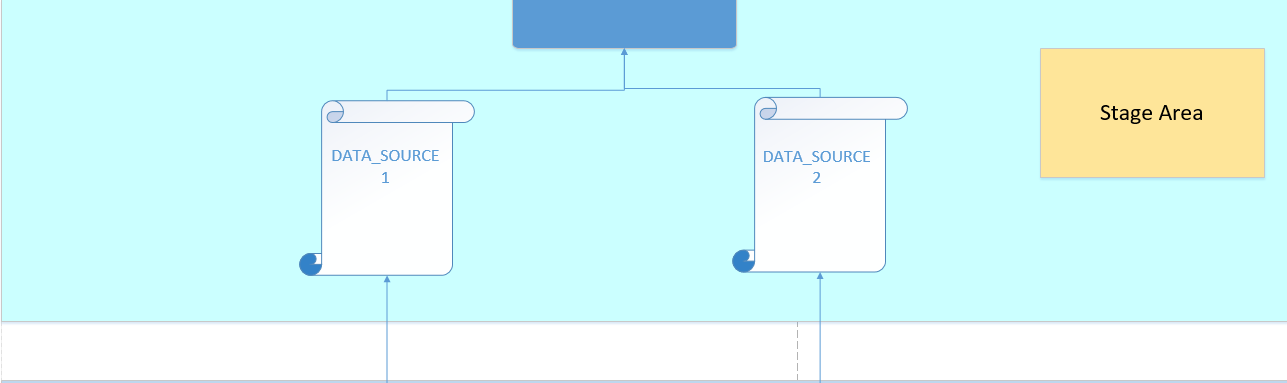
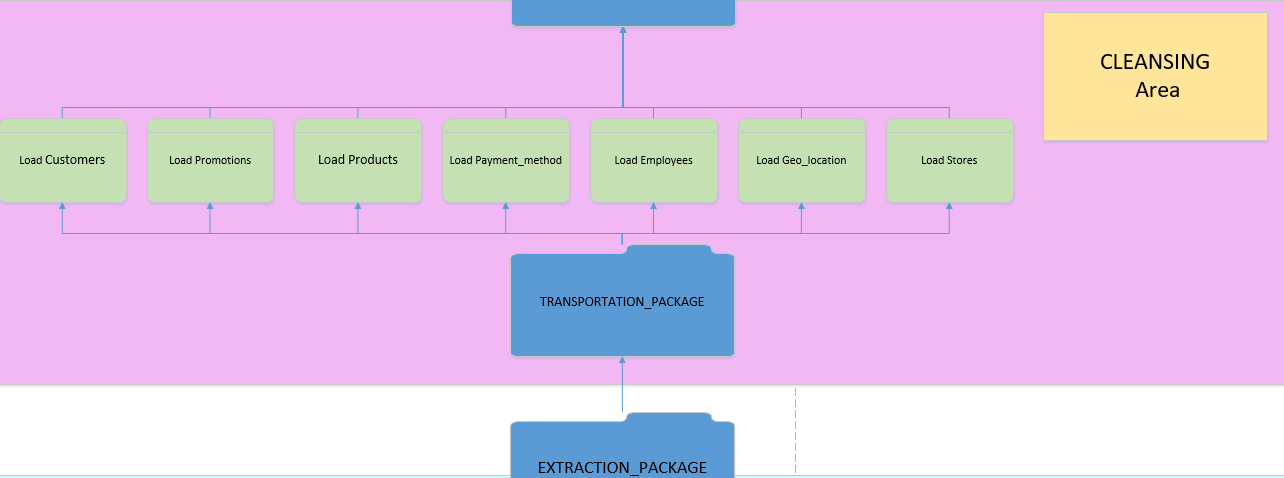
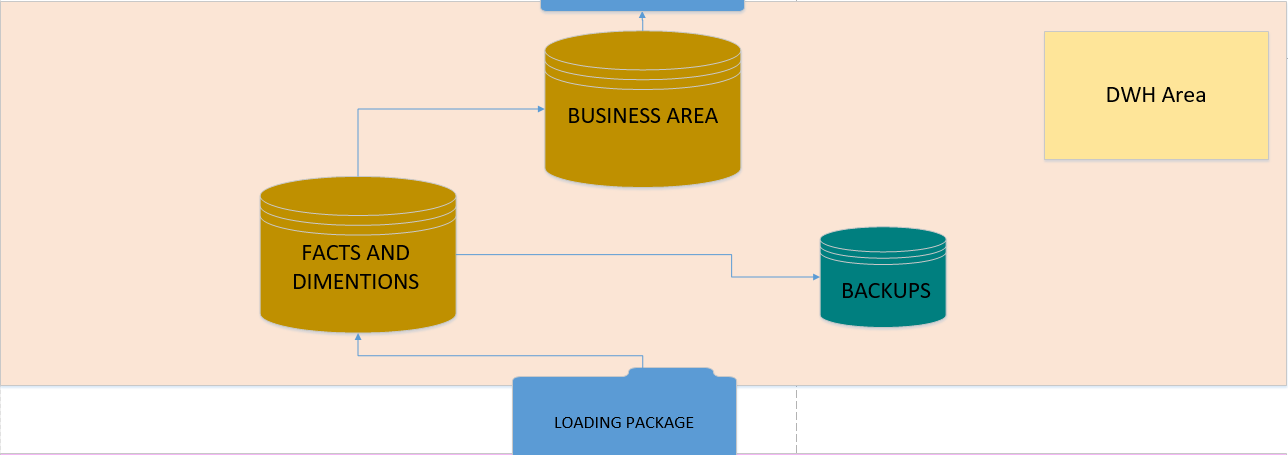
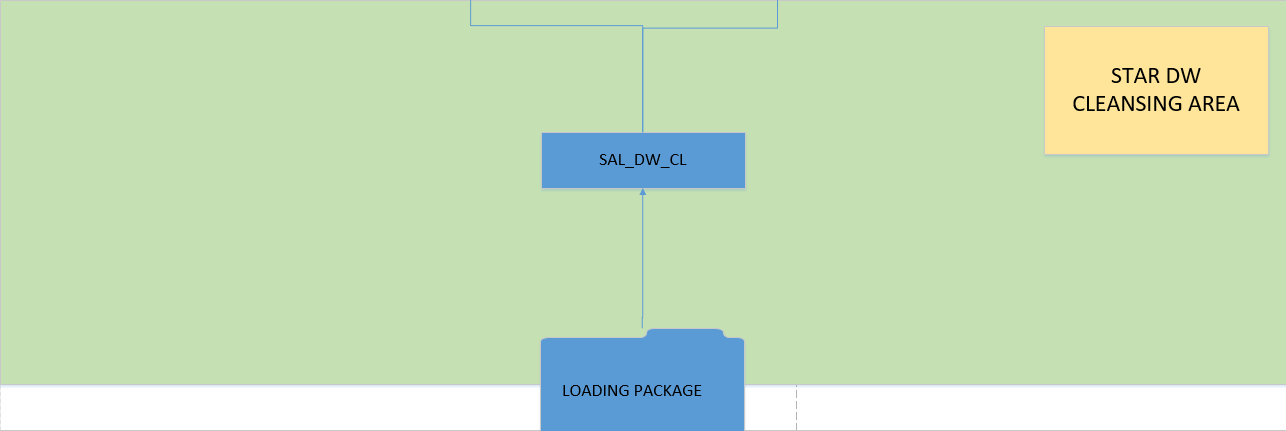
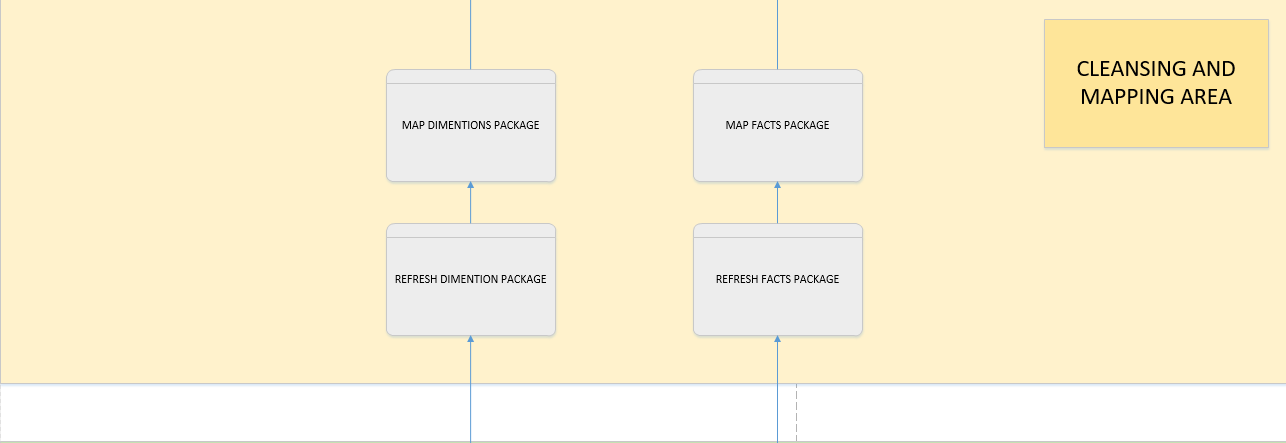
* + Facts Aggregations for table FACT\_PRODUCT\_BALANCES

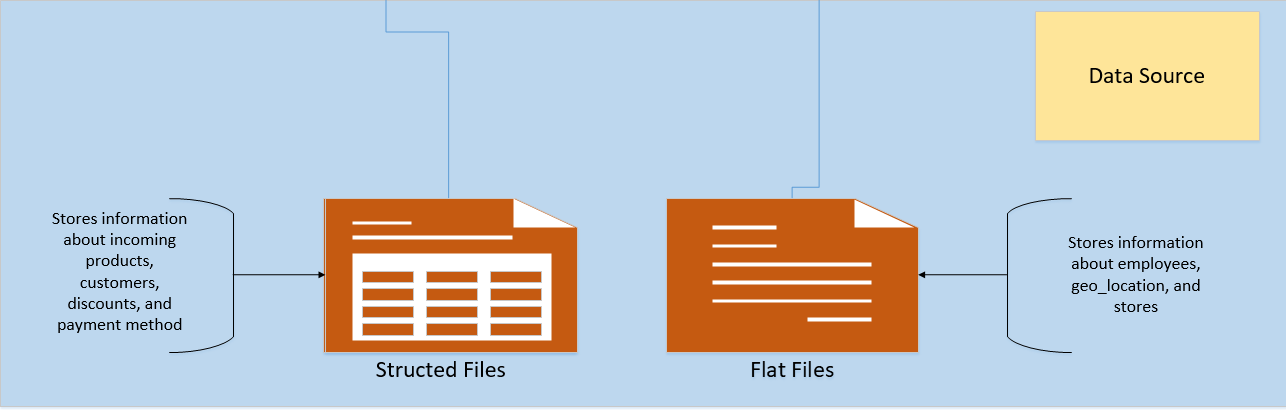
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Code | Table Name | Additive | Descriptions |
| тumber of products in stock | STOCK\_VALUE | PRODUCT\_FACT\_BALANCES | + | Calculate amount of products in stock |
| total cost of the remaining product | TOTAL\_PRODUCT\_BALANCE | PRODUCT\_FACT\_BALANCES | + | Сalculate the total cost of the remaining product |

## 5.5. Dataflow Diagram

DataFlow Diagram to describe refresh process of mine Business STAR:







## 

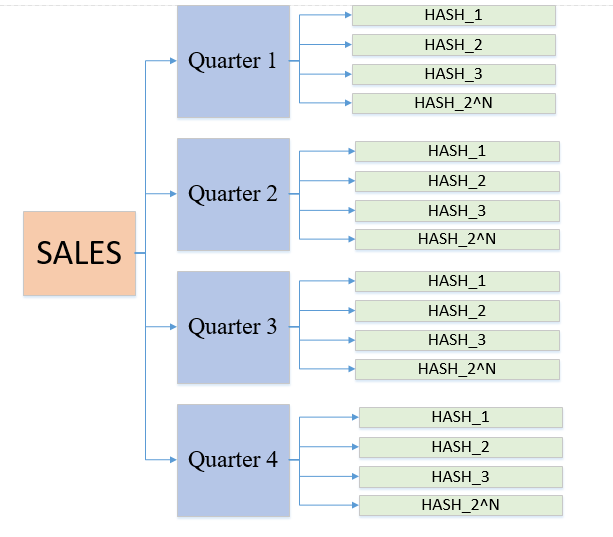
## 5.6. Partitioning Facts

The query execution speed, which works with fact table, can be increased with partitioning.

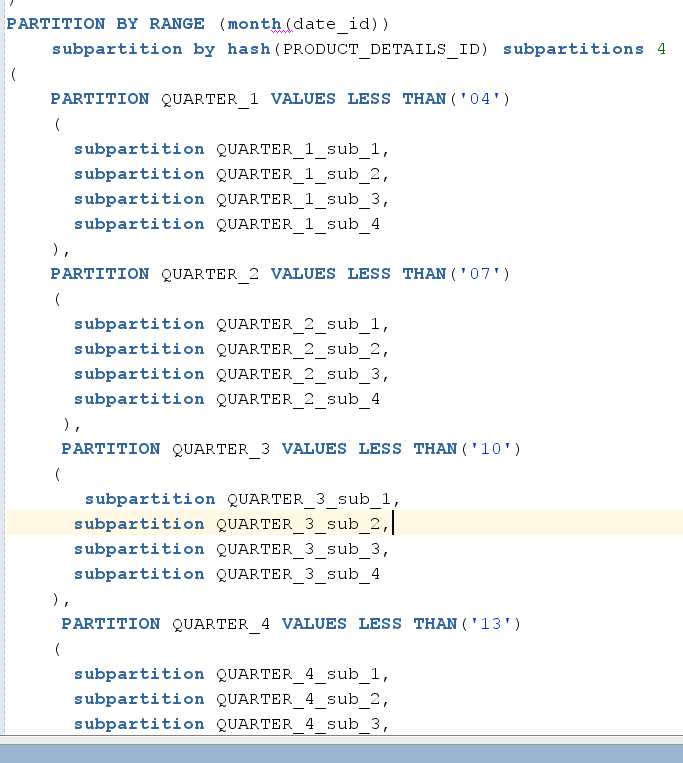
In my opinion, the most optimal way is to divide our data into 4 parts (quarterly-information for 3 months).

* Range Partition by DATE\_ID. The sales are divided by quarters.
* Hash partition of every quarter by PRODUCT\_DETAILS\_ID (number of sub partitions is 2^N(in my case I chose 4), because it should be the degree of 2)

Example of a result table of facts with partition taken into account:



Here is a table with the suggested partition option:



## 5.7. CREATE Strategy of Parallel execution

In my data warehouse system, the large RETAIL\_FACT\_Sales, DIM\_PRODUCTS, DIM\_CUSTOMERS, and DIM\_GEN\_PERIOD tables must be updated periodically with new or changed data from the production system. I can do this effectively by using parallel DML in conjunction with updatable connection views. This also increases the DDL speed.

The data that needs to be updated will be loaded into the table before the update process begins. This table will contain either new rows or rows that have been updated since the data store was last updated. After all changes are made, the data will be sent for public viewing.

# 6. Open Questions and Risks

The main issues that arise when working with storage:

* How information will be accessed and when data will be uploaded/updated
* Security of user information, such as passport data, payment method(card number, Bank transaction number)
* Risk of data loss due to storage system failures
* Validation of data received from various sources
* Effective search for product information (fast).

In our case, you need to understand that a properly organized data warehouse minimizes the above risks. All responsibility for the correct operation of the final system is on the developer.

# 7. Estimates

Thanks to a comprehensive backup approach, we will not only be able to manage our data as we see fit, but we will also be able to rest assured that it is always available, protected, and meets a set of standards. Using this system, our business can reduce risks and maintain critical information. The selected Star schema provides high query performance (because the star schema database has a small number of tables and clear connection paths, queries are executed faster) and broad support for a large number of business intelligence tools.