|  |
| --- |
| EPAM Systems, RD Dep. |
| MTN.\*NIX.07 Oracle DB. Introduction to DWH |
| Vasili Korzun Solution Concept |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| REVISION HISTORY | | | | | |
| Ver. | Description of Change | Author | Date | Approved | |
| Name | Effective Date |
| 1.0 | Initial status of document | Vasili\_Korzun | July 23, 2013 |  |  |
|  |  |  |  |  |  |

*Contents*

[1. Overview 3](#_Toc362587892)

[1.1. Business Background 3](#_Toc362587893)

[1.2. Benefits 4](#_Toc362587894)

[2. Requirements 4](#_Toc362587895)

[2.1. Business Requirements 4](#_Toc362587896)

[2.2. Technical Requirements 4](#_Toc362587897)

[3. Solution Sketch 4](#_Toc362587898)

[3.1. Source Tables structure 5](#_Toc362587899)

[3.2. Summarize Data Plan 7](#_Toc362587900)

[4. DWH Solution Concept 8](#_Toc362587901)

[4.1. Logical Diagram 9](#_Toc362587902)

[4.2. Physical diagram 11](#_Toc362587903)

[4.3. Dimensions 11](#_Toc362587904)

[4.4. Facts 12](#_Toc362587905)

[4.5. Dataflow Diagram 13](#_Toc362587906)

[4.6. Partitioning rules 14](#_Toc362587907)

[4.7. Strategy of Parallel execution 14](#_Toc362587908)

# Overview

Online shopping becomes the most popular way to buy something We’ve been improving our online shop for 2 years and there is another step to be done. The DWH solution was developed and the first analytical question was asked. This solution concept is designed to explore analytical details about payment and delivery systems are used by customers of our online shop.

## Business Background

. Our company provides internet access to the catalog of various products, mostly books, collectibles, electronics and media. Customers are able to register themselves at online store, read and post reviews, order goods, pay them in any convenient manner and then get them home by courier.

Payments are made by a set of payment systems, with whom we have entered into agreements. The delivery is made by a courier and delivery services, with whom we have entered into agreements too. Customers from many countries order and receive good the same way.

## Benefits

### Having all business information gathered stored and stored in ‘right’ way we’ll be able to continue solving various analytical problems.

### With information about most popular among our customers payment systems we can provide better support for some systems, thus we’ll receive more money from customers.

### Knowing the preferences of customers we can negotiate more favorable agreements with some payment systems.

### Information about sales on promo periods will be helpful to develop more attractive discounts and promotions.

### Top managers will be able to track and control sales of specific products or products categories in historical context.

# Requirements

## Business Requirements

### Daily information about all products ordered with specific payment in a specific country with additional information about used delivery systems and business period.

### Daily calculated count of certain sales.

### Daily calculated revenue for certain sales.

### Daily count of certain rejected or failed orders.

### Information about payment system should include it type, description and code. Information about delivery system should include just code and description. Information about product should include it description, code, category and subcategory.

### Information about country should include it’s code, description, region code and description, continent description.

## Technical Requirements

### Minimal time period – day.

### It shall be possible to add multi-language translations of this analytics.

### History about payment systems should be tracked and saved.

### History about geographical changes must be tracked and saved.

### There is no need to track history changes of delivery systems.

### There is no need to track history changes in products.

# Solution Sketch

## Source Tables structure

### Transactions

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Comment |
| TRANSACTION\_ID | NUMBER(15) | Transaction unique number. |
| DATE\_ID | DATE | Date of transaction. |
| USER\_ID | NUMBER(15) | User that purchase something. |
| PRODUCT\_ID | NUMBER(15) | Product that user ordered. |
| LOCATION\_ID | NUMBER(10) | Point of delivery. |
| PAYMENT\_SYSTEM\_ID | NUMBER(10) | Payment system is used to order. |
| DELIVERY\_SYSTEM\_ID | NUMBER(10) | Delivery system. |
| COST | NUMBER(15) | Cost of product including delivery cost. |
| STATUS\_CODE | NUMBER(15) | Status of order (success/fail/reject) |

### Users

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Comment** |
| USER\_ID | NUMBER (20) | User’s unique identifier |
| FIRST\_NAME | VARCHAR2 (32) | User’s First Name |
| LAST\_NAME | VARCHAR2 (32) | User’s Last Name |
| GENDER | VARCHAR2 (5) | User’s Gender (M/F) |
| YEAR\_OF\_BIRTH | NUMBER (10) | User’s date of birth |
| EMAIL\_ADDR | VARCHAR2(32) | User’s email |

### Payment systems

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Comment |
| PAYMENT\_SYSTEM\_ID | NUMBER (10) | Payment system unique code |
| PAYMENT\_SYSTEM\_CODE | VARCHAR2 (15) | Payment system short code |
| PAYMENT\_SYSTEM \_DESC | VARCHAR2 (32) | Payment system description |
| PAYMENT\_SYSTEM\_TYPE\_ID | NUMBER (5) | Type of payment system |
| PAYMENT\_SYSTEM\_TYPE\_DESC | VARCHAR2 (32) | Description of payment system |

### Delivery systems

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Comment |
| DELIVERY\_SYSTEM\_ID | NUMBER (10) | Delivery system unique code |
| DELIVERY \_SYSTEM\_CODE | VARCHAR2 (15) | Delivery system short code |
| DELIVERY \_SYSTEM \_DESC | VARCHAR2 (32) | Delivery system description |
| DELIVERY \_SYSTEM\_TYPE\_ID | NUMBER (5) | Type of delivery system |
| DELIVERY \_SYSTEM\_TYPE\_DESC | VARCHAR2 (32) | Description of delivery system |

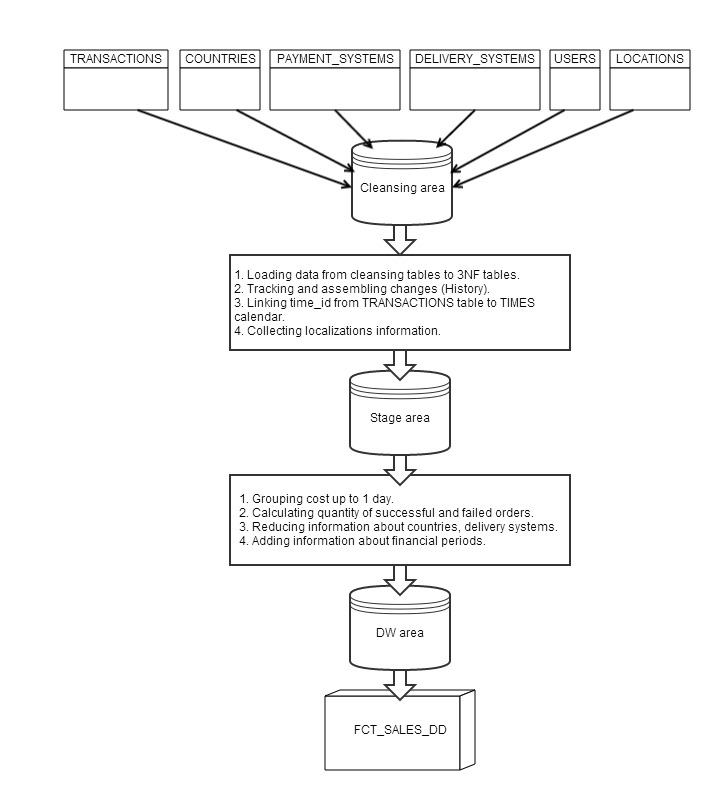
### Countries

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Comment |
| COUNTRY\_ID | NUMBER (10) | Country unique code |
| COUNTRY\_DESC | VARCHAR2 (32) | Country name |
| COUNTRY\_SHORT\_DESC | VARCHAR2 (5) | Country code |
| SUBREGION\_ID | NUMBER (5) | Subregion id |
| SUBREGION\_CODE | VARCHAR2 (32) | Subregion code |
| REGION\_ID | NUMBER (5) | Region id |
| REGION\_CODE | VARCHAR2 (32) | Region code |
| CONTINENT\_ID | NUMBER (5) | Continent id |
| CONTINENT\_CODE | VARCHAR2 (32) | Continent code |

### Locations

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Comment |
| LOCATION\_ID | NUMBER (10) | Location unique code |
| STREET\_ADDRESS | VARCHAR2 (100) | Street address (also includes all delivery information) |
| CITY | VARCHAR2 (5) | City name |
| STATE\_PROVINCE | NUMBER (5) | Province name |
| COUNTRY\_ID | VARCHAR2 (32) | Reference to country |

## Summarize Data Plan

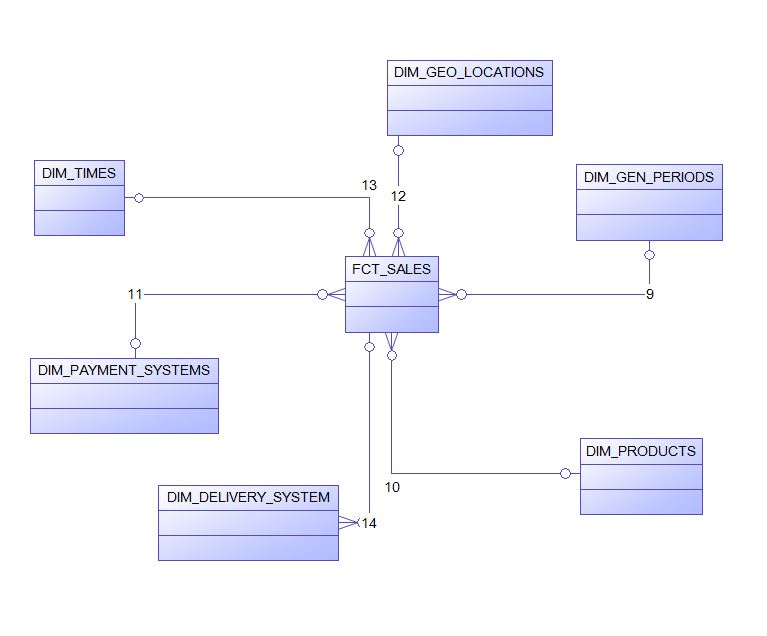


# DWH Solution Concept

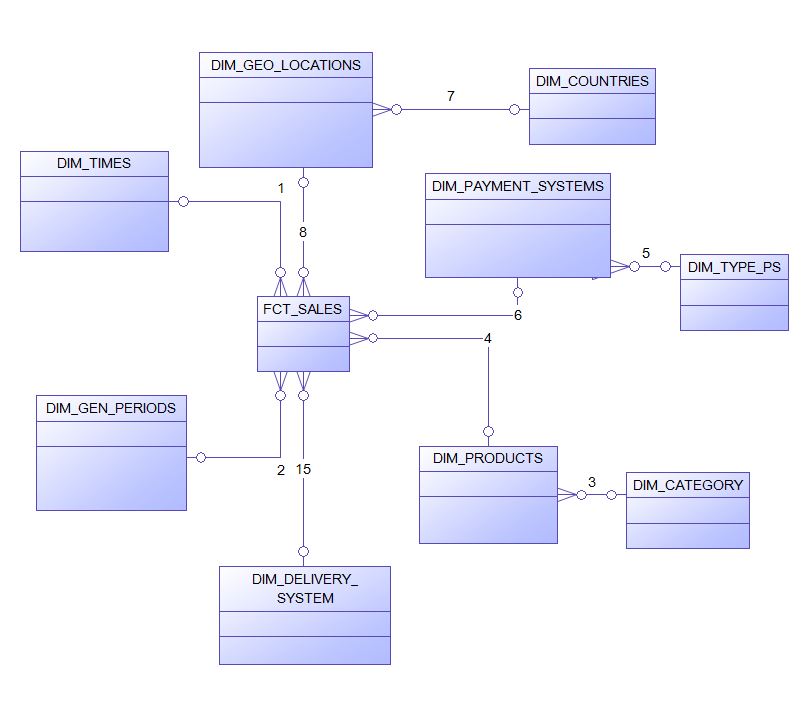
|  |  |  |  |
| --- | --- | --- | --- |
| Level Type | Object Name | Tablespace | Desctiption |
| Storage level  ST\_\* | ST\_DATA | ts\_st\_data\_01 | 3NF data loaded from external sources. |
| DW - Cleansing Level | DW\_CL | ts\_dw\_cl\_01 | Transformation and loading data to DW tables. Security enforcement. |
| DW – Level | DW | ts\_dw\_01 | Core DW. Dimensions, facts and general information. |
| DW– Prepare Star Cleansing Level | SAL\_DW\_CL | dw\_cl\_01 | Specific views for merged dimensions and facts tables. |
| STAR - Cleansing | SAL\_CL | ts\_star\_cl\_01 | Security layer.Deleting non-useful data.Preparing The Star. |
| STAR – Level | SAL\_DIM |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Logical Diagram

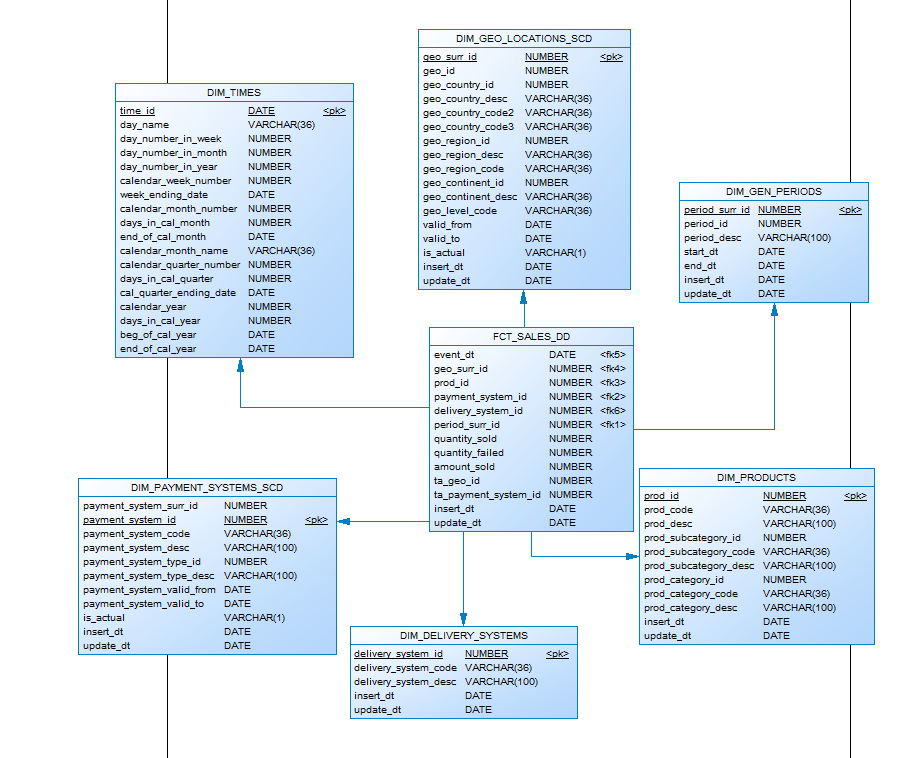
### Star schema



### Snowflake schema



## Physical diagram



## Dimensions

### Dimension Types

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Type | Size | DW – Merged Dimensions | Descriptions |
| DIM\_TIMES | SCD1 | BIG | DW.T\_DAYS, DW.T\_WEEKS, DW.T\_MONTHS, DW.T\_QUARTERS,  DW.T\_YEARS | Calendar. |
| DIM\_GEO\_LOCATIONS | SCD2 | BIG | DW.T\_COUNTRIES  DW.T\_REGIONS  DW.T\_CONTINENTS | Full information about countries, regions and continents. |
| DIM\_GEN\_PERIODS | SCD2 | SMALL | DW.T\_PROMOTIONS  DW.T\_PERIOD\_BEG  DW.T\_PERIOD\_END | Time periods for analytics, mostly with business logic. Used to additional facts grouping. |
| DIM\_PAYMENT\_SYSTEMS | SCD2 | SMALL | DW.T\_PS\_TYPES  DW.T\_PS\_DESC | Detailed information about payment systems. |
| DIM\_DELIVERY\_SYSTEMS | SCD1 | SMALL | DW.T\_DS\_TYPES  DW.T\_DS\_DESC | Detailed information about delivery systems. |
| DIM\_PRODUCTS | SCD1 | BIG | DW.T\_CATEGORIES  DW.T\_SUBCATEGORIES  DW.T\_PRODUCTS | Detailed information about products. |

### Dimension Hierarchies

**DIM\_GEN\_TIME:**

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

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| DAYs | DAY | Store day at the calendar | DAY\_ID |
| WEEKs | WEEK | Store weeks at the calendar year | WEEK\_ID |
| MONTHs | MONTH | Store months at the calendar | MONTH\_ID |
| QUARTERs | QUARTER | Store all quarters at the calendar | QUARTER\_ID |
| YEARs | YEAR | Store years at the calendar | YEAR\_ID |

**DIM\_PRODUCTS:**

**Hierarchy PRODUCT- SUBCATEGORY-CATEGORY**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| PRODUCTs | PRODUCT | Products in subcategory | PRODUCT\_ID |
| SUBCATEGORYs | SUBCATEGORY | Subcategories in category. | SUBCATEGORY\_ID |
| CATEGORYs | CATEGORY | Information about category. | CATEGORY\_ID |

**DIM\_GEO LOCATIONS:**

**Hierarchy COUNTRY –REGION -- CONTINET**

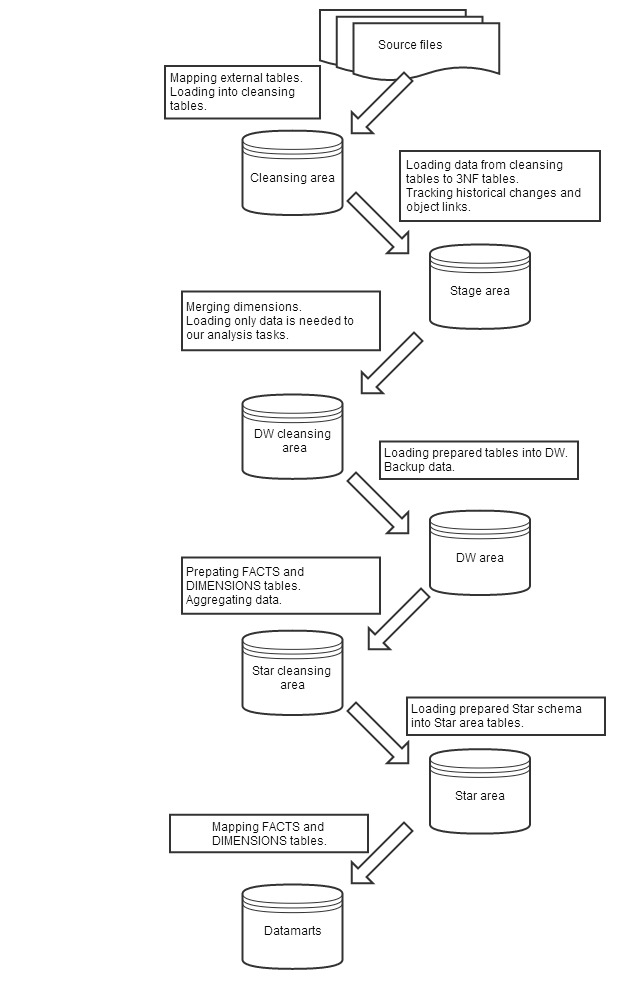
|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| COUNTRIES | COUNTRY | Countries in region | GEO\_COUNTRY\_ID |
| REGIONS | REGION | Regions in continent | GEO\_REGION\_ID |
| CONTINENTS | CONTINENT | Continents in the world. | GEO\_CONTINENT \_ID |

## Facts

### Facts Aggregations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Code | Table Name | Additive | Descriptions |
| Sales amount aggregation. | AMOUNT\_SOLD | FCT\_SALES\_DD | YES | Calculating daily sales for specific country, product, payment and delivery system. |
| Failed/rejected transactions aggregation. | QUANTITY\_FAILED | FCT\_SALES\_DD | YES | Calculating quantity of failed/rejected transactions |
| Successful transactions aggregation. | QUANTITY\_SOLD | FCT\_SALES\_DD | YES | Calculating quantity of successful transactions. |

## Dataflow Diagram



## Partitioning rules

**FCT\_SALES\_DD:**

To facilitate maintenance of Facts table, it should be partitioned in several segments.

Main partitioning is Range by event\_dt. Administrator should service partitions, archive old and add new in time. Subpartitioning is Hash into 4 tablespaces.

*Partitioning by event\_dt with granularity 2-3 month allows me to faster refresh transactions table, using MINUS operation only on small datasets, not on all fact table.*

## Strategy of Parallel execution

ETL loading packages should use parallel queries to access big cleansing tables.

Selection from star should be performed using parallel queries.

Updating the Facts table should be performed using parallel DML operations.

*All examples of my parallel execution strategy are represented in ETL packages scripts.*