|  |
| --- |
| EPAM Systems, RD Dep. |
| Analysis of Amazone.com Sales |



Contents

[1. Business Description 3](#_Toc498071105)

[1.1. Business background 3](#_Toc498071106)

[1.2. Problems because of poor data management 3](#_Toc498071107)

[1.3. Benefits from implementing a Data Warehouse 3](#_Toc498071108)

[2. Dimensions of a Business 3](#_Toc498071109)

[2.1. Select the Business Process 3](#_Toc498071110)

[2.2. Declare the Grain 3](#_Toc498071111)

[2.3. Identify the Dimensions 3](#_Toc498071112)

[2.4. Identify the Facts 3](#_Toc498071113)

[3. Logical Scheme 3](#_Toc498071114)

[4. Data Flow 3](#_Toc498071115)

[5. Fact Table Partitioning Strategy 3](#_Toc498071116)

[6. Strategy of Parallel Load 3](#_Toc498071117)

[7. Report Layouts 3](#_Toc498071118)

# Business Description

## Business background

Amazon (Amazon.com) is the world’s largest online retailer. Amazon offers its services all over the world. It operates its own websites in the UK, Germany, China, Australia, as well as many South American and Asian countries. Amazon.com serves 20 million customers every day. Its annual income is over 34 billion dollars. Amazon is a business empire that sells almost anything cheaper and faster than anyone else.

## Problems because of poor data management

It may be occurred such problems as:

* Technical data not recorded properly;
* Data not stored properly;
* Data Sharing Difficulties.

## Benefits from implementing a Data Warehouse

DWH structure is going to meet the following benefits:

* Quick and easy access to data;
* Improving the performance;
* Consistency of data;
* The opportunity to forecast;
* The opportunity for marketing campaign.

# Dimensions of a Business

## Select the Business Process

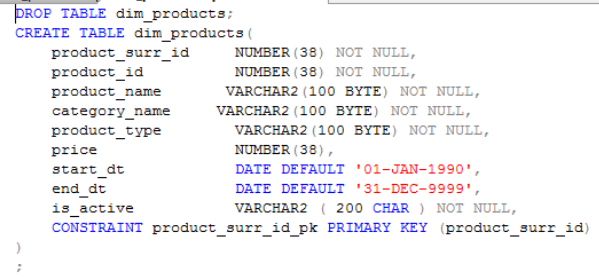
Business Process for analysis is Amazon’s sales per different metrics.

## Declare the Grain

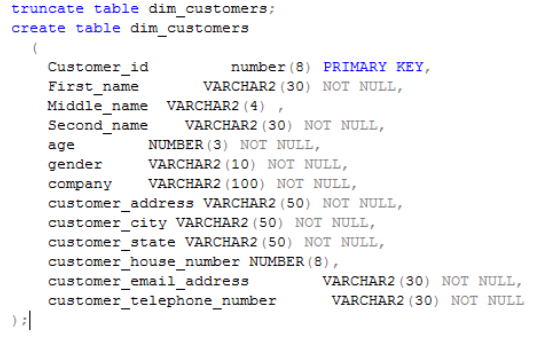
Each row in the table shows product that have been ordered by customer.

## Identify the Dimensions

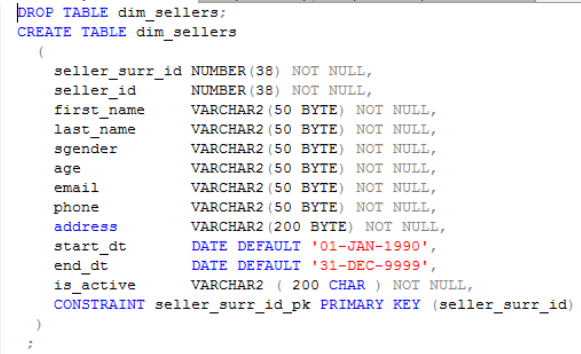
## Dim\_Products contains information about products, their manufacturers and other useful information.



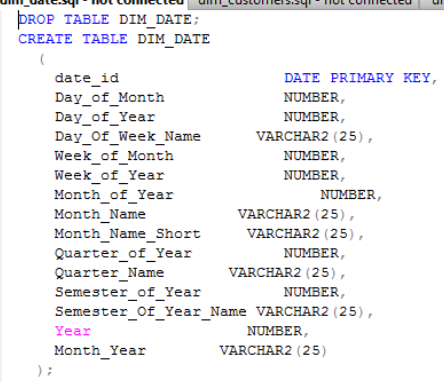
## Dim\_Customers contains information about customers, their contacts and their locations.



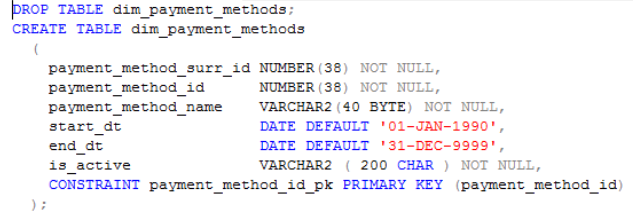
## Dim\_Sellers contains information about company’s employees, their contacts and their locations.



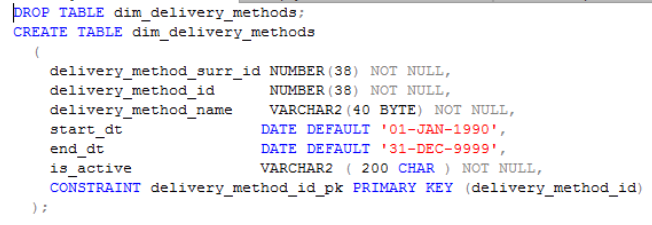
## Dim\_Date contains date information.



## Dim\_PaymentMethod contains information about methods of payment of orders.



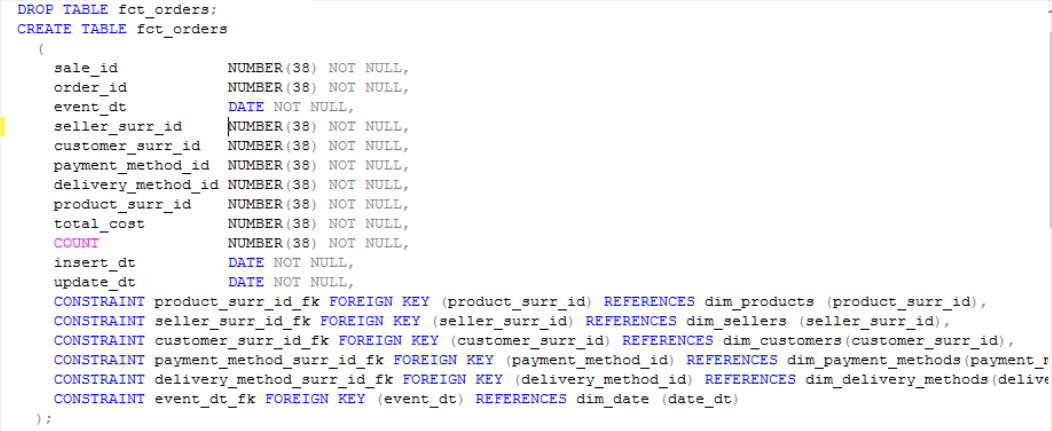
## Dim\_DeliveryMethod contains information about methods of delivering of orders.



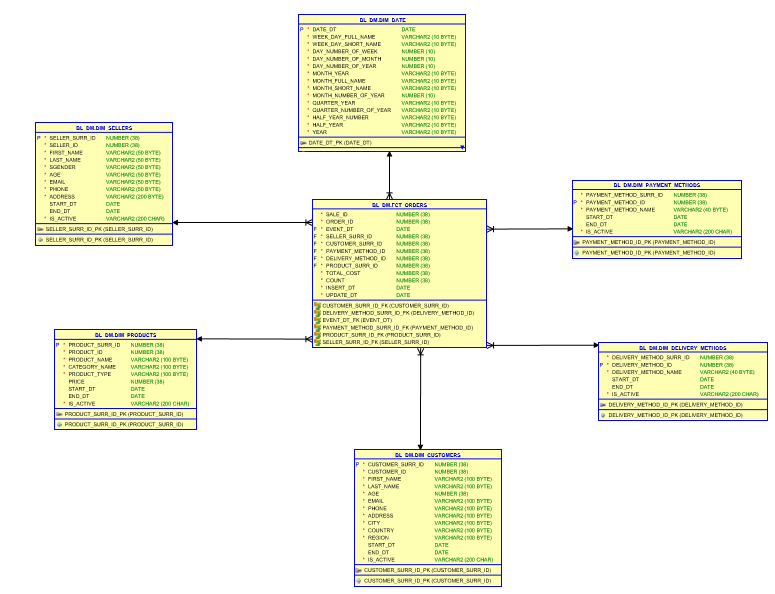
## Identify the Facts

Fact\_Orders:

* SALE\_ID
* ORDER\_ID
* EVENT\_DT
* SELLER\_SURR\_ID
* CUSTOMER\_SURR\_ID
* PAYMENT\_METHOD\_ID
* DELIVERY\_METHOD\_ID
* PRODUCT\_SURR\_ID
* TOTAL\_COST
* COUNT
* INSERT\_DT
* UPDATE\_DT

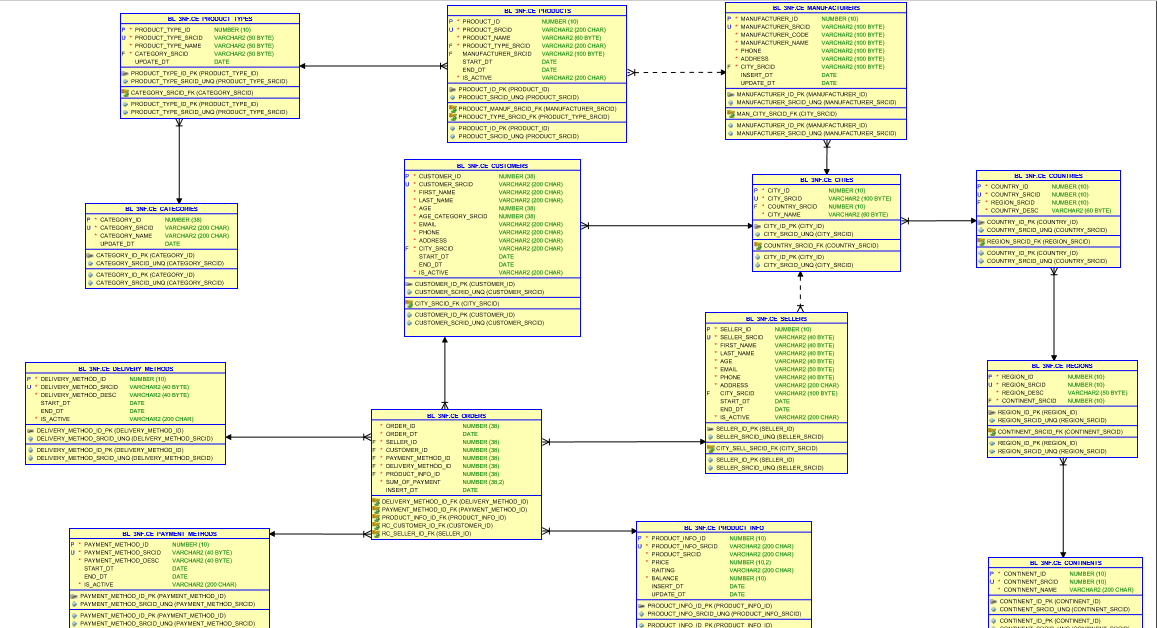


Star Model:



# Logical Scheme

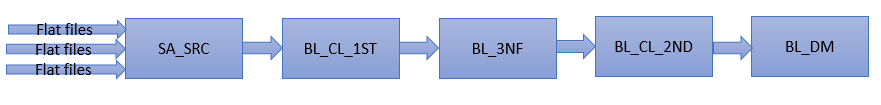
Tables in 3NF:



Tables:

* Ce\_products;
* Ce\_categories;
* Ce\_product\_type;
* Ce\_customers;
* Ce\_sellers;
* Ce\_product\_info;
* Ce\_cities;
* Ce\_countries;
* Ce\_regions;
* Ce\_countries;
* Ce\_orders;
* Ce\_delivery\_methods;
* Ce\_payment\_methods.

# Data Flow



**Staging Layer (sa\_src).**The main purpose of the Staging Layer is to load source data into the Data Warehouse environment for further processing (the processing from source to staging). In other words, the Staging Layer is responsible for the physical movement of data from the source into the Data Warehouse platform.

**Cleansing Layer (bl\_cl).** The Cleansing Layer is used for data cleansing, filtering wrong data, replace missing values with singletons and performing transformations like code lookups or currency conversions. As the Staging Area, the Cleansing Area contains only data of the last delivery, and data from different sources is not integrated.

**Core Layer(bl\_3NF).** The Core has two purposes: First, it is the integration layer for the data from all source systems of the Data Warehouse. The data in the Core is not stored separately for each source like in the previous layers, but stored in a "subject-oriented" form. The second purpose is to store the history of all data changes in a form that allows to retrieve previous states of the data at any time in the past. This is an important requirement for traceability of data changes, and there are different approaches to store historic data.

**Data Marts (bl\_dm).** The Data Marts provide the data in an adequate form that allows easy access for the front-end applications of the BI platform.

# Fact Table Partitioning Strategy

Partitioning of fact table will be made by date of orders for every year and every month.

# Strategy of Parallel Load

The parallel load didn’t be used in the project because it slows data loading.

# Report Layouts

There is a comparison of the performance of 3NF Layer and DM Layer in this chapter.

**Performance of the 3NF Layer:**

This script provides us with opportunity to analyze company’s sales in the previous year with grouping by quarter and month.

**SELECT DECODE(GROUPING\_ID(TO\_CHAR(order\_dt,'YYYY'), upper(TO\_CHAR(order\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(order\_dt,'Q'), upper(TO\_CHAR(order\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(order\_dt,'Mon') ), order\_dt), 7, 'GRAND TOTAL FOR '**

**|| TO\_CHAR(order\_dt,'YYYY'), ' ') AS YEAR,**

**DECODE(GROUPING\_ID(TO\_CHAR(order\_dt,'YYYY'), upper(TO\_CHAR(order\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(order\_dt,'Q'), upper(TO\_CHAR(order\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(order\_dt,'Mon') ), order\_dt), 3, 'GRAND TOTAL FOR '**

**|| upper(TO\_CHAR(order\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(order\_dt,'Q'), ' ') AS quarter,**

**DECODE(GROUPING\_ID(TO\_CHAR(order\_dt,'YYYY'), upper(TO\_CHAR(order\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(order\_dt,'Q'), upper(TO\_CHAR(order\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(order\_dt,'Mon') ), order\_dt), 1, 'GRAND TOTAL FOR '**

**|| upper(TO\_CHAR(order\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(order\_dt,'Mon') ), ' ') AS MONTH,**

**DECODE(GROUPING(order\_dt), 1, ' ', order\_dt) AS DAY,**

**TO\_CHAR(SUM(sum\_of\_payment), '999,999,999,999') AS sales**

**FROM ce\_orders dt**

**WHERE TO\_CHAR(order\_dt,'YYYY') = 2016**

**GROUP BY ROLLUP( TO\_CHAR(order\_dt,'YYYY'), upper(TO\_CHAR(order\_dt,'YYYY'))**

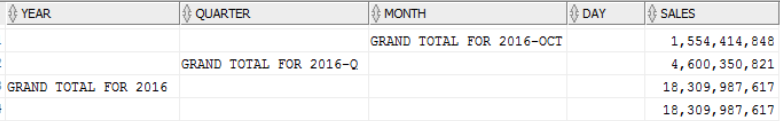
**|| '-'**

**|| 'Q'**

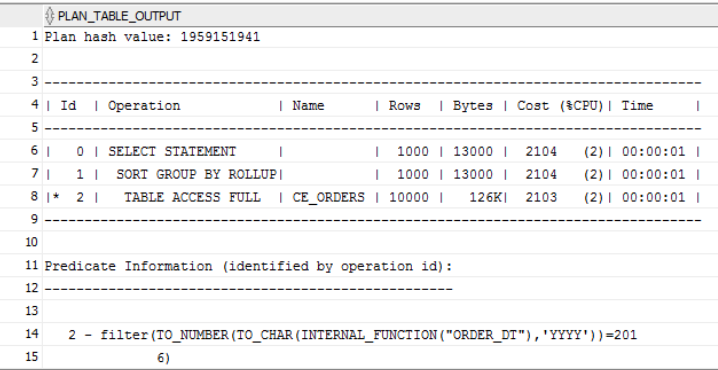
**|| TO\_CHAR(order\_dt,'Q'), upper(TO\_CHAR(order\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(order\_dt,'Mon') ), order\_dt );**



EXPLAIN PLAN for the script:



**Performance of the DM Layer:**

**SELECT DECODE(GROUPING\_ID(TO\_CHAR(event\_dt,'YYYY'), upper(TO\_CHAR(event\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(event\_dt,'Q'), upper(TO\_CHAR(event\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(event\_dt,'Mon') ), event\_dt), 7, 'GRAND TOTAL FOR '**

**|| TO\_CHAR(event\_dt,'YYYY'), ' ') AS YEAR,**

**DECODE(GROUPING\_ID(TO\_CHAR(event\_dt,'YYYY'), upper(TO\_CHAR(event\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(event\_dt,'Q'), upper(TO\_CHAR(event\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(event\_dt,'Mon') ), event\_dt), 3, 'GRAND TOTAL FOR '**

**|| upper(TO\_CHAR(event\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(event\_dt,'Q'), ' ') AS quarter,**

**DECODE(GROUPING\_ID(TO\_CHAR(event\_dt,'YYYY'), upper(TO\_CHAR(event\_dt,'YYYY'))**

**|| '-'**

**|| 'Q'**

**|| TO\_CHAR(event\_dt,'Q'), upper(TO\_CHAR(event\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(event\_dt,'Mon') ), event\_dt), 1, 'GRAND TOTAL FOR '**

**|| upper(TO\_CHAR(event\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(event\_dt,'Mon') ), ' ') AS MONTH,**

**DECODE(GROUPING(event\_dt), 1, ' ', event\_dt) AS DAY,**

**TO\_CHAR(SUM(total\_cost), '999,999,999,999') AS sales**

**FROM fct\_orders dt**

**WHERE TO\_CHAR(event\_dt,'YYYY') = 2016**

**GROUP BY ROLLUP( TO\_CHAR(event\_dt,'YYYY'), upper(TO\_CHAR(event\_dt,'YYYY'))**

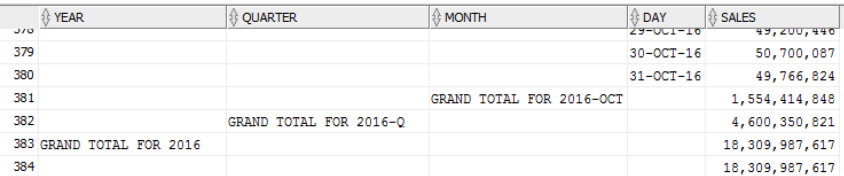
**|| '-'**

**|| 'Q'**

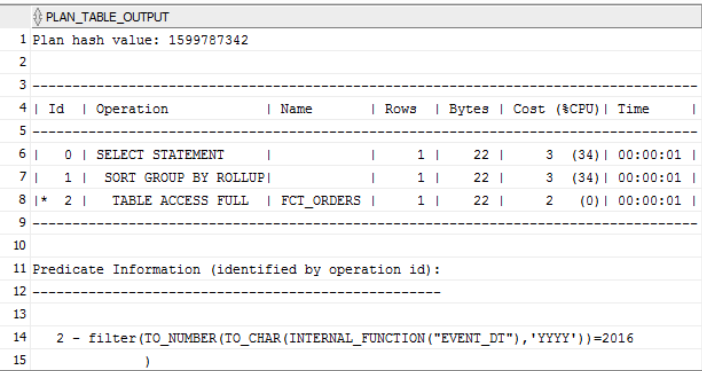
**|| TO\_CHAR(event\_dt,'Q'), upper(TO\_CHAR(event\_dt,'YYYY')**

**|| '-'**

**|| TO\_CHAR(event\_dt,'Mon') ), event\_dt );**

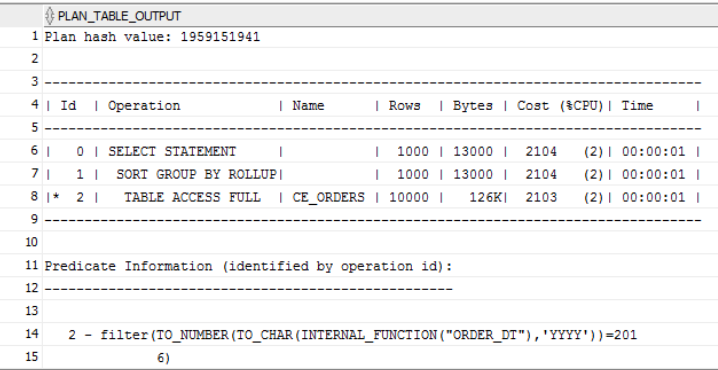


EXPLAIN PLAN for the script:

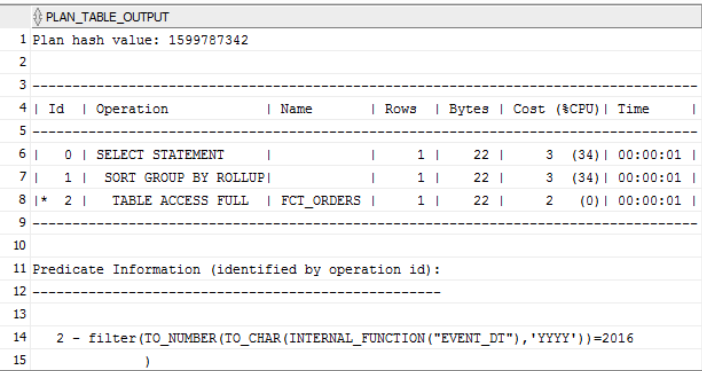


Comparison of the results:

**3NF Layer:**



**DM Layer:**



The comparison of the result shows that the cost of the script from 3NF Layer is much more than the cost of the script of the DM Layer.