|  |
| --- |
| EPAM Systems, RD Dep. |
| MTN.\*NIX.07 Oracle DB. Introduction to DWH |
| MTX.\*NIX.07. Introduction to DWH - Exit Task |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ver. | Description of Change | Performed by | Date | Approved | |
| Name | Effective Date |
| 1.0 | Initial status of document | **Aliaksandr Ahushevich** | 04-MAR-2013 |  |  |
|  |  |  |  |  |  |

*Contents*

[1. Exit Task 3](#_Toc320774863)

[2. Overview 3](#_Toc320774864)

[2.1. Business Background 3](#_Toc320774865)

[2.2. Benefits 3](#_Toc320774866)

[3. Requirements 3](#_Toc320774867)

[3.1. Business Requirements 3](#_Toc320774868)

[3.2. Technical Requirements 4](#_Toc320774869)

[4. Solution Sketch 4](#_Toc320774870)

[4.1. Source Tables structure 4](#_Toc320774871)

[4.2. Summarize Data Plan 5](#_Toc320774872)

[5. DWH Solution Concept 5](#_Toc320774873)

[5.1. Logical Diagram 5](#_Toc320774874)

[5.2. Physical diagram 6](#_Toc320774875)

[5.3. Dimensions 7](#_Toc320774876)

[5.3.1. Dimension Types 7](#_Toc320774877)

[5.3.2. Dimension Hierarchies 7](#_Toc320774878)

[5.4. Facts 8](#_Toc320774879)

[5.4.1. Facts Aggregations 8](#_Toc320774880)

[5.5. Dataflow Diagram 9](#_Toc320774881)

[5.6. Partitioning rules 10](#_Toc320774882)

[5.7. Strategy of Parallel execution 10](#_Toc320774883)

# Exit Task

Create next separate documents:

* BI Solution Proposal + Solution Concept

Chapter’s structures of all documents below:

# Overview

Creating an analytics tool that helps website owners understand how visitors engage with their website. Customer can view a variety of reports about how visitors interact with their website so that they can improve it.

## Business Background

Site owners want to get detailed information about site traffic so that they can understand their visitor's behavior.

The tools helps owners to analyze the traffic that is important for targeting website visitors in advertising and marketing. It could also help with site development and maximizing profit by promoting the efficient use of resources.

Tools provides detailed information in context of time on who visits site, when they visit, where they visit from geographically, how many pages they viewed and what browser or operation system they use.

Also to look at historical and temporal patterns: count of visitors daily (monthly); in what time site most or less visited, average count of pages views for each age category; how many users signed up in the last month; percent of authorized users.

For example, Firefox users behave differently from Internet Explorer users. Computer users spend more time on website than smartphone users. The maximum number of users is reached between 7pm and 10pm hours. And users from Russia spend more time on site

## Benefits

## Tools allow to analyze the behavior of site users in different countries, in different age categories.

* Site owners will be able to make better use of their servers.
* The tool enables a more efficient use of targeted advertising.

## Business Requirements

* count of visits in context of time (hour, day, month)

1/2/2010| 19:00-20:00| 25352 visits

1/2/2010| 20:00-21:00| 35323 visits

1/2/2010| 21:00-22:00| 30953 visits

………………………………

Total for 1/2/2010: |189898 visits

* count of visits in context of geolocation (City, Country, Subregion, Region)

1/2/2010| Germany | 5352 visits

1/2/2010| France |3323 visits

1/2/2010| Italy | 1953 visits

1/2/2010| Russia | 10352 visits

…………………………..

Total for Europe: |99834 visits

* Period with max count of visits
* Period with min count of visits
* Page views in context of time (Hour, day month)
* Unique visitors in context of time (hour, day)
* Percentage of authorized users
* Average length of visit
* Average length of visit of an authorized users
* Average length of visit of an unauthorized users
* The number of registered users in context of time (day, month, year)
* The number of registered users in context of geolocation (City, Country, Subregion, Region)
* Bounce Rates (viewing the site less 1 minute)

## Technical Requirements

* DB Oracle 11g R2
* Statistics should be updated hourly
* Ability to create reports on the set parameters
* Storage of statistics for the last 5 years

# Solution Sketch

## Source Tables structure

## *User File*

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Description** |
| USER\_ID | NUMBER | User unique key. |
| NICKNAME | VARCHAR2 | Displayed name |
| FIRST\_NAME | VARCHAR2 | User first name |
| LAST\_NAME | VARCHAR2 | User last name |
| GENDER | VARCHAR2 | User genderM/F |
| BIRTH\_DATE | DATE | User birth date |
| EMAIL | VARCHAR2 | User email address |
| NATIVE\_COUNTRY | VARCHAR2 | User native country |
| NATIVE\_LANGUAGE | VARCHAR2 | User native language |
| USER\_STATUS | VARCHAR2 | Adminstrator/Moderator/User |
| REG\_DATE | DATE | User registration date |

*Geo\_location file*

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Description** |
| CITY\_CODE | NUMBER | City digit code |
| CITY\_NAME | VARCHAR2 | City name |
| COUNTRY\_NAME | VARCHAR2 | Country name |
| COUNTRY\_CODE | NUMBER | Country digit code |
| SUBREGION | VARCHAR2 | Subregion |
| SUBREGION\_CODE | NUMBER | Subregion digit code |
| REGION | VARCHAR2 | Region |

*Site\_traffic*

|  |  |  |
| --- | --- | --- |
| Name | Data type | **Description** |
| IP\_address | VARCHAR2 | User ip address |
| User\_id | NUMBER | Authorized User ID. |
| Time\_id | TIMESTAMP | Beginning of the session to within hour. |
| Beg\_session | TIMESTAMP | Beginning of the session to within seconds |
| End\_session | TIMESTAMP | End of the session to within seconds |
| OS | VARCHAR2 | User operation system |
| Browser | VARCHAR2 | User browser |
| Pages\_views | NIMBER | Count of pages viewed |

## Summarize Data Plan

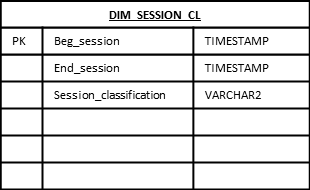
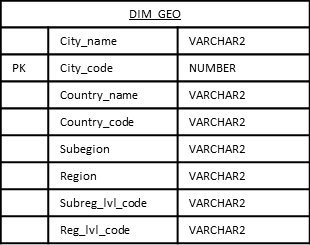
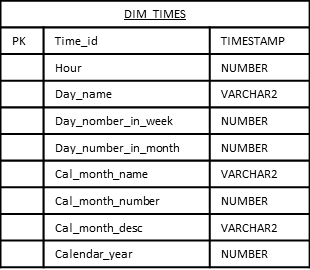
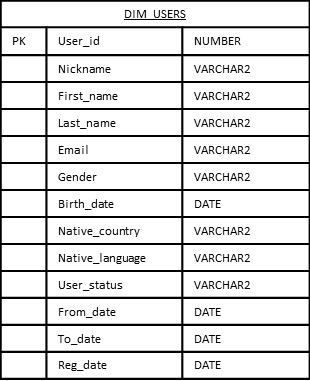
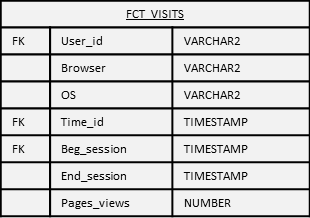


# DWH Solution Concept

## Logical Diagram



## Physical diagram



## Dimensions

### Dimension Types

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Type | Size | DW – Merged Dimensions | Descriptions |
| DIM\_TIMES | SCD1 | BIG | DW.T\_HOURS | Dimension table with time parameters |
| DW.T\_DAYS |
| DW.T\_MONTH |
| DW.T\_YEARS |
| DIM\_USERS | SCD2 | BIG | DW.T\_GENDER | Details information about authorized users |
| DW.T\_USER\_STATUS |
| DW.T\_BIRTH\_DATE |
| DW.T\_USERS |
| DIM\_COUNTRIES | SCD1 | SMALL | DW.T\_COUNTRIES | Full information about countries |
| DW.T\_CNTR\_GROUPS |
| DW.T\_CNTR\_SUB\_GROUPS |
| DW.T\_GEO\_TYPES |
| DW.T\_GEO\_SYSTEMS |
| DW.T\_GEO\_PARTS |
| DW.T\_GEO\_REGIONS |
| DW.T\_GEO\_OBJECTS |
| DW.T\_CNTR\_GROUP\_SYSTEMS |
| DW.LC\_COUNTRIES |
| DW.LC\_CNTR\_SUB\_GROUPS |
| DW.LC\_CNTR\_GROUPS |
| DW.LC\_GEO\_TYPES |
| DW.LC\_GEO\_SYSTEMS |
| DW.LC\_GEO\_PARTS |
| DW.LC\_GEO\_OBJECTS |
| DW.LC\_ GEO\_REGIONS |
| DW.LC\_CNTR\_GROUP\_SYSTEMS |
| DIM\_ SESSION\_CL | SCD1 | SMALL | DW.T\_SESSION\_CL | Technical dimension which stores session's classifications by time |

### Dimension Hierarchies

DIM\_TIMES

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| HOURS | HOURS | All hourly periods per day | TIMES\_HOURS\_ID |
| DAYS | DAYS | All daily periods per month | TIMES\_DAYS\_ID |
| MONTH | MONTH | All monthly periods per year | TIMES\_MONTH\_ID |

DIM\_GEO

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| COUNTRIES | COUNTRIES | All cities of the same country | GEO\_COUNTRY\_ID |
| SUBREGIONS | SUBREGIONS | All countries of the same subregion | GEO\_SUBREGION\_ID |
| REGIONS | REGIONS | All countries of the same region | GEO\_\_REGION\_ID |

## Facts

### Facts Aggregations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Code | Table name | Additive | Descriptions |
| Count of pages viewed | Pages\_views | FCT\_VISITS | + | Calculate count of pages viewed |
|  |  |  |  |  |

## Dataflow Diagram



## Partitioning rules

DIM\_TIMES table should be partitioned

Partition by CALENDAR\_YEAR and subpartition by CAL\_MONTH\_NUMBER

Part\_2010\_1

Part\_2010\_2

…

Part\_2010\_12

## Strategy of Parallel execution

Using parallel queries enable the capability to perform a single query using many operating system processes or threads.

Parallel DML is used in reference to performing modifications (INSERT, UPDATE, DELETE, and MERGE) it's can save a lot of time to update FACT and DIMENSION tables

Parallel DDL provide the ability to perform large DDL operations in parallel. It's save a lot of time during updating structure of DWH and star levels