

*KORYTKO Ineternet CASINO*

solution Concept Document

Version number 1.0

|  |  |
| --- | --- |
| **Submission Date** | *04/24/2012* |
| **Requested By** | *Kyril Bulcha* |
| **Business Owner** | *Alexander Korytko* |
| **Contact Info.** | *Alexander\_Korytko@epam.com* |

# Overview

For the correct creation of a system of analysis at first we need necessary to determine the Data of interest. Our system will receive data from other tables that were structured based on data received from the files. The main objects of our system are:

|  |  |  |
| --- | --- | --- |
| N/N | Data of interest | Description |
| A1 | USERS | User personal information (First name, Last name, gender, day of birth, money balance) |
| A2 | TIMES | Time period for analysis: year, month, quarter (last 12 month) |
| A3 | CURRENCY | Type of users and transactions currency (EUR, USD, RUB and ect.) |
| A4 | LOCATIONS | Regions, subregions, countries |
| A5 | OPERATIONS | Types of transactions – DEPOSIT and WITHDRAWAL |
| A6 | TRANSACTIONS | Full information about operation in one time (owner, time, type and etc.) |

## Business Background

# KORYTKO Internet CASINO provides a selection of the world’s favorite games which can be downloaded for free and play for fun or for real money.

KORYTKO Internet CASINO should always have complete statistics by countries of the total value of deposit and withdrawal money to the users account monthly over the past 12 months. In addition, for each country should be analyzed one the most popular method of depositing and withdrawing money from user account.

## Benefit

# With this our statistics collected by countries can be very easy and fast to identify the most profitable countries for CASINO in each month;

# Managers can always have full statistic about total sum amount of all deposits and withdrawals by countries in each month;

# Using this analyzing system managers of company will be able to more accurately analyze the most frequently used methods of deposit and withdrawal money from the user accounts by countries, regions and subregions.

# Statistics for each month allows very accurately determine the profitability of the company over the last year and will help to predict next year's income in each of 12 months.

# Requirements

## Business Requirements

|  |  |
| --- | --- |
| N/N | Business Requirements |
| B1 | Monthly calculated amount of total Deposits by country |
| B2 | Monthly calculated amount of total Withdrawal by country |
| B3 | Top used methods of deposit money to user account for each country |
| B4 | Top used methods of withdrawal money from user account for each country; |
| B5 | Types of transactions – DEPOSIT and WITHDRAWAL |
| B6 | Full information about operation in one time (owner, time, type and etc.) |
| B7 | Total amount of deposit per day by country should be calculated at the end of each day |
| B8 | Total amount of withdrawal per day by country should be calculated at the end of each day |
| B9 | Total amount of each operation method by country should be calculated at the end of each day |

## Technical Requirements

|  |  |
| --- | --- |
| N/N | Technical Requirements |
| C1 | Statistic should be calculated monthly for each country (at the end of each month); |
| C2 | Statistic period - last 12 month; |
| C3 | User information should consist gender information (Name, surname, gender and countries); |
| C4 | Should be possible to see statistic by month and quarter; |
| C5 | Should be possible to see statistic by country, region and subregion; |
| C6 | The average count of transactions per day is 1000 and around 20 000 per month; |

# Solution Sketch

## Source Tables structure

### *USERS FILE*

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Comment** |
| **USER\_ID** | NUMBER (32) | User unique code |
| **FIRST\_NAME** | VARCHAR2 (32) | User First Name |
| **LAST\_NAME** | VARCHAR2 (32) | User Last Name |
| **GENDER** | VARCHAR2 (5) | User Gender (M/F) |
| **YEAR\_OF\_BIRTH** | NUMBER (10) | Users date of birth |
| **COUNTRY** | NUMBER (32) | User country code |
| **BALANCE** | NUMBER (32) | Users current money balance |

### *TIMES FILE*

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Comment** |
| **TIMES\_ID** | NUMBER (32) | Time unique code |
| **DATE** | DATE | Day number in month |
| **MONTH** | NUMBER (5) | Month number |
| **MONTH\_DESCRIPTION** | VARCHAR2 (10) | Name of month |
| **QUARTER** | NUMBER (5) | Quarter number (1/2/3/4) |
| **QUARTER\_DESCRIPTION** | VARCHAR2 (5) | Quarter description (first, second, third, fourth) |
| **YEAR** | NUMBER (32) | Year number |
|  |  |  |

### *CURRENCY*

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Comment** |
| **CURRENCY\_ID** | NUMBER (32) | Currency unique code |
| **CURRENCY\_NAME** | VARCHAR2 (32) | Currency name (EUR, USD, RUB and etc.) |
| **CURRENCY\_TYPE\_ID** | NUMBER (5) | Convertible currency type or not (1/2) |
| **CURRENCY\_TYPE\_NAME** | VARCHAR2 (32) | Convertible currency type (CONVERTIBLE, NOT CONVERTABLE) |
| **CURRENCY\_SCD\_TYPE\_ID** | NUMBER (32) | Slow Changed Dimension Type 2 code (1,2) |

### *OPERATIONS*

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Comment** |
| **OPERATION\_ID** | NUMBER (32) | Operation unique code |
| **OPERATION\_TYPE\_NAME** | VARCHAR2 (32) | Operation name (DEPOSIT / WITHDRAWAL) |
| **DATE** | DATE | Date of operation |
| **OPERATION\_METHOD\_NAME** | VARCHAR2 (5) | Name of method (Credit Cards, Direct Deposit, Check, Webmoney) |
| **OPERATION\_METHOD\_ID** | NUMBER (32) | Operation unique code |
| **OPERATION\_MAX\_AMOUNT** | VARCHAR2 (32) | Max amount of operation |
| **OPERATION\_MIN\_AMOUNT** | NUMBER (32) | Min amount of operation |

### *LOCATIONS*

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Comment** |
| **COUNTRY\_ID** | NUMBER (32) | Country unique code |
| **COUNTRY\_DESCRIPTION** | VARCHAR2 (32) | Country Name |
| **COUNTRY\_SHORT\_DESC** | VARCHAR2 (5) | Country abbreviation |
| **SUBREGION** | VARCHAR2 (32) | Subregion Name |
| **SUBREGION\_CODE** | NUMBER (32) | Subregion code |
| **REGION** | VARCHAR2 (32) | Region name |
| **REGION\_CODE** | NUMBER (32) | Region code |

## Logical and Phisical STAR schema

## 

|  |  |
| --- | --- |
| N/N | LOGICAL SCHEMA |
| D1 | FCT\_TRANSACTION is a fact table. It’s get information from all other dimensions such as USERS, LOCATIONS, TIMES, OPERATIONS, CURENCY |
| D2 | USERS – there is full information about user |
| D3 | TIEMS – List of all possible periods for analysis (Year, month and quarters) |
| D4 | LOCATIONS - List of all possible divisions by countries |
| D5 | OPERATIONS - List of operations types |
| D6 | CURRENCY List of all possible currency name and types |

## Logical SNOW Flakes schema

There are couple more tables, because sometimes we don’t need to use all data from all tables, for example quarters.

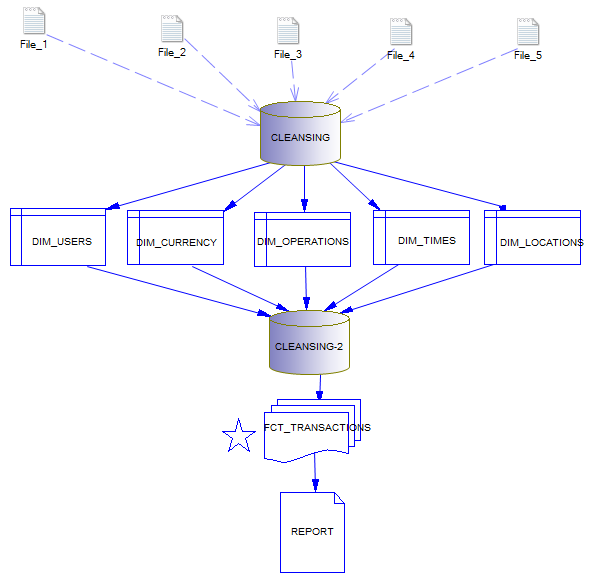
### 

|  |  |
| --- | --- |
| N/N | LOGICAL SCHEMA |
| E1 | FCT\_TRANSACTION is a fact table. It’s get information from all other dimensions such as USERS, LOCATIONS, TIMES, OPERATIONS, CURENCY |
| E2 | USERS – there is full information about user |
| E3 | TIEMS – List of year and month |
| E4 | QUARTERS – List of quarters fro TIMES table |
| E5 | LOCATIONS - List of all possible divisions by countries |
| E6 | COUNTRIES – List of Countries for LOCATION TABLE |
| E7 | OPERATIONS - List of operations |
| E8 | OPERATIONS\_TYPES\_METHODS – List of operations methods for OPERATION table |
| E9 | CURRENCY List of all possible currency name and types |

## PHISICAL STAR SCHEMA

## 

## Summarize Data Plan



# DWH ARCHITECTURE

|  |  |  |  |
| --- | --- | --- | --- |
| **Level Type** | **Object Name** | **Tablespace** | **Desctiption** |
| **STAGE AREA** | U\_SA\_USERS | ts\_sa\_users\_data\_01  (SIZE 100M  AUTOEXTEND ON NEXT 50M  SEGMENT SPACE MANAGEMENT AUTO  LOGGING;) | Contains users information. U\_SA\_USERS user should be able to load data from source files, create tables and view in this area. Upload data from source files to these tables and views an send it to next area. |
| U\_SA\_OPERATIONS | ts\_sa\_operations\_data\_01  (SIZE 100M  AUTOEXTEND ON NEXT 50M  SEGMENT SPACE MANAGEMENT AUTO  LOGGING;) | Contains information about operations and currencies.  U\_SA\_OPERATIONS user should be able to load data from source files, create tables and view in this area. Upload data from source files to these tables and views an send it to next area. |
| **DW CLEANSING AREA** | U\_DW\_CL | ts\_dw\_cl  (SIZE 200M  AUTOEXTEND ON NEXT 50M  SEGMENT SPACE MANAGEMENT AUTO  NOLOGGING;) | Contains all information from stage area. We are cleaning our data before upload to DW. NOLOGGING because there are many not useful information for us. U\_DW\_CL user should be able to create tables and views , upload and select clean data. |
| **DATA WAREHOUSE AREA** | U\_DW | ts\_dw\_data\_01  (SIZE 300M  AUTOEXTEND ON NEXT 100M  SEGMENT SPACE MANAGEMENT AUTO  LOGGING;) | Contains clean and normalized data from DW CLEANING AREA. LOGGING should be enabled for this area. U\_DW user should be able to create Tables and views, and make possible to select data from DW cleansing area. |
| **STAR DW CLEANSING AREA** | U\_SAL\_DW\_CL | ts\_sal\_dw\_cl  (SIZE 100M  AUTOEXTEND ON NEXT 50M  SEGMENT SPACE MANAGEMENT AUTO  NOLOGGING;) | There are data loaded from DW AREA. Create views from merged data of DWA.  U\_SAL\_DW\_CL user should able to select data from DW Area, create tables, delete some data from this tables, and create views with clean data for the next AREA. |
| **STAR CLEANSING AREA** | U\_SAL\_CL | ts\_sal\_cl  (SIZE 100M  AUTOEXTEND ON NEXT 50M  SEGMENT SPACE MANAGEMENT AUTO  NOLOGGING;) | Loading data from SAL-DW-CL AREA and then removing all not useful data. Consist view like in previous area, but only with data for our analysis. There should be only data with are useful for our analysis.  U\_SAL\_CL user should be able to select data from the view on STAR DW CLEANSING AREA, create tables and view (view for STAR AREA) and insert data to these tables and view. |
| **STAR AREA** | U\_STR\_DATA | ts\_ str\_data  (SIZE 100M  AUTOEXTEND ON NEXT 50M  SEGMENT SPACE MANAGEMENT AUTO  LOGGING;) | Receiving data from STAR CLEANSING AREA. There are data for the FACT Tables. U\_STR\_DATA user should be able to select data from tables and view from STAR CLEANSING AREA and create tables and view for data marts |
| U\_STR\_REFERENCES | ts\_ str\_references  (SIZE 100M  AUTOEXTEND ON NEXT 50M  SEGMENT SPACE MANAGEMENT AUTO  LOGGING;) | Receiving data from STAR CLEANSING AREA. There are data for the DIMENSIONS Tables. U\_STR\_REFERENCES user should be able to select data from tables and view from STAR CLEANSING AREA and create tables and view for data marts |

## DIMMENSION 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 | Dimension table with list of time parameters for analytics |
| DIM\_USERS | SCD1 | BIG | DW.T\_USERS  DW.T\_GENDER  DW.T\_YEAR\_OF\_BIRTH | Dimension table with details information about of users |
| DIM\_GEO | SCD2 | SMALL | DW.T\_COUNTRIES  DW.T\_CNTR\_GROUPS  DW.T\_CNTR\_SUB\_GROUPS  DW.LC\_CNTR\_GROUPS  DW.T\_GEO\_TYPES  DW.T\_GEO\_SYSTEMS  DW.LC\_GEO\_SYSTEMS  DW.T\_GEO\_PARTS  DW.T\_GEO\_REGIONS  DW.T\_GEO\_OBJECTS  DW.T\_CNTR\_GROUP\_SYSTEMS  DW.LC\_CNTR\_GROUP\_SYSTEMS  DW.LC\_CNTR\_SUB\_GROUPS  DW.LC\_ GEO\_PARTS  DW.LC\_COUNTRIES  DW.LC\_ GEO\_REGIONS | Dimension table with full information about all countries, regions, subregions, economic types and some other classifications etc. |
| DIM\_GEN\_PERIOD | SCD1 | SMALL | DW.T\_PERIOD\_DESC  DW.T\_PER\_START  DW.T\_PER\_END  DW.T\_LEVEL\_CODE | Dimension table with information how to group our facts |
| DIM\_CURRENCY\_SCD | SCD2 | SMALL | DW.T\_CURRENCY  DW.T\_CURRENCY\_TYPES | SCD2 Type dimension with list of all currencies and types of them (convertible or not) |
| DIM\_OPERATIONS | SCD1 | SMALL | DW.T\_OPERATIONS\_TYPES  DW.T\_OPERATIONS\_METHODS | Dimension table with details information about operations (DEPOSIT/WITHDRAWAL) and used methods for them (Credit Cards, Check, Webmoney and etc) |

## 4.2 DIMMENSION HIERARHIES

**DIM\_GEN\_TIME:**

**Hierarchy DAY-WEEK-MONTH-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 year | MONTH\_ID |
| YEARs | YEAR | Store years at the calendar year | YEAR\_ID |

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

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

**DIM\_CURRENCY:**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| CURRENCY\_TYPE | CURRENCY\_TYPE | Describe currency type (Convertible or not) | CURRENCY\_TYPE\_ID |

**DIM\_GEO LOCATIONS:**

**Hierarchy COUNTRY –REGION -- CONTINETS**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| COUNTRIES | COUNTRY\_NAME | All countries in regions | GEO\_COUNTRY\_ID |
| REGIONS | COUNTRY\_REGION | All regions on continents | GEO\_COUNTRY\_REGION\_ID |
| CONTINENTS | COUNTRY\_CONTINENT | All continents in the world. | GEO\_COUNTRY\_ CONTINENT \_ID |

**DIM\_OPERATION:**

**Hierarchy** OPERATION\_TYPE -- OPERATION\_METHODS

|  |  |  |  |
| --- | --- | --- | --- |
| Name | LEVEL\_CODE | LEVEL\_DESC | LEVEL\_NATURAL\_KEY |
| OPERATION\_TYPE | OPERATION\_TYPE | Deposit/Withdrawal | OPERATION\_TYPE\_ID |
| OPERATION\_METHODS | OPERATION\_METHODS | Credit Cards, Bank transfer, Check, Webmoney | OPERATION\_METHOD\_TYPE\_ID |

## FACT AGGREGATIONs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Code | Table Name | Additive | Descriptions |
| Counts total amount of transactions | FCT\_AMOUNT | FCT\_AMOUNT | + | Calculate distinct values of Transaction at the EVENT\_DT period. |

## dataflow diagram

## 

## PARTITIONING RULES

There is one table that should be partitioned: DIM\_FACTS\_MONTH.

It’s should be range partitions by Event\_dt column by years and and month. Move all partitions to different tablespaces:

CREATE TABLE FCT\_TRANSACTIONS

(

TRANSACTION\_ID NUMBER,

OPERATION\_TYPE\_ID NUMBER,

OPERATION\_TYPE\_METHOD\_ID NUMBER,

COUNTRY\_ID NUMBER,

USER\_ID NUMBER (32),

EVENT\_DT DATE,

CURRENCY\_ID NUMBER (32),

FCT\_AMOUNT NUMBER (32)

)

partition by range (EVENT\_DT)

(

partition p1 values less than (to\_date('01-04-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M1,

partition p2 values less than (to\_date('01-05-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M2,

partition p3 values less than (to\_date('01-06-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M3,

partition p4 values less than (to\_date('01-07-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M4,

partition p5 values less than (to\_date('01-08-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M5,

partition p6 values less than (to\_date('01-09-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M6,

partition p7 values less than (to\_date('01-10-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M7,

partition p8 values less than (to\_date('01-11-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M8,

partition p9 values less than (to\_date('01-12-2011','dd-mm-yyyy')) tablespace TS\_SAL\_M9,

partition p10 values less than (to\_date('01-01-2012','dd-mm-yyyy')) tablespace TS\_SAL\_M10,

partition p11 values less than (to\_date('01-02-2012','dd-mm-yyyy')) tablespace TS\_SAL\_M11,

partition p12 values less than (to\_date('01-03-2012','dd-mm-yyyy')) tablespace TS\_SAL\_M12,

partition p13 values less than (MAXVALUE) tablespace TS\_SAL\_M13

);

## STRATEGY of PARALLEl EXECUTION

|  |  |
| --- | --- |
| N/N | STRATEGY of PARALLEl EXECUTION |
| F1 | Parallel query should be used to perform SELECT operation to prevent a long waiting of a response from the database |
| F2 | Using parallel in DDL operations (create tables as select, create index, alter index rebuild) saving a lot of time during updating structure of our DWH and star levels because in this case OS using more hardware resources to execute it, but not DB. |
| F3 | Using parallel in DML operations (update, insert and etc.) can help to save a lot of time to update FACTs and DIMMENSIONs tables. |

5. **Extraction properties and strategy.**

**5.1 Logical Extraction Methods in Data Warehouses**

|  |  |
| --- | --- |
| N/N | **Logical Extraction Methods in Data Warehouses** |
| G1 | In our project we are using Incremental ExtractionLogical methods, because some tables are really big and new data come very fast, so it’s very expansive to load full data every time. |
| G2 | Incremental ExtractionLogical methods mean that at a specific point in time, only the data that has changed since a well-defined event back in history will be extracted. This event is a last time of extraction. |
| G3 | To identify this delta change there must be a possibility to identify all the changed information since this specific time event. This information should be either provided by the source data itself such as an application column, reflecting the last-changed timestamp. In our case should be column with name “TIMESTAMPS”. New data should be added and old data should stay without changing with old timestamp |

**5.2 Physical Extraction Methods in Data Warehouses**

|  |  |
| --- | --- |
| N/N | **Physical Extraction Methods in Data Warehouses** |
| H1 | **Online Extraction** - The data is extracted directly from the source system itself. The extraction process can connect directly to the source system to access the source tables themselves (connect to tables on CL Layer to download data from them true the package). |

6.Transformation Description**.**

|  |  |
| --- | --- |
| N/N | **Transformation Description** |
| K1 | The data-transformation logic for our project will consist of multiple steps. For example, in transforming new records to be inserted into a sales table, there may be separate logical transformation steps to validate each dimension key.  When data is loaded into the Oracle db we can prepare it using several techniques for our project we will choose Transformations Using SQL and PL/SQL:   * The CREATE TABLE ... AS SELECT ... statement * MERGE * UPDATE * MULTITABLE insert * PL/SQL Procedures * PL/SQL Packages |

|  |  |
| --- | --- |
| N/N | **Refresh Scenarios** |
| L1 | New data comes every day. Data warehouse contains 12 month of data, (MIN TIME 04.04.2011 and MAX TIME 23.04.2012). The solution is to partition by month and year (as appropriate). So the FACT table should be with partitions and Inserts can be parallelized.  All DIMs tables are loaded by MERGE mechanism from previous layers, so on each layer we have actual data.  Use parallel SQL operations (such as SELECT … FROM) to separate the new data from the data in previous time periods. |