Swedbank Data Quality Visualization

Meeskonna liikmed ja tööjaotus:

- Projektijuht - Reimo Vellemaa
- UIX - Kristina Õim
- Server/Repo - Ragnar Pääslane
- Tehnoloog/Arhitekt - Lauri Roomere
- Dokumentatsioon - Andres Heiduk
- Andmebaas - Erki Sidron
- API - Thomas Toodo

-Kanban töölaud ja ülesannete jagamine: Taiga.io

https://tree.taiga.io/project/reimovellemaa-data-quality-visualization/kanban

-Meeskonnavaheline vestlus: Fleep.io

-Andmebaas PostgreSQL peal

PostgreSQL
starterpack: https://www.postgresql.org/docs/9.5/static/index.htmlhttps://www.postgresql.org/docs/9.5/static/tutorial.html

-Näidisandmed (80 000+ rida) Swedbanki poolt saadetud kujul:

```
INSERT INTO
MESA.PL_MEASURE_FACT_PRT(Query_Id,Run_Id,Fact_Row_No,Fact_Col_No,Measure_Fact_Date
,Run_Country_ShortName,Measure_Group_ID,'Measure_Param_Id,Measure_Text,Measure_Num
eric,Measure_Amt,Measure_Cnt,Measure_Date,Measure_Data_Type_Code,Validation_Rule_I
d,Measure_Type,Measure_Type_Name,DBObject_Object_Id,DBColumn_Object_Id,Database_Na
me,Object_Name,Column_Name,DBObject_Service_Shortname,Entity_Concept_ID,Entity_Con
cept_Name,Validation_Service_Shortname,Validation_ShortName,Context_Type,Context_V
alue,Context_Type_Desc,Context_Value_Desc,Context_Service_ShortName,Context_Applic
ation_Name,Country_ShortName,Run_DtTm)VALUES(5509,180898,613,1,'2017-01-
31','GR',32381,4482,'0',0,NULL,0,NULL,3,5510,321,'DOMAIN_TYPE_321',22057,6221808,'
DBSCHEMA_16771', 'OBJECT_22057','COLUMN_4482','SERVICE_73',5130907,'AGREEMENT
METRIC','SERVICE_73','AM_GR_M',3,'66','DOMAIN_TYPE_3','DOMAIN_VALUE_66','SERVICE_1
8','APPLICATION_42','GR','2017-02-06 08:14:25');
```

- PL_MEASURE_FACT_PRT - konkreetne dokumentatsioon puudub, veergude nimetused on tekitatud koosolekute tulemusena ja tuletatud mutuaalsest arusaamisest, mis võib korraga esineda.

-Kõik näidisandmed edukalt andmebaasis 12.03 seisuga

-Võimalikud dashboardid: *Tableau *Kibana *Compass Collibra *BIRT *Razorflow *Zoho *ClicData *Bootstrap raamistik koos templatega

-Dashboardiks valitud Bootstrap raamistik koos templatega 12.03 seisuga

Dashboardi live link: sidron.ee/Swedbank

-Serverina kasutusel AWS - Amazon Web Services

Serveri peale sai installitud Linux Ubuntu. Olemas nii DNS kui ka public IP.

-Githubi link ehk kus kood asub:

https://github.com/reimovellemaa/Swedbank

-Iga meeskonna liige saab oma lokaalis teha muudatusi ning siis need ühisesse GITi pushida. Sellest omakorda toimub automaante laadimine serverisse (AWSi peal hetkel antud funktsioon ei tööta).

Loodava tarkvara nõuded Swedbanki poolt (inglise keeles):

1 Objectives

1.1 Effect objectives

It will allow speeding up resolution times of Data Quality issues. It will create capability to identify relations, patterns, which would otherwise be harder to detect. Bringing down overall Issue Management related costs. Ensuring there is transparency to the Data Quality health status of published data.

1.2 Objectives of the assignment

The main objective is that we create out of our DQ metrics, which we calculate in the Data Warehouse, a visual, interactive and useful Data Quality dashboard which provides us clear & immediate info if there are DQ issues as it is much harder to find this out based on numerical reports. As an extra, we could aim to create a mobile version of the Data Quality dashboard.

2 Scope

Before going into deliverables, it is needed to explain the current setup. We use the concept of Services (see definitions) to govern our data, in other words a set of data is related to a service for which we have appointed a responsible person to monitor the quality of data. We would like to use this service concept as a way to publish DQ metrics in the report/dashboard. To check the quality of data for a particular BI Service we run one or several Validations. A Validation is a set of Data Quality Rules. One Validation can belong to one or several BI services, one DQ rule can belong to one or several validations. We would like to present now, on different levels, the quality of the data which we validate, using interactive visualizations.

For this we want:

- Application with a user-friendly interface
- Application can load Data Quality Validation results (format to be decided)
- Application needs to communicate DQ metrics (DQ validation results) based on different groupings/aggregation levels: BI Service Level DQ dimension level (completeness, conformance ...) (see Dashboard_Swedbank) Validation level Business DQ Rule level Country level (Estonia, Latvia, Lithuania or all) – filter
- Service is the highest level for which we want to see the status of the Data Quality per DQ dimension. The overall DQ Validation of a service is made of by several Validations, which would be good if they can be selected/highlighted/... to drill down on this Validation level.
- Also on Validation level we can have DQ metrics presented by DQ dimensions.
- Related Validations could e.g. be presented using a heat map, indicating what their overall Data Quality Level is and will attract us to the problem areas.
- Lowest level is the actual DQ Rules which make up the Validations; we should be able to drill down to this lowest level to actually identify which DQ rule is causing an error.
- On all levels business semantics should be used to create the reports otherwise the users don't understand what they are looking at: To achieve this there are business definitions available
- On lowest level we should be presenting DQ metrics in a numerical (amounts) and percentages way. E.g. 1000 records are validated and 100 are erroneous as they miss the data we are checking, in that case we have 90% completeness for this specific Data Quality Rule. How to present is up to the designers, but the users need to be able to see total amount of validated records, total amounts of errors and corresponding percentages.
- On lowest level and for analysis purpose we need to be able to see Object name, Column name, Validation name, DQ Rule name.
- Application should allow us to change the period we are looking at (different time periods, in length and in time).

- The application should also use thresholds to indicate if data is not meeting the set criteria. Thresholds can be applicable on all levels/dimensions. It can be used while color-coding info or using other design tricks.
- Thresholds can be used to filter out the records where the DQ criteria are not met.
- A trend line or trend indication let us understand if DQ is getting worse or better, the app should have such a capability to visualize the trend for a BI service, Validation, Dimension, DQ Rule or country level. One should be able to create the trend by any aggregation/dimension chosen. It would be good to have it as a default on the Service and Validation level for each DQ dimension (completeness, conformance ...).
- The Data Quality trend of a service should also be possible to be compared with other services by adding an extra service to the trend line graphic.

In general, the Dashboard should provide different type of charts, choose most relevant or provide the capability that users can shift between options:

- Trend over time (different periods: 1 month, 1 year, 5 year etc.; different periods in time):
- Trend lines
- Area charts
- Bar charts
- Comparison & Ranking
- Correlation
- Distribution
- Heat maps
- If geographical data is provided - use a map!

As a conclusion we can state that the application should be flexible enough for the user to bring forward any visualized DQ statistic based on his/her chosen level/dimension/grouping. Management will rather like to see high level trends based on service level, while a Data Steward wants to dig deeper and select the validations he/she is responsible for and drill down to actual Data Quality rules and start to analyse errors more in detail. Create an operational and strategical dashboard and in second step combine them using interactive features within the application.

- Operational dashboards are used by data analysts, data stewards on daily, weekly basis.
- Strategical dashboards are used by management to get a clear understanding what the quality of the current data is and shift (data/system) enhancement priority activities.