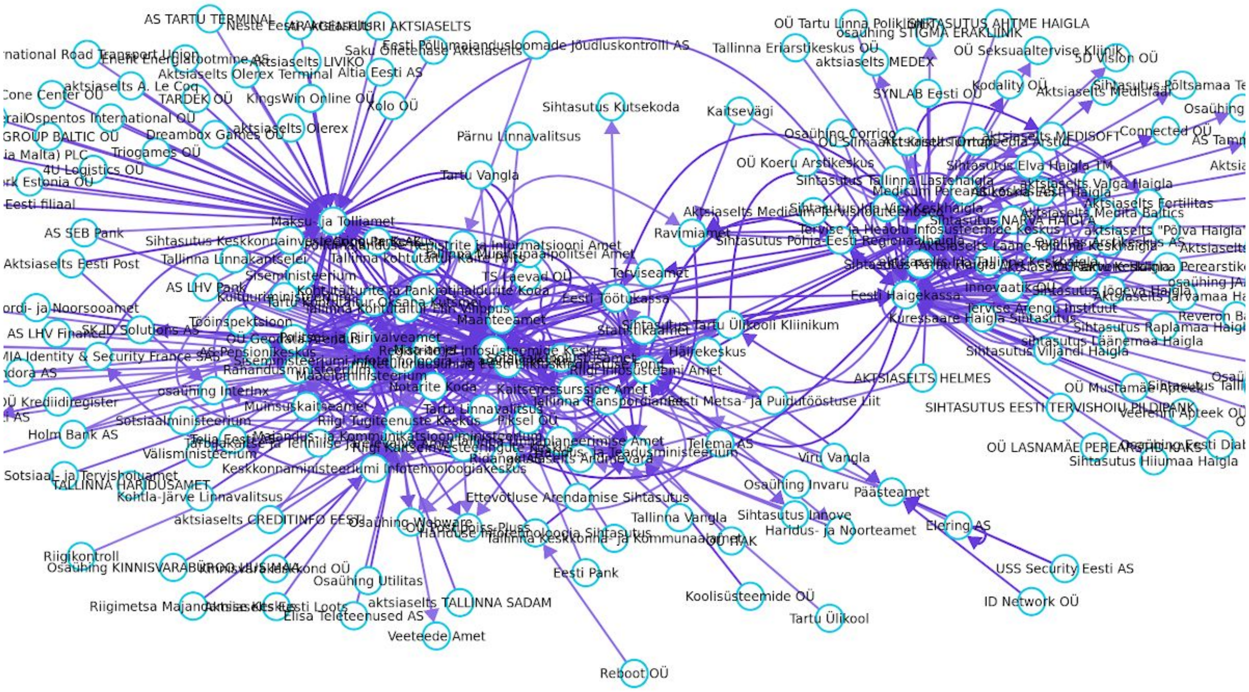

Project A4:
EESTI-X-TEE-HICCUPS

30.11.2020

Joosep Tavits

Mari Seeba

Task 1. Setting up (0.5 points)

- Add the link of the repository also to the report.

Link of the repository: https://github.com/joosepgit/EESTI-X-TEE-HICCUPS

Task 2. Business understanding (1 point)

Business understanding report

This project is not meant to directly benefit anyone, although it could benefit everyone that uses services related to X-tee servers. In other words, it could potentially benefit Estonian society by finding patterns in system usage based on external events that could be foreseen. For example whether services that are used for sending packages during Christmas time experience more hiccups in availability than during any other month of the year. If we manage to discover such occurrences where the systems availability suffers, or worse, doesn't correspond to its reliability demands fixed in Estonian Information and Security systems law, the systems would be demanded to find a solution. For example just to enhance the systems performance capabilities at a time of peak usage, thus creating a better user experience for everyone across the nation.

Our goal is to find out if and how much external events affect the availability of several Estonian systems that are used nationwide on an every-day basis and proof of concept that with some simple techniques and obvious examples, it is possible to estimate the systems availability needs depending on external events. This project's business success criteria would be finding several examples similar to the one mentioned above. This would show that the affected systems need to be capable of boosting their availability at certain times of the year where they reach peak usage.

We have direct contact with people who handle the usage logs of such systems for a living, meaning that we have access to their knowledge and potentially, data covering the logs of all queries made to all systems using X-tee from a few years ago to yesterday. In terms of hardware, we don't have much computing power, but considering our current plan, we will not need more than we have access to for sure. As for requirements, we have a deadline on 16th of December 2020 for finishing the whole project. In terms of security obligations, the data is actually not necessarily confidential as long as it is in anonymized form. The requirements for an acceptable finished work would be to at least have found 3 examples of such systems and having visualized and stated at exactly what time it suffers most and what could be done to handle such events better or prevent them from happening overall. One of the most realised risks, we still do not have access to the data archive because of COVID19. That is because there is a need for physical contact with the supervisor of the data, as it is stored in very large amounts. Instead we are using open data resources which only give an overview of the last 6 months.

Terminology

X-tee : A secure server system used to connect its users' systems without creating one large database

ISKE : Infosüsteemide Kolmeastmelise Etalonturbe Süsteem -> rough translation: Three Layer Security System for Information Systems

This project does not necessarily have monetary benefits as the people administering the above mentioned systems are mainly government officials who have fixed salaries anyway. It does have the benefit of creating a better user experience for all citizens using these systems and making these systems more reliable overall.

As for data-mining goals, we want to establish that the amount of queries and real-life events are strongly correlated and that it could be visualized through monitoring X-tee logs. In addition to that we want to determine the most optimal timing for system maintenance, which would be a time of least queries within the average week for example. We also want to offer service providers ISKE security classes. The success criteria for data-mining here is choosing the correct time periods to analyze.

Task 3. Data understanding (2 points)

Gathering data

Data Requirements (min):

1. A-dataset: X-tee request between Patsiendiportaal and Digilugu at the COVID-19 period 2020
2. B-dataset: X-tee request from eesti.ee to Eksamiinfosüsteem containing period from May to August 2020
3. C-dataset: COVID-19 tests trends statistics (daily positive, negative tests counts)
4. A- and B-datasets: Data should contain DateTime (hourly) to estimate the availability class and suitable maintenance time

Data availability

A- and B-datasets are possible to download by day from https://logs.x-tee.ee/EE/gui/ (JSON format).

C-dataset is downloadable https://koroonakaart.ee/et (CSV format).

Selection criteria

Downloadable data field description is here: https://github.com/ria-ee/X-Road-opmonitor/blob/master/docs/opendata/user_guide/ug_opendata_interface.md

To select the needed data client service code was used: https://www.x-tee.ee/service-catalog

Also, to understand the system itself was used RIHA (State IS management system).

To download data (A- and B-datasets), we used external knowledge (set of scripts) and used the outcome. The constraints for these scripts where defined by our team as follows.

A-dataset:

From https://www.riha.ee/Infosüsteemid?searchText=patsiendiportaal&sort=meta.update_timestamp&dir=DESC we got the name of client system and code of client system (`70009770-tis-patsiendiportaal`), also looking on data of `digilugu`. Used constraints:

```
constraints=[{"column":"clientMemberCode","operator":"=","value":"7000977
0"},
{"column":"clientSubsystemCode","operator":"=","value":"tis-patsiendiport
aal"},{"column":"serviceSubsystemCode","operator":"=","value":"digilugu"}
,{"column":"serviceMemberCode","operator":"=","value":"70009770"}]
```

B-dataset:

Filtering data of exam results, where filtered based on RIHA search results: https://www.x-tee.ee/catalogue/EE search word “`eksam`” gave us subsystems: `EE/NGO/90008287/eis-adapter`, as we do not narrow it more, because e-tunnistus, riigieksam and some more results are not needed to separate, because they show services of the same server which availability we study:

```
constraints=[{"column":"clientSubsystemCode","operator":"!=","value":"mon
itoring"},{"column":"serviceMemberCode","operator":"=","value":"90008287"
}]
```

Additional information to B-dataset time constraints: https://www.innove.ee/eksamid-ja-testid/riigieksamid/eksamitulemustest-teavitamine/ which says:

> *The results of the National Examination are known no later than 30 June 2020 and electronic state Examination Certificates will be issued to students no later than 1 July 2020. Since then the electronic state Examination Certificate can also be downloaded from the State Portal.*

C-dataset:

https://koroonakaart.ee/et > Teste päevas (information section) > All (timecontrains) > Abs (% or count) > Download CSV (Format)

Data Description

A- and B-datasets were transformed into CSV format.

A-dataset

File: pats-mai-praegu.csv
RangeIndex: 11974427 entries
Memory usage: 1.5+ GB
B-dataset
File: eksam.csv

RangeIndex: 783130 entries
memory usage: 96.3+ MB

Both A- and B-datasets Data columns (total of 17 columns) (Descriptions of features are based on https://github.com/ria-ee/X-Road-opmonitor/blob/master/docs/opendata/user_guide/cfg_lists/field_data.yaml):

Iindex	Feature	Type	Description	Comments, importance to project
0	clientMemberCode	int64	Member code of the X-Road member (client)	A-dataset: 70009770
1	clientSubsystemCode	object	Subsystem code of the X-Road member (client)	A-dataset: tis-patsiendiportaal
2	producerDurationProducerView	float64	The time it takes for a producer to generate a response and send it, once the request has arrived	Result depends from which side the log record is made (from request or response)
3	requestAttachmentCount	float64	Number of attachments of the request	
4	requestInTs	datetime64[ns]	In the client's security server: the Unix timestamp in milliseconds when the request was received by the client's security server. In the service provider's security server: the Unix timestamp in milliseconds when the request was received by the service provider's security server	One main feature to estimate availability needs
5	requestMimeSize	float64	Size of the MIME-container of the request (sum of the SOAP request message and attachments data size in bytes)	
6	requestSoapSize	int64	Size of the request (bytes)	
7	responseAttachmentCount	float64	Number of attachments of the response	
8	responseMimeSize	float64	Size of the MIME-container of the response (sum of the SOAP response message and attachments data size in bytes)	
9	responseSoapSize	float64	Size of the response (bytes)	
10	securityServerType	object	Type of the security server	
11	serviceCode	object	Code of the service	
12	serviceMemberCode	int64	Member code of the X-Road member (service provider)	A-dataset: 70009770 B-dataset: 90008287
13	serviceSubsystemCode	object	Subsystem code of the X-Road member (service provider)	A-dataset: digilugu B-dataset: eis-adapter
14	serviceVersion	object	Version of the service	

15	succeeded	bool	True, if request mediation succeeded, false otherwise.	Important feature to estimate availability
16	totalDuration	float64	Request duration from sending the request to getting a response from the client's perspective	Important feature to estimate availability

C-dataset

File: teste-pevas.csv

RangeIndex: 253 entries
memory usage: 10.0+ KB

Data columns (total 5 columns):

Index	Feature	Type	Description	Comment
0	DateTime	object	Date of COVID-19 tests results	To find correlation with A-dataset date
1	Positiivne	int64	Positive tests count	
2	Negatiivne	int64	Negative tests count	
3	Positiivsete testide %	float64	Percent of positive tests of all tests	
4	Testide_arv	int64	Sum of tests (Positiivne + Negatiivne)	To find correlation with A-dataset date

Exploring data

A-dataset shows clearly how people request their results from Patsiendiportaal -> Digilugu. Data does not show exactly what results were requested. At a moment in our project we make assumption, that main request and response are related with COVID-19 test results. For that we have C-dataset. We need to estimate the correlation of day based results (counts per day). Both datasets need date_time transformation into comparable format. Preliminary results are shown in figure 1.

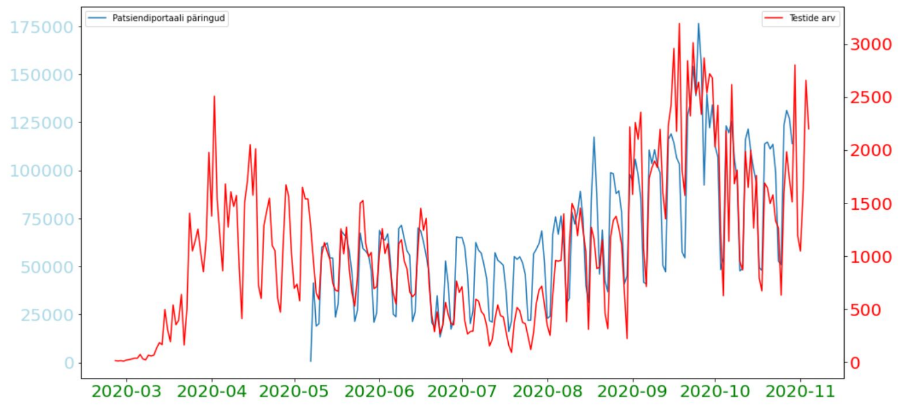

Figure 1. Preliminary results of A and C dataset visualization

B-dataset shows requests and responce between Riigiportaal and Eksamiinfosüsteem. We do not extract the State exam results and Certificate requests and responses to show the real system workflow. We hope to show the hiccups when the results are made available to pupils. For that we need datetime and date of exam results publishing.Preliminary results are shown in figure 2.

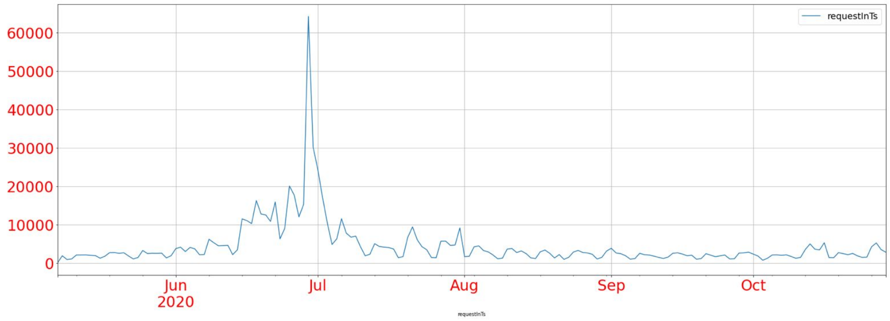

Figure 2. Preliminary results of B-dataset

Also we can find the request (Client) and response (Producer) duration. Duration can be NaN, if the response and request are not directly connected in monitoring server. Duration can also related with attachments and sizes.

To estimate the maintenance timing we need to group data of A and B dataset weekday based and hourly based.

To estimate the availability class we need to understand the distribution of requests and find the possible maximum service interruption duration with minimum influence to clients.

Data Quality verification

Known problem: request and response records are not clearly to connect to each other because of X-tee servers software update conflict between versions. It does not disturb to make the proof of concept models and find needed correlations between real life events and X-tee monitoring log records.

We miss archive data to find better comparable models for longer time period.

At a moment we use only last 6 month data to get results. As we provide simple proof of concept models, then main idea is possible to show. If we can have more data (request is givet to RIA monitoring service owner), we can improve our results.

Task 4. Planning your project (0.5 points)

Each task contains domain knowledge study, consultation with domain experts at least 2 h to both team members. Team mostly works together in pair programming style.

Main tasks:

1. Data gathering and understanding (domain study) - 6h +6h

   Asking RIA personnell for the data, how to look at it, which attribute describes what real-life feature etc.

2. Environment and Data preparation...... 4h + 4h

   What do we need to do in order to get the data into analyzable form, which parts of the data to ignore, which time periods to prioritize etc.

3. Working with data: the correlation of real-time events and logs, for all datasets.... Visualize 6h+6h

   In this step and the next we will be applying methods learned during this course in order to bring the above-mentioned goals to life.

4. Heatmap to find suitable maintenance time... (features) 3h+3h
5. Availability class estimation (ISKE availability class) 3h+3h

   Comparing our results for specific systems with the demands presented for certain ISKE availability classes.

6. Results presentation 6h+6h