SOEN 6611 - SOFTWARE MEASUREMENT: THEORY AND PRACTICE

Project Report on Task 4

Summer 2022

Course Instructor: Dr. Olga Ormandjieva

Source: SEI *Implementing Goal-Driven Measurement* course material (adapted).

Team 11	
#Student ID	**Name**
40198687	Hasandeep Singh
40159259	Anushka Sharma
40218417	Jasleen Kaur
40205476	Kavleen Kour Sidhu

NOTE: the scope of the Steps 3, 4 and 5 is reduced to Validity, Vincularity and Veracity.

Project Step 4 /S22 (5 points, due before midnight on July 29th)

Objective: Planning of the measurement process

Summary of Step 4.

The objective of this step 4 is to identify and plan the activities that must be accomplished in order to collect, store, process, and report the measurements necessary to build your 3V's indicators.

To help you with this portion of the job, here are some guidelines (the order may differ from the listed below):

a) Review the action checklist in section 1;

b) Analyze the tasks in the checklist to see if they are sufficient to collect, store, analyze, etc. the required measures (data elements) for your indicators.

Specific tasks should be defined for:

- Prepare [specific data collection]

- Collect [defined data]

- Analyze [the results]

- Report [the results]

c) Document your tasks using the template provided below. Label each measurement task as MTXX (XX is the sequential number of the task). Trace it to the corresponding DAXX / INXX / MGXX. [DAXX is the label of the corresponding Data Element, INXX is the label of the corresponding Indicator, MGXX is the label of the corresponding measurement goal).

You must remain consistent with all of the base and derived measures defined in the previous step 3.

1. **Checklist to complete:**

#	**Checklist**	
a.	List and label as DAXX the data elements (base measures) (XX is the sequential number of the data element).	☑
b.	Define the intervals of time and frequency when the data recorded would be measured during measurement process	☑
c.	Define the time frames required and used for moving measurement results from the points of collection to databases or users	☑
d.	Define how the data is to be stored and how the data will be accessed.	☑
e.	Create methods and procedures (or forms, or tools) for collecting and recording the data to be measured (Base measures)	☑
f.	Identify who is responsible for designing the database (or tool), and for entering data, retaining data, and managing this data.	☑
g.	Determine how the data will be viewed, analyzed and reported	☑
h.	Determine on what basis different timeline data would be compared and validated.	☑
i.	Identify the supporting tools that must be developed or acquired to help you automate and administer the measurement process.	☑
j.	Prepare a short process guide for collecting, analyzing, and reporting the data	☑

2. Measurement Plan Checklist:

2.1) Labels

Measurement Goals	Labels
Increasing the **Validity** of the Big data at regular time intervals	**MG01**
Enhancing the **Vincularity** of the Big Data	**MG02**
Increasing the **Veracity** of the big data over the given time frames	**MG03**

Indicators	Labels
Mval	**I01**
Mvin	**I02**
Mver	**I03**

Base Measures	Labels
Nds(MDS)- Number of datasets	**BM01**
Nds_cr(MDS) - Number of credible Datasets	**BM02**
Nrec_comp- Number of compliant records in a Dataset	**BM03**
Rec_Trace (MDS)-Provides the total number of records that are traceable in MDS	**BM04**
Ldst(MDS) - length of dataset	**BM05**

Rec_no_null (MDS) -Frequency of records (in MDS) with no null values	**BM06**
Lbd-Total Number of records in MDS	**BM07**
Rec_acc_age(MDS) - total number of records with ages that fall within the acceptable range	**BM08**
Pj - Provides the total number of duplicate items and their specific count in each dataset	**BM09**
N_succ_req - Number of successful requests	**BM10**
N_req(MDS) - Number of Requests	**BM11**
Time	**BM12**

2.2) Frequency of Data Collection

Initial dataset: Once the requirement is established and the initial dataset is identified(T1)

Incremental: Dataset to be collected for the new incremental data. Here the dataset was divided into 3 subsets and collected at T1, T2, and T3 timeframes. (Where T2-T1 = T3-T2)

2.3) TimeLine

Planned: [min 70 person-hours, max: 90 person-hours]

2.4) Procedure for collecting and recording data.

Dataset is hosted on Kaggle, and we can download the same and split it into 3 datasets for analysis. As this dataset is not big enough and used for prototyping only, a local filesystem is used to store the data, and python/pandas are used to analyze the same.

As the dataset grows and a filesystem is not enough for storing the same, the team may decide to move to a distributed file systems like Hadoop, and Spark for storing the same.

2.5) Data storage strategy.

Data is stored as it is and in memory preprocessing is done using python.

2.6) Role and responsibility

Role	**Responsibility**
Product Owner/Project Manager	● Identify scope and user requirement ● All Resources identification ● Assign the Roles and responsibilities ● Evaluate and measurement process
Data Scientist/Developer	● Identify Dataset which fulfills requirement ● Analyze report ● Communicates results ● Evaluates measurement tasks ● Develop analytical code to identify data for analysis ● Develops report and documentations
QA Analyst	● Execute codes developed by Developer ● Do manual verifications on the correctness of analysis ● Verifies correctness of documentation

3. **Plan tasks / activities:**

T1 = Day 1, T2 = T1+2 days, T3 = T2+3 DAYS

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
MT01	Identify the stakeholders who are interested	BM01- BM02	Product owner/project manager		3	24	During the planning phase	Based on the survey	Party involved who will have a commitment towards quality improvement
MT01	Data Collection	BM01 - BM12	Developers/ Analysts	Product Managers	2	16	The data available is ready for use at the source	Kaggle	The data means to be collected to perform the measurement steps involved in Big Data Project
MT02	Divide the allocated data into different timeframes	BM01 - BM12	Developers/ Analysts	Product Managers	0.5	4	At the iteration, the data collected and is analyzed	Microsoft Excel. Jupyter Notebook	This helps in creating datasets and helps in creating comparison which are better to perform analysis
MT03	Compare and analyze data over	BM01 - BM12	Developers/ Analysts	Product Managers	2	16	At the iteration, the data	Microsoft Excel. Jupyter Notebook	The data separated into several time frames need to be consistent and

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
	different time frames						collected and is analyzed		competent at different time schedule
MT04	Calculate the veracity of Collected BIG Data	BM07, BM09,BM06, BM08,BM10, BM11, BM02, I03,MG03	Product Manager	Strategic Manager	1.5	12	After collection of data and when base measures have been calculated over different time frames.	Microsoft Excel. Jupyter Notebook	Calculating the collected big data with characteristics of accuracy, completeness, currentness and availability of data
MT05	Calculate the validity of Collected BIG Data on timeframes	BM02, BM03, BM01,BM12, I01, MG01	Product Manager	Strategic Manager	1	8	when base measures have been calculated over different time frames	Microsoft Excel. Jupyter Notebook	Validity has been characterized with compliance and credentiality of the big data being used.
MT06	Calculate the vincularty of the big data over different time frames	BM01,BM04, BM05,BM12	Product Manager	Strategic Manager	1	8	After calculation of required data elements. Calculate over different time frames	Microsoft Excel. Jupyter Notebook	Vincularity is comparable to traceable access of audit data.
MT07	Calculate entropy and MAX	BM07,BM09, I03,MG03	Analysts	Product Managers	0.5	4	After calculation of	Microsoft Excel.	The max entropy will be calculated

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
	entropy of the multiple datasets accuracy (Veracity).						required data elements. Calculate over different time frames	Jupyter Notebook	and can be indicated further to find accuracy
MT08	Calculate the completeness and report for veracity calculation	BM07, BM06,I03, MG03	Analysts	Product Managers	0.5	4	After calculation of required data elements. Calculate over different time frames	Microsoft Excel. Jupyter Notebook	Completeness indicates help to characterize better veracity calculation
MT09	Calculate the currentness for calculating veracity.	BM07, BM08,I03, MG03	Analysts	Product Managers	0.5	4	After calculation of required data elements. Calculate over different time frames	Microsoft Excel. Jupyter Notebook	The output percentage relates to the acceptable range of data when using boxplot criteria
MT10	Availability of big data to calculate veracity	BM02, BM11, I03, MG03	Analysts	Product Managers	0.5	4	After calculation of required data elements. Calculate over different time frames	Microsoft Excel. Jupyter Notebook	The output percentage signifies the successful requests when compared to the total requests made.

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
MT11	Credibility calculations of big data and report for validity calculation	BM)!,BM02, I01, MG01	Product Managers	Strategic Managers	0.5	4	After calculation of required data elements. Calculate over different time frames	Microsoft Excel. Jupyter Notebook	Accurate and precise records help to increase credibility
MT12	Compliance of big data and reporting for validity calculation	BM03, BM01,I01,MG01	Product Manager	Strategic Manager	0.5	4	After calculation of required data elements. Calculate over different time frames	Microsoft Excel. Jupyter Notebook	Compliance is the degree of records that are accurate with the expected information of the dataset.
MT13	Traceability and report for vincularity calculation	BM04, BM05	Product Manager	Strategic Managers	1	8	After calculation of required data elements. Calculate over different time frames	Microsoft Excel. Jupyter Notebook	Traceability signifies the degree to which records can be backtracked to their specific context.
MT14	Length of Big Data Calculation	BM07, I03, MG03	Analysts	Product Managers	1	8	It can be calculated after analyzing the data for timeframes (T1, T2, T3)	Microsoft Excel. Jupyter Notebook	This can be defined as the total number of records in the dataset

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
MT15	Count the total number of duplicate items (Pj)	BM09, I03, MG03	Analysts	Product Managers	1	8	After three iterations of data over time frame T1,T2,T3, analyze the data	Microsoft Excel. Jupyter Notebook	This can be used to find out the count of duplicate records in dataset
MT16	Calculate rec_no_null (MDS)	BM06, I03, MG03	Analysts	Product Managers	1	8	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	Indicates total number of non-null values in the database.
MT17	Calculate rec_acc_age(MDS)	BM08, I03, MG03	Analysts	Product Managers	1	8	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the number of records that lie within the acceptable range.

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
MT18	Calculate N_succ_req	BM10, I03,MG03	Analysts	Product Managers	1	8	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the number of successful requests
MT19	Calculate N_req(MDS)	BM11, I03, MG03	Analysts	Product Managers	1	8	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the total number of requests.
MT20	Calculate Nds_cr(MDS)	BM02, I01, MG01	Analysts	Product Managers	0.5	4	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the total number of credible datasets in big data.

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
MT21	Calculate Nrec_comp (MDS)	BM03, I01, MG01	Analysts	Product Managers	0.5	4	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the total number of compliant records in the dataset.
MT22	Calculate Nds(MDS)	BM01, I02, I03, MG01, MG02	Analysts	Product Managers	0.5	4	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the total number of datasets.
MT23	Calculate Rec_trace(MDS)	BM04, I02, MG02	Analysts	Product Managers	0.5	4	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the total number of records that are traceable.

#	Task / activity (what / how)	Trace to DAXX / INXX / MGXX	Responsible (who)	Participants (with whom)	Estimated duration (In days)	Estimated effort (In person-hours)	Schedule (when)	Tool (With what)	Rationale
MT24	Calculate Ldts(MDS)	BM05, I02, MG02	Analysts	Product Managers	1	8	It can be calculated after analyzing the data for timeframes (T1, T2, T3) over three iterations	Microsoft Excel. Jupyter Notebook	This indicates the total number of occurrences of elements in the dataset.
MT25	Record and analyze time to calculate over different time frames	BM12, I01, I02, I03, MG01, MG02, MG03	Analysts	Strategic Managers	1	8	It will be calculated 3 times at start, mid and end	Microsoft Excel. Jupyter Notebook	Time will indicate the different time frames over which data is collected and analyzed
	Total :				**25**	**200**			

Assumption: One working day consists of 8 hours

4. Measurement process guide

Write a measurement data collection guide, how the data are to be stored and how the data will be accessed, how the data will be analyzed and reported. to make it easier for the different people involved to collect/analyze/report measurement data / results. This guide can be organized by time of data collection/analysis/reporting (daily, specific days of the week, start or end of an iteration, etc.). This short guide should be used as a reminder and should fit in one page.

Concerned Personnel	Activity	Description	Time Frame
Data Scientist / Developer	Data Collection	**Base Measures** (i.e., Nds, Ldst, Rec_no_null) are to be collected here. Few of the base measures like **Nds, Lbd,** etc. are collected manually, whereas other measures like **Rec_trace, Nds_cred,** etc. are collected/ calculated via implementing methods in **Python code.** A list of base measures has been provided in Step 3 (along with formulas).	This activity is performed at the end of each time frame. This is to ensure derived measures can be presented at the end of each frame.
Data Scientist / Developer	Data Measurement	**Derived Measures** are calculated after base measures are collected. The formulae for **Mval, Mvin, Mver** and rest of the derived measures are presented in step 3 (along with formulas). These final calculations form the basis for the later process of analyzing.	The task is performed at the end of each time frame after the collection of base measures are done.
QA Analyst	Analysis / Interpretation	Analysis / Interpretation model developed in Step 2, helps us to understand the meaning of **Mval, Mvin and Mver** (other derived measures also) values given by previous tasks. The model helps us recognize the trends/ variations in different time frames. Note that the interpretation model can be changed upon **feedback** from reports, and we would have to repeat this task.	The task is performed two ways. Firstly, at the end of each frame, to comprehend them individually, and then after all the frames are done, to compare and interpret the findings.
Data Scientist	Reporting	The cognitive model of understanding from the previous task is applied to the English language and formed as a report. It is formulated in such a way to highlight the interpretation into words common to software/ topic provided. Any discrepancy found in the interpretation model (in relation to topic) is reported as **feedback** to the previous step and a new report is generated after correction.	Performed after analysis of all measures are done. Can be performed repeatedly if any improvements are suggested to the analysis model.
Team Lead / Manager	Results	The final report after n number of iterations is mapped to goals described in step 1 & operationalized in step 2. The final result of the tasks and its observations are formulated and presented as end results.	After the report in the previous step is finalized.