SOEN 6611 - SOFTWARE MEASUREMENT: THEORY AND PRACTICE

Project Report on Task 4

WINTER 2022

Course Instructor: Dr. Olga Ormandjieva

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

TEAM 5	
Student #	**Name**
40163582	Mohammod Suhel Firdus
40157109	Vivekananda Reddy Gottam
40184906	Saswati Chowdhury
40156971	Milesh Kotadia

PROJECT STEP 4

NOTE: the scope of Steps 3, 4 and 5 is reduced to the 3V's only (Volume, Variety, Velocity)

Project Step 4 /W22 (5 points, due on April 3rd): Planning of the measures

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]
- Check the quality [of the data collected, for instance, remove outliers where applicable]
- Analyze [the results]
- Report [the results]

c) Identify what else the organization has to do in order to complete the above tasks.

d) Document your tasks using the template provided below and list the rationale for each. 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. If necessary, you can improve these measures at the end of this document.

1. Checklist

#	Checklist	
a.	List and label as DAXX the data elements (base measures) (XX is the sequential number of the data element).	☑
b.	Define the frequency of collection and the points in the process where the measurements will be made.	☑
c.	Define the timeline required for moving measurement results from the points of collection to databases or users	☑
d.	Create procedures (or forms, or tools) for collecting and recording the measured data	☑
e.	Define how the data are to be stored and how the data will be accessed.	☑
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 analyzed and reported.	☑
h.	Identify the supporting tools that must be developed or acquired to help you automate and administer the measurement process.	☑
i.	Prepare a short process guide for collecting the data.	☑

2. Measurement Plan Checklist:

2.1 Labels

Measurement Goals	
Measurement Goals	**Labels**
Increasing the **Volum**e of big data sets	**MG01**
Accelerate the Big Data set **Velocity**	**MG02**
Enhancing **Variety** in Big Data	**MG06**

Indicators:	
Indicators	**Labels**
Mvol	**I01**
Mvel	**I02**
Mvar	**I03**

Base Measures:	
Base Measures	**Labels**
Ndde - Number of Distinct Data Elements	**DA01**
Lbd: Length of Big Data (Number of Records)	**DA02**
Nds: Number of Datasets	**DA03**
Time	**DA04**

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 system 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	Student # [1]
Product Owner/Project Manager	• Identify scope and requirement • Resource identification	40156971

	• Role and responsibility assignment • Evaluate measurement process	
Data Scientist/Developer	• Identify Dataset which fulfills requirement • Analyses report • Communicates results • Evaluates measurement tasks • DevelopTe analytical code to identify data for analysis • Develops report and documentations	40163582, 40184906
QA Analyst	• Execute codes developed by Developer • Do manual verifications on the correctness of analysis • Verifies correctness of documentation	40157109

[1] - These roles to student ID mapping are just indicative.

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	MG01/ MG02/ MG06	Product owner/project manager		3 Days	24[a]	During the planning phase	Based on the survey	Party involved who will have a commitment towards quality improvement

MT04	Define source of data collection, Frequency, Process of data collection	MG01/ MG02/ MG06	Product owner/ Project manager	Data Scientist/ Developers	2 days	16	During the planning phase	Collect the sources from trusted sources	Data sources are to be identified and data providers are to be notified
MT01	Calculate the number of distinct data elements (Ndde) present in the dataset	DA01	Data scientist/ Developer	Team of Data Scientists/ Developers	0.5 day	4	At the beginning of time frame T1, T2, T3	Google collab	It will be used to evaluate the value of big data volume for all three phases: Extraction, preprocessing, processing in different time frames.
MT02	Calculate the length of the dataset (Lbd)	DA02	Data scientist/ Developer	Team of Data Scientists/ Developers	0.5 day	4	At the beginning of the time frame T1, T2, T3	Google collab	It will be used to evaluate the variety present in the dataset and to compare it in different time frames
MT03	Calculate the number of datasets (Nds)	DA03	Data scientist/ Developer	Team of Data Scientists/ Developers	0.5 day	4	At the beginning of time frame T1, T2, T3	Manual calculation	It will give the count of the various dataset which are present to evaluate the variety of big data
MT04	Record the time period (T) for analyzing the velocity of big data.	DA04	Data scientist/ Developer	Team of Data Scientists/ Developers	0.5 day	4			It will be used to find the speed at which the volume of big data is increasing and the speed of processing that data.
MT05	Calculate the volume characteristic by substituting	I01/ DA01	Data scientist/ Developer	QA Analyst	1 day	8	At the beginning of time frame T1, T2, T3	Google collab	Generates the value of Mvol derived measure which depicts the volume of big data.

	the values of its base measures in the formula								
MT06	Calculate the velocity characteristic by substituting the values of its base measures in the formula	I02/ DA01/ DA04	Data scientist/ Developer	QA Analyst	1 day	8	At the beginning of time frame T1, T2, T3	Google collab	Generates the value of Mvel derived measure which depicts the relative growth of big data over a time period
MT07	Calculate the variety characteristic by substituting the values of its base measures in the formula	I06/ DA01/ DA02/ DA03	Data scientist/ Developer	QA Analyst	1 day	8	At the beginning of time frame T1, T2, T3	Google collab	Generates the value of Mvar derived measure which depicts the variety in big data.
MT07	Miscellaneous Activity: Report Generation and Communication	NA	Data scientist/ Developer	Project/Product manager.	1 day	8	At the end of each time frame T1, T2, T3	Graph, plots using report generation tool(excel/matplotlib) Communication channel of the organization(email/announcements)	It provides the visualization of the changes happening in the values of derived measures in different time frames.
	Total :				**11**	**88**			

[a] One working day consists of 8 hours

4. Data collection guide

Write a data collection guide to make it easier for the different people involved to collect data. This guide can be organized by role and/or by the time of data collection (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.

People who collect the data	**Data Collection**	**Role of the collected data**	**Time of data collection**
Data Scientist/ Developer	Number of distinct data elements (Ndde)	This base measure provides the count of all the unique data elements in the dataset and it is calculated using **python code in Google collab**. This data will be used to determine the volume and variety of the dataset	The Ndde needs to be calculated at the beginning of each time frame.
Data Scientist/ Developer	Length of the Big Data (Lbd)	It gives the number of records present in the dataset and it is calculated using **python code in Google collab**. This data will be passed to data scientists to calculate the variety in big data.	The Lbd needs to be calculated at the beginning of each time frame.
Data Scientist/ Developer	Number of datasets (Nds)	This base measure gives the count of the number of datasets and it is **calculated manually.** This data will be used to calculate the variety.	The Nds need to be updated in each time frame.
Data Scientist/ Developer	Volume of BigData (Mvol)	Volume of big data will show the change in the volume of the dataset across different time periods. It will be determined by the collected information of the number of distinct data elements present in that dataset (Ndde) and	It will be evaluated for three different phases of the big data pipeline: Extraction, preprocessing, and processing in each time frame.

		using the formula. ***Mvol (MDS) = Ndde (MDS) * log_2 ((Ndde (NDS)))***	
Data Scientist/ Developer	Velocity of Big Data (Mvel)	Velocity will show the rate of change in volume of Big Data in different time frames. To measure, it will require the volume of the dataset in the present in the different time frames and applied to the formula. ***Mvel(MDS) = ((Mvol(MDS_{T2}) - Mvol(MDS_{T1})) / Mvol(MDS_{T1}) * 100***	This data will be generated for each time frame to visualize the change in the volume over time.
Data Scientist/ Developer	Variety in Big Data (Mvar)	Variety will indicate the various forms of data present in the dataset during each time frame. It will need the data of Ndde, Lbd and Nds during each time period to calculate it through the formula. ***Mvar (MDS) = Ndde (DE) * W_{Ndde} + Lbd (MDS) * W_{Lbd} + Nds(MDS) + W_{Nds}***	Variety in the dataset will be collected in each time frame.