SOEN 6611
SOFTWARE MEASUREMENT
INSTRUCTOR: DR. OLGA ORMANDJIEVA

PROJECT
TASK 4

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

SUBMITTED ON : 21-October-2022

Declaration:
We, the members of the team, have read and understood the Fairness Protocol and the Communal Work Protocol, and agree to abide by the policies therein, without any exception, under any circumstances, whatsoever.

TEAM-7

SIDDHARTHA NANDA	40200496
BARIQ ISHTIAQ MOHAMMED	40208194
RAJAT KUMAR	40201807
VIKYATH SRINIVISAULU	40218245

TEAM LEADER

SIDDHARTHA NANDA	40200496
Email ID:	siddhartha.nanda@mail.concordia.ca

PROJECT STEP 4(Due before noon on Nov. 4th**)**

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): 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.	Identify the supporting tools that must be developed or acquired to help you automate and administer the measurement process.	☑
d.	Prepare a short process guide for collecting the data, how the data are to be stored and how the data will be accessed, how the data will be analyzed and reported.	☑

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 the Distinct Data Elements	**DA01**
Lbd: Length of the Big Data (Number of Records)	**DA02**
Nds: Number of Datasets	**DA03**
Time	**DA04**

2.2 Frequency of Data Collection

Initial dataset: The initial dataset is identified once the requirement of data is established(T1)
Incremental dataset: The Dataset to be collected for the new incremental data. It was here split into 3 subsets and is collected at time frames T1, T2 and T3. (T2-T1 = T3-T2)

2.3 Time-Line

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

2.4 Procedure for collecting and recording data.

We may get the dataset from Kaggle and divide it into 3 separate datasets for study. A local disc is used to store the data because it is not large enough to be used for anything other than prototyping, and Python and Pandas are used to analyze it.

The team may opt to switch to a distributed file system like Hadoop or Spark for storing the dataset when it increases and a file system is no longer enough for doing so.

2.5 Data storage strategy.

Python is used to perform preprocessing on the data while it is stored in the memory.

2.6 Role and responsibility

Role	Responsibility	Student #
Product Owner/Project Manager	• Identification of the scope and requirement • Identification of resources	4020####
	• Assigning role and responsibility to team members • Evaluating the measurement process	

Data Scientist/Developer	• Identifying Dataset that meets the requirement • Analyzing the reports • Communicating the results • Evaluating the measurement tasks • Developing analytical code to identify the data for analysis • Doing the report and documentation	
QA Analyst	• Execution of codes developed by the developer • Manually verifying the correctness of the analysis • Checking the correctness of the documentation and the report	

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 interested stakeholders	MG01/ MG02/ MG06	Product owner/Project manager		3 Days	24	During the planning phase	Based on the survey	Party who will be dedicated to 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	Collection of the resources from the sources trusted	Identification of data sources and notification of data providers are required.
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	The value of the big data volume will be assessed for each of the three phases—extraction, preprocessing, and processing throughout various time frames.
MT02	Calculating the dataset length (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 applied to assess the dataset's diversity and compare it across various 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 provide a count of the different datasets that are available to assess 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 determine how quickly big data is being processed and how quickly its volume is growing.

MT05	Calculate the volume characteristic by substituting the values of its base measurements in the formula	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.

MT06	By changing the values of its basic measures in the formula, calculate the velocity characteristic	I02/ DA01/ DA04	Data scientist/ Developer	QA Analyst	1 day	8	At the beginning of time frame T1, T2, T3	Google collab	creates the value of a measure developed from Mvel that shows the relative growth of big data over time
MT07	By changing the values of its basic measures in the formula, calculate the velocity characteristic.	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: Communiction and the generation of report	NA	Data scientist/ Developer	Project/Product manager.	1 day	8	At the end of each time frame T1, T2, T3	Report generating software (Excel/Matplotlib) graphs and charts Organizational announcements and email channels of communication	It offers a depiction of the changes in derived measure values over various time frames.
	Total :				**11**	**88**			

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 collecting 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, which is calculated using Python code in Google Colab, gives the total number of distinct data elements in the dataset. This information will be used to estimate the size and scope of the dataset.	This Ndde needs to be calculated at the startof each time frame.
Data Scientist/ Developer	Length of the Big Data (Lbd)	It provides the dataset's record count and is calculated in Google Colab using python code. To determine the variety in big data, this data will be sent to data scientists.	The Lbd needs to be calculated at the startof each time frame.
Data Scientist/ Developer	Number of datasets (Nds)	This fundamental metric, which is manually calculated, counts the amount of datasets. The variety will be determined using this information.	The Nds need to be updated in each time frame.
Data Scientist/ Developer	Volume of BigData (Mvol)	Big data volume will demonstrate how the dataset's volume has changed over time. The information gathered regarding the number of distinctive data elements existing in that dataset (Ndde) will be used to make this determination.	It will be assessed for each time frame's extraction, preparation, and processing phases, which make up the three stages of the big data pipeline.

		The formula: ***Mvol (MDS) = Ndde (MDS) * log_2 ((Ndde (NDS)))***	
Data Scientist/ Developer	Velocity of Big Data (Mvel)	In various time frames, Velocity will display the rate of change in the volume of Big Data. It will be necessary to apply the formula to the volume of the dataset that is currently available for the various time frames in order to measure. The formula: ***Mvel(MDS) = ((Mvol(MDS_{T2}) - Mvol(MDS_{T1})) / Mvol(MDS_{T1}) * 100***	To show how the volume changes over time, data will be generated for each time frame.
Data Scientist/ Developer	Variety in Big Data (Mvar)	Variety will show the many types of data that were present in the dataset at each point in time. To compute it using the formula, it will need the data of Ndde, Lbd, and Nds for each time period. The formula: ***Mvar (MDS) = Ndde (DE) * W_{Ndde} + Lbd (MDS) * W_{Lbd} + Nds(MDS) + W_{Nds}***	Each time frame will yield a different collection of data for the dataset.