**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 frequency of collection and the points in the measurement 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. |

**The labels associated with measures and indicators below**

| **Measure/Indicator** | **Corresponding type** | **Label** |
| --- | --- | --- |
| Measuring Veracity of Big Data | Measurement Goal | MG01 |
| Measuring Vincularity of Big Data | Measurement Goal | MG02 |
| Measuring Validity of Big Data | Measurement Goal | MG03 |
| M(vin) - measure of vincularity | Indicator | IN01 |
| Percentage change in vincularity | Indicator | IN02 |
| M(ver) - measure of veracity | Indicator | IN03 |
| Percentage change in veracity | Indicator | IN04 |
| M(val) - measure of validity | Indicator | IN05 |
| Nds : Number of datasets | Base | DA01 |
| Nds\_cr : number of credible datasets | Base | DA02 |
| Nrec\_comp: number of compliant records | Base | DA03 |
| Rec\_trace: number of traceable records | Base | DA04 |
| Ldst : length of record | Base | DA05 |
| Lbd : number of records in dataset | Base | DA06 |
| Rec\_acc\_age : number of records within acceptable ranges | Base | DA07 |
| N\_succ\_req : number of successful requests | Base | DA08 |
| N\_req : total number of requests | Base | DA09 |
| Rec\_no\_null : record with no null values | Base | DA10 |
| P(j) : number of duplicate items in specific dataset. | Base | DA11 |

1. **Plan tasks / activities**

| **#** | **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** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Identifying the need to improve the quality of big data. | MG01, MG02. MG03 | Project Manager | Data Engineer, Data Analyst, Developers | 1 | 8 | At the start of the project | Excel Report | Need for a standardized quality measurement model |
| 2 | Identifying the requirement needed for the measurement process | MG01, MG02. MG03 | Project Manager | Data Engineer, Data Analyst, Developers | 2 | 16 | At the start of the project | Excel Report | Need for a standardized quality measurement model |
| 3 | Determine the person/Stakeholder involved in this measurement process. | MG01, MG02,MG03 | Project Manager | Data Engineer, Data Analyst, Developers,End Users | 2 | 16 | Planning Phase of the Project | Based on the requirement elicitation document. | All the people involved in this measurement process are committed to project quality of V’s of Big Data. |
| 4 | Identifying and defining the data elements. | MG01,MG02,MG03,IN01,IN02,IN03,DA01-DA13 | Data Engineer | Project, Manager, Data Analyst, Developers,End Users | 2 | 16 | Planning Phase | Excel | Identify the data elements to perform measurement on it. |
| 5 | Determining base measure and derived measure from the combination of one or more base measures. | MG01,MG02,MG03,IN01,IN02,IN03,DA01-DA13 | Project Manager | Data Engineer,Data Analyst, Developers | 2 | 16 | Planning Phase | Formula Editor | Identify base measure and derive the measures from it which will be used for Indicators |
| 6 | Identify the source of datasets | MG01,MG02,MG03,IN01,IN02,IN03,DA01-DA13 | Data Engineer | Project, Manager, Data Analyst, Developers,End Users | 1 | 8 | Planning Phase | From Kaggle Website | Data Sources are identified and downloaded. |
| 7 | Define Data Collection Frequency | MG01,MG02,MG03,IN01,IN02,IN03,DA01-DA13 | Project Manager | Data Engineer, Data Analyst, Developers,End Users | 1 | 4 | Planning Phase | Discussion with team | Team agrees upon the frequency to collect data from the sources |
| 8 | Define Data Collection Procedure | MG01,MG02,MG03,IN01,IN02,IN03,DA01-DA13 | Data Engineer | Data Analyst, Developers | 3 | 24 | Development Phase | Using the Data Extraction Tools and Programming languages | Specify how to extract data for quality measurements |
| 9 | Data Collection | MG01,MG02,MG03,IN01,IN02,IN03,DA01-DA13 | Data Engineer | Developers | 2 | 16 | Development Phase | Using Data Extraction Tools | Collecting data allows us to store and analyze important information. |
| 10 | Data Processing | IN01,IN02,IN03,DA01-DA13 | Developer, Data Analyst,  Data Scientist | Data Scientist,  Project Manager | 2 | 16 | Development Phase | Using Programming Language or Excel | Data processing is essential for organizations to create better business strategies and increase their competitive edge. |
| 11 | Perform Required Calculations on Data | ,IN01,IN02,IN03,DA01-DA13 | Developer | Data Engineer, Developer, Data Analyst | 2 | 16 | Development Phase | Using Programming Language or Excel | To derive the quantitative value of measures and indicators. |
| 12 | Generate Reports | IN01, IN02, IN03 | Data Analyst | Data Engineer, Developer | 3 | 24 | Development Phase | Using Visualization Tools | Prepare reports with appropriate charts to interpret results |
| 13 | Circulating reports to various stakeholders | MG01, MG02, MG03 | Project Manager | Data Analyst | 1 | 8 | Testing | Email or PDF reports | Sharing the result and analytics of the measurement process to the stakeholder who is involved in the measurement process. |
| 14 | Postmortem analysis of all the V’s of Big Data Set. | MG01, MG02, MG03 IN01, IN02, IN03 | Project Manager | Data Engineer, Developer, Data Analyst, End User | 1 | 4 | End of the Project | Power Point Presentation | Assessments of decisions are made based on the information generated during the analysis phase |
| 15 | Communicate with stakeholders with further improvement in the measurement process. | MG01, MG02, MG03 | Project Manager | Investors,  Clients,  Senior Management | 1 | 6 | End of the Projec | Email or Power Point Presentation | Communicating with the stakeholder to get feedback and working on it until they are satisfied. |
|  | **Total :** |  |  |  | **26** | **198** |  |  |  |

1. **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.

The aim for the collection guide is to give a short process overview of the data collection, storage and analysis. The data is collected from source and further ingested into the ETL system where it is preprocessed after which analysis is performed on the preprocessed data and reports are generated for decision making.

**Data Source:** <https://www.kaggle.com/datasets/samuelcortinhas/credit-card-classification-clean-data>

**File Format:** CSV

| Data Collection Phase | Role (Responsibilities) | Description | Frequency |
| --- | --- | --- | --- |
| Data Extraction (collection) | Data Engineer, Data Analyst | Data is collected from credible sources through extraction scripts which retrieve data from API and stored in server as data files.  Since in our case the data is static in nature we divide the dataset into three parts assuming each part as a separate timeframe. | Daily: we receive data of credit card transactions daily from front systems. |
| Data Ingestion (storage) | Data Engineer, Data Analyst | The data files from server is read and the data elements are extracted and transformed as per relational database schema of data warehouse and the records are saved into data warehouse. This process is done with the help of ETL tools like Informatica, IBM DataStorage, SAS Data Managment. | Daily : data files needs to be ingested daily into data warehouse. |
| Data Processing (processing) | Data Engineer, Data Analyst | The data is pre processed and processed in this stage where removing unnecessary rows , handling null values, removing duplicate data and removing unnecessary columns are performed on the data to make the data useful for use in analysis. This can be performed by python scripts or ETL tools like Informatica, IBM DataStorage, SAS Data Managment. | Daily: data needs to be cleaned and processed daily. |
| Analysis and Reporting (analysis) | Data Analyst, Business Analyst, Machine learning engineer, Data Scientist | The data is used to derive meaningful insights using visulaization tools like PowerBI, Excel etc. Insights could be in the form of plots or report which contain derived measured values which are interpreted to make business decisions. | Weekly, monthly or yearly: reports of credit card transaction are generated on weekly, monthly and yearly basis for identifying anomalies and making business decisions. |