Concordia University
Engineering and Computer Science

SOEN 6611 – Software Measurement

PROJECT STEP 3

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

Objective: Operationalize Goals, Derive Success Criteria and Indicators, derived measures, and base measures

Submitted to - Prof. Dr. Olga Ormandjieva

Team – 7

Rutwikkumar Sunilkumar Patel – (40160646)
Charit Pareshbhai Patel – (40160658)
Deep Pareshkumar Patel – (40185585)
Bhoomi Shah – (40169655)

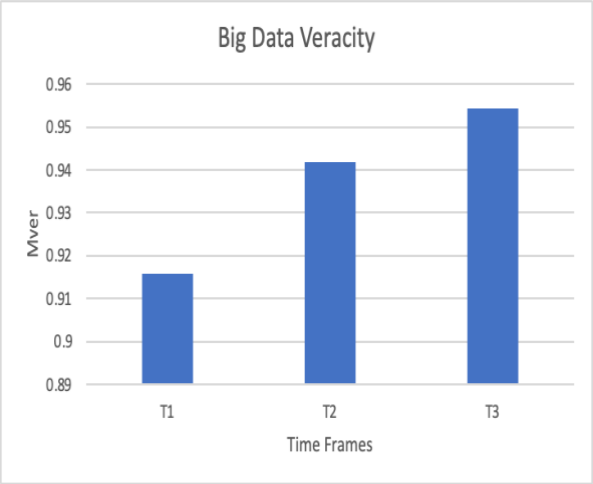
Table of Contents		
Index	**Content**	**Page**

Part 1 (6 points): Derive Success Criteria and Indicators (for Validity, Vincularity, and Veracity)

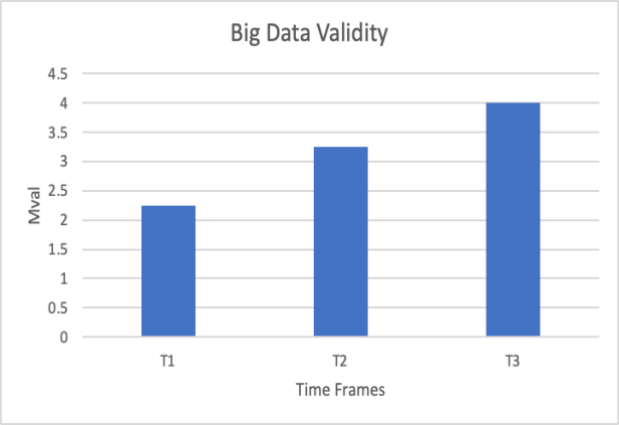
The objective of Part 1 is to develop success criteria and success indicators.
Success (answering the measurement question within the desired timeframe) can only be achieved when certain conditions are in place. indicators will allow you to answer the questions quantitatively and then communicate the results to others.

For each measurement question related to Validity, Vincularity, and Veracity, develop success criteria that will allow you to answer the measurement questions quantitatively.

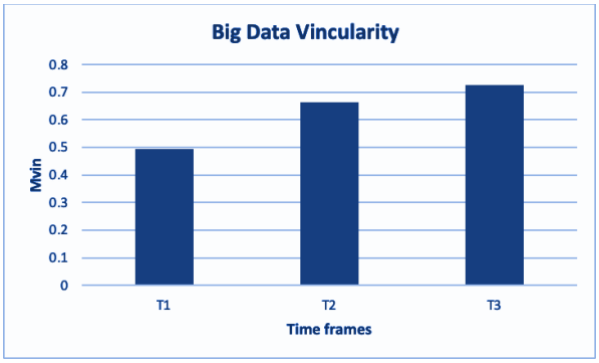
Measurement Question Label / Operationalized Goal Label	**MG3 - Veracity** Measurement Question Label (Q3) ● What are the possible states that the algorithm can work with? ● In what state the data is? Operationalized Goal Label Improve the state of the data in the dataset and use the necessary and relevant data with context to specific system needs.
Success Criteria Label and description	**The success criteria Label: SCver** The trustworthiness of the data source, accuracy, type, and processing must be higher in time frame T2 than calculated in time frame T1, where T2 > T1.
Indicator Label and description	**The indicator label <I1>: Mver** Mver describes how exact, reliable, and accurate data is. Not only is the data itself accurate, but also the reliability of the data's source, kind, and processing.
Indicator Analysis Model and Interpretation	**Indicator Analysis:** Veracity levels that are lower are thought to be weaker than ones that are higher. Given that all weights added together equal 1, and all sub-values of Veracity fall between 0 and 1.0, the optimal value for Veracity is 1.0.

	Interpretation: We use the above-mentioned algorithm with weights specified by the data practitioner. All weights are set to 1/4 by default. If, for example, the data practitioner wants to prioritize Accuracy, the weight might be increased, allowing them to better notice improvements in that specific indicator.
Indicator Sketch	

Measurement Question Label / Operationalized Goal Label	**MG5 - Validity** <u>Measurement Question Label (Q5)</u> ● What is the value of accuracy and correctness of the data during different time frames? <u>Operationalized Goal Label</u> Improve the quality of the data used for different purposes in the application to be authorized and should be of high level.
Success Criteria Label and description	**The success criteria Label: SCval** The credibility and compliance should be increased from time frame T1 to time frame T2, where T2 > T1.
Indicator Label and description	**The indicator label <I2>: Mval** Mval of big data is defined in terms of accuracy and correctness for the purpose of usage.
Indicator Analysis Model and Interpretation	**Indicator Analysis:** Lower Validity levels are perceived as being weaker than higher ones. Given that all weights added together equal 1, and all sub-

	values of Validity fall between 0 and 1.0, the best value for Veracity that is conceivable is 1.0. **Interpretation:** All weights are set to 1/4 by default. For instance, the weight may be raised to make Compliance more significant, enabling the data practitioner to notice changes in that particular indicator more clearly.
Indicator Sketch	

Measurement Question Label / Operationalized Goal Label	**MG6 - Vincularity** <u>Measurement Question Label - Q6</u> • What is the traceability value for different data records in different datasets at different time intervals? <u>Operationalized Goal Label</u> Improve the connectivity, traceability, and linkage between big data.
Success Criteria Label and description	**The success criteria Label: SCvin** The amount of data traceable in time frame T2 should be more than the data traceable in time frame T1, where T2 > T1.
Indicator Label and description	**The indicator label <I3>: Mvin** Mvar pertains to how the data is connected and linked in the dataset.

Indicator Analysis Model and Interpretation	**Indicator Analysis:** Vincularity indicator, which spans from 0-100 and represents the proportion of data that can be traced across all datasets, shows that all records are traceable across MDS. **Interpretation:** We calculate the degree of vinculality using the method above and use the thresholds established by the data practitioner to show the extent to which our data satisfies traceability standards.
Indicator Sketch	

Step 3 - Part 2 - The objective of Part 2 is to define all measures required to derive your V's indicators (for Validity, Vincularity, and Veracity) and decide on the achievement of the corresponding operationalized goals.

3.2.1 Identification of the V's measures (for Validity, Vincularity, and Veracity), tracing them to the corresponding indicators, their availability, and source

For each of the V's indicators (for Validity, Vincularity, and Veracity), identify all required measures (derived and base). The table below will be used to complete each of these measures in sections 3.2 and 3.3. It is also recommended that you review and complete this table after all measures have been defined.
This table, therefore, gives a good summary of all the measurements to be collected and analyzed.

Base Measures

1. **Length of Big Data (LBD)** - Total number of records in MDS(across multiple datasets)
2. **Rec_no_null (MDS**) - Frequency of records in MDS (Multiple Datasets) with no null values.
3. **Rec_cc_age (MDS)** - Provides the total number of records with ages that fall within the acceptable range based on the upper and lower quartiles of the Box and Whisker.
4. **N_succ_req (MDS)** - Number of successful requests (from an API, server, datastore, origins of data, etc).
5. **N_req (MDS)** - Number of requests
6. **Number of Distinct Data Elements (Ndde)** - Across multiple data sets
7. **Nrec_comp** - Number of compliant records in a Dataset
8. **Nds_cr** - Number of credible datasets
9. **Nds** - Number of datasets
10. **Length of the record in dataset (Ldst) -** Total number of occurrences of data elements in dataset
11. **Rec_Trace** - Provides the total number of records that are traceable in MDS

Derived Measures

1. **Accuracy (MDS)**

$$H_{acc}(MDS) = \log_2(Lbd) - (1 / Lbd) \times \sum_{j=\{1\ldots k\}} p_j \log_2 (p_j)$$

$$H_{max}\ (MDS) = \log_2(Lbd)$$

Where,

H_{acc} = Entropy of multiple datasets

H_{max} = Max entropy

$$Accuracy\ (MDS) = \frac{H_{acc}}{H\_\max}$$

2. **Completeness (MDS)**

$$Com_m\ (MDS) = \frac{[rec\_no\_null\ (MDS)]}{Lbd(MDS)}$$

3. **Currentness (MDS)**

$$Currentness\ (MDS) = \frac{[rec\_acc\_age\ (MDS)]}{Lbd(MDS)}$$

4. **Availability (MDS)**

$$Availability\ (MDS) = \frac{[n\_succ\_req\ (MDS)]}{n\_req(MDS)}$$

5. **Big Data Veracity (Mver) (MDS)**

$$Mver\ (MDS) = Accuracy\ (MDS) * W_{Acc} + Completness(MDS) * W_{Comp} + Currentness\ (MDS) * W_{Curr} + Availability * W_{Avail}$$

Where

W_{acc} : Weight of Ndde (Set to 1/4 by default)
W_{comp} : Weight of Lbd (Set to 1/4 by default)
W_{curr} : Weight of Nds (Set to 1/4 by default)
W_{avail} : Weight of Nds (Set to 1/4 by default)
Sum of all weights is equal to 1

6. **Compliance (MDS)**

$$Compliance\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Nrec_{comp}(DS)}{Nds(MDS)}$$

7. **Credibility (MDS)**

$$Credability\ (MDS) = \frac{Nds_{cr}(MDS)}{Nds\ (MDS)}$$

8. **Big Data Validity (Mval) (MDS)**

$$Mval\ (MDS) = Credability\ (MDS) * W_{Cred} + Compliance(MDS) * W_{Compli}$$

Where

W_{Cred} : Weight of Credibility (Set to 1/2 by default)
W_{Compli} : Weight of Compliance (Set to 1/2 by default)
Sum of all weights is equal to 1

9. **Traceability of dataset (DS)**

$$Traceability\ (DS) = \frac{Rec_{Trace}(DS)}{Ldst\ (DS)}$$

10. Big Data Vincularity (Mvin)

$$Mvin\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Traceability\ (DS)}{Nds\ (MDS)}$$

Summary of all the measurements to be collected and analyzed

Measures					Indicator(s) label		
#	**Identification (name of the measure)**	**Type**	**Availability**	**Source**	**<I1> <Mver>**	**<I2> <Mval>**	**<I3> <Mvin>**
1	Length of Big Data (LBD)	Base	A	Dataset	X		
2	Rec_no_null (MDS)	Base	C	Dataset	X		
3	Rec_cc_age (MDS)	Base	C	Dataset	X		
4	N_succ_req (MDS)	Base	A	Dataset	X		
5	N_req (MDS)	Base	A	Dataset	X		
6	Number of distinct elements (Ndde)	Base	C	Dataset	X		
7	Nrec_comp	Base	C	Dataset		X	
8	Nds_cr	Base	C	Dataset		X	
9	Nds	Base	A	Dataset		X	X
10	Length_of_the_record_in_dataser (Ldst)	Base	C	Dataset			X

11	Rec_Trace	Base	C	Dataset			X
12	Accuracy	Derived	C	Dataset	X		
13	Completeness	Derived	C	Dataset	X		
14	Currentness	Derived	C	Dataset	X		
15	Availability	Derived	C	Dataset	X		
16	Compliance	Derived	C	Dataset		X	
17	Credibility	Derived	C	Dataset		X	
18	Traceability	Derived	C	Dataset			X
19	Big Data Veracity (Mver)	Derived	B	Dataset	X		
20	Big Data Validity (Mval)	Derived	B	Dataset		X	
21	Big Data Vincularity (Mvin)	Derived	B	Dataset			X

***Type**: "Derived" or "Base".*

Availability:

"A": Already available and collected;

"B": Can be derived from other data fairly directly;

"C": Possibly obtained with minor effort;

"D": Not available at the moment;

"E": Very difficult, if not impossible to obtain at the moment.

***Source**: Place or tool where data is collected. In the case of base measures, this is obvious; in the case of derived measures, it depends on where the base data is stored after collection.*

***Indicator (s):** Mark an "X" when this measurement is required for each of your indicators.*

3.2.2 3V's Derived measures: definitions and operationalization

<table>
<tr><td colspan="4">Derived measure or indicator: Big data veracity (Mver)</td></tr>
<tr><td>#1</td><td>Derived Measure or indicator
Mver</td><td colspan="2">Formula
$$Mver\ (MDS) = Accuracy\ (MDS) * W_{Acc} + Completness(MDS) * W_{Comp} + Currentness\ (MDS) * W_{Curr} + Availability * W_{Avail}$$
Where
W_{acc} : Weight of Ndde (Set to 1/4 by default)
W_{comp} : Weight of Lbd (Set to 1/4 by default)
W_{curr} : Weight of Nds (Set to 1/4 by default)
W_{avail} : Weight of Nds (Set to 1/4 by default)
Sum of all weights is equal to 1</td></tr>
<tr><td colspan="2">Link with the measurement goal (which goal)
MG3: Comparing and analyzing the rate at which the veracity changes at different time frames in big data pipeline</td><td>Responsible (Who Analyzes)
Data Scientist</td><td>Stakeholder (Who Uses)
Technical Team (Product Manager, Developer, etc.)</td><td>Frequency (When)
In every phase of data extraction, data preprocessing, and data processing for each time frame.

Also, every time when new data is added to the big data pipeline.</td></tr>
<tr><td colspan="2">Data source (where the measurement data will</td><td>Storage of the result</td><td colspan="2">Data interpretation rules
Veracity comes up in the amount of</td></tr>
</table>

be extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	(where data will be stored after the extraction) **Internal disk or any external storage systems or devices**	**Accuracy, Completeness, Currentness, and Availability calculated from our dataset. We can assign the weights according to our preference and can calculate veracity using the above formula. To analyze the veracity, check accuracy, completeness, currentness, and availability either increasing or decreasing with different time frames.** **Veracity values that are lower are thought to be weaker than ones that are higher. Given that all weights added together equal 1, and all sub-values of Veracity fall between 0 and 1, the optimal value for Veracity is 1.**

Analysis procedure

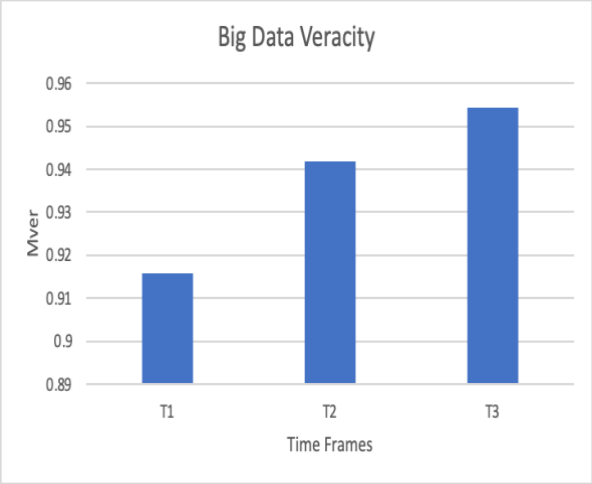
Plot a bar graph and compare the values of veracity, calculated using the formula given above, for different phases of different time frames. Veracity increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

We apply the above formula mentioned above with the default weight of ¼.

As our goal is to increase the value of veracity over different time frames and as veracity is proportional to accuracy, completeness, currentness, and availability, all these values need to be increased.

To calculate all the required terms of formula, i.e. accuracy, completeness, currentness, and availability, we will use the below formulas,

Presentation of the results (sketch illustrating what it looks like):



$$H_{acc}\ (MDS) = \log_2(Lbd) - \frac{1}{Lbd * \sum_{j=[1\ldots k]} p_j \log_2(p_j)}$$

$$H_{max}\ (MDS) = \log_2(Lbd)$$

$$Com_m\ (MDS) = \frac{[rec\_no\_null\ (MDS)]}{Lbd(MDS)}$$

$$Currentness\ (MDS) = \frac{[rec\_acc\_age\ (MDS)]}{Lbd(MDS)}$$

$$Availability\ (MDS) = \frac{[n\_succ\_req\ (MDS)]}{n\_req(MDS)}$$

<u>At time frame T1</u>,

LBD = 4096
P_j = Total number of duplicate items = 128
Rec_no_null = 3864
Rec_acc_age = 3524
N_succ_req = 3418
N_req = 3930

So,

H_{acc} = 12 - 0.1976 = 11.8024

H_{max} = 12

Therefore, Accuracy = 0.9835
Completeness = 0.9433
Currentness = 0.8604
Availability = 0.8764

Mver = ¼ * (0.9835 + 0.9433 + 0.8604 + 0.8764)
Mver = 0.9159

<u>At time frame T2</u>,

LBD = 8192 **P_j = Total number of duplicate items = 256** **Rec_no_null = 7868** **Rec_acc_age = 7397** **N_succ_req = 7218** **N_req = 7800** **So,** **H_{acc} = 13 - 0.2006 = 12.7994** **H_{max} = 13** **Therefore, Accuracy = 0.9845** **Completeness = 0.9550** **Currentness = 0.9029** **Availability = 0.9253** **Mver = ¼ * (0.9845 + 0.9550 + 0.9029 + 0.9253)** **Mver = 0.9419** **<u>At time frame T3</u>,** **LBD = 12000** **P_j = Total number of duplicate items = 314** **Rec_no_null = 11256** **Rec_acc_age = 11189** **N_succ_req = 10234** **N_req = 11456** **Therefore, Accuracy = 0.9898** **Completeness = 0.9678** **Currentness = 0.9143** **Availability = 0.9456** **Mver = ¼ * (0.9898 + 0.9678 + 0.9143 + 0.9456)**	

Mver = 0.9543	
Potential decision making depending on the results **Bar chart helps us to determine the veracity of big data at different time frames. Veracity increases as new data is added to the big data pipeline in consecutive time frames. This helps us to improve the quality of big data as veracity improves over time.**	

Derived measure or indicator: **Big Data Validity (Mval)**				
#**2**	Derived Measure or indicator **Mval**	Formula $Mval\,(MDS) = Credability\,(MDS) * W_{Cred} + Compliance(MDS) * W_{Compli}$ $Compliance\,(MDS) = \frac{\sum_{\forall\, DS \in MDS} Nrec_{comp}(DS)}{Nds(MDS)}$ $Credability\,(MDS) = \frac{Nds_{cr}(MDS)}{Nds\,(MDS)}$ **Where,** **W_{Cred} : Weight of Credibility (Set to 1/2 by default)** **W_{Compli} : Weight of Compliance (Set to 1/2 by default)** **Sum of all weights is equal to 1**			
Link with the measurement goal (which goal) **MG5: Increasing the accuracy and correctness for the purpose of usage.**		Responsible (Who Analyzes) **Data Scientist**	Stakeholder (Who Uses) **Technical Team (Product Manager, Developer, etc.)**	Frequency (When) **In each time frame, of big data pipeline.** **Also, when the new data is added in the dataset.**
Data source (where the measurement data will be		Storage of the result	Data interpretation rules	

extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	(where data will be stored after the extraction) **Internal disk or any external storage systems or devices**	**Validity comes up in the amount of compliance, and credibility calculated from our dataset. We can assign the weights according to our preference and can calculate validity using the above formula. If the validity value in the data pipeline rises with time, the data is reliable and compliant for its intended use. Lower Validity levels are perceived as being weaker than higher ones. Given that all weights added together equal 1, and all sub-values of Validity fall between 0 and 1.0, we know that 1.0 is the ideal value for Validity.**

Analysis procedure

Plot a bar graph and compare the values of validity, calculated using the formula given above, for different phases of different time frames. Validity increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

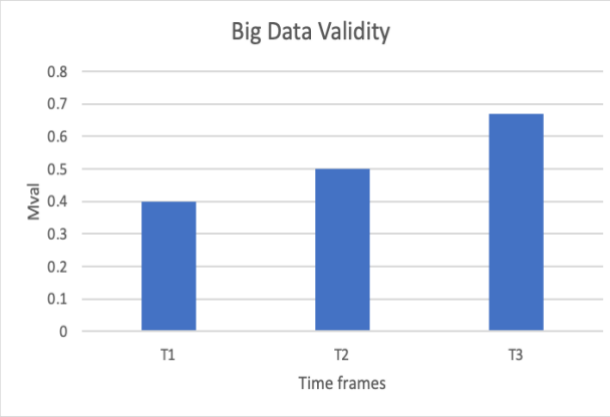
We apply the above formula mentioned above with the default weights of ½.
To calculate all the required terms of formula, i.e. compliance and credibility, we will use below formulas,

$$Compliance\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Nrec_{comp}(DS)}{Nds(MDS)}$$

$$Credability\ (MDS) = \frac{Nds_{cr}(MDS)}{Nds\ (MDS)}$$

<u>At time frame T1</u>,

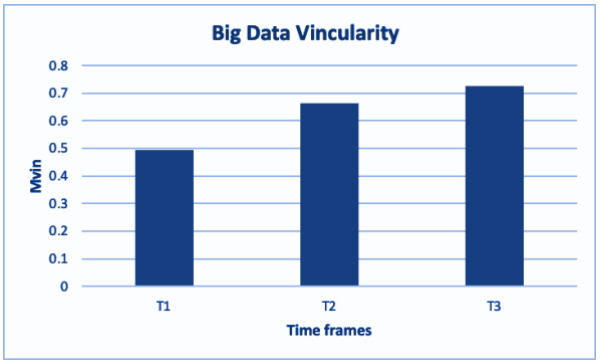
Presentation of the results (sketch illustrating what it looks like):



Nrec_comp(DS1) = 5 **Nrec_comp(DS2) = 4** **Nds_cr = 2** **Nds = 2** **Therefore,** **Compliance = (5 + 4) / 2 = 4.5** **Credibility = 2 / 2 = 1** **Mval = ½(4.5 + 1) = 2.25** **<u>At time frame T2</u>,** **Nrec_comp(DS1) = 6** **Nrec_comp(DS2) = 5** **Nds_cr = 2** **Nds = 2** **Therefore,** **Compliance = (6 + 5) / 2 = 5.5** **Credibility = 2 / 2 = 1** **Mval = ½(5.5 + 1) = 3.25** **<u>At time frame T3</u>,** **Nrec_comp(DS1) = 8** **Nrec_comp(DS2) = 6** **Nds_cr = 2** **Nds = 2** **Therefore,** **Compliance = (8 + 6) / 2 = 7** **Credibility = 2 / 2 = 1** **Mval = (7 + 1) /2 = 4**	

Potential decision making depending on the results **Bar chart helps us to determine the validity of big data at different time frames. Validity increases as new data are added to the big data pipeline in consecutive time frames. This helps us improve big data quality as validity improves over time.**	

Derived measure or indicator: **Big Data Vincularity (Mvin)**				
#**3**	Derived Measure or indicator **Mvin**	Formula $$Mvin\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Traceability\ (DS)}{Nds\ (MDS)}$$ Where, Nds = Number of datasets.		
Link with the measurement goal (which goal) **MG6: To increase traceability by increasing connectivity and linkages of data.**		Responsible (Who Analyzes) **Data Analyst/Data Scientist**	Stakeholder (Who Uses) **Technical Team (Product Manager, Developer, etc.)**	Frequency (When) **In each time frame, of big data pipeline.** **Also, when the new data is added in the dataset.**
Data source (where the measurement data will be extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**		Storage of the result (where data will be stored after the extraction) **Internal disk or any external storage systems or devices**	Data interpretation rules **Vincularity is measure in terms of traceability of the records in the dataset. We can use the above formula to calculate the vincularity of the big data in the dataset. To analyze the vincularity, we have to see the traceability value with respect to a number of datasets, either increasing or decreasing with different time frames.**	

Analysis procedure	Presentation of the results (sketch illustrating what it looks like):
We apply the above-given equation to determine vincularity. **Considering we have two datasets and each dataset has three records.** **For time frame T1,** **Ldst(T1) (DS1, DS2) = 3** **Nds(T1) (MDS) = 2** **rec_trace(T1) (DS1) = 2** **rec_trace(T1) (DS2) = 1** **Trace(T1) (DS1) = ⅔ = 0.66** **Trace(T1) (DS2) = ⅓ = 0.33** **Mvin(T1) (MDS(T1)) = (0.66 + 0.33)/2 = 0.495** **For time frame T2,** **Ldst(T2) (DS1, DS2) = 6** **Nds(T2) (MDS) = 2** **rec_trace(T2) (DS1) = 5** **rec_trace(T2) (DS2) = 3** **Trace(T2) (DS1) = 5/6 = 0.83** **Trace(T2) (DS2) = 3/6 = 0.5** **Mvin(T2) (MDS(T2)) = (0.83 + 0.50)/2 = 0.665** **For time frame T3,** **Ldst(T3) (DS1, DS2) = 9** **Nds(T3) (MDS) = 2** **rec_trace(T3) (DS1) = 7** **rec_trace(T3) (DS2) = 6** **Trace(T3) (DS1) = 7/9 = 0.78** **Trace(T30) (DS2) = 6/9 = 0.67**	

Mvin(T2) (MDS(T2)) = (0.78 + 0.67)/2 = 0.725	
Potential decision making depending on the results **It can be determined how the vincularity of big data is changed over different time periods. By meticulously observing we can analyze the traceability of the record in every time interval.**	

Derived measure or indicator: **Accuracy**				
# **4**	Derived Measure or indicator **Accuracy**	Formula $H_{acc}(MDS) = \log_2(Lbd) - \frac{1}{Lbd * \sum_{j=[1\ldots k]} p_j \log_2(p_j)}$ $H_{max}(MDS) = \log_2(Lbd)$ Where, H_{acc} = Entropy of multiple datasets H_{max} = Max entropy $Accuracy(MDS) = \frac{H_{acc}}{H\_\max}$		
Link with the measurement goal (which goal) **MG3: Comparing and analyzing the rate at which the veracity changes at different time frames in big data pipeline**		Responsible (Who Analyzes) **Data Analyst / Data Scientist**	Stakeholder (Who Uses) **Technical Team (Product Manager, Developer, etc.)**	Frequency (When) **In each time frame or in every phase, of big data pipeline.**
Data source (where the measurement data will be extracted from) **IBM Analytics Dataset** https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset		Storage of the result (where data will be stored after the extraction)	Data interpretation rules **Accuracy refers to the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.**	

	Internal disk or any external storage systems or devices	**(ISO/IEC 25012, 2008.)** **As accuracy is used to calculate veracity and veracity lies between 0 and 1, the accuracy also lies between 0 and 1. Value 0 indicates less accuracy and value 1 indicates high accuracy.**

Analysis procedure

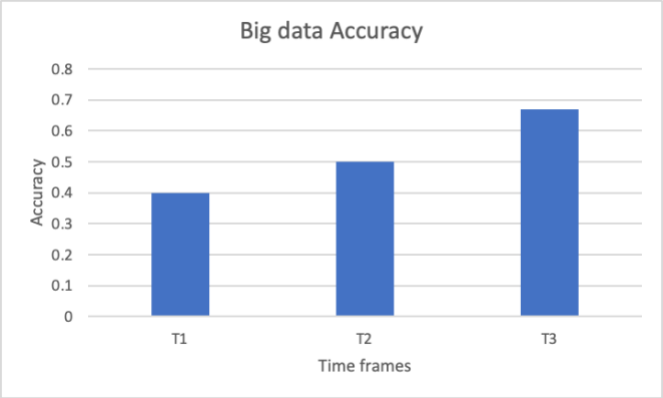
Plot a bar graph and compare the values of accuracy, calculated using the formula given above, for different phases of different time frames. Accuracy increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

We apply the formula mentioned above.

Dataset 1 : [1,2,2,3,3,4,5]
Dataset 2 : [1,2,6,7,4,8]
Pj_D1 = {'2':2, '3':2}, k_D1 = 5
Pj_D2 = {}, k_D2 = 6
Lbd = 13

1log2(1) = 0
2log2(2) + 2log2(2) = 4
H_acc = log2(13) – 4/13 = 3.39
H_max = log2(13) = 3.7
Acc = 3.39 / 3.7 = 0.916

Presentation of the results (sketch illustrating what it looks like):

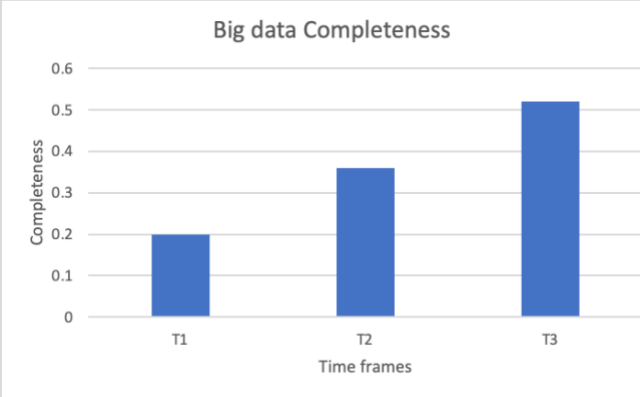


Potential decision making depending on the results

Bar chart helps us to determine the accuracy of big data at different time frames. Accuracy increases as new data are added to the big data pipeline in consecutive time

frames. This helps us improve big data veracity as accuracy improves over time.	

Derived measure or indicator: **Completeness**			
# **5**	Derived Measure or indicator **Completeness**	Formula $Com_m\ (MDS) = \frac{[rec\_no\_null\ (MDS)]}{Lbd(MDS)}$	
Link with the measurement goal (which goal) **MG3: Comparing and analyzing the rate at which the veracity changes at different time frames in big data pipeline**	Responsible (Who Analyzes) **Data Analyst / Data Scientist**	Stakeholder (Who Uses) **Technical Team (Product Manager, Developer, etc.)**	Frequency (When) **In each time frame or in every phase, of the big data pipeline.**
Data source (where the measurement data will be extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Storage of the result (where data will be stored after the extraction) **Internal disk or any external storage systems or devices**	Data interpretation rules **Completeness refers to degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use. (ISO/IEC 25012, 2008.)** **As completeness is used to calculate veracity and veracity lies between 0 and 1, the completeness also lies between 0 and 1. Value 0 indicates less completeness and value 1 indicates high completeness.**	
Analysis procedure **Plot a bar graph and compare the values of completeness, calculated using the formula given above, for different phases of**	Presentation of the results (sketch illustrating what it looks like):		

different time frames. Completeness increases or decreases based on the data operation performed on big data and any change in the structure of data sets. **We apply the formula mentioned above.** **Dataset 1: [1, null, 2,3,4]** **Rec_no_null = 1** **Lbd = 5** **Com = 1/5**	
Potential decision making depending on the results **Bar chart helps us to determine the completeness of big data at different time frames. Completeness increases as new data are added to the big data pipeline in consecutive time frames. This helps us improve big data veracity as completeness improves over time.**	

Derived measure or indicator: **Currentness**			
# **6**	Derived Measure or indicator **Currentness**	Formula $$Currentness\ (MDS) = \frac{[rec\_acc\_age\ (MDS)]}{Lbd(MDS)}$$	
Link with the measurement goal (which goal) **MG3: Comparing and analyzing the rate at which the veracity changes at different time frames in big data pipeline**	Responsible (Who Analyzes) **Data Analyst / Data Scientist**	Stakeholder (Who Uses) **Technical Team (Product Manager, Developer, etc.)**	Frequency (When) **In each time frame or in every phase, of the big data pipeline.**
Data source (where the	Storage of	Data interpretation	

measurement data will be extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	the result (where data will be stored after the extraction) **Internal disk or any external storage systems or devices**	rules **Currentness refers to degree to which data has attributes that are of the right age in a specific context of use. (ISO/IEC 25012, 2008.)** **As currentness is used to calculate veracity and veracity lies between 0 and 1, the currentness also lies between 0 and 1. Value 0 indicates less currentness and value 1 indicates high currentness.**

Analysis procedure

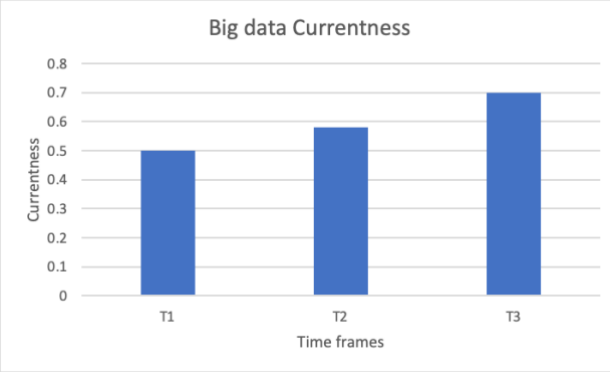
Plot a bar graph and compare the values of currentness, calculated using the formula given above, for different phases of different time frames. Currentness increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

We apply the formula mentioned above.

Order Values from least to greatest:

- **5, 19, 19, 19, 19, 21, 21, 21, 22, 22, 22, 25, 25, 26, 27, 33, 33, 33, 33, 35**
- **Median: 22**
- **Lower and Upper Quartile (fourth's) : 19, 27**
- **Box Length: 10**
- **Upper and Lower Tail Range:**
- **[Box_Length*1.5 - Lower Quartile, Box_length + Upper_Quartile]**
- **= [10*1.5 - 19, 10*1.5+27]**
- **= [15-19,15+27] = [-4, 42]**
- **Lower Tail: 5**

Presentation of the results (sketch illustrating what it looks like):



• **Upper Tail: 35** • **Rec_acc_age (MDS) = 10 (Range between lower and upper quartile)** • **Lbd (MDS) = 20 UIDR** • **Curr (MDS) = rec_acc_age (MDS) / Lbd (MDS) = 10/20 =0.5**	
Potential decision making depending on the results **Bar chart helps us to determine the currentness of big data at different time frames. Currentness increases as new data are added to the big data pipeline in consecutive time frames. This helps us improve big data veracity as currentness improves over time.**	

Derived measure or indicator: **Availability**			
#**7**	Derived Measure or indicator **Availability**	Formula $$Availability\ (MDS) = \frac{[n\_succ\_req\ (MDS)]}{n\_req(MDS)}$$	
Link with the measurement goal (which goal) **MG3: Comparing and analyzing the rate at which the veracity changes at different time frames in big data pipeline**	Responsible (Who Analyzes) **Data Analyst / Data Scientist**	Stakeholder (Who Uses) **Technical Team (Product Manager, Developer, etc.)**	Frequency (When) **In each time frame or in every phase, of the big data pipeline.**
Data source (where the measurement data will be extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Storage of the result (where data will be stored after the extraction) **Internal disk or**	Data interpretation rules **Availability refers to the degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of**	

	any external storage systems or devices	**use. (ISO/IEC 25012, 2008.)** **As availability is used to calculate veracity and veracity lie between 0 and 1, the availability also lies between 0 and 1. Value 0 indicates less currentness and value 1 indicates high availability.**

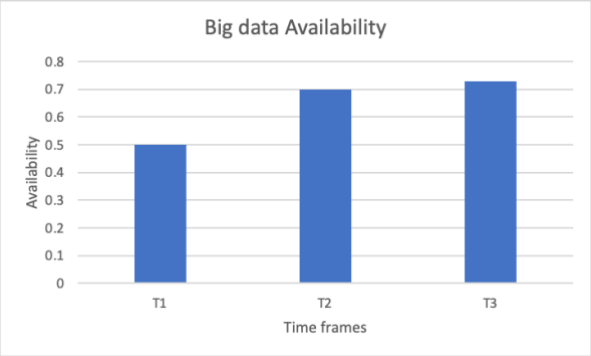
Analysis procedure

Plot a bar graph and compare the values of availability, calculated using the formula given above, for different phases of different time frames. Availability increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

We apply the formula mentioned above.

Dataset 1: The first request fails, the second request succeeds
N_succ_req = 1
N_req = 2
Availability = 1/2 = 50%

Presentation of the results (sketch illustrating what it looks like):



Potential decision making depending on the results

Bar chart helps us to determine the availability of big data at different time frames. Availability increases as new data are added to the big data pipeline in consecutive time frames. This helps us improve big data veracity as availability improves over time.

Derived measure or indicator: **Compliance**

<table>
<tr><td>#8</td><td>Derived Measure or indicator
Compliance</td><td colspan="3">Formula
$$Compliance\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Nrec_{comp}(DS)}{Nds(MDS)}$$</td></tr>
<tr><td colspan="2">Link with the measurement goal (which goal)
MG5: Increasing the accuracy and correctness for the purpose of usage.</td><td>Responsible (Who Analyzes)
Data Analyst/Data Scientist</td><td>Stakeholder (Who Uses)
Technical Team (Product Manager, Developer, etc.)</td><td>Frequency (When)
In each time frame, of the big data pipeline.

Also, when the new data is added to the dataset.</td></tr>
<tr><td colspan="2">Data source (where the measurement data will be extracted from)
IBM Analytics Dataset

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset</td><td>Storage of the result (where data will be stored after the extraction)
Internal disk or any external storage systems or devices</td><td colspan="2">Data interpretation rules
degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use. (ISO/IEC 25012, 2008.)</td></tr>
<tr><td colspan="3">Analysis procedure
Plot a bar graph and compare the values of compliance, calculated using the formula given above, for different phases of different time frames. Compliance increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

We apply the formula mentioned above.

<u>At time frame T1</u>,
Nds =3
rec_compT1 (DS1) = 2, rec_compT1 (DS2) = 1,</td><td colspan="2">Presentation of the results (sketch illustrating what it looks like):
 
</td></tr>
</table>

rec_compT1 (DS3) = 3 **DS_compT1 (DS1) = 0.66,** **DS_compT1 (DS2) =0.33, DS_compT1 (DS3) = 1.00** **MDS_compT1 (MDS) = 0.66** **<u>At time frame T2</u>,** **Nds =3** **rec_compT2 (DS1) = 4, rec_compT2 (DS2) = 3,** **rec_compT2 (DS3) = 6** **DS_compT2 (DS1) = 0.66,** **DS_compT2 (DS2) = 0.50,** **DS_compT2(DS3) = 1.00** **MDS_compT2 (MDS) = 0.72** **As expected,MDS_CompT2 (MDS) > MDS_CompT1 (MDS)**	
Potential decision making depending on the results **Bar chart helps us to determine the compliance of big data at different time frames. Compliance increases as new data are added to the big data pipeline in consecutive time frames. This helps us improve big data validity as compliance improves over time.**	

Derived measure or indicator: **Credibility**				
#9	Derived Measure or indicator **Credibility**	Formula $Credability\ (MDS) = \frac{Nds_{cr}(MDS)}{Nds\ (MDS)}$		
Link with the measurement goal (which goal) **MG5: Increasing the accuracy and**		Responsible (Who Analyzes) **Data**	Stakeholder (Who Uses) **Technical Team (Product**	Frequency (When) **In each time frame, of the big**

correctness for the purpose of usage.	**Analyst/Data Scientist**	**Manager, Developer, etc.)**	**data pipeline.** **Also, when the new data is added to the dataset.**
Data source (where the measurement data will be extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Storage of the result (where data will be stored after the extraction) **Internal disk or any external storage systems or devices**	Data interpretation rules **Degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, and commitments). (ISO/IEC 25012, 2008.)**	

Analysis procedure

Plot a bar graph and compare the values of credibility, calculated using the formula given above, for different phases of different time frames. Credibility increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

We apply the formula mentioned above.

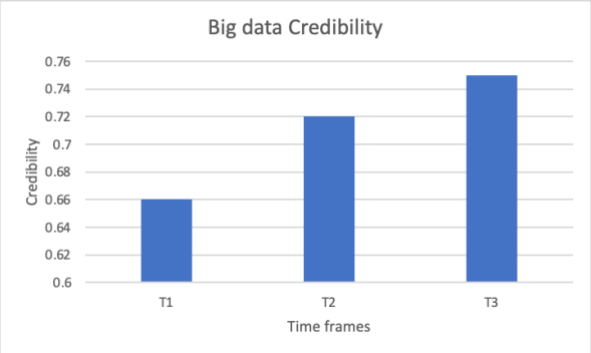
Nds = 3.

- **Assume that cre_source (DS1) = cre_source (DS3) = 1 and cre_source (DS2) = 0.**
- **Assume that the credibility of the DS at times T1 and T2 remain the same.**

- **=> Cre (MDS) =⅔(or 66%) - proportion of credible DS).**
- **New assumption: cre_source (DS'2) = 1 at time T2,**
- **We expect the credibility of the MDST2 to increase:**

o As expected, cre (MDST2') = 1 > ⅔

Presentation of the results (sketch illustrating what it looks like):



Potential decision making depending on the results **Bar chart helps us to determine the credibility of big data at different time frames. Credibility increases as new data are added to the big data pipeline in consecutive time frames. This helps us improve big data validity as credibility improves over time.**	

Derived measure or indicator: **Traceability**				
# **10**	Derived Measure or indicator **Traceability**	Formula $$Traceability\ (DS) = \frac{Rec_{Trace}(DS)}{Ldst\ (DS)}$$		
Link with the measurement goal (which goal) **MG6 - To increase traceability by increasing connectivity and linkages of data.**		Responsible (Who Analyzes) **Data Analyst/Data Scientist**	Stakeholder (Who Uses) **Technical Team (Product Manager, Developer, etc.)**	Frequency (When) **In each time frame, of the big data pipeline.** **Also, when the new data is added to the dataset.**
Data source (where the measurement data will be extracted from) **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**		Storage of the result (where data will be stored after the extraction) **Internal disk or any external storage systems or devices**	Data interpretation rules **degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use. (ISO/IEC 25012, 2008.)**	
Analysis procedure **Plot a bar graph and compare the values of traceability, calculated using the formula given above, for different phases of different**		Presentation of the results (sketch illustrating what it looks like):		

time frames. Traceability increases or decreases based on the data operation performed on big data and any change in the structure of data sets.

We apply the formula mentioned above.

<u>At time frame T1</u>,
Let assume dataset with below information:
LdstT1(DS1,DS2,DS3) = 3
NdsT1 (MDS) = 3
rec_traceT1 (DS1) = 2,
rec_traceT1 (DS2) = 1,
rec_traceT1 (DS3) = 3

Then traceability would be:-
TraceT1 (DS1) = 2/3 = 0.66.
TraceT1 (DS2) = 1/3 = 0.33
TraceT1 (DS3) = 3/3 = 1.00.

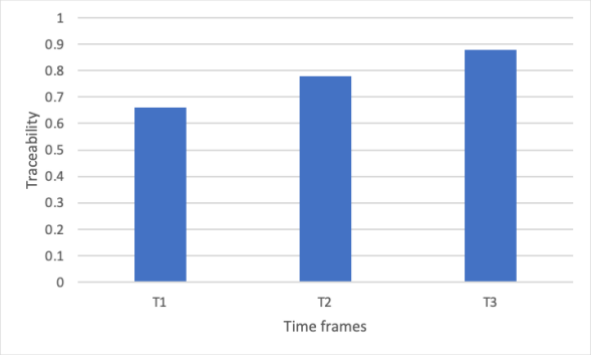
MvinT1 (MDST1) = (0.66+0.33+1)/3 = 0.66

<u>At time frame T2</u>,
LdstT2(DS1,DS2,DS3) = 6
NdsT2 (MDS) = 3
Rec_traceT2 (DS1) = 5
Rec_traceT2 (DS2) = 3
Rec_traceT2 (DS3) = 6

TraceT2 (DS1_T2) = 5/6 = 0.83
TraceT2 (DS2_T2) = 3/6 = 0.50
TraceT2 (DS3_T2) = 6/6 = 1.00.

Mvin (MDST2) = (0.83+0.5+1)/3= 0.78.

As expected we have,
Mvin (MDST2) > Mvin (MDST1)



Potential decision making

depending on the results **Bar chart helps us to determine the traceability of big data at different time frames. Traceability increases as new data are added to the big data pipeline in consecutive time frames. This helps us improve big data vincularity as traceability improves over time.**	

3.2.3 Validity, Vincularity, and Veracity: Base measures definitions and operationalization

<table>
<tr><td colspan="6">Base measure: LBD</td></tr>
<tr><td>#1</td><td colspan="2">Measure (what: entity, attribute)
LBD: Length of Big data

Entity: Dataset
Attribute: Size of dataset</td><td>Scale Type
Absolute</td><td colspan="2">Applicability
The value of LBD represents the length of the dataset is used to calculate the Veracity of the dataset.</td></tr>
<tr><td>Who measures?
Technical Team (Product Manager, Data Scientist, Developer, etc.)</td><td>Source of Measurement
IBM Analytics Dataset

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset</td><td>Where to store the result?
Internal disk or any external storage systems or devices</td><td>Tool

Jupyter Notebook, Pandas, and NumPy libraries in python</td><td>Time(when to measure)

During each time frame, when the new record is added or deleted, the size (length) of the dataset is calculated.</td></tr>
<tr><td colspan="2">Collection procedure(how to collect the data)
Counting number of data records in the entire dataset.</td><td colspan="3">Notes or comments:
This measure is generally used to calculate the veracity of the big data.</td></tr>
</table>

<table>
<tr><td colspan="4">Base measure: Rec_no_null</td></tr>
<tr><td>#2</td><td>Measure (what: entity, attribute)
Rec_no_null: frequency of records in multiple datasets with no null values.

Entity: Datasets
Attribute: Number of records with no null value</td><td>Scale Type
Absolute</td><td>Applicability
The value is used to calculate the veracity of the big data (Mver)</td></tr>
</table>

Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **While calculating Completess**
Collection procedure(how to collect the data) **Counting number of data with no null values in datasets**		Notes or comments: **This measure is generally used to calculate the veracity of the big data**		

Base measure: **Rec_cc_age**				
#3	Measure (what: entity, attribute) **Rec_cc_age: total number of records with ages that falls within acceptable range based on the upper and lower quartiles of Box and Whisker.** **Entity: Datasets** **Attribute: Number of data that falls within the acceptable range.**	Scale Type **Absolute**	Applicability **The value is used to calculate the veracity of big data (Mver)**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **While calculating currentness**

Collection procedure(how to collect the data) **Counting the total numbers of data that falls within an acceptable range of the right age in a specific use of context**	Notes or comments: **This measure is generally used to calculate the veracity of the big data**

Base measure: **N_succ_req**				
#**4**	Measure (what: entity, attribute) **N_succ_req: Number of successful requests** **Entity: Datasets** **Attribute: Number of successful requests from API, server, datastore, etc.**	Scale Type **Absolute**	Applicability **To keep the count of successful requests made by API calls, servers, etc .**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **To calculate availability by getting the number of successful authorization requests of data through API calls, server, etc.**
Collection procedure(how to collect the data) **Count the number of successful requests for data in a specific context of use.**		Notes or comments: **This base measure is used to calculate the Veracity of the big data.**		

Base measure: **N_req**			
#**5**	Measure (what: entity, attribute)	Scale Type	Applicability **To count the number of**

	N_req: Number of requests to a dataset **Entity: Datasets** **Attribute: Number of all requests to a dataset.**	**Absolute**	**all successful and unsuccessful requests to a dataset.**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **To calculate the availability of the dataset.**
Collection procedure(how to collect the data) **Count a total number of requests from an API, data source, etc. for data in a specific context of use.**		Notes or comments: **This base measure is used to calculate the veracity of big data.**		

Base measure: **Ndde**				
#6	Measure (what: entity, attribute) **Ndde: Number of Distinct Data Elements** **Entity: Dataset** **Attribute: Number of unique data records**	Scale Type **Absolute**	Applicability **The value of Ndde represents the number of unique data records in the dataset and is used to calculate the Veracity of Big data (Mver).**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any**	Tool **Jupyter Notebook, Pandas and NumPy**	Time(when to measure) **During each time frame, when the new**

		external storage systems or devices	**libraries in python**	**record is added or deleted, numbers of unique records are calculated.**
Collection procedure(how to collect the data) **Counting number of distinct data records in the entire dataset.**		Notes or comments: **This measure is generally used to calculate the veracity of the big data.**		

Base measure: **Nrec_comp**				
#**7**	Measure (what: entity, attribute) **Nrec_comp**	Scale Type **Absolute**	Applicability **The value of Nrec_comp is used to measure the compliance of the dataset, which is then used to calculate the validity of big data.**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **While measuring the compliance of the big data.**
Collection procedure(how to collect the data) **Count the number of accurate and complete records in the dataset.**		Notes or comments: **This base measure is used to calculate the Validity of big data.**		

Base measure: **Nds_cr**					
#8	Measure (what: entity, attribute) **Nds_cr: Number of credible datasets** **Entity: Datasets** **Attributes: Number of credible datasets**		Scale Type **Absolute**	Applicability **The count of the credible dataset is used to find the credibility of the dataset, which is used to find the validity of the dataset.**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**		Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **While calculating the credibility of the dataset**
Collection procedure(how to collect the data) **Count the number of valid and credible datasets**			Notes or comments: **This base measure is used to calculate the Validity of big data.**		

Base measure: **NDS**					
#9	Measure (what: entity, attribute) **NDS: Number of Datasets** **Entity: Dataset** **Attribute: Number of Datasets**		Scale Type **Absolute**	Applicability **The value of NDS represents the number of datasets which are present for analysis and is used to calculate the Validity (Mval) and Vincularity (Mvar) of Big data.**	
Who measures? **Technical Team (Product Manager,**	Source of Measurement **IBM Analytics Dataset**		Where to store the	Tool **Jupyter Noteboo**	Time(when to measure)

Data Scientist, Developer, etc.)	**https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	result? **Internal disk or any external storage systems or devices**	**k, Pandas, and NumPy libraries in python**	**During each time frame, the number of datasets are calculated while calculating credibility and compliance.**
Collection procedure(how to collect the data) **Counting the number of datasets available for analysis**		Notes or comments: **This measure is generally used to calculate the Validity (Mval) and Vincularity (Mvar) of Big data**		

Base measure: **Ldst**				
#10	Measure (what: entity, attribute) **Ldst: Length of the record in the dataset (Total number of occurrences of data elements in dataset)** **Entity: Dataset** **Attributes: Length of the record in the dataset**	Scale Type **Absolute**	Applicability **The value of Ldst is used to monitor traceability of the data in the dataset, which ultimately helps in finding the vincularity of dataset.**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **While calculating the traceability of the dataset**

Collection procedure(how to collect the data) **Count the number of occurrences of data elements in the dataset**	Notes or comments: **This base measure is used to calculate the Vincularity of big data.**

Base measure: **Rec_trac**				
#**11**	Measure (what: entity, attribute) **Rec_trac: Total number of traceable records in the dataset** **Entity: Dataset** **Attribute: Traceability of records**	Scale Type **Absolute**	Applicability **The value of Rec_trac is used to monitor traceability of the data in the dataset, which ultimately helps in finding the vincularity of the dataset.**	
Who measures? **Technical Team (Product Manager, Data Scientist, Developer, etc.)**	Source of Measurement **IBM Analytics Dataset** **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**	Where to store the result? **Internal disk or any external storage systems or devices**	Tool **Jupyter Notebook, Pandas, and NumPy libraries in python**	Time(when to measure) **While calculating the traceability of the dataset**
Collection procedure(how to collect the data) **Count the total number of traceable data elements**		Notes or comments: **This base measure is used to calculate the Vincularity of big data.**		

References

1. Ormandjieva, Olga et al. “Measuring the 3V's of Big Data: A Rigorous Approach.” IWSM-Mensura (2020).
2. Lecture 9 and 10 Notes for performing Step 3 of Project.
3. Dave Bharadwaj, “Measurement Framework for Assessing Quality of Big Data (Mega) in Big Data Pipeline”.
4. Dava Bhardwaj, “Big data quality”.