SOEN-6611 SOFTWARE MEASUREMENT

Big Data Quality Measurement

TEAM 5

Team Lead:

Vivekananda Reddy Gottam - 40157109

vivekanandareddy95@gmail.com

Mohammod Suhel Firdus 40163582

Saswati Chowdhury 40184906

Milesh Kotadia 40156971

Introduction

What is Big Data Quality

- Vs controls Quality

Data Set Chosen

- Most popular superhero TV shows
- IMDB rating
- With 8 features

- Name :- Most Popular Superhero TV Shows
- **Source :- Kaggle**
- We had to divide the dataset into 3 subsets to assume we collected data on three distinct times **T1,T2, T3**
- Size of Dataset: 164.41 kB
- Structure of data: 8 columns and 750 rows (text_format - String)
- **No of records: 750, No of Unique records: 740**

Image courtesy : https://www.analyticsvidhya.com/blog/2021/05/what-is-big-data-introduction-uses-and-applications/

Business Goals

Business Goal : Improve Big Data Quality

Subgoals:

- Increasing the **Volume** of big data sets (**BSG-01)**
- Enhancing **Variety** in Big Data (**BSG-02)**
- Accelerate the Big Data set **Velocity** (**BSG-03)**
- Enhancing **Veracity** of Big Data set (**BSG-04)**
- Validate Big Date set **Validity** (**BSG-05)**
- Continuously monitoring Big Data set **Vincularity** (**BSG-06)**

Measurement Stakeholders

- ➔ **Product Owner/Project Manager**
- ➔ **Developers/Data Scientist**
- ➔ Testers/QA
- ➔ Sales and Marketing Team
- ➔ End Users

Image Source: Google Images

Measurement Goals

MG1 - To select a dataset with a sufficient volume of records

MG2 -The goal is to quickly acquire that dataset and process it

MG3 -Useful in improving the validity of the dataset instead of selecting a dataset that has more noise

MG4 -Increasing the trust and authenticity of data by removing the incomplete data

MG5 -Big data sets that have records that are of similar nature and can be used for comparison can be used for analytics.

MG6 -To select the dataset with different types (but same genre) of records helping to categorize and segregate data.

GQM 

Project/Product Manager

- Is the **Volume** of the big data set sufficient enough for developing a model? How many records are available in the data set?
- Is the dataset **frequently changing**? How easy is it to source a newly changed dataset?
- Are we selecting the **right data source** for our business requirements?
  - What is current **customer satisfaction**? Is this Dataset going to impact customer satisfaction?
  - Does the selected dataset have **connected/related data** required for developing a product?
  - Does the dataset contain any incomplete information? What will be the level of **complexity to remove that data**?

Image Source: Google Images

Step : 1 & 2

GQM

Developer

- Do we have pre-processed data in a **recognizable format**? Is there sufficient data for processing? Is it too huge to process and current programming techniques can handle the same?
- Does the algorithm consider the new stream of data? Is it synchronized or not? Do we need to **adjust our code for a changed dataset**?
- To what extent the dataset is valid? Is it **meaningful data** that can be used as input for ML?
- Do we have all the correct data information? Does the collected dataset lead us to get **insightful information**? Which method will be effective for removing outliers?
- In what way can results be evaluated when the processed **data is compared** with the other data?
- Is the dataset organized? How to **categorize** the different forms of data?

Image Source: Google Images

Operationalised Goals

- Select the dataset with **high-quality Volume**.
- **Improve the decision** by frequently gathering data and processing it faster.
- **Improving the quality** by checking the correctness of big dataset.
- Improving the **veracity** by analysing and processing only that data which is specific to need and removing unwanted data.
- To improve the **traceability** and the **relationship** between data.
- The purpose of **Variety** is to classify and separate data through categorization and segmentation.

Image Source: Google Images

Step :3

Success Criteria

VOLUME :- The percentage of data which can be preprocessed increases with the increase of volume of dataset and decreases with the decrease in the volume

The percentage of data which can be preprocessed increases with the increase of volume of dataset and decreases with the decrease in the volume.

VELOCITY :- An extreme difference in the velocity of the dataset during different time frames indicates an outdated dataset.

The amount of meaningful data which can be processed increases with respect to the increase in volume with respect to time and vice versa.

VARIETY :- When amount of data, the number of records, and the number of datasets increases from previously used dataset

The amount of meaningful data which can be processed increases with respect to the increase in volume with respect to time and vice versa.

Base and Derived Measures

Base Measures

Ndde - Number of Distinct Data Elements
Lbd: Length of Big Data (Number of Records)
Nds: Number of Datasets
Time : Time Frame (T1, T2, T3)

Derived Measures

Mvol
Mvel
Mvar

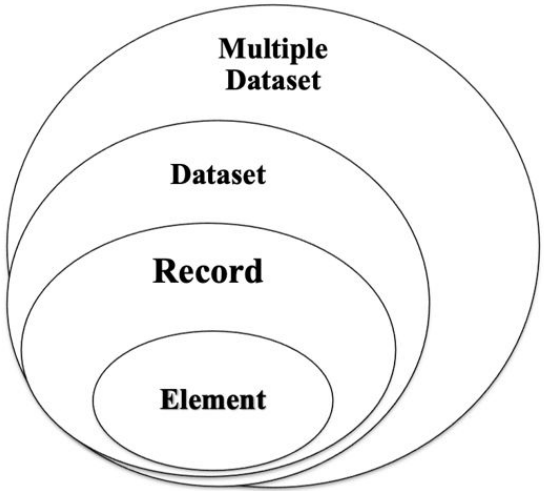


Image Source: Lecture Slides

Planning of the Measures

- Defined the **Labels**
- **Frequency** of Data Collection(**T1,T2,T3**)
- **TimeLine** (**min - 70** , **max 90** person hours)
- Procedure for collecting and recording data. (**Kaggle , Pandas DataFrame**)
- Data storage strategy (memory preprocessing).
- Role and responsibility (Defined and Distributed)

Measurement Goals	
Measurement Goals	**Labels**
Increasing the **Volume** 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

Plan Tasks and Activities

- Various Tasks are identified
- They are traced back respective labels
- Captured the stakeholders responsible
- Note the participants
- Estimated Duration is captured in Days
- Estimated Effort in person hours
- Schedule - When to plan the particular task
- Tool to be used
- Rationale behind the task performed

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

#	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 have a commitmen towards quality improvement

Data Collection

Guide

- People **who** collect the data (The stakeholders involved)
- Data Collection (Measures calculated and Data collected)
- Role of the collected data (**How** the collected data helps and its use cases)
- **Time** of data collection (Beginning , End or During the Time Frames)

Base Measures Data Collection Procedure

- Data analysed using Python code and manually
- Measure Ndde,Lbd,Nds
- Manually counts all datasets at T1,T2,T3
- The dataset is split for three different timeframes - T1, T2, T3

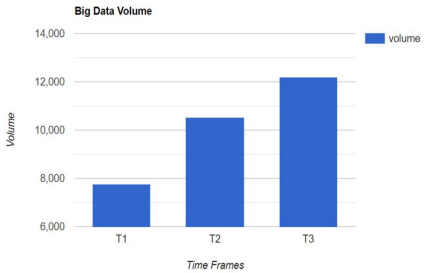
Measures

BASE MEASURES

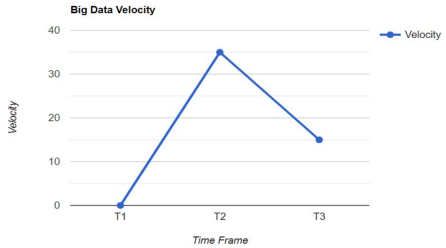
Data Collected	T1	T2	T3
Ndde	805	1050	1195
Lbd	200	248	300
Nds	1	1	1

DERIVED MEASURES/INDICATORS

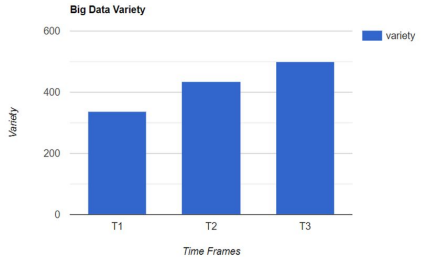
Volume(Mvol)



Velocity(Mvel)



Variety(Mvar)



Conclusions

There are clear trends in terms of the **volume, velocity, and variety** of datasets over the 3 time frames

From T1 to T3, the volume of big data has been observed to **increase gradually**

There hasn't been any **significant change in variety**

The **velocity measure varied** the most during the time frames

Even though the dataset is not big enough, we used this dataset for prototyping and found the dataset suitable for exercising measurement activities

Dataset has structured **data suitable for processing** using a machine learning algorithm

As new data can be added anytime we conclude that the data is suitable for machine learning algorithm

Retrospective Analysis

What we can improve

- We can take advantage of all the other V's to achieve better result
- As data source grows in future, move to Distributed File System for storing

What went well

- Understanding the Requirement and Business Goals
- Availability of structured dataset
- Team Collaboration
- Split of Roles and Responsibilities

What should we stop doing?

- The procrastinating behavior
- Postponement in learning new concepts

What have we learned?

- Big Data Quality Indicators
- Measuring V's
- Team Work
- Usage of collaboration tools
- Learning from mistakes

What still puzzles us ?

A benchmark of dataset size which is ideal

Top three issues and Recommendations

Issues:

- Data set is too small
- Doesn’t contain many outliers to capture variety
- Only one version of dataset available

Recommendation:

- Select a bigger Data set
- Split the dataset at different timeframes with incremental records
- Data Noise to be monitored frequently

Thank You!
Any Questions