# DAT 325 Project One

# Data Quality Plan

### Purpose Statement:

High quality data is important to all businesses. It helps ensure outputs that are useful to the organization. The data needs to act in a way that is expected to be useful for the business. Data that is lower quality increases the risk to the business. This can mean loss of profit, loss of productivity, poor business decisions, and greater workload to fix the data errors.

With high quality data, informed business decisions can occur. This can include knowing the on-hand stock of an item or which areas of the globe have the highest number of consumers of certain products, which can, in turn, lead to targeted advertising and a better investment of the marketing budget. As well, knowing the customer base can lead to greater satisfaction since their needs can be known and anticipated. This will lead to a competitive edge over other businesses with poorer quality data.

### Organizational Goals:

### The organizational goals for the data quality will first include identifying what the business needs are for the data that is coming over from Wayne Enterprises. Knowing what data is needed, what format it is needed in, and how it will be used is important for the creation and integration of the database.

### This will also involve designating a data quality team. This will be the team of individuals who are going to oversee the data quality process. Evaluating the data for the degree to which the data follows the correct format and patterns, is consistent and complete, as well as highlighting any abnormalities will head off potential issues later (Shen, 2019). This will also be repeated as data is entered periodically, but upon this initial transfer of data between companies the higher the quality of incoming data the easier the maintenance of high-quality data will be going forward.

A second over-reaching goal will be for the inspection of the data as it is coming in, utilizing data profiling and visualizations, will help with identifying the quality of the data that is coming over from Wayne Enterprises. By creating summary statistics about the data, ensuring that it meets standards, identify if there are missing values, ensuring the tables have primary and foreign keys, it allows for a greater understanding of the data and what changes need to occur in order to have it meet the needs of Bruce, Inc. This will help guide where the cleaning needs to occur and ensure that the data will meet the high-quality data standards expected.

A clear ETL pipeline will be developed with the ability to enable data lineage tracing. This is an important part of the data governance that will allow for more reliable data outputs and an easier identification of problem areas that may become evident later (Shen, 2019). Knowing from where the data comes allows for a greater understanding of what the data means and whether it is representing what it intends to represent as well.

### Data Quality Characteristics and Procedures:

The core dimensions that define data quality are completeness, validity, uniqueness, timeliness, accuracy, and consistency. These are the dimensions that are the industry standard for rating the quality of the data, which allows for confidence in business decisions as well as ensuring adherence to laws and guidelines of the industry.

The completeness of the data speaks to whether all the necessary information is present. An example of this would be if there are missing elements that are needed, such as a phone number or ISBN or other information that is a requirement for the business. Incomplete data can lead to inaccurate stock counts or inability to contact a customer, or other failing in the business strategy.

Uniqueness, sometimes called conformity, is how well the data conforms to the formats that are needed. This can mean that the constraints that are established are met, whether it be a mandatory column being filled in, a foreign-key constraint that must match a primary key in the referencing table, or patterns that must be met for fields such as phone numbers (Elgabry, 2020). The validity of the data ensures that it can be read, and it will correspond with other data points.

The validity of the data is also known as the integrity of the data. This speaks to the ability of the data to be traced and connected to other data, and not orphaned (Elgabry, 2020). The integrity of the data allows for connections to be made for the analysis.

Timeliness of data is the ability for it to be available when it is needed. If there are business constraints that require outputs in a specified timeframe then the data must be able to meet those requirements. This can be an expectation for a bank balance to be updated instantaneously with each transaction or a quarterly report aggregated by a certain date.

Accuracy is one of the most important of the core dimensions. This is measuring whether the data is true. Errors with accuracy will inevitably lead to poor outcomes from any data analysis as an accurate analysis cannot be made reliably, or may be undefendable, with inaccurate data.

Consistency of data is whether the data is consistent across the same data set or with multiple linked data sets. In one area data may indicate a person lives in a rental property, but in another area they have a mortgage payment listed. This would be flagged as contradictory information due to the inconsistencies, potentially leading to erroneous outcomes.

### Security and Personnel Responsibility Plan:

### Privacy of data as it is being transferred from one organization to another will be important. The servers on which data are stored increase the risk to the data. This can include confidentiality issues, data integrity problems, as well as availability complications. The servers should be accessible only to those who need to access it. This will involve a requirement for transactions to follow ACID Properties (atomicity, consistency, isolation, and durability) to ensure that the data integrity remains intact (IBM, 2020).

The data that has been received from Wayne Enterprises is also subject to privacy controls that were instituted upon the collection of that data. It will be important to ensure that future use of that data falls within what is within the agreement of any individual’s data given to Wayne Enterprises initially. If there is a limitation to the scope of use for the data when gathered, that will be the limitation for Bruce, Inc. as well.

The security of the data will be the responsibility of all who have any access to it. This will include following protocols to ensure individual safeguards for the data, such as secure passwords, using secure multi-level encryption for the server, and having all access points protected with antivirus software.

References

Elgabry, O. (2020, April 18). *The Ultimate Guide to Data Cleaning - Towards Data Science*. Medium. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

IBM. (2020, November 2). *IBM Knowledge Center*. https://www.ibm.com/support/knowledgecenter/SSGMCP\_5.4.0/product-overview/acid.html

Shen, S. (2019, July 29). *7 Steps to Ensure and Sustain Data Quality - Towards Data Science*. Medium. https://towardsdatascience.com/7-steps-to-ensure-and-sustain-data-quality-3c0040591366