**Foo.Bar Data Governance Plan**

Laurence T. Burden

Purdue Global University

IN498: Bachelor’s Capstone in Analytics

Robert Kayl

January 1, 2024

**Roles and Responsibilities**

**Data Owner**

The data owner shall be a senior member of the organization. Initially, this will be the Chief Information Officer (CIO). The data owner has the final say on all actions taken with the dataset titled *IN498\_Unit6\_Results.docx* and the related 11 CSV files, though most day-to-day responsibilities will be delegated to the data custodians and data steward.

**Data Custodians**

Any member of the company that has access to the data will be considered a data custodian. All data custodians must adhere to the policies outlined in this document. The data team shall oversee the security of the dataset. They will also be responsible for implementing any new requirements.

**Data Steward**

The data steward will be the head of the data team. They will be responsible for ensuring the quality and security of the data. They will oversee the storage, retrieval rules, and formatting of the dataset.

**Privacy**

The dataset does not contain any PII, but all relevant laws must still be understood and adhered to where they apply. The data shall be considered company property and kept private from those outside of the organization. A clear need-to-know must be shown before the data is released to others within the organization.

**Data Collection and Storage**

**Collection and Control of New Data Sets**

New data will be collected from the Google Play Store© every first business day of the month. The query will request data for only the previous month. The new files will be titled with the following format: retained\_installers\_com.foo.bar\_YYYYMM\_country.csv. The data will be integrated into the combined\_results.csv through the use of a standard Python script on the same day as it is retrieved.

**Metadata**

Metadata is information about the data. The organization has a central location where all metadata is listed for each project. The Foo.Bar project will utilize this location to store its metadata for easy retrieval by anyone that needs it. The data steward shall be the point of contact on this endeavor.

**Confidentiality of Data**

The data shall be kept confidential from anyone outside of the organization. Any member of the organization must show a need-to-know before being granted access to the data. The data steward has the authority to allow anyone in the organization access to the data, and the data owner must give approval to any requests for outside entities to have access to the data.

**Data Storage Types and Locations**

The main copy of the data shall reside on the company’s internal servers. It will be located in a secure folder with proper access controls. The data may reside on any individual’s computer that has the proper authority to access the data. This copy of the data will never be considered the master copy of the data and any changes made during an individual’s analysis will not be integrated into the master set.

The data shall not be stored on removable media (CDs, thumb drives, etc.). The primary method of transferring the data will be conducted over the company’s secure network or VPN for external connections. The data steward may authorize a deviation from this policy if circumstances make it necessary.

**Data Governance Plan**

**Scope**

This document sets out the formalized plan on governing information associated with data analysis of the Foo.Bar Android app. This application was created to track the user’s moods throughout a given time period. Any analysis done to further the business venture of this application will be stored according to this document.

All data collected on the Foo.Bar application must be movable between relevant business units. This document applies to following departments: customer service, IT, financial, and all operational departments. This document covers activities pertaining to external data exchanges, financial records, and logistics.

**Vision Statement**

*Operate at an optimized level of data management to the benefit of Foo.Bar’s continued growth.*

**Business Value and Alignment**

This plan must align with the rest of the company’s values and business strategy. The primary goal is to show a successful return on investment from the data analysis project for the Foo.Bar application. All policies laid out in this document must align with the organization’s master data governance plan. The master data plan will take precedence in the event of any conflicts between the two documents.

**Functional Design**

***Legal***

The storage and use of all data collected within this project will follow all applicable laws and regulations of each country and region that we service. This primarily includes the United States and European Union.

***Privacy***

The privacy of our users shall be maintained as much as possible. Any PII that is collected must follow all applicable laws. Any dataset that contains PII will be anonymized to preserve our users’ privacy.

***Ethical Concerns***

The data currently collected for this project does not have the ability to identify any specific individual. This allows us to use the information for any internal AI and machine learning projects without concern of running afoul on ethical concerns. The data will not be shared with any outside organization for the purpose of furthering their AI and ML projects, except where we have a fully recognized partnership and stake in the project.

**Communication Matrix**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Executive Leadership | Enterprise Data Stewards | Operational Data Stewards | IT | Partners | All Others |
| Documentation |  | **X** | **X** | **X** |  |  |
| Performance Metrics | **X** |  | **X** |  | **X** | **X** |
| Alters and Triggered Events |  | **X** | **X** | **X** | **X** |  |
| Data Governance Council Minutes | **X** | **X** | **X** |  |  |  |

**Road Map**

This data governance plan shall be implemented immediately. The organization already has a robust data governance plan and this plan shall be implemented using it. Any changes that need to be made to this document shall be coordinated with the data owner and data steward.

**Roll Out**

This plan will be rolled out using existing company assets. The existing shared drive will have a new folder created for the project and it will be properly secured to allow only the organizational data team, the IT team, and the assigned data team to have access. The IT team will also provision a new virtual server to contain the necessary technical requirements for the project.

These technical requirements consist of a database and data analysis tools. The database will be PostgreSQL 13.2.X. The data analysis environment will consist of Python 3.12 with the following packages installed: Panda, Seaborn, Matplotlib, Numpy, and Scikit-Learn.