Sembian Balasubramanian

UIUC CS598 Foundations of Data Curation Final Project

Part 2

Part 2 Objective

Write a convincing memo (650 word max) explaining why data curation services are important. Assume that the memo is written for your new director, who is not familiar with data curation, and not convinced whether to keep funding this work. You will want to make sure to introduce data curation within the broader context of data science. You will need to cover the key areas that you think are the most important for data curation at your company. We ask that you incorporate at least two of the following topics into your memo: Provenance, Policy, Metadata, and/or Preservation.

Part 2 Memo

As you are aware, many of our agency divisions have begun to hire new data scientists to analyze the gather insights into the wealth of data we have within our department. I want to share the critical role our team has to manage and curate the data within the area of data science for the creation, management, analysis, and communication of data.

We are building out a centralized data lake to make enterprise datasets securely available across the corporation by establishing an organized workflow for datasets as part of our data preservation strategy. As multiple heterogeneous datasets from different companies are ingested, we are capturing metadata to ensure datasets are searchable utilizing a data catalog, ensuring that data can be efficiently and reliably found and reused over time.

The metadata that we are storing captures critical attributes that provide security and data lineage to their source systems using an industry-standard definition of a dataset to ensure provenance to enable trust and reproducibility of data analysis. As datasets further generated from our initial ingested data sets, we are capturing information pertaining to their generation to ensure computational provenance. Our project provides a technical solution to data ingestion and includes policies and governance around the datasets to ensure we are protecting the corporation and building out business value.

Currently, the data scientists hired have challenges to find the data they need to perform their tasks. Using a data catalog, we will enable dataset exploration by data scientists in a secure environment. Our data scientists have the objective of extracting useful knowledge from data through exploratory data analysis, but they need data to have reliable data available to analyze. The scientific nature data science itself is based on the concept of data curation and data analysis. The valuable data analysis can only be performed if datasets can be reliably and efficiently available.

The metadata will provide discoverability of our datasets for their use and reuse. The security component incorporated into our metadata will ensure only authorized users will have access to these datasets.

Our organized workflow will provide a solution to ensure that datasets are not only stored and searchable, but that can be marked as an approved dataset, which will mean that our governance team has approved this data set that it has gone through our validation process to ensure it has the correct information and that our governance team has supported the inputs and calculated values found within the dataset. We will also ensure the datasets use the approved and standard data model and provide trust in the data being used for valuable analysis. This will ensure that data scientists across the corporation use the same base datasets with consistent data modeling to enable their research.

The solution we are working on is a scalable strategy for managing not just the information in our current customer complaints system but data generated across our cooperation. For this solution to scale, we need to ensure our workflow will provide prospective provenance, meaning that data scientists will have expectations regarding how the data will be stored and searched.

During this mainstreaming of data preservation, we also need to provide tools to ensure retrospective provenance to visualize and analyze the data that is outputted from our workflow solution and visualize and analyze the workflow itself.

By implementing the data curation practice, we will achieve efficiency and reliability, ensuring that the data we have is finable, usable, and secure, thereby supporting the use and adaptation of data management methods to meet our user's needs and support data analytics.