Megan Masanz

CS598 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

Jon,

As you are aware, many of our divisions have begun to hire data scientists to find value within the wealth of data we have within our organization. I want to share with you the import role our team has to play within the area of data science. I would first like to provide our teams vision and then provide insight into the business value delivered by this strategy.

We are building out a data lake to make enterprise datasets securely available across the corporation through establishing an organized workflow for datasets as part of our data preservation strategy. As datasets are ingested we are capturing meta data to ensure datasets are searchable utilizing a data catalog. The metadata that we are storing captures key attributes, provides security, as well as 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 are further generated from our initial ingested data sets, we are capturing information pertaining to their generation to ensure computational provenance. Our project does not just provide a technical solution to data ingestion but includes policies and governance around the datasets to ensure we are protecting the corporation and building out business value.

Currently the data scientists hired are struggling to find the data they need to perform their tasks. Using a data catalog, we will be able to enable dataset exploration by data scientists in a secure environment. Our data scientists have an objective of extracting useful knowledge from data, but they need data to have reliable data available to analyze. It may surprise you to know that 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 a approved dataset which will mean that this data set has been approved by our governance team that it has gone through our validation process to ensure it has the

correct information and that the inputs and calculated values found within the dataset have been approved by our governance team. We will also ensure the datasets are using 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 are using the same base datasets with consistent data modeling to enable their analysis.

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 main streaming our of data preservation, we also need to provide tools to ensure retrospective provenance so that we can visualize and analyze not only the data that is outputted from our workflow solution, but also visualize and analyze the workflow itself.

The cloud provides our governmental agency with unlimited computing resources, but with the cloud, the saying goes – “you get what you pay for”. Without this project, every team and business group will be pumping their own version of datasets without ensuring the security, reliability, reuse or searchability of these data sets. If our agency is to enable data science at scale, this project and proposal is key to our success.