MEMORANDUM

To: Director
From: Data Curation Lead
Date: 11/28/2019
Subject: Importance of Data Curation

I'm writing this to you inform you about the importance of Data Curation and why it is necessary for our organization.

"**You can have data without information, but you cannot have information without data."**

— Daniel Keys Moran

According to IDC, the world's data will grow from about 29 zettabytes (ZB) in 2018 to 175 ZB by 2025, a growth rate of 66% per year. 45-90% of data becomes "cold" within weeks of its creation. Yet, most organizations continue to treat all this data the same. People have and continue to gather, maintain, and archive data at ever greater volumes, and they always have. Data curation is a means of managing data that makes it more useful for users engaging in data discovery and analysis. With the rapid growth of data, organizations have lot of more divergent data sources to extract data from, making it much more difficult to maintain a consistent method to curate data. Further complicating the problem is the fact that much of today's data is created in an ad hoc way that can't be anticipated by the people intended to use data for analysis.

Data curation is the management of data throughout its lifecycle, from creation and initial storage to the time when it is archived for posterity or becomes obsolete and is deleted. The main purpose of data curation is to ensure that data is reliably retrievable for future research purposes or reuse.

Enterprises often struggle with getting things done and operationalizing Big Data. Without access to good Data Curation, business effectiveness decreases. Risks of poor or no Data Curation include factually inaccurate information, incorrect guidelines, and

knowledge gaps. Being able to present the data in an effective manner is also extremely important. It's important to know the context of the data before it can be trusted.

Data curation is important in today's world of data sharing and self-service analytics, but it is a frequently misused term. Often data in data lakes and data warehouses are considered as curated data, believing that it is curated because it is stored as shareable data. Curating data involves much more than storing data in a shared database. Collecting and storing data in a sharable location is only the starting for data curation activities. It involves more activities like organizing the data using appropriate data model and standards, preserving the data such that is usable for future, ensuring that the data is secure from tampering and follow legal and regulatory polices etc.

Our organization being a government agency, it is important to perform certain data curation activities listed below but are not limited to.

- **Compliance** – This ensures that the data follows all the legal, regulatory and local **policy** requirements. Ensuring that the data is compliant makes sure that we are prepared for Audits.

- **Preservation** - Ensures that data will be understandable and useable in the future. This helps us retrieve data easily in the future either for analysis or audit purposes.

- **Provenance** – It is often necessary to understand what are the calculations and manipulations that are performed on data to ensure that the data is correctly represented. Because datasets are used and reformulated or reworked to create new data, provenance is important to trace newly designed or repurposed data back to their original datasets. Provenance support identifying what inputs, processes, and calculations are responsible for data values.

- **Organization** – Organization in data curation ensures that an appropriate data model and standards are applied in storing the data. Data curation is more concerned with maintaining and managing the **metadata** rather than the database

itself and, to that end, a large part of the process of data curation revolves around ingesting metadata such as schema, table and column popularity, usage popularity, top joins/filters/queries.

In summary, not everyone in data deals with information the same way. Data analyst uses technology to make sense of the available information. Data Scientists gather the data through various means. Data curation bridges the divide between these two by providing information about how the data is used.

External References: (Not part of memo, for assessment)

https://www.alation.com/what-is-data-curation/
https://www.dataversity.net/data-curation-101/
https://en.wikipedia.org/wiki/Data_curation
https://cloudian.com/blog/getting-a-handle-on-data-growth/
https://www.ringlead.com/blog/20-inspirational-quotes-about-data/