To Mr. A,

I am pleased to have you on the team and excited for us to work together. I understand joining new organization might be at first overwhelming. As lead data scientist of data curation project, I’d like to fill you in with the foundation and circumstances of our work.

As you may know, we have an ongoing project of switching the customer complaint data to a new database system. Due to the fundamental differences from old to new data structures, we discovered a number of data contract/schema inconsistencies between them. This is a blocker which considerably hindered the new platform from being put into service, while the old is decommissioned. With the continued growth of our company comes large amount of data to be transmitted which demand quality assurance attention. Should we fail to identify the heterogeneity and data independence matter, crisis is prone to happen. Therefore, Data Curation initiative was introduced to preserve the integrity and sufficiency as we migrate the dataset to its new home. The work focuses on metadata, preservation system and accurate workflow.

“Data Curation? Meta Data? Provenance? Etc”. You might wonder what they are and how they benefit our work…

Data Curation can be judged from a Data Science perspective. It combines data analysis and curation into a single stack. While analysis means acquiring knowledge of data using queries and business domain specifics, curation is a companion to ensure things are valid and reliable across different departments. Important factors of data curation are documentation, validation, transformation and integration.

When data volume continues to accelerate, they need to conform to certain metadata rule, or in another word, data contract. Metadata is the “data of data”. Just as it explains the characteristics of a data segment, so too it tracks different versions of that segment of data. *One example of meta data is the channel (web/phone/email) describing the communication mean via which a customer complaint was submitted. Similarly. each complaint encapsulates more line items: who complained, when it was received, what was our response, etc*. By designing and maintaining a robust metadata, we can achieve documentation factor of Data Curation. It assists identify how a datum was originated, where it located and the detailed information it entailed. Accordingly, these success criteria enhance organization understanding over the management lifetime. Likewise, it helps interpret data with familiar attribution’s semantics, improves quality and integrity, avoids illegal usage and helps formalize the integration process.

In addition to metadata, a preservation system is mandated to reliably persist data for future references. With this, the intrinsic information of a particular subject is stable even if business domain fluctuates. It’s necessary to track data versions and keep them up to date with new schema (if there is). Hence, we need software stack and procedures to convey the preservation process.

As also, we need a workflow to support data transformation. It’s in place to supervise what’s going on with each data iteration, identify what is being done or overlooked, the inputs and outputs at each node of the whole chain. Should an issue emerged, a well design workflow let us quickly pinpoint the at-fault unit and resolve the problem promptly.

For a customer centric company, retainable and reliable user data are our assets. Attempting a new platform without asserting old data compatibility poses a risk of damaging that asset. We need methodology to evaluate our data, find potentials and keep good practice with maintaining and transforming them. With the points I laid out above, Data Curation is the practical solution to this subject matter. I hope the proposal clearly presented the priority not only in the data themselves, but also in the workflow that orchestrate this asset among different sectors. Therefore, I think it’s crucial that you reprioritize our projects and keep funding this important work.