# Data Management Plan

Loic Niragire

TIM-8130 v4: Data Curation

Dr. Ulrich Vouama

March 6, 2022

1. Architecture role in the storage process

A successful data management strategy requires the alignment of business functionalities and their associated data. This includes needed data to accomplish a given task and produced data and metadata. Additionally, design considerations must ensure data quality, accessibility, reliability, and security. Finally, the role of architecture in the storage process is to assert that data remain in an actionable state where organizations can effectively act upon it with greater confidence.

Design must ensure that required data is available and accessible at the point of request, which could be key business personnel making a decision or an automated business process. This is a challenging design goal requiring thoughtful architectural considerations. I highlight three reasons that make this design a challenge in the following sections.

1.1 Data vs. Context

Data is dynamic and assumes different shapes based on the context. A customer entity in an accounting context, for instance, will have another representation in a shipping context even though both representations reference the same customer. The shipping system may not need to know a customer's bank information. Likewise, the accounting system may not care about a customer's delivery preferences. Different concerns within a system dictate variations in data representation. It is the role of the architect to craft context boundaries within an organization carefully.

### 1.2 Data Ownership

The context boundary challenge discussed in the above section leads to the ownership challenge. Notably, what processes can operate on a given piece of data within a given context? In this case, operating on a piece of data means assuming knowledge. For instance, does the shipping system assumes the knowledge of a customer's bank information from the accounting system? In other words, who owns the various parts of a customer entity, and how are modifications to those aspects mitigated. Are specific security measures required for those aspects? Architecture defines domains within a system and assigns data ownership within those domains.

### 1.3 Data vs. Metadata

Bounded context is often used within the micro-service literature to refer to a business domain data entity within a specific context. These concepts are illustrated in the previous two sections using a customer entity in the shipping and accounting context. The challenge I address in this section is due to this split representation – namely, metadata. Does metadata follow the same rules as the data it describes? What about ownership, usage, collection, and storage? Are there different policies on metadata? Architecture plays a significant role in addressing these concerns and asserts that metadata management is adequately considered.

## 2. Data Repository Scenario

### 2.1 General goals and objectives

The data repository use-case scenario I selected for support is concerned with pulmonology and respiratory medicine. The goal is to provide a data management plan that allows researchers to conduct descriptive, predictive, diagnostic, and prescriptive analysis sessions on curated data. The primary dataset is currently made up of 46,032 rows and 45 columns.

### 2.2 Data Management Plan

I propose a micro-service-based architecture to approach the above data management scenario as a solution approach. This architecture facilitates data encapsulation by aligning micro-services to business functionalities, which allows a granular security and access control model (Anastasiou, Lin, He, Chiang, & Shahabi, 2019). The process begins with an extensive data modeling phase involving business analysts and software developers, with the primary goal of understanding organizational data. This is an essential step as it helps uncover available data and the processes that act upon them. Additionally, it helps bridge the business language with that of software developers. The outcome should be a detailed data model illustrating data relationships and essential metadata to be captured and maintained. Obtained data model helps align business functionalities to micro-services by establishing bounded contexts.

Due to the non-relational characteristics of the dataset, proposed data modeling will follow a NoSQL scheme and leverage a column-oriented database. The choice of column-oriented data store is largely due to their efficient disk I/O performance and memory

utilization, which makes them suitable for data analytics. Moreover, column-oriented databases have greater data evolution support than row-oriented databases (Liu, He, Hsiao, & Chen, 2011). Data access will be managed by exposing Restful Web APIs built on top of corresponding micro-services. It is worth mentioning that these micro-services are designed for CRUD operations in service of analytics applications. Metadata and data catalogs will be managed through a data warehouse. Whereas policies and procedures guiding the creation and preservation of new data will be implemented through a business rule engine, which is a helpful strategy for detecting data security vulnerabilities (Rouf, et al., 2019)

References

Anastasiou, C., Lin, J., He, C., Chiang, Y.-Y., & Shahabi, C. (2019). ADMSv2: A Modern Architecture for Transportation Data Management and Analysis. ARIC'19: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances on Resilient and Intelligent Cities (pp. 25-28). ACM.

Liu, Z., He, B., Hsiao, H.-I., & Chen, Y. (2011). Efficient and Scalable Data Evolution with Column Oriented Databases. EDBT/ICDT'11: Proceedings of the 14th International Conference on Extending Database Technology (pp. 105-116). ACM.

Rouf, Y., Mukherjee, J., Fokaefs, M., Shtern, M., Le, J., & Litiou, M. (2019). Rule-based security management system for data-intensive applications. CASCON'19: Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering (pp. 254-263). ACM.