NIST Big Data Lifecycle Management

# 1. Overview

Lifecycle Management is responsible of managing data coming into the system, residing within the system, and going out of the system for application usage. In other words, the role of Lifecycle Management is to ensure that the data are accessible by other Provider Components throughout the lifecycle of the data, since the moment they are ingested into the system by the Data Provider, and until the data are dispositioned. Moreover, this accessibility has to comply with policies, regulations, and security requirements. In the context of Big Data, Lifecycle Management has to deal with the three V characteristics: Volume, Velocity and Variety. As such, Lifecycle Management and its components will have to interact with other components of the Big Data Reference Architecture, such as Capability Provider, Transformation Provider, and Vertical Orchestrator.

# 2. Activities

Lifecycle Management activities include:

* Metadata Management. Metadata Management is the enabler of Lifecycle Management, since metadata are used to store information that governs the lifecycle management of the data within the system. Metadata also contains critical information such as persistent identification of the data, the fixity, and the access rights.
* Accessibility Management
  + Data Masking for security privacy. Privacy information has to be anonymized prior to the data analytics process. For instance, demographic data can be aggregated and analyzed to reveal data trends, but specific personal identifiable information (PII) with names and social security numbers have to be masked. This masking managed by Lifecycle Management depends on the type of application usage and the authorization usage specified by Security and Privacy.
  + Accessibility of data may change over time. For instance, Census data can be made available to the public after 75 years. In that case, Lifecycle Management is responsible of triggering the update of the accessibility of the data or sets of data according to the policy and legal requirements. Normally, data accessibility information is stored in the metadata.
  + Data Recovery. Data management should also include recovering data that were lost due to disaster, or system/storage fault. Traditionally, this data recovery can be achieved using backup and restore mechanisms. But, in order to cope with the large volume of Big Data, this should be embedded in the architectural design, and the exploitation of modern technologies within the Big Data Capability Provider.
* Preservation Management. At the basic level, the system needs to ensure the integrity of the data so that the veracity and velocity of the analytics process are fulfilled. Due to the extremely large volume of Big Data, Preservation Management is responsible to disposition aged data contained in the system. Depending on the retention policy, these aged data can be deleted or migrated to archival storage. On the other hand, in the case where data need to be retained for years, decades and even centuries, there will be a need to have a preservation strategy so they can be accessed by the Provider Components if required. This will invoke the so-called long-term digital preservation that can be performed by Transformation Provider using the resources in Capability Provider.

# 3. Interaction with Other Components

In order to perform its activities, Lifecycle Management will interact with the other Provider Components of the Big Data Reference Architecture:

* Data Provider to manage the metadata from the entry of data into the Big Data system;
* Transformation Provider to perform data masking and format transformations for preservation purpose;
* Capabilities Provider to perform basic bit-level preservation and data recovery;
* Security and Privacy to keep the data management up to date according to new security policy and regulations. In the other direction, Security and Privacy also utilizes information coming from Lifecycle Management with respect to data accessibility. Assuming that Security and Privacy controls access to the functions and data usage produces by the Big Data system, this data access control can be informed by the metadata managed and updated by Lifecycle Management.