|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Item #** | **Type** | **Page #** | **Line #** | **Section** | **Comment (with rationale)** | **Suggested Change** |
| 1 | G | vi |  |  | Quyen Nguyen is now with US Census Bureau | Change NARA to US Census Bureau |
| 2 | T | 21 |  | 4.1 | Fig. 7: The Management component is missing the Big Data Lifecycle Management sub-component (BDLM) | Add BDLM to the Management Fabric. |
| 3 | T |  |  | 4.2.4 |  | Add a subsection for managing data according to the Functional Component for Big Data Lifecycle Management as followed: Data Lifecycle Manager Lifecycle Data Management components are necessary to manage the lifecycle of the data ingested into the system, stored and preserved in the system, and accessed for processing or dissemination purposes:  **Data Catalog** is the inventory of all data records and data sets in the system. It should contain the model for the foundational concept of “unit” of data, whether it is a database record (e.g. key-value pair or relational table row), or a data set (e.g. database export file). Each data unit has characteristics maintained in the associated metadata, which should include at least a unique identifier and timestamp indicating when the data was created and/or ingested. These timestamps will help the Data Lifecycle Manager to monitor the “age” of the data within the system. Moreover, the Data Catalog will have to support data discovery that is necessary for data access, and data governance.  **Data Tracker** tracks the movement of data throughout the system, from the ingestion point to the dissemination or destruction point. Data Tracker component has to handle the Volume and Variety characteristics inherent to Big Data. There are two kinds of movements:   * Ingress and egress movement tracks data entering and exiting the system. Data exiting means that the data are dispositioned to satisfy the retention policy, which can originate from either the need of the Big Data application or Preservation policy. Indeed, some applications may require “fresh” data for analytical purpose. The degree of freshness depends on the specific requirements of the business applications, and can be influenced by policy and regulations. For instance, while the visual analytics application monitoring the approval or disapproval feedback during a presidential election debate requires realtime and most recent tweet and blog data, the study of the trend of household income over the past 50 years needs both recent and archived Census data. On the other hand, records management laws and policies may dictate the retention time for the data, and hence impact the Data Preservation. * Intra-system. Due to the large Volume of Big Data, the Big Data Framework Provider will likely have multi-tier storage for cost efficiency and scalability. Within that storage environment, data have to made available to the analytics processes managed by the Big Data Application Provider. Commercial vendors have offered different storage categories with different pricing models. Examples are Microsoft Azure Hot, Cool, and Archive tier storage, Amazon Web Services (AWS) Elastic Block Storage, S3 and Glacier, and Google cloud Multi-regional, Regional, Nearline, and Cold storage classes. The action of making available to processes and applications could be realized by physically moving the data to a storage where the processing software can operate. However, recent paradigm is to move computation and processing capabilities to where data are located instead to circumvent the large data transfer between storage tiers. Whatever approach or methodology is used, the Data Tracker has to manage the so-called “checkout” of data, for lack of better terms.   **Data Preservation** component is applied to both permanent and temporaneous data. Its responsibility is to continuously inspect the “age” of data in the system, and take action on the data based on the retention policy. For permanent data, Data Preservation will perform the Preservation Plan, which can consist of migrating data to a long-term preservation format, periodically refreshing of the storage hardware, or maintaining emulation environments used to read the archived data. Data Preservation will leverage the multi-tier storage which satisfies data durability requirement, and achieves cost efficiency. If data are deemed to have limited lifetime, then Data Preservation will apply appropriate disposition methods to purge them from the system. The purge methods will depend on the security policy to ensure data confidentiality. |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |