|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | National Archives and Records Administration Accession NARA Accession, Search, Retrieve, Preservation | |
| **Vertical (area)** | | Digital Archives | |
| **Author/Company/Email** | | Quyen Nguyen & Vivek Navale (NARA) | |
| **Actors/Stakeholders and their roles and responsibilities** | | Agencies’ Records Managers  NARA’s Records Accessioners  NARA’s Archivists  Public users | |
| **Goals** | | Accession, Search, Retrieval, and Long term Preservation of Big Data. | |
| **Use Case Description** | | 1. Get physical and legal custody of the data. In the future, if data reside in the cloud, physical custody should avoid transferring big data from Cloud to Cloud or from Cloud to Data Center. 2. Pre-process data for virus scan, identifying file format identification, removing empty files 3. Index 4. Categorize records (sensitive, unsensitive, privacy data, etc.) 5. Transform old file formats to modern formats (e.g. WordPerfect to PDF) 6. E-discovery 7. Search and retrieve to respond to special request 8. Search and retrieve of public records by public users | |
| **Current**  **Solutions** | **Compute(System)** | | Linux servers |
| **Storage** | | NetApps, Hitachi, Magnetic tapes. |
| **Networking** | |  |
| **Software** | | Custom software, commercial search products, commercial databases. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed data sources from federal agencies.  Current solution requires transfer of those data to a centralized storage.  In the future, those data sources may reside in different Cloud environments. |
| **Volume (size)** | | Hundred of Terabytes, and growing. |
| **Velocity**  **(e.g. real time)** | | Input rate is relatively low compared to other use cases, but the trend is bursty. That is the data can arrive in batches of size ranging from GB to hundreds of TB. |
| **Variety**  **(multiple datasets, mashup)** | | Variety data types, unstructured and structured data: textual documents, emails, photos, scanned documents, multimedia, social networks, web sites, databases, etc.  Variety of application domains, since records come from different agencies.  Data come from variety of repositories, some of which can be cloud-based in the future. |
| **Variability (rate of change)** | | Rate can change especially if input sources are variable, some having audio, video more, some more text, and other images, etc. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Search results should have high relevancy and high recall.  Categorization of records should be highly accurate. |
| **Visualization** | | TBD |
| **Data Quality** | | Unknown. |
| **Data Types** | | Variety data types: textual documents, emails, photos, scanned documents, multimedia, databases, etc. |
| **Data Analytics** | | Crawl/index; search; ranking; predictive search.  Data categorization (sensitive, confidential, etc.)  PII data detection and flagging. |
| **Big Data Specific Challenges (Gaps)** | | Perform pre-processing and manage for long-term of large and varied data.  Search huge amount of data.  Ensure high relevancy and recall.  Data sources may be distributed in different clouds in future. | |
| **Big Data Specific Challenges in Mobility** | | Mobile search must have similar interfaces/results | |
| **Security & Privacy**  **Requirements** | | Need to be sensitive to data access restrictions. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | . | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |