**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Mendeley – An International Network of Research | |
| **Vertical (area)** | | Commercial Cloud Consumer Services | |
| **Author/Company/Email** | | William Gunn / Mendeley / william.gunn@mendeley.com | |
| **Actors/Stakeholders and their roles and responsibilities** | | Researchers, librarians, publishers, and funding organizations. | |
| **Goals** | | To promote more rapid advancement in scientific research by enabling researchers to efficiently collaborate, librarians to understand researcher needs, publishers to distribute research findings more quickly and broadly, and funding organizations to better understand the impact of the projects they fund. | |
| **Use Case Description** | | Mendeley has built a database of research documents and facilitates the creation of shared bibliographies. Mendeley uses the information collected about research reading patterns and other activities conducted via the software to build more efficient literature discovery and analysis tools. Text mining and classification systems enables automatic recommendation of relevant research, improving the cost and performance of research teams, particularly those engaged in curation of literature on a particular subject, such as the Mouse Genome Informatics group at Jackson Labs, which has a large team of manual curators who scan the literature. Other use cases include enabling publishers to more rapidly disseminate publications, facilitating research institutions and librarians with data management plan compliance, and enabling funders to better understand the impact of the work they fund via real-time data on the access and use of funded research. | |
| **Current**  **Solutions** | **Compute(System)** | | Amazon EC2 |
| **Storage** | | HDFS Amazon S3 |
| **Networking** | | Client-server connections between Mendeley and end user machines, connections between Mendeley offices and Amazon services. |
| **Software** | | Hadoop, Scribe, Hive, Mahout, Python |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed and centralized |
| **Volume (size)** | | 15TB presently, growing about 1 TB/month |
| **Velocity**  **(e.g. real time)** | | Currently Hadoop batch jobs are scheduled daily, but work has begun on real-time recommendation |
| **Variety**  **(multiple datasets, mashup)** | | PDF documents and log files of social network and client activities |
| **Variability (rate of change)** | | Currently a high rate of growth as more researchers sign up for the service, highly fluctuating activity over the course of the year |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Metadata extraction from PDFs is variable, it’s challenging to identify duplicates, there’s no universal identifier system for documents or authors (though ORCID proposes to be this) |
| **Visualization** | | Network visualization via Gephi, scatterplots of readership vs. citation rate, etc |
| **Data Quality** | | 90% correct metadata extraction according to comparison with Crossref, Pubmed, and Arxiv |
| **Data Types** | | Mostly PDFs, some image, spreadsheet, and presentation files |
| **Data Analytics** | | Standard libraries for machine learning and analytics, LDA, custom built reporting tools for aggregating readership and social activities per document |
| **Big Data Specific Challenges (Gaps)** | | The database contains ~400M documents, roughly 80M unique documents, and receives 5-700k new uploads on a weekday. Thus a major challenge is clustering matching documents together in a computationally efficient way (scalable and parallelized) when they’re uploaded from different sources and have been slightly modified via third-part annotation tools or publisher watermarks and cover pages | |
| **Big Data Specific Challenges in Mobility** | | Delivering content and services to various computing platforms from Windows desktops to Android and iOS mobile devices | |
| **Security & Privacy**  **Requirements** | | Researchers often want to keep what they’re reading private, especially industry researchers, so the data about who’s reading what has access controls. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | This use case could be generalized to providing content-based recommendations to various scenarios of information consumption | |
| **More Information (URLs)** | | <http://mendeley.com> <http://dev.mendeley.com> | |
| **Note:** <additional comments> | | | |