**Note: No proprietary or confidential information should be included**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Pathology Imaging/digital pathology | |
| **Vertical (area)** | | Healthcare | |
| **Author/Company/Email** | | Fusheng Wang/Emory University/fusheng.wang@emory.edu | |
| **Actors/Stakeholders and their roles and responsibilities** | | Biomedical researchers on translational research; hospital clinicians on imaging guided diagnosis | |
| **Goals** | | Develop high performance image analysis algorithms to extract spatial information from images; provide efficient spatial queries and analytics, and feature clustering and classification | |
| **Use Case Description** | | Digital pathology imaging is an emerging field where examination of high resolution images of tissue specimens enables novel and more effective ways for disease diagnosis. Pathology image analysis segments massive (millions per image) spatial objects such as nuclei and blood vessels, represented with their boundaries, along with many extracted image features from these objects. The derived information is used for many complex queries and analytics to support biomedical research and clinical diagnosis. Recently, 3D pathology imaging is made possible through 3D laser technologies or serially sectioning hundreds of tissue sections onto slides and scanning them into digital images. Segmenting 3D microanatomic objects from registered serial images could produce tens of millions of 3D objects from a single image. This provides a deep “map” of human tissues for next generation diagnosis. | |
| **Current**  **Solutions** | **Compute(System)** | | Supercomputers; Cloud |
| **Storage** | | SAN or HDFS |
| **Networking** | | Need excellent external network link |
| **Software** | | MPI for image analysis; MapReduce + Hive with spatial extension |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Digitized pathology images from human tissues |
| **Volume (size)** | | 1GB raw image data + 1.5GB analytical results per 2D image; 1TB raw image data + 1TB analytical results per 3D image. 1PB data per moderated hospital per year |
| **Velocity**  **(e.g. real time)** | | Once generated, data will not be changed |
| **Variety**  **(multiple datasets, mashup)** | | Image characteristics and analytics depend on disease types |
| **Variability (rate of change)** | | No change |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | High quality results validated with human annotations are essential |
| **Visualization** | | Needed for validation and training |
| **Data Quality** | | Depend on pre-processing of tissue slides such as chemical staining and quality of image analysis algorithms |
| **Data Types** | | Raw images are whole slide images (mostly based on BIGTIFF), and analytical results are structured data (spatial boundaries and features) |
| **Data Analytics** | | Image analysis, spatial queries and analytics, feature clustering and classification |
| **Big Data Specific Challenges (Gaps)** | | Extreme large size; multi-dimensional; disease specific analytics; correlation with other data types (clinical data, -omic data) | |
| **Big Data Specific Challenges in Mobility** | | 3D visualization of 3D pathology images is not likely in mobile platforms | |
| **Security & Privacy**  **Requirements** | | Protected health information has to be protected; public data have to be de-identified | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Imaging data; multi-dimensional spatial data analytics | |
| **More Information (URLs)** | | <https://web.cci.emory.edu/confluence/display/PAIS>  <https://web.cci.emory.edu/confluence/display/HadoopGIS> | |
| **Note:** <additional comments> | | | |