We expect other WG to comment on and probably edit the use case proposal that follows.

There are 5 existing use cases (The last 3 of these use cases need minor updates for changed template)

* Web Search
* Remote Sensing of Ice Sheets
* NIST/Genome in a Bottle Consortium
* Particle Physics
* Netflix

We got volunteers to collect use cases

* Yuri Demchenko ( Use case (UvA1): LifeWatch – European Infrastructure for Biodiversity and Ecosystem Research; Use case (UvA2): Humanities and language research infrastructure )
* William Miller (Cargo Shipping)
* Gary Mazzaferro sent template to OOI (Ocean Observatory Initiative)
* Fox will do Astronomy

We need others to contribute

**Current Draft:**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Genomic Measurements | |
| **Vertical (area)** | | Healthcare | |
| **Author/Company/Email** | | Justin Zook/NIST/jzook@nist.gov | |
| **Actors/Stakeholders and their roles and responsibilities** | | NIST/Genome in a Bottle Consortium – public/private/academic partnership | |
| **Goals** | | Develop well-characterized Reference Materials, Reference Data, and Reference Methods needed to assess performance of genome sequencing | |
| **Use Case Description** | | Integrate data from multiple sequencing technologies and methods to develop highly confident characterization of whole human genomes as Reference Materials, and develop methods to use these Reference Materials to assess performance of any genome sequencing run | |
| **Current**  **Solutions** | **Compute(System)** | | 72-core cluster for our NIST group, collaboration with >1000 core clusters at FDA, some groups are using cloud |
| **Storage** | | ~40TB NFS at NIST, PBs of genomics data at NIH/NCBI |
| **Networking** | | Varies. Significant I/O intensive processing needed |
| **Software** | | Open-source sequencing bioinformatics software from academic groups (UNIX-based) |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Sequencers are distributed across many laboratories, though some core facilities exist. |
| **Volume (size)** | | 40TB NFS is full, will need >100TB in 1-2 years at NIST; Healthcare community will need many PBs of storage |
| **Velocity**  **(e.g. real time)** | | DNA sequencers can generate ~300GB compressed data/day. Velocity has increased much faster than Moore’s Law |
| **Variety**  **(multiple datasets, mashup)** | | File formats not well-standardized, though some standards exist. Generally structured data. |
| **Variability (rate of change)** | | Sequencing technologies have evolved very rapidly, and new technologies are on the horizon. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | All sequencing technologies have significant systematic errors and biases, which require complex analysis methods and combining multiple technologies to understand, often with machine learning |
| **Visualization** | | “Genome browsers” have been developed to visualize processed data |
| **Data Quality** | | Sequencing technologies and bioinformatics methods have significant systematic errors and biases |
| **Data Types** | | Mainly structured text |
| **Data Analytics** | | Processing of raw data to produce variant calls. Also, clinical interpretation of variants, which is now very challenging. |
| **Big Data Specific Challenges (Gaps)** | | Processing data requires significant computing power, which poses challenges especially to clinical laboratories as they are starting to perform large-scale sequencing. Long-term storage of clinical sequencing data could be expensive. Analysis methods are quickly evolving. Many parts of the genome are challenging to analyze, and systematic errors are difficult to characterize. | |
| **Big Data Specific Challenges in Mobility** | | Physicians may need access to genomic data on mobile platforms | |
| **Security & Privacy**  **Requirements** | | Sequencing data in health records or clinical research databases must be kept secure/private, though our Consortium data is public. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | I have some generalizations to medical genome sequencing above, but focus on NIST/Genome in a Bottle Consortium work. Currently, labs doing sequencing range from small to very large. Future data could include other ‘omics’ measurements, which could be even larger than DNA sequencing | |
| **More Information (URLs)** | | Genome in a Bottle Consortium: www.genomeinabottle.org | |
| **Note:** <additional comments> | | | |

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