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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Use Case Title** | | Pharmaceutical and Biomedical Research, Discovery and Diagnostics Systems | | |  |
| **Vertical (area)** | | Pharmaceutical and Biomedical companies | | |  |
| **Author/Company/Email** | | Albert Landeck, Perkin Elmer, Inc.(PKI) – Informatics Group - albert.landeck@perkinelmer.com | | |  |
| **Actors/Stakeholders and their roles and responsibilities** | | Pharmaceutical companies (Pharmaceutical chemists, Quality Assurance, Regulators, Manufacturing, Intellectual Property/Legal), Government (Certifying bodies, Regulators), BioMedical Research (Geneticists, Molecular Biology, Cell Pathology), Clinical Trials (Subject management) | | |  |
| **Goals** | | Enabling comprehensive lab discovery analytics for the pharmaceutical and life science industries | | |  |
| **Use Case Description** | | To tie together data from the drug discovery process, through assay testing, drug trials, analysis of genetic and cellular results with biomarkers, as well as manufacturing and chemical registry and inventory management | | |  |
| **Current**  **Solutions** | **Compute(System)** | | Each company has their own custom hardware installation that includes a core Electronic Notebook system either single instance or clustered cloud deployed (Amazon AWS or Verizon Terramark), a cluster of middle tier application servers, an ad hoc middle tier layer using Spotfire, and numerous external data sources | |  |
| **Storage** | | A core set of database file systems and a large file store | |  |
| **Networking** | | Corporate internet and cloud based VPC | |  |
| **Software (Identify COTS, open source products** | | Oracle, PostgreSQL, MongoDB, SciDB, Cassandra, PKI Chemical Search analytics, Genetic Indexing, Cellular imaging spatial and contrast data | |  |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | A core data set with affiliated distributed data sets (Chemistry, Genetics/Genomic, Imaging, Spectroscopy) | |  |
| **Volume (size)** | | Core data set as large as 10 TB, with federated datasets as large as 100TB | |  |
| **Velocity**  **(e.g. real time)** | | No special real-time requirements | |  |
| **Variety**  **(multiple datasets, mashup, how various)** | | Traditional database data as well as analytics for searching of Base64-CDX chemistry, Genomic data (including biomarker signatures), and PAIS (Pathology Imaging) | |  |
| **Variability (rate of change)** | | No special requirements | |  |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Strict control of core data for regulatory compliance reasons. Federated data sources have variable veracity (from internal controlled data to open source data) | |  |
| **Visualization** | | Representation within an application as well as 3-D molecular model viewing or chemical editing, graphs/histograms of assay data, viewing of cellular images with displayed GIS metadata. | |  |
| **Data Quality** | | The purpose of the system is to provide workflows for the curation of pharma and life sciences data. | |  |
| **Data Types** | | Base64-CDX chemical formulas, Genome sequences (complete and biomarker), TCGA/PAIS Cancer Genomic Sample sets | |  |
| **Data Analytics** | | Chemical search analytics (find a chemical that is like another chemical), Genome matching | |  |
| **Big Data Specific Challenges (Gaps)** | | The wide open and extensible nature of data including complex binary data with specific database indexing technologies. | | |  |
| **Big Data Specific Challenges in Mobility** | |  | | |  |
| **Security & Privacy**  **Requirements** | | **ITEM** | | **RESPONSE** | **CLARIFICATION** |
|  | | Investigator provenance | | A complete audit trail is maintained (including read-only access). A complex ACL and workflow model is also maintained. | Method used to associate researchers to digital artifacts |
| Sponsor disclosures | | Tunable rules by Group Access (departmental, regulatory, administrative) | Method used to associate researchers to digital artifacts; permissions for redisclosure |
| Investigator interests | | Application Virtualization for Contract Research Organizations (CRO) | Investigators may be required to disclose potential conflicts of interest |
| Institution where performed | | Internal controls strictly maintained to track at departmental and research group levels | Method used to associate responsible institution to digital artifacts; redisclosure rules; point in time considerations |
| Investigator affiliations | | Not an issue since the data is all single owner | Method for associating digital artifacts with investigator affiliations; one to many; point in time |
| Human Subject Data | | Not in core product. Limited HIPAA/HL7 data may be present when managing clinical trial data | Yes/No |
| IRB traceability | | FDA Title 21 regulations. See <http://en.wikipedia.org/wiki/Title_21_CFR_Part_11> , regulations also vary by region and country | Institution-specific event(s), digital records, US-specific regulation |
| Data / analytics / meta-result publication rights | | Results are typically corporate proprietary with highest levels of control of all data. | Open publisher; traditional publisher; white paper; working paper; IP issues |
| Results repository | | Results are immutable. Multiversioning of datasets is maintained with flashback capability | Immutable store for data collected, results |
| Reference data | | Some use of Chemical Abstract Service data especially with respect to standardization of reagents and solvents, NIH Human Genome data, NIH National Cancer Institute – Cancer Genomic Atlas | Third party dependency for ref - Census or geospatial data could be basis for independent variables |
| Delegated rights | | Since this is an internal proprietary data source, there are no delegated rights issues to be solved | Distributed delegation:: legal, governance, provenance covenants |
| Intellectual property | | All results generated within this system are to be treated as potential IP. | Includes COTS, open source EXE, collection artifacts |
| Third party privacy notices | |  | Voluntary or mandated privacy act notices (US - FTC implications) |
| Reidentification risk | | No efforts being made to identify re-tagged proprietary data at the present time | Risk assessment by: Data Provider; covenants imposed on Data Consumers |
| Instrumentation and protocols | | Numerous plug-ins that collect and incorporate added data sources. Limited controls within the system itself to maintain calibration. Built in tabular aggregation methods are common. | “Procedure” in some academic paradigms, but considerable domain-specific elaboration may be needed. Sensor provenance, calibration, propagation, audit, aggregation |
| Primary meaning: Digital reproduceability.Secondary: simulation | |  | Full digital forward-construction, backward deconstruction of experiment, data collection, video, other digital artifacts for reproduceability |
|  | | Data Life-cycle | | Current projects are the most sensitive. Projects that have been completed and signed off often have public visibility because of clinical trial requirements and so access to that data is looser once it reaches that stage. | Identify any legal mandates for data “destruction” |
|  | | Disclosure-on-demand | | It is assumed that all data in the system be available for regulatory or audit disclosure in an on-demand basis. This includes audit trail by project, by group, by individual with simple of time based analysis and reporting. | Requirements for mandated or voluntary data disclosure; consumer or data owner;; may be regulation-, court-ordered, veracity motivated |
|  | | Recommended data security / privacy levels | | Data security requirements are usually driven by the individual pharmaceutical company with levels ranging from most secure (internal systems not on the internet) through hosted cloud based environments. | For template, see [HL7 Privacy Segmentation for Privacy](http://wiki.siframework.org/Data+Segmentation+for+Privacy+Paper) |
|  | | Dependency Analytics | |  | Measures in place to assure data integrity and regulatory compliance |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | 1. Federation of queries joining between heterogeneous data sources 2. Federation of queries across geographically distributed homogeneous data sets. 3. Addition of arbitrary heterogeneous data feeds. Examples could include DNA sequencers, chromatographic data, or cellular image analysis. | | |  |
| **More Information (URLs)** | | <http://www.cambridgesoft.com/>  <http://www.perkinelmer.com/technologies/default.xhtml> | | |  |
| **Note:** <additional comments> | | | | |  |