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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Individualized Diabetes Management | |
| **Vertical (area)** | | Healthcare | |
| **Author/Company/Email** | | Peter Li, Ying Ding, Philip Yu, Geoffrey Fox, David Wild at Mayo Clinic, Indiana University, UIC; dingying@indiana.edu | |
| **Actors/Stakeholders and their roles and responsibilities** | | Mayo Clinic + IU/semantic integration of EHR data  UIC/semantic graph mining of EHR data  IU cloud and parallel computing | |
| **Goals** | | Develop advanced graph-based data mining techniques applied to EHR to search for these cohorts and extract their EHR data for outcome evaluation. These methods will push the boundaries of scalability and data mining technologies and advance knowledge and practice in these areas as well as clinical management of complex diseases. | |
| **Use Case Description** | | Diabetes is a growing illness in world population, affecting both developing and developed countries. Current management strategies do not adequately take into account of individual patient profiles, such as co-morbidities and medications, which are common in patients with chronic illnesses. We propose to approach this shortcoming by identifying similar patients from a large Electronic Health Record (EHR) database, i.e. an individualized cohort, and evaluate their respective management outcomes to formulate one best solution suited for a given patient with diabetes.  Project under development as below  **Stage 1**: Use the Semantic Linking for Property Values method to convert an existing data warehouse at Mayo Clinic, called the Enterprise Data Trust (EDT), into RDF triples that enables us to find similar patients much more efficiently through linking of both vocabulary-based and continuous values,  **Stage 2**: Needs efficient parallel retrieval algorithms, suitable for cloud or HPC, using open source Hbase with both indexed and custom search to identify patients of possible interest.  **Stage 3**: The EHR, as an RDF graph, provides a very rich environment for graph pattern mining. Needs new distributed graph mining algorithms to perform pattern analysis and graph indexing technique for pattern searching on RDF triple graphs.  **Stage 4**: Given the size and complexity of graphs, mining subgraph patterns could generate numerous false positives and miss numerous false negatives. Needs robust statistical analysis tools to manage false discovery rate and determine true subgraph significance and validate these through several clinical use cases. | |
| **Current**  **Solutions** | **Compute(System)** | | supercomputers; cloud |
| **Storage** | | HDFS |
| **Networking** | | Varies. Significant I/O intensive processing needed |
| **Software** | | Mayo internal data warehouse called Enterprise Data Trust (EDT) |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | distributed EHR data |
| **Volume (size)** | | The Mayo Clinic EHR dataset is a very large dataset containing over 5 million patients with thousands of properties each and many more that are derived from primary values. |
| **Velocity**  **(e.g. real time)** | | not real-time but updated periodically |
| **Variety**  **(multiple datasets, mashup)** | | Structured data, a patient has controlled vocabulary (CV) property values (demographics, diagnostic codes, medications, procedures, etc.) and continuous property values (lab tests, medication amounts, vitals, etc.). The number of property values could range from less than 100 (new patient) to more than 100,000 (long term patient) with typical patients composed of 100 CV values and 1000 continuous values. Most values are time based, i.e. a timestamp is recorded with the value at the time of observation. |
| **Variability (rate of change)** | | Data will be updated or added during each patient visit. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Data are annotated based on domain ontologies or taxonomies. Semantics of data can vary from labs to labs. |
| **Visualization** | | no visualization |
| **Data Quality** | | Provenance is important to trace the origins of the data and data quality |
| **Data Types** | | text, and Continuous Numerical values |
| **Data Analytics** | | Integrating data into semantic graph, using graph traverse to replace SQL join. Developing semantic graph mining algorithms to identify graph patterns, index graph, and search graph. |
| **Big Data Specific Challenges (Gaps)** | | For individualized cohort, we will effectively be building a datamart for each patient since the critical properties and indices will be specific to each patient. Due to the number of patients, this becomes an impractical approach. Fundamentally, the paradigm changes from relational row-column lookup to semantic graph traversal. | |
| **Big Data Specific Challenges in Mobility** | | Physicians and patient may need access to this data on mobile platforms | |
| **Security & Privacy**  **Requirements** | | Health records or clinical research databases must be kept secure/private. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Data integration: continuous values, ontological annotation, taxonomy  Graph Search: indexing and searching graph  Validation: Statistical validation | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |