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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Truthy: Information diffusion research from Twitter Data | |
| **Vertical (area)** | | Scientific Research: Complex Networks and Systems research | |
| **Author/Company/Email** | | Filippo Menczer, Indiana University, [fil@indiana.edu](mailto:fil@indiana.edu);  Alessandro Flammini, Indiana University, [aflammin@indiana.edu](mailto:aflammin@indiana.edu);  Emilio Ferrara, Indiana University, [ferrarae@indiana.edu](mailto:ferrara@indiana.edu); | |
| **Actors/Stakeholders and their roles and responsibilities** | | Research funded by NFS, DARPA, and McDonnel Foundation. | |
| **Goals** | | Understanding how communication spreads on socio-technical networks. Detecting potentially harmful information spread at the early stage (e.g., deceiving messages, orchestrated campaigns, untrustworthy information, etc.) | |
| **Use Case Description** | | (1) Acquisition and storage of a large volume of continuous streaming data from Twitter (~100 million messages per day, ~500GB data/day increasing over time); (2) near real-time analysis of such data, for anomaly detection, stream clustering, signal classification and online-learning; (3) data retrieval, big data visualization, data-interactive Web interfaces, public API for data querying. | |
| **Current**  **Solutions** | **Compute(System)** | | Current: in-house cluster hosted by Indiana University. Critical requirement: large cluster for data storage, manipulation, querying and analysis. |
| **Storage** | | Current: Raw data stored in large compressed flat files, since August 2010. Need to move towards Hadoop/IndexedHBase & HDFS distributed storage. Redis as a in-memory database as a buffer for real-time analysis. |
| **Networking** | | 10GB/Infiniband required. |
| **Software** | | Hadoop, Hive, Redis for data management.  Python/SciPy/NumPy/MPI for data analysis. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed – with replication/redundancy |
| **Volume (size)** | | ~30TB/year compressed data |
| **Velocity (e.g. real time)** | | Near real-time data storage, querying & analysis |
| **Variety (multiple datasets, mashup)** | | Data schema provided by social media data source. Currently using Twitter only. We plan to expand incorporating Google+, Facebook |
| **Variability (rate of change)** | | Continuous real-time data-stream incoming from each source. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | 99.99% uptime required for real-time data acquisition. Service outages might corrupt data integrity and significance. |
| **Visualization** | | Information diffusion, clustering, and dynamic network visualization capabilities already exist. |
| **Data Quality (syntax)** | | Data structured in standardized formats, the overall quality is extremely high. We generate aggregated statistics; expand the features set, etc., generating high-quality derived data. |
| **Data Types** | | Fully-structured data (JSON format) enriched with users meta-data, geo-locations, etc. |
| **Data Analytics** | | **Stream clustering**: data are aggregated according to topics, meta-data and additional features, using ad hoc online clustering algorithms. **Classification**: using multi-dimensional time series to generate, network features, users, geographical, content features, etc., we classify information produced on the platform. **Anomaly detection**: real-time identification of anomalous events (e.g., induced by exogenous factors). **Online learning**: applying machine learning/deep learning methods to real-time information diffusion patterns analysis, users profiling, etc. |
| **Big Data Specific Challenges (Gaps)** | | Dealing with real-time analysis of large volume of data. Providing a scalable infrastructure to allocate resources, storage space, etc. on-demand if required by increasing data volume over time. | |
| **Big Data Specific Challenges in Mobility** | | Implementing low-level data storage infrastructure features to guarantee efficient, mobile access to data. | |
| **Security & Privacy**  **Requirements** | | Twitter publicly releases data collected by our platform. Although, data-sources incorporate user meta-data (in general, not sufficient to uniquely identify individuals) therefore some policy for data storage security and privacy protection must be implemented. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Definition of high-level data schema to incorporate multiple data-sources providing similarly structured data. | |
| **More Information (URLs)** | | <http://truthy.indiana.edu/>  <http://cnets.indiana.edu/groups/nan/truthy>  <http://cnets.indiana.edu/groups/nan/despic> | |
| **Note:** <additional comments> | | | |