**Current Draft:**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Intelligence Data Processing and Analysis | |
| **Vertical (area)** | | Defense (Intelligence) | |
| **Author/Company/Email** | | David Boyd/Data Tactics/dboyd@data-tactics.com | |
| **Actors/Stakeholders and their roles and responsibilities** | | Senior Civilian/Military Leadership  Field Commanders  Intelligence Analysts  Warfighters | |
| **Goals** | | 1. Provide automated alerts to Analysts, Warfighters, Commanders, and Leadership based on incoming intelligence data. 2. Allow Intelligence Analysts to identify in Intelligence data    1. Relationships between entities (people, organizations, places, equipment)    2. Trends in sentiment or intent for either general population or leadership group (state, non-state actors).    3. Location of and possibly timing of hostile actions (including implantation of IEDs).    4. Track the location and actions of (potentially) hostile actors 3. Ability to reason against and derive knowledge from diverse, disconnected, and frequently unstructured (e.g. text) data sources. 4. Ability to process data close to the point of collection and allow data to be shared easily to/from individual soldiers, forward deployed units, and senior leadership in garrison. | |
| **Use Case Description** | | 1. Ingest/accept data from a wide range of sensors and sources across intelligence disciplines (IMINT, MASINT, GEOINT, HUMINT, SIGINT, OSINT, etc.) 2. Process, transform, or align date from disparate sources in disparate formats into a unified data space to permit:    1. Search    2. Reasoning    3. Comparison 3. Provide alerts to users of significant changes in the state of monitored entities or significant activity within an area. 4. Provide connectivity to the edge for the Warfighter (in this case the edge would go as far as a single soldier on dismounted patrol) | |
| **Current**  **Solutions** | **Compute(System)** | | Fixed and deployed computing clusters ranging from 1000s of nodes to 10s of nodes. |
| **Storage** | | 10s of Terrabytes to 100s of Petabytes for edge and fixed site clusters. Dismounted soldiers would have at most 1-100s of Gigabytes (mostly single digit handheld data storage sizes). |
| **Networking** | | Networking with-in and between in garrison fixed sites is robust. Connectivity to forward edge is limited and often characterized by high latency and packet loss. Remote comms might be Satellite based (high latency) or even limited to RF Line of sight radio. |
| **Software** | | Currently baseline leverages:   1. Hadoop 2. Accumulo (Big Table) 3. Solr 4. NLP (several variants) 5. Puppet (for deployment and security) 6. Storm 7. Custom applications and visualization tools |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Very distributed |
| **Volume (size)** | | Some IMINT sensors can produce over a petabyte of data in the space of hours. Other data is as small as infrequent sensor activations or text messages. |
| **Velocity**  **(e.g. real time)** | | Much sensor data is real time (Full motion video, SIGINT) other is less real time. The critical aspect is to be able ingest, process, and disseminate alerts in NRT. |
| **Variety**  **(multiple datasets, mashup)** | | Everything from text files, raw media, imagery, video, audio, electronic data, human generated data. |
| **Variability (rate of change)** | | While sensor interface formats tend to be stable, most other data is uncontrolled and may be in any format. Much of the data is unstructured. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Data provenance (e.g. tracking of all transfers and transformations) must be tracked over the life of the data.  Determining the veracity of “soft” data sources (generally human generated) is a critical requirement. |
| **Visualization** | | Primary visualizations will be Geospatial overlays and network diagrams. Volume amounts might be millions of points on the map and thousands of nodes in the network diagram. |
| **Data Quality (syntax)** | | Data Quality for sensor generated data is generally known (image quality, sig/noise) and good.  Unstructured or “captured” data quality varies significantly and frequently cannot be controlled. |
| **Data Types** | | Imagery, Video, Text, Digital documents of all types, Audio, Digital signal data. |
| **Data Analytics** | | 1. NRT Alerts based on patterns and baseline changes. 2. Link Analysis 3. Geospatial Analysis 4. Text Analytics (sentiment, entity extraction, etc.) |
| **Big Data Specific Challenges (Gaps)** | | 1. Big (or even moderate size data) over tactical networks 2. Data currently exists in disparate silos which must be accessible through a semantically integrated data space. 3. Most critical data is either unstructured or imagery/video which requires significant processing to extract entities and information. | |
| **Big Data Specific Challenges in Mobility** | | The outputs of this analysis and information must be transmitted to or accessed by the dismounted forward soldier. | |
| **Security & Privacy**  **Requirements** | | Foremost. Data must be protected against:   1. Unauthorized access or disclosure 2. Tampering | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Wide variety of data types, sources, structures, and quality which will span domains and requires integrated search and reasoning. | |
| **More Information (URLs)** | | <http://www.afcea-aberdeen.org/files/presentations/AFCEAAberdeen_DCGSA_COLWells_PS.pdf>  http://stids.c4i.gmu.edu/papers/STIDSPapers/STIDS2012\_T14\_SmithEtAl\_HorizontalIntegrationOfWarfighterIntel.pdf  <http://stids.c4i.gmu.edu/STIDS2011/papers/STIDS2011_CR_T1_SalmenEtAl.pdf>  <http://www.youtube.com/watch?v=l4Qii7T8zeg>  <http://dcgsa.apg.army.mil/> | |
| **Note:** <additional comments> | | | |