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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Statistical Survey Response Improvement (Adaptive Design) | |
| **Vertical (area)** | | Government Statistical Logistics | |
| **Author/Company/Email** | | Cavan Capps: U.S. Census Bureau/cavan.paul.capps@census.gov | |
| **Actors/Stakeholders and their roles and responsibilities** | | U.S. statistical agencies are charged to be the leading authoritative sources about the nation’s people and economy, while honoring privacy and rigorously protecting confidentiality. This is done by working with states, local governments and other government agencies. | |
| **Goals** | | To use advanced methods, that are open and scientifically objective, the statistical agencies endeavor to improve the quality, the specificity and the timeliness of statistics provided while reducing operational costs and maintaining the confidentiality of those measured. | |
| **Use Case Description** | | Survey costs are increasing as survey response declines. The goal of this work is to use advanced “recommendation system techniques” using data mashed up from several sources and historical survey para-data to drive operational processes in an effort to increase quality and reduce the cost of field surveys. | |
| **Current**  **Solutions** | **Compute(System)** | | Linux systems |
| **Storage** | | SAN and Direct Storage |
| **Networking** | | Fiber, 10 gigabit Ethernet, Infiniband 40 gigabit. |
| **Software** | | Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Survey data, other government administrative data, geographical positioning data from various sources. |
| **Volume (size)** | | For this particular class of operational problem approximately one petabyte. |
| **Velocity**  **(e.g. real time)** | | Varies, paradata from field data streamed continuously, during the decennial census approximately 150 million records transmitted. |
| **Variety**  **(multiple datasets, mashup)** | | Data is typically defined strings and numerical fields. Data can be from multiple datasets mashed together for analytical use. |
| **Variability (rate of change)** | | Varies depending on surveys in the field at a given time. High rate of velocity during a decennial census. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Data must have high veracity and systems must be very robust. The semantic integrity of conceptual metadata concerning what exactly is measured and the resulting limits of inference remain a challenge |
| **Visualization** | | Data visualization is useful for data review, operational activity and general analysis. It continues to evolve. |
| **Data Quality (syntax)** | | Data quality should be high and statistically checked for accuracy and reliability throughout the collection process. |
| **Data Types** | | Pre-defined ASCII strings and numerical data |
| **Data Analytics** | | Analytics are required for recommendation systems, continued monitoring and general survey improvement. |
| **Big Data Specific Challenges (Gaps)** | | Improving recommendation systems that reduce costs and improve quality while providing confidentiality safeguards that are reliable and publically auditable. | |
| **Big Data Specific Challenges in Mobility** | | Mobile access is important. | |
| **Security & Privacy**  **Requirements** | | All data must be both confidential and secure. All processes must be auditable for security and confidentiality as required by various legal statutes. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Recommender systems have features in common to e-commerce like Amazon, Netflix, UPS etc. | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |

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