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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | NIST Information Access Division analytic technology performance measurement, evaluations, and standards | |
| **Vertical (area)** | | Analytic technology performance measurement and standards for government, industry, and academic stakeholders | |
| **Author/Company/Email** | | John Garofolo (john.garofolo@nist.gov) | |
| **Actors/Stakeholders and their roles and responsibilities** | | NIST developers of measurement methods, data contributors, analytic algorithm developers, users of analytic technologies for unstructured, semi-structured data, and heterogeneous data across all sectors. | |
| **Goals** | | Accelerate the development of advanced analytic technologies for unstructured, semi-structured, and heterogeneous data through performance measurement and standards. Focus communities of interest on analytic technology challenges of importance, create consensus-driven measurement metrics and methods for performance evaluation, evaluate the performance of the performance metrics and methods via community-wide evaluations which foster knowledge exchange and accelerate progress, and build consensus towards widely-accepted standards for performance measurement. | |
| **Use Case Description** | | Develop performance metrics, measurement methods, and community evaluations to ground and accelerate the development of advanced analytic technologies in the areas of speech and language processing, video and multimedia processing, biometric image processing, and heterogeneous data processing as well as the interaction of analytics with users. Typically employ one of two processing models: 1) Push test data out to test participants and analyze the output of participant systems, 2) Push algorithm test harness interfaces out to participants and bring in their algorithms and test them on internal computing clusters. Developing approaches to support scalable Cloud-based developmental testing. Also perform usability and utility testing on systems with users in the loop. | |
| **Current**  **Solutions** | **Compute(System)** | | Linux and OS-10 clusters; distributed computing with stakeholder collaborations; specialized image processing architectures. |
| **Storage** | | RAID arrays, and distribute data on 1-2TB drives, and occasionally FTP. Distributed data distribution with stakeholder collaborations. |
| **Networking** | | Fiber channel disk storage, Gigabit Ethernet for system-system communication, general intra- and Internet resources within NIST and shared networking resources with its stakeholders. |
| **Software** | | PERL, Python, C/C++, Matlab, R development tools. Create ground-up test and measurement applications. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Large annotated corpora of unstructured/semi-structured text, audio, video, images, multimedia, and heterogeneous collections of the above including ground truth annotations for training, developmental testing, and summative evaluations. |
| **Volume (size)** | | The test corpora exceed 900M Web pages occupying 30 TB of storage, 100M tweets, 100M ground-truthed biometric images, several hundred thousand partially ground-truthed video clips, and terabytes of smaller fully ground-truthed test collections. Even larger data collections are being planned for future evaluations of analytics involving multiple data streams and very heterogeneous data. |
| **Velocity**  **(e.g. real time)** | | Most legacy evaluations are focused on retrospective analytics. Newer evaluations are focusing on simulations of real-time analytic challenges from multiple data streams. |
| **Variety**  **(multiple datasets, mashup)** | | The test collections span a wide variety of analytic application types including textual search/extraction, machine translation, speech recognition, image and voice biometrics, object and person recognition and tracking, document analysis, human-computer dialogue, and multimedia search/extraction. Future test collections will include mixed type data and applications. |
| **Variability (rate of change)** | | Evaluation of tradeoffs between accuracy and data rates as well as variable numbers of data streams and variable stream quality. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | The creation and measurement of the uncertainty associated with the ground-truthing process – especially when humans are involved – is challenging. The manual ground-truthing processes that have been used in the past are not scalable. Performance measurement of complex analytics must include measurement of intrinsic uncertainty as well as ground truthing error to be useful. |
| **Visualization** | | Visualization of analytic technology performance results and diagnostics including significance and various forms of uncertainty. Evaluation of analytic presentation methods to users for usability, utility, efficiency, and accuracy. |
| **Data Quality (syntax)** | | The performance of analytic technologies is highly impacted by the quality of the data they are employed against with regard to a variety of domain- and application-specific variables. Quantifying these variables is a challenging research task in itself. Mixed sources of data and performance measurement of analytic flows pose even greater challenges with regard to data quality. |
| **Data Types** | | Unstructured and semi-structured text, still images, video, audio, multimedia (audio+video). |
| **Data Analytics** | | Information extraction, filtering, search, and summarization; image and voice biometrics; speech recognition and understanding; machine translation; video person/object detection and tracking; event detection; imagery/document matching; novelty detection; a variety of structural/semantic/temporal analytics and many subtypes of the above. |
| **Big Data Specific Challenges (Gaps)** | | Scaling ground-truthing to larger data, intrinsic and annotation uncertainty measurement, performance measurement for incompletely annotated data, measuring analytic performance for heterogeneous data and analytic flows involving users. | |
| **Big Data Specific Challenges in Mobility** | | Moving training, development, and test data to evaluation participants or moving evaluation participants’ analytic algorithms to computational testbeds for performance assessment. Providing developmental tools and data. Supporting agile developmental testing approaches. | |
| **Security & Privacy**  **Requirements** | | Analytic algorithms working with written language, speech, human imagery, etc. must generally be tested against real or realistic data. It’s extremely challenging to engineer artificial data that sufficiently captures the variability of real data involving humans. Engineered data may provide artificial challenges that may be directly or indirectly modeled by analytic algorithms and result in overstated performance. The advancement of analytic technologies themselves is increasing privacy sensitivities. Future performance testing methods will need to isolate analytic technology algorithms from the data the algorithms are tested against. Advanced architectures are needed to support security requirements for protecting sensitive data while enabling meaningful developmental performance evaluation. Shared evaluation testbeds must protect the intellectual property of analytic algorithm developers. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Scalability of analytic technology performance testing methods, source data creation, and ground truthing; approaches and architectures supporting developmental testing; protecting intellectual property of analytic algorithms and PII and other personal information in test data; measurement of uncertainty using partially-annotated data; composing test data with regard to qualities impacting performance and estimating test set difficulty; evaluating complex analytic flows involving multiple analytics, data types, and user interactions; multiple heterogeneous data streams and massive numbers of streams; mixtures of structured, semi-structured, and unstructured data sources; agile scalable developmental testing approaches and mechanisms. | |
| **More Information (URLs)** | | www.nist.gov/itl/iad/ | |
| **Note:** <additional comments> | | | |