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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey | |
| **Vertical (area)** | | Scientific Research: Astronomy | |
| **Author/Company/Email** | | S. G. Djorgovski / Caltech / george@astro.caltech.edu | |
| **Actors/Stakeholders and their roles and responsibilities** | | The survey team: data processing, quality control, analysis and interpretation, publishing, and archiving.  Collaborators: a number of research groups world-wide: further work on data analysis and interpretation, follow-up observations, and publishing.  User community: all of the above, plus the astronomical community world-wide: further work on data analysis and interpretation, follow-up observations, and publishing. | |
| **Goals** | | The survey explores the variable universe in the visible light regime, on time scales ranging from minutes to years, by searching for variable and transient sources. It discovers a broad variety of astrophysical objects and phenomena, including various types of cosmic explosions (e.g., Supernovae), variable stars, phenomena associated with accretion to massive black holes (active galactic nuclei) and their relativistic jets, high proper motion stars, etc. | |
| **Use Case Description** | | The data are collected from 3 telescopes (2 in Arizona and 1 in Australia), with additional ones expected in the near future (in Chile). The original motivation is a search for near-Earth (NEO) and potential planetary hazard (PHO) asteroids, funded by NASA, and conducted by a group at the Lunar and Planetary Laboratory (LPL) at the Univ. of Arizona (UA); that is the Catalina Sky Survey proper (CSS). The data stream is shared by the CRTS for the purposes for exploration of the variable universe, beyond the Solar system, lead by the Caltech group. Approximately 83% of the entire sky is being surveyed through multiple passes (crowded regions near the Galactic plane, and small areas near the celestial poles are excluded).  The data are preprocessed at the telescope, and transferred to LPL/UA, and hence to Caltech, for further analysis, distribution, and archiving. The data are processed in real time, and detected transient events are published electronically through a variety of dissemination mechanisms, with no proprietary period (CRTS has a completely open data policy).  Further data analysis includes automated and semi-automated classification of the detected transient events, additional observations using other telescopes, scientific interpretation, and publishing. In this process, it makes a heavy use of the archival data from a wide variety of geographically distributed resources connected through the Virtual Observatory (VO) framework.  Light curves (flux histories) are accumulated for ~ 500 million sources detected in the survey, each with a few hundred data points on average, spanning up to 8 years, and growing. These are served to the community from the archives at Caltech, and shortly from IUCAA, India. This is an unprecedented data set for the exploration of time domain in astronomy, in terms of the temporal and area coverage and depth.  CRTS is a scientific and methodological testbed and precursor of the grander surveys to come, notably the Large Synoptic Survey Telescope (LSST), expected to operate in 2020’s. | |
| **Current**  **Solutions** | **Compute(System)** | | Instrument and data processing computers: a number of desktop and small server class machines, although more powerful machinery is needed for some data analysis tasks.  This is not so much a computationally-intensive project, but rather a data-handling-intensive one. |
| **Storage** | | Several multi-TB / tens of TB servers. |
| **Networking** | | Standard inter-university internet connections. |
| **Software** | | Custom data processing pipeline and data analysis software, operating under Linux. Some archives on Windows machines, running a MS SQL server databases. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed:   1. Survey data from 3 (soon more?) telescopes 2. Archival data from a variety of resources connected through the VO framework 3. Follow-up observations from separate telescopes |
| **Volume (size)** | | The survey generates up to ~ 0.1 TB per clear night; ~ 100 TB in current data holdings. Follow-up observational data amount to no more than a few % of that.  Archival data in external (VO-connected) archives are in PBs, but only a minor fraction is used. |
| **Velocity**  **(e.g. real time)** | | Up to ~ 0.1 TB / night of the raw survey data. |
| **Variety**  **(multiple datasets, mashup)** | | The primary survey data in the form of images, processed to catalogs of sources (db tables), and time series for individual objects (light curves).  Follow-up observations consist of images and spectra.  Archival data from the VO data grid include all of the above, from a wide variety of sources and different wavelengths. |
| **Variability (rate of change)** | | Daily data traffic fluctuates from ~ 0.01 to ~ 0.1 TB / day, not including major data transfers between the principal archives (Caltech, UA, and IUCAA). |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | A variety of automated and human inspection quality control mechanisms is implemented at all stages of the process. |
| **Visualization** | | Standard image display and data plotting packages are used. We are exploring visualization mechanisms for highly dimensional data parameter spaces. |
| **Data Quality (syntax)** | | It varies, depending on the observing conditions, and it is evaluated automatically: error bars are estimated for all relevant quantities. |
| **Data Types** | | Images, spectra, time series, catalogs. |
| **Data Analytics** | | A wide variety of the existing astronomical data analysis tools, plus a large amount of custom developed tools and software, some of it a research project in itself. |
| **Big Data Specific Challenges (Gaps)** | | Development of machine learning tools for data exploration, and in particular for an automated, real-time classification of transient events, given the data sparsity and heterogeneity.  Effective visualization of hyper-dimensional parameter spaces is a major challenge for all of us. | |
| **Big Data Specific Challenges in Mobility** | | Not a significant limitation at this time. | |
| **Security & Privacy**  **Requirements** | | None. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | * Real-time processing and analysis of massive data streams from a distributed sensor network (in this case telescopes), with a need to identify, characterize, and respond to the transient events of interest in (near) real time. * Use of highly distributed archival data resources (in this case VO-connected archives) for data analysis and interpretation. * Automated classification given the very sparse and heterogeneous data, dynamically evolving in time as more data come in, and follow-up decision making given limited and sparse resources (in this case follow-up observations with other telescopes). | |
| **More Information (URLs)** | | CRTS survey: <http://crts.caltech.edu>  CSS survey: <http://www.lpl.arizona.edu/css>  For an overview of the classification challenges, see, e.g., <http://arxiv.org/abs/1209.1681>  For a broader context of sky surveys, past, present, and future, see, e.g., the review <http://arxiv.org/abs/1209.1681> | |
| **Note:**  CRTS can be seen as a good precursor to the astronomy’s flagship project, the Large Synoptic Sky Survey (LSST; <http://www.lsst.org>), now under development. Their anticipated data rates (~ 20-30 TB per clear night, tens of PB over the duration of the survey) are directly on the Moore’s law scaling from the current CRTS data rates and volumes, and many technical and methodological issues are very similar.  It is also a good case for real-time data mining and knowledge discovery in massive data streams, with distributed data sources and computational resources. | | | |

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