21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Contribution ID: **364** Type: **oral presentation**

Analysis Traceability and Provenance for HEP

Tuesday, April 14, 2015 3:00 PM (15 minutes)

In complex data analyses it is increasingly important to capture information about the usage of data sets in addition to their preservation over time in order to ensure reproducibility of results, to verify the work of others and to ensure appropriate conditions data have been used for specific analyses. This so-called provenance data in the computer science world is defined as the history or derivation of a data (or process) artifact. Many scientific workflow based studies are beginning to realise the benefit of capturing the provenance of their data and the activities used to process, transform and carry out studies on that data. This is especially true in scientific disciplines where the collection of data through experiment is costly and/or difficult to reproduce and where that data needs to be preserved over time. With the increase in importance of provenance data many different models have emerged to capture provenance such as PROV or the OPM models. However, these are more for interoperability of provenance information and do not focus on the capture of provenance related data. There is a clear and emerging requirement for systems to handle the provenance of data over extended timescales with a emphasis on preserving the analysis procedures themselves and the environment in which the analyses were conducted alongside the processed data sets.

A provenance data system that has been built in house at CERN since early 2000 is called CRISTAL. CRISTAL was used to capture the provenance resulting from the design and construction of the CMS ECAL detector over the period 2002-2010. The CRISTAL Kernel (V3.0) has now been launched as open source under the LGPL (V3) licencing scheme and it is available for use by the wider communities including teams involved in the offline analysis of physics data, whether at CMS or other experiments. In addition, in the EC funded neuGRID and N4U projects the original developers have been using CRISTAL to capture the provenance of analyses for neuroscientists running complex pipelines of algorithms in the study of biomarkers for the onset of Alzheimer's disease. In this paper this application is presented with a focus on how its approach can be customised for use in the high energy physics data analysis community at large. The main focus of this is a set of analysis tools (persistency, browsing/querying, visualising and analysis tracking services) which together with a generic analysis model backend can be used to capture the information required to support complex analyses.
The Analysis Tools comprise the following interfaces :

- **The Analysis Web Serivce** – Which is for advanced users, these users are able to programmatically create analyses on the fly.
- **The Analysis Command-line Interface** – This is for users that are intermediate/advanced. It allows users to create analysis using a shell like interface.
- **The Analysis Web Portlet Interface** – This is for novice users, it is a visual interface which is portlet based and allows users to browse datasets and pipelines. It also allows users to create and deploy their analyses in a visual manner.
- **The Analysis Core** – These are a core set of objects used by the above interfaces. These objects connect directly to a customised analysis aware CRISTAL instance which is provenance aware.

During the provenance capture phase the Analysis tools are able to capture :

- **who** ran an analysis, this is a user name,
- for **what** purpose, what their analysis is supposed to achieve,
- **when** they ran it this is a timestamp which denotes when it started and when it finished,
- **where** it was run this is GRID and Cloud related information,

- **which** datasets and algorithms were used to create and run their analyses,
- **how** it was executed, this more detailed infrastructure information
- and lastly **why** the analysis was run, this is a justification from the user.

Also in this paper, we present the case for using the Analysis Services developed using CRISTAL as an avenue for long timescale data preservation. The tools are able to store the provenance metadata surrounding the analyses in a human readable form (XML). This is a light-weight and queryable manner of storing provenance as well as analysis results. In CRISTAL everything is recorded and nothing is thrown away. So another user would be able to replicate the experiment at a later date and time. Besides reproducibility of experiments, users can also share their experiments with other users using the provenance related information.

The analysis tools run currently in a GRID and a Cloud infrastructure. As well as collecting analysis provenance information, they are able to provenance of the infrastructure. There is strong novelty in this work which facilitates allows more precise reproducibility of experiments. This information is known as *infrastructure provenance*. It is currently being collected in the course of the N4U project. This infrastructure provenance can also be applied to HEP to aid in the reproducibility of results. For example, if performance is a factor it can be sent to the same compute node.

Concerning the future of the analysis tools there is currently an emerging standard known as PROV which is used for *provenance interoperability*. In the near future, the analysis provenance information that we have collected with be exported to PROV. This work has already begun, we are looking for mapping patterns to aid in our cause. The reasoning for this is so that people can study and use provenance information in a standard and commonly understood format. This will also allow users to publish the provenance generated from their analyses onto the ever growing linked data cloud as well.

Primary author: SHAMDASANI, Jetendr (University of the West of England (GB))

Co-authors: BRANSON, Andrew (University of the West of England (GB)); Prof. MCCLATCHEY, Richard (University of the West of England (GB)); KOVACS, Zsolt (University of the West of England (GB))

Presenter: SHAMDASANI, Jetendr (University of the West of England (GB))

Session Classification: Track 5 Session

Track Classification: Track5: Computing activities and Computing models