Avoiding the Tower of Babel syndrome: An integrated issue-based quality assurance system [vers. 4]

G. Van Buren*, M. DePhillips, L. Hajdu, J. Lauret, BNL, Upton, NY 11973-5000, USA
L. Ray, UT, Austin, TX 78712, USA

Abstract

Samples of data acquired by the STAR Experiment at RHIC are examined at various stages of processing for quality assurance (QA) purposes. As STAR continues to mature and utilize new hardware and software, it remains imperative to the experiment to work cohesively to insure the quality of STAR data so that the collaboration may continue to produce many new physics results in an efficient and timely manner. Correlating detector subsystem expert-specific information, shift crew reports, online QA, and offline reconstruction information would pose a daunting challenge to any collaboration. Presentation of QA results in an organized and integrated fashion has proven vital to establishing robust communication of issues to both operators and users. We will present in this paper the integrated QA system developed to achieve these goals within the STAR Experiment, from detector operations to data production and analysis.

GOALS

The primary goal of QA in the STAR Experiment [1] is to ensure effective and efficient progress towards achieving physics-quality data. Achieving this goal requires identification of issues in the data, reporting, communicating, and archiving those issues, and providing access to the QA and issue reports. Issue identification is performed using browsers which provide detailed information at the earliest opportunity possible, implying a time-staged approach to QA, while electronic logs archive and simplify report lookup in a ”one-stop-shopping” manner.

The first stage of QA is performed online at the experiment's control room. Its goals are to provide immediate assessment of hardware functionality and performance directly to the experiment operators (shift crew). Experience has taught us that it also serves well as a trouble-shooting aid for other run-time issues. The second stage of QA uses an initial ”fast offline” reconstruction of the data to give short-time-scale assessment on the general validity of the data (i.e. whether it can be reconstructed), and feedback to the shift crew on effectiveness of operations. Final production of the data also receives a round of QA to ensure use of proper reconstruction code and validity of calibrations.

* gene@bnl.gov

ONLINE QA

As is typical in many experiments STAR provides online QA browsing through histograms of raw (unreconstructed) information from detectors in the data stream. The information comes from the STAR data acquisition system (DAQ) [2] directly to an online event pool, wherein full events are available, but only for a small fraction of the events collected by the experiment. Histograms are continually updated from the event pool data, and are saved to disk and then reset between runs. Further, PDF renditions of the histograms are saved in a database pending approval of the shift crew.

Identification of issues is aided by comparison to a reference histogram set from data which subsystem experts have declared to be valid. Descriptions of these issues are immediately entered into the STAR Electronic ShiftLog (ESL), and can be classified by relevant subsystems (e.g. detectors, DAQ, trigger, or general) and run numbers. ESL entries also allow for attachments, such as screen captures of histograms, to facilitate information flow. Once created, an online issue remains open until its status is specifically changed. Further details of the ESL will be described later in this paper.

OFFLINE QA

To standardize offline QA, histograms of reconstructed data (e.g. hits, tracks, vertices) are generated during the reconstruction process in the same manner for the fast offline and production stages. This enables the offline QA to aggregate information from entire DAQ files. Analysis is then performed using PostScript or PDF documents produced on-demand for only a fraction of these files, where the selection strategy is meant to sample at least some data from each run within the time it takes a user to perform the analysis.

For the final reconstruction the generated histograms are stored for any potential subsequent analyses, while fast offline only maintains recently generated information and is run for only a small fraction of the data files. To expedite turn-around, fast offline processing is done without waiting for final calibrations. Analyzers must be aware of this, but can then also provide feedback on possible calibration issues.

Reporting of offline QA issues uses a separate dedicated web-based mechanism which can be used from anywhere over the internet. Users establish sessions when working on reports, which allows flexibility to return to previously

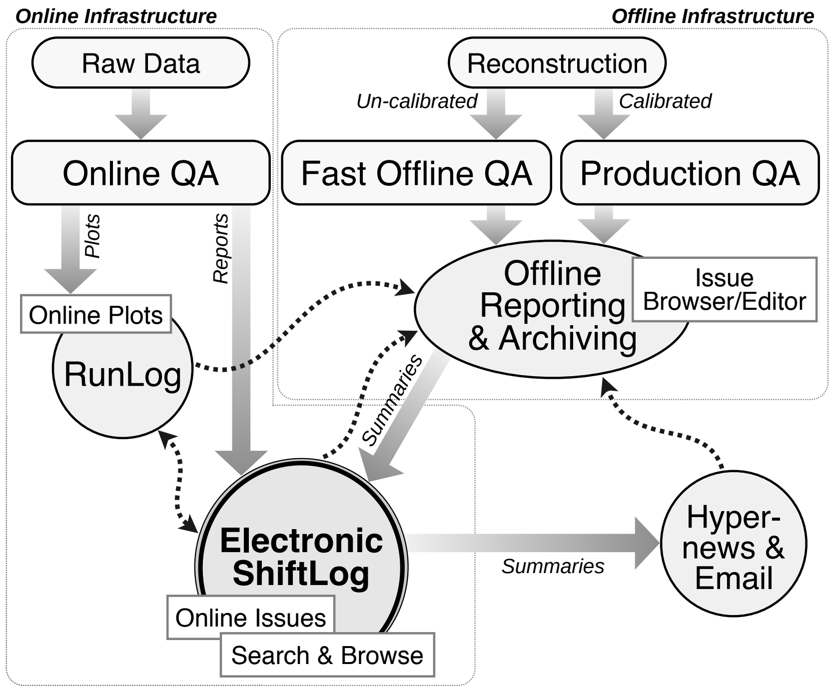

Figure 1: Schematic diagram of QA information flow. Large shaded arrows represent data and analysis results, while dotted arrows show interconnectivity of information access through hyperlinks.

started work and permits multiple simultaneous users. Reports maintain distinction between data types, including the aforementioned fast offline and production types, as well as code sanity tests or simulation productions. Users create "data entries" for each histogram file examined, and all entries for their QA shift are compiled into one full shift report.

Each offline QA data entry is constructed as an issue-centric object. A header with details of data identity is accompanied by a body containing an enumeration of associated active issues and a section for additional comments.

The offline issues themselves are entities maintained through an accompanying browser/editor, though they may be a continuation of online issues and may refer to such. Each issue has an identification number, a brief description, one or more data type tags (e.g. fast offline or simulation), and a full description which can be edited for corrections and appended with time-stamped notes and comments. Issues may persist for several runs or appear sporadically, becoming inactive for some datasets, and then becoming active again, and may eventually become closed/resolved.

QA users are provided a list of existing issues and their status when creating a data entry: by default, an issue which was previously active stays active, and one which was previously inactive stays inactive. The user can then decide whether an issue is indeed active for a particular dataset and should become associated with the data entry for it. The user may even re-open an issue which had been considered closed/resolved.

When a QA user submits their report at the conclusion of a shift, the full report is archived on a webserver where it is indexed by filing time and by the runs examined for simplified lookup. A summary of the report is also generated which provides a hyperlink to the full report, and highlights the brief descriptions of only those issues whose status has changed since the last submitted report, each hyperlinked to their full details in the issue browser/editor. Our experience in interacting with people operating the experiment has shown that concise summaries have a greater chance of being digested and addressed than full reports.

The fast offline QA summaries are submitted to the ESL and are subsequently emailed to the operations shift crew for immediate perusal. Other offline QA summaries are directed to a hypernews forum on productions [3]. All summaries also indicate the runs and files examined in the report.

THE ELECTRONIC SHIFTLOG

The information provided by the QA process in STAR must serve the needs of a variety of users, including experiment operators, subsystem coordinators, data processing administrators, and data analyzers. A single, unified starting point for access to that information is potentially the best means to ensure not only open availability to all, but more importantly a healthy channel for information flow and communication: all parties have an interest in keeping that channel functional.

Beyond the information submitted by those performing QA (direct reports from online QA, and summaries from fast offline QA), the ESL contains summaries of each operation shift of the experiment, notes on individual subsystem issues, and the brief descriptions of purpose for each run. The ESL entries for runs are hyperlinked to the STAR Run-Log system which provides a wealth of details regarding each particular run and subsequent links to both online QA plots and offline QA reports related to that run. Each ESL entry includes an author and a time-stamp tag. Combining all of these assets allows the ESL to serve as a central information hub of STAR QA.

To realize this, the ESL is designed to be accessed in a variety of ways. ESL entries can be browsed by time period, subsystem, QA entries, or in several combinations of those categories. Full text searches permit retrieval of entries tied to particular runs or specific topics. These basic but flexible methods organize information in a manner proficient for navigation and understanding. For example, the task of studying the history of an issue is simplified by presenting time-stamped information over its lifetime and permitting study of other concurrent and possibly related events and issues. Figure 1 illustrates this flow of information to and from the ESL as the central hub of QA.

SUMMARY

Effective progress within a complicated and diverse organization such as the STAR Experiment requires efficient identification and communication of issues that arise. The STAR QA effort attempts to ensure this progress through a staged process of issue identification which promotes increased degree of inspection as the extent of data reconstruction increases. Issues are managed as entities through all stages to enhance traceability, and communication of those issues is integrated into a unified central hub of information pertinent to QA.

This effort has benefited from several years of operation during which suggestions and constructive criticisms have helped shape QA to improve its proficiency. The STAR QA program is meeting the needs of a mature experiment today and is helping STAR reach its own goals.

ACKNOWLEDGEMENTS

This work was supported by the Offices of Nuclear Physics and High Energy Physics within the U.S. Department of Energy Office of Science, and the U.S. National Science Foundation. We would like to thank the organizers of the CHEP 2006 conference.

REFERENCES

[1] K.H. Ackerman et al., Nucl. Instr. and Meth. A499 (2003) 624; http://www.star.bnl.gov.

[2] J.M. Landgraf et al., Nucl. Instr. and Meth. A499 (2003) 762.

[3] http://www.hypernews.org.