Offline access to archived protoDUNE DAQ configuration and run history information

Kurt Biery, Andrew Norman (16-Mar-2018)

**Introduction**

As part of the protoDUNE Single-Phase data acquisition (DAQ) system, there are facilities to manage the configuration of the detector electronics and other parts of the DAQ system. This configuration management system handles both the sets of *Available Configurations* and the sets of *Archived Configurations* that are needed by the experiment. Available Configurations comprise those sets of parameters and information that operators running the detectors, may choose from when preparing to take data. Archived Configurations are the historical sets of parameters that were actually used in the acquisition of each data-taking run. The configuration management system also stores *Run History* information such as the start time, end time, and number of events in each run.

The Archived Configuration information and Run History information are need by collaborators working on both online and offline data analysis. There is currently not a well defined interface that has been adopted by the ProtoDUNE experiment to provide to provide access to this data from the offline environments, specifically a storage and retrieval system that can scale to the concurrency levels that are required to analyze/process the ProtoDUNE data.

The intention of this note is to provide an overview and background regarding the systems as they exist currently, and what functionality or planning needs to be added to adapt these systems to use in the offline analysis environments, without directly exposing the critical DAQ and data taking components of the system to resource contention that might occur from access or traffic from offline analysis users.

Appendices at the end of this document provide definitions of terms that are used in this document and describe some of the known changes that are needed with the current online configuration management system in order to implement the online components of what is described.

**Current DAQ Infrastructure**

In the online environment, ProtoDUNE uses the *artdaq-database* software package for configuration management, and a MongoDB based storage technology for storing configurations. There are command-line and graphical tools available for viewing and modifying configuration parameters in these contexts.

Within the MongoDB storage, the configuration information is stored in the form of JSON documents. The members of the protoDUNE DAQ team typically interact with FHiCL files on disk that have been exported from the DB or are being prepared to be imported into the DB, and the *artdaq-database* tools handle the translation between JSON and FHiCL.

The *artdaq-database* package allows for considerable flexibility in how configuration data is stored, and each experiment can make different choices. The details of how the data for the Available Configurations are stored in the database in protoDUNE is not described in this document, and is not relevant to the need to publish parameters to the offline/analysis environments, but we will describe a little about how the Archived Configuration data is stored.

In contrast, the Archived Configuration data is needed in the offline environment and its storage model is important to the access methods that are used to retrieve it. Currently, each Archived Configuration is identified by two primary keys, the run number and the name of the Run Configuration. The run number is assumed to be a unique integer that is assigned by the DAQ system. The name of the Run Configuration corresponds to the string identifier of the configuration that the operator selected within in the Run Control interface before starting the run. The two keys are combined with a “/” delimiter to form the full identifier for the Archived Configuration (e.g. for run 800, the list of Archived Configurations has a single element: “800/artdaq\_v3\_testing00019”). The inclusion of the configuration names, along with run numbers, in the identifiers for the archived configuration data sets permits the expansion of the types of configuration data sets that are stored/associated with each run number.

The contents of each Archived Configuration are copies of the FHiCL documents that were sent to the *artdaq* processes prior to the start of the run. Each FHiCL document within an Archived Configuration is identified by the process that received it. Currently, these identifying strings are composed from the *artdaq* process type, the hostname where the process was running, and the port on which the process was listening for control commands (e.g. BoardReader-np04-srv-012-8000.fcl). In future evolutions of the software, the identifiers will be switched to names that correspond by the component name (e.g. timing.fcl). This is considered a more useful offline enumeration type. The contents of each FHiCL document includes parameters that are needed by the *artdaq* processes as well as parameters that are programmed into the upstream electronics.

There are tools in the artdaq-database package that provide the ability to export a configuration from the database to files on disk. In the current software design, these tools are the primary way for online users to get access to the Archived Configuration information.

**Support of Offline Access**

Based on extensive experience with other DAQ/Offline interfaces, the follow model is considered to provide adequate support and access to the Archived Configuration data from the offline environment.

1. An offline-accessible copy of the Run History information. Because the underlying organization of the information is both well structured and fairly static with time, there is a desire to provide and present this data in a relational form. This data is often queried to directly select physics/detector data based on combinations of records. Given both its structure and primary uses, the Run History information is well suited to storage in a relational database which is both accessible and scalable to the levels need by the offline grid computing environment.
2. An offline-accessible copy of the Archived Configuration data. Because the underlying organization of the information is already in a well structured, key based format, this information could be usable by offline applications in its native format as a JSON document. The information also easily allows for it to be time indexed, and as such is suitable for storage in an “interval of validity” (validity context) organized database. Moreover, the data itself could be extracted into more traditional relational schema.
3. Software libraries providing both interface layers to interpret the Archived Configuration data and to provide access to its underlying storage. Offline analysis applications require well defined interface which can bridge the access protocols and provide well behaved class objects within the analysis environment. Specifically supporting interpretation of the FHiCL language or database access from the C++ and Python contexts.
4. Replication of online (DAQ) instances of source data, into databases/datasources which service and scale to the offline environments. Enhancements to the artdaq-database tools can be implemented to support this, and there have been some preliminary discussions about techniques for doing this in the past, but nothing has been formalized.

**Appendix A: Terms Used in this Document**

This section defines some of the configuration-related terms that are used in this document and are relevant for protoDUNE DAQ.

* Dataflow Configuration – the parameters that define the layout (or architecture) of the DAQ processes and the technologies that are used to communicate between them. It includes the number and location (host computer) of each type of DAQ process that is being used at any given time. Different Dataflow Configurations are appropriate at different times, based on which subset of the readout electronics is working (or under study), the amount of disk-writing capacity is needed (number of Data Loggers), and other factors.
* Run Configuration – the parameters that prepare the DAQ system and readout electronics for data taking. They include things like threshold settings or channel readout lists.
* Available Configurations – these are sets of configurations (either Dataflow or Run) that are available for use when a DAQ system operator goes through the process of starting up the DAQ system. This term may also be used to describe parameter sets that can be used by software tools to support the creation of a Dataflow or Run configuration. For example, a list of computers that are available to host Data Logger processes could be used by a software tool to present a list to an operator for his/her choice.
* Archived Configurations – these are copies of the configuration data that was used for each data-taking run. They are indexed by run number and a string that identifies the configuration.

It should be noted that there have been discussions about separating the current Run Configuration information into two or more different types. These may be Trigger Configurations or *art* Process Configurations. If and when any new types of configurations are added, they will be added to the data that is stored with each Archived Configuration, and we will need to ensure that suitable access is provided to them.

**Appendix B: Changes Needed in the Online**

The following changes are needed in the online configuration management system to support the availability of configuration data in the offline:

1. Add support for accessing the Run History information.
2. Update the strings that are used to identify FHiCL documents inside Archived Configurations to use component or logical names.
3. Improve the storage of archived Dataflow Configuration information, and provide additional tools to access it, if needed.