Overview of the LHC Data Model
DASPOS Technical Report #1

Peter Ivie, Anna Woodard, Matthias Wolf,
Douglas Thain, Kevin Lannon, Michael Hildreth, and Rob Gardner

June 2014

The DASPOS project is an NSF-funded collaborative research project designed to explore the technical means of preserving the data and software components necessary to reproduce and build upon published scientific results. This technical report series consists of short reports that outline background, case studies, and technical results related to data preservation, with a focus on high energy physics.

Introduction

This technical report provides a high level summary of the Large Hadron Collider (LHC) data model from the perspective of a computer scientist for the purposes of data preservation and continued access. While much of what is described here may be familiar to the high energy physics (HEP) community, this brief summary serves to orient contributors that do not have an HEP background and to highlight the essentials of the data preservation problem.

The Large Hadron Collider (LHC) is the highest energy particle collider ever built. Its construction by the European Organization for Nuclear Research (CERN) was completed in 2008. The goal is to create new understanding of the physics of subatomic particles at very high energies, particularly that of the (now discovered) Higgs Boson. The LHC is a circular particle accelerator with a 27 kilometer circumference housed in a 3.8-meter-wide tunnel buried 50 to 175 meters underground. Two separate **beamlines** in the tunnel allow for particles to be circulated in opposite directions at high energy. The beamlines are filled with particles (a process that may take a few hours) and superconducting magnets are used to control the velocity and concentration of the particles. Once the beamlines are sufficiently filled, intense bunches of protons collide within 4 intersection points (or **collision regions**) in the tunnel [2]. If all goes well, one of these **fills** can last approximately 24 hours.

Massive detectors placed around these intersection points generate data about what happens during the collisions. There are four major detectors -- ATLAS, CMS, ALICE, and LHCb -- and a variety of smaller experiments. Each of the four major detectors is supported by an international organization of scientists. Generally speaking, a single detector can produce over a petabyte of data every second in full operation, but only about 25 petabytes per year get stored for analysis. The LHC undergoes periods of full operation separated by upgrades, adjustments, and maintenance during which no data is

generated. It ran roughly March through November of 2011 and March 2012 through February 2013. It is currently being upgraded to run at higher energy starting in 2015.

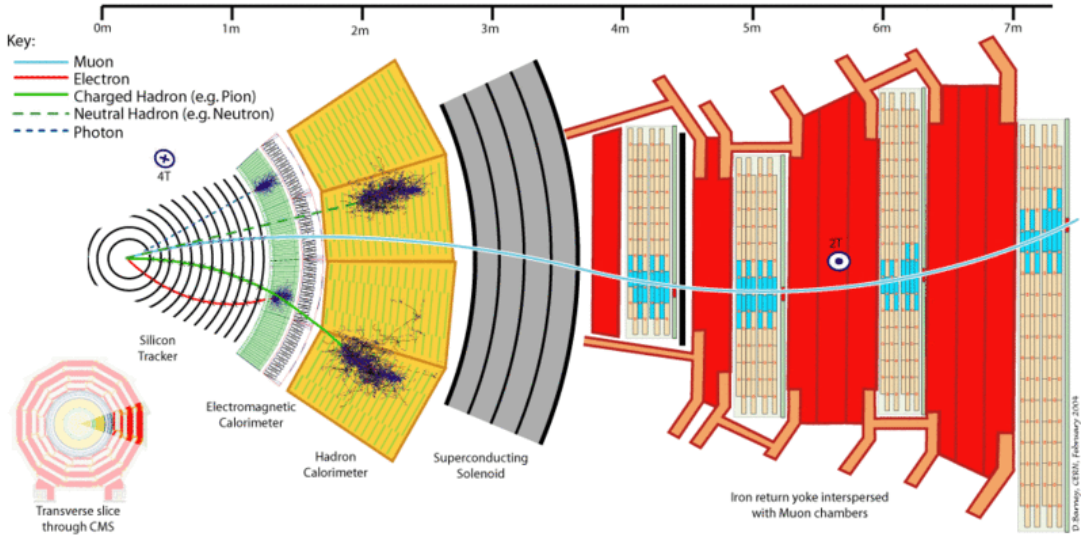
The sheer size of the data has caused the HEP community to develop a variety of technologies to collect, share, and move the data between the various sites involved in the work. However, the exact technologies used to deal with the data have changed year by year, while the fundamental structure has not. Thus, in this paper, we look past the current technologies in use, and focus on the **logical structure** of the data, the **reduction stages**, and the **physical distribution** that is necessary to accommodate the enormous data sizes. While the various experiments differ in scientific goals, they all share a common high level structure for data distribution and analysis.

Logical Structure

The speed and number of particles travelling through the LHC contribute (with other factors) to the **luminosity**, which is a measure of the total particle production rate in the system. At a design luminosity (of 10^{34} $cm^{-2}s^{-1}$) 112 trillion protons of 7 TeV per nucleon, travel the entire ring 11,000 times per second at only 3 meters per second less than the speed of light. The protons are also grouped into 2,808 bunches as an artifact of the mechanism used to accelerate them. The synchrotron is designed to redirect the opposing particle beams so that they will collide with each other. A **crossing** is when a pair of bunches passes through an intersection point. Bunches can collide as frequently as 25 nanoseconds apart. Overall, the system is designed to generate 40 million crossings per second with about 20 collisions at each crossing. This means that to handle the luminosity the LHC is designed to generate, the system that processes the data should be able to handle about 800 million collisions can occur every second.

[Image from http://scienceblogs.com/startswithabang/2009/05/01/the-lhc-black-holes-and-you/]
Caverns around the 4 intersection points contain equipment or detectors designed to observe different properties of the interactions. General purpose particle detection is performed by the "A Toroidal LHC Apparatus" **(**ATLAS) experiment and the "Compact Muon Solenoid" (CMS) experiment. ATLAS and

CMS are very similar in capabilities and purpose, looking for the origins of mass (Higgs boson), extra dimensions, and the nature of dark matter. The other detectors are more specialized. ALICE studies quark-gluon plasma. LHCb searches for antimatter. There are also a few other smaller detectors for very specialized research. Each detector has different components designed to meet their specific goals.



[Image from http://en.wikipedia.org/wiki/File:CMS_Slice.gif]

A detector is made of various different types of sub-detectors layered around the collision region to track each collision and the results. Each detector consists of many individual sensors or channels. The values for each channel are sampled over time at its own appropriate frequency. Each one is designed to detect a different type of behavior. For example, particles passing by the silicon tracker trigger electrical signals that can be used to determine the timing and path a particle travelled. A electromagnetic calorimeter measures the energy of electrons and photons by absorbing them.

Given the sampling rates and the number of channels, a huge amount of data is generated very quickly. In the design of the system it was estimated that it would take about 25 megabytes to store all the details for a single event in ATLAS detector. For 40 million events per second, this would be 1 PB/s. Even after compressing events by doing things like ignoring silent channels (zero suppression), each event would still be about 1-1.6 megabytes for both ATLAS and CMS. This reduces the throughput to ~60 TB/s, but that is still too difficult to manage. The events would get stored on a tape archive system that can handle on the order of 100 MB/s or O(100) events, so 40 million events would need to get reduced down to only O(100) of the most interesting.

L1 Triggers

Even with a large data processing center it would be a challenge to process 60 TB/s of data with this level of complexity. A electronic system (Level 1 or L1 triggers) was designed to watch the stream of

collisions and only output an event if something interesting appeared to occur during a collision.

The L1 triggers output about 100,000 events through per second, or on the order of 100 GB/s. Different triggers watch for different interesting behavior. One of the L1 triggers just picks random collisions at a certain rate and outputs events for those collisions. These random events become part of the minimum bias (MinBias) dataset, and can be used to evaluate whether the L1 trigger is skipping events that it should have passed through.

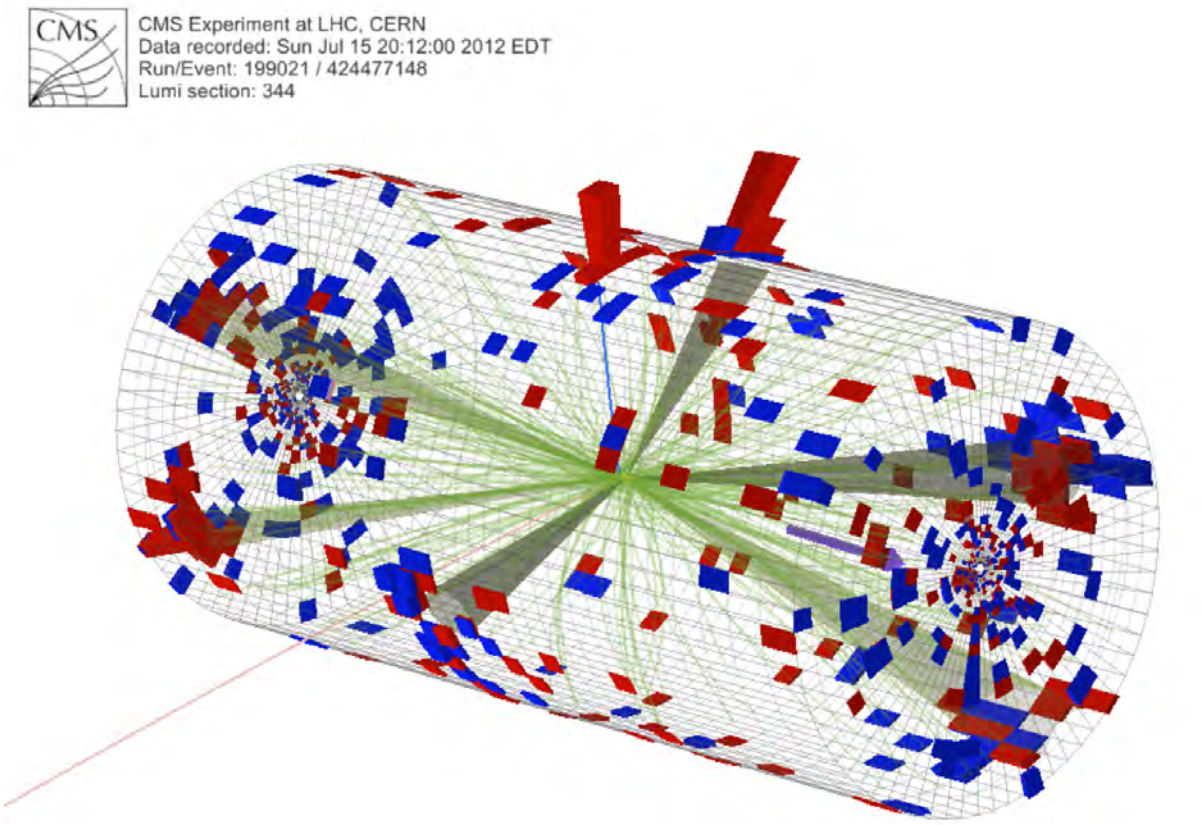
Some sensors take longer than 25 nanoseconds to indicate that something happened. This means another crossing may have occurred before the results of the last crossing are reported. The observed data during collisions is buffered for the L1 electronics, so that when an event is triggered, all relevant data (after and even before the crossing) can be read from the buffer and be used to describe the event.

HLT triggers

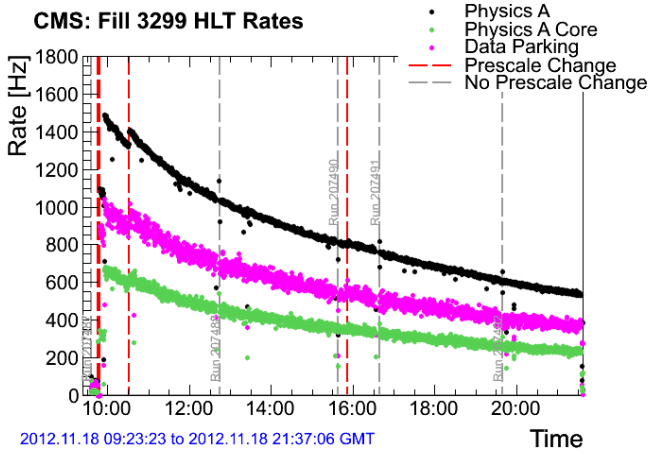
The next stage is a CPU farm of a few thousand nodes called the High Level trigger (HLT), which is used to perform a much more complex analysis of the events. Atlas has L2 and L3 triggers, but at CMS these are both folded together into a single High Level Trigger. The HLT selects about 100 of the most interesting events per second and stores them on tapes for future analysis, or 100MB/s. Events with similar interesting qualities (the same trigger) are stored on tape together as a group. A single event might be interesting for and stored in more than one dataset.

In normal operation, there are a few hundred triggers, or sets of criteria, used to decide whether or not to record a collision event. A single event may satisfy the requirements of more than one trigger. Each trigger is associated with a **primary dataset** according to what type of requirements it imposes. For example, triggers which require the presence of more than one jet are all associated with the MultiJet primary dataset, while triggers which require the presence of two or more electrons are associated with the DoubleElectron dataset. The results are written out to files according to the primary dataset associated with the trigger it passed. There are approximately 20 primary datasets, but in a given analysis, one might examine up to two or three.

For each **fill** of the beamlines, the collisions are divided up into a number of **runs**. Each run is identified by a number independent of the fill when it occurred. Runs are divided up into multiple parts in order to keep file sizes more manageable. These parts are called **lumi sections** and are a period of time short enough that the luminosity can be approximated as constant (about 23 seconds for the CMS). Each event within a lumi section gets an event number. To distinguish one event from all others, the run number, lumi section number and event number are all needed. The following figure shows event number 424,477,148 in lumi section number 344 in run number 199,021.



This next image shows that during a particular fill, the number of interesting events starts out high (left side of the figure). But as collisions occur the number of particles that continue to travel around the rings decreases, which results in fewer and fewer interesting events (right side of the figure).



Data Transformation and Reduction Formats

These events are stored in a format called **RAW**. In practice, a RAW event consumes only about 300 KB (much less than the 1-1.6 MB that was originally anticipated). A single RAW file normally holds many RAW events that exhibit similar behavior (the same trigger). As with any instrument, the raw data produced by the LHC must be prepared before it is suitable as input for analysis tools. As shown in the table below, data flows through multiple stages of conversion, calibration, selection, and reduction until it reaches the hands of an individual researcher with a particular research agenda. Generally speaking, the formats given at the top of the table are used for complete, long-term storage at CERN and other major centers. Proceeding down the table, each format becomes more and more specialized to a particular scientific goals, and is used for smaller amounts of data that are re-processed or re-created more frequently.

Summary of Data Formats

Format	Event Size	Purpose
RAW	~300 KB	Raw sensor state surrounding an event.
RECO	~1 MB	Complete logical reconstruction of an event.
AOD	~ 300 KB	Most commonly used fields from the RECO format.
NTuples	~12 KB	Flattened into selected properties of events.
BEANS		Example of NTuple variant used by Notre Dame CMS group.
Trees		Post-analysis format for visualization.

For example, RAW and RECO data are comprehensive, or in other words, all available information and all events are included in the datasets. AOD data is comprehensive in the number of events, but the information in each event is a subset of what is in a RECO event. The information available in AOD data is designed to be suitable for perhaps 90% of all research needs. The remaining researchers must use the larger RECO files directly for their research. Ntuple data is specialized to a particular analysis area, and even more specialized BEAN (Boson Exploration Analysis Ntuple) files are used by a handful of researchers at Notre Dame interested in studying the production of Higgs Bosons in association with top quarks.

The RAW event data is difficult to understand by itself. It contains information like values for the electric signals in the tracking detector channels. This information can be the basis for drawing conclusions about what was actually happening, such as what particles traveled where. But this is a very complex process, both computationally and intellectually, and is performed only a few times a year. The resulting data is

called **RECO** (RECOnstructed) data. In general, all RAW events are converted to RECO, and each RECO event includes all possible conclusions about the collision based on the RAW data and the best interpretation available when the reconstruction occurred. As experts discover better interpretations of the RAW data, RECO data can be re-reconstructed, but this is only done rarely. Using current hardware, each event takes somewhere on the order of 10 seconds to be processed from RAW to RECO. A single RECO event takes up about 1MB. This is larger than the RAW event, but it is more logical for processing since the data attempts to describe results of the event like the velocity, spin, energy, etc of each particle leaving the interaction site, instead of simply the behavior of the sensors.

The RECO format is very comprehensive and contains information that may not be interesting for a particular line of research. Because there is so much data involved, researchers find it beneficial, at this point, to remove the parts of the RECO data that are not needed for their particular research. This allows them to run analyses faster and with fewer resources. For example, the detailed hit-by-hit resolution for tracking particle trajectories might not be relevant for an analyses, so it could be removed. The result after these removals is called **AOD** data. Each event at this stage is approximately 300 KB. As mentioned above, AOD files are designed to be suitable for 90% of all researchers. However this is just a design target, not a measured result.

Several terms are used in the community to describe the ways in which data is reduced for analysis.

- **Skimming** is the removal of entire events that match certain criteria from a dataset; the result is a dataset with fewer events, but the same level of detail.
- **Slimming** or **thinning** is the removal of a set of objects and/or attributes from all events in a dataset; the result is a dataset with the same number of events, but each event contains less data. (There seems to be some disagreement about the meaning of slimming and thinning, and they are not always used consistently.)

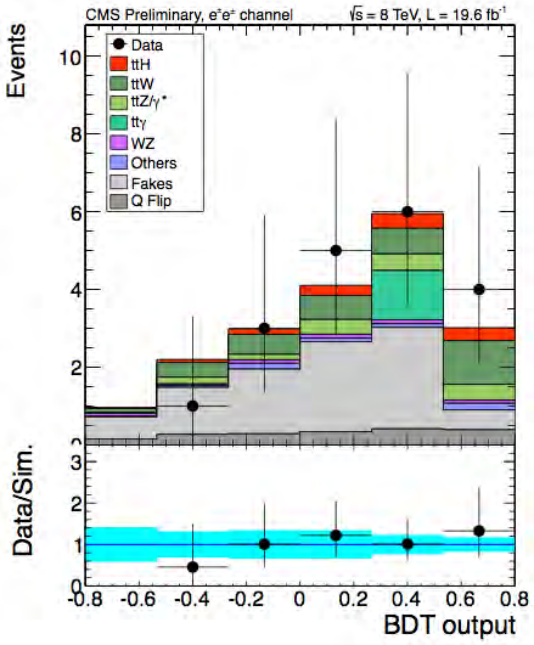
For example, there could be a collection of charged particle trajectories or **tracks** and associated information like quality-of-fit or the number of hits on sensitive detector elements. Slimming/thinning might remove the latter two objects and all but the helix parameters from the tracks, but the number of events would remain the same.

The next step usually involves removing most of the hierarchy from the AOD data in favor of a much more flattened tabular format called NTuples. This could involve thinning and/or slimming as unnecessary data is removed. For example, the detailed track information for all particles could be reduced down to just the tracks for a specific particle. Without the hierarchal structure of AOD, the NTuples can be in any form and the researcher that created them would separately keep track of what each field contains. For at least one researcher in the Notre Dame CMS group, events at this level consumed ~12 KB of disk space.

Not all of the events may be relevant to a particular analysis. Eliminating unnecessary events is known as **skimming**. For example, one could select only those NTuples where total energy is greater than some threshold. In the particular case we looked at, the resulting data was called skimmed BEANs, which are used by the Notre Dame CMS group and some close collaborators. Other researchers would have their own approach in generating data at this level.

The next stage is where most of the analysis occurs. Physicists apply **cuts**, where events not matching some condition are removed, in order to exclude background noise so that only **signals** are left for consideration. Corrections can also be made to the data such as **scaling factors** that adjust data generated by models to more closely match actual results, or re-weighting of events based on conclusions about the accuracy of the detectors when the event occurred.

The data may then be further processed to create graphs depicting behavior or attributes that might be interesting, such as estimates of the level of background noise remaining. In the case we looked at, this was called **tree data**. The following image is a plot of one of the types of variables that might be saved in a tree. In this case the plot shows how closely aligned some generated models are with the actual measured events. The black dots indicate the observed behavior, with vertical error bars. The stacked boxes represent estimates, based on models, of the things that should be contributing to the observed behavior.



[Image from http://cds.cern.ch/record/1604480/files/HIG-13-020-pas.pdf]

Simulated Data

Physics analysis depends on very detailed simulations of physical processes that both create high energy particles and those that describe their interaction with the detector. The simulation begins with a choice of the particle physics process to be simulated. Monte Carlo simulations use theoretical probability distributions describing particle production to generate interactions. Pythia is an example of one of these generators. [2]

The particles that are produced by the generator are propagated through a description of the particle detectors, including detector construction materials and geometry. This is often done using the GEANT software package [3]. The simulated energy loss of these particles is converted into simulated electronic signals which are effectively simulated RAW data. This RAW data is processed through the same reconstruction code as collision data. The subsequent data formats are identical to collision data, except that information about the physics and particle content of the generated interaction is kept for diagnostic purposes.

Physical Distribution

The term Worldwide LHC Computing Grid (WLCG) refers to a conglomerate of computing services revolving around the LHC data. It is divided up into 4 tiers, each with a specific set of services.

Tier 0 (CERN data center): Less than 20% of the computing capacity of the WLCG, yet all LHC data passes through this tier. This Tier provides both raw and reconstructed (at least partially reconstructed) data to computers on Tier 1. When the LHC is not operating, these resources are used for further reconstruction.

Tier 1: These 11 computer centers are responsible to share raw and reconstructed data with the computers on Tier 2. They also perform additional processing on the data and make the results available. They each have a 10 Gb/s fiber with Tier 0.

Tier 2: Approximately 140 universities and other scientific institutes produce and process simulated data. They share data with those on Tier 3.

Tier 3: Individual researchers can store and process data on their individual computer or other local resources such as a department cluster. There is no formal agreement between Tier 3 users and the WLCG.

Generally speaking, as data flows through the physical layers, it is reduced and transformed to

correspond to the interests of that particular data center. For example, Tier 0 stores all of the RAW data, Tier 1 centers collectively produce all the RECO data, AOD data is used for distribution in Tier 2 and Tier 3, and custom data formats like BEAN would typically be used in Tier 3.

Implications for Data Preservation

The complexity of the LHC data model was born of the necessity to deliver large amounts of data in an efficient manner to a widely distributed community of users. However, it presents challenges for identifying the relevant parameters, analysis paths and software used to derive meaning from the data. The scale of the data prevents simple archiving and/or sharing of a researcher's workspace, and the scale of varied research interests and expertise compounds the issue further.

One objective of the DASPOS project will be to develop a logical framework that expresses in a compact form how derived data was created, using the RAW experimental (or simulated) data as the ultimate source. In prose, a simple specification might be "Select all proton-proton events with energy greater than threshold E and reduce to ntuple format." Of course a wide variety of tools for doing individual transformations exist, but a comprehensive data framework is missing.

Doing so will have several benefits. First, it reduces the total amount of data that must be preserved, so that widely shared data can be preserved with few copies, and unique derived data can be automatically from shared data. Second, it preserves the *semantics* of a dataset, which makes explicit the many assumptions, opinions, and objectives that appear to be necessary at every level of analysis. By preserving the semantics behind a dataset, other researchers can easily reproduce a given result, and then precisely modify it to yield a new but comparable result. Defining a common framework for abstracting transformations applied to a common base of data seems to be a promising way to reuse efforts within the high energy physics community.

[1] Coll, C. M. S. "CMS physics technical design report, Vol. 1." CERN/LHCC 1 (2006): 2006.

[2] T. Sjo¨strand et al., Computer Phys. Commun. 135, 238 (2001).

[3] S. Agostinelli et al. (GEANT4), Nucl. Instrum. Methods A506, 250 (2003).

Acknowledgements

This work was supported in part by NSF grant PHY-1247316.