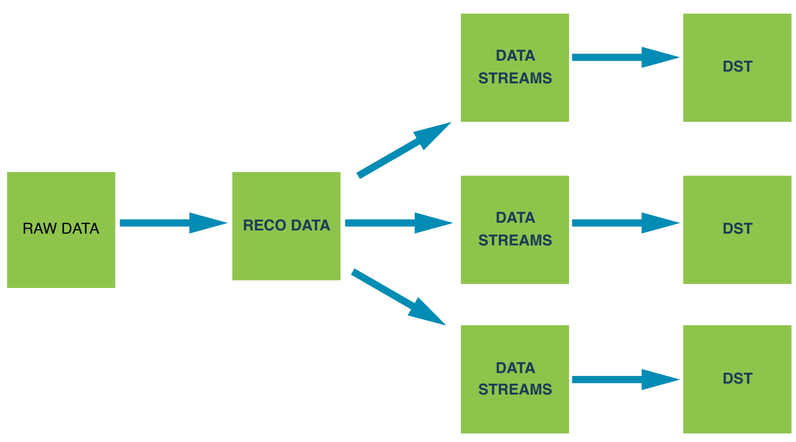
Dynamic and on demand data streams

Matteo Duranti[a], Valerio Formato[a], Valerio Vagelli[a,b]
[a] Istituto Nazionale di Fisica Nucleare, Perugia (IT)
[b] Department of Physics and Geology, University of Perugia (IT)

Replicability and efficiency of data processing on the same data samples are a major challenge for the analysis of data produced by HEP experiments. High-level data analyzed by end-users are typically produced as a subset of the whole experiment data sample to study interesting selection of data (streams). For standard applications, streams may be eventually copied from servers and analyzed on local computing centers or user machine clients. The creation of streams as copy of a subset of the original data results in redundant information stored in filesystems and may be not efficient: if the definition of streams changes, it may force a reprocessing of the low-level files with consequent impact on the data analysis efficiency.

We propose an approach based on a database of lookup tables intended for dynamic and on-demand definition of data streams. This enables the end-users, as the data analysis strategy evolves, to explore different definitions of streams with minimal cost in computing resources. We also present a prototype demonstration application of this database for the analysis of the AMS-02 experiment data

HEP data formats and data streams



Production of data streams - Static approach

- Process reconstructed data
- Select interesting events
- Copy subset of events to data stream

Drawbacks:

- Changes in the definition of the stream or in the selection force a new data reprocessing
- Data processing can be resource demanding
- Data streams are copied to analysis centers and the information is duplicated

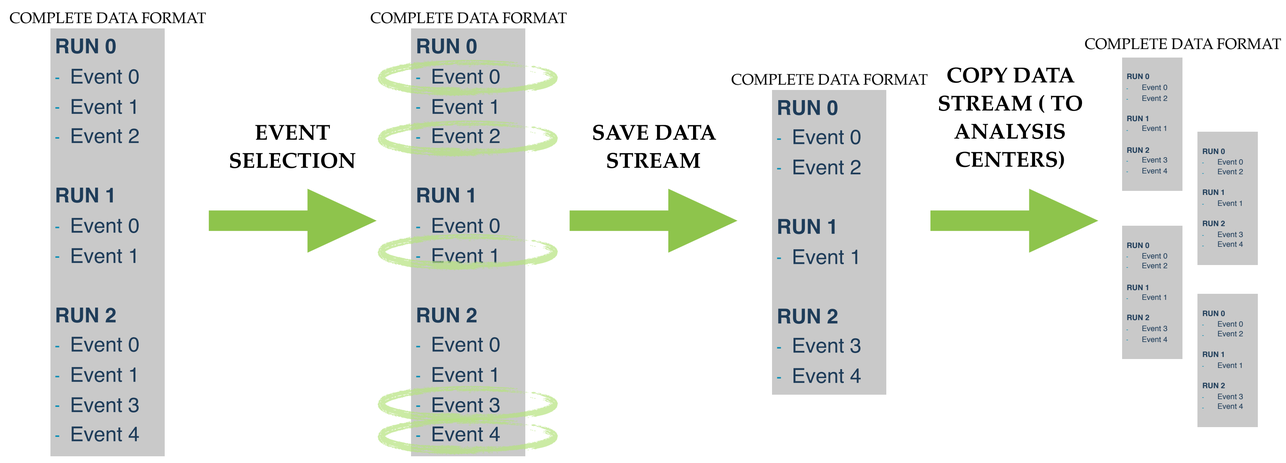
Typical process for production of end-user analysis data

RAW data
- Low level instrument information, no physics objects

Reconstructed data
- High level physics objects
- Organized in *events* as part of *runs*

Data streams
- Subset of events or runs selected for a specific analysis

DST data
- Simplified data structure with a selection of variables and high level information



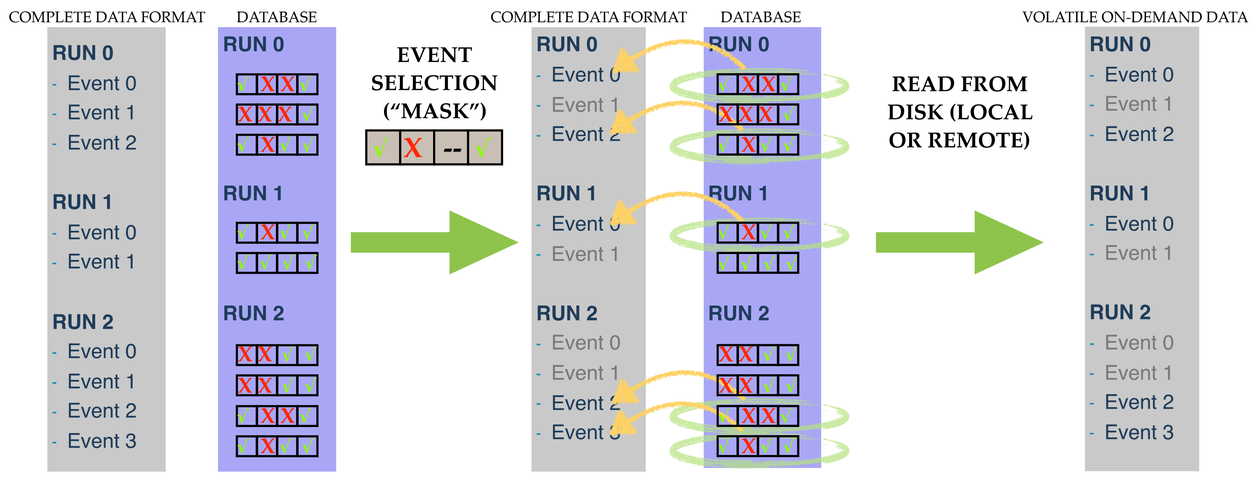
The PreselectionDB approach

Creation of a database for dynamic definition of data streams

- Access to reconstructed data information and test all pre-defined selection criteria
- Store the information on a database indexed by run & event number
- Index the database with the data ntuples

Definition of stream

- Inspect the database for events that fulfill the selection criteria
- Interesting events are automatically selected
- Non interesting events are not access from filesystem at all



All selection criteria are tested against the whole dataset only during the creation of the PreselectionDB
During the data analysis, the stream is identified dynamically: the PreselectionDB provides the user a unique identifier of the events fulfilling the selection. **The user retrieves on demand from file only the selected events**

Less I/O demands
Less disk space
Flexible and dynamic definition of data streams

Use case: PreselectionDB applied to AMS-02 data analysis

Concept tested for analysis of AMS-02 data
AMS-02 data are stored in a *AMSRoot* ROOT TTree, events are uniquely identified by Run + Event numbers
PreselectionDB implemented in the form of a ROOT TTree indexed to AMSRoot via Run & Event

Creation of the PreselectionDB

- Boolean selection criteria are tested for all events and stored as bits in a word
- Boolean selection criteria are tested for all *Particle* objects in each event and stored as bits in a word
- The value of interesting continuous variables is discretized and stored in a dedicated word

Selection of events for data analysis

- The user defines which selections should be checked and assigns bitwise words (i.e. "masks")
- The user inputs the selection mask to the PreselctionDB
- The PreselectionDB provides the Run and Event numbers of all events fulfilling the user requirements
- The user loops on the whole datasets and retrieves (**locally** or **remotely**) from file only the interesting events

Applications tested and applied successfully:
- Dynamic selection of data from the complete AMS-02 dataset as on-demand streams
- Istantaneous comparison of data selections
- Calculation of efficiencies or ratios

Event selection criteria
[0] = "Event-Reconstruction-Success"
[1] = "Science-Run"
[2] = "No-HW-Failures"
[3] = "Good-Livetime"
[4] = "Phys-Trigger"
[5] = ……

Particle selection criteria
[0] = "Particle-Reconstruction-Success"
[1] = "Downgoing-Particle"
[2] = "Inner-Tracker-Acceptance"
[3] = "Particle-Above-GM-cutoff"
[4] = "Good-Track-Reconstruction"
[5] = …..

Example of PreselectionDB entry

```
UInt run = 1234567890

UInt event = 123456

ULong event_word = 0x3E9 = (0011 1110 1001)2
```
12 selections tested for the event

```
vector<ULong> particle_word = <0x29, 0x71> = <(0010 1001)2, (0111 0001)2>
```
2 particles in the events, 8 selections tested for each particles

```
vector<ULong> particle_energy = <0xA1, 0x32> = <(1010 0001)2, (0011 0010)2>
```
2 particles in the event, energy value discretized in a 8 bin interval