

# Dataset

Collection of event data (EDO's)

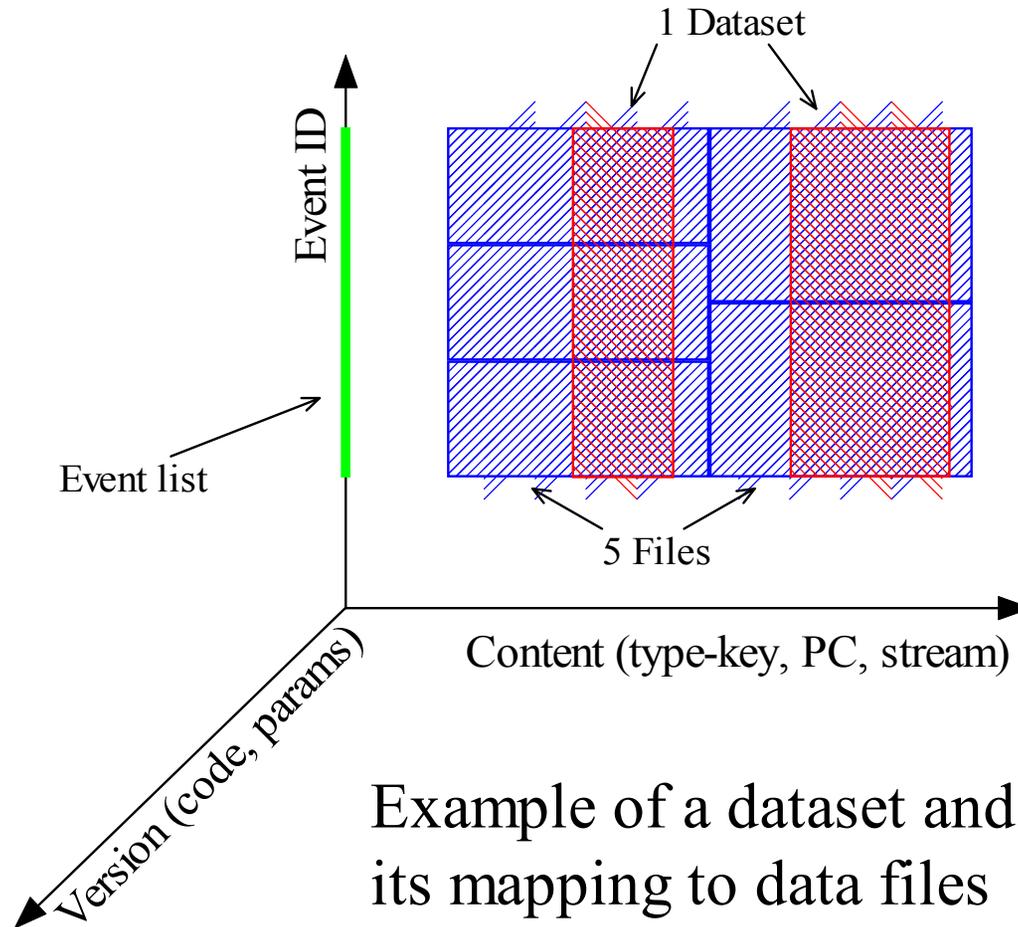
A dataset provides:

- List of event (beam crossing) ID's
- Same content (track, jets, tag) for each event
- Means to access the data
  - Logical file collection(s)

Datasets are composite:

- Merge same content and different events
- Merge same events and different content
- Event or content selection

# Dataset example



# DIAL

(Distributed Interactive Analysis of Large datasets)

## User

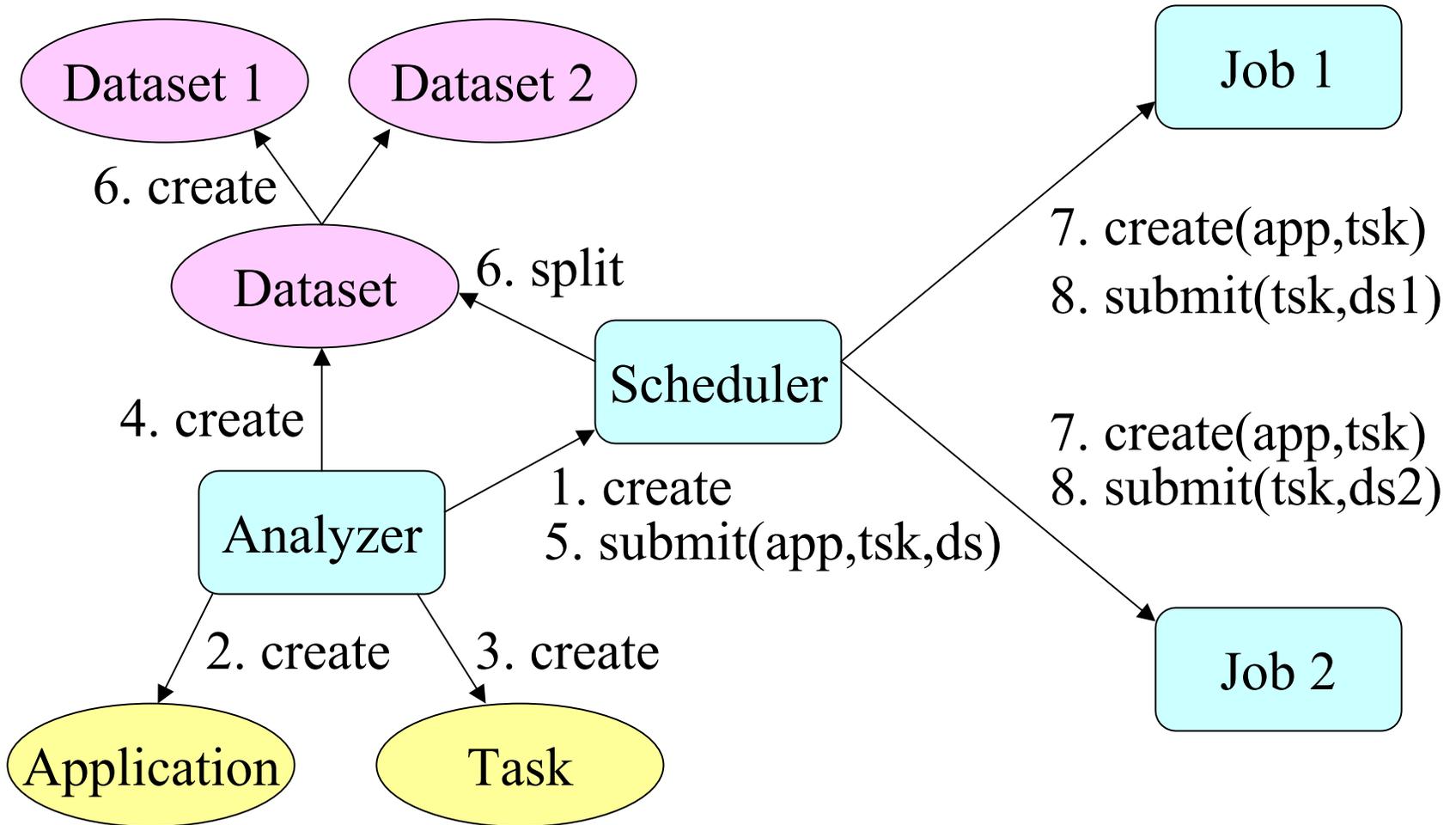
- Works inside a C++ analysis framework
  - e.g. ROOT
- Creates a job description
  - Specifies application (e.g. athena)
  - Creates task
    - > result definition (e.g. an empty histogram, or event list)
    - > code to apply to each event and fill result
  - Specifies a dataset (which event data to process)
- Submits job description to a scheduler for interactive analysis

# DIAL (cont)

## Scheduler

- Divides dataset by event into sub-datasets
- Finds CPU's close to the event data in each sub-dataset
- Creates a job to run application and apply task to each sub-dataset
- Combines the results for each job and returns combined result to the user
- May return job status or partial results
- GRID awareness is inside scheduler

# DIAL



# Datasets and DIAL in DC1

Both can easily be ready for DC1 phase 2

- prototypes for phase 1

Datasets can provide means to

- catalog DC1 data
  - Extension of existing “datasets”
  - Give users easy access to particular data samples
- identify sub-samples (event selections)

DIAL can provide

- distributed interactive analysis of large samples
  - Event selections, filling of histograms or ntuples