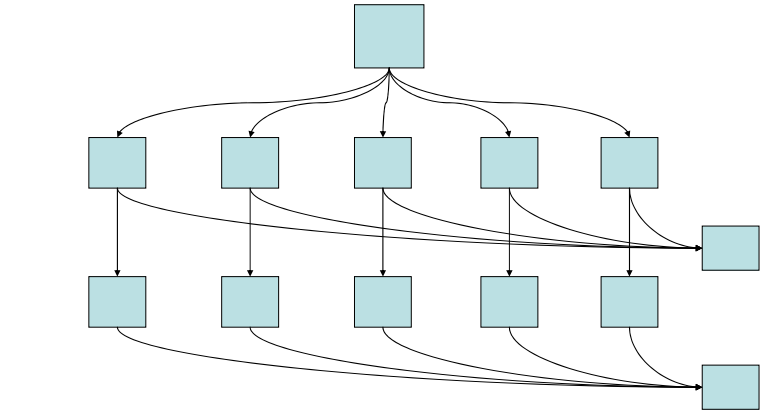
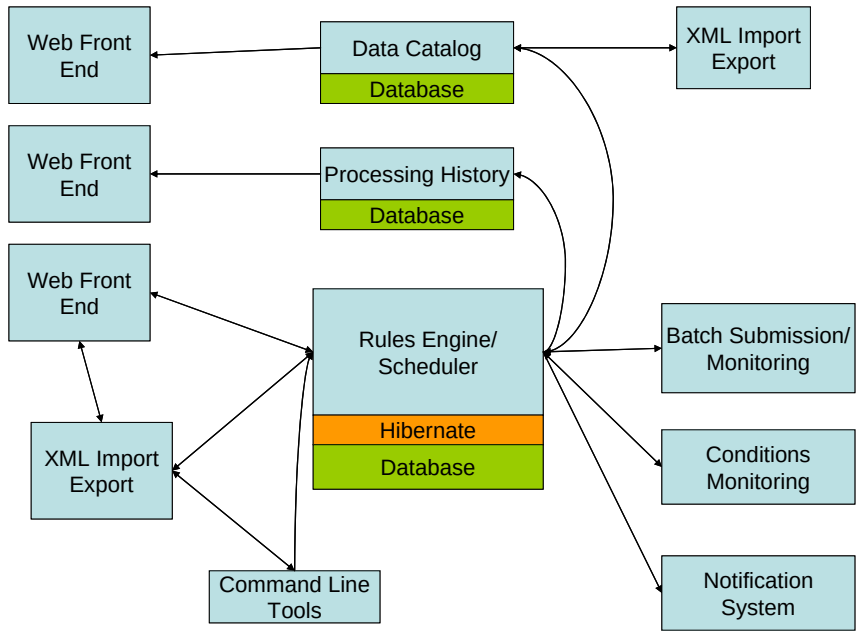
Data Handling Developers Workshop Highlights, March 2005

- Pipeline
  - Status/Plans for current Pipeline
  - Future Pipeline
  - Data Catalog
- Data Server
  - Short term plans
  - Medium term plans
- Full agenda, presentations, discussion:
  - http://confluence.slac.stanford.edu/display/ds/Developers+Workshop

Next Generation Pipeline

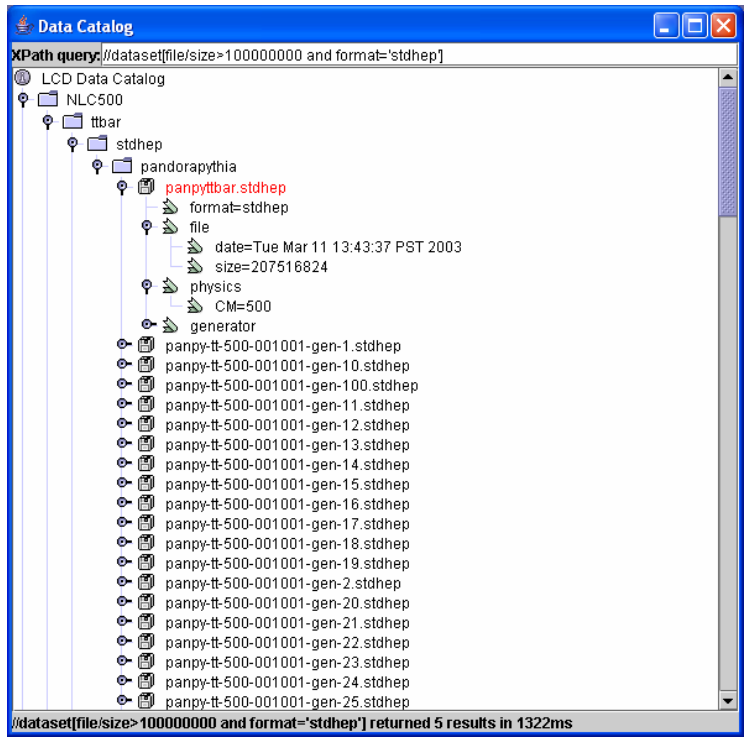
- Plan to remove linear workflow limitations of current pipeline
  - Arbitrary graphs of tasks
  - Flexible “rules” based scheduler
  - Support user data
    - E.g. runtime specification of contents of tar files
- Split implementation into smaller components
  - Work first on interfaces between components
  - Subsequently development and test independently





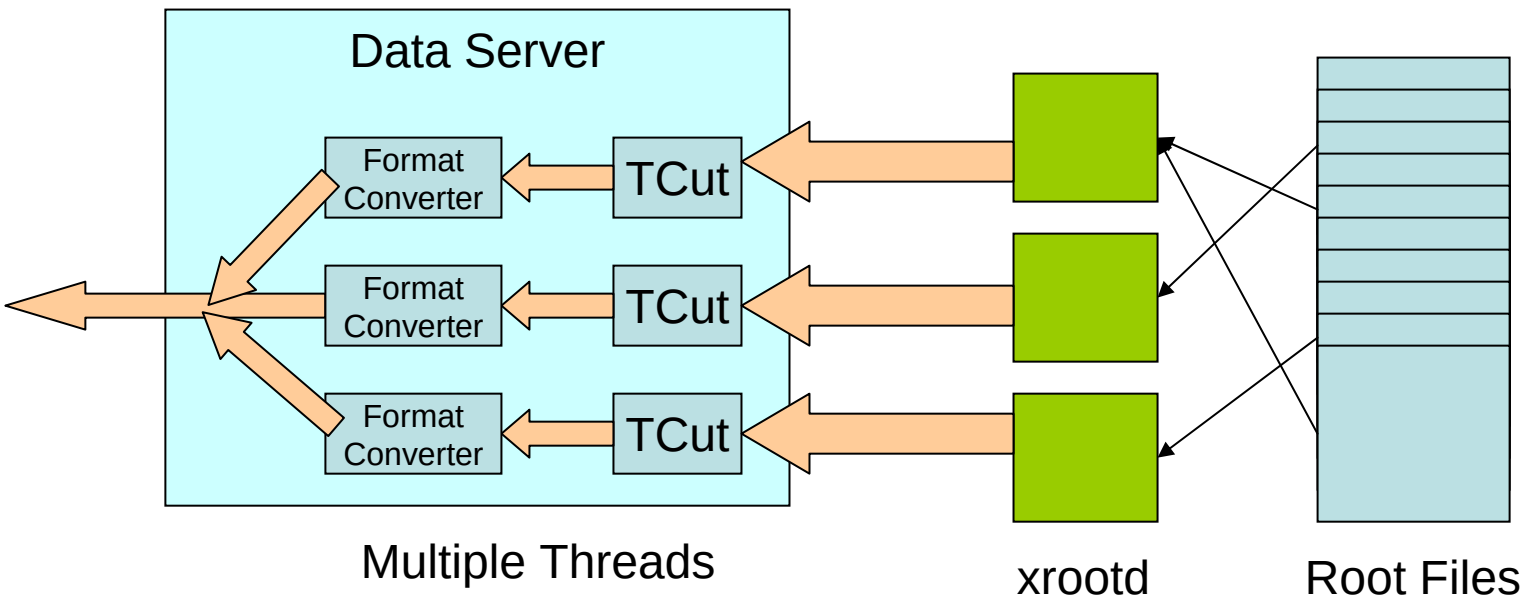
Data Catalog

- Currently data catalog is part of pipeline
- In new architecture data catalog would be separate component
  - Allow arbitrary meta-data to be associated with datasets
  - Support import (export) of data via XML
    - For example MC generated outside pipeline
  - Web based catalog browse/search
- Need to solicit requirements from users
- Need to understand how would interface to existing systems
  - e.g ELogbook



Data Server

- In a few weeks plan to have initial version of data server for DC2
  - Will support access to tuples by run-number, TCut
  - Will support access to full root tree by run/event number
  - Data delivered (after some delay) via FTP
- Incrementally improve over next few months
  - More flexible event selection
    - Fast selection based on photon position, time, energy, quality
    - Enhanced web interface
    - Web based event display (WIRED)
  - Real time data streaming



Tasks

- Data Server
  - Simple NTuple server
    - Tony, Richard, Tom, Navid – 2 weeks
  - Extend to handle full Root trees
    - Tom - ?
  - Web interface
    - Jean Paul - End-to-end 4 weeks?
  - WIRED event display
    - Mark
  - Study larger datasets in MySql, Oracle 10g
  - Study streaming xrootd/etc
- Current Pipeline
  - New batch submission mechanism (Dan, Navid)
  - Archiving (Dan, Navid)'
  - Ongoing web interface improvements (Matt)
  - Improved logging (Igor)
- Next generation Pipeline
  - Implement batch submission interface (also needed ASAP for Data Server)
    - (Based on Navids batch submission system)
  - Write extended schema for next generation pipeline import/export
  - Data Catalog
    - Poll user community for data catalog requirements, using CS11 meta-data catalog as example
    - Investigate XPath support in Oracle
    - Develop XML schema for import/export from data catalog
    - Implement Data Catalog + web front-end
  - Conditions/Resource checker
    - Probably just extend Navid's disk space etc checker to also put data into database
  - Investigate use of Rules engine (JESS and similar)
    - Create simple pipeline mockup using JESS