A Flexible Distributed Event-level Metadata System for ATLAS

David Malon*, Jack Cranshaw, Kristo Karr (Argonne),

Julius Hrivnac, Arthur Schaffer (LAL Orsay)

CHEP’ 06, Mumbai

13-17 February 2005

Acknowledgment

- ✴ Thanks to Caitriana Nicholson (Glasgow) for presenting on our behalf
- ✴ Some of us would rather be in Mumbai!

Event-level metadata in the ATLAS computing model

- ATLAS Computing Model proposes an event-level metadata system--a “tag” database--for rapid and efficient event selection
- Budget allows for approximately 1 kilobyte of “payload” metadata per event, so storage requirements are at the scale of a small number of terabytes
  - Should be widely replicable in principle--all Tier 1s and most Tier 2s should be able to accommodate it if its unique resource demands are not onerous

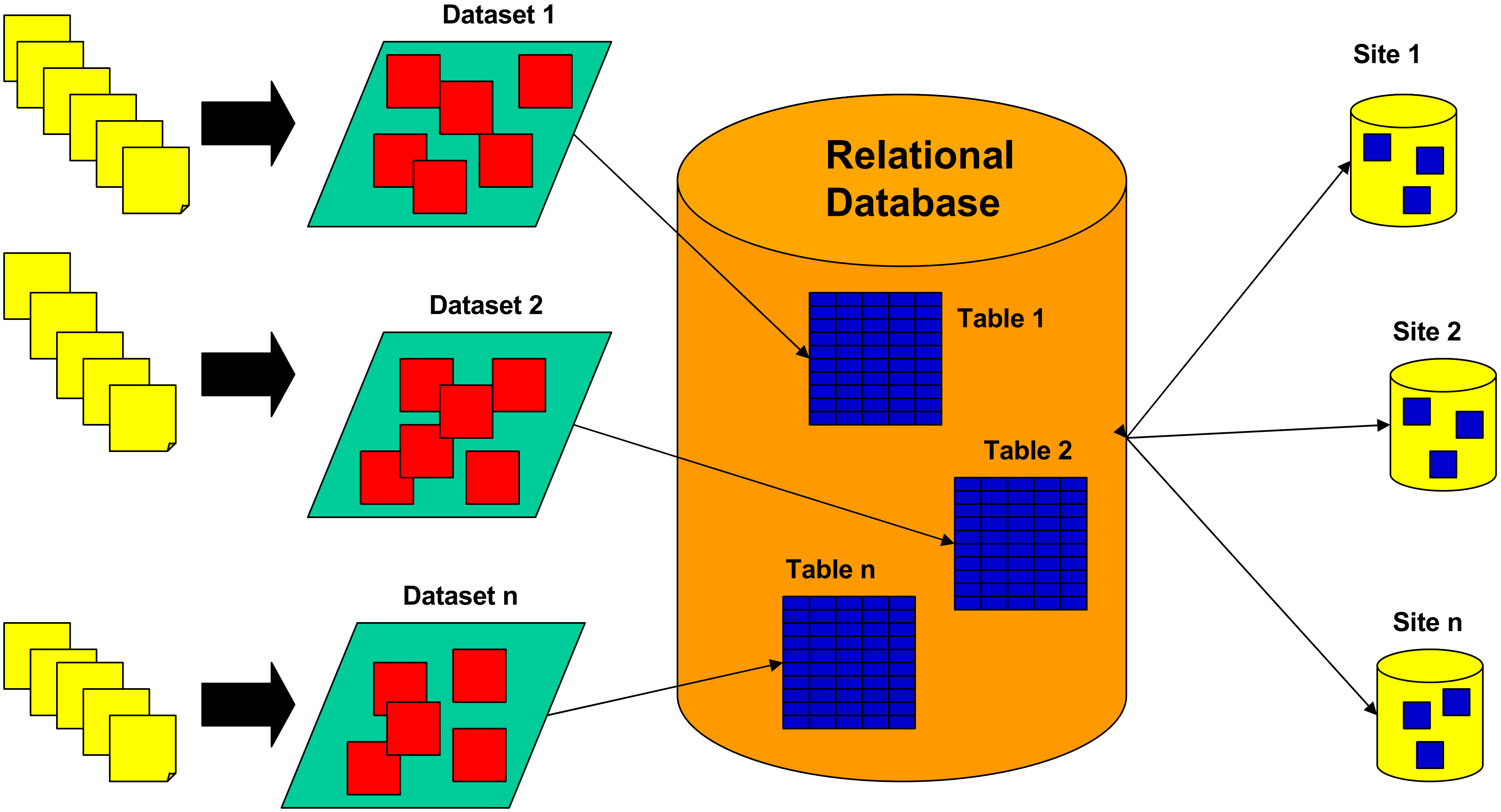
Underlying technology

- Persistence technology for tags are currently based upon POOL collections
- Collections store references to objects, along with a corresponding attribute list upon which one might base object-level selection
- Implemented in ROOT and in relational database backends

Data flow

- Event tags are written into ROOT files when Analysis Object Data (AOD) are produced at Tier 0
  - ❑ Strictly speaking, tags are produced when (relatively small) AOD files are merged into larger files
- File-based tags are bulk loaded into relational database at CERN
- File-based tags may not be discarded--they may serve as indices for simple attribute-based selection and direct addressing of specific events in the corresponding data files
- Tags are sent from Tier 0 to Tier 1s, and thence to Tier 2s
  - ❑ May send only file-based tags to Tier 2s: depends on Tier 2 capabilities

Rome Workshop Event Tag Production



AOD Files (1000 evts each)

Root Collections (1 Per AOD File)

Event Tag Database (Tier 0)

Replica Databases (Tier 1)

Machinery and middleware

- Queries return lists of references to events, grouped by id (GUID) of the containing files
  - ATLAS infrastructure stores references both to AOD and to upstream processing stages (e.g., ESD) and can return any or all of these
- Utility also returns the list of distinct file GUIDs for use by resource brokers and job schedulers
- (Some) ATLAS distributed analysis prototypes are already capable of splitting the event list on these file GUID boundaries and spawning multiple jobs accordingly, to allow parallel processing of the sample

ATLAS physics experience with tags

- Tags were put into the hands of ATLAS physicists in advance of the June 2005 ATLAS Rome Physics Workshop
  - Physicists defined tag “ schema” and provided content
  - Event store group ensured that tags contained, not only pointers to events in the latest processing stage, but to upstream data as well
- Rome data production was globally distributed; only datasets that were moved back to CERN had tags inserted into collaboration-wide tag database
- Just under 3 million events in master tag database
- Feedback was positive: triggered initiation of a collaboration-wide tag content review in Fall 2005
  - Report and recommendations due late February 2006

Performance tests and experience

- Rome tag production with genuine tag content (from simulated data) provided a testbed for many things, including implementation alternatives, scalability, and performance tests
- Used for tests of indexing strategies, technology comparisons, …; details on ATLAS Twiki and elsewhere
- Performance was “adequate” for a few million events
- Conclusions: some grounds for optimism, some grounds for concern about scalability of a master tag database
  - ❑ Clearly not ready yet for 10**9 events
  - ❑ Room for divide-and-conquer strategies (horizontal partitioning, e.g., by trigger stream, vertical partitioning (e.g., by separating trigger from physics metadata), as well as indexing and back-end optimizations

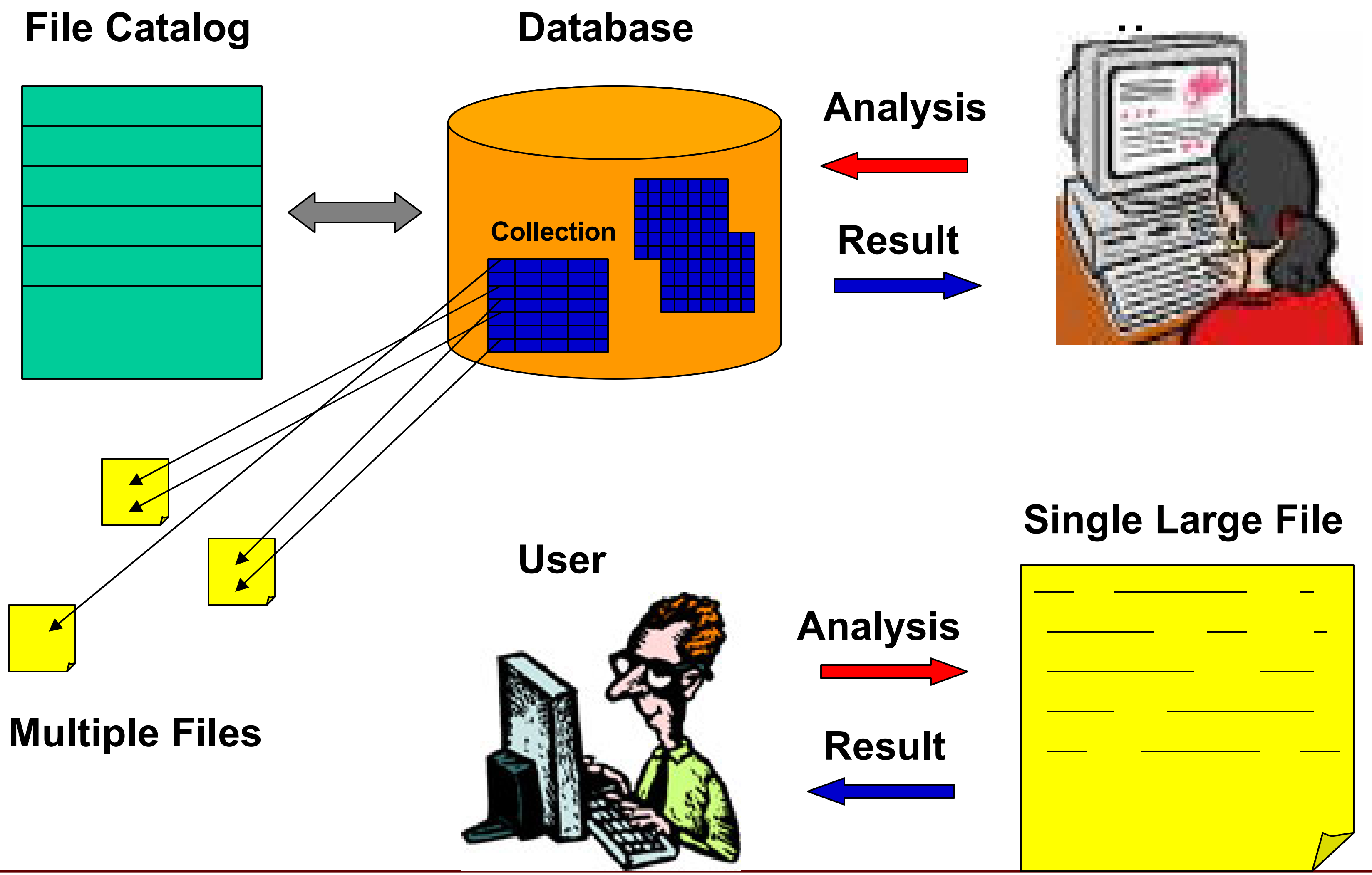
Event collections and streaming

- One natural use case: instead of extracting into a set of files the events that satisfy your selection criteria (a “ skim” in the parlance of some experiments), what about simply building a list of references to those events?
- Should you be disappointed if you are a loser in the collaboration-wide negotiations--your “ skim” is not one of the standard ones--and all you have instead is a list of event references?
  - Note that you can always use your event collection to extract the events you want into your own files on your own resources--we have utilities for this.
- What are the consequences of using reference lists to avoid the storage waste of building overlapping streams?
  - e.g., you get references to events that satisfy your selection criteria but “ belong” to someone else’ s stream

Event collections versus streams tests

- Used the Rome tag database to test the performance implications of iterating through N events scattered uniformly through M files versus iterating through them after they had been gathered into a single file (or file sequence)
- Results: Cost is approximately the cost of opening M-1 additional files--navigational overhead was too small to measure
  - ❑ Rule of thumb for ATLAS users: ~1 second per additional file in CERN Castor; much smaller on AFS, though space constraints are more stringent; don’ t have quotable SRM/dCache figures yet
- ATLAS has this month commissioned a streaming work group to decide about stream definitions and the processing stages at which streaming will be done:
  - ❑ Prototyping based upon tag database work will be integral to the strategy evaluation and comparison process

Collection vs. Direct File Access



Distributed data management (DDM) integration

- ATLAS has, since the Rome physics workshop, substantially altered its distributed data management model to be more “ dataset” -oriented
- Poses challenges to the event-level metadata system, and to the production system as well
- Tag payloads today are references to events, which, following the LCG POOL model, embed file ids. This was “ easy” for the distributed data management system in the past: take the output list of unique file GUIDs, and find the files or the sites that host them.

Distributed data management integration issues

- "Dataset" questions
  - What is the corresponding dataset? Is it known at the time the tag is written? Does this imply that a job that creates an event file needs to know the output dataset affiliation in advance (true in general for production jobs, but in general?)?
  - How are datasets identified? Is versioning relevant? Is any (initial?) dataset assignment of an event immutable?
- Result of a query to an event collection is another event collection, which can be published as a "dataset" in the DDM sense
- How is the resulting dataset marshalled from the containing (file-based) datasets?
- We now have answers to many of these questions, and integration work is progressing in advance of the next round of commissioning test.
- Glasgow group (including our substitute presenter, Caitriana--thanks!) is addressing event-level metadata/DDM integration

Replication

- Used Octopus-based replication tools for heterogeneous replication (Oracle to MySQL) CERN-->Brookhaven
- Plan to use Oracle streams for Tier 0 to Tier 1 replication in LHC Service Challenge 4 tests later this year

Stored procedures

- Have done some preliminary work with Java stored procedures in Oracle for queries that are procedurally simple, but complicated (or lengthy) to express in SQL
  - ❑ Capabilities look promising; no performance numbers yet
- We may use this approach for decoding trigger information (don’ t know yet--Rome physics simulation included no trigger simulation, and hence no trigger signature representation)

Ongoing work

- Need POOL collections improvements if POOL is to provide a basis for a genuine ATLAS event-level metadata system
  - Much work is underway: bulk loading improvements, design for horizontal partitioning (multiple table implementation), and for tag extensibility
- Production system and distributed data management integration already mentioned
- We haven’t yet investigated less-than-naïve indexing strategies, bit-sliced indexing (though we have some experience with it from previous projects), or any kind of server-side tuning
- Computing System Commissioning tests in 2006 will tell us much about the future shape of event-level metadata systems in ATLAS