ATLAS EventIndex Data Collection
Supervisor and Web Interface

A. Fernández Casani, C. García Montoro, J. Sánchez

On behalf of the ATLAS Collaboration

Instituto de Física Corpuscular (IFIC)
CSIC- Universitat de València

CGW’16 Kraków, Poland – 24-26 October 2016
Outline

• LHC and ATLAS
• Events in ATLAS
  • How ATLAS production works
• What is ATLAS EventIndex
  • Current Messaging
  • New Data Flow
• ATLAS EventIndex Data Collection Supervisor
• ATLAS EventIndex Web Interface
• Conclusions and Future Work
LHC and ATLAS

Large Hadron Collider (LHC) is a particle accelerator and collider located at CERN in the border of Switzerland and France.

ATLAS is one the 4 detectors at LHC, devoted to test the predictions of the Standard Model, and to physics beyond the Standard Model.

“ATLAS is an experiment at CERN designed to explore the secrets of the universe.”
Events in ATLAS

An event is a p+/p+ collision
Around 1 billion collisions/sec.

- Separate Triggers (downward arrows)

Raw Data:
Data from detector
- 1000 evt/s
- 1 MB/evt
- 1 GB/s
- 10 PB/year

Reconstructed Data
Tracks, showers,...
- 2.6 MB/event
- 26 PB/year

Physics Data:
Particle Types, Momenta, Energies,...
- 0.3 MB/evt
- 3000 TB/year (AOD)

Dataset:
Set of files

File:
Set of physics events (GUID)
Worldwide LHC Computing Grid (WLCG)

Tier 0
CERN (Geneva) + Wigner (Budapest)
- **First copy** of raw data.
- First pass reconstruction.
- **Distribution** of raw and reconstruction data to Tier 1.

Tier 1
13 centres (11 for ATLAS)
- **Safe-keeping** of a proportional share of raw and reconstructed data.
- Large-scale reprocessing and safe-keeping the output.
- **Distribution** to T2 and safe-keeping of a share of T2-produced simulated data.

Tier 2
Around 160 sites
- Computing power for specific **analysis** tasks.
- Handle analysis requirements and proportional share of simulated event **production** and reconstruction.
Some Key Components

**Tier0**
- Processes from raw data to first pass reconstruction.
- First copy of data.

**PanDA**
- Production ANd Distributed Analysis system.
- A manager of tasks and jobs for analysis and production.

**Rucio**
- Distributed Data Management.
- A catalogue of files, datasets and containers.
What is ATLAS EventIndex

• Just in 2015, ATLAS has produced:
  – 12 billion real events in one million files.
  – 5 billion Monte Carlo simulation events in 8 million files.

• EventIndex aims to build a complete catalogue of all ATLAS physics events, real and Monte Carlo simulation, for all processing stages.

• To allow:
  – **Event picking**: Give the reference to specific events depending on constraints:
    • Order of hundreds of concurrent users requesting from 1 event to 30k.
    • Provenance: Chain of processes.
  – **Production consistency checks**:
    • Duplicate event checking and overlap detection.
  – **Trigger checks and event skimming**:
    • Count or give an event list based on trigger selection.
    • Trigger Overlap: number of events in real data that satisfies two different triggers.
Current ATLAS EventIndex Messaging

**Message size** is set small 1-10kB to keep broker queues agile.

**Producers tag messages by group (JMSXGroupID)**
Ensures that all messages from 1 Producer will go to 1 Consumer.

**Atomic transactions on Producers:**
if connections breaks no partial processing occurs.

**Status messages** are sent from producers and consumers to an alternate queue.

Stomp performance in our benchmarks reaches over 350 msg/s and 10Mb/s per Producer. Measured performance for sending real events reached 200K event/s and 60Mb/s (1Broker, 6 Prod/s, 4 Cons, 50K events/job)

**Files stored on Mapfile format usable by Hadoop Core Services**
Redesigning ATLAS EventIndex

• Current EventIndex is working reliably, but...
  – We have to cope with future necessities, i.e. more data to be indexed.
  – It has a significant messaging overhead.
  – Requires frequent human intervention of developers and experts.

• A new data flow architecture prototype is ongoing:
  – To reduce messaging complexity.
  – To ease operational procedures.
New ATLAS EventIndex Data Flow

Data production is shared between producers and consumers by means of CERN's Object Store facility based on Ceph, using its Amazon S3 compatible interface.

Validation objects are stored on Object Store too.

Control messages:
- STOMP + ActiveMQ
- Production stats + URL
- Consumer stats

Producers

Object Store

STOMP brokers

EventIndex Supervisor

Web Interface

Consumers

DB

Rucio

Tier0

Panda

Validation objects are stored on Object Store too.

Data production is shared between producers and consumers by means of CERN's Object Store facility based on Ceph, using its Amazon S3 compatible interface.
EventIndex Data Collection Supervisor

• Python + SQLAlchemy:
  - Not tied to a particular DBMS.
  - Shared models with web interface.
  - Thread-local scoped sessions with the database

• Asynchronous multithreaded application with queues:
  - Threads retrieve info from external resources asynchronously
    - Main supervisor loop.
    - Message loop.
    - Active Task loop.
    - Rucio loop.
    - Validation loop.
  - Queues store small transient entities, mainly ids, on their own tables
    - Active tasks queue.
    - Rucio queue.
    - Validation queue.
EventIndex Data Collection Supervisor

- STOMP broker
  - from producer
  - Message loop
- Rucio queue
- Active task queue
- Active tasks loop
  - TIER0
  - PANDA
- Rucio loop
- Validation queue
- Validation loop
- STOMP broker
to consumer

Main supervisor loop

RUCIO

- from producer
- Message loop
- STOMP broker
- Validation loop
- STOMP broker
ATLAS EventIndex web interface

• Shows the data flow and processing state.
  – Able to alert when some problems are detected.

• Provides information about the supervisor and the indexing tasks.

• Written in Python with Flask framework and SQLAlchemy.
  – Shared models with supervisor.

• Uses Apache's GridSite module to authenticate through certificates.
  – Will ease interactions with external applications.

• JSON is used to provide data to clients.
EventIndex Web Interface: Components

- MySQL
- ORACLE
- python
- SQLAlchemy
- EventIndex Supervisor
- Python
- Flask
- SQLAlchemy
- Jinja
- jQuery
- Apache
- Foundation

In web development, one drop at a time.
ATLAS EventIndex Web Interface

### Tier0 Tasks in EventIndex Queues

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TaskId</th>
<th>Creation</th>
<th>Task State</th>
<th>EventIndex State</th>
<th>Total</th>
<th>Done</th>
<th>Run.</th>
<th>Proc.</th>
<th>Produced</th>
<th>Consumed</th>
</tr>
</thead>
<tbody>
<tr>
<td>notknownyxe</td>
<td>11</td>
<td>2016-09-23 19:31</td>
<td>FINISHED</td>
<td>EI_REGISTERED</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012354.physics_Main.merge.AOD.9999_m9999</td>
<td>10</td>
<td>2016-09-23 19:31</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012353.physics_Main.merge.AOD.9999_m9999</td>
<td>9</td>
<td>2016-09-23 19:31</td>
<td>RUNNING</td>
<td>EI_RUNNING</td>
<td>20</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012352.physics_Late.merge.AOD.9999_m9999</td>
<td>8</td>
<td>2016-09-23 18:59</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>35</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012351.physics_Late.merge.AOD.9999_m9999</td>
<td>7</td>
<td>2016-09-23 18:58</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012350.physics_Main.merge.AOD.9999_m9999</td>
<td>6</td>
<td>2016-09-23 18:49</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>35</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012349.physics_Main.merge.AOD.9999_m9999</td>
<td>5</td>
<td>2016-09-23 18:48</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012348.physics_zeroBias.merge.AOD.9999_m9999</td>
<td>4</td>
<td>2016-09-23 17:53</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012347.physics_Main.merge.AOD.9999_m9999</td>
<td>3</td>
<td>2016-09-23 20:13</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>data16_13TeV.00012346.physics_Late.merge.AOD.9999_m9999</td>
<td>2</td>
<td>2016-09-23 20:13</td>
<td>FINISHED</td>
<td>EI_VALID</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 11 entries

Rucio Queue

<table>
<thead>
<tr>
<th>Total</th>
<th>Waiting</th>
<th>Recoverable</th>
<th>Unrecoverable</th>
<th>Active Tasks Queue</th>
<th>Validation Queue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

• ATLAS EventIndex handles billions of events in millions of files per year.

• Though current EventIndex is working reliably, we are making it future-proof.
  – Object Store will reduce messaging considerably.
  – New supervisor and web interface will reduce operational efforts.

• The web interface already implements some REST endpoints, but further development is required.
Backup: Map of ATLAS WLCG Sites
Backup: Some ATLAS facts

From [http://atlasexperiment.org/fact_sheets.html](http://atlasexperiment.org/fact_sheets.html)

- 3,000 scientists, 38 countries, 180 universities and labs.
- An event is a proton-proton collision.
  - Around 1 billion collisions per second.
  - Trigger and Event Filter select about 1000 *interesting* events/sec.
  - Raw data: The data after trigger 1GB/s, 1MB/event, 10PB/year.
Backup: Some EventIndex Numbers

- Around 200 producer tasks simultaneously running on Tier0.
- Around 20 events/message.
- Typically around 300 messages/file.
- In some unusual case, up to 20,000 events/file.
- 5 ActiveMQ Brokers.
- Consumers = 2 x Brokers = 10.