Developing of Hybrid SQL/NoSQL Metadata Storage

M. Grigorieva, M. Golosova

Laboratory of Big Data Technologies for mega-science projects
Hybrid (SQL/NoSQL) Storage research areas:

- Implementation of the internal logic of the Hybrid Metadata Storage (M. Golosova)

- Adaptation of BigPanDAmon for interaction with NoSQL databases (M. Grigorieva)
Hybrid Storage Internal Logic

- Storage management inner tasks:
  - Provide consistence of SQL and NoSQL parts
  - Provide inner consistency within NoSQL
    - Detect unsynchronized data
    - Synchronize data
    - Do not affect the whole system performance
- Clean SQL database from archive records

Hybrid Storage SQL/NoSQL Data Synchronization

1) **Criteria**
   - Modification time
   - Not copied yet
     - [date – number of jobs]
2) **Policy**
   - When? [Daily]
     - days limit: ?

```
<table>
<thead>
<tr>
<th>date</th>
<th>count</th>
<th>state</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-01-26</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>2014-01-27</td>
<td>1096887</td>
<td></td>
</tr>
<tr>
<td>2014-01-28</td>
<td>2210772</td>
<td></td>
</tr>
<tr>
<td>2014-01-29</td>
<td>75862</td>
<td></td>
</tr>
</tbody>
</table>
```

```sql
SQL
```
```nosql
NoSQL
```

SERVICETABLES
Hybrid Storage Inner NoSQL Data Synchronization

NoSQL Data Model

1) Criteria:
   - Modification time
   - Inner consistency broken

2) Policy:
   - When?
   - How much?

<table>
<thead>
<tr>
<th>Table</th>
<th>Date</th>
<th>Key</th>
<th>Value</th>
<th>Count</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs</td>
<td>2015-01-27</td>
<td>TaskId</td>
<td>1516868</td>
<td>201</td>
<td>-</td>
</tr>
<tr>
<td>Task</td>
<td>2015-01-27</td>
<td>TaskId</td>
<td>1516868</td>
<td>201</td>
<td>Ok</td>
</tr>
<tr>
<td>Jobs</td>
<td>2015-01-27</td>
<td>TaskId</td>
<td>1516869</td>
<td>723</td>
<td>-</td>
</tr>
<tr>
<td>Task</td>
<td>2015-01-27</td>
<td>TaskId</td>
<td>1516869</td>
<td>512</td>
<td>?</td>
</tr>
<tr>
<td>Jobs</td>
<td>2015-01-27</td>
<td>Cloud</td>
<td>UK</td>
<td>692</td>
<td>-</td>
</tr>
<tr>
<td>Cloud</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
BigPanDA monitor

- Separates data access layer and visualization
- Built around common key PanDA objects: *jobs, resources, etc.*
- BigPanDAMon based on **django** Framework
- Runs on top of SQL DB backends
- Modular and reusable monitoring:

**Goal**

Adapt BigPanDA Monitor to work with both SQL and NoSQL DB backends (SQL – operational data, NoSQL – historical data):

**Methods:**

1. Enchance Django ORM to interact with both: SQL and NoSQL
2. Integration of Hybrid SQL/NoSQL Storage in BigPanDA Monitor

- Analyze job failures
- Monitor progress of analysis/activity of a PanDA resource
- Organize/visualize data
BigPanDA Monitor with NoSQL backend

1. Enhance Django ORM to interact with NoSQL
2. Use Cassandra database wrapper
3. Integration with Hybrid Storage API
Many Thanks!

- Eygene Ryabinkin
- Gancho Dimitrov
- Mario Lassnig