Studies of Big Data metadata segmentation between relational and non-relational databases

M.Golosova (NRC KI), G.Dimitrov (CERN), M.Grigorieva (NRC KI), A.Klimentov (BNL), M.Potekhin (BNL), E.Ryabinkin (NRC KI)

National Research Centre «Kurchatov Institute»
CERN
Brookhaven National Laboratory

CHEP2015, Okinawa, Japan, April 13-17 2015
Outline

• **PanDA WMS**
  • Metadata
  • Monitoring
  • Data specification

• **NoSQL as archive**
  • Precalculation for speedup

• **Study**
  • Test set-up
  • Scaling
  • Precalculation effect
  • NoSQL archive performance
PanDA WMS

Metadata storage

Production system

PresentDB

Job

Panda Server

Submitter

Job

Pilots

Request for a job

User

courtesy picture from T.Maeno

Scale of the challenge: up to 1.7M completed jobs per day

Metadata analysis and monitoring: PanDA Monitor

Related CHEP talks: T.Maeno,

The Future of PanDA in ATLAS

Distributed Computing (CHEP ID 144)
PanDA monitoring

- **PanDA Monitor**
  Web-based analytical interface providing information about jobs and tasks within the system

- **Page «Errors»**
  Page of PanDA Monitor providing information about job errors that have occurred in the system

http://bigpanda.cern.ch/errors/

13/04/2015
Golosova Marina, CHEP 2015
Job errors monitoring

1) Get list of jobs from DB (according to request parameters)
2) Summaries by parameters (overall, computing site, user, task)
3) Get additional information from DB
4) Generate output

CERN
Oracle Monitor instance (4 GB RAM)
ATLAS Integration DB
JobsArchived

Page generation time

T, sec
0 50 100 150
N jobs, 10^3

0 20 40 60

Total
DB request

13/04/2015
Golosova Marina, CHEP 2015
Specificity of the metadata storage

- Number of requested records (ATLAS PanDA Archive: ~900M of jobs)
- RDBMS as archive backend (ACID standard)

**Key points:**
- fast reads and writes
- strong consistency

**Daily work**

**WMS (PanDA)**

**Archive metadata**

**Scheduled data transfer**

- single write

**Short-term analytics**

**Long-term analysis**

**ACID is too strong**

**BASE is enough**

**Key points:**
- no writes
- bulk reads (long-term)
- fast reads (short-term)
Test set-up

CERN
Oracle Monitor instance
4 GB RAM

Oracle (ATLAS Integration DB)
RAM: 128 GB
CPU: 16 cores, 2.00GHz

Client
Request 1
Request 2

NRC KI
Oracle + NoSQL Monitor instance
48 GB RAM

Cassandra cluster
2 nodes - RAM: 24 GB
CPU: 4 cores, 2.4 GHz,
1 node – RAM: 48 GB
CPU: 8 cores, 2.4 GHz

Jobs
Precalculation
synchronization

... JediTasks

day_site_errors
day_site_errors_30m
day_errors_30m

DB Tables
Request-specific, aggregated data

PanDA Monitor, adapted for work with both Cassandra (for a limited set of requests) and Oracle archives
Cassandra scaling test

Total (page generation) and DB request time vs. number of jobs matching the request

Sample A: 4.2 M records (118 MB)
Sample B: 1.4 M records (83 MB)

Cassandra DB
(2 months archive):
- metadata for ~60 M jobs
  (~100 GB per replica)
**Aggregation effect test**

**Page generation and DB request time for plain (not aggregated) and aggregated data**

- **No aggregation** — total
- **No aggregation** — DB request
- **30 min** — total
- **30 min** — DB request

**No aggregation**: DB record ~ 1 error

**30 min**: DB record ~ all errors (with the same error code) within given 30 minutes
**NoSQL Archive performance**

Values for Cassandra

\[ T = 4 \times (t - t_{add}) + t_{add}, \text{ where:} \]

- \( t \) – total page generation time
- \( t_{add} \) – not a summary-specific DB requests
- 4 – number of specific summaries
Summary

- It is hardly possible to perform long-term metadata analysis without any precalculation.

- Replacing RDBMS with NoSQL in archive part of metadata storage can be used to improve availability of historical data.

- Prototype of NoSQL (Cassandra) archive was created and tested on a 2-month slice of metadata from ATLAS PanDA Archive.

- First results look promising. NoSQL archive will be extended to confirm this belief in new tests.

- Adaptation of PanDA Monitor for work with NoSQL archive will be continued.
Acknowledgment

• This talk drew on presentations, discussions, comments, input from many our CERN, ATLAS, NRC-KI Colleagues, thanks to all people working in PanDA project

• This work was funded in part by the Russian Ministry of Science and Education under Contract N14.Z50.31.0024
Back up slides
Synchronization & Precalculation

Data transformation (to Cassandra format)

Import of 1M rows took ~16 min
Precalculation: ~6 min
Daily synchronization (~1.5M jobs) will be taking ~35 min
## Cassandra Column Families

<table>
<thead>
<tr>
<th>Primary Key</th>
<th>Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Partition Key</td>
<td>□ Common</td>
</tr>
<tr>
<td>□ Clustering Key</td>
<td>□ Counter</td>
</tr>
</tbody>
</table>

### day_site_errors_cnt_30m

<table>
<thead>
<tr>
<th>date</th>
<th>computingSite</th>
<th>base_mtime</th>
<th>errcode</th>
<th>errdiag</th>
<th>err_count</th>
<th>job_count</th>
</tr>
</thead>
</table>

### day_site_errors

<table>
<thead>
<tr>
<th>date</th>
<th>computingSite</th>
<th>errcode</th>
<th>pandaid</th>
<th>errdiag</th>
</tr>
</thead>
</table>

### day_errors_30m

<table>
<thead>
<tr>
<th>date</th>
<th>base_mtime</th>
<th>count</th>
</tr>
</thead>
</table>
Compaction types

Compaction type defines the way Cassandra stores data on disk.
Aggregation effect test (DB records)

Page generation and DB request time for plain (not aggregated) and aggregated data vs. the number of read DB records

No aggregation — total
No aggregation — DB request
30 min — total
30 min — DB request

No aggregation: DB record ~ 1 error
30 min: DB record ~ all errors (with the same error code) within given 30 minutes