Unlocking Big Data at CERN

Matthias Braeger CERN, Manish Devgan Software AG (Terracotta)
4:15pm Thursday, 10/16/2014
Hadoop in Action
Location: 1 C03/1 C04
Speakers & Agenda

- Big Data @ CERN
- In-Memory Data Management
- In-Memory @ CERN

Matthias Braeger
Software Engineer
CERN
matthias.braeger@cern.ch

Manish Devgan
Product Management
Software AG (Terracotta)
manish.devgan@softwareag.com
Physics data

Metadata of physics data

Sensor Data of technical installations

Log data

Configuration data

Documents

Media data

Others
European Organization for Nuclear Research

- Founded in 1954 (60 years ago!)
- 21 Member States
- ~ 3’360 Staff, fellows, students...
- ~ 10’000 Scientists from 113 different countries
- Budget: 1 billion CHF/year

http://cern.ch
From Physics to Industry
The world's biggest machine

Generated 30 Petabytes in 2012
> 100 PB in total!
LHC - Large Hadron Collider

27km ring of superconducting magnets

Started operation in 2010 with 3.5 + 3.5 TeV, 4 + 4 TeV in 2012

Since early 2013 in Long Shutdown 1 (machine upgrade)

Restart early 2015 at 6.5 + 6.5 TeV
Some ATLAS facts

- 25m diameter, 46m length, 7'000 tons
- 100 million channels
- 40MHz collision rate (~ 1 PB/s)
- Run 1: 300 Hz event rate after filtering
- Run 2: up to 1 kHz
Is Hadoop used for storing the ~30 PB/year of **physics data**?

No ;-(

Experimental data are mainly stored on tape

CERN uses Hadoop for storing the **metadata** of the experimental data
Physics Data Handling

- **Run 1**: 30 PB per year demanding 100,000 processors with peaks of 20 GB/s writing to tape spread across 80 tape drives

- **Run 2**: > 50 PB per year

CERN’s Computer Center (1st floor)
Physics Data Handling

2013 already more than 100 PB stored in total!

- > 88 PB on 55’000 tapes
- > 13 PB on disk
- > 150 PB free tape storage waiting for Run 2
Physics Data Handling

- Cost of tape storage is a lot less than disk storage
- No electricity consumption when tapes are not being accessed
- Tape storage size = Data + Copy
  Hadoop storage size = Data + 2 Copies
- No requirement to have all recorded physics data available within seconds
3 HBase Clusters

- CASTOR Cluster with ~10 servers
  - ~ 100 GB of Logs per day
  - > 108 TB of Logs in total
- ATLAS Cluster with ~20 servers
  - Event index Catalogue for experimental Data in the Grid
- Monitoring Cluster with ~10 servers
  - Log events from CERN Computer Center
Metadata from physics event

Metadata are created upon recording of the physics event

Examples 1:

- Tape Storage event log
  - On which tape is my file stored?
  - Is there a copy on disk?
  - List me all events for a given tape or drive
  - Was the tape repacked?
## Example 1: Tape Storage event log

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Instance</th>
<th>Hostname</th>
<th>Daemon</th>
<th>PID</th>
<th>TID</th>
<th>Message text</th>
<th>Request ID</th>
<th>Tape ID</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-05-21</td>
<td>c2repack</td>
<td>c2repactm4021</td>
<td>staged</td>
<td>6295</td>
<td>6322</td>
<td>Request processed</td>
<td>88040a8b-882d-430d-92e0-007030b0c0a4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2014-09-09</td>
<td>c2repack</td>
<td>c2repactm4021</td>
<td>nad</td>
<td>7083</td>
<td>7102</td>
<td>Processing complete</td>
<td>2674859-8035-4955-8ff6-144a3d5ac5fe</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2014-09-09</td>
<td>c2repack</td>
<td>c2repactm4021</td>
<td>staged</td>
<td>0</td>
<td>6321</td>
<td>selfRecalced: db updated after full reci completed</td>
<td>036a4a6d-471c-76dc-9e02-6b670100a008</td>
<td>140480</td>
<td></td>
</tr>
<tr>
<td>2014-09-13</td>
<td>c2repack</td>
<td>c2repactm4021</td>
<td>nad</td>
<td>0</td>
<td>6321</td>
<td>checkRecalced: created missing directories in the namespace</td>
<td>036a4a6d-471c-76dc-9e02-6b670100a008</td>
<td>140480</td>
<td></td>
</tr>
<tr>
<td>2014-09-21</td>
<td>c2repack</td>
<td>c2repactm4021</td>
<td>staged</td>
<td>6401</td>
<td>6527</td>
<td>Wrier: file to recall retrieved from db</td>
<td>-</td>
<td>140480</td>
<td></td>
</tr>
<tr>
<td>2014-07-09</td>
<td>c2repack</td>
<td>c2repactm3011</td>
<td>staged</td>
<td>7430</td>
<td>7467</td>
<td>Request processed</td>
<td>88900d89-882d-430d-92e0-007030b0c0a4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2014-07-09</td>
<td>c2repack</td>
<td>c2repactm3011</td>
<td>staged</td>
<td>0</td>
<td>7464</td>
<td>createRecallCandidate: create new MigrationJob to migrate</td>
<td>036a4a6d-471c-76dc-9e02-6b670100a008</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Example 1: Tape Storage event log

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Severity</th>
<th>Instance : Hostname</th>
<th>Daemon</th>
<th>PID</th>
<th>TID</th>
<th>Message text</th>
<th>Request ID</th>
<th>Tape ID</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-06-28 02:47:59.239042</td>
<td>info</td>
<td>c2repack-c2repacksv401</td>
<td>tapegatesyd</td>
<td>0</td>
<td>30405</td>
<td>setFileMigrated db updates after full migration completed</td>
<td>fcdb7403-114a-70ae-6043-a708100ab1c2</td>
<td>T52505</td>
<td></td>
</tr>
<tr>
<td>2014-06-28 02:47:58.536147</td>
<td>info</td>
<td>c2repack-c2repacksv401</td>
<td>msd</td>
<td>0</td>
<td>30405</td>
<td>New segment information</td>
<td>fcdb7403-114a-70ae-6043-a708100ab1c2</td>
<td>T52505</td>
<td></td>
</tr>
</tbody>
</table>

Page generated in 0.118740 sec. Data fetched from HBase in 0.115751 sec. Estimated size of the full data set: 6048Bytes.
Metadata from physics event

Metadata are created upon recording of the physics event

Examples 2:

- Information about
  - Event number
  - Run number
  - Timestamp
  - Luminosity block number
  - Trigger that selected the event, etc.
Example 2: ATLAS EventIndex catalogue

Prototype of an event-level metadata catalogue for all ATLAS events

- In 2011 and 2012, ATLAS produced 2 billion real events and 4 billion simulated events
- Migration from former solution by the end of this year

Data are read from the brokers, decoded and stored into Hadoop.
Example 2: ATLAS EventIndex catalogue

The major use cases of the EventIndex project are:

- **Event picking:**
  give me the reference (pointer) to “this” event in “that” format for a given processing cycle.

- **Production consistency checks:**
  technical checks that processing cycles are complete (event counts match).

- **Event service:**
  give me the references (pointers) for “this” list of events, or for the events satisfying given selection criteria.
A lot of ongoing research for treating Big Data

Big Data “at-rest”
- Oracle DB + Hadoop + for advanced analytics
  - `{swirl}`: Learn R, in R (http://swirlstats.com)

Big Data “in-motion”
- Complex Event Processing (CEP), e.g. Esper
- In-Memory frameworks built on JCache (JSR-107)
Speakers & Agenda

- Big Data @ CERN
- In-Memory Data Management
- In-Memory @ CERN

Matthias Braeger
Software Engineer
CERN
matthias.braeger@cern.ch

Manish Devgan
Product Management
Software AG (Terracotta)
manish.devgan@softwareag.com
Growth of Data

Transactions, Sensors, Logs, M2M, ..
The value of *real* time

Latency Matters
Uptime, SLAs, HA

Performance and Scale
The Shift

90% of Data in Disk-based Databases

90% of Data in In-Memory
Why now?

Steep drop in price of RAM

Explosion in volume and velocity of data
In-Memory Data Platforms

- Scale of NoSQL
- Low latency of In-Memory databases
- Reliability & Fault Tolerance
- Transactional Guarantees

*Fast* Big Data
Tiered Storage

Latency
- Micro-seconds
- Milli-seconds
- Seconds

Speed (TPS)
- 2,000,000+
- 1,000,000
- 100,000

External Data Source
(e.g., Database, Hadoop, Data Warehouse)

- Tier 1: Local BigMemory
- Tier 2: Distributed BigMemory Server RAM or Flash/SSD
- Tier 3: BigMemory

- 4 GB Process Memory
- 32 GB – 12 TB
- 100s GB – 100s TB

- Micro-seconds
- Micro-seconds
- Milli-seconds
- Seconds

1,000s

App Server
- Application
- BigMemory
- Application
- BigMemory

RDBMS

#strataconf #hadoopworld
Scale with data and processing needs

- **Scale Up**
  - Increase Data in Memory
  - App Server
    - API
    - BigMemory
  - Database

- **Scale Up & Out**
  - App Server
    - API
    - BigMemory
  - Terracotta Server Array
  - Database

- **Elastic Scale**
  - App Server
    - API
    - BigMemory
  - Terracotta Server Array
  - Database

Reduce Database Reliance
HA, Extreme Resiliency

- Active Mirror
- No Single point of failure
- Fast Restartable Storage (SSD/Flash)
Use cases

Influencing operations and decisions
In-Memory Data Fabric: Operationalize Hadoop

Transactions, “Events”, At risk accounts information, ..

Streaming insights into In-Memory Operational Store
In-Memory Data Fabric: Streaming Analytics

High Speed resilient data access across shared time windows
Speakers & Agenda

- Big Data @ CERN
- In-Memory Data Management
- In-Memory @ CERN

Matthias Braeger
Software Engineer
CERN
matthias.braeger@cern.ch

Manish Devgan
Product Management
Software AG (Terracotta)
manish.devgan@softwareag.com
Access Control

Network and Hardware Controls

Safety Systems

Electricity

Cryogenics

Cooling
C2MON - CERN Control and Monitoring Platform

- Allows the rapid implementation of high-performance monitoring solutions
- Modular and scalable at all layers
- Optimized for High Availability & big data volume
- Based on In-Memory solution

Currently used by two big systems at CERN: TIM & DIAMON

http://cern.ch/c2mon
Raw data filtering on DAQ layer
C2MON Server

- Authentication
- Logging
- Client communication
- Alarm
- Rules
- Benchmark
- Video access
- Lifecycle
- Configuration
- DAQ supervision
- In-Memory Store (JCache - JSR-107)
  - TERRACOTTA
- Cache persistence
- Cache loading
- Cache DB access
- DAQ in
- DAQ out
TIM – Technical Infrastructure Monitoring

- Operational since 2005
- Used to monitor and control infrastructure at CERN
- 24/7 service
- ~ 100 different main users at CERN
- Since Jan. 2012 based on new server architecture with C2MON

CERN Control Center at LHC startup
Cooling Safety Systems

Electricity Access Network and Hardware Controls

TIM (Business Layer)

Client Tier

Alarm Console Data Analysis TIM Viewer Web Apps Access Management Video Viewer

> 120k data sensors
> 41k alarms

TIM (Business Layer)

> 1200 commands
> 1300 business rules

Data Acquisition & Filtering

Cooling Safety Systems Electricity Access Network and Hardware Controls Cryogenics

> 1200 commands
> 1300 business rules

#strataconf #hadoopworld
TIM (Business Layer)

Client Tier

- Alarm Console
- Data Analysis
- TIM Viewer
- Web Apps
- Access Management
- Video Viewer

> 120k data sensors
> 41k alarms

TIM

Data Acquisition & Filtering

ca. 400 million raw data per day

Filtering

> 1200 commands
> 1300 business rules

ca. 1.5 million updates
Scenario 1: High availability

- moderate data size
- average throughput
- min service interrupts
- high availability
Scenario 2: High requirements

- large data set
- high throughput
- min service interrupts
- high availability
C2MON Roadmap

- Offering C2MON to the Open Source community [http://cern.ch/c2mon](http://cern.ch/c2mon)
- Introduction of Complex Event Processing (CEP) module
- Providing NoSQL log storage solution for high data throughput scenario
Takeaways

- **Data and High Availability** services are more important than ever before for all modern organizations.
- Deriving **value** from collected data is key to success.
- **In-Memory** platforms are essential for high value & high velocity data storage and processing.
Credits & References

Many thanks to CERN & Software AG:

- Sebastien Ponce (CERN), for providing information about CASTOR
- Rainer Toebbacke (CERN), for providing information about CERN HBASE service
- Jan Iven (CERN), for being helpful finding information about existing CERN Hadoop projects
- Software AG/Terracotta Product & Engineering Team

References:

- C2MON: http://cern.ch/c2mon
- The ATLAS EventIndex: https://cds.cern.ch/record/1690609
- Agile Infrastructure at CERN - Moving 9'000 Servers into a Private Cloud, Helge Meinhard (CERN): http://vimeo.com/93247922
- CRAN, The Comprehensive R Archive Network: http://cran.r-project.org
- Software AG Terracotta: http://www.terracotta.org
Related Information

Office Hours with Manish Devgan
(In-Memory Data Management & Computing)
5:05pm Thursday, 10/16/2014, Location: Table D

- Technology landscape for in-memory data management platforms
- Convergence of In-Memory, NoSQL, Hadoop, and other “Big Data” solutions
- Real-world deployments and use cases leveraging In-Memory Data Management

Follow up questions
- Software AG Booth #458
Questions?

Thank you for coming!