Strata+ Hadoop WORLD

PRESENTED BY

O'REILLY cloudera

strataconf.com

#strataconf

#hadoopworld

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

Log data

Metadata of physics data

Configuration data

Documents

Physics data (>100 PB)

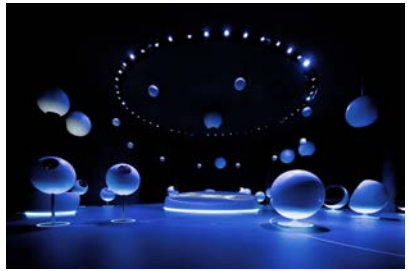
Sensor Data of technical installations

Media data

Others

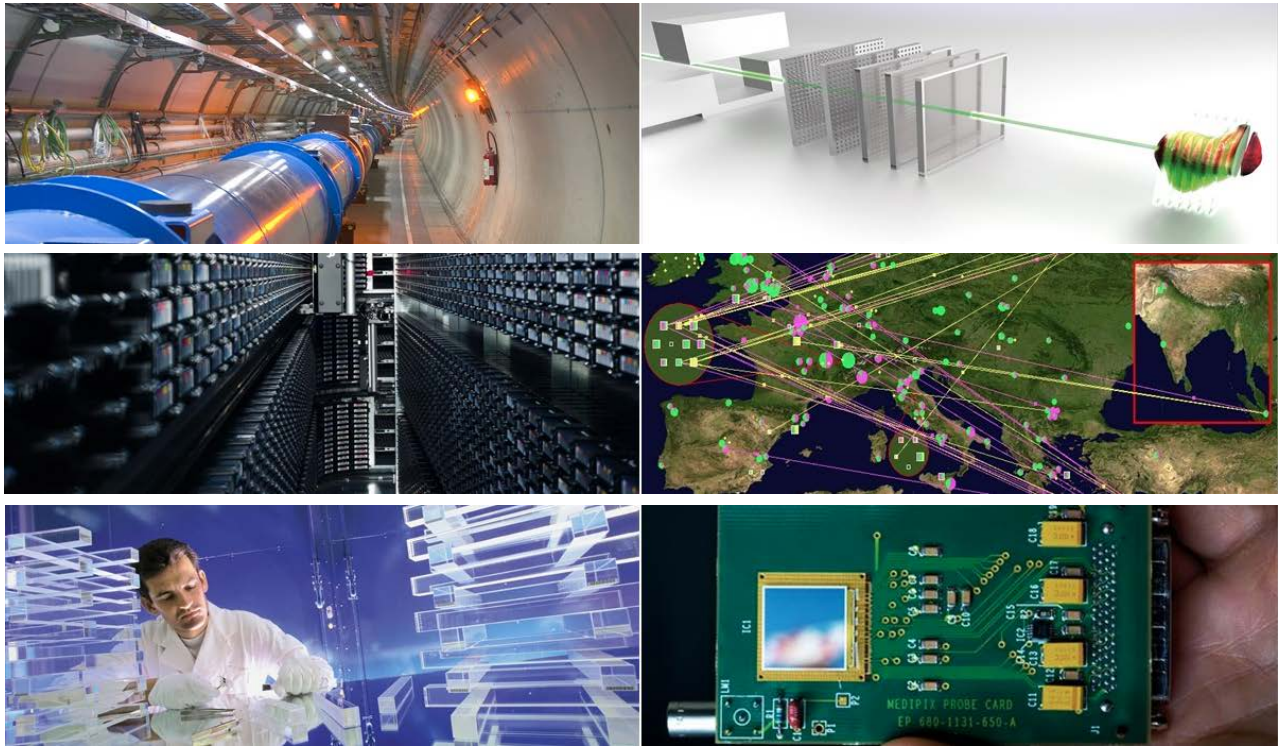
European Organization for Nuclear Research

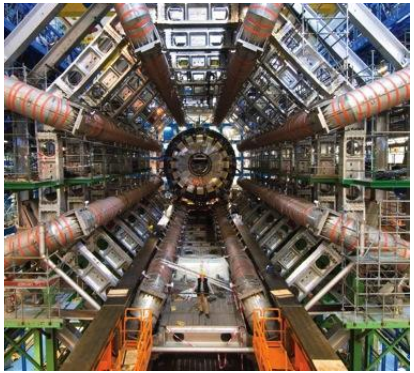
- Founded in 1954 (60 years ago!)
- 21 Member States
- ~ 3'360 Staff, fellows, students...
- ~ 10'000 Scientists from
  113 different countries
- Budget: 1 billion CHF/year



http://cern.ch

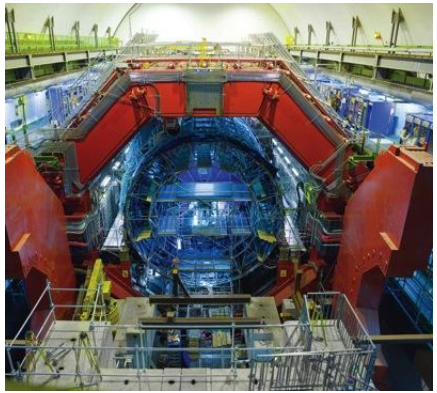
From Physics to Industry





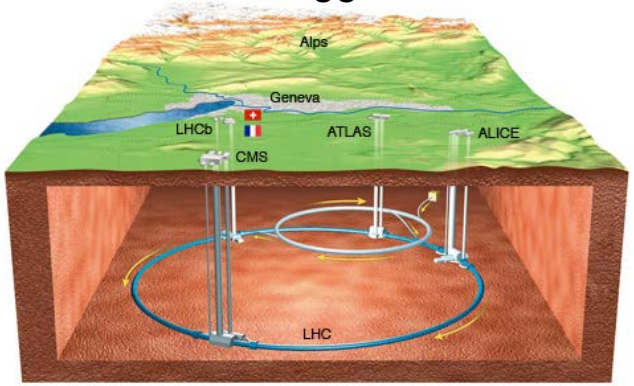
ATLAS

Alice

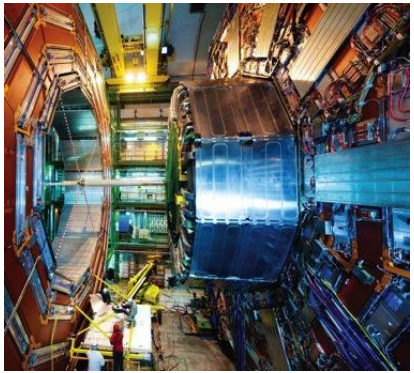


LHC

The worlds biggest machine



Generated 30 Petabytes in 2012
> 100 PB in total!



CMS

LHCb

LHC - Large Hadron Collider

27km ring of superconducting magnets

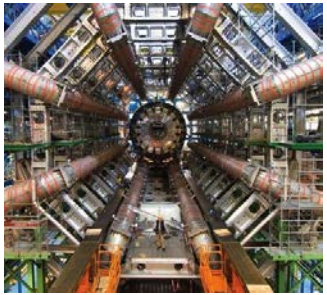
Started operation in 2010 with 3.5 + 3.5 TeV,
4 + 4 TeV in 2012

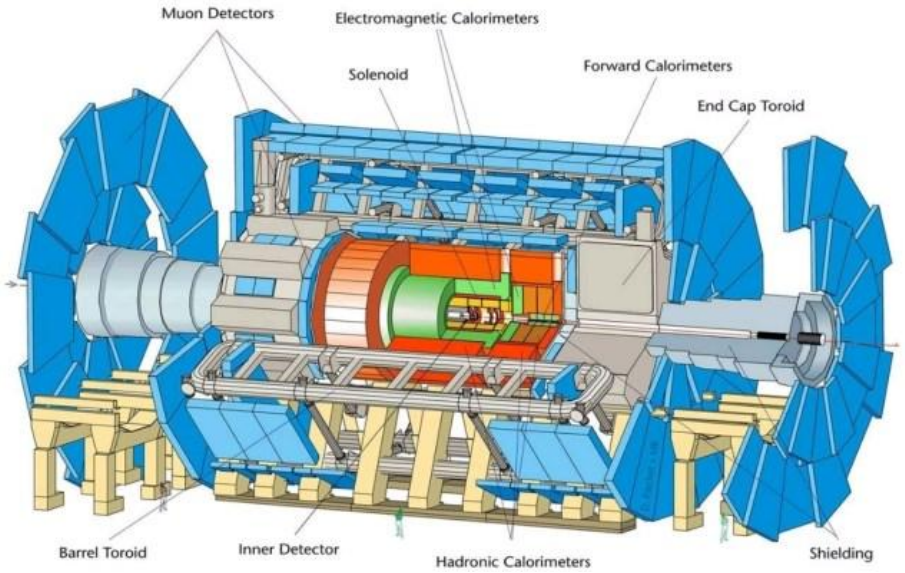
Since early 2013 in Long Shutdown 1
(machine upgrade)

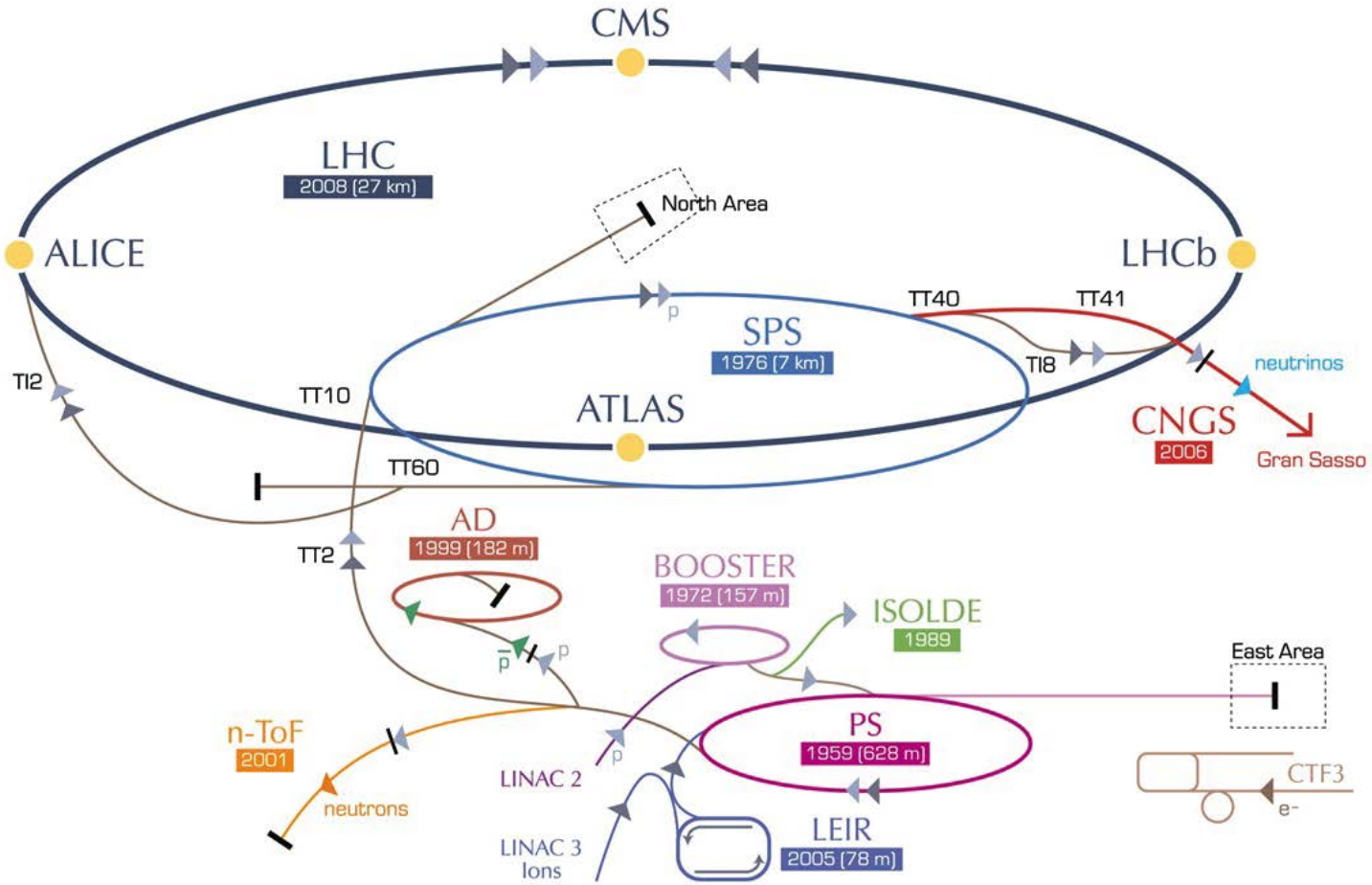
Restart early 2015 at 6.5 + 6.5 TeV

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






CMS
LHC
2008 (27 km)
North Area
ALICE
LHCb
TT40
TT41
p
SPS
1976 (7 km)
TI8
neutrinos
TI2
TT10
ATLAS
CNGS
2006
Gran Sasso
TT60
AD
1999 (182 m)
TT2
BOOSTER
1972 (157 m)
ISOLDE
1989
East Area
n-ToF
2001
PS
1959 (628 m)
LINAC 2
CTF3
e-
neutrons
LINAC 3
Ions
LEIR
2005 (78 m)

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

CERN's tape robot

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

CERN's tape robot

APACHE HBASE @ CERN

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

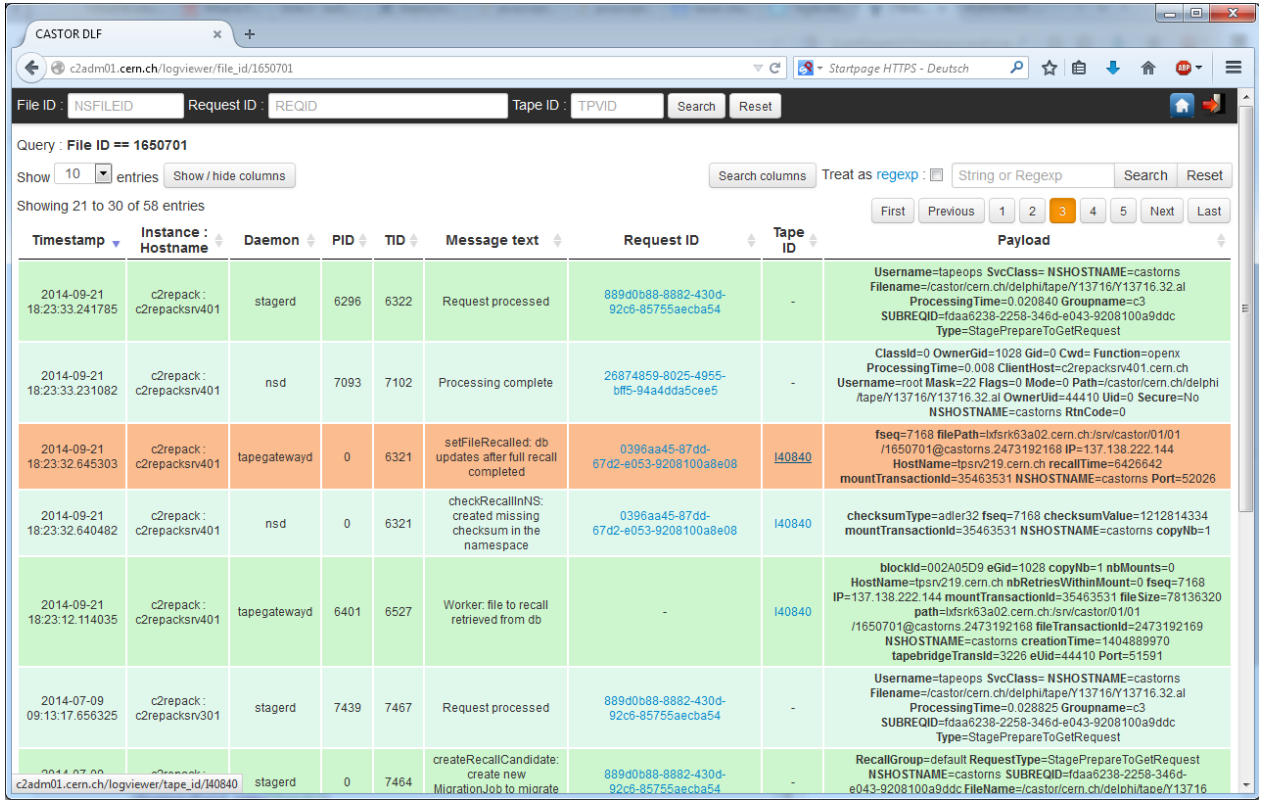
Metadata from physics event

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



Example 1: Tape Storage event log

Timestamp	Severity	Instance : Hostname	Daemon	PID	TID	Message text	Request ID	Tape ID	Payload
2014-06-28 02:47:59.239042	Info	c2repack : c2repacksrv401	tapegatewayd	0	30405	setFileMigrated: db updates after full migration completed	fcdb7403-1f4a-70ae-e043-a708100ab1c2	T52505	**IP**=137.138.223.25 **migrationTime**=8379 **HostName**=tpsrv689.cern.ch **mountTransactionId**=31426158 **NSHOSTNAME**=castorns **Port**=47385
2014-06-28 02:47:58.536147	Info	c2repack : c2repacksrv401	nsd	0	30405	New segment information	fcdb7403-1f4a-70ae-e043-a708100ab1c2	T52505	**Compression**=100 **blockId**=00867CB0 **copyNb**=2 **ChecksumType**=adler32 **gid**=1160 **fseq**=16433 **Repack**=True **NSHOSTNAME**=castorns **SegmentSize**=43699200 **ChecksumValue**= **creationTime**=1030782952

First Previous 1 2 3 Next Last

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

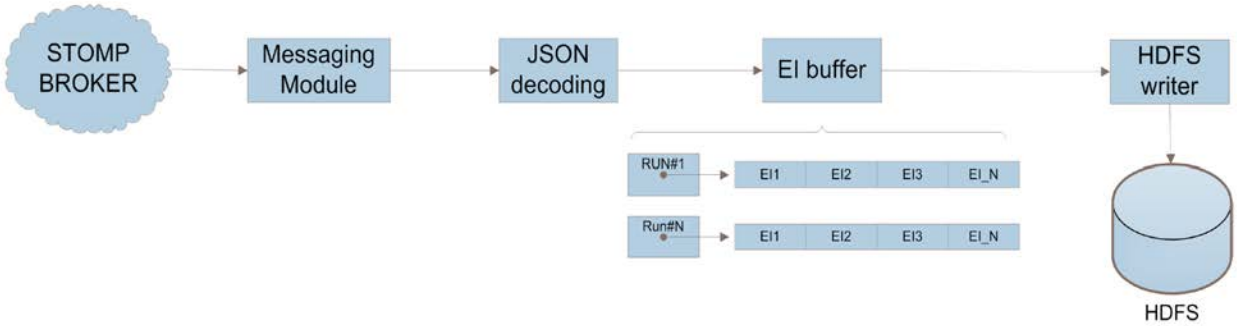
Examples 2:

- 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)

In-Memory

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, ..

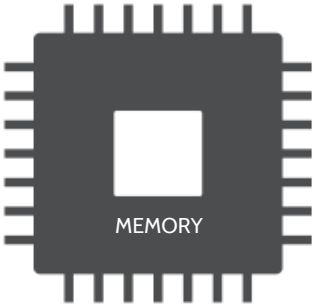
The value of *real* time

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

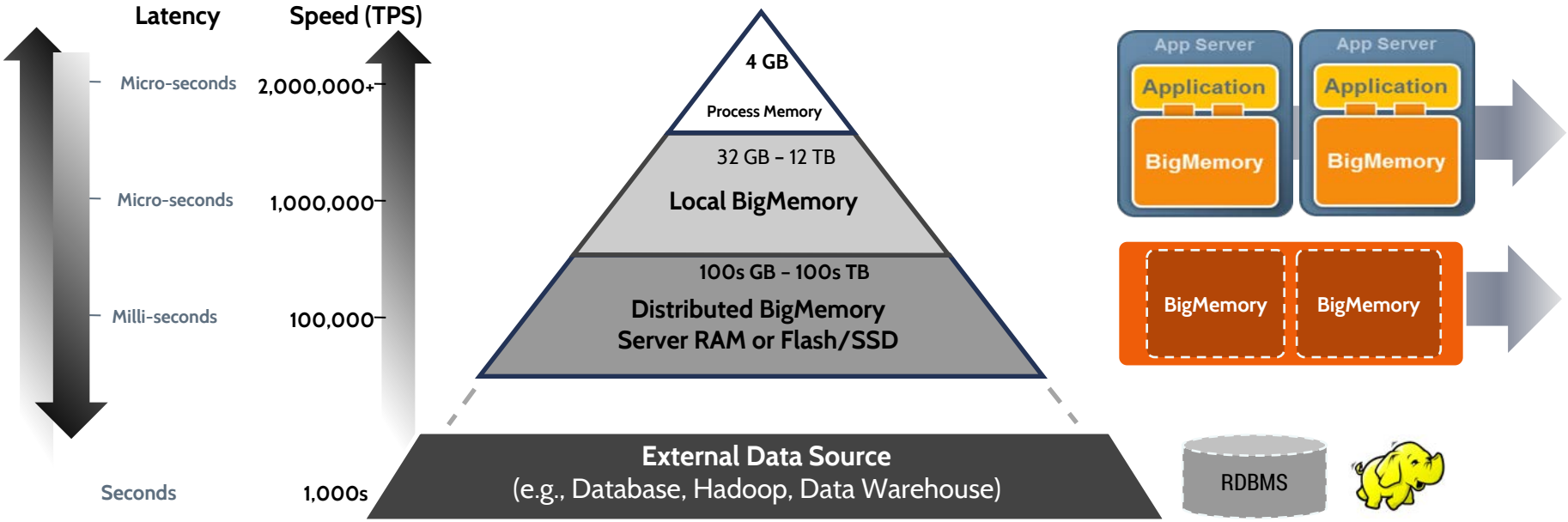
Explosion in volume and velocity of data

In-Memory Data Platforms

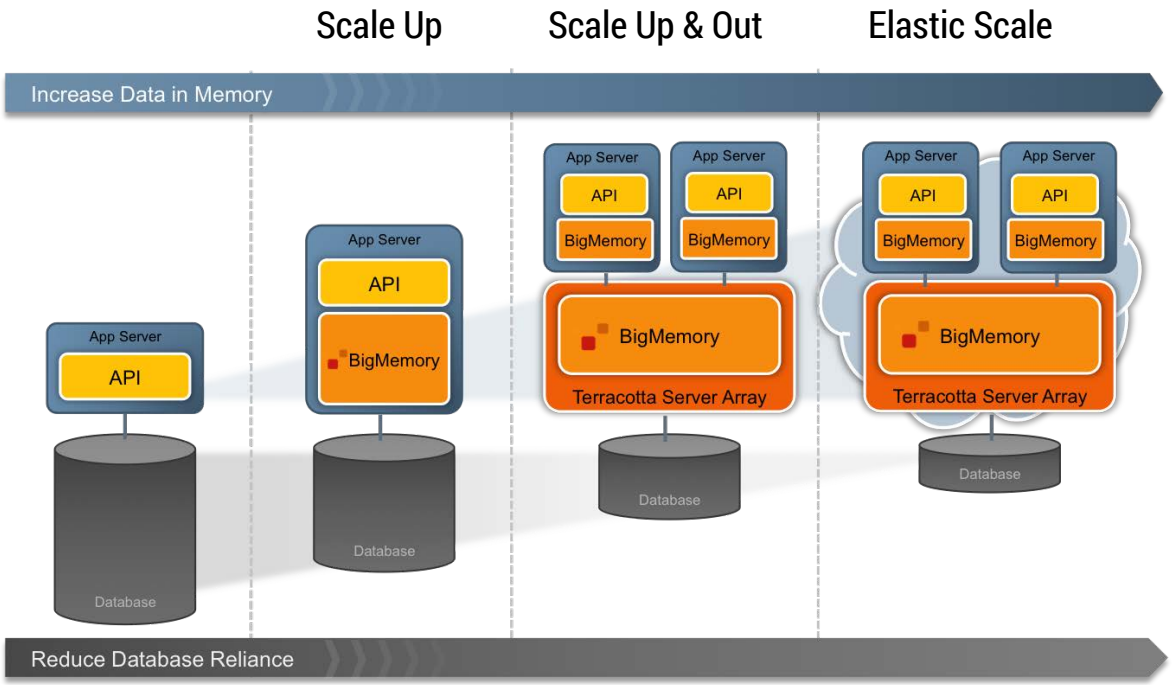
- Scale of NoSQL
- Low latency of In-Memory databases
- Reliability & Fault Tolerance
- Transactional Guarantees

Fast Big Data

Tiered Storage

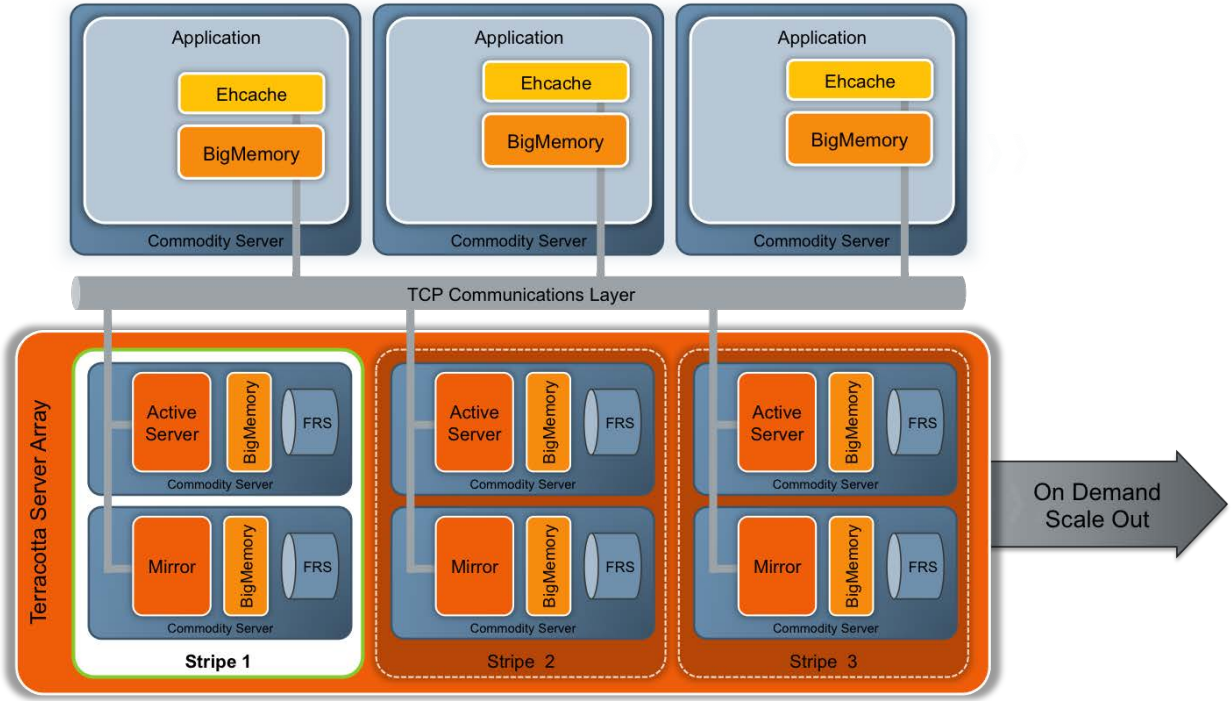


Scale with data and processing needs



HA, Extreme Resiliency

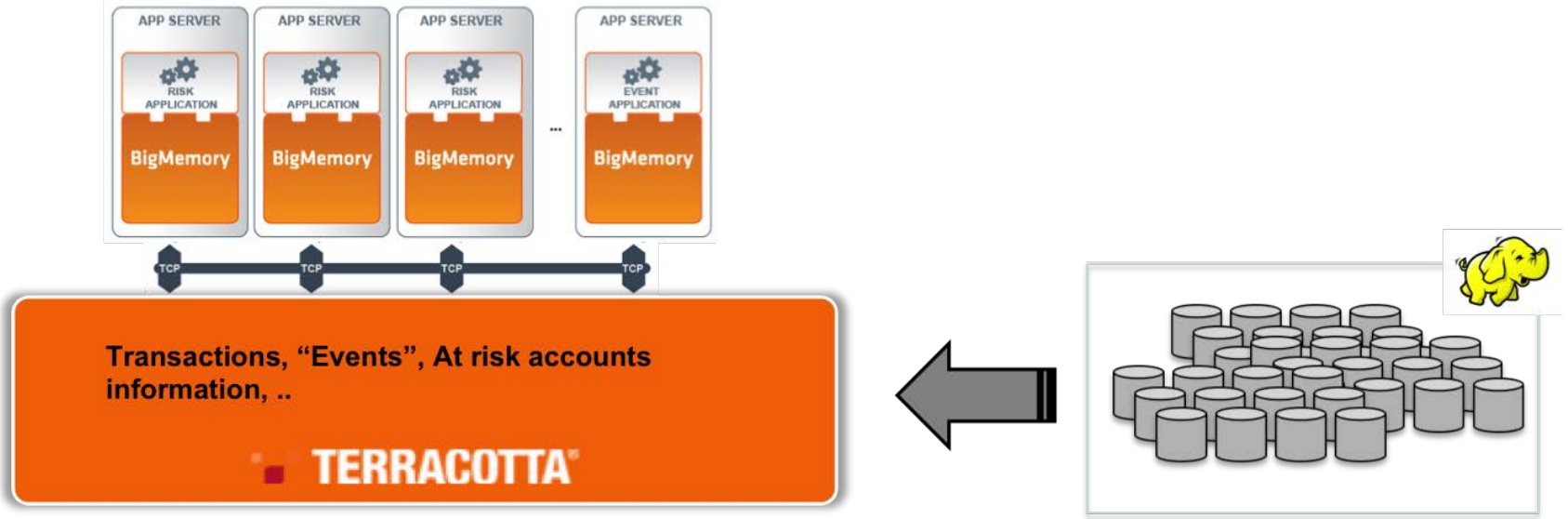
- Active Mirror
- No Single point of failure
- Fast Restartable Storage (SSD/Flash)



Use cases

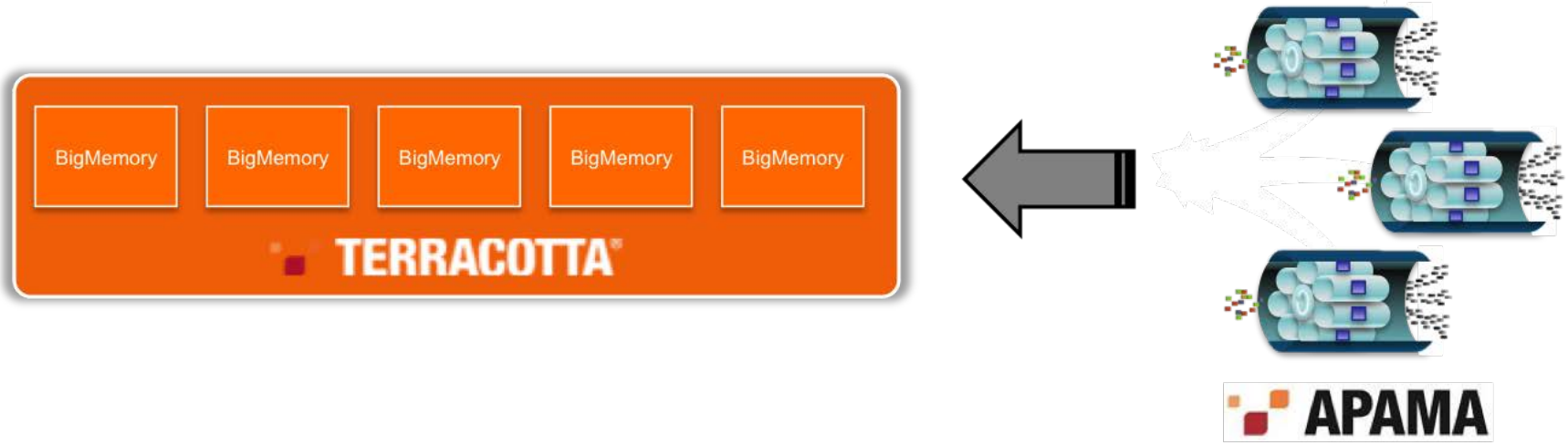
Influencing operations and decisions

In-Memory Data Fabric: Operationalize Hadoop*



Streaming insights into In-Memory Operational Store

In-Memory Data Fabric: Streaming Analytics



High Speed resilient data access across shared time windows


FORODWAITH
Angmar
Ered Mithrin
Iron Hills
The Lost Realm Of
ARNOR
HITHAEGLIR
Erebor
(The Lonely
Mountain)
Eriador
Lake
Forlindon
Weather Hills
Rhudaur
ARTHEDAIN
Old Forest
The Shire
Bree
Rivendell
Mirkwood
East Bight
RHÛN
Blue Mountains
Harlindon
ERIADOR
Eregion
Dur Guldur
MINHRIATH
Dunland
Lorien
RHOVANION
Fangorn
Isengard
The Brown Lands
ROHAN
Helms
Deep
Ered Lithui (Ash Mountians)
N. Ithilien
Orodruin
Barad dûr
Plateau of
Gorgoroth
MORDOR
GONDOR
NURN
Anfalas
Lebennin
Hadoop
Spark
In-Memory
NoSQL
NoSQL
The Big Data Land
#strataconf
#hadoopworld
Strata Hadoop
WORLD

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

Cryogenics

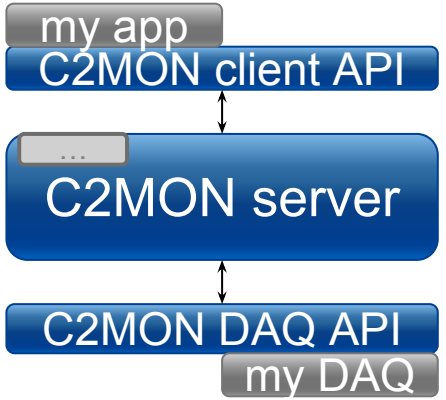
Electricity

Safety Systems

Cooling

C2MON - CERN Control and Monitoring Platform

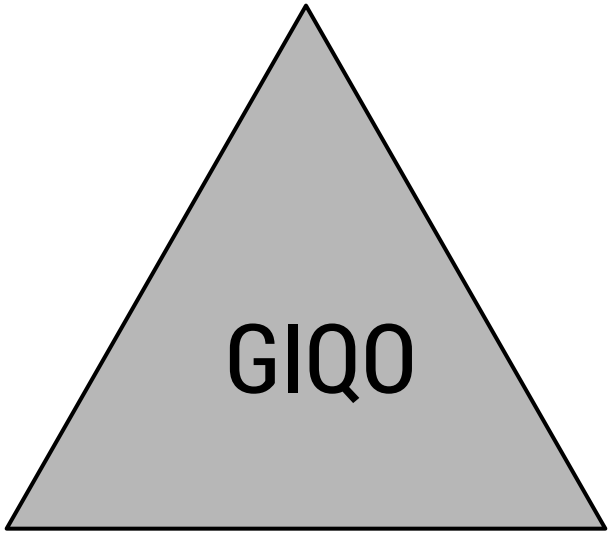
- Allows the rapid implementation of high-performance monitoring solutions
- Modular and **scalable at all layers**
- Optimized for High Availability & big data volume
- Based on In-Memory solution



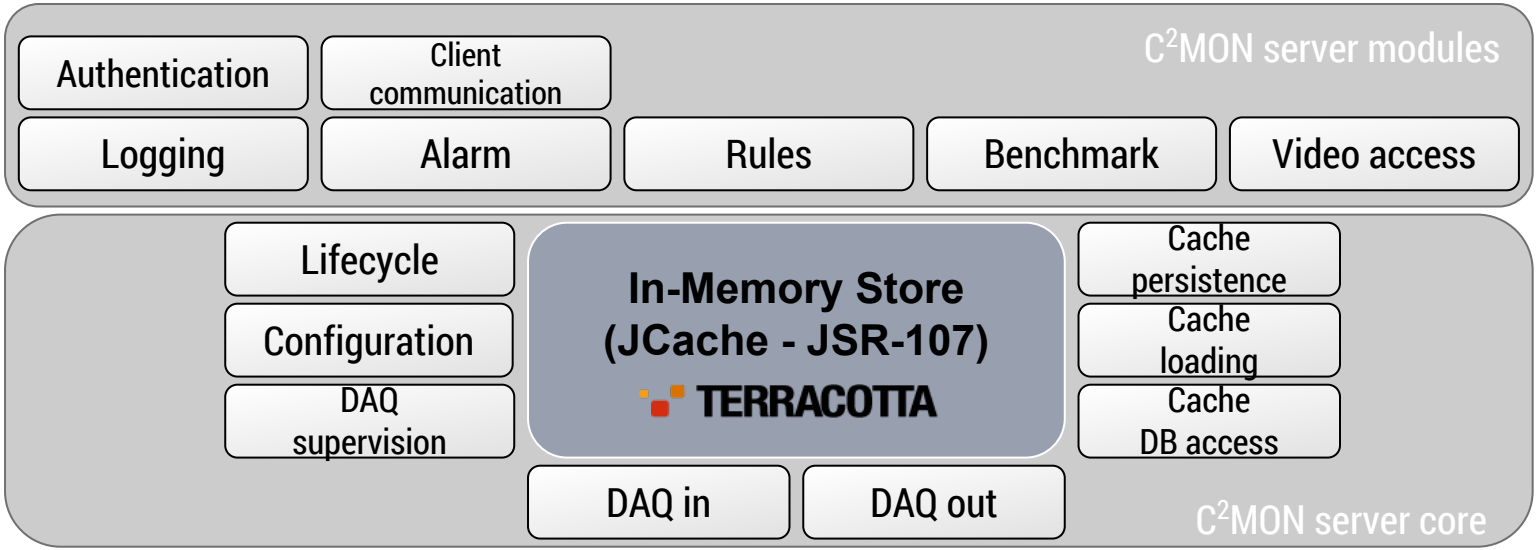
Currently used by two big systems at CERN: TIM & DIAMON

http://cern.ch/c2mon

Raw data filtering on DAQ layer



C2MON Server



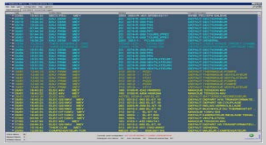
TIM – Technical Infrastructure Monitoring

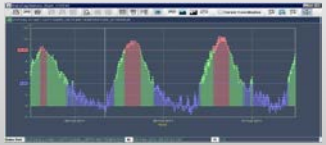
- Operational since 2005
- Used to monitor and control infrastructure at CERN
- **24/7** service
- ~ 100 different main users at CERN

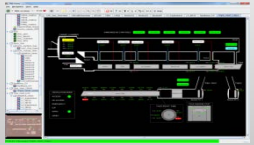
- Since Jan. 2012 based on
  new server architecture with C2MON

CERN Control Center at LHC startup

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

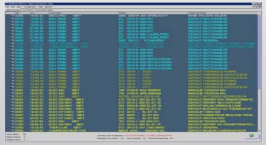

Electricity

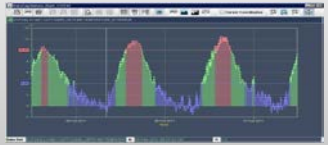

Access

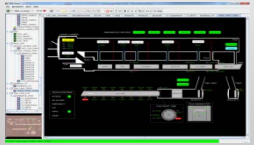

Network and Hardware Controls


Cryogenics

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

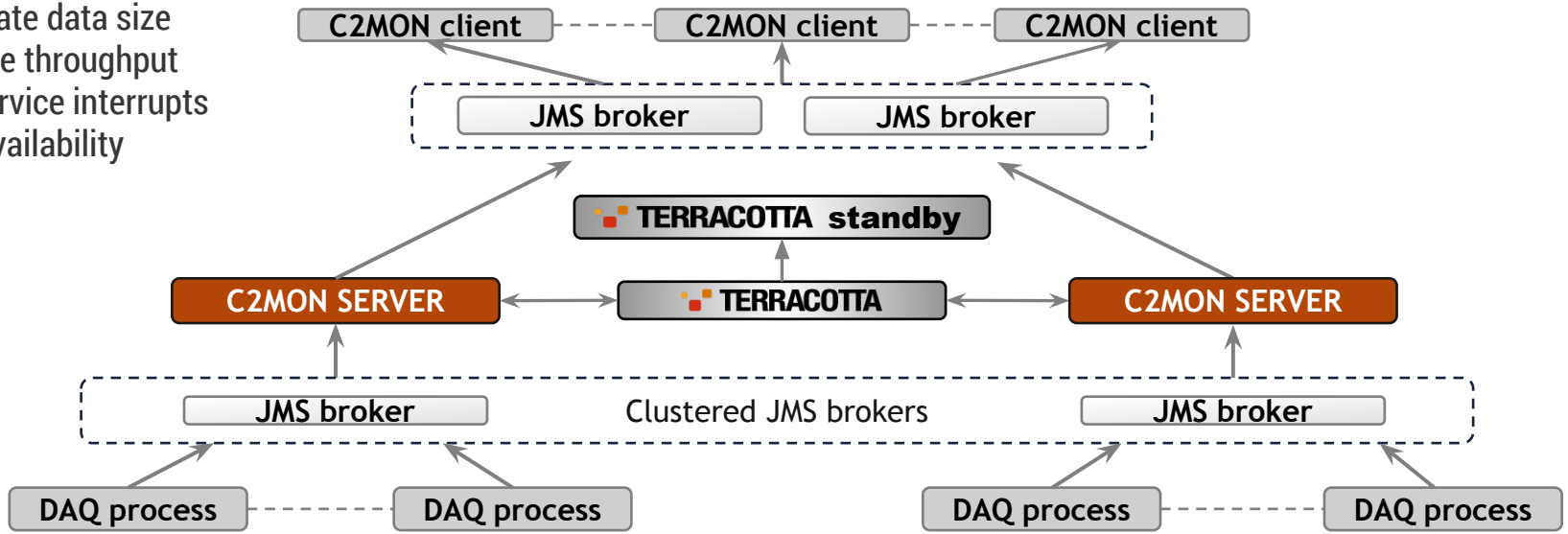
ca. **400 million**
raw data per day

Filtering

ca.**1.5 million updates**

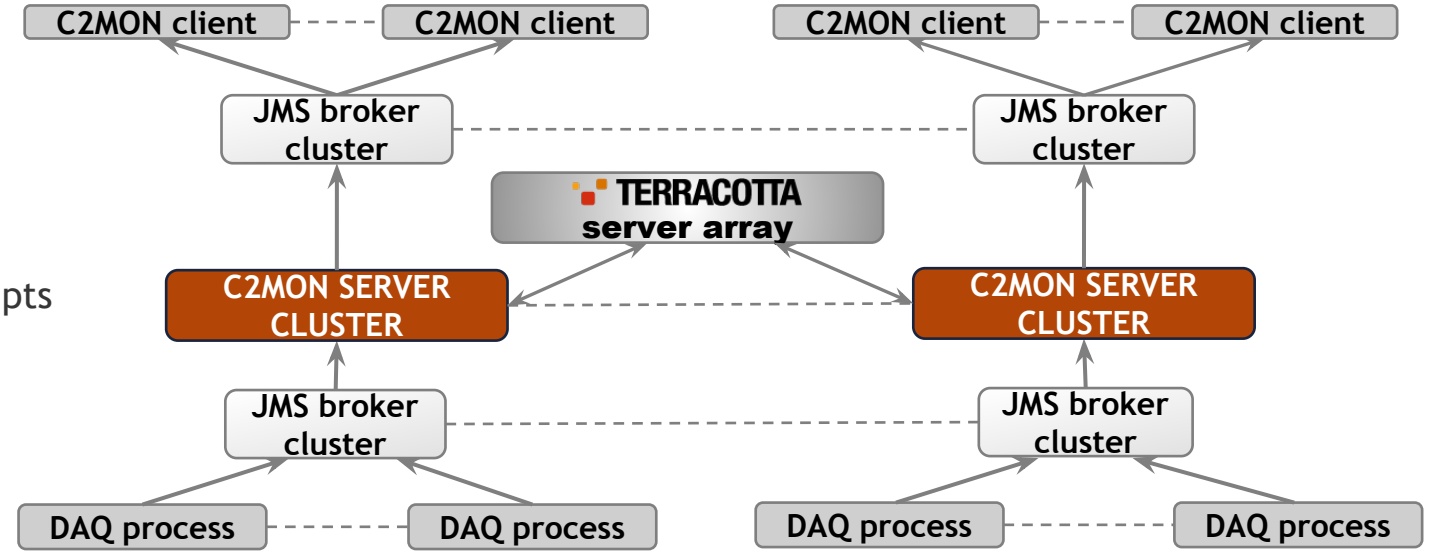
Scenario 1: High availability

- 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
- 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 Toebbicke (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!