

Database Services at CC

6th Feb 2019

Osman AÏDEL





Physics experiments























LHCb

2

RDBMS

- Mysql (ver. 5.6)
- MariaDB Cluster (ver. 10.2) with Galera (ver. 25.10)
- PostgreSQL (ver. 10.6)
- Oracle (ver. 11.2 -> 12 on progress)
- Nosql
 - Mongodb (ver. 3.6)
 - ElasticSearch
- Monitoring
- Spark



Mysql / Mariadb

- Shared platform
 - 6 galera clusters (3 nodes)
 - 12 Mariadb / Mysql standalone instances
 - 610 accounts
- Each instance is reachable through its own VIP
- Keepalived is used for security and high availabiliy
 - Kernel patch in rolling mode
 - Major / Minor upgrade in rolling mode
 - Node failures
- Daily backup :
 - Mysql : SQL Dump plus archived logs
 - Local copy with a retention period of 1 day
 - Tape storage : TSM with a retention period of 1 month
 - TSM : Tivoly Storage System is a backup solution on tape
- Some specific configurations :
 - One-way Replication (Hess Dchooz)





MYSQL/Mariadb



PostgreSQL architecture

- Shared platform
 - 12 PostgreSQL standalone instances
 - 76 accounts
- Each instance is reachable through its own VIP
- Resource manager (Pacemaker)
 - Kernel patch in rolling mode
 - Minor upgrade in rolling mode
 - Node failures
- Daily Backup :
 - SQL dump / binary copy
 - Local copy with a retention period of 1 day
 - Tape storage : TSM with a retention period of 1 month
- Some specific configurations :
 - UGE (batch system) slave on read only



PostgreSQL architecture



8

Database volumetry





server1	server2	server3	server4						
ATLAS									
AMI									

server7	server8	server9	server9		
SYM	IOD	BASTET			
AMON	MECA	NIQUE	AMON		
RM	AN	DEDWEN			
IRODS	DEV	11G	IRODS		

494 accounts



Real Application Cluster database

- N actives nodes simultenaously
- Clients only access to one public VIP
- Node failure does not impact the availability of the database
- Data is shared via Automatic Storage Management (ASM)
- ASM makes transparent the physical devices
 - Renewing of storage does not require downtime
 - Data may be replicated at logical level



Dataguard

- Oracle's standby database solution for disaster recovery
- Exact copy of the database production
- All clients can point to the standby db without any change at the application level
- Switchover very fast (some minutes)
- Major upgrade only requires some minutes of unavailability
- Severe crash on production might involve a small lost (few transactions)



- Golden Gate (streams)
 - Oracle replication solution for heterogeneous databases
 - Used for the Atlas metadata catalog (AMI)
 - Replication at user account / table level
 - DBAMI : one-way replication from CC-IN2P3 to CERN



Backup

- Recovery MANager
- RMAN Catalog
- Weekly full backup and incremental backup every day to TSM
- Retention depends on databases



OPERA Project

OPERA

- Oscillation Project with Emulsion tRacking Apparatus
- 34 institutes spread out 12 countries
- Decommissioning at CC on progress



MongoDB

- 3 Nodes in replication
 - Kernel patch in rolling mode
 - Minor upgrade in rolling mode
 - Node failures



- 3 nodes in replication
 - One active node and 2 passives
 - MongoDB driver manages connections in case of failure
 - Kernel patch in rolling mode
 - Major and Minor upgrade in rolling mode
 - Node failures
- Usage (12 accounts)
 - INEE (vigiechiro)
 - Openstack (ceilometer)
 - UGE (Batch system)
 - •
- Daily backup to TSM

Database volumetry



Nagios

Thruk									
General Home Documentation	Current I Last Update: Fri Feb 1 10 Thruk 2.26-2 Logged in as <i>Osman Aid</i>	Network Status 0.45:10 CET 2019 (=90s)					Host Star Up Down 35 0	tus Totals Unreachable Penfing 0 0	Service Status Totals OK Warning Unknown Critical Pending 64 5 8 0 0
Current Status Tactical Overview Map	View History For all hosts View Notifications For All Hosts View Host Status Detail For All Hosts						0	aa types 25	All Types 13 77
Hosts Services Host Groups Summary (Grid)	F Service Status Details For All Host Market Hand A Service Status Details for All Host								
Summary (Grid)				-	to a start of		select all (hosts) - unselect a	all - all problems - all with downtime	
Mine Map	Host AV	Service AV	1	Status 🗸	Last Check AV	11d 18b 40m 7a	Attempt A	Status Information AV	
Problems Services (Unhandled) Hosts (Unhandled) Network Outages		Chack mariadh churter even h	actor No.	UNICATION	01-00-02	#3d 9h #5m Re	1/1 #1	contract sources as up	
		Check mariadh cluster intra h	arian 🐴 1	UNKNOWN	00.00.02	43d 10h 45m 8s	1/1 #1	certhoa05 will not backup cluster intra'	
		Chack mariadh cluster my co	haring 4x -	OK.	2019-01-31 22:40:00	43d 11h 51m 44s	1/1	OK	
		Sampler post data	1 M		10:45:02	1d 20h 40m 5s	1/1	doc.count:114	
Reports	corboa06	Service status	1 64	OK	10:45:03	11d 18b 40m 7s	1/1	service status is 'uo'	
Availability		Check mariadb cluster dchoo	z01 backup 4	OK	03:05:21	69d 7h 37m 58s	1/1	OK C	
Trends Alerts		Check mariado cluster dohoo	z03 backup I	oK	02:41:00	69d 8h 1m 56s	1/1	96	
History (Summary) Notifications Event Log Business Process Reporting		Check mariadb cluster expe b	ackup 💫 🗎 🗒 🗖	UNKNOWN	01:00:02	98d 4h 35m 13s	1/1 #1	ccdbga06 will not backup cluster 'expe'	
		Check mariadb cluster intra b	ackup	OK	00.32.26	69d 10h 0m 26s	1/1	OK .	
		Check mariadb cluster my oc	backup 😜 🕹 🖬 🖬	UNKNOWN	2019-01-31 22:00:02	98d 4h 35m 13s	1/1 #1	ccdbga06 will not backup cluster 'my_cc'	
		Sampler post data	- 1	OK	10:45:02	2d 0h 40m 59s	1/1	doc_count:111	
System	ccdbga07	Service status	- <u>1</u>	OK	10:45:03	11d 18h 40m 7s	1/1	service status is 'up'	
Comments Downtimes Recurring Downtimes Performance Info Scheduling Queue		Check mariadb cluster ami bi	ickup 🕹	oĸ	01:37:06	43d 9h 0m 52s	1/1	OK	
		Check mariadb cluster dchoo	z02 backup	OK	2019-01-31 21:24:43	43d 13h 19m 27s	1/1	OK	
		Check mariadb cluster expe b	ackup 1	OK	09:05:28	43d 1h 6m 29s	1/1	ОК	
		Check mariadb cluster intra b	ackup 🕷 🚺	UNKNOWN	00:00:02	43d 10h 45m 8s	1/1 #1	ocdbga07 will not backup cluster 'Intra'	
Bookmarks		Check mariadb cluster my cc	backup 💘 🕹	UNKNOWN	2019-01-31 22:00:03	43d 12h 45m 8s	1/1 #1	ccdbga07 will not backup cluster 'my_cc'	
CCNAGIOS - service NOK		Sampler post data	¥0 ⊥	OK	10:45:02	1d 20h 40m 5s	1/1	doc_count:114	
SYSNAGIOS - service NOK	ccdbga08	Service status	1	OK	10:45:03	11d 18h 40m 7s	1/1	service status is 'up'	

Coloss



Grafana



Kibana interface



SPARK

The rise of new framework for BIG DATA / DATA SCIENCE.

Spark

- Large scale data processing
- High availability
- Best suited for batch applications that apply the same operation to all elements of a dataset
- Spark features
 - Mllib (Machine Learning lib)
 - Streaming
 - Spark SQL
 - GraphX



SPARK : data processing



SPARK : Use case

Storage metrics

Storage system



SPARK : Use case

- The main idea is to provide an unified view on storage systems for users, experiment manager, CC support, storage experts and direction.
- User view
 - How storage is used at the experiment or user level ?
 - Number of files, used space, file size
 - How files are distributed through CC storage systems
 - • •
- Experiment Manager
 - How much space is allocated / used by the experiment ?
 - Which user consumes too much storage ?
 - •••
- CC Support view
 - Users leaving an experiment do not cleanup their space. The support in collaboration with the experiment manager need to identify
 user files and what to do with it. Which files is involved ? Are they relevant ?
 - Optimize the storage use by locating files no more accessed for one month / year in order to spill them down to a cheap storage (ex. tape).
- Expert view
 - For some storage system, it may be interesting to know the number of links or empty files because the number of inodes is not unlimited.
 - Small and Large files have an impact on system performance and in resource consumption. Indeed, a system configured with large block are not adapted for small files, a small file could consume more space it requires. It's why, it is important for an expert to know the distribution of file size n order to adapt the best block size in order to minimize the wasted space.
- Direction view
 - How much space is allocated to each experiment ?
 - How files are spread through CC storage systems ?
 - Do we respect experiment agreements

Storage metrics

Storage system



QUESTIONS?

