Migration of Gratia Active Archive Project Accounting to GRACC

Mariano Basile

28/09/2017

Supervisor: Kevin Retzke

In partnership with:

UNIVERSITÀ DI PISA

The Active Archive Project

The Active Archive Facility

- Fermilab provides a custodial active archive for customers, either on-site and not, for long-term storage of hundreds of petabytes of scientific data.
- The archive facility at Fermilab is capable of providing access to these data over a 100 Gb/s network.
- Tape storage is assumed to be the permanent means of custodial for data. Customers can also opt for a dedicated disk cache integrated with the tape storage.
- The Active Archive Facility is indeed an hierarchical storage system consisting of **tape storage**, **Enstore**, with a front-end **disk cache** called **dCache.**

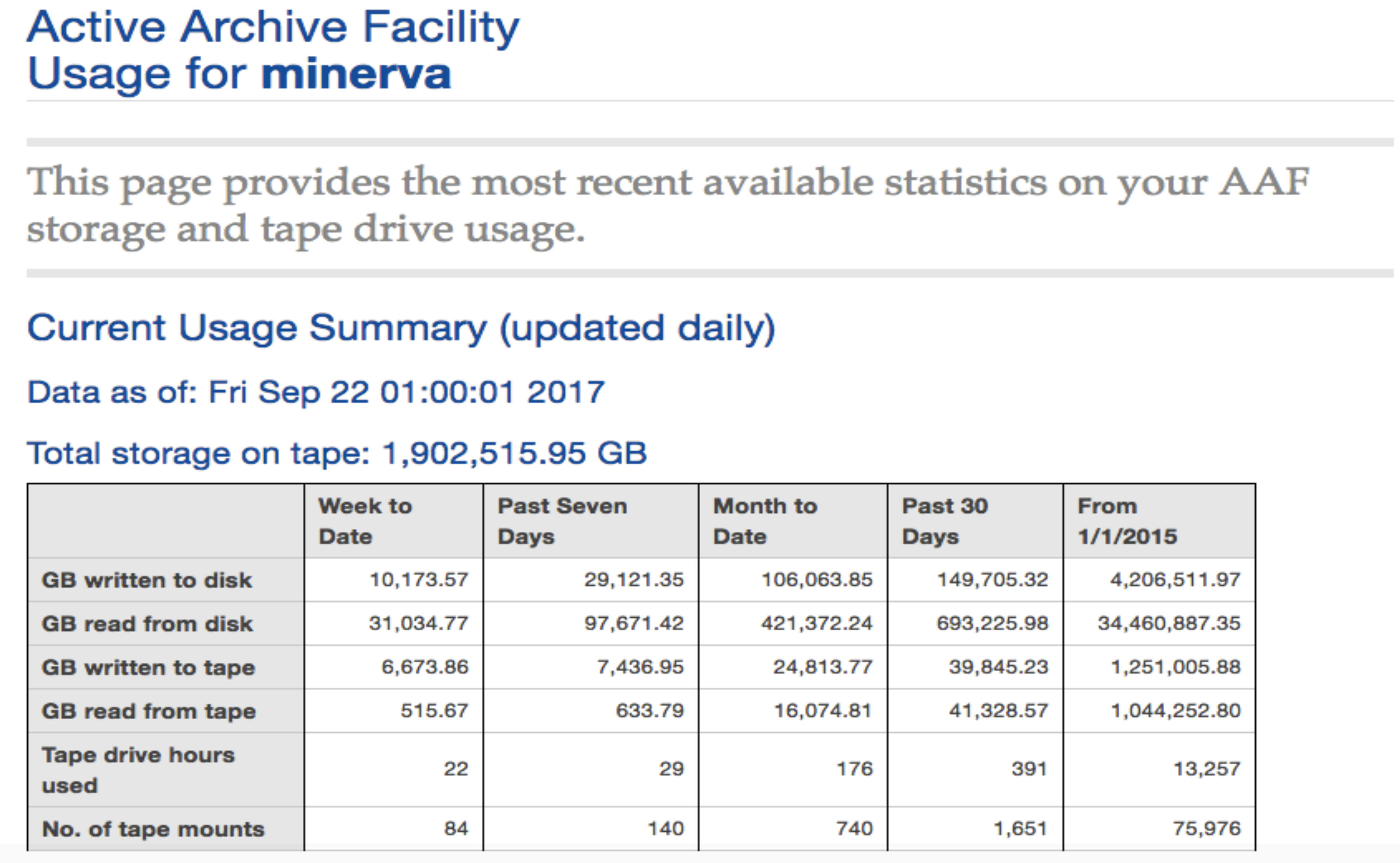
The Active Archive Facility:(I)

- Daily and monthly usage metrics to customers have been provided so far by means of the GRATIA accounting system.

- AAF usage metrics consist of:
  - *amount of data read/written from/to disks*
  - *amount of data read/written from/to tapes*
  - *total amount of data on tapes*
  - *No. of tape mounts and tape drive hours used*

- Usage metrics are provided to the customers through a web portal: https://archive.fnal.gov/

Gratia Active Archive Project: Web Portal



Active Archive Facility Usage for **minerva**

This page provides the most recent available statistics on your AAF storage and tape drive usage.

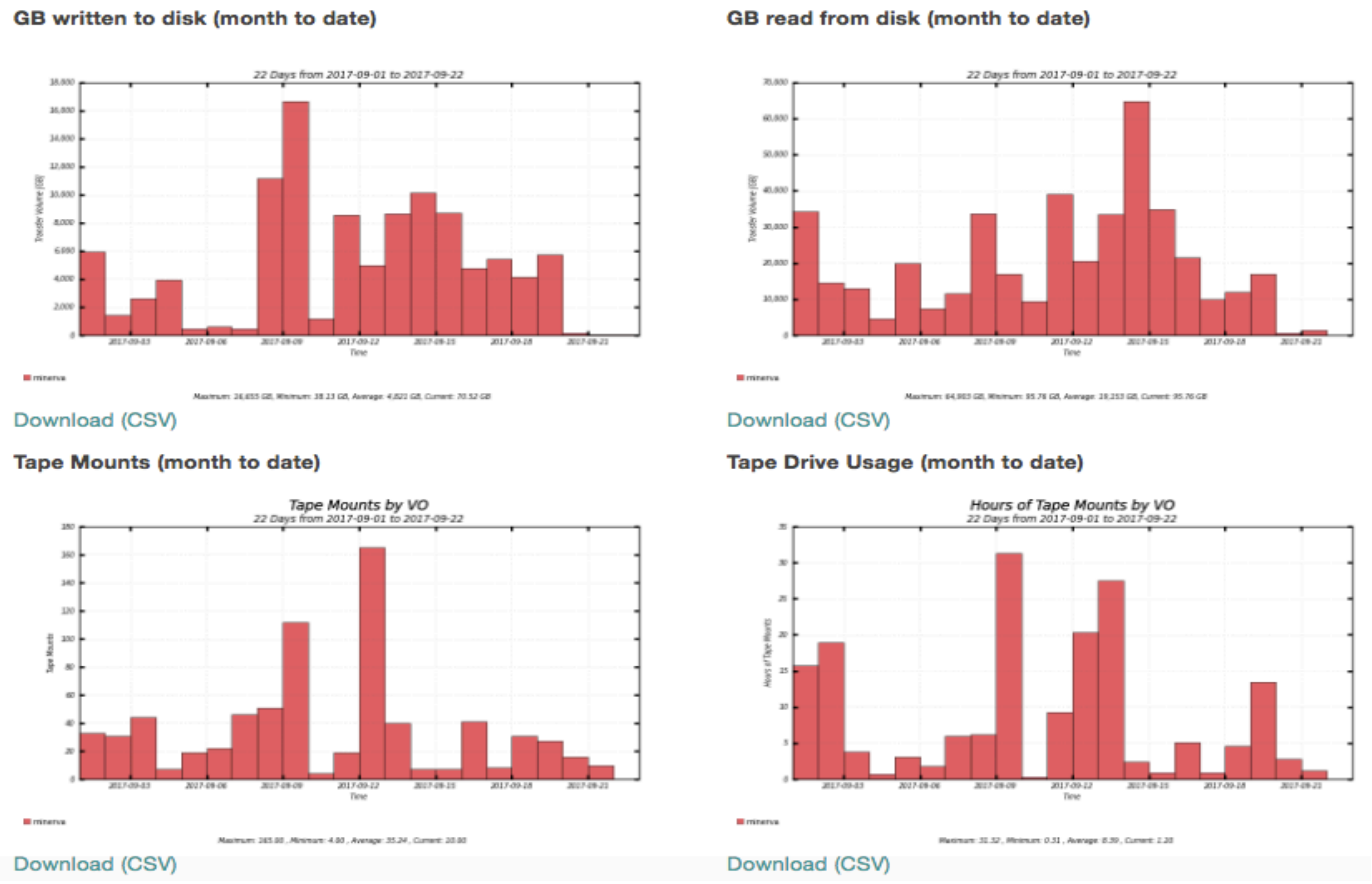
Current Usage Summary (updated daily)

Data as of: Fri Sep 22 01:00:01 2017

Total storage on tape: 1,902,515.95 GB

	Week to Date	Past Seven Days	Month to Date	Past 30 Days	From 1/1/2015
GB written to disk	10,173.57	29,121.35	106,063.85	149,705.32	4,206,511.97
GB read from disk	31,034.77	97,671.42	421,372.24	693,225.98	34,460,887.35
GB written to tape	6,673.86	7,436.95	24,813.77	39,845.23	1,251,005.88
GB read from tape	515.67	633.79	16,074.81	41,328.57	1,044,252.80
Tape drive hours used	22	29	176	391	13,257
No. of tape mounts	84	140	740	1,651	75,976

Gratia Active Archive Project: Web Portal



GB written to disk (month to date)

Maximum: 16,655 GB, Minimum: 38.13 GB, Average: 4,821 GB, Current: 70.52 GB

Download (CSV)

GB read from disk (month to date)

Maximum: 64,903 GB, Minimum: 95.76 GB, Average: 19,253 GB, Current: 95.76 GB

Download (CSV)

Tape Mounts (month to date)

Maximum: 165.00 , Minimum: 4.00 , Average: 35.24 , Current: 10.00

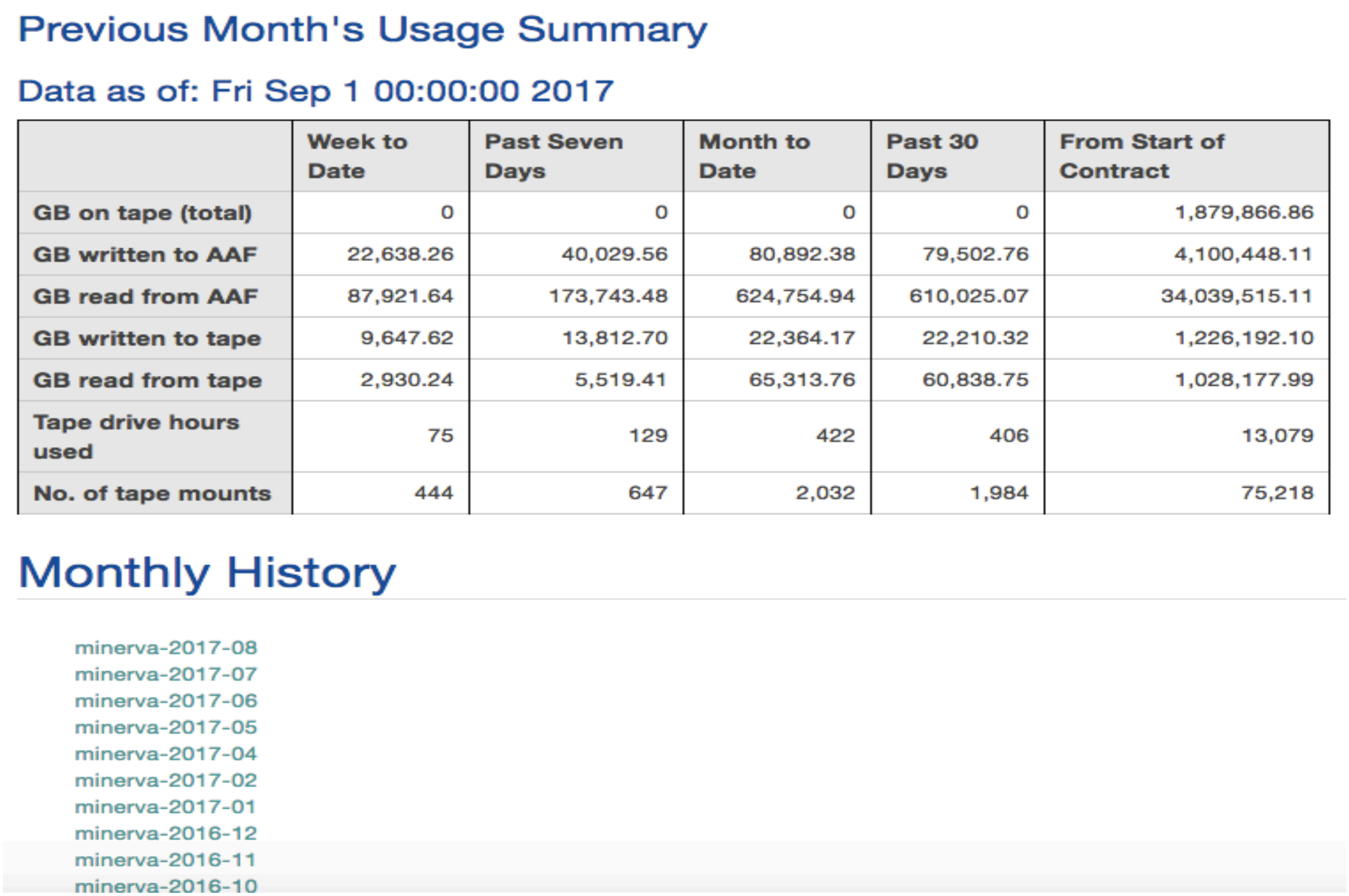
Download (CSV)

Tape Drive Usage (month to date)

Maximum: 31.32 , Minimum: 0.31 , Average: 8.39 , Current: 1.28

Download (CSV)

Gratia Active Archive Project: Web Portal



Previous Month's Usage Summary

Data as of: Fri Sep 1 00:00:00 2017

	Week to Date	Past Seven Days	Month to Date	Past 30 Days	From Start of Contract
GB on tape (total)	0	0	0	0	1,879,866.86
GB written to AAF	22,638.26	40,029.56	80,892.38	79,502.76	4,100,448.11
GB read from AAF	87,921.64	173,743.48	624,754.94	610,025.07	34,039,515.11
GB written to tape	9,647.62	13,812.70	22,364.17	22,210.32	1,226,192.10
GB read from tape	2,930.24	5,519.41	65,313.76	60,838.75	1,028,177.99
Tape drive hours used	75	129	422	406	13,079
No. of tape mounts	444	647	2,032	1,984	75,218

Monthly History

minerva-2017-08
minerva-2017-07
minerva-2017-06
minerva-2017-05
minerva-2017-04
minerva-2017-02
minerva-2017-01
minerva-2016-12
minerva-2016-11
minerva-2016-10

The Active Archive Facility:(II)

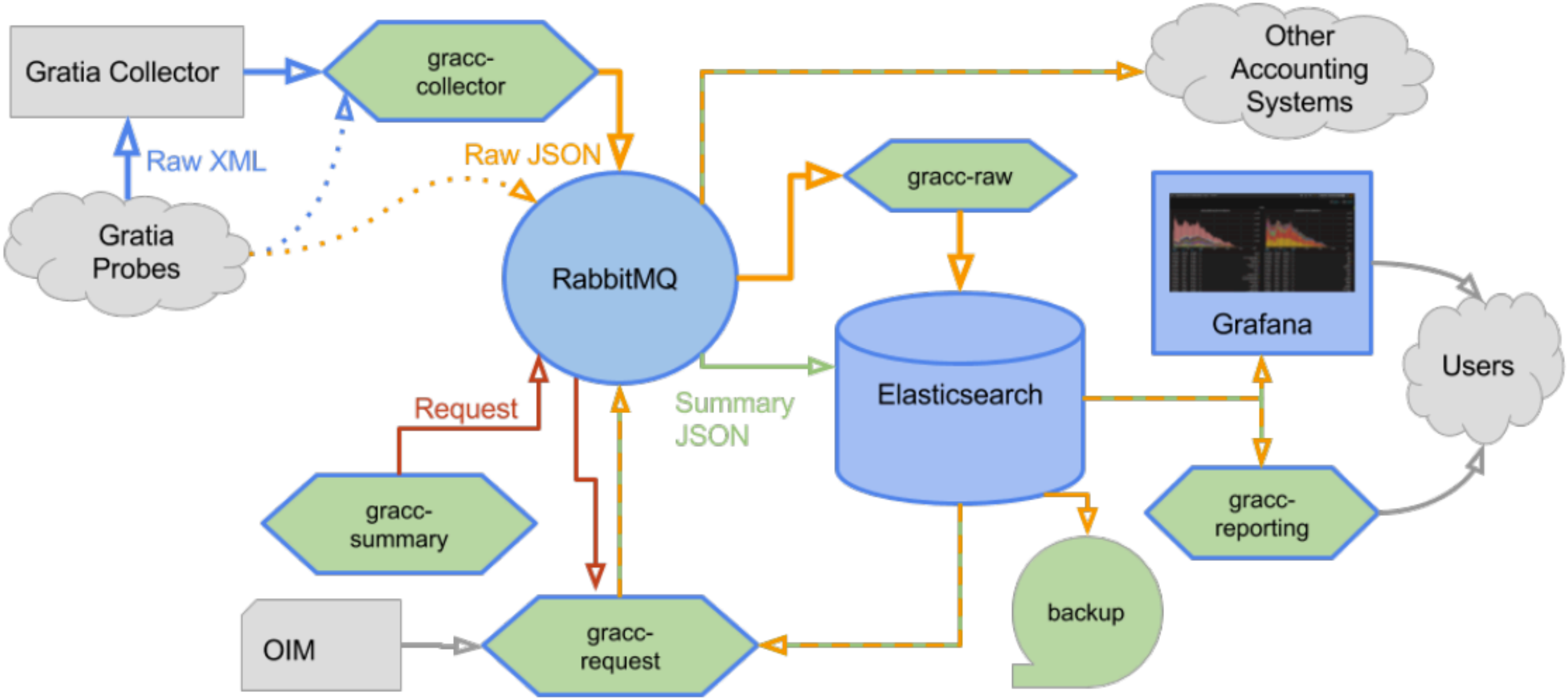
Because of the GRATIA shutdown, a migration of the active archive project to GRACC, the new Grid ACCounting system was required.

GRACC: New Generation of the OSG Accounting

GRACC: New Generation of the OSG Accounting

- The requirements laid out by the OSG technical committee steered the GRACC investigation towards a small set of existing technologies:
  - **Elasticsearch** as a data storage and query platform as it provides a fault-tolerant distributed architecture, flexible schema, partitionable indices, and powerful search and aggregations.
  - **RabbitMQ** for data exchange between services
  - **Grafana** for the primary user interface. It supports a rich ecosystem of data sources and graph plugins and provides a powerful interface to create and share dashboards.
  - **Kibana** for ad-hoc analytics
  - **Prometheus**, as the monitoring platform: system and service monitoring is included in GRACC as a first-class citizen, as it is critical to understand the performance and limitations of the system.

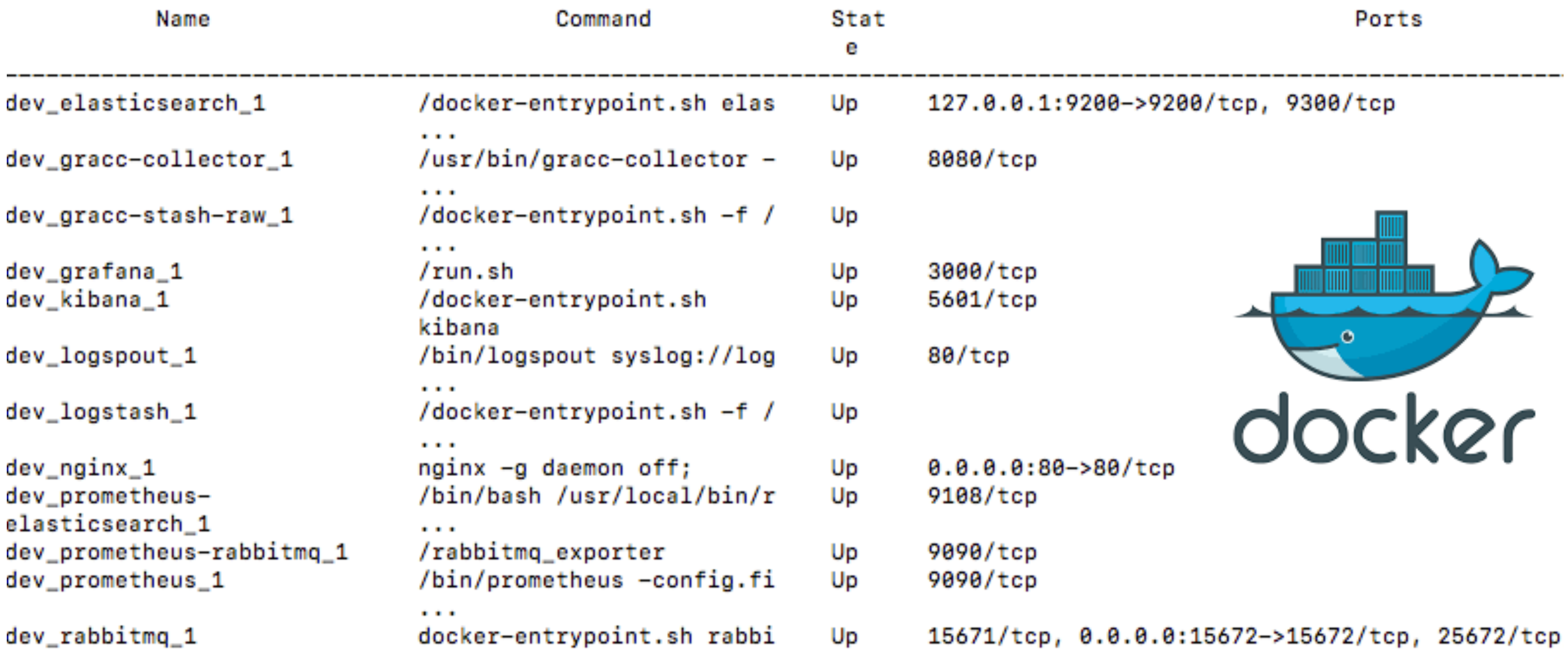
GRACC Architecture Overview



Migration Of AAF To GRACC

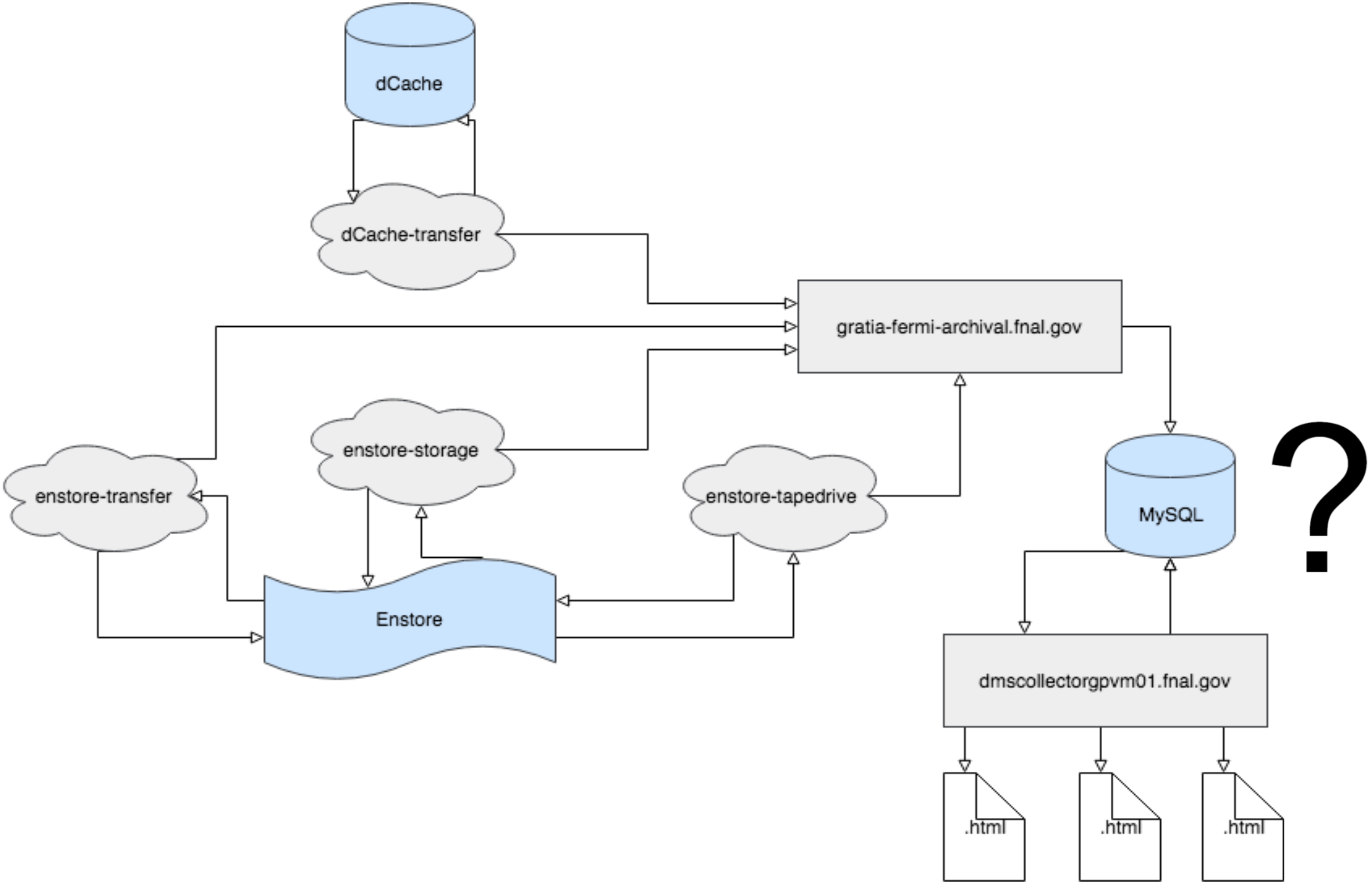
Migration of AAF To GRACC: Dev Environment

- A remote FermiCloud VM has been used as dev environment
- The entire GRACC stack has been deployed *as within containerized environment with docker.*

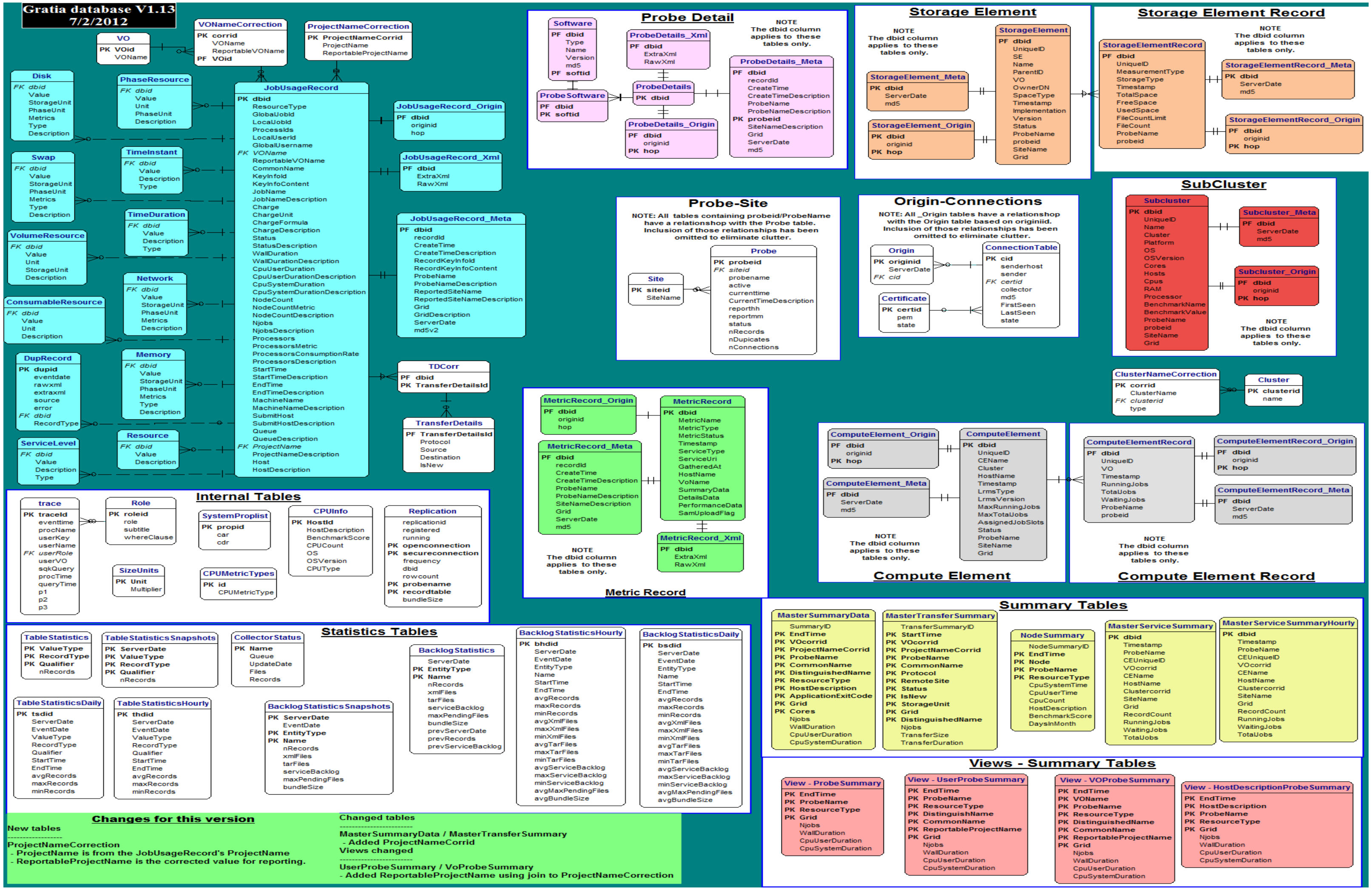


```
             Name                         Command                Stat                 Ports
                                                                  e
----------------------------------------------------------------------------------------------------------------------
dev_elasticsearch_1           /docker-entrypoint.sh elas   Up      127.0.0.1:9200->9200/tcp, 9300/tcp
                              ...
dev_gracc-collector_1         /usr/bin/gracc-collector -   Up      8080/tcp
                              ...
dev_gracc-stash-raw_1         /docker-entrypoint.sh -f /   Up
                              ...
dev_grafana_1                 /run.sh                      Up      3000/tcp
dev_kibana_1                  /docker-entrypoint.sh        Up      5601/tcp
                              kibana
dev_logspout_1                /bin/logspout syslog://log   Up      80/tcp
                              ...
dev_logstash_1                /docker-entrypoint.sh -f /   Up
                              ...
dev_nginx_1                   nginx -g daemon off;         Up      0.0.0.0:80->80/tcp
dev_prometheus-               /bin/bash /usr/local/bin/r   Up      9108/tcp
elasticsearch_1               ...
dev_prometheus-rabbitmq_1     /rabbitmq_exporter           Up      9090/tcp
dev_prometheus_1              /bin/prometheus -config.fi   Up      9090/tcp
                              ...
dev_rabbitmq_1                docker-entrypoint.sh rabbi   Up      15671/tcp, 0.0.0.0:15672->15672/tcp, 25672/tcp
```

Migration of AAF To GRACC: AAF In A Nutshell



Migration of AAF To GRACC: Data Migration



Gratia database V1.13
7/2/2012

VO
PK VOid
VOName

VONameCorrection
PK corrid
VOName
ReportableVOName
PF VOid

ProjectNameCorrection
PK ProjectNameCorrid
ProjectName
ReportableProjectName

Disk
FK dbid
Value
StorageUnit
PhaseUnit
Metrics
Type
Description

PhaseResource
FK dbid
Value
Unit
PhaseUnit
Description

Swap
FK dbid
Value
StorageUnit
PhaseUnit
Metrics
Type
Description

TimeInstant
FK dbid
Value
Description
Type

TimeDuration
FK dbid
Value
Description
Type

VolumeResource
FK dbid
Value
Unit
StorageUnit
Description

Network
FK dbid
Value
StorageUnit
PhaseUnit
Metrics
Description

ConsumableResource
FK dbid
Value
Unit
Description

DupRecord
PK dupid
eventdate
rawxml
extraxml
source
error
FK dbid
RecordType

Memory
FK dbid
Value
StorageUnit
PhaseUnit
Metrics
Type
Description

ServiceLevel
FK dbid
Value
Description
Type

Resource
FK dbid
Value
Description

JobUsageRecord
PK dbid
ResourceType
GlobalJobId
LocalJobId
ProcessIds
LocalUserId
GlobalUsername
FK VOName
ReportableVOName
CommonName
KeyInfoId
KeyInfoContent
JobName
JobNameDescription
Charge
ChargeUnit
ChargeFormula
ChargeDescription
Status
StatusDescription
WallDuration
WallDurationDescription
CpuUserDuration
CpuUserDurationDescription
CpuSystemDuration
CpuSystemDurationDescription
NodeCount
NodeCountMetric
NodeCountDescription
Njobs
NjobsDescription
Processors
ProcessorsMetric
ProcessorsConsumptionRate
ProcessorsDescription
StartTime
StartTimeDescription
EndTime
EndTimeDescription
MachineName
MachineNameDescription
SubmitHost
SubmitHostDescription
Queue
QueueDescription
FK ProjectName
ProjectNameDescription
Host
HostDescription

JobUsageRecord_Origin
PF dbid
originid
hop

JobUsageRecord_Xml
PF dbid
ExtraXml
RawXml

JobUsageRecord_Meta
PF dbid
recordId
CreateTime
CreateTimeDescription
RecordKeyInfoId
RecordKeyInfoContent
ProbeName
ProbeNameDescription
ReportedSiteName
ReportedSiteNameDescription
Grid
GridDescription
ServerDate
md5v2

TDCorr
PF dbid
PK TransferDetailsId

TransferDetails
PF TransferDetailsId
Protocol
Source
Destination
IsNew

Probe Detail

NOTE
The dbid column applies to these tables only.

Software
PF dbid
Type
Name
Version
md5
PF softid

ProbeSoftware
PF dbid
PK softid

ProbeDetails_Xml
PF dbid
ExtraXml
RawXml

ProbeDetails
PK dbid

ProbeDetails_Origin
PF dbid
originid
PK hop

ProbeDetails_Meta
PF dbid
recordId
CreateTime
CreateTimeDescription
ProbeName
ProbeNameDescription
PK probeid
SiteNameDescription
Grid
ServerDate
md5

Storage Element

NOTE
The dbid column applies to these tables only.

StorageElement_Meta
PK dbid
ServerDate
md5

StorageElement_Origin
PK dbid
originid
PK hop

StorageElement
PF dbid
UniqueID
SE
Name
ParentID
VO
OwnerDN
SpaceType
Timestamp
Implementation
Version
Status
ProbeName
probeid
SiteName
Grid

Storage Element Record

NOTE
The dbid column applies to these tables only.

StorageElementRecord
PF dbid
UniqueID
MeasurementType
StorageType
Timestamp
TotalSpace
FreeSpace
UsedSpace
FileCountLimit
FileCount
ProbeName
probeid

StorageElementRecord_Meta
PK dbid
ServerDate
md5

StorageElementRecord_Origin
PF dbid
originid
PK hop

Probe-Site

NOTE: All tables containing probeid/ProbeName have a relationshop with the Probe table. Inclusion of those relationships has been omitted to eliminate clutter.

Site
PK siteid
SiteName

Probe
PK probeid
FK siteid
probename
active
currenttime
CurrentTimeDescription
reporthh
reportmm
status
nRecords
nDupicates
nConnections

Origin-Connections

NOTE: All _Origin tables have a relationshop with the Origin table based on originiid. Inclusion of those relationships has been omitted to eliminate clutter.

Origin
PK originid
ServerDate
FK cid

Certificate
PK certid
pem
state

ConnectionTable
PK cid
senderhost
sender
FK certid
collector
md5
FirstSeen
LastSeen
state

SubCluster

Subcluster
PK dbid
UniqueID
Name
Cluster
Platform
OS
OSVersion
Cores
Hosts
Cpus
RAM
Processor
BenchmarkName
BenchmarkValue
ProbeName
probeid
SiteName
Grid

Subcluster_Meta
PF dbid
ServerDate
md5

Subcluster_Origin
PF dbid
originid
PK hop

NOTE
The dbid column applies to these tables only.

ClusterNameCorrection
PK corrid
ClusterName
FK clusterid
type

Cluster
PK clusterid
name

Metric Record

MetricRecord_Origin
PF dbid
originid
hop

MetricRecord_Meta
PF dbid
recordId
CreateTime
CreateTimeDescription
ProbeName
ProbeNameDescription
SiteNameDescription
Grid
ServerDate
md5

MetricRecord
PK dbid
MetricName
MetricType
MetricStatus
Timestamp
ServiceType
ServiceUri
GatheredAt
HostName
VoName
SummaryData
DetailsData
PerformanceData
SamUploadFlag

MetricRecord_Xml
PF dbid
ExtraXml
RawXml

NOTE
The dbid column applies to these tables only.

Compute Element

ComputeElement_Origin
PF dbid
originid
PK hop

ComputeElement_Meta
PF dbid
ServerDate
md5

ComputeElement
PK dbid
UniqueID
CEName
Cluster
HostName
Timestamp
LrmsType
LrmsVersion
MaxRunningJobs
MaxTotalJobs
AssignedJobSlots
Status
ProbeName
SiteName
Grid

NOTE
The dbid column applies to these tables only.

Compute Element Record

ComputeElementRecord
PF dbid
UniqueID
VO
Timestamp
RunningJobs
TotalJobs
WaitingJobs
ProbeName
probeid

ComputeElementRecord_Origin
PF dbid
originid
PK hop

ComputeElementRecord_Meta
PF dbid
ServerDate
md5

NOTE
The dbid column applies to these tables only.

Internal Tables

trace
PK traceId
eventtime
procName
userKey
userName
FK userRole
userVO
sqkQuery
procTime
queryTime
p1
p2
p3

Role
PK roleid
role
subtitle
whereClause

SizeUnits
PK Unit
Multiplier

SystemProplist
PK propid
car
cdr

CPUMetricTypes
PK id
CPUMetricType

CPUInfo
PK HostId
HostDescription
BenchmarkScore
CPUCount
OS
OSVersion
CPUType

Replication
replicationid
registered
running
PK openconnection
PK secureconnection
frequency
dbid
rowcount
PK probename
PK recordtable
bundleSize

Statistics Tables

TableStatistics
PK ValueType
PK RecordType
PK Qualifier
nRecords

TableStatisticsSnapshots
PK ServerDate
PK ValueType
PK RecordType
PK Qualifier
nRecords

CollectorStatus
PK Name
Queue
UpdateDate
Files
Records

TableStatisticsDaily
PK tsdid
ServerDate
EventDate
ValueType
RecordType
Qualifier
StartTime
EndTime
avgRecords
maxRecords
minRecords

TableStatisticsHourly
PK thdid
ServerDate
EventDate
ValueType
RecordType
Qualifier
StartTime
EndTime
avgRecords
maxRecords
minRecords

BacklogStatisticsSnapshots
PK ServerDate
EventDate
PK EntityType
PK Name
nRecords
xmlFiles
tarFiles
serviceBacklog
maxPendingFiles
bundleSize

BacklogStatistics
ServerDate
PK EntityType
PK Name
nRecords
xmlFiles
tarFiles
serviceBacklog
maxPendingFiles
bundleSize
prevServerDate
prevRecords
prevServiceBacklog

BacklogStatisticsHourly
PK bhdid
ServerDate
EventDate
EntityType
Name
StartTime
EndTime
avgRecords
maxRecords
minRecords
avgXmlFiles
maxXmlFiles
minXmlFiles
avgTarFiles
maxTarFiles
minTarFiles
avgServiceBacklog
maxServiceBacklog
minServiceBacklog
avgMaxPendingFiles
avgBundleSize

BacklogStatisticsDaily
PK bsdid
ServerDate
EventDate
EntityType
Name
StartTime
EndTime
avgRecords
maxRecords
minRecords
avgXmlFiles
maxXmlFiles
minXmlFiles
avgTarFiles
maxTarFiles
minTarFiles
avgServiceBacklog
maxServiceBacklog
minServiceBacklog
avgMaxPendingFiles
avgBundleSize

Summary Tables

MasterSummaryData
SummaryID
PK EndTime
PK VOcorrid
PK ProjectNameCorrid
PK ProbeName
PK CommonName
PK DistinguishedName
PK ResourceType
PK HostDescription
PK ApplicationExitCode
PK Grid
PK Cores
Njobs
WallDuration
CpuUserDuration
CpuSystemDuration

MasterTransferSummary
TransferSummaryID
PK StartTime
PK VOcorrid
PK ProjectNameCorrid
PK ProbeName
PK CommonName
PK Protocol
PK RemoteSite
PK Status
PK IsNew
PK StorageUnit
PK Grid
PK DistinguishedName
Njobs
TransferSize
TransferDuration

NodeSummary
NodeSummaryID
PK EndTime
PK Node
PK ProbeName
PK ResourceType
CpuSystemTime
CpuUserTime
CpuCount
HostDescription
BenchmarkScore
DaysInMonth

MasterServiceSummary
PK dbid
Timestamp
ProbeName
CEUniqueID
VOcorrid
CEName
HostName
Clustercorrid
SiteName
Grid
RecordCount
RunningJobs
WaitingJobs
TotalJobs

MasterServiceSummaryHourly
PK dbid
Timestamp
ProbeName
CEUniqueID
VOcorrid
CEName
HostName
Clustercorrid
SiteName
Grid
RecordCount
RunningJobs
WaitingJobs
TotalJobs

Views - Summary Tables

View - ProbeSummary
PK EndTime
PK ProbeName
PK ResourceType
PK Grid
Njobs
WallDuration
CpuUserDuration
CpuSystemDuration

View - UserProbeSummary
PK EndTime
PK ProbeName
PK ResourceType
PK DistinguishName
PK CommonName
PK ReportableProjectName
PK Grid
Njobs
WallDuration
CpuUserDuration
CpuSystemDuration

View - VOProbeSummary
PK EndTime
PK VOName
PK ProbeName
PK ResourceType
PK DistinguishedName
PK CommonName
PK ReportableProjectName
PK Grid
Njobs
WallDuration
CpuUserDuration
CpuSystemDuration

View - HostDescriptionProbeSummary
PK EndTime
PK HostDescription
PK ProbeName
PK ResourceType
PK Grid
Njobs
WallDuration
CpuUserDuration
CpuSystemDuration

Changes for this version

New tables

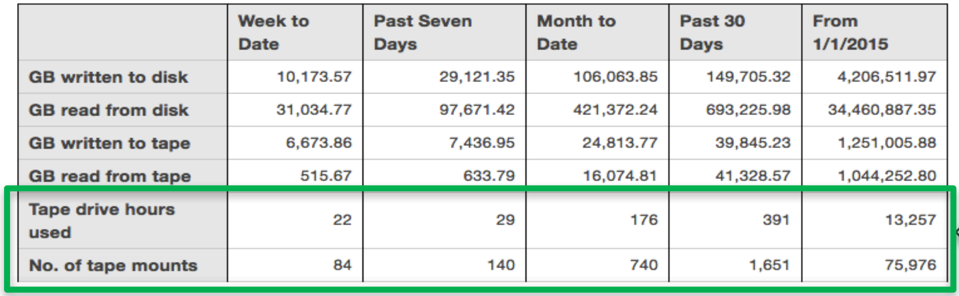
ProjectNameCorrection
- ProjectName is from the JobUsageRecord's ProjectName
- ReportableProjectName is the corrected value for reporting.

Changed tables

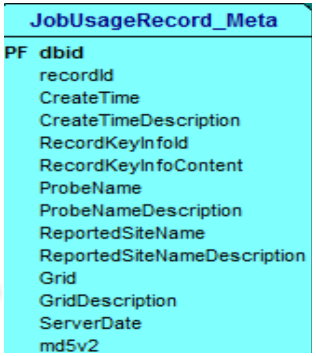
MasterSummaryData / MasterTransferSummary
- Added ProjectNameCorrid
Views changed

UserProbeSummary / VoProbeSummary
- Added ReportableProjectName using join to ProjectNameCorrection

Migration of AAF To GRACC: Data To Migrate



	Week to Date	Past Seven Days	Month to Date	Past 30 Days	From 1/1/2015
GB written to disk	10,173.57	29,121.35	106,063.85	149,705.32	4,206,511.97
GB read from disk	31,034.77	97,671.42	421,372.24	693,225.98	34,460,887.35
GB written to tape	6,673.86	7,436.95	24,813.77	39,845.23	1,251,005.88
GB read from tape	515.67	633.79	16,074.81	41,328.57	1,044,252.80
Tape drive hours used	22	29	176	391	13,257
No. of tape mounts	84	140	740	1,651	75,976

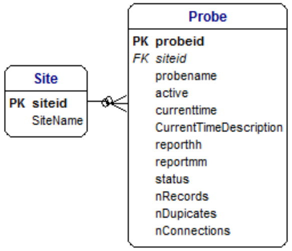


JobUsageRecord_Meta

PF dbid
recordId
CreateTime
CreateTimeDescription
RecordKeyInfoId
RecordKeyInfoContent
ProbeName
ProbeNameDescription
ReportedSiteName
ReportedSiteNameDescription
Grid
GridDescription
ServerDate
md5v2

Probe-Site

NOTE: All tables containing probeid/ProbeName have a relationshop with the Probe table. Inclusion of those relationships has been omitted to eliminate clutter.

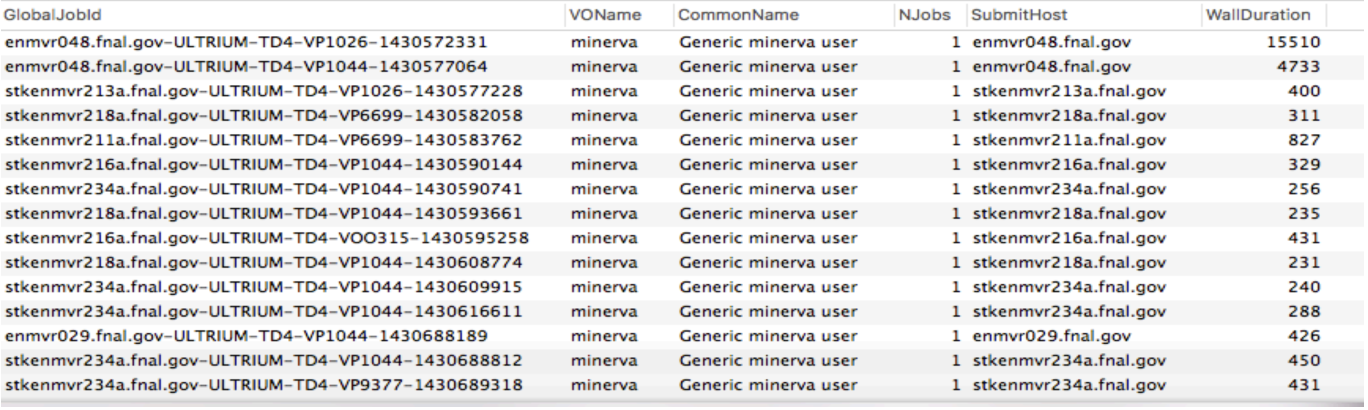


Site

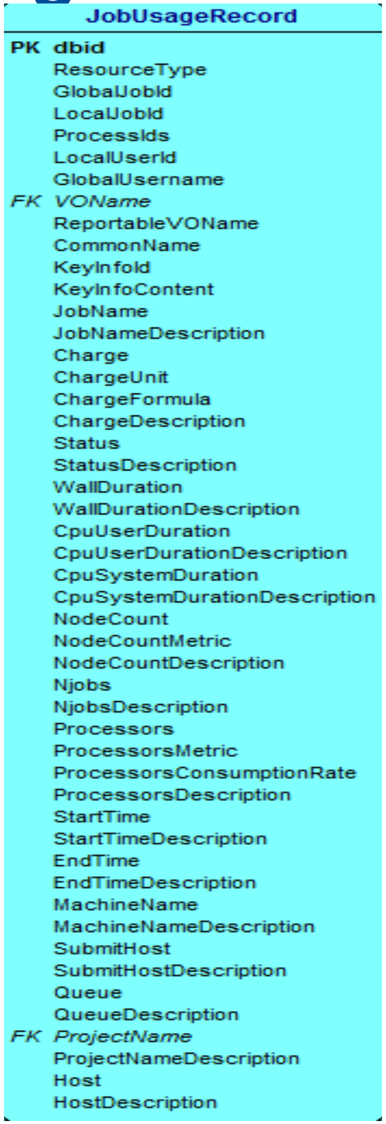
PK siteid
SiteName

Probe

PK probeid
FK siteid
probename
active
currenttime
CurrentTimeDescription
reporthh
reportmm
status
nRecords
nDupicates
nConnections



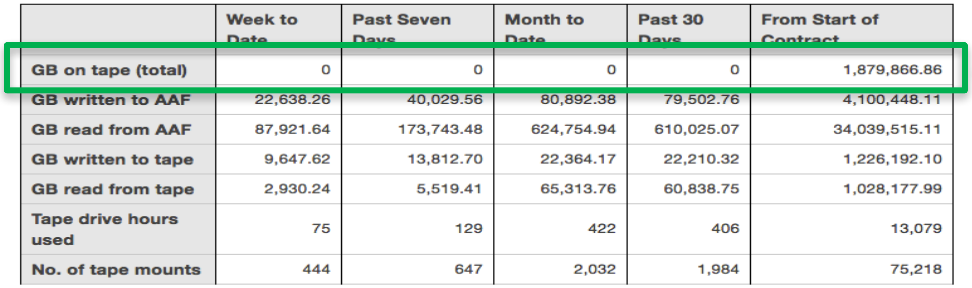
GlobalJobId	VOName	CommonName	NJobs	SubmitHost	WallDuration
enmvr048.fnal.gov-ULTRIUM-TD4-VP1026-1430572331	minerva	Generic minerva user	1	enmvr048.fnal.gov	15510
enmvr048.fnal.gov-ULTRIUM-TD4-VP1044-1430577064	minerva	Generic minerva user	1	enmvr048.fnal.gov	4733
stkenmvr213a.fnal.gov-ULTRIUM-TD4-VP1026-1430577228	minerva	Generic minerva user	1	stkenmvr213a.fnal.gov	400
stkenmvr218a.fnal.gov-ULTRIUM-TD4-VP6699-1430582058	minerva	Generic minerva user	1	stkenmvr218a.fnal.gov	311
stkenmvr211a.fnal.gov-ULTRIUM-TD4-VP6699-1430583762	minerva	Generic minerva user	1	stkenmvr211a.fnal.gov	827
stkenmvr216a.fnal.gov-ULTRIUM-TD4-VP1044-1430590144	minerva	Generic minerva user	1	stkenmvr216a.fnal.gov	329
stkenmvr234a.fnal.gov-ULTRIUM-TD4-VP1044-1430590741	minerva	Generic minerva user	1	stkenmvr234a.fnal.gov	256
stkenmvr218a.fnal.gov-ULTRIUM-TD4-VP1044-1430593661	minerva	Generic minerva user	1	stkenmvr218a.fnal.gov	235
stkenmvr216a.fnal.gov-ULTRIUM-TD4-VOO315-1430595258	minerva	Generic minerva user	1	stkenmvr216a.fnal.gov	431
stkenmvr218a.fnal.gov-ULTRIUM-TD4-VP1044-1430608774	minerva	Generic minerva user	1	stkenmvr218a.fnal.gov	231
stkenmvr234a.fnal.gov-ULTRIUM-TD4-VP1044-1430609915	minerva	Generic minerva user	1	stkenmvr234a.fnal.gov	240
stkenmvr234a.fnal.gov-ULTRIUM-TD4-VP1044-1430616611	minerva	Generic minerva user	1	stkenmvr234a.fnal.gov	288
enmvr029.fnal.gov-ULTRIUM-TD4-VP1044-1430688189	minerva	Generic minerva user	1	enmvr029.fnal.gov	426
stkenmvr234a.fnal.gov-ULTRIUM-TD4-VP1044-1430688812	minerva	Generic minerva user	1	stkenmvr234a.fnal.gov	450
stkenmvr234a.fnal.gov-ULTRIUM-TD4-VP9377-1430689318	minerva	Generic minerva user	1	stkenmvr234a.fnal.gov	431



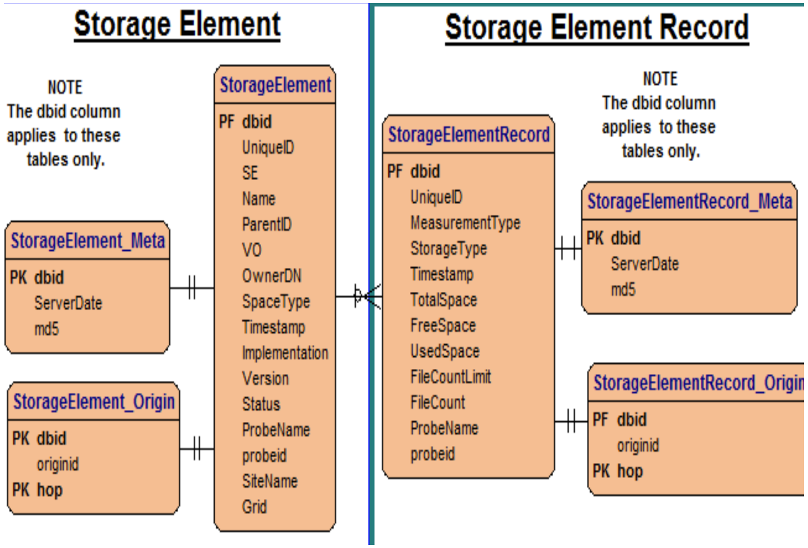
JobUsageRecord

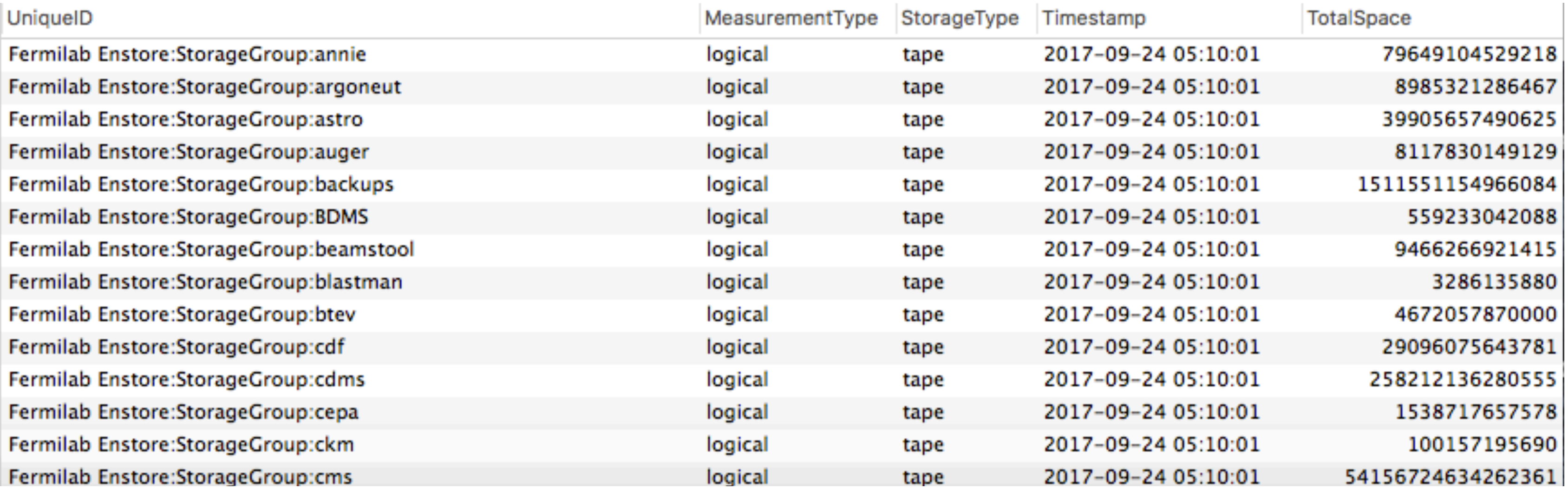
PK dbid
ResourceType
GlobalJobId
LocalJobId
ProcessIds
LocalUserId
GlobalUsername
FK VOName
ReportableVOName
CommonName
KeyInfoId
KeyInfoContent
JobName
JobNameDescription
Charge
ChargeUnit
ChargeFormula
ChargeDescription
Status
StatusDescription
WallDuration
WallDurationDescription
CpuUserDuration
CpuUserDurationDescription
CpuSystemDuration
CpuSystemDurationDescription
NodeCount
NodeCountMetric
NodeCountDescription
Njobs
NjobsDescription
Processors
ProcessorsMetric
ProcessorsConsumptionRate
ProcessorsDescription
StartTime
StartTimeDescription
EndTime
EndTimeDescription
MachineName
MachineNameDescription
SubmitHost
SubmitHostDescription
Queue
QueueDescription
FK ProjectName
ProjectNameDescription
Host
HostDescription

Migration of AAF To GRACC: Data To Migrate



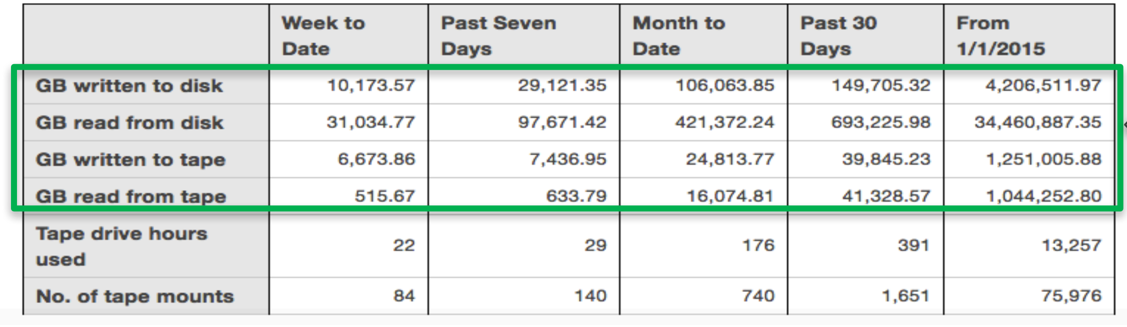
	Week to Date	Past Seven Days	Month to Date	Past 30 Days	From Start of Contract
GB on tape (total)	0	0	0	0	1,879,866.86
GB written to AAF	22,638.26	40,029.56	80,892.38	79,502.76	4,100,448.11
GB read from AAF	87,921.64	173,743.48	624,754.94	610,025.07	34,039,515.11
GB written to tape	9,647.62	13,812.70	22,364.17	22,210.32	1,226,192.10
GB read from tape	2,930.24	5,519.41	65,313.76	60,838.75	1,028,177.99
Tape drive hours used	75	129	422	406	13,079
No. of tape mounts	444	647	2,032	1,984	75,218



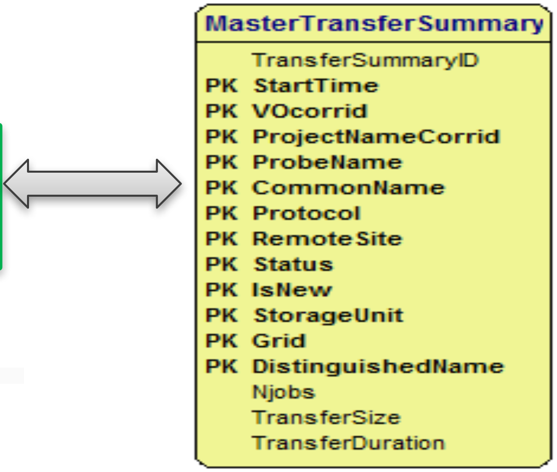


UniqueID	MeasurementType	StorageType	Timestamp	TotalSpace
Fermilab Enstore:StorageGroup:annie	logical	tape	2017-09-24 05:10:01	79649104529218
Fermilab Enstore:StorageGroup:argoneut	logical	tape	2017-09-24 05:10:01	8985321286467
Fermilab Enstore:StorageGroup:astro	logical	tape	2017-09-24 05:10:01	39905657490625
Fermilab Enstore:StorageGroup:auger	logical	tape	2017-09-24 05:10:01	8117830149129
Fermilab Enstore:StorageGroup:backups	logical	tape	2017-09-24 05:10:01	151155154966084
Fermilab Enstore:StorageGroup:BDMS	logical	tape	2017-09-24 05:10:01	559233042088
Fermilab Enstore:StorageGroup:beamstool	logical	tape	2017-09-24 05:10:01	9466266921415
Fermilab Enstore:StorageGroup:blastman	logical	tape	2017-09-24 05:10:01	3286135880
Fermilab Enstore:StorageGroup:btev	logical	tape	2017-09-24 05:10:01	4672057870000
Fermilab Enstore:StorageGroup:cdf	logical	tape	2017-09-24 05:10:01	29096075643781
Fermilab Enstore:StorageGroup:cdms	logical	tape	2017-09-24 05:10:01	25821213628055
Fermilab Enstore:StorageGroup:cepa	logical	tape	2017-09-24 05:10:01	1538717657578
Fermilab Enstore:StorageGroup:ckm	logical	tape	2017-09-24 05:10:01	100157195690
Fermilab Enstore:StorageGroup:cms	logical	tape	2017-09-24 05:10:01	54156724634262361

Migration of AAF To GRACC: Data To Migrate



	Week to Date	Past Seven Days	Month to Date	Past 30 Days	From 1/1/2015
GB written to disk	10,173.57	29,121.35	106,063.85	149,705.32	4,206,511.97
GB read from disk	31,034.77	97,671.42	421,372.24	693,225.98	34,460,887.35
GB written to tape	6,673.86	7,436.95	24,813.77	39,845.23	1,251,005.88
GB read from tape	515.67	633.79	16,074.81	41,328.57	1,044,252.80
Tape drive hours used	22	29	176	391	13,257
No. of tape mounts	84	140	740	1,651	75,976



MasterTransferSummary

	TransferSummaryID
PK	StartTime
PK	VOcorrid
PK	ProjectNameCorrid
PK	ProbeName
PK	CommonName
PK	Protocol
PK	RemoteSite
PK	Status
PK	IsNew
PK	StorageUnit
PK	Grid
PK	DistinguishedName
	Njobs
	TransferSize
	TransferDuration

- *MasterTransferSummary records* are obtained by means of a MySQL stored procedure.

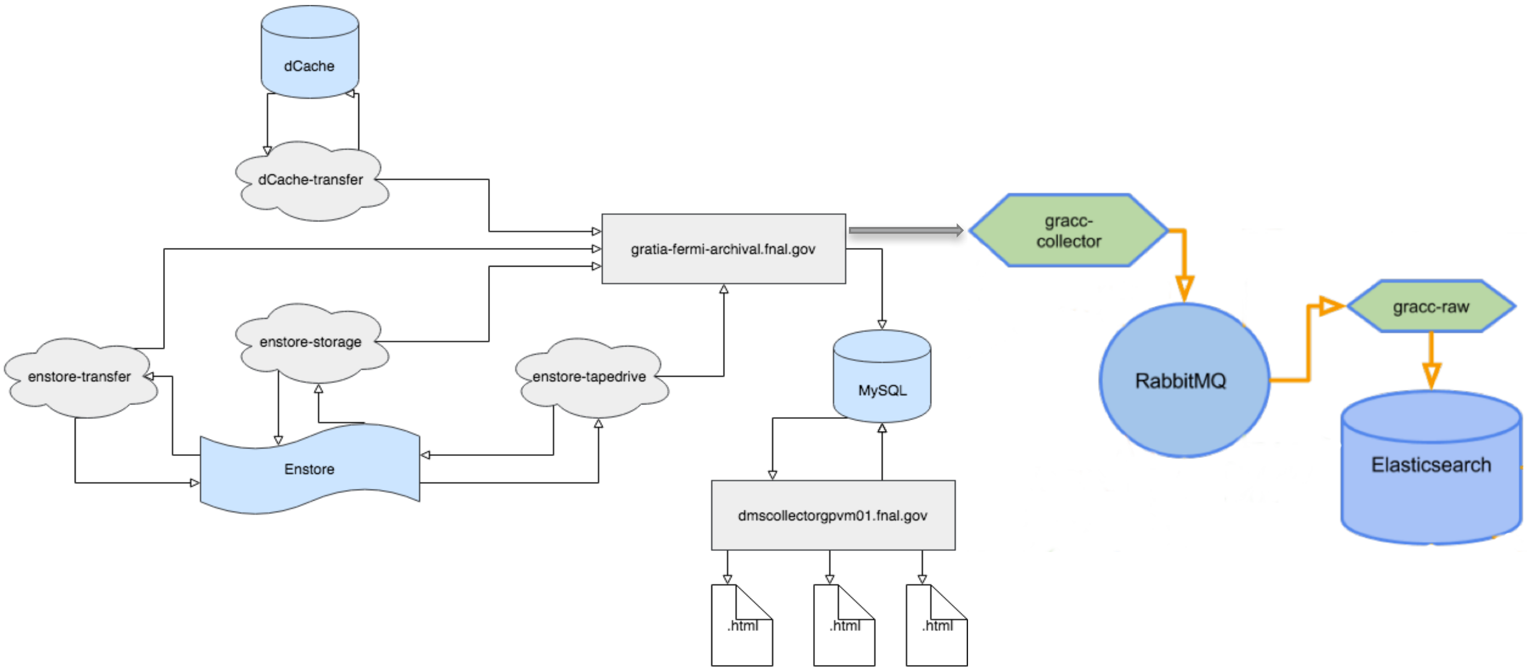
- *The stored procedure generates daily summaries that collate usage metrics across different dimensions so that to reduces the number of records by two to three orders of magnitude*.

- This enable faster analytics of these records over different time periods.

Migration of AAF To GRACC: Migration Plan

- *"…Gratia supports hierarchical collectors' structure and permits forwarding and filtering between collectors…"*

- *For the near-term, the **gracc-collector** service was developed as a transitory endpoint that is compatible with existing Gratia collectors and probes.*

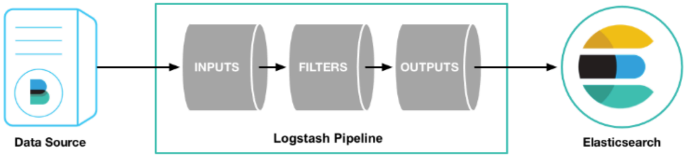


Migration of AAF To GRACC: Migration Plan(I)

- *JobUsageRecord, StorageElement and StorageElementRecord were forwarded relying on the Gratia forwarding mechanism.*

- There was no possibility for *MasterTransferSummary* records to be forwarded to GRACC by using the replication mechanism.

- *At that time Logstash has been taken into account:*
  - *An* open source, server-side data processing pipeline that ingests data from a multitude of sources, transforms it, and then sends it to the favorite "stash".



  - *A logstash.conf file has been defined:*
    - *An jdbc input plugin fetches data from the MasterTransferSummary table*
    - *An output plugin indexes MasterTransferSummary records into Elasticsearch*

Migration of AAF To GRACC: AAF Porting to ES

- Once all data have been stored inside Elasticsearch it has been necessary to implement the AAF's porting to GRACC.
- The porting has involved:
  - The development of a python utility class to interact with Elasticsearch**.**
  - Modification of the Python reporting scripts to query Elasticsearch.
  - The development of a bash script to check that no differences are present between the .html files generated between GRATIA and GRACC AAF.

Migration of AAF To GRACC: AAF Porting to ES

- An initial comparison via the script has shown that GRACC's reports may differ wrt to Gratia's ones, as far as the amount of data read/written from/to disk/tape is concerned.

- *Issue was due to the fact that in Gratia the TransferSize field in the MasterTransferSummary table, which is used to account the data read/written from/to disk/tape, is of type "double" whereas in GRACC was of type "float" because no template was defined at time of migration.*

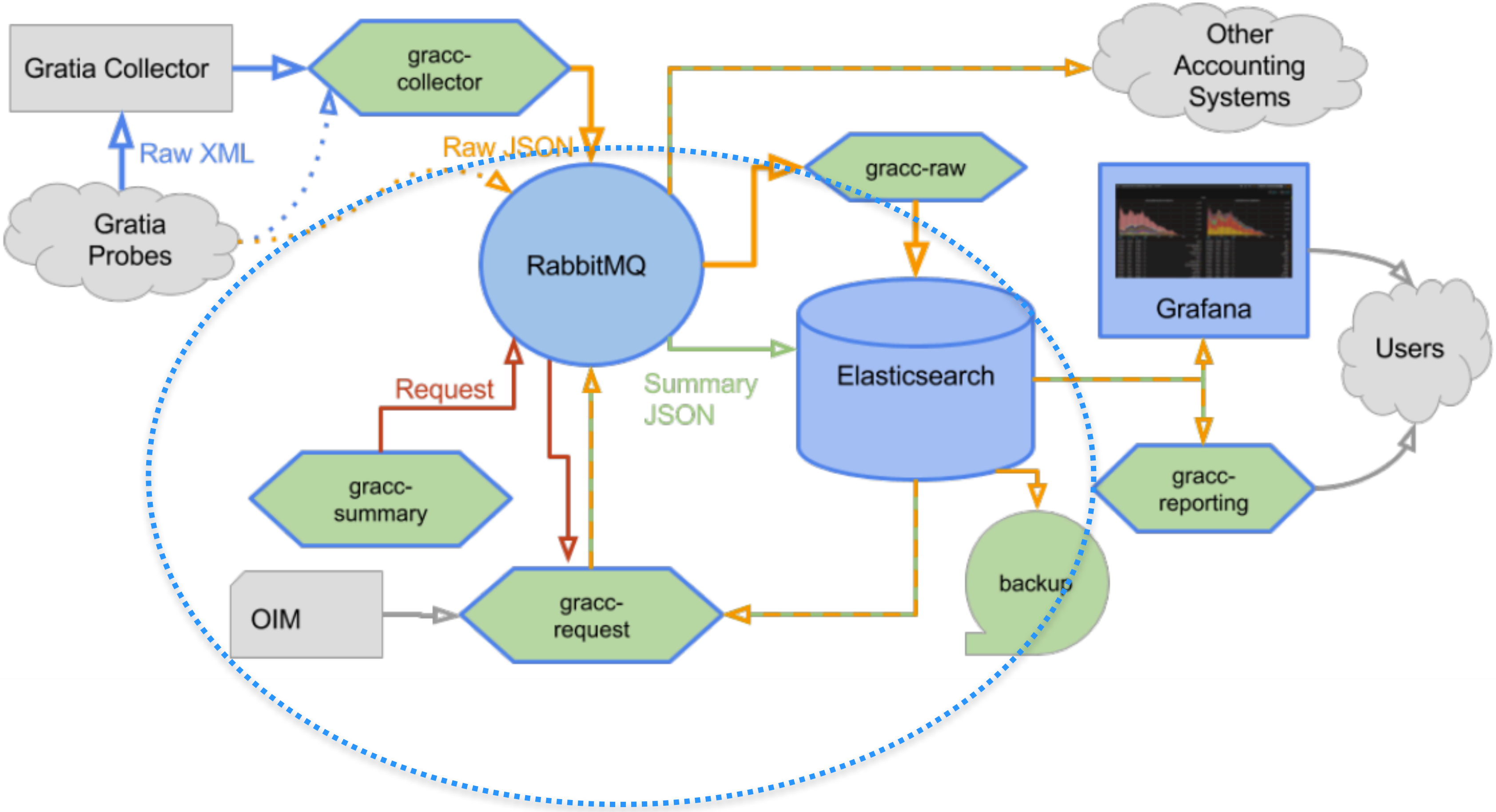
- The solution was to explicitely define the schema mapping and apply it in the output filter of the logstash.conf.

GRACC SUMMARIZATION

GRACC Summarization

- GRACC needs to support the same use cases currently supported by Gratia, e.g an equivalent record summarization service is required.
- Record summarization is performed by a process that listens for requests through RabbitMQ.
- Summary requests specify the time period to summarize and the RabbitMQ exchange to send the summarized records to.
- The summarization is done by aggregation query to Elasticsearch, formatting the results into summary records, performing name corrections/mappings, and further enriching the records by looking up corresponding data in OIM.

GRACC Summarization



GRACC Summarization: Dev Deploy

- The gracc-request agent, the gracc-summary agent and the gracc-stash-summary.transfer were deployed with docker.
- VOName corrections migration to GRACC was also required. Migration has been possibile thanks to logstash.
- At that point we needed to be sure that the GRACC summary procedure actually "*sees*" the same set of raw records seen by the GRATIA one.
- This aims at verifying that the two summary procedures execute on the same initial raw data.

GRACC Summarization: Raw Records Comparison

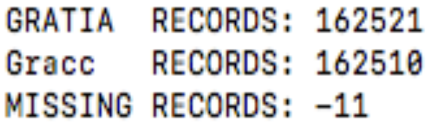
- Raw data forwarding has been enabled on 26th August 2017 in Gratia. The replication mechanism has required also to specify a record id to actually start the replication.

- A bunch of on purpose developed python scripts have been expoited for the data comparison task.

1ST RUN

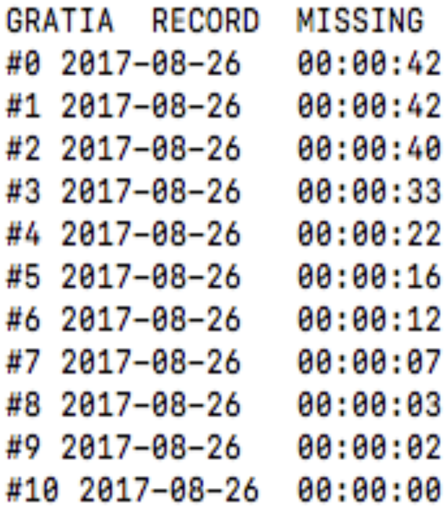
Time Interval:

2017-08-26 to 2017-08-27



```
GRATIA  RECORDS: 162521
Gracc   RECORDS: 162510
MISSING RECORDS: -11
```

?



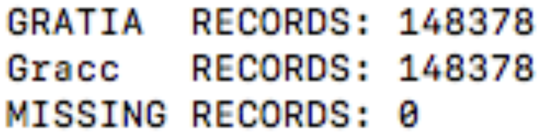
```
GRATIA  RECORD   MISSING
#0 2017-08-26    00:00:42
#1 2017-08-26    00:00:42
#2 2017-08-26    00:00:40
#3 2017-08-26    00:00:33
#4 2017-08-26    00:00:22
#5 2017-08-26    00:00:16
#6 2017-08-26    00:00:12
#7 2017-08-26    00:00:07
#8 2017-08-26    00:00:03
#9 2017-08-26    00:00:02
#10 2017-08-26   00:00:00
```

GRACC Summarization: Raw Records Comparison(I)

Time Interval: 2017-08-26 to 2017-08-27

- *Querying GRATIA with the results provided by the script it has been found out that the deficit actually involves:*
  - *10 records that have a value for the StartTime field which is greater than the one specified at replication init time but whose record id is smaller.*
  - *Plus the record specified at replication init time itself.*

Time Interval: 2017-08-27 to ..



```
GRATIA  RECORDS: 148378
Gracc   RECORDS: 148378
MISSING RECORDS: 0
```

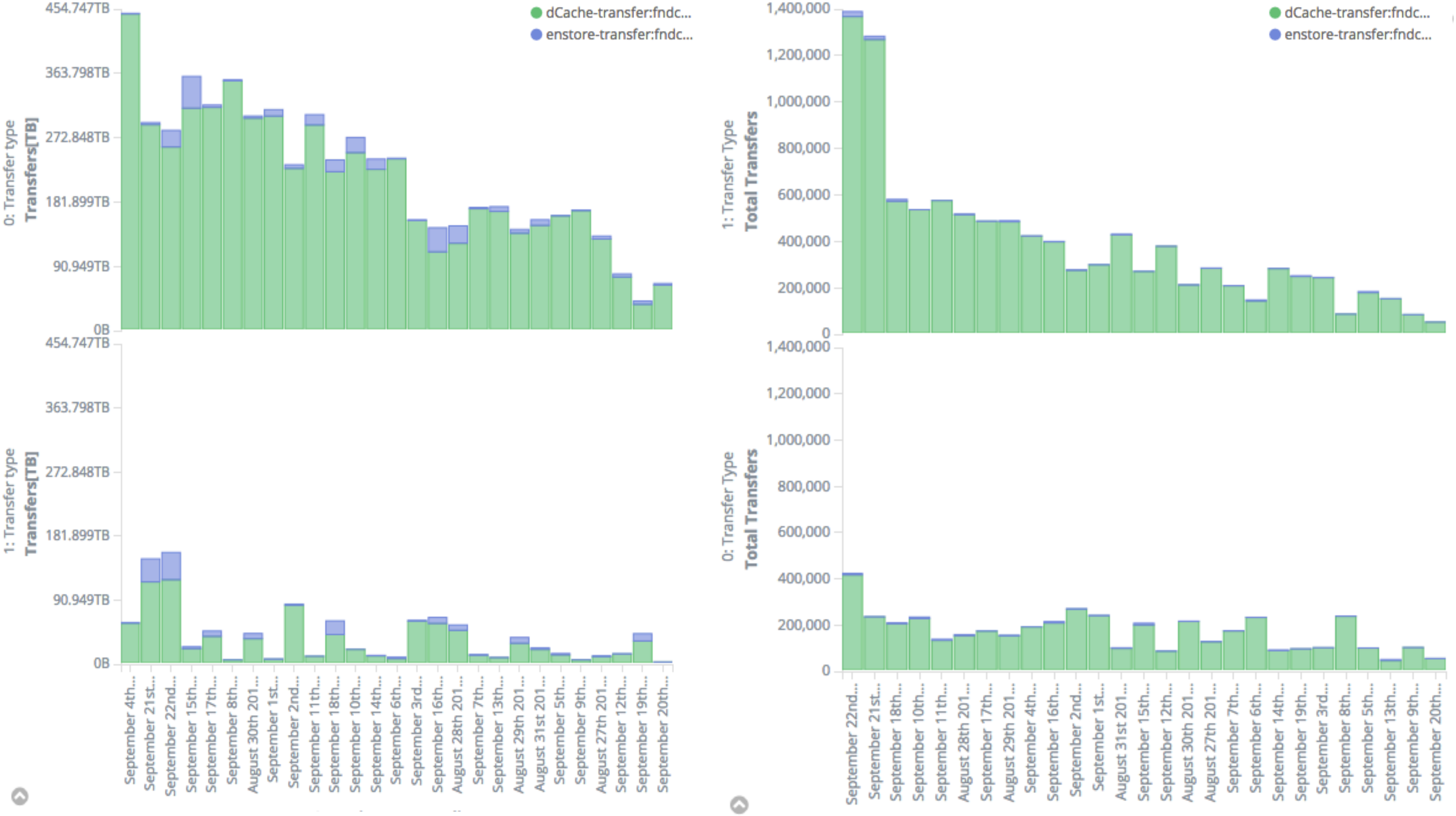
GRACC Summarization: Modification required

- Next step has been verifying that the GRACC summarization procedure outputs the same records migrated out of GRATIA if the same time interval is passed.

- To accomplish this aim different modification were required to (the already developed) GRACC python summary scripts:
  - Let the aggregations being consistent between GRATIA and GRACC queries.
  - Let summary records coming out of logstash be indexed in GRACC according to UTC date format.
  - Refine the Elasticsearch query to avoid *summarizing twice the last day of the week (The entire time interval passed as summarization period is splitted in weeks)*

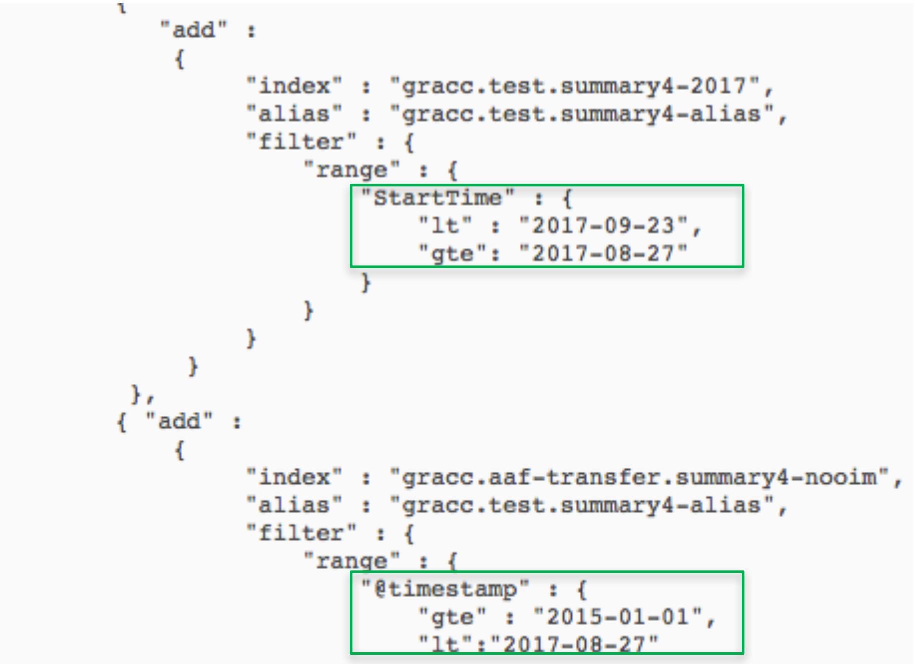
- Different time period have been submitted and visual inspected and no difference has been appreciated.

GRACC Summarization



GRACC Reports: Final assessment

- What the GRACC reports was supposed to output is zero-difference with respect to what the GRATIA reports actually output.
- Since the outstading issue (record missing in GRACC on 26th August) one last necessary step was required in order to overcome:
  - Want we need is a different *“view” of the index being queried while retrieving transfers to/from disk/tape.*
  - To do that we rely on filtered aliases in Elasticsearch:

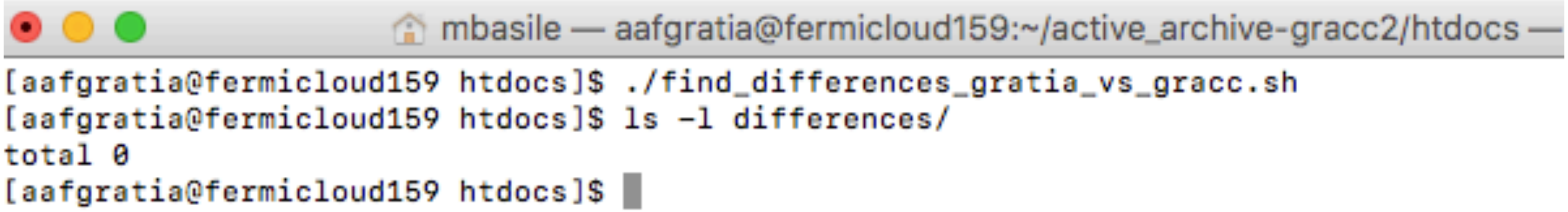


```
        {
            "add" :
             {
                  "index" : "gracc.test.summary4-2017",
                  "alias" : "gracc.test.summary4-alias",
                  "filter" : {
                      "range" : {
                          "StartTime" : {
                              "lt" : "2017-09-23",
                              "gte": "2017-08-27"
                          }
                      }
                  }
             }
        },
        { "add" :
             {
                  "index" : "gracc.aaf-transfer.summary4-nooim",
                  "alias" : "gracc.test.summary4-alias",
                  "filter" : {
                      "range" : {
                          "@timestamp" : {
                              "gte" : "2015-01-01",
                              "lt":"2017-08-27"
```

GRACC Reports: Final assessment(I)

- Only at that time GRACC reports have been generated.

- The *already developed bash script used to find differences between Gratia and Gracc reports has been carried out and this is the content of the differences folder:*

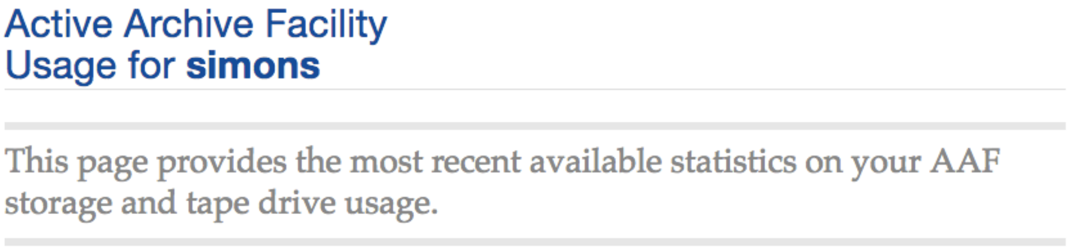


mbasile — aafgratia@fermicloud159:~/active_archive-gracc2/htdocs —

```
[aafgratia@fermicloud159 htdocs]$ ./find_differences_gratia_vs_gracc.sh
[aafgratia@fermicloud159 htdocs]$ ls -l differences/
total 0
[aafgratia@fermicloud159 htdocs]$ 
```

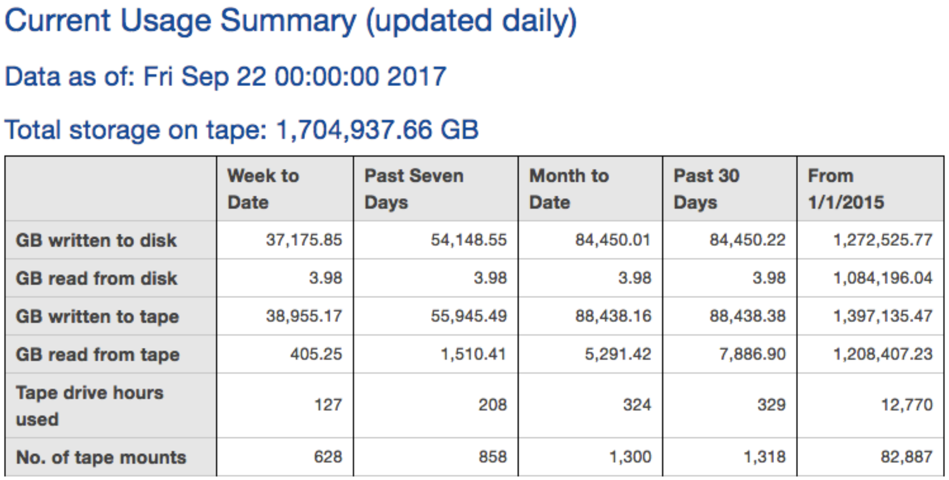
- *We can conclude that GRACC reports are accurate ☺.*

GRACC Reports: Final assessment(I)



Active Archive Facility
Usage for **simons**

This page provides the most recent available statistics on your AAF storage and tape drive usage.



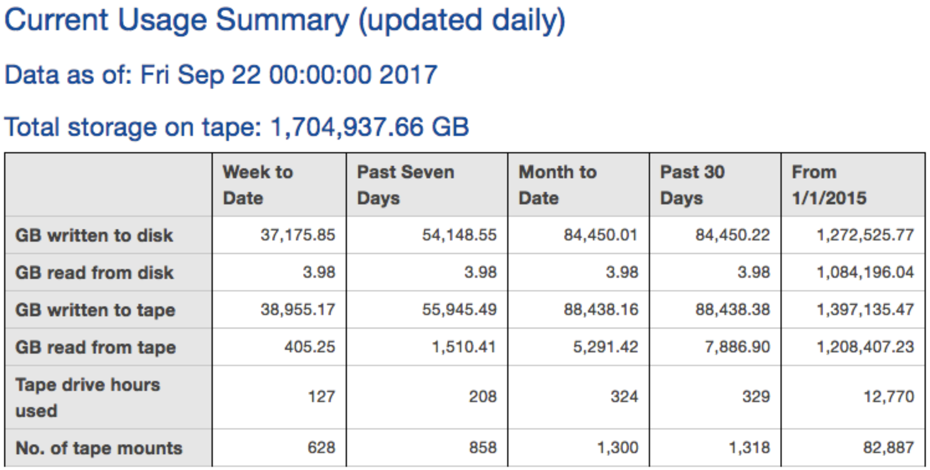
Current Usage Summary (updated daily)

Data as of: Fri Sep 22 00:00:00 2017

Total storage on tape: 1,704,937.66 GB

	Week to Date	Past Seven Days	Month to Date	Past 30 Days	From 1/1/2015
GB written to disk	37,175.85	54,148.55	84,450.01	84,450.22	1,272,525.77
GB read from disk	3.98	3.98	3.98	3.98	1,084,196.04
GB written to tape	38,955.17	55,945.49	88,438.16	88,438.38	1,397,135.47
GB read from tape	405.25	1,510.41	5,291.42	7,886.90	1,208,407.23
Tape drive hours used	127	208	324	329	12,770
No. of tape mounts	628	858	1,300	1,318	82,887

Gratia Report



Current Usage Summary (updated daily)

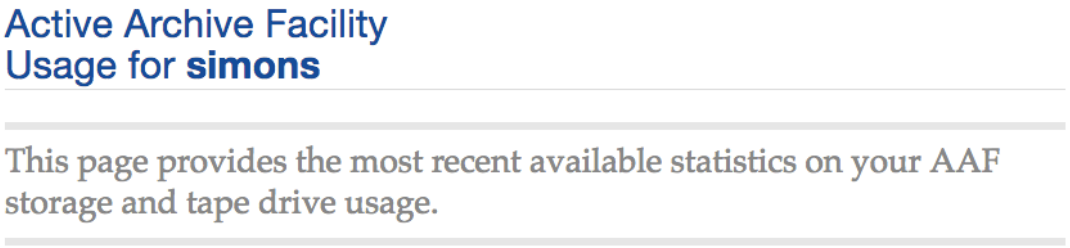
Data as of: Fri Sep 22 00:00:00 2017

Total storage on tape: 1,704,937.66 GB

	Week to Date	Past Seven Days	Month to Date	Past 30 Days	From 1/1/2015
GB written to disk	37,175.85	54,148.55	84,450.01	84,450.22	1,272,525.77
GB read from disk	3.98	3.98	3.98	3.98	1,084,196.04
GB written to tape	38,955.17	55,945.49	88,438.16	88,438.38	1,397,135.47
GB read from tape	405.25	1,510.41	5,291.42	7,886.90	1,208,407.23
Tape drive hours used	127	208	324	329	12,770
No. of tape mounts	628	858	1,300	1,318	82,887

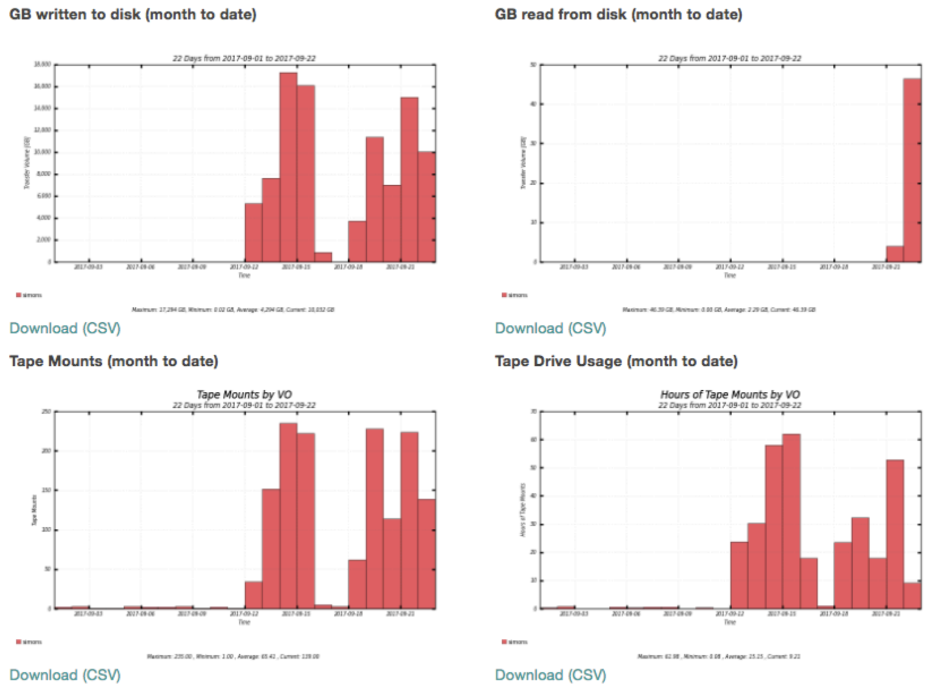
Gracc Report

GRACC Reports: Final assessment(I)



Active Archive Facility
Usage for **simons**

This page provides the most recent available statistics on your AAF storage and tape drive usage.



GB written to disk (month to date)

22 Days from 2017-09-01 to 2017-09-22
Transfer Volume [GB]
Time
simons
Maximum: 17,294 GB, Minimum: 0.02 GB, Average: 4,294 GB, Current: 10,032 GB

Download (CSV)

GB read from disk (month to date)

22 Days from 2017-09-01 to 2017-09-22
Transfer Volume [GB]
Time
simons
Maximum: 46.39 GB, Minimum: 0.00 GB, Average: 2.29 GB, Current: 46.39 GB

Download (CSV)

Tape Mounts (month to date)

Tape Mounts by VO
22 Days from 2017-09-01 to 2017-09-22
Tape Mounts
Time
simons
Maximum: 235.00 , Minimum: 1.00 , Average: 65.41 , Current: 139.00

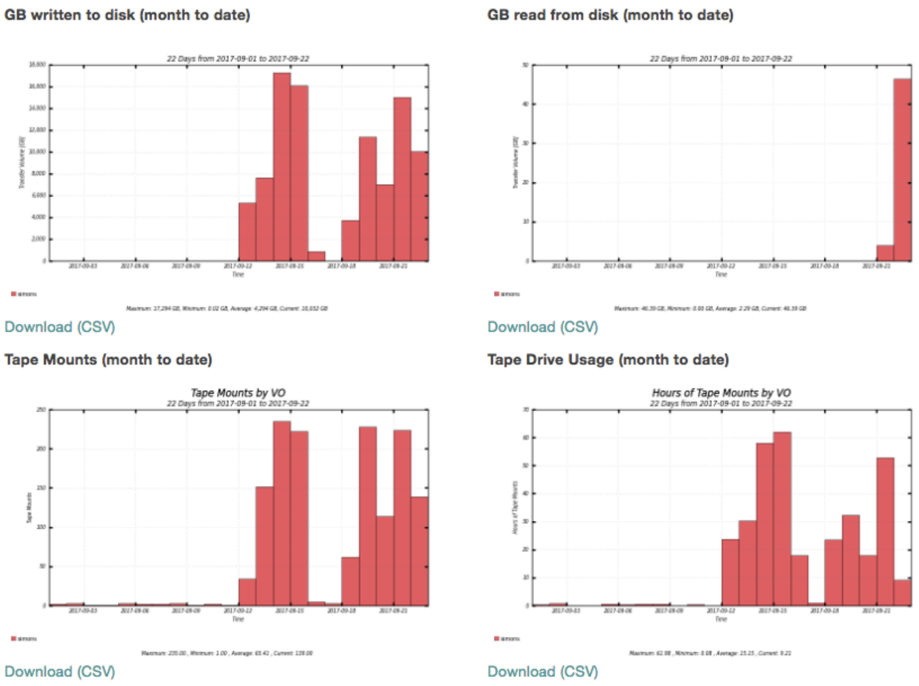
Download (CSV)

Tape Drive Usage (month to date)

Hours of Tape Mounts by VO
22 Days from 2017-09-01 to 2017-09-22
Hours of Tape Mounts
Time
simons
Maximum: 62.98 , Minimum: 0.08 , Average: 25.25 , Current: 9.21

Download (CSV)

Gratia Report



GB written to disk (month to date)

22 Days from 2017-09-01 to 2017-09-22
Transfer Volume [GB]
Time
simons
Maximum: 17,294 GB, Minimum: 0.02 GB, Average: 4,294 GB, Current: 10,032 GB

Download (CSV)

GB read from disk (month to date)

22 Days from 2017-09-01 to 2017-09-22
Transfer Volume [GB]
Time
simons
Maximum: 46.39 GB, Minimum: 0.00 GB, Average: 2.29 GB, Current: 46.39 GB

Download (CSV)

Tape Mounts (month to date)

Tape Mounts by VO
22 Days from 2017-09-01 to 2017-09-22
Tape Mounts
Time
simons
Maximum: 235.00 , Minimum: 1.00 , Average: 65.41 , Current: 139.00

Download (CSV)

Tape Drive Usage (month to date)

Hours of Tape Mounts by VO
22 Days from 2017-09-01 to 2017-09-22
Hours of Tape Mounts
Time
simons
Maximum: 62.98 , Minimum: 0.08 , Average: 25.25 , Current: 9.21

Download (CSV)

Gracc Report

Lesson Learned

- **32-bit vs 64-bit precision:**
  Picking the *wrong* floating point representation may result in unexpected results. Summing a lot of data using a higher precision makes a difference.

- **Time Zone awareness:**
  Distributed systems may use different time zones. When exchanging data, time zone bugs may lead to inconsistencies. Whenever passing a time or date a time zone would've passed too.

- **Lags awarness:**
  Lags need to be taken into account. A time delay between the time an event has occurred and the time the same event has been recorded may cause having different snapshots of the same system as far as different time instants are considered.

THANK YOU!

BACKUP SLIDES

The Grid Accounting Service: GRATIA

- One of the crucial components of a cyber-infrastructure is an **accounting service that collects data related to resource utilization.**

- **Gratia** originated as an accounting system for batch systems and Linux process accounting at FNAL.

- Starting from 2007 Gratia has been adopted by the Open Science Grid as a **distributed, grid-wide accounting system**.

- Collected data includes information about:
  - **Batch Jobs and Glide-in Jobs**
  - **Grid Transfers**
  - **Storage Usage and Allocation**
  - **Cloud Accounting**
  - **Grid Services Availability**
  - **....**

GRATIA Architecture Overview

Gratia Architecture Overview: Probes

- **Gratia collects resource utilization records by means of probes** running on remote sites.

- Probes usually consist of:
  - A python script (the probe itself)
  - ProbeConfig file
  - A cron job which schedules the probe's execution

- **Probes at services** or resources:
  - Batch (HTCondor, PBS, LSF, SGE, SLURM)
  - Transfer (gridftp, hadoop, dCache, xrootd)
  - Storage (dCache, xrootd, hadoop)
  - Cloud Accounting (OpenNebula)
  - Unix Accounting

Gratia Architecture Overview: Collector & Reporter

- **Probes** sends collected data formatted according to the OGF usage record format to the a **Gratia Collector** specified inside the ProbeConf file by using the Gratia API

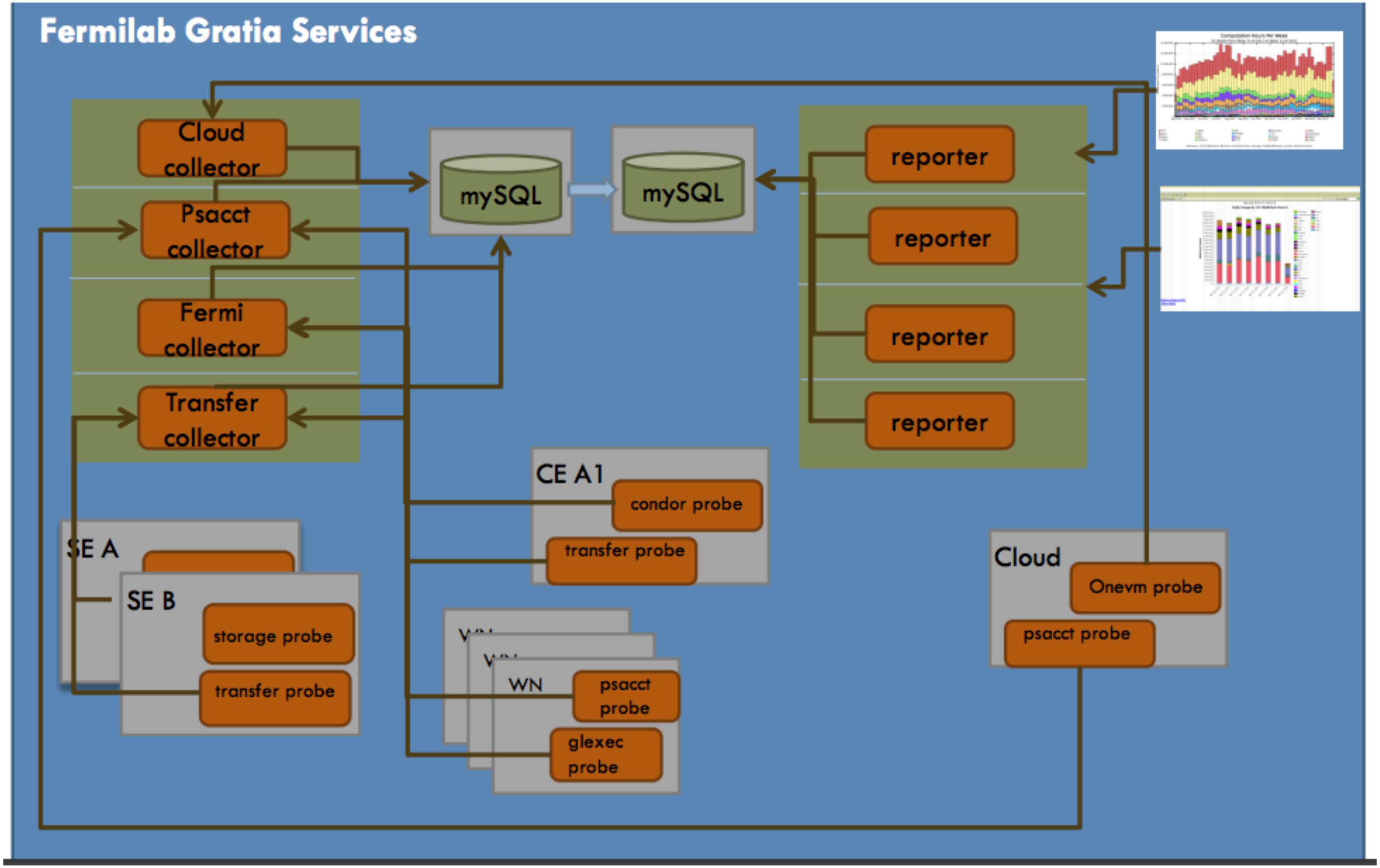
- The **collector** after some data validation **stores the information** in permanent storage which is a MySQL database.

- Gratia supports hierarchical collectors' structure and **permits forwarding and filtering between collectors**.

- So far **more than *1 billion*** job usage records have been collected.

- Often collocated with a collector there's a **reporter**, which gives access to the data through a web user interface.

Gratia Architecture Overview



Gratia Limitations

- Over the years, **the requirements of what types of usage should be tracked, and what information should be stored, have evolved** as scientific computing has evolved:
  - expansion into new types of resources:
    - public and private clouds
    - migration to multicore environment

- **Gratia has struggled to keep up with these changes, due to inflexible record models and storage**.

- Additionally, the **Gratia service architecture has struggled to scale as the OSG usage** and hence record rate **has increased**.

- Finally, **the user interface and visualization tools have fallen behind the state-of-the-art**, due to large effort being required to build interfaces with the Gratia system.

Gratia: Limitations(II)

- For these reasons and more, in early 2016 the OSG and Fermilab decided to investigate re-designing Gratia.

- The idea was to provide:
  - **a more flexible architecture and data storage format**
  - **easier integration with open-source data exploration and visualization tools**.

- **The investigation settled on a microservice-based architecture called GRACC**.

Enstore

- Enstore provides access to data on tape to/from a user's machine on-site over the local area network, or over the wide area network through the dCache disk caching system.

- When used with caching/buffering system, files first get written to disks which then get migrated from disk to Enstore tapes.

- For file read requests, if the files do not reside in the disk cache, they first get retrieved from Enstore to the cache.

- Direct access to Enstore is limited to on-site machines – dCache is required for off-site access.

dCache

- dCache decouples the low latency, high speed disk access over the network from the high latency sequential access of tapes. The cache provides high performance access to frequently accessed files.

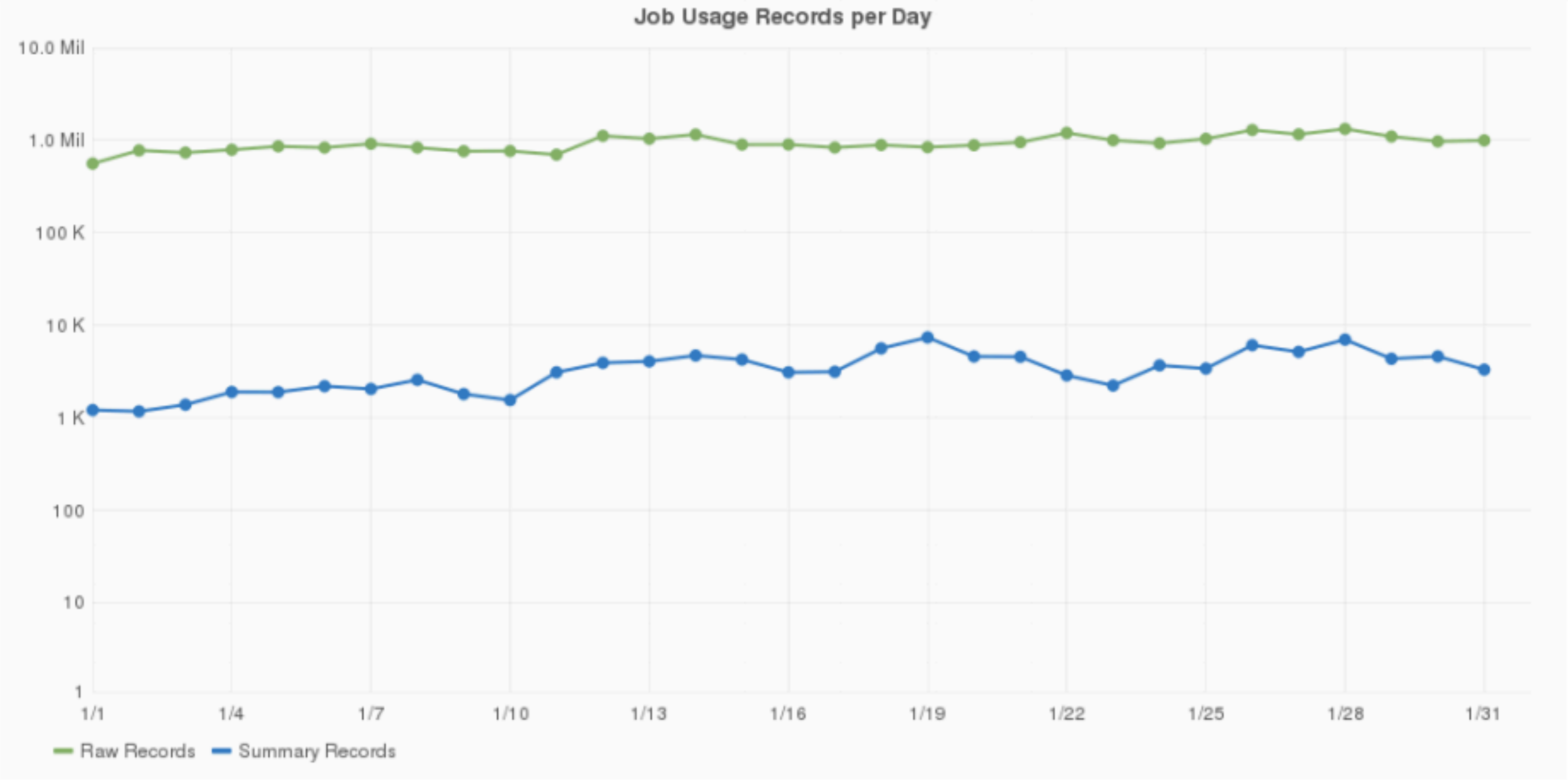
- Whether the file already exists in the disk cache, or needs to be first retrieved from tape is transparent to the user.

- Fermilab dCache systems use raided disk in redundant configurations to reliably store users' files.

- Files in dCache can be accessed with several different protocols.
  Local users can access data through dcap, kerberized FTP, GridFTP, and NFSV4.1.

- Users needing to access files on tape from off-site computers must do so through dCache.

Migration of AAF To GRACC: Data To Migrate



Count of Open Science Grid job usage records received each day in January 2016 and corresponding summary records