LIneA
Laboratório Interinstitucional de e-Astronomia

DES Science Portal readiness for Y1A1

Angelo Fausti, Luiz da Costa
and LineA IT Team

1 Feb 2014

Changes

Version	Prepared by	Revised by	Date	Comments
1.4	Angelo Fausti	-	30/07/14	Comparing execution times between LineA and FNAL installations (In prep.)
1.3	Angelo Fausti	-	05/05/14	Reviewed execution times for Y1A1 based on Y1P1
1.2	Angelo Fausti	Luiz da Costa	02/10/14	---
1.1	Angelo Fausti	Marcio Maia	02/04/14	---
1.0	Angelo Fausti and IT team		02/01/14	First version

Index

1. Summary

This document describes the Science Portal processes used in the validation of the DESDM releases and in the production of Value-added Catalogs (VACs). We present the execution time and data sizes for the different processes in order to evaluate the Science Portal readiness for the DES first year data (Y1A1). The results presented here are for the portal installation at LineA but we include also the execution times collected from the portal running at Fermilab for comparison (in preparation)

We find that data transfer from NCSA and the database ingestion are the main bottlenecks. We also find that the scalability of the system installed at LIneA is degraded by our current scratch area and database solution limiting the number of parallel jobs to a 100. We propose solutions to improve the overall system performance in the short term allowing us to validate the Y1A1 release[1]. Finally, we plan the system expansion for Y2.

The main actions for improving the current system are:

- to reinstall Lustre, increase the number of OSTs and consider the SGI solution for the scratch area;
- to install pgpool in the parallel mode with more backends servers or to consider the parallel database solutions being investigated by LNCC
- to increase the number of cores in the cluster by a factor of two and evaluate the use o infiniband to improve the internal bandwidth (from1Gbps to 10Gbps);

In addition, the transferring of the data center from the POP-RJ to IDC-BSB will guarantee network, power stability and a connectivity of 10Gbps.

[1]) Y1A1 is the first DES annual release, assuming a nominal area of ~1500 sq deg, (~3000 tiles) scheduled for mid of August, 2013

2. Science Portal processes

Fig.1 illustrates the processing stages at the Science Portal. The validation of DESDM releases is part of the **Data Installation** stage by the results produced by the QA pipeline and by the visual inspection of the images and associated catalogs using the Tile Viewer tool.

In the **Data Preparation** stage we compute photo-z, perform S/G classification and compute galaxy properties running all the codes available in the portal. The results are stored in the database and are used during the **VAC Creation** stage. VACs can be published and downloaded from the portal or be used as input for the **Science Analysis** stage. A QA pipeline is also present for the characterization of the VACs.

The **Science Analysis** stage will be discussed in another document, but it includes workflows being developed for the following working groups and applications: GE (mass, luminosity, correlation evolution), Cluster (different cluster finders), LSS (ACF), TCP (des-pes), GA (FindSat, Milkyway fit)

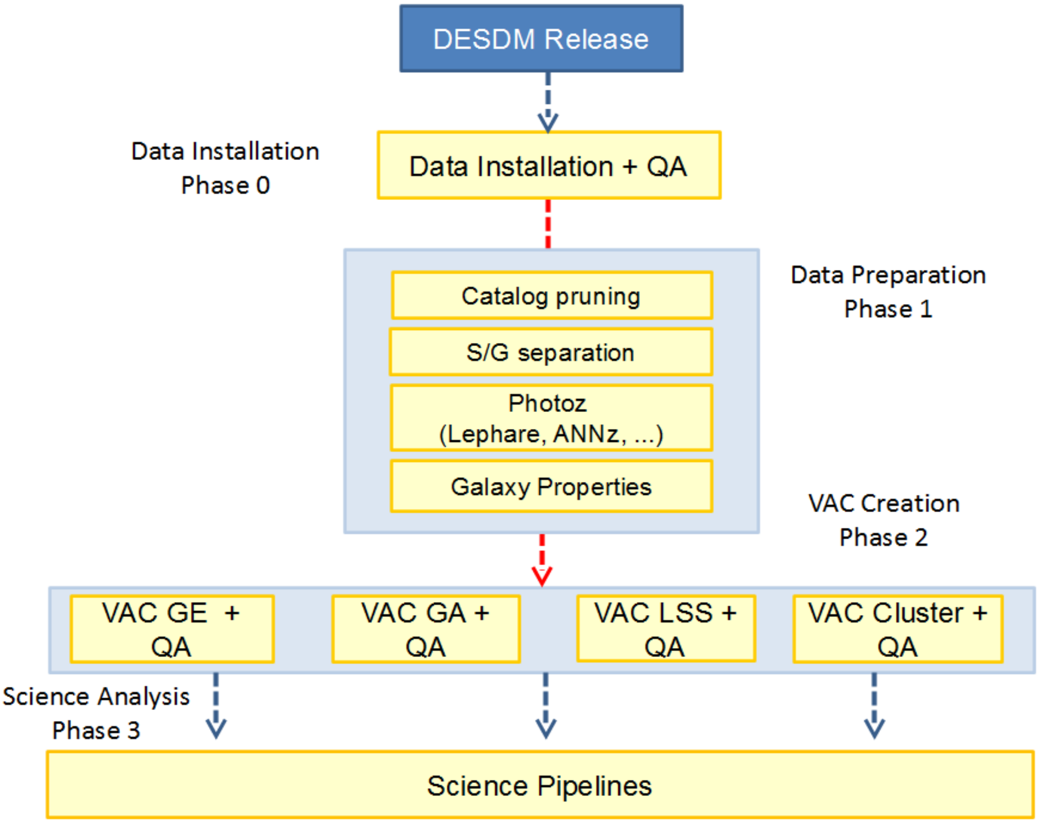


Fig 1. Science portal processes for the validation of DESDM releases, creation of Value-Added Catalogs and execution of Science Analysis pipelines

In the stages described above, the processing unit is the tile. Tiles are processed in parallel whenever possible but there are tasks like data transfer, database ingestion and the consolidation of the results which are serial. Due to this combination of parallel and serial tasks the “execution time per tile” is not a good estimator for the execution time. A key parameter is the maximum number of tiles that can be executed in parallel which depends on the characteristics of the system as discussed in the appendix A. In the Appendix B we show how the execution time of a given process is estimated.

In the next sections we'll describe each processing stage in more detail and will estimate the execution times for Y1A1 based on measurements from Y1P1 and SVA1 releases.

2.2 Data Installation

The goal of this stage is to automate all the steps needed to install and validate a data release in the portal. Because the portal is not directly attached to the DESDM Science

Database the Data Installation it includes the data transfer from NCSA and the ingestion of the catalogs into our PostgreSQL database.

1. Data Transfer: transfer of coadd images and coadd catalogs,
2. Image installation: prepare full resolution color image, PNG thumbnails (400x400 pixels) and JPGEs for the Tile Viewer with eight different resolutions for zoomming in/out + WCS information
3. Catalog installation: ingest the coadd catalog and mask into the portal catalog DB, creates a healpix representation of the catalog, execute pre-matching with comparison catalogs (in preparation)
4. Run the QA pipeline
5. Produces the input for the Data Preparation pipeline

Table 1 summarizes the time estimates to install Y1A1 releases based on the time measurements using SVA1_COADD_SPTE (505 tiles) and Y1P1_COADD_STRIPE82 (285 tiles). For each serial or parallel task, the **effective time per tile** is obtained from the prescription presented in the Appendix B.

Table 1 - Time estimates for the Data Installation process in the current system

Process/Time	**Per tile (s)**	**Y1A1 (3000 tiles) (h)**
Data Transfer	50.4	42
Data Ingestion	36	30
Image Installation (*)	1440	12
QA (*)	600	5
QA consolidation	10	8
Handling mask(**)	500	4
Save & Publication	10	8
Total	**----**	**110h**

(*) tasks performed in parallel

(**) mangle masks are not fully integrated in the science portal yet, this is a crude estimate for the mask installation and handling

From the time estimates for Y1A1 it is clear that the data transfer and data ingestion are the main bottlenecks as we increase the number of tiles. The mean transfer rate between NCSA and LineA is about 250Mbps and data ingestion cannot be performed in parallel unless a parallel database solution is used.

2.3 Data Preparation

The Data Preparation pipeline first cleans and applies corrections to the sample, then separates stars from galaxies, compute photometric redshifts and other galaxy properties. It creates one table per algorithm in the database. Table 2 summarizes the time estimates to execute this pipeline on Y1A1 release.

Table 2 - Time estimates for the Data Preparation processes in the current system

Process/Time	Per tile (s)	Y1A1 (3000 tiles) (h)
Catalog prunning	5	4.5
S/G Classification (*)	50	0.45
Photo-z (LePhare) (*)	1800	15
Photo-z (ANNz) (*)	100	0.9
Consolidation	20	15
Total	**----**	**36h**

(*) tasks performed in parallel

- Catalog prunning is executed in the database so it scales almost linearly with the number of tiles

- The other Photo-z algorithms installed in the portal will be included in this table as soon as we have the time estimates, all of them are optionally executed during Data Preparation

- Consolitation includes the ingestion of the classification, photo-z and galaxy properties tables into the databae. That is not performed in parallel so it also scales linearly with the number of tiles

2.4 VAC creation

Table 3 - Time estimates for the VAC creation processes in the current system

Process/Time	Per tile (s)	Y1A1 (3000 tiles)(h)
Query execution	5	4
Catalog Characterization	500	4
Consolidation	10	8
Save & Publication	5	4
Total	**-----**	**20h**

- The time estimates for VAC creation does not include mask handling which is not fully integrated in the portal yet.
- Save and publication times means moving files to archive and to the FTP area and making the association of the tiles with other tools in the portal.

3 Data size estimates

In addtion to the processing requirements we need to consider data volumes.

Table 4 – Data volumes

Data/Size	Per tile (GB)	Y1P1 (1000 tiles) (GB)	Y1A1 (3000 tiles) (GB)
FITS images	4	4000	12000
FITS catalog	0.3	300	900
PNGs / JPEGs (a)	1	1000	3000
QA	0.2	200	600
Data Preparation	-	-	-
VACs	-	-	-
Total	**5.5GB**	**5.5 TB**	**16.5 TB**

(a) thumbnails, full resolution PNGs and zommable JPEGs

- From data size estimates, we find that Data Installation only will require ~16.5 TB for Y1A1

- For Y1, we assume a factor 1.5 or ~25TB for Data Preparation and VACs. This will be reviewed when we have better size estimates for the remaining processes

- A release of the whole survey will require ~150TB plus the space to keep all data from previous releases

TODO: distinguish sizes in the database and archive for coadds

Conclusions

We have estimated the execution times for the different stages involved in the validation of the releases and in the production of value-added catalogs.

Considering the goal of installing and validating Y1A1 release in about 48h, based on the results presented above we have to improve the overall system performance by a factor of 2 at least. In order improve the system scalability and use the 500 cores available to reduce the execution times the main actions are:

- Continue the work on the science DMZ and network tuning to guarantee a sustainable transfer rate of at least ~500 Mbps from NCSA to LineA; This will reduce the data transfer times presented in table 1 by a factor of 2.

- A high performance file system such as Lustre is crucial for the system scalability, reinstall Lustre following NCSA recommendation (see *Science Portal installation at external sites* for details);

- A short term solution is to fix the load balance problem and include more backends in the replication mode – replace the mass storage by servers with ~30TB local/fast disks in total

- A mid term solution is to install a database cluster it should also allow parallel ingestion which is needed to reduce the data ingestion times (see *table 1*).

- The the internal network performance and disk I/O on both the mass storage and the data transfer server are very important during processing and also for the publication of the results. This can be improved by removing intermediate files (reducing the the process size) and increasing the internal network bandwidth using 10 Gbps using infiniband.

Appendix A

A1 Current system at LIneA

The current system is detailed in the *Science Portal installation at external sites* document.

A1.1 Data transfer

Current Data transfer rate (NCSA-LIneA) : 250 Mbps

A1.2 Processing framework

Number of cores: 504 (84 nodes)
Using Condor as job scheduler

A1.3 Catalog DB

The database solution is based on pgpool in the replication mode. Currently, load balancing is not working between the two backends. This problem is being investigated. Data Ingestion is performed in one of the pgpool backends and the second backend is synchronized via streaming replication.

The processing is limited to ~100 parallel tasks due to a problem in the Lustre file system (we are temporarily using NFS for the scratch area) and due to limitations in the database solution which cannot handle > 100-150 queries in parallel when the processing nodes execute data retriever.

A2 Current system at Fermilab

A2.1 Data transfer

A1.2 Processing framework

Number of cores: 30 (single machine)
Using pardal as job schedules (implementation based on python threads)

A1.3 Catalog DB

Appendix B - Time profiling

The execution time to process N tiles in parallel, Tp, is given by the formula:

Tp = t*max(1, N/p) (1)

where t is the execution time per tile and p is the maximum number of parallel jobs that can be submitted to the cluster.

As discussed in the appendix A, p=100 in the current system. Thus if t=10s we can process N=1000 tiles in N/p=10 blocks of t=10s thus Tp=100s in total.

The execution time for serial tasks is just

Ts=N*t (2)

The total execution time T is the sum of the Tp and Ts for the different execution blocks of the workflow.

NOTE: we also refer to t as the effective time per tile to distinguish from the mean time per tile which is simply the total execution time divided by N. Of course if T is dominated by Tp and N < p the effective time per tile and the mean time per tile are the same, but they are totally different if N>p and if we have a combination of serial and paralled tasks in the process.

B1. Photo-z (Lephare)

Here we show the case of Photo-z in which the execution time is dominated by Tp.

Figure C1 shows an example of the Photo-z computing with 50 parallel jobs submited to the cluster.

The execution time is dominated by the Lephare algorithm. From eq. 1 Tp = t and we obtain t ~1800s.

Now in order to estimate the execution time for N=1000 tiles assuming p=100 for the current system we conclude that it would be computed in N/p=10 blocks of t=1800s or 5h.

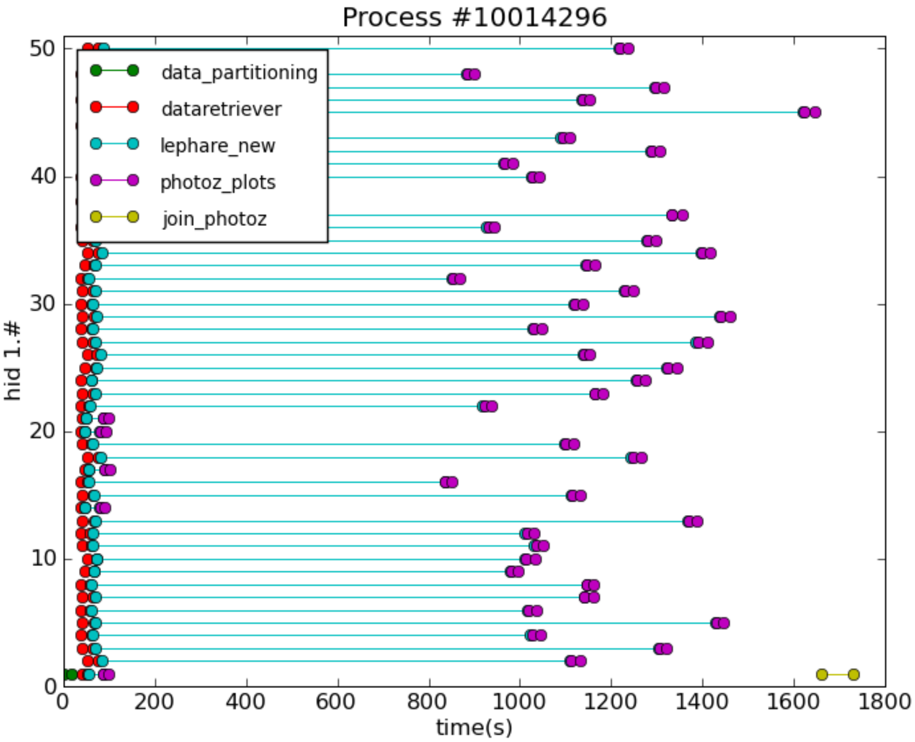


Fig B1. Example of process where the execution time is dominated Tp

B2. QA pipeline

The case of the QA pipeline is interesting because the tiles are processed in parallel but a significant fraction of the execution time is spent to consolidate the results.

We take as example the process 10014148 as part for the SVA1 release validation which ran over 467 tiles (with complete spatial coverage) of SVA1_COADD_SPTE. Figure B2 shows the time profile for this process.

An important finding from the tile profiling was the large initialization time of ~1000s where tiles are evaluated as complete or incomplete and the input for the subsequent tasks is prepared. That step can be parallelized.

Note also that the consolidation time is equivalent to the total time of the individual parallel tasks ~3000s.

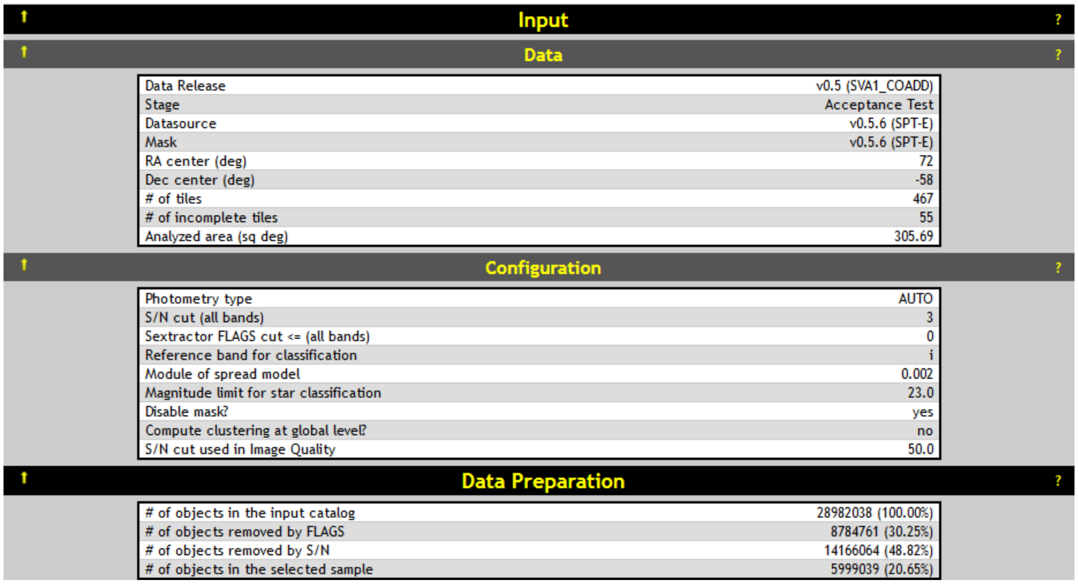
The dead time seen between the end of the parallel tasks and the beginning of the consolidation (~1500s) is also significatn and it is due to the network communication among the computer nodes, we can incorportate that in the consoliation time Ts

From the time profiling we have:

Tp~3000s and from eq. 1 t=600s per tile

Ts~5000s and from eq. 2 t=10s per tile serial.

The estimated time for N= 1000 tiles is thus T = Tp + Ts = 1.6h + 2.8h = 4.4h



Input

Data

Data Release	v0.5 (SVA1_COADD)
Stage	Acceptance Test
Datasource	v0.5.6 (SPT-E)
Mask	v0.5.6 (SPT-E)
RA center (deg)	72
Dec center (deg)	-58
# of tiles	467
# of incomplete tiles	55
Analyzed area (sq deg)	305.69

Configuration

Photometry type	AUTO
S/N cut (all bands)	3
Sextractor FLAGS cut <= (all bands)	0
Reference band for classification	i
Module of spread model	0.002
Magnitude limit for star classification	23.0
Disable mask?	yes
Compute clustering at global level?	no
S/N cut used in Image Quality	50.0

Data Preparation

# of objects in the input catalog	28982038 (100.00%)
# of objects removed by FLAGS	8784761 (30.25%)
# of objects removed by S/N	14166064 (48.82%)
# of objects in the selected sample	5999039 (20.65%)

Fig. C2 – Summary of QA process 10014148

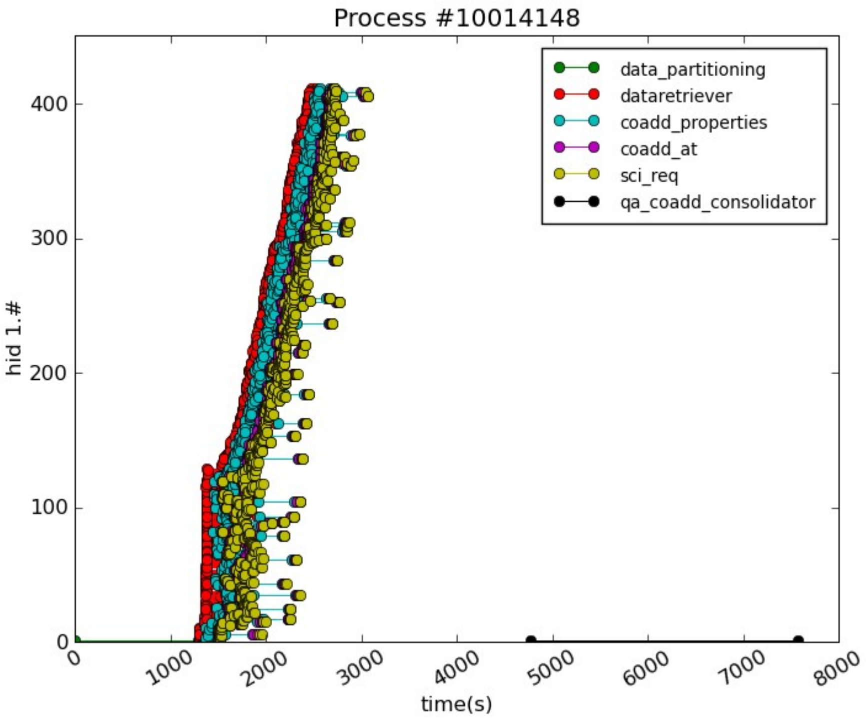
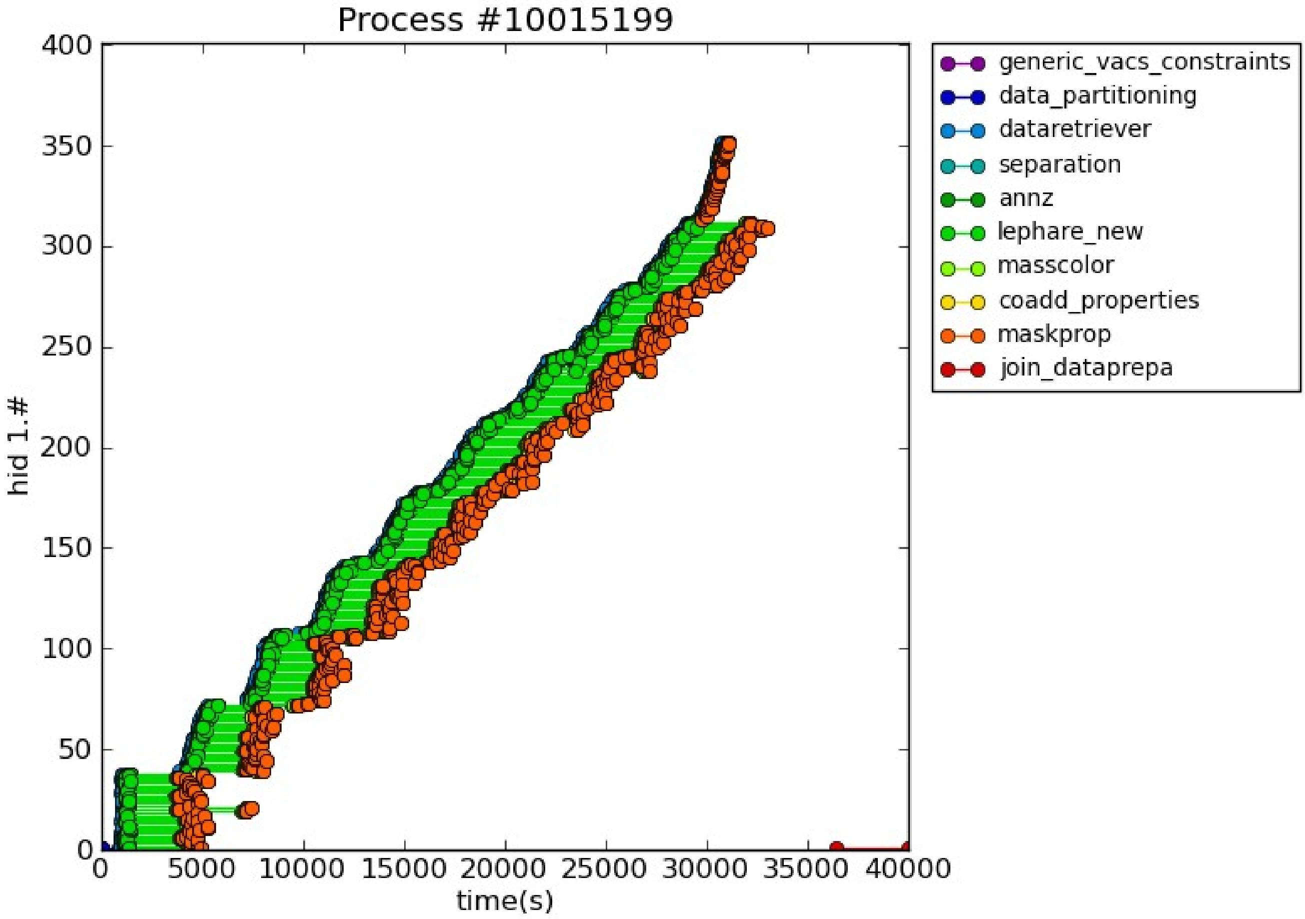
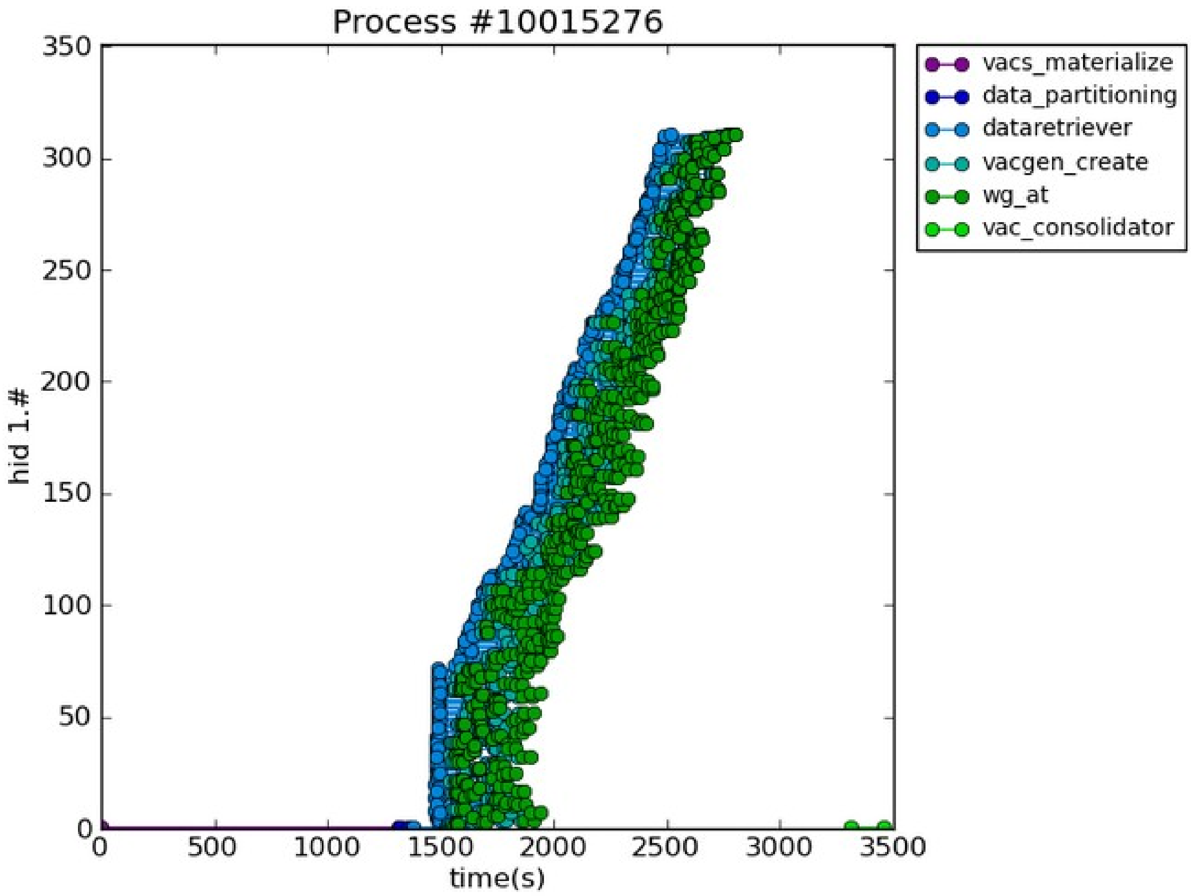


Fig. B2 Time profiling for QA process 10014148

Data Preparation process for Y1P1 SPTW (352 tiles) 12h 42 cores., Fraction of time used by Lephare 70%



VAC Creation Process for Y1P1 SPTW (352 tiles) 1h using 75 cores


Process #10015276
hid 1.#
time(s)
vacs_materialize
data_partitioning
dataretriever
vacgen_create
wg_at
vac_consolidator