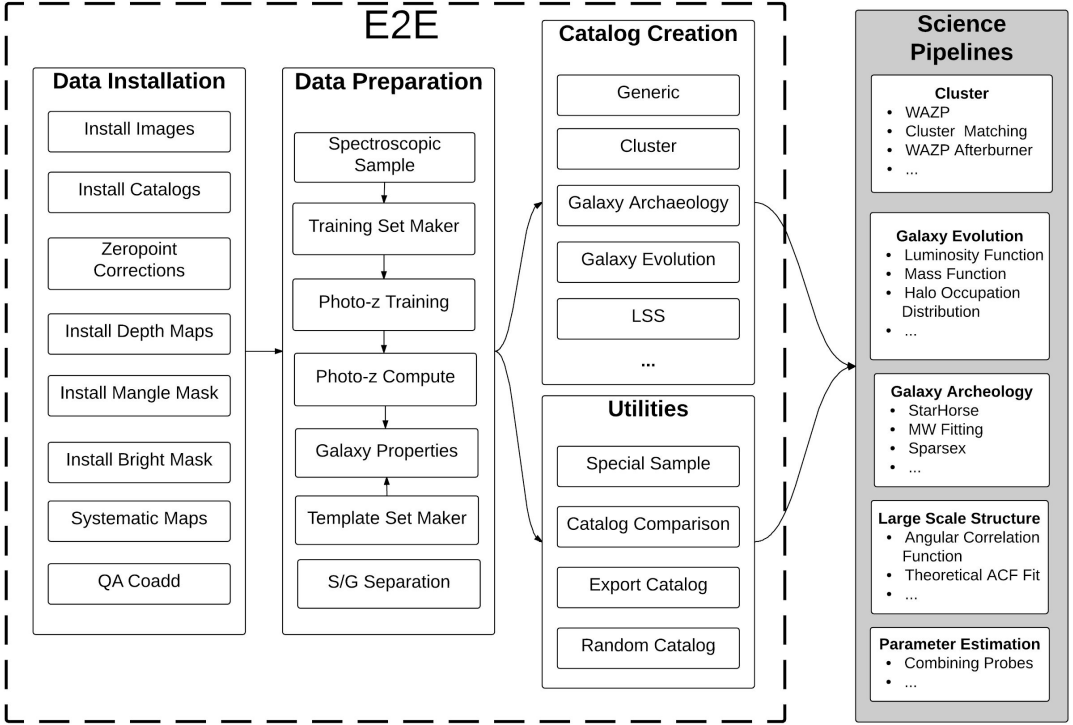
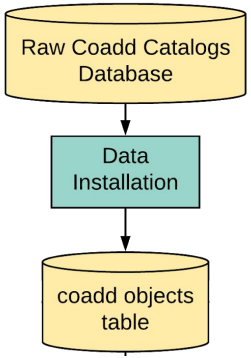
33 - DES Portal: Data Preparation

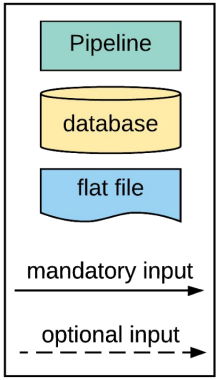
Julia Gschwend

Recap end-to-end (E2E)

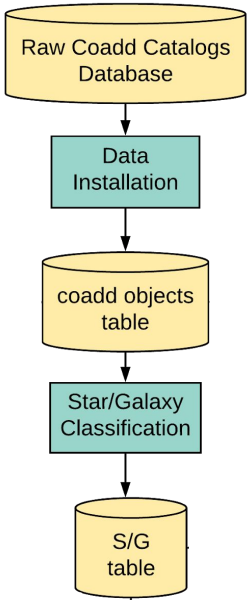


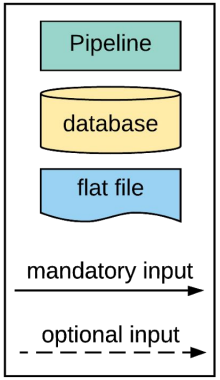
End-to-end (E2E)



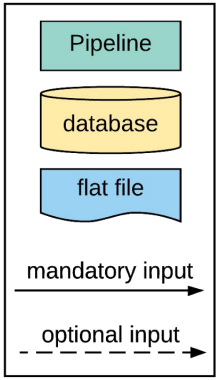


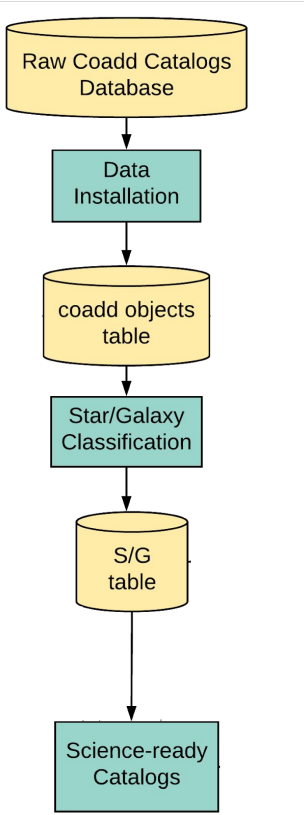
End-to-end (E2E)



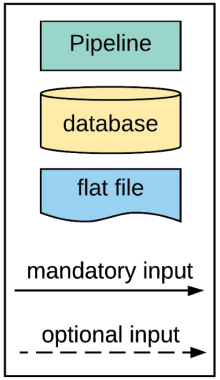


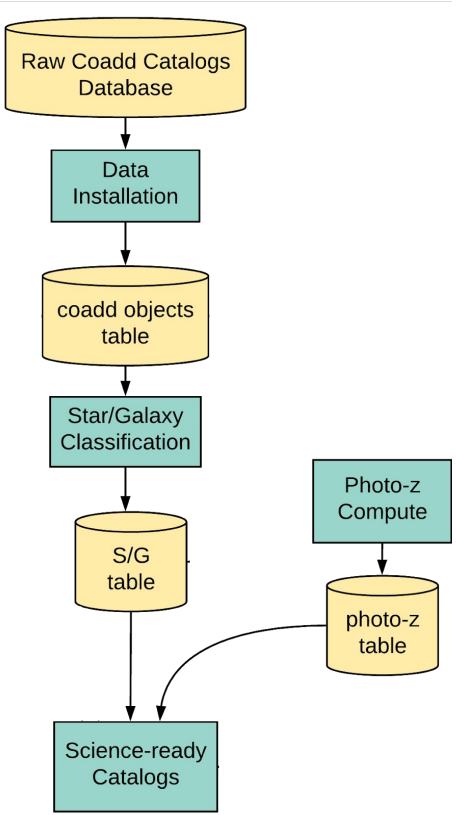
End-to-end (E2E)



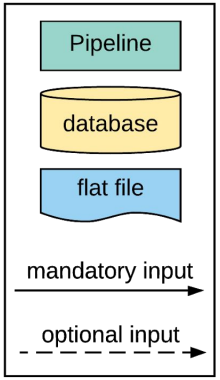


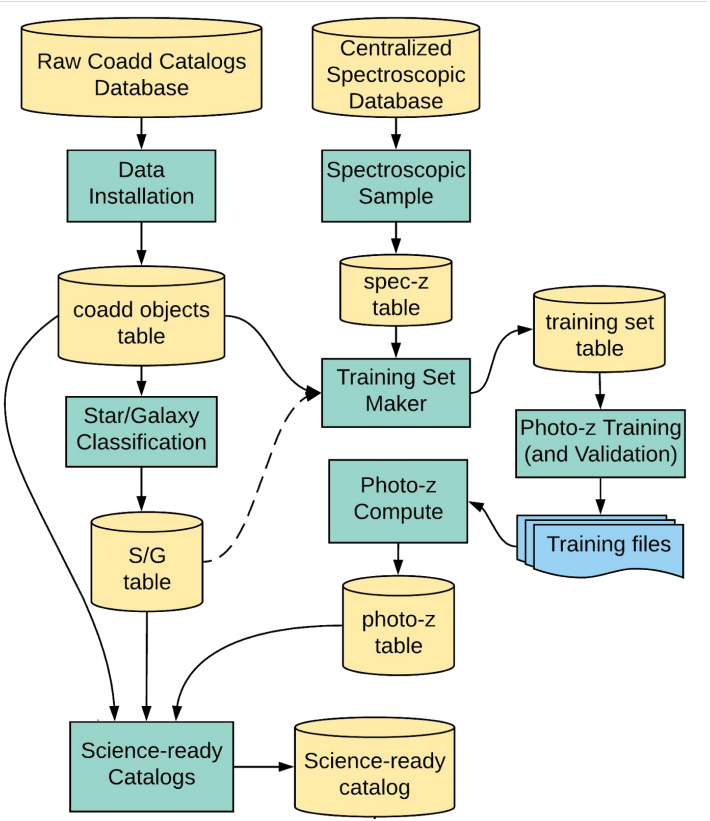
End-to-end (E2E)



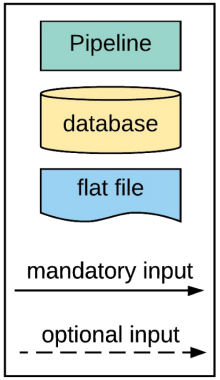


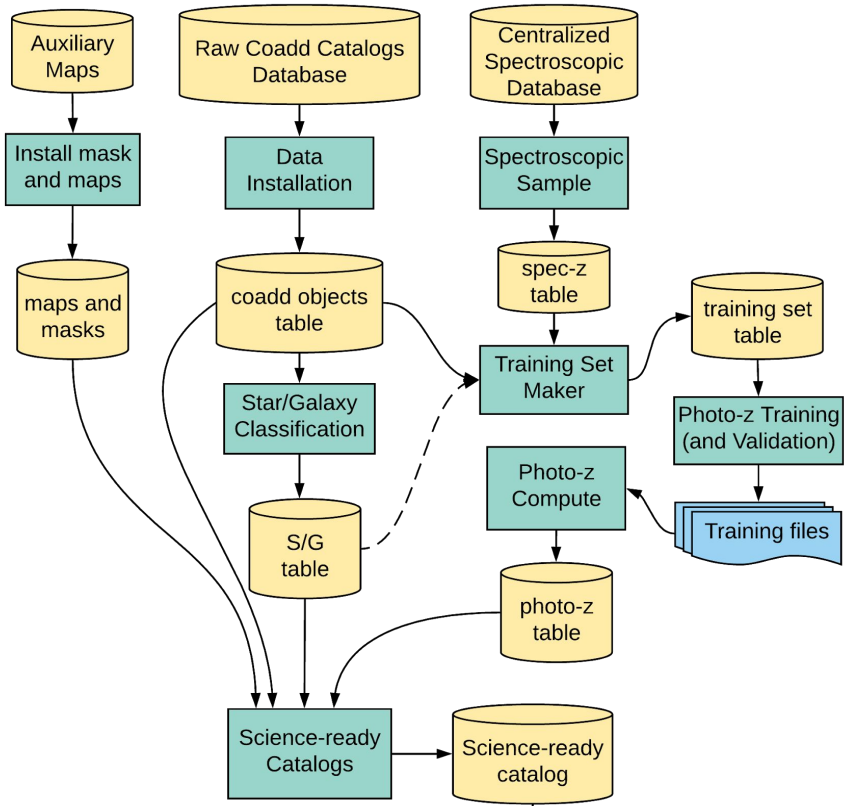
End-to-end (E2E)





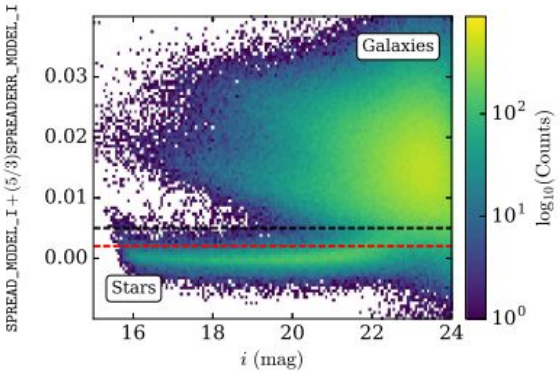
End-to-end (E2E)

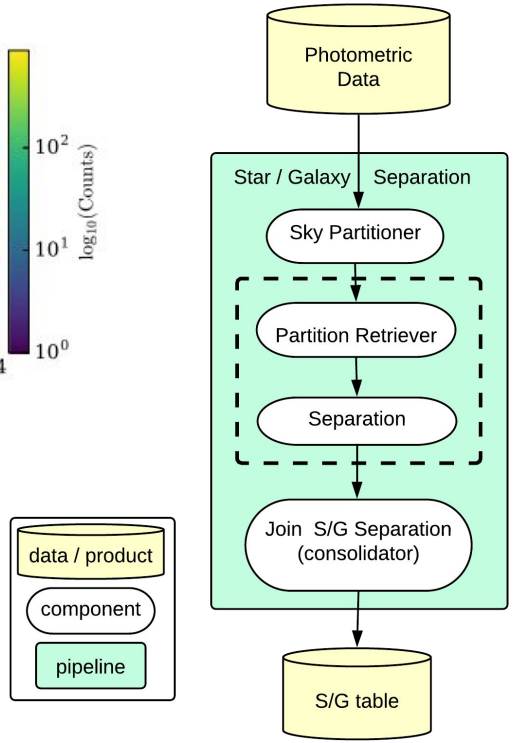




Star/galaxy separation

- Objetivo
  - Classificar objetos quanto a sua natureza





Star/galaxy separation

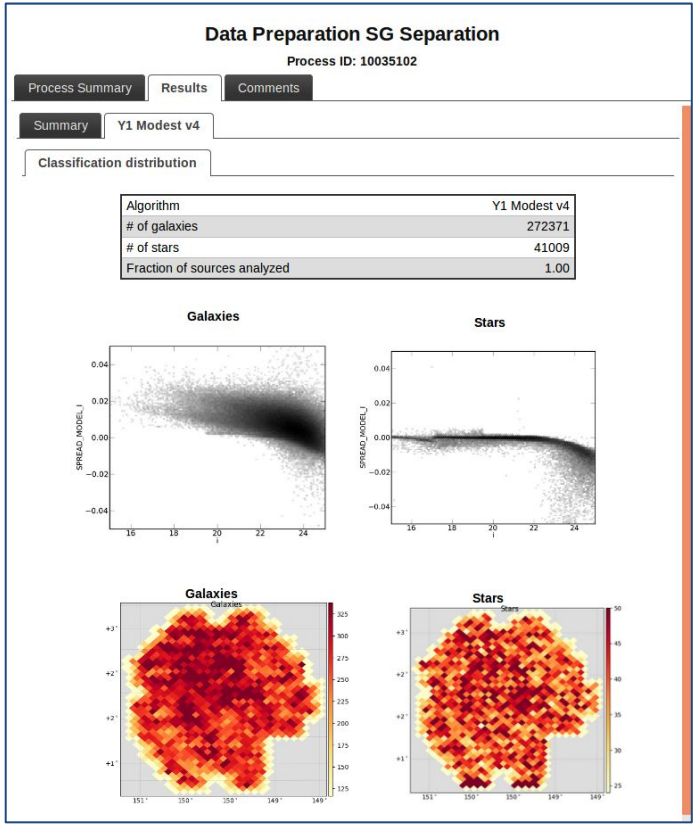
- Objetivo
  - Classificar objetos quanto a sua natureza
- Inputs
  - Tabela coadd (HDFS)
- Outputs
  - Tabela com classificação (id, class), onde class: 0=galáxia e 1=estrela

```
<input>
    <product class="coadd_objects" />
    <product class="slr" optional="True"/>
</input>
```

```
<output>
    <product class="class_star" />
    <product class="spread_model" />
    <product class="modest" />
    <product class="y1_modest_v1" />
    <product class="y1_modest_v2" />
    <product class="y1_modest_v3" />
    <product class="y1_modest_v4" />
    <product class="y3_extended_class_spread" />
    <product class="y3_extended_class_wavg_spread" />
    <product class="y3_extended_class_mof" />
    <product class="y3_extended_class_sof" />
</output>
```

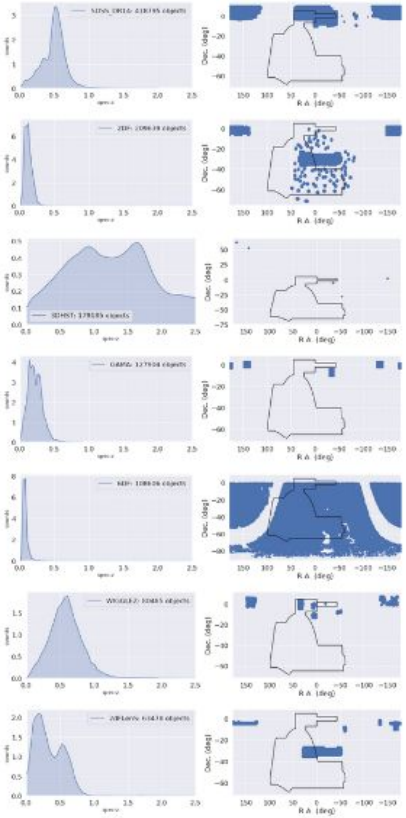
Star/galaxy separation

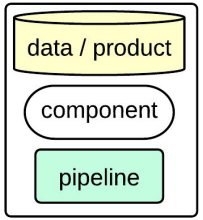
- Objetivo
  - Classificar objetos quanto a sua natureza
- Inputs
  - Tabela coadd (HDFS)
- Outputs
  - Tabela com classificação (id, class), onde class: 0=galáxia e 1=estrela
- Configurações relevantes
  - Paralelização
  - Diferentes métodos

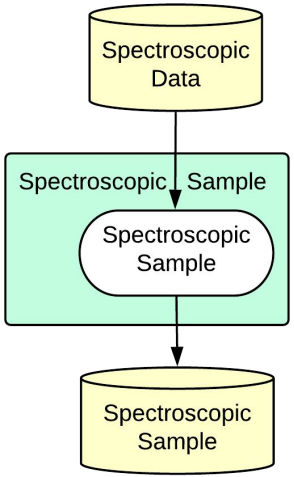


Spectroscopic Sample

- Objetivo
  - Reunir e padronizar dados espectroscópicos públicos obtidos de diversos levantamentos que servirão para o treinamento de códigos de photo-z







Spectroscopic Sample

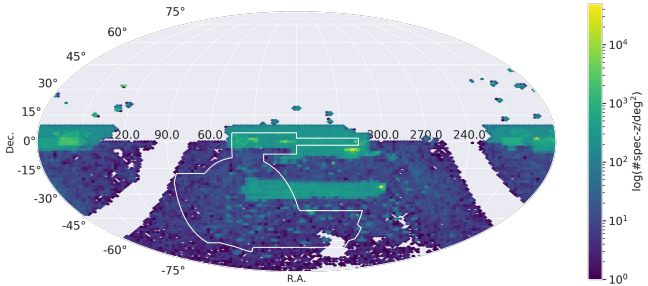
- Objetivo
  - Reunir e padronizar dados espectroscópicos públicos obtidos de diversos levantamentos que servirão para o treinamento de códigos de photo-z
- Inputs
  - Tabelas de spec-z dos surveys (DB)
- Outputs
  - Spectroscopic sample (id, z_spec, z_spec_err, flag, source)

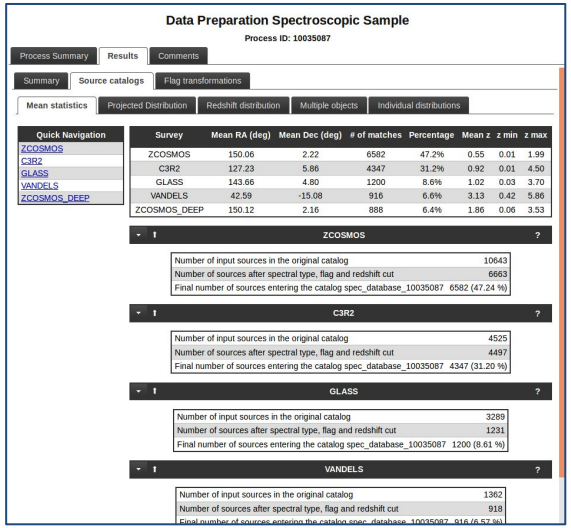
```
<input>
    <product class="external_spectroscopy" multivalue="True" />
</input>
```

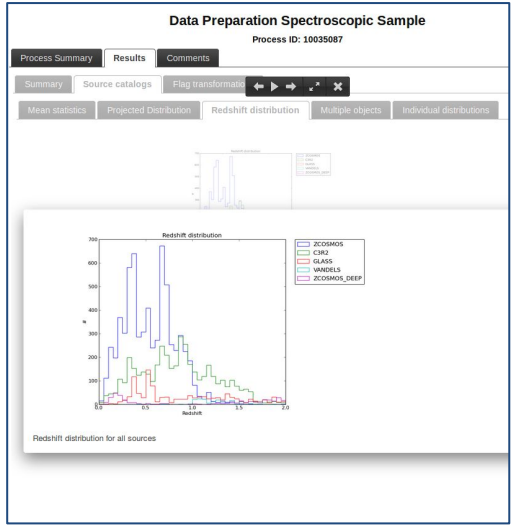
```
<output>
    <product class="training_set_spectroscopy" />
</output>
```

Spectroscopic Sample

- Objetivo
  - Reunir e padronizar dados espectroscópicos públicos obtidos de diversos levantamentos que servirão para o treinamento de códigos de photo-z
- Inputs
  - Tabelas de spec-z dos surveys (DB)
- Outputs
  - Spectroscopic sample (id, z_spec, z_spec_err, flag, source)
- Configurações relevantes
  - Como lidar com múltiplas medidas da mesma galáxia
  - Threshold de qualidade

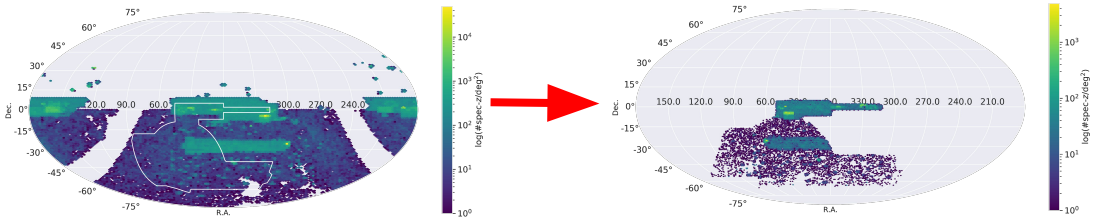


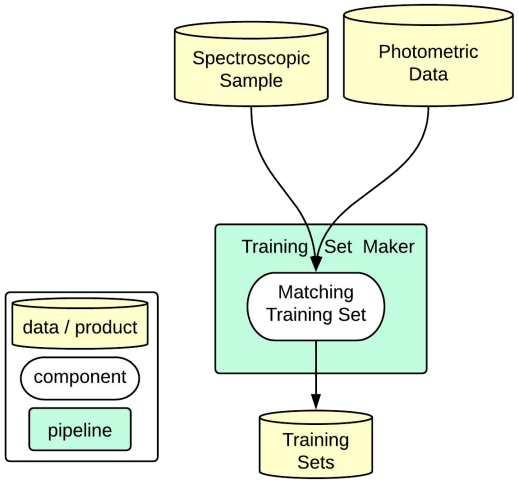




Training Set Maker

- Objetivo
  - Associar dados do DES (magnitudes e erros) às galáxias do Spec. Sample





Training Set Maker

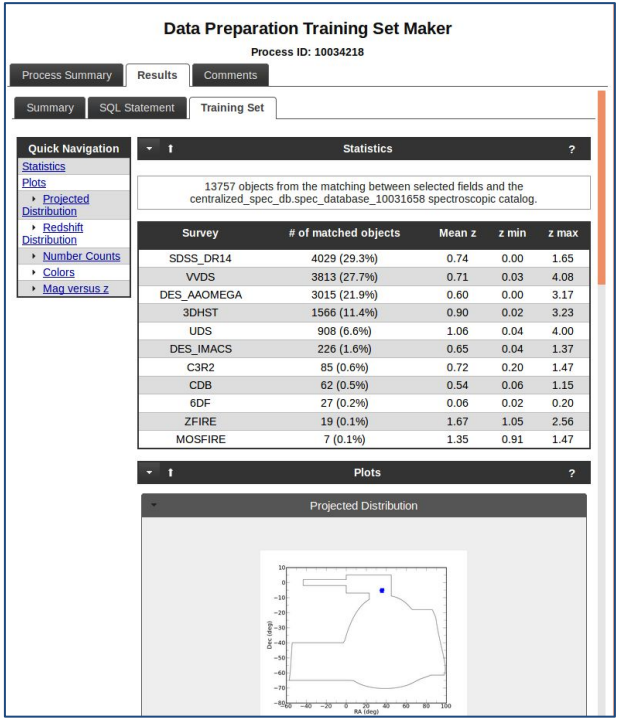
- Objetivo
  - Associar dados do DES (magnitudes e erros) às galáxias do Spec. Sample
- Inputs
  - Tabela coadd (DB)
  - Spec. Sample (DB)
- Outputs
  - Training Set

```
<input>
    <!-- Photo Sample -->
    <product class="coadd_objects" multivalue="True" optional="True"/>
    <product class="bcc_objects" optional="True" />
    <product class="mof_objects" optional="True" />
    <product class="slr" optional="True" multivalue="True" />
    <!-- S/G Separation -->
    <product class="class_star" optional="True" multivalue="True"/>
    <product class="spread_model" optional="True" multivalue="True"/>
    <product class="modest" optional="True" multivalue="True"/>
    <product class="y1_modest_v1" optional="True" multivalue="True"/>
    <product class="y1_modest_v2" optional="True" multivalue="True"/>
    <product class="y1_modest_v3" optional="True" multivalue="True"/>
    <product class="sg_y3_mesh" optional="True" multivalue="True"/>
    <product class="y3_extended_class" optional="True" multivalue="True"/>
    <!-- Spec Sample -->
    <product class="training_set_spectroscopy" />
</input>
```

```
<output>
    <product class="training_set" />
</output>
```

Training Set Maker

- Objetivo
  - Associar dados do DES (magnitudes e erros) às galáxias do Spec. Sample
- Inputs
  - Tabela coadd (DB)
  - Spec. Sample (DB)
- Outputs
  - Training Set
- Configurações relevantes
  - Raio de busca
  - Cortes de qualidade



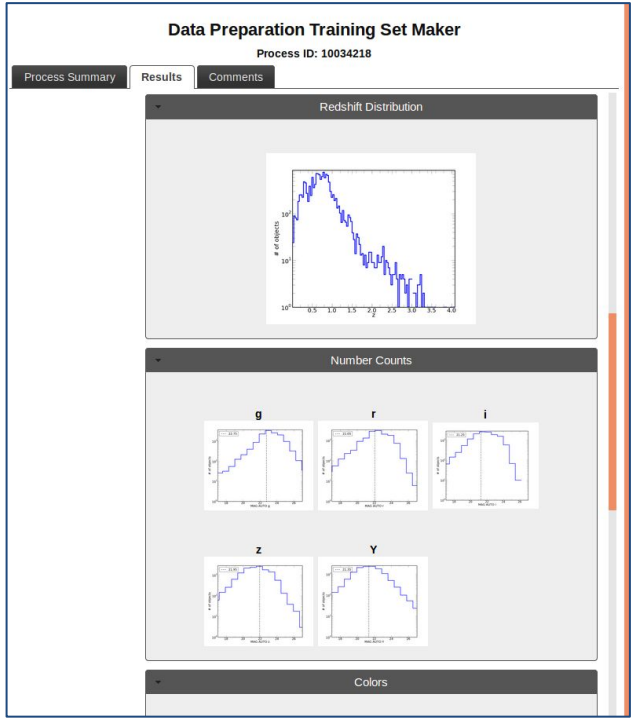


Photo-z Training

- Objetivos
  - Treinar algoritmos de empíricos de photo-z
  - Testar a performance dos códigos
  - Validar os resultados

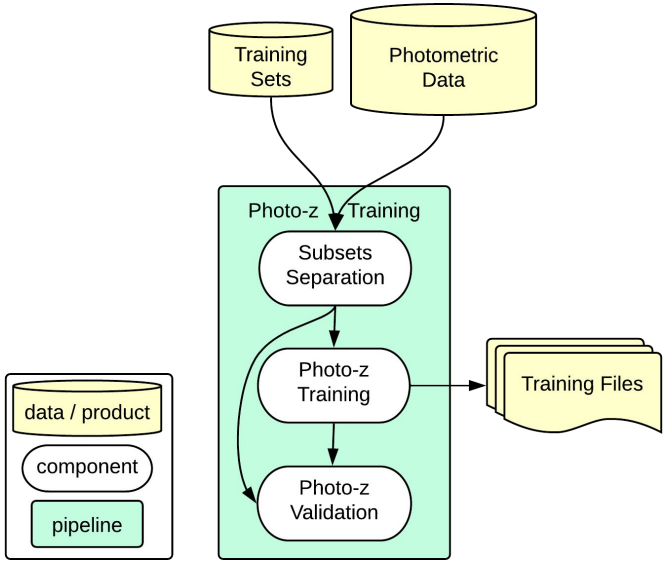


Photo-z Training

- Objetivos
  - Treinar algoritmos de empíricos de photo-z
  - Testar a performance dos códigos
  - Validar os resultados
- Inputs
  - Training Set (id, z_spec, mags, errs) (DB)
- Outputs
  - Training files (formato depende do código)

```
<input>
    <product class="coadd_objects" optional="True" multivalue="True" />
    <product class="mice_objects" optional="True"/>
    <product class="bcc_objects" optional="True"/>
    <product class="mof_objects" optional="True"/>
    <product class="training_set" optional="True" multivalue="True" />
</input>
```

```
<output>
    <product class="annz_training" />
    <product class="annz2_training" />
    <product class="arborz_training" />
    <product class="bpz_training" />
    <product class="dnf_training" />
    <product class="lephare_training" />
    <product class="mlz_training" />
    <product class="pofz_training" />
    <product class="skynet_training" />
</output>
```

Photo-z Training

- Objetivos
  - Treinar algoritmos de empíricos de photo-z
  - Testar a performance dos códigos
  - Validar os resultados
- Inputs
  - Training Set (id, z_spec, mags, errs) (DB)
- Outputs
  - Training files (formato depende do código)
- Configurações relevantes
  - Separação em subconjuntos para treinamento e validação
  - Pesos baseados na amostra fotométrica
  - Configurações específicas dos códigos

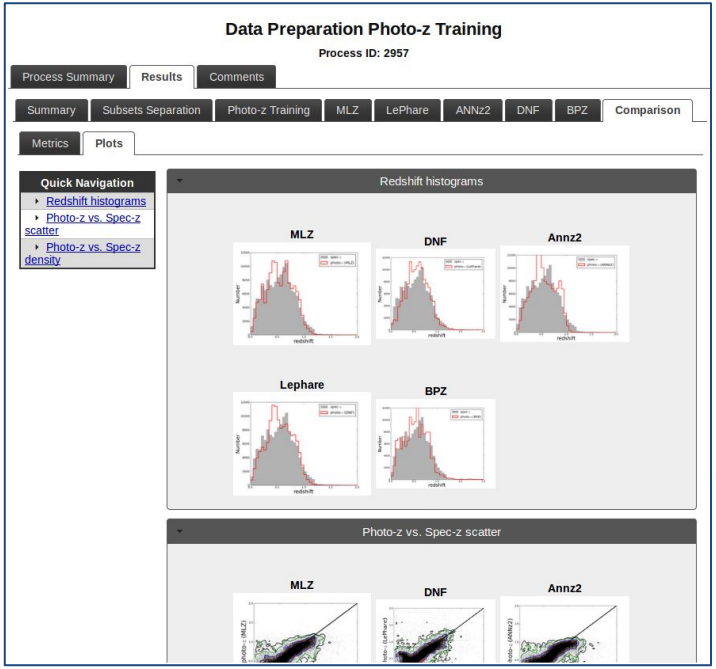
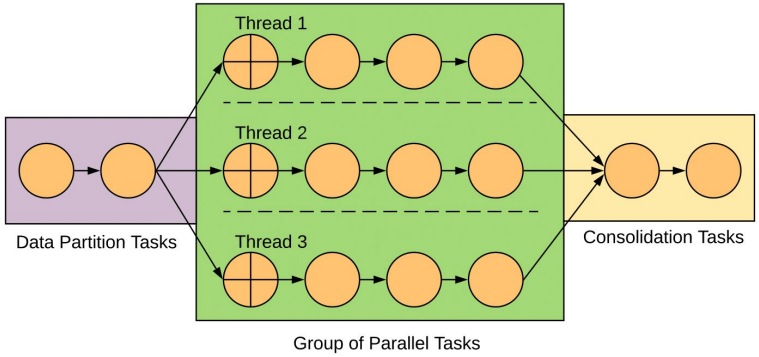


Photo-z Compute

- Objetivo
  - Calcular photo-z para grandes datasets



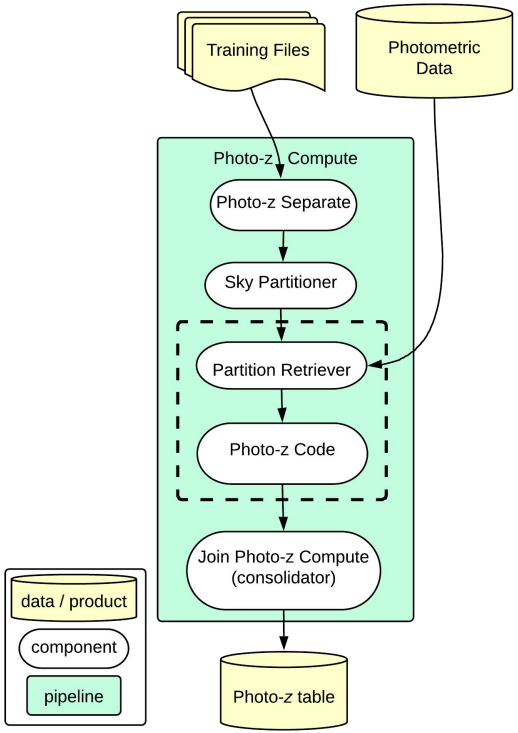


Photo-z Compute

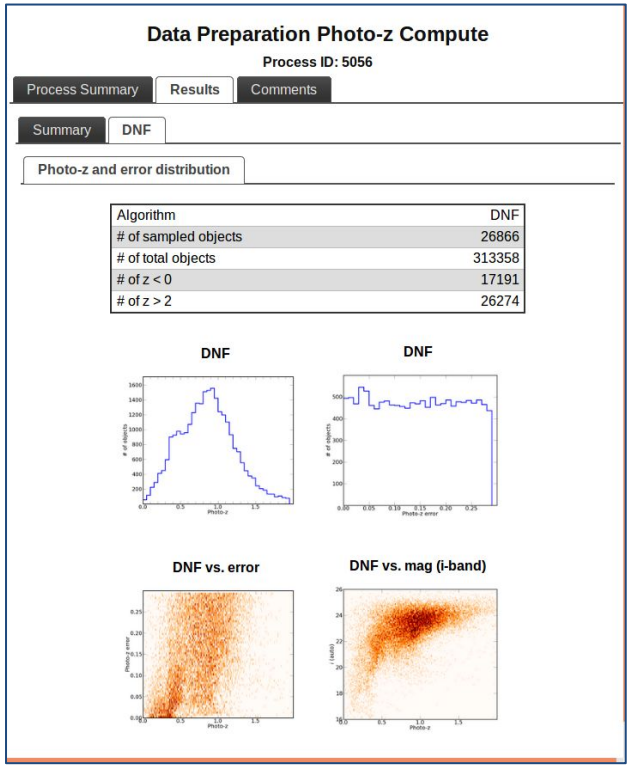
- Objetivo
  - Calcular photo-z para grandes datasets
- Inputs
  - Tabela coadd (HDFS)
  - Training files (archive)
- Outputs
  - Tabela de photo-z
    (id, z_best, z_err)

```
<input>
    <product class="coadd_objects" optional="True"/>
    <product class="mice_objects" optional="True"/>
    <product class="dnf_training" optional="True"/>
</input>
```

```
<output>
    <product class="annz" />
    <product class="annz2" />
    <product class="dnf" />
    <product class="mlz" />
    <product class="bpz" />
    <product class="lephare_photoz" />
</output>
```

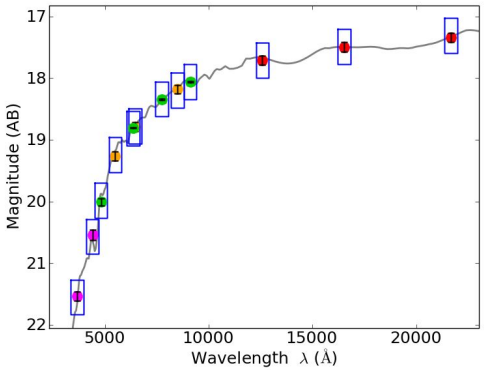
Photo-z Compute

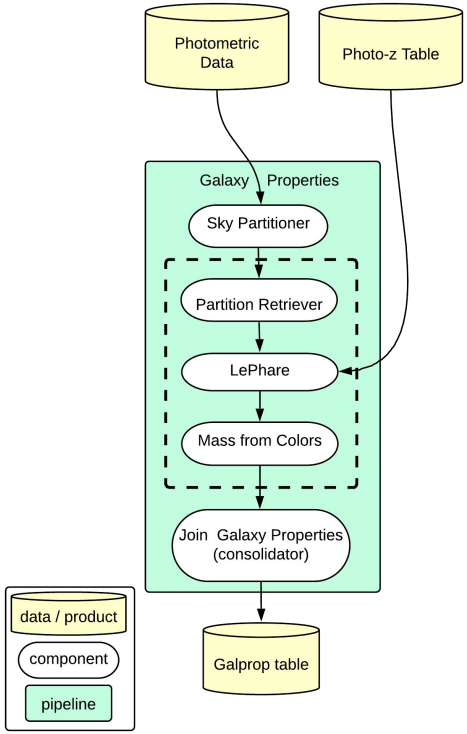
- Objetivo
  - Calcular photo-z para grandes datasets
- Inputs
  - Tabela coadd (HDFS)
  - Training files (archive)
- Outputs
  - Tabela de photo-z
    (id, z_best, z_err)
- Configurações relevantes
  - Paralelização
  - Configurações específicas dos códigos são herdadas do processo anterior (training)



Galaxy Properties

- Objetivo
  - Calcular propriedades adicionais de galáxias (M, mass, spectral type, E(B-V), age, metallicity, SFR, SSFR, SFH, etc) a partir de um photo-z previamente estimado





Galaxy Properties

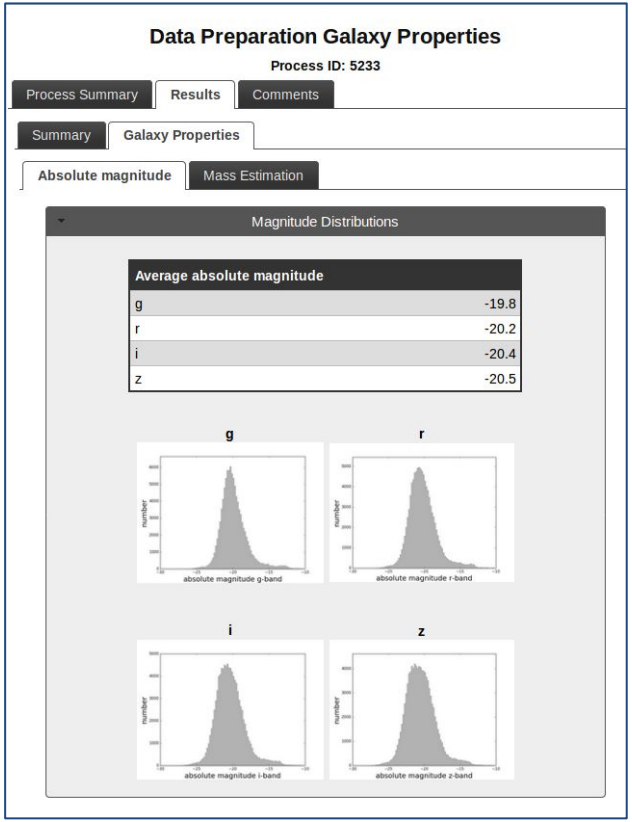
- Objetivo
  - Calcular propriedades adicionais de galáxias (M, mass, spectral type, E(B-V), age, metallicity, SFR, SSFR, SFH, etc) a partir de um photo-z previamente estimado
- Inputs
  - Tabela coadd (HDFS)
  - Tabela photo-z (DB)
- Outputs
  - Tabela galprop

```
1 <input>
2     <!-- Objects Catalog -->
3     <product class="coadd_objects" optional="True"/>
4     <product class="bcc_objects"  optional="True"/>
5     <product class="mice_objects"  optional="True"/>
6     <!-- Photoz -->
7     <product class="annz2" optional="True"/>
8     <product class="dnf"  optional="True"/>
9 </input>
```

```
1 <output>
2     <product class="lephare" />
3 </output>
```

Galaxy Properties

- Objetivo
  - Calcular propriedades adicionais de galáxias (M, mass, spectral type, E(B-V), age, metallicity, SFR, SSFR, SFH, etc) a partir de um photo-z previamente estimado
- Inputs
  - Tabela coadd (HDFS)
  - Tabela photo-z (DB)
- Outputs
  - Tabela galprop
- Configurações relevantes
  - Paralelização
  - Configs específicas LePhare / BPZ



Value-added catalog (VAC) → science-ready catalogs

- Objetivo
  - Criar catálogos customizados para os pipelines científicos, contendo apenas os dados necessários e incluindo produtos criados no estágio de Data Preparation

Value-added catalog (VAC) → science-ready catalogs

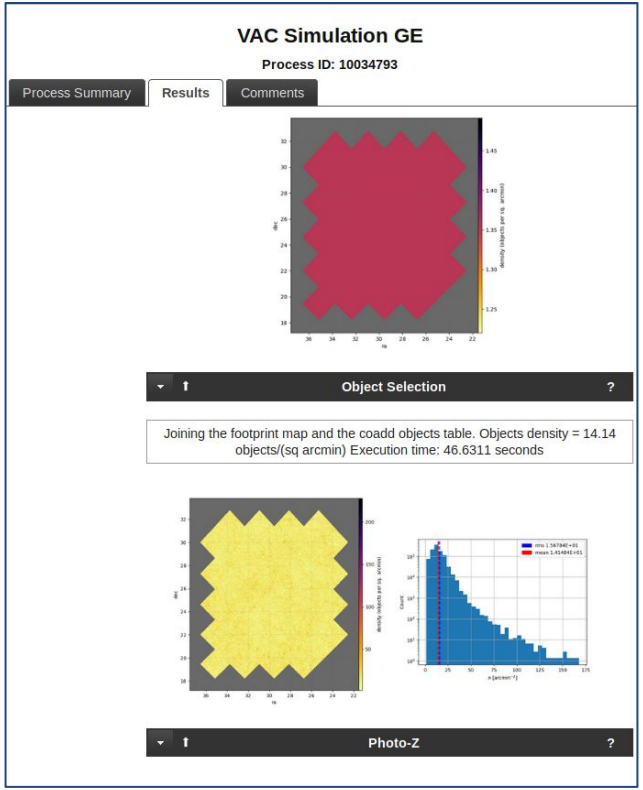
- Objetivo
  - Criar catálogos customizados para os pipelines científicos, contendo apenas os dados necessários e incluindo produtos criados no estágio de Data Preparation
- Inputs
  - Mapas (DB)
  - Tabela coadd (DB)
  - Tabela SG sep (DB)
  - Tabela photo-z (DB)
  - Tabela galprop (DB)
- Outputs
  - VAC
  - Mapa de footprint

```
1 <output>
2     <product class="vac_generic" />
3     <product class="vac_lss" />
4     <product class="vac_ge" />
5     <product class="vac_ga" />
6     <product class="vac_cluster" />
7     <product class="footprint_map" />
8 </output>
```

```
1 <input>
2
3     <!-- Objects Catalog  -->
4     <product class="coadd_objects" optional="True" />
5     <product class="bcc_objects" optional="True" />
6     <!-- Photoz -->
7     <product class="annz" optional="True" />
8     <product class="lephare_photoz" optional="True" />
9     <product class="mlz" optional="True" />
10    <product class="dnf" optional="True" />
11    <product class="skynet" optional="True" />
12    <product class="zebra" optional="True" />
13    <product class="arborz" optional="True" />
14    <product class="annz2" optional="True" />
15    <product class="pofz" optional="True" />
16    <product class="bpz" optional="True" />
17    <!-- S/G Separation -->
18    <product class="class_star" optional="True" />
19    <product class="spread_model" optional="True" />
20    <product class="modest" optional="True" />
21    <product class="y1_modest_v1" optional="True" />
22    <product class="y1_modest_v2" optional="True" />
23    <product class="y1_modest_v3" optional="True" />
24    <product class="y1_modest_v4" optional="True"/>
25    <product class="y3_extended_class_spread" optional="True"/>
26    <product class="y3_extended_class_wavg_spread" optional="True"/>
27    <product class="y3_extended_class_mof" optional="True"/>
28    <product class="y3_extended_class_sof" optional="True"/>
29    <!-- Galaxy Properties -->
30    <product class="lephare" optional="True" />
31
32 </input>
```

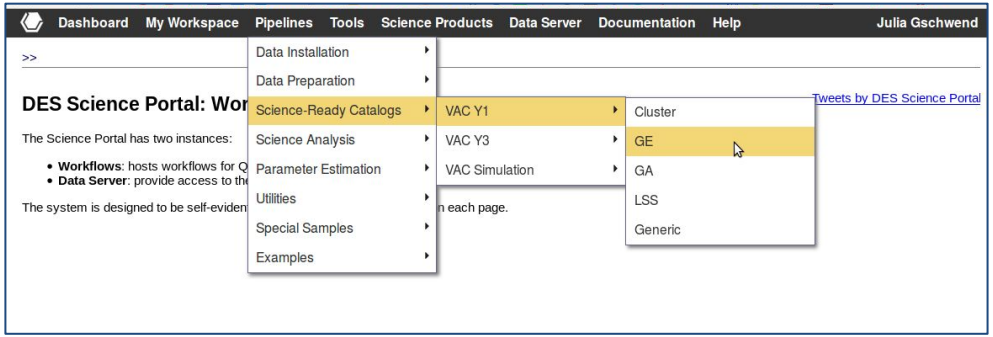
Value-added catalog (VAC) → science-ready catalogs

- Objetivo
  - Criar catálogos customizados para os pipelines científicos, contendo apenas os dados necessários e incluindo produtos criados no estágio de Data Preparation
- Inputs
  - Mapas (DB)
  - Tabela coadd (DB)
  - Tabela SG sep (DB)
  - Tabela photo-z (DB)
  - Tabela galprop (DB)
- Outputs
  - VAC
  - Mapa de footprint
- Configurações relevantes
  - Seleção de regiões
  - Seleção de objetos
  - Seleção de colunas



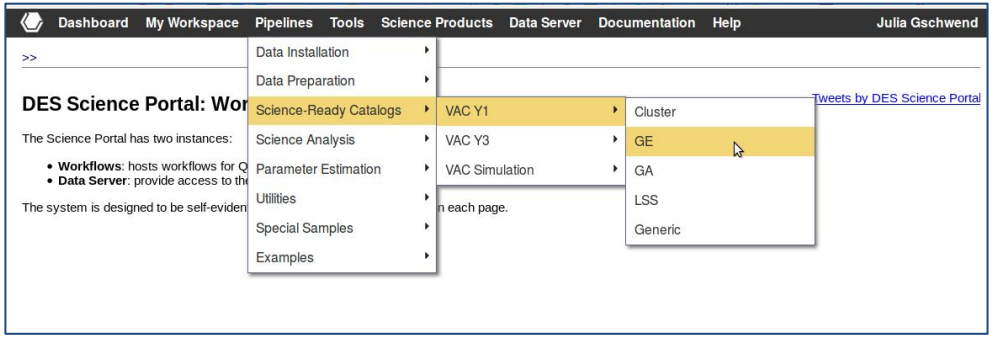
Value-added catalog (VAC)

- Um pipeline para cada grupo de trabalho
  - VAC Cluster
  - VAC GA
  - VAC GE
  - VAC LSS
  - VAC Generic



Value-added catalog (VAC)

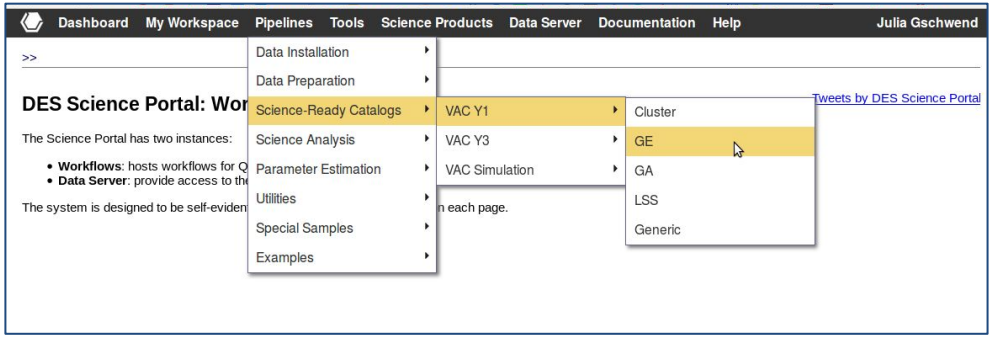
- Um pipeline para cada grupo de trabalho
  - VAC Cluster
  - VAC GA
  - VAC GE
  - VAC LSS
  - VAC Generic



- Todos compartilham os mesmos componentes (variam apenas os inputs e configurações default)
  - query builder
  - catalog properties

Value-added catalog (VAC)

- Um pipeline para cada grupo de trabalho
  - VAC Cluster
  - VAC GA
  - VAC GE
  - VAC LSS
  - VAC Generic



- Todos compartilham os mesmos componentes (variam apenas os inputs e configurações default)
  - query builder
  - catalog properties

Query Builder

1. Region Selection (maps and masks combined)
2. Object Selection (quality cuts, sample selection)
3. Value added quantities (extra columns)

Value-added catalog (VAC)

Ancilliary maps

Value-added catalog (VAC)

Mangle Fraction Map i ?

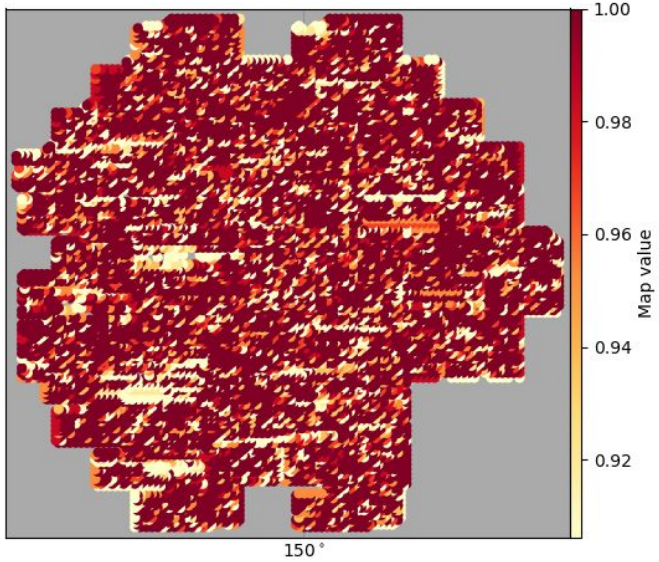
Regions with mangle detfract larger than 0.8 in i. table: vac_cluster.mangle_maps_area_fraction_i_10029880
Execution time: 3.0176 seconds

```
SELECT a.pixel,
a.signal,
a.ra,
a.dec
FROM y1a1_coadd_cosmos_d04.mangle_maps_4096_ring_frac_i_10028293 a
WHERE signal >= 0.8
```

Value-added catalog (VAC)

Ancilliary maps

Ex. 1: detection fraction map



Value-added catalog (VAC)

Depth Map i snr 5 ?

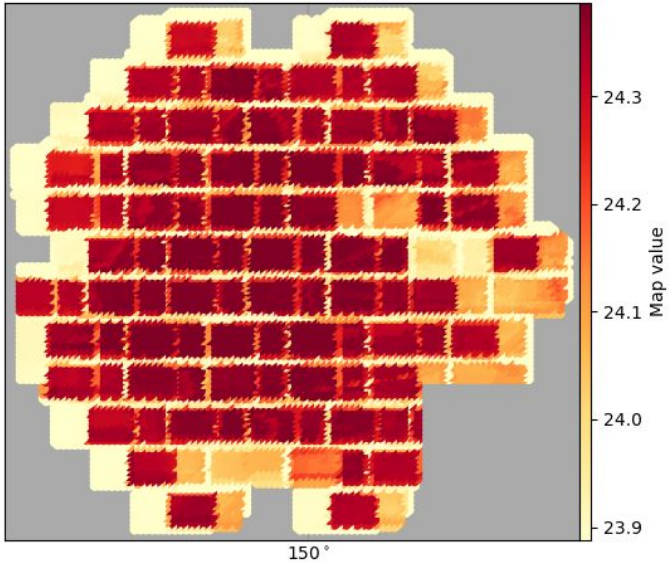
Regions with depth at 5 sigma; larger than 22 in i. table: vac_cluster.mag_auto_i_10029880 Execution time: 2.7850 seconds

```
SELECT a.pixel,
a.signal,
a.ra,
a.dec
FROM depth_maps.y1a1_coadd_cosmos_d04_5_auto_4096_ring_i_10024579 a
WHERE signal >= 22
```

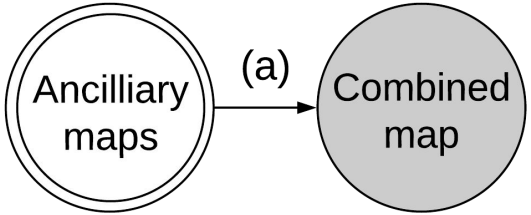
Value-added catalog (VAC)

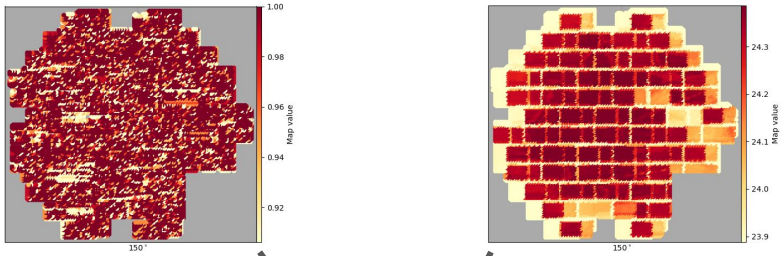
Ancilliary maps

Ex. 2: depth map (magnitude limit 10σ *i*-band)



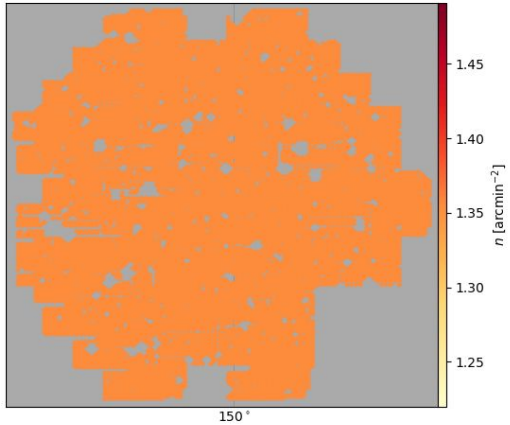
Value-added catalog (VAC)



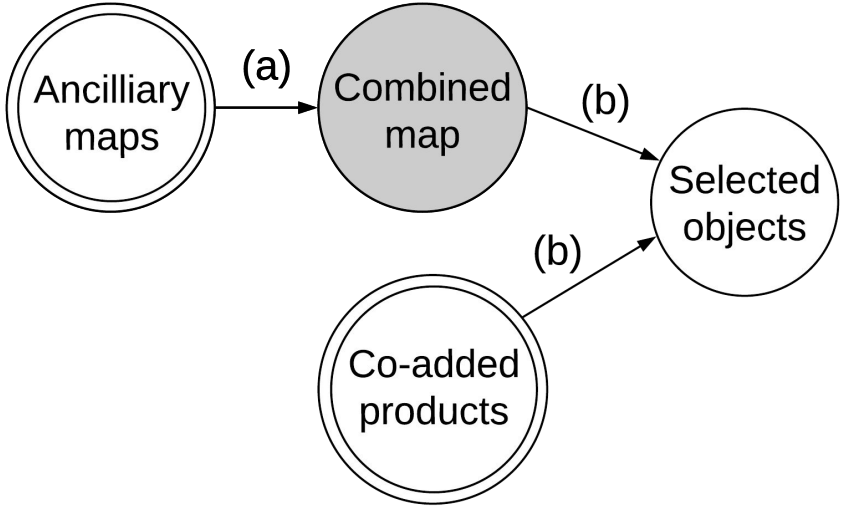


Ex.:
depth >= 22.0
Fracdet >=0.8

<u>Footprint Map</u> (combined map):
0: pixels out
1: pixels in



Value-added catalog (VAC)



Value-added catalog (VAC)

Sextractor: reference flags

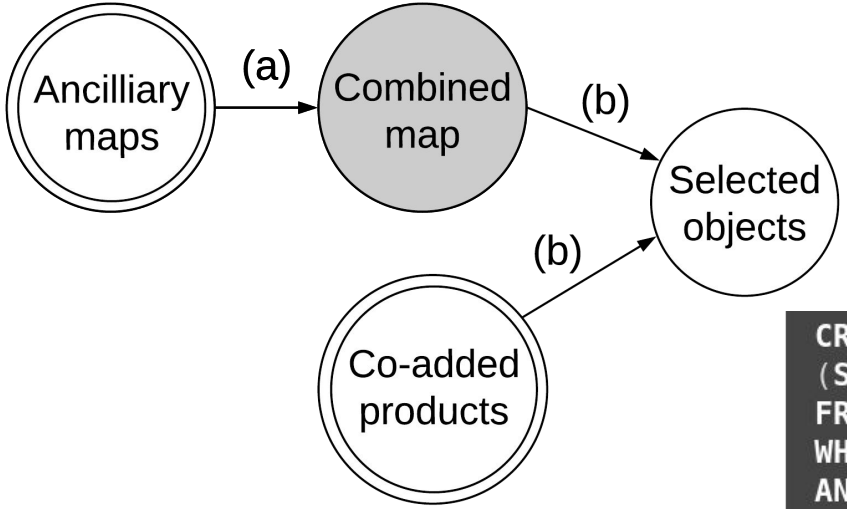
- [x] 0 - Clean object
- [x] 1 - The object has neighbours, bright and close enough to significantly bias the MAG_AUTO photometry, or bad pixels
- [x] 2 - The object was originally blended with another one.
- [ ] 4 - At least one pixel of the object is saturated (or very close to)
- [ ] 8 - The object is truncated (too close to an image boundary)
- [ ] 16 - Object aperture data are incomplete or corrupted
- [ ] 32 - Object isophotal data are incomplete or corrupted
- [ ] 64 - A memory overflow occurred during deblending
- [ ] 128 - A memory overflow occurred during extraction

Value-added catalog (VAC)

Magnitude: color cuts

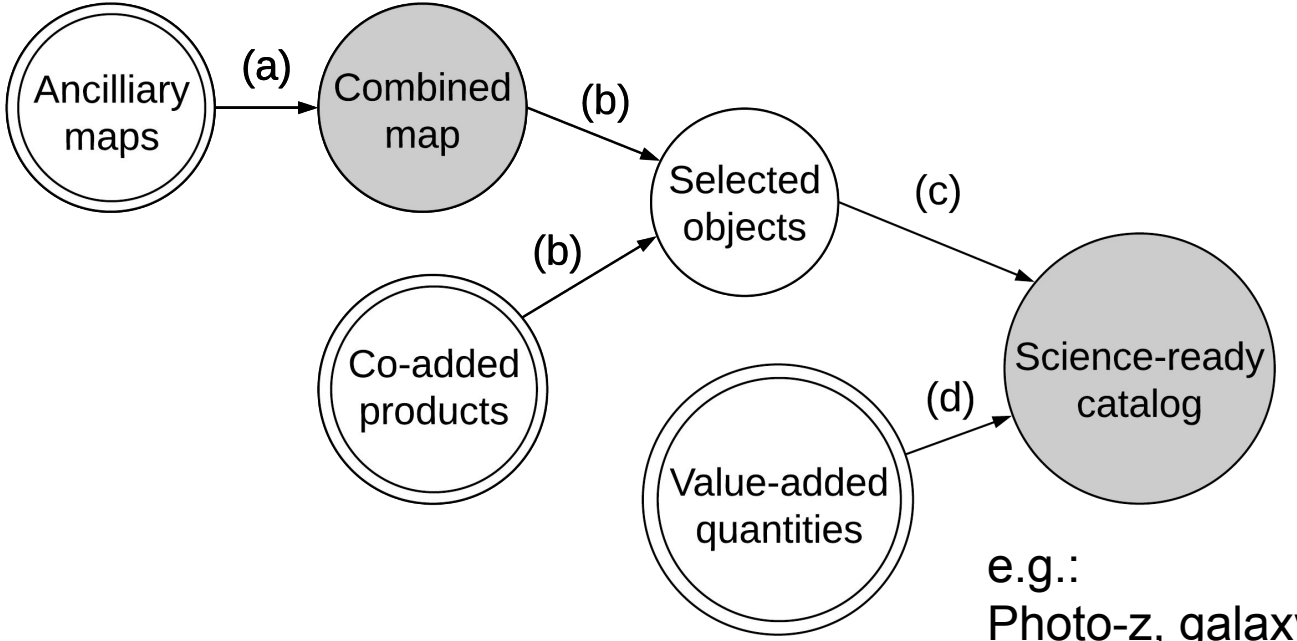
g-r color limits	-2.0, 4.0
r-i color limits	-2.0, 4.0
i-z color limits	-2.0, 4.0
z-Y color limits	-2.0, 4.0

Value-added catalog (VAC)




```
CREATE TEMP TABLE temp_cuts_10031782 AS
(SELECT *
FROM temp_cuts_10031782_inner
WHERE (mag_auto_i < 22)
AND (mag_auto_i > 15)
AND (mag_auto_g - mag_auto_r BETWEEN -2.0 AND 4.0
AND mag_auto_r - mag_auto_i BETWEEN -2.0 AND 4.0
AND mag_auto_i - mag_auto_z BETWEEN -2.0 AND 4.0
AND mag_auto_z - mag_auto_Y BETWEEN -2.0 AND 4.0));
```

Value-added catalog (VAC)



e.g.:
Photo-z, galaxy properties

Value-added catalog (VAC)

S/G Classification

- [ ] Select stars
- [x] Select galaxies

Reference band used by the classifier: i

Galaxy Separation ?

Density map after star/galaxy separation. table: vac_cluster.galaxy_separation_10029880 Execution time: 3.4172 seconds

```
SELECT a.*
FROM vac_cluster.object_selection_map_10029880 a
INNER JOIN sg_separation.y1_modest_v3_10028487 b ON a.coadd_objects_id= b.coadd_objects_id
WHERE b.i='0'
```

Value-added catalog (VAC)

Photo z

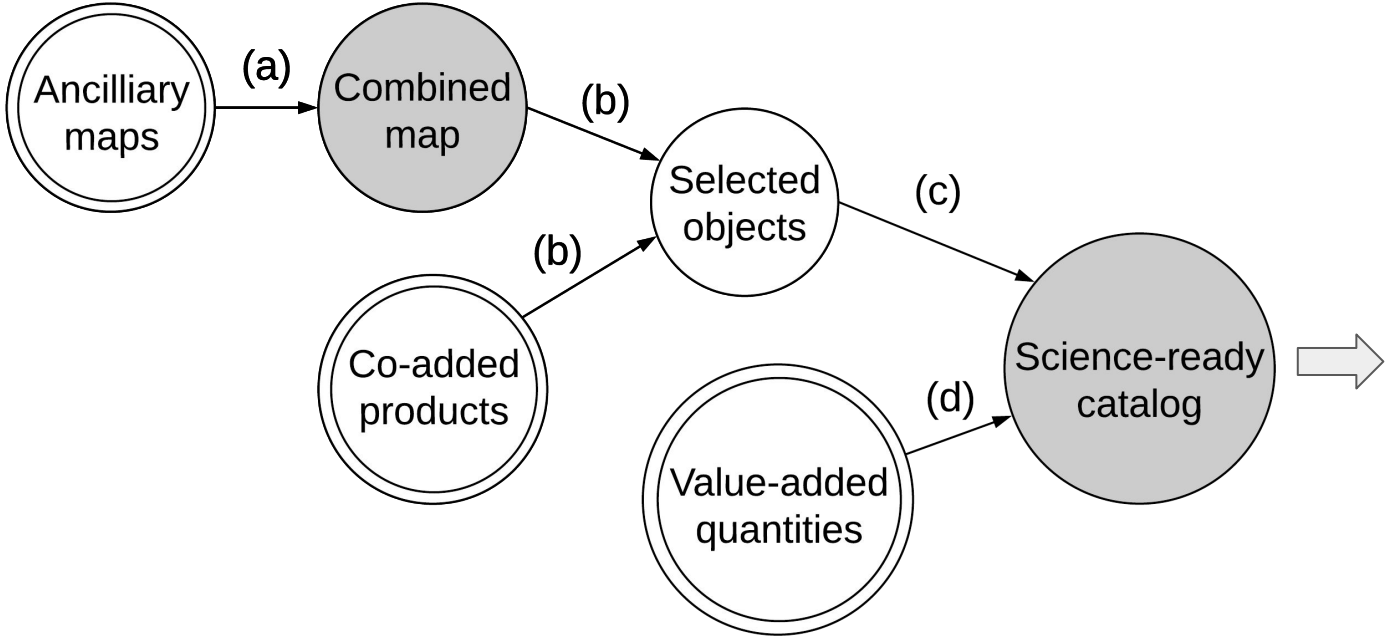
Z min	0
Z max	2.0

Photo-Z ?

Density map after Photo-z. table: vac_cluster.photo_z_10029880 Execution time: 3.4715 seconds

```
SELECT a.*,
z_best,
err_z
FROM vac_cluster.galaxy_separation_10029880 a
INNER JOIN photoz_compute.dnf_10029496 b ON a.coadd_objects_id= b.coadd_objects_id
WHERE b.z_best > 0
AND b.z_best < 2.0
```

Value-added catalog (VAC)



Science Pipelines

Cluster
- WAZP
- Cluster Matching
- WAZP Afterburner
- ...

Galaxy Evolution
- Luminosity Function
- Mass Function
- Halo Occupation Distribution
- ...

Galaxy Archeology
- StarHorse
- MW Fitting
- Sparsex
- ...

Large Scale Structure
- Angular Correlation Function
- Theoretical ACF Fit
- ...

Parameter Estimation
- Combining Probes
- ...

Value-added catalog (VAC)

Exemplo de product log de um VAC:

Processo: 5464

Produto: GE 21

VAC Y3 GE

Process ID: 5464

Process Summary | Results | Comments

Input Data	
Release	Y3A2
Dataset	Y3 Gold Full
Objects Catalog	Coadd Objects 1
Galaxy Properties	Lephare 3
Photo-z	DNF 7

Additional Information	
Zp correction	None
Extinction map	None
Footprint area:	3437.75 sq deg
Mean density:	13.98 objects/(sq arcmin)
Pipeline name:	VAC GE

Process Information	
Stage	Science-Ready Catalogs
Process ID	5464
User	Julia Gschwend
Used Config	Custom
Start	2019-07-15 19:07:40
End	2019-07-15 22:59:43
Execution Time	3h 52m 03s
Expiration Date	2019-07-22 22:59:43
Size	8229
Status	Success
Code Viewer	
Overall Success Rate	100%
Total Number of Jobs	2
Time Profiler	

Output Data	
Region Selection	Footprint map 33
Value-Added Catalogs	GE 21

Module	Duration	Config	Error Log	Pipeline Out	Log	Condor Log	NC	Success Rate	Status
Query Builder	3:13:50						null	-	✔
Catalog Properties	0:36:53						null	-	✔

Dúvidas?

Extra

Sessão 33 - DES Portal - Data Preparation

Calendário do Bootcamp

	2	3	4	5	6
9:00 - 9:30	**1- Introdução ao LIneA** L. Nicolaci - 45'	**9- Introdução ao LIneA II** L. Nicolaci - 45'	**17- Usando o Santos Dumont** V. Lima - 30'	**Colaborações Científicas** **25- DES** - B. Santiago - 25' **26- DESI** - F. Oliveira - 25' **27- SDSS** - J. Schimoia - 25' **28- TON** - J. Camargo - 25'	**37- Documentação** C.Silveira - 30'
9:30 - 10:00	**2- LIneA: Como se registrar** G. Soares - 30'	**10- Trabalhando no LIneA** A. Nunes - 45'	**18- Usando o Santos Dumont: Casos de Uso** H. Camacho, A. Pieres - 30'		**38-Transferência de Dados** J. Souza - 20' **39- Máquinas Virtuais, Containers e Kubernetes** J. Souza - 20'
10:00 - 10:30	**Coffee break** 15'		**Foto e Coffee break** 30'		**40- Rede Nacional de Ocultação** G. Rossi - 30'
10:30 - 11:00	**3- LIneA: Infraestrutura** C. Adean - 45'	**Coffee break** 15' **11- Ferramentas Colaborativas** F. Massena - 30'	**19- Uso Avançado da Infraestrutura** C. Adean - 45'	**Coffee break** 20' (inicio 10:40)	**Coffee break** 20' (inicio 10:40)
11:00 - 11:30	**4- Landing Page do LIneA** C. Silveira - 30'	**12- Ferramentas Astronômicas I** C. Singulani - 30'	**20- Portal Cientifico: Introdução** J.Gschwend - 45'	**Colaborações Científicas** **29- LSST** - L. Nicolaci - 30'	**41- Ciberinfraestrutura para Apoio à e-Ciência** A. Moura - 30'
11:30 - 12:00				**30- Atividades Científicas** A. Pieres - 30'	**42- Atividades de EPO** E. Nogueira - 30'
12:00 - 12:30	**Almoço** 75'	**Almoço** 75'	**Almoço** 60'	**Almoço** 60'	**Almoço** 60'
12:30 - 13:00					
13:00 - 13:30	**5- Spotlight talks** R. Boufleur - 90'	**13- Ferramentas Astronômicas II** G. Costa - 60'	**21- Criando um Pipeline no Portal I** C. Singulani, L. Nunes - 60'	**31- Analisando DES Y6** R. Gruendl - 60´	**43- Planejando o Futuro** A. Nunes, A. Pieres - 150'
13:30 - 14:00					
14:00 - 14:30		**14- Jupyter Notebooks: Criando** J.Gschwend - 30'	**22- Portal Científico: Workflows Científicos** A. Pieres, M. Aguena - 45'	**32- DES - Primary Products** A. Pieres - 45´	
14:30 - 15:00	**6- Serviços** F. Massena - 30'	**15- Jupyter Notebooks: Usando e Desenvolvendo** A. Pieres, M. Aguena - 60'	**23- Criando um Pipeline no Portal II** J.Gschwend - 60'	**33- DES Portal - Data Preparation** J.Gschwend - 45´	
15:00 - 15:30	**7- Jupyter Notebooks: Introdução** G. Soares - 30'				
15:30 - 16:00	**Coffee break** 30'	**Coffee break** 30'	**Coffee break** 30'	**Coffee break** 30'	**44- Perspectivas do LIneA** L. Nicolaci - 30'
16:00 - 16:30	**8- Jupyter Notebooks: Exemplos** A. Pieres - 60'	**16- Jupyter Notebooks: Casos de Uso** F. Oliveira, R. Boufleur - 60'	**24- Requisitos para o Portal do LSST** M.Sanchez - 45'	**34- Validando Imagens DES Y6** A. Pieres - 30´	**Encerramento** 30'
16:30 - 17:00				**35- Validando Catálogos DES Y6** M. Aguena - 30´	**Legenda** Apresentação Tutorial Hacksession/Hands on
17:00 - 17:30	**Q&A** 30'	**Q&A** 30'	**Q&A** 30'	**36- Explorando Dados DES Y6** M. Aguena - 30´	

Value-added catalog (VAC)

Exemplo de science workflow utilizando GE 21:

Processo: 5543

Galaxy Evolution LF

Process ID: 5543

Process Summary | Results | Comments

Input Data	
Release	Y3A2
Dataset	Y3 Gold Full
Value-Added Catalogs	GE 21

Additional Information	
Total area:	3858.88 sq. deg

Process Information	
Stage	Science Analysis
Process ID	5543
User	Julia Gschwend
Used Config	Custom
Start	2019-08-04 01:45:54
End	2019-08-04 14:55:57
Execution Time	13h 10m 03s
Expiration Date	2019-08-11 14:55:57
Size	231110524
Status	Success
Code Viewer	
Overall Success Rate	100%
Total Number of Jobs	3282
Time Profiler	

Output Data	
Measurements	LF Measurement 31
Measurements	LF Median 30

Module	Duration	Config	Error Log	Pipeline Out	Log	Condor Log	NC	Success Rate	Status
Data Slicer	0:07:39						null	-	
Data Retriever	7:56:07	-	-	-	-	-		100%	
Luminosity Function	12:18:40		-	-	-	-		100%	
LF Consolidator	0:02:52						null	-	