
Site consistency tools

Rob Gardner • University of Chicago • USATLAS • MWT2

Facilities Jamboree
CERN, December 3, 2014

Site concerns

- Lost data
 - Consistency
 - Dark data
 - Deletion process
 - Hot and cold data
-

CCC - complete consistency checker

At MWT2 we have developed a tool to check consistency - “CCC”

`ccc_pnfs_rucio.py` will download the list of datasets listed in Rucio for a site, compute Rucio paths for each file in the datasets, and check that against what is in PNFS. It will additionally check PNFS against disk. A report is generated in the output directory with a summary of discrepancies found.

Available in Github: <https://github.com/DHTC-Tools/ATLAS/tree/master/Complete%20Consistency%20Check>

Report - flags:

GHOST: A file that is present in a higher level storage catalog, but not in a lower one or on disk. Ghost files often cause job failures when a job runs at the site and fails to fetch the missing input file.

ORPHAN: A file that is missing in the higher level storage catalog, but is present in the lower one or on disk. Orphans do not cause job failures. They are typically 'dark data', and may need to be removed manually.

DAMAGED DATASET: A dataset that is listed as having a complete replica at the site, but lower level catalogs or disk show missing files. Damaged datasets may cause job failures, for the same reason as ghost files. **MISSING DATASET:** A dataset that is listed as having a complete replica at the site, but lower level catalogs or disk show all files are missing. Missing datasets may cause job failures, for the same reason as ghost files.

INCOMPLETE DATASET: A dataset is listed as having an incomplete replica at the site, and the lower level catalog and file system confirm that. These are not a problem.

EMPTY DATASET: A dataset listed as having 0 files in the catalog. These are not a problem.

UNKNOWN DATASET: The catalog lists the dataset as present at the site, but when the catalog was queried for a list of files it returned an 'unknown dataset' error. If a dataset appears in this category repeatedly it may be broken in the catalog and need attention by the catalog admins.

OK DATASET: A dataset that is listed as having a complete replica at the site, and the lower level catalogs and file system show all files present. Most datasets should be in this category.

Example report

Started at Thu Nov 20 01:25:05 2014

ccc_pnfs_rucio.py running on uct2-int.mwt2.org

Checking pools against PNFS... [183 PNFS orphans \(101.8G\)](#)

Checking PNFS against pools... [64 PNFS ghosts](#), [32040 duplicates](#)

Checking DQ2 against LFC/Rucio

MWT2_UC_PHYS-HIGGS... 18129 replicas, 17838 complete

[149 replicas EMPTY](#)

[11 replicas DAMAGED](#)

[142 replicas INCOMPLETE](#)

[17827 replicas OK](#)

MWT2_UC_LOCALGROUPDISK... 36555 replicas, 36275 complete

[1 replica UNKNOWN](#)

[3 replicas EMPTY](#)

[16 replicas MISSING](#)

[42 replicas DAMAGED](#)

[277 replicas INCOMPLETE](#)

[36216 replicas OK](#)

MWT2_UC_USERDISK... 296364 replicas, 273856 complete

[60 replicas UNKNOWN](#)

[428 replicas EMPTY](#)

[2 replicas MISSING](#)

[54 replicas DAMAGED](#)

[22097 replicas INCOMPLETE](#)

[273723 replicas OK](#)

MWT2_UC_PRODDISK... 9506 replicas, 8552 complete

[603 replicas UNKNOWN](#)

[576 replicas EMPTY](#)

[105 replicas MISSING](#)

[437 replicas DAMAGED](#)

[441 replicas INCOMPLETE](#)

[7344 replicas OK](#)

MWT2_UC_PHYS-TOP... 1526 replicas, 1526 complete

[1526 replicas OK](#)

MWT2_DATADISK... 10665 replicas, 10617 complete

[69 replicas UNKNOWN](#)

[16 replicas EMPTY](#)

[1320 replicas MISSING](#)

[4 replicas DAMAGED](#)

[31 replicas INCOMPLETE](#)

[9225 replicas OK](#)

[804488 dq2 orphans \(67.3T\) \(dark data\)](#)

Finished at Fri Nov 21 01:01:54 2014

Status and goals

- Goal is to integrate closely with Rucio tools and distribute via CVMFS
 - Use daily dumps from central catalogs to perform efficient checks
 - Hope to extend to provide local popularity & # replicas
 - - useful for site deployment (hot/cold data)
 - Input, volunteers welcome! (rwg@uchicago.edu)
-