Journal of Physics: Conference Series **219** (2010) 042053 doi:10.1088/1742-6596/219/4/042053

Data Management tools and operational procedures in ATLAS : Example of the German cloud

Cédric Serfon

Department für Physik, LMU München Am Coulombwall 1, D-85748 GARCHING, GERMANY

E-mail: Cedric.Serfon@physik.uni-muenchen.de

Abstract. A set of tools have been developed to ensure the Data Management operations (deletion, consistency checks) within the German cloud for ATLAS. These tools are described hereafter and presented in the context of the operational procedures of the German cloud. A particular emphasis is put on the consistency checks between the different catalogs (LFC, DQ2 Central Catalogs) and the files stored on the Storage Element. These consistency checks are crucial to be sure that all the data stored in the sites are actually available for the users and to get rid of non registered files also known as Dark Data.

1. Introduction

Distributed Data Management (DDM) is one of the key component in ATLAS that is used by many other domains (Production system, user analysis...). In ATLAS, DDM is performed through the use of a software called DQ2 [1] that interacts with different services and catalogs. DQ2 is composed of many differents parts :

- Data movement service : Export and register data between sites.
- Staging service : Stage data on tapes.
- Deletion service : Perform deletion consistently on the Storage Elements and the various catalogs, based on deletions requests.
- Accounting service : Monitor the evolution of disk space versus time and different metadata (group, user, pattern).

In order to provide functionalities that were missing in DQ2, tools have been written, tested and are now used in the German Cloud. Some of these tools are now integrated in DQ2. All of them have some commonalities :

- Written in python.
- Object oriented.
- Logging of outputs.
- Retries procedures.

In chapter 2 is presented the local deletion module that has been added to the deletion service. In chapter 3 is introduced the problem of consistency ; the tools developed in parallel with the procedures applied in the German Cloud [2] are then exposed.

2. Local deletion tool

As already mentioned, the deletion service deletes files both from the Storage Element and from the Catalogs. Deletion from Storage Element is made using the SRM[1] interface (SRMs are Grid storage services providing interfaces to storage resources). This access via a layer above the Storage System has unfortunately some drawbacks :

- The deletion rate is limited to O(1 Hz)
- It puts some load on the SRM interface which is also used for data transfers.

To get rid of these issues, a local deletion has been implemented for most of the Storage Element flavours (dCache, Castor) and is fully integrated into the deletion service. Switch between remote and local deletion is done easily via a configuration file. Local deletion is now used regularly to delete data at CERN and at GridKa (the German Tier 1). The performance of the deletion service reach then up-to 30 Hz when no activity happens (see Figure 1) and O(15 Hz) when there is a heavy load due to data transfers. This approach of local deletion also allows directories deletion which is not implemented yet in the remote deletion.

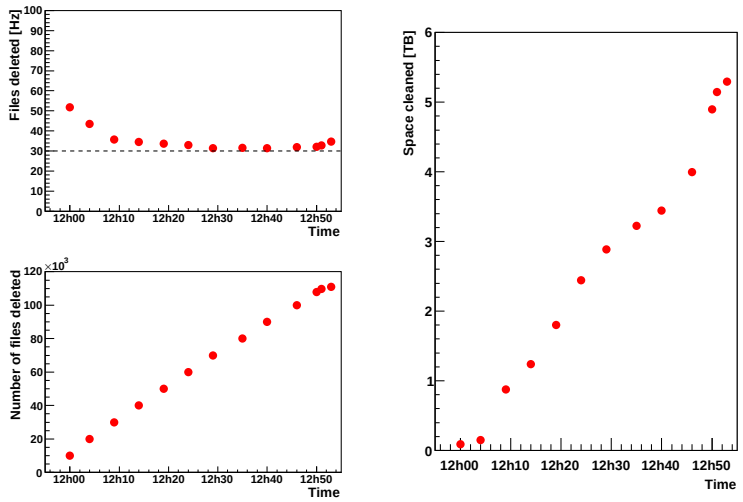

Figure 1. Deletion of a batch of ∼115 000 files representing 5.5 TB at GridKa using local deletion. Deletion rate versus time (left upper plot) and deletion rate integrated over time (left bottom plot). Space cleaned versus time (right plot).

3. Consistency tools

3.1. Sources of inconsistencies

Distributed Data Management, in ATLAS, relies on 2 different catalogs :

- LCG File Catalog (or LFC) : There are 17 different LFCs (1 at CERN, 10 on the Tier1s, and 6 at US Tier2s).
- DQ2 Central Catalogs : Located at CERN. There are many different catalogs, one of them which is called the location catalog lists the datasets names, their replicas, etc.

[1] Storage Resource Manager

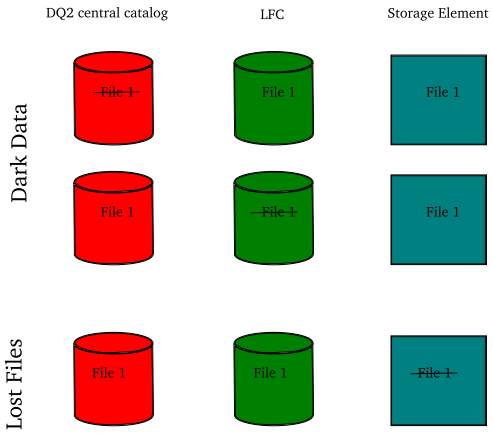

Figure 2. Description of the various sources of inconsistencies. In the first 2 cases, the files are on the Storage Element (SE) but cannot be found on one of the catalogs (Dark Data that waste disk space since they cannot be used by analysis tools). In the last case, the files are in the catalogs but cannot be found on the Storage Element (Lost Files which make users' jobs crash).

These catalogs in particular list the files that are available on the Storage Elements. Keeping the consistency between catalogs, that are the primary source of information for the grid tools and for users, and the Storage Element is therefore crucial. Unfortunately, problems can occur and break this consistency. Figure 2 describes the various types of inconsistencies that have been observed.

3.2. Consistency between SE and LFC

The first level of consistency (SE↔LFC) is ensured by using Storage Element dumps provided regularly by sites. The format of these dumps has been standardized for various Storage Elements flavours according to the format developed for dCache[3]. These dumps are downloaded by an agent that compares them to the list of files recorded in the LFC. 8 000 000 entries can be checked in less than 2 hours ; this time is actually dominated by the query of the LFC (the actual comparison only lasts about 30 minutes). For sites that have never been checked before the amount of files on the Storage Element but not registered in the LFC can reach up-to 3%. These files come mostly from failed transfers attempts.

In the German cloud, every site is asked to provide the 7^{th} day of each month a dump that they upload into a dedicated place. 2 or 3 days after the consistency check is run on this dump. A list of files to delete is then produced and the files are deleted remotely if their number does not exceed O(1 000). If a higher number of files is found, the list is sent to the site for local deletion and investigation is performed to understand where these errors come from.

3.3. Consistency between LFC and DQ2

This second level of consistency is checked by a direct comparison of the files registered in DQ2 Catalogs and files registered in the LFC. This check does not need any action from the sites and can be done remotely. Various sources have been identified that can produce such kind of inconsistencies :

- Use of lcg commands outside of DQ2. This is unfortunately unavoidable in the user area, even if users are strongly encouraged to register all their files in DQ2.

- Bad use of DQ2 commands (dataset definition removed without cleaning-up of Storage Element + LFC).
- Problem due to the deletion procedure of transient datasets from the Production System where some files were not attached to the final datasets. After identifying the problem, the procedure to get rid of these transient datasets has been changed and the problem is fixed.

This tool induces many LFC queries and therefore needs to be run close to the LFC server to reduce the Round Trip Time. The speed also highly depends on the number of datasets and files in the site. Once the files not registered in DQ2 are identified, they are attached to a "trash dataset" that is then declared to the deletion service that automatically clean the data from the Storage Element and from the LFC.

In the German cloud, this tool is run once or twice a month. Figure 3 shows the effect of running the tool on one site (FZK-LCG2_MCDISK).

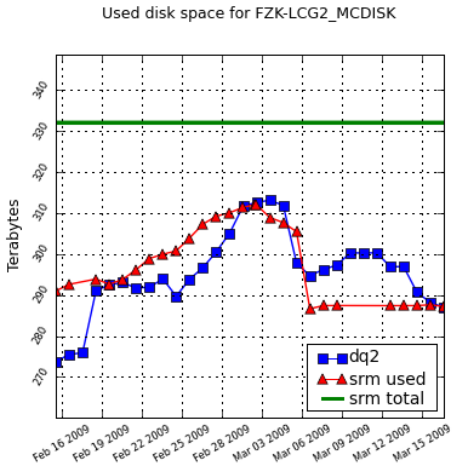

Figure 3. Evolution of the Space seen on the site FZK-LCG2_MCDISK versus time : in red information from the Storage System provided by SRM and blue information from the DQ2 location catalog. The effect of consistency checks (20th February, 1st March, 16th March) can be seen.

4. Conclusion

A set of tools have been developed to identify the inconsistencies between the Storage Element and the catalogs used in DDM. These tools are regularly used in the German cloud and start to be used in other clouds (France, CERN). They allow to keep the consistency that is needed for a smooth use of the Grid.

In addition, the DQ2 deletion service has been extended to run local deletion that allows to delete efficiently and quickly (about one order of magnitude faster than SRM) the files from catalogs and Storage Elements.

References

[1] M. Branco *et al.* Managing ATLAS data on a petabyte-scale with DQ2, J. Phys. Conf. Ser. **119**, 062017 (2008).
[2] J.Kennedy *et al.* ATLAS operation in the GridKa Tier1/Tier2 cloud, Proceedings CHEP 2009
[3] P. Millar Dealing with orphans: catalogue synchronisation with SynCat, Proceedings CHEP 2009