ATLAS Logical File Name and Directory Paths convention

eris, M.Branco, D.Cameron, S.Campana, A.Klimentov, D.Liko, P.Nevski, T.Wenaus

ATL-COM-SOFT-2006-017 December 1st, 2006
revised Feb 12, 2009

ATL-SOFT-PUB-2006-010
06 May 2009

Logical File Name (LFN)

The following convention is used to form local files entries in file catalogs for data processed or simulated using ATLAS Production System. ATLAS online files naming convention isn't covered by this document.

Logical File Name (LFN) will have 2 parts separated by '._' (dot-underscore) symbols. The first part of LFN corresponds to the DDM dataset name, as it is described in [1]. The first part contains two fields separated by dot, data type (dataset field 6, f.e. 'AOD') and production task unique id used to produce file f.e 022009.), the second part is the file name. In this case the catalog entry looks like :
datasetType.TaskID._fileName.
LFN is defined during the first file registration in the catalog.
LFN is not modified if the file is added to more than one dataset.
Data reprocessing and MC production follows this convention, and production LFNs have format :
AOD.022009._00293.pool.root.1

Output Directory Path

Since year 2008 ATLAS uses *LFC* (Local File Catalog) for all Grid flavours. LFC has a hierarchical namespace with a directory structure. The physical storage at sites also has a hierarchical structure, which means that for this and LFC we need to provide a convention on directory structure for physical and logical file names created by DDM.
On LFC the output directory path is always preceeded by the path */grid/atlas/dq2*.
On storages the output directory path is always preceeded by the path for the storage defined in Tiers of ATLAS.

Current Conventions

All files (MC, DAQ, etc) are registered in LFC and stored by DDM system under the directory tree
/project/datasetType/datasetName/fileName, where

- *project* - name of project (f.e. *csc11, mc12, larg,...*), the first field in dataset name,
- *datasetType* - dataset type (f.e. *RDO, AOD, RAW, ...*), the 5th field in dataset name,
- *datasetName* - DDM dataset name. In the case of a catalog registration following a subscription, the dataset name is the name of the dataset that was subscribed that led to this file being replicated,
- *fileName* - For LFC: the DDM "flat" LFN. For storage: the physical file name of the original source file.

Example

If we take the example LFN above *LFN : AOD.022009._00293.pool.root.1*
in the dataset *mc08.007447.singlepart_mupt50.recon.AOD.e315_s422_r408_tid022009*
the output directory path becomes
/mc08/AOD/mc08.007447.singlepart_mupt50.recon.AOD.e315_s422_r408_tid022009/AOD.022009._00293.pool.root.1

Datasets outside convention

Any dataset with 5 or more dots in the name is assumed to follow the convention described above, where *project* is the first field and *dataset type* is the fifth field.

Datasets with between 1 and 4 dots inclusive will have the following output directory path:
/project/datasetName/filename
where these terms have the same meaning as above.

Datasets with no dots will have the following output directory path:
/other/datasetname/filename
where these terms have the same meaning as above.

Examples

A file with the physical and logical name *joetestfile.1* in a dataset named *user.joeuser.datasettest1* has the output directory path

/user/user.joeuser.datasettest1/joetestfile.1
A file with the physical and logical name *file1* in a dataset named *mydataset1* has the output directory path
/other/mydataset1/file1

Proposed Conventions

Increasing granularity for Tape Families

In order to increase the granularity for tape families at some sites, it has been proposed to extend the current convention to include an extra level of hierarchy in the directory structure. The proposal is to modify the current convention above and register in LFC and store via DDM system under the following directory tree:

/project/datasetType/physicStream/datasetName/fileName, where

- *project* - name of project (f.e. *csc11, mc12, larg,...*), the first field in dataset name,
- *datasetType* - dataset type (f.e. *RDO, AOD, RAW, ...*), the 5th field in dataset name,
- *physicStream* - the physics stream (f.e. *physics_IDCosmic, calibration_LArCells ...*), the 3rd field in dataset name,
- *datasetName* - DDM dataset name. In the case of a catalog registration following a subscription, the dataset name is the name of the dataset that was subscribed that led to this file being replicated,
- *fileName* - For LFC: the DDM “flat” LFN. For storage: the physical file name of the original source file.

concerns : it may be useful for ATLAS RAW data staging from tape, if it will be implemented by the majority of Tier-1s. For the moment only one Tier-1 is willing to implement such tape family. At the same time, the second field of MC dataset name is the longest field, there are up to 100 characters, and using it as directory and LFC path will increase complexity of using any CLI commands.

Increasing granularity

In order to increase the granularity and to reduce number of subdirectories associated with project and data type (now we have up to 5000 subdirectories), it has been proposed to extend the current convention to include an extra level of hierarchy in the directory structure. The proposal is to modify the current convention above and register in LFC and store via DDM system under the following directory tree:

/project/datasetType/DSN/RunNumber/datasetName/fileName, where

- *project* - name of project (f.e. *csc11, mc12, larg,...*), the first field in dataset name,
- *datasetType* - dataset type (f.e. *RDO, AOD, RAW, ...*), the 5th field in dataset name,
- *DSN or RunNumber* - dataset number or run number (f.e. *007447*, the 2nd field in dataset name,
- *datasetName* - DDM dataset name. In the case of a catalog registration following a subscription, the dataset name is the name of the dataset that was subscribed that led to this file being replicated,
- *fileName* - For LFC: the DDM "flat" LFN. For storage: the physical file name of the original source file.

concerns : this convention won't help to group files on tapes, but it will decrease number of sub-directories and it may be useful in the future when number of datasets per (project)/(data type) will increase

Group and User Datasets

To place datasets used by different ATLAS groups or/and users the following convention is proposed for *group* and *user* datasets.
/project/(Group or User)info/datasetName/fileName, where

- *project* - name of project (f.e. *group08, user08, user,...*), the first field in dataset name,
- *(user or group)info* - the second field in dataset name (f.e. *JohnBakken* or *EGamma*)
- *datasetName* - DDM dataset name.
- *fileName* - For LFC: the DDM "flat" LFN. For storage: the physical file name of the original source file.

Transient (_sub,_dis) datasets

Directory and LFC path of datasets with limited lifetime (f.e. PanDA *_dis and _sub* datasets) is formed differently. Only basic part of dataset name (everthing before _dis or _sub subfields) is used to form directory and LFC path.

F.e. file *AOD.022009._00293.pool.root.1*
from dataset
/mc08/AOD/mc08.007447.singlepart_mupt50.recon.AOD.e315_s422_r408_tid022009_sub0123455
will be stored in directory
/mc08/AOD/mc08.007447.singlepart_mupt50.recon.AOD.e315_s422_r408_tid022009/
and _subXX part will be also removed from LFC path

Time scale

The proposed new convention will be used starting from Mar 1st 2009.

References

[1] "ATLAS Dataset Definition". ATLAS SWING internal note 8 Mar 2006;