HDFS Data Deletion Strategy

Older ingestion/processed dataset needs to be deleted from HDFS periodically.

Also need a mechanism where its easy to add/remove new directories, configure the days of data to keep in each directory and also see when the last deletion happened and which all directories were deleted preferably via UI.

The proposed solution is to create a configurable deletion script(details below) and schedule it in Airflow so that its each to change and track the scheduled runs of the script.

Deletion script

Deletion script should take a list of directories to be deleted along with directory date pattern and days to keep. The list of directories can be passed as command line parameter or a file with list of directories. The script should support two modes of operation one being the date pattern based deletion which is mentioned above and the other being modification time where it would check the modified time of files within the directory and remove it if older than configured days to keep. In this mode the directories won't be deleted. Also there is a bit of unpredictability with this approach as if one file within a directory was modified by renaming or replacing then that file alone will remain in that directory and others deleted.

Date pattern

deletion_script.py --dir <dir1> --dir <dir2> --pattern 'date=YYYY-MM-DD' --days 30 --mode pattern

or

deletion_script.py --dirfile <directory_file> --pattern 'date=YYYY-MM-DD' --days 30 --mode pattern

Eg: *deletion_script.py --dir /data/fun2/stage/ingestion/version=*/ --pattern 'date=YYYY-MM-DD' --days 15 --mode date_pattern*

The deletion script reads all directories immediately under /data/fun2/stage/ingestion/version=*/ and then extracts the 'date=2020-05-02' part from it and compares it against current date and if greater than 15 days will remove it.

Modification time

deletion_script.py --dir <dir1> --dir <dir2> --days 30 --mode modification_time

or

deletion_script.py --dirfile <directory_file> --days 30 --mode modification_time

For deleting different paths with different days to keep execute script multiple times with different set of directories and days to keep in Airflow.

Auditing

The deleted directories needs to be printed via stdout of the script and Airflow configured to log it so it would serve as an audit log.

Log line format

<date which hopefully airflow will add> [HDFS_DELETION] [LOG_LEVEL] <message: directory name or error name>

So we can search for HDFS_DELETION in logs to get the required lines.