archiving strategy

This is another item we will need very soon. Navid has made a perl interface to the archive system. But now the issue is how and when to archive.

Since the I&T pipelines are parallel, they have several different named tasks operating on the same run, writing to the same directory. Additionally, not all files are reported to the pipeline, but they are wanted to be archived. So we have to archive the entire directory. Ideally we would prevail on everyone to identify every file they want archived, but we seem to be on the losing end of that one!

I think this means the archiving has to happen asynchronously to the pipeline. I'd be curious to see a comment to this blog item from Dan with his thoughts on the algorithm for figuring out what to archive and when.

Note that SCS asks that we keep the files larger than 500 MB to use the tapes efficiently. So we had been thinking to make tar files. Navid keeps track in his archiver db of the file content inside the tar file, so he can ask the archive system for the right tar file when someone asks for an individual file.

I'm at a bit of a loss at the moment to divine a way to know when one can archive in a general way. It would be nice not to need custom archiving per group of tasks.