ARCHIVER Project

Technical Summary

EMBL CLOUD CACHING

Problem Definition:
As we expand our computational activities into the cloud, we expect that it won't be cost-effective to put all our data into the cloud immediately. Our data volume is doubling every two years, so half our data is less than two years old, and that data is likely to be actively used much of the time. Even the older data has a long tail of access, so very little of our data is 'never' used anymore.

We want to be able to start using our more popular data from the cloud, while not paying to store data that's not currently used as much. This is a classic caching problem, but a simple 'least recently used' algorithm could be far from optimal.

We want the cache size to be optimised as a function of cost, tracked over time, i.e. to deduce which data is cost-effective to cache at a given time, and size the cache accordingly, with cost as input and some measure of efficiency as output.

Cache lifetime and size should be variable, driven by usage patterns. For example, the quarterly release of new versions of ENSEMBL, one of our big data sources, results in a lot of downstream processing. At that time, it may be necessary to increase the cache size and/or the lifetime of certain files in the cache at the expense of other files.

As the computing activity in the cloud increases, some of our data will be in constant use. This data should then be 'cached' forever.

We can provide hints to the caching algorithm about the expected use of our data. A typical example is the new release of a version of a given dataset, such as a reference genome, which means that its older version is likely to decline in use. Such hints may be able to improve the efficiency.

Lifecycle - Workflow Characteristics:

Our users come from all over Europe, and beyond. Currently we receive about 2 billion download requests per month, for 1 PB of data. The download pattern is relatively flat throughout the day, there are no peak-hours as such. Users needing large amounts of data will be using scripted tools, such as anonymous FTP, or ASPERA, to get their data. Other users download data through web interfaces. In both cases, users typically use web portals to discover the data they need, based on domain-specific metadata.

Authentication and Management Functions:
For data retrieval, most of our data is publicly available, no authentication required. Some of our data is restricted-access, so users are authenticated and given access to separately encrypted copies. The encryption process, and key management, is outside the scope of ARCHIVER, however the cache will be required to authenticate users and check their authorisation for certain data. We manage authentication and authorisation using standard protocols (OAUTH, SAML)

Data and Metadata Characteristics:
Data consists of files from several kB up to 500 GB, with the vast majority in the range of 1 to a few GB. The current archive is 20 PB in size, and is currently doubling every two years. We expect that growth rate to continue for at least the next decade.

Domain-specific metadata will be managed as today, in specific portals that we will continue to own and control. File-specific metadata (such as creation-time, access-time, checksum, size etc.) can be held and used by the cache. As noted above, some of our data is restricted access, using standard RBAC techniques.

One interesting aspect of this use case will be to explore how we can provide domain-specific metadata to a caching algorithm in a way that it can use to extract meaningful relationships between files, such as identifying related files that should be cached or purged together. If successful, this could even be used to pre-fetch data for the cache to improve performance.

For example, noting that certain files were produced by the same research project may be valuable. The challenge is to extract useful relationships *without* having to fully understand the domain knowledge.

Interface Characteristics:
As a cache, there will be essentially no interface for users. However, where workflows use data, or create data that is then added to the EBI data store, they could be instrumented to give cache hints. E.g. marking files as single-use (therefore not worth caching), or as needed for downstream processing in the near future (therefore worth keeping in the cache).

Such hints could either be given automatically (by instrumenting workflows), or perhaps deduced automatically, from process-mining and analysis of workflow structure.

Reliability Requirements:
Reliability of the infrastructure is not as critical as for a normal archive. The impact to the user of the cache going down should be limited to falling back to the main data store. This should preferably be transparent to the user.

Compliance and Verification:
Data is not manipulated or owned by the cache, so there are no compliance and verification concerns for this use case.

Cost Requirements:
The goal is to deduce the size of the cache as a function of budget, so we can set budgets as we wish over time and have the cache size itself and operate itself for maximal efficiency.

We should be able to know the efficiency we are achieving, so we can decide to increase or decrease the budget accordingly. The exact definition of 'efficiency' here is open for discussion, though some form of 'effective size' that can be costed as if it were real disk would seem to be appropriate.

Initial Data Management Plan:

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	The data to be used is primarily DNA sequence data from the European Nucleotide Archive (ENA), located in the EMBL-EBI data centres. Data is submitted to ENA by research teams around Europe.
Documentation and data quality	Data quality is measured as part of the DNA sequencing process, the quality measurement is included with the raw data. The origin of the data (organism, sample etc) forms part of the accompanying metadata collected during the submission process.
Storage and backup during the research process	Data is stored in the FIRE (FIle REplication) archive, consisting of one copy on a distributed object-store and one on a tape archive, all hosted on EMBL-EBI data centres
Legal and ethical requirements, codes of conduct	Much of the data in ENA is freely available, by anonymous FTP. Some data is tightly controlled, since it contains highly sensitive personal information - e.g. cancer tumour

	DNA sequences that can be traced to an individual. However, for the ARCHIVER project, we will only use the freely available data
Data sharing and long-term preservation	This use-case is about caching data in the cloud, there is no long-term preservation issue. Similarly, being a cache, the data will be transparently and freely available, given that the source data will be the freely available subset of ENA.
Data management responsibilities and resources	Management of the original copy of the data will remain the responsibility of EMBL-EBI. We make redundant copies of the data on different storage technologies (tape and object store) to minimise risk of loss.