HEIDS Research Data Management Seminar: Active Data Services and Storage Infrastructure

Tony Weir

Information Services

UoE Context, some history ...

- History of dealing with big data:
  - EDINA National Data Centre
  - Scottish Digital Library Consortium
  - Central unstructured file-stores
  - Central large research infrastructure

UoE Context, more recently ...

- Very rapid growth in unstructured data
  - Instruments : gene sequencers, medical imaging, climate data, LHC
  - Simulation : computational output (Physics, Engineering, Chemistry)
- Rise of digital humanities …
- Individual projects with very large data needs (200TB, 0.5PB, 10PB …)
- Majority of data sets are relatively small and held in common formats (spreadsheets)

UoE planned provision

- “baseline 0.5TB per researcher"
- principle that no researcher should be inhibited by the lack of storage space
- the university recognises the value of its data and that it should be held securely

UoE – direction ...

Data storage working group identified service requirements for:

- “Archiving of research data”

- “Globally accessible cross-platform file store” “ … facilitating extra-institutional collaboration”

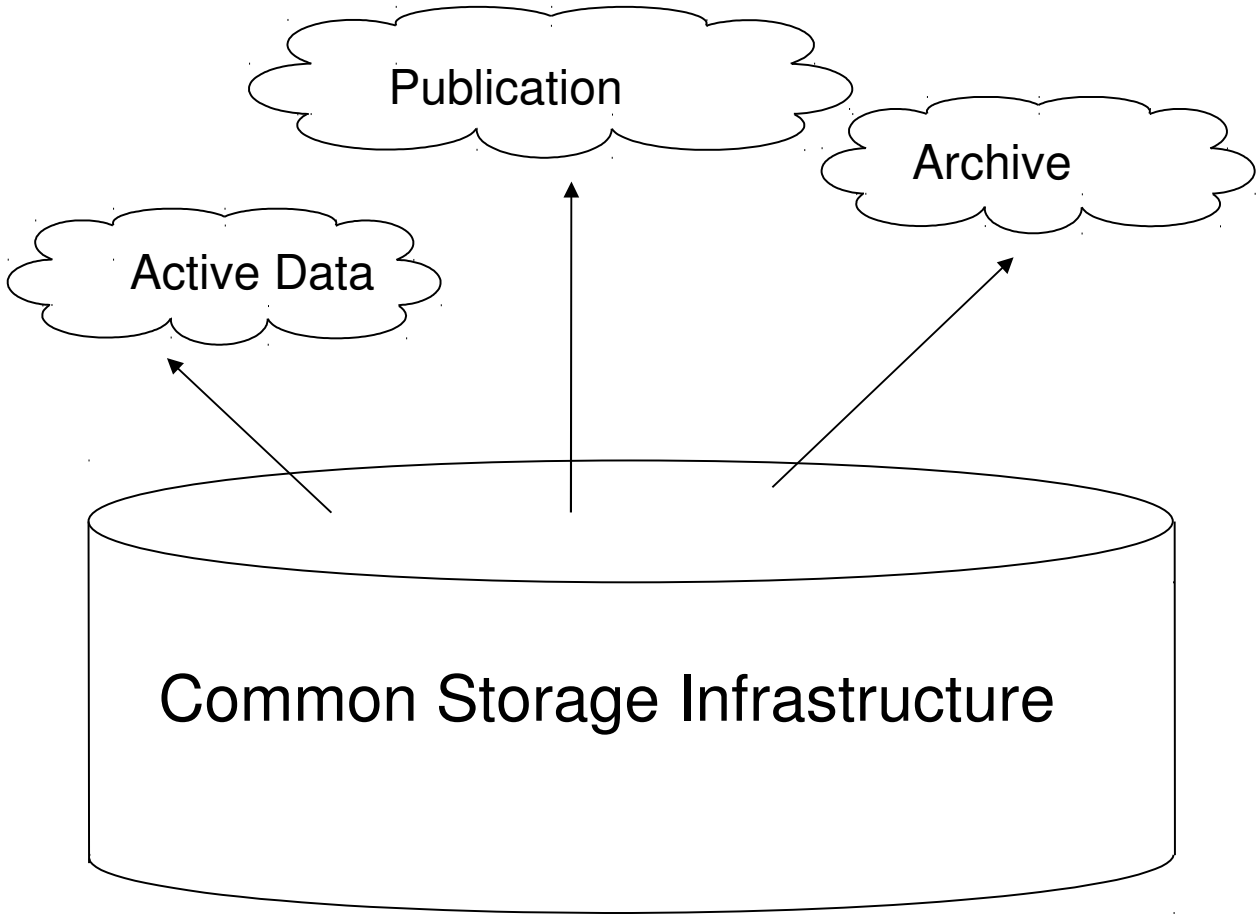
- Wider RDM storage requirements for data publication

Common data requirements ...

Across all data services there are common requirements:

- Data must be held securely
- Have to handle data at scale
- Expansion must be easy
- Highly available
- Flexible presentation
- Timely recovery
- … all at appropriate cost

Common Storage Infrastructure

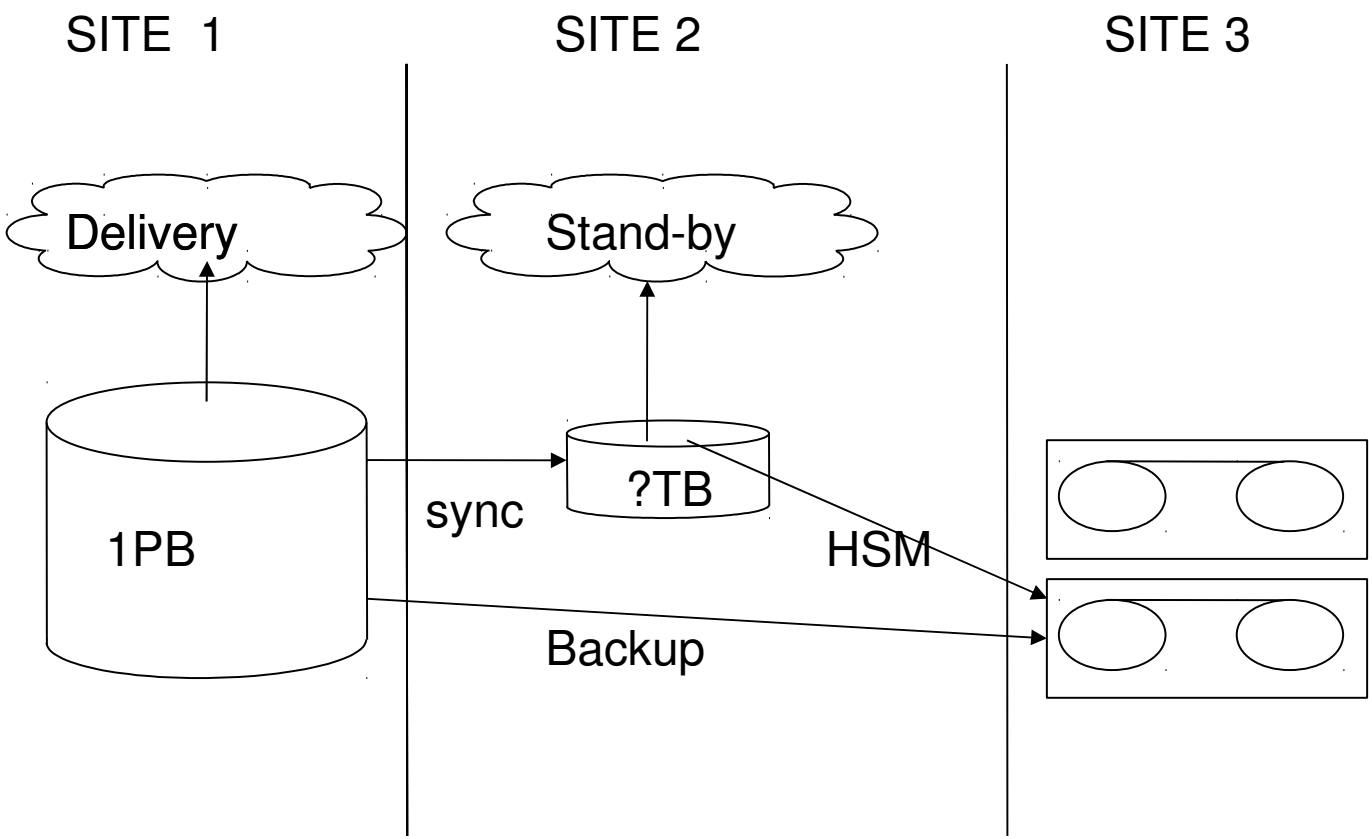


Storage technologies ...

IBM GPFS:

- Lots of experience gained from HPC – bringing this to general file stores
- Scale-out (presentation and storage)
- Single name space (if we want)
- Rich policy engine
- Integrates with HSM / supports integration of multiple storage tiers

File-store



File-store

- provides working file-space
- implements 0.5TB per researcher
- many access mechanisms:
  - Commonly present the same file – (CIFS, NFS, sshfs)
  - May present the file unique to that presentation (e.g. WebDAV)
- daily file versioning
- technologies: samba, linux nfs, ctdb, sshfs (expandrive)

Archive – future ...

- Data Vault (archive):
  - Requirements less certain
  - Golden copy data – safe – could this be mirror of publication ?
  - Consider work-flow from live to publication and/or vault

Dropbox-like

- Synchronisation of mobile devices
- Requirements :
  - Secure access – concerns raised on commercial services
  - Easy for collaboration – like dropbox
  - Good integration with all clients – like dropbox
  - Sharing seems more important that off-line syncronisation
- Pathfinder project: owncloud