



Data Management

- ▶ **What it is about?**
- ▶ **Status**
- ▶ **Current work**
- ▶ **Next developments**



- ▶ Definition of available resources:
 - ▶ StorageElements, FileCatalogs, FileTransferService,...
- ▶ Tools to:
 - ▶ Upload your data to SEs and register in in FCs.
 - ▶ Retrieve your data from SEs and FCs.
 - ▶ **Resources/Storage**
 - ▶ **Resources/Catalog**
 - ▶ **DataManagementSystem**
- ▶ Mechanisms for:
 - ▶ Making data available for Payloads
 - ▶ Managers to replicate, remove, archive data
 - ▶ **RequestsManagementSystem**
 - ▶ **StorageManagementSystem**



Update in Resource Definition

- ▶ **StorageElements:**
 - ▶ Will be defined as part of a given Site (like CEs).
 - ▶ Can be declared Close to other Sites for Job Scheduling.
 - ▶ Capabilities will be included in the definition:
 - ▶ Capacity
 - ▶ Disk/Tape
 - ▶ StageCache
 - ▶ State will be kept by Resource Status System:
 - ▶ Read, Write, Remove, Stage
- ▶ **Catalogs, FTS servers**
 - ▶ Will be moved to a new Services top section in Configuration
 - ▶ Need to find the way defined which catalog is used in different contexts



- ▶ Basically OK
- ▶ Need to be adapted to new Configuration Schema
- ▶ The required functionality is covered by:
 - ▶ `DataManagementSystem/Client/ReplicaManager.py`
 - ▶ `Interfaces/API/Dirac.py`
 - ▶ Review of duplicated functionality/code
- ▶ Command line tools unification
- ▶ New commands to define/display/update definitions of Resources.



- ▶ **Currently relying on other components**
 - ▶ Massive replication implies:
 - ▶ Transformation + Requests + Agents
 - ▶ Indirection
 - ▶ Complicates Monitoring
- ▶ **Provide a Real Data Management functionality (like Workload Management):**
 - ▶ Tasks will be declared/submitted/executed/monitored
 - ▶ Started with Staging, next one Replication/Removal
 - ▶ Aim is to improve traceability, monitoring, reliability



Other work in progress

- ▶ **DIRAC File Catalog optimization**
 - ▶ Under request from ILC
 - ▶ Need a proper comparison with LFC + metadata FC
 - ▶ ILC: Metadata Query → 10k Replicas in 15 s.
- ▶ **Storage Element Usage Accounting**
 - ▶ Under request from LHCb
 - ▶ Basic framework can be extended with community dependent metadata information.
- ▶ **Consistency checks**
 - ▶ Under request from LHCb



In the pipeline and ...

- ▶ **Concurrent download from a single SE**
 - ▶ Can cause destructive interference
 - ▶ Means to limit (not provided by SRM)
- ▶ **Interfaces to File Catalogs**
 - ▶ Web portal
- ▶ **Dynamic Data replication based on data popularity**
 - ▶ Requirement from LHCb to improve access to hot data
 - ▶ Based on improved advanced Tools



- ▶ Data Management is evolving from a collection of useful tools to a real management system.
- ▶ As in the past with WMS, LHCb is the driving force (providing the manpower).
- ▶ The integrated Replica and MetaData DIRAC File Catalog is a powerful tool that will keep improving.
- ▶ Ideas, contributions, requirements are welcome

