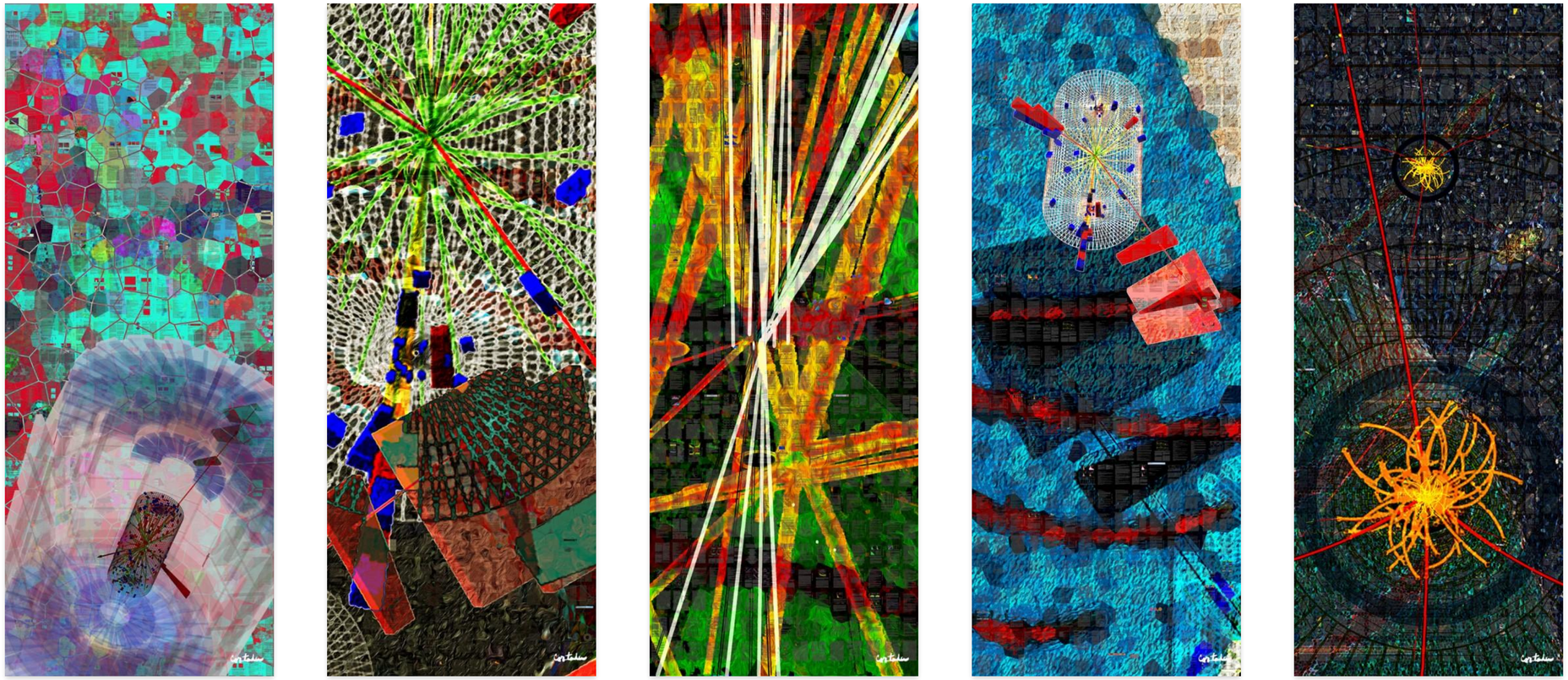
U.S. CMS Operations Program

CMS Data Management

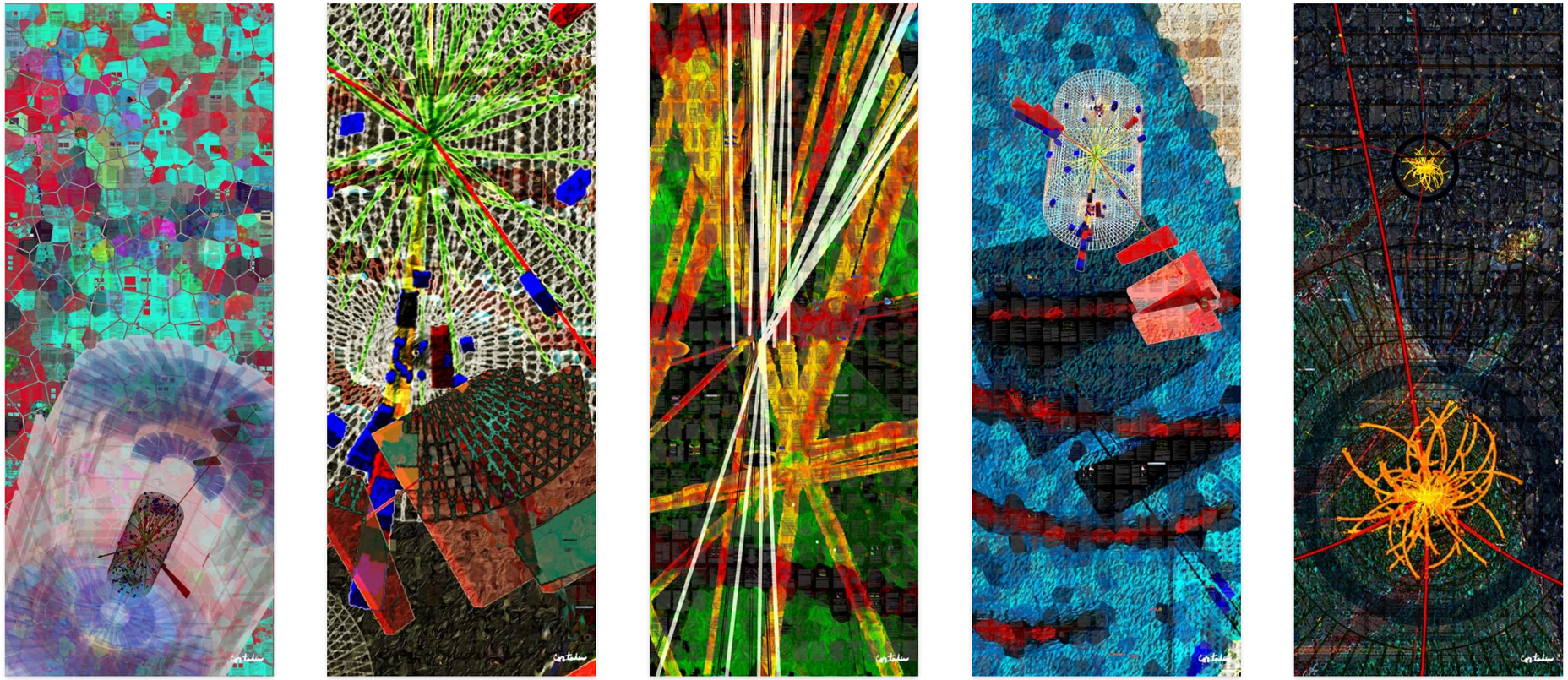
Andres Manrique (University of Wisconsin-Madison)
Jhonatan Amado (FNAL)



Agenda

1. General Overview
   a. CMS Data Management Systems
2. Content Creation
   a. WMAgent (input data placement)
   b. WMAgent (output data placement)
   c. MSRuleCleaner (close workflow)
3. Content Replication and Access
4. Content Invalidation & Deletion
   a. Invalidations
      i. Consistency check
      ii. Rucio Daemons
   b. Deletion process
5. Rubin Approach

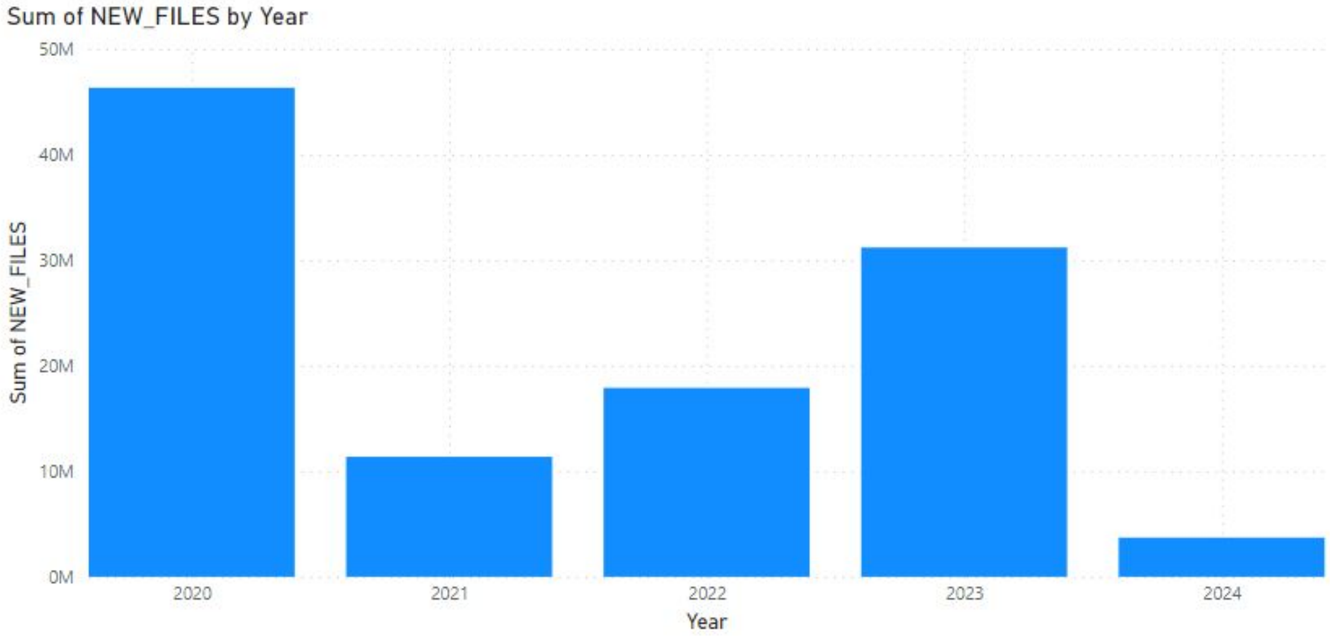
General Overview



General Overview

Data growing quickly

- 110M Registered Files
  - 29 TB of DB storage
- 649.59 PB of Replicas storage
- ~692 TB being moved daily
  - ~8GB/s average



Data Management Systems

- Organize, manage, and access
- Data Cataloging and basic metadata (adler32 and size)
- Data Replication and Distribution
- Data Access Control and Authorization
- Data Lifecycle Management

DBS

- Centralized catalog of all event data used
- Files metadata: Stores parentage, data definitions, such as, run number, algorithms, etc

Data Management Systems

SCIENTIFIC DATA MANAGEMENT

RUCIO

Rucio UI | Monitoring | Data Transfers (R2D2) | pattern OR name OR rule id | Search

You are here: Rule [edf4b78280884ea19bfbf8c39204ff10:]

Rule metadata

account	tier0_prod
activity	T0 Tape
comments	CMSTRANSF-751:HI data transfer
copies	1
created_at	Fri, 16 Feb 2024 18:07:13 UTC
did_type	CONTAINER
expires_at	never
grouping	DATASET
id	edf4b78280884ea19bfbf8c39204ff10
ignore_account_limit	false
ignore_availability	false
locked	true
locks_ok_cnt	0
locks_replicating_cnt	0
locks_stuck_cnt	0
name	/HIForward10/HIRun2023A-PromptReco-v1/AOD

Data Aggregation System (DAS): Home | Services | Keys | Bug report | Status | CLI | FAQ | Help

results format: list, 50 results/page, dbs instance prod/global, Search, Reset

/ParkingBPH1/Run2018A-20Jun2021_UL2018-v1/AOD

Show DAS keys description

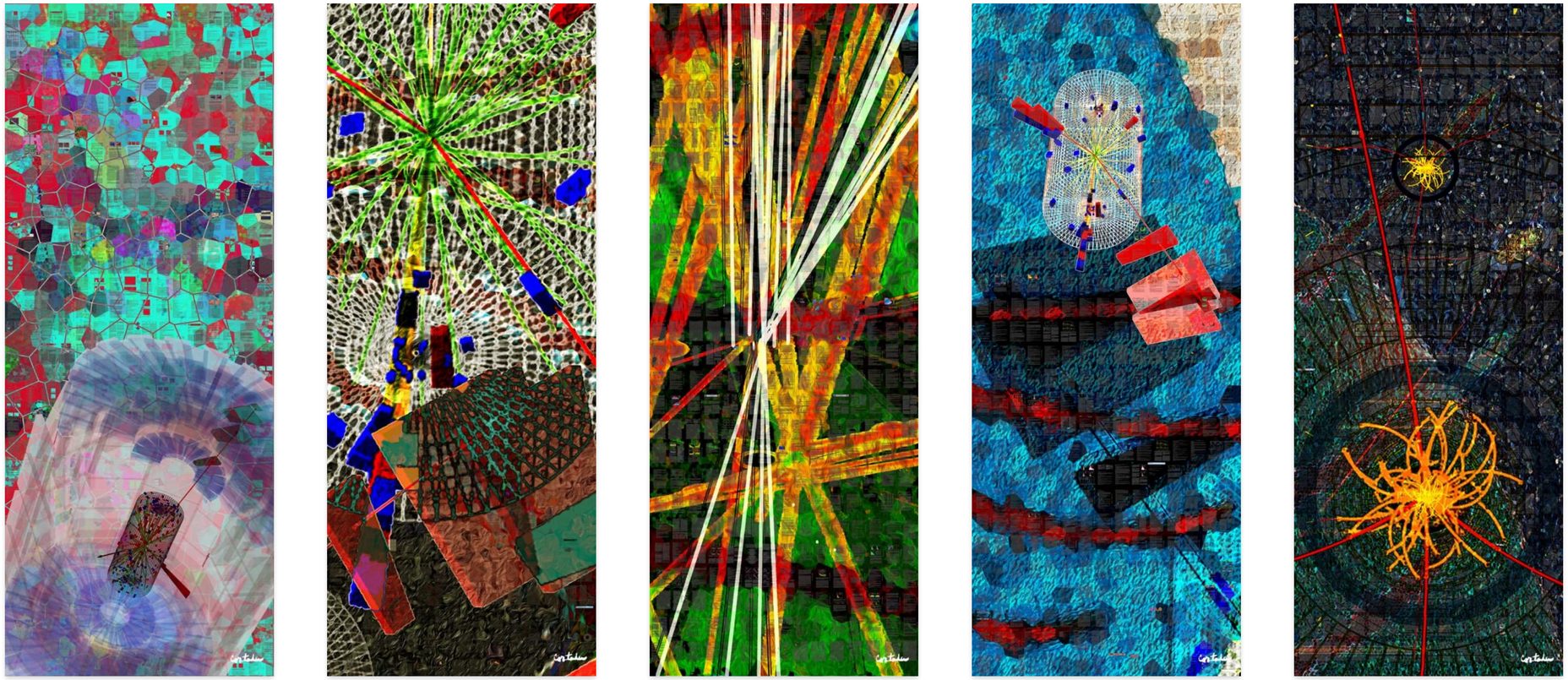
Showing 1—1 records out of 1.

Dataset: /ParkingBPH1/Run2018A-20Jun2021_UL2018-v1/AOD
Dataset size: 60726283970953 (60.7TB) Number of blocks: 89 Number of events: 205962879 Number of files: 18923 Creation time: 2021-10-16 06:18:42 Physics group: NoGroup Status: VALID Type: data
Release, Blocks, Files, Runs, Configs, Parents, Children, Sites, Physics Groups XSDB Sources: dbs3 show

Showing 1—1 records out of 1.

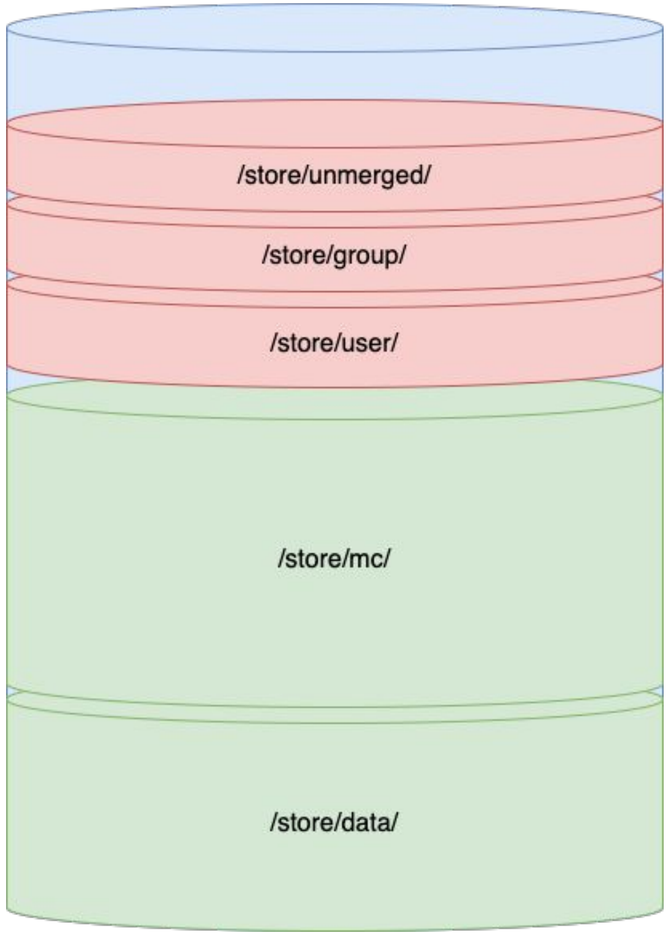
DBS

Content Creation



Rucio Storage Element (RSE) - Site

T1_US_FNAL		
	US-FNAL-CMS	Name of the WLCG federation containing the site pledge
	1.000	Fraction of WLCG federation pledge fulfilled by site
	320000	CPU pledge in HS06 (auto-calculated from Rebus and core performance)
	12.600	Average HS06 performance of a core at the site
	30000	Number of cores usable by CMS
	38300	Max number of cores used recently by CMS (auto-filled from gWMSmon)
	24000	Number of cores for production (auto-set to 80% or 50% of usable cores)
	24000	Max number of cores to be used for CPU intensive jobs
	3000	Max number of cores to be used for I/O intensive jobs
	39200.0	Disk space pledge in TBytes (auto-filled from Rebus if federation is set)
	39200.0	Disk space in TBytes usable by CMS
	36200.0	Disk space in TBytes available for experiment central operations (used by Rucio)
	92.34	Percent of disk space for experiment use (auto-calculated)
	750.0	Disk space in TBytes available for local use (used by Rucio)
	3695.00	Disk space in TBytes reserved for /store/unmerged (auto-filled from Rucio)
	126400.0	Tape space pledge in TBytes (auto-filled from Rebus)
	126400.0	Tape space in TBytes usable by CMS
	Update Information	(previous update: *2023-Aug-16 17:57:34 by dmason*)



Workflow Management Agent (input)

Campaign is announced

1. Create dataset rules to disk expr: disk,ddm>0
2. Once dataset is replicated > [95-98]%
3. WF starts at that particular site
   - What about files not on disk?
     - Rucio retries the transfer
     - AAA as a fallback (Any data, Any where, Any time) - generic implementation of xrootd
       - Access to the content of a file only knowing lfn through a proxy interface (Remote read)
4. JobsCreator (type:[Repack,Express,Production...others]) starts generating data at ../store/unmerged

Workflow Management Agent (output)

1. JobCreator
   a. Merge
      i. DBSUpload - Upload information to DBS (not injected)
         - Name must be unique (UUID)
         - Name data+ metadata
      ii. Move file from /store/unmerged to /store/TYPE/
   b. CleanUp /store/unmerged
2. RucioInjector
   a. Pull not injected data from DBS
      i. Create DIDs (Create container, dataset, files)
         - Insert file in Rucio with mandatory metadata (lfn, adler,size)
      ii. Register replica at site (Available State)
      iii. Mark file DBS as injected
      iv. Attach DIDs
   b. Protect files on creation site (create dataset level rules)
   c. Create container rule to ship data according to policies

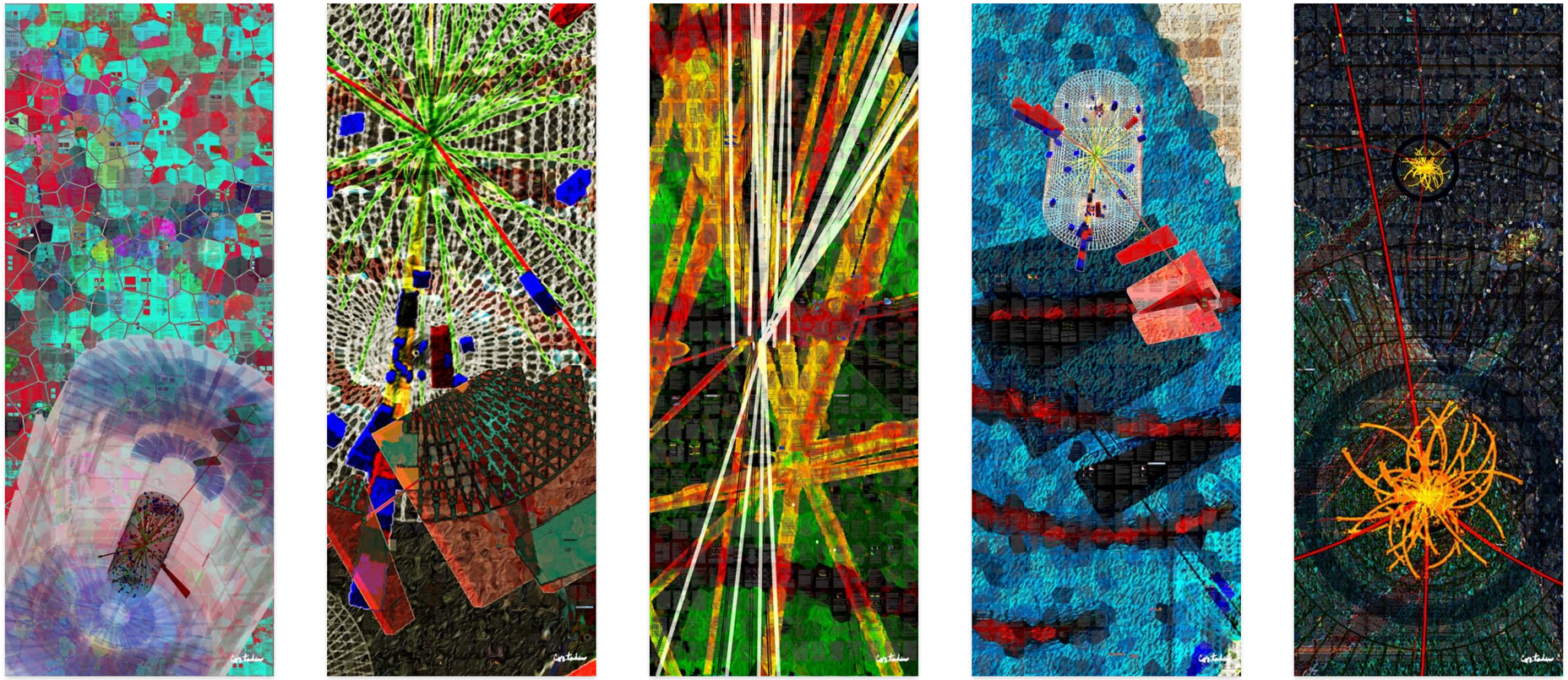
Workflow Management Agent (close)

WMCore Daemon

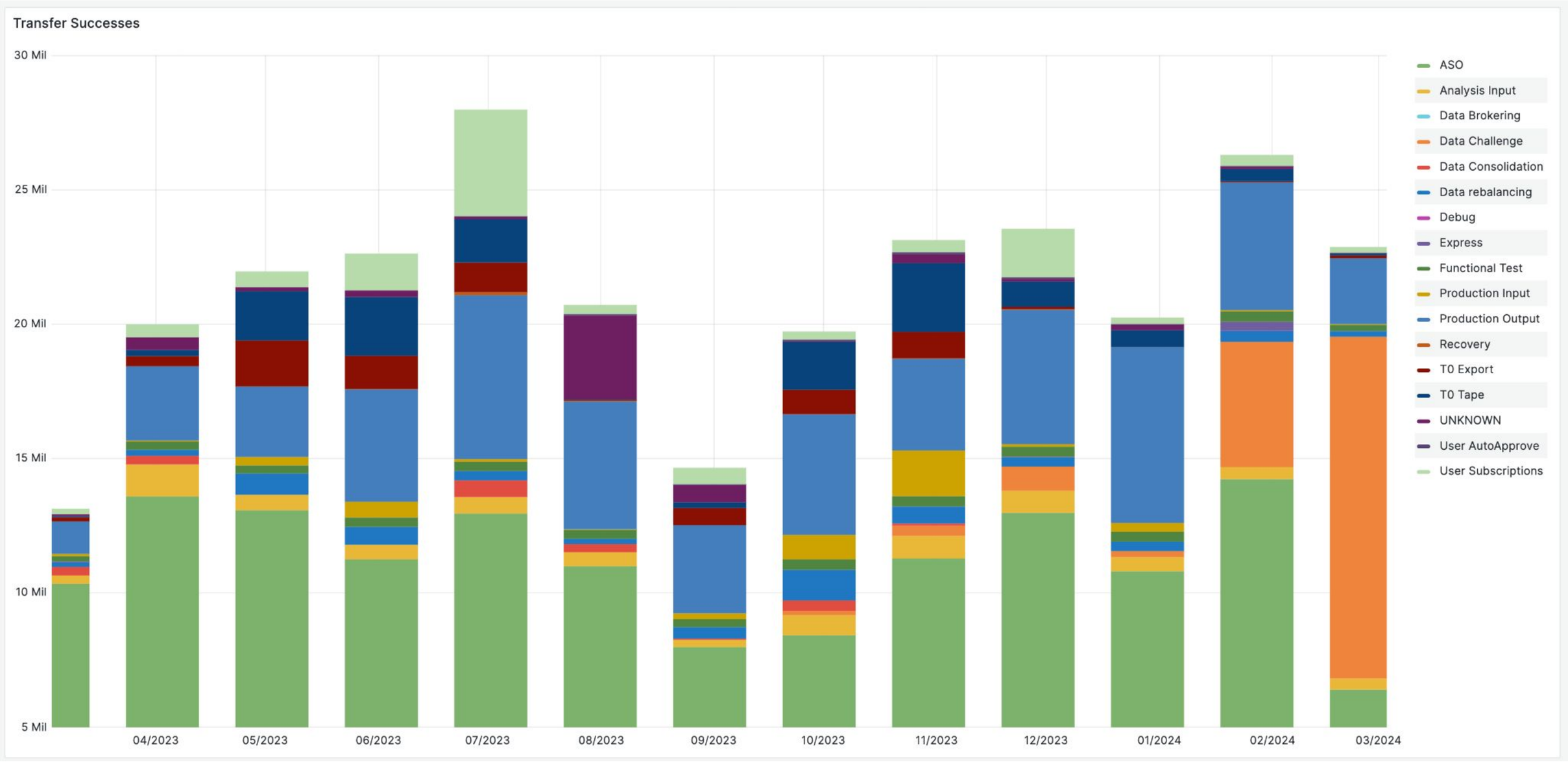
MSRuleCleaner

- Check state of all container rules for a particular wf
  - If all [ok] -> delete dataset rules from Account_A -> Mark workflow as complete
  - If not -> ping DM

Content Replication and Access



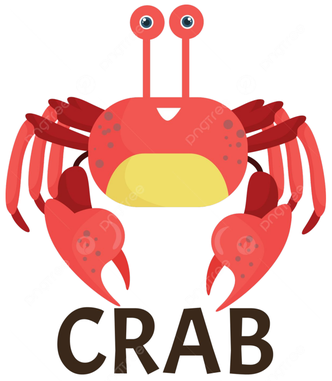
FTS Transfers



Replication and Access

Replication rules for DIDs matching a RegEx

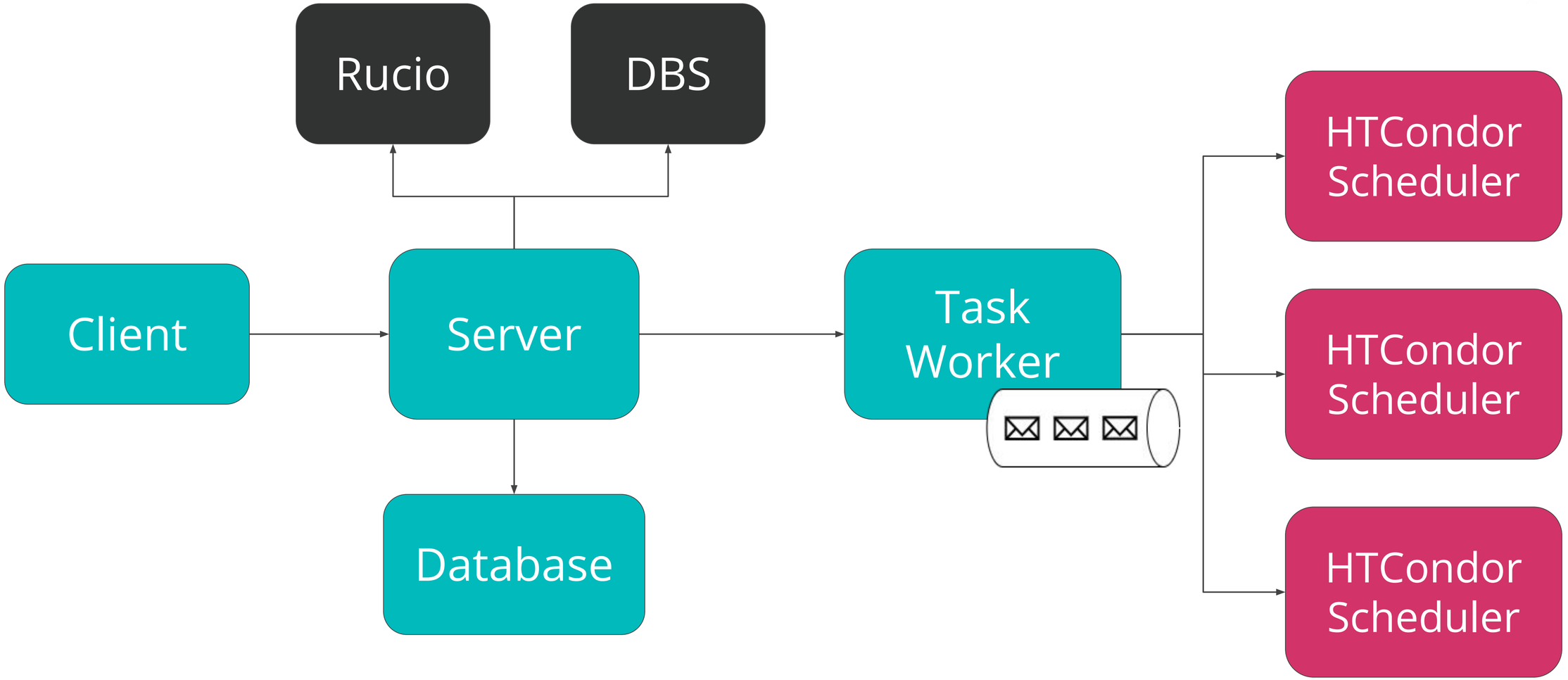
- WMAgent uses it
- Users create rules via auto-approval activity
  - Lifetime <= 6 months
  - 1 PB when RSEs are not specified
  - 50 TB for single RSEs



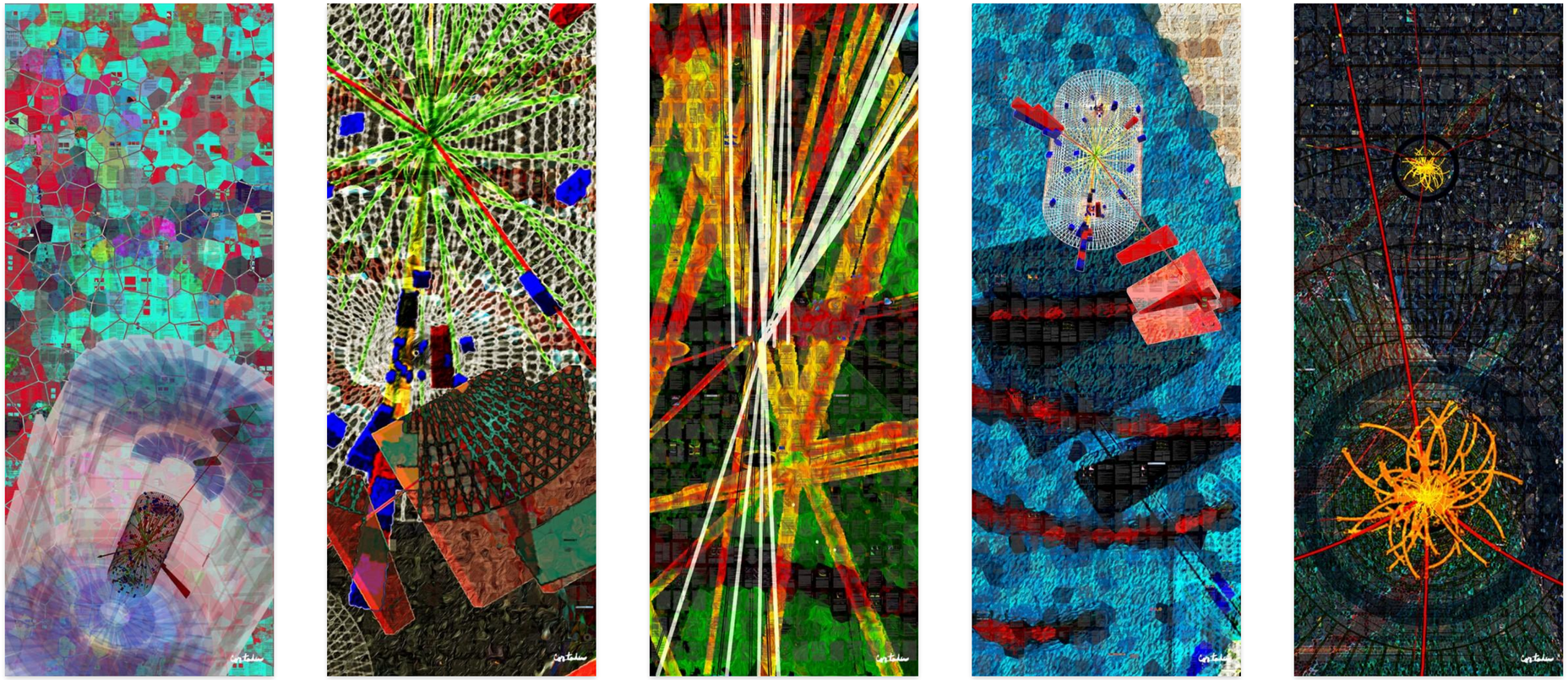
CRAB is a utility to submit CMSSW jobs to distributed computing resources

- Access Raw data and Monte-Carlo
- Exploit the CPU and storage resources

CRAB Architecture



Content Invalidation and Deletion



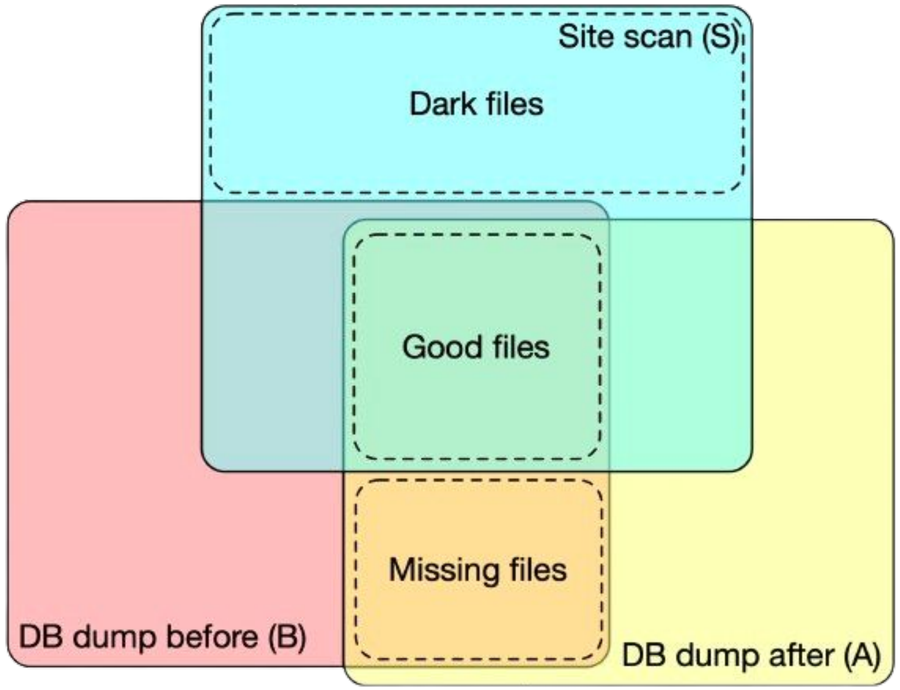
Planned Deletions

Unexpected Deletions

UD - Consistency Check

Inconsistency detection:

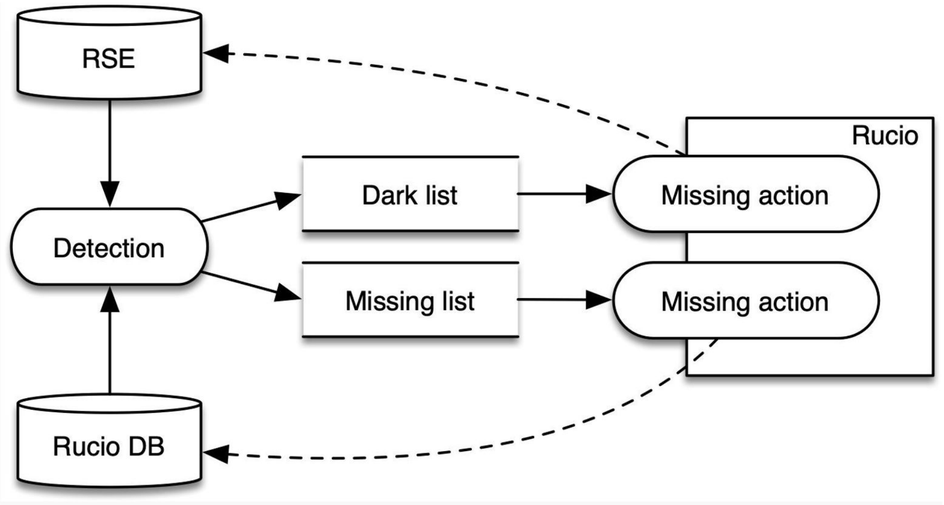
1. Dump site contents -> list of replica paths
2. Dump Rucio database “replicas” table -> list of paths or LFNs
3. Compare site contents to DB contents
   a. Dark replicas
      i. A and B - all replicas (including BAD, SUSPICIOUS, etc.)
   b. Missing replicas
      i. A and B - only AVAILABLE replicas



UD - Consistency Check

Inconsistency correction actions:

1. Declare missing replicas as BAD (scopes, names)
   - Force rucio to re-transfer replicas
2. Quarantine dark replicas (paths)
   - Dark Reaper daemon will remove those replicas



UD - Consistency Check

Safeguards

If number of missing or dark files > 5% (configurable per RSE) of total number of files found in the RSE, the corresponding action is aborted

- Admin review is required

Dark replica is quarantined depending on site if

- It appears as dark in all runs since X1 days ago
- There were at least X2 runs since X3 days ago
- The fist run is at least X4 days old

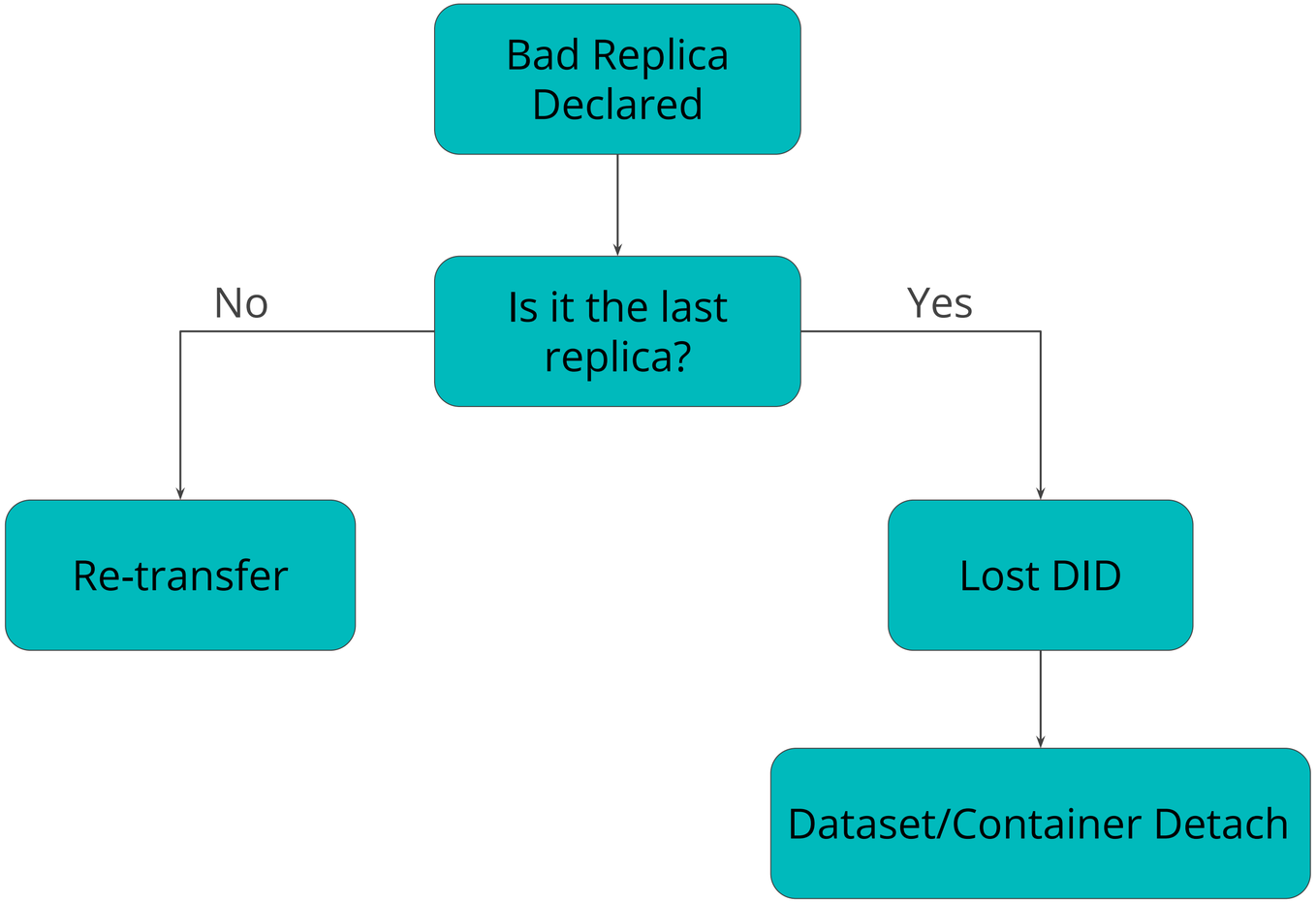
UD - Rucio Daemons

- **Conveyor-finisher** can declare files suspicious based transfer errors when given a set of patterns
  - Works for source replica
- Kronos can declare files suspicious based on the “stateReason” field of file traces when given a set of patterns

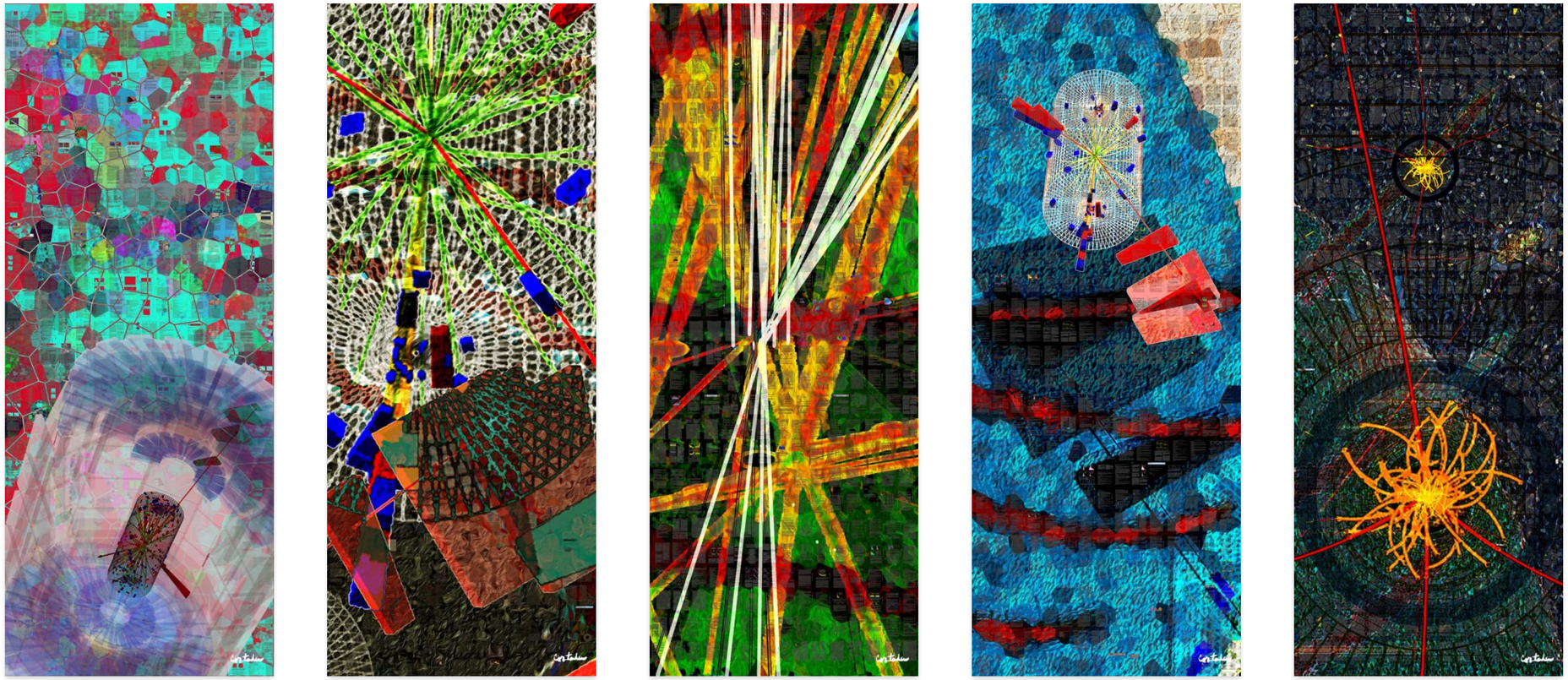
Suspicious Replica Recoveror

Declares suspicious replicas that are available on other RSE as bad. Consequently, automatic replica recovery is triggered via **necromancer daemon**.

Deletion Process



Rubin Approach



Example of T0 WMAgent config. (Flavor from default WMAgent)

addDataset(tier0Config, "Default", -> includes RAW DATA

archival_node="T0_CH_CERN_MSS", -> container *rule to rse*

tape_node="T1_US_FNAL_MSS", -> container *rule to rse*

disk_node="T1_US_FNAL_Disk", -> container *rule to rse*

*dataset_lifetime=3*30*24*3600 -> container rule lifetime default (disk)*

addExpressConfig(tier0Config, "ExpressCosmics",

diskNode="T2_CH_CERN", -> container *rule to rse*

alca_producers=["SiStripPCLHistos", "SiStripCalZeroBias", ..] -> subproducts

*dataset_lifetime=12*30*24*3600 -> container rule lifetime (disk)*

PromtReco Input is RAW Data.

DATASETS = ["Cosmics"]

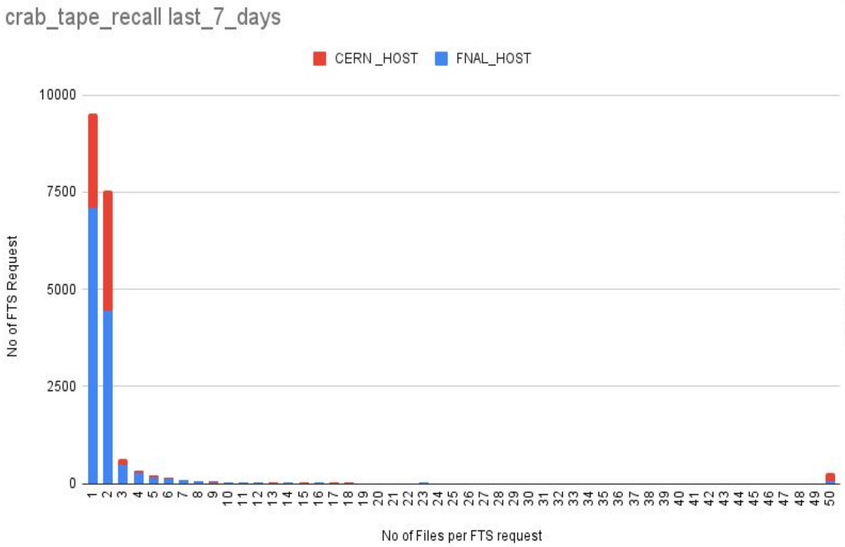
for dataset in DATASETS:

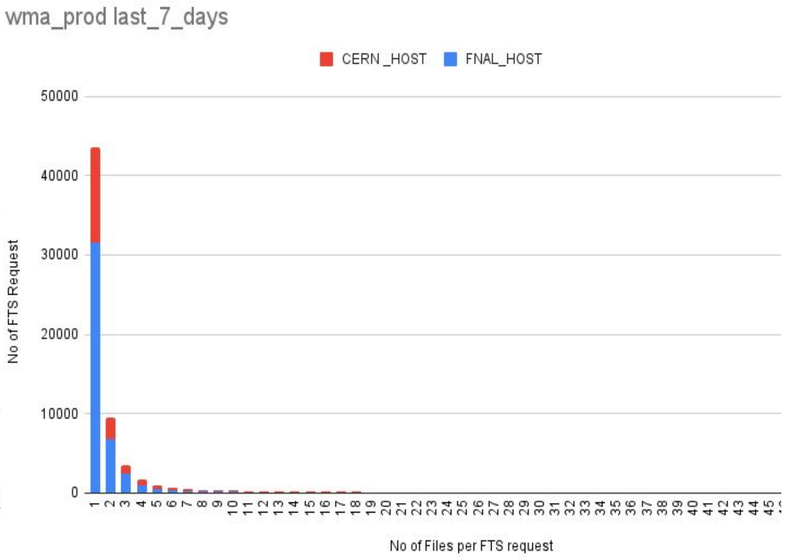
addDataset(tier0Config, dataset,

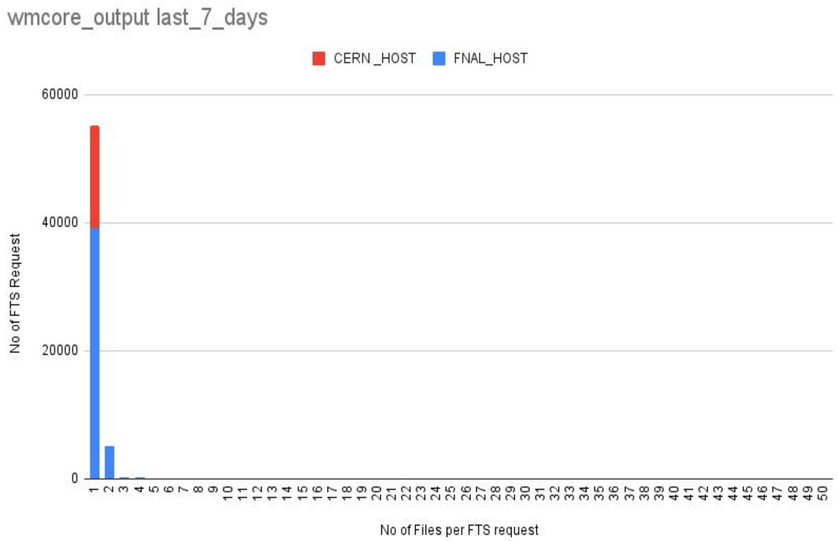
alca_producers=["SiStripCalCosmics", "SiPixelCalCosmics",..] ->subproducts

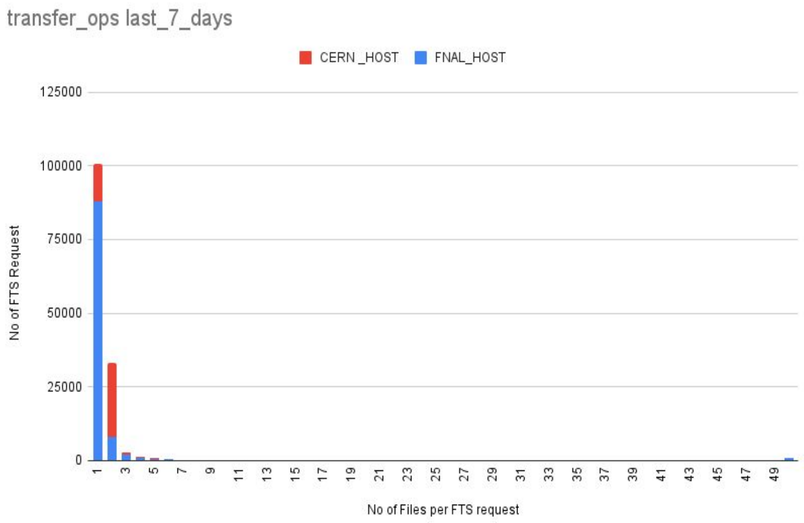
physics_skims=["CosmicSP", "CosmicTP"...] -> sub-subproduct

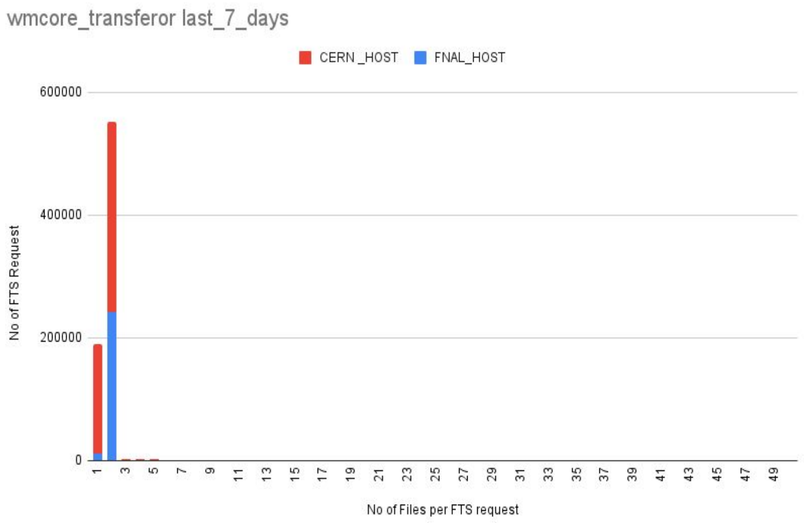
FTS granularity ~ 1.03M





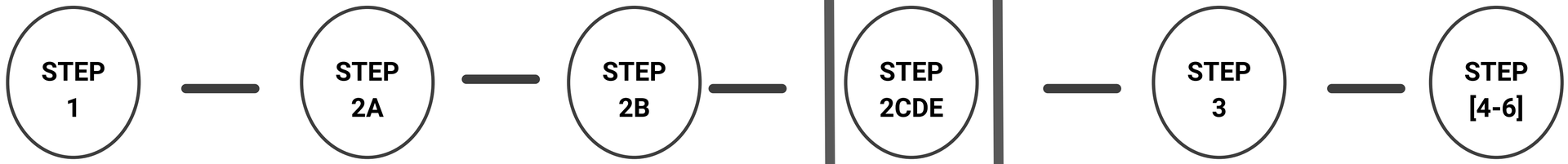






DRP (1 year survey)

~200k chips
~200 days
40M files



USDF

FRDF

UKDF

>28M files to transfer.
X fts request

> fts request

>fts requests

> fts requests

>fts requests

>fts requests

Rucio Subscription_1

Rucio Subscription_2

Rucio Subscription_3

Rucio Subscription_4 **Only metadata**

Rucio Subscription_5

Rucio Subscription_6

Questions?
Comments?