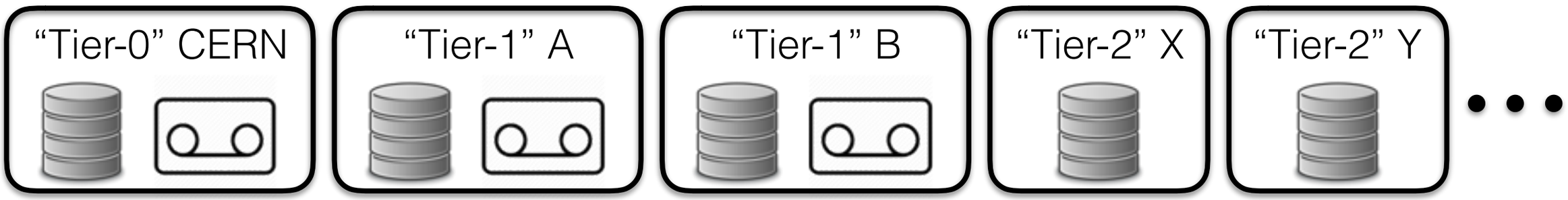
# Dynamo

## The dynamic data management system for the distributed CMS computing system

Maxim Goncharov, Yutaro Iiyama, Benedikt Maier, Christoph Paus (MIT)
for the CMS Collaboration

# CMS storage pools and tools

50+120 PB of collision and simulation data distributed across ~60 disk+tape pools around the world.




Data managed in units of datasets and data blocks.
~300k datasets, >3M blocks

Dataset Bookkeeping System (DBS):

- Maps datasets and data blocks to file names
- Handles dataset metadata

Physics Experiment Data Export (PhEDEx):

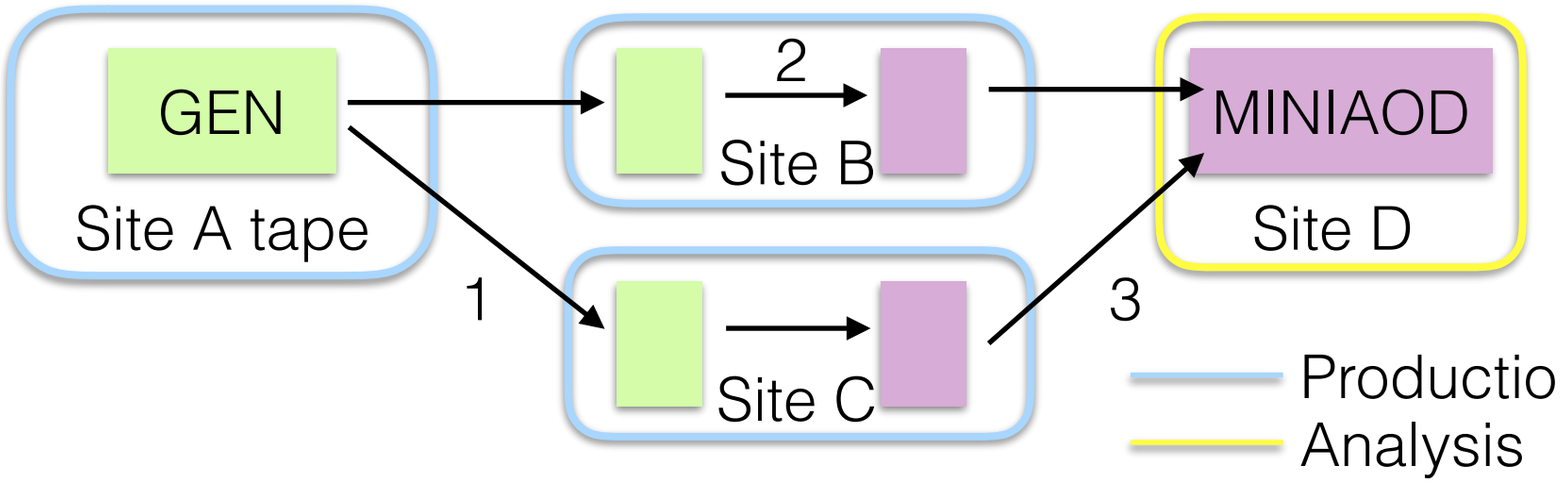
- File catalog = tracks physical placement of dataset replicas
- Orchestrates file copies between sites and deletions

# CMS storage pools management

Storage pool divided in partitions:

- Production = Input / output of data processing
- Analysis = Reconstructed data for physicists
- Other, purpose-specific (software release validation etc.)

Typical data flow




1. Input dataset is staged to processing sites
2. Sites process parts of dataset
3. Output dataset is merged at the final host site

Data management needs:

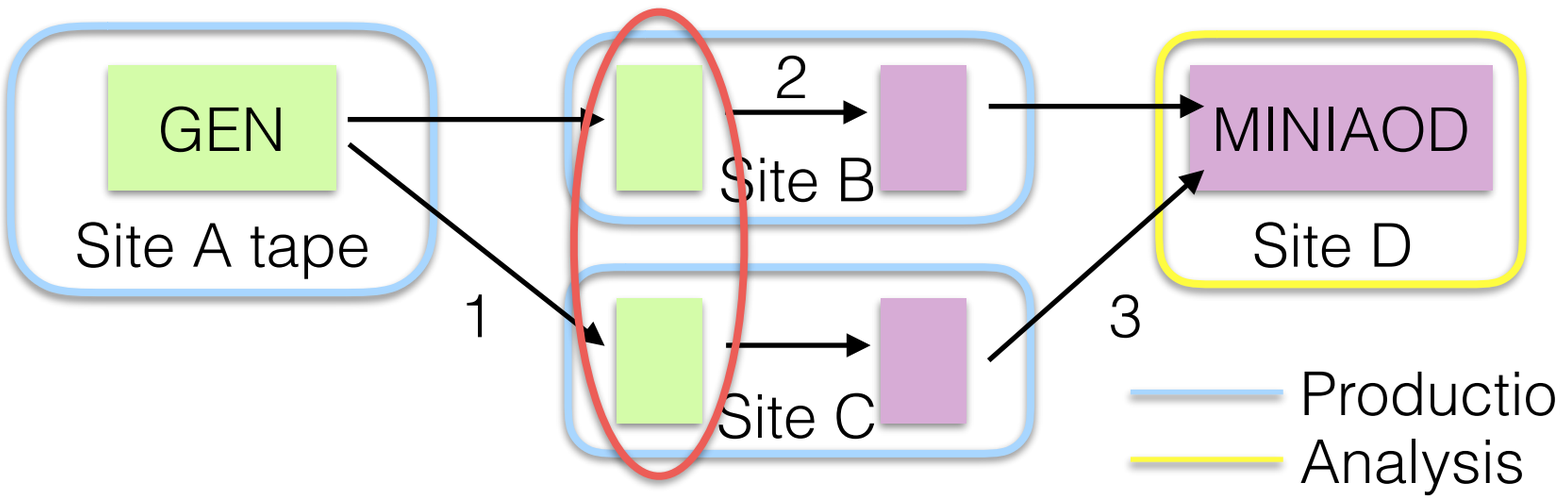
- Delete unneeded data
- Replicate popular analysis data

# CMS storage pools management

Storage pool divided in partitions:

- Production = Input / output of data processing
- Analysis = Reconstructed data for physicists
- Other, purpose-specific (software release validation etc.)

Typical data flow




1. Input dataset is staged to processing sites
2. Sites process parts of dataset
3. Output dataset is merged at the final host site

Data management needs:

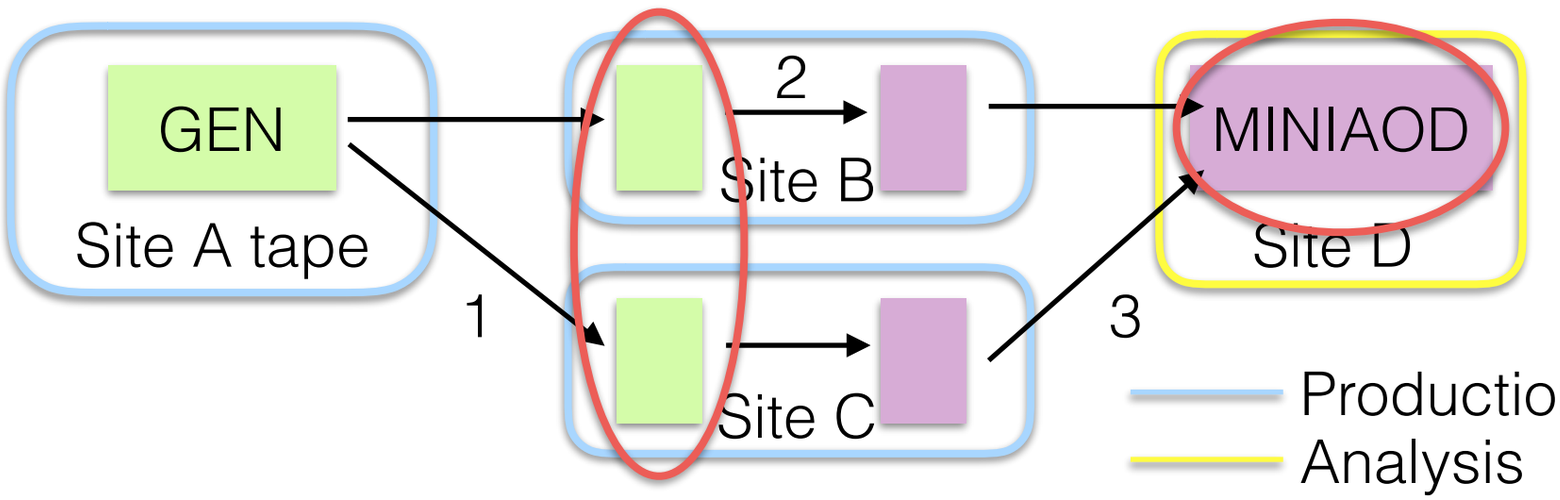
- Delete unneeded data
- Replicate popular analysis data

# CMS storage pools management

Storage pool divided in partitions:

- Production = Input / output of data processing
- Analysis = Reconstructed data for physicists
- Other, purpose-specific (software release validation etc.)

Typical data flow




1. Input dataset is staged to processing sites
2. Sites process parts of dataset
3. Output dataset is merged at the final host site

Data management needs:

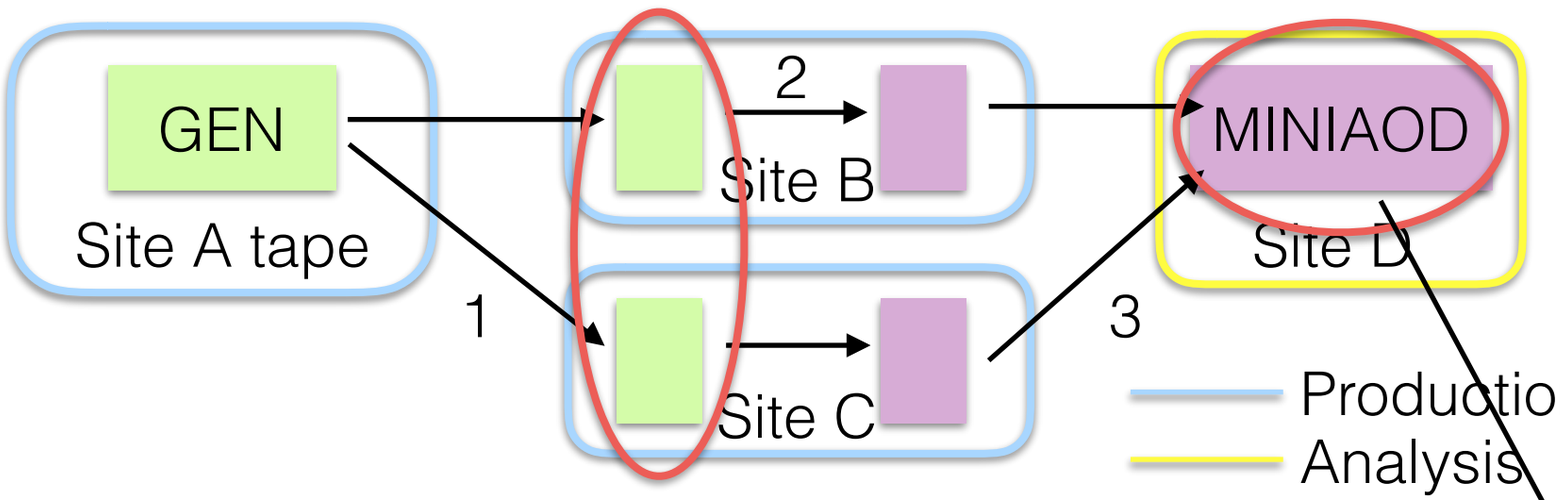
- Delete unneeded data
- Replicate popular analysis data

# CMS storage pools management

Storage pool divided in partitions:

- Production = Input / output of data processing
- Analysis = Reconstructed data for physicists
- Other, purpose-specific (software release validation etc.)

Typical data flow




1. Input dataset is staged to processing sites
2. Sites process parts of dataset
3. Output dataset is merged at the final host site

Data management needs:

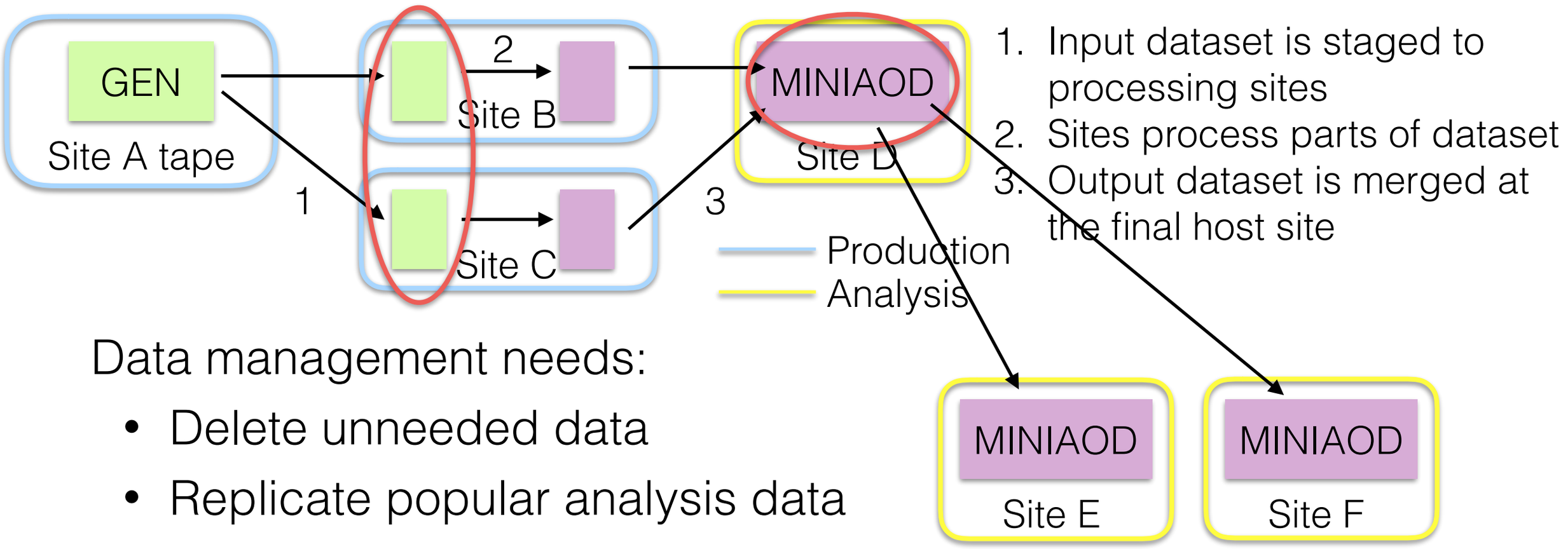
- Delete unneeded data
- Replicate popular analysis data

# CMS storage pools management

Storage pool divided in partitions:

- Production = Input / output of data processing
- Analysis = Reconstructed data for physicists
- Other, purpose-specific (software release validation etc.)

Typical data flow




Data management needs:

- Delete unneeded data
- Replicate popular analysis data

# Dynamo

Dynamic & automatic data management (DDM) should:

- Delete data as needed according to set policies
- Analyze data usage trend and replicate popular datasets

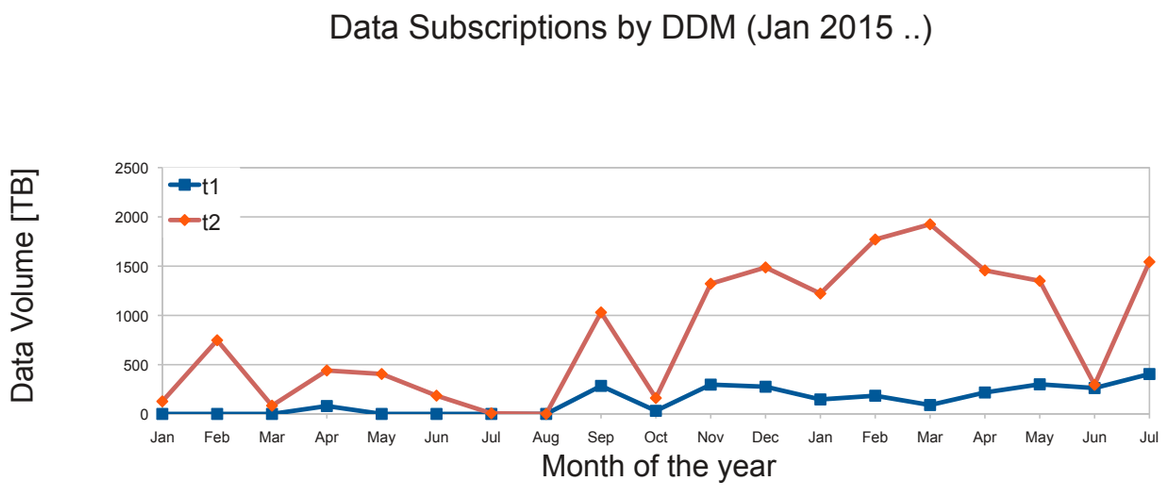
First version of CMS DDM deployed in April 2014.
Scope expansion, policy complexation … → Upgrade required

Second version: **Dynamo** (since July 2016)

- Modular and expandable software architecture.
- Flexible, config-file based policies. Versioned and traceable.

Two components

- Detox: data deletion
- Dealer: data distribution



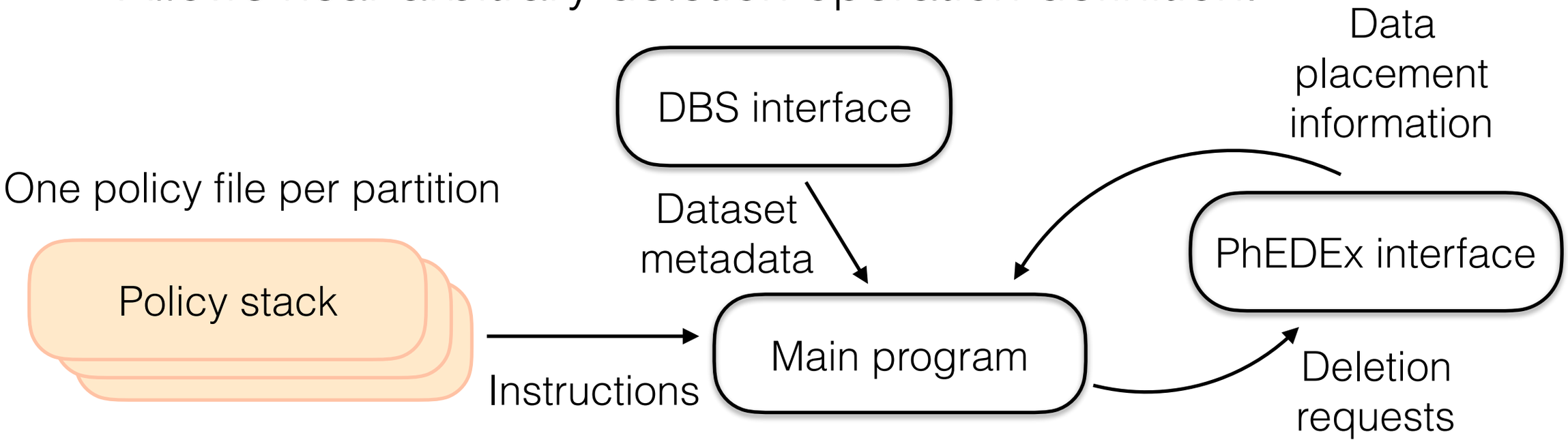
# Detox - data deletion

Main program driven by a plain-text config file (“policy stack”) that defines:

1. Datasets and sites to consider
2. Data classification into “Protect”, “Can delete”, and “Delete”
3. When to delete “can delete” data
4. Priority of deletion

Conditions written in terms of site and dataset properties, and translated to python code.

→ Allows near-arbitrary deletion operation definition.



# Detox policy stack

## Example 1: Routine Analysis partition cleaning

```
### Target sites
On site.name in [T1_*_Disk T2_*] and site.status == READY

### Deletion trigger
When site.occupancy > 0.9
Until site.occupancy < 0.85

### Replica protection / deletion policies
Delete dataset.status == DEPRECATED and dataset.last_update older_than 1 week ago
Delete dataset.status == INVALID and dataset.last_update older_than 1 week ago
Protect replica.incomplete
...
Dismiss dataset.usage_rank > 400
Protect dataset.name == /*/*/MINIAOD* and dataset.num_full_disk_copy < 3
Protect dataset.num_full_disk_copy < 2

# Default decision
Dismiss

### Candidate ordering
Order decreasing dataset.usage_rank
```

1. Where

3. When

2. Classification

4. Prioritization

# Detox policy stack

## Example 2: “Greedy” deletion of obsolete backups

```
### Target sites
On site.name == T*_MSS and site.status == READY

### Deletion trigger
When always
Until never

### Replica protection / deletion policies
# Protect all Spring16 and Summer16
Protect dataset.name == /*/*Spring16*/*
Protect dataset.name == /*/*Summer16*/*
...

Dismiss dataset.name == /*/*/GEN-SIM-RAW
Dismiss dataset.name == /*/*/GEN-RAW
...

## All else are protected
Protect

### Candidate ordering
Order none
```

1. Where

2. Classification

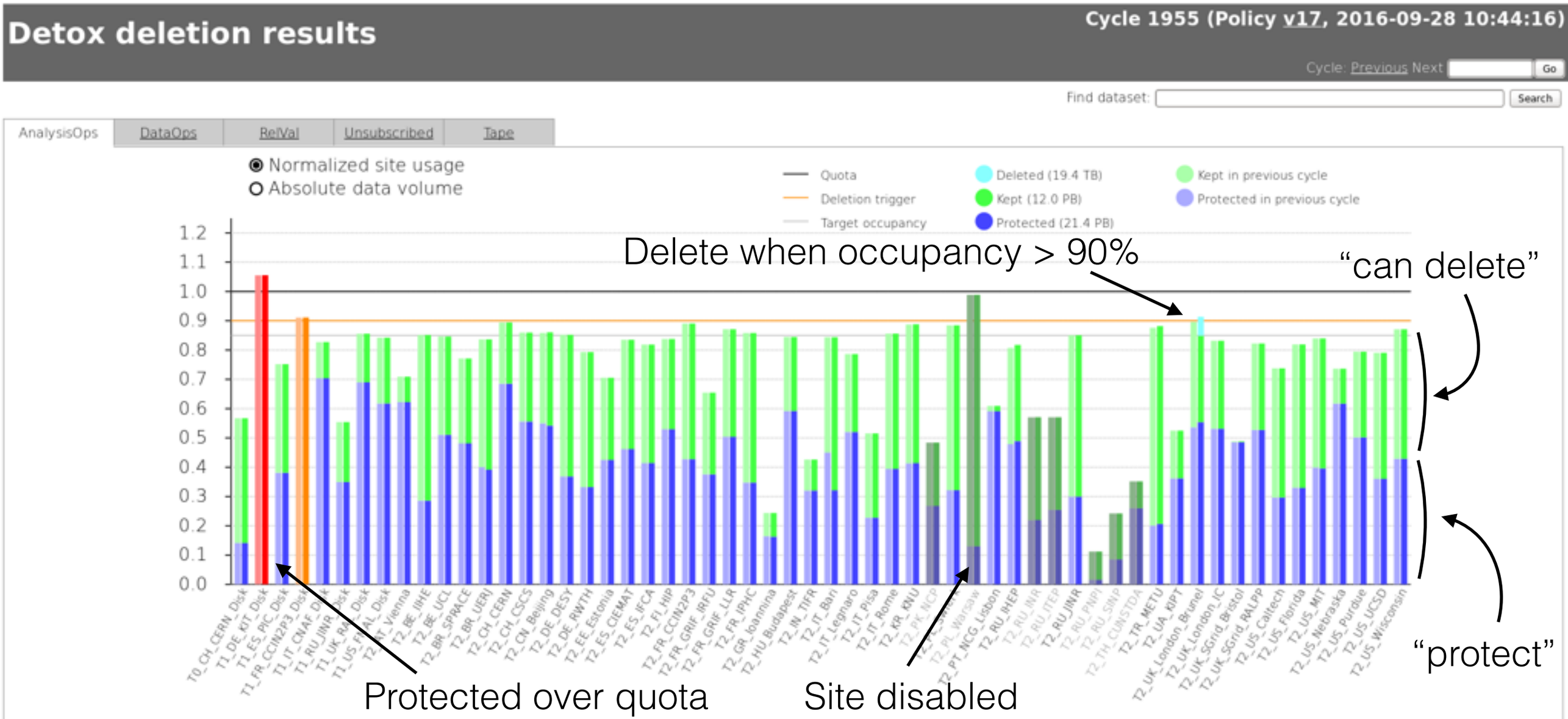
3. When

4. (no) Prioritization

# Detox web interface

Site usage history monitored via a web interface.

Dataset classification and their reasons at each program execution also made available.



# Dealer - data distribution

Currently focused on replicating popular datasets.

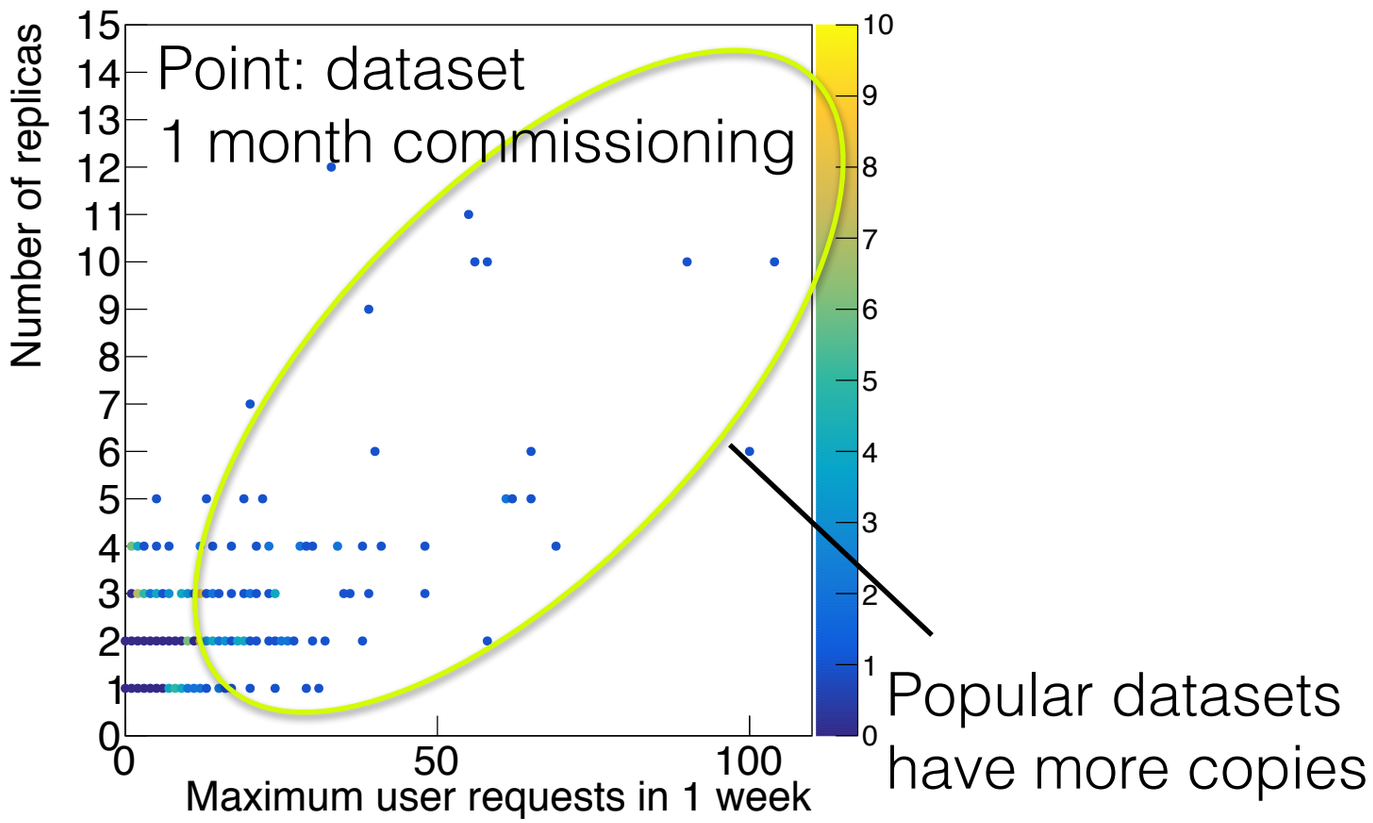
Queries HTCondor scheduler of the CRAB* system
→ Real-time dataset popularity

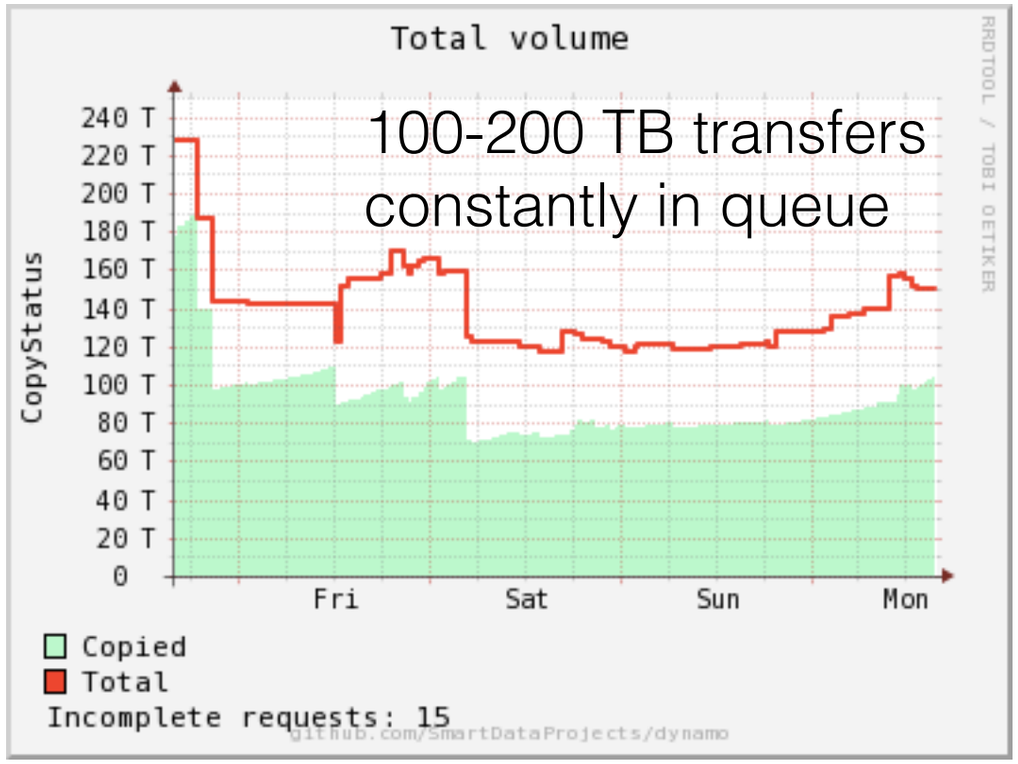
* CMS Remote Analysis Builder

Using a simple algorithm for replication.

For each dataset:

1. Compute time-weighted number of requests $n$
2. Make copies until number of replicas $R > C \cdot n$ ($C = 1.75$)



# Going beyond

Extending Dealer roles:

- Site occupancy balancing
- Automatic recovery of missing data
- Offloading data from sites with troubles

Active development ongoing.
2-3 months timescale to complete the extension.

Goal of Dynamo is to eliminate the need for human intervention in CMS data management and optimize our resource usage.

# Conclusion

- CMS manages 170 PB of data split in millions of blocks distributed across 60 sites worldwide.
- Intelligent and automatic storage resource optimization: Dynamic Data Management.
- Dynamo features
  - Modular architecture
  - Flexible, configurable operation policy
- Active development ongoing for fully automated data management and optimized resource usage.