



UiO : **University of Oslo**

Data placement issues

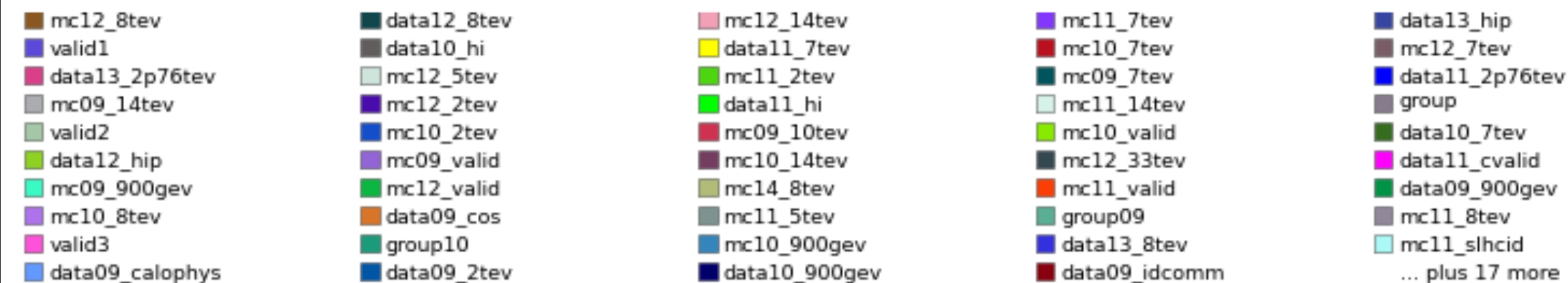
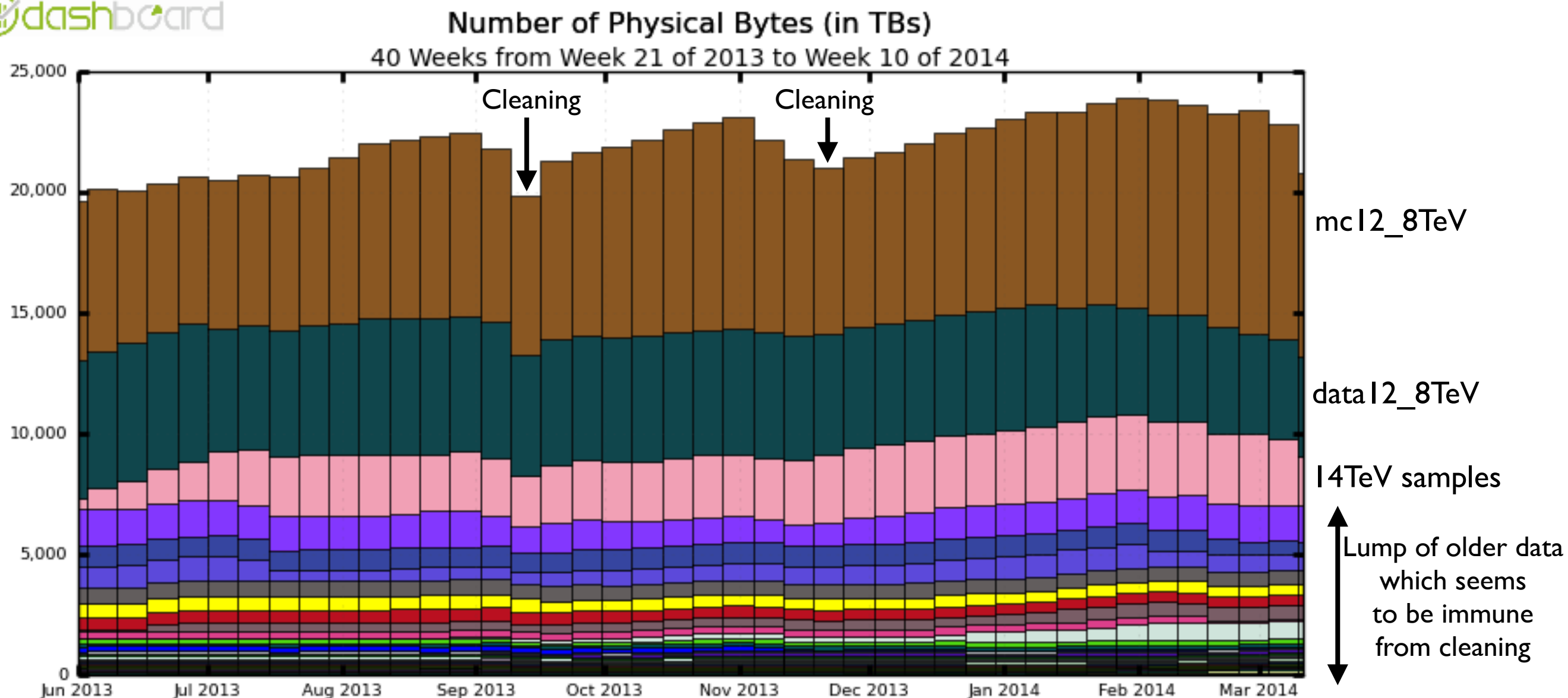
James Catmore (UiO)

- Replica status
 - ▶ **custodial**: permanent replicas on tape
 - ▶ **primary**: replicas on DISK, not subject to the automatic cleanup. A lifetime can be defined.
 - ▶ **secondary**: subject to the automatic cleanup in case of disk space shortage in the site. Least popular datasets get deleted first within a site
- Space tokens managed by DDM
 - ▶ **DATADISK**: centrally managed space allocated to the official datasets. These replicas can be written in this place by central production, PD2P or DaTRI. Automatic cleaning of secondary replica is triggered if full
 - ▶ **DATATAPE/MCTAPE**: centrally archived tape copies of the centrally produced datasets
- **Replication Policy**: list of rules on exactly what data/MC should be where, how many replicas and their status
 - ▶ <https://twiki.cern.ch/twiki/bin/viewauth/Atlas/ReplicationPolicy2013>
 - ▶ A thrilling read... too long to summarize but will refer to it throughout

The problem:

steady increase in primary data on Tier-1 datadisk

3



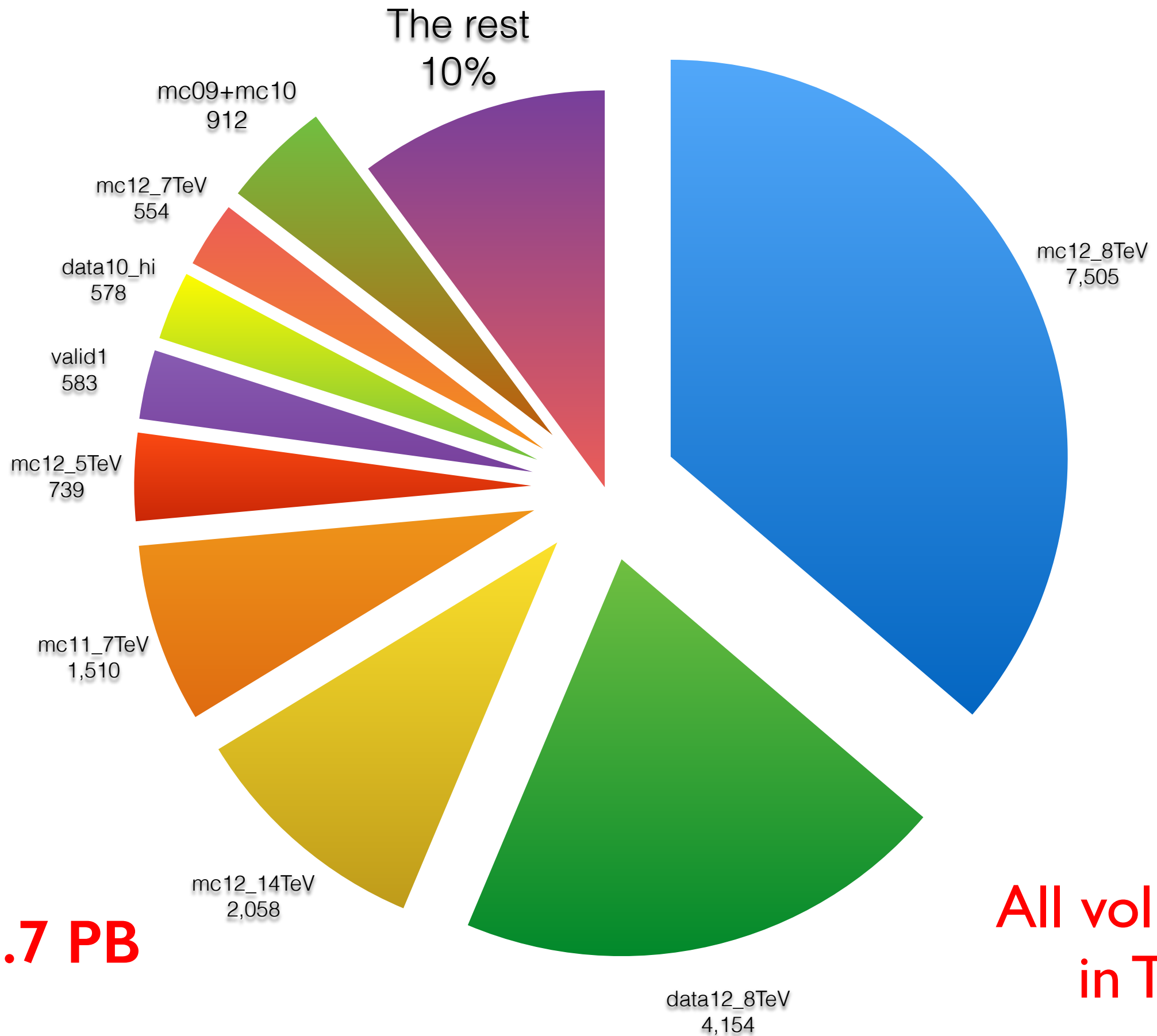
Maximum: 23,910 , Minimum: 0.00 , Average: 21,423 , Current: 20,797

0th March 2014

- Last week, many of the Tier-I sites ended up with less than 100TB of free space on their data disks.
- ADC was limited in how much space it could free because the data was primary
- Since each sites needs ~300TB of free space to support ongoing operations (reprocessing, MC production) we hit the wall
- ATLAS was not able to run the NTUP_COMMON v2 production

- Several reasons
 - ▶ Large productions done in the last few months
 - 22% of the primary data was created in the last 90 days
 - ▶ Replication policy has too much emphasis on keeping Tier-I primary copies
 - ▶ Old primary datasets sitting around on Tier-I not being used

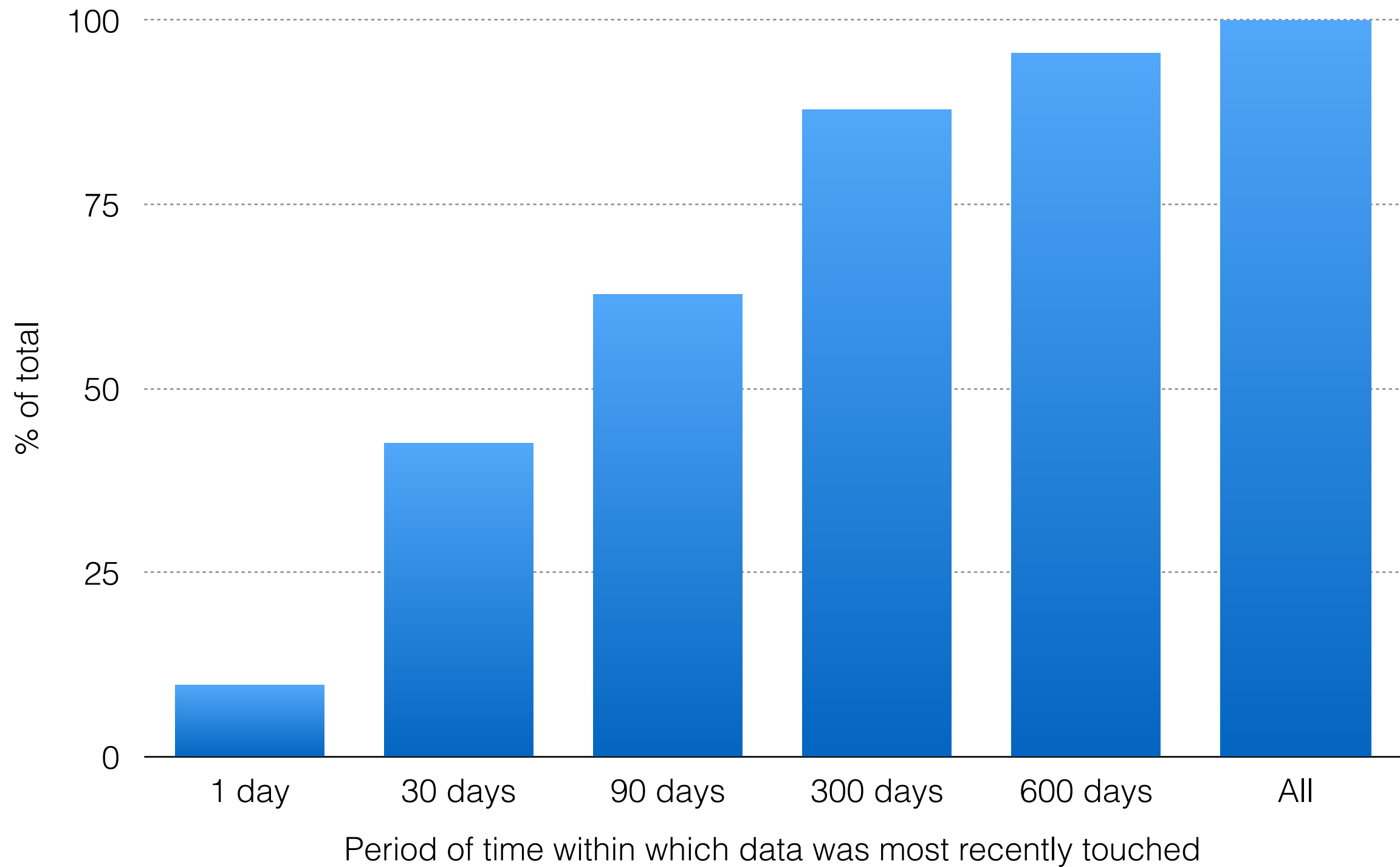
Overall picture: primary datasets on T1 datadisk



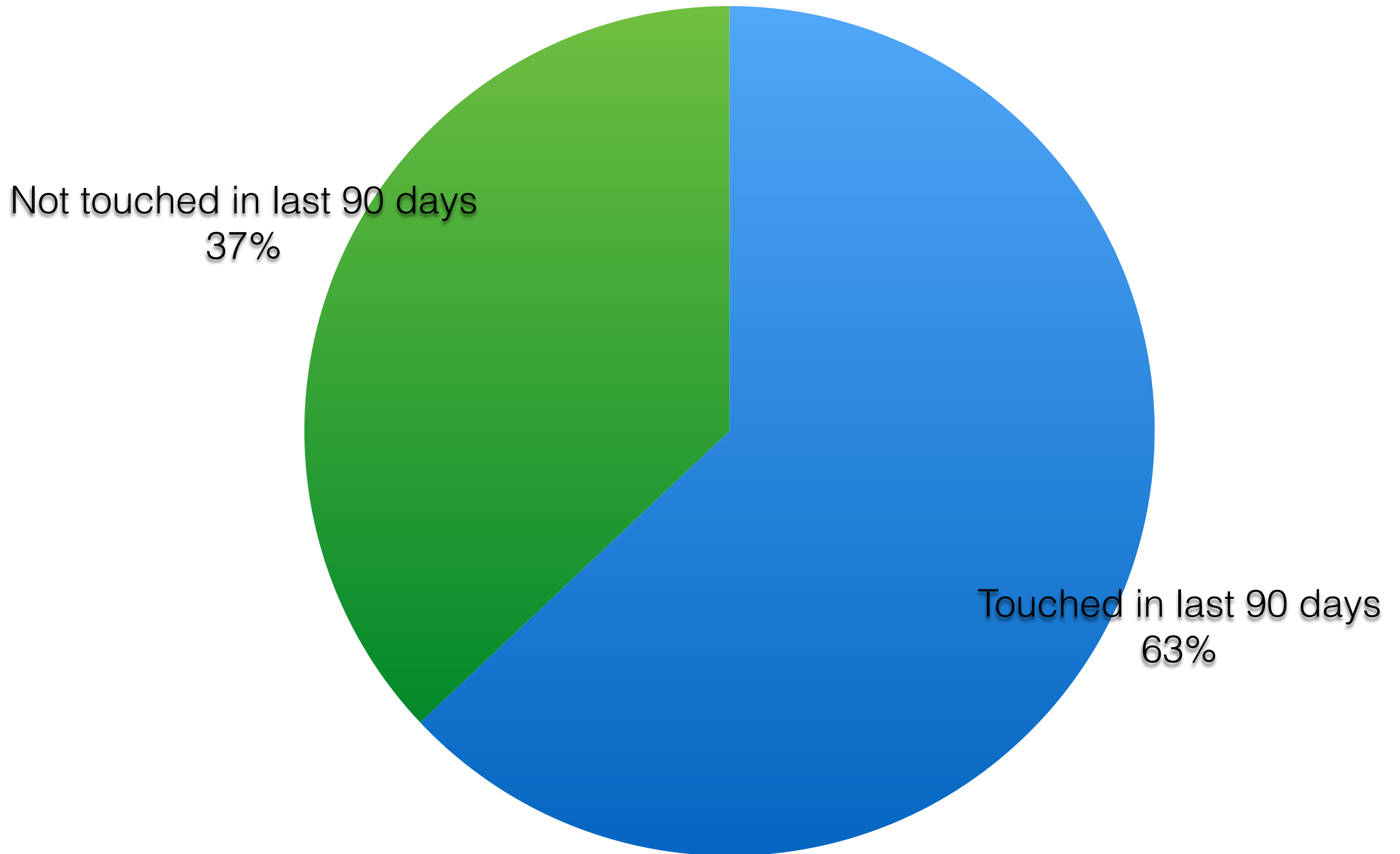
Total: 20.7 PB

**All volumes
in TB**

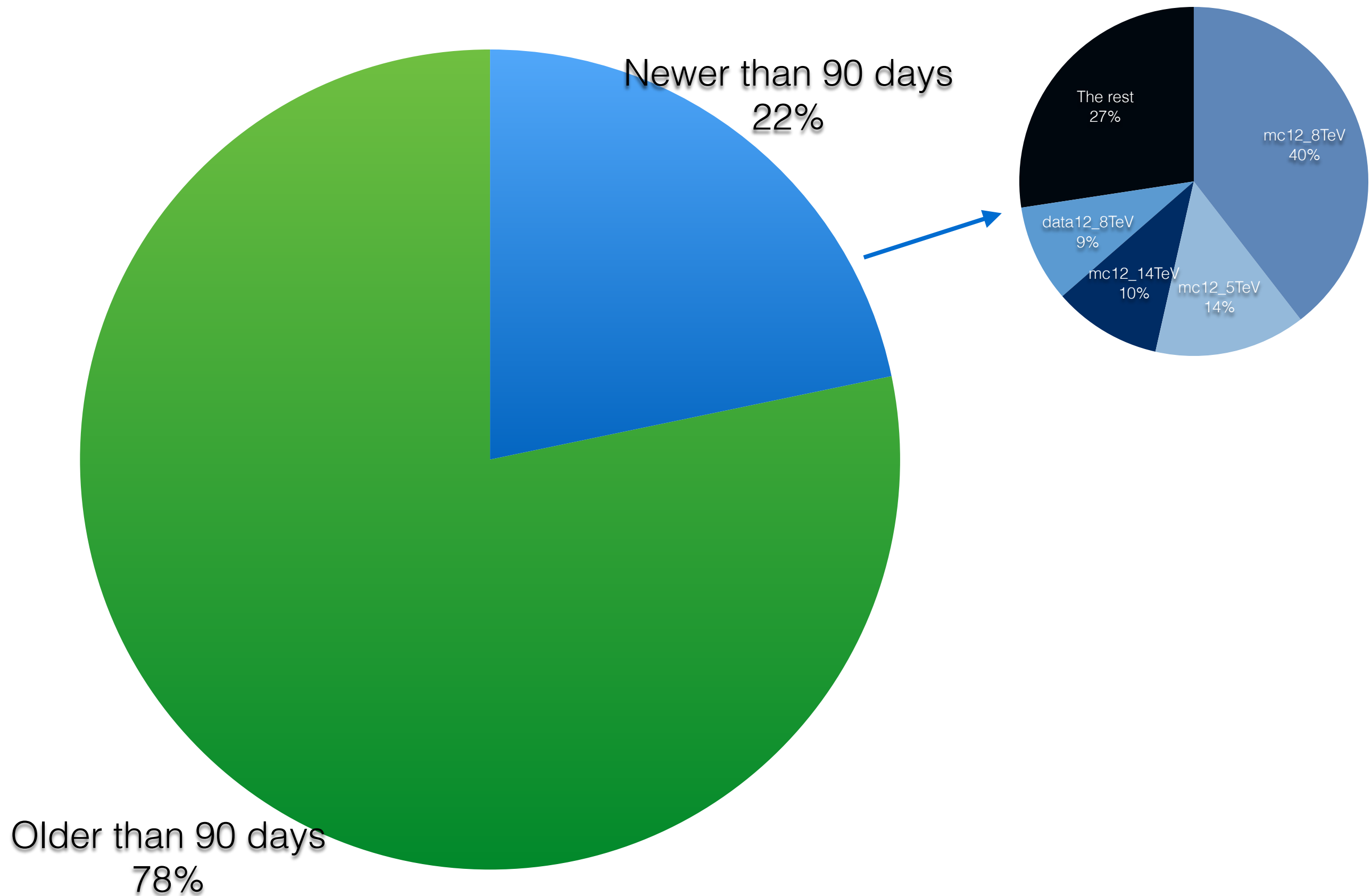
All T1 primary datadisk



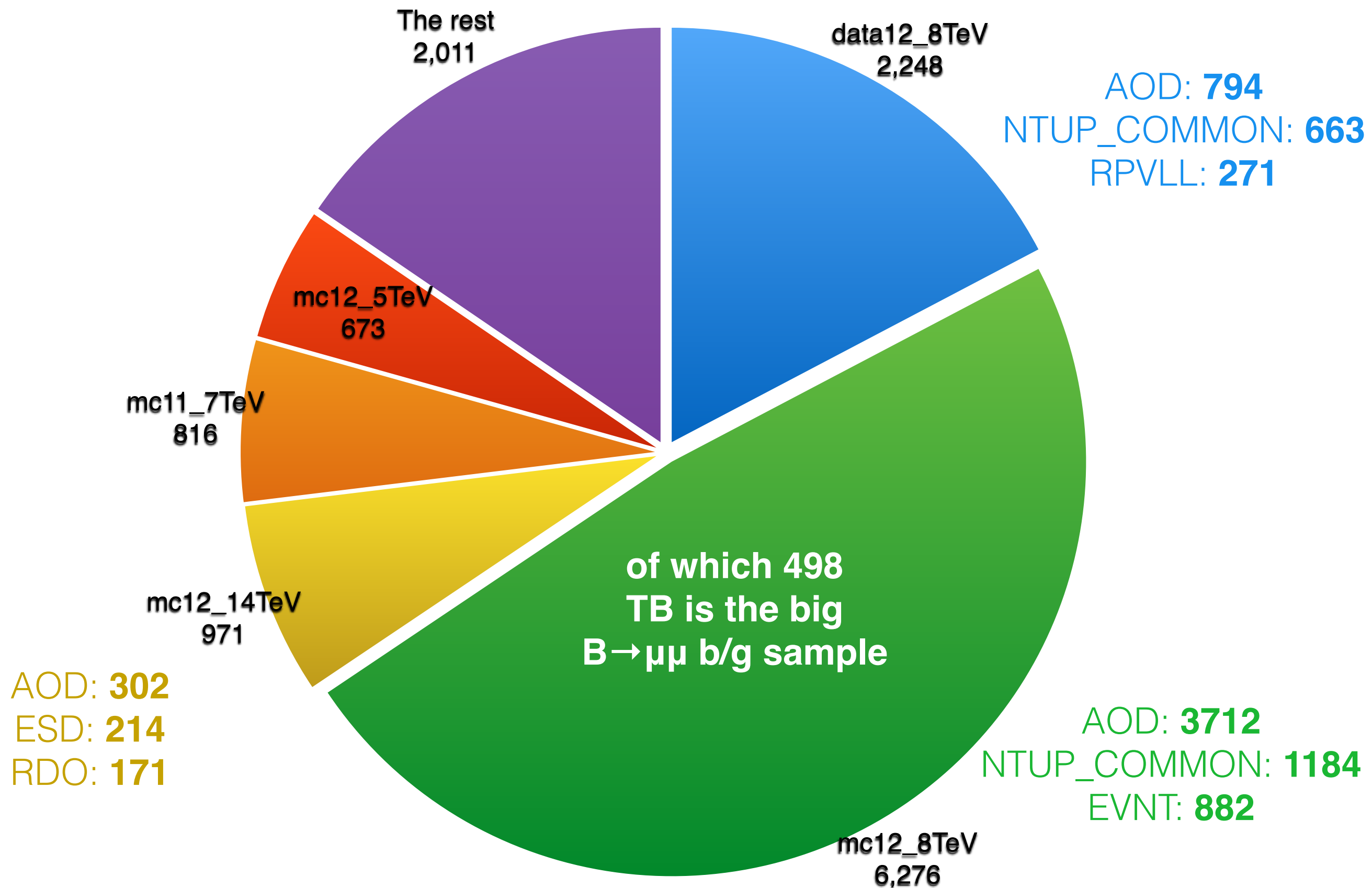
All T1 primary datadisk



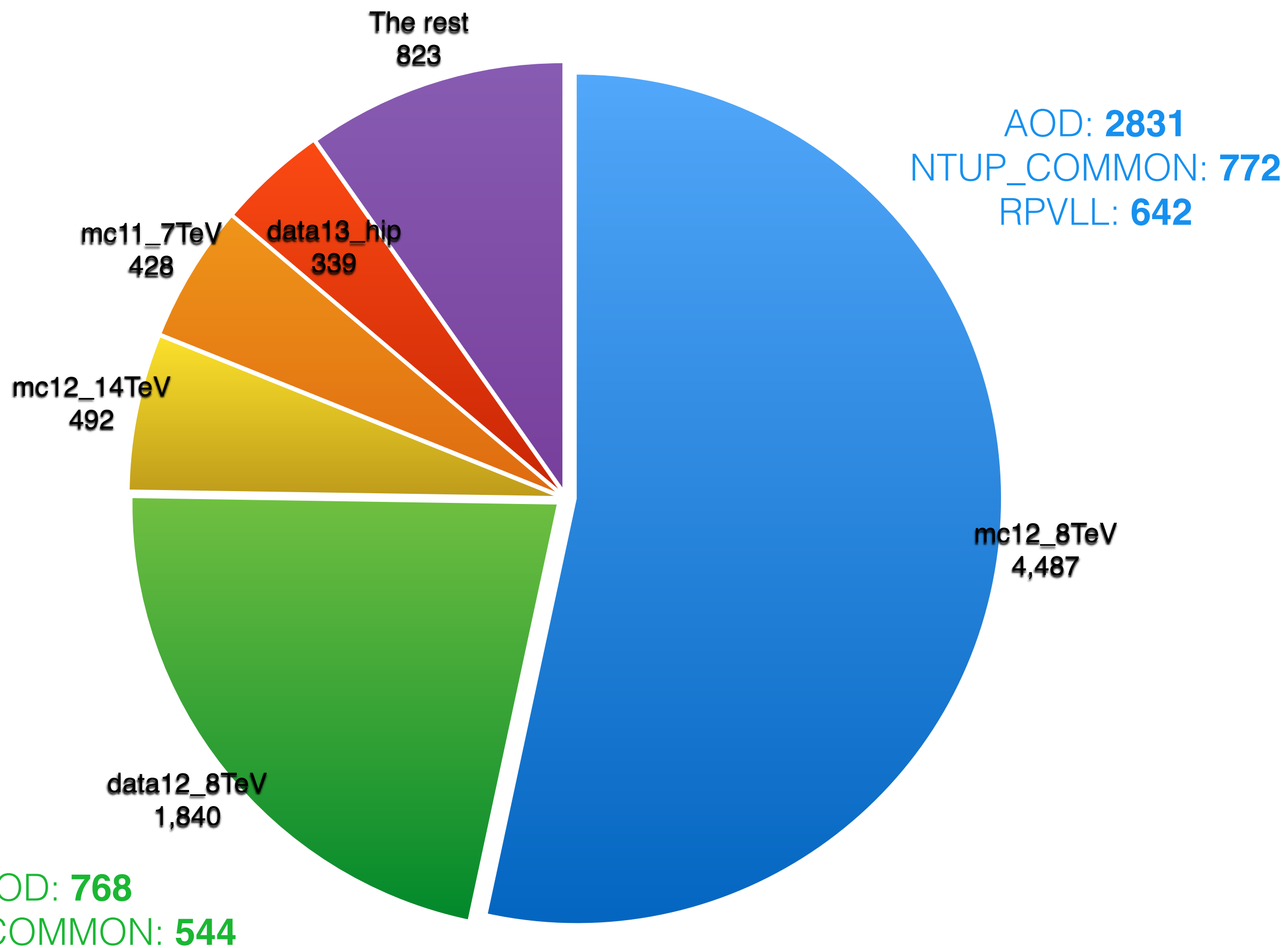
All T1 primary datadisk



Files touched in the last 90 days



Files older than 90 days touched in the last 90 days

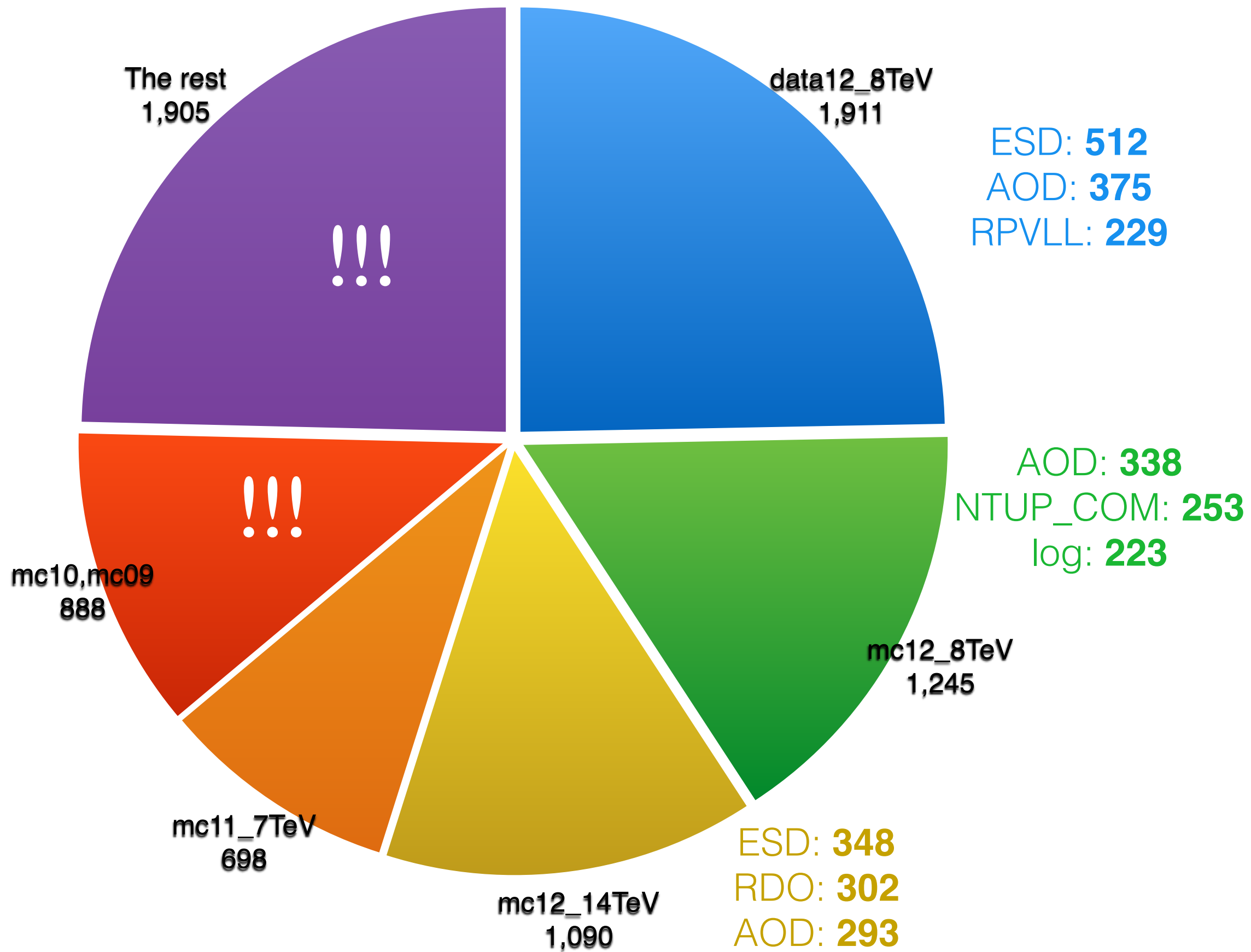


Older than 90 days touched in the last 90 days:
mc12_8TeV AOD sorted by DSN

DSN	What?	Size
208010	$J/\psi \rightarrow \mu\mu$	55TB
105200	t/tbar	52TB
147913	Di-jet slice	51TB
18539 others, < 50TB		2673TB

Huge range of datasets in use

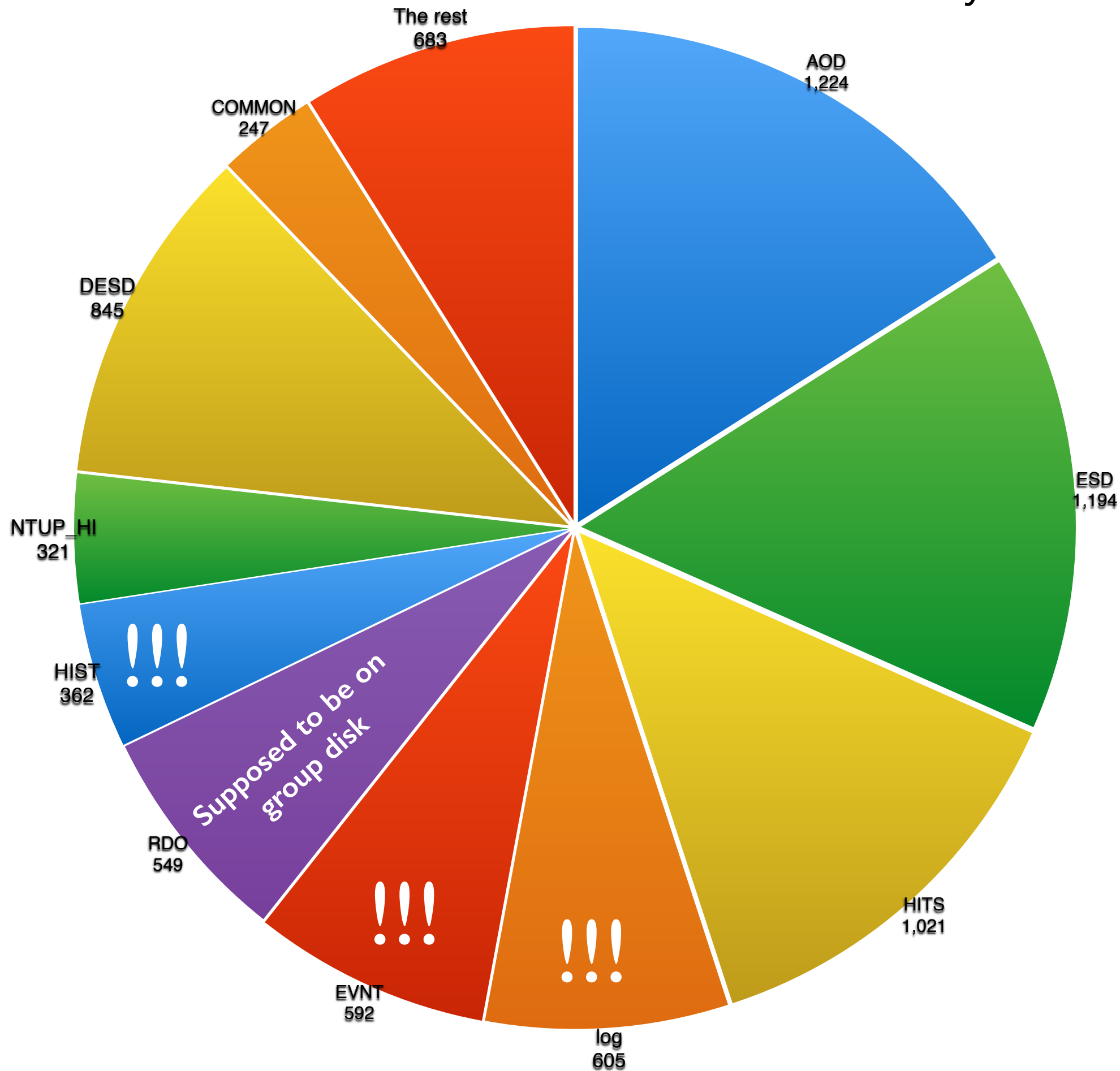
Files **not** touched in the last 90 days



What is “the rest”?
(only datasets > 1TB)

	Size (TB)
group10	2
data11_hi	3
mc11_8TeV	6
group09	6
mc11_5TeV	7
data09_cos	7
data11_cvalid	9
mc11_valid	14
data09_900GeV	15
data11_2p76TeV	15
mc12_valid	15
valid2	17
mc12_33TeV	20
data13_2p76TeV	22
data12_hip	26
data10_7TeV	42
group	44
mc12_2TeV	53
mc12_5TeV	65
mc11_14TeV	73
mc11_2TeV	161
valid1	167
data13_hip	199
mc12_7TeV	257
data11_7TeV	267
data10_hi	359

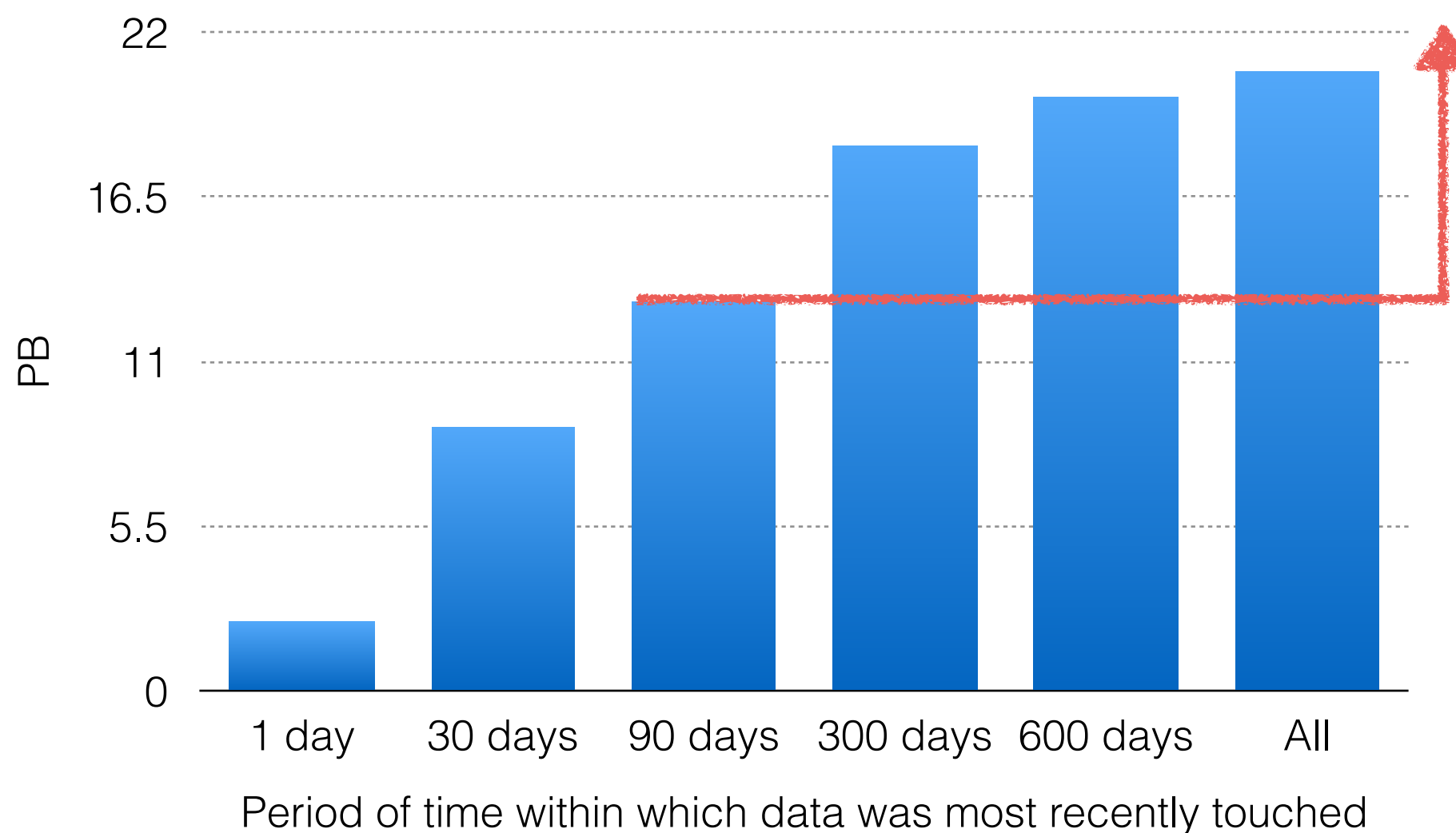
Files **not** touched in the last 90 days



*How much of this would be perfectly
OK on Tier-2 disk or on tape???*

- New production is not a big part of the primary data (20%)
 - ▶ so not really correct to say that the problems have been caused by recent big MC requests
 - ▶ although clearly when there is so little headroom to start with, an extra 500TB of MC is going to be problematic
 - ▶ so we need to make more headroom!
- Old data which has been accessed recently is a much bigger piece
 - ▶ but this data is clearly useful so not a problem
- Old data which is not being used is the thing we should attack - roughly 40% of the total

- **Emergency fix:** any primary data not touched for more than three months should be set to secondary, provided they are found elsewhere (e.g. on tape)
 - ▶ Would automatically clean up the very old datasets with little manual intervention
 - ▶ Manually chasing up the owner of every single old dataset isn't an option
- Could send a list of datasets on the “hit list” to data prep, physics coordination and upgrade coordination and ask for them to identify any datasets that need to be kept



- Longer term: revise the data placement policy
 - ▶ Fewer/no primary copies at Tier-1 sites?
 - ▶ What about the Tier-2 sites?
 - ▶ More use of custodial copies?
 - ▶ Fewer exceptions?
 - Too many lines say “do X except for case A,B,C,D,E” - this is why we end up with legacy primary data
 - ▶ Obvious policy problems: EVNT, log files, old MC
- This work is under way now
 - ▶ Line-by-line analysis of the policy to assess the impact of each part on our overall use of disk resources
 - ▶ Will have preliminary discussion in the CREM on Thursday

- Where we put our files is as important as how much of them there are
- Long-term storage on Tier-1 disk of files which are rarely read is **not efficient**
 - ▶ It can have serious effects on ongoing operations
- Reduction of replicas **does not** mean that the data are not accessible or not usable
 - ▶ for production single custodial tape copy is often used
 - ▶ for analysis we have DaTRi just for this: users can make the requests
 - we don't have statistics today on how often it happens that data are recalled from tape by users, it is believed to be very low activity
 - moreover today the tape systems are not heavily used so even if we get a bit more requests, we can sustain the load
 - on the other hand, deleting data from tape is a complex procedure, and tape isn't an infinite resource
- We clearly need to re-think our data placement policy in general
 - ▶ and for Monte Carlo in particular
 - ▶ This work is under way now