Proposal for Sharing of Data on Data Access

Diego Davila, June 2019

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Outline

- The Goal
- The 3 datasets we propose
- Where is the data? and What is available?
- How are they grouped/compressed?
- How they look? and How big they are?

The Goal

To create a set of datasets that:

- can be used to understand and find patterns of data access
- can be shared among the WLCG collaboration
- is minimal in size

The 3 datasets that we propose

1. **input_data**: This dataset describes the data accessed by the jobs
2. **analysis_jobs**: Describes the ‘analysis’ jobs that read data
3. **production_jobs**: Describes the ‘production’ jobs that read data

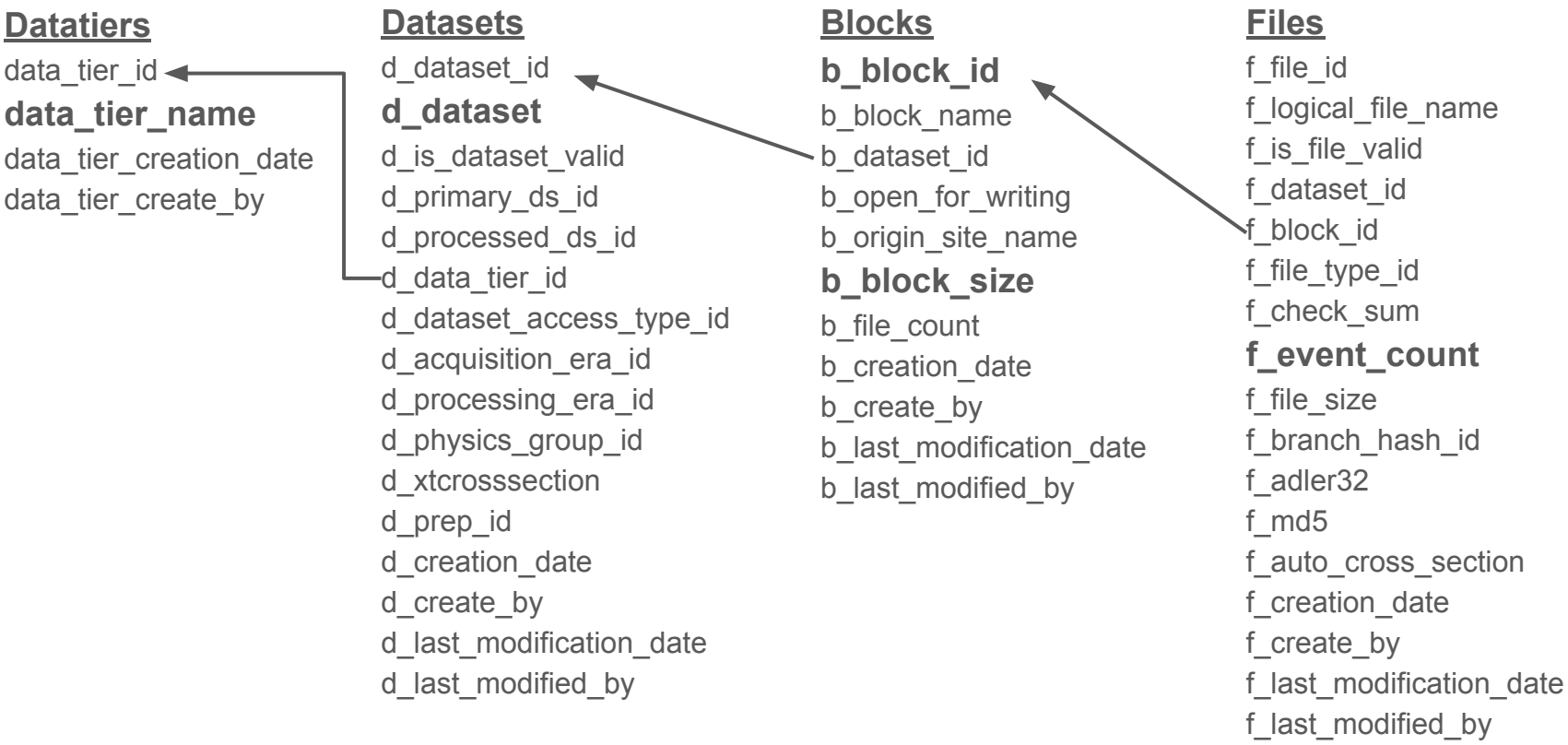
input_data - The source

- The **Data Bookkeeping Service(DBS)** provides **a catalog of event metadata** for Monte Carlo and recorded data of CMS
- Comprises all necessary information for **tracking datasets, their processing history and associations between runs, files and datasets**
- All kind of **data-processing** as well as **physics analysis** done by the users are heavily relying on the information stored in DBS.

input_data - DBS structure

- **Datatier**. 63 different
  AOD, MINIAOD, GEN-SIM
  - **Dataset**. +900K different
    /EGamma/Run2018A-17Sep2018-v2/MINIAOD
    - **Block**. +9.5M different
      /EGamma/Run2018A-17Sep2018-v2/MINIAOD#c460460b-a4ac-454a-ab42-723e6c418826
      - **File**. +135M
        /store/data/Run2018A/EGamma/MINIAOD/17Sep2018-v2/100000/FBEED00E-DB6B-E948-A774-936B3074776A.root

input_data - DBS tables



input_data - The fields that are kept

- **b_block_id**. The unique identifier of a block
- **b_block_size**. Total size of the block in Bytes
- **ds_logical_name.** First part of the dataset name
- **ds_campaign.** Campaign name and second part of the dataset name
- **ds_campaign_sufix.** Third part of the dataset name
- **ds_datatier**. Name of the datatier, the fourth part of the dataset name
- **num_events(*)**. Sumatory of the number of events on all the files belonging to the block

(*) Grouped value

input_data - The final dataset (159 MB)

b_block_id	b_block_size	ds_logical_name	ds_campaign	ds_datatier	num_events
16986783	8.342055e+09	DM_ScalarWH_Mphi-1000_Mchi-450_gSM-1p0_gDM-1p0...	RunIISummer16DR80Premix	AODSIM	27215
19179262	2.158620e+07	JetHT	Run2018D	USER	96183
20551050	2.666179e+08	HeavyNeutrino_trilepton_M-600_V-0p01_tau_NLO_a...	RunIISummer16DR80Premix	AODSIM	800
17630677	2.129800e+09	RelValTenMuExtendedE_200_5000	CMSSW	AODSIM	9000
18271220	3.136090e+08	ZeroBias3	Run2017H	USER	5946445
12108346	2.369607e+09	WminusToMuNu_M-50To250_ew-BMNNP_7TeV-powheg-py...	Summer11Leg	GEN-SIM	4400
14953094	2.226808e+08	QCD_HT1500to2000_GenJets5_TuneCUETP8M1_13TeV-m...	RunIISummer15GS	GEN-SIM	153
15884928	7.041365e+10	DisplacedJet	Run2016B	AOD	326990
16053947	1.260662e+10	QCD_Pt-300to470_MuEnrichedPt5_TuneCUETP8M1_13T...	RunIISummer15GS	GEN-SIM	11223
18800140	2.786129e+09	HcalNZS	Run2018A	MINIAOD	84621

analysis_jobs - The source

- These job records are collected by a monitoring system that queries the **HTCondor infrastructure** of the CMS Global and Tier0 pools.
- The records are sent to an Elasticsearch instance and later on are backed up into HDFS where we read them
- No data structure, just a set of key-value pairs (many of them!)

Notes:

- Jobs that do no read any data are not considered
- Only jobs going through the Global and CERN pools are recorded
- Only jobs submitted by the CRAB system will be taken into account

analysis_jobs - The fields that are kept

- **day**. The day where the job was completed
- **b_block_id**. The id of the block read by the job (the link to ‘input_data’)
- **OverflowType.** The type of read done by the job:
  - FrontedOverflow - Remote from the same region
  - IgnoreLocality - remote from anywhere
  - None - Onsite
- **site_name**. The site where the job ran
- **exitCode**. One of the following: ‘Success’ or type of error: 'Environment' 'Executable', 'Stageout', 'Publication', 'JobWrapper', 'FileOpen', 'FileRead', 'OutOfBounds', 'Other'

analysis_jobs - The fields that are kept (grouped values)

The following fields have aggregated values of all the jobs that share the same values of the previous fields: **day, b_block_id, OverflowType, site_name** and **exitCode**

- **num_jobs**. Number grouped jobs
- **sum_CpuTimeHr**. Sumatory of the CPU time per hour of all grouped jobs
- **sum_CoreHr.** Sumatory of the (WallTime * Cores) of all grouped jobs
  - a job running for 2 hrs using 4 cores will show 8 CoreHr
- **num_users.** Number of different users in the group

analysis_jobs - The final dataset (94 MB)

	day	b_block_id	OverflowType	site_name	exitCode	num_jobs	sum_CpuTimeHr	sum_CoreHr	num_users
2713	1.555459e+09	19082936	FrontendOverflow	T2_US_Vanderbilt	Success	2	32.478056	38.457500	1
7064	1.555373e+09	18669473	IgnoreLocality	T2_US_Caltech	Success	30	9.781944	16.243056	1
3508	1.555114e+09	19631682	None	T1_ES_PIC	Success	15	9.066944	9.775278	2
6710	1.555027e+09	19142549	IgnoreLocality	T2_CH_CERN	Success	13	92.464444	96.866944	1
8662	1.554595e+09	20165222	None	T2_US_Caltech	Other	2	0.000000	0.095833	1
9824	1.554509e+09	19085634	None	T2_US_Florida	Success	65	178.993889	241.549167	3
4279	1.556323e+09	19328189	IgnoreLocality	T2_UK_SGrid_Bristol	Success	12	19.546111	48.425278	1
9733	1.555373e+09	19994702	None	T2_US_Florida	Success	1	0.714722	0.768889	1
8832	1.554077e+09	17318210	FrontendOverflow	T2_US_Vanderbilt	Success	6	9.235833	10.506944	1
9421	1.554336e+09	18273597	IgnoreLocality	T2_UK_London_Brunel	Success	6	27.440833	58.592778	1

production_jobs

In principle this should be very similar to ‘analysis_jobs’ but given the more complicated data processes present in production jobs (i.e. different types of processing jobs) we will have to deal with that at some later time.

Questions?

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

References

[1] Data Bookkeeping Service 3 - Providing event metadata in CMS
https://cds.cern.ch/record/1623287

[2] Job monitoring
https://github.com/dmwm/cms-htcondor-es