Workflow & Data Processing Requirements in 2008

- Computing-physics needs
  - Dataflow
  - Requirements:
    - Reconstruction
    - Stripping
    - Simulation
    - Analysis

Dataflow

Real Data:

RAW data produced via Event Filter Farm at the pit - 280 TB

Categorise: b-exclusive; dimuon; D*; b-inclusive

Simulation:

Simulated data hits produced with GEANT4-based application (Gauss). Hits digitised & spillover added (Boole application)

Format of simulated data are similar to those from DAQ - additional history info. also stored

Dataflow

RAW data is reconstructed:

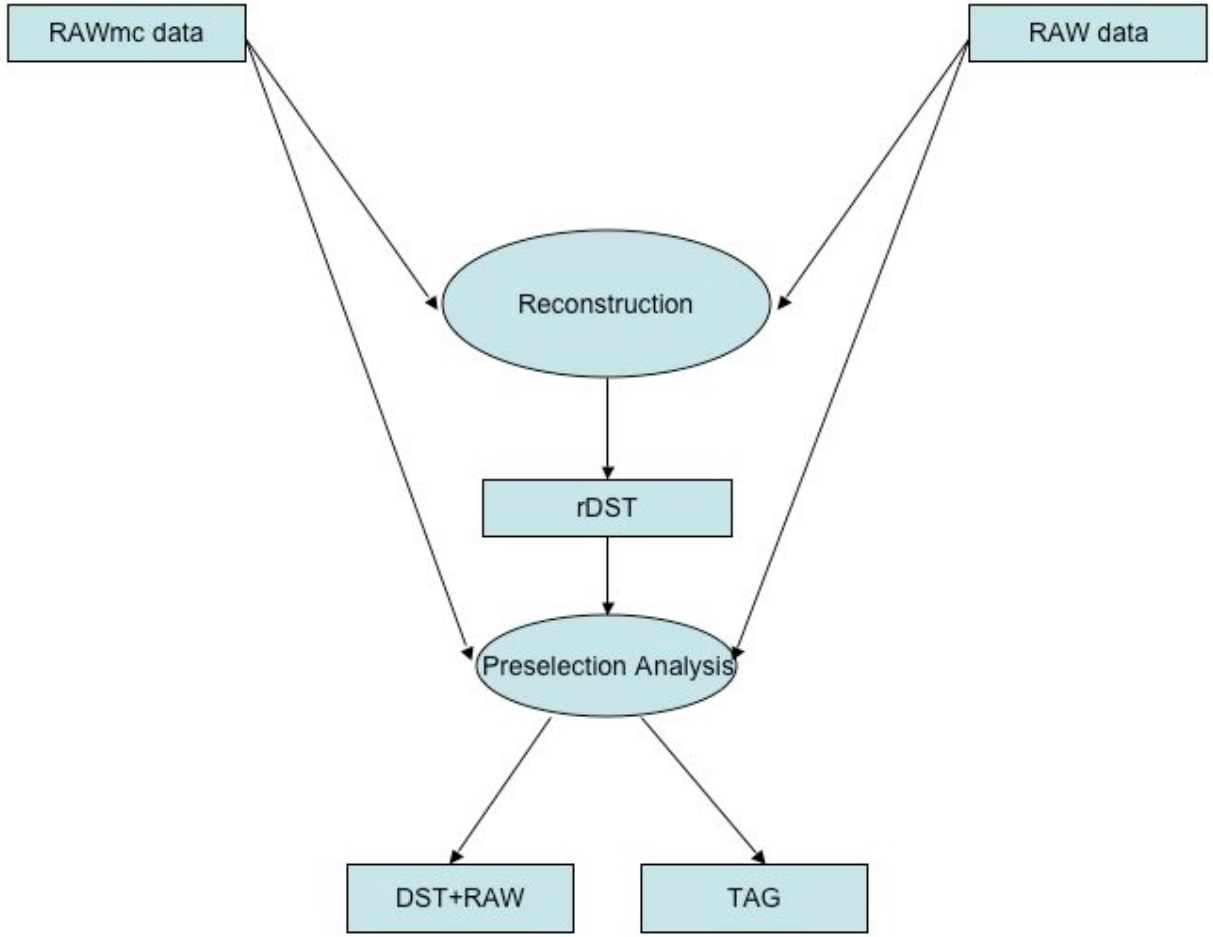
e.g.

Calo. Energy clusters

Particle ID

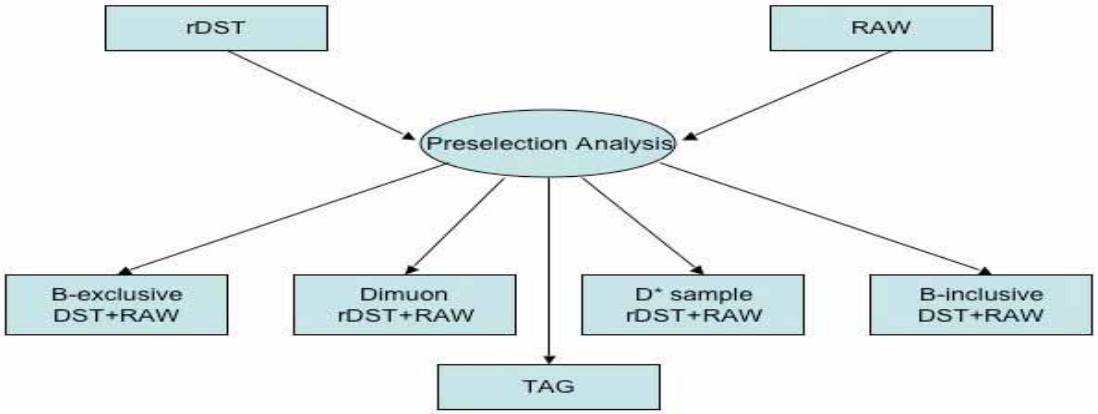
Track momentum

…



At recons time only enough info is stored to allow physics pre-selection to run at a later stage - reduced DST (rDST) - stored separately from RAW

Dataflow - Stripping



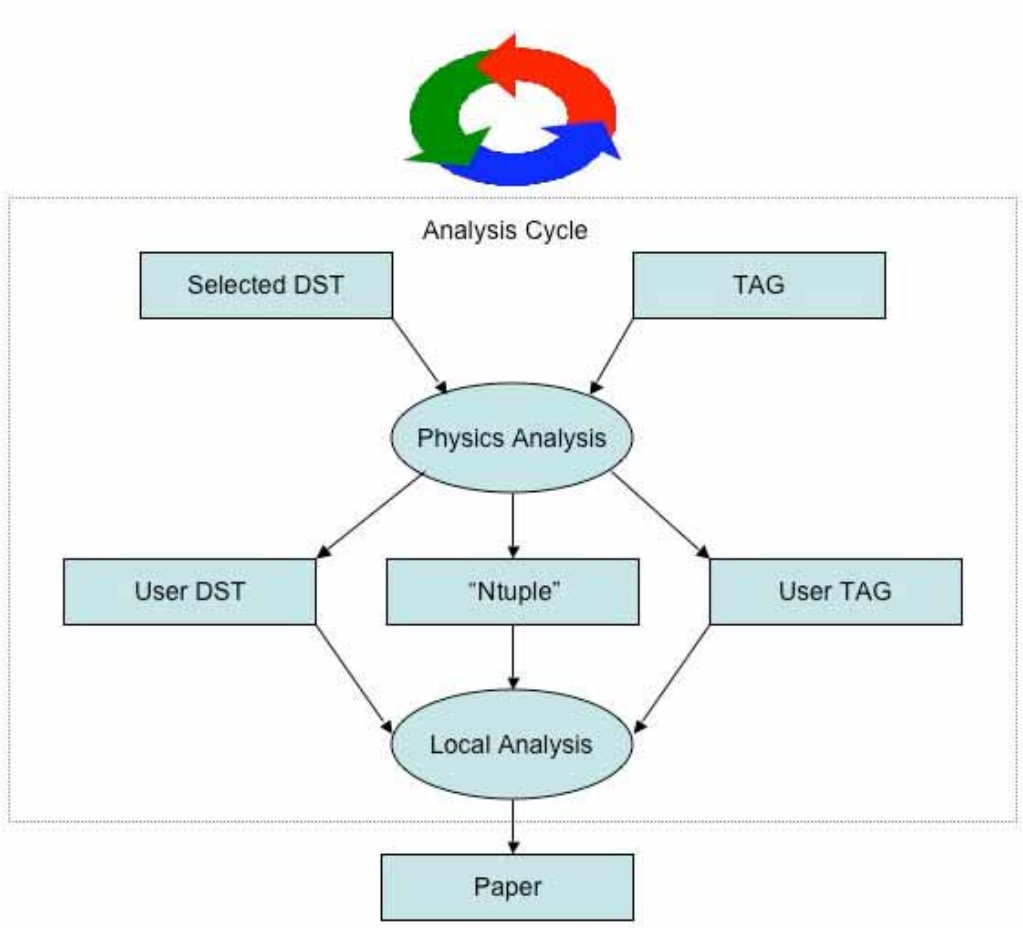
rDST is analysed in production-mode → event streams for further analysis; 4 high level categories (≡ min. 4 streams)

Algorithm developed by physics working groups - use as i/p rDST & RAW

Event to be output will have additional reconstructed info added: (full) DST+ RAW data

Event Tag Collection - created to allow "quick" access to data; contain "metadata"

Dataflow - Analysis



User physics analysis will be primarily performed on the output of the stripping

Output from stripping is self-contained i.e. no need to navigate between files

Analysis generates quasi-private data e.g. Ntuple and/or personal DSTs

Data publicly accessible - enable remote collaboration

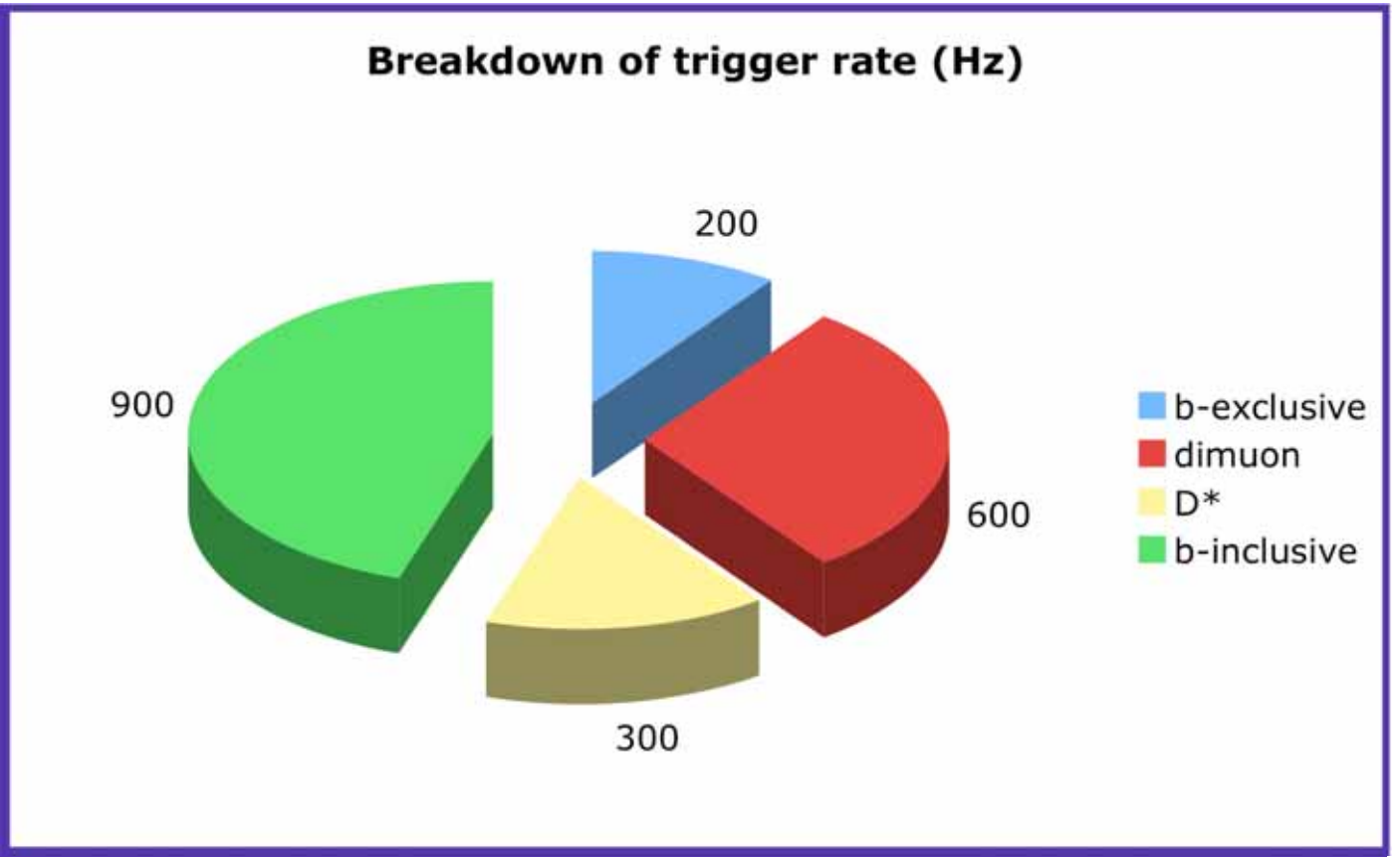
Event Parameters

	TDR estimate	Current estimate
Event Size	kB	
RAW	25	35
rDST	25	20
DST	100	110
Evt processing	kSI2k.s	
Reconstruction	2.4	2.4
Stripping	0.2	0.2
Analysis	0.3	0.3

All sizes corresponds on storage

Implementation of rDST didn't exist at time of TDR

Online



Rates are independent inputs to the model

Reconstruction

	b-exclusive	Dimuon	D*	b-inclusive	Total
Input fraction	0.1	0.3	0.15	0.45	1.0
Number of events	8×10^8	2.4×10^9	1.2×10^9	3.6×10^9	8×10^9
MSS storage (TB)	16	48	24	72	160
CPU (MSI2k.yr)	0.15	0.45	0.23	0.68	1.52

2 passes per year:

1 quasi real time over ~100 day period (2.8 MSI2k)

re-processing over 2 month period of shutdown (4.3 MSI2k)

Make use of Filter Farm at pit (2.2 MSI2k) - data back to the pit!

Stripping

Stripping 4 times per year - 1 month production outside of recons

Stripping has *at least* 4 output streams

Only rDST stored for "non-b" channels+RAW i.e. 55 kB; RAW+full DST for "b" channels* - i.e. 110kB

	Exclusive-b	dimuon	D*	Inclusive-b	Total
Input fraction	0.1	0.3	0.15	0.45	1.00
Reduction factor	10	5	5	100	9.57
Event yield per stripping	8×10^{7}	4.8×10^{8}	2.4×10^{8}	3.6×10^{7}	8.4×10^{9}
CPU (MSI2k.year)	0.02	0.06	0.03	0.02	0.11
Storage requirement per stripping (TB)	9	26	13	4	52
TAG (TB)	1	2	1	4	8

Simulation

- studies to measure performance of detector & event selection in particular regions of phase space
- use large statistics dimuon & D* samples for systematics - reduced Monte Carlo needs

	Application	Nos. of events	CPU time/evt (kSI2k.s)	Total CPU (MSI2k.year)
Signal	Gauss	8×10^8	75	1.9
	Boole	8×10^8	1	0.03
	Brunel	8×10^7	2.4	0.01
Inclusive	Gauss	8×10^8	75	1.9
	Boole	8×10^8	1	0.03
	Brunel	8×10^7	2.4	0.01
Total				3.87

Simulation

- Simulation still dominate LHCb CPU needs
- Current evt size for Monte Carlo DST (with truth info) is ~400kB/evt;
- Total storage needs 160 TB

	Output	Nos. of events	Storage/evt (kB)	Total Storage (TB)
Signal	DST	8×10^7	400	32
	TAG	8×10^7	1	0.1
Inclusive	DST	8×10^7	400	32
	TAG	8×10^7	1	0.1
Total				64

Analysis

- user analysis accounted in model predominantly batch - ~30k jobs/year
- predominantly analysing ~10^6 events
- CPU of 0.3 kSI2k.s/evt
- Analysis needs grow linearly with year in early phase of expt

Nos. of physicist performing analysis	14
Nos. of analysis jobs per physicist/week	4
Event size reduction factor after analysis	5
Number of "active" Ntuples	10
2008 CPU needs (MSI2k.years)	0.31
2008 Disk storage (TB)	80

25% of collaboration submitting analysis jobs

Anticipated some analyses will run toy-Monte Carlo for sensitivity studies

Table is for both data & Monte Carlo analysis

Summary

- CPU needs dominated by simulation needs
- Reconstruction performed of current year's data performed twice
  - quasi-real time
  - 2 month period after data taking
- Production analysis (stripping) to create base analysis sets performed 4 times a year
  - Twice associated with reconstruction
  - Twice in a 1-month period

Summary

Including efficiencies:

Data on tape 2008 (TB)	RAW	rDST	Stripped	Simulation	Analysis
	560	320	483	128	-

Data on disk 2008 (TB)	RAW	rDST	Stripped	Simulation	Analysis
	76	43	775	375	114

CPU needs in 2008 (MSI2k.yr)	Recons.	Stripping	Simulation	Analysis
	1.4	0.5	4.6	0.5