Brainstorming Data Lake Challenge CMS use cases

Frank Wuerthwein

UCSD/SDSC

July 15th 2020

Basic HL-LHC Use Cases

1) (Central) Reconstruction
   - 1.1) RAW to AOD/MINI/NANO
   - 1.2) AOD to MINI
   - 1.3) MINI to NANO
2) (Central) Simulation
   - 2.1) Gen-Sim-Digi-Reco as a single workflow
   - 2.2) Gen-FastSim
3) Data Analysis

Only 1.3) and 3) are relevant for Data Lake model.

Reminder on Scale (I)

- 120 Billion events per year
  - 56Billion events per year of LHC running
  - 64Billion events per year of simulation
- Event sizes:
  - RAW = 6.5MB => 364PB per year (tape only)
  - AOD = 2MB => 240PB per year and version (mostly on tape)
  - MINI = 250kB => 30PB per year and version
  - NANO = 2kB => 240TB per year and version
- Disk space across 7 US T2s in 2020 = 35PB useable

Each lake can have a copy of each relevant NANO version
Most lakes will have a copy of relevant MINI version(s)

Reminder on scale (II)

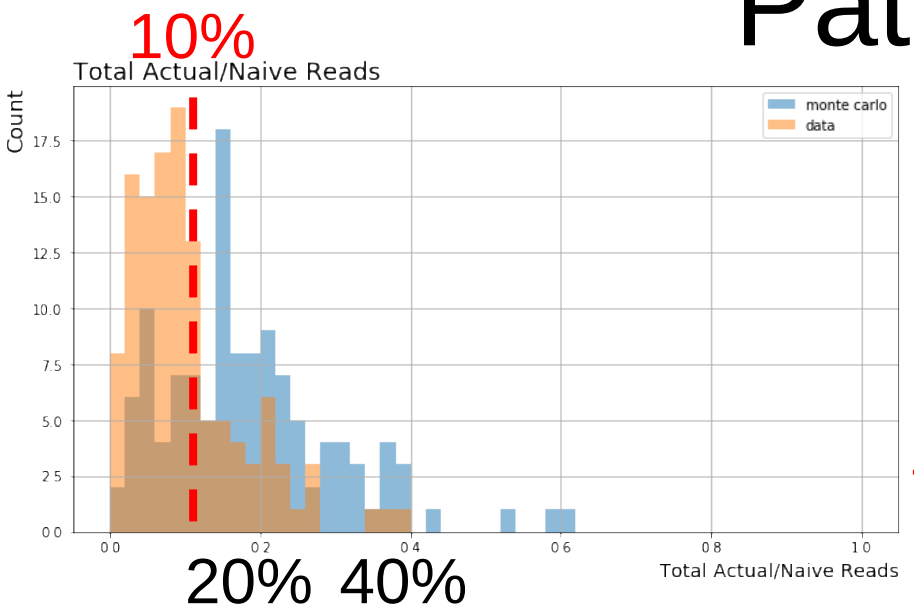
- Processing times per event
  - RAW to AOD ~ 6 kHS06
  - Gridpack to AOD ~ 2 kHS06 (sim&digi) + 6 kHS06 (reco) = 8 kHS06
  - AOD to MINI ~ O(0.1) kHS06
  - MINI to NANO ~ O(0.01) kHS06
  - NANO analysis ~ O(0.0001) kHS06

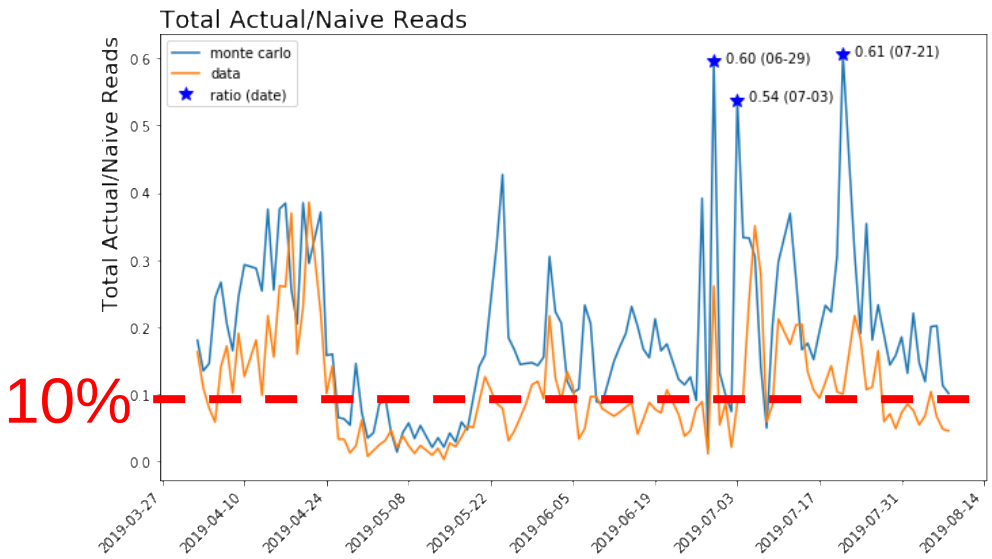
educated guesses

~ 3 orders of magnitude in sizes

~ 8 orders of magnitude in processing times !!!

Reminder on Access Patterns





MINI as measured on Xcache@SoCal

MINI is optimized for remote reads rather than partial file reads.
NANO is optimized for partial file reads.

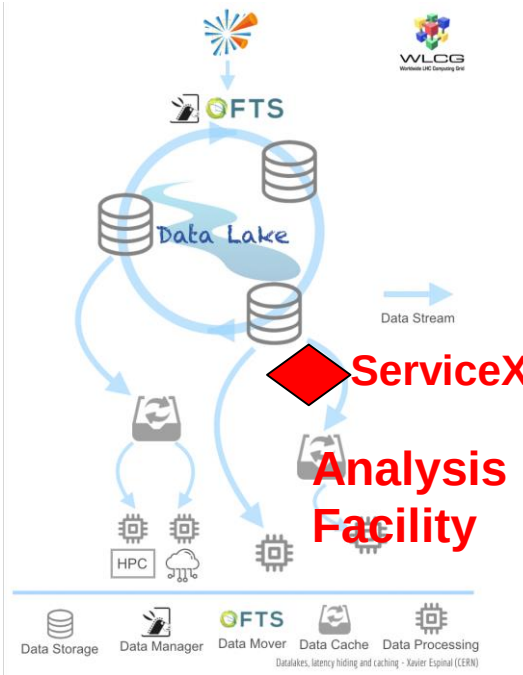
=> Expect NANO to have even stronger partial file read patterns.

Data Analysis

- Yesterday:
  - Make custom user level Ntuple with CRAB
  - Analyze Ntuple on laptop/desktop/workstation.
- Tomorrow:
  - Filter NANO using NANOTOOLS
  - Analyze filtered NANO on laptop/desktop/workstation
- In a few years:
  - Work on “Analysis Facility” (AF)

AF in a few years

- Request data ingest “over night” via ServiceX
  - Filters NANO
  - adds some analysis specific floats
  - Optimizes data layout for columnar analysis in AF
- Request additions to existing custom NANO in AF
  - add info from MINI to custom& filtered NANO
  - Correct “branches” in custom NANO
- Analyze custom NANO at AF interactively or batch.
- Extract custom NANO from AF to laptop.



Lot’s of details to be worked out

Ideas for Data Lake Prototyping Goals

- Benchmark scalability of federated data origins (dCache ?)
  - Input via FTS & Rucio & TPC non-gridftp protocol (HTTPS ?)
    - Replication between lakes
  - Output to Xcache (Rel 5)
    - Reading within lake
  - Balancing internally within the federation
    - Exercise deletions as well as reading & writing
- Benchmark the entire token based authz chain at scale
  - From submission to data access via cache to cache-miss handling to output storing.
- Measure NANOAOD data access patterns for a range of existing analyses.
- Expose architecture discussions
  - What part of software needs knowledge of where what data is?
    - How much does Rucio need to know?
    - How much does submission infrastructure need to know?
  - How does AF integrate with Data Lake ?

Ideas for Data Lake Prototyping Goals

- Benchmark scalability of federated data origins (dCache ?)
  - Input via FTS & Rucio & TPC non-gridftp protocol (HTTPS ?)
    - Replication between lakes
  - Output to Xcache (Rel 5)
    - Reading with Xcache
  - Balancing internally within the federation
    - Exercise deletions as well as reading & writing
- Benchmark the entire token based authz chain at scale
  - From submission to data access via cache to cache-miss handling to output storing.
- Measure NANOAOD data access patterns for a range of existing analyses.
- Expose architecture discussions
  - What part of software needs knowledge of where what data is?
    - How much does Rucio need to know?
    - How much does submission infrastructure need to know?
  - How does AF integrate with Data Lake ?

Are these goals better met with multiple independent pilots in parallel rather than one comprehensive one?

Comments & Questions