

2022 Plans for PHYSLITE columnar data analysis

Nikolai Hartmann

LMU Munich

ATLAS google meeting, January 19, 2022



Plans for columnar data analysis

- Reminder about the dataset:
 - 100 TB `DAOD_PHYSLITE` (future analysis format) with full Run2 data
 - Distributed across 260k files, 18e9 events in total
 - Stored in **ROOT** format, but now also fully converted to **parquet** on google cloud
 - analyzed using python tools (uproot/awkward array)
- Want to run analysis test on all parquet files (done so far with all root files and 10% parquet files)
- Want to try this also with BigQuery (preferably directly loading data from GCS)
- If time allows:
 - Investigate further the scaling limits
 - Can we get around limitations of range requests
 - full file downloads seemed to be still faster than range requests (for ROOT files where we have to read many ranges)