



# Open Data implementation in ALICE: Status & Plan

Stefano Piano  
on behalf of the ALICE Collaboration

# CERN Open Data Policy

- Endorsed by ALICE Collaboration Board in November 2020
- The policy commits to publicly releasing level 3 scientific data:
  - Input to most physics studies (AOD or derived data formats)
  - To be released alongside the software and documentation needed to use the data
  - Allowing high-quality analysis (needed to be also released MC AOD)
- Public data releases expected periodically
- Needed appropriate latency period to allow:
  - Thorough understanding of the data
  - Reconstruction and calibrations
  - The scientific exploitation of the data by the collaboration
- Aim to commence data releases **within five years** of the conclusion of the run period
- Size of the released datasets commensurate with the amount of data collected
- Full datasets will be made available at the end of the collaboration

# ALICE plans for Open Data Implementation

- Set up CERN Open Data Portal with sample of ALICE data:
  - Current status: 5% (7%) of Pb-Pb (pp) 2010 ESD datasets released
  - Preparation of new data format for Run 1 and Run 2 data and simulated data sets:
    - Run 3 AOD or derived data set (nanoAOD)
    - Open Data Quality Control
- (Simple/Run 3) ALICE analysis demonstrator in CERN Open Data portal
  - As already done for Run 1 & 2 AliPhysics Analysis framework:
    - VM and docker container to ease portability
    - Integration with REANA to run analysis directly on Open Data
- Documentation

# New ALICE Open Data Format

- Working to a new AOD format to be published as open data
- Based on the new data format and software framework developed by ALICE O2 project for Run 3 and 4:
  - It will ensure data preservation of Run 1 and Run 2 data
  - Significantly reduced size per collision (factor 16 wrt Run 2 ESD and factor 5 wrt Run 2 AOD)
  - New flat data model optimized for fast IO
  - Possible to adopt (skimmed) derived data set like nanoAOD format to compress further
- Such a refurbishment requires a long conversion production of all Run 1 and Run 2 ESDs and AODs into new AOD format for both data and MC
  - Conversion production is expected to be accomplished by the end of the year
  - Expected to populate CERN Open Data Portal with the Run 1 and 2 data from 2022

# New ALICE Analysis Framework

- The Run 3/4 ALICE Analysis Framework is designed to process the significantly increased amount of data:
  - Wrt Run 2 analysis: factor 3-10 higher event throughput (still being optimized)
  - Streamline data model, trade generality for speed: flatten data structures
  - Recompute quantities on the fly rather than storing them: CPU cycles are cheaper
  - The new analysis model is declarative and imperative:
    - The users specify inputs/outputs and filters
  - Topology, parallelism and rate limiting auto-optimized by Data Processing Layer
- Possible to produce highly targeted derived data (in terms of information needed and selected events of interest):
  - Skimmed derived data sets with higher compression factor wrt AODs
  - We expect the usage of skimmed data to bring additional x10 speed-up

# Expected Open Data release in the next 5 years

- Based on current estimates for the new AOD format, to release Run 1 and Run 2 pp and Pb-Pb data we expect roughly:
  - 2022: 20 TB
  - 2023: 100 TB
  - 2024: 200 TB
  - 2025: 300 TB
- In Run 2 ALICE inspected  $\sim 1 \text{ nb}^{-1}$  Pb-Pb data, while for Run 3 & 4 ALICE plans to collect  $\sim 13 \text{ nb}^{-1}$  of Pb-Pb collisions:
  - In 2029 we will start releasing Run 3 data:
    - $\sim 2 \text{ PB/year}$  publishing AODs  $\Rightarrow \sim 1 \text{ PB/year}$  publishing skimmed derived data sets (under the assumption that the skimmed derived data sets are 2x more compressed wrt Run 3 AOD format)

# Summary & Outlook

- Up to now most of ALICE efforts focused on the software upgrade for Run 3
- ALICE Open Data will benefit from the new Analysis Framework improvements:
  - New Run 3 AOD format suitable for Run 1 and 2 Open Data
  - Skimmed derived data sets will provide
    - Further data compression crucial to make public Run 3 data
    - Higher event throughput and reduction of needed CPU wall time
- Dedicated human resources for the ALICE Open Data implementation will be allocated soon:
  - ALICE committed to publicly releasing level 3 scientific data
  - Organization of new ALICE Open Data working group
  - ALICE aims at making public Run 1 and 2 Open Data from 2022