Open Data implementation in ALICE: Status & Plan

David Dobrigkeit Chinellato,
Alexandru Florin Dobrin, Stefano Piano
on behalf of ALICE

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

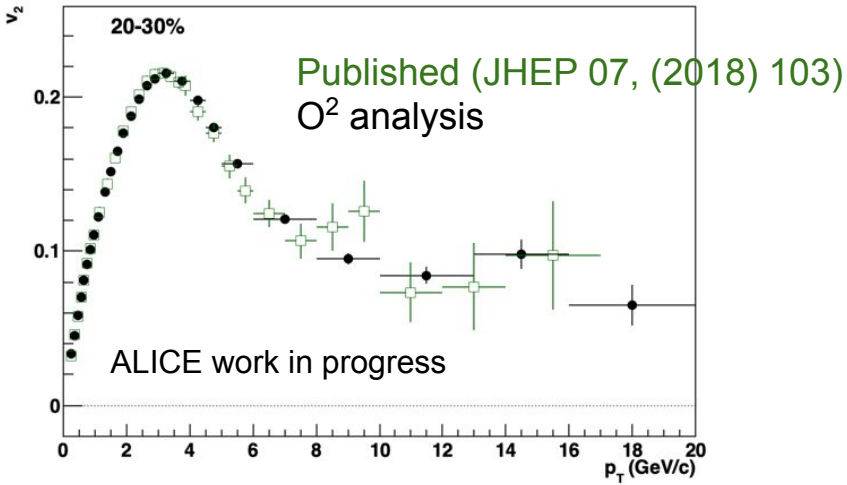
- Working to a new AOD format to be published as open data
- Based on the new data format and software framework developed by ALICE O^2 project for Run 3 and 4:
  - It will ensure data preservation of Run 1 and Run 2 data
  - Significantly reduced size per collision (factor 16 wrt Run 2 ESD and factor 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

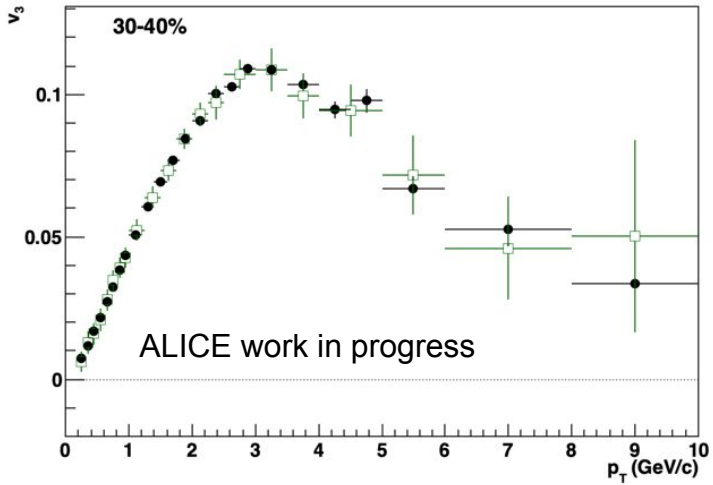
- Such a refurbishment required a long conversion production of all Run 1 and Run 2 ESDs and AODs into new AOD format for both data and MC
  - Conversion was done in 2022, final validation ongoing
  - Expected to populate CERN Open Data Portal with the Run 1 data in the next months

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: 10x higher event throughput confirmed, optimization ongoing
  - 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

Lightweight analysis test: flow measurement





- v_n coefficients measured using the scalar product method
- Good agreement with published results
- Lightweight analysis framework being finalised

$$v_n = \frac{\langle u^{\eta<-0.5} Q_n^{\eta>0.5} / M_Q^{\eta>0.5} \rangle}{\sqrt{\langle Q_n^{\eta<-0.5} Q_n^{\eta>0.5} / M_Q^{\eta<-0.5} / M_Q^{\eta>0.5} \rangle}}$$

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:
  - 2023 35 TB (50% Run 1)
  - 2024 105 TB (10% Run 2)
  - 2025 - 2028 105 TB/year to reach 50% of Run 2 in 2028

- In Run 2 ALICE inspected ~1 nb^{-1} Pb-Pb data, while for Run 3 & Run 4 ALICE plans to collect 13 nb^{-1} of Pb-Pb collisions:
  - In 2029 we will start releasing Run 3 data:
    - ~2 PB/year publishing AODs ⇒ ~1 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 Run 2 Open Data
    - Conversion of Run 1+2 already done!
    - Open data software framework + documentation: work in progress
- Dedicated human resources for the ALICE Open Data implementation has been allocated:
  - ALICE committed to publicly releasing level 3 scientific data
  - ALICE aims at making public Run 1 and Run 2 Open Data from 2023