Offline QC data sets for ML in ALICE

Marian Ivanov
Jens Wiechula

ML in QA Workshop – 3./4. Dec 2018

Outline

- Considerations on input sources
- Types of sources
- Origin of sources
- Usage examples
- Data sizes (TPC)
- Summary

Considerations on input sources

- Provide homogeneous access to heterogeneous sources
  - Data bases (e.g. logbook, MonALISA, DCS, …)
  - Root files (QA, calibration, …)
  - Text files (logs, …)

    → Easy correlation of different sources
- Reduce load on experts
  - Automation of data extraction (no time costly manual intervention for each new data input)

    → Faster feedback → Trouble shooting of problems in the data quality
  - Allow experts to spend more time on important tasks
- Standardized validated data sources
  - Automation of plotting
  - Automation of decisions on data quality
  - Automation of alarm levels
  - Enable non-experts to make high-level data analysis

Relevant input sources

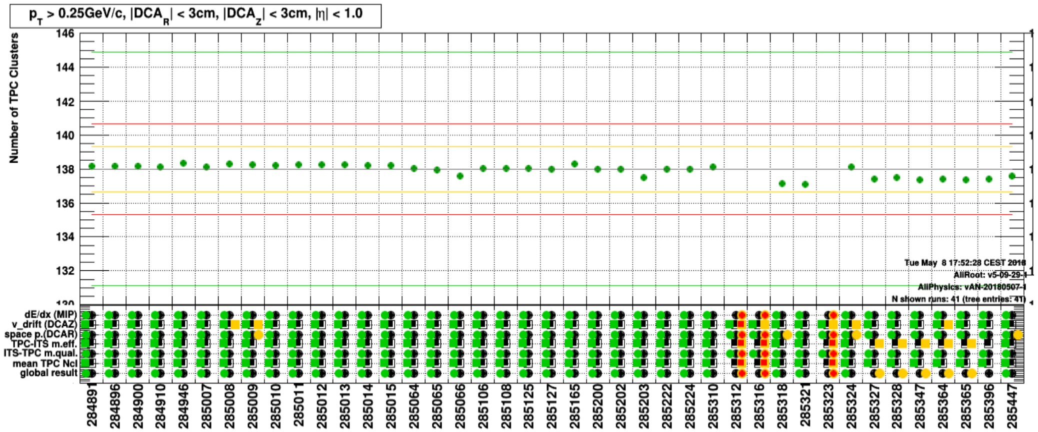
- Offline/Online QA (root files) – automatized
- Calibration (OCDB – root files, partially contains DCS information via the shuttle) – automatized
- Logbook (SQL) – automatized
- MonALISA (web interface) – automatized
- DCS (Oracle – DARMA interface) – manual
  - Scriptable access would be very welcome
- Custom local analysis → used by experts e.g. during development
- Logfile summaries (text files) → not yet in production
- High level user analysis → proposed but not used

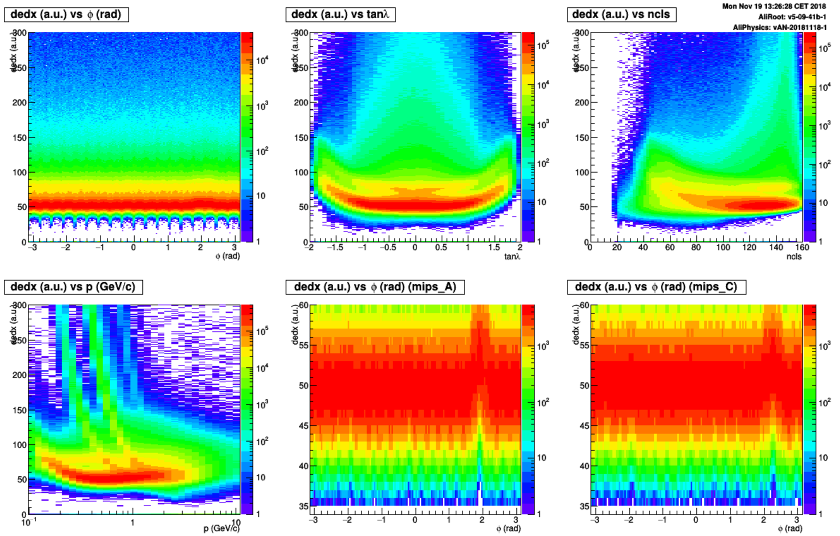
Offline QA procedure

- CPass0 / CPass1 / PPass automatically triggered after run finished
- QA analysis automatically started after CPass1 / PPass → QAresults(_barrel).root
  - (Histogrammed) results of all detectors
- QA postprocessing run as cron job, checking for new QAresults(_barrel).root available
  - Produce typical QA per run plots
  - Produce run and period summary (trending) information
  - Produce QA web pages
  - Trigger (automatic) alarms

Standard QA visualisation examples

- *trending.root* trees **automatically** created during the QA
- Run centrally at CERN
- **Used by** (almost) **all detectors**

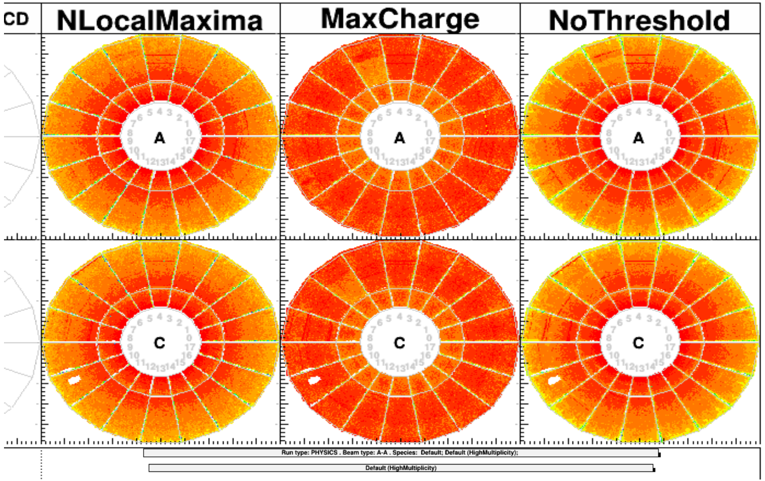


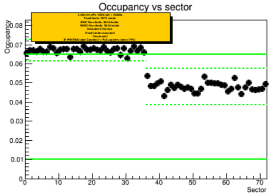


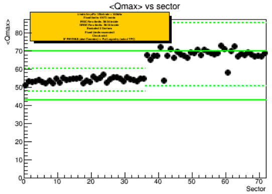
http://aliqa[DET].web.cern.ch/aliqa[DET]/{data,sim}/[year]/[period]/[pass]/trending.root

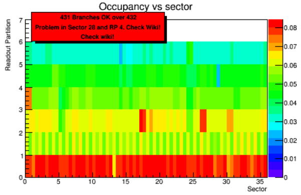
Online QA procedure – TPC example

- Process raw data in DAs
- Run simple cluster finding
- Store average cluster properties per readout pad in OCDB
- Provide projections to the online QA shifter via Amore









OCDB trending

- Automatically go through all relevant OCDB objects and produce summary trees
  - Calibration data
  - DCS data (via shuttle)
  - Online DA output (Pedestals, simple cluster finder, ...)
- Private cron job run at GSI
  - Output copied to QA web pages
- Done for TPC since run 1, activities for TRD started

Logbook and MonALISA

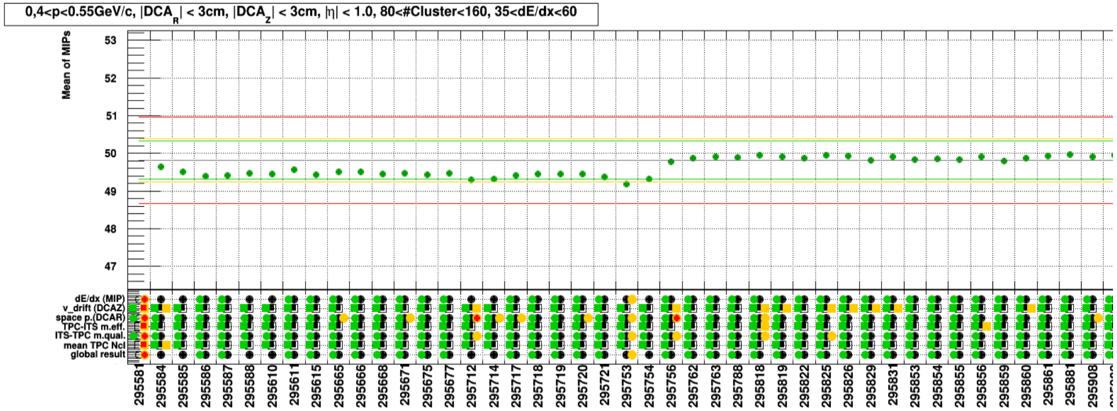
- Logbook
  - Root trees created in cron jobs, published on QA web site
  - e.g. data rates, data volumes, trigger information, …
- MonALISA
  - Root trees created via web page interface
  - e.g. production information (data, MC)

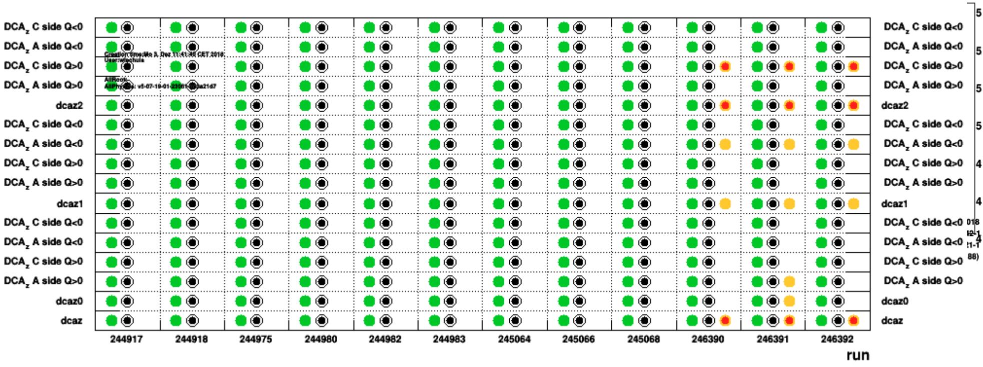
Usage examples: Combining data sources

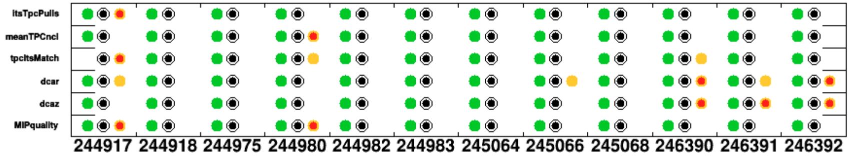
```
AliExternalInfo externalInfo;
treeQA = externalInfo->GetTree("QA.TPC", period, pass,
"QA.TPC;QA.TRD;QA.TOF;QA.ITS;Logbook;QA.rawTPC");
```

- Interface to the most relevant data source predefine in the default configuration file of the AliExternalInfo.cfg
  - Automatic connect between different sources
  - Easy to add new sources
  - Possibility to overwrite default locations
    - E.g. for development purposes

Usage examples: Default trending plots



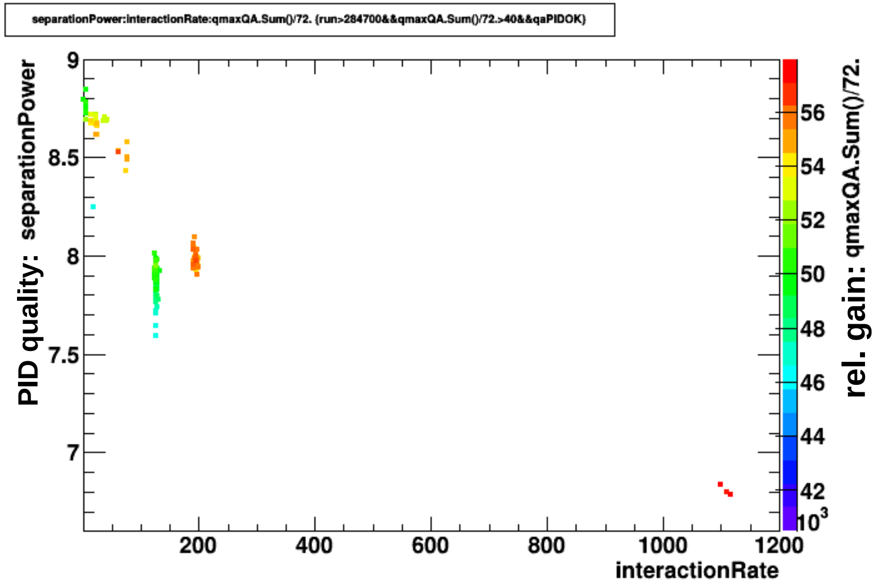




- Standardized tool (AliTreeTrending) to create simple trending plots and status matrices
- Manually define limits for warning, outlier, physics acceptable
- Configuration of status information

More details can be found here

Usage examples: Expert QA



```
Materialized view - usage example:
AliExternalInfo info;
treeCalib = info.GetChain("QA.rawTPC","LHC18*",
            "cpass1_pass1","QA.TPC;QA.EVS;Logbook”);
treeCalib->SetAlias("separationPower","2*(meanMIPele-meanMIP)/
          (resolutionMIP*meanMIP+resolutionMIPele*meanMIPele)");
treeCalib->SetAlias("qaPIDOK","resolutionMIP>0&&resolutionMIPele>0");
treeCalib->Draw("separationPower:interactionRate:qmaxQA.Sum()/72.",
                "run>284700&&qmaxQA.Sum()/72.>30&&qaPIDOK","colz")
```

Simple access to sources

- No manual interaction needed → trending info created automatically

Connection of several sources

- Pid quality (QA.TPC)
- interaction rate (QA.EVS)
- gain (QA.rawTPC)
  - qmaxQA (OCDB – Amore)
- B field (Logbook)

Data sizes in the present system

- QA (TPC, all years, all periods, all passes): 2G
- OCDB trending (TPC, all years, all periods, one pass): ~600MB
- Logbook (All years): 100MB
- MonALISA (All years, all periods, all passes): <100MB
- DCS currents (TPC + TRD): e.g. LHC15o ~50MB (relevant summary data – averages etc.), ~3GB (raw data)

Summary

- QA relevant information are gathered from many (heterogeneous) sources
- Gathered run (or sub-run wise) in root trees
  - Common format for easy connects
  - Most physicists familiar with usage
- Provide common interfaces for combining and plotting the information
- Output used for
  - QA website generation
  - “Automatic” QA decision
  - Expert queries
  - Input for ML techniques