ATLAS software tools to handle ROOT RNTuple

Mete, A. S.[1] Nowak, M.[2] Rybkin, G.[3] Van Gemmeren, P.[1]
on behalf of the ATLAS Computing Activity

[1]Argonne National Laboratory (US), [2] Brookhaven National Laboratory (US),
[3] Université Paris-Saclay, CNRS/IN2P3, IJCLab (FR)

Abstract

The software of the ATLAS experiment [1] at the CERN LHC accelerator contains a number of tools to inspect (validate, summarize, peek into etc.) all its official data formats recorded in ROOT files. These tools [2] — mainly written in the Python programming language — handle the ROOT TTree which is currently the main storage object format of ROOT files. However, the ROOT project [3] has developed an alternative to TTree, called RNTuple [4]. The new storage format offers significant improvements and ATLAS plans to adopt it in LHC Run 4. Work is ongoing to enhance the tools in order to handle the RNTuple storage format in addition to TTree in a transparent for the user way. The work is aided by modern and detailed APIs provided by RNTuple. We will present the progress made and lessons learnt.

Validate event by event

1. read/load all fields of an entry
2. repeat for each entry/event

```
reader=RNTupleReader.Open(ntuple)
for i in reader:
  reader.LoadEntry(i)
```

```
DEBUG Checking ntuple of key EventData ...
DEBUG   contains 10 events
DEBUG Checking 10 entries ...
DEBUG NTuple of key EventData looks ok.
```

Summarize data content

1. use RNTupleInspector, RNTupleDescriptor, RFieldDescriptor API to obtain on disk and in memory size of objects

Snippet of PoolFile.py

```
inspector = RNTupleInspector.Create(ntuple)
descriptor = inspector.GetDescriptor()
fieldZeroId = descriptor.GetFieldZeroId()
for fieldDescriptor in descriptor.GetFieldIterable(fieldZeroId):
  fieldId = fieldDescriptor.GetId()
  fieldTreeInspector = inspector.GetFieldTreeInspector(fieldId)
  diskSize = fieldTreeInspector.GetCompressedSize()
  memSize = fieldTreeInspector.GetUncompressedSize()
  typeName = fieldDescriptor.GetTypeName()
  fieldName = fieldDescriptor.GetFieldName()
```

2. summarize sizes by data object type and category

Validate field by field

1. for a top level field bulk read all values belonging to a cluster
2. repeat for each cluster
3. repeat the above for each top level field

Snippet of trfValidateRootFile.py

```
reader=RNTupleReader.Open(ntuple)
descriptor = reader.GetDescriptor()

model = reader.GetModel()
fieldZero = model.GetFieldZero()
subFields = fieldZero.GetSubFields()
for field in subFields:
  bulk = model.CreateBulk(field.GetFieldName())

  for clusterDescriptor in descriptor.GetClusterIterable():
    clusterIndex = RClusterIndex(clusterDescriptor.GetId(), 0)
    size = int(clusterDescriptor.GetNEntries())
    maskReq = array('b', (True for i in range(size)))
    values = bulk.ReadBulk(clusterIndex, maskReq, size)
```

```
DEBUG Checking ntuple of key EventData ...
DEBUG   contains 10 events
DEBUG ntupleName=EventData
DEBUG Top level fields number 863
DEBUG fieldName=index_ref typeName=std::uint64_t
DEBUG     cluster #0 firstEntryIndex=0 nEntries=10
DEBUG         values array at <cppyy.LowLevelView object at 0x7ffa97df1fb0>
DEBUG fieldName=xTrigDecisionAux: typeName=xAOD::TrigDecisionAuxInfo_v1
DEBUG     cluster #0 firstEntryIndex=0 nEntries=10
DEBUG         values array at <cppyy.LowLevelView object at 0x7ffa979884b0>
DEBUG fieldName=METAssoc_AntiKt4EMPFlowAux: typeName=xAOD::↘
 ↳MissingETAuxAssociationMap_v2
...
DEBUG fieldName=DiTauJetsAux::subjet_f_core typeName=std::vector<std::↘
 ↳vector<float>>
DEBUG     cluster #0 firstEntryIndex=0 nEntries=10
DEBUG         values array at <cppyy.LowLevelView object at 0x7ffa901e0d30>
DEBUG NTuple of key EventData looks ok.
```

References

1. Collaboration ATLAS (2008) The ATLAS experiment at the CERN Large Hadron Collider. *JINST* 3:S08003. https://doi.org/10.1088/1742-6596/331/7/072034
2. ATLAS Software: https://gitlab.cern.ch/atlas/athena
3. ROOT project: https://root.cern
4. RNTuple Format Specification: https://github.com/root-project/root/blob/master/tree/ntuple/v7/doc/specifications.md

Data content summary

Snippet of checkxAOD.py output

```
================================================================================
       Event data
================================================================================
   Mem Size        Disk Size        Size/Evt      Compression  Items  Container Name (Type)
--------------------------------------------------------------------------------
   195.312 kb         0.082 kb         0.000 kb     2380.952   200000  IsSimulation (bool) [EvtId]
   195.312 kb         0.082 kb         0.000 kb     2380.952   200000  IsCalibration (bool) [EvtId]
  ...
102751.795 kb     10447.442 kb         0.052 kb        9.835   200000  AnalysisElectrons (DataVector<xAOD::Electron_v1>) [AnalysisElectrons]
 44689.328 kb     12590.621 kb         0.063 kb        3.549   200000  GSFTrackParticles (DataVector<xAOD::TrackParticle_v1>) [egamma]
 84394.000 kb     13683.679 kb         0.068 kb        6.167   200000  TruthNeutrinos (DataVector<xAOD::TruthParticle_v1>) [Truth]
  ...
347359.531 kb     81513.889 kb         0.408 kb        4.261   200000  HardScatterParticles (DataVector<xAOD::TruthParticle_v1>) [Truth]
198828.125 kb    108201.033 kb         0.541 kb        1.838   200000  EventInfo (xAOD::EventInfo_v1) [EvtId]
250332.719 kb    133026.684 kb         0.665 kb        1.882   200000  HLTNav_RepackedFeatures_Particle (DataVector<xAOD::Particle_v1>) [Trig]
653769.254 kb    298982.955 kb         1.495 kb        5.531    22825  DataHeaderForm (DataHeaderForm_p6) [MetaData]
251935.105 kb    401949.173 kb         2.010 kb        5.603   200000  AnalysisJets (DataVector<xAOD::Jet_v1>) [AnalysisJets]
814643.835 kb    403361.655 kb         2.017 kb        2.020   200000  InDetTrackParticles (DataVector<xAOD::TrackParticle_v1>) [InDet]
  ...
```

Encountered issues

- it turned out that sometimes seemingly bogus numbers were reported for RNTuple, e.g.,

```
9389707.954 kb 18014398508157516.000 kb 90071992540.788 kb        0.000   200000  HLTNav_Summary_DAODSlimmed (DataVector<xAOD::TrigComposite_v1>) [Trig]
```

- after investigation, an issue and a pull request to fix it were opened
- with the pull request accepted, the problematic line becomes

```
9389707.954 kb  2869832.889 kb       14.349 kb        3.272   200000  HLTNav_Summary_DAODSlimmed (DataVector<xAOD::TrigComposite_v1>) [Trig]
```

Grigori.Rybkine at cern.ch