# ATLAS RNTuple Update

**Alaettin Serhan Mete**
*Argonne National Laboratory*

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

**HEP-CCE-SOP Meeting**
10/16/2024

Argonne NATIONAL LABORATORY

# Impact of Split vs Unsplit Columns

- **100k MC ($t\bar{t}$) events in DAOD PHYSLITE format:**
  - Jonas Hahnfeld: Splitting of columns (byte shuffling) can have negative impact on compression
  - A single field (`vector<vector<uint>>`): **60 MB (split) → 22.5 MB (unsplit), i.e. 60+% reduction**

```
*.....................................................................*
*Br  515 :HLTNav_Summary_DAODSlimmedAuxDyn.decisions :                 *
*         | vector<vector<unsigned int> >                            *
*Entries :   100000 : Total  Size=  239180609 bytes  File Size  =   35748601 *
*Baskets :     1980 : Basket Size=     131072 bytes  Compression=   6.69     *
*.....................................................................*
```

```
.............................................................
  HLTNav_Summary_DAODSlimmedAuxDyn:decisions [#0]  --  SplitIndex64{id:766}
    # Elements:          100000
    # Pages:             21
    Avg elements / page: 4761
    Avg page size:       4272 B
    Size on storage:     89723 B
    Compression:         8.92
.............................................................
  HLTNav_Summary_DAODSlimmedAuxDyn:decisions._0 [#0]  --  SplitIndex64{id:767}
    # Elements:          6154734
    # Pages:             170
    Avg elements / page: 36204
    Avg page size:       12723 B
    Size on storage:     2162926 B
    Compression:         22.76
.............................................................
  HLTNav_Summary_DAODSlimmedAuxDyn:decisions._0._0 [#0]  --  SplitUInt32{id:768}
    # Elements:          53217539
    # Pages:             636
    Avg elements / page: 83675
    Avg page size:       94475 B
    Size on storage:     60086325 B
    Compression:         3.54
.............................................................
```

**Default (Split)**

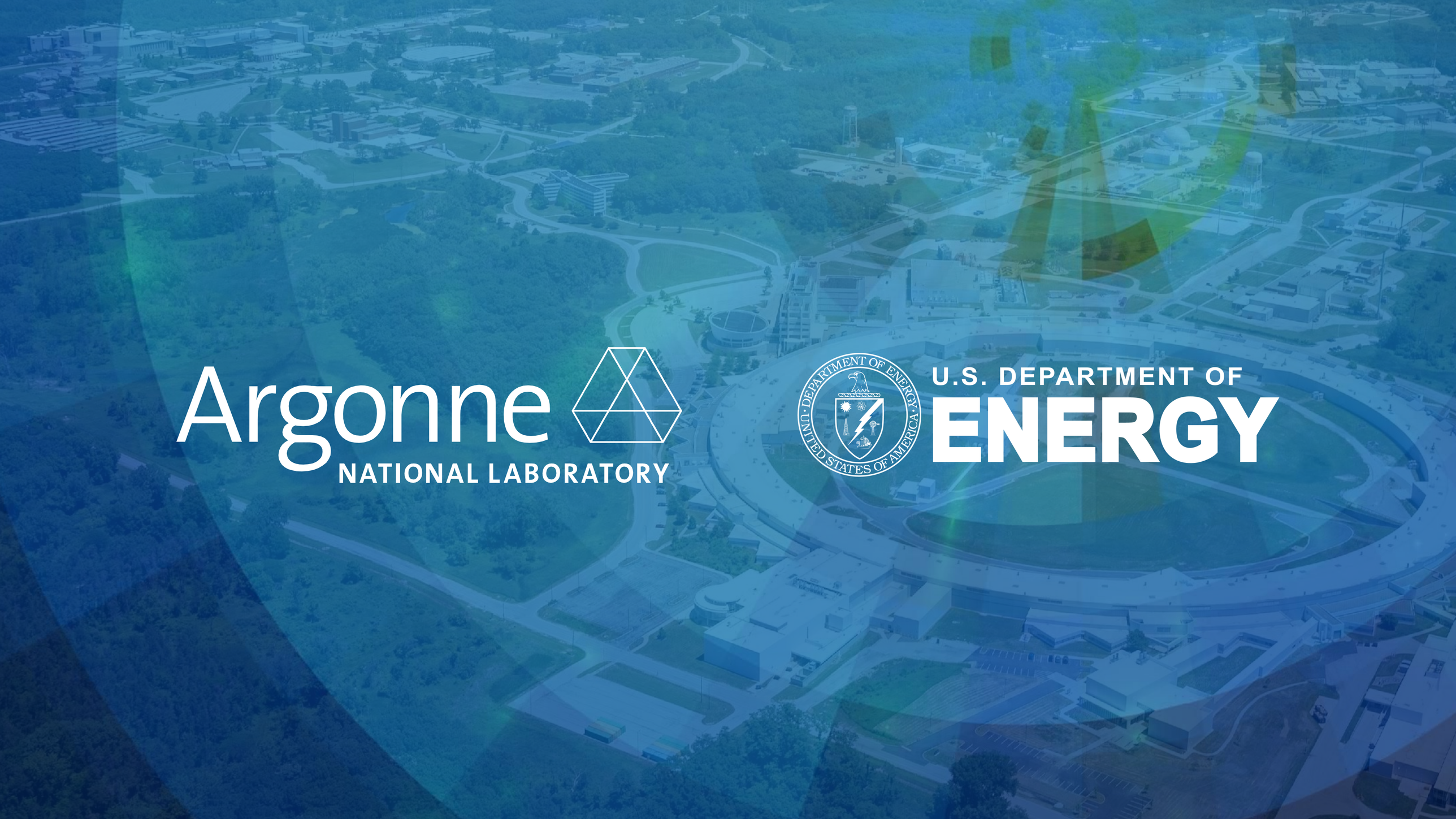
```
.............................................................
  HLTNav_Summary_DAODSlimmedAuxDyn:decisions [#0]  --  SplitIndex64{id:766}
    # Elements:          100000
    # Pages:             20
    Avg elements / page: 5000
    Avg page size:       4480 B
    Size on storage:     89608 B
    Compression:         8.93
.............................................................
  HLTNav_Summary_DAODSlimmedAuxDyn:decisions._0 [#0]  --  SplitIndex64{id:767}
    # Elements:          6154734
    # Pages:             174
    Avg elements / page: 35372
    Avg page size:       12436 B
    Size on storage:     2164008 B
    Compression:         22.75
.............................................................
  HLTNav_Summary_DAODSlimmedAuxDyn:decisions._0._0 [#0]  --  UInt32{id:768}
    # Elements:          53217539
    # Pages:             661
    Avg elements / page: 80510
    Avg page size:       34064 B
    Size on storage:     22516354 B
    Compression:         9.45
.............................................................
```

**Unsplit**

# Impact of Split vs Unsplit Columns

- **Some practical considerations:**
  - Each field/column can be configured separately, which is nice!
  - This has to be done **before** the first write
    - How about model extensions?
  - Is there a nicer way to do this per column?
    - For example, only unsplit the innermost data
    - For this `vector<vector<...>>` case, I ended up doing:

```
   auto field = RFieldBase::Create(field_name, field_type).Unwrap();
+  if (field_name.find("decisions")!=std::string::npos) {
+    ((field->GetSubFields()[0])->GetSubFields()[0])->SetColumnRepresentatives({{ROOT::Experimental::EColumnType::kUInt32}});
+  }
```


Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT OF
ENERGY