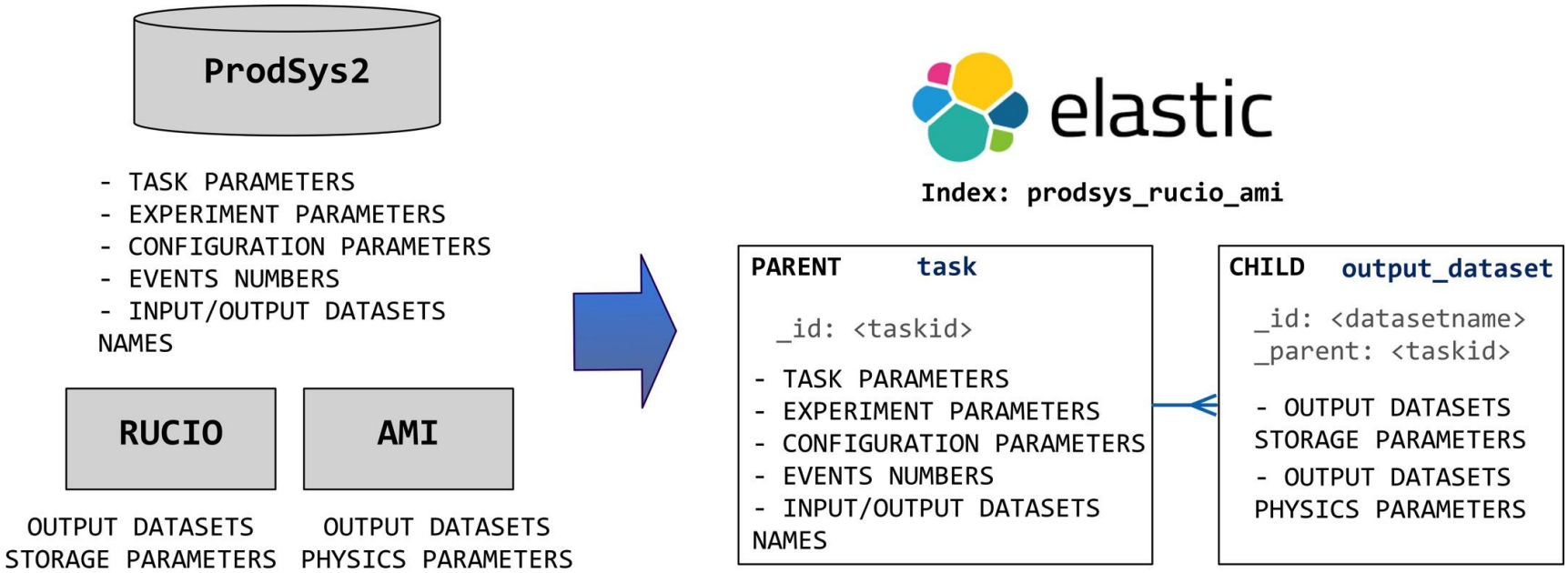
DKB/DCC STATUS REPORT

Grigorieva M., Golosova M., Borodin M.

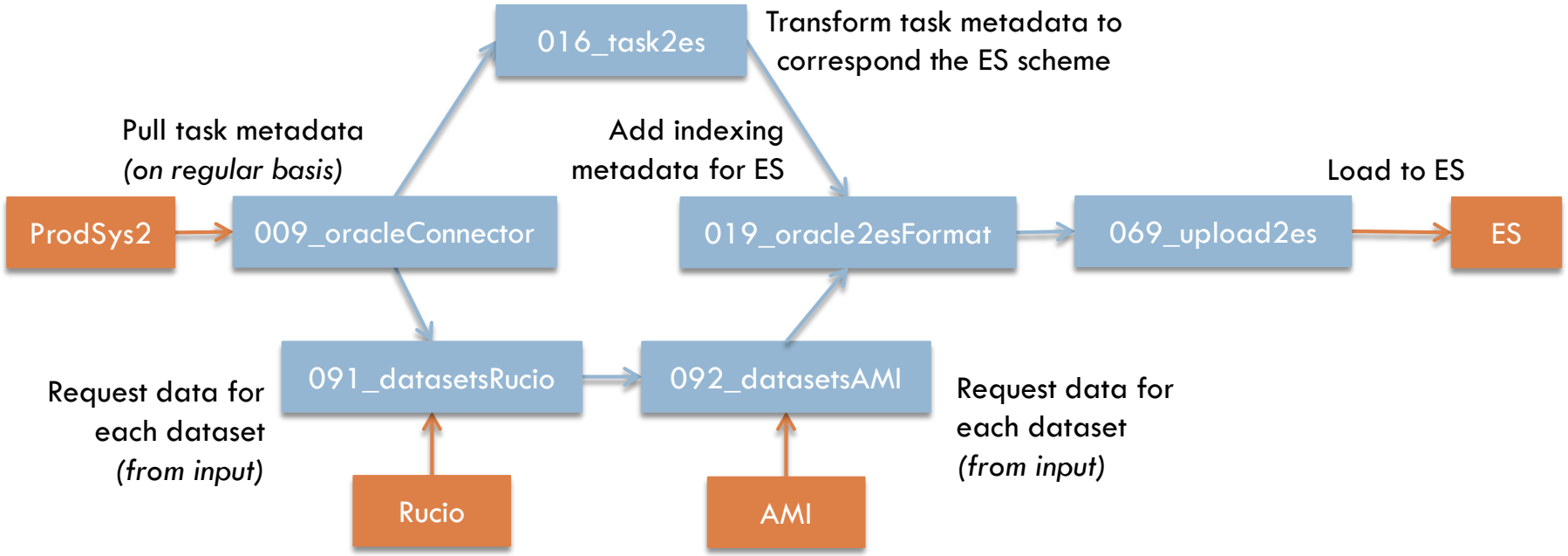
DKB/DCC Outline

- Developed the following data processing scripts:
  - Extract metadata from the initial data sources:
    - ProdSys2 (bulk (timestamp-based) extraction)
    - AMI (by dataset name)
    - Rucio (by dataset name)
  - Preparation of the metadata to import into the ElasticSearch
    - adding the indexing meta information to the records
  - Bulk ES import procedure
  - Supervising script that chains the scripts, extracting metadata from ProdSys2, AMI and Rucio together
- Metadata from ProdSys2, Rucio, AMI was integrated and imported to the ElasticSearch Storage
- Data update is possible only basing on timestamp parameter from ProdSys2

DKB/DCC Data Model



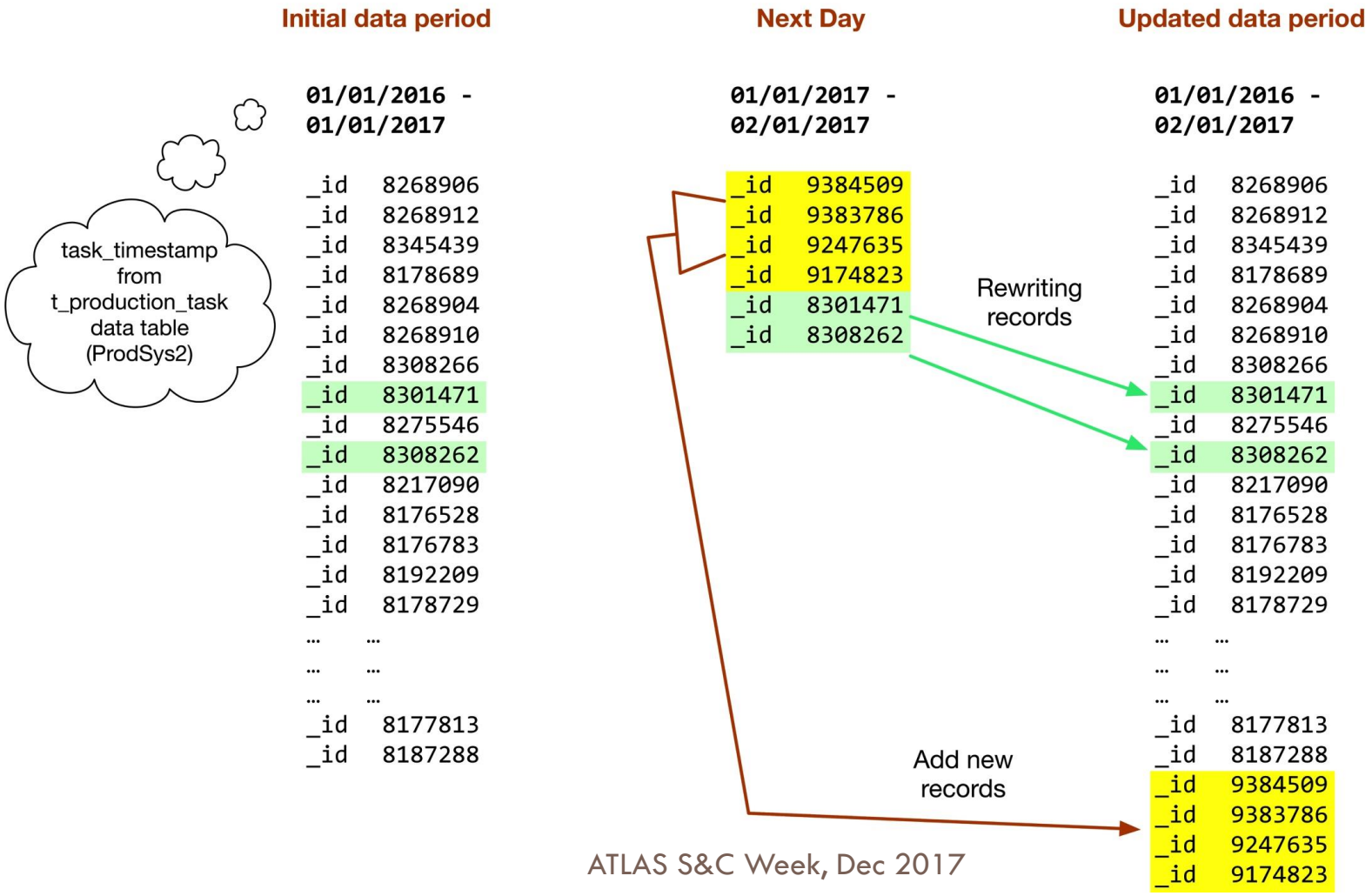
Data Processing Pipeline



Supervising script that chains, extracting data from ProdSys2, AMI and Rucio together, organizing a continuous pipelines to ES:

- ProdSys2 records with task metadata are appended with categorizartion (physics category) information and go to ES;
- in parallel, from task metadata are taken output dataset names to query data from Rucio and AMI;
- information from AMI is appended to the data from Rucio;
- records with data from both AMI and Rucio go to ES.

Data Synchronization

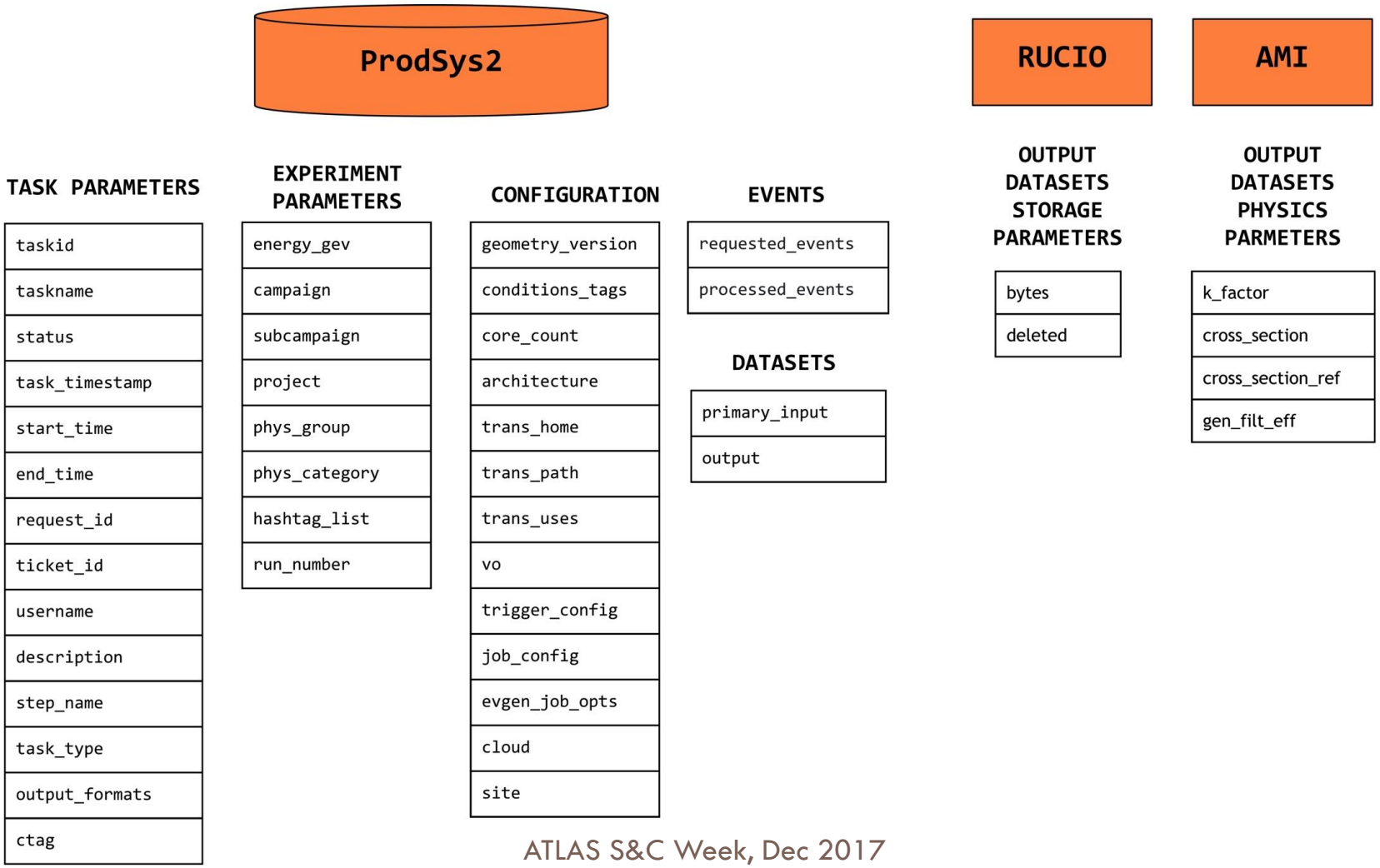


DKB/DCC Conclusions

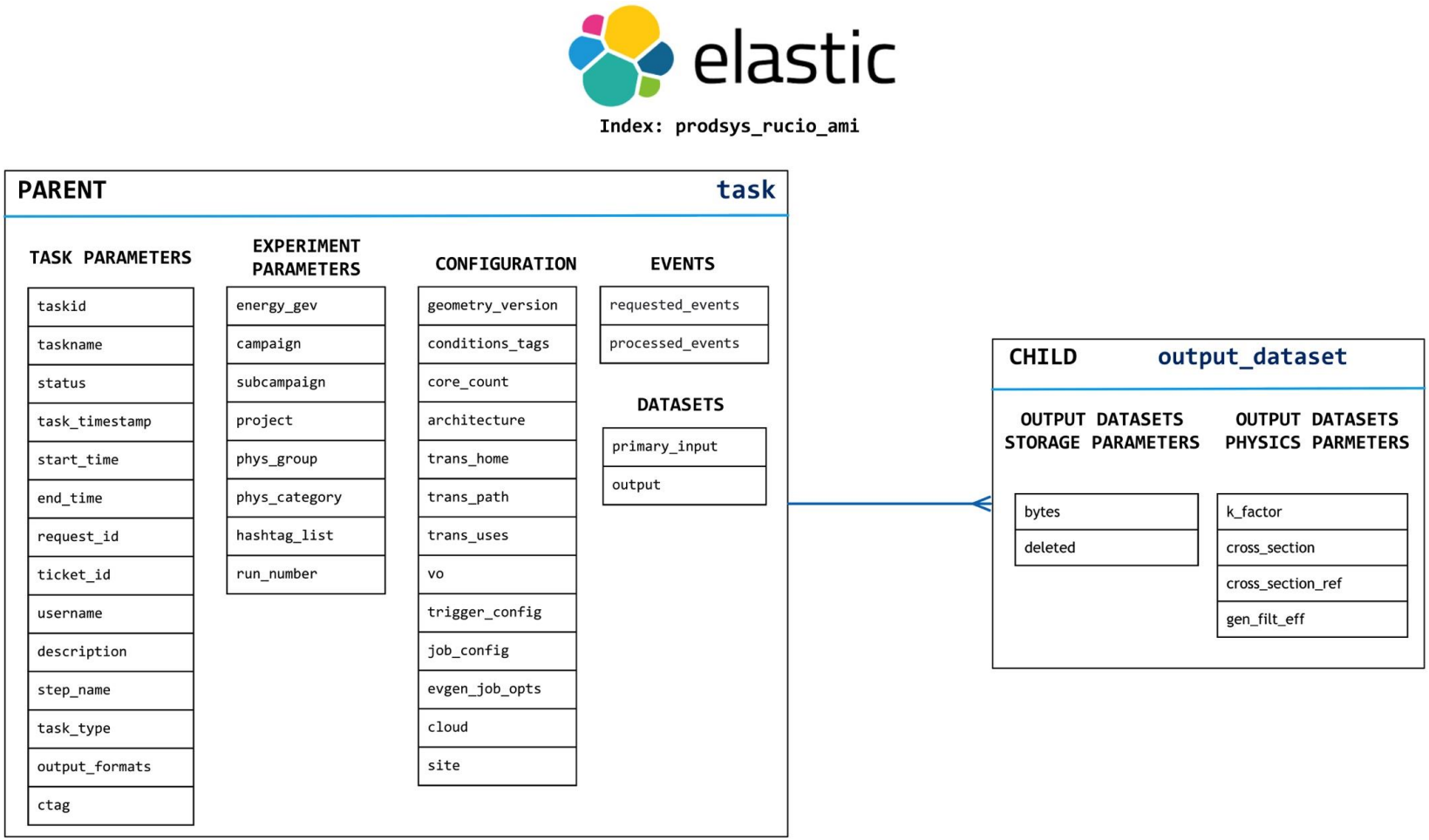
- Metadata since 01/01/2016 for tasks and datasets are loaded into the ElasticSearch of DKB instance at CERN
  - aiatlas171.cern.ch:9200
- Synchronization:
  - Data update process for tasks and related datasets is initiated when the task timestamp in ProdSys2 is updated to the "current".
  - How do we get updates from AMI and Rucio?
    - The easiest way would be to have a possibility to ask for data "changed since moment X" on a regular basis.
    - We also thought about receiving notifications about changes from the external systems in more or less real-time mode (e.g. via some REST API).
- Future Plans:
  - Automatic deployment via Puppet
  - Kafka-driven data processing flow

Addition slides

Metadata in ProdSys2, Rucio and AMI



Data model in ElasticSearch



Parent-Child Relationship in ElasticSearch

- Allows to associate one entity to another in one-to-many relationship, and all entities live within separate documents.
  - The parent document can be updated without reindexing the children.
  - Child documents can be added, changed or deleted without affecting either the parent or other children.
  - Child documents can be returned as the result of a search request.

- ElasticSearch Index Mapping General Structure:

```
{
  "template": "prodsys_rucio_ami»,
  "settings": {
    "number_of_shards": 4
  },
  "mappings" : {
    "task" : {
          "properties" : {...}
    },
    {
    "output_dataset": {
        "_parent": {
             "type": «task»
         },
         "properties": {...}
    }
}
```