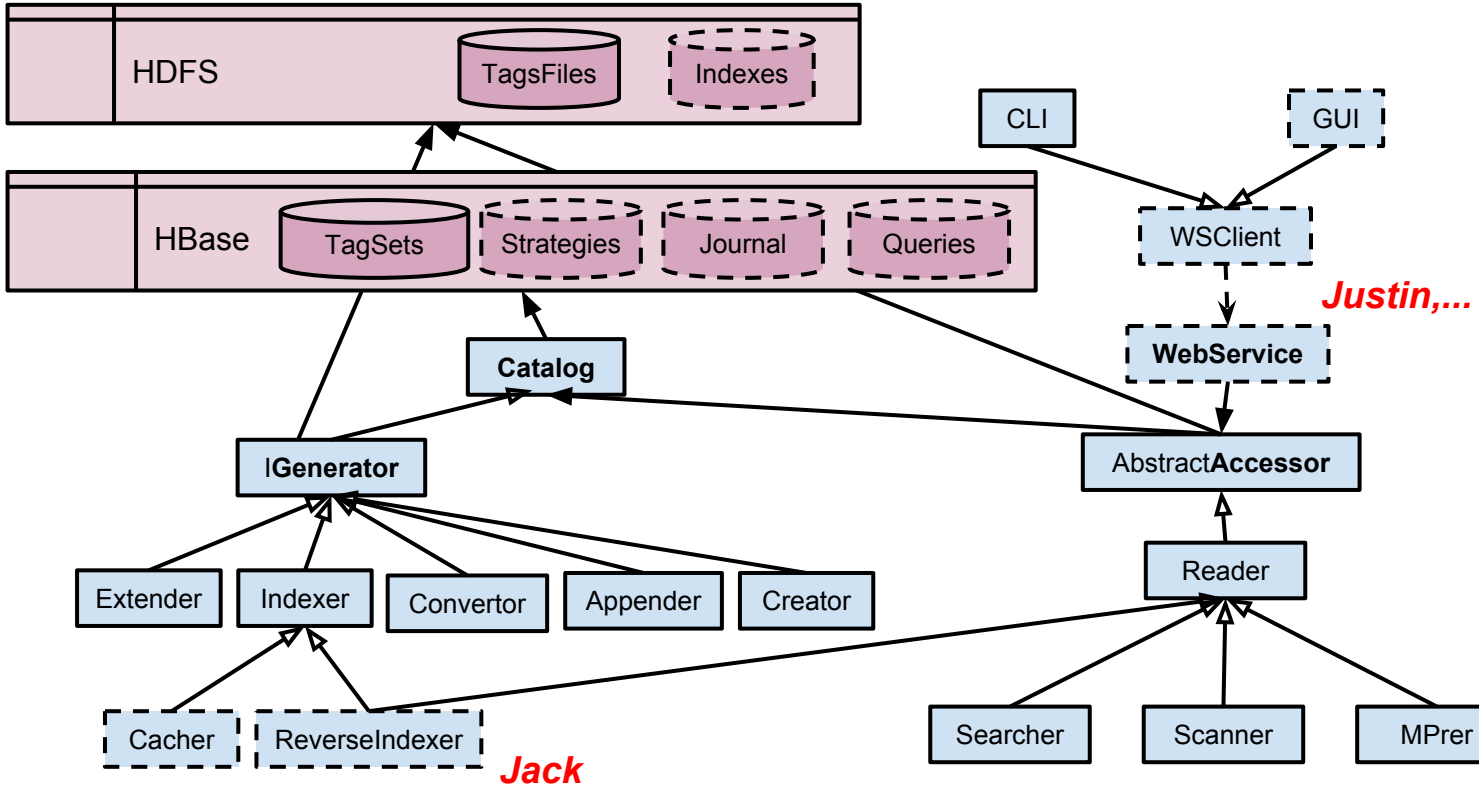
Event Index - Core

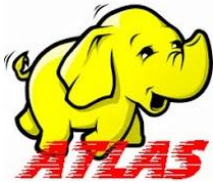
- What's done
  - Catalog
  - EI
- What will be done

(comments welcomed)

Domains


HDFS
TagsFiles
Indexes
HBase
TagSets
Strategies
Journal
Queries
CLI
GUI
WSClient
Justin,...
WebService
Catalog
IGenerator
AbstractAccessor
Extender
Indexer
Convertor
Appender
Creator
Reader
Cacher
ReverseIndexer
Searcher
Scanner
MPrer
Jack
Andrea, Fedor,...

Catalog - CLI



hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.CatalogCLI [-help] [-catalog <catalog name>] [-query <query>] [-modify <modification query>]

-help:
Gives this help.
-catalog:
If catalog name is not specified, the default catalog is used.
-query:
Query should be specified as a blank separated list of key-value pairs (those separated by ":"). They are considered as being connected with AND, i.e. only TagSets with all correct fields will be given. If query is not specified, the whole table is shown.
-modify:
The TagSets satisfying <query> will be modified with <modification query>. This argument can be used to switch TagSets on/off by modifying appropriate field. Only descriptions can be modified in this way. All modifications which don't correspond to a description will be set as an attribute.

comments to names/syntax welcomed

Examples:

hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.CatalogCLI **-query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979**
gives all TagFiles with specified path.

hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.CatalogCLI **-query format:map**
gives all TagFiles which are written as maps.

wildcards, etc not yet implemented

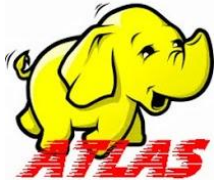
hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.CatalogCLI
gives the full dump of the default Catalog.

hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.CatalogCLI **-query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -modif format:map**
set format=map to all TagFiles with specified path.

hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.CatalogCLI **-query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -modif disabled:true**
set/add attribute disabled=true to all TagFiles with specified path.

only some properties are modifiable

Catalog - Notes



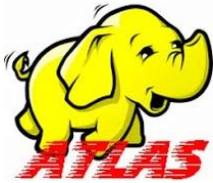
in HBase

- each entry corresponds to one **TagFile**
- contains three families:
  - descriptions
    - contains schema = column names+types
  - relations
  - attributes: user can add/modify any attribute
- master TagFile = **TagSet** = set of TagFiles (not yet fully implemented)
- interrogations / modifications via
  - API (automatic when using **Generator** package)
  - CatalogCLI (only some operations allowed)
  - GUI (not yet done)
- ToBeDone:
  - consistency selftest
  - backup
  - rollback
- private catalog via **-catalog atlas.<username>.fileset**

```
description: 002, data11_7TeV_physics_Muons_f403_m980_m979, EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979, tags, map, imported tags,
             LumiBlockN=long BunchID=long EventTime=long EventTimeNanoSec=long EventWeight=float ID=long L1PassedTrigMaskTAP=String ...
relations:   001, 003, null, null, null
attributes:  {import date=Tue Feb 25 01:44:43 CET 2014, disabled=true}
```

EI - Help

EI CLI is an extension of Catalog CLI



hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI [-help] [-catalog <catalog name>] [-query <query>]
[[-key <key>] [[-scan|-mr] <formula>] [-filter <column list>]]

-help:

Gives this help.

-catalog:

If catalog name is not specified, the default catalog is used.

-query:

Query should be specified as a blank separated list of key-value pairs (those separated by ":"). They are considered as being connected with AND, i.e. only TagSets with all correct fields will be given. If query is not specified, the whole table is shown.

-key:

The list of keys to be searched for.

-scan:

The (Java-correct) formula to be evaluated to boolean value against data in TagFile. Evaluation will be done using full-scan job.

-mr:

The (Java-correct) formula to be evaluated to boolean value against data in TagFile. Evaluation will be done using Map/Reduce job.

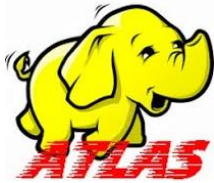
-filter:

The blank-separated list of variable names to be printed out. If a file is created with result entries, it will continue full records even when filter is specified. If no filter is specified, all values will be print out. If filter is empty, no values will be print out (just the number of events).

comments to names/syntax welcomed

scan will probably disappear

EI - Examples




```
hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI -query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -key '189184-10000235 189184-10000183'
```

gives all tags with specified path and keys.

```
hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI -query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -scan 'ID==705598'
```

gives all tags with specified path and satisfying specified formula using full-scan.

```
hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI -query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -mr 'ID==705598'
```

gives all tags with specified path and satisfying specified formula using Map/Reduce job.

can use any correct Java construct
and call Evaluator functions and scripts

```
hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI -query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -mr 'ID==705598 && LumiBlockN==133'
```

gives all tags with specified path and satisfying specified formula using Map/Reduce job.

```
hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI -query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -mr 'ID==705598' -filter 'ID StreamRAW_ref_2'
```

gives all tags with specified path and satisfying specified formula using Map/Reduce job. Report will contain only variables selected by filter.

```
hadoop jar EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI -query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -mr 'ID==705598' -filter ' '
```

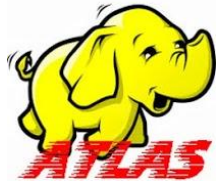
gives the number of tags with specified path and satisfying specified formula using Map/Reduce job.

EI - Result

```
$ jar ../lib/EIHadoop.jar net.hep.atlas.Database.EIHadoop.Apps.EICLI -query path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979 -mr 'ID==705598' -filter 'ID StreamRAW_ref_2'
14/02/25 10:36:58 INFO Apps.CatalogCLI: CLI arguments: {-filter=ID StreamRAW_ref_2, -mr=ID==705598, -query=path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979}
   12 INFO  (Util.Init                    :  31) : Initialised, version: 1.0.1 [24/Feb/2014 at 15:49:20 CET]
   14 INFO  (Apps.CatalogCLI              :  95) : Opening the default Catalog ...
 1226 INFO  (Catalog.Catalog              :  70) : Opening Catalog atlas.hrivnac.filesets
 1226 INFO  (Apps.EICLI                   :  79) : Searching in path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979
 1228 INFO  (Apps.CatalogCLI              : 119) : Searching for path:EIHadoop/data11_7TeV/physics_Muons/f403_m980_m979
 1390 INFO  (Accessor.MRer                :  79) : Searching with FormulaQuery(ID==705598)
87595 INFO  (Accessor.AbstractResult      :  73) : Setting report filter: ID StreamRAW_ref_2
87596 INFO  (Apps.EICLI                   : 115) : FileResult(FormulaQuery(ID==705598)):
### MRer/Tue_Feb_25_10_37_00_CET_2014
==========
ID = 705598
StreamRAW_ref_2 = 81553260
All 1 results are available from files: [MRer/Tue_Feb_25_10_37_00_CET_2014]
87679 INFO  (Apps.EICLI                   :  58) : Task took 87s
```

- up-to 5 results shown as requested via -filter
- all results stored in TagFile(s)
  - with the same schema as primary TagFiles
  - will be registered in Catalog / cached / bookmarked
- number of found entries reported

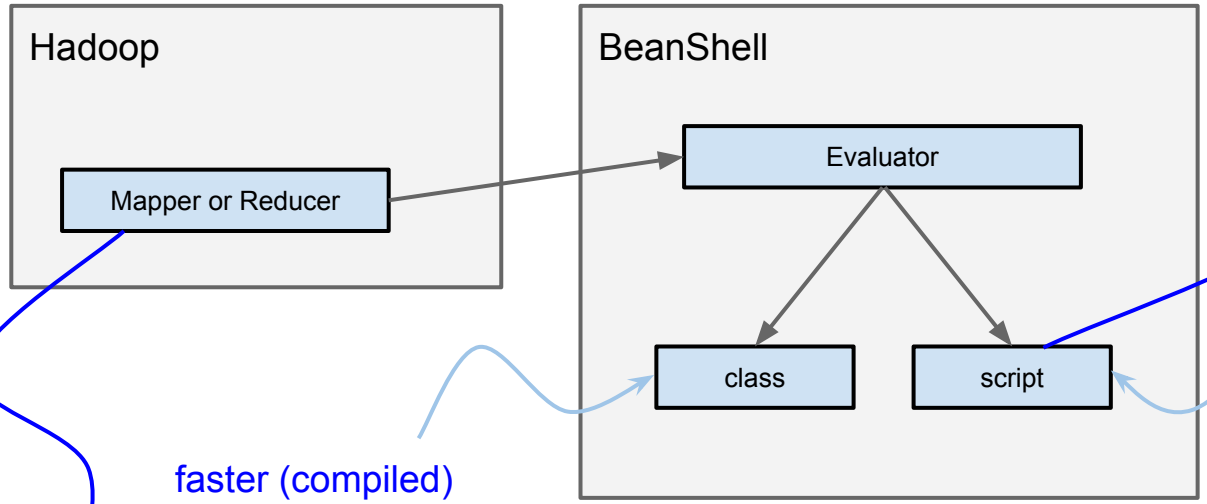
EI - Notes



- EIHadoop.jar is in TagConvertor/lib
- two-step search:
  - Catalog -query
  - EI -key/-scan/-mr & -filter
- -key search is immediate (based on memory-resident index)
  - key = RunNumber-EventNumber
  - will add possibility to search for a sequence of keys
  - more general search (wildcards,...) doesn't have sense, use -scan / -mr instead
  - will look at MapFile subformats, maybe implement some
- -scan / -mr accept the same <formula>, but process it in different ways
  - -scan performs full scan, -mr executes MapReduce job
  - -mr is much faster
  - maybe will abandon -scan
- any Java-correct <formula> is accepted, it should evaluate to boolean
- more complex <formula> can be easily implemented

EI - Formula Evaluation

```
public static long eventNumber() {
  String r = RunNumber_EventNumber.substring(RunNumber_EventNumber.indexOf("-") + 1, RunNumber_EventNumber.length());
  return new Long(r).longValue();
  }
```



faster (compiled)

Trigger Decoding etc.should go here

simpler (interpreted)
has direct access to variables
(like closure)

```
if (_evaluator.evalBoolean(varNames, varTypes, varValues, formula)) {
  context.write(key, value);
  }
```

```
-mr “eventNumber() ==
123”
```

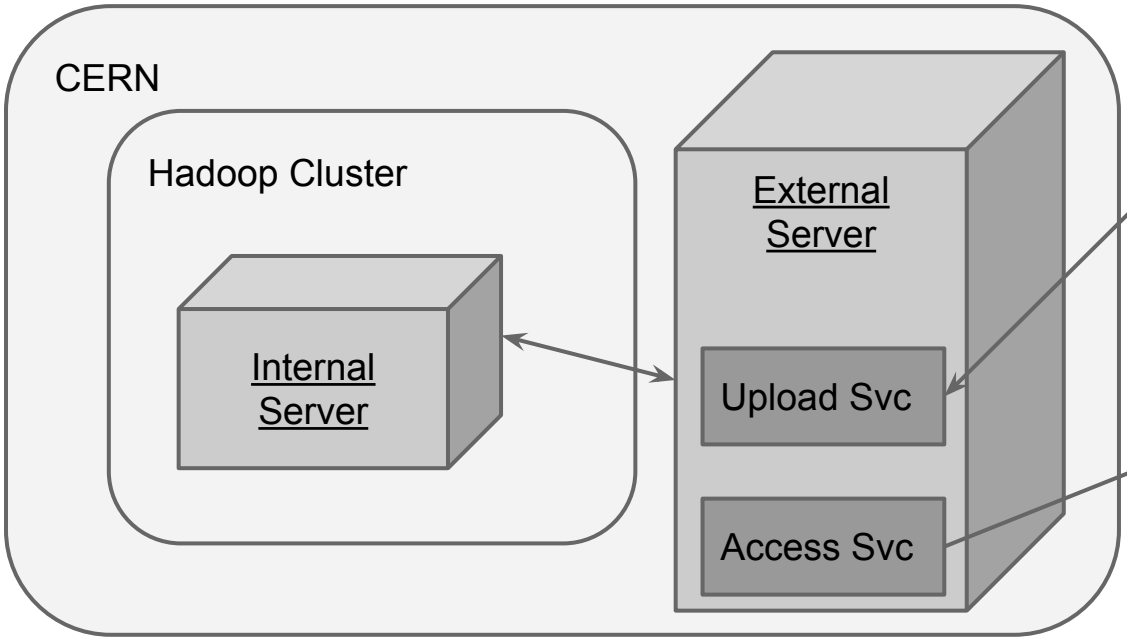
EI - Formula Evaluation - Future Optimisation

```
public static long eventNumber() {
  String r = RunNumber_EventNumber.substring(RunNumber_EventNumber.indexOf("-") + 1, RunNumber_EventNumber.length());
  return new Long(r).longValue();
  }
```

- (interpreted) scripts → (compiled) class
- (dynamically) wrap formula in (compiled) class - compilation is cheap
- store pre-calculated intermediate results in parallel tables (*slaves*) - space is cheaper than CPU
- replace BeanShell with faster (but poorer) scripting engine (Java 7+)

```
-mr “eventNumber() == 123”
```

External Relations



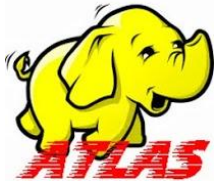
Jason over **ActiveQM**

PanDA

http

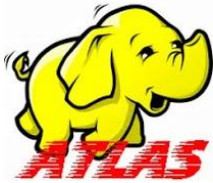
User
Web Service
&
CLI/GUI clients

Requests



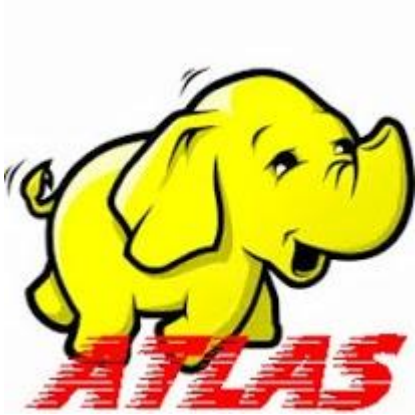
- Urgent
- Later
  - Interval searches should be supported for keys in Accessor.
  - Results should be cached and cataloged to be used in subsequent searches.
  - Error reports should be descriptive.
  - Long queries should provide progress report (for -scan).
  - Catalog -query request should accept wildcards.
- Maybe
  - -scan/-rm should allow sourcing script from file.
- If Reasonably Possible
  - Wildcards should be supported for keys in Accessor - should look at MapFile extensions.
  - It should be possible to mix -key with -scan/-mr request.

Next Steps



- Optimisation:
  - Speed
- Service:
  - Currently runs as CLI on Hadoop cluster
  - Should be available as a service (need https server in Hadoop) and CLI/GUI should connect to it (three-tier architecture)
- Logistics:
  - Access (shared & private data/Catalog)
  - Data/Catalog protection
  - Data import
  - Catalog backup
- Catalog:
  - TagSets (collections of TagFiles)
  - Register results (= cache, bookmarks)
  - Register indexes
- GUI:
  - Web (Justin)
  - Real GUI ?
  - Android ?
  - Visual Catalog editor (or at least browser)
- Trigger:
  - Function available to Evaluator (Fedor, Claudia)
- Indexes:
  - Creation (Jack,...)
  - Use & Catalog

Links



- Home:
  - http://cern.ch/hrivnac/Activities/Packages/TagConvertor
- JavaDoc:
  - http://cern.ch/hrivnac/Activities/Packages/TagConvertor/JavaDoc
    - **ElHadoop**: current package
    - TagHadoop: legacy package (but useful code)
- SourceDoc:
  - http://cern.ch/hrivnac/Activities/Packages/TagConvertor/Src
- SVN:
  - svn+ssh://svn.cern.ch/reps/atlasoff/Database/TAGHadoop/TagConvertor
- New build targets:
  - ant check-catalog
  - ant check-ei