Towards enhanced databases for HEP

Andrea Ceccarelli, Andrea Cioni,
Maria Vittoria Garzelli, Piergiulio Lenzi, Laura Redapi

[1] Università degli Studi di Firenze, Dipartimento di Matematica e Informatica
[2] Università degli Studi di Firenze, Dipartimento di Fisica e Astronomia
[3] Eberhard Karls Tuebingen Universität, Institut für Theoretische Physik
[4] GARR Consortium

XXVII International Workshop on DIS and Related Subjects,
Torino, 8 - 12 April, 2019

Which HEP data ? [DPHEP Group, 1205.4667]

* **Raw data** (level 4):

- $\mathcal{O}(Petabyte)$ scale for a single modern HEP collision experiment.
- organized in billion/trillion events.
- indexing fundamental for data retrieval.

* **Analysis level data** (level 3):

- sufficient to perform a complete re-analysis

* **Simplified event level data** (level 2):

- 4-vectors describing the detected particles

* **Published data** (level 1):

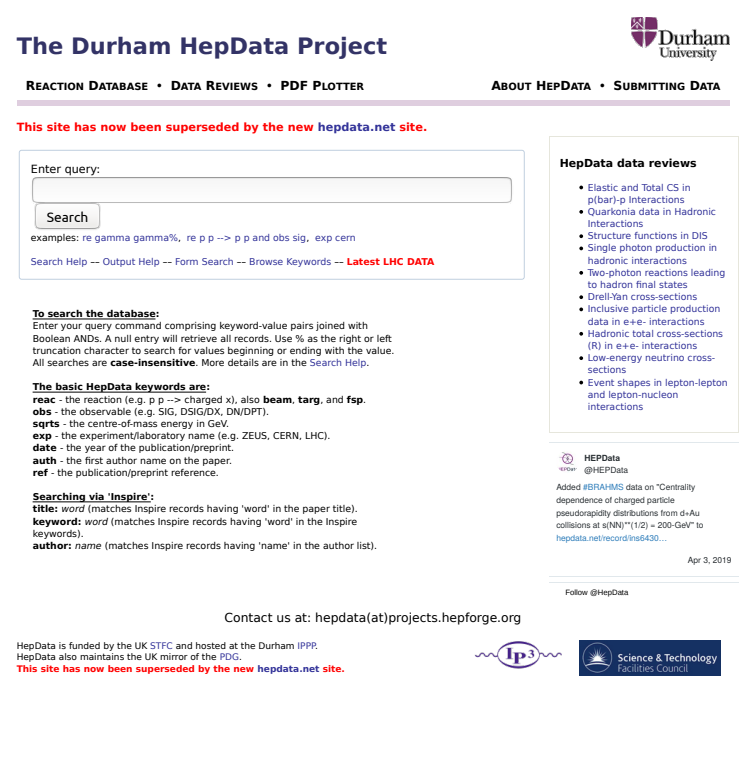
- far less than raw data, but rapidly increasing as well.
- multisource (many experiments).
- heterogeneous (different data structures).
- multidimesional (e.g. 3D structure of protons at the EIC).

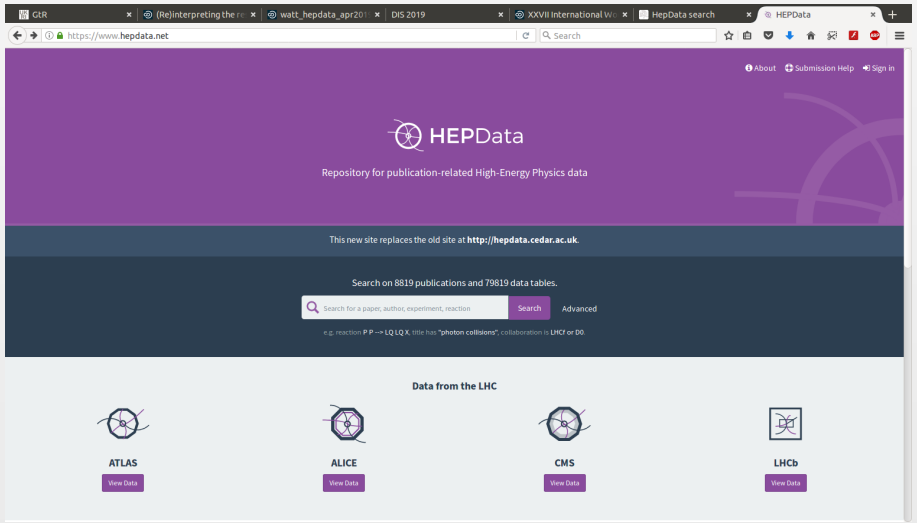
Many challanges for database organization and access !
In the following we focus on published data (level 1).

This work is based on a dump of the HepData *database, kindly provided us by R. K. Ellis, F. Krauss and G. Watt (IPPP Durham).*

`hepdata.cedar.ac.uk` `www.hepdata.net`

HepData search http://hepdata.cedar.ac.uk/





The HEPData project

The HEPData project was born in the middle seventies, to include data of high-energy scattering reactions, as published in Physics papers (level 1).
The oldest data included date back to the fifties (1954).

1) Legacy version [W. J. Stirling, M. R. Whalley et al.]:
   * Database according to the hierarchical Barkeley DBMS.
   * Tree-like structure, record unit = paper including its experimental data.

2) Intermediate HepData version, devoloped within the CEDAR project [A. Buckley, M. R. Whalley, J. and S. Butterworth et al.]:
   * Database according to the relational DBMS MySQL.
   * Complex structure, which reflects the structure of (paper + data).
   * Data entered as text files, creation of the HepML data format
   * User access through a Web interface written in Java.
   * Developments towards the automatic comparison of Shower Monte Carlo predictions to experimental data (JetWeb, RIVET), interest for SMC tuning.

3) Recent HEPData version [G. Watt, E. Maguire, L. Heinrich et al.]:
   * Database according to the object-relational DMBS PostgreSQL.
   * Data entered as `.yaml` (superset of `.json`) files directly by the experimental authors. Format for semi-structured data used, but still RDBMS.
   * Invenio3 open-source framework for large-scale digital repositories.
   * Developments towards the inclusion of more data, not only limited to HEP reactions. BSM physics, statistical models.

The MineHEP project

* Interdisciplinary project financed by the University of Firenze.
  Proposal prepared and submitted in 2017. Funded for 2018 - 2019.

* **Aims**:
  Organize the already available HEP data in such a way
  to automatically extract as much information as possible.

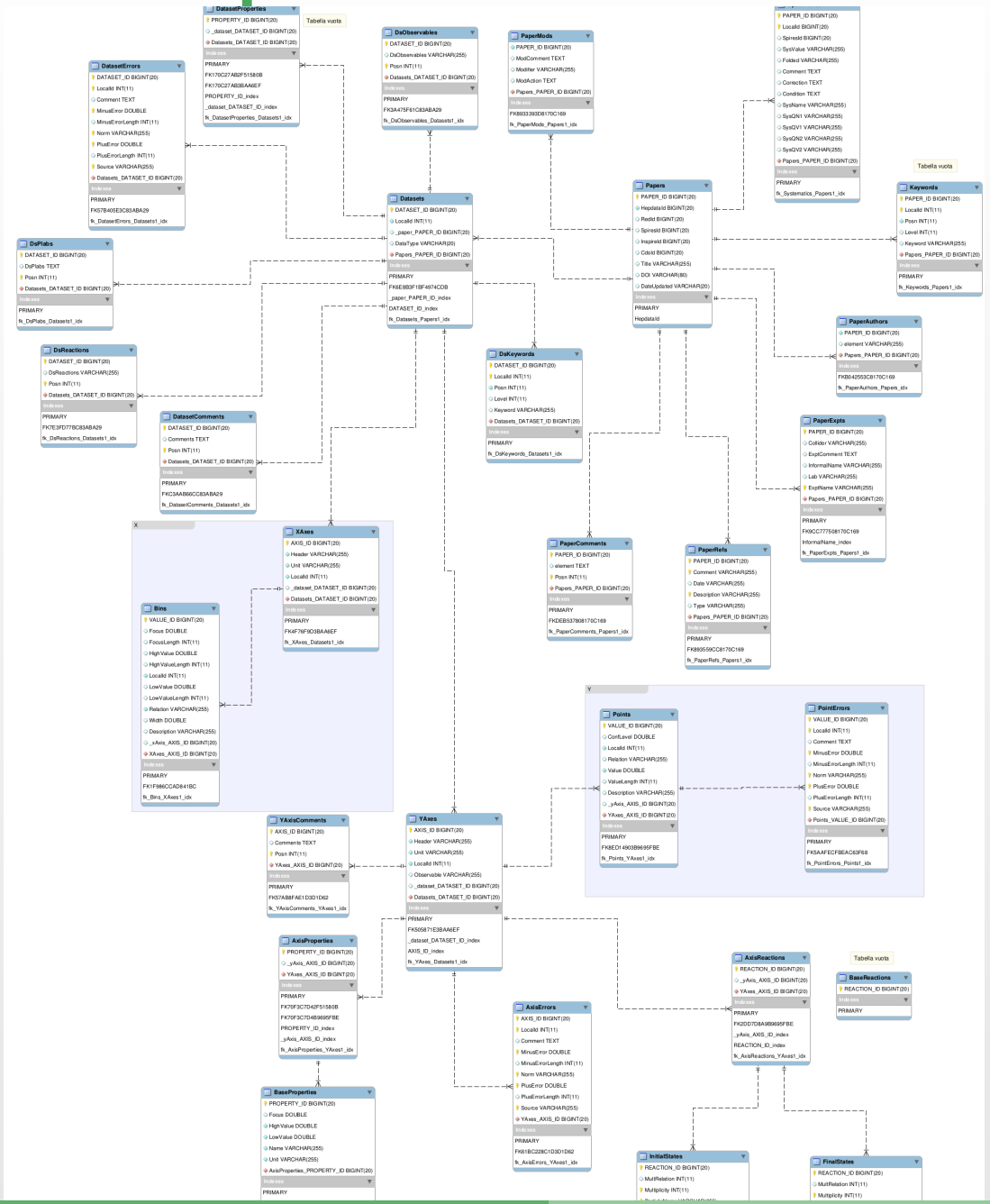
  Tests of the Standard Model and searches for New Physics.

* HEP **data input** for MineHEP:
  the data included in the HEPData database.

  As a matter of fact we started from the dump of the HepData database
  in the file `hepdatapublic.dmp.gz_01Apr18`

The MineHEP database adopts MySQL, like the HepData version from which it is derived. Migrating to PostgreSQL (which has support for `.json` data, which is the most popular format for semi-structured data commonly stored in NoSQL systems) is not expected to be difficult.

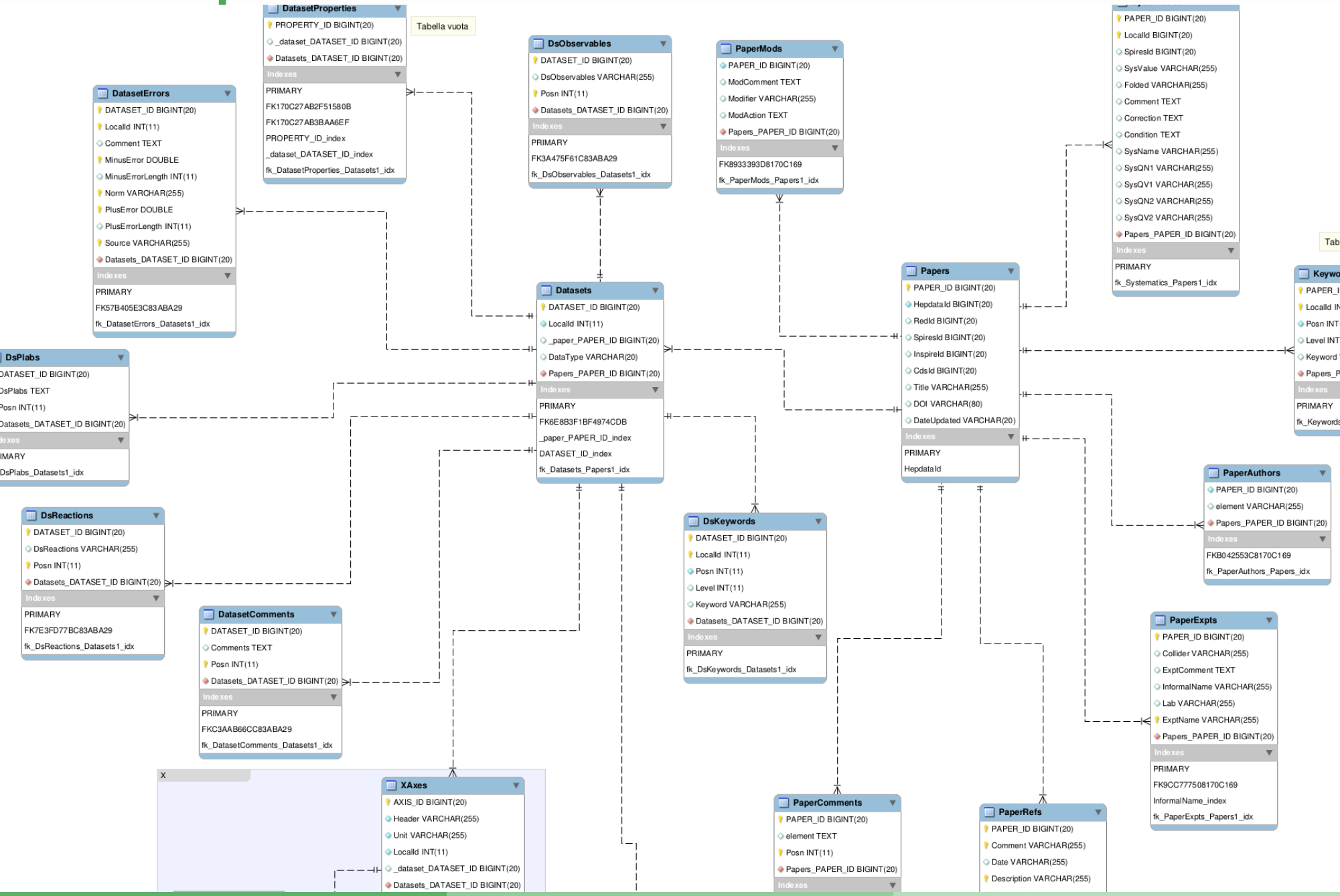
The HepData relational database structure



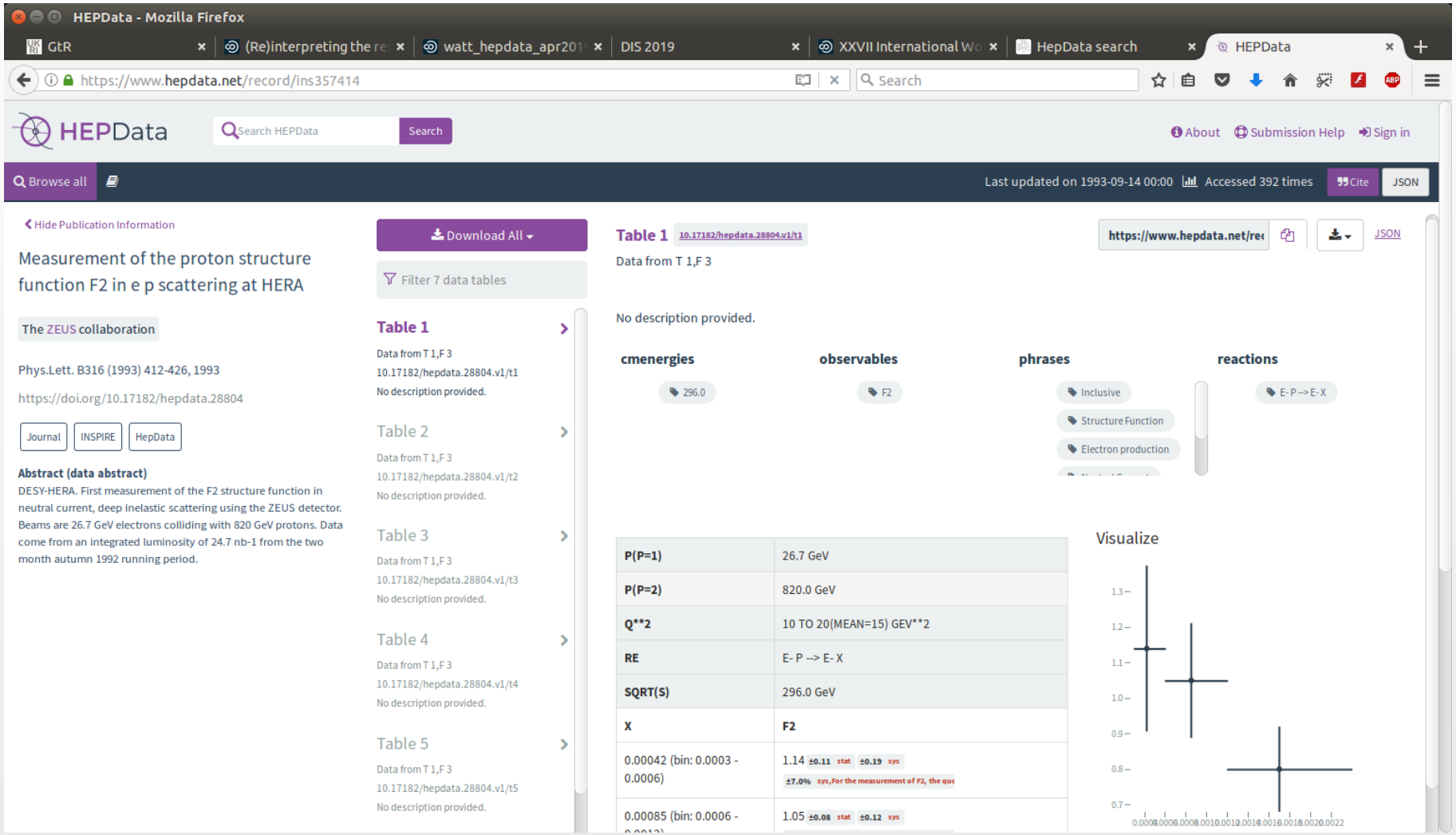
Relational Database:

- data organized in Tables organized in structured fields (Columns);
- complex relations between Tables are allowed.

The HepData database structure: zoom



Example of an HEPData webpage for a paper



a `.json` file for each paper: data with semi-structured format.

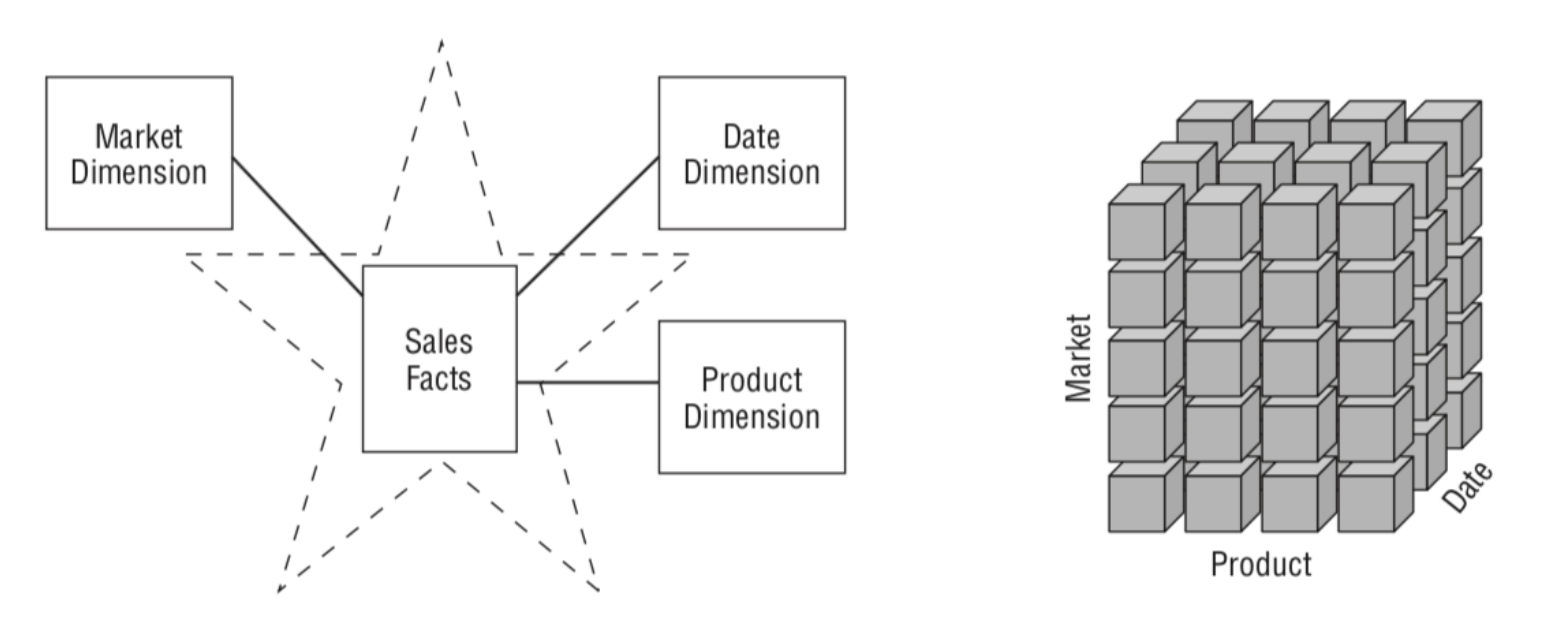
Flexibility: possibility to add new information a posteriori in a file.

MineHEP and OLAP data warehouses

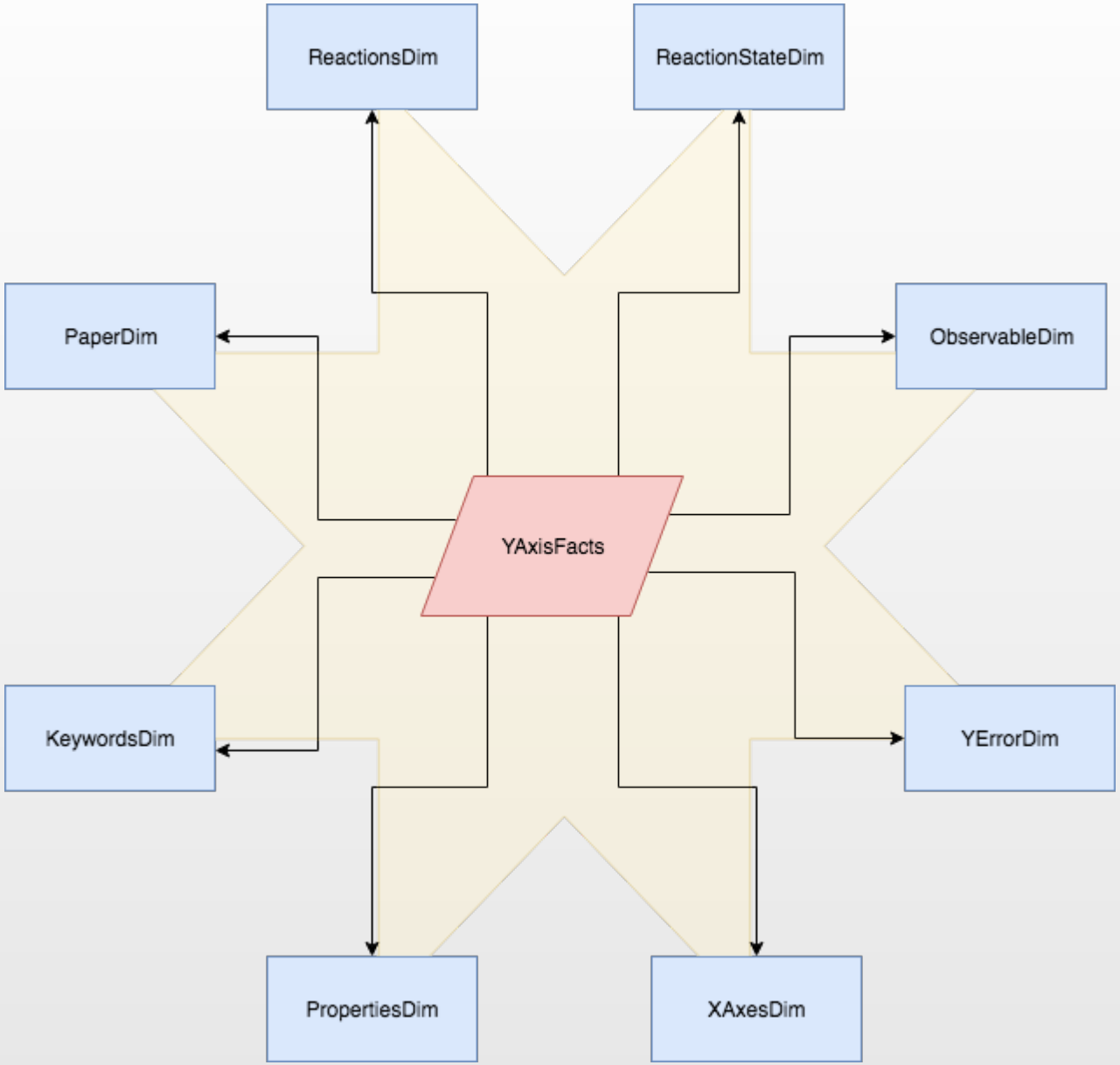
Organization of HEP data in structures inspired by the OLAP techniques, used in business analyses:

* a star schema,
* from which one can build a multidimensional OLAP cube,
* to allow for complex queries for data extraction and analysis which can involve simultaneously various dimensions.

Slicing, dicing, roll-up and drill-down operations possible on the cube.



The MineHep star schema



The migration from the HepData schema to the MineHEP schema

Make use of **temporary intermediate tables**, to store the information derived from different initial HepData tables, before populating the final tables of the star-schema.

Complex procedure, due to the inhomogeneity of the various HepData records, not all of them include the same filled fields.

Spotted errors in some of the HepData records.

Still **open issues** (at least for us):

- lack of some information (e.g. there is some info on the analysis cuts given only in the paper texts, but not stored in the database),
- many different names for a same variable (make the queries more difficult),
- different units of measurement.... (solvable at the plotting/analysis level).

Tests of the success of the migration

* **Automatization of the migration procedure**: given a new HepData dump, it is now possible to **populate the MineHEP database** by a sequence of scripts MySql in a **few hours** on a modern laptop.

* **Tests**: specific queries on MineHEP, whose output was checked against the output of HEPData, which allows for a search for papers selecting a specific reaction, observable, cmenergy, keywords, phrases, collaboration, title, abstract, date, authors, institutions, doi.

* **Tests**: more complex queries, whose results were checked manually against the info in the papers and in HEPData.

Example of a relatively complex query

1) Let's search for all the experimental differential distributions $d\sigma/dp_T$ *for the reaction* $pp \to B^+ + X$ *at a fixed center of mass energy, for* p_T *in the range between* 0 *and* 250 GeV *and* $|y| < 5$.

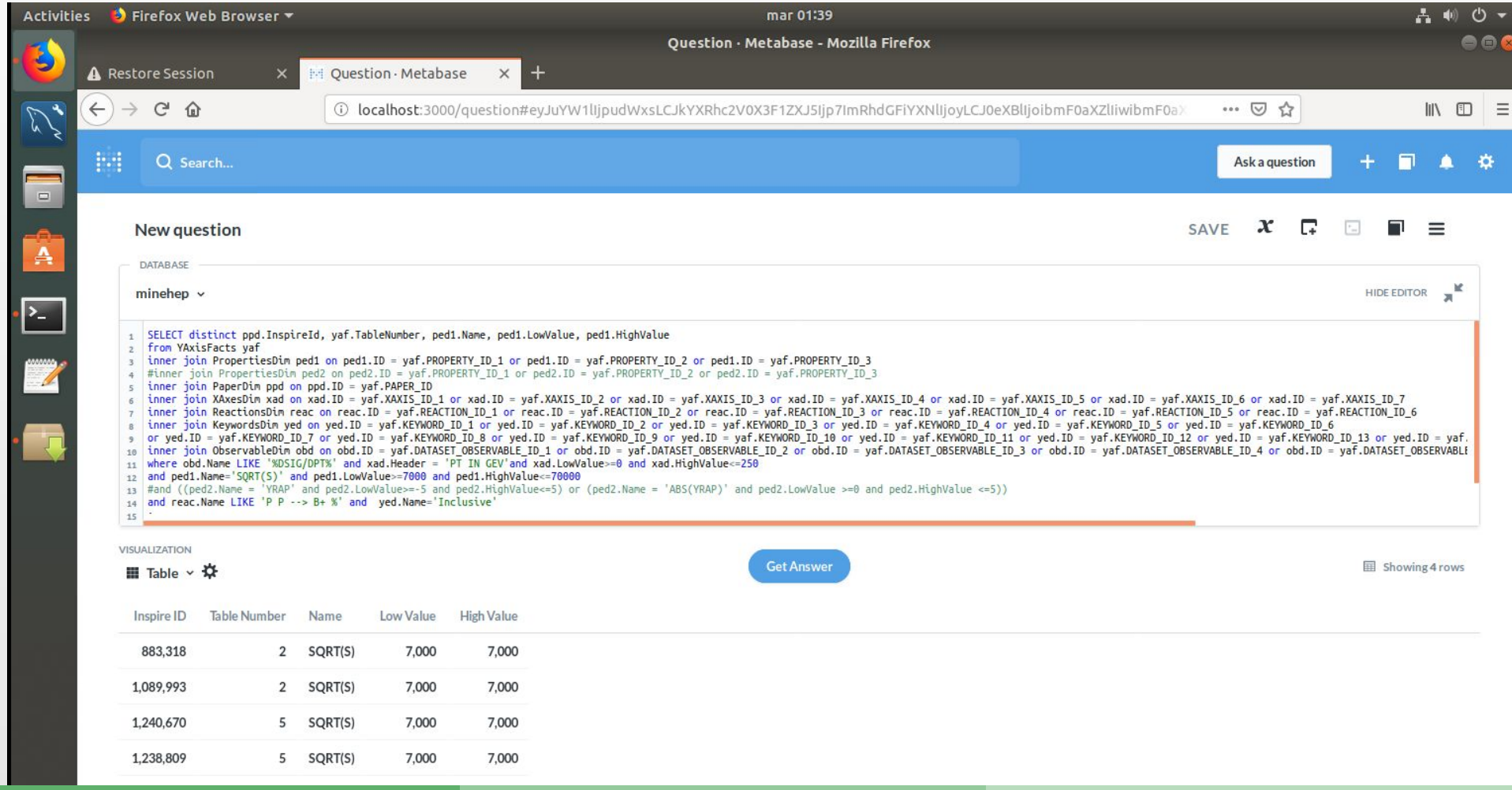
2) Then, let's plot all these distributions, and superimpose them.

As a matter of fact, it is possible to execute the query 1) on the MineHEP database, obtaining as an output a single table with the experimental data $d\sigma/dp_T$ belonging to different papers and satisfying the p_T and y cut in the query. For each experimental point, we additionally store the indication of the paper from which it is extracted.

Reading the data from that table, it should then be possible to make a plot with the distributions of the various papers superimposed.

MineHEP interfaced with Metabase: the dashboard

Let's search for all papers which include the $d\sigma/dp_T$ *distribution for the reaction* $pp \to B^+ + X$ *at a fixed center of mass energy* $\sqrt{s} = 7\,TeV$*, for* p_T *in the range between* 0 *and* 250 *GeV.*

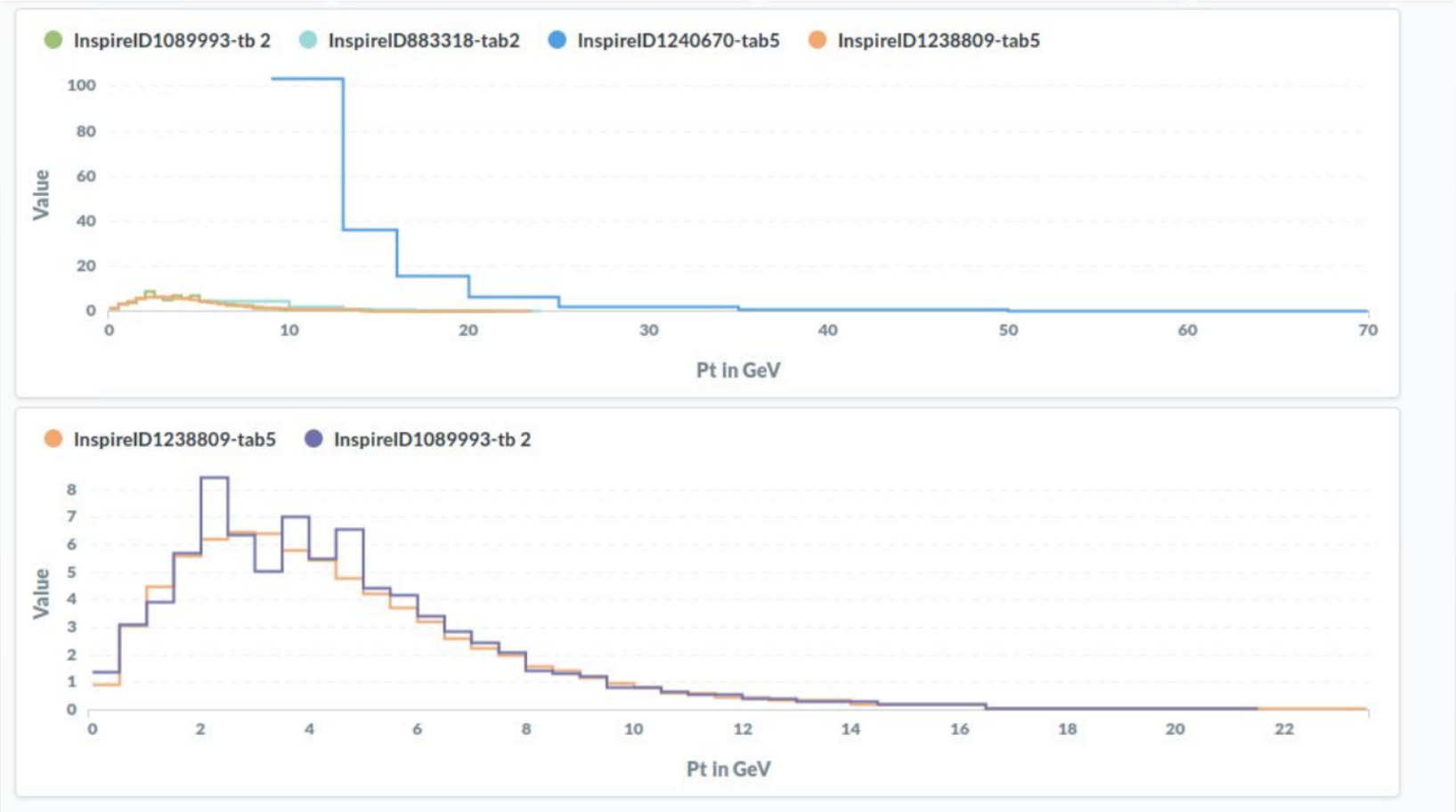


MineHEP interfaced with Metabase: the dashboard

For each paper selected by the previous query, let's plot the $d\sigma/dp_T$ distribution satisfying the cuts.

Minehep interfaced with Metabase: the dashboard

Finally, let's superimpose the plots.



The user will observe the difference by a large factor between the ATLAS distribution and the LHCb ones. Besides noticing that they are related to different rapidity cuts, he will then go to read the paper and discover that ATLAS distribution includes a non-trivial multiplicative factor by a branching ratio. Unfortunately this info is not yet in MineHEP Tables.

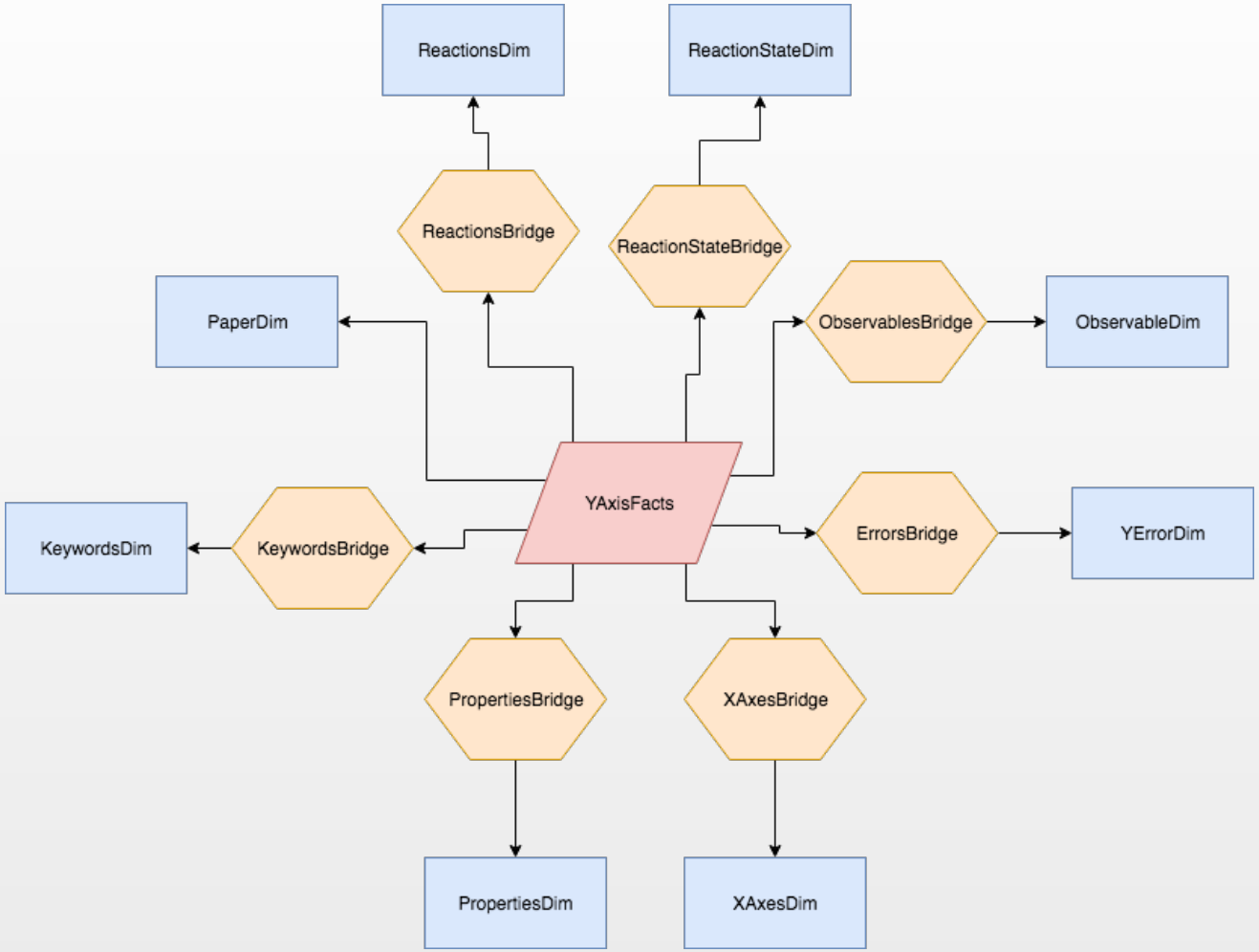
MineHEP table population from the new HEPData

* Python script to **import** into the MineHEP tables the information stored in **the** `.json` **file** associated to each HEPData publication, under development.

* So far we imported a single publication, which showed us that automatic import can be done. Generalization, however, will require some work.

* Possible issue of **scalability** of the star-schema when some of the dimensions will grow exagerately (e.g. the dimension of the keywords, if many new HEP papers will include more and more new keywords).

MineHEP alternative schema: the snowflake schema



* Improved scalability.
* A version of MineHEP populated according to this schema also already exists. We are now testing it.

Conclusions

* Likely future **explosion of the amount of data** in HEP experiments
⇒ importance of data preservation, longevity, accessibility and reusability.

* Properties to be evaluated in the **design of a HEP database**:
  - user needs
  - scalability
  - agility (quick adaptation to changes)
  - easy monitoring
  - security
  - availability / fault tolerance

* Different **schema designs** are possible:
  SQL - OLAP - No SQL

* We show that **OLAP techniques** can be used **to re-organize the contents of HepData, allowing for complex queries** and comparisons between experimental data from different sources.

* What to do next ? Wish-list from the DIS community for an HEP database ?