Karlsruhe Institute of Technology

Steinbuch Centre for Computing

Evaluation of NoSQL database MongoDB for HEP analyses

Christopher Jung, Paul Jäger, Jörg Meyer – Steinbuch Centre for Computing

Motivation

Most analyses in experimental high-energy physics (HEP) are based on the data analysis framework ROOT. Monte Carlo and data events are stored in ROOT trees. A typical analysis loops over events in ROOT files and selects relevant events according to certain selection criteria for further processing.
With the emergence of NoSQL databases that allow to store large amount of data on a horizontally scaling infrastructure, the feasibility of an alternative approach for HEP analyses is being evaluated.

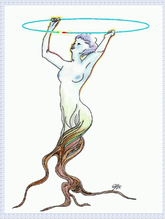
MongoDB vs. ROOT

Properties of MongoDB:

- document based NoSQL database
- schemaless: collections may contain JSON-documents with any structure
- MongoDB drivers for more than 40 programming languages
- open source, developed by MongoDB, Inc.
- http://www.mongodb.org

Properties of ROOT:

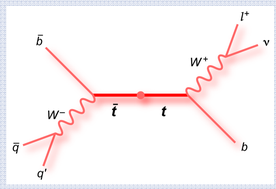
- analysis framework to handle large amount of data data
- trees, histograms, graphs, fits, …
- C++ libraries, C++ interpreter, pyroot, rubyroot
- open source, developed at CERN
- https://root.cern.ch/drupal/



Toy Analysis

sample:

- MC@NLO ttbar production, 1.96TeV

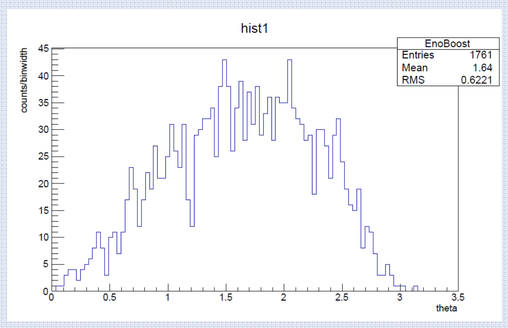


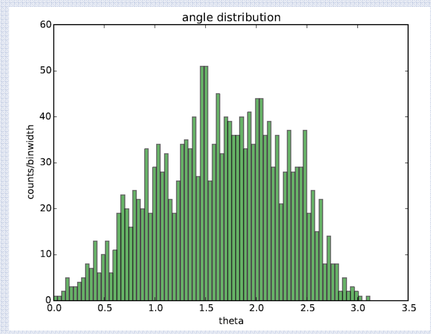
events saved in ROOT trees and as JSON documents:

```
{ eventnumber: 42,
  mu_px: [25.5, 34.1, 42.2, 3.5],
  mu_py: [44.3, 23.3, 89.2, 5.3],
  …
  e_px: […]
  …
}
```

- toy analysis steps
  - event selection
  - calculate angle theta between W and lepton in W restframe
  - match truth leptons (e, mu) to reconstructed electrons and muon
- methods
  - ROOT trees + ROOT analysis (C++)
  - ROOT tress + PyROOT analysis
  - MongoDB + pymongo + matplotlib

Results



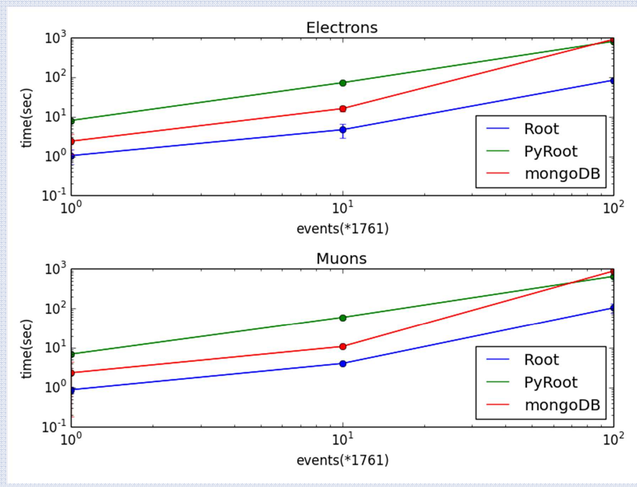


- left: ROOT histogram, right: plot by matplotlib
- calculation of theta: Lorentz boost to W restframe
  - ROOT: TLorentzVector class
  - MongoDB: JavaScript code for MapReduce
- calculation of non-trivial quantities (like theta) inside database possible via query language or MapReduce functions
  - MapReduce: submit JavaScript code to server to process data inside database

Performance

performance measurements:

- single computer
- I/O from local disk
  - files in file system
  - local MongoDB
- task: calculation of lepton resolutions



Conclusion

Analysis with MongoDB:

- even non-trivial analysis steps doable with MongoDB
- MapReduce extends ability for complicated queries
- works with all modern programing languages
- reasonable performance comparable to scripting languages

Analysis with ROOT:

- as expected ROOT outperforms Mongo-analysis
- ROOT trees have much better storage efficiency
- multi-purpose framework providing various statistical tools
- requires C++ skills, practice and training

When to use MongoDB?

- interesting alternative for people/students without special knowledge in C++/ROOT
- education
- getting into HEP analysis

LSDMA