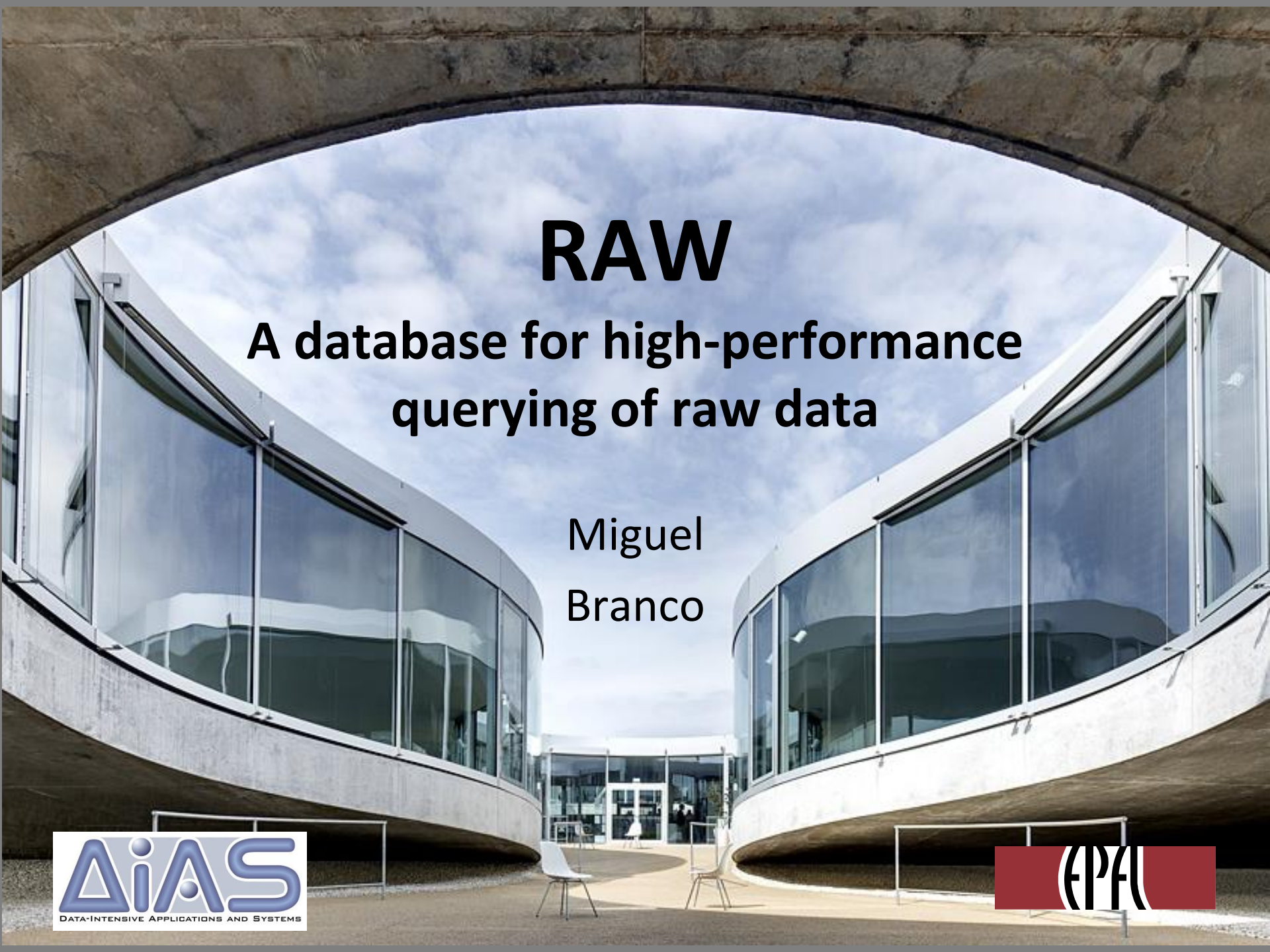


RAW

A database for high-performance querying of raw data

Miguel Branco

DATA-INTENSIVE APPLICATIONS AND SYSTEMS

EPFL

What is RAW?

RAW is a ***database*** **that answers queries directly from the user's files.**

Why use a database?

Databases make querying easy.

```
SELECT event
 FROM root:/data1/mbranco/ATLAS/*.root
 WHERE
 ( event.EF_e24vhi_medium1 OR event.EF_e60_medium1 OR …) AND
 event.muon.mu_ptcone20 < 0.1 * event.muon.mu_pt AND
 event.muon.mu_pt > 20000. AND
 ABS(event.muon.mu_eta) < 2.4 AND
….
```

Databases make querying fast.

Column-stores & vectorized execution use h/w efficiently.

Why query files?

Plenty of files around!

Flexible & easy to use.

Own data format means no vendor lock-in.

(And with many files you get to rewrite *"ls"* and *"cp"* ☺)

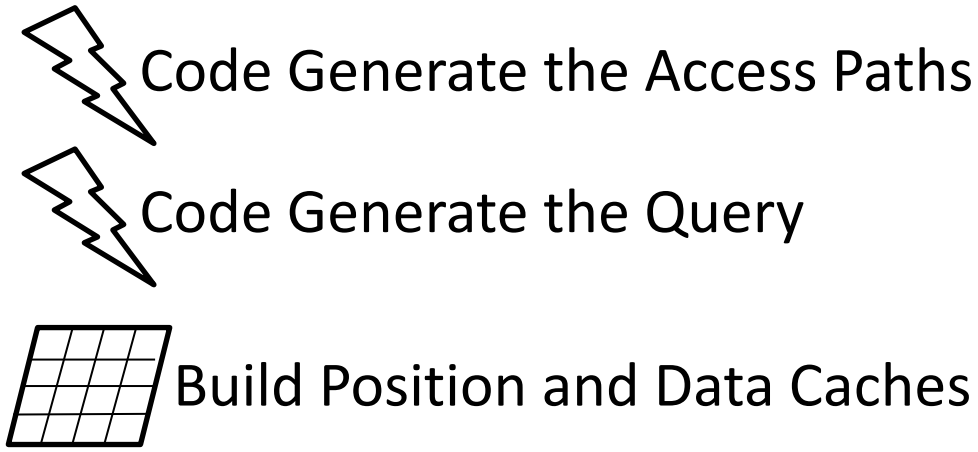
RAW combines best of both worlds:

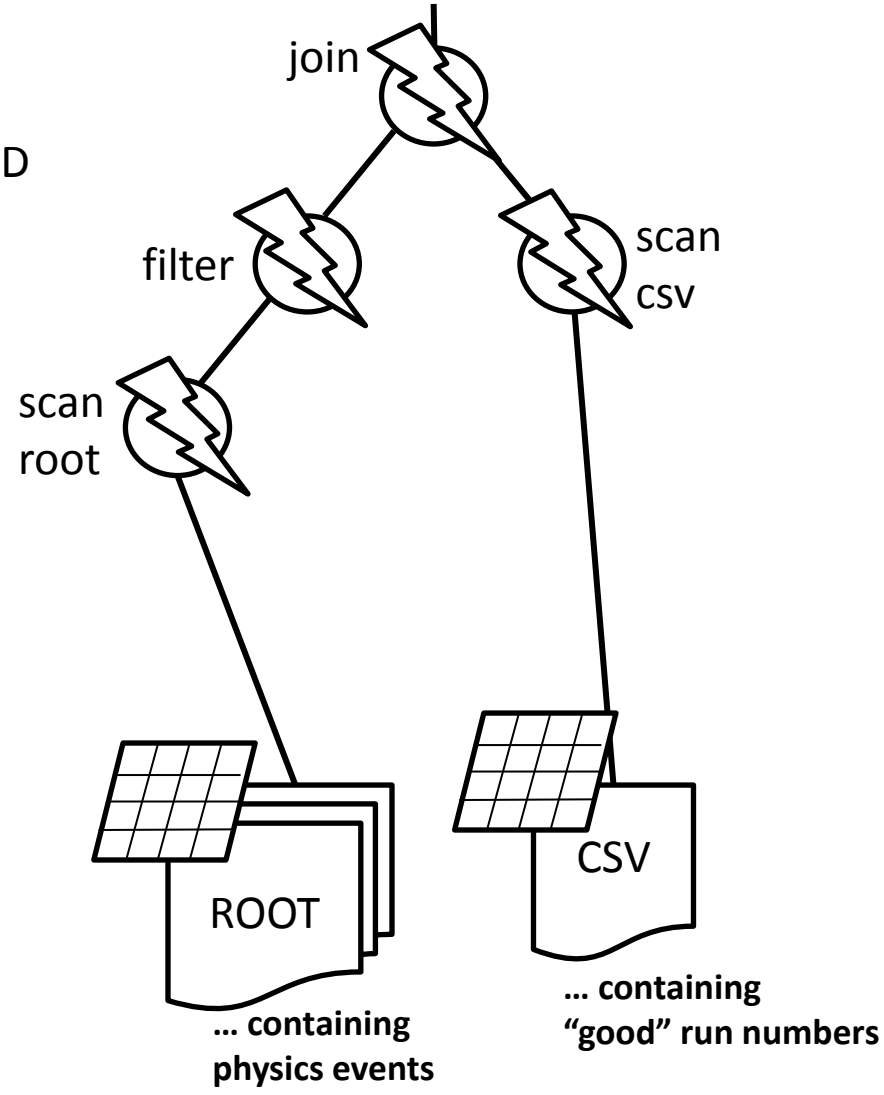
High-performance querying capabilities….

…. while keeping your own data formats, files, and scripts.

How does RAW work?

```
SELECT event.jet…
FROM csv, root
WHERE csv.RunNumber == root.RunNumber AND
        root. EF_2mu13 == TRUE AND …
```





A *Higgs Analysis* in RAW

***~900 GB** of ROOT physics data (127 files)*

ATLAS Experiment

(Thanks to ATLAS Info. Officer - Dario Barberis)

1 CSV file with “Good Runs”

Compare handwritten C++ ROOT query

(Thanks to Maaike Limper from CERN/IT)

… with RAW

	ROOT	RAW
Cold Caches	1499 s	1431 s
Warm Caches	52 s	575 *ms*

8 x 10-Core Intel Xeon CPU E7-L8867 @ 2.13GHz
192GB memory
2 x 1TB 2.5'' SASII 24x7 Disk Drive 7’200rpm

DATA-INTENSIVE APPLICATIONS AND SYSTEMS


RAW team, DIAS @ EPFL:
Manos Karpathiotakis
Miguel Branco
Anastasia Ailamaki
DIAS
Data-Intensive Applications and Systems
EPFL