UNIVERSITÀ
DEGLI STUDI
DI PADOVA

J. Pazzini
PADOVA UNIVERSITY

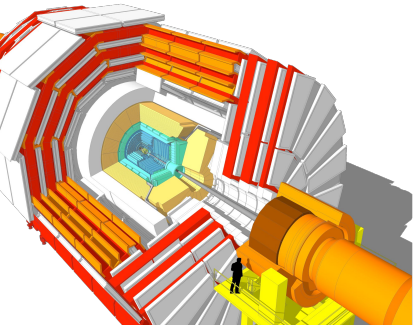
0 - INTRODUCTION

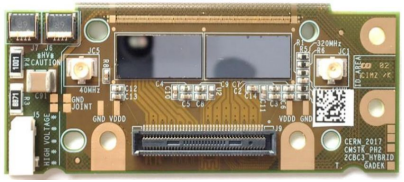
Management and Analysis of Physics Datasets - Module B

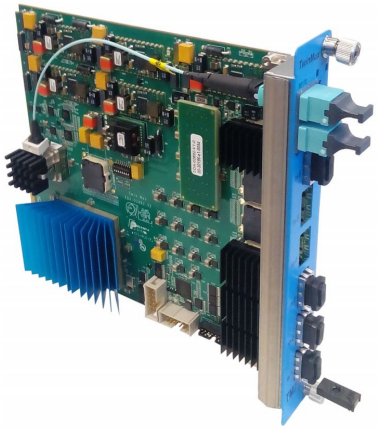
Physics of Data

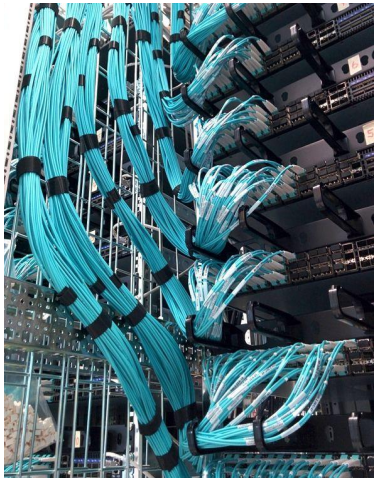
A.A. 2023/2024

FROM SENSORS TO DATA









SENSORS → FRONTEND ELECTRONICS → READOUT ELECTRONICS → TRIGGER & DAQ → DATA

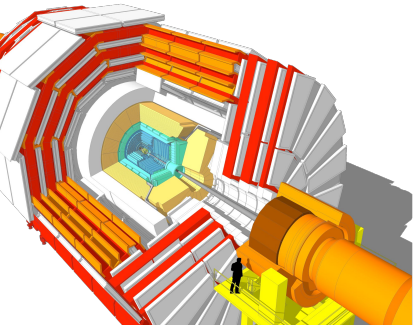
Sensors, sensing elements, sources of information, …

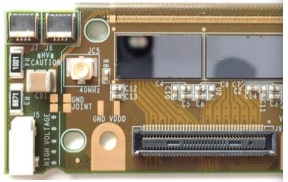
Amplification, discrimination, digitalization, …

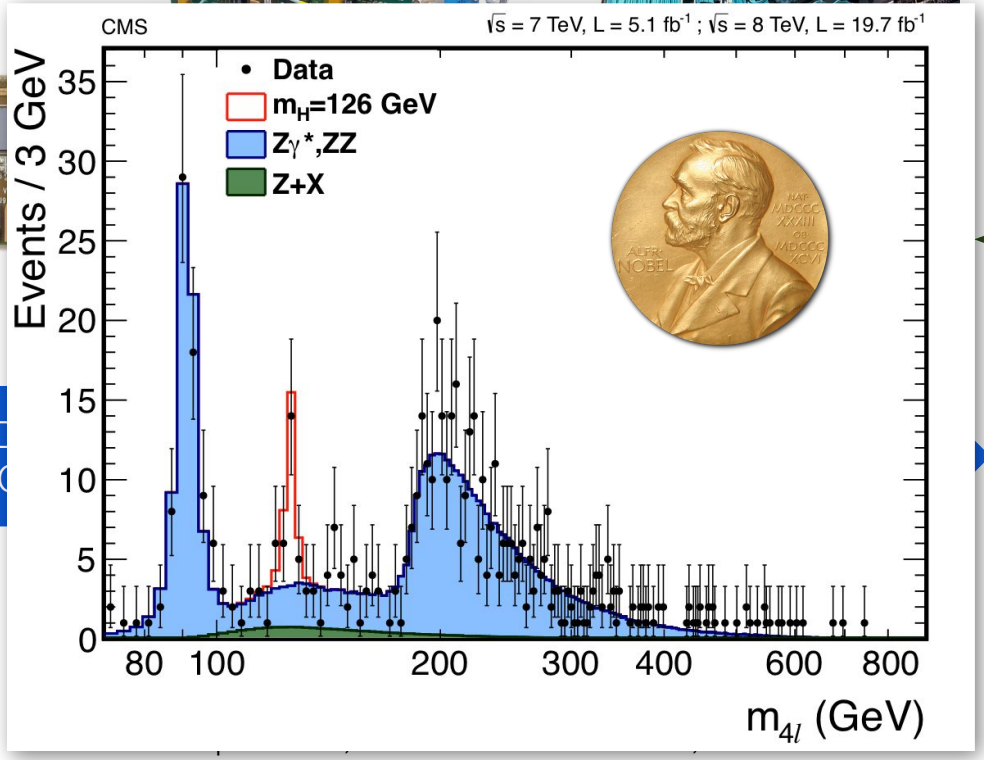
Data concentration, low-to-high level information, fast computations, …

Building of higher-level data, high-level computations, filtering, selection, …

FROM SENSORS TO DATA







SENSORS

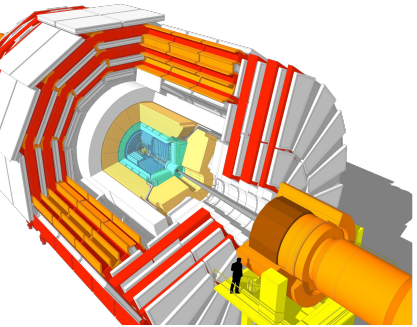
FRONTEND ELECTRONIC

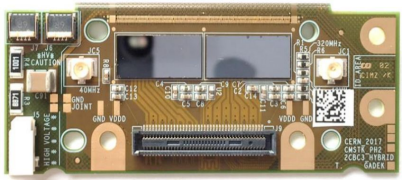
DATA

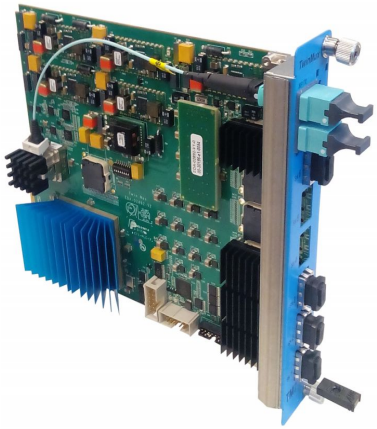
Sensors, sensing elements, sources of information, …

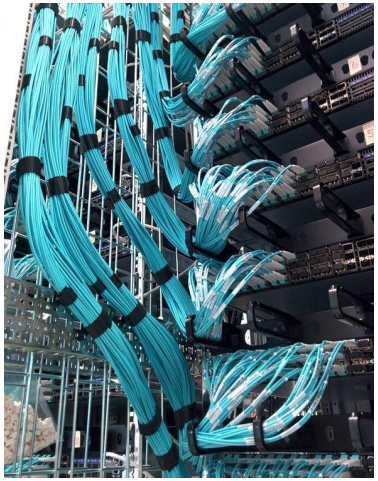
Amplification, discrimination, digitalization, …

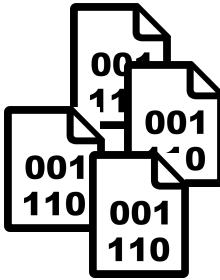
FROM SENSORS TO DATA











SENSORS → FRONTEND ELECTRONICS → READOUT ELECTRONICS → TRIGGER & DAQ → "RAW" DATA

Sensors, sensing elements, sources of information, …

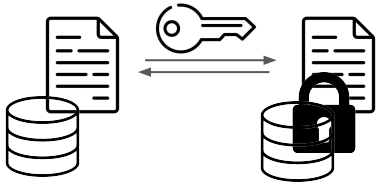
Amplification, discrimination, digitalization, …

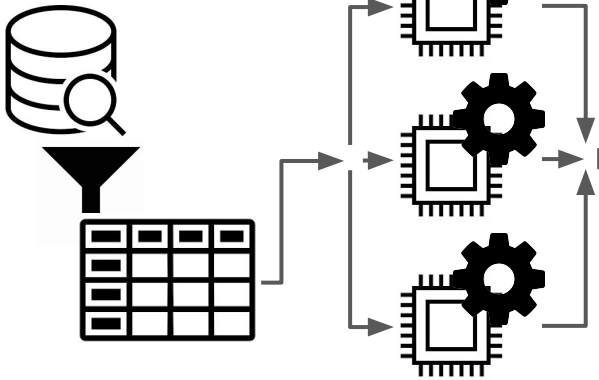
Data concentration, low-to-high level information, fast computations, …

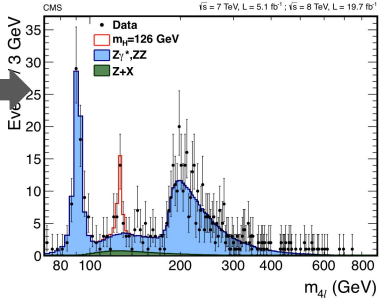
Building of higher-level data, high-level computations, filtering, selection, …

"DATA ACQUISITION SYSTEM"

FROM DATA TO INFORMATION







"RAW" DATA → STORAGE → RELIABILITY SECURITY → QUERYING FILTERING → PROCESSING → INFORMATION

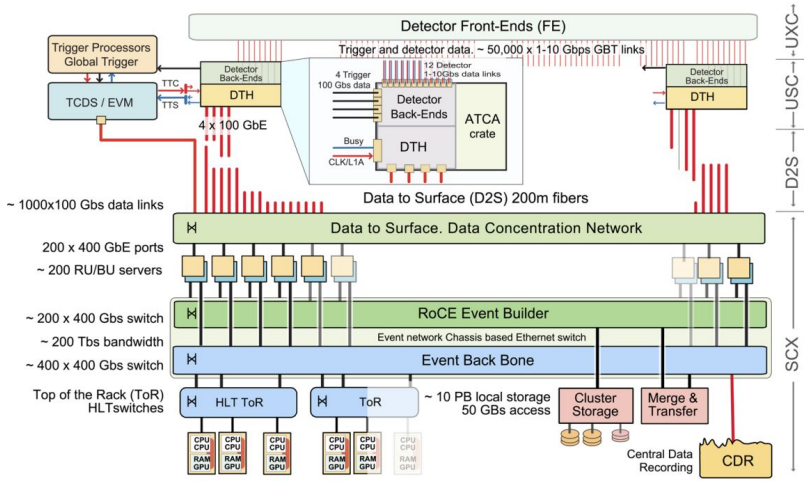
STORAGE	RELIABILITY SECURITY	QUERYING FILTERING	PROCESSING
From data to datasets, storage, file-system, …	Data preservation, reliability, authentication, authorization, …	Querying and filtering of datasets, data-mining, feature-enrichment, …	Higher-level computation, training and testing of algorithms, …

"COMPUTING MODEL"

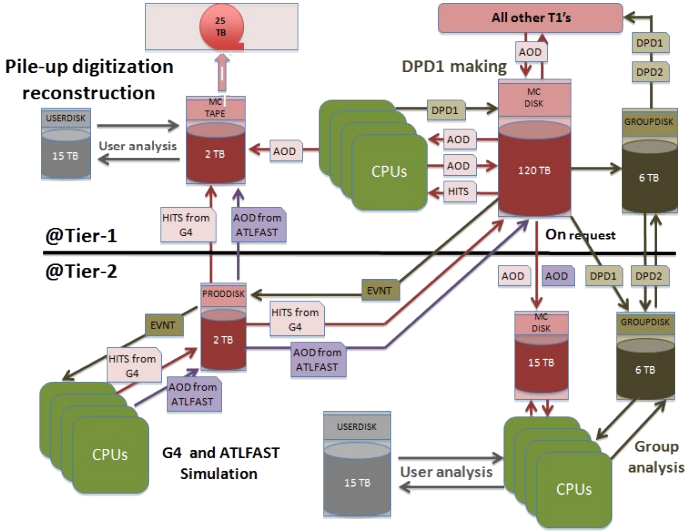
DAQ AND COMPUTING MODELS IN PHYSICS

SENSORS | FRONTEND ELECTRONICS | READOUT ELECTRONICS | TRIGGER & DAQ | STORAGE | RELIABILITY SECURITY | QUERYING FILTERING | PROCESSING

"DATA ACQUISITION SYSTEM"



"COMPUTING MODEL"



NOT ONLY PHYSICS EXPERIMENTS

- RPM
- Power / Current
- Bearings' temperature
- Shaft vibrations
- ...

- Monitoring
- Residual Useful Life
- Working parameters' optimization
- Anomaly detection
- ...

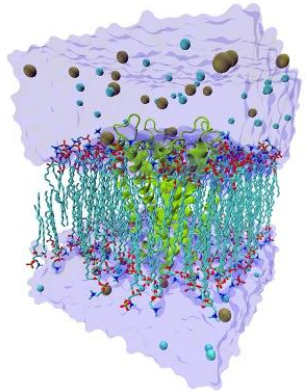
DATA SOURCES

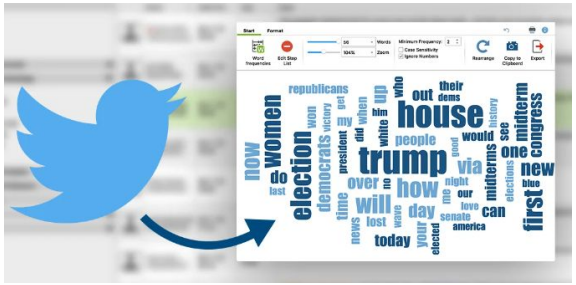
INFORMATION

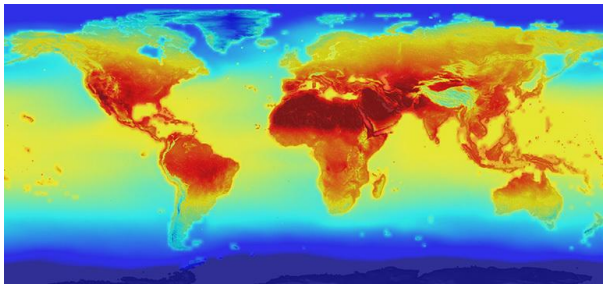
SENSORS | FRONTEND ELECTRONICS | READOUT ELECTRONICS | TRIGGER & DAQ | STORAGE | RELIABILITY SECURITY | QUERYING FILTERING | PROCESSING

DIFFERENT SOURCES - SIMILAR PATTERNS

- Sensors
  (e.g.: experiment / IoT / …)
- Simulated data
  (e.g.: MonteCarlo / …)
- Text
  (e.g.: log-files / documents / …)
- Heterogeneous sources
  (e.g.: GPS location + image + …)
- …





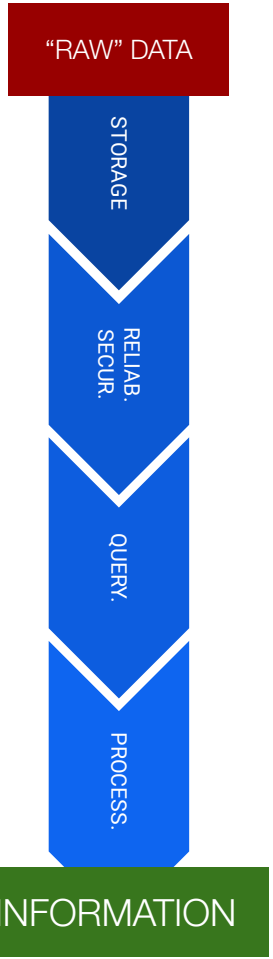


ANY KIND OF DATA SOURCE → "RAW" DATA → STORAGE → RELIAB. SECUR. → QUERY. → PROCESS. → INFORMATION

THE SIMPLEST "COMPUTING MODEL"

Let's assume we have as "RAW" data a few files produced by a given source

perhaps collected by an experiment DAQ system, or simulated events of a given model, or from users



A few .csv files of O(~1 MB) each

Some high-level result of your data processing (e.g.: "final" plot / result of ML application)

THE SIMPLEST "COMPUTING MODEL"

Let's assume we have as "RAW" data a few files produced by a given source

A few .csv files of O(~1 MB) each

Stored on the file system of our own machine

No data replication

No access to data for other user / No security if accessible by others

Data can be queried and filtered by accessing it directly from the file system

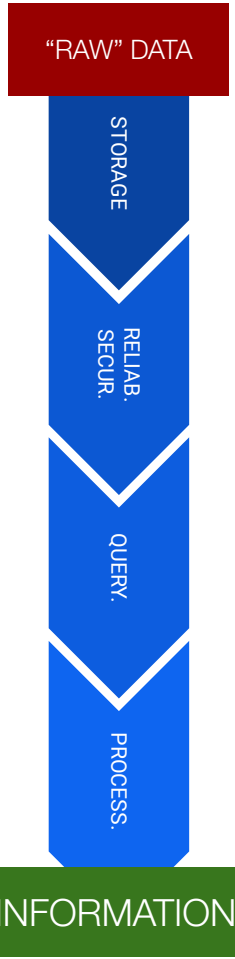
Pre-processed data can be saved and stored

1- or multi-staged single-core applications (e.g. using Python, pandas, etc)

Some high-level result of your data processing (e.g.: "final" plot / result of ML application)

THE SIMPLEST "COMPUTING MODEL"

Let's assume we have as "RAW" data a few files produced by a given source



A few .csv files of O(~1 MB) each

Sto

No

No

Da

Pre

1- or multi-staged single-core applications (e.g. using Python, pandas, etc)

HOW DOES THIS SCALE?

- With the amount of raw data
- With the need of securing and preserving the data for long periods of time
- With the number of users accessing the data at the same time
- With the complexity of the processing
- …

Some high-level result of your data processing (e.g.: "final" plot / result of ML application)

A LOT OF DATA!


SKA Science Archive
searches on Google 98PB
uploads to facebook 180PB
LHC Raw Data 100 PB
LOFAR Long Term Archive 23PB
YouTube 15PB
CERN 15PB
6PB
Census 4PB
NASDAQ 3PB
LIBRARY OF CONGRESS 5PB
SKA Phase1 Science Archive 300PB
PER YEAR
1 Petabyte

Management and

Analysis of

Physics

Datasets

Management and

Analysis of

Physics

Datasets

Dataset and metadata

Structured vs unstructured data

Management and

Analysis of

Physics

Datasets

Data Storage and Preservation

File Systems

Databases

Management and
Processing
~~Analysis~~ of

Physics

Datasets

Data *processing* in a broader sense than *data analysis (LCP mod. A and B)*

Parallel programming

Distributed architectures

WHAT THIS COURSE IS ABOUT

Management and Processing ~~Analysis~~ of

***(not exclusively)* Physics**

Datasets

PROGRAM

- **Data Management**
  - Datasets and Data Structure
  - Data Storage and Scalability
  - Data Reliability & Security
  - Local and Distributed File Systems
  - Databases and Relational DBs → ***mySQL***
  - Distributed and non relational DBs

Mid-semester Data Management Written Test

- **Data Processing**
  - From single- to parallel- to distributed-processing
  - Basics of parallel programming
  - Intro to distributed processing
  - Hadoop & the Map-Reduce programming paradigm
  - Distributed computing frameworks → ***Apache Spark*** *+ (a brief introduction to)* ***Dask***
  - Distributed streaming platforms → ***Apache Kafka***

Containers → ***Docker***

GOALS

- The course is intended as an **introduction and** an **overview** of recurrent ideas and issues that are going to be found extensively in Physics and in Data Science

- The main goal is to expose you to terms and concepts commonly dealt with when working with data **outside the scope of small projects**

- The class aims at providing a basic **knowledge of tools commonly used in real-life applications** (both in- and outside academia), such as databases and large scale computing frameworks

CAVEATS AND PREREQUISITES

- What this course **IS** intended to be:
  - An overview of these topics with a "bird's-eye view" of the underlying *(usually vast)* complexity
  - With some "deep-dives" into the most relevant topics
  - Including hands-on sessions (live coding + discussion) to grasp the basics of few selected tools

- What this course **IS NOT** intended to be:
  - An introduction to programming
  - Or, an extensive and exhaustive "computer-science level" course on all these topics

---

- What I'm going to **take for granted**:
  - Familiarity with the Python language
  - Some basic knowledge of Unix shell commands
  - Anything git-related

FINAL PROJECT / EXAMS

- ≥ 85% lesson attendance is required to access the exam

- The exam will be comprised of two parts:
  1. Data Management : written exam on data management and database topics
     ⇒ open (argumentative) questions and exercises on databases
  2. Data Processing : processing of a dataset using distributed computing techniques
     ⇒ 3/4-people group project with presentation and oral discussion

- The overall MAPD Mod. B grade will be the 50%-50% combination of both parts (iff each ≥ 18/30)

- The Data Management and Data Processing exams can be taken independently
  - e.g. Data Management first, then Data Processing, or vice versa

- A mid-semester date will be arranged as an early ("extra") chance to pass the Data Management test

- Next semester's exam dates will be on:
  26-27 June **10-11 July** **04-05 September** **18-19 September**

ADMINISTRATIVA

- Lessons will be on these unfortunate timeslots:
  - **Wed 08:30-10:30**
  - **Thu 08:30-10:30**

  } All lessons will be held in **PRESENCE** in **LabP104**

- Slides and lab. sessions' material has been prepared with the help of many nice people, especially:
  - Matteo Migliorini (teaching assistant for this year)
  - Stefano Campese
  - Federico Agostini

- Official communications *(from me to all of you)* will be sent through Moodle Announcements
  - https://stem.elearning.unipd.it/course/view.php?id=7989
  - All course material (slides and links) will be uploaded to the Moodle page as well

- In case of any need for direct communication *(from you to me)*, ping me before/after class, contact me via email, or just knock at the door:
  - jacopo.pazzini@unipd.it
  - Room 134, Via Marzolo 8