Esperienza di un'analisi dati CMS nell'INFN "Analysis Facility" framework

Tommaso Diotalevi *(Università di Bologna, INFN Bologna)*

Workshop sul calcolo nell'INFN
Loano, 22-26 Maggio 2023

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

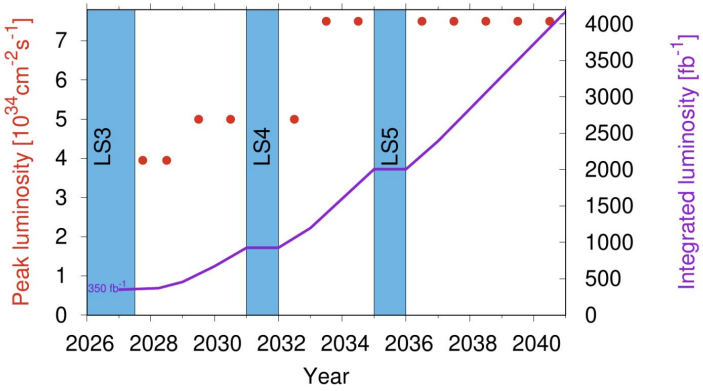
Outline

- Motivazioni principali
- Come funziona l'Analysis Facility (AF)
- Esperienze di analisi dati su Analysis Facility: l'analisi Heavy Neutral Lepton (HNL)
  - Workflow dell'analisi
  - Porting su Analysis Facility
- Pro e contro dell'Analysis Facility, vista da un utente
- Sinergia con ICSC: Centro Nazionale di ricerca in HPC, Big Data e Quantum Computing
- Prospettive future e conclusioni

Motivazioni principali

R&D per analisi ad High Luminosity LHC:

- ottimizzare uso di CPU e storage
- promuovere l'utilizzo di formati di dati ridotti (NanoAOD)
- sperimentare nuovi paradigmi di analisi
-



Fare analisi in maniera efficiente sarà cruciale!

A tale scopo:

- Test di software basato su programmazione dichiarativa e workflow interattivi
- Calcolo distribuito sulle risorse disponibili

Come funziona l'Analysis Facility (AF)

Istituto Nazionale di Fisica Nucleare

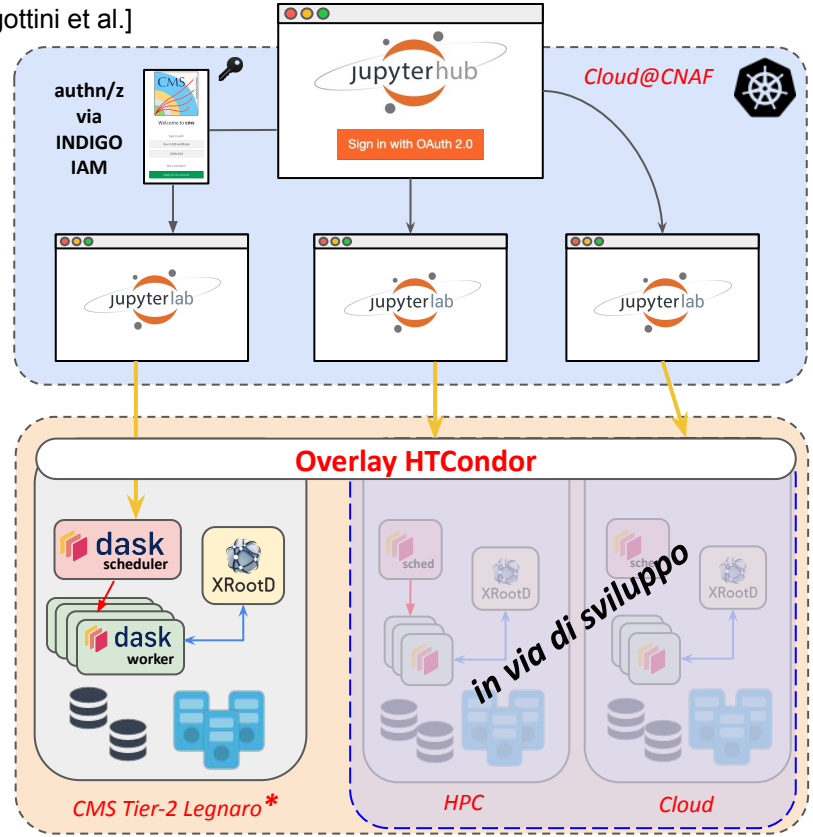
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

[Documentazione](#) - Aspetti più tecnici già presentati nel [talk CCR Paestum 2022](#)

[D.Ciangottini et al.]

- Accesso al singolo HUB e autenticazione via token (INDIGO-IAM)
- Basato su tecnologie standard dell'industria
- Kernel python configurabile
- Containerizzazione dell'ambiente di lavoro specifico

- Overlay basato su HTCondor (utilizzabile anche standalone)
- Libreria DASK (python) per calcolo distribuito
  - Scalare l'esecuzione da 1 a N cores
- Possibilità di implementare su risorse eterogenee (HTC/HPC/[Cloud - "Datacloud"](#))
- Interfacciabile con WLCG (xrootd, WebDAV, ...)



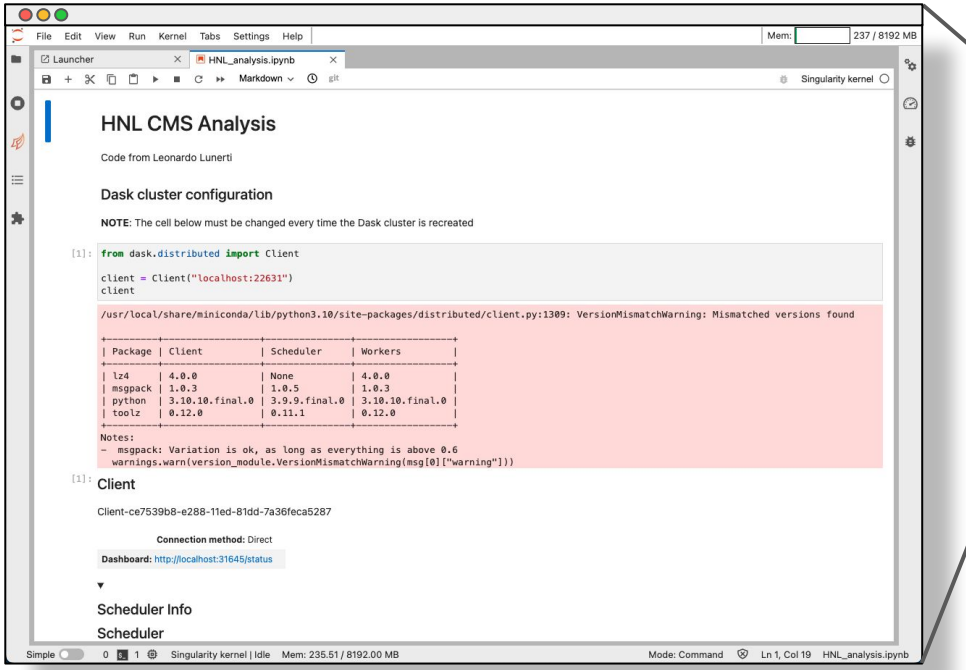
* 3 nodi, ognuno con 32 logical CPU - 128 GB RAM - 1 Gb/s

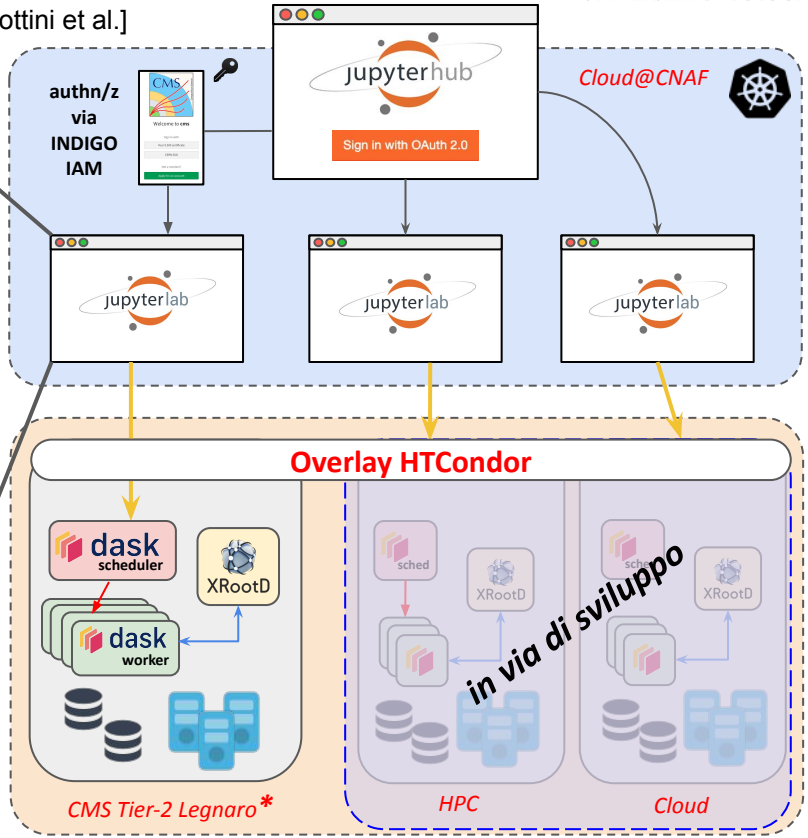
Come funziona l'Analysis Facility (AF)

Documentazione - Aspetti più tecnici già presentati nel talk CCR Paestum 2022

[D.Ciangottini et al.]

Cosa vede l'utente





* 3 nodi, ognuno con 32 logical CPU - 128 GB RAM - 1 Gb/s

Come funziona l'Analysis Facility (AF)

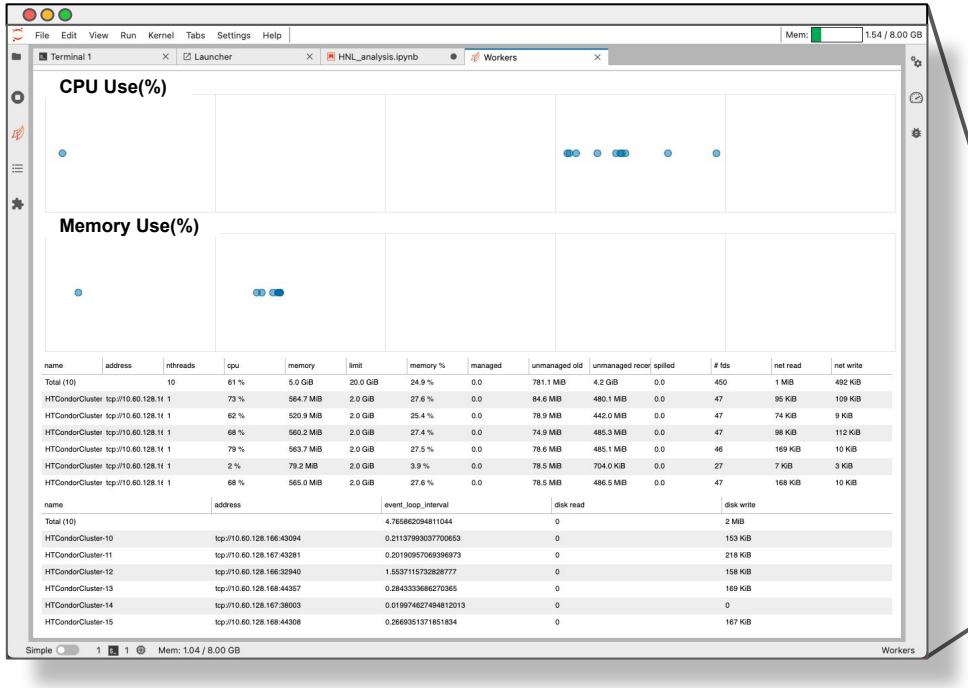
Istituto Nazionale di Fisica Nucleare

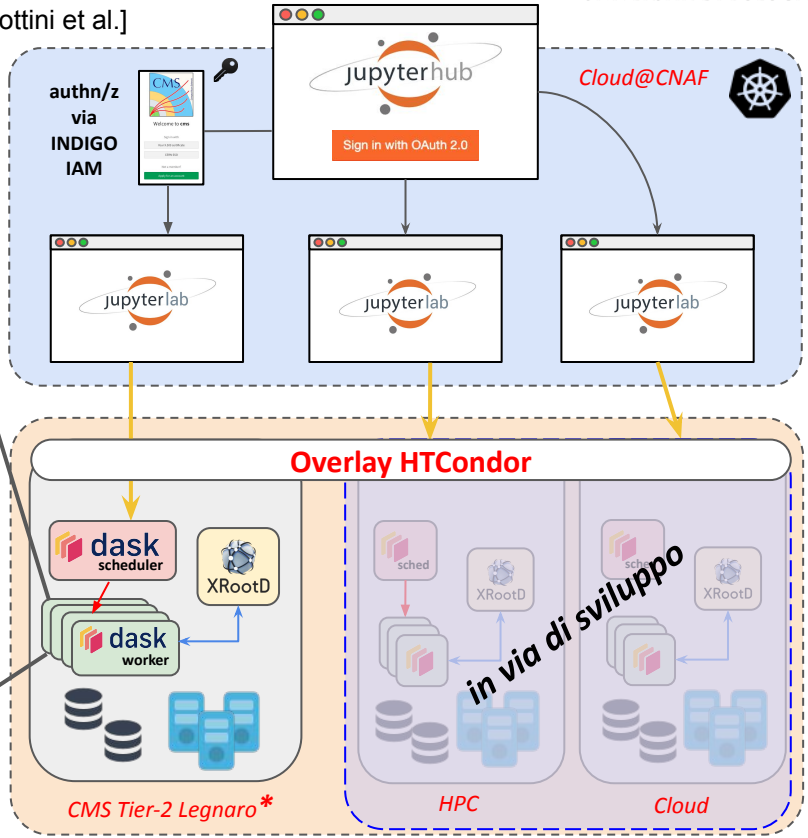
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

[Documentazione](#) - Aspetti più tecnici già presentati nel [talk CCR Paestum 2022](#)

[D.Ciangottini et al.]

Cosa succede dietro le quinte





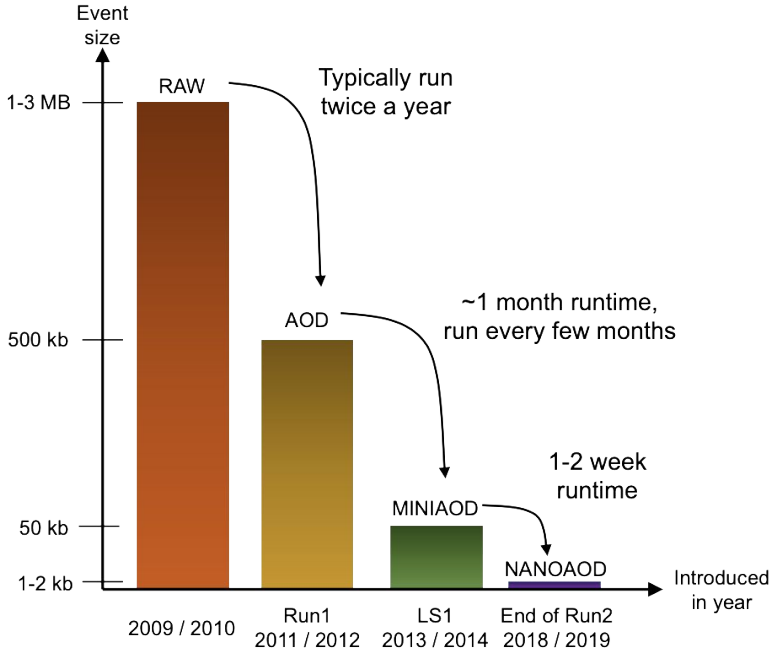
* 3 nodi, ognuno con 32 logical CPU - 128 GB RAM - 1 Gb/s

Processamento dei dati in CMS

La totalità dei dati CMS, prodotti ufficialmente, vengono salvati in ROOT file, in diversi formati:

- **RAW**: Una raccolta degli output dell'elettronica del detector. Dimensione di un evento: 1-3 MB
- **AOD**: Analysis Object Data. Trasformazione dei dati RAW in entità utilizzabili nelle analisi come jet, muoni, elettroni,etc… Dimensione di un evento: 500kB
- **MiniAOD**: Riduzione della dimensione degli AOD, rendendoli più compatti al costo di perdere informazione (e.g. azzerando bit di numeri float). Dimensione di un evento: 50kB
- **NanoAOD**: Ulteriore riduzione dei MiniAOD, salvato su un ROOT file colonnare. Questo nuovo formato, utilizzando tipi fondamentali (int, float), esce dall'ecosistema CMS e richiede un numero ristretto di dipendenze per essere analizzato. Dimensione di un evento: 1-2 kB.

Principalmente, le analisi dati in CMS utilizzano i formati MiniAOD e NanoAOD, in base alle variabili fisiche che necessitano.



(immagine presa da: link to CHEP19 contribution)

ROOT RDataframe

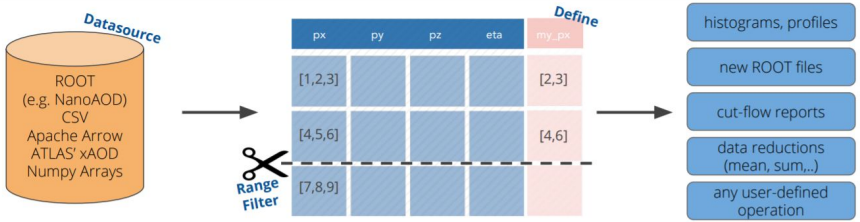
RDataFrame (RDF) è l'interfaccia di alto livello di ROOT per l'analisi dei dati archiviati in TTree, CSV e altri formati di dati. È caratterizzata da:

- multi-threading;
- ottimizzazioni di basso livello (parallelizzazione e caching).

I calcoli sono espressi in termini di una catena di azioni e trasformazioni, che costituiscono un grafo computazionale.

L'esecuzione del grafo può essere effettuata in maniera distribuita sfruttando back-end quali Spark e Dask.

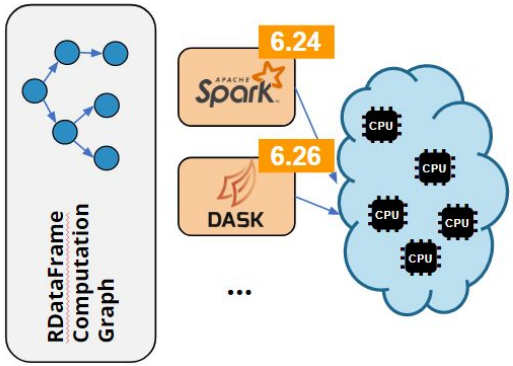
Grazie all'estensione "Distributed" di RDF, attiva in via sperimentale.




```
# enable multi-threading
ROOT.EnableImplicitMT()
df = ROOT.RDataFrame(dataset)
```

```
df = df.Range(2)
      .Define("my_px", "px[eta > 0]")
```

```
# filled in a single loop
h1 = df.Histo1D("my_px", "w")
h2 = df.Histo1D("px", "w")
```



Immagini riprodotte da: talk PyHEP 2021

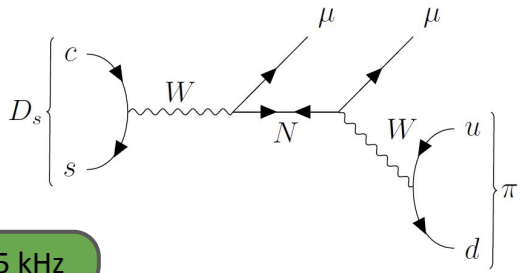
Esperienze di analisi dati su AF

Prima attività di “benchmarking” di una reale analisi dati in CMS

- Scattering (VBS) di due bosoni W (same-sign) in tau adronico e leptone leggero (INFN e Università, Perugia)
  - Middle-size analysis (~1-2 TB) based on NanoAOD data format
  - Porting e benchmarking da analisi “legacy” ad analisi “RDataFrame” con supporto a calcolo distribuito con Dask Link presentazione CHEP23 [D.Spiga et al.]

Un’ulteriore analisi in corso in CMS, con caratteristiche differenti:

- Analisi HNL nella ricerca neutrini pesanti in decadimenti D_S:
  - Ricerca di un leptone pesante neutro N (*bump hunt*), proveniente dal mesone D_S, nello stato finale contenente un **μ**, un **π** ed un **μ** aggiuntivo, proveniente dal **W** iniziale
  - Analisi nativa in python, con adozione dalle origini di RDataFrame
  - Analisi sul dataset intero **MiniAOD** **B-Parking**



Informazioni per effettuare refit dei vertici assenti nei NanoAOD

Esplorazione di un pattern di accesso ai dati diverso, con rate di trigger fino a 5 kHz

Workflow dell'analisi HNL

1. Data skimming e preprocessing

Input: Formato **MiniAOD**, proveniente dal dataset B-Parking (41.5 fb^{-1}). → **Size**: ~700TB (Data), ~14TB (MC); **Eventi**: ~12mld (Data), ~300M (MC)

Operazioni: pre-selection, data reduction, vertex refitting.

Output: flat ntuple (i.e. simili a NanoAOD), momentaneamente salvate al Tier-2 di Legnaro. → **Size**: ~0.5TB (Data) - < 1TB (Data+MC)

2. Selezione del miglior candidato HNL

Input: Output step 1.

Operazioni: Selezione del miglior candidato HNL.

Output: Flat ntuple, più leggere di quelle in input (scelta dovuta ai tempi di esecuzione).

3. Event selection e categorizzazione

Input: Output step 2.

Operazioni: Categorizzazione, applicazione criteri di selezione per categoria, applicazione pesi (PU, MC, SFs)

Output:
- Flat ntuple per categoria;
- Istogrammi di controllo;
- File CSV per compatibilità con tool esterni

4. Tool statistici

Input: Output step 3.

Operazioni: Fit e analisi statistica (e.g. Combine)

Output: Risultati fisici (istogrammi, limiti, …)

Creazione di un container singularity con tutte le dipendenze necessarie per l'esecuzione dell'analisi sui worker Dask.

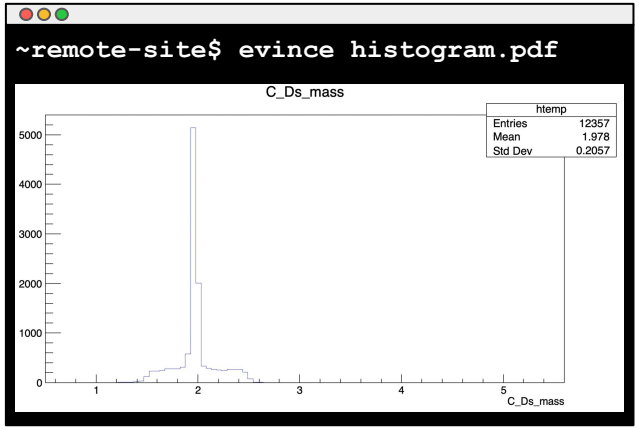
- CMS Distributed analysis (CRAB)
- Analisi Interattiva (Analysis Facility con Dask)
- Analisi Interattiva (Analysis Facility in locale)

Porting su notebook interattivo con Dask

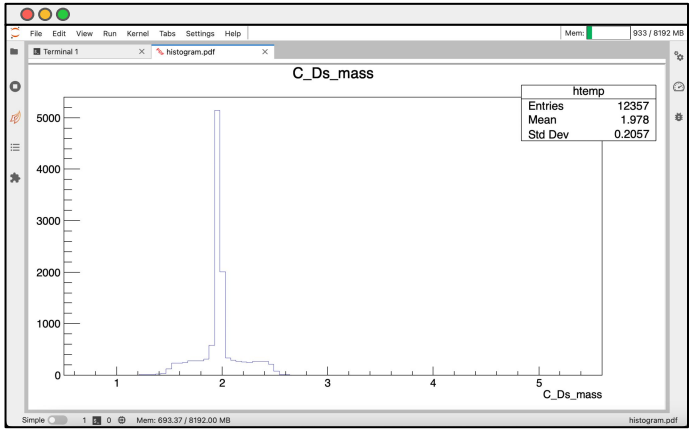
- Porting da RDataFrame “locale” (nello specifico, il Tier-3 di Bologna) ad RDataFrame distribuito, con supporto a DASK.

2. Selezione del miglior candidato HNL

Variable estratta dal tree in output dello Step 2



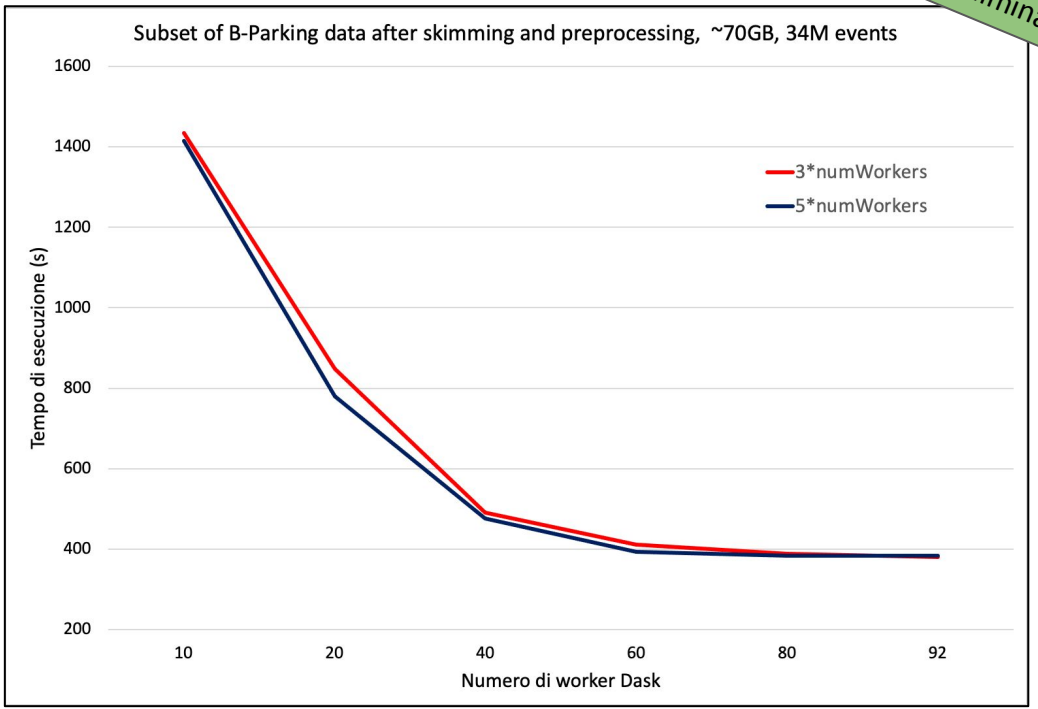
porting



Primi test di speedup analisi

- Con l'utilizzo di Dask, possibilità di scalare l'analisi su N nodi.

preliminare



- All'aumentare del numero di nodi utilizzati, diminuzione del tempo di esecuzione.
- Granularità nella configurazione di RDataFrame in lettura: 3*numWorkers e 5*numWorkers.

Pro e contro dell'AF - Vista da un utente

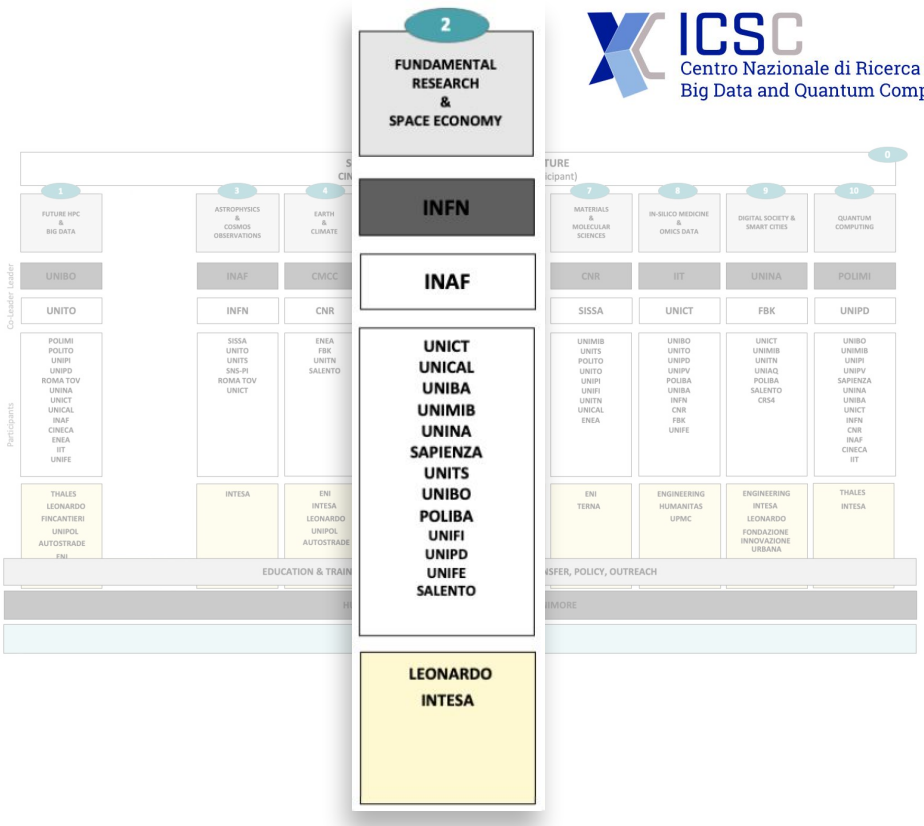
Pro:

- Tutto il codice dell'analisi in un **unico notebook**, con controllo dell'esecuzione interattivo
- Scalare l'esecuzione da 1 a N cores, con la **libreria Dask**, in maniera trasparente
- Libertà di utilizzo delle proprie librerie, in base alle esigenze specifiche di ogni use case (e.g. ROOT, Combine, moduli Python)
- **Indipendente dall'esperimento**: nato come progetto all'interno di CMS, estendibile a tutte le collaborazioni

Contro:

- Framework non ancora in produzione: riscontrati talvolta errori prontamente risolti con patch dedicate
- Necessaria esperienza su containers per impacchettare l'ambiente di lavoro in un'immagine Singularity (apptainer): non immediato per tutti gli utenti, risolvibile con la documentazione dettagliata
- L'interfaccia RDataFrame "distribuita" è ancora in beta: tutte le analisi dati HEP, che hanno un utilizzo nativo di RDataFrame "base", ed utilizzano metodi non supportati dalla controparte distribuita potrebbero riscontrare showstopper importanti. → Fondamentale la collaborazione tra devs e ROOT team

Sinergia con il Centro Nazionale ICSC



ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Spoke 2

Diviso in 6 Work Packages (WP):

WP1: Theoretical Physics.
WP2: Experimental High Energy Physics.
WP3: Experimental Astroparticle Physics.

WP4: Boosting the computational performance of Theoretical and Experimental Physics algorithms.
WP5: Architectural Support for Theoretical and Experimental Physics Data Management on the Distributed CN infrastructure.
WP6: Cross–domain Initiatives

Sinergia con il Centro Nazionale ICSC

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Spoke 2

Lavoro in corso (su attuali risorse fuori da ICSC)

- Valutare le **metriche di benchmarking** (timing, uso di risorse).
- Estendere il porting a tutti i dati dell'analisi, identificando bottleneck e stress-test del framework.

WP2: Experimental High Energy Physics.

Porting e benchmarking di un'analisi dati fisica nel framework "Analysis Facility"

WP5: Architectural Support for Theoretical and Experimental Physics Data Management on the Distributed CN infrastructure.

- Test del framework e confronto con gli sviluppatori per feedback
  (setup corrente + setup in corso WP5 da UniNa)
- Contributo alla documentazione

Conclusioni

- La sfida presentata dalle prossime fasi di LHC, richiede un forte lavoro di sviluppo di nuovi strumenti per fare analisi dati in modo più efficiente e con strumenti moderni
- In questo senso nasce l'INFN-CMS Analysis Facility, sfruttando standard in Data Science come Jupyter (per approccio interattivo) e Dask (per calcolo distribuito), assieme a strumenti più moderni di analisi nel mondo HEP come RDF
- Lo use-case di un'analisi HNL nella ricerca di neutrini pesanti dal decadimento del mesone D_S:
  - analisi già in RDF "locale", con porting necessario per l'esecuzione con Dask e RDF distribuito
  - lavoro in corso: studi di benchmarking e upscaling, per evidenziare eventuali bottleneck o debugging con dev team
- ICSC: WP2 in sinergia con WP5 per l'upscaling della facility con maggiori risorse

Ringraziamenti:

- **Gruppo R&D CMS Computing Italia**, in particolare Daniele Spiga (INFN), Diego Ciangottini (INFN) e Tommaso Tedeschi (Università di Perugia, INFN).

- **CMS Bologna**, in particolare Alessandra Fanfani (Università di Bologna, INFN) e Leonardo Lunerti (INFN).

This work is partially supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU

Backup

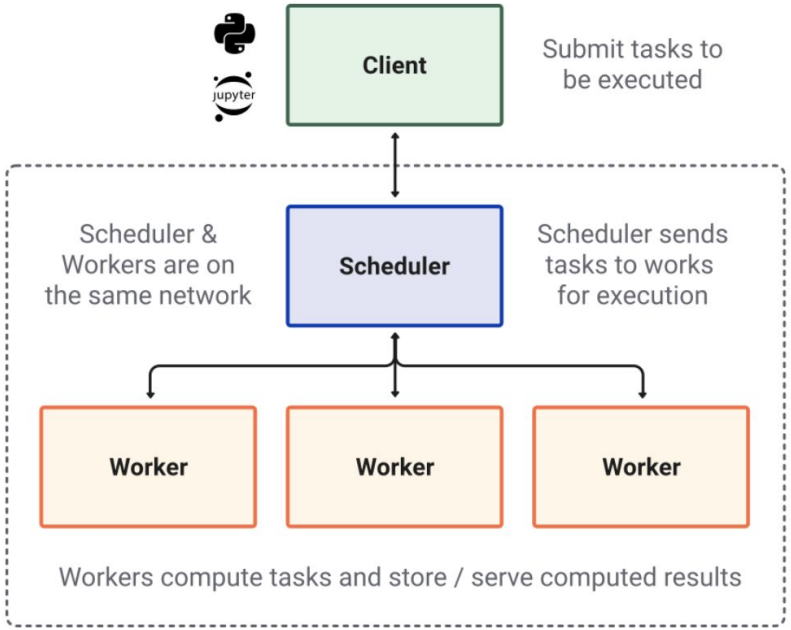
Collaboration with ROOT devs

- **introduction of new useful distributed RDataFrame features**:
  - Delegation of automatic file-splitting to workers
  - DefinePerSample method for sample-wise column definitions
  - The Vary and VariationFor methods for systematic variations
  - The Redefine method for column redefinition
  - The resilience to empty TTrees in a chain
  - Additional monitoring features for benchmarking purposes

Dask

There are many parts to the “Dask” the project:

- Collections/API also known as “core-library”.
- Distributed – to create clusters
- Integrations and broader ecosystem



Dask Cluster

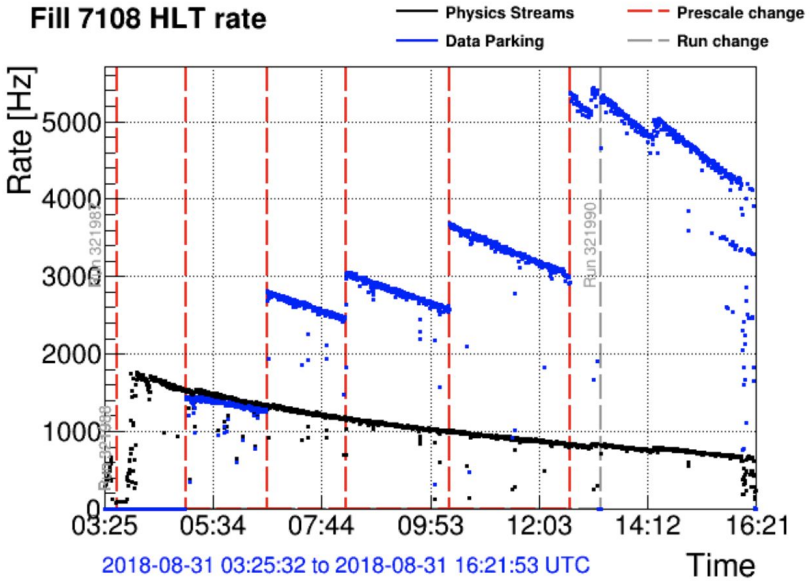
B-parking dataset

Il dataset B-Parking, sottogruppo del flusso di dati *parking,* consiste in dati RAW salvati su Tape al CERN (in singola copia) e ricostruiti una-tantum durante la pausa natalizia (fino al formato **MiniAOD**, in quanto il formato NanoAOD taglierebbe informazioni necessarie allo studio di eventi con b quarks).

Gli eventi vengono salvati con picchi di trigger rate di 50 e 5 kHz a L1 e HLT, rispettivamente, e scritti (parked) su Tape ad un rate medio di 2 GB/s.

- Il rate del Physics Stream (in nero) decade partendo da 2kHz;
- Il rate del Parking Stream (in blu) aumenta a gradini, con i cambiamenti del prescaling durante il run, raggiungendo fino a 5kHz.

Il dataset B-parking è stato popolato durante il 2018. Contiene ~10 miliardi di decadimenti unbiased di adroni contenenti quark b, con una luminosità integrata pari a 41.5 ± 1.0 fb^{-1}

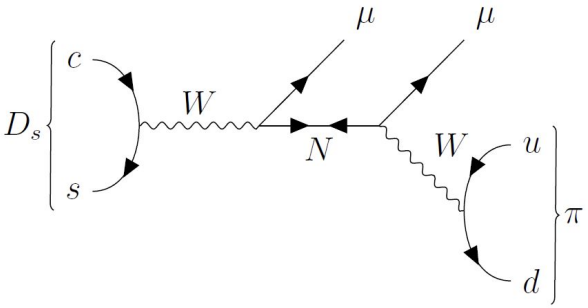


Analisi HNL - ricerca neutrini pesanti in decadimenti D_s

Talk di Leonardo Lunerti al CMS working meeting B-Physics: link (CMS restricted)

Ricerca di un leptone pesante neutro N (*bump hunt*), proveniente dal mesone **D_s**, nello stato finale contenente un **μ** e un **π**.

Alla firma sperimentale viene inoltre considerato un **μ** aggiuntivo, proveniente dal **W** iniziale. I due muoni possono avere sia stesso-segno che segno-opposto.



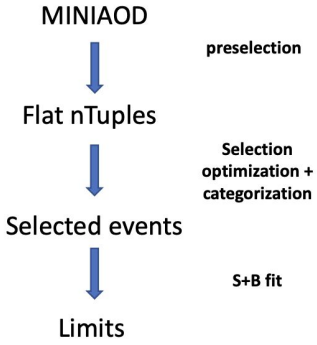
Strategia:

- Analisi sul dataset intero **MiniAOD B-Parking** Ultra-Legacy re-reco (RunII).
- Discriminazione segnale-fondo:
  - Campioni QCD μ-enriched, per modellare la shape del fondo;
  - Campioni di segnale per m_{HNL}=1.0,1.5 GeV e $c\tau_{HNL}$=10,100,1000 mm.

Stima del background data-driven, per ogni ipotesi di massa del neutrino.

Stima del segnale rispetto ad un canale di normalizzazione: $D_S \to \phi(\to \mu\mu)\pi$

Questa analisi nasce in PyROOT, con adozione dalle origini di RDataFrame.

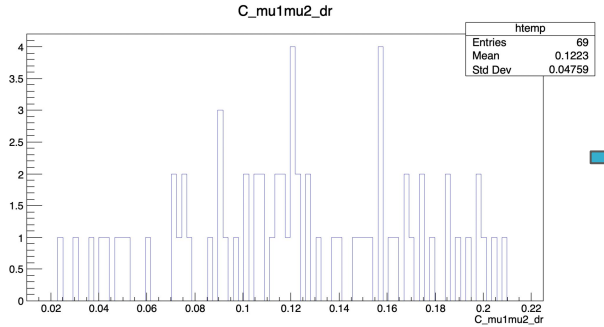


Porting su notebook interattivo con Dask

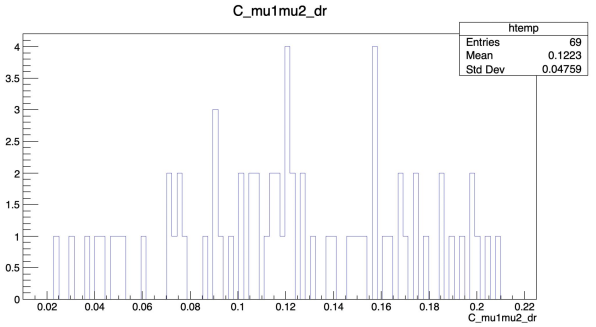
3. Event selection e categorizzazione

Variable estratta dal tree in output dello Step 3

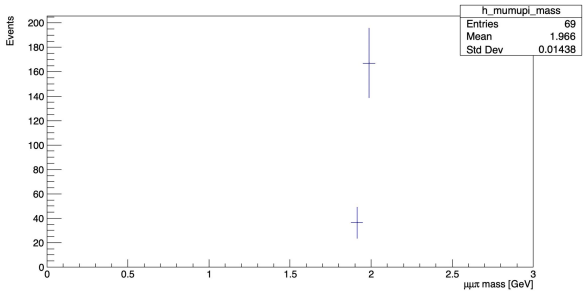
workflow originale



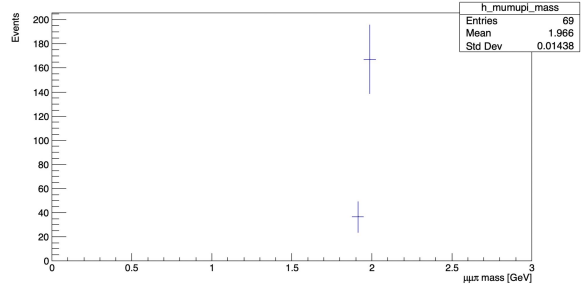
porting



Variable estratta dagli istogrammi in output dello Step 3



porting



workflow su AF con DASK