ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Attività UniBO per esplorazione analisi interattiva

Tommaso Diotalevi

WP5 Meeting 12/04/2023

Use case utilizzato

Analisi dati, in corso, della Collaborazione CMS: "*Heavy Neutral Lepton (HNL) search in D_S decays*".

Lavoro portato avanti da **Leonardo Lunerti**, PhD dell'Università di Bologna, che ringrazio per la disponibilità e per avermi fornito il codice in maniera riproducibile e sequenziale!

Dataset utilizzato:

B-parking dataset, contenente eventi RAW salvati immediatamente su Tape storage, durante la fase di High-Level Trigger (HLT). Attraverso questa metodologia è possibile salvare eventi con rate maggiori rispetto a quelli di trigger, in quanto non viene fatta una ricostruzione prompt.

Enorme mole di dati in input.

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

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Workflow dell'analisi

CMS Distributed Analysis (CRAB)

Interactive Analysis (Analysis Facility)

1. Data skimming e preprocessing

Tempo di esecuzione: ≈ 2-3 giorni (una tantum)

Input: Formato "heavy" (MiniAOD), proveniente dal dataset B-Parking. → **Size**: ≈ 700TB (Data)

Output: flat ntuple, momentaneamente salvate al Tier-2 LNL. → **Size**: ≈ 0.5TB (Data) - <1TB (Data+MC)

2. Selezione del candidato HNL migliore

Tempo di esecuzione 2.+ 3. pre-AF ≈ qualche ora

Input: Output step 1.

Output: Flat ntuple, più leggere di quelle in input.

3. Analisi e calcolo limiti

Input: Output step 2.

Output: Risultati fisici (istogrammi, limiti, …)

Step considerati per use-case Analysis Facility*

* Workflow dell'analisi ad ora non definitivo. Possibile futura eliminazione di step 2, con l'utilizzo di tutti i candidati HNL.

A che punto ero?

- Esecuzione corretta del workflow dell'analisi originale su HTCondor dell'Analysis Facility. Per questioni di semplicità, il workflow è stato girato per ora solo su un segnale Monte Carlo piuttosto leggero.
- L'esecuzione su Dask del workflow dell'analisi originale è quasi completo: ultimi debug in corso, per problemi con immagine custom (ROOT 6.28) -> vedi prossima slide.
    - Possibile problema futuro: altre mancanze di RDF Experimental che potrebbero bloccare l'esecuzione del codice originale. Nel caso: trovare workaround nell'analisi o richiedere funzionalità RDF Experimental.

Immagine singularity con ROOT 6.28

- Immagine singularity/apptainer, con framework ROOT alla versione 6.28/00.
- Cronologia delle versioni:
  - 🟥 v1.0.0: ROOT compilato correttamente da source, sopra un'immagine Jupyterlab con versione DASK worker: 2022.9.2
    Problema: Incompatibilità con la versione del DASK scheduler (2021.11.1), su un'immagine utilizzata dentro l'AF durante l'avvio del cluster Dask.
  - 🟥 v1.1.0: ROOT compilato correttamente da source, sopra un'immagine diversa Jupyterlab con versione inferiore DASK worker (2021.11.1), per farla combaciare con lo scheduler DASK.
    Problema: ROOT 6.28 supporta una versione minima di DASK pari a 2022.8.1, per RDF Distributed.
  - ☑ v1.2.0: ROOT installato attraverso Conda, sopra un'immagine con versione dei DASK worker 2022.9.2. La stessa immagine viene anche usata dall'AF per configurare il DASK scheduler (2022.9.2). Necessaria patch all'avvio AF.
  - ☑ v1.3.0 (latest): miglioramento dell'immagine, con dipendenze necessarie per girare correttamente il notebook.

Workflow analisi su Jupyter Notebook con DASK

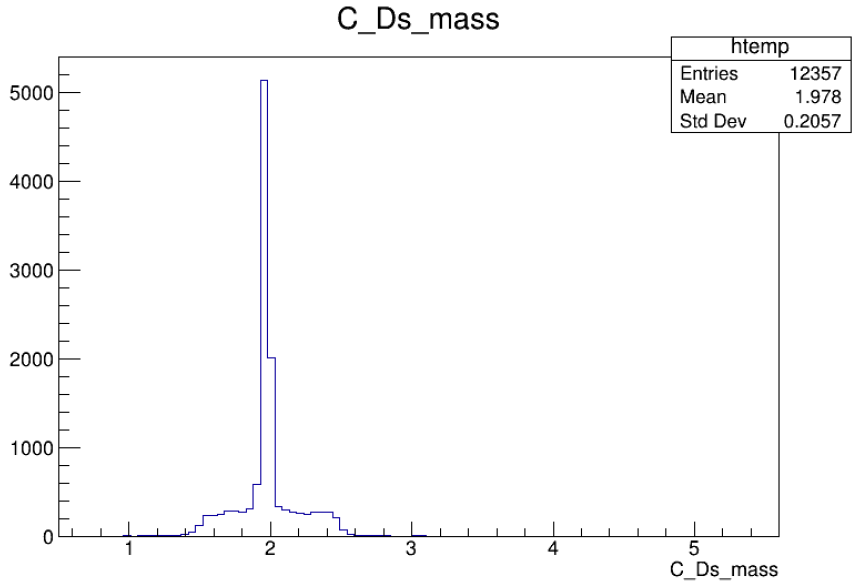
Interactive Analysis (Analysis Facility)

2. Selezione del candidato HNL migliore → Porting completato

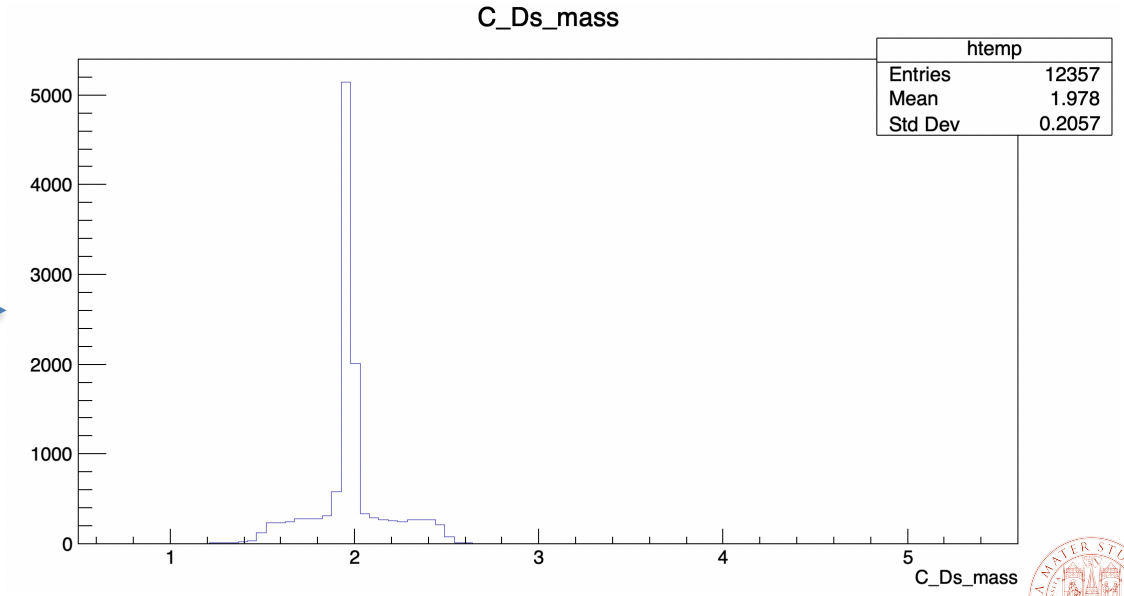
Input: Output step 1 (fuori da AF).

Output: Flat ntuple, più leggere di quelle in input.

Variable estratta dal tree in output dello Step2



Workflow originale



Workflow AF

* Workflow dell'analisi ad ora non definitivo. Possibile futura eliminazione di step 2, con l'utilizzo di tutti i candidati HNL.

Workflow analisi su Jupyter Notebook con DASK

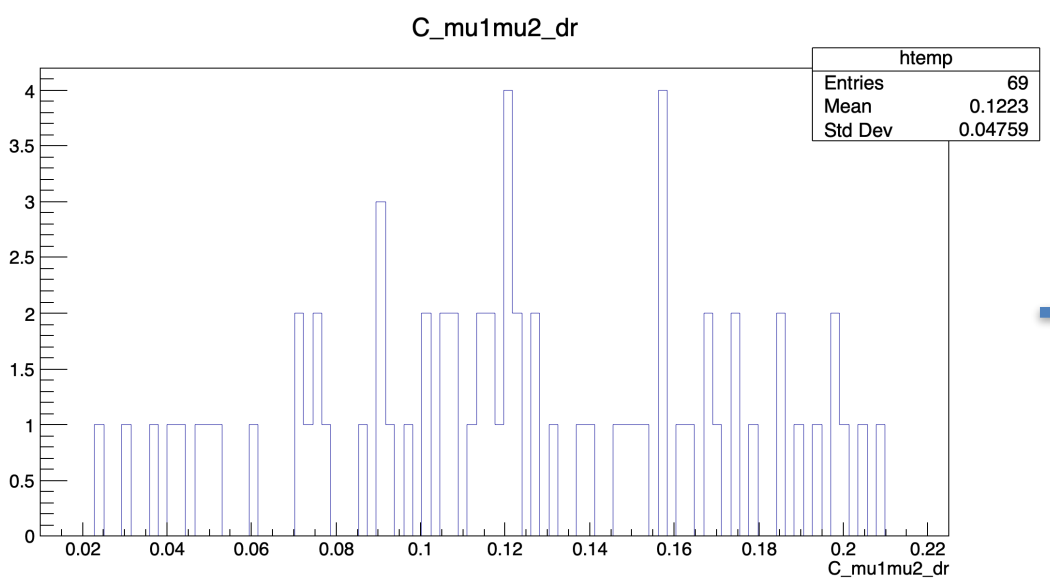
Interactive Analysis (Analysis Facility)

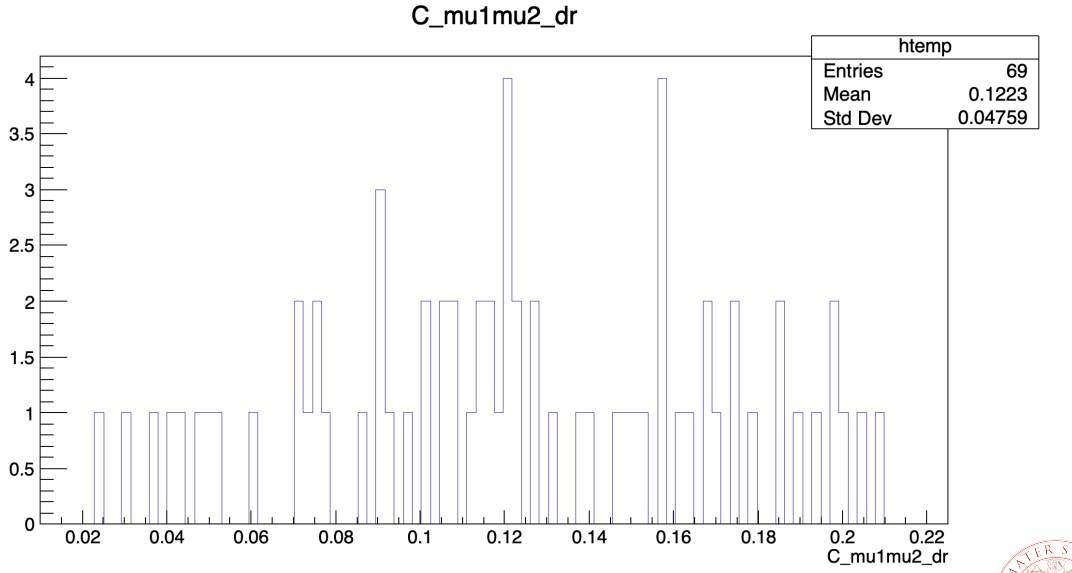
3. Analisi e calcolo limiti → Porting quasi completato

Input: Output step 2.

Output: Risultati fisici (istogrammi, limiti, …)

Variable estratta dal tree in output dello Step3





Workflow originale

Workflow AF

* Workflow dell'analisi ad ora non definitivo. Possibile futura eliminazione di step 2, con l'utilizzo di tutti i candidati HNL.

Prossimi passi

- L’ultimo step richiede qualche input dal lato fisico, per fare un porting completo e totale. In corso.
- Scalare su tutti i dati/MC dell’analisi.
- Monitoring: una volta “pronti”, sarà necessario valutare delle metriche ben precise (timing, uso di risorse,…)
  - Potrebbe essere un buon momento per iniziare a guardare queste metriche.