Esplorazione analisi interattiva con l'esperimento CMS

A. Cagnotta –WP5 meeting 24/05/2023

Use case

- Analisi dati in corso della Collaborazione CMS → top quark+MET analysis
- All'analisi sta lavorando il Gruppo CMS di Napoli → Antimo Cagnotta (50%), Orso Iorio (10%)
- fase di sviluppo dell'analisi → al momento stiamo lavorando su dataset MC della collaborazione
- Il framework di base da cui partiamo è in pyROOT → intenzione di portarlo in RDataFrame - prototipo funzionante già realizzato

Workflow analisi (STEP 1)

STEP 1 (NON in porting all'analisi interattiva)

- Data preprocessing, evaluation tramite modello ML
  - Input data: formato NanoAOD (ROOT-pla quasi plain di CMS)
  - Operazioni: preselezione minima, evaluation candidati top quark con modello Machine Learning dedicato
  - Output: nuovo file formato NanoAOD di size leggermente inferiore all'originale (stesso ordine di grandezza) salvato su Tier2
  - Strumenti utilizzati: CMS NanoAOD tools (pyROOT-based) e CRAB

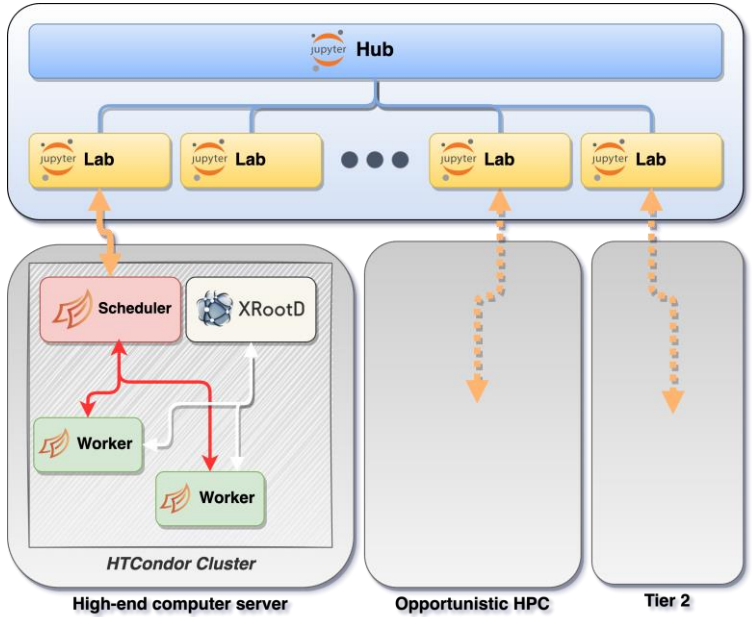
Modello ML **allenato separatamente** su una parte dei MC utilizzati.
A questo step accediamo al solo modello finale salvato in formato h5

Workflow analisi (STEP 2)

Work in progress

STEP 2: in porting all'analisi interattiva

- Skimming e selezione
  - Input: output dello step1 (file presi dal tier)
  - Operazioni: selezione eventi di interesse, estrazione candidato top, calcolo variabili di interesse + calcolo sistematiche
  - Output: file di istogrammi per combine e snapshot da RDataFrame per eventuali controlli
  - strumenti utilizzati: RDataFrame con Dask su INFN AF



- Ultimo step (fit, calcolo limiti) → pyROOT (Combine?)

Implementazione STEP 2

- Grazie al supporto di Diego Ciangottini, Tommaso Diotalevi e Tommaso Tedeschi, abbiamo iniziato il porting di questo step su JupyterLab (https://infn-cms-analysisfacility.readthedocs.io/en/latest/introduction/)
- Cosa fa al momento il notebook implementato:
  1. Legge i file in input dal Tier2 di Pisa tramite global redirect “root://cms-xrd-global.cern.ch/” in un RDataFrame
  2. Viene girata la selezione e definite variabili utili per l’analisi tramite cluster Dask
  3. In output produce file ROOT in cui vengono salvati gli istogrammi per controlli/step successive
  4. Guadagno: con RDataFrame girando su metà dei file circa 2h, usando il cluster su tutti i file circa 30minuti
- Cosa manca :
  - creazione di snapshot per controlli vari → e conseguente gestione della memoria,da implementare trasferimento file utilizzando davix (?), al momento I file di output sono di piccola size
  - calcolo pesi e sistematiche
  - ottimizzazione delle performance tramite supporto degli esperti
- Implementazione STEP 1 tramite analisi interattiva: non nei piani immediati