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

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

Studio di metriche su analisi Heavy Neutral Lepton

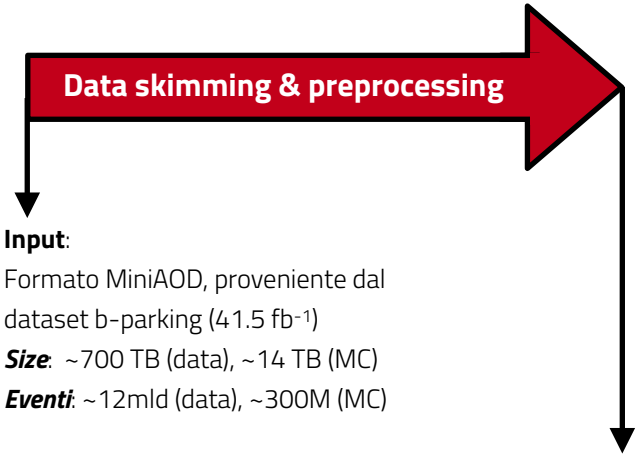
Tommaso Diotalevi (UniBO)

WP5 Meeting, 03/11/2023

Workflow dell'analisi

Operazioni

- Data reduction
- File format change



Input:
Formato MiniAOD, proveniente dal dataset b-parking (41.5 fb^{-1})
Size: ~700 TB (data), ~14 TB (MC)
Eventi: ~12mld (data), ~300M (MC)

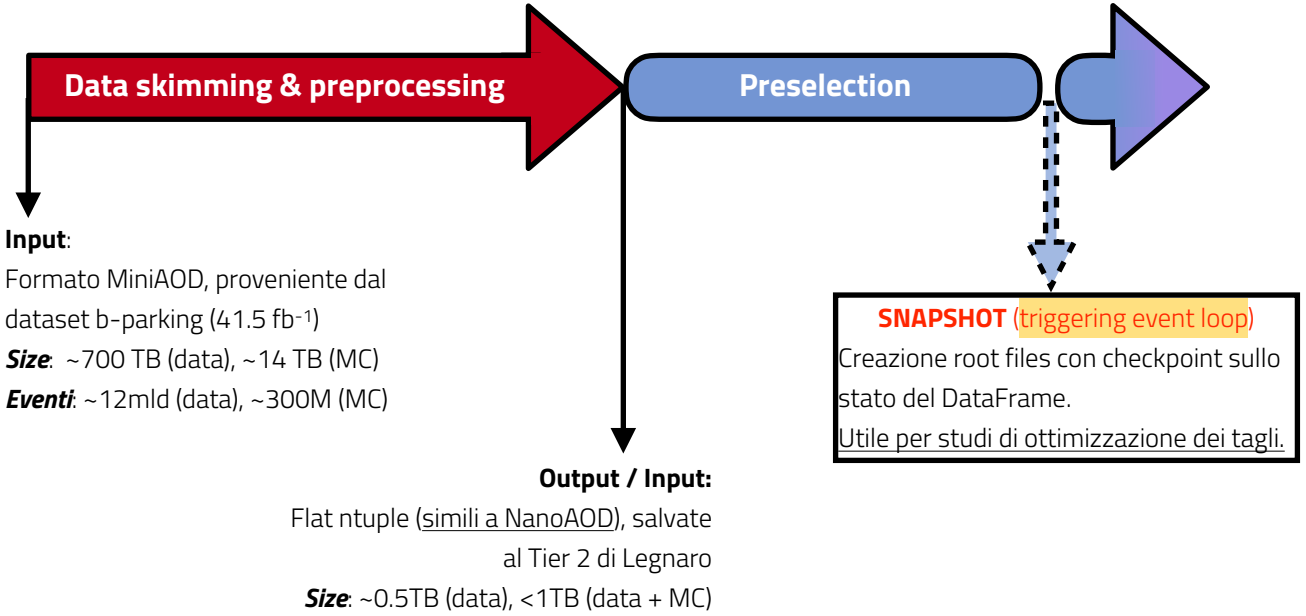
Output / Input:
Flat ntuple (simili a NanoAOD), salvate al Tier 2 di Legnaro
Size: ~0.5TB (data), <1TB (data + MC)

Workflow dell'analisi

Operazioni

- Data reduction
- File format change

- Tagli di preselezione
- Applicazione pesi (MC, SF trigger, SF muon, PU)



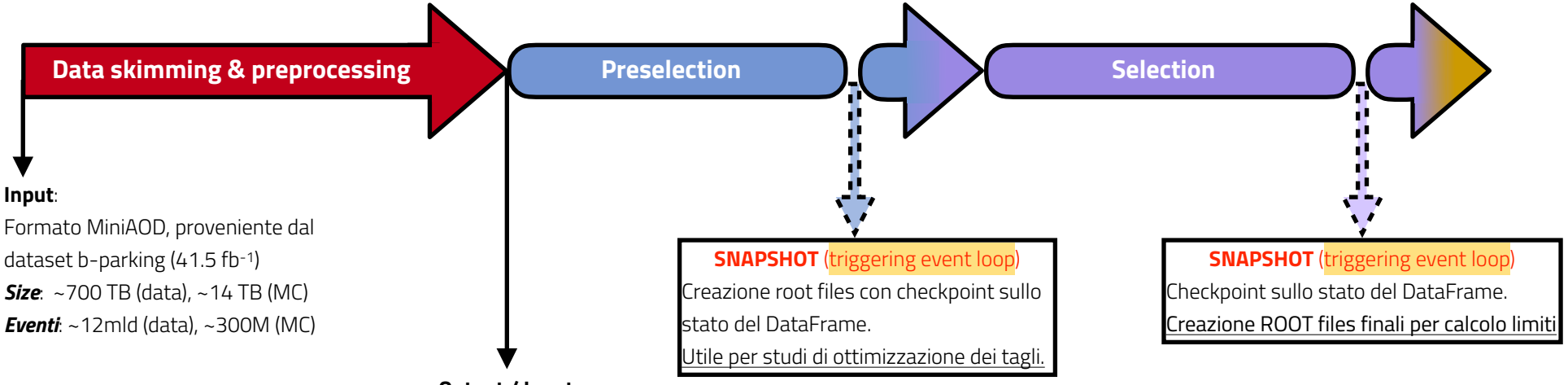
Workflow dell'analisi

Operazioni

- Data reduction
- File format change

- Tagli di preselezione
- Applicazione pesi (MC, SF trigger, SF muon, PU)

- Selezione ottimizzata per categoria;
- Selezione miglior candidato HNL
- Ridefinizione pesi totali



Workflow dell'analisi

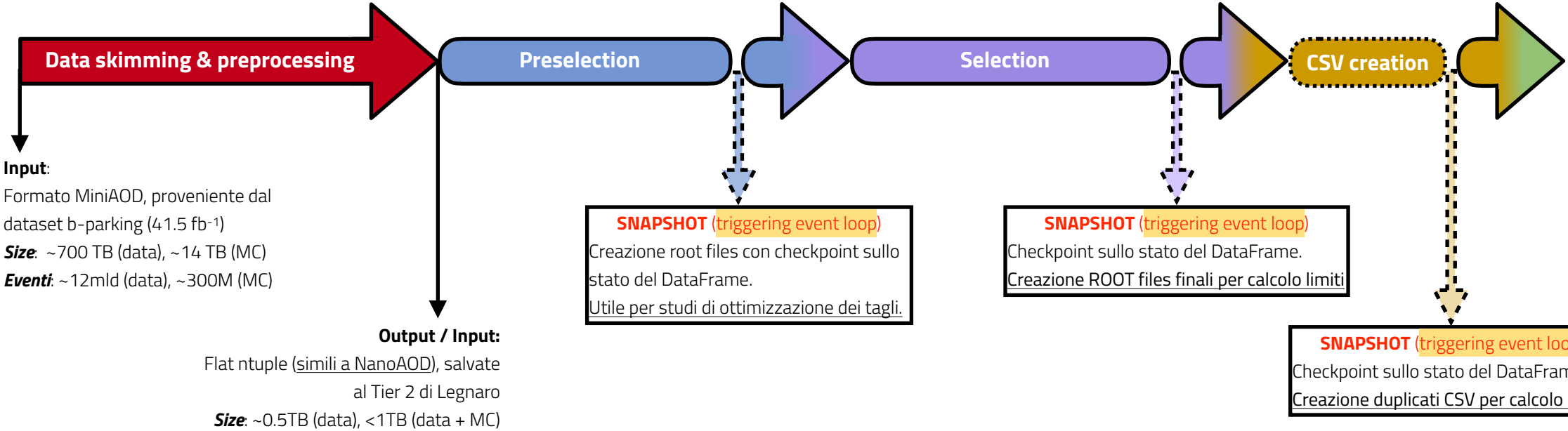
Operazioni

- Data reduction
- File format change

- Tagli di preselezione
- Applicazione pesi (MC, SF trigger, SF muon, PU)

- Selezione ottimizzata per categoria;
- Selezione miglior candidato HNL
- Ridefinizione pesi totali

- *CSV temporaneo per incompatibilità con tool statistici (combine)*



Workflow dell'analisi

Operazioni

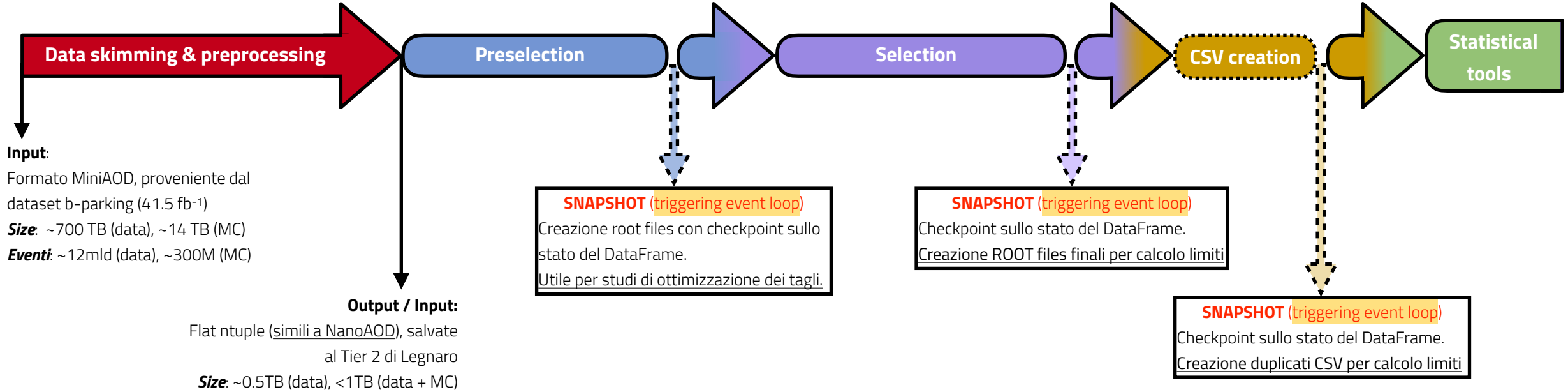
- Data reduction
- File format change

- Tagli di preselezione
- Applicazione pesi (MC, SF trigger, SF muon, PU)

- Selezione ottimizzata per categoria;
- Selezione miglior candidato HNL
- Ridefinizione pesi totali

- *CSV temporaneo per incompatibilità con tool statistici (combine)*

- Fit e analisi statistica



Workflow dell'analisi

Operazioni

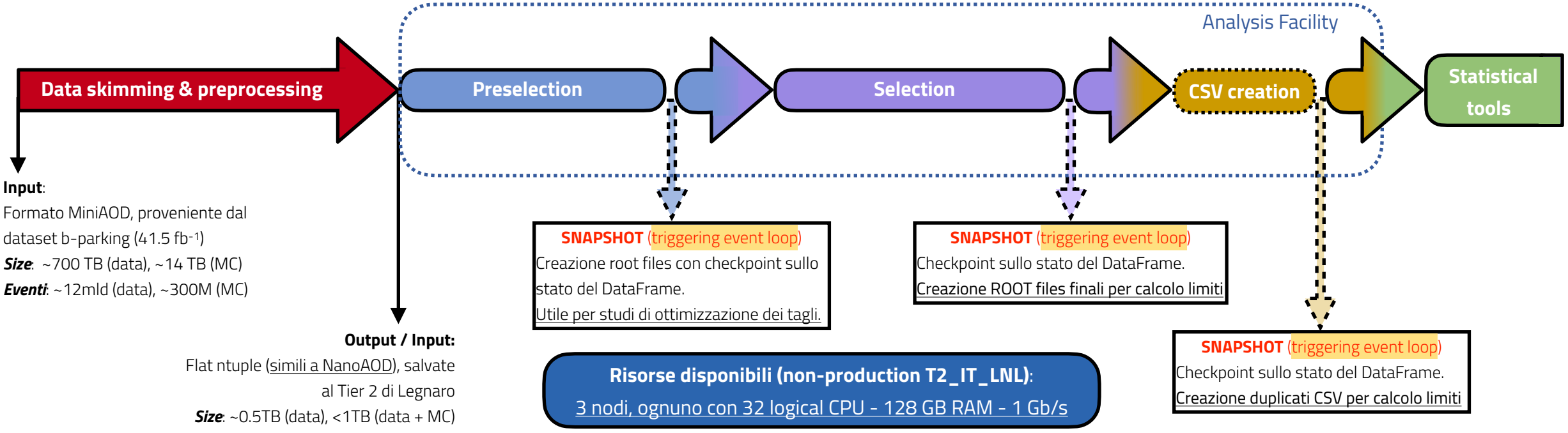
- Data reduction
- File format change

- Tagli di preselezione
- Applicazione pesi (MC, SF trigger, SF muon, PU)

- Selezione ottimizzata per categoria;
- Selezione miglior candidato HNL
- Ridefinizione pesi totali

- *CSV temporaneo per incompatibilità con tool statistici (combine)*

- Fit e analisi statistica



Sorgenti metriche

Le metriche utilizzate per questi studi provengono da differenti sorgenti, con diverse finalità:

- **Timestamps** sul Jupyter Notebook contenente l'analisi HNL. Visione "utente" dei tempi di esecuzione, suddivisi nelle varie fasi del workflow;

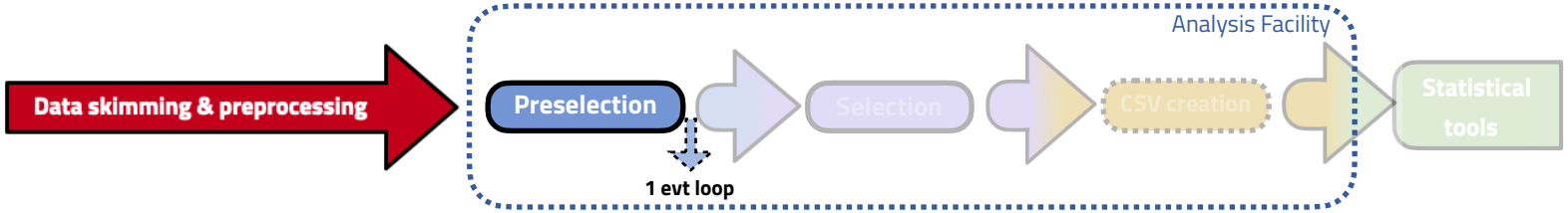
- Metriche sull'utilizzo delle risorse (CPU, memory, network), con granularità sulle singole partizioni (task) interne a Dask, utilizzando alcune features "*beta*" di **ROOT RDataFrame distributed** (v6.27):
  - Immagine specifica per i worker Dask, `root-in-docker:ubuntu22-kernel-v1-monitoring`, costruita utilizzando una versione dev di ROOT apposita.
  - Le metriche di ogni task sono salvate sui worker, in files csv. Al termine dell'esecuzione, viene fatto *stage-out* su risorse LNL.
  - Problema: I file csv vengono sovrascritti al termine di ogni event loop. Nelle configurazioni in cui sono attivati più loop, viene registrato solo l'ultimo! —> Risolto con alcune modifiche dei nomi dei csv, dentro i worker, durante l'esecuzione.

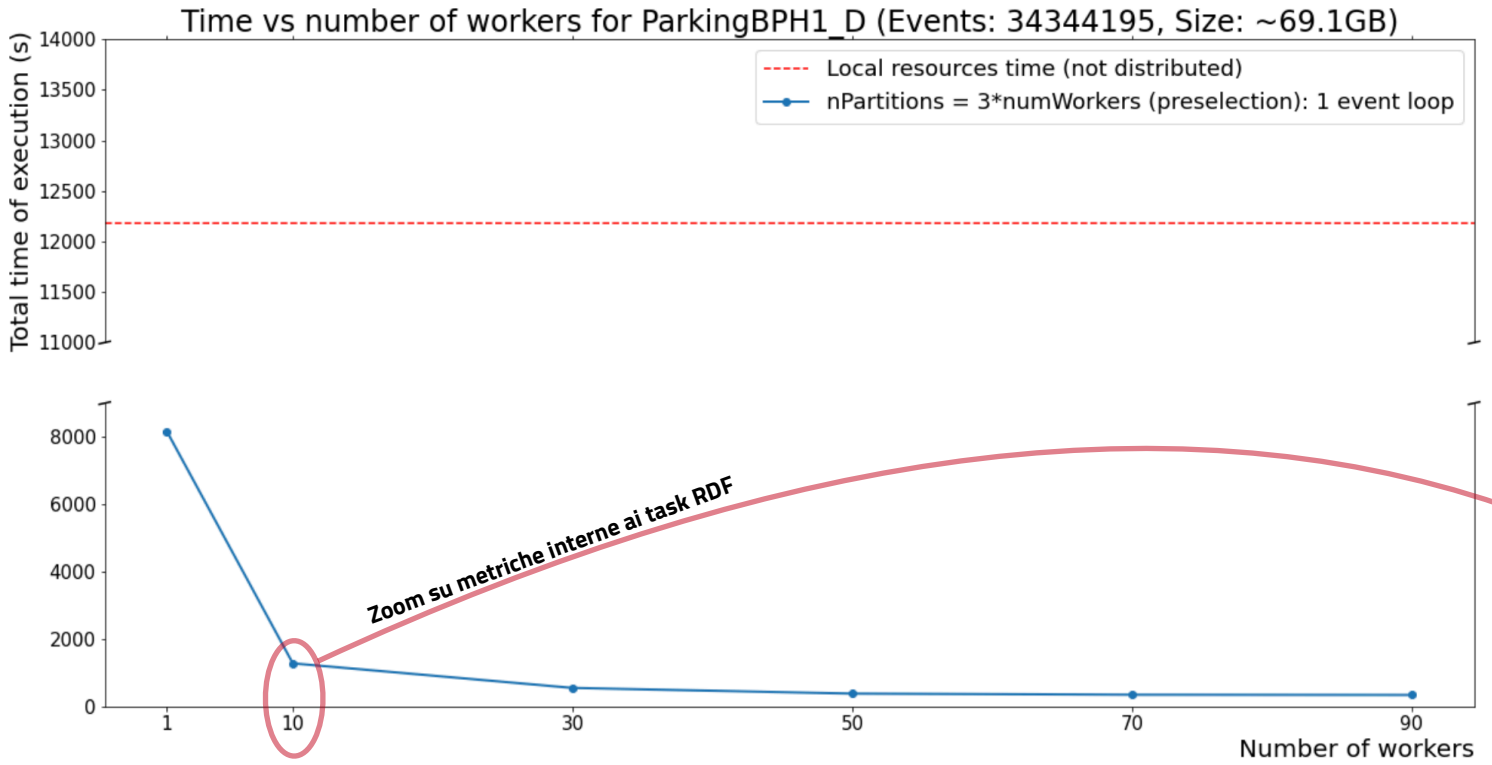
- Metriche dei 3 nodi T2_LNL_PD, salvate su istanza **InfluxDB@CNAF**. Visione di insieme del cluster, sommata su tutti gli utenti AF.
  - Problema: ~~Momentaneamente l'istanza non ha persistenza! Salvare le metriche immediatamente dopo il test!~~

Lo studio delle metriche viene effettuato su **SWAN@CERN**, per comodità (salvataggio csv, via davix/https, direttamente su EOS).

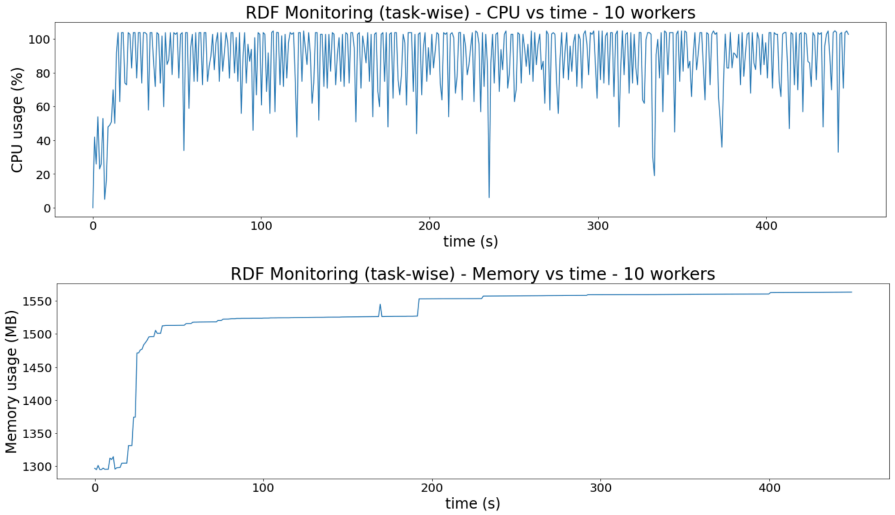
Time vs Number of DASK workers

B-Parking fraction analysed:
69GB / 500GB



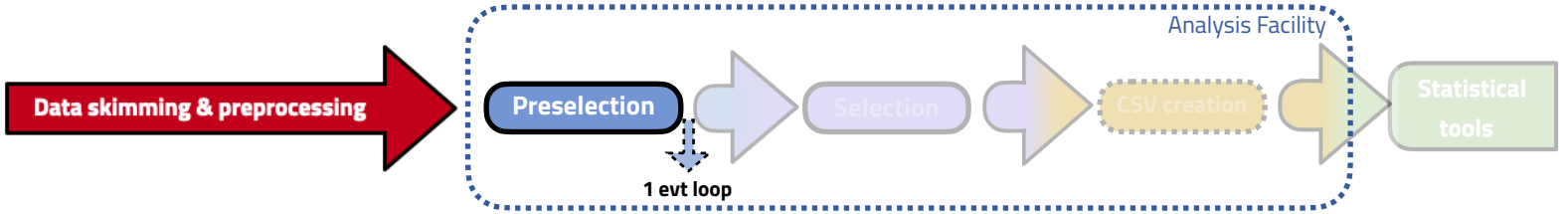


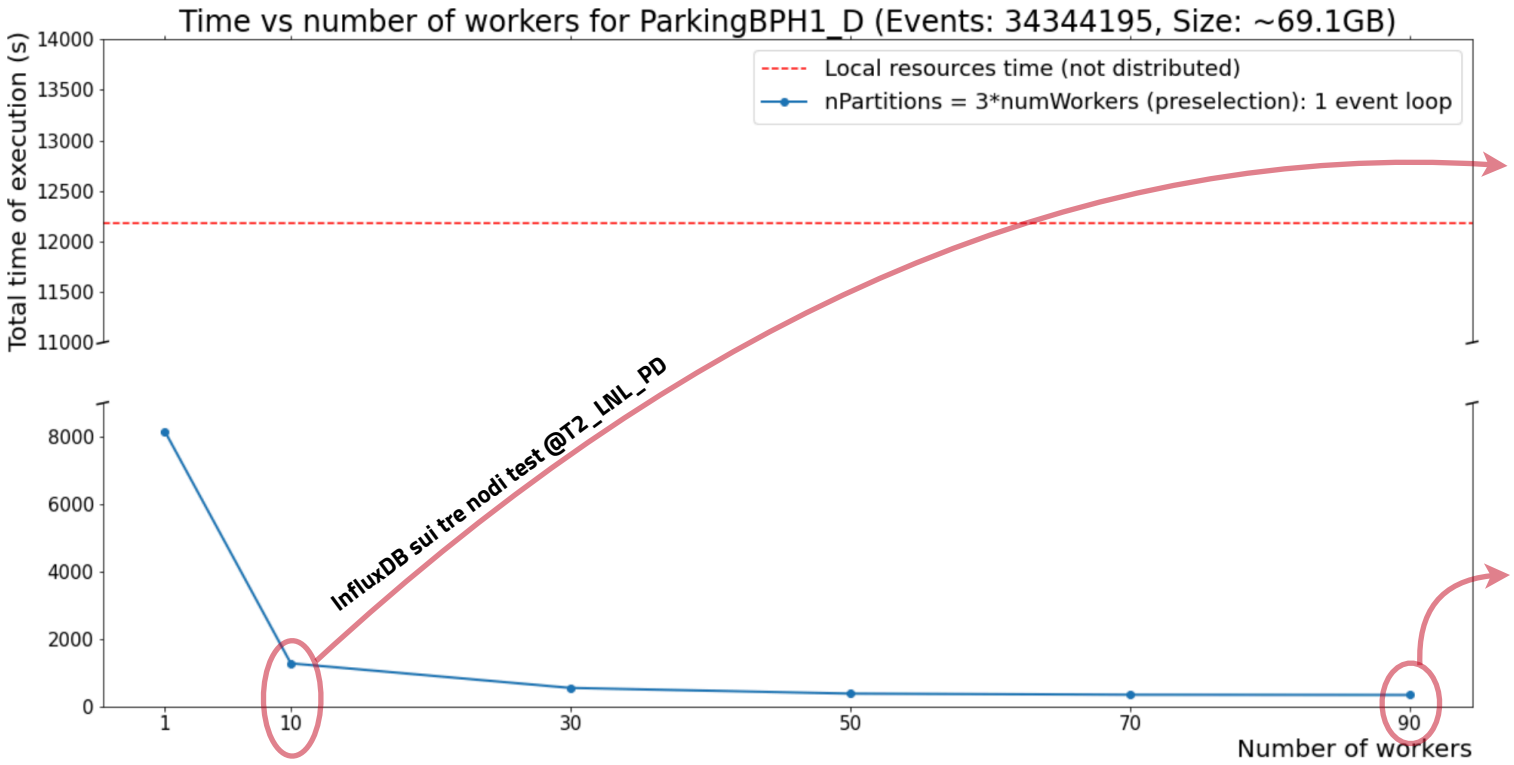
Number of workers = DASK workers (nodi) che condividono la computazione (max 92 - risorse disponibili su T2_IT_LNL).
1 worker = 1 logical CPU, 4GB RAM.
nPartitions = granularità (tasks) nella configurazione di RDF in lettura.
In rosso, il tempo di esecuzione dello stesso workflow, su risorse HTcondor locali (T3_IT_BO - senza Dask).



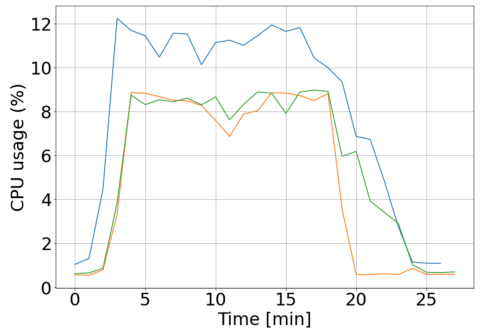
In-site metrics (CPU, memory, network)

B-Parking fraction analysed:
69GB / 500GB

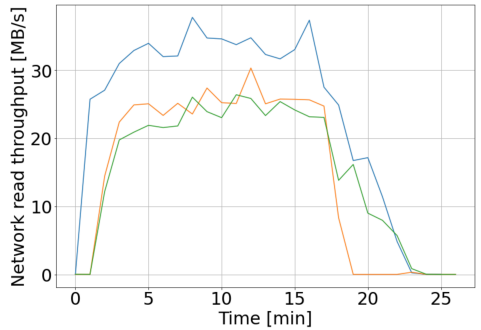


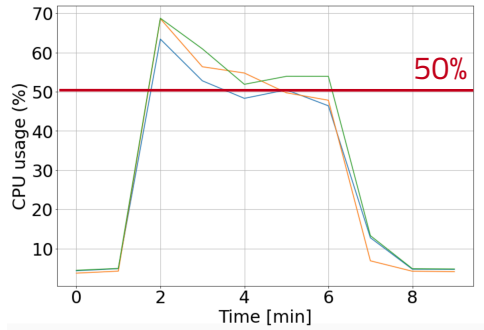


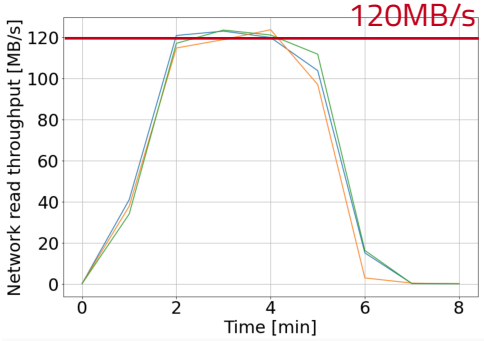
CPU



Network read

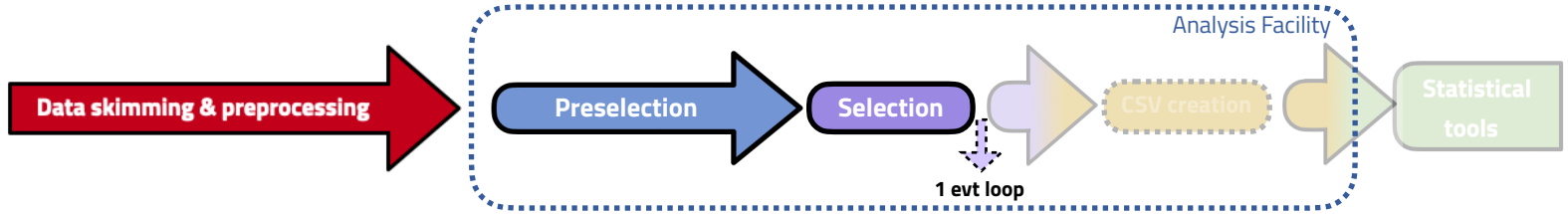


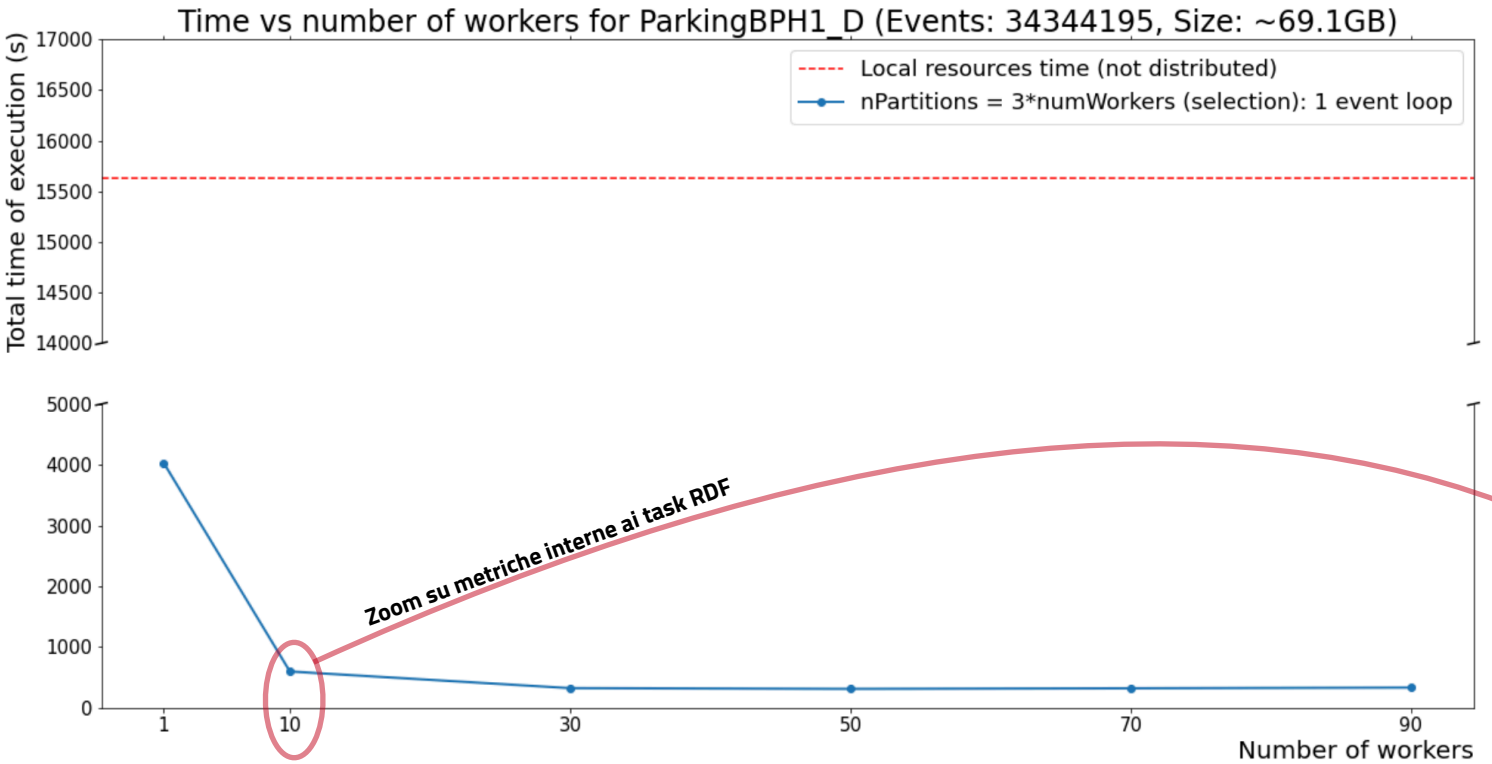




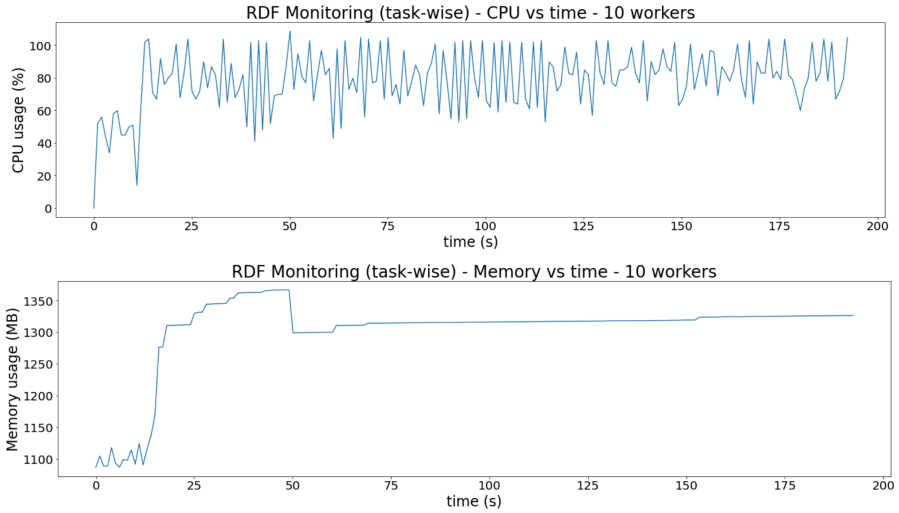
Time vs Number of DASK workers

B-Parking fraction analysed:
69GB / 500GB



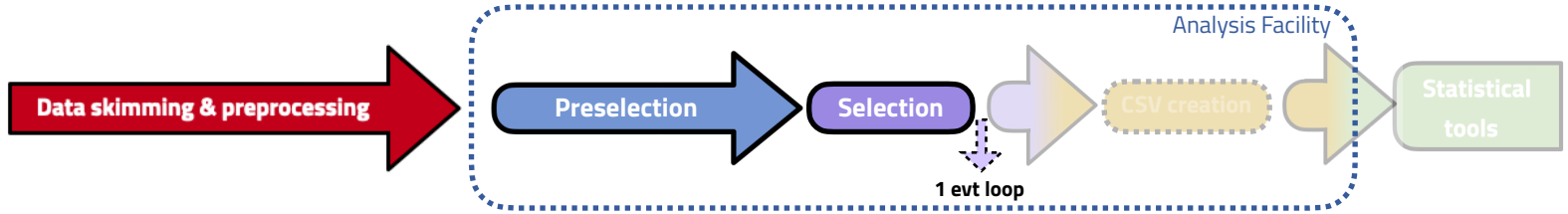


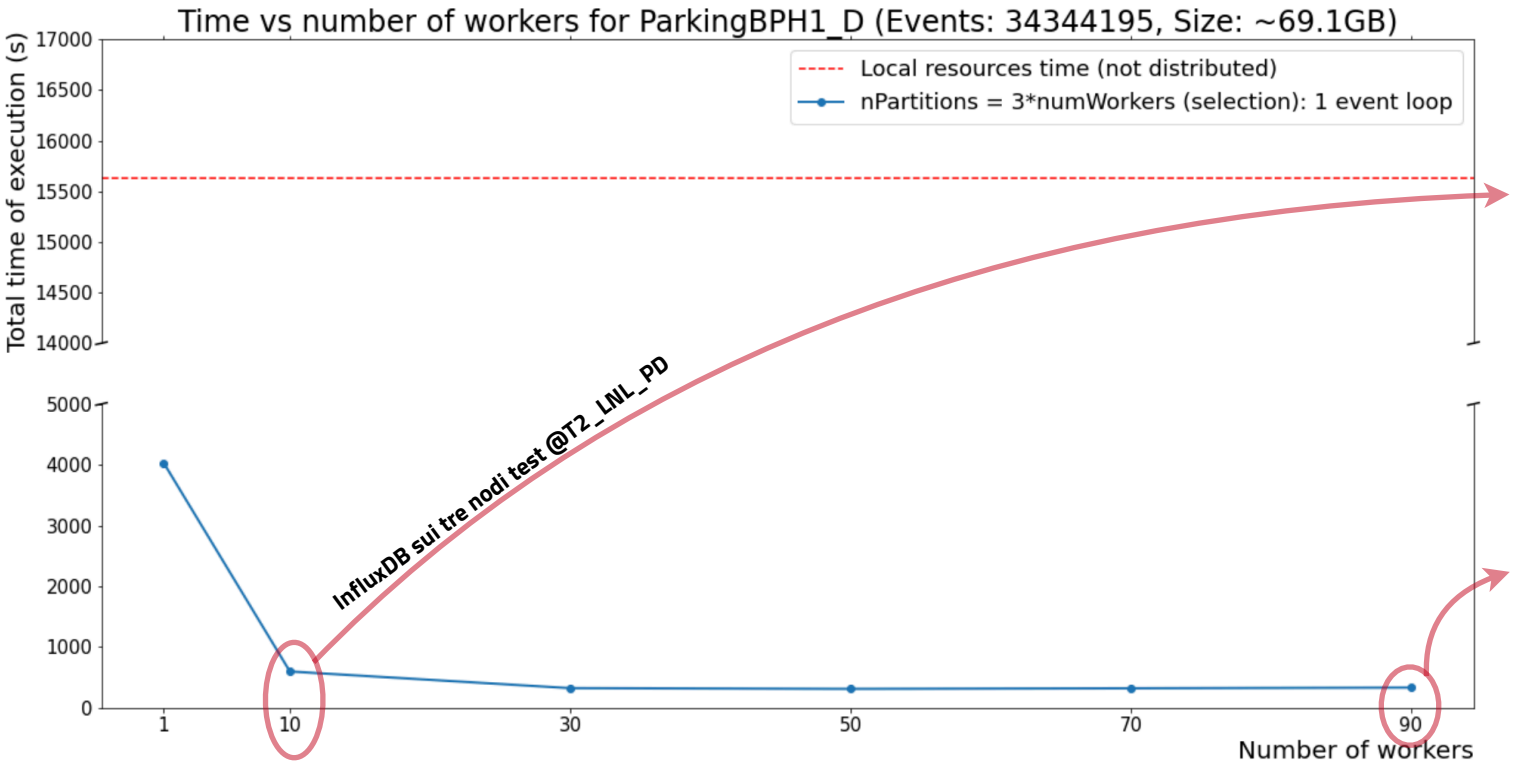
Number of workers = DASK workers (nodi) che condividono la computazione (max 90 - risorse disponibili su T2_IT_LNL).
1 worker = 1 logical CPU, 4GB RAM.
nPartitions = granularità (tasks) nella configurazione di RDF in lettura.
In rosso, il tempo di esecuzione dello stesso workflow, su risorse HTcondor locali (T3_IT_BO - senza Dask).



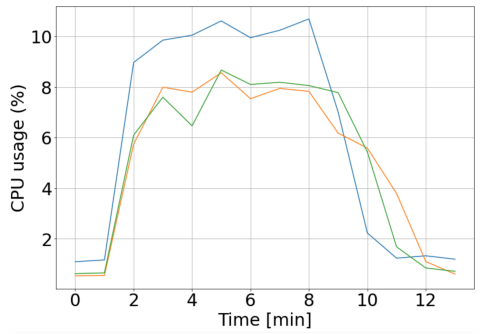
In-site metrics (CPU, memory, network)

B-Parking fraction analysed:
69GB / 500GB

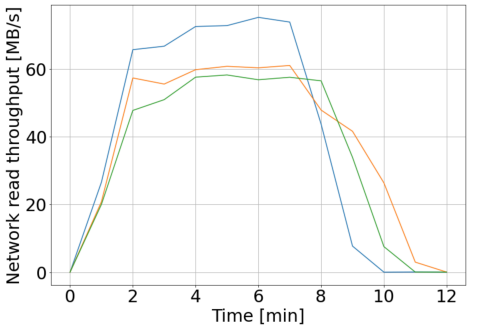


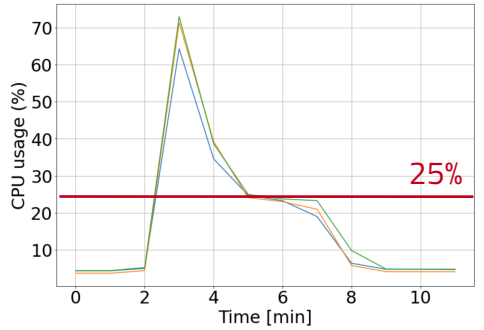


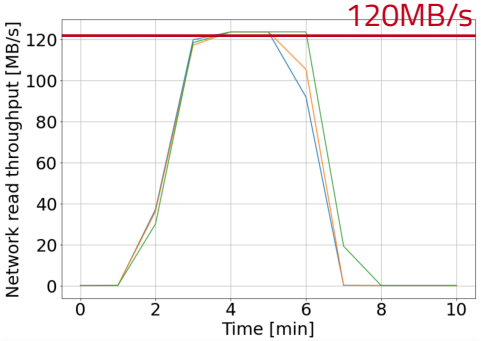
CPU



Network read

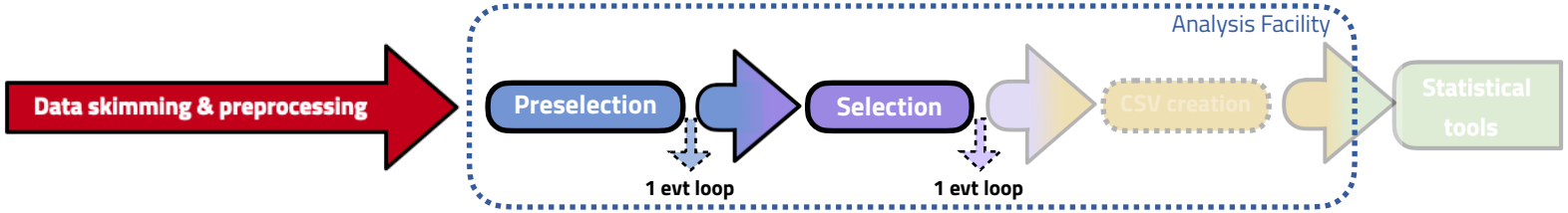


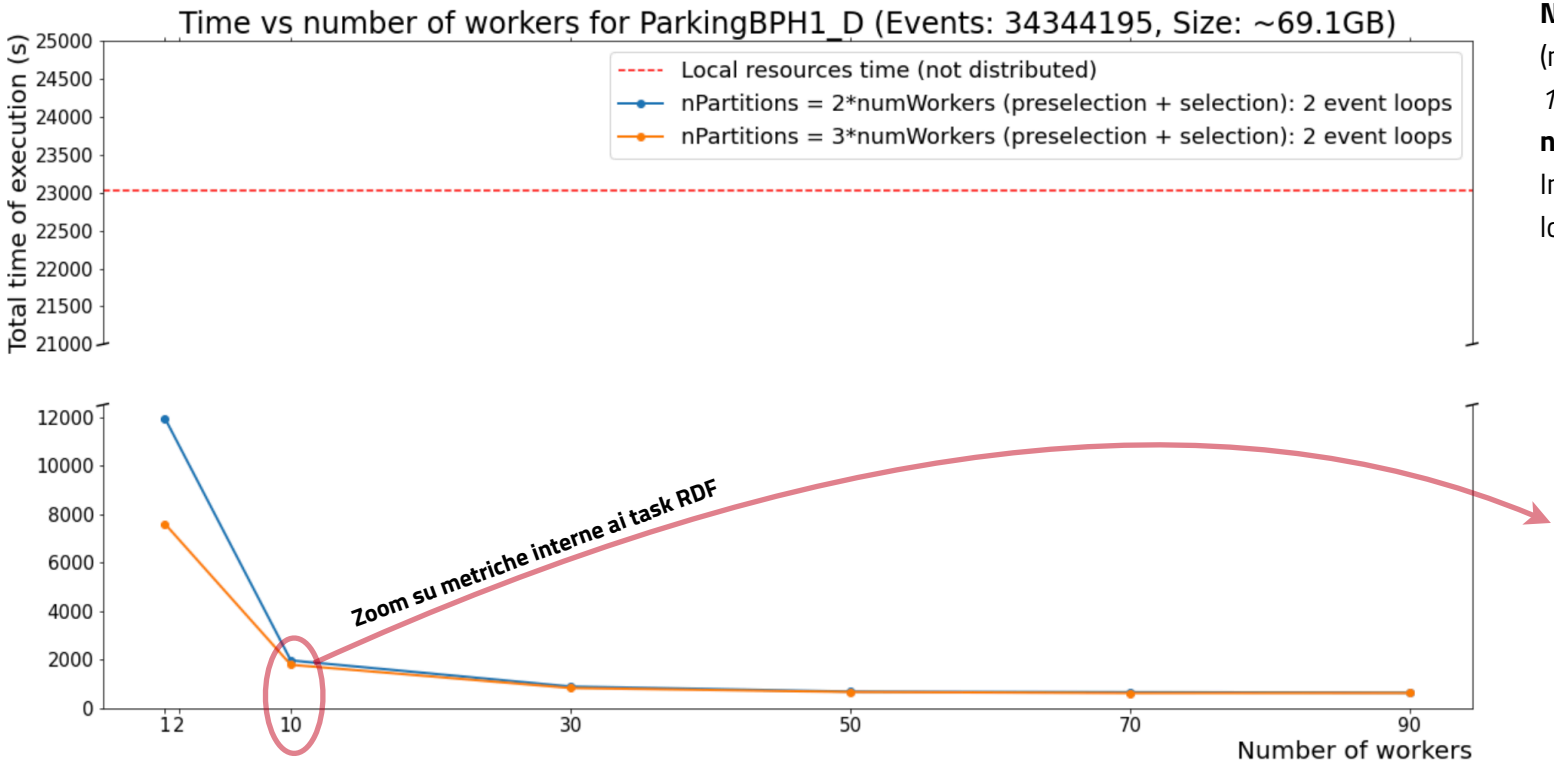




Time vs Number of DASK workers

B-Parking fraction analysed: **69GB / 500GB**



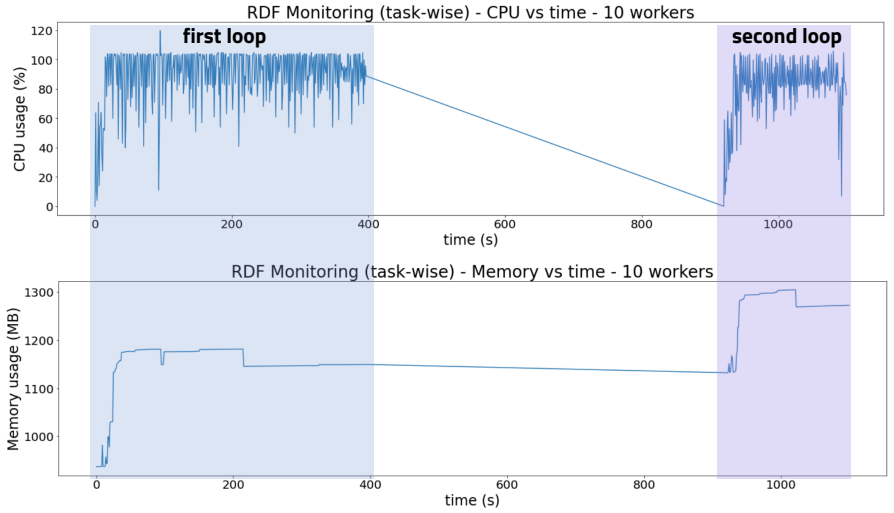


Number of workers = DASK workers (nodi) che condividono la computazione (max 90 - risorse disponibili su T2_IT_LNL).

1 worker = 1 logical CPU, 4GB RAM.

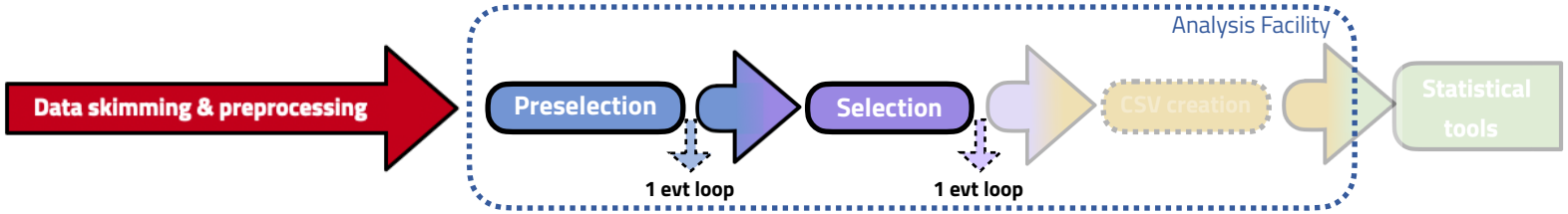
nPartitions = granularità (tasks) nella configurazione di RDF in lettura.

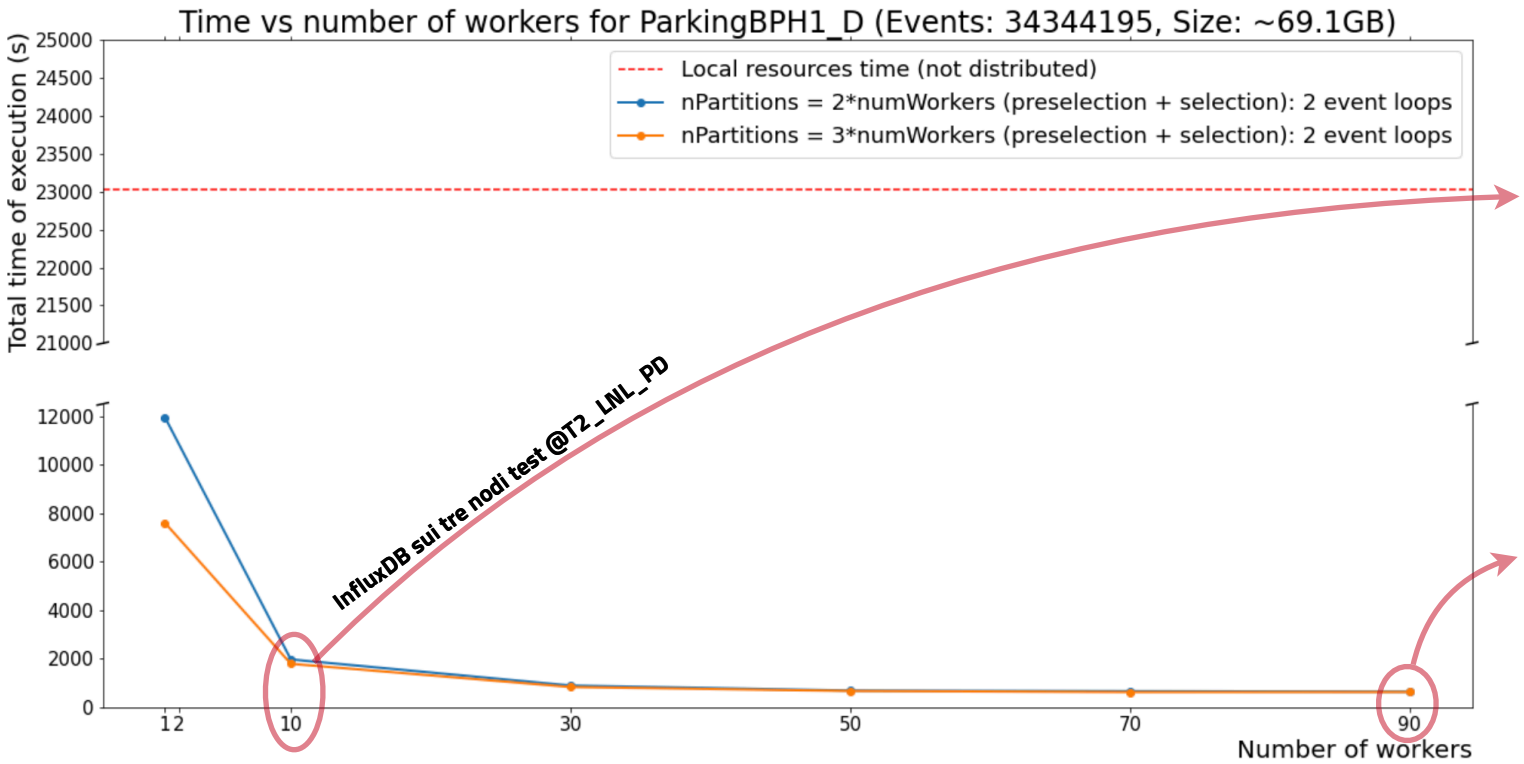
In rosso, il tempo di esecuzione dello stesso workflow, su risorse HTcondor locali (T3_IT_BO - senza Dask).



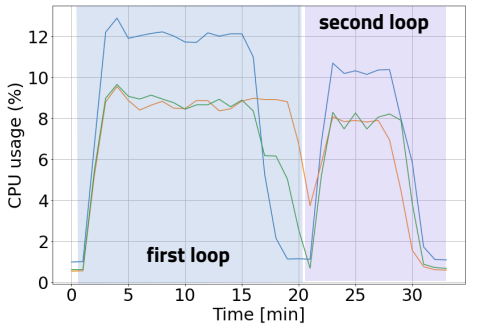
In-site metrics (CPU, memory, network)

B-Parking fraction analysed:
69GB / 500GB

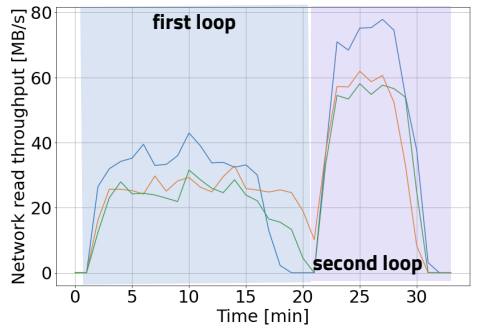


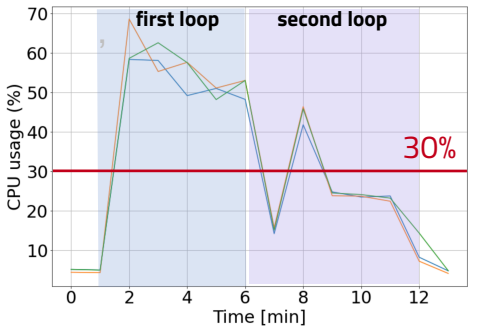


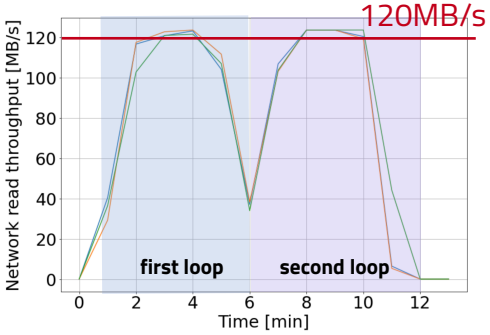
CPU



Network read

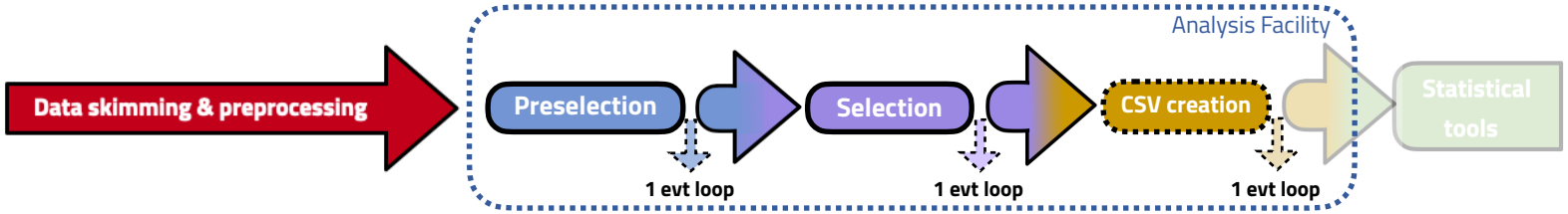


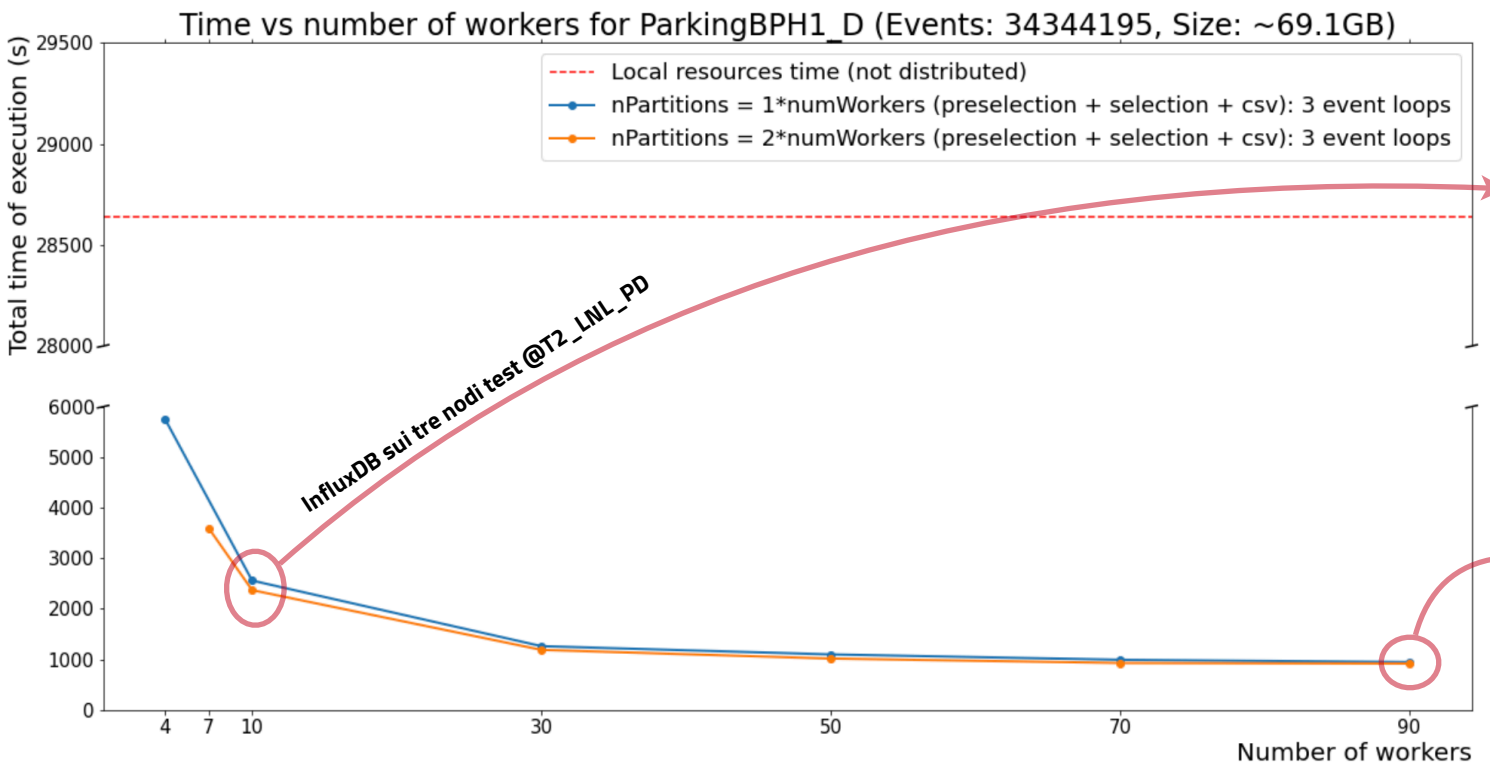




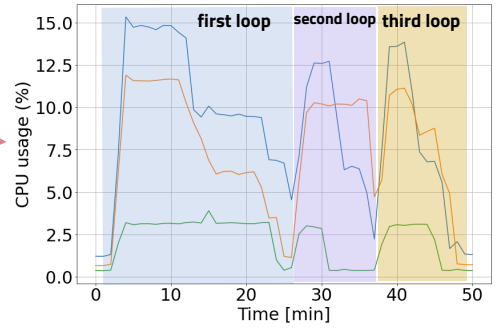
In-site metrics (CPU, memory, network)

B-Parking fraction analysed:
69GB / 500GB

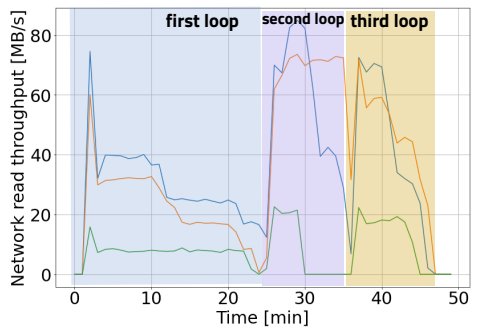


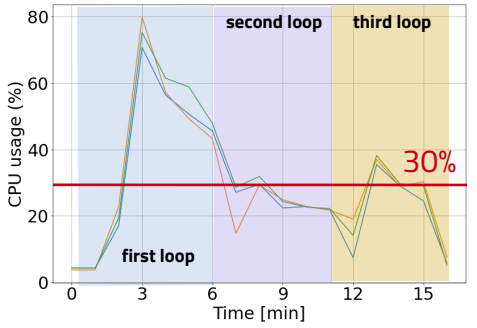


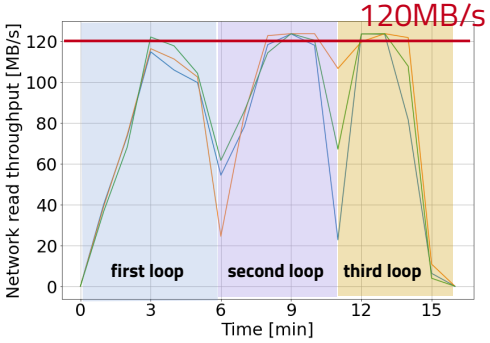
CPU



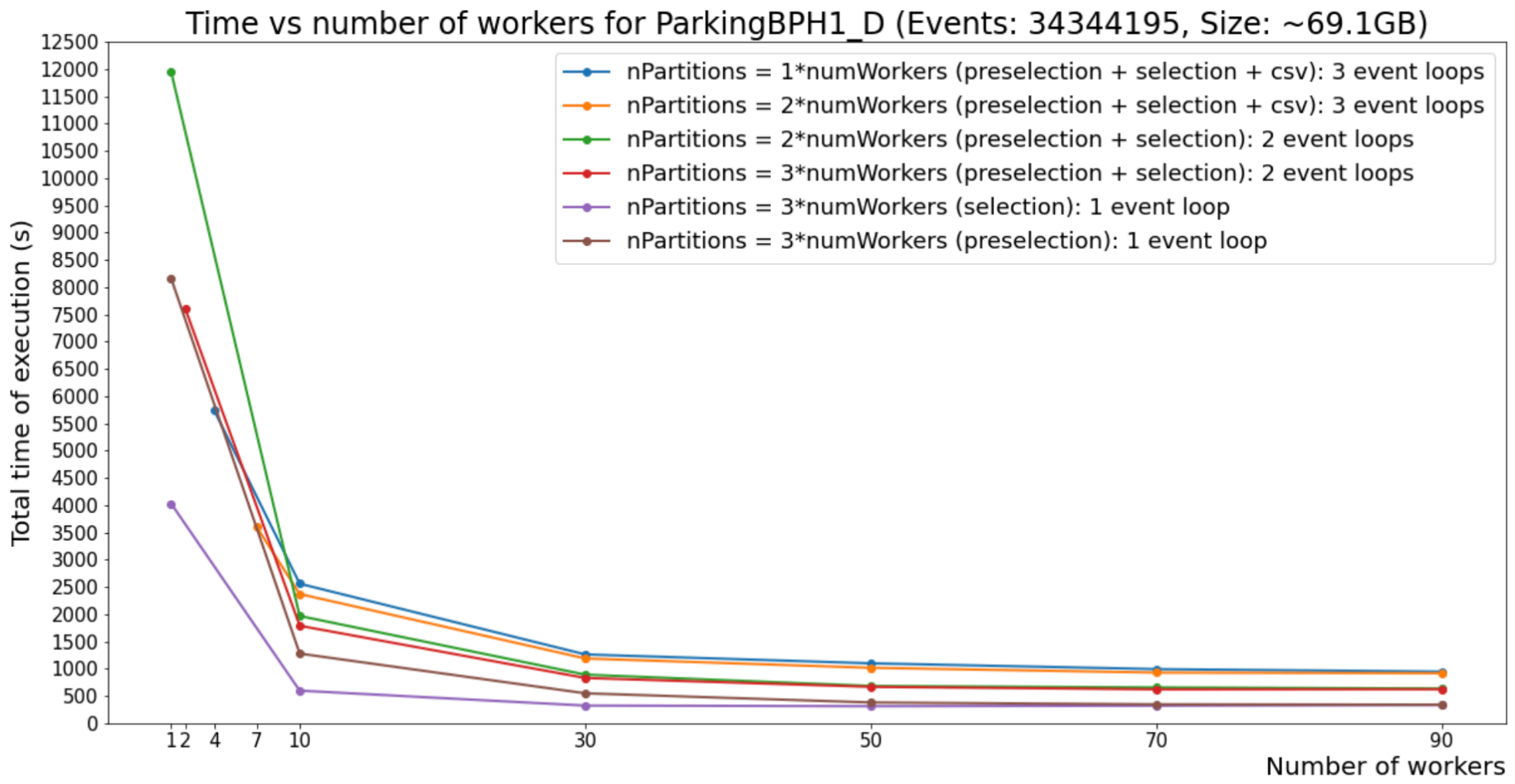
Network read







Time vs Number of DASK workers - Total



Bottleneck Matrix

Nella griglia di test effettuati, alcuni sono falliti a causa di problemi di saturazione della memoria dei worker (visibile anche da plugin DASK).

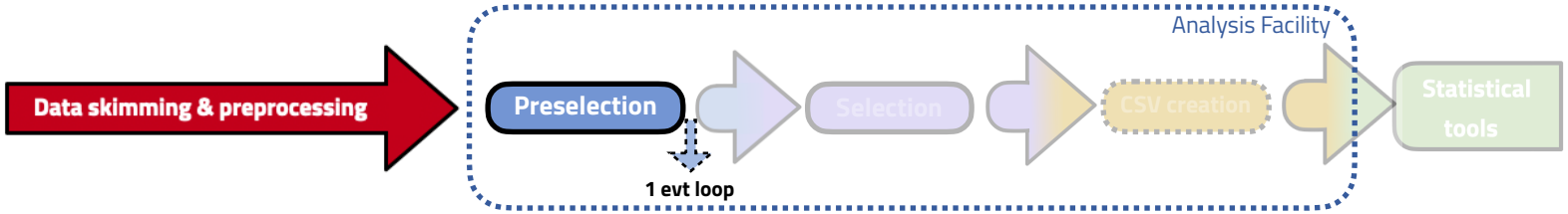
# workers / partitions	workflow	1	2	4	7	10	30	50	70	90
1 (*numWorkers)	P / S	—	—	—	—	—	—	—	—	—
	P+S	—	—	—	—	—	—	—	—	—
	P+S+C	❌	❌	✅	✅	✅	✅	✅	✅	✅
2 (*numWorkers)	P / S	—	—	—	—	—	—	—	—	—
	P+S	✅	✅	✅	✅	✅	✅	✅	✅	✅
	P+S+C	❌	❌	❌	✅	✅	✅	✅	✅	✅
3 (*numWorkers)	P / S	✅	✅	✅	✅	✅	✅	✅	✅	✅
	P+S	❌	✅	✅	✅	✅	✅	✅	✅	✅
	P+S+C	❌	❌	❌	❌	❌	❌	❌	❌	❌

Legenda:

P: preselection (1 event loop)
S: selection (1 event loop)
P+S: preselection + selection (2 event loops)
P+S+C: preselection + selection + csv (3 event loops)

✅ Test superato con successo
❌ Test fallito (saturazione della memoria)
— Test non fatto (considerato come superato, in quanto più semplice di test più intensi superati con successo)

Upscaling preliminary tests

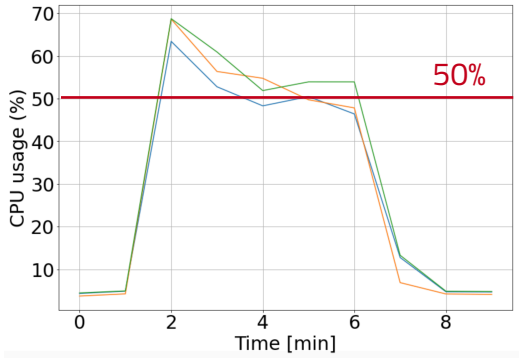


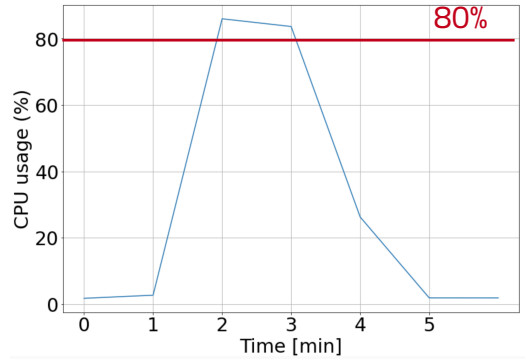
Non-production 3 nodes with 1Gbit/s network @T2_IT_LNL

Production 1 node with 10Gbit/s network @T2_IT_LNL

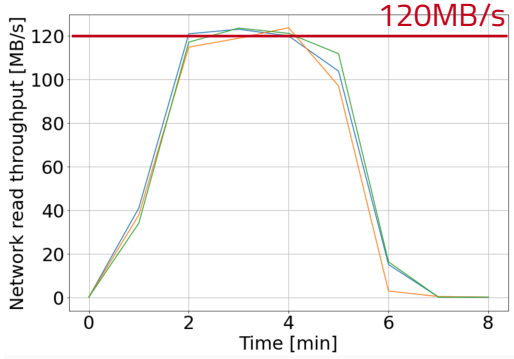
New!

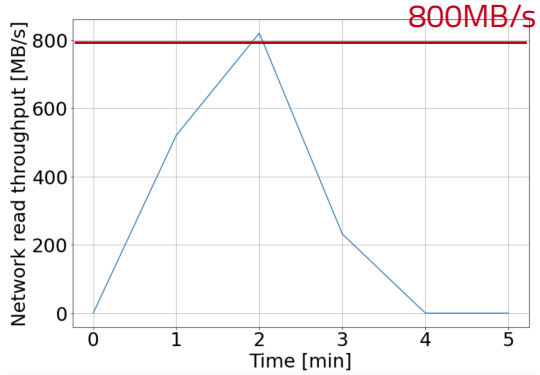
CPU





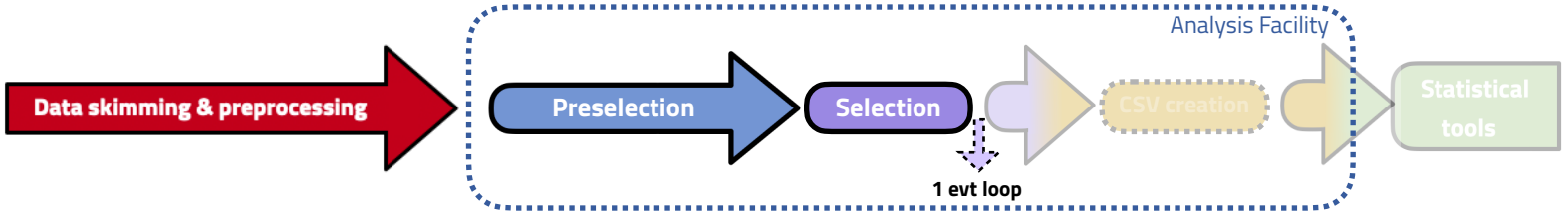
Network read





New!

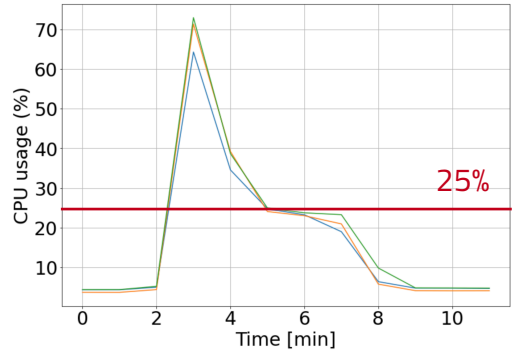
Upscaling preliminary tests

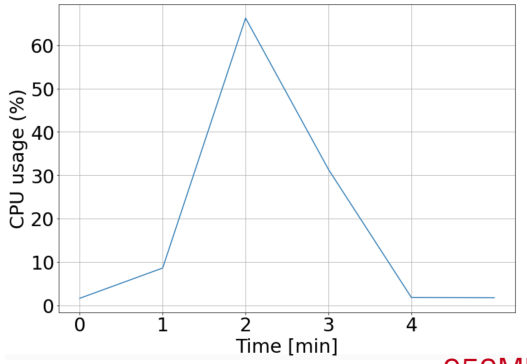


Non-production 3 nodes with 1Gbit/s network @T2_IT_LNL

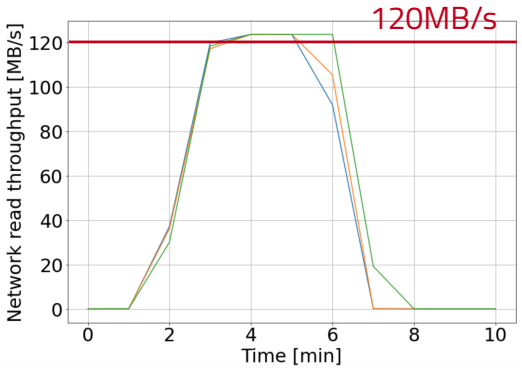
Production 1 node with 10Gbit/s network @T2_IT_LNL

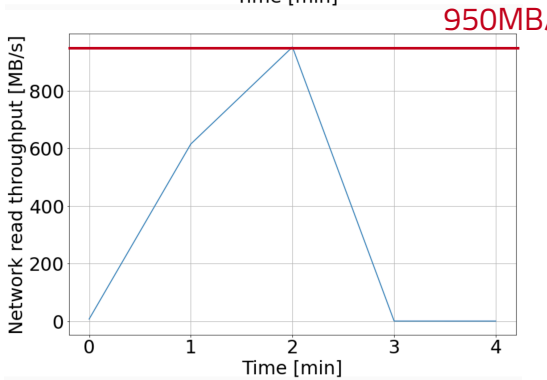
CPU





Network read





Conclusioni

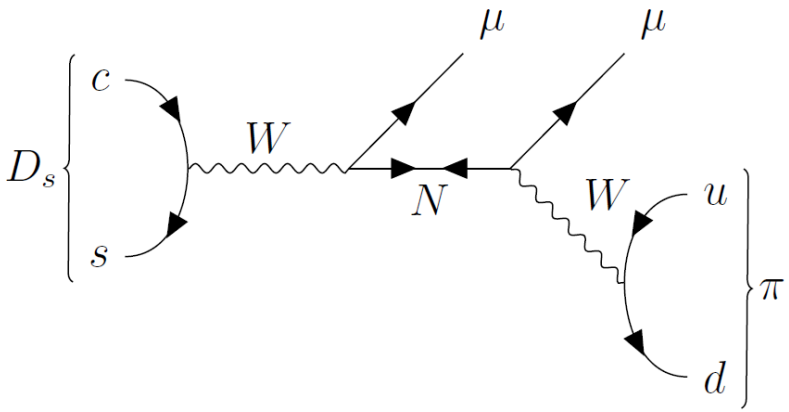
- Dai test svolti con diverse configurazioni di *worker*, *partitions*, e numero di step dell'analisi, sono state raccolte metriche provenienti da varie sorgenti (timestamps, RDF, InfluxDB).
- Solamente una frazione del dataset b-parking a disposizione è stata utilizzata per questo studio (69GB / 500GB):
  - Utilizzando tutte le risorse a disposizione (3 nodi, ognuno con 32 logical CPU - 128 GB RAM - 1 Gb/s), si riscontrano problemi di saturazione di rete quasi immediatamente dopo l'esecuzione del workflow. Questo indipendentemente dal numero di event loop della configurazione.
  - Alcune configurazioni, inoltre, falliscono a causa di saturazione della memoria allocata.
- Fare up-scaling a tutto il dataset b-parking non è quindi possibile, date le risorse *non-production* attuali. Primissimi test in corso su un nodo *production* del Tier-2 LNL (risultati preliminari), con miglioramenti causati dalla miglior connettività (10Gbit/s).
  - Next steps: fare ulteriori test (configurazioni a più loop) su queste risorse con più bandwidth (se e quando possibile)!

BACKUP

Analisi Heavy Neutral Lepton

Talk di Leonardo Lunerti al CMS General Meeting: link (CMS restricted).

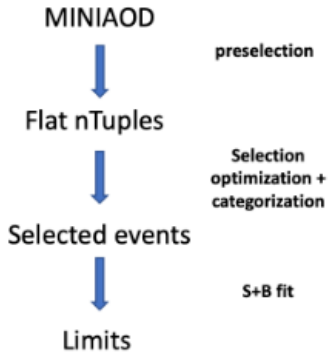
Ricerca di un leptone pesante neutro N (bump hunt), proveniente dal mesone **D_S**, nello stato finale contenente un **μ** e un **π**.
Alla firma sperimentale viene inoltre considerato un **μ** aggiuntivo, proveniente dal **W** iniziale. I due muoni possono avere sia stesso-segno che segno-opposto.



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.

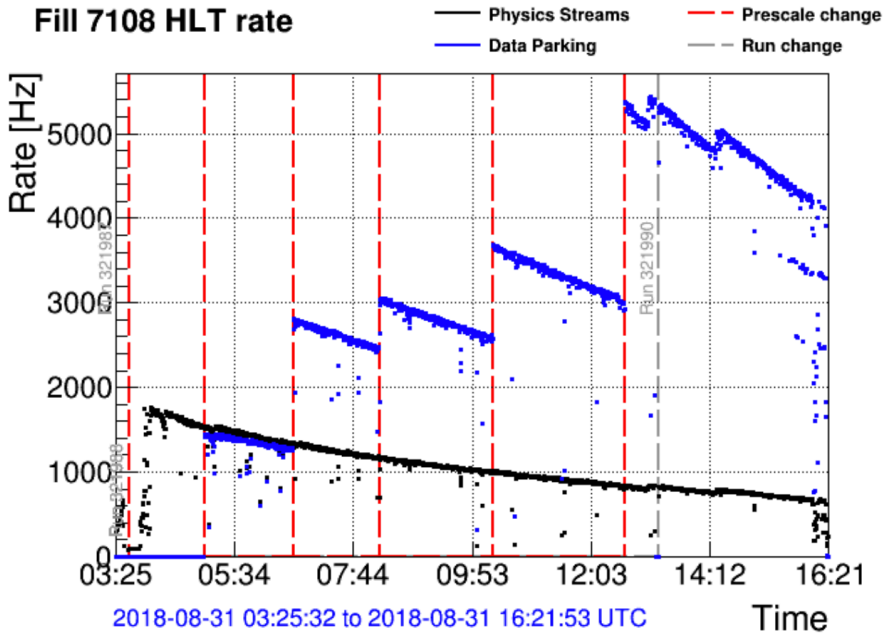
B-Parking dataset

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 pre-scaling 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}



(immagine presa da: link)

ROOT RDataFrame

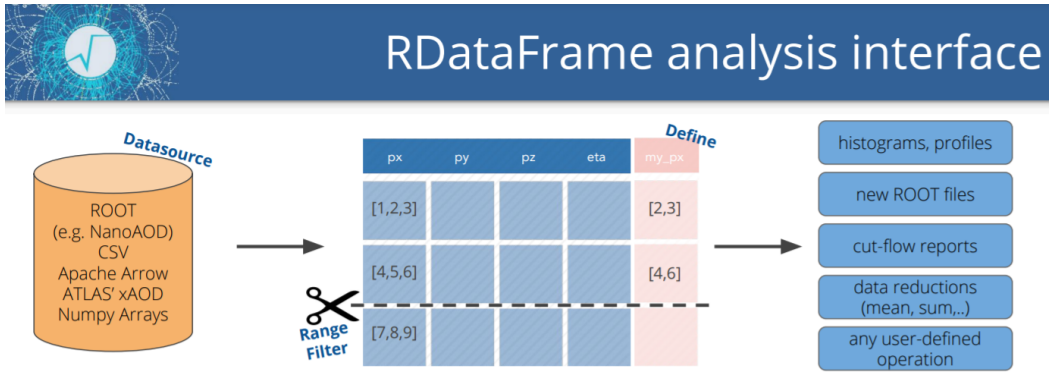
RDataFrame (RDF) è l'interfaccia di alto livello di ROOT per l'analisi dei dati archiviati in TTree, CSV e altri formati di dati. È caratterizzata da:

- 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, che attiva la computazione (*lazy*) i.e. esecuzione dell'event loop, può essere effettuata in maniera distribuita sfruttando back-end quali Spark e Dask.

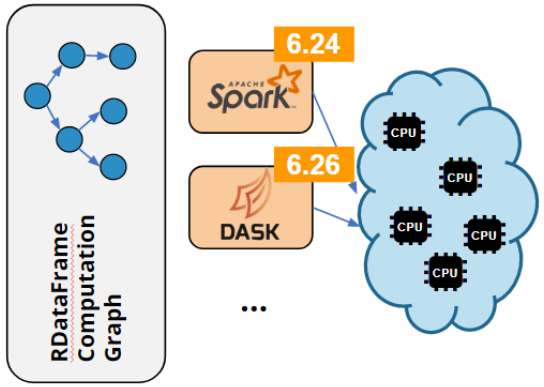
Grazie all'estensione "Distributed" di RDF, attiva in via sperimentale.




```
# 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")
```

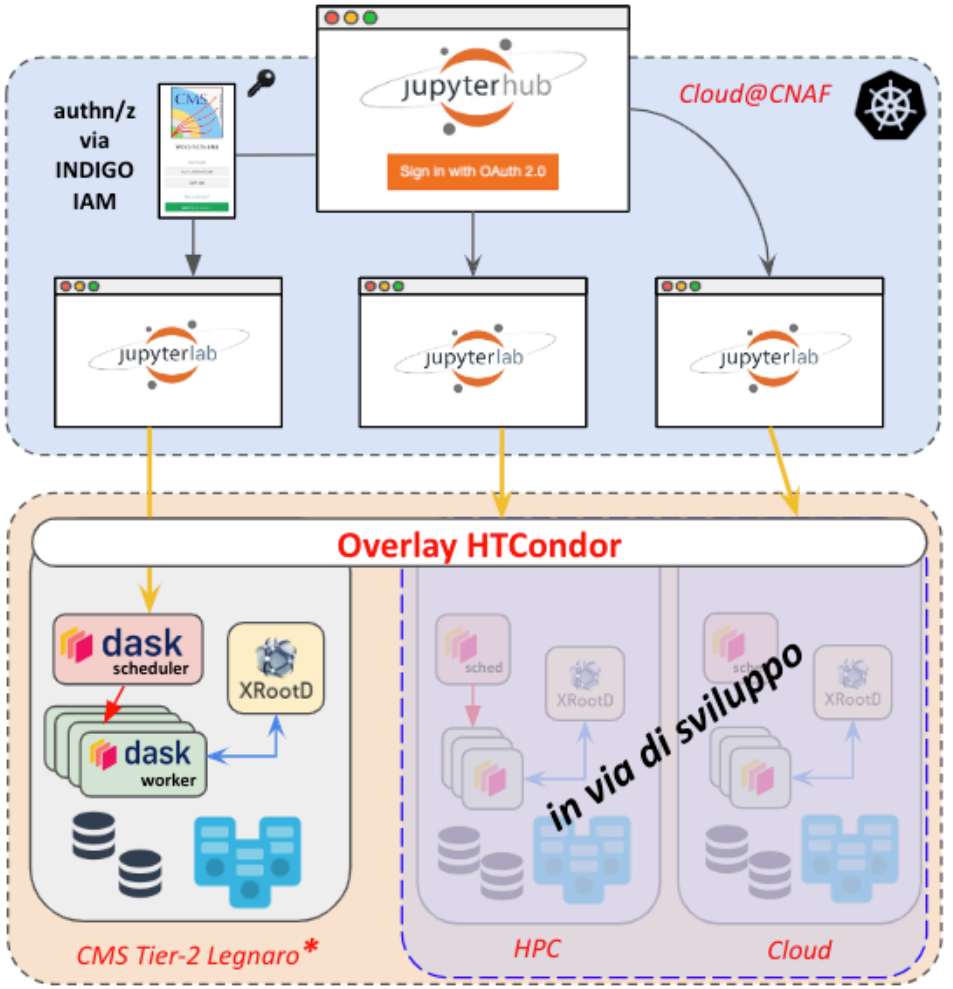


Immagini riprodotte da: talk PyHEP 2021

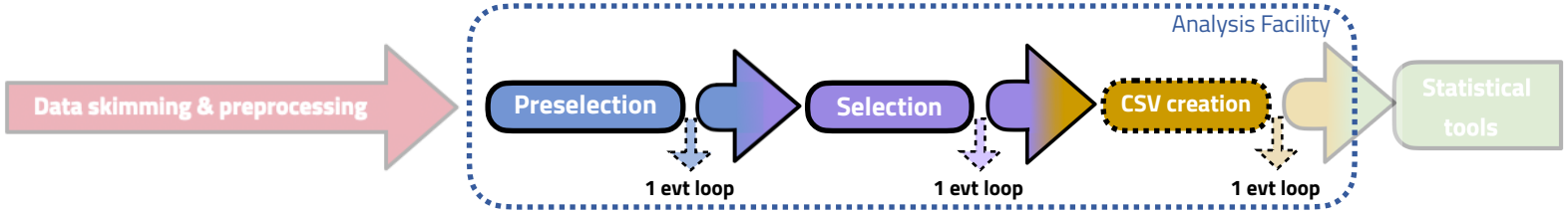
CMS Analysis Facility

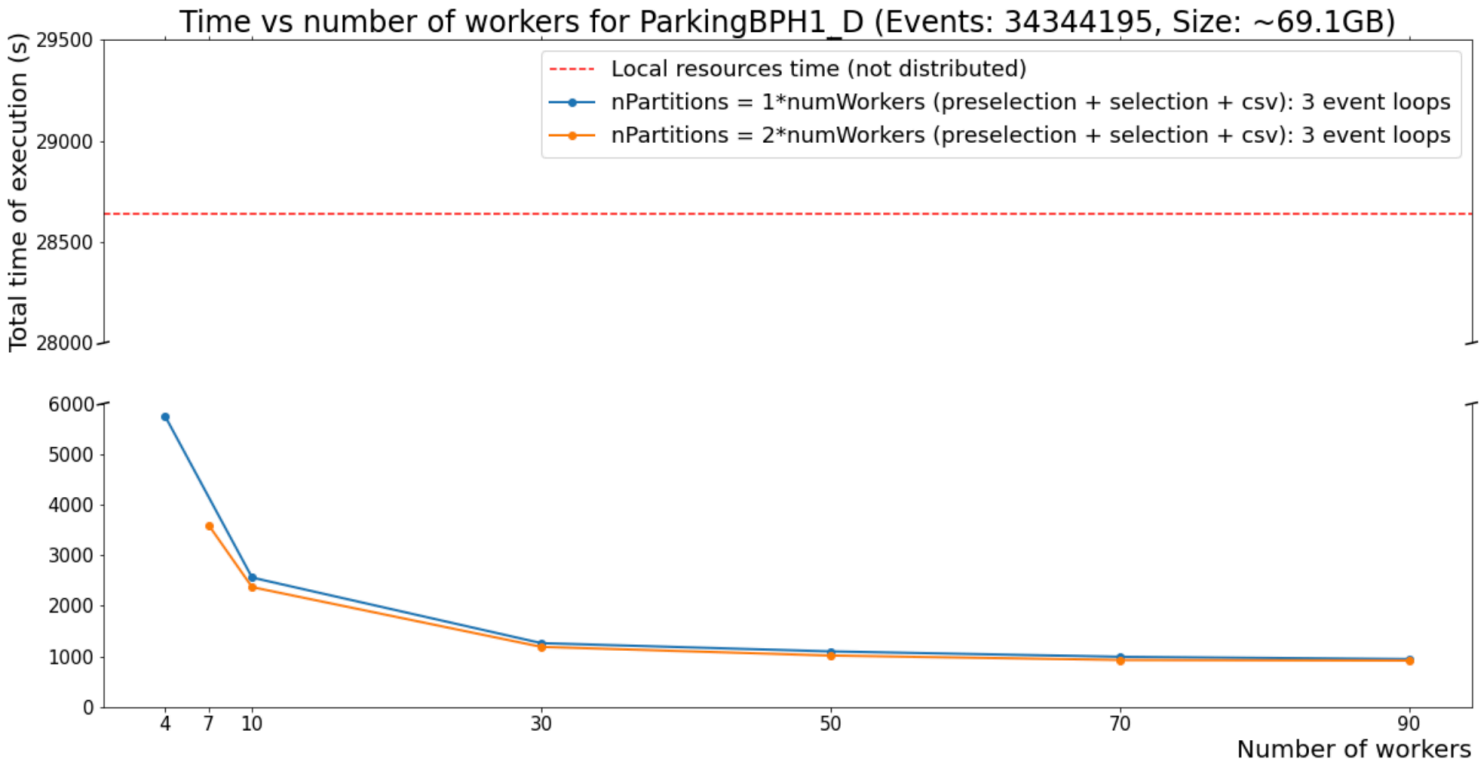
- Accesso al singolo HUB e autenticazione via token (INDIGO-IAM)
- Basato su tecnologie standard dell'industria
- Kernel python configurabile
- Containerizzazione dell'ambiente di lavoro specifico

- 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, ...)



Time vs Number of DASK workers





Number of workers = DASK workers (nodi) che condividono la computazione (max 90 - risorse disponibili su T2_IT_LNL).
1 worker = 1 logical CPU, 4GB RAM.
nPartitions = granularità (tasks) nella configurazione di RDF in lettura.

In rosso, il tempo di esecuzione dello stesso workflow, su risorse HTcondor locali (T3_IT_BO - senza Dask).