Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea in Fisica

Studi di Data Popularity nell'analisi distribuita su Grid dell'esperimento CMS a LHC

Relatore:
Prof. Daniele Bonacorsi

Presentata da:
Matteo Neri

Sessione III
Anno Accademico 2013/2014

Indice

Sommario

L'esperimento CMS a LHC ha raccolto ingenti moli di dati durante Run-1, e sta sfruttando il periodo di shutdown (LS1) per evolvere il proprio sistema di calcolo. Tra i possibili miglioramenti al sistema, emergono ampi margini di ottimizzazione nell'uso dello storage ai centri di calcolo di livello Tier-2, che rappresentano - in Worldwide LHC Computing Grid (WLCG)- il fulcro delle risorse dedicate all'analisi distribuita su Grid.

In questa tesi viene affrontato uno studio della popolarità dei dati di CMS nell'analisi distribuita su Grid ai Tier-2. Obiettivo del lavoro è dotare il sistema di calcolo di CMS di un sistema per valutare sistematicamente l'ammontare di spazio disco scritto ma non acceduto ai centri Tier-2, contribuendo alla costruzione di un sistema evoluto di data management dinamico che sappia adattarsi elasticamente alle diversi condizioni operative - rimuovendo repliche dei dati non necessarie o aggiungendo repliche dei dati più "popolari" - e dunque, in ultima analisi, che possa aumentare l'"analysis throughput" complessivo.

Il Capitolo 1 fornisce una panoramica dell'esperimento CMS a LHC.

Il Capitolo 2 descrive il CMS Computing Model nelle sue generalità, focalizzando la sua attenzione principalmente sul data management e sulle infrastrutture ad esso connesse.

Il Capitolo 3 descrive il CMS Popularity Service, fornendo una visione d'insieme sui servizi di data popularity già presenti in CMS prima dell'inizio di questo lavoro.

Il Capitolo 4 descrive l'architettura del toolkit sviluppato per questa tesi, ponendo le basi per il Capitolo successivo.

Il Capitolo 5 presenta e discute gli studi di data popularity condotti sui dati raccolti attraverso l'infrastruttura precedentemente sviluppata.

L'appendice A raccoglie due esempi di codice creato per gestire il toolkit attraverso cui si raccolgono ed elaborano i dati.

Capitolo 1

La fisica delle alte energie a LHC

1.1 Il Large Hadron Collider: una panoramica

Il Large Hadron Collider (LHC) [1, 2] è l'ultimo anello del complesso di acceleratori del CERN [3], a Ginevra, ed è il più grande e potente acceleratore di particelle ad oggi costruito. Ciascun elemento di questa catena inietta il fascio, dopo averlo accelerato, nell'elemento successivo, dando luogo ad un incremento progressivo della sua energia che raggiunge il massimo all'interno di LHC, in cui ciascuna particella è accelerata fino a raggiungere (nel caso in cui le particelle siano protoni) l'energia nominale di 7 TeV.
LHC consiste in un anello circolare di 27 km di lunghezza, diviso in otto settori sostanzialmente indipendenti, progettato per accelerare protoni e ioni pesanti. L'accelerazione avviene su due fasci che viaggiano in direzioni opposte in condotti separati e in condizioni di vuoto ultraspinto. I fasci sono controllati da elettromagneti superconduttori che li mantengono nella loro traiettoria e li portano fino alla loro energia di regime.
Le principali caratteristiche tecniche di LHC sono riportate in Tabella 1.1. Nel seguito, verranno illustrate singolarmente alcune delle sue componenti.

1.1.1 Il sistema per il vuoto

Il sistema per il vuoto di LHC[2, Cap. 5][1, Cap. 12] è, con più di 104 chilometri di condotti sotto vuoto, tra i più grandi e avanzati al mondo. Esso ha, essenzialmente, due funzioni: la prima è di evitare collisioni tra le particelle del fascio e le molecole d'aria presenti nei condotti, ricreando una condizione di vuoto ultraspinto a pressione pari a 10^{-13}atm mentre la seconda è quella di annullare lo scambio di calore degli elementi dell'acceleratore che necessitano di operare a temperature estremamente basse, in modo da massimizzarne l'efficienza.
Il sistema per il vuoto è composto da tre parti, indipendenti tra loro:

- un sistema per il vuoto di isolamento per i criomagneti;
- un sistema per il vuoto di isolamento per la linea di distribuzione dell'elio;
- un sistema per il vuoto per i fasci.

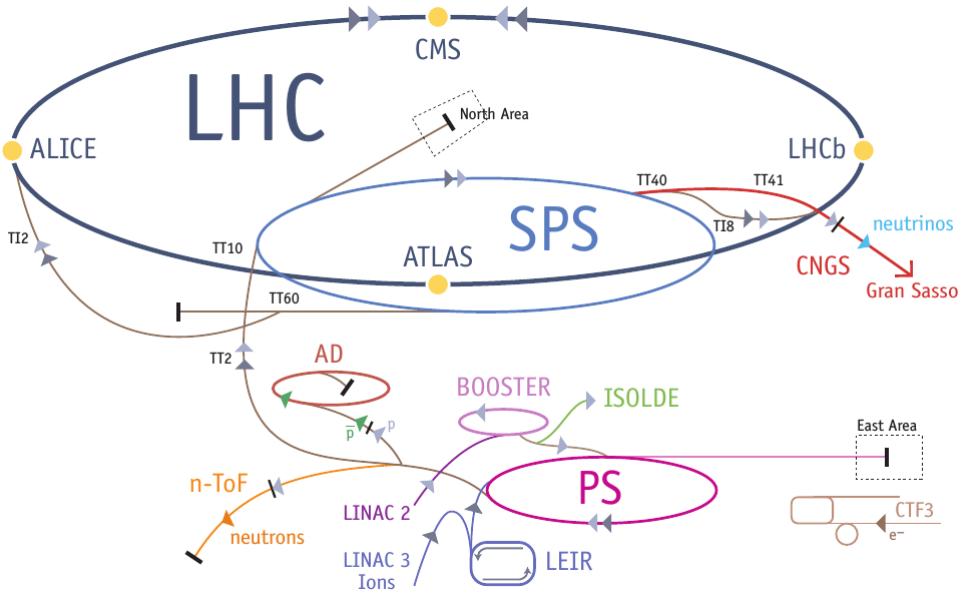


Figura 1.1: La catena di acceleratori del CERN.

1.1.2 Gli elettromagneti

Gli elettromagneti [2, Cap. 3][1, Cap. 7] sono le componenti che si occupano di guidare i fasci nel loro percorso, modificando la traiettoria delle singole particelle, nonchè collimandole tra loro in modo da aumentare la probabilità di collisione. Ci sono più di cinquanta diversi tipi di magneti in LHC, per un numero complessivo di circa 9600 magneti. I dipoli principali, i più grandi, generano un campo magnetico con intensità massima di 8.3 T. Gli elettromagneti usano una corrente di 11 850 A in condizioni di superconduttività in modo da azzerare le perdite di energie dovute alla resistenza. Ciò avviene grazie ad un sistema di distribuzione di elio liquido che mantiene i magneti a circa 1.9 K. A questa temperatura, inferiore a quella necessaria affinchè le bobine dei magneti operino in condizioni di superconduttività, l'elio assume la proprietà di superfluidità: essa lo dota di una conducibilità termica molto elevata che fa sì che esso possa essere impiegato come sistema di raffreddamento per magneti.

1.1.3 Le cavità a radiofrequenza e la divisione in pacchetti

Le cavità a radiofrequenza [2, Cap. 4][1, Cap. 6] sono camere metalliche con al loro interno un campo elettromagnetico. Esse hanno il ruolo di dividere i protoni in pacchetti e mantenerli raggruppati tra loro in modo da garantire una luminosità alta nel punto di collisione e quindi massimizzare il numero di collisioni stesse. Inoltre forniscono potenza al fascio durante l'accelerazione fino all'energia di regime per poi mantenerla costante. LHC usa otto cavità per fascio: ciascuna fornisce 2 MV a 400 MHz, frequenza a cui viene fatto oscillare il campo elettromagnetico al loro interno. Le cavità operano a 4.5 K e sono raggruppate in quattro criomoduli (due per fascio).

Tabella 1.1: Parametri tecnici principali di LHC.

Quantità	**valore**
circonferenza (m)	26 659
temperatura di lavoro dei magneti (K)	1.9
numero di magneti	9593
numero di dipoli principali	1232
numero di quadrupoli principali	392
numero di cavità a radiofrequenza per fascio	8
energia nominale, protoni (TeV)	7
energia nominale, ioni (TeV/nucleone)	2.76
intensità massima campo magnetico (T)	8.33
luminosità di progetto ($\mathrm{cm^{-2}\,s^{-1}}$)	10×10^{34}
numero di pacchetti di protoni per fascio	2808
numero di protoni per pacchetto (in partenza)	1.1×10^{11}
minima distanza tra i pacchetti (m)	~ 7
numero di giri per secondo	11 245
numero di collisioni per secondo (milioni)	600

In LHC, in condizioni di regime, ciascun fascio di protoni è diviso in 2808 pacchetti, ciscuno contenente circa 10^{11} protoni. La loro dimensione non è costante lungo LHC quando viaggiano lontani dal punto di collisione: la loro grandezza è dell'ordine di qualche centimetro di lunghezza e un millimetro di larghezza; invece nei punti di collisione i protoni vengono collimati e i pacchetti sono compressi fino a dimensioni di circa 16 nm (un capello umano è spesso circa 50 nm). Alla piena luminosità LHC usa pacchetti spaziati tra loro di 25 ns (circa 7 m) che scontrandosi tra loro generano circa 600 milioni di collisioni per secondo.

1.2 I rivelatori ("detector") a LHC

Il Large Hadron Collider accelera i fasci fino alla loro energia di regime per poi farli collidere in quattro punti, intorno ai quali sono stati costruiti e installati quattro rivelatori di particelle. L'obiettivo è riuscire a tracciare e caratterizzare tutti i tipi di particelle prodotte in ogni collisione in modo da poter ricostruire tutti i processi fisici che le hanno generate.
Sono quattro gli esperimenti maggiori istallati a LHC:

- A Large Ion Collider Experiment (ALICE)
- A Toroidal LHC ApparatuS (ATLAS)
- Compact Muon Solenoid (CMS)
- Large Hadron Collider beauty (LHCb)

Essi sono istallati in quattro grandi camere costrite attorno ai quattro punti di collisione dei fasci. Oltre ad essi, esistono anche due esperimenti minori: TOTEM, istallato vicino a CMS e LHCf, installato vicino ad ATLAS.

ALICE

ALICE [4, 5] è un rivelatore specializzato nell'analisi delle collisioni tra ioni pesanti. Le sue caratteristiche principali sono illustrate in tabella 1.2. ALICE, ricreando condizioni simili a quelle presenti pochi istanti dopo il Big Bang, studia le proprietà del plasma di quark e gluoni, stato della materia in cui quark e gluoni non sono più confinati in adroni. L'obiettivo è osservare come il plasma si espande e si raffredda dando origine progressivamente alle particelle che costituiscono l'universo attuale. La collaborazione di Alice conta più di 1500 persone (dato aggiornato a Ottobre 2014).

Tabella 1.2: Caratteristiche del rivelatore ALICE.

Dimensione	lunghezza: 26 m, altezza: 16 m, larghezza: 16 m
Peso	10 000 tonnellate
Design	central barrel plus single arm forward muon spectometer
Costo dei materiali	115 MCHF
Posizione	St. Genis-Pouilly, Francia

ATLAS

ATLAS [6, 7] è un detector costruito per coprire molti dei settori della fisica delle alte energie studiate a LHC e il più grande detector al mondo mai costruito. Le sue caratteristiche principali sono illustrate in tabella 1.3. Un totale di più di 3000 persone lavorano attivamente in ATLAS nel processamento e anaisi dei dati da esso raccolti (dato aggiornato a Febbraio, 2012). La caratteristica principale è il suo enorme sistema di magneti. Esso consiste in otto bobine magnetiche superconduttrici lunghe 25 metri a forma cilindrica, lungo il cui asse passano i fasci.

Tabella 1.3: Caratteristiche del rivelatore ATLAS.

Dimensione	lunghezza: 46 m, altezza: 25 m, larghezza: 25 m
Peso	7000 tonnellate
Design	barrel plus andcaps
Costo dei materiali	540 MCHF
Posizione	Meyrin, Switzerland

CMS

CMS [8, 9] è un detector a carattere generale come ATLAS, ma che si differenzia da quest'ultimo per le differenti soluzioni tecniche e per la sua struttura. É costruito attorno a un unico grande solenoide superconduttore di forma cilindrica in grado

di generare un campo magnetico di 4 T. Le sue caratteristiche principali sono illustrate in tabella 1.4. Nella prossima sessione sarà oggetto di una descrizione più approfondita.

Tabella 1.4: Caratteristiche del rivelatore CMS.

Dimensione	lunghezza: 21 m, altezza: 15 m, larghezza: 15 m
Peso	12 500 tonnellate
Design	barrel plus end caps
Costo dei materiali	500 MCHF
Posizione	Cessy, France

LHCb

LHCb [10, 11] è specializzato nello studio dell'asimmetria debole tra materia e antimateria presente nelle interazioni di particelle contenenti il quark bottom. Le sue caratteristiche principali sono illustrate in tabella 1.5. LHCb è costituito da una serie di rivelatori: il primo è costruito attorno al punto di collisione e gli altri sono disposti in serie lungo una lunghezza di 20 metri. LHCb conta circa 700 membri (dato aggiornato a Ottobre 2013).

Tabella 1.5: Caratteristiche del rivelatore LHCb.

Dimensione	lunghezza: 21 m, altezza: 10 m, larghezza: 13 m
Peso	5600 tonnellate
Design	forward spectometer with planar detectors
Costo dei materiali	75 MCHF
Posizione	Ferney-Voltaire, France

LHCf e TOTEM

LHCf [12, 13] è piccolo esperimento che si occupa di raccogliere dati sulle particelle prodotte molto vicino alla direzione del fascio durante le collisioni p-p. Lo scopo principale dell'esperimento è di stimare l'energia primaria dei raggi cosmici ad alta energia. Ha due rivelatori posizionati a 140 m dal punto di collisione attorno al quale è costruito ATLAS. La collaborazione di LHCf conta 30 membri provenienti da 9 istituti in 5 stati (dato aggiornato a Novembre, 2012).

TOTEM [14, 15] ha come obiettivo quello di misurare l'effettiva dimensione della sezione d'urto dei protoni a LHC. Per far ciò si sono costruiti dei detector in delle apposite camere a vuoto chiamate "roman pots" connesse ai condotti in cui trasitano i fasci, in modo da permettere a TOTEM di rivelare le particelle prodotte durante gli urti con direzione prossima a quella del fascio. Ci sono otto roman pots disposte a coppie in quattro posizioni vicino al punto di collisione dell'esperimento CMS. TOTEM conta circa 100 membri provenienti da 16 istituti in 8 stati (dato aggiornato a Agosto, 2014).

1.3 L'esperimento CMS a LHC

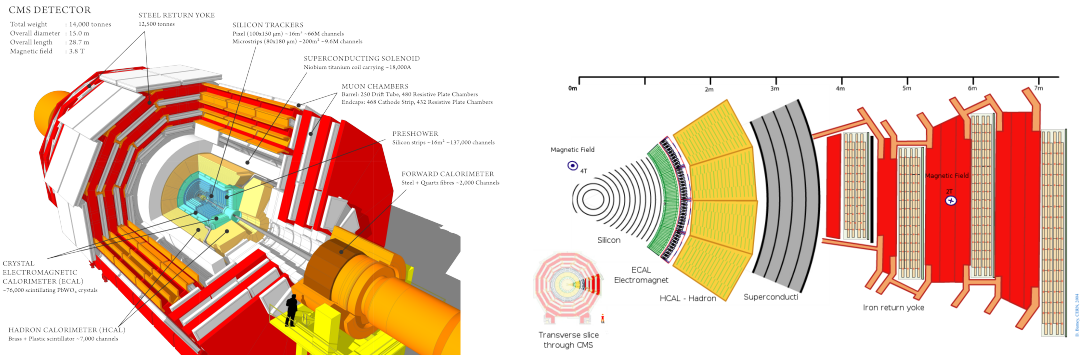
L'esperimento CMS conta circa 4300 persone tra fisici, ingegneri, tecnici, studenti che lavorano attivamente all'esperimento, provenienti da 182 Istituti divisi in 42 Stati nel mondo (Febbraio, 2014). L'esperimento ha come scopo investigare tutta la fisica che viene studiata a LHC, dalla fisica del Modello Standard alla ricerca della nuova fisica. I dati raccolti provengono da un rivelatore, strutturato a strati e di forma cilindrica. Ogni strato del rivelatore si occupa di effettuare misure su particelle di tipi differenti.

1.3.1 Struttura del detector

Il detector è costituito da diversi layer, raffigurati in Figura 1.2. Ognuno di essi è progettato per fermare, tracciare o misurare il percorso e le proprietà fisiche di un differente tipo di particella subatomica generata dalle collisioni (p-p o ioni pesanti). Esso è costruito attorno a un grande solenoide che prende la forma di una bobina cilindrica formata da un cavo superconduttore, che opera alla temperatura di 4.4 K e genera un campo magnetico di 4 T.
I prodotti delle collisioni si muovono inizialmente attraverso un Tracker[9, pp. 26-89], fatto interamente di silicio, che mappa la loro posizione con la risoluzione di 10 nm mentre si muovono dentro al detector. In questo primo modo viene misurato il momento delle particelle cariche. Successivamente esse incontrano due calorimetri, l'Electromagnetic Calorimeter (ECAL) [9, pp. 90-121] e l'Hadron Calorimeter (HCAL) [9, pp. 122-155] disposti in serie. Il primo misura l'energia di fotoni ed elettroni, il secondo degli adroni.
Il Tracker è progettato per interferire il meno possibile con le particelle che lo attraversano, mentre i calorimetri sono appositamente disegnati per fermare le particelle. La bobina superconduttrice è circondata da 12 ferromagneti divisi in tre strati e intervallati ai rivelatori di muoni[9, pp. 162-246], le uniche particelle che, insieme ai neutrini, riescono a superare i calorimetri. Questi ferromagneti sono detti "return yoke" e contengono e modificano il campo magnetico. I neutrini, essendo neutri e non interagendo con il detector, riescono a "fuggire" dal rivelatore. Sono le uniche particelle di cui non si può misurare energia e momento, ciononostante vengono "rivelati" assegnando loro l'energia e il momento mancante nelle interazioni che avvengono con gli urti, è in questo modo possibile stabilirne l'effettivo passaggio e la loro traccia.

Tracker

Il Tracker è uno degli elementi cruciali di CMS poichè la misura del momento delle particelle avviene attraverso il rivelamento della loro posizione: maggiore è la curvatura del loro percorso attraverso un campo magnetico, minore è il momento che esse possiedono. Il Tracker può ricostruire il percorso dei muoni, degli elettroni e degli adroni così come riesce a ricostruire le tracce prodotte dal decadimento di molte particelle con un tempo di vita molto basso, come il quark beauty.
Le caratteristiche principali che lo contraddistinguono sono il basso grado di interferenza con le particelle che lo attraversano e l'alta resistenza alle radiazioni.

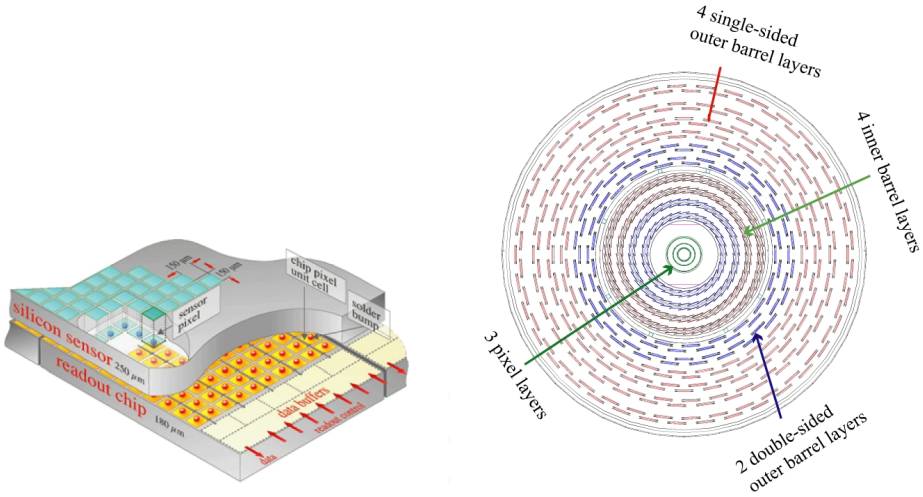


(a) Vista in sezione.

(b) Vista lungo il piano perpendicolare alla direzione del fascio.

Figura 1.2: Il Compact Magnetic Solenoid.

Essendo infatti posizionato nella parte più interna del rivelatore esso riceve anche il maggior volume di particelle. Le interferenze del Tracker con le particelle, necessarie per poter "vedere" la loro posizione, avvengono sono in pochi e predefiniti punti che consentono comunque di tracciare il loro percorso con una precisione molto elevata grazie all'accuratezza con cui vengono effettuate le misure, dell'ordine dei 10 nm. Il design finale di questo layer (Figura 1.3b consiste in un Tracker costruito interamente di silicio: internamente, al centro del detector, si trovano tre livelli di pixel (pixel detector) e dopo questi le particelle attraversano dieci livelli di strip detectors, fino a raggiungere un raggio di 130 cm dalla beam pipe.
Il pixel detector, raffigurato in Figura 1.3a, contiene circa 65 milioni di pixel, i suoi tre livelli sono posti a un raggio rispettivamente di 4, 7 e 11 cm. Essendo così prossimo al fascio il flusso di particelle è massimo: a 8 cm è di circa 10 milioni di particelle per centimetro ogni secondo. Il pixel detector è in grado di identificare e ricostruire le loro tracce con alta precisione. Ogni livello è diviso in piccole unità, ciascuna contenente un sensore di silicio di 150 nm $\times$ 150 nm. Quando una particella carica attraversa una di queste unità deposita abbastanza energia da far sì che venga rilasciato un elettrone e creare di conseguenza una lacuna. Ciascun pixel è collegato con un chip che riceve questo segnale e lo amplifica. Poichè esistono 3 livelli è possibile ricostruire un'immagine 3-D usando layer bidimensionali. A causa dell'enorme numero di canali di alimentazione presenti (uno per ogni pixel), la potenza di ogni pixel deve essere mantenuta al minimo poichè, anche se ciascuno di essi genera all'incirca 50 µW, la potenza totale non è irrilevante. Per questo motivo i pixel sono stati montati in tubi mantenuti a bassa temperatura.
Gli strip detectors consistono invece in dieci livelli divisi in quattro barriere interne e sei barriere esterne. Questa parte del Tracker contiene un totale di 10 milioni di detector strip divisi in 15 200 moduli, letti da 80 000 chip. Ciascun modulo consiste di tre elementi: un insieme di sensori, la sua struttura di supporto e l'elettronica necessaria per raccogliere i dati. I sensori sono studiati per ricevere molte particelle in un piccolo spazio grazie alla loro rapidità di risposta e alla buona risoluzione spaziale. Lavorano in modo molto simile ai pixel: rilevando le correnti generate dalle particelle interagenti con essi, amplificandole e inviando i dati raccolti attraverso gli step successivi. Anche questa parte del detector è mantenuta a bassa



(a) CMS Silicon pixel detector.

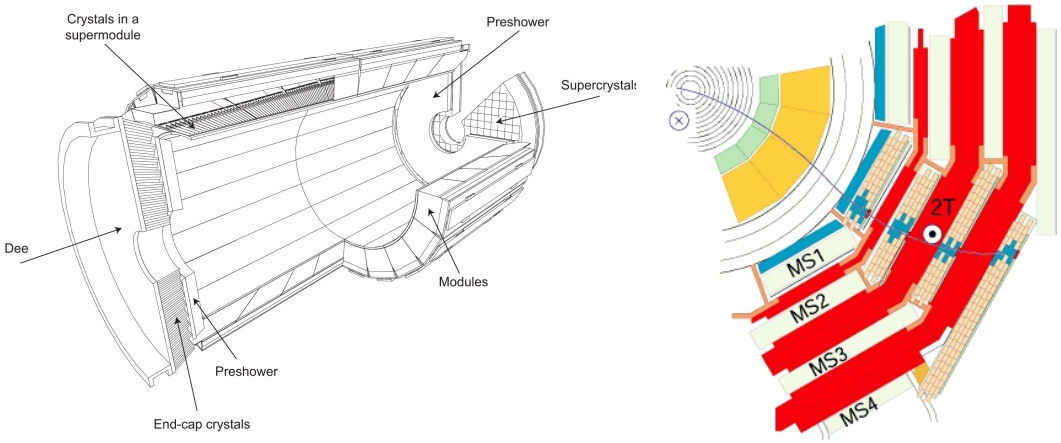
(b) I layer del tracker visti nel piano perpendicolare ai fasci.

Figura 1.3: Il Tracker Detector.

temperatura ($-20\,°C$), in modo da "congelare" i danni alla struttura del silicio dovuti alle radiazioni.

I calorimetri

I due calorimetri presenti in CMS si occupano di misurare l'energia di elettroni, fotoni e adroni.
I primi due tipi di particelle vengono misurate e fermate dal calorimetro elettromagnetico (ECAL). La misurazione avviene all'interno di un forte campo magnetico, in presenza di un alto livello di radiazioni e nei soli 25 ns esistenti tra una collisione e la seguente. Il materiale usato è costituito da tungstenio di piombo ($PbWO_4$), in grado di produrre luce in proporzione all'energia delle particelle. É un cristallo ad alta densità e la produzione di luce avviene in un tempo molto breve e definito che permette una rivelazione rapida, precisa e molto efficace grazie a appositi fotodetector costruiti per lavorare in presenza di un forte campo elettromagnetico. Il calorimetro elettromagnetico è composto da un corpo cilindrico detto "barrel" e dalle due estremità dette "endcaps" (Figura 1.4a) e forma uno strato tra il Tracker e il calorimetro adronico. Il calorimetro per adroni (HCAL) è appositamente costruito per misurare l'energia delle particelle che sono soggette ad interazioni forti e fornisce gli strumenti per ottenere misure indirette di particelle quasi per niente interagenti come i neutrini. Quando gli adroni decadono possono produrre nuove particelle che non lasciano traccia della loro presenza in nessuna parte del detector CMS. Per supplire a ciò l'HCAL è ermetico, cioè cattura ogni particella proveniente dalla collisione, in questo modo si riesce a rivelare, attraverso una violazione apparente della conservazione del momento e dell'energia, la presenza di particelle invisibili. Il calorimetro ad adroni è fatto di una serie di strati fortemente assorbenti. Ogni volta che una particella prodotta dai vari decadimenti a catena attraversa uno strato provoca l'emissione di luce di colore blu-viola. Questa luce viene assorbita da fibre



(a) Layout del calorimetro elettromagnetico che mostra la disposizione dei moduli di cristalli.

(b) Un muone, nel piano perpendicolare ai fasci, che lascia una traiettoria curva nelle quattro stazioni per muoni.

Figura 1.4: L'*Electromagnetic CALorimeter* e il *muon detectors*.

ottiche del diametro di circa 1 mm che cambiano la lunghezza d'onda della luce nella regione verde dello spettro in modo che questa possa essere trasportata da cavi ottici in grado di rivelarla ed essereo trasformata in dati. La somma ottica della luce prodotta lungo la scia di una particella è una misura dell'energia e/o può essere un indicatore del tipo di particella, essa viene effettuata da appositi fotodiodi ibridi appositamente progettati per questo scopo. L'HCAL è diviso in varie parti, una serie di sezioni cilindriche (barrel), le endcaps e due ulteriori sezioni dette "forward sections", posizionate alle estremità di CMS, per rivelare anche le particelle che dopo l'urto si muovono nelle regioni prossime al fascio. Queste ultime due sezioni ricevono gran parte dell'energia totale presente nel sistema delle particelle del fascio e sono costruite con materiali differenti dalle altre sezioni, pensati per renderle maggiormente resistenti alle radiazioni.

I rivelatori di muoni

Il rivelatore di muoni è posto all'esterno del solenoide poichè i muoni sono fortemente penetranti e non vengono fermati dai precedenti calorimetri. In questo modo essi rimangono le sole particelle in grado di lasciare un segnale in tali strati attivi di CMS. Il loro momento viene misurato ricreando il loro percorso lungo quattro stazioni per muoni, intervallate con dei ferromagneti ("return yoke"), mostrati in rosso nella Figura 1.4b e sincronizzando i dati così ottenuti con quelli raccolti nel Tracker. Ci sono 1400 camere a muoni: 250 "drift tubes" (DTs) e 540 "cathode strip chambers" (CSCs) tracciano le posizioni della particelle e agiscono contemporaneamente come trigger, mentre 610 "resistive plate chambers" (RPCs) formano un ulteriore sistema di trigger che decide velocemente se conservare i dati appena acquisiti o no. Le DTs e gli RPOCs sono disposti in cilindri concentrici attorno alla linea del fascio, mentre gli CSCs e gli RPCs formano le endcaps che chiudono le estremità del rivelatore.

1.3.2 Trigger e Data Acquisition

Quando CMS opera a regime sono circa un miliardo le interazioni p-p che hanno luogo ogni secondo all'interno del detector. Poichè l'intervallo temporale tra un urto e il successivo è di 25 ns e una nuova serie di collisioni ha inizio prima che quelle prodotte dallo scontro precedente siano completamente uscite dal detector, i dati vengono conservati in pipelines che possono trattenere e processare le informazioni provenienti da molte interazioni allo stesso tempo. Per non confondere le particelle di due eventi diversi il detector è stato progettato per avere un'ottima risoluzione temporale e un'ottima sincronizzazione dei segnali provenienti da tutti i canali, il cui numero è dell'ordine dei milioni.
Esistono un sistema di trigger[9, pp. 247-282] organizzato su due livelli. Il primo livello è hardware, basato su un processo di selezione dei dati estremamente veloce e completamente automatico che seleziona in base a valori di interesse per la fisica, come ad esempio un alto valore di energia o una combinazione non usuale di particelle interagenti in un evento. Questo trigger agisce asincronicamente e nella fase di ricezione dei segnali ed effettua una prima selezione che riduce la frequenza di dati da acquisire a qualche centinaia di eventi per secondo. Essi vengono poi trasferiti e salvati in appositi spazi di storage per le analisi successive.
Il livello successivo è software, e opera invece in seguito alla sincronizzazione delle informazioni provenienti da tutto il detector, dopo quindi che gli eventi sono stati ricostruiti, analizzandoli in un'apposita farm. Qui gli eventi vengono processati in una farm di processore. La frequenza degli eventi che esce dal secondo livello di trigger è di circa 100 Hz.

Capitolo 2

Il CMS Computing Model

Gli "eventi", che ciascuno dei quattro detector operanti a LHC rivela, sono alla base dei modelli di calcolo, sono registrazioni dei segnali generati dai prodotti delle singole collisioni protone-protone (*p*-*p*) o Pb-Pb. La frequenza delle collisioni *p*-*p* in ciascun rilevatore é 10^9 Hz che equivale approssimativamente alla generazione di 1 PB al secondo di dati per ciascun detector. Il flusso di dati viene selezionato attraverso un sistema di trigger, come già descritto nel precedente capitolo, che riduce la mole di dati prodotti a qualche centinaio di megabytes per secondo. I sistemi di storage del CERN sono in grado di salvare questo flusso di dati: esso viene infatti indirizzato al *CERN Computer Center* per essere archiviato.
Inclusi i dati generati dalle simulazioni fisiche e dei rilevatori, LHC Computing deve gestire approssimativamente 15 PB di dati ogni anno di piena presa dati. Oltre a questa ingente quantità di dati, ogni modello di calcolo a LHC[16] deve garantire l'accesso e la possibilità di avviare job in ogni parte del mondo a ogni utente di LHC, senza che essi si trovino necessariamente al CERN. I requisiti base necessari per costruire un sistema di calcolo del genere sono quindi:

- poter gestire un grande volume di dati in modo efficace
- avere lo spazio per immagazzinare e trattare i *raw data* (CPU e storage)
- acconsentire a migliaia di utenti di accedere ai dati
- poter archiviare i dati ottenuti

Inoltre il *computing environment* deve essere in grado di gestire processi di analisi, in quella che viene chiamata *chaotic user analysis*. Si è così sviluppata l'idea di creare un'infrastruttura di risorse, servizi e strumenti basati sul calcolo distribuito, chiamata oggi Worldwide LHC Computing Grid[17] (WLCG), a cui ciascun esperimento possa aggiungere un proprio layer delle applicazioni servendosi di un unico middleware comune a tutti, fornito dai progetti Grid asiatici, europei ed americani (EGEE[18], EGI[19], OSG[20]).
Ciascun esperimento che opera su Grid deve dotarsi di un modello di calcolo (Computing Model) adeguato. Esso comprenderà l'insieme di tutte le componenti hardware, software che sono state sviluppate per far fronte alla raccolta, distribuzione e analisi della grandissima mole di dati prodotta e la gestione e l'interazione

di ciascuna di queste componenti attraverso un certo numero di strumenti e servizi mantenuti in tempo reale.

I progetti di middleware sopra citati forniscono alla comunità *High-Energy Physics* (HEP) (e non solo) il cosiddetto middleware, che consente di operare su Grid. Ciascun esperimento HEP, tra cui gli esperimenti LHC (e dunque anche CMS), utilizzano una serie di componenti tra quelli esistenti a questo livello, che chiemeremo "middleware layer", ma ciascun esperimento aggiunge anche del software autonomamente progettato e sviluppato per svolgere funzioni di specifico interesse di quell'esperimento: tale software rappresenta quello che nel seguito chiameremo "application layer". Le soluzioni comuni a piu' esperimenti, cosi' come quelle specifiche di un solo esperimento, devono tuttavia essere tutte ugualmente capaci di operare in modo coerente (e insieme agli altri esperimenti, o Virtual Organizations) su risorse che consistono in centri di calcolo sparsi in tutto il mondo.

Il Computing Model di CMS [21, 22], di interesse in questo lavoro di tesi, viene descritto nelle successive sezioni, con particolare attenzione all'uso delle risorse e al settore del Data Management.

2.1 CMS Computing Resources

Le infrastrutture di calcolo su cui si appoggiano i servizi della Grid e che sono parte integrante di WLCG sono divise in *Tier*: per Tier si intende un centro di calcolo che fornisce capacità di storage, potenza di CPU e connettività di rete. La struttura a Tier e' stata formalizzata dal modello MONARC [23], che prevede una gerarchia abbastanza rigida. A ogni Tier è associato un numero: minore è tale numero e maggiori sono la quantità dei servizi e delle funzionalità offerte, nonché anche (almeno tipicamente) la conseguente dimensione del sito stesso in termini di storage, CPU e network, ed anche la cosiddetta "*availability*" richiesta (24/7 o inferiore). Un centro di calcolo che partecipa a WLCG per uno o piu' esperimenti LHC, potrà dunque teoricamente ricoprire ruoli diversi e svolgere funzioni differenti per ATLAS rispetto a quelle che svolge per CMS (ad esempio, essere un Tier-1 in WLCG sia per ATLAS che per CMS, ma svolgere le funzioni di un Tier-2 solo per CMS), oppure scegliere di servire le necessità di un solo esperimento (ad esempio, essere un Tier-1 interamente dedicato a CMS e non usato da altri esperimenti). Attualmente il modello di calcolo distribuito di CMS utilizza le seguenti risorse in WLCG:

- 1 Tier-0 Centre al CERN (T0);
- la CMS Analysis Facility al CERN (CMS-CAF);
- 8 Tier-1 (T1), considerando anche le funzioni di T1 svolte dal CERN;
- 52[1] Tier-2 (T2).

Originariamente nel modello esistevano anche Tiers di ordine superiore. Oltre ai siti sopra-citati, in CMS oggi si parla anche di centri Tier-3, risorse interamente

[1] A seguito dell'uscita da CMS del Tier-2 di Taiwan in Gennaio 2015, nel lavoro eseguito nel Capitolo 4 il numero dei Tier-2 analizzati è pari a 53.

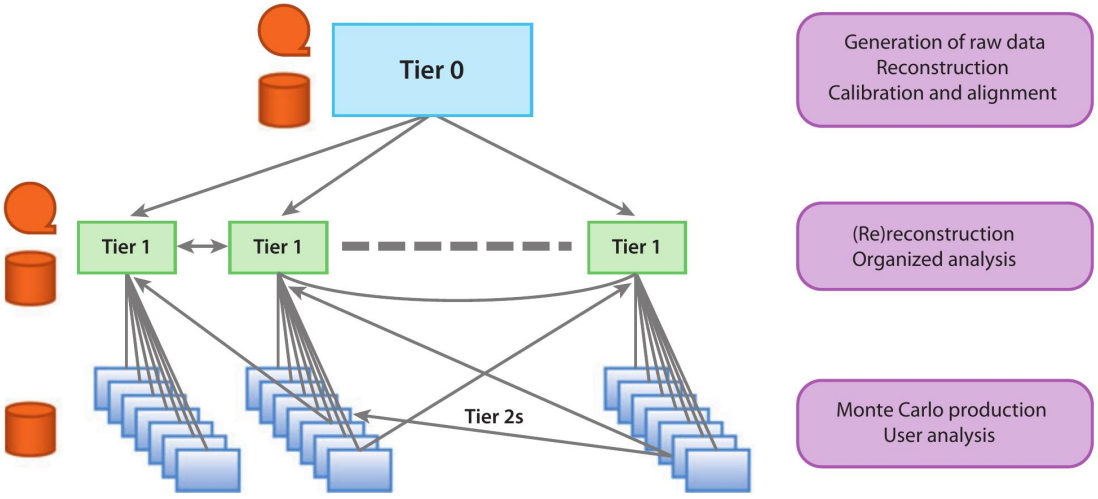


Figura 2.1: Struttura a tier del CMS computing model [16]. Oggi il numero e la tipologia dei link presenti tra i vari tier è fortemente aumentato.

Grid-compliant, ma che non firmano alcun *Memorandum of Understanding* e dunque non sono vincolate a garantire alcun livello di "update" dei servizi nel tempo. Si tratta di risorse dedicate quasi interamente all'analisi dell'utente finale e al supporto delle comunità' di analisti locali; in alcuni casi si tratta di risorse anche ingenti e molto flessibili, ma tipicamente posso avere dimensioni variabili fino a farm locali relativamente piccole. La maggior parte degli utenti che fa analisi distribuita finora si è relazionata principalmente con le risorse dei Tier-2 e i Tier-3 per eseguire le loro analisi, mentre i Tier-1 e il Tier-0 sono risorse dedicate alle attività di processing "schedulato".

2.1.1 I Tiers di CMS

Tier-0 e CMS-CAF

Il Tier-0 e la CMS CERN Analysis Facility (CAF) si trovano al CERN. Dal 2012 il Tier-0 è stato esteso collegandolo al Wigner Research Centre for Physics a Budapest, che opera in remoto e garantisce una maggiore *availability* consentendo al T0 di essere operativo anche nel caso di problemi nella sede principale. Il ruolo del Tier-0 è di ricevere i RAW data dal rilevatore e di archiviarli su nastri, ragguppandoli in flussi di dati e avviando una prima ricostruzione rapida degli eventi. Il T0 di CMS classifica i dati ricostruiti (raggruppati in RECO) in ulteriori 50 dataset primari e li rende disponibili ad essere trasferiti ai T1. Il meccanismo attraveso cui le policy di *integrity* e *availability* sono applicate prevede che i dati siano immagazzinati in modo sicuro in due copie, una conservata al CERN ("*cold copy*") e una distribuita ai Tier-1 ("*hot copy*"). Il compito di inviare la copia ai T1 spetta sempre al T0: per questo proposito la capacità di trasferimento dei dati è fondamentale e si è creata una infrastruttura di rete su fibra ottica dedicata che consente di raggiungere velocità di circa 120 Gbps con i Tier 1.
La CMS-CAF ha lo scopo di fornire supporto a tutte quelle attività che richiedono un tempo di latenza molto ridotto e un accesso asincrono molto rapido ai RAW

data al T0, come la diagnostica del rilevatore, servizi relativi alla gestione della performance del trigger o calibrazioni.

Tier-1

Gli otto Tier-1 sono dislocati in varie parti del mondo (CERN, Germania, Italia, Francia, Spagna, Inghilterra, U.S.A e Russia). I T1 hanno la responsabilità dello storage dei dati e del loro riprocessamento. Il loro ruolo principale è quello di mettere a disposizione spazio fisico in cui conservare le copie "attive" dei dati di CMS reali e simulati, in modo che essi possano essere velocemente acceduti. É inoltre necessario che abbiano una notevole potenza di calcolo e una cache disco veloce. Ciò permette che i dati possano essere riprocessati ad ogni ricalibrazione, selezionati in modo da scegliere quelli più utili ai fini dell'analisi fisica e trasferiti velocemente ai T2, attraverso una WAN con capacità di trasferimento di circa 10 Gbps.

Tier-2

I 52 Tier-2 hanno come compito primario l'analisi dei dati da parte degli utenti via Grid e la produzione di eventi Monte Carlo. Essi hanno bisogno essenzialmente di storage disco performante, di CPU e di una buona connessione (1-10 Gbps). Quest'ultima è necessaria per poter sostenere il flusso di dati che hanno con i Tier-1, comprendente sia la ricezione di dati sia l'invio degli eventi simulati prodotti.

2.2 Il CMS Data Model e Simulation Model

Il workflow della "data reconstruction" in CMS consiste nel passaggio dalle informazioni contenute nei RAW data, attraverso successivi stadi di processamento, fino a formati contenenti oggetti di interesse per le analisi fisiche. Tipicamente, sia nella ricostruzione "prompt" immediata (effettuata al T0) sia nei successivi riprocessamenti dei dati nel corso del tempo (effettuata ai T1), viene eseguita un'applicazione abbastanza simile, e che produce piu' output che vengono successivamente processati in uno stadio di "skimming" che raggruppa in "dataset" specifici gli eventi di interesse per particolari tipologie di analisi. Lo stadio finale e' dunque costituito da tipologie di dati derivati ("derived data") che contengono tutte le informazioni utili e necessarie alla larga maggioranza degli analisti finali.

Il workflow della "simulation reconstruction", invece, prevede che venga eseguito un primo step ("kinematics") basato su vari generatori di eventi Monte Carlo, seguito poi da un secondo step ("simulation") che simula la risposta del rivelatore alle interazioni generate, e infine da un terzo step ("reconstruction") in cui, per simulare un bunch crossing reale, l'interazione singola viene combinata con eventi di pile-up e poi ricostruita. Quest'ultimo step, aggiungendo eventi dalle collisioni precedente e successiva, puo' richiedere l'aggiunta di centinaia di eventi di minimum-bias, e dunque risulta uno step molto impegnativo dal punto di vista dell'I/O sulle risorse di calcolo.

Al termine di queste fase, il CMS data model prevede che i formati di riferimento siano chiamati AOD e AODSIM rispettivamente. Maggiori dettagli sui formati intermedi si possono trovare in tabella 2.1.

Tabella 2.1: Formati principali dei dati raccolti da CMS.

RAW	Raw Data, sono i dati grezzi così come escono dal rilevatore
ESD o RECO	Event Summary Data, contengono tutte le informazioni ottenute dopo la ricostruzione, incluso tutto ciò che c'è nei RAW.
AOD	Analysis Object Data, contengono tutti gli eventi, ma hanno solo la parte delle informazioni dei RECO maggiormente usate nell'analisi.
AODSIM	Simulated Analysis Object Data, contengono tutti gi eventi generati dalle simulazioni.

2.2.1 CMS data organization

I RAW data prodotti da CMS, una volta rifiniti dai vari livelli di ricostruzione, sono soggetti anche a una divisione in "gruppi di dati" detti *dataset*. Essi rappresentano un insieme coerente definito principalmente dai criteri con cui è stato processato e dalla sua processing history. La loro dimensione varia considerevolmente, solitamente nel range di 0.1-100 TiB[2]. Al loro interno presentano un'ulteriore struttura interna: i file contenuti in un dataset sono organizzati in blocchi di 10-1000 file, per aumentare la scalabilità del sistema. Questi insiemi di file, detti *fileblock*, sono il gruppo di dati più piccolo spostabile attraverso la Grid, e dunque rappresentano de-facto la granularità del data management di CMS.

2.2.2 CMS data Flow

Il *data flow* della Grid è correlato con la sua architettura a Tier (Figura 2.2). Il Tier-0 ha il compito di ricostruire gli eventi a partire dai RAW data che gli vengono trasmessi in tempo reale dalla DAQ. La dimensione dei RAW data che CMS ha prodotto durante il Run-1 è di circa 4.5 PB all'anno. Una copia dei RAW viene immagazzinata nel T0 e un'altra impacchettata insieme al file Event Summary Data prodotto nella ricostruzione (a cui in CMS ci si riferisce anche come RECO[26]) e spedita ai T1. Il file RECO contiene un output più dettagliato dell'output

[2] Il TiB (*tebibyte*), ***te**ra **bi**nary **byte***, è un'unità di misura comunemente usata per l'informazione. I prefissi per multipli binari sono stati istituiti dalla International Electrotechnical Commission (IEC) [24] come uno standard per l'informatica nel 1998.
Il *tebibyte* rappresenta $1\,\text{TiB} = 2^{40}\text{byte} \simeq (1+10\%)\text{TB} = (1+10\%)10^{12}\text{byte}$

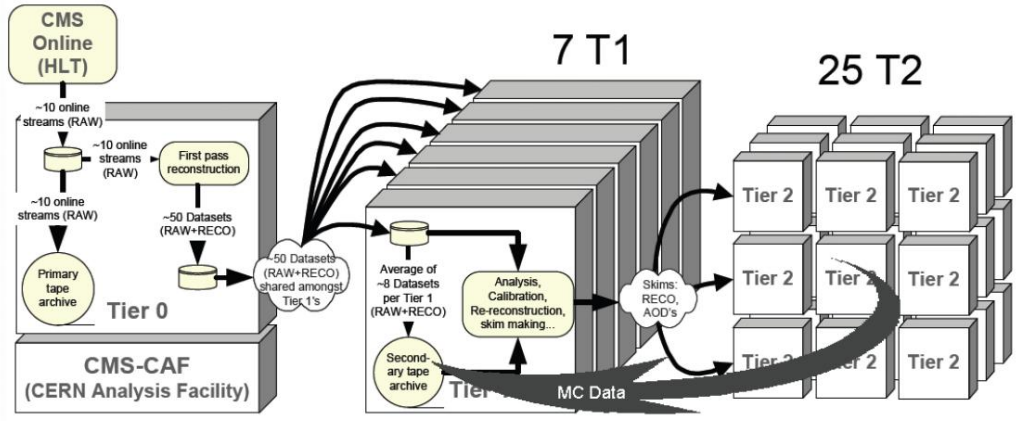


Figura 2.2: flusso dei dati, MC e detector data, attraverso i Tier [25, p. 13].

del rilevatore, ha al suo interno tutte le informazioni aggiuntive fornite dalla ricostruzione degli eventi. I Tier-1 eseguono un ulteriore lavoro di ricostruzione e scrematura dei dati, creando gli Analysis Object Data (AOD[27]), questi contengono tutti gli eventi e le informazioni principali che riassumono ogni evento ricostruito e che servono per le analisi più comuni e che vengono conservati sui T1 e in larga parte anche sui T2, che li utilizzano per analisi e simulazioni Monte Carlo.

2.2.3 CMS data location

CMS non utilizza un catalogo centrale per conservare le informazioni relative alla posizione di ciascun file in ciascun sito. Ad ogni file corrisponde biunivocamente un *Logical File Name*. Si ottiene la conoscenza dell'esistenza e della posizione di un file attraverso un *mapping*, effettuato e conservato dai *central services* della grid, di questo LFN con i siti in cui esistono repliche del file. Ciascun sito contenente repliche ha al suo interno un catalogo locale, il *Trivial File Catalog* (TFC), che contiene a sua volta un mapping locale del LFN con il *Physical File Name* (PFN). I vantaggi legati a questa implementazione e gestione dei dati sono dovuti alla libertà lasciata ai vari siti relativamente alla gestione del loro storage: non è infatti necessario contattare i *central services* della Grid per operazioni interne al sito e in ogni caso in cui il mapping LFN-LFN venga modificato, in quanto ad ogni sito basta aggiornare il proprio TFC locale.

2.3 CMS computing services and operations

Il CMS Computing Model segue un paradigma *data-driven* dove sono i jobs a muoversi verso i dati, e si possono identificare due sistemi distinti, per quanto direttamente interagenti: il *CMS Workload Management System*[28] che ha lo scopo di gestire il flusso dei jobs e il *CMS Data Management System*[29] a cui è destinata la gestione dei dati. A essi si aggiungono altri servizi più specifici volti a eseguire compiti particolari nel merito dei quali in questa tesi non si entrerà.

2.3.1 CMS Workload Management System

Il *Workload Management System* (WMS) di CMS si appoggia in larga misura su servizi centrali (middleware-layer) offerti da Grid per la gestione dei job, quali le infrastruttura con job "pilota" ("pilot jobs"), il sistema informativo, eccetera. Tuttavia, vi sono strumenti di esperimenti, sviluppati e mantenuti da CMS, sia per le attività di processing schedato (quali il riprocessamento degli eventi) sia per le attività di analisi distribuita. L'insieme di questi strumenti consente a CMS di eseguire un numero dell'ordine di 100'000 jobs sui Tiers di WLCG che supportano l'esperimento CMS. Tali job sono basati sul *CMS software framework* (CMSSW) [30], e l'output di tali job viene registrato del *Data Management System* (vedi sezione successiva). Alcune delle caratteristiche principali del WMS di CMS a cui si è arrivati attraverso un percorso composto da miglioramenti successivi sono: l'ottimizzazione degli accessi in lettura, la minimizzazione della dipendenza dei job dai Worker Nodes (WN) di Grid e quindi la riduzione di effetti bottle-neck su di essi, la distribuzione dei job in base alle policy che devono soddisfare e alla priorità all'interno della Virtual Organization. Nel presente lavoro di tesi ci si concentra principalmente su aspetti legati alla parte di data management, descritta nella sezione successiva.

2.3.2 CMS Data Management System

Il CMS Data Management System (DMS) si appoggia sia sui servizi grid sia su servizi di CMS. Il suo scopo è garantire un'infrastruttura e dei tool atti a trovare, accedere e trasferire i diversi tipi di dati ottenuti da CMS. I compiti che si incarica di gestire sono:

- mantenere un *data bookeeping catalog* che descrive i contenuti dei dati in termini fisici,
- mantenere un *data location catalog* che conserva in memoria la locazione e il numero di repliche dei dati,
- gestire il *data placement* e il *data transfer* dei dati.

I componenti, principali, che si occupano di adempiere a questi compiti sono:

- **PhEDEx** [31, 32], il *data transfer and location system*. Gestisce il trasporto dei dati nei vari siti di CMS e tiene traccia di quali dati esistono e dove si trovano.
- **DBS** [33], il *Data Bookkeeping Service*, un catalogo di metadati relativi alle simulazioni di montecarlo e ai dati provenienti dall'esperimento. Contiene record di quali dati esistono, le informazioni di provenienza, la relazione tra dataset e file, in modo da consentire di trovare ogni particolare sottinsieme di eventi dentro un dataset su un totale di circa 200 000 dataset e più di 40 milioni di file.
- **DAS**, il *Data Aggregation System* [34], creato per fornire agli utenti una interfaccia [35] uniforme e coerente a informazioni di data management registrate su sorgenti multiple.

Tutte le componenti del DMS sono progettate e implementate separatamente. Esse interagiscono tra di loro e con gli utenti della Grid come web services.

2.4 PhEDEx

PhEDEx, acronimo di *Physics Experiment Data Export*, ha il compito di gestire i trasferimenti di CMS attraverso la Grid in maniera sicura, affidabile e scalabile. É basato su un cluster di database Oracle[36] situato al CERN, il *Transfer Management Data Base* (TMDB): esso contiene informazioni sulla posizione delle repliche dei dati e sui tasks attivi su tali dati. Dispone di due interfacce: un sito web[37], attraverso cui gli utenti possono richiedere il trasferimento di dataset o di fileblock interattivamente, e un web data service[38] creato per l'interazione di PhEDEx con altre componenti del Data Management. Una volta effettuata una richiesta attraverso una delle interfacce, PhEDEx si connette al TMDB per ottenere i metadati necessari e riscrive i risultati ottenuti una volta effettuato il task. Il TMDB è progettato in modo da minimizzare le *locking contention* tra i diversi agenti e demoni eseguiti da PhEDEx in parallelo ed è disegnato per ottimizzare l'utilizzo della cache, evitando il più possibile problemi di coerenza al suo interno. Gli agenti eseguiti centralmente al CERN eseguono la maggior parte del lavoro di individuazione e ottimizzazione del *data routing*, calcolando il percorso meno costoso in termini di performance. Ciò avviene prendendo in considerazione le performance dei link usati precedentemente per connettere il sito di destinazione con quello di origine dei dati da copiare, basandosi sulle percentuali di trasferimenti riusciti tra essi in una determinata finestra temporale. Una volta effettuata la scelta del percorso, gli agenti responsabili per il dowload, in esecuzione su ogni sito, ricevono i metadati necessari dal TMDB e iniziano il trasferimento usando plugin specifici a seconda di quale dei vari middleware Grid si sta usando. Il successo di ogni trasferimento o di ogni cancellazione di dati in un sito della grid è verificato indipendentemente per ogni fileblock e in caso di fallimento si attivano altri agenti che cercano di completare la richiesta. Tutti i dati relativi alle performance vengono costantemente registrati nel TMDB e possono essere visualizzati attraverso la dashboard di PhEDEx.

2.5 Il CMS Remote Analysis Builder

Il CMS Remote Analysis Builder (CRAB) [39, 40] è un tool sviluppato per l'analisi distribuita [41] di CMS. Si occupa di creare, sottomettere e monitorare i job di analisi di CMS su Grid. É disegnato per prendersi cura del rapporto con ogni singolo componente Grid, consentendo così un utilizzo trasparente di tutto il sistema senza rendere necessario da parte dell'utente conoscere il dettaglio della sua complessità e massimizzando così il tempo del fisico dedicato all'analisi. CRAB fornisce ad ogni fisico l'accesso a tutti i dati raccolti e prodotti dall'esperimento indipendentemente dalla loro posizione geografica (ovvero in quale Grid storage element risiedono).
Il paradigma di analisi dei dati in CMS è, come già detto nel paragrafo precedente,

data-location driven; sono le analisi degli utenti ad essere effettuate dove i dati sono conservati. Gli step previsti da CRAB nell'effettuare l'analisi distribuita sulla Grid sono:

1. eseguire localmente i codici di analisi su samples locali in modo da testare la correttezza del loro workflow;

2. selezionare l'insieme di dati completo su cui effettuare l'analisi;

3. avviare l'analisi: CRAB trasporta il codice nel sito in cui si trovano i dati da analizzare e restituisce all'utente il risultato della propria analisi e i log dei propri jobs.

Attualmente CMS è in una fase di transizione tra la versione 2 e la nuova versione 3 che implementa una serie di ottimizzazioni a livello di codice e di nuove funzionalità. Una prima versione di CRAB 3 [42] è già utilizzata dalla comunità di analisti CMS, per quanto molti utenti usino ancora CRAB 2: si prevede che la migrazione avrà un boost entro il 2015, prima dell'inizio della presa dati a Run-2.

2.5.1 CRAB3 Architecture and Workflow

La figura 2.3 mostra uno schema semplificato dell'architettura di CRAB3. Gli step del workflow con cui CRAB opera, dalla sottomissione dei job da parte dell'utente alla pubblicazione dei risultati sul database del DBS, sono:

1. Il client di CRAB sottomette la richiesta al server di CRAB.

2. Il server di CRAB inserisce la richiesta in Task (Oracle) Database (Task DB).

3. Un sottocomponente del CRAB server chiamato Task Worker monitora costantemente il Task DB alla ricerca di nuove richieste.

4. Il Task Worker riceve le nuove richieste e le inoltra all'infrastruttura che si occupa della loro sottomissione (GlideInWMS).

5. GlideInWMS cerca dei Worker Node disponibili e vi sottomette i jobs.

6. Il Worker Node, una volta finito di eseguire job, copia gli output files nel *temporary storage* del sito.

7. Il servizio AsyncStageOut trasferisce gli output files dallo storage temporaneo al loro storage permanente.

8. AsyncStageOut, una volta finito il trasferimento, pubblica gli output in DBS.

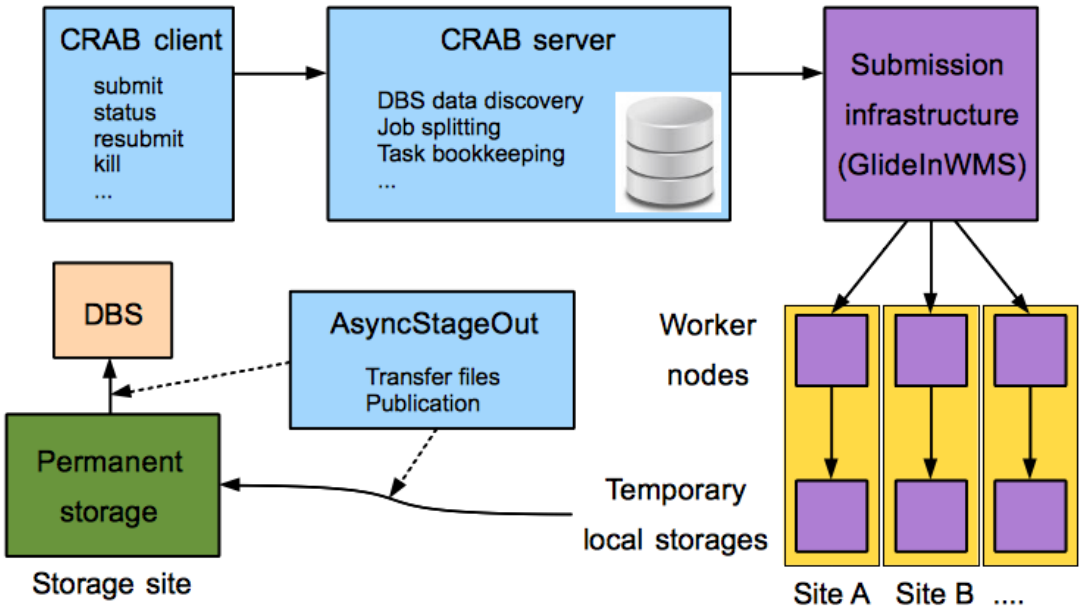


Figura 2.3: Schema semplificato dell'architettura di CRAB3 [43].

2.6 CMS Computing as a fully-connected mesh

Il CMS Computing Model ha subito una profonda trasformazione dalla sua creazione ad oggi. L'evoluzione della sua architettura tende sempre di più a un modello fortemente meno gerarchico di quello originariamente pensato nel Technical Design Report, basato su MONARC, che prevedeva, seppur lasciando una certa flessibilità, che ogni Tier-2 fosse connesso ad un singolo Tier-1 "regionale" [25, p. 20]. Ciò è mostrato anche in figura 2.1, dove ogni Tier-2 è collegato ad ogni singolo Tier-1 e ciascun Tier-1 è collegato solo con un numero fissato e limitato di T2. La *network topology* pensata per il modello originale (figura 2.4) prevedeva:

- un data-flow unidirezionale in uscita dal Tier-0 verso i Tier-1;
- un data-flow tra i Tier-1 regolare e un data-flow tra ciascun Tier-1 e i Tier-2 ad esso collegati;
- ciascun Tier-2 connesso con uno e un solo Tier-1. Il data-flow comprende in entrata i dati di CMS e in uscita le simulazioni di Monte Carlo prodotte e rimandate ai Tier-1 per divenire *custodial* ed essere distribuite sulla Grid.
- i link tra i Tier-2 sono permessi, ma l'aspettativa era che il data-flow in questi link fosse un piccolissima parte del carico complessivo della rete.

Il numero complessivo di link bidirezionali creati da un tale modello è dell'ordine del centinaio.
Durante il Run-1, tuttavia, la flessibilità permessa nello spostamento dei dati attraverso i siti è divenuta una parte importante del modello di calcolo, dinamicizzandone l'architettura. Come si vede in figura 2.5, i link tra gli stessi T2 non sono una parte irrilevante del totale; essi hanno creato la possibilità di accedere ai dati in modo

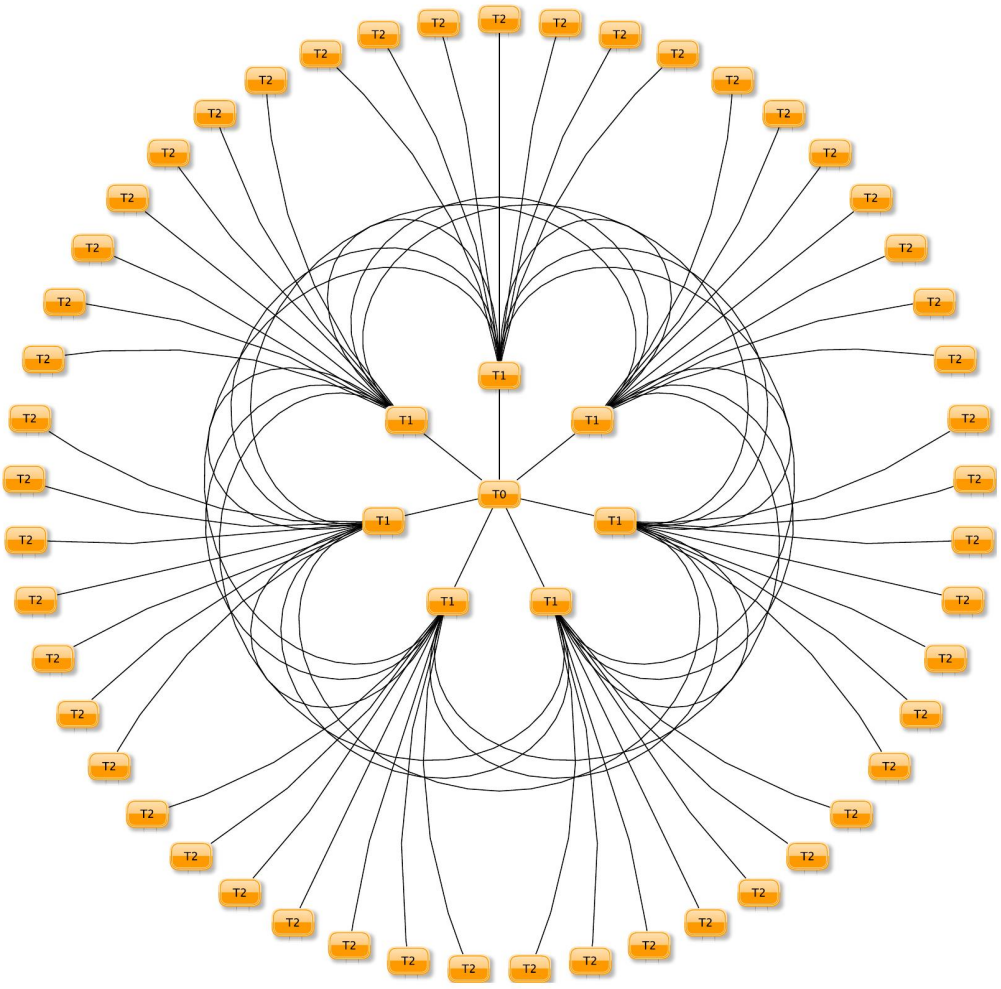


Figura 2.4: La topologia network originale del sistema di calcolo di CMS. Sono presenti $O(100)$ links. [44]

veloce e fluido non ricorrendo necessariamente a un passaggio attraverso i T1. Il CMS Computing Model sta optando sempre di più per una maggiore flessibilitá in modo da poter trattare siti indipendenti come un singolo centro. In questo modo è possibile ottimizzare l'uso delle risorse e processare ogni workflow più velocemente. Alcuni metodi che si stanno sviluppando consistono nel dar la possibilità di leggere in remoto dati da analizzare, senza doverli copiare su un disco locale o di eseguire un singolo processo da diversi siti utilizzando l'accesso ai dati in remoto. Questa evoluzione è già iniziata in Run-1, ma per il Run-2 si prevede una flessibilità maggiore sul luogo in cui i processi sono eseguiti e un accesso ai dati più trasparente [45, pp. 39-48]. L'evoluzione in un modello attuale che può essere ancora bene rappresentato con il termine inglese *fully-connected mesh* trova nella disponibilità di bandwith il suo punto di forza. Oggi l'insieme dei dati trasferiti nei siti appartenenti a CMS raggiunge oltre 1 PiB di traffico ogni settimana e la banda della maggior parte dei link presenti è ancora lontana dall'essere saturata. La connessione presente tra il Tier-0 e i Tier-1, l'*LHC Optical Private Network* (LHCOPN) [46] raggiunge velocità di 120 Gbps. Il numero di link presenti tra i Tier-2 è inoltre incrementato

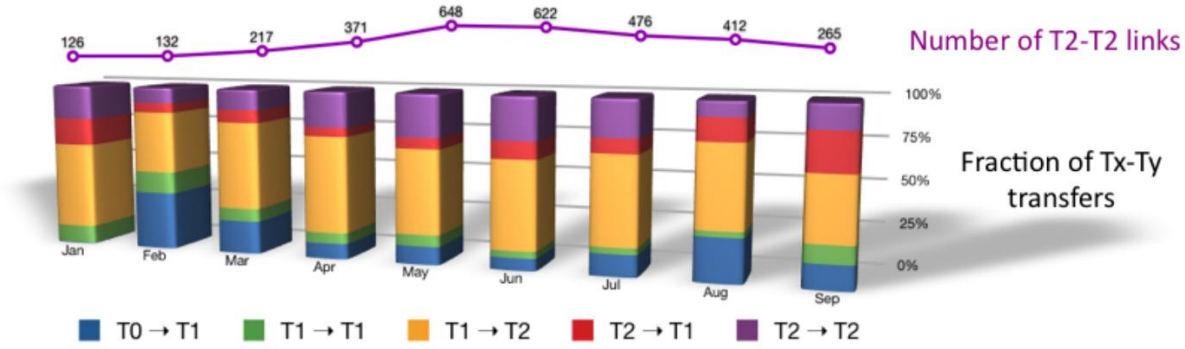


Figura 2.5: Volume trasferito in percentuale sul totale nel corso del 2010 [21].

fino a raggiungere quella che è la *network topology* attuale [44, 47] (rappresentata in figura 2.6) costituita da una rete i cui elementi sono fortemente interconnessi gli uni gli altri: il Tier-0 è collegato direttamente sia con ogni Tier-1 sia con la maggior parte dei Tier-2, con circa 60 link in totale; inoltre i Tier-1 e i Tier-2 sono quasi tutti connessi tra loro, bidirezionalmente, per un ammontare di circa 100-120 link per sito. Sono pochi i siti scarsamente collegati con il resto di CMS e si punta a incrementare la loro connessione. Il numero totale di link è di circa 2000. In questo contestoci si aspetta che un miglior uso della rete porterebbe a un miglioramento effettivo della capacità e velocità di analisi [44, 47]. L'attuale stima della performance è basata su metriche relative a quantità come il tempo di trasferimento di un dataset, il tasso di fallimento in un determinato link o dei batch jobs, il tempo di completamento di un insieme di job o l'efficienza delle CPU.

2.6.1 CMS Data access su WAN e data popularity

I passi avanti in questa direzione compiuti da CMS sono diversi. Nel breve periodo esistono due progetti "pronti". Il primo di questi è il progetto *Any data, Anytime, Anywhere* (AAA) [48] che si pone come obiettivo quello di incrementare l'efficienza delle CPU risolvendo il fallimento dei batch job, causati dalla non lettura di un dato file, attraverso la lettura remota del file attraverso la WAN. Il calo di performance dovute alla lettura sulla WAN anzichè sulla LAN è minore rispetto a quello provocato dal fallimento del job stesso. Il secondo progetto, che verrà descritto in modo più approfondito nel capitolo 3, poichè strettamente legato allo studio effettuato in questa tesi, è quello del *CMS Data Popularity Service*. Misurando e conservando metadati relativi ai pattern di accesso è possibile individuare quali dati rimangono inutilizzati per lunghi periodi di tempo nella Grid, eliminandone le repliche non popolari e ottimizzando l'uso dello storage. Nel lungo periodo, invece, CMS sta investigando, mediante piccoli progetti pilota, la fattibilità di una integrazione più profonda tra le applicazioni di DM di CMS e la network, con l'obiettivo di agire direttamente (application-layer) su parametri di rete durante le operazioni di pianificazione ed esecuzione dei trasferimenti. Si tratta di attività tuttora in fase di R&D e non si prevede si abbiano soluzioni pronte per entrare in produzione prima di Run-3.

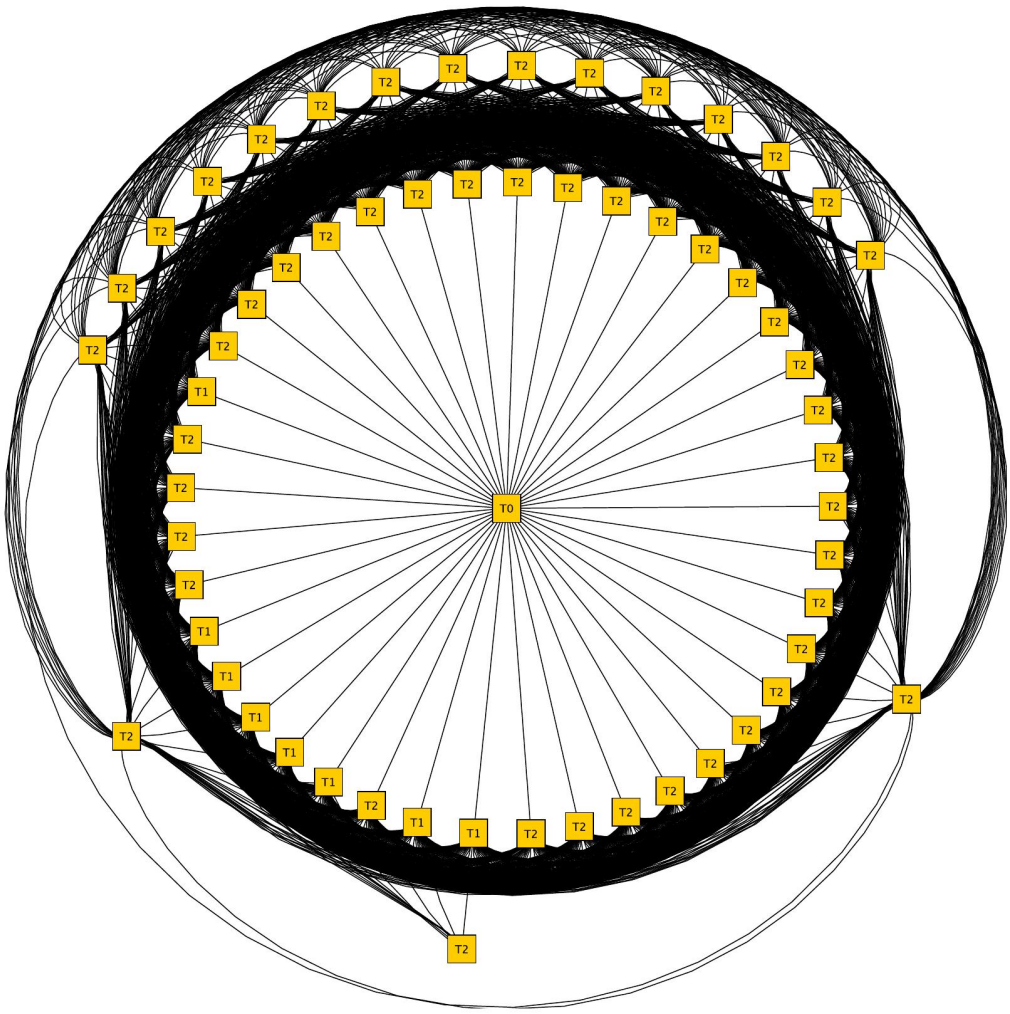


Figura 2.6: La topologia network così come è oggi nel CMS Computing System. Sono presenti $O(2000)$ links. [44]

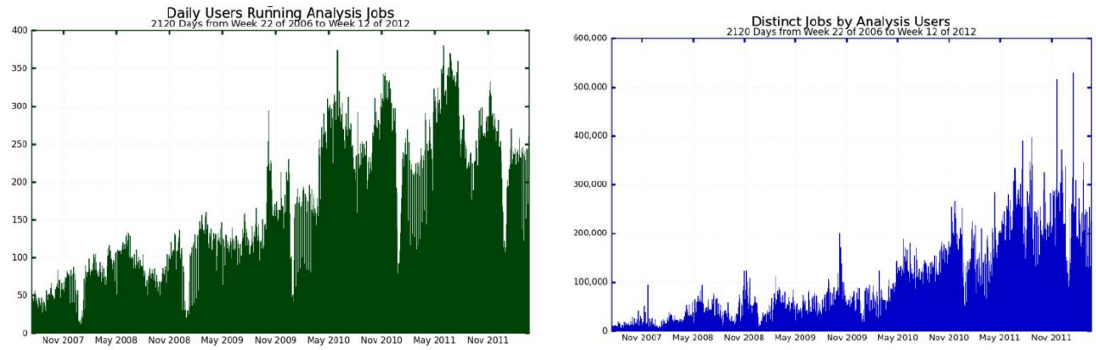
Capitolo 3

Il CMS Popularity Service

Una delle necessità per un esperimento come CMS che fa un ampio uso di risorse distribuite è di disegnare modelli di calcolo basati su soluzioni che consentono di ottimizzare l'uso della rete e dello storage disponibile.
Come mostrato in figura 3.1 [49, p. 2] centinaia di utenti sottomettono circa 200 000 jobs ogni giorno fornendo alla Grid un carico di lavoro che esegue un numero di jobs dell'ordine della decina di milioni. Allo stesso modo la gestione delle risorse di storage riguarda una quantità di dati di circa 15 PiB [49, p. 3](figura 3.2), quantità che si prevede saranno ancora maggiori in occasione dell'ormai prossimo Run-2. La gestione dello spazio fisico e dei job è inoltre fortemente interconnessa poichè le analisi effettuate sulla Grid seguono un paradigma *data-driven*, in cui i dati necessari in input vengono già distribuiti nei siti preventivamente. Oltre a ciò è importante considerare che l'analisi nell'esperimento di CMS è coordinata attraverso un ampio numero di differenti gruppi di analisi (dell'ordine di 20). Ciascun gruppo può usare un determinato spazio in specifici Tier-2, per un totale di più di cento associazione gruppo-sito.
Data la dimensione del sistema di calcolo di CMS, il lavoro volto a gestire e liberare lo spazio richiede un notevole sforzo umano e l'automatizzazione delle procedure gioca un ruolo chiave nel ridurre il carico di lavoro complessivo.

Attualmente non esiste ancora un sistema in CMS in grado di localizzare inefficienze a livello globale, in termini di spazio allocato e inutilizzato dall'analisi distribuita. Ciò provoca nelle operazioni dei sistemi di calcolo una riduzione progressiva dello spazio di storage effettivamente disponibile, di difficile individuazione. Per questo e con l'obiettivo di arrivare ad un data placement dinamico che ottimizzi l'allocazione delle risorse è stato sviluppato il CMS Popularity Service, originariamente ispirandosi all'esperienza di ATLAS [50]. In CMS si introduce il concetto di popolarità dei dati ("data popularity") come un'osservabile in grado di quantificare l'interesse della comunità degli analisti per campioni di dati o simulazioni Monte Carlo specifici, sulla base del numero di accessi, locali o remoti, di successo o falliti, che i job utenti hanno effettuato sui file che storano tali dati su disco. La funzione del Popularity Service è di monitorare quali dati sono più usati misurando la loro popolarità nel tempo relativamente a tutti i livelli possibili di aggregazione dei dati (file, blocchi, dataset).
Le informazioni sono fornite tracciando l'evoluzione nel tempo di:

- nome del dataset;



(a) Utenti attivi giornalmente nella Grid. **(b)** Analysis jobs per utente.

Figura 3.1: CMS Dashboard monitoring.

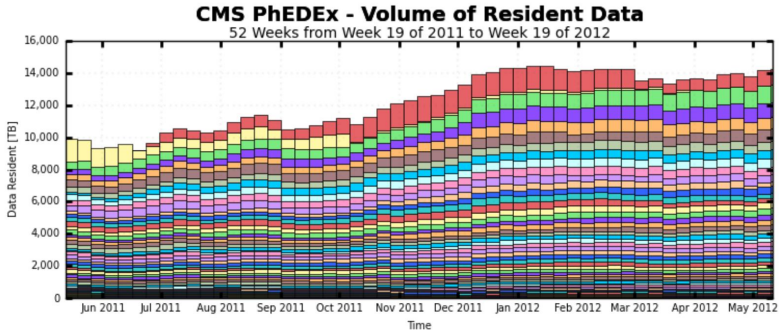


Figura 3.2: Quantità di dati salvati nella grid - PhEDEx Dashboard

- numero di accessi;
- esito degli accessi avvenuti in lettura (successo o fallimento);
- ore di lavore delle CPU nell'accederli;
- numero di utenti unici che eseguono l'accesso;

Il Popularity Service è inoltre stato disegnato in modo da poter fornire dati ad altri servizi, esterni ad esso.

3.1 Architettura del CMS Popularity Service

Il design complessivo dell'infrastruttura del CMS Popularity Service è progettato per rendere il servizio semplice, ottimizzato al data mining. L'infrastruttura raccoglie informazioni sull'utilizzo dei file da differenti servizi, sia specifici di esperimento, come CRAB (descritto nel paragrafo 2.5]), sia non specifici di esperimento, come il protocollo Xrootd [51] per l'accesso diretto ai file da locale e da remoto. Grazie a questo design, il servizio resta relativamente slegato dalle sorgenti di dati da raccogliere, dunque è garantita "by design" una notevole flessibilità ed è possibile estendere il numero delle sorgenti da cui vengono raccolti i dati.
L'architettura del CMS Popularity Service si può suddividere essenzialmente in tre componenti: una relativa alla raccolta dati, una al loro stoccaggio e un'ultima che

rende accessibili questi dati all'esterno attraverso diverse API. La prima componente di questa infrastruttura è composta da una serie di plugins modulari in grado di interfacciarsi alle varie sorgenti dati che vengono eseguiti giornalmente. La seconda componente, centrale in questa infrastruttura e attorno alla quale sono costruite le altre, è un "popularity database" relazionale implementato con un RDBMS Oracle [36] come back-end. Tale database è in grado di aggregare i dati a differenti livelli di granularità (file/blocchi/dataset) correlandoli con gli altri attributi (Tier, user, numero di user, numero di ore di CPU). Le aggregazioni di dati hanno quindi una struttura gerarchica con un primo livello che esegue un'aggregazione sui dati raw e un secondo livello che esegue un'ulteriore aggregazione sul risultato. Sia i dati raw che le viste indicizzate vengono aggiornate su base giornaliera in modo che le informazioni contenute nel database siano sempre aggiornate al giorno precedente. I dati sono infine resi disponibili attraverso un web layer che implementa interfacce multiple da cui poter accedere ai dati.

3.1.1 Data sources per il CMS Popularity Service

L'attività di monitoraggio e di raccolta dati del CMS Popularity Service riguarda la popolarità dei dati di CMS acceduti in analisi. Dato che l'analisi distribuita in CMS si basa essenzialmente sul livello Tier-2, la raccolta dei dati ai fini di studi di popularity si concentra sugli accessi ai dati residenti su uno dei 52 tier2 di cms da parte di ogni job di utenti CMS che fanno analisi distribuita su Grid. I popularity data sono raccolti sfruttando i due strumenti usati per accedere e/o spostare fileblock e dataset su Grid:

CRAB (vedi Paragrafo 2.5) è abilitato per inviare alla Dashboard del CERN [52, 53] le informazioni di tutte le attività dei job. Queste informazioni includono anche le statistiche relative all'utilizzo dei file, e rimangono solo temporaneamente nell'archivio della Dashboard, fino a quando non vengono trasferite nel database del Popularity Service.
Il Popularity Service è completamente indipendente dalla Dashboard ed essa è solo una delle fonti da cui attingere i dati con cui popolare il database. Sono infatti molto diversi gli scopi per cui è stata creata la Dashboard, più incentrata sul monitoraggio in tempo reale dei job piuttosto che sull'aggregazione e l'elaborazione di pattern di utilizzo dei dati acceduti.

Xrood è una suite generica, completa e modulare che si propone di fornire un accesso ai dati in modo veloce e con una latenza molto bassa. Attualmente diversi Tier-2 hanno implementato l'accesso remoto ai dati in essi archiviati attraverso questo protocollo, in modo da fornire una soluzione alternativa ("fall-back") per accedere a dati in tutta la Grid qualora la copia locale fosse temporaneamente non disponibile. In tal modo si aumenta la sicurezza che un job si completi con successo. Le informazioni dei file acceduti tramite questo servizio sono inviate attraverso CRAB al CMS Popularity Service e consentono di monitorare questo comportamento.
Xrootd viene implementato anche nello storage data service dalla CAF di CMS del CERN, basato sul sistema EOS [54].

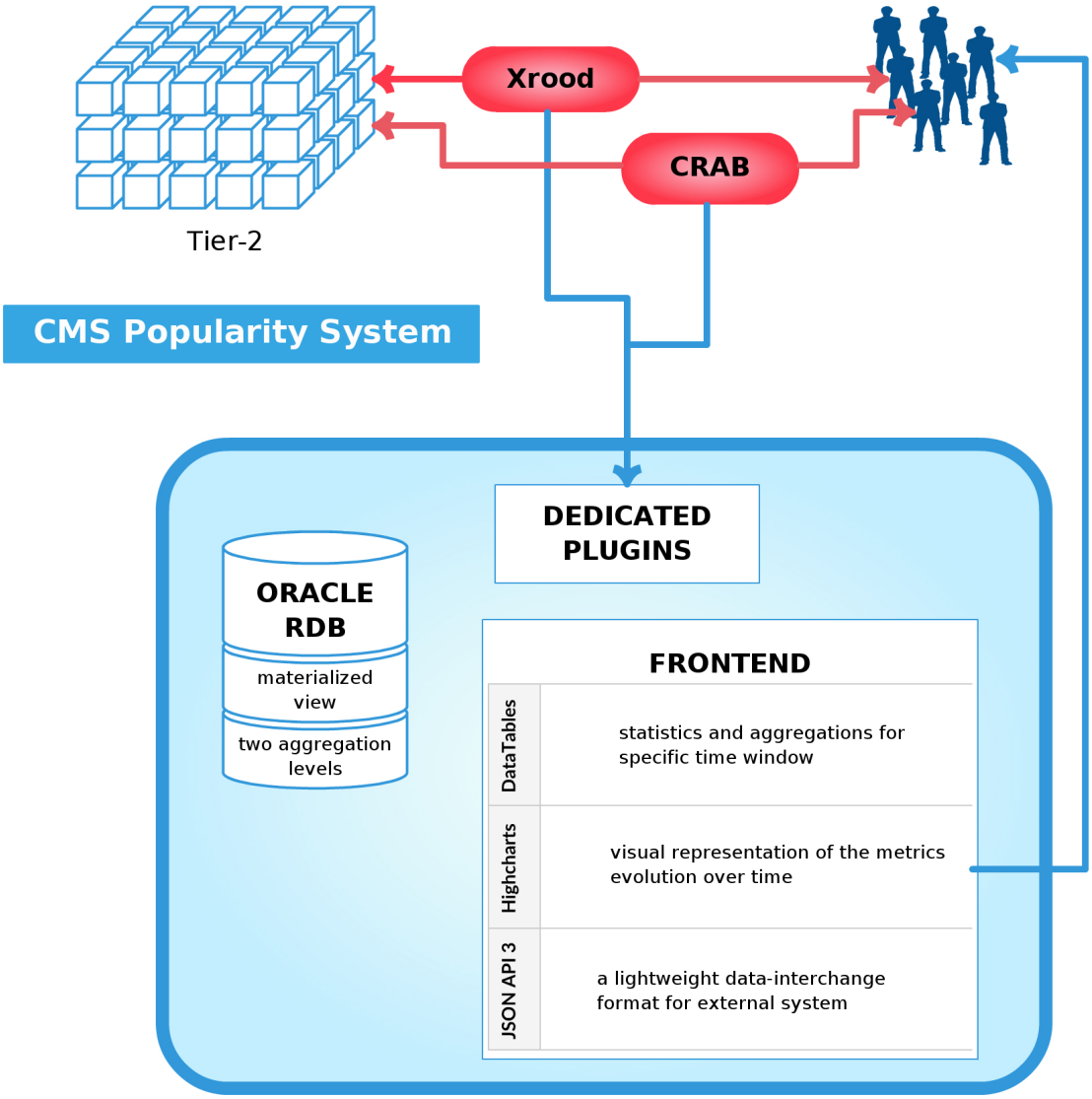


Figura 3.3: Architettura del CMS Popularity Service

3.1.2 Il frontend del CMS Popularity Service

La user interface è implementata attraverso un web service. Esso fornisce attraverso un sito web una visualizzazione grafica delle informazioni, con tavole interattive e grafici in modo che l'applicazione sia facilmente accedibile attraverso un ordinario browser, senza necessità per l'utente finale di installare ulteriori componenti. Il web service fornisce inoltre una API web che consente di recuperare le informazioni sulla popularity formattate in JSON [55], un formato che garantisce interoperabilità con la maggior parte dei linguaggi di programmazione di alto livello garantendo così un alto grado di usabilità. In questo modo è possibile accedere e sfruttare i dati di popularity come sorgente per altri servizi, quali il Victor Site Cleaning Agent (vedi Sezione 3.4 o per l'analisi come quella effettuata nel prossimo Capitolo.

Il web service è sviluppato con Django [56], un framework open source per applicazioni web. Per migliorarne la velocità con cui effettua l'update del grande

volume di dati gestiti, è stato sviluppato un memory-based data caching system (memcached [57]). Il web service è strutturato in modo da seguire un approccio modulare in modo da facilitarne il mantenimento e lo sviluppo. Il nucleo del servizio contiene dei moduli python che forniscono funzionalità quali la validazione degli utenti, gestiscono le comunicazioni http e le connessioni al database. Le librerie javascript DataTables [58] e Highcharts [59] sono successivamente usata per la rappresentazione grafica dei dati (figura 3.4).

3.2 I dati raccolti

Il CMS popularity Service, come descritto precedentemente, fornisce le statistiche d'uso relative a dataset e datatiers acceduti dagli utenti. I dati sono presentati attraverso:

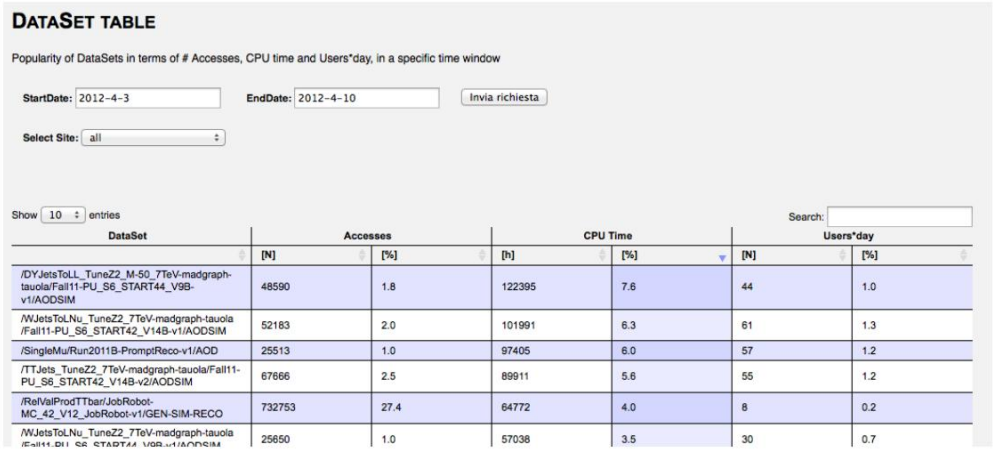
- **tabelle:** l'insieme completo delle statistiche è mostrato in una finestra temporale tramite l'aggregazione dei dataset in un gruppo determinabile di Tier;
- **grafici:** la visualizzazione nel tempo dei dati avviene attraverso vari livelli di aggregazione;
- **JSON API:** permettono di ottenere i dati di popularity per studi specifici o come input ad altri sistemi.

Le statistiche raccolte sono altamente personalizzabili in termini di valori su cui si indaga, periodo di tempo e locazione geografica. I valori possono essere scelti tra ore di CPU e numero di accessi.

Attraverso l'implementazione del Popularity Service è stato possibile creare una serie di strumenti che contribuisco alla gestione dello storage, quali l'identificazione dei file corrotti che causano fallimenti ai job sottomessi da GlideInWMS in CRAB.

3.3 Identificazione dei file corrotti

Il Popularity Service è in grado di identificare i file corrotti che sono causa di fallimenti nei jobs gestiti da CRAB. Identificare velocemente i file corrotti consente il loro replacement prima che altri jobs si interrompano per lo stesso motivo. Questa identificazione non è possibile attraverso i sistemi di monitoraggio dei jobs in CRAB poichè il fallimento viene riportato a livello dei job, in termini di exit code, senza che sia fornita alcuna informazione che identifica il file che l'ha causato. Il Popularity Service ha identificato nei file corrotti la causa di fallimento di circa il 3% dei job con exit code non nullo. Il data mining dei fallimenti è inoltre necessario per garantire un feedback affidabile e evitare falsi allarmi. I file corrotti vengono trovati controllando se il loro accesso risulta sempre errato durante gli ultimi job in cui tali file sono oggetto dell'analisi. Il time range in cui viene effettuata la ricerca di file corrotti è attualmente di dieci giorni, è configurabile e garantisce di non marcare come corrotti quei file che sono temporaneamente non leggibili per un problema tecnico di altro tipo.



DATASET TABLE

Popularity of DataSets in terms of # Accesses, CPU time and Users*day, in a specific time window

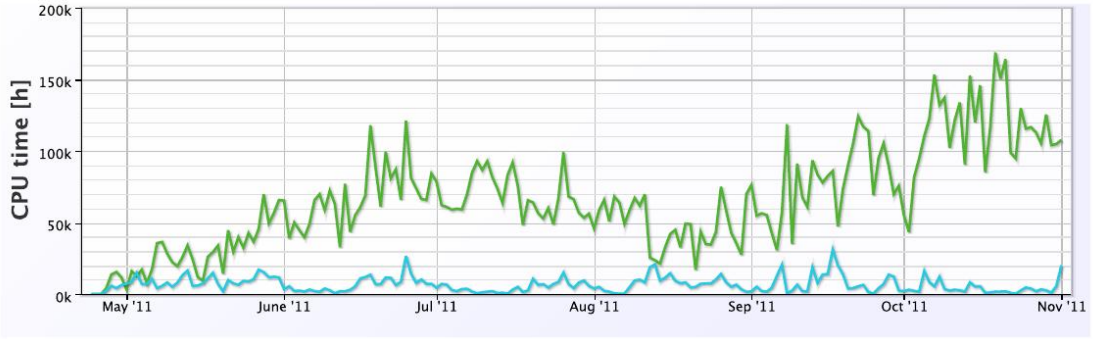
StartDate: 2012-4-3 EndDate: 2012-4-10 Invia richiesta

Select Site: all

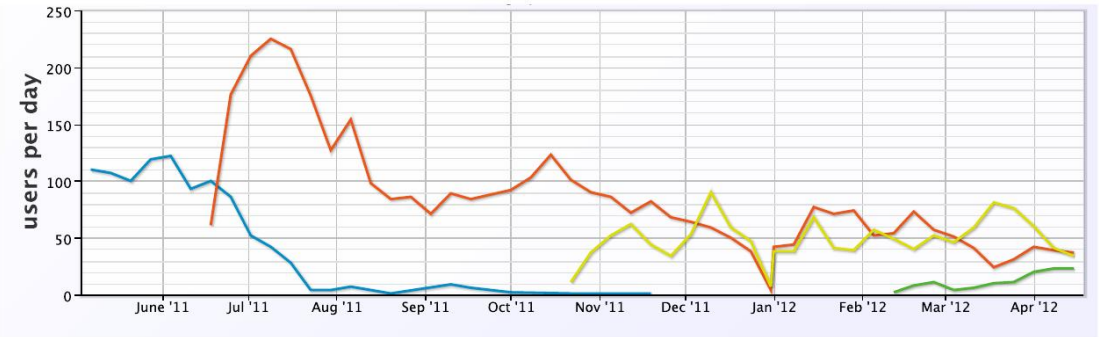
Show 10 entries Search:

DataSet	Accesses		CPU Time		Users*day	
	[N]	[%]	[h]	[%]	[N]	[%]
/DYJetsToLL_TuneZ2_M-50_7TeV-madgraph-tauola/Fall11-PU_S6_START44_V9B-v1/AODSIM	48590	1.8	122395	7.6	44	1.0
/WJetsToLNu_TuneZ2_7TeV-madgraph-tauola/Fall11-PU_S6_START42_V14B-v1/AODSIM	52183	2.0	101991	6.3	61	1.3
/SingleMu/Run2011B-PromptReco-v1/AOD	25513	1.0	97405	6.0	57	1.2
/TTJets_TuneZ2_7TeV-madgraph-tauola/Fall11-PU_S6_START42_V14B-v2/AODSIM	67666	2.5	89911	5.6	55	1.2
/RelValProdTTbar/JobRobot-MC_42_V12_JobRobot-v1/GEN-SIM-RECO	732753	27.4	64772	4.0	8	0.2
/WJetsToLNu_TuneZ2_7TeV-madgraph-tauola	25650	1.0	57038	3.5	30	0.7

(a) Tabella che mostra i dataset più popolari con relativi valori di uso. Il gruppo di Tiers e la finestra temporale sulla quale effettuare la domanda al database sono modificabili dall'utente.



(b) Tempo di utilizzo della CPU speso nell'analisi su due differenti datatier: AOD (verde) e RECO (blu).



(c) Numero di utenti per giorno per giorno in funzione del tempo su quuattro differenti collezioni.

Figura 3.4: Popularity Service web frontend. Esempi di tabelle e grafici ottenuti su diversi livelli di aggragazione.

3.4 Victor Site Cleaning Agent

Victor è un servizio, popularity-based, implementato sull'uso dei dati raccolti dal CMS Popularity Service con lo scopo di esaminare i Tier-2 che hanno raggiunto la loro quota massima di storage e individuare i dati vecchi e meno popolari che possono essere rimossi in modo sicuro senza avere un impatto negativo sull'attività di analisi.

3.4.1 Workflow di Victor

Victor è un'applicazione che viene lanciata ogni giorno su macchine dedicate per trovare repliche non popolari per i gruppi che hanno raggiunto la loro quota massima di storage in un determinato sito.
Il workflow di Victor è mostrato in figura 3.5. Il primo step è di identificare i gruppi di analisi sui cui dati operare. Le informazioni riguardo lo spazio usato provengono da PhEDEx. Le soglie sono configurabili, ma come default un sito viene considerato pieno se esso ha a disposizione ("free space") meno del 10% o alternativamente 15 TiB di *free space*. Nel secondo step Victor prova a selezionare le repliche in modo da ridurre lo spazio fino al 30% (alternativamente 25 TiB) dello spazio libero per poi fermarsi. Nel selezionare le repliche non popolari Victor legge i dati relativi alle repliche possedute dai gruppi in ciascun sito dal CMS Popularity Service e da PhEDEx, ed esegue gli step seguenti:

1. ordina i blocchi in ordine cronologico rispetto all'ultimo accesso;
2. filtra le repliche da cancellare applicando una serie di condizioni:
   (a) i blocchi devono essere più vecchi di 90 giorni (poichè si vuole evitare la cancellazione di dati recenti che non hanno ancora avuto l'opportunità di essere acceduti);
   (b) i blocchi non devono avere la flag *custodial* e devono avere una copia *custodial* in un Tier-1 (in modo da evitare di cancellare l'unica copia dei dati presente nell'infrastruttura);
   (c) (FIX ME: PARLARNE DI QUESTO PUNTO) il numero degli accessi ai file negli ultimi 30 giorni deve essere minore di `{1, 10, 100}`, dove la soglia degli accessi viene incrementata con questi step fino a raggiungere la quota di spazio libero richiesto.

3.4.2 Web interface di Victor

Tutte le informazioni riguardardanti i possibili target della pulizia sono salvate in un database Oracle, al quale è possibile accedere per visualizzare i dati, attraverso una web interface. Questa web page presenta un sistema di storage accounting (Figura 3.6a) e fornisce un interfaccia per i gruppi di fisica e gli amministratori dei Tier che possono selezionare le repliche suggerite (Figura 3.6b) per poi inoltrare a PhEDEx la richiesta della loro eliminazione. La web interface fornisce una interfaccia da cui visualizzare una serie di grafici che mostrano lo stato dello spazio

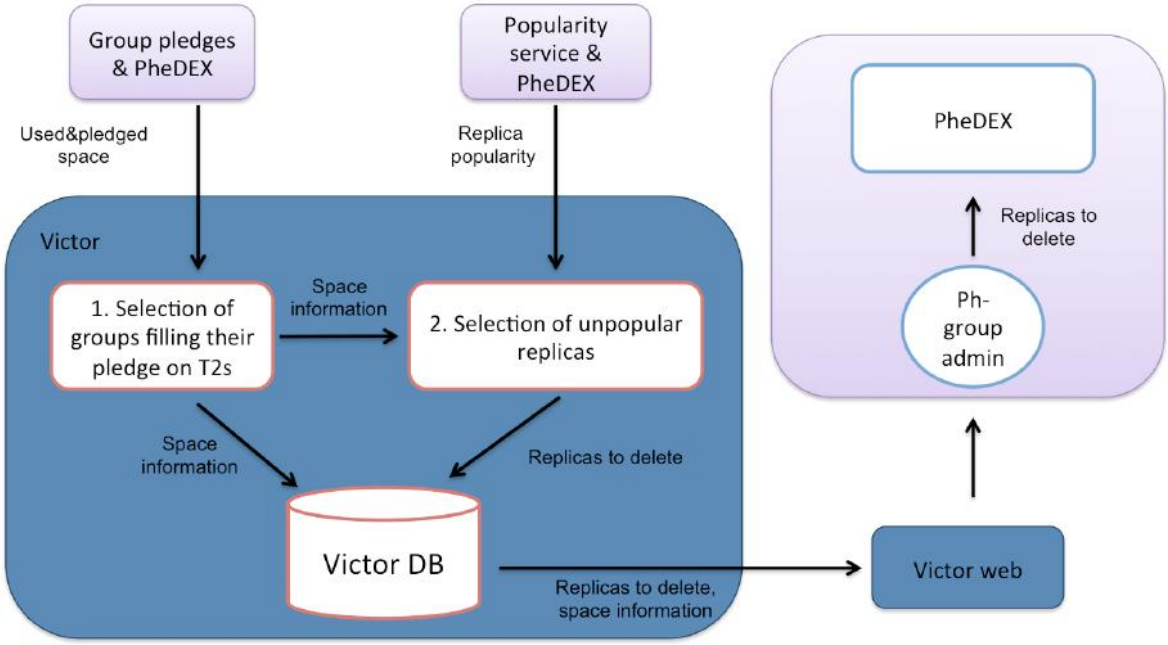
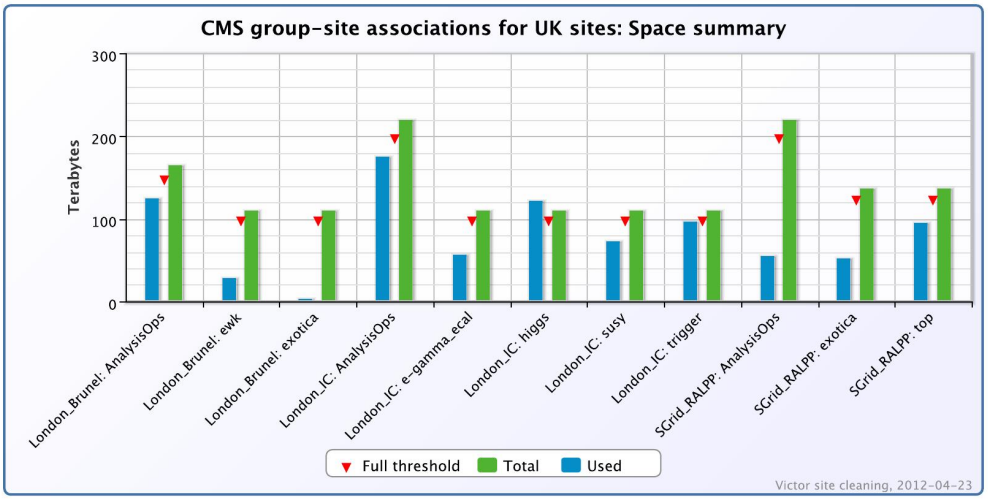


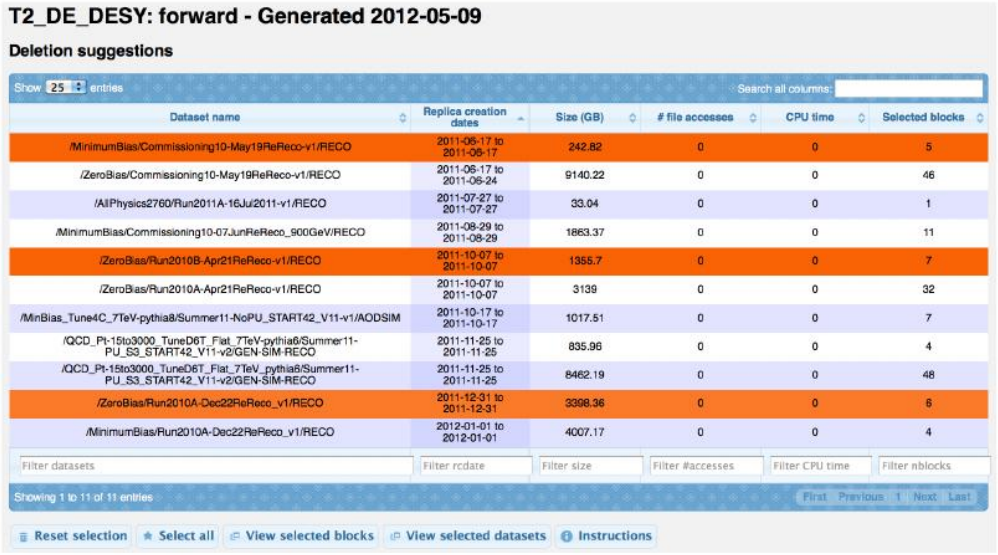
Figura 3.5: Victor Site cleaning workflow [49].

usato sullo spazio totale per ogni associazione sito-gruppo di fisica ad esso connesso, oppure l'evoluzione dello spazio occupato da ciascun gruppo nel totale dei siti, così come su un singolo sito (Figura 3.6c).

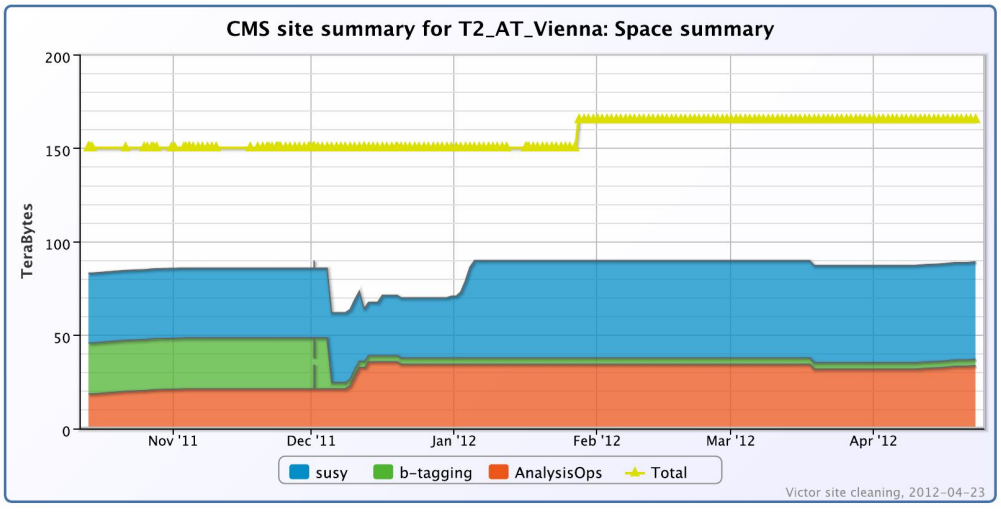
La quantità di dati suggeriti da Victor per l'eliminazione dipende da diversi fattori come lo stato dell'attività dell'analisi in ciascun gruppo di fisica, la prossimità a conferenze importanti o alla eventuale disponibilità di nuove e recenti versioni di dati ripocessati che rendono obsoleti dati di versione precedente (che possono dunque essere considerati per la cancellazione). Come stima dell'utilità dell'introduzione del servizio del CMS Popularity Service connesso all'uso che ne fa Victor, si consideri che solo nei primi giorni di esecuzione Victor ha identificato circa 1.5 PiB di dati potenzialmente obsoleti alla fine del Run-1 che sarebbero potuti essere cancellati dai Tier-2.



(a) Utilizza dello spazio disco per i Tier-2 inglesi. Sono raffigurati, per ogni associazione sito-gruppo, lo spazio allocato (verde), lo spazio utilizzato (blu) e lo spazio limite (rosso) raggiunto il quale Victor inizia il processo di pulizia.



(b) Tabella che mostra i suggerimenti di Victor per l'eliminazione delle repliche non popolari, dataset per dataset.



(c) Evoluzione nel tempo dello spazio usato dai gruppi su un dato Tier-2 (esempio del Tier-2 austriaco).

Figura 3.6: Victor web interface

Capitolo 4

Costruzione dell'infrastruttura per studi di data popularity

4.1 Obiettivo

Gli studi di data popularity all'interno del sistema di calcolo distribuito di CMS, presentati in questa tesi, si pongono all'interno del processo di evoluzione che il CMS Computing System sta effettuando sul lungo termine, in preparazione ai prossimi Run di LHC. Un dynamic data placement che gestisca la distribuzione dei dati in base alla loro importanza per i fisici è in corso di sviluppo in questi mesi. Sul più lungo termine l'evoluzione della *network topology* della Grid, descritta nel paragrafo 2.6, potrebbe aprire la strada a nuovi sviluppi che potenzialmente miglioreranno le prestazioni dei servizi offerti all'analisi distribuita su Grid, Il Popularity Service, descritto nel capitolo precedente, rappresenta un primo passo: è il primo servizio creato per raggiungere questo scopo e i metadati da esso raccolti forniscono la base per un ulteriore sviluppo in questa direzione.

L'obiettivo che ci si pone è la raccolta e lo studio degli schemi di accesso dei dati di CMS in relazione alla loro popolarità[1]. In CMS si introduce il concetto di popolarità dei dati ("data popularity") come un'osservabile in grado di quantificare l'interesse della comunità degli analisti per campioni di dati o simulazioni Monte Carlo specifici, sulla base del numero di accessi, locali o remoti, di successo o falliti, che i job utenti hanno effettuato sui file che storano tali dati su disco. Il prodotto di questo lavoro è una semplice ma completa infrastruttura, del tutto operativa, che accede ai dati di popolarità di CMS e può essere personalizzata per produrre indagini ad-hoc e visualizzazione dei risultati, con un duplice scopo: da un lato, offrire a esperti e operatori di CMS Computing un mezzo di indagare aspetti interessanti legati ai pattern di accesso ai dati CMS in analisi distribuita, onde esplorare possibili ottimizzazioni e automatizzazioni del sistema (dynamic data placement, utilizzo di tecniche di Big Data analytics e ricerca di correlazioni, creazione di un modello con potenzialità predittive, etc); dall'altro lato, fornire - secondo schemi concordati - grafici ufficiali che CMS può usare nelle discussioni con i vari organismi che controllano l'uso delle risorse da parte degli esperimenti

[1] La parola popolarità verrà di seguito usata con un significato collegato al termine inglese "*popular*", con l'accezione di essere molto acceduto e quindi di notevole importanza per l'analisi.

LHC (ad esempio, nelle interazioni tra il CMS Resource Management Office e il Computing Resource Scrutiny Group) [65]. Con il presente lavoro ci si focalizza sui datatiers di tipo AOD e AODSIM e solo su quelli presenti a livello dei Tier-2. Il servizio creato fornisce accesso grafici contenenti informazioni già elaborate, nonchè strumenti per effettuare ulteriori elaborazioni.

4.2 Data sources

Una parte dei dati usati nell'analisi sono metadati raccolti dal CMS Popularity Service di CMS e ottenuti mediante API. Questa informazioni vengono coplementate con altri metadati, ottenuti tramite una query al database di PhEDEx, che contengono anche le informazioni relative ai quei fileblock esistenti nel momento della query, ma ancora non acceduti. Le informazioni presenti nel CMS Popularity Service da sole non sono infatti sufficienti per effettuare statistiche che contemplino anche i dati di cui non si è ancora effettuato l'accesso nè tramite CRAB nè tramite Xrootd. I dati su cui si effettueranno gli studi sono valori relativi ad ogni fileblock presente in ogni Tier-2 di CMS al momento delle query e comprendono informazioni relative ad una finestra temporale che può variare da una settimana a un anno indietro nel tempo. Il formato di interscambio dei dati è il *JavaScript Object Notation* (JSON) [55] supportato attraverso librerie dalla quasi totalità dei linguaggi di alto livello più diffusi. Ogni file JSON è formato da un oggetto composto che racchiude al suo interno informazioni sulla stessa query effettuata e sui fileblock presenti nel Tier di cui si sta indagando, come mostrato nella mindmap di Figura 4.1. I dati relativi ad ogni fileblock, contenuti in ogni JSON relativo ad un singolo Tier-2, sono

group : il nome del gruppo di lavoro a cui "appartiene" il fileblock, ovvero il gruppo di lavoro la cui attività è più fortemente pertinente con i dati contenuti in esso;

popularitycpu : il numero di ore di lavoro delle CPU che hanno lavorato sul fileblock in loco (l'analisi segue il paradigma *data-driven*);

custodial : è un booleano che indica se il fileblock è una copia di sicurezza dei dati; normalmente sui Tier-2 non sono presenti dati *custodial*, ma solo ai T1 e al T0;

creation_time : l'istante in cui è stato creato il fileblock;

nfiles : il numero di file di cui è composto il fileblock;

popularitynacc : il numero di accessi in lettura da parte di analisti via Grid;

size : la dimensione totale del fileblock;

4.3 Architettura e Workflow

L'architettura con cui è stato progettato il sistema è stata pensata in modo modulare, così da poter essere facilmente e velocemente ampliata e da garantire

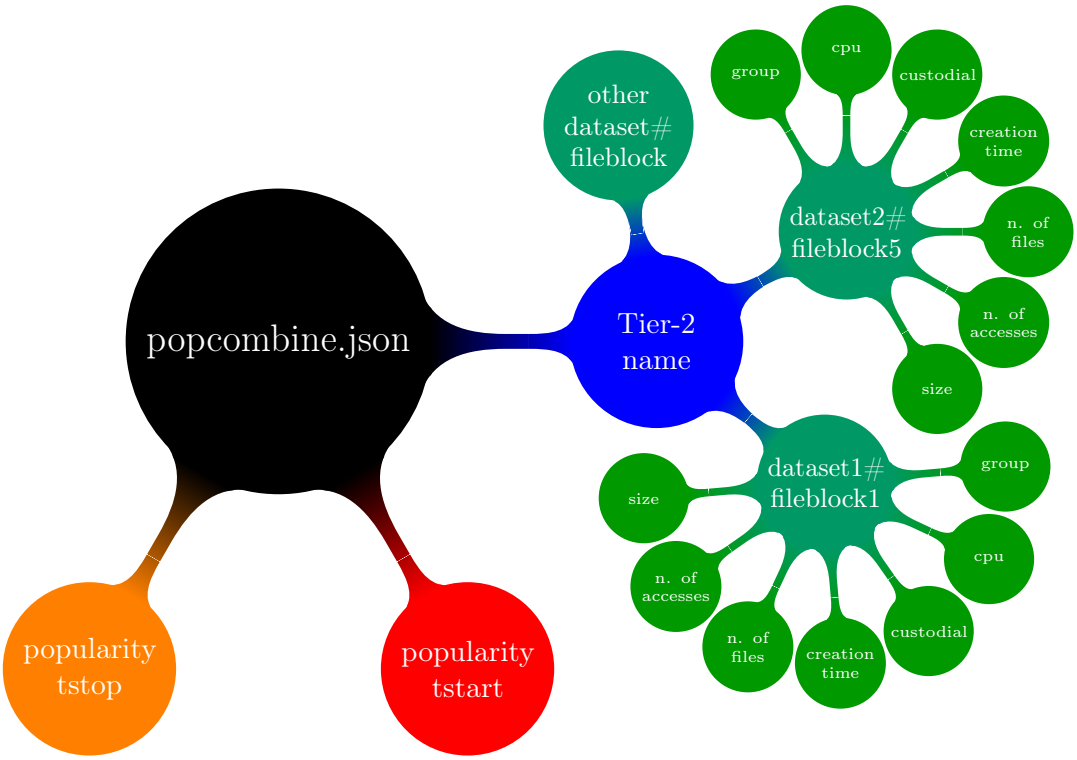


Figura 4.1: Organizzazione di un file popcombine.json

una notevole flessibilità nell'utilizzo degli strumenti creati. Tale architettura può essere divisa in tre blocchi, a ciascuno dei quali è associato uno step fondamentale nel workflow. Il primo step consiste nella richiesta dei dati al popularity DB. I dati vengono successivamente, attraverso due ulteriori step, riprocessati. Il secondo si occupa del parsing dei dati, mentre il terzo step riempie l'istogramma cercato.
Per quanto il flusso del programma sia sequenziale e passi attraverso i tre blocchi appena elencati, è tuttavia importante notare che esso non è continuo: ad ogni step l'output dello step precedente viene salvato e controllato, prima di essere iniettato nel blocco successivo. Questa implementazione presenta il vantaggio di poter sviluppare e migliorare asincronamente i vari step, e permette eventualmente l'inserimento di ulteriori moduli. Il workflow è raffigurato in Figura 4.2. Di seguito i tre step verranno descritti in maggior dettaglio.

4.3.1 Step-0

Le queries al popularity database sono state avviate a partire da Luglio 2014, con cadenza settimanale, attraverso l'API "`popdbcombine`". Essa restituisce, interrogando il CMS Popularity DB e PhEDEx, i metadati raccolti relativi ad ogni Tier-2 con le seguenti granularità temporale:

- 7 giorni;
- 1 mese;
- 3 mesi;

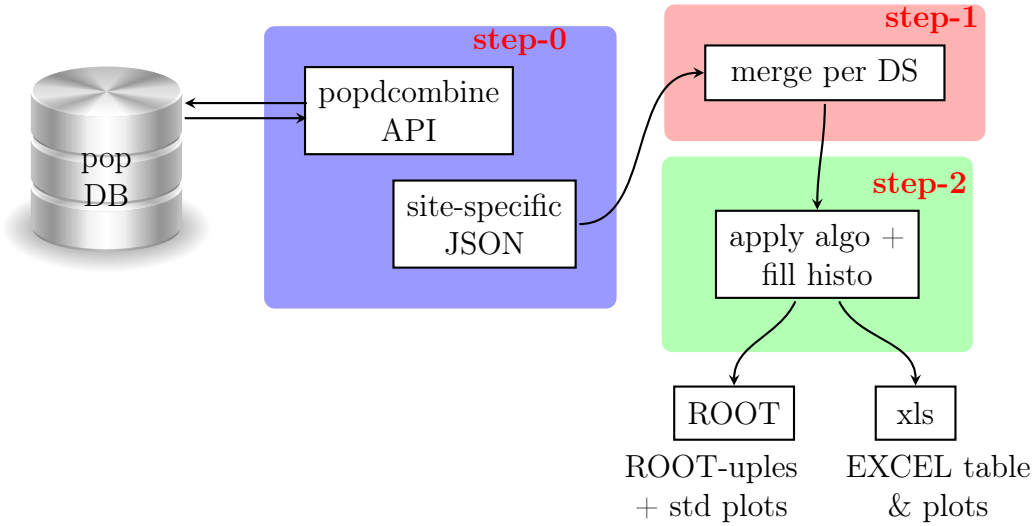


Figura 4.2: Descrizione del workflow. Vedi testo per dettagli.

- 6 mesi;
- 9 mesi;
- 12 mesi.

Le informazioni ottenute e contenute nel popularity DB nel momento in cui la query è stata effettuata vengono conservate in due copie, congelando così i metadati ottenuti ad un preciso momento temporale. Per quanto detto nel Paragrafo 4.2 i dati raccolti ogni volta che si effettua la query sono relativi soltanto ai fileblock presenti al momento della stessa e quindi non comprendono i dati che fanno sì parte della finestra temporale analizzata, ma solo in un intervallo di tempo che si conclude prima che lo Step-0 venga effettuato. Questi dati non sono in realtà importanti in uno studio che osservi i pattern di accesso di dati presenti oggi, ma è importante conoscerne la loro assenza nella statistica effettuata.

Il formato e il contenuto dei dati ottenuti è quello già presentato nel paragrafo 4.2. Nel seguito si farà riferimento all'output dello Step-1 come a "popcombine".

4.3.2 Step-1

Lo Step-1 ha come scopo quello di riorganizzare i dati aggregandoli non per sito, ma per tipologia di dati. Tutti i metadati relativi ad ogni sito vengono rielaborati per ottenere le statistiche di ogni fileblock ed ogni dataset relative sia al loro utilizzo complessivo nella Grid sia alla loro dimensione totale. Viene inoltre tenuta in considerazione la presenza di repliche dei fileblock all'interno della Grid calcolando, per i valori inerenti la dimensione, sia i valori totali sia i valori del fileblock singolo. Tra tutti i dati disponibili vengono mantenuti, grazie a un'apposita funzione di filtro, solo i datatier AOD e AODSIM. Le impostazioni con cui è possibile modificare il workflow dello Step-1 sono configurabili separatamente attraverso un file di configurazione (di cui si riporta un esempio in appendice A) che viene importato dal modulo al suo avvio. Le opzioni configurabili sono:

- la directory da cui leggere i metadati ottenuti dallo Step-0;
- le opzioni della funzione di filtro. É possibile filtrare i dati per:
  - datatiers,
  - gruppi di lavoro in CMS;
- la scelta relativa a quali dataset analizzare.

I popcombine vengono ricombinati sfruttando due istanziazioni della classe dizionario di Python [60]. La prima istanziazione, che nel seguito verrà chiamata `blockDict`, viene riempita assegnando a ogni fileblock presente al livello dei Tier-2 i valori relativi alla sua dimensione, al numero totale di ore di lavoro delle CPU, al numero di accessi effettuati e al numero di file presenti nel blocco, relativamente al periodo di tempo su cui la query è stata effettuata (indipendentemente dal sito). I valori propri di ogni fileblock vengono ottenuti eseguendo prima un ciclo su tutti i fileblock presenti in un Tier e poi un ciclo su ogni Tier, operando come segue:

- `blockDict[nomeblocco]=(size, tcpu, nacc, nfiles, sizetot, nfilestot):`
  - `size = valore_‘‘size’’_del_blocco`
  - `tcpu = sum(valori_‘‘popularitycpu’’)`
  - `nacc = sum(valori_‘‘popularitynacc’’)`
  - `nfiles = valore_‘‘nfiles’’_di_nomeblocco`
  - `sizetot = sum(valori_‘‘size’’_del_blocco)`
  - `nfilestot = sum(valore_‘‘nfiles’’_di_nomeblocco)`

dove le sommatorie sono effettuate su tutte le istanze di `nomeblocco` incontrate. La seconda istanziazione della classe dizionario, `DSDict`, viene invece riempita con i valori della dimensione e del numero di accessi totali sugli interi dataset, come segue:

- `DSDict[DSname][Tier]=(size,tcpu,nacc,nfiles,sizetot,nfilestot)`
  - `size = sum(blockDict[nomeblocco][size])`
  - `tcpu = sum(blockDict[nomeblocco][tcpu])`
  - `nacc = sum(blockDict[nomeblocco][nacc])`
  - `nfiles = sum(blockDict[nomeblocco][nfiles])`
  - `sizetot = sum(blockDict[nomeblocco][sizetot])`
  - `nfilestot = sum(blockDict[nomeblocco][nfilestot])`

dove le sommatorie sono effettuate su tutte le istanze `nomeblocco` appartenenti ad `DSname`.

L'output dello Step-1 viene anch'esso salvato in modo che l'analisi sui dati da esso ottenuti possa essere eseguita asincronamente, senza dover rilanciare tutto il

modulo. Lo Step-1, peraltro, ha il carico computazionale più elevato: per velocizzare l'elaborazione dei dati raccolti settimanalmente nel caso di un miglioramento di questo e del successivo modulo si è provveduto a creare uno script che lancia in parallelo tutti i processi. Il formato dei metadati in uscita da questo modulo sono due: il formato di interscambio JSON e un file ROOT [61] in cui i dati sono salvati all'interno di un `TTree` [62]. Lo schema dell'output JSON è mostrato in Figura 4.3.

4.3.3 Step-2

Lo Step-2 riempie i bin di un istogramma e automaticamente crea un foglio excel con tabelle e plot. L'istogramma viene creato riempiendo un'istanziazione della classe di istogrammi uno-dimensionali progettata all'interno di ROOT: `TH1F` [63], usando il metodo `Fill(x/w,w)`, dove:

- $x = (\#\text{ accesses}) \cdot (< \text{dataset file size} >) = (\#\text{ accesses}) \cdot \frac{(\text{dataset tot size})}{(\#\text{ files in dataset})}$
- $w = (\#\text{ replicas}) \cdot (\text{dataset tot size})$

Viene inoltre riempito il primo bin dell'istogramma, quello che da ora in poi chiameremo "*Bin-0*", con il volume dei dataset sui quali non sono stati effettuati accessi in lettura. É possibile, agendo sul file di configurazione dello script che esegue lo Step-2, riempire i bin dell'istogramma sia considerando le repliche di ogni dataset presenti ai T2 sia considerare i dataset singolarmente, senza repliche.

4.4 Considerazioni sulla raccolta e presentazione dei dati

Il procedimento con cui vengono riempiti gli istogrammi si basa sul fatto che ogni accesso in analisi distribuita di CMS avviene su un singolo file e la lettura di ciascun file equivale alla lettura di tutti i byte che lo compongono (ovvero di tutti gli eventi in esso contenuti). Inoltre, nel conto della dimensione totale di un daaset si può tenere conto anche delle sue repliche.
La variabile x rappresenta il numero di accessi per la dimensione media di
Per chiarire meglio il concetto supponiamo come esempio che, dato un DS composto da quattro files della dimensione di 3 GiB che abbia due repliche in tutta la grid, la sua popularity ci dica che è stato effettuato un accesso su di esso negli ultimi sette giorni (ovvero che un file è stato letto). Si riempirà l'istogramma al valore $x/w = 1 \cdot \frac{4 \cdot 3}{4} \cdot \frac{1}{2 \cdot 3 \cdot 4}$, dove $w = 24$ GiB e la dimensione media di un file è 3 GiB. Si può approssimativamente immaginare che questo dataset sia distribuito su otto file con una dimensione standard (otto poichè stiamo tenendo conto anche delle copie) e che solo uno di questi file sia stato letto, quindi si avrà un istogramma in cui 24 GiB sono acceduti 0.8 volte. Per estremizzare il ragionamento, se tutti i dati di CMS consistessero di un solo DS e tutti i bits di questo dataset fossero acceduti esattamente una volta, questo approccio realizzerebbe un plot con il solo bin $x = 1$ riempito e di altezza pari allo spazio occupato dal dataset. Poichè i DS sono più di uno, si può comunque vedere il plot finale come la somma dei plot di ogni dataset. Questo approccio ci assicura quindi, per costruzione, che, integrando

il plot, si ottiene la dimensione totale dei file presenti nei Tier-2.

Una conseguenza di questa procedura è che i valori dell'asse x non siano interi. Ad ogni modo, i valori >1 significano che il numero di totale di accessi supera il numero di file presenti nel dataset (incluse le repliche), mentre i valori <1 indicano il contrario. Un dataset molto grande (supponiamolo di 1000 file), con un singolo accesso porta a un valore di $x = 0.001$ molto piccolo. Per questa ragione si ritiene che il modo più corretto per presentare i risultati sia attraverso una scala logaritmica. Questa procedura porta ad un'ottima valutazione degli accessi relativi ai vari dataset, non fornisce invece informazioni di quale copia fisica sia stata acceduta, nè quale file in un dataset. Quindi in un dataset formato da due file il valore $x/w = 1$ può derivare da situazioni che non possono essere distinte, per esempio:

- un file letto due volte, l'altro letto zero volte;
- entrambi i file acceduti una volta.

La situazione di indistinguibilità si ripresenta anche nel caso siano presenti repliche. Se abbiamo un dataset con due repliche e e l'istogramma è riempito a $x/w = 1$, non possiamo distinguere se una replica è stata acceduta due volte e l'altra non è mai stata acceduta o le letture siano state distribuite su entrambe le repliche. L'effetto della scelta del modo di graficare i valori è di confinare il volume non acceduto nel Bin-0. Se si ragiona in termini di file fisici, tutti i file acceduti in un dato periodo di tempo dovrebbero non rientrare in questo bin. Poichè la presente analisi lavora per dataset l'altezza del Bin-0 rappresenterà il volume totale dello storage che non è mai stato acceduto, poichè i dataset meno acceduti (ma comunque acceduti) rientrerebbero, anche se ad essere letto fosse un solo file su un numero molto grande, nei bin successivi, ben distinguibili grazie alla scala logaritmica.
Come già osservato precedentemente il CMS Popularity Service, non ha modo di conoscere l'esistenza di quei DS che non sono mai stati acceduti e di conseguenza il bin-0 risulterebbe vuoto. La soluzione sfruttata per dar consistenza al bin-0 è di utilizzare le informazioni relative alla dimensione totale dello spazio occupato da tutti i file presenti nei Tier-2 di CMS, che il DB di PhEDEx conosce, e sottrarre da esse la dimensione dei dataset di cui noi sappiamo di avere qualche accesso. Ciò che rimane sono dataset che:

- esistono oggi nello storage;
- hanno zero accessi.

Non abbiamo tuttavia modo di contare i dataset che:

- sono stati aggiunti da un periodo di tempo breve (l'ultima settimana) e che non sono ancora stati acceduti per niente (rientrerebbero nel bin-0, ma non possiamo distinguerli);
- sono appena stati aggiunti (verrebbero contati allo stesso modo dei DS non acceduti da un grande periodo di tempo).

É possibile ciònonostante analizzare il bin-0 e dividere i DS ivi presenti in base alla loro data di creazione (si potrebbero creare per esempio tre categorie: DS creati nei trascorsi 3/6/9 mesi) e in questo modo visualizzare un trend per ogni diversa categoria. Una soluzione possibile potrebbe essere quella di effettuare un PhEDEx dump su base settimanale con il quale ottenere i dati necessari per effettuare studi in ogni finestra temporale che termina in una delle settimane in cui lo Step-0 è stato lanciato. In questo modo sarebbe possibile migliorare ulteriormente la completezza dei dati raccolti mediante questa architettura.

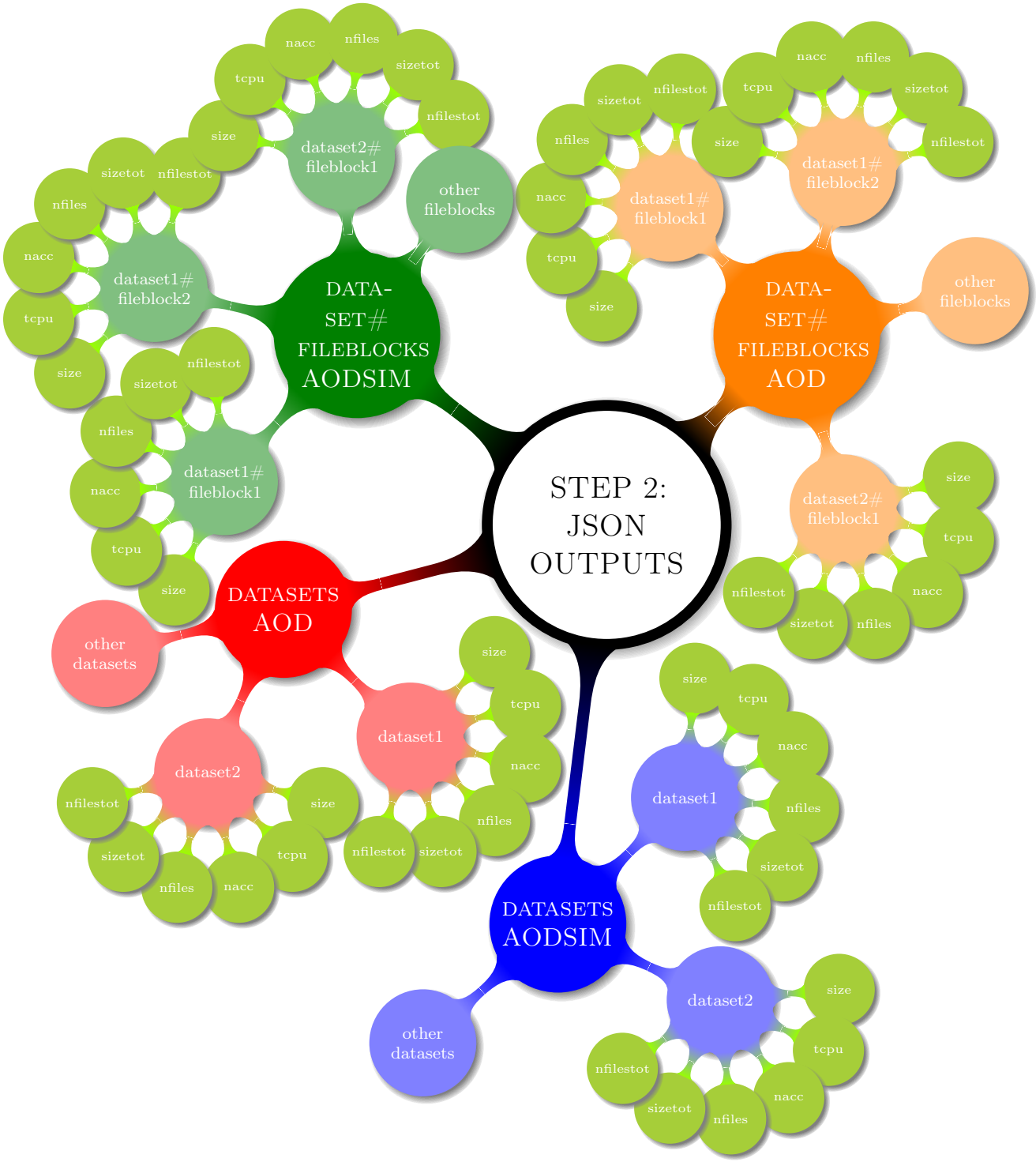


Figura 4.3: Output dello Step-2 effettuato su un unico periodo temporale. Vengono istanziate 4 classi JSON contenente ognuna un ulteriore classe, rispettivamente: classi dataset e classi fileblock. Ognuna di queste classi è descritta attraverso i suoi attributi, riempiti tramite il parsing dei dati proveniente dai popcombine.

Capitolo 5

Discussione dei risultati degli studi di data popularity

5.1 Introduzione

Alla luce della grande quantità e della varietà dei dati di popolarità collezionati, della continuità nella raccolta degli stessi nel tempo, del dettaglio delle informazioni (ad esempio sito per sito) e della granularità delle finestre temporali di indagine, è pensabile affrontare studi anche piuttosto complessi sui pattern di accesso degli analisti CMS ai dati AOD e AODSIM sui Tier-2 dell'esperimento. Inoltre, tali studio potrebbero essere estesi considerando l'utilizzo di tecniche di Big Data analytics per studiare, ad esempio, possibili correlazioni con eventi esterni, quali boost nelle attività di analisi per prossimità a conferenze, correlazioni con problemi tecnici in alcuni siti o regioni, cambiamento di tali pattern a seconda di modifiche nell'ambiente di calcolo che circonda e supporta l'analisi distribuita, eccetera.

In questo capitolo, vengono presentati gli studi che sono stati completati nel contesto del presente lavoro di tesi, nonchè l'utilizzo e l'impatto che stanno avendo nel sistema di calcolo di CMS. Vengono inoltre presentati e discussi altri aspetti che potrebbero far parte del lavoro successivo, e di essi viene discussa la potenziale importanza.

5.2 Primo studio: popolarità globale

La raccolta settimanale dei dati di popolarità iniziata a Luglio 2014, con finestre temporali che si estendono fino a un anno indietro nel tempo, consente di scegliere un qualsiasi periodo di interesse su cui focalizzare l'attenzione. Come esempio di questo approccio, si è scelto di considerare i dati raccolti la prima volta che si è lanciato lo Step-0 (vedi Paragrafo 4.3.1) nel Gennaio 2015, e di scegliere un anno intero come finestra temporale all'indietro: in tal modo si possono sostanzialmente analizzare i pattern di accesso degli analisti CMS a datatier AOD e AODSIM, esistenti nel momento in cui è stata effettuata la richiesta, approssimativamente in tutto l'anno solare 2014. Nel riempimento degli istogrammi si considerano anche repliche multiple degli stessi dataset di CMS: si ritiene che questi dati rappresentino i valori "veri" relativi all'utilizzo dello storage su T2 poichè effettivamente le repliche

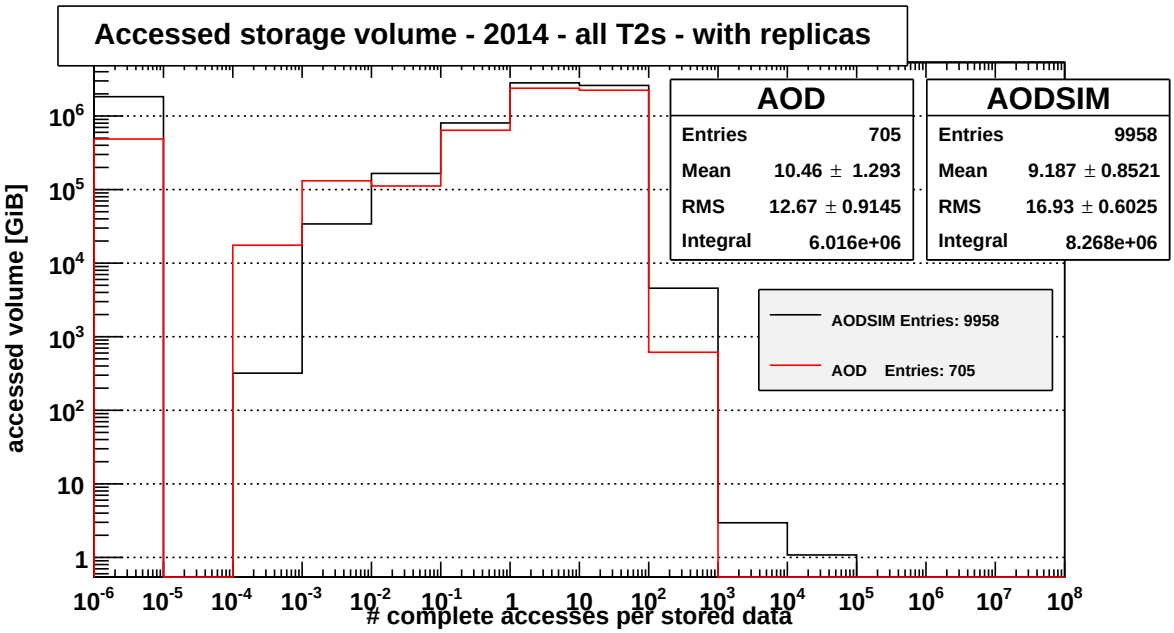


Figura 5.1: Plot del patter degli accessi ai dataset di tipo AOD (rosso) e AODSIM (nero) presenti nei Tier-2 di CMS il 04-01-2015. Le statistiche fornite dall'istogramma e dai suoi valori principali, mostrati nella tabella, sono inerenti ad una finestra temporale di 12 mesi e quindi a tutto l'anno 2014.

dei dataset sono presenti e sono parte integrante e importante del meccanismo di distribuzione e accesso ai dati nella Grid.

I dati di popolarità negli accessi a AOD e AODSIM sui T2 di CMS nel corso del 2014, mostrati nell'istogramma di Figura 5.1, sono tabulati nella relativa Tabella 5.1. Il numero totale di dataset AOD e AODSIM che sono residenti - o per tutta la durata dell'anno o almeno per una parte di esso (per il motivo spiegato nei Paragrafi 4.2 e 4.3.1) - su almeno uno dei T2 di CMS sono quasi 11000, di cui il 93% circa è rappresentato da AODSIM e solo il restante 7% da AOD. Lo spazio da essi occupato su disco ai T2 di CMS risulta di 7.9 PiB e di 5.7 PiB rispettivamente per AODSIM e AOD che corrispono a circa il 58% (AODSIM) e al 42% (AOD) del volume totale.
Il volume complessivo mostra che i dati messi a disposizione degli analisti sono circa 13.6 Pebibyte, ma di questa quantità solo ~ 11.4 PiB (84% del totale) sono effettivamente acceduti in analisi e i restanti ~ 2.2 PiB non sono invece mai stati acceduti. Risulta dunque pari al 16% lo spazio totale occupato da CMS ai T2 nel 2014 con datatier AOD e AODSIM che ha occupato lo storage senza che su di esso avvenissero accessi in lettura effettuati da programmi di analisi. Tale quantità non è trascurabile, ed è peraltro abbastanza coerente con quanto misurato da altri esperimenti a LHC: essa indica la necessità di esplorare modi per ottimizzare l'occupazione dello storage, in uso per le attività di analisi, da parte di tutta la comunità LHC, non solo in CMS, e verrà discusso in dettaglio nel Paragrafo 5.4.

In tabella 5.1 vengono anche mostrate ("numero medio di accessi"), le medie dei numeri di accessi effettuati su AOD e AODSIM in tutti i Tier-2 di CMS: come si può vedere, ogni bit che CMS ha scritto per gli analisti sullo storage ai T2 nel

Tabella 5.1: Valori più importanti relativi all'istogramma di figura 5.1.

Accessed data storage volume			
2014 - all T2s	**AOD**	**AODSIM**	**sum**
n. dataset:	705	9958	10 663
Bin-0 volume (GiB)**:**	486 888	1 831 380	2 318 268
other bins volume (GiB)**:**	5 528 615	6 436 473	11 965 089
total volume (GiB)**:**	6 015 504	8 267 853	14 283 357
numero medio di accessi:	10.5	9.2	
RMS:	12.7	16.9	

2014, è stato acceduto in media 10.5 volte nel caso di AOD (e 9.2 volte nel caso di AODSIM). Nel relativo istogramma, raffigurato in Figura 5.1, si può osservare la distribuzione degli accessi per unità di informazione, distinti per tipo di datatier su cui si è effettuato l'accesso: AOD (in rosso) e AODSIM (in nero). Come già osservato più dettagliatamente nel Paragrafo 4.4, l'ascissa del grafico rappresenta il numero di accessi "completi" su un byte di un DS normalizzato al numero totale di bytes di quel dataset, ovvero mostra quanto sia stato acceduto un determinato dataset pesando il numero di accessi con la dimensione del dataset. Sull'ordinata del grafico viene mostrata invece la dimensione totale del dataset in lettura comprese le eventuali repliche. Si ottiene così, integrando su tutto l'istogramma, il volume totale occupato AOD e AODSIM nei Tier-2 di CMS. Si noti che il plot è su scala logaritmica, il che permette di mettere in evidenza il volume dei dati contenuti nel Bin-0 (vedi Paragrafo 4.3.3).

In figura 5.2 è mostrata l'evoluzione nel tempo del numero medio di accessi completi per byte su storage per AOD e AODSIM rispettivamente.

Verranno di seguito illustrati alcuni comportamenti dei datatier spiegabili attraverso un approccio di analisi di data popularity come questo.
Osservando la Figura 5.1 si nota che il Bin-0 assume una dimensione maggiore nel caso in cui si prenda in considerazione il datatier AODSIM. Dall'altra parte, osservando quale tra AOD e AODSIM sia il datatier più acceduto si vede che il valore medio degli accessi è più alto nel caso degli AOD, nonostante il volume dei bin che contengono datatier acceduti almeno una volta sia maggiore per gli AODSIM (come si vede dalla tabella 5.1). Eseguendo lo stesso tipo di osservazione sui dati presi in un diverso momento temporale o su periodo di ampiezza temporale diversa si constata che quanto appena notato continua ad essere vero se vengono considerate finestre temporali maggiori di tre mesi. Ciò è dovuto alla differente tipologia dei datatier considerati. Gli AODSIM sono dati provenienti dalle simulazioni Monte Carlo; nel loro utilizzo nell'analisi solo pochi fileblock tra i tanti ricreati attraverso le ricostruction (vedi Paragrafo 2.2), di cui tra l'altro si succedono nel tempo varie versioni, vengono acceduti molto. Ne consegue che gli AODSIM "vecchi" finiscono più facilmente nel Bin-0 rispetto ai datatier AOD, infatti mentre l'analisi eseguita sugli AOD non può fare una selezione dei dati da analizzare, ma deve processare tutti i file interessati, l'analisi eseguita sugli AODSIM non ha bisogno di tutte le simulazioni prodotte, ma si focalizzerà su alcune di esse che riterrà migliori,

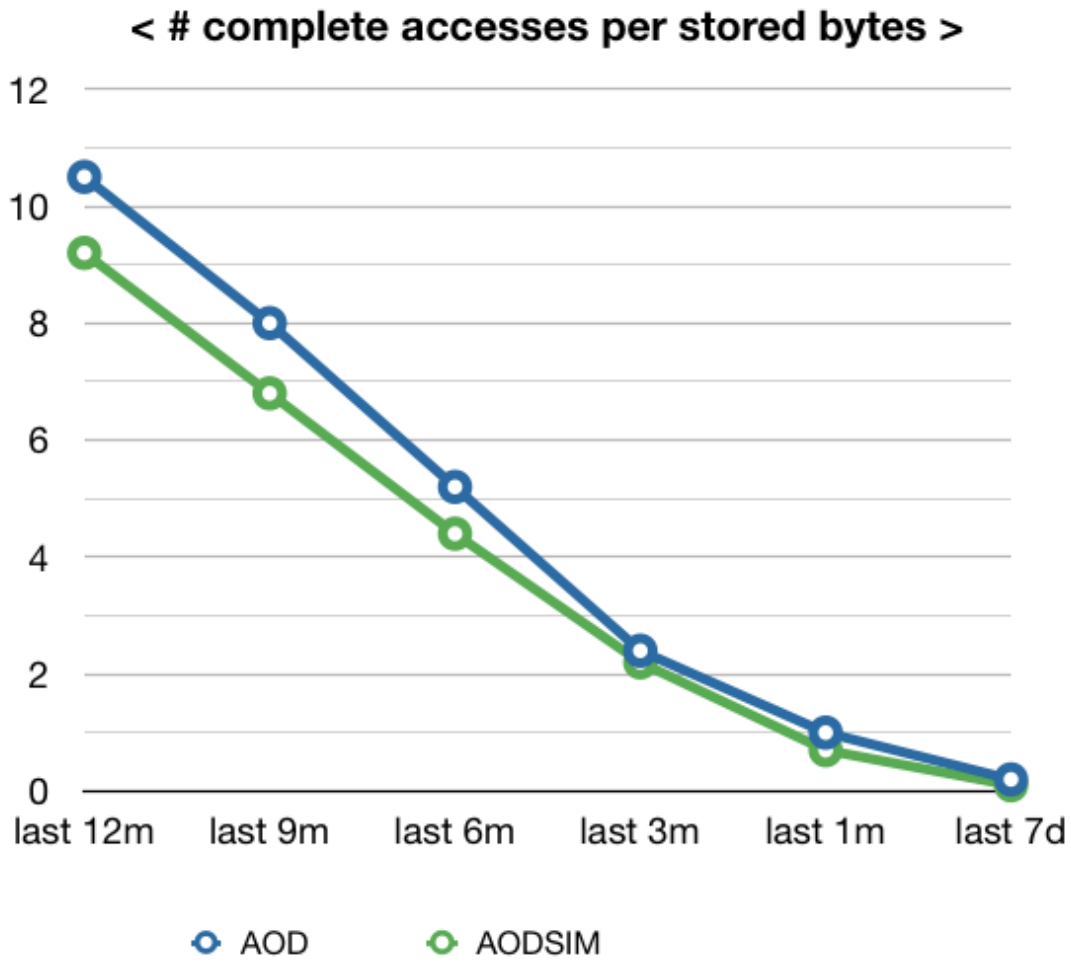


Figura 5.2: Evoluzione nel tempo del numero medio di accessi completi per byte su storage per AOD e AODSIM.

concentrandosi a migliorare l'analisi stessa che verrà poi eseguita sui dati "reali" e quindi sugli AOD. Il volume degli AODSIM appartenenti al Bin-0, di un ordine di grandezza maggiore rispetto a quello degli AOD, abbassa quindi la media del numero di accessi per byte storato più di quanto il volume dei fileblock AODSIM più popolari riesca ad alzarla. Risulta quindi che la media degli accessi agli AODSIM è sempre minore di quella degli AOD.

Si procede ora ad un analisi dell'evoluzione del Bin-0 in funzione della finestra temporale, indietro nel tempo, nella quale si analizzano gli accessi. In Figura 5.3 sono rappresentate le evoluzioni temporali del Bin-0 e dei bin contenenti dataset con almeno un accesso sia in valore percentuale rispetto al volume totale (Grafici A e B), sia in volume assoluto (Grafici C e D). Si nota che la percentuale del volume dei dataset con zero accessi sul volume totale aumenta quando si vanno ad analizzare gli accessi su un periodo di durata minore. In effetti il numero di dataset acceduti in una finestra temporale all'indietro più piccola rispetto ad un altra, saranno minori. Ci si aspetta una funzione monotona quale è quella raffigurata nel grafico. Si nota che il volume del Bin-0 passa da un 81% (60%) del totale degli AODSIM (AOD), osservando i dati nella settimana precedente alla query, al circa 20% (9%) se si va indietro di un anno. Il tempo in cui i dataset divengono popolari è quindi un aspetto importante da considerare che cambia notevolmente il contenuto di Bin-0. Più il Data Placement avrà la capacità di rispondere a una sottoscrizione di un determinato dataset più potrà considerare come Bin-0 il volume dei dati non acceduti in finestre temporali all'indietro minori. Nei grafici C e D dove è mostrato il volume e non la frazione di accessi si nota che il volume del Bin-0 degli AODSIM è sempre maggiore del volume degli AOD. Guardando invece il volume acceduto nel grafico 5.3 gli AODSIM dominano solo su larghe finestre temporali. Se le finestre temporali sono invece minori e si studia il recente passato l'andamento probabilmente dipende dal periodo in cui la query dello Step-0 è stata effettuata. In questa query dominano gli AOD. Sarà interessante estendere in modo continuativo questo studio durante Run-2.

5.3 Secondo studio: popolarità regionale

5.3.1 Panoramica generale

Nel Paragrafo precedente si sono fatte considerazioni sull'intero livello Tier-2 di CMS e ci si è basati su medie globali per l'esperimento, con la sola eccezione della distinzione di AOD e AODSIM. Alla luce di quanto evidenziato, si è ritenuto tuttavia interessante procedere a indagare in maggior dettaglio alcune specificità regionali, aumentando la granularità dello studio. In questo Paragrafo verrà infatti studiato il comportamento dei T2 di diverse nazioni che collaborano all'esperimento, al fine di valutarne la prestazione in termini di efficienza per l'analisi, nelle loro strategie di occupazione dello storage con dati di utilità alla fisica. Le metriche precedentemente mostrate in tabella 5.1 per tutti i Tier-2, sono analogamente mostrate nella tabella 5.3 per le singole regioni. In particolare, data la forte componente statunitense nell'esperimento CMS (8 Tier-2 di grosse dimensioni), si è ritenuto opportuno mostrare non solo i dati relativi agli accessi nei Tier-2 in USA,

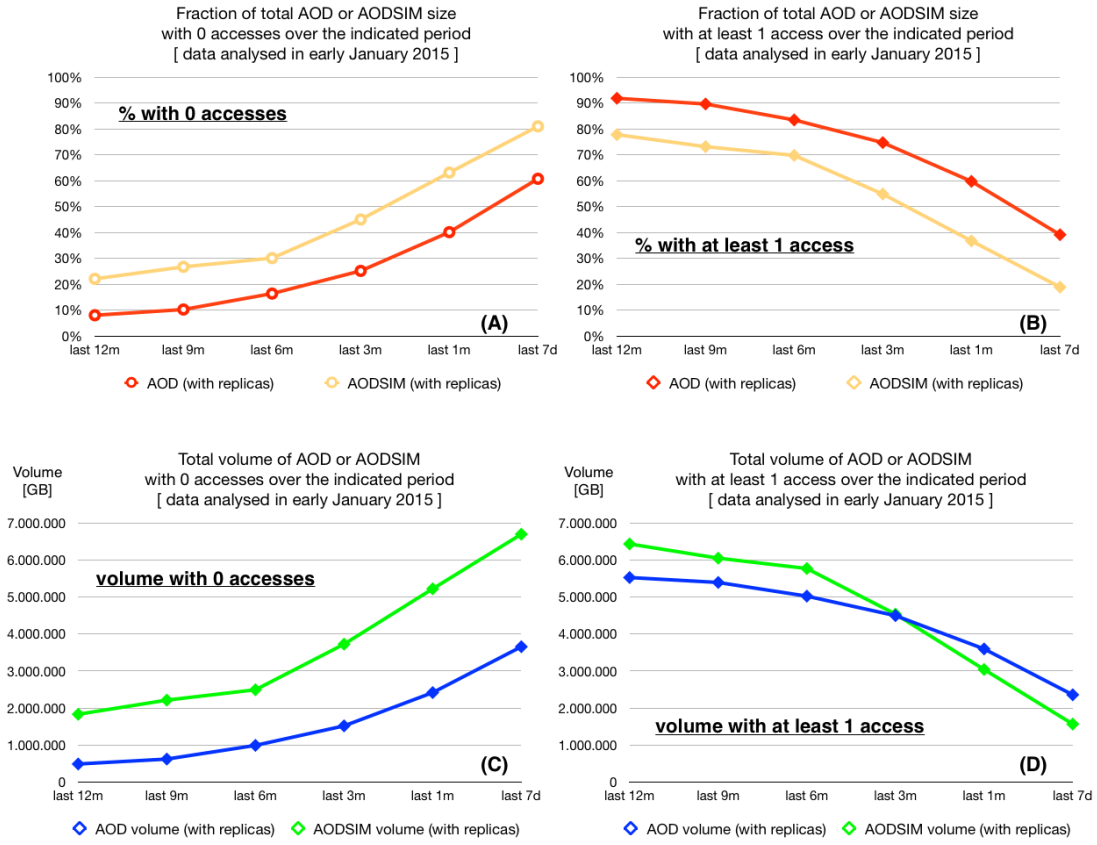


Figura 5.3: Evoluzione nel tempo del Bin-0 e dei restanti bin sia in percentuale sul totale sia in valore assoluto.

ma anche i dati relativi all'intero livello dei Tier-2 di CMS esclusi gli USA, in modo aggragato, così da valutare meglio l'impatto della regione statunitense sull'intero esperimento. Consideriamo in questo studio lo stesso campione di dataset sul quale si è effettuato lo studio nel Paragrafo precedente, ovvero i dataset presenti sui T2 la prima volta che si è lanciato lo Step-0 nel mese di Gennaio 2015 relativamente a una finestra temporale di un anno indietro nel tempo, considerando le repliche. In presenza di repliche dei dati, non è indifferente che uno o più dataset che esistano e vengano acceduti su un Tier-2 A in una data regione esistano anche (come repliche, appunto) e vengano analogamente acceduti su un Tier-2 B in una regione differente di WLCG. Questo implica che considerazioni fatte su valori di una determinata regione in relazione al totale siano correlate tra loro in funzione dei dataset condivisi tra tali regioni. Nella misura in cui questa correlazione entra in gioco, diventa complesso fare confronti tra le regioni rispetto al totale. Un esempio evidente si ha osservando la tabella 5.3: la somma del numero di AOD conservati negli Stati Uniti e del numero di AOD conservati nel resto del mondo dà come risultato un valore che supera il numero totale di AOD presenti in CMS, pari a 85 dataset; quindi 85 è il numero di AOD presenti sia negli Stati Uniti che nel resto di CMS e quindi probabilmente molto acceduti. Seguendo lo stesso ragionamento è possibile inoltre affermare che sono 51 gli AOD che a livello dei Tier-2 sono presenti solo negli USA.

Tabella 5.2: Percentuale del volume del Bin-0 di ogni regione rispetto al volume del Bin-0 totale in CMS (prime due colonne), e percentuale del volume del Bin-0 di ogni regione rispetto al volume totale in quella regione

	% Bin-0 vol. over tot Bin-0 vol		% Bin-0 vol. over total vol.	
	AOD	AODSIM	AOD	AODSIM
ALL	100.0%	100.0%	8.09%	22.15%
ALLnoUS	96.1%	80.2%	9.68%	22.59%
US	9.6%	25.9%	3.96%	26.84%
RU	29.6%	4.3%	34.04%	41.67%
DE	15.0%	13.3%	8.87%	16.33%
UK	24.1%	15.0%	14.21%	36.43%
FR	12.9%	8.5%	19.37%	28.52%
ES	15.0%	7.7%	17.13%	22.72%
IT	6.6%	12.6%	8.26%	30.48%

5.3.2 Analisi delle repliche

La discussione finora effettuata si basa sui dati di popolarità raccolti considerando anche le repliche dei dati ai siti T2. Nelle tabelle e nei grafici precedenti ogni dataset che compare in più siti sottoforma di replica contribuisce allo studio con una size maggiore. Questo rappresenta il modo migliore di presentare il pattern di accesso ai dati, poichè le repliche vengono effettuate precisamente per mettere a disposizione degli analisti una maggiore scelta dei siti su cui leggere i dati in input ai job. Qualora un dataset risulti acceduto, tuttavia, lo sarà solo in una delle repliche esistenti, rendendo dunque inevitabilmente non acceduta (almeno da quell'utente) le altre, pur presenti simultaneamente nello storage di un T2. Qualora un dataset con N repliche venga acceduto in una delle sue copie, le altre risultano non accedute ma è possibile che sia tatticamente corretto lasciare che tali repliche esistano; tuttavia, nel caso i dataset trasferiti in più repliche su più siti non risultino acceduti per lunghi periodi, la presenza di tali repliche implica un fattore moltiplicativo in più allo spazio disco sprecato ai fini dell'analisi.

É dunque interessante studiare le stesse distribuzioni di accesso ai dati senza considerare le repliche dei dataset: il confronto diretto tra lo studio precedente e quello in discussione in questa sezione permette peraltro di verificare in modo diretto il numero di repliche di dati AOD e AODSIM presenti sui siti WLCG in uso da CMS. In tabella 5.4 sono mostrati i valori riportati negli studi precedenti, senza questa volta considerare le repliche nel riempimento dell'istogramma rappresentante i pattern di accesso ai dataset per unità di informazione. Valutando questi dati in correlazione con quelli della tabella 5.3 si ottengono diverse statistiche interessanti al nostro scopo: il confronto può essere anche osservato visivamente sui rispettivi plot, mostrati in Figura 5.4.

Consideriamo inizialmente solo i dati ottenuti relativamente a tutti i Tier-2 di CMS. La prima osservazione che si può fare è valutare il numero medio di repliche presenti in tutto CMS su tutti i dataset sul livello di aggregazione di AOD e AODSIM. Il numero di repliche può essere ottenuto attraverso due procedimenti:

Tabella 5.3: Sono rappresentati di seguito i dati relativi ai plot dei patter degli accessi ai dataset di tipo AOD e AODSIM esistenti al livello dei Tier-2 di CMS il 04-01-2015 aggregati per stato (ALL: tutti i Tier-2, ALLnoUS: tutti i Tier-2 eccetto quelli presenti negli Stati uniti; US, RU, DE, UK, FR, ES, IT: rispettivamente i Tier-2 presenti negli Stati Uniti, in Russia, in Germania, nel Regno Unito, in Francia, in Spagna e in Italia). Le statistiche fornite dall'istogramma e dai suoi valori principali, mostrati nella tabella, sono inerenti ad una finestra temporale di 12 mesi e quindi a tutto l'anno 2014 e al caso in cui si riempe l'istogramma considerando le repliche.

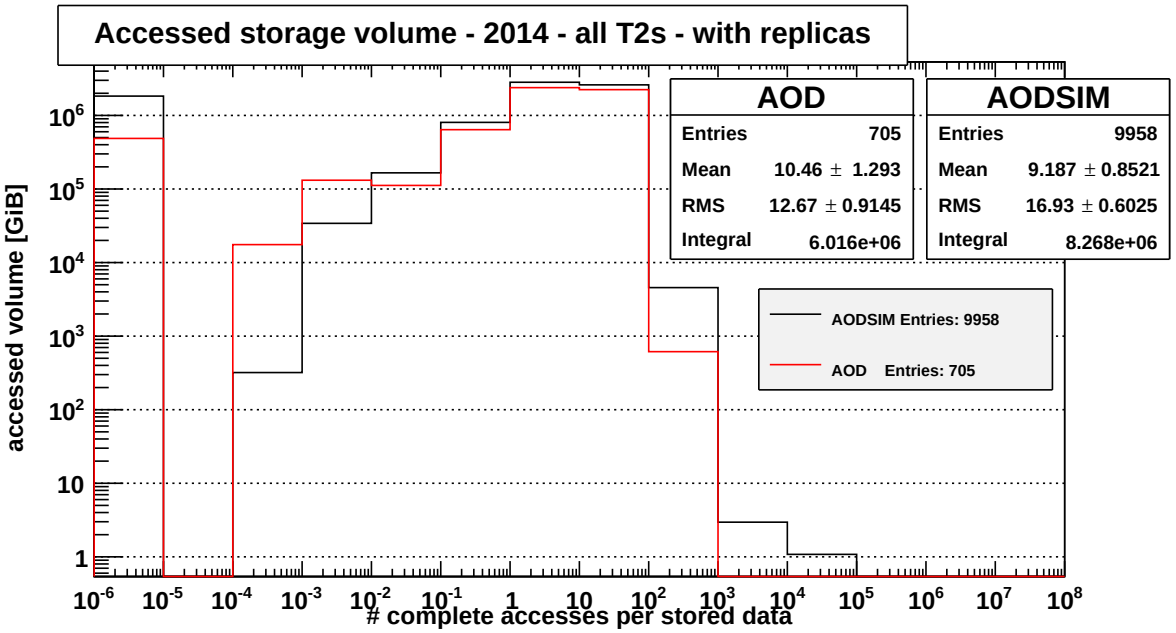
	n. DS			mean access number		RMS	
T2	**AOD**	**AODSIM**	**sum**	**AOD**	**AODSIM**	**AOD**	**AODSIM**
ALL	705	9958	10 663	10.5	9.2	12.7	16.9
ALLnoUS	654	8213	8867	7.2	8.8	9.1	17.7
US	136	2581	2717	23.7	10.5	25.7	17.7
RU	105	502	607	4.3	2.2	8.7	6.3
DE	128	1082	1210	15.2	13.4	17.0	17.6
UK	135	1867	2002	6.9	11.8	8.8	24.2
FR	52	987	1039	3.5	7.7	6.7	15.4
ES	101	780	881	8.2	7.9	10.9	10.1
IT	63	1066	1129	9.6	9.3	12.6	21.1

	Bin-0 volume			other bins volume			total volume		
T2	**AOD**	**AODSIM**	**sum**	**AOD**	**AODSIM**	**sum**	**AOD**	**AODSIM**	**sum**
ALL	486 888	1 831 380	2 318 268	5 528 615	6 436 473	11 965 089	6 015 504	8 267 853	14 283 357
ALLnoUS	468 071	1 469 221	1 937 292	4 367 621	5 034 126	9 401 747	4 835 692	6 503 347	11 339 039
US	46 748	473 599	520 347	1 133 064	1 290 907	2 423 971	1 179 812	1 764 506	2 944 318
RU	144 062	78 277	222 340	279 182	109 576	388 758	423 244	187 854	611 098
DE	72 894	243 078	315 972	749 129	1 245 819	1 994 948	822 023	1 488 897	2 310 920
UK	117 581	274 463	392 044	709 841	478 904	1 188 746	827 422	753 368	1 580 790
FR	62 976	155 075	218 051	262 121	388 694	650 816	325 097	543 769	868 866
ES	73 161	141 749	214 911	353 958	482 096	836 054	427 120	623 845	1 050 965
IT	32 305	231 517	263 822	358 802	528 127	886 929	391 107	759 644	1 150 751

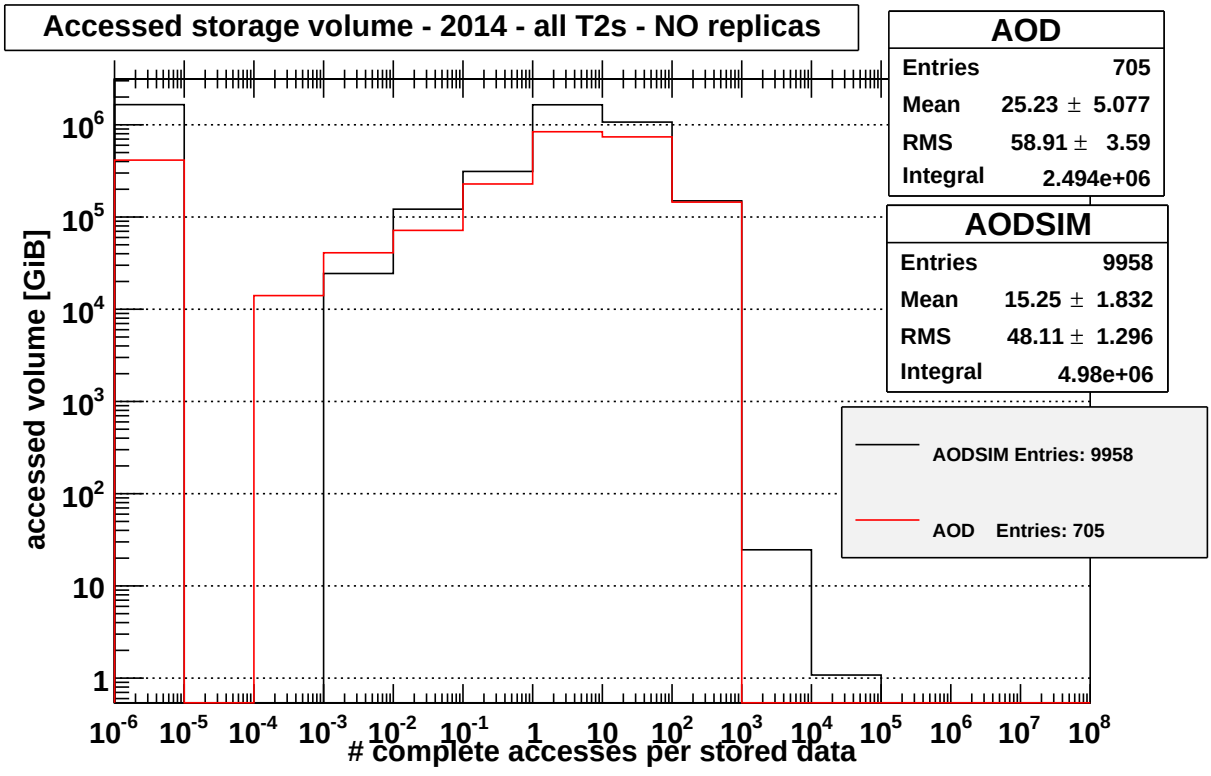
Tabella 5.4: Sono rappresentati di seguito i dati relativi ai plot dei patter degli accessi ai dataset di tipo AOD e AODSIM esistenti al livello dei Tier-2 di CMS il 04-01-2015 aggregati per stato (ALL: tutti i Tier-2, ALLnoUS: tutti i Tier-2 eccetto quelli presenti negli Stati uniti; US, RU, DE, UK, FR, ES, IT: rispettivamente i Tier-2 presenti negli Stati Uniti, in Russia, in in Germania, in Gran Bretagna, in Francia, in Spagna e in Italia). Le statistiche fornite dall'istogramma e dai suoi valori principali, mostrati nella tabella, sono inerenti ad una finestra temporale di 12 mesi e quindi a tutto l'anno 2014 e al caso in cui si riempe l'istogramma NON considerando le repliche.

	n. DS			mean access number		RMS	
T2	**AOD**	**AODSIM**	**sum**	**AOD**	**AODSIM**	**AOD**	**AODSIM**
ALL	705	9958	10 663	25.2	15.3	58.9	48.1
ALLnoUS	654	8213	8867	14.8	13.6	31.4	40.1
US	136	2581	2717	30.8	11.8	46.9	29.6
RU	105	502	607	4.3	2.2	8.7	6.3
DE	128	1082	1210	15.6	14.1	18.9	19.5
UK	135	1867	2002	7.1	12.1	9.3	24.4
FR	52	987	1039	3.6	8.3	7.1	18.6
ES	101	780	881	9.4	8.1	12.2	10.6
IT	63	1066	1129	10.4	9.6	19.0	21.8

	Bin-0 volume			other bins volume			total volume		
T2	**AOD**	**AODSIM**	**sum**	**AOD**	**AODSIM**	**sum**	**AOD**	**AODSIM**	**sum**
ALL	414 056	1 653 985	2 068 041	2 080 165	3 326 117	5 406 282	2 494 222	4 980 101	7 474 323
ALLnoUS	400 421	1 326 164	1 726 585	1 963 541	2 900 289	4 863 830	2 363 962	4 226 453	6 590 415
US	46 736	471 809	518 545	860 814	1 099 606	1 960 421	907 550	1 571 415	2 478 965
RU	144 062	78 277	222 340	278 627	109 179	387 806	422 690	187 456	610 146
DE	72 894	243 068	315 963	726 243	1 167 861	1 894 104	799 137	1 410 930	2 210 067
UK	117 581	274 458	392 039	690 120	462 115	1 152 235	807 701	736 573	1 544 275
FR	62 976	154 962	217 938	255 427	349 819	605 246	318 403	504 781	823 184
ES	73 161	141 336	214 497	298 220	465 452	763 672	371 381	606 788	978 169
IT	32 305	227 806	260 111	327 964	507 042	835 006	360 269	734 848	1 095 117



(a) con repliche



(b) senza repliche

Figura 5.4: Pattern degli accessi ai datatier AOD (rosso) e AODSIM (nero) nei Tier-2 di CMS nel 2014, (a) considerando le repliche e (b) non considerando le repliche. Vedi testo per ulteriori informazioni.

1. sfruttando il numero di accessi medio per datatier
2. sfruttando il volume totale dei due datatier (includendo e escludendo le repliche)

Entrambi i procedimenti portano allo stesso risultato, coerentemente col metodo con cui è stata implementata l'intera infrastruttura di processamento dei dati, ma contribuiscono ad evidenziare aspetti differenti nello studio. Partiamo considerando il numero medio di accessi. Si nota subito che, mentre la media degli accessi per unità di storage nella finestra temporale considerata è di 10.5 (9.2) per AOD (AODSIM) se si contano le repliche, tali medie divengono 25.2 (15.3) se non si contano le repliche. Il numero di accessi aumenta poichè tutti gli accessi su un dataset si concentrano sulla sua dimensione singola. Minore è la discrepanza tra i due valori e più il numero di repliche si avvicina a una replica soltanto. Qualora il numero di accessi medio (per unità di informazione), considerando l'istogramma riempito contando le repliche, fosse uguale al numero di accessi medio dell'istogramma riempito non considerando le repliche il numero di repliche medio sarebbe esattamente pari a 1. Allo stesso risultato si giunge considerando invece i volumi totali di dati, suddivisi per datatier: il numero di repliche medio in questo caso è tanto maggiore quanto maggiore è il volume totale occupato da quel datatier. La media delle repliche per AOD (AODSIM) sullo storage dei T2 di CMS risulta di 2.41 (1.66). Tale valore è stato confermato come corretto dai responsabili del CMS Computing Data Management development team [64] e dalle statistiche interne al TMDB di PhEDEx. La verifica della correttezza di cifre globali di cui questa è solo un esempio aumenta il livello di confidenza sull'affidabilità dello strumento che è stato realizzato nell'ambito di questa tesi e fornisce il presupposto per approfondire l'osservazione del numero di repliche in correlazione con il Bin-0. Ciò può essere fatto sfruttando i dati trovati relativamente al volume del Bin-0 e al volume dei restanti bin. É interessante vedere il Bin-0 in termini di numero di repliche e ottenere un confronto con i restanti bin, poichè è significativo visualizzare il comportamento del data placement di CMS in relazione al numero di repliche di dataset allocati, correlando il dato con l'importanza che quei dataset hanno avuto per l'analisi. Potenzialmente le repliche dei dataset non acceduti affatto sono le prime da cancellare. Guardando i valori relativi al totale dei dataset presenti in tutti i Tier-2 di CMS il numero di repliche dei dataset che si trovano nel Bin-0 è pari a 1.18 (1.11) contro le 2.66 (1.94) repliche in media sui dataset che sono stati acceduti almeno una volta. Questo mostra, come ci si aspetterebbe, un numero di repliche maggiore laddove i dati sono più richiesti. É comunque rilevante la presenza di più di una replica di dataset non acceduti per niente nel corso di un intero anno.

Il numero medio di repliche pari a circa 2.4 (1.7) per AOD (AODSIM) rispettivamente è, come commentato in precedenza, una media globale su tutta CMS. A questo punto può essere interessante allontanarsi dalla media globale ed esplorare invece le differenze in questa metrica tra regione e regione. Si prenda ad esempio la regione italiana. Sui 4 T2 di CMS Italia, il numero medio di accessi è 9.6 (9.3) contando le repliche, mentre è 10.4 (9.6) non contando le repliche: si tratta di variazioni relativamente piccole, conseguenza del fatto che in generale sui T2 italiani i dati utili agli analisti è noto che vengono copiati mediamente in singola copia e senza farne un numero eccessivo di repliche, per mancanza di sufficienti

Tabella 5.5: Valori ottenuti del numero medio di repliche presenti in ogni singola regione

	n. repliche Bin-0		**n. repliche negli altri bin**		**n. repliche totali**	
	AOD	**AODSIM**	**AOD**	**AODSIM**	**AOD**	**AODSIM**
ALL	1.18	1.11	2.66	1.94	2.41	1.66
ALLnoUS	1.17	1.11	2.22	1.74	2.05	1.54
US	1.00	1.00	1.32	1.17	1.30	1.12
RU	1.00	1.00	1.00	1.00	1.00	1.00
DE	1.00	1.00	1.03	1.07	1.03	1.06
UK	1.00	1.00	1.03	1.04	1.02	1.02
FR	1.00	1.00	1.03	1.11	1.02	1.08
ES	1.00	1.00	1.19	1.04	1.15	1.03
IT	1.00	1.02	1.09	1.04	1.09	1.03

risorse di storage. Se si effettua lo stesso controllo sui T2 statunitensi, invece, il numero medio di accessi è 23.7 (10.5) contando le repliche, mentre è 30.8 (11.8) non contando le repliche. Questo indica una strategia completamente differente nella popolazione dello storage ai T2 per gli analisti, dove gli AOD sono replicati anche il 30% di più di quanto non lo siano stati in Italia (vedi tabella 5.5). In generale, osservando la situazione singolarmente in tutte le regioni che non siano gli USA, il numero di repliche è di poco superiore a 1 (ad esempio, 1.00 per RU, 1.02 per FR e UK, fino a 1.15 per ES). Quindi, gli USA sono l'unica regione in CMS che può permettersi - per disponibilità di risorse di storage - di ospitare un maggior numero di repliche a vantaggio della flessibilità nell'allocazione dei job, e dunque nel "throughput" complessivo dell'analisi degli utenti USA e di tutta CMS. Va infatti anche considerato che gli USA sono i principali attori nell'attività di fornitori di "remote access" via xrootd ai dati, che contribuisce alla popolarità degli stessi dal punto di vista globale di CMS, ma il cui contributo al momento non si può distinguere dagli accessi locali: non è al momento possibile ancora verificare questa ipotesi quantitativamente, ma si suppone che gli USA ospitino al momento più repliche (1.30) e riescano a tenere il bin-0 relativamente più basso di altre regioni anche in virtù degli accessi remoti da centri di calcolo fuori dagli USA, anche se questo comportamento sembra avere un maggior peso soprattutto sugli AOD. Per quanto riguarda gli AODSIM il Bin-0 statunitense ha un peso non trascurabile sul volume totale di questo datatier: è infatti circa il 27%.

5.4 Interazioni con il Computing Scrutiny

Tra i possibili "consumer" di informazioni legate alla popolarità dei dati degli esperimenti LHC e delle capacità di usare tali informazioni per ottimizzare l'utilizzo dello storage ai Tier-2 sono degni di nota anche i membri di comitati non CMS, quali il Computing Resources Scrutiny Group di WLCG (C-RSG) [65]. Si tratta di un gruppo di una dozzina di persone che rappresentano le "Funding Agencies" (FA) di nazioni diverse, ed ha lo scopo di dialogare con i progetti di Software e Computing degli esperimenti LHC ed informare il Computing Resources Review Board (C-RRB) in modo che vengano prese decisioni informate e ponderate su finanziamento e

allocazione delle risorse negli anni successivi. Ogni anno, gli esperimenti presentano richieste di acquisizione di risorse per il futuro che vengono sottoposte ad accurato "scrutiny" da parte del C-RSG: il dialogo tra il C-RSG e gli esperimenti è volto a capire sostanzialmente se le richieste sono adeguatamente motivate, e culmina in una serie di meeting annuali con il C-RRB in cui le richieste vengono ratificate o ridimensionate. Nel dettaglio, il C-RSG è chiamato a controllare: l'utilizzo delle risorse da parte di ogni esperimento LHC nell'anno solare precedente; le concrete richieste di risorse per l'anno successivo, e previsioni per ulteriori due anni; l'adeguato matching tra le richieste che gli esprimenti fanno e le capacità delle FA nazionali di soddisfarle; la valutazione di eventuali raccomandazioni ulteriori.

Nel corso del 2014, è emersa dal C-RSG la richiesta di comprendere le modalità di accesso ai dati dei vari esperimenti, onde studiare se non vi fossero eccessive inefficiente soprattutto nell'uso dello spazio disco ai T2, una risorsa costosa e cruciale per tutti gli esperimenti. É divenuto normale effettuare questi controlli, per tutti gli esperimenti, in modo leggermente differente (essendo diversi i modelli di calcolo, il numero dei T2, le tipologie di datatiers in uso in analisi, etc), ma basandosi per tutti su un concetto comune: il pattern di data popularity, appunto. A questo scopo, è stato richiesto di produrre graficamente tale informazione sotto forma di un unico plot - rigorosamente analogo per tutti gli esperimenti - che mostri il volume di storage acceduto 0, 1, 2,.. fino a >14 volte in tre diverse finestre temporali nel passato, ovvero gli ultimi 3 mesi, 6 mesi e 12 mesi. Lo studio condotto in questa tesi ha permesso di avere già a disposizione tutti i dati per produrre tale plot, e di intercettare dunque con facilità la richiesta e soddisfarla velocemente. Un esempio del tipo di plot che è stato prodotto e usato ufficialmente da CMS nelle discussioni con il C-RSG verso Novembre 2014 è mostrato in Figura 5.5. Come si vede, una consistente porzione dei dati di CMS interessanti all'analisi è acceduta di frequente nell'anno precedente la data di riferimento, mentre la quantità di dati residenti sui T2 che non sono mai stati letti in quel periodo ammonta a poco più di 2 PB. Tale numero è consistente con quanto prodotto a Gennaio 2015 negli studi discussi in precedenza in questo capitolo, e il leggero aumento è confermato dal fatto che da Novembre non sono state effettuate consistenti azioni di pulizia di dati raramente acceduti.

L'esperienza dell'interazione con il C-RSG ha dunque mostrato che il toolkit sviluppato in questa tesi è risultato affidabile dal punto di vista tecnico, riproducibile per quanto riguarda le informazioni prodotte, e di una certa importanza per attività anche manageriali dell'esperimento CMS.

5.5 Pubblicazione dei risultati e accesso all'infrastruttura

Questo lavoro permette di mettere a disposizione di qualunque utente CMS un piccolo framework stabile e costantemente monitorato che raccoglie ed elabora dati sui pattern di accesso ai dati di CMS e li "digerisce" fornendone una rappresentazione di facile accesso via web e in base a opzioni altamente configurabili dall'utente, dunque flessibile in base ai casi d'uso del "consumer" di tali informazioni (sia egli un analista di fisica, un project manager di un'area Software/Computing di

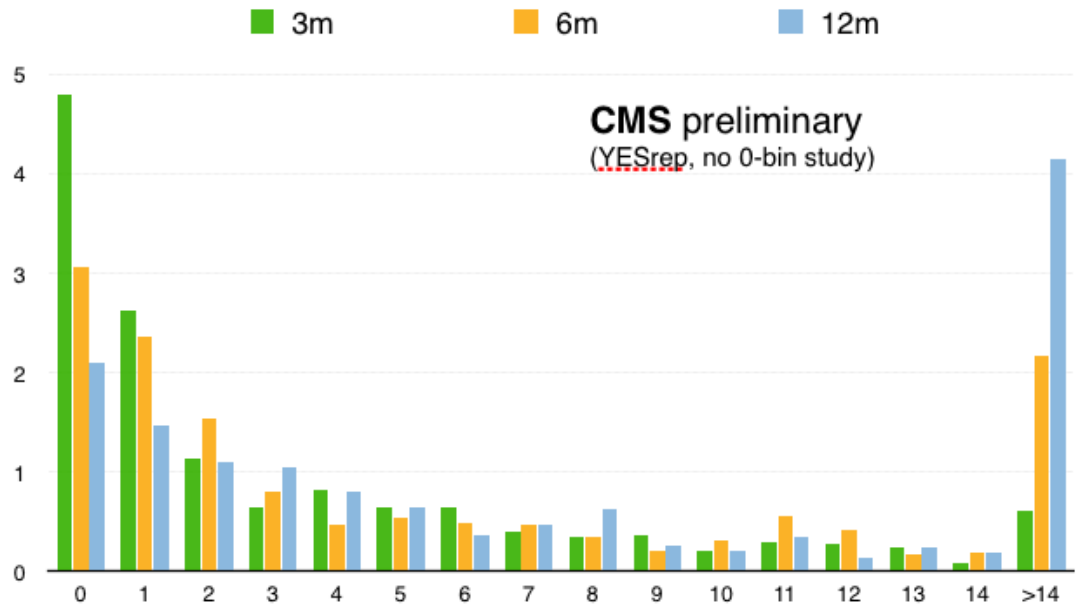


Figura 5.5: Esempio di plot presentato e usato ufficialmente da CMS nelle discussioni con il C-RSG

CMS, o un amministratore di un centro Tier-2). La flessibilità dello strumento deriva dall'architettura del toolkit stesso, in particolare la personalizzazione che caratterizza ogni Step della sua architettura lo rende un ottimo candidato a poter essere ulteriormente sviluppato. Dal momento che vari colleghi in CMS hanno interesse ad accedere a tali informazioni dopo un primo processamento delle stesse per renderle facilmente fruibili, è stato ritenuto opportuno preparare un sito web che, una volta selezionata la data di interesse e la finestra temporale di interesse nel passato fino a tale data, offra un display di tutti i plot principali che il toolkit produce in modo regolare. L'utilizzo di questa pagina, mostrata in Figura 5.6, temporaneamente ospitata nello spazio web CERN del relatore della presente tesi, verra ulteriormente migliorata e valutata da colleghi di CMS per l'inserimento in pagine pubbliche di esperimento.

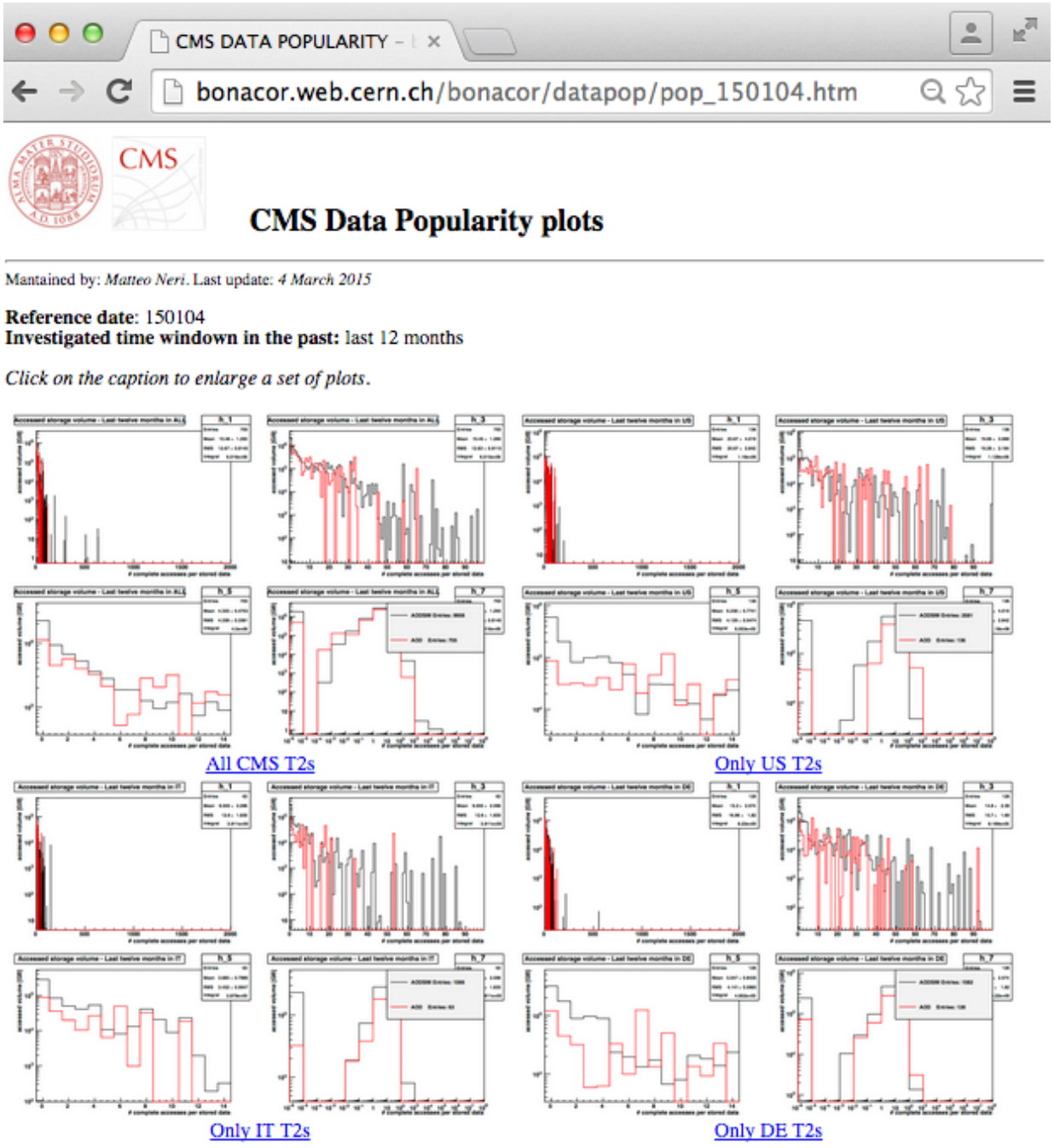


Figura 5.6: Pagina html attraverso la quale è possibile accedere ai plot prodotti dal toolkit sviluppato in questo lavoro di tesi.

Appendice A

Esempio di file di configurazine e di uno script di avvio

Di seguito si riportano i file di configurazione dello Step-1 e lo script con cui viene lanciato in parallelo lo Step-1. Il mutex, necessario affinchè non avvengano modifiche al file di configurazione dello Step-1 mentre viene caricato da uno dei job avviati, è creato attraverso la creazione e la distruzione nei momenti opportuni di un file che agisce come un semaforo per rendere l'accesso al file di configurazione sequenziale.

Listing A.1: Step-1 configuration file

```
#
# CONFIGURATION FILE
#


#IMPOSTAZIONI GENERALI

# directory in cui sono  i file (str).
# nome base dei file: directory+'/'+base_name_file % tiers+'.
   json'
input_directory = "/gpfs_data/local/cms/bonacor/forMatteoNeri
   /150104/data_150104_9m"
base_name_file = 'popcombine_%s'

# output: 'output_directory/HistoMaker_block_'+collection+'
   _output.format
date = input_directory.split("_")[-2]
output = input_directory.split("_")[-1]
T2_output = "Wisconsin"
output_directory = "/gpfs_data/local/cms/CMS_data_popularity/
   Wisconsin_HistoMaker/%s/%s" %(date, date+"_"+output+"_"+
   T2_output)


# elenco tiers
```

```
# tiers = ["T2_AT_Vienna","T2_BE_IIHE","T2_BE_UCL","
    T2_BR_SPRACE","T2_BR_UERJ","T2_CH_CSCS","T2_CN_Beijing","
    T2_DE_DESY","T2_DE_RWTH","T2_EE_Estonia","T2_ES_CIEMAT","
    T2_ES_IFCA","T2_FI_HIP","T2_FR_CCIN2P3","T2_FR_GRIF_IRFU
    ","T2_FR_GRIF_LLR","T2_FR_IPHC","T2_GR_Ioannina","
    T2_HU_Budapest","T2_IN_TIFR","T2_IT_Bari","T2_IT_Legnaro
    ","T2_IT_Pisa","T2_IT_Rome","T2_KR_KNU","T2_MY_UPM_BIRUNI
    ","T2_PK_NCP","T2_PL_Warsaw","T2_PT_NCG_Lisbon","
    T2_RU_IHEP","T2_RU_INR","T2_RU_ITEP","T2_RU_JINR","
    T2_RU_PNPI","T2_RU_RRC_KI","T2_RU_SINP","T2_TH_CUNSTDA","
    T2_TR_METU","T2_TW_Taiwan","T2_UA_KIPT","
    T2_UK_London_Brunel","T2_UK_London_IC","
    T2_UK_SGrid_Bristol","T2_UK_SGrid_RALPP","T2_US_Caltech","
    T2_US_Florida","T2_US_MIT","T2_US_Nebraska","T2_US_Purdue
    ","T2_US_UCSD","T2_US_Vanderbilt","T2_US_Wisconsin","
    T2_CH_CERN"]
tiers = ["T2_US_Wisconsin"]


# IMPOSTAZIONI ISTOGRAMMI PER DATASET E BLOCCHI
block_ds_bool = True

# type ['/AOD#', '/AODSIM#']     #tiene solo i ds dei tipi
    contenuti nell'array
ds_type_bool = True
ds_types = ["/AODSIM#"]

# group 'DataOps' 'FacOps'       #come per i types
group_bool = False
groups = ["DataOps"]


# IMPOSTAZIONI ANALISI PER DS SINGOLI
singleDS_bool = False

# DS names
singleDS_names = ["/SingleMu/Run2012D-22Jan2013-v1/AOD", "/
    SUSY_sqgo_mH1_85_mH2_125p3_MSUSY_1000_8TeV_Pythia6/
    Summer12_DR53X-PU_S10_START53_V7C-v1/AODSIM"]

collection = ""
if ds_type_bool == True :
    for type in ds_types:
        collection += '_%s' % type[1:-1]

if group_bool == True :
    for group in groups:
        if collection:
            collection += '_%s' % group
        else:
```

```
            collection = group


print "reading data from: "+input_directory
print "writing output in: "+output_directory
print "output for time: "+output+"\n"

Semaphore = open("GREEN_LED","w")
Semaphore.close()
```

Listing A.2: launch Step 1

```
#!/bin/bash

LOG_FILE="/gpfs_data/local/cms/CMS_data_popularity/LOG_HM"
date >> $LOG_FILE
DATA_FOLDER="/gpfs_data/local/cms/bonacor/forMatteoNeri/" #
   necessario l'ultimo slash!!!
START_TYPE="ds_types = [\"\/AOD#\"]"

echo @@@@@@@@@@@@@@@@@@@@@@@@
date
echo @@@@@@@@@@@@@@@@@@@@@@@@

if [ ! $# == 2 ]; then
    echo "Usage: $0 start_date end_date"
    exit 1
fi

if [ ! -e GREEN_LED ]; then
    touch GREEN_LED
fi

for data in ${DATA_FOLDER}*
do
    (
        echo "
           ############################################################"
        echo $data":"
        echo
        echo "
           ############################################################"
            >> $LOG_FILE    #DB
        echo $data":"
                    >> $LOG_FILE    #DB
        echo

           >> $LOG_FILE    #DB
        for data_period in ${data}/*
        do
```

```
if [[ -d $data_period ]]; then
    if [ "${data_period:${#data_period}-3:3}" ==
       12m ]; then
        DATE=${data_period:${#data_period}-10:6}
    else
        DATE=${data_period:${#data_period}-9:6}
    fi
    if (("$DATE" >= "$1")) && (("$DATE" <= "$2"))
       ; then
        while [ ! -e GREEN_LED ]
        do
            :
        done
        if [ -e GREEN_LED ]; then
            echo "
               ------------------------------------
               "
            echo "
               ------------------------------------
               "
            echo
            echo "
               ------------------------------------
               " >> $LOG_FILE
            INPUT_DIRECTORY="input_directory = \"
               ${data_period}\""
            echo $INPUT_DIRECTORY
            sed -i "36s!.*!${INPUT_DIRECTORY}!"
               config.py
            sed -i "56s/.*/$START_TYPE/" config.
               py
            rm GREEN_LED
            echo _____________
            echo PARSING AOD
            echo
            date >> $LOG_FILE
            ./HistoMaker_Matteo3.py 2>> $LOG_FILE
                &
        fi
        while [ ! -e GREEN_LED ]
        do
            :
        done
        if [ -e GREEN_LED ]; then
            sed -i "56s/AOD/AODSIM/" config.py
            rm GREEN_LED
            echo _______________
            echo PARSING AODSIM
            echo
            date >> $LOG_FILE
```

```
                        ./HistoMaker_Matteo3.py 2>> $LOG_FILE
                            &
                    fi
                else
                    echo skipping $data_period
                fi
            fi
        done
        echo ".......... waiting .........."
        wait
    )
done

echo "-------------------------------------------"

echo "-------------------------------------------" >> $LOG_FILE
   #DB
date                                               >> $LOG_FILE
   #DB
echo DONE                                          >> $LOG_FILE
   #DB

# done!

exit 0
```

Conclusioni

L'uso ottimizzato dello storage ai centri di calcolo di livello Tier-2 sarà cruciale, in tempi di ristrettezza di risorse, ai fini dell'efficienza del sistema di calcolo di un esperimento LHC a Run-2. In CMS, studi come quello presentato in questa tesi consentono di dotare l'esperimento di strumenti capaci di fornire, regolarmente nel tempo, informazioni preziose su quali sono i dati più popolari o - al contrario - sulle porzioni di storage occupate da dati che non vengono acceduti, in ogni finestra temporale di interesse.

L'impatto di questo lavoro è forte in almeno tre direzioni. La prima direzione consiste nel dotare i gruppi di Software e Computing dell'esperimento CMS di un toolkit che permette di acquisire consapevolezza dell'effettiva efficacia delle scelte di "static data placement" effettuate dal modello di calcolo, ed eventualmente correre al riparo con lo sviluppo (attualmente in corso) di modelli più evoluti di "dynamic data placement". La seconda direzione consiste nel fornire a chi si occupa di allocazione e controllo d'uso delle risorse uno strumento quantitativo che consenta di fare scelte strategiche informate. In questo senso, questo lavoro si è mostrato estremamente utile nelle interazioni tra il CMS Resource Management Office e il WLCG Computing Resource Scrutiny Group. La terza direzione consiste nell'offrire a CMS una maggiore consapevolezza sulla natura e il significato più profondo di alcuni dei metadati relativi alle operazioni di calcolo, una delle classi di dati non-fisici (ovvero non provenienti dalle collisioni) la cui comprensione viene spesso erroneamente ritenuta troppo complessa e/o di secondaria importanza, ma che costituisce una ingente mole di informazioni preziose che - se trattate in futuro in modo opportuno con tecniche quali quelle di Big Data analytics - potrebbe rappresentare la base di uno strumento real-time dalle preziose potenzialità predittive.

Bibliografia

[1] Oliver Sim Brüning et al. *LHC Design Report*. A cura di CERN library copies. Vol. 1, 2, 3. 2012. URL: http://ab-div.web.cern.ch/ab-div/Publications/LHC-DesignReport.html (cit. alle pp. 1, 2).

[2] Lyndon Evans e Philip Bryant. «LHC Machine». In: *Journal of Instrumentation* 3.08 (2008). A cura di IOPscience, S08001. URL: http://iopscience.iop.org/1748-0221/3/08/S08001 (cit. alle pp. 1, 2).

[3] *CERN*. URL: http://www.cern.ch (cit. a p. 1).

[4] *The Alice Collaboration*. URL: http://aliceinfo.cern.ch (cit. a p. 4).

[5] The ALICE Collaboration et al. «The ALICE experiment at the CERN LHC». In: *Journal of Instrumentation* 3.08 (2008). A cura di IOPscience, S08002. URL: http://iopscience.iop.org/1748-0221/3/08/S08002 (cit. a p. 4).

[6] *The Atlas Collaboration*. URL: http://atlas.web.cern.ch/Atlas/Collaboration (cit. a p. 4).

[7] The ATLAS Collaboration et al. «The ATLAS Experiment at the CERN Large Hadron Collider». In: *Journal of Instrumentation* 3.08 (2008). A cura di IOPscience, S08003. URL: http://iopscience.iop.org/1748-0221/3/08/S08003 (cit. a p. 4).

[8] *The CMS Collaboration*. URL: http://cms.web.cern.ch (cit. a p. 4).

[9] The CMS Collaboration et al. «The CMS experiment at the CERN LHC». In: *Journal of Instrumentation* 3.08 (2008), S08004. URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08004 (cit. alle pp. 4, 6, 10).

[10] *The LHCb Collaboration*. URL: http://lhcb.web.cern.ch/lhcb (cit. a p. 5).

[11] The LHCb Collaboration et al. «The LHCb Detector at the LHC». In: *Journal of Instrumentation* 3.08 (2008). A cura di IOPscience, S08005. URL: http://iopscience.iop.org/1748-0221/3/08/S08005 (cit. a p. 5).

[12] *The LHCf experiment*. URL: http://home.web.cern.ch/about/experiments/lhcf (cit. a p. 5).

[13] The LHCf Collaboration et al. «The LHCf detector at the CERN Large Hadron Collider». In: *Journal of Instrumentation* 3.S08006 (2008). A cura di IOPscience. URL: http://iopscience.iop.org/1748-0221/3/08/S08006 (cit. a p. 5).

[14] *The TOTEM Collaboration*. URL: http://totem.web.cern.ch/Totem/ (cit. a p. 5).

[15] The TOTEM Collaboration et al. «The TOTEM Experiment at the CERN Large Hadron Collider». In: *Journal of Instrumentation* 3.S08007 (2008). A cura di IOPscience. URL: http://iopscience.iop.org/1748-0221/3/08/S08006 (cit. a p. 5).

[16] Ian Bird. «Computing for the Large Hadron Collider». In: *Annual Review of Nuclear and Particle Science* 61.1 (2011), pp. 99–118. DOI: 10.1146/annurev-nucl-102010-130059. eprint: http://dx.doi.org/10.1146/annurev-nucl-102010-130059. URL: http://dx.doi.org/10.1146/annurev-nucl-102010-130059 (cit. alle pp. 11, 13).

[17] *WLCG Project*. URL: http://www.cern.ch/lcg (cit. a p. 11).

[18] *Enabling Grind for E-sciencE (EGEE)*. URL: http://www.eu-egee.org (cit. a p. 11).

[19] *European Grid Infrastructure (EGI)*. URL: http://www.egi.eu/ (cit. a p. 11).

[20] *Open Science Grid (OSG)*. URL: http://www.opensciencegrid.org (cit. a p. 11).

[21] Daniele Bonacorsi (for the CMS Computing Model). «Experience with the CMS Computing Model form commissioning to collision». In: *Journal of Physics*. Conference Series 331.7 (2010). A cura di IOPscience, p. 072005. URL: http://iopscience.iop.org/1742-6596/331/7/072005 (cit. alle pp. 12, 22).

[22] Daniele Bonacorsi (on behalf of the CMS Collaboration). «The CMS Computing Model». In: *Nuclear Physics B - Proceedings Supplements*. Conference Series 172.53-56 (2007). A cura di Science Direct. URL: http://www.sciencedirect.com/science/article/pii/S092056320700552X (cit. a p. 12).

[23] *MONARC project*. URL: http://monarc.web.cern.ch/MONARC (cit. a p. 12).

[24] *International Electrotechnical Commission*. URL: http://www.iec.ch/ (cit. a p. 15).

[25] G. L. Bayatyan et al. *CMS computing: Technical Design Report*. Technical Design Report CMS. Submitted on 31 May 2005. Geneva: CERN, 2005 (cit. alle pp. 16, 20).

[26] *RECO Data Format Table*. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideRecoDataTable (cit. a p. 15).

[27] *AOD Data Format Table*. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideAodDataTable (cit. a p. 16).

[28] M. Cinquilli et al. «The CMS workload management system». In: *Journal of Physics*. Conference Series 396.3 (2012). A cura di IOPscience, p. 032113. URL: http://iopscience.iop.org/1742-6596/396/3/032113 (cit. a p. 16).

[29] M. Giffels et al. «The CMS Data Management System». In: *Journal of Physics*. Conference Series 513.4 (2014). A cura di IOPscience, p. 042052. URL: http://iopscience.iop.org/1742-6596/513/4/042052 (cit. a p. 16).

[30] Sudhir Malik et al. *CMSSW Application Framework*. 16 Giu. 2014. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSSWFramework (cit. a p. 17).

[31] Tony Wildish et al. «From toolkit to framework - the past and future evolution of PhEDEx». In: *Journal of Physics*. Conference Series 396.3 (2012). A cura di IOPscience, p. 032118. URL: http://iopscience.iop.org/1742-6596/396/3/032118 (cit. a p. 17).

[32] J. Rehn et al. «PhEDEx high-throughput data transfer management system». In: *CHEP06* (2006). A cura di GridPP. URL: http://www.gridpp.ac.uk/papers/chep06_tuura.pdf (cit. a p. 17).

[33] M Giffels, Y Guo e D Riley. «Data Bookkeeping Service 3 – Providing event metadata in CMS». In: *Journal of Physics: Conference Series*. Conference Series 513.4 (2014), p. 042022. URL: http://stacks.iop.org/1742-6596/513/i=4/a=042022 (cit. a p. 17).

[34] G Ball et al. «Data Aggregation System - a system for information retrieval on demand over relational and non-relational distributed data sources». In: *Journal of Physics: Conference Series* 331.4 (2011), p. 042029. URL: http://stacks.iop.org/1742-6596/331/i=4/a=042029 (cit. a p. 17).

[35] *DAS web page*. URL: https://cmsweb.cern.ch/das/ (cit. a p. 17).

[36] *ORACLE*. URL: http://www.oracle.com/ (cit. alle pp. 18, 27).

[37] R. Egeland et al. «The PhEDEx next-gen website». In: *Journal of Physics*. Conference Series 396.3 (2012). A cura di IOPscience, p. 032117. URL: http://iopscience.iop.org/1742-6596/396/3/032117 (cit. a p. 18).

[38] R. Egeland, C.-H. Huang e T. Wildish. «PhEDEx Data Service». In: *Journal of Physics*. Conference Series 219.6 (2010). A cura di IOPscience, p. 062010. URL: http://iopscience.iop.org/1742-6596/219/6/062010 (cit. a p. 18).

[39] Giuseppe Codispoti et al. «CRAB: A CMS Application for Distributed Analysis». In: *Nuclear Science Symposium Conference Record* N02.79 (2008). A cura di IEEE. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4774652 (cit. a p. 18).

[40] *Software Guide on CRAB*. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideCrab (cit. a p. 18).

[41] Alessandra Fanfani et al. «Distributed Analysis in CMS». In: *CMS-NOTE-2009-013, CERN-CMS-NOTE-2009-013* (2009). URL: http://inspirehep.net/record/875969 (cit. a p. 18).

[42] M. Cinquilli et al. «CRAB3: Establishing a new generation of services for distributed analysis at CMS». In: *Journal of Physics*. Conference Series 396.3 (2012). A cura di IOPscience, p. 032026. URL: http://iopscience.iop.org/1742-6596/396/3/032026 (cit. a p. 19).

[43] Andres Tanasijczuk. *CRAB3 architecture and task workflow*. 23 Ott. 2014. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/CRAB3TaskFlow (cit. a p. 20).

[44] Daniele Bonacorsi e Tony Wildish. «Challenging data and workload management in CMS Computing with network-aware systems». In: CMS-CR-2013-373 (ott. 2013). URL: https://cds.cern.ch/record/1626815/ (cit. alle pp. 21–23).

[45] Ian Bird et al. «Update of the Computing Models of the WLCG and the LHC Experiments». In: *Nuclear Physics B - Proceedings Supplements* LCG-TDR-002 (15 apr. 2014). A cura di CERN-LHCC-2014-014. URL: http://cds.cern.ch/record/1695401/files/LCG-TDR-002.pdf?version=1 (cit. a p. 21).

[46] *LHCOPN*. URL: http://lhcopn.web.cern.ch (cit. a p. 21).

[47] Daniele Bonacorsi e Anthony Wildish. *Challenging Data Management in CMS Computing with Network-aware Systems*. Rapp. tecn. CMS-CR-2013-426. Geneva: CERN, nov. 2013. URL: https://cds.cern.ch/record/1977895/ (cit. a p. 22).

[48] Kenneth Bloom e the Cms Collaboration. «CMS Use of a Data Federation». In: *Journal of Physics: Conference Series* 513.4 (2014), p. 042005. URL: http://stacks.iop.org/1742-6596/513/i=4/a=042005 (cit. a p. 22).

[49] F. H. Barreiro Megino et al. «Implementing data placement strategies for the CMS experiment based on popularity model». In: *Journal of Physics. Conference Series* 396.3 (2012). A cura di IOPscience, p. 032047. URL: http://iopscience.iop.org/1742-6596/396/3/032047 (cit. alle pp. 25, 32).

[50] Angelos Molfetas et al. «Popularity framework to process dataset traces and its application on dynamic replica reduction in the ATLAS experiment». In: *Journal of Physics. Conference Series* 331.6 (2011). A cura di IOPscience, p. 062018. URL: http://iopscience.iop.org/1742-6596/331/6/062018 (cit. a p. 25).

[51] *The Scalla Software Suite: xrootd/cmsd*. URL: http://xrootd.slac.stanford.edu (cit. a p. 26).

[52] J. Andreeva et al. «Experiment Dashboard for Monitoring of the LHC Distributed Computing Systems». In: *Journal of Physics. Conference Series* 331.7 (2011). A cura di IOPscience, p. 072001. URL: http://iopscience.iop.org/1742-6596/331/7/072001/ (cit. a p. 27).

[53] J. Andreeva1 et al. «Dashboard for the LHC experiments». In: *Journal of Physics. Conference Series* 119.6 (2008). A cura di IOPscience, p. 062008. URL: http://iopscience.iop.org/1742-6596/119/6/062008 (cit. a p. 27).

[54] Andreas J. Peters e Lukasz Janyst. «Exabyte Scale Storage at CERN». In: *Journal of Physics. Conference Series* 331.5 (2011). A cura di IOPscience, p. 052015. URL: http://iopscience.iop.org/1742-6596/331/5/052015 (cit. a p. 27).

[55] *JSON*. URL: http://www.json.org/ (cit. alle pp. 28, 36).

[56] *The Django framework*. URL: https://www.djangoproject.com (cit. a p. 28).

[57] *memcached*. URL: http://memcached.org (cit. a p. 29).

[58] *DataTables*. URL: http://datatables.net (cit. a p. 29).

[59] *Highcharts*. URL: www.highcharts.com (cit. a p. 29).

[60] *Python*. URL: https://www.python.org/ (cit. a p. 39).

[61] *ROOT Data Analysis Framework*. URL: https://root.cern.ch/drupal (cit. a p. 40).

[62] *ROOT TTree*. URL: https://root.cern.ch/root/html/TTree.html (cit. a p. 40).

[63] *ROOT TH1F*. URL: https://root.cern.ch/root/html/TH1F.html (cit. a p. 40).

[64] CMS Computing Data Management developement lead Nicolo Magini. private communication. 2015 (cit. a p. 55).

[65] *Scrutiny Group*. URL: http://wlcg.web.cern.ch/collaboration/management/computing-resources-scrutiny-group.