DOI: 10.26314/GARR-Conf19-proceedings-19

# MineHEP: Data Mining in High Energy Physics

Laura Redapi [1,5], Andrea Cioni[2], Maria Vittoria Garzelli [1,3,4], Andrea Ceccarelli[2], Piergiulio Lenzi[1,3], Vitaliano Ciulli[1,3]

[1] Università degli Studi di Firenze, Dipartimento di Fisica e Astronomia,
[2]Università degli Studi di Firenze, Dipartimento di Matematica e Informatica,
[3]INFN, Firenze, [4] Tuebingen Universitaet, Institute fur Theoretische Physik,
[5]Consortium GARR

**Abstract.** MineHep è un progetto interdisciplinare che nasce dalla collaborazione tra il Dipartimento di Fisica e Astronomia ed il Dipartimento di Matematica e Informatica dell'Università di Firenze, con l'obiettivo di fornire una serie di strumenti software a supporto dell'archiviazione e dell'analisi dei dati raccolti dalla comunità della Fisica delle Alte Energie negli ultimi sessant'anni. L'obiettivo innovativo di MineHep è fornire strumenti per estrarre informazioni rilevanti in modo coerente da un'enorme mole di dati, utilizzando avanzate tecniche di data mining e data warehousing. Queste tecniche, solitamente applicate in campo aziendale, vengono utilizzate qui in Fisica delle particelle per la prima volta. Questo progetto mira a facilitare la libera e pubblica diffusione di dati e dei risultati scientifici a supporto delle infrastrutture di ricerca. Il Consortium GARR ha partecipato al progetto in questa prima fase finanziando una borsa di studio di un anno

**Keywords.** Data Mining, Data Warehousing, Open Data, High-Energy Physics

## Introduzione

Nonostante il successo del Modello Standard (MS), evidenziato dalla scoperta del bosone di Higgs avvenuta nel 2012, i fisici continuano a ritenere che si tratti di una teoria incompleta, in quanto non spiega alcuni aspetti del cosmo (come, ad esempio, la presenza di materia oscura) e non risponde ad alcune questioni concettuali fondamentali (come, ad esempio, la gerarchia tra le masse fermioniche). Sebbene tutte le ricerche esplicite di particelle diverse da quelle contenute nel MS ad oggi non abbiano portato alla scoperta di alcuna particella di Nuova Fisica (D. Kazakov, 2017), la presenza di Nuova Fisica potrebbe essere rimasta invisibile alle singole analisi, ciascuna ristretta ad un sottoinsieme limitato di dati, poiché sparsa tra canali diversi, esperimenti diversi e collider diversi. Al giorno d'oggi, all'interno della comunità della Fisica alle Alte Energie (HEP), non abbiamo modo di gettare uno sguardo di insieme che abbracci contemporaneamente tutti i dati raccolti, da una prospettiva e con uno sforzo di interpretazione globale.

MineHEP si propone come un nuovo approccio per estrarre informazioni dall'intero insieme di dati pubblicati in articoli scientifici dalla comunità di HEP. MineHEP è un cambio di paradigma nella ricerca di Nuova Fisica. Mentre solitamente l'analisi dei dati in HEP è organizzata come studio della compatibilità di un particolare modello di Nuova Fisica con specifici set di dati sperimentali contenuti in un unico documento scientifico, Mi-

neHep beneficia della disponibilità dei dati già presenti nel database HepData (e dei molti altri che vi confluiranno in futuro) per testare le ipotesi in modo globale, consentendo l'analisi simultanea di più dati e la reinterpretazione di vecchie informazioni nel contesto di nuovi modelli, rendendo i dati raccolti più utili nel tempo.

Dimostreremo che le tecnologie esistenti per Data Warehousing (DW) solitamente usate in campo aziendale possono essere portate con successo in nuovi settori, ottenendo risultati innovativi e utili a risolvere diverse criticità.

## 1. Da HepData a MineHEP

### *1.1 Pregi e limiti di HepData/HEPData*

HepData (High Energy Physics Database) è stato costruito presso la Durham University in Inghilterra, a partire dagli anni sessanta, e attualmente contiene i dati di migliaia di pubblicazioni, incluse pubblicazioni relative alle collisioni al Large Hadron Collider (LHC). Una revisione portata avanti nel decennio 2000-2010 nell'ambito del progetto CEDAR, ha condotto a una versione intermedia, ancora consultabile pubblicamente (http://hepdata.cedar.ac.uk), che prevedeva un database di tipo relazionale (MySQL) e offriva la possibilità, ad un utente esperto, di accedere a quei dati attraverso un'interfaccia web basata su Java. A causa della struttura particolarmente complessa ed articolata, risulta difficile effettuare azioni di data mining su questa versione. Questo è stato uno dei motivi che ha portato ad importanti modifiche a favore di nuove tecnologie e strumenti più adatti alla analisi di dati.

Una nuova versione (https://www.hepdata.net), che fa uso di PostgreSQL è stata rilasciata nel 2016. Ad oggi la ricerca su HEPData degli articoli scientifici di interesse, che rispondono a valori fissati di determinati parametri, risulta molto veloce ed efficiente, ma allo stesso tempo, la ricerca è limitata a query sulla base di questi metadati, che producono come output liste di articoli scientifici per i quali i metadati hanno i valori richiesti dall'utente. Risulta quindi possibile una ricerca per articoli, ma non per dati interni agli articoli stessi, limitando anche la possibilità di visualizzare contemporaneamente in un medesimo grafico i dati provenienti da analisi diverse per eventuali confronti. Per rispondere alle nostre esigenze di ricerca e analisi, abbiamo identificato e selezionato tecniche di analisi Online Analytical Processing (Kimball Group website) che sono oggi strumenti consolidati per logiche decisionali di azienda, e le abbiamo esportate e applicate al nostro problema in ambito HEP.

### *1.2 Estrazione di dati e pulitura del database HepData*

I membri del gruppo HEP dell'Università di Durham ci hanno fornito un dump del database MySQL che contiene tutti i dati già presenti in HepData fino ad Aprile 2018.

Questa versione ci ha permesso di accedere allo schema di tabelle in cui i dati sono organizzati e di comprende la relazione tra le varie entità del database, indispensabile per gli steps successivi. Sono state fatte diverse operazioni tra cui l'eliminazione di tabelle vuote e chiavi duplicate, l'eliminazione di alcune informazioni ridondanti, oltre a una prima riorganizzazione dei dati che ha evidenziato l'importanza di rendere omogenei i loro formati.

### *1.3 Progettazione e popolamento del nuovo database MineHep*

Il modello dimensionale risponde a due requisiti fondamentali per la presentazione di dati analitici: fornire l'informazione in maniera chiara e permettere all'utente di accedervi velocemente. La semplicità assicura che gli utenti possano facilmente comprendere i dati, ma anche che i software possano navigare e consegnare i risultati velocemente ed efficientemente. Una volta pulito il database originale, abbiamo progettato uno schema a stella del tipo mostrato in Figura 1. In uno schema a stella, la tabella dei fatti è l'unica a possedere collegamenti multipli con le altre tabelle, che si concretizzano attraverso gli identificatori univoci (campi chiavi) in essa contenuti. Le dimensioni hanno tutte un unico collegamento alla tabella centrale, con lo scopo di minimizzare il numero di istruzioni richieste per ciascuna query e facilitare l'interrogazione del database da parte di un utente.

Del nostro schema si nota la centralità della tabella dei fatti, YAxisFacts, che contiene i dati misurati di ogni esperimento e, intorno a questa, otto dimensioni che caratterizzano il singolo dato numerico.

Una volta progettato il nuovo database, l'intera procedura di importazione è stata automatizzata attraverso l'uso di Docker e Docker Compose prevedendo l'integrazione con sistemi di generazione dashboard e altri software.

Sono stati infine scritti ed eseguiti manualmente 112 test per verificare che il nuovo database contenesse informazioni coerenti con quelle contenute nel database HepData originale. I test effettuati hanno dato esito positivo.

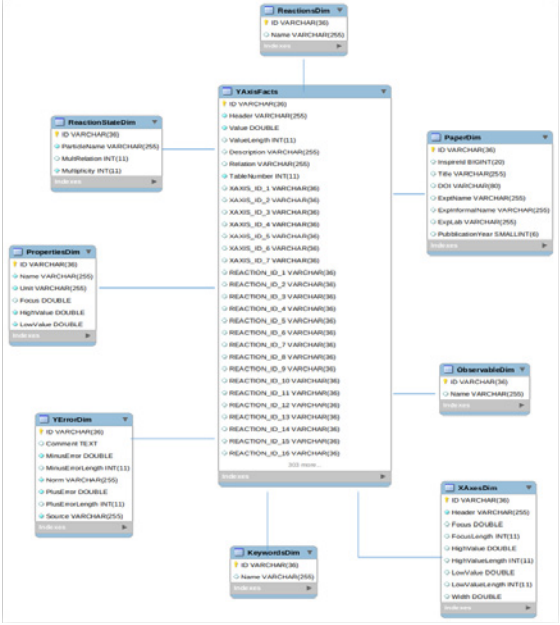


Fig.1
Diagramma a stella del database MineHep

## 2. Sviluppo della query interface/dashboard per la visualizzazione dei dati di MineHep

Una prima esigenza dell'utente è suddividere in diverse serie i risultati ottenuti e inserirli in un unico grafico per poter fare dei confronti. Il software di DW scelto per la costruzione di

una dashboard accessibile via web, per permettere una prima presentazione dei risultati, è stato Metabase (Metabase website). Metabase è uno strumento open source che permette di interrogare ed estrarre conoscenza dai dati in modo semplice, fornendo un'interfaccia SQL. Abbiamo preimpostato in Metabase delle query in linguaggio SQL, che l'utente può completare attraverso semplice menu a tendina. All'utente è inoltre possibile salvare le queries in una dashboard, dove confrontare risultati. Di seguito riportiamo un esempio per l'esplorazione e il confronto di dati estratti da articoli scientifici diversi:
Cerchiamo tutti i dati sperimentali relativi a distribuzioni differenziali $d\sigma/dp_T$ per la reazione $pp \rightarrow B^+ + X$, con produzione di mesoni $B^+$ di impulso trasversale nell'intervallo compreso tra 0 e 250 GeV e rapidità < 5, a fissata energia di collisione, e riportiamoli in un unico grafico.


Fig.2
Output di Metabase alla query di interesse riportata nel testo

Usando Metabase la risposta cercata può essere ottenuta dall'esecuzione di più query in serie: la prima seleziona i papers e le tabelle in cui i dati relativi all'osservabile di interesse sono riportati, mentre le successive estraggono l'informazione, che può essere visualizzata tramite tabella o tramite plot. Infine, è possibile sovrapporre i grafici ottenuti, permettendo di evidenziare differenze o analogie tra i vari set di misure in esame. L'output di Metabase alla successione di query descritta è riportato in Figura 2.

## 3. Conclusioni

Abbiamo dimostrato come la riorganizzazione del database HepData nel database MineHep tramite tecniche di DW, permetta il confronto automatico tra serie di dati sperimentali presentate in articoli scientifici differenti, ottenute sotto le medesime condizioni di analisi. Questa nuova potenzialità può permettere di evidenziare in maniera semi-automatica anomalie di specifici set di dati, che potrebbero indicare la presenza di Nuova

Fisica. L'estensione di tecniche di DW alla riorganizzazione di database di interesse in altri campi scientifici è possibile e può permettere un uso più proficuo dell'informazione già raccolta, migliorando la comprensione di grandi quantità di dati.

## Riferimenti bibliografici

D. Kazakov, "Beyond the Standard Model '17", lectures at the European School of High-Energy Physics 2017, CERN Yellow Reports School Proc. 3 (2018), 83-131

Kimball Group - Star Schemas and OLAP Cubes, https://bit.ly/2FfF4AY

Metabase website URL: https://metabase.com/

## Autori

**Laura Redapi** - laura.redapi@gmail.com

Laura Redapi è specializzanda in Fisica medica presso l'Università degli Studi di Firenze. Ha vinto una borsa GARR per l'anno 2018/2019 durante il quale si è occupata del progetto MineHep estendendo l'utilizzo delle tecniche di DW anche a database medici.

**Andrea Cioni** - andrea.cioni6@stud.unifi.it

Andrea Cioni si è laureato in Informatica presso l'Università degli Studi di Firenze nel 2019, discutendo una tesi sull'uso di tecniche di DW per lo sviluppo del progetto MineHEP. Attualmente è software developer presso Florence Consulting Group.

**Maria Vittoria Garzelli** - maria.vittoria.garzelli@cern.ch

Maria Vittoria Garzelli è borsista post-dottorato nell'ambito del progetto MineHep, esperta in vari aspetti di fenomenologia delle particelle elementari e nello sviluppo di software per produrre predizioni teoriche da confrontare con i dati sperimentali.

**Andrea Ceccarelli** - andrea.ceccarelli@unifi.it

Andrea Ceccarelli è Ricercatore a Tempo Determinato (RTDb) in Informatica presso l'Università degli Studi di Firenze. I suoi interessi principali di ricerca riguardano la progettazione e validazione di sistemi informatici disponibili e sicuri.

**Piergiulio Lenzi** - piergiulio.lenzi@unifi.it

Piergiulio Lenzi è Ricercatore a Tempo Determinato (RTDb) in Fisica presso l'Università degli Studi di Firenze. Partecipa all'esperimento CMS a LHC ed è esperto nelle tecniche statistiche di analisi dati agli acceleratori. È Principal Investigator del progetto MineHEP presso l'Università di Firenze.

**Vitaliano Ciulli** - vitaliano.ciulli@unifi.it

Vitaliano Ciulli è Professore Associato in Fisica presso l'Università degli Studi di Firenze e Vicedirettore del Dipartimento di Fisica ed Astronomia. Partecipa all'esperimento CMS a LHC per il quale è stato coordinatore del gruppo dei generatori di eventi Monte Carlo.