DOI: 10.26314/GARR-Conf17-proceedings-01

Una "Big Data Open Platform" italiana per la ricerca e l'innovazione

Roberta Turra

Cineca

Abstract. Cineca, l'infrastruttura di supercalcolo più importante d'Italia, ha avviato un processo di sviluppo per la realizzazione di una piattaforma per la gestione ed elaborazione di grandi moli di dati per la ricerca scientifica e l'innovazione industriale. La nuova piattaforma conta sulle più avanzate risorse di calcolo e archiviazione, e sulla collaborazione delle comunità scientifiche che già raccolgono grandi quantità di dati da sensori e dispositivi e ne producono di nuovi attraverso le simulazioni computazionali.
La piattaforma supporta gli scienziati nella gestione dei dati durante tutto il ciclo di vita del progetto e mette insieme diversi modelli di utilizzo. Un team di esperti aiuta i ricercatori a ottimizzare l'uso delle risorse attraverso lo sviluppo e la selezione di componenti hardware e software appropriate. La piattaforma del Cineca, utilizzata da un numero sempre crescente di progetti nel settore Big Data, è stata di recente riconosciuta come innovation space (i-Space) da parte della BDVA (Big Data Value Association).

Keywords. Big Data, Data Life Cycle, HPC, Simulazioni, Deep Learning

Introduzione

La crescita esponenziale di dati generati e raccolti in quasi tutti i campi di attività apre la strada a processi di innovazione e a scoperte scientifiche "data driven" che necessitano di supporto in termini di competenze, potenza di calcolo, strumenti e servizi.
In questo contesto il Consorzio interuniversitario Cineca ha avviato lo sviluppo di una piattaforma abilitante, mettendo a frutto sia una lunga tradizione in abito di gestione dati, ontologie, data mining e business intelligence, sia le riconosciute competenze in ambito di calcolo ad alte prestazioni. Queste ultime costituiscono, in effetti, un elemento caratterizzante e distintivo rispetto ad altre analoghe piattaforme per i "big data".

La strategia di sviluppo si basa sulla partecipazione a progetti finanziati e sull'attivazione di accordi di collaborazione e di ricerca congiunta con centri di rilevanza nazionale e consolidate comunità scientifiche per la raccolta dei requisiti, lo sviluppo e la validazione di strumenti, servizi e risorse ad hoc. Vista la sua natura trasversale e intrinsecamente complessa, questa attività è svolta coniugando diverse competenze e analizzando i singoli problemi con un approccio end-to-end in stretta collaborazione con gli utenti e/o i clienti finali. Consapevoli del fatto che non esistono soluzioni universali, l'approccio seguito è quello di analizzare i casi singolarmente cercando di individuare classi di soluzioni e intrecciando competenze orizzontali, di dominio e tecnologiche.

Per dare un quadro generale dell'approccio utilizzato, di seguito viene presentato il ciclo di vita del dato declinando la presentazione nel contesto della ricerca scientifica. Ven-

gono inoltre descritte le caratteristiche della piattaforma Big Data allo stato attuale di sviluppo e gli obiettivi verso cui è indirizzata la sua evoluzione.

1. Il ciclo di vita del dato

La gestione efficiente dei dati scientifici costituisce un aspetto cruciale per consentire la ricerca, l'analisi e l'utilizzo dei dati raccolti e/o generati in quanto ne assicura l'organizzazione, l'identificazione e la descrizione. Inoltre, una buona gestione del dato ne garantisce la qualità, la protezione, la condivisione, la riproducibilità, la conservazione per un uso nel lungo periodo e il riuso. Questo ruolo fondamentale è stato riconosciuto dalla Commissione Europea che richiede, per ogni progetto finanziato, un documento descrittivo dei dati generati e della loro gestione (data management plan).

Il ciclo di vita del dato è una rappresentazione ad alto livello dei passaggi e dei processi che sono coinvolti nella gestione del dato. È utile per identificare e pianificare tutte le operazioni che devono essere implementate. Ne esistono diverse versioni a seconda delle prassi in vigore in ciascun dominio e comunità scientifica. Un esempio di riferimento è quello fornito dal Data Observation Network for Earth (https://www.dataone.org/data-life-cycle) che definisce otto componenti: la pianificazione della gestione dati, la raccolta dati, la valutazione della qualità, la descrizione mediante metadati, l'archiviazione, l'identificazione, l'integrazione e l'analisi. In questo schema l'analisi è il fine ultimo della raccolta e gestione dati e i suoi risultati possono dare luogo a nuovi progetti e nuove raccolte dati.

All'interno del progetto EUDAT (European Collaborative Data Infrastructure – www.eudat.eu), la definizione in vigore è quella del ciclo di vita del dato della ricerca scientifica dell'UK Data Service (http://www.data-archive.ac.uk/create-manage/life-cycle) che vede nell'ordine: 1) la pianificazione e raccolta, 2) il trattamento (che comprende inserimento, controllo qualità, pulizia, anonimizzazione, descrizione e archiviazione), 3) l'analisi e la produzione di risultati, 4) la conservazione, 5) la condivisione, 6) il riuso (Figura 1). In questo schema l'analisi dati è una fase intermedia e l'accento è posto sulla condivisione e il riuso. Cineca, come membro di EUDAT, e grazie anche agli altri progetti e collaborazioni, mette a disposizione strumenti e servizi che coprono tutte le fasi della gestione e analisi dati.

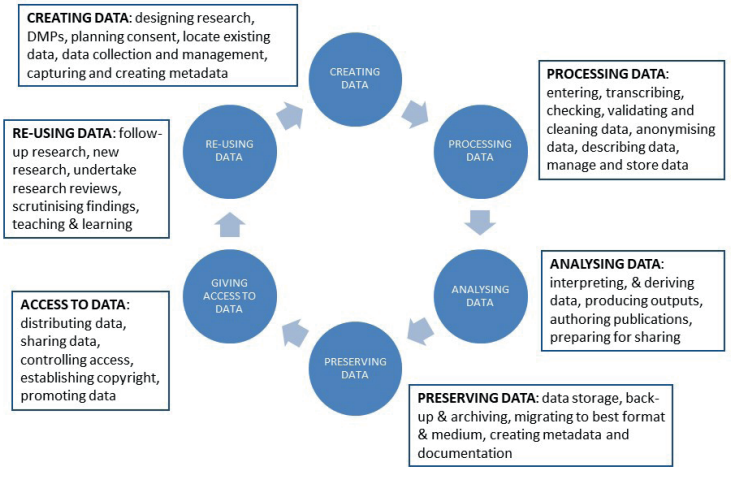


Figura 1
Ciclo di vita del dato della ricerca scientifica in uso in EUDAT
(fonte: UK Data Service)

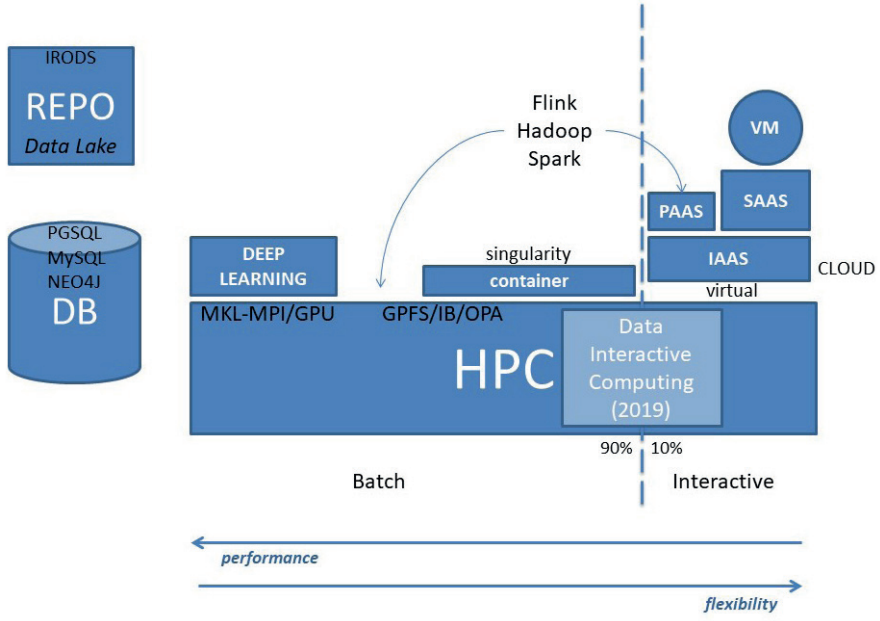
2. La piattaforma Big Data

Sviluppata a partire dal 2014, la piattaforma per i Big Data è utilizzata principalmente dalle comunità di bioinformatica e di calcolo industriale (per quanto riguarda la predictive maintenance e IND4.0), ma ospita anche numerosi progetti in altri ambiti: beni culturali e digital humanities, assicurazioni, media, energy e e-Government.

Attualmente, la piattaforma mette a disposizione risorse di calcolo che privilegiano la flessibilità e abilitano diversi modelli di utilizzo: da carichi computazionali intensi che fanno uso di GPU sull'infrastruttura HPC, in modalità batch, per gli utenti più esperti, all'uso più interattivo di risorse in modalità "container", e ottimizzate per obiettivi di analisi specifici, fino alla PaaS disponibile su risorse cloud in modalità interattiva e liberamente e autonomamente configurabile secondo le esigenze dell'utente (Figura 2).

Anche lo spazio di archiviazione offre diverse opzioni, dal data lake, uno spazio di storage per dati non strutturati, al database, uno spazio con gestione dei metadati, fino al servizio di repository, per dati strutturati.

Figura 2
Disegno logico
della piattaforma Big Data



Oltre agli strumenti e ai servizi di gestione, annotazione e analisi dati, la piattaforma mette a disposizione anche servizi di consulenza (indipendente dai fornitori di tecnologie), training e supporto utenti e competenze che vanno dall'ottimiz-zazione di codice alla data science e alla visualizzazione.

Eccellenza dell'infrastruttura, qualità dei servizi e trend crescente nel numero di progetti big data supportati, hanno consentito a questa piattaforma di ottenere la label di innovation space, i-Space, da parte della BDVA (Big Data Value Association). (http://www.bdva.eu/?q=node/790).

3. Le linee di sviluppo

L'evoluzione dei sistemi tende verso l'erogazione delle risorse, anche di supercalcolo, attraverso il paradigma del cloud computing, per garantire maggiore flessibilità agli utenti senza ridurre le prestazioni, per supportare diversi carichi computazionali e fornire ambienti isolati sicuri e interattivi. Tende inoltre verso il potenziamento dell'infrastruttura dedicata e ottimizzata per processi di deep learning per ridurre i tempi di addestramento delle reti neurali.

Il deep learning è infatti l'ambito dove maggiormente si sposano le necessità di dati e calcolo. Grandi quantità di dati devono essere disponibili per l'addestramento e grande potenza di calcolo deve essere disponibile per valorizzare l'enorme quantità di parametri (pesi) che il modello richiede. Il fatto di poter procedere in parallelo consente inoltre di esplorare diverse architetture di reti neurali simultaneamente e giungere in maniera tempestiva a identificare i modelli più efficaci.

Questa direzione di sviluppo, volta a rendere più efficienti (e quindi anche più efficaci) i processi di deep learning è originato dall'ambito big data classico, che fondamentalmente tenta di modellare il comportamento umano. Per quanto riguarda l'ambito del calcolo scientifico e la modellazione del mondo fisico, si possono porre due obiettivi:

- individuare sinergie tra l'approccio computazionale (simulazioni) e l'approccio data driven,
- sviluppare simulazioni che incorporano il deep learning / machine learning per aumentarne efficienza (accorciare il time-to-science) ed efficacia.

Nel primo caso, si tratta di accoppiare il dato simulato con quello reale, proveniente da sensori, per correggere il modello e migliorare i risultati della simulazione. Questo approccio trova applicazione anche in ambito industriale nella realizzazione di un digital twin sempre più fedele all'oggetto reale con la conseguente possibilità di modificarne il disegno e la produzione. In questo contesto è indicato l'uso del machine learning (non necessariamente del deep learning) per generare modelli empirici del funzionamento degli oggetti reali e prevederne i guasti.

Nel secondo caso si tratta invece di sostituire la parte di simulazione che assorbe maggiore potenza di calcolo con un modello di machine learning addestrato a riprodurre gli stessi risultati, date le condizioni di partenza, della simulazione. In fase di applicazione, un modello di machine learning non richiede infatti grandi potenze di calcolo e può accorciare i tempi e ridurre il consumo energetico.

4. Conclusioni

La necessità di gestire enormi e sempre crescenti quantità di dati, di diversa tipologia e in maniera tempestiva e l'opportunità di derivarne nuove chiavi di lettura della realtà sono aspetti trasversali che permeano sia le discipline scientifiche che il mondo produttivo. La condivisione dei dati, degli strumenti e delle best practices è fondamentale per l'innovazione, per nuove scoperte scientifiche e per affrontare le grandi sfide economico sociali. Per rispondere a questa esigenza è nata la piattaforma Big Data del Cineca, un ambiente che si arricchisce del contributo delle diverse comunità scientifiche cui dà supporto e che,

mettendo a disposizione gli strumenti più opportuni, consente una buona gestione di tutto il ciclo di vita del dato e favorisce la condivisione e il riuso dei dati stessi.
L'elemento distintivo rispetto ad altre piattaforme risiede nella potenza di calcolo e nelle competenze che consentono di sfruttarla al meglio, per questo motivo lo sviluppo strategico va nella direzione di ottimizzare i processi di deep learning, sia a beneficio delle applicazioni big data (data-driven), sia per inglobarli nei processi di simulazione.

Autori

Roberta Turra - r.turra@cineca.it

Roberta Turra coordina il team di Big Data Analytics del dipartimento HPC al Cineca. Si è laureata in Scienze Statistiche ed Economiche all'Università di Bologna nel 1991 e lavora al Cineca dal 1994 dove sviluppa applicazioni di data mining e text mining. Ha partecipato a numerosi progetti di ricerca finanziati a livello nazionale ed europeo e rappresenta Cineca presso la PPP BDVA (Big Data Value Association).