Bringing intelligence to the big data lake through smart metadata management

# Proposta di progetto

I *big data* sono una realtà consolidata a cui molte aziende sono oramai costrette ad adeguarsi. Le caratteristiche dei big data (le note 4V: volume, velocità, varietà e veracità [1]) richiedono di appoggiarsi su opportune architetture distribuite (*cluster computing*), siano esse *in-house* o nel *cloud*, le quali consentono la memorizzazione e l’elaborazione di grandi quantità di dati attraverso framework e software specializzato (ad esempio, Apache Hadoop, Apache Spark). Il *data lake* identifica l’area del file system distribuito nel quale confluiscono i dati che devono essere gestiti dall’azienda. In un contesto di big data, le fonti a cui il sistemi attinge sono spesso numerose, e i dati ingeriti presentano notevoli differenze in termine di natura (dati strutturati e non) e di dominio (dati aziendali, reperiti dai social network o da una rete di sensori). Inoltre, in fase di preparazione e analisi vengono spesso prodotte nuove versioni dei dataset esistenti, che vanno ad incrementare ulteriormente il volume dei dati da gestire.

Le comunità scientifiche e professionali concordano nella necessità di dotare il data lake di una struttura interna (in particolare, un insieme di zone) che consenta di separare i dataset rispetto al loro utilizzo all’interno del ciclo di vita del dato (dato grezzo, trasformazioni intermedie, dato integrato, ecc.). Tuttavia, questo non è sufficiente a supportare appieno il *data scientist* nel capire il contenuto dei dati e nel reperire i dataset di proprio interesse – in particolare nella zona riservata ai dati grezzi. A tal proposito diventa fondamentale la raccolta ed il mantenimento di metadati che tengano traccia di tutte le informazioni utili sui dataset, in termini di tipologia, provenienza, contenuto e tipologia di trasformazioni a cui sono soggetti. La meta-conoscenza che si va così a costruire può essere utilizzata per scopi molteplici. Ad esempio:

* Riconoscere automaticamente la struttura e la semantica dei dati acquisiti.
* Suggerire i dataset correlati ad un dataset di interesse.
* Automatizzare le procedure di preparazione del dato (*data wrangling*).
* Supportare le attività di analisi, sia a livello di query OLAP che di data mining (*auto machine learning*).

L’obiettivo di ricerca preposto consiste nel definire la struttura di riferimento per la meta-conoscenza e nel progettare un approccio in grado di popolarla, mantenerla e sfruttarla per gli scopi indicati, ponendo particolare interesse verso l’aspetto semantico dei dati (il quale risulta spesso trascurato negli approcci esistenti in letteratura.

# Attività di progetto

Nell’ambito dell’anno di lavoro, l’assegnista sarà coinvolto nelle seguenti attività:

1. Analisi dei possibili casi d’uso ed individuazione dei requisiti della meta-conoscenza
2. Analisi della letteratura e valutazione delle eventuali soluzioni esistenti
3. Definizione della struttura di riferimento per la meta-conoscenza necessaria alla gestione del dato rispetto agli obiettivi preposti e ai requisiti individuati
4. Definizione e prototipazione di un approccio per l’estrazione automatica dei metadati relativi ai dataset acquisiti
5. Definizione di tecniche capaci di sfruttare la meta-conoscenza acquisita in funzione degli obiettivi preposti
6. Studio della metafora OLAP applicabile nel contesto del progetto