

Titolo: Confronto tra Strumenti di Data Stream Processing su Sistemi Scale-up

Proposta

Il paradigma del Data Stream Processing studia l'elaborazione efficiente di computazioni che operano su sequenze infinite di dati trasmessi ad alta velocità (es. dati finanziari, reti di sensori, social media). Le elaborazioni compiute possono andare dal calcolo di statistiche (streaming analytics) ma anche elaborazioni più complesse come il processo di inferenza su modelli precedentemente addestrati oppure tecniche di learning continuo. L'elaborazione di computazioni di streaming avviene mediante sistemi in cui le applicazioni sono sviluppate come grafi dataflow di operatori, ciascuno rappresentante uno stadio intermedio di trasformazione dei dati.

Il lavoro di tesi verte sull'analisi di sistemi di Data Stream Processing ed il loro confronto prestazionale. Particolare enfasi viene posta nei sistemi per soluzioni *scale-up*, ovvero in grado di sfruttare singole macchine server equipaggiate con più CPU multicore. L'analisi prestazionale si baserà su benchmark di streaming esistenti (come ad esempio come quelli disponibili nel repository [StreamBenchmarks](#)) e considererà la libreria C++ [WindFlow](#) sviluppata dal Dipartimento di Informatica contro altre soluzioni disponibili allo stato dell'arte. Una di queste, di particolare interesse per il lavoro, è la libreria [FS2](#) in grado di fornire un supporto per il data streaming concorrente nel linguaggio di programmazione Scala.

Strumenti

Il candidato prenderà dimestichezza con il linguaggio di programmazione Scala la cui esecuzione a runtime è basata sulla Java Virtual Machine (JVM). In modo particolare sarà di interesse l'apprendimento delle API fornite da FS2 per implementare applicazioni di stream processing concorrenti su architetture multicore.

Obiettivo

Lo studente dovrà estendere il repository StreamBenchmarks (anche solo una parte da concordare) in modo da avere il giusto numero di applicazioni di streaming scritte per FS2 in modo da confrontare le prestazioni contro WindFlow sulle macchine messe a disposizione per la tesi. Ogni applicazione dovrà essere valutata come minimo sugli aspetti prestazionali quali banda di elaborazione (throughput) e latenza, così come già fatto per le applicazioni presenti nel repository (implementate per WindFlow e per altri framework di streaming esistenti ma non per FS2).

Prerequisiti

- Architetture degli Elaboratori e Sistemi Operativi
- Ingegneria del Software
- Una buona base di programmazione (con linguaggio Java/Scala)