Hands-on Apache Spark

Emiliano Pelella

AGENDA
Apache Spark

- Introduzione ad Apache Spark
  - Perché un nuovo paradigma?
  - Big Data framework
  - Architettura
  - Spark Core e stack
- RDD
  - Resilient Distribuited Dataset
  - DAG (Direct Acyclic Graph)
  - Trasformazioni ed azioni
  - Persistenza
- Spark SQL
  - Dataset e Dataframe
- Hands-on Spark

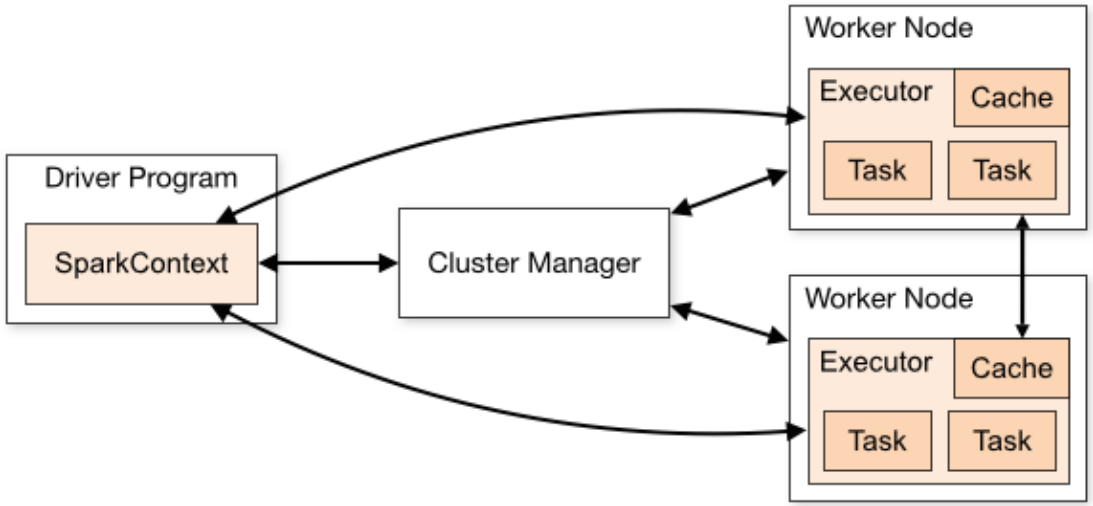
Perché un nuovo paradigma?
Limitazioni MapReduce

- Limitato: difficile implementare tutte le operazioni come combinazione di Map e Reduce.
- Nessun supporto nativo alle iterazioni. Queste vengono riprodotte tramite Map+Reduce+Scrittura su disco.
- Continue scritture su disco che comporta una perdita d'efficienza. Utilizzo della memoria RAM altamente inefficiente.
- Non pronto e utilizzabile per il data stream processing.

Apache Spark

Big Data Framework

Introduzione Apache Spark
Architettura



Architettura Master-Worker.
Lo SparkContext ha il compito di coordinare l'esecuzione di un'applicazione che gira su un cluster.

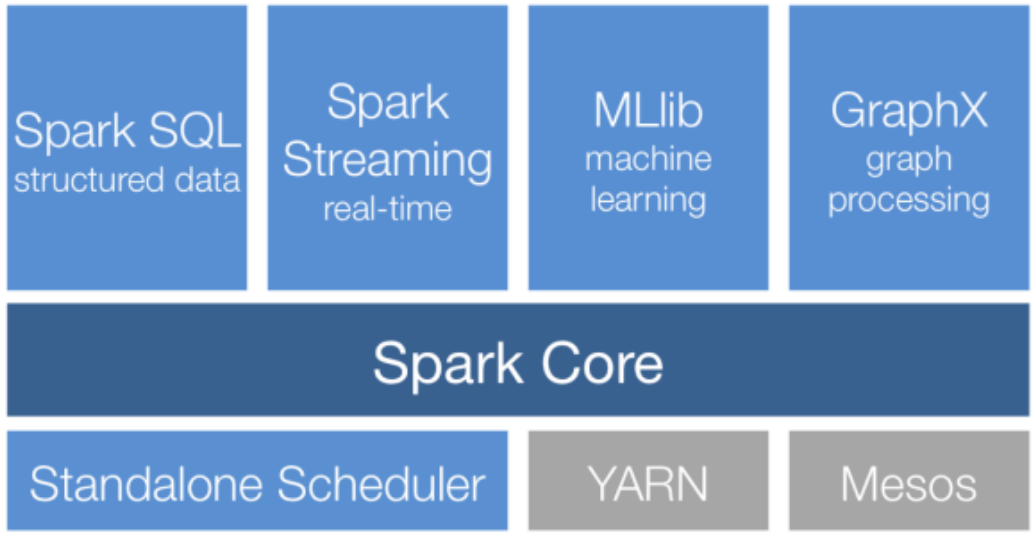
Ogni applicazione ha il suo executor che ha il compito di far girare i task in thread multipli.

Per lavorare su un cluster di macchine lo SparkContext ha bisogno connettersi ad un cluster manager (YARN, Mesos)

Una volta connesso, Spark acquisisce gli executors su ogni macchina ed invia il codice da eseguire ad ogni task.

Introduzione Apache Spark

Spark Core e lo stack



Spark SQL: libreria di Spark per il processamento di dati strutturati che permette l'esecuzione «on top» di query SQL.

Spark Streaming: estensione di Spark per l'elaborazione ed il processamento di dati streaming.

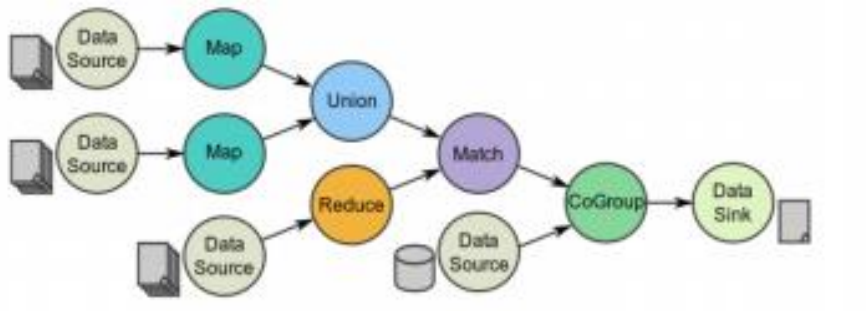
SparkML: libreria di machine learning.

GraphX: API per la computazione parallela di grafi.

Resilient Distribuited Dataset

Introduzione e DAG

- È una collezione di oggetti distribuiti su una serie di nodi di computazione in grado di elaborare in parallelo
- Un RDD è immutabile, partizionato, tollerante ai guasti ed elaborabile in parallelo.
- La descrizione del flusso dei dati viene fatta attraverso un grafo diretto aciclico.



Resilient Distribuited Dataset

Trasformazioni ed azioni

- Le trasformazioni sono operazioni «lazy» che possono creare nuovi RDD.
- Le azioni sono operazioni con il compito di far ritornare un valore al driver a seguito di una computazione.
- L'insieme di tutte le trasformazioni accumulate viene elaborato nel momento in cui applica un'azione sul dato.
- La persistenza dei dati può essere fatta su diversi livelli: Memoria, Disco

Spark SQL
Dataset e dataframe

- Libreria di Spark che estende le API per RDD per lavorare con dati strutturati.
- Permette l'analisi dei dati tramite l'esecuzione di query SQL.
- Due strutture dati utilizzabili: Dataset e Dataframe.
- Si possono generare tramite lettura da file (CSV, Parquet) o dalla trasformazione da RDD.