1. Spiegare quelli che sono i componenti principali di Hadoop

I componenti principali di hadoop sono diversi e hanno diverse funzionalità:

HDSF: Hadoop Distributed File system ( File System distribuito di hadoop), si basa su nameNode e dataNode i quali sono rispettivamente il server dove gira l’applicazione richiesta dal client e il server dove sono contenuti fisicamente i dati. I data node sono su stack diversi in modo da garantire la presenza dei dati, i quali sono ridondanti su piu cluster diversi.

 ZooKeeper fornisce un servizio di configurazione distribuita, un servizio di sincronizzazione e un registro dei nomi per i sistemi distribuiti.

È un sistema molto simile a un normale file system, ma reso molto affidabile dai servizi ridondanti offerti da ZooKeeper.

MapReduce: il suo compito consiste nel dividere in partizioni che andranno poi ad essere mappate e ridotte.

Hive consente di eseguire attività di riepilogo, query e analisi dei dati. Le query di Hive sono scritte in HiveQL, linguaggio di query simile a SQL. Ottiene in ingresso dati non strutturati e dopo averne dato una struttura utilizza le query sql

Yarn: ha come obiettivo l’ottimizzazione dell’utilizzo della memoria; consente così di suddividere i carichi tra tutti i nodi di un cluster; è l’evoluzione del map reduce.

1. Spiegare cos’è il “meccanismo” di MapReduce

Map Reduce: algoritmo che avviene in due fasi: la fase di mapping e la fase di reduce.

La prima per ogni record crea una coppia intermedia di chiave-valore e il risultato viene riordinato (nella fase di shuffle) e mandato ad un riduttore. Una volta che tutti gli input sono stati mappati si procede con la riduzione. Bisogna precisare che tutte queste operazioni vengono svolte in maniera parallela su diversi dispositivi.

1. Cos’è Pig Latin?

Pig: è un linguaggio di scripting, creato per velocizzare e snellire le operazioni di sviluppo del mapreduce, per il quale è necessario scrivere lunghe funzioni ecc

Principalmente viene utilizzato per data sampling ovvero riduzione dei dataset, analisi web ed elaborazioni per estrazione trasformazione e caricamenti nei datawarehouse

1. Immaginando di avere un file con il seguente contenuto  
   **Dear, Bear, River, Car, Car, River, Deer, Car ,Bear**mostrare in modo concettuale come andrebbe a lavorare MapReduce.  
   (NON FATE CODICE VOGLIO SOLAMENTE SAPERE QUALI SONO I PASSAGGI E COME VENGONO FATTI IN PSEUDOCODICE)

**Mapping**

**Dear 1, Bear 1, River 1, Car 1, Car 1, River 1, Deer 1, Car 1 ,Bear 1**

**Shuffle and sort**

**Dear 1, Bear 1,1 ; river 1,1; Car 1,1; Deer 1;**

**Reduce**

**Dear 1, Bear 2 ; river 2; Car 2; Deer 1;**

1. Cos’è ZLIB?

Zlib è una libreria opensource per la compressione e decompressione di file, garantendo l’integrità dei dati ed evitandone la perdita; ha 10 livelli di compressione da 0 non compresso a 9 massima compressione

1. Definizione e utilizzo del K-MEANS

K-means è un algoritmo di clustering, uno dei più diffusi per le elevate performance, che si basa sull’utilizzo dei centroidi, ovvero il punto medio delle osservazioni all’interno dello stesso cluster; è un algoritmo iterativo che, dopo aver inizializzato dei centroidi, ad ogni iterazione calcola la distanza minima tra un centroide e tutti gli elementi, ogni elemento viene assegnato al cluster il cui centroide ha la distanza minore; l’algoritmo si ferma quando non ci sono stati cambiamenti dall’iterazione precedente.

(NB in questo caso ho utilizzato come metrica la distanza minima, può essere utilizzata distanza media, massima, minima, euclidea, di ward ecc)

1. Quali sono i file utilizzabili in Hive? I file utilizzabili in avro sono molteplici:

SEQUENCEFILE Memorizza i dati in SEQUENCEFILE se i dati devono essere compressi.

ORC Il formato di file Ottimized Row Columnar (ORC) fornisce un modo estremamente efficiente per archiviare i dati Hive. È stato progettato per superare i limiti degli altri formati di file Hive.

PARQUET: Formato di archiviazione colonnare

AVRO Usa JSON per definire tipi di dati e protocolli e serializza i dati in un formato binario compatto.

File di testo