Test Hadoop – Michela Ginestri

1. **Spiegare quelli che sono i componenti principali di Hadoop +**
2. **Spiegare cos’è il “meccanismo” di MapReduce**

Hadoop è un software open-source per l’archiviazione di grandi quantità di dati divisi in cluster. Mette a disposizione diverse memorie virtuali per raccogliere, monitorare e modificare dati di grandi volumi e di qualsiasi tipologia. Hadoop è quindi un framework che utilizza diverse attività di memorizzazione e di analisi dati ed è per questo diviso in tre diverse componenti di base: HDFS, MapReduce e YARN.

HDFS: (Hadoop Distributed File System) è un sistema di elaborazione distribuita, è il framework di Hadoop che ha il compito di archiviare i set di dati che sono stati eseguiti in modo persistente. L’HDFS monitora i server in uso attraverso una funzionalità di checksum per verificare non ci siano blocchi danneggiati. Le operazioni principali che si possono fare su un HDFS sono quelli di lettura (:Il client HDFS contatta il namenode, che restituisce i vari datanode che possiedono una copia dei primi blocchi interrogandoli direttamente) e scrittura (:HDFS comunica la creazione di un file al namenode e, se autorizzato, quest’ultimo crea il record del file). HDFS offre però una memorizzazione di file poco affidabile, per questo Hadoop si basa anche sul MapReduce.

MapReduce: è il sistema di elaborazione parallela di dati che permette di eseguire un gran numero di elaborazioni contemporaneamente. Inizialmente le due funzioni “Map” e “Reduce” erano divise. Il primo (Map) è quello che prende un grande insieme di dati e lo converte in un altro insieme di dati, in cui i singoli elementi sono restituiti in forma chiave-valore. Dopo il lavoro di mappatura si ha il lavoro di riduzione (Reduce), che prende l'output da una mappa chiave-valore come input e combina queste coppie di dati in un insieme più piccolo. Il framework si occupa inoltre dei dettagli tecnici riguardanti il partizionamento dei dati, il controllo dei fallimenti dei nodi e la gestione della comunicazione tra le macchine, il tutto in modo automatico permettendo una certa semplicità nell’utilizzo.

Yarn: (Yet Another Resource Negotiator) è quello che rende definitivi i passaggi precedenti andando a lavorare sulla memoria, gestendo la schedulazione dei lavori e allocando le risorse del cluster alle applicazioni in esecuzione, avendo come obiettivo l’ottimizzazione dell’utilizzo della memoria all’interno del cluster in questione. E’ diviso in Resource Manager (che riceve le richieste dai client e alloca le risorse necessarie) ed Application Master (che monitora tutte le attività).

1. **Cos’è Pig Latin?**

Pig è uno strumento esterno di Hadoop realizzato per integrarsi con le sue tre componenti principali. Serve per accedere velocemente a dataset estesi. Utilizzando un linguaggio proprio dal nome PigLatin, crea programmi MapReduce rendendo l’originale linguaggio di programmazione più SQL-like. In altre parole, Pig fornisce un linguaggio per trasformare set di dati di grandi dimensioni in un linguaggio di programmazione di livello più semplice rispetto al Java iniziale.

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)

Per riuscire a lavorare in parallelo, si può dividere il contenuto in tre gruppi in modo da lavorare su diversi nodi. Ad esempio: Deer, Bear,River ----- Car Car River -----Dear, Car, Bear. Adesso, in fase di Mapping I tre gruppi vengono indicizzati, cioè ad ogni parola contenuta una sola volta verrà assegnato “1” . Si ha quindi Deer(1),Bear(1),River(1) per il primo gruppo, Car(1),Car(1),River(1) per il secondo e Dear(1),Car(1),Bear(1) per il terzo. A questo punto vengono smistate e raggruppate le parole coincidenti, ovvero la coppia chiave-valore con stessa chiave. In questo modo ogni chiave sarà univoca e con un elenco di valori corrispondi alla chiave stessa. In particolare si avrà Bear(1,1), Car(1,1,1) Deer(1,1), Dear(1),River(1,1). A questo punto si va alla parte di reducing, ovvero si raggruppano in Bear, (2), Car(3), Deer(2), River(2), Dear(1). Il numero tra parentesi conta quante volte i valori sono presenti nell’elenco iniziale. A questo punto uscirà l’output finale espresso come unico Bear,2 – Car,3 – Deer,2 – River,2 – Dear,1

1. **Cos’è ZLIB?**

ZLib è una libreria Hadoop utile per comprimere e decomprimere i file senza perdita di memoria. E’ divisa in livelli di prestazione, salendo aumenta il rapporto di compressione ma diminuisce la velocità di realizzazione di quest’ultima. ZLib quindi codifica i dati e li decomprime.

1. **Definizione e utilizzo del K-MEANS**

Il K-Means è un algoritmo utilizzato come tecnica di partizione, che genera dei cluster specifici e sferici. Il K-Means è di tipo iterativo e basato sul concetto di centroide (o punto medio). Permette in questo modo di suddividere un insieme di dati in k gruppi sulla base dei loro attributi.

1. **Quali sono i file utilizzabili in Hive?**

In Hive si possono utilizzare File testuali (quindi file di testo semplici), file ORC (Optimazed Row Columnar, quindi file di ottimizzati tipo riga-colonna), file di tipo Parquet (per metadati da incolonnare), RC File (file binari chiave-valore) e Avro (formato a righe write-intensive).