1. Spiegare quelli che sono i componenti principali di Hadoop

2. Spiegare cos'è il "meccanismo" di MapReduce

3. Cos'è Pig Latin?

4. 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)

5. Cos'è ZLIB?

6. Definizione e utilizzo del K-MEANS

7. Quali sono i file utilizzabili in Hive?

RISPOSTE

1) Hadoop è un framework per scrivere applicazioni che elaborano grandi quantità di dati in parallelo, usando cluster di grandi dimensioni (migliaia di nodi) e assicurando una elevata affidabilità e disponibilità (fault-tolerant).
Per garantire queste caratteristiche, Hadoop usa macrosistemi come **HDFS**, un file system distribuito, progettato per immagazzinare grandi quantità di dati, organizzati in file di grandissime dimensioni.
Un'altra componente fondamentale è Map Reduce, che è la parte di elaborazione dei dati.
Hadoop offre librerie che permette suddivisione dei dati da elaborare direttamente sui nodi, quindi i dati sono subito disponibili. Il framework garantisce affidabilità, in quanto i problemi (esempio file corrotto o assenza di corrente) sono risolti a livello applicativo (software), invece che usare hardware esterno.
Un'altra caratteristica di Hadoop è la scalabilità che è realizzabile semplicemente aggiungendo nodi al cluster.
Un **Cluster** è un insieme di computer connessi tramite rete. Lo scopo di un cluster è di permettere l'elaborazione complessa di un dati gestita su più computer. I cluster Hadoop consistono di 3 parti: il **master node** è responsabile della memorizzazione dei dati in HDFS e dell'esecuzione del calcolo parallelo; è formato da un NameNode, un NameNode secondario e un JobTracker. I primi tengono traccia delle informazioni sui file e ne fa un backup, il JobTracker monitora il calcolo parallelo con MapReduce.
I **Worker Node** sono responsabili dell'esecuzione dei calcoli.
I **Client Node** sono responsabili del caricamento dei dati e del recupero dei risultati.

MapReduce serve per distribuire i dati sul cluster**:** divide i dati in partizioni, che vengono poi mappati (trasformati) e ridotti (aggregati). Inoltre, fornisce un meccanismo di gestione dei guasti.

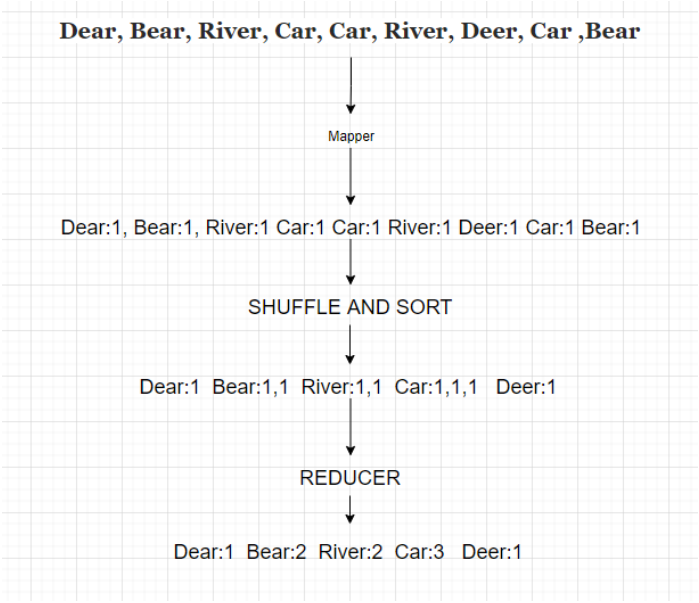
Un'altra componente fondamentale di Hadoop è **YARN**, che si occupa della gestione delle risorse, quindi del loro allocamento e dello scheduling dei vari jobs nei nodi del cluster. Permette quindi anche di ottimizzare l'uso della memoria e il funzionamento di Hadoop in generale.

2) Il meccanismo di MapReduce è uno delle principali componenti di Hadoop, che permette di suddividere le attività di calcolo in blocchi, distribuibili poi sul cluster grazie al HDFS.
MapReduce è diviso in 2 fasi: la fase di Map e la fase di Reduce. Nella fase di Map, i dati vengono mappati in coppie chiave-valore; nella fase di Reduce si implementa l'effettiva logica scelta, ad esempio un aggregamento e un sort.

3) **PIG** permette di eseguire programmi scritti in un linguaggio compatibile, che permette di semplificare il concetto di MapReduce. In pratica permette, attraverso un linguaggio più semplice, di creare funzionalità di MapReduce. Il linguaggio utilizzato si chiama Pig Latin ed è SQL like, ed è molto più espressivo del MapReduce, evitando inoltre la gestione della sincronia dei processi. I programmi Pig seguono un modello generale diviso in tre fasi:

- Caricamento dei dati dal disco
- Trasformazione dei dati
- Archiviazione per successiva elaborazione

4)



5) Zlib è una libreria utilizzata nei file ORC per la compressione-decompressione lossless dei file. È molto utilizzata nei servizi che necessitano modalità streaming dei dati. Ha 10 livelli di compressione, che hanno prestazioni di rapporto compressione/velocita diversi. Il livello 0 corrisponde a nessuno compressione, mentre il 9 ha il rapporto di compressione più alto, quindi con velocita di esecuzione minore.

6) Il K-mean è una tecnica di clusterizzazione dei dati, utilizzata in numerosi campi.
È un algoritmo iterattivo che raffina la suddivisione dei gruppi ad ogni iterazione.
Il concetto di base dell'algoritmo è abbastanza semplice: si inizia inserendo k centroidi nel grafico dei dati; si assegna ogni oggetto al cluster con distanza minore dal centroide. Si ricalcola il punto medio per ogni cluster e si sposta qui il centroide del cluster. Si ripetono i passaggi in modo iterativo e ci si ferma quando non si hanno più cambiamenti o se si raggiunge il limite massimo di iterazioni.

7) Con HIVE è possibile utilizzare moltissimi formati di dati, oltre ai più comuni CSV, file testuali, JSON, XML etc.

- ORC file: Optimazed Row Colonna, quindi una matrice (una tabella normale di un sql) che permette operazioni di compressione.
- Parquet è di tipo colonnare, quindi una riga e moltissime colonne
- RC file è simile a ORC file, ma quando eseguo una query, questo evita di recuperare le colonne che non servono. Se faccio una query con select nome, RC considera solo la colonna nome.
- AVRO è nato per Hadoop e minimizza lo spazio di archiviazione dei dati. All'interno di questo tipo di file esista una porzione di intestazione e un di dati.