1. Spiegare quelli che sono i componenti principali di Hadoop

## Apache Hadoop è un framework per applicazioni distribuite, utilizzato per l’archiviazione di dati strutturati, semi-strutturati e non strutturati. Le sue componenti principali sono un sistema di elaborazione distribuita, ovvero l’Hadoop Distributed File System (HDFS), e un file di archivio. L’HDFS ha come obbiettivo quello di assicurare sempre la disponibilità dei dati, sia in lettura che in scrittura, tramite un continuo monitoraggio dei server, effettuato utilizzando algoritmi di checksum.

## Componenti fondamentali di Hadoop sono anche il meccanismo di MapReduce e il modulo Yarn, il quale lima i punti deboli del MapReduce, migliorando la gestione delle risorse del sistema.

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

## Il Map Reduce è un sistema di elaborazione parallela di grosse moli di dati che, in modalità batch, raccoglie, accoppia e quindi riduce la quantità di dati. In particolare, l’operazione di Map trasforma i dati in input in delle coppie intermedie chiave-valore. Queste vengono poi raccolte dalla funzione Reduce che le ordina per chiave e raggruppa i valori associati a una stessa chiave.

1. Cos’è Pig Latin?

PigLatin è un linguaggio di scripting usato da Apache Pig per analizzare i dati nel contesto Hadoop. E’ un linguaggio testuale, basato su una serie di comandi che si prestano ad attività di amministrazione e che permette l’accesso veloce a dataset estesi e di effettuare su essi elaborazioni articolate.

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)

Il file in input viene inizialmente diviso in gruppi più piccoli, ad esempio da tre elementi, distribuendo così il lavoro su più map nodes. Ad esempio si ottengono i seguenti gruppi: {(dear, bear, river), (car, car, river), (deer, car, bear)}.

In seguito inizia la fase di Mapping. In ogni gruppo, ad ogni elemento verrà associato un valore di default pari a 1 (numero di occorrenze dell’elemento), creando una lista di coppie chiave-valore, in cui la chiave è l’elemento stesso: {[(dear:1), (bear:1), (river:1)], [(car:1), (car:1), (river:1)], [(deer:1), (car:1), (bear:1)]}

A questo punto, tramite un processo di shuffling, tutte le coppie individuate dalla stessa chiave vengono raggruppate, generando una lista di coppie caratterizzate da una chiave univoca e una lista di valori pari al numero delle occorrenze: {[dear:1], [bear:1,1], [river:1,1], [car: 1, 1, 1], [deer:1]}.

A questo punto il processo di Reduce conterà per ogni chiave il numero di valori presenti nella lista, generando in output delle coppie del tipo chiave-somma dei valori: {[dear:1], [bear:2], [river:2], [car:3], [deer:1]}

Alla fine le coppie key-value sono raccolte nell’output finale.

1. Cos’è ZLIB?

ZLIB è una libreria pensata per effettuare compressione e decompressione di dati senza perdita.

Permette di usare dieci livelli di compressione da 0 a 9, caratterizzati da diversi livelli di rapporto di compressione e velocità di compressione. In particolare, il livello 0 rappresenta l’assenza di compressione, 9 il massimo livello di compressione ma con minima velocità. In generale il livello di compressione di default è 6.

1. Definizione e utilizzo del K-MEANS

Il K-means è una tecnica di clusterizzazione basata su un algoritmo iterativo che permette di partizionare in gruppi un insieme di oggetti sulla base dei loro attributi. Ad ogni step l’algoritmo assegna ogni oggetto al cluster il cui centroide ha la distanza euclidea più bassa, raffinando la suddivisione degli oggetti a ogni ciclo.

1. Quali sono i file utilizzabili in Hive?

Hive supporta diversi formati, permettendo di leggere diverse tipologie di files.

I file testuali, ovvero file di testo semplici e leggibili per righe; gli ORC file, file in struttura riga colonna ottimizzata; i file Parquet, file con approccio colonnare, tipicamente usati in framework per big data; file RC, file binari in forma chiave-valore, utitlizzati per ottimizzare le analisi svolte tramite meccanismo MapReduce; file Avro, costituiti da una parte di intestazione in formato JSON e una di dati in struttura a righe.