Laboratorio Big Data and Reasoning

Teoria di Mazzotta

Che cos'è hadoop?

Hadoop è un ecosistema che permette di gestire task da svolgere in parallelo su un cluster di macchine.
Viene usato quando ci sono grosse quantità di dati da gestire. La cosa principale è che puoi usare hardware economico per scalare
le applicazioni.

Cos'è HDFS

HDFS significa Hadoop Distributed File System ed è un file system distribuito che permette di gestire file di grandi dimensioni. I file vengono divisi in blocchi e questi blocchi vengono replicati in modo da garantire la tolleranza ai guasti.

Ci sono nodi:

- Namenode: gestisce i metadati e coordina i datanode
- Datanode: gestisce i blocchi di dati e risponde alle richieste di lettura e scrittura

Cos'è YARN

Yet Another Resource Negotiator. E' un framework che permette di gestire le risorse di un cluster e di schedulare i task da eseguire.

Ci sono vari elementi che compongono YARN:

- ResourceManager: gestisce le risorse del cluster
- Container: astrazione che rappresenta le risorse di un nodo come CPU, RAM, disco e rete
- ApplicationMaster: gestisce le risorse di un'applicazione e coordina l'esecuzione delle applicazioni client.
- NodeManager: gestisce le risorse di un nodo e risponde alle richieste del ResourceManager. E' un servizio Slave

Cos'è HIVE?

Hive è un framework che salva tabelle come file nel hadoop file system. Supporta diversi formati per indicizzare le righe.
Usa database relazionali per salvare i metadati degli schemi e delle tabelle e puoi usare linguaggoi simil SQL

Cos'è HBASE

HBase è un database NoSQL che permette di salvare dati strutturati in tabelle direttamente nel hdfs. I dati sono distribuiti in **region server** ed è comodo per applicazinoi mapreduce

Cos'è SQOOP

Un tool di ETL che permette di
importare i dati nel RDBMS dall'HDFS e viceversa. Utilizza mapreduce per salvare i dati.

Come funzionano le letture nell'HDFS?

Quando si vuole leggere dal file system ci sono i seguenti passaggi:

1. Il client chiama il metodo **open** sul filesystem
2. Il filesystem chiede al namenode dove si trova il blocco di dati
3. Il namenode ritorna la lista di datanode che hanno una copia del blocco
4. I datanode sono ordinati in base alla metrica di distanza
5. Un inputstream FSDataInputStream viene creato
6. Mentre il client chiama il metodo **read**, l'inputstream chiede al datanode i dati del blocco e li ritorna al client
7. Quando i blocchi sono consumati l'inputstream chiede al namenode i prossimi blocchi

Come funzionano le scritture nell'HDFS?

Ancora, passaggi:

1. Il client chiama il metodo **create** sul filesystem
2. Il filesystem chiede al di creare un file senza blocchi associati
3. Il namenode controlla che il file non esista, i permessi, ecc
4. Un FSDataOutputStream viene creato
5. Il client chiama il metodo **write** sull'outputstream e i dati vengono divisi in pacchetti salvati in DataQueue
6. La DataQueue è consumata da un DataStreamer

7. Il DataStreamer chiede al namenode dove salvare i blocchi
8. Il namenode ritorna la lista di datanode dove una replica del blocco deve essere salvata
9. Il DataStreamer scrive i blocchi nei datanode
10. Il Icient chiude l'outputstream
11. Il DataStreamer chiude la DataQueue e manda un segnale al namenode
12. Il namenode attende che la replicazione minimale sia raggiunta e ritorna il controllo al client

Cos'è un blocco?

Un blocco è un file di dimensione fissa che viene salvato in un datanode. La dimensione di default è 128MB ma può essere cambiata.

Cos'è un DataNode?

Un DataNode è un servizio che gestisce i blocchi di dati e risponde alle richieste di lettura e scrittura.

Cos'è un NameNode?

Un NameNode è un servizio che gestisce i metadati e coordina i DataNode.

Cos'è un Container?

Un container è un'astrazione che rappresenta le risorse di un nodo come CPU, RAM, disco e rete.

Tipi di dati di HIVE

- Primitivi: stringhe, int, float, boolean, date, timestamp
- Complessi: Sono costruiti in base ai primitivi e permettono il nesting: array, map, struct

Cos'è un database in HIVE?

Un'insieme di tabelle. In HDFS i database sono cartelle

Cos'è una tabella in HIVE?

Dati che condividono lo stesso schema e appartengono ad un daabase. In HDFS le tabelle sono cartelle dentro il database
, cioè una sottocartella

Cos'è una partizione in HIVE?

Una partizione è una suddivisione di una tabella in base ai valori di una o più colonne. Le partizioni sono salvate in cartelle

Cos'è un bucket in HIVE?

Colonne esistengi che dividono i dati in un numero fissato di file in base ad una funzione di hashing

Cos'è una view in HIVE?

Strutture dati logiche che vengono usate per fare query. Sono definite nel metastore. Se si fanno modifiche su queste tabelle non vengono salvate fisicamente e non influiscono sulle tabelle fisiche.

Cos'è un metastore in HIVE?

Un database relazionale che salva i metadati delle tabelle e degli schemi. Può essere un database MySQL o Derby.

Cos'è un fixed schema in HIVE?

E' una collection di dati che hanno le stesse colonne.

Differenza tra managed ed external in HIVE

- Managed: i dati sono salvati in HDFS nella cartella della tabella. Quando si droppa la tabella i dati vengono cancellati.
- External: i dati sono salvati in HDFS in una cartella esterna alla tabella. Quando si droppa la tabella i dati rimangono.

Formati di file in HIVE

- PlainText: Formato normale di testo e i dati non sono indicizzati
- Parquet: Formato compresso che funziona bene su HDFS.
  Dato una tabella di N colonne, le righe sono raggruppate in M gruppi di righe. I dati sono memorizzati in un formato simile a una matrice NxM. Per ogni gruppo di righe, le colonne sono memorizzate sequenzialmente insieme ai metadati della colonna. I metadati del file contengono le posizioni delle colonne. Questo formato supporta una lettura sequenziale veloce. I metadati seguono i dati: i scrittori appendono direttamente i metadati dopo i dati, mentre i lettori leggono prima i metadati e poi possono accedere facilmente ai dati.

- ORC: Formato indicizzato, usato per alcuni readere e writers specifici.
  ORC è un formato di file binario non leggibile dall'uomo. Contiene strisce da 250 MB, indipendenti tra loro, con dati di indice, dati di riga e un piè di pagina. Il piè di pagina del file descrive il contenuto; il numero di righe, i tipi di dati delle colonne, le statistiche e l'elenco delle strisce. Il postscript contiene informazioni per interpretare il file.

Cos'è un bucket in HIVE?

Un bucket è una divisione di una tabella in base ad una funzione di hashing. I bucket sono salvati in cartelle.

Cos'è HBASE?

HBase è un database NoSQL che permette di salvare dati strutturati in tabelle direttamente nel HDFS. I dati sono distribuiti in **region server** ed è comodo per applicazioni mapreduce. Viene usato spesso quando i dati sono eterogenei tra loro. Permette lo sharding dei dati e il versionamento.

Cos'è una colonna in HBASE?

Una colonna è una coppia del tipo *famiglia / qualificatore*

Una famiglia di colonne gruppa insieme un insieme di colonne. Tutte le colonne di una famiglia hanno lo stesso prefisso.

Una o più colonne formano una riga e ogni riga ha un identificatore univoco.

Un insieme di righe forma una tabella.

Un namespace è una collezione di tabelle.

Esempio:

- Name
- Surname
- Age
- City
- Province
- List
  - Skill represent a programming language and a confidence level

In formato HBASE sarebbe:

```
RowKey: 1
Column Family: Personal
    Name: John
    Surname: Doe
    Age: 30
    City: Rome
    Province: Rome
Column Family: Skills
    Skill: Java
    Skill: 5
    Skill: Python
    Skill: 3
```

Cos'è HMaster in HBASE?

Servizio Master di un cluster hbase. Assegna le regioni ai region server e gestisce le richieste di lettura e scrittura.

Cos'è un region server in HBASE?

Servizio Slave di un cluster hbase. Serve e gestisce le regioni.

Cos'è Zookeeper in HBASE?

Zookeeper è un servizio che permette di coordinare i servizi di un cluster. Sceglie il cluster master per mantenere i metadati del cluster.

Com'è l'architettura di Hbase

Le tabelle di HBASE sono divise in regioni. Le regioni salvano un sottoinsieme delle righe e sono hostate in un region server. Ogni regione è salvata in HFile diversi in base alle famiglie di colonne. Gli HFile sono salvati in HDFS. I RegionServer usano un Write-Ahead-Log per tenere traccia delle operazioni non salvate in modo permanente. I regionserver compattano diversi HFile

Cos'è la tabella hbase:meta

E' una tabella speciale che contiene i metadati delle regioni. Contiene le informazioni di tutte le regioni del cluster.

Struttura:

- rowkey: tablename, startkey, regionid
- info:regioninfo (contiene le informazioni della regione)
- info:server (contiene l'indirizzo del region server che contiene la regione)
- info:serverstartcode (contiene il timestamp dell'ultima volta che il region server è stato avviato)

Cos'è SQOOP?

E' un tool per fare trasferimenti efficienti di grandi quantità di dati tra hadoop e risorse esterne. Usa MapReduce per fare questi trasferimenti

Che cos'è Spark?

Spark è un framework che lavora su un cluster progettata per essere veloce ed eseguire le operazioni in **memoria ram**. Offre le API per vari linguaggi di programmazione. Gira su Hadoop. Facile da usare, utilizzabile con altre librerie ed estendibile

Com'è formata la Stack di Spark?

- SparkCore: Gestisce tutte le funzionalità
- SparkSQL: Leggere e scrivere da diverse fonti ed eseguire query SQL su esse
- SparkMLlib: Fornisce diversi modelli di machine learning, permette di creare pipeline per lavorare in modo unificato su più dataframe
- Spark Structured Streaming: Implementata sulla base di SparkSQL e permette di gestire flussi di dati in tempo reale
- GraphX: Permette di gestire grafi e fare operazioni su di essi

Workflow differenze tra Spark e MapReduce

- MapReduce: I dati intermedi vengono salvati sul disco delle macchine mapper e poi letti dalle macchine reducer. Alla fine del processo viene salvato l'output sulla memoria globale
- Spark: Non ci sono salvataggi sul disco, ma tutto avviene in memoria.

Quali sono i componenti di Spark?

- Spark Driver: Negozia le risorse, trasforma e schedula le operazioni sugli Executors
- Spark Session: Permette di leggere i dataset e leggere dati da diverse fonti.
- Cluster Manager: Gestisce le risorse del cluster
- Spark Executor: Esegue le task di spark sulle macchine worker

Flusso di esecuzione di Spark

1. Definizione dell'applicazione: Definire con le API di Spark cosa fare
2. Spark Job: Un job spark è ciò che viene spawnato quando va eseguita un'azione
3. Spark Stage: Dato un job, uno stage è un insieme di task che possono essere eseguiti in parallelo
4. Spark Task: Una singola unità di laovro assegnato ad un Executor Spark

Che cos'è un RDD

Significa Resilient Distributed Dataset. E' un tipo di dataset che viene condiviso in memoria sulle macchine del cluster

Che cos'è un DataFrame

E' un dataset che contiene righe e colonne. Può essere visto come una tabella di un database relazionale. Può essere creato da un RDD, da un file o da un database.

Perché è importante l'immutabilità dei dataframe?

Dato un dataframe, quando si effettua una modifica non viene modificato quello ma viene creato uno nuovo basandosi su quello originale. Viene materializzato solo quando c'è necessità.

Differenza tra Transformation e Action

Una trasformazione è una modifica al dataset (anche se non viene modificato in modo diretto). Narrow Dependencies e Wide Dependencies significano rispettivamente:

- Narrow: Le righe di output dipendono da una sola partizione di input
- Wide: Le righe di output dipendono da più partizioni di input

Un'azione è un'operazione che ritorna un risultato. Quando viene chiamata un'azione, vengono eseguite tutte le trasformazioni necessarie per arrivare al risultato.

Cosa significa che le trasformazioni sono lazy?

Significa che le trasformazioni vengono eseguite solo quando viene chiamata un'azione, e non vengono materializzate fino a quel momento.

Cosa fa spark submit?

Permette di lanciare il programma spark sul cluster.

Cos'è SparkML

SparkML è una libreria che permette di fare machine learning su Spark. Offre diversi modelli di machine learning e permette di creare pipeline per lavorare in modo unificato su più dataframe.

Cos'è la classe Transformer in SparkML

Una classe che permette di eseguire trasformazioni su un dataset, ritornandone uno nuovo. Cose tipo aggiungere colonne, rimuovere colonne, ecc.

Cos'è la classe Estimator in SparkML

Estimator è l'algoritmo di machine learning, che impara dai parametri di un dataset e ritorna un modello è che un Transformer.

Cos'è la classe Pipeline in SparkML

Una pipeline è un'insieme di Transformer ed Estimator che vengono eseguiti uno dopo l'altro

Cosa fa il VectorAssembler in SparkML

Il vector assembler trasforma un insieme di colonne in un unico vettore. E' un Transformer.
Ad esempio:

Colonna 1	Colonna 2	Colonna 3
1	2	3
4	5	6
7	8	9

Viene trasformato in:

Colonna 1	Colonna 2	Colonna 3	Vettore
1	2	3	[1,2,3]
4	5	6	[4,5,6]
7	8	9	[7,8,9]

Per usare i modelli serve PER FORZA avere una colonna che contiene le colonne da considerare per il training.

Come gestire le features categoriche in SparkML

Prima si enumerano i possibili valori di quella colonne e si associa un indice in base al valore. Poi si crea un vettore di dimensione #possibiliValori e si mette 1 nella posizione corrispondente al valore della colonna. [One Hot Encoding]

Si può allenare in parallelo Decision Tree, Random Forest ecc?

Si, si può fare con SparkML. Perché sono algoritmi che non dipendono da altri dati. Ogni worker si fa i suoi split per conto suo. Si prende lo split migliore alla fine.

Per il random forest stessa cosa, si fa il training in parallelo e poi si fa il voting.

E hyperparameter tuning?

Si può fare con SparkML.Un esempio è con la k-fold cross validation. Si divide il dataset in k parti e si usa una parte per il testing e le altre per il training. Si fa questo k volte e si calcola la media delle performance. Si fa questo per ogni combinazione di parametri e si sceglie quella con le performance migliori.