1TD069, Data Engineering 1

UPPSALA
UNIVERSITET

Assignment 2

MapReduce and Hadoop

Ludvig Westerholm

February 10, 2023

Contents

1 Task 1.1

1.1 a

I got 2 files in the output folder:

- _SUCCESS
- part-r-00000

When doing the "cat" command on "part-r-00000", a list of words appeared and a number next to the word. So part-r-00000 is probably a list of occurrences of each word in the text.

As for the "_SUCCESS"-file, it indicates that the completion of the job was successful.

1.2 b

The word 'Discovery' appeared 5 times.

1.3 c

The cluster is simulated in pseudo-distributed mode compared to local mode.

2 Task 1.2

2.1 a

core-site.xml is where you configure the hadoop core, such as IP and port for the namenode.

hdfs-site.xml is where you configure specific properties to the hdfs, for example block size and directories where the data to be stored.

2.2 b

The namenode manages the filesystem namespace and stores the file system tree which contains the metadata about all files in the tree. It also keeps the locations of the datanodes.

The datanode is where the data is stored and handles read/write requests from the namenode.

3 Task 1.3

3.1 a

- WordCount starts by finding all different words that are separated by a space.
- IntSumReducer counts the occurrence of the words found by the tokenizer in wordcount.

3.2 b

Basically HDFS are designed for big data workloads, whereas a local filesystem on a VM is better suited for smaller, less complex data storage and processing needs.

4 Task 1.5

4.1 a

Semi-structured data has some form of structure but does not necessarily conform to the strict structure of traditional, such as relational database. Tweets have a structure, like created_at, text etc but the content within those fields can be unstructured and free-form, like the text of a tweet. Therefore we can classify Twitter's JSON-format as semi-structured data.

4.2 b

Structured Query Language (SQL) databases are based on the relational database model and store data in tables with well-defined relationships between them. Some pros of using SQL databases are:

1. SQL databases are well-suited for structured data, where data can be easily organized into tables with defined columns and relationships.
2. SQL databases can very easily be scaled simply by adding more hardware, such as servers and or storage.

A consequence is that it can have a hard time when storing semi-structured or unstructured data, such as JSON-files.

Not only SQL (NoSQL) databases are designed to handle unstructured data and semi-structured and do not rely on fixed schema like SQL. Some pros of NoSQL are:

1. High flexibility and can easily handle a wide variety of data types and structures, like graph data and JSON-files.
2. Easy to scale as they scale horizontally which allows for easy and efficient distribution of data across multiple servers.

Some consequences are:

1. Lack of standardization is both the perk but at the same time can lead to data coming all different sizes and shapes which may make it difficult to integrate and query.
2. NoSQL databases may not enforce strict consistency which can lead to data inconsistencies and potential errors.

SQL databases are suitable fro finacial systems where accurate and consistent data is critical, whereas NoSQL are good for social media platforms, since they tend to generate much unstructured data.