VE472 — Methods and tools for big data

Project 1

Manuel — UM-JI (Summer 2021)

Goals of the project

- Work with Hadoop, Drill, and Spark
- Compare MapReduce and Spark
- Perform advanced data analysis on big data
- Develop presentations skills (slides + poster)

1 Introduction

After the great success of his new cinema concept your friend Reapor Rich decides to invest in a music platform. To ensure a new triumph he first wants to analyse a huge amount of songs and artists in order to come up with better suggestions and higher prediction accuracy than all his future competitors.

In this endeavour he calls his friends Krystor, Frank, and Simon to the rescue. The four of them decide to setup a Hadoop cluster running the usual MapReduce, but also Spark and Drill. As soon as their cluster is ready they start looking into the Million Song Dataset (MSD) [1].

When importing the data in Hadoop they realise the files are very small compared to an HDFS block and do not know how to handle the situation as it will (i) waste much space, and (ii) increase the amount of RAM needed by the namenode.

As they are stuck on this problem they come to you for advice. You recall homework 7 where several solutions to this kind of problem were investigated. After a bit you thinking you decide to apply the Avro approach and compact many small files into larger ones.

2 Tasks

The compaction being quickly completed you decide to write a simple program to read the Avro files and easily extract the information. Since the binary parts of the Avro files are in the `hdf5` file format you look into the two main Java libraries allowing to open it, namely hdf5-java and sis-jhdf5-java.

Before you have even finished pronouncing the name of the second library Simon has already grabbed the keyboard and typed `apt install libhd5-j` in the console. So you stop him and recall that it is important to first discuss what tasks need to be completed and check what differentiate the two libraries. You conclude by telling Simon that in the end you might even realise that yet another library is better suitable for the tasks...

As Reapor explains his idea it seems that the tasks split into two main categories: (i) simple database queries and (ii) more advanced data analysis. He also declares that he has some more ideas to really beat the competitors but he first needs to refine them and obtain preliminary information before he can share them with you.

Among the basic information that can be retrieved with database queries he wants to know the range of dates covered by the songs in the dataset, i.e. the age of the oldest and of the youngest songs.

When Reapor formulates his next request you all look at him, eyes wide open: find the hottest song that is the shortest and shows highest energy with lowest tempo. As he argues that it is for marketing

purpose, you still feel this is a weird request but anyway Drill should be able to handle that without any problem, so you just do it.

When seeing how easy this is Kristor and Frank also want to try their own queries, so they ask you to find the name of the album with the most tracks, as well as the name of the band who recorded the longest song.

Now Reapor looks both excited and impatient: he wants you to work on one of the special features of his new platform. While most music platforms suggest similar artists, he would like to determine the distance between two artists such as to not only propose similar songs with distance one but also provide more diverse recommendations.

At this stage everybody agrees that for the first part of this task it suffices to construct an adjacency matrix to model the graph, and for each node store the distance from the current song node, as well as a backpointer to it. As Kristor directly concludes that Dijkstra will do the job, he looks at you and say "You seem sceptical, is everything OK?" To what you laconically reply "Good luck to make Dijkstra parallel..."

After a bit of thinking you suggest to go with Breadth First Search (BFS) since it should be possible to easily parallelise it and even render it compatible with MapReduce. Besides there is no need to care about the weight of the edges because only similar songs are considered, i.e. each edge has weight one.

Now starts a heated discussion, MapReduce or Spark? Based on what you learnt in your *Big data* course you feel Spark should be a better option as it will not lead to much input/output compared to MapReduce. As a result, to know how large the gap is between the two you decide to implement both and time how much faster Spark is compared to MapReduce.

Soon Reapor's phone rings and he leaves you after having quietly explained that it was investors contacting him about his platform. The four of you are now left with all his recommendations and a Hadoop cluster. Lets start!

While you are all busy working, Reapor feels a bit bad: this is his idea, you are all working for him, and he is unable to help you. He really regrets his credit overloading for his last semester in JI, without that he could have taken VE472 and be of a real help. Simon makes fun of him, saying that he is a real boss: not able to do anything aside of asking others to work!

Frank, more constructively notices: "Well guys, we can do the best work ever, if nobody can present it well and sell it, then it will just be a waste... So Reapor, why don't you work on your presentations skills, and prepare something that will blow up the mind of your investors?"

Feeling useful again Reapor starts thinking of the best way to impress his audience, i.e. the investors. Clearly he should put the emphasis on the quality of the work but also show them the superiority, flexibility, and range of application on the solutions. The goal should not be to loose the audience with too advanced unimportant technical details, but rather to ensure the global approach is clear and easy to follow. If the investors really want to get more information on advanced technical details they can do it at the end of the presentation during the Q&A time.

Once happy with his content choice Reapor thinks of the best way to package the work of the team. Not exactly able to choose between slides and poster he decides to prepare both. The poster will be a self-explanatory A1-page that mainly emphasises the significance of the work, with a short description of the methodology. Its goal is to stir the interest of the audience and encourage it to spend time on the more advanced technical part as they see a clear benefit in doing so. The slides of their side can provide

more details without being too advanced as initially decided, their goal also being to show the quality and significance of the work without drowning the investors.

Being a perfectionist Reapor wants his slides and poster to look professional. Therefore this directly rules out Microsoft PowerPoint and similar software putting the emphasis on useless effects that distract the audience from the content of the presentation. A logical choice is hence to use LaTeX.

After some research the best option appears to be Beamer for the slides and Beamerposter for the poster. To ensure the highest possible quality for his work he starts by reading section 5, "Guidelines for creating presentations", from the Beamer user guide. As figures and tables are of a major importance for both slides and poster, he also refers to section 7, "Guidelines on Graphics", of the TikZ & PGF manual, and section 2, "The layout of formal tables", of the Booktabs documentation.

Knowing how demanding the tasks is and how picky the investors are, nobody procrastinates. All immediately start working hard, with the promise of a future success in mind.

References

[1] Thierry Bertin-Mahieux et al. "The Million Song Dataset". In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*. 2011.