Project 3 : Data Analysis using Hadoop Map/Reduce

Done by Team-19:
Abhiram Vempati(1001843957)
Sankarshana Harish Rao (1001846315)

Overall status

The code was developed locally and has been moved, executed on expanse system. In this project, we have used Expanse to run the code instead of local Hadoop framework.

We have used 3 mappers and 3 reducers to generate the output. Mappers function maps the input data and produces intermediate data in the form of <key,value> pairs. This data is inputted to Reducers. Reducers will check if same person has acted and directed in the same title type and next, reducers will check if the title belongs to the 3 selected genres. The filtered output is then written to output file. Since there are 3 reducers, there will be 3 output files.

Analysis

We have done the analysis in three ways, 3 mappers/1 Reducer, 3 mappers/3 reducers, and 3 mappers/5 reducers. Based on the running times, we can state that for our logic 3M/3R took lesser time to compute the result and increasing the number of reducers to 5 did not reduce the compute time.

Following are the screenshots of the results running times,

3M/1R

IMDb Analysis using Hadoop Map Reduce

3M 1R

totalLines - 37699

- `part-r-00000` - 37699

Acted and Directed (Genre - `Romance` , `Comedy` , `Action`)

- `tvMovie` - 1299
- `movie` - 10259
- `short` - 26150

Time	Duration
startTime	21/12/09 00:41:20
endTime	21/12/09 00:43:10
totalTime	0d0h1m50s

Project 3 : Data Analysis using Hadoop Map/Reduce

Done by Team-19:
Abhiram Vempati(1001843957)
Sankarshana Harish Rao (1001846315)

3M/3R

3M3R

totalLines - 37699

- `part-r-00000` - 12426
- `part-r-00001` - 12565
- `part-r-00002` - 12708

Acted and Directed (Genre - `Romance` , `Comedy` , `Action`)

- `tvMovie` - 1299
- `movie` - 10259
- `short` - 26150

Time	Duration
startTime	21/12/09 00:56:06
endTime	21/12/09 00:57:38
totalTime	0d0h1m32s

3M/5R

3M5R

totalLines - 37699

- `part-r-00000` - 7631
- `part-r-00001` - 7522
- `part-r-00002` - 7564
- `part-r-00003` - 7396
- `part-r-00004` - 7586

Acted and Directed (Genre - `Romance` , `Comedy` , `Action`)

- `tvMovie` - 1299
- `movie` - 10259
- `short` - 26150

Time	Duration
startTime	21/12/09 01:10:23
endTime	21/12/09 01:12:05
totalTime	0d0h1m42s

Project 3 : Data Analysis using Hadoop Map/Reduce

Done by Team-19:
Abhiram Vempati(1001843957)
Sankarshana Harish Rao (1001846315)

Methods

1) Main Method – Main method sets the input split size and initializes the mapper instances, sets the input files into the mappers. It also configures the job and sets the memory settings for the mappers and reducers. It also manages the number of reducers and the output paths.

2) Map – There are 3 mappers classes. ActorMapper will add the actor's name and pick Title ID as key. It also sends the name and ID of the actor in the value part of the map. DirectorMapper will add the director's name along with the title ID as the key. TitleMapper will add the title year, genre. Some amount of sanity check of the data is done here to eliminate '//N' from the the input file. The string will be the value in <key,value> pair.

3) Reducer – Reducer method will add the director names into director list, actor names into actor list. Next, we check if an element in directors list is present in actors list. On the filtered data, reducer will check if the title genre belongs to the 3 selected genres. We have selected Romance, action, and comedy as genres in this project. If a record satisfies all the three conditions, then it written to output file. It also filters based on the selected title types.

Files Descriptions

1) IMDbActDir.build -
2) IMDbActDir.distr.run -
3) IMDbActDir.java – Contains class declaration of IMDbActDir class. This class contains the subclasses for the above described mappers and reducer. The IMDBActDir class also contains the main function which configures the job and handles the input and output paths.
4) IMDbActDir.distr.out – Contains the output written by the reducer function
5) IMDB_Datasets – IMDB dataset that is provided we need to compute on.

Project 3 : Data Analysis using Hadoop Map/Reduce

Done by Team-19:
Abhiram Vempati(1001843957)
Sankarshana Harish Rao (1001846315)

Division of Labor

Both of team members individually explored the code and we came to a general understanding of the classes. We individually developed the methods and regularly met to discuss and share our findings.

Configuration Details

We have used Hadoop library for the following: -

1) To import Job, default mapper and reducer class .
2) Packages for handling multiple inputs -
3) Packages for handling input formats
4) Hash map for quick lookup if an element is present. (For example, if an actor's name is present in director's list.)

Additional Results

Number of tuples in output for each title type:-

1) TvMovie – 1299
2) movie - 10259
3) short - 26150