**COMP30770 Programming for Big Data**

**Extracting Trends in Movie Metadata**

**Project Report**

**Group Members**

|  |  |
| --- | --- |
| Daniel Fallon | 22393263 |
| Colin Wang | 22343536 |
| Harry McCormack | 22513539 |

**Code Link:** <https://github.com/Daniel7Fallon/BigDataProject>

**Section 1**

**The Movies Dataset**

This dataset integrates 45,000 movies from TMDB (metadata, cast/crew, keywords) with 26 million user ratings from GroupLens. It includes structured CSVs (e.g., movies\_metadata.csv, ratings.csv) and semi-structured JSON-embedded files (e.g., credits.csv, keywords.csv), enabling cross-domain analysis like recommendation systems, box office prediction, and genre/director impact studies.

Source: <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download>

**Volume:**

Size: 943.76 MB

Columns: 43

System Specs:

Some execution times from section 3.

**Variety:**

Dataset contains 7 files, ordered by descending size:

|  |  |
| --- | --- |
| ratings.csv | 692,921 KB |
| credits.csv | 185,467 KB |
| movies\_metadata.csv | 33,638 KB |
| keywords.csv | 6,086 KB |
| ratings\_small.csv | 2,382 KB |
| links.csv | 966 KB |
| links\_small.csv | 180 KB |

The dataset’s variety stems from structural and functional differences across files. Structured data like *ratings.csv* (flat numerical columns) contrasts sharply with semi-structured files like *credits.csv*, where nested JSON fields describe actors’ roles and crew departments.

These files can be combined to extract correlations between themes, cast, revenue and public reception. This is invaluable information to the film industry.

**Section 2**

**Objective:**

This project aims to analyse the resource requirements to extract trends and key insights into the film industry and film marketplace.

**Value:**

Companies within the film industry can use these insights to consult and guide them in their decisions of what projects to pursue.

Specifically, the application of Big Data processing is essential to this application as:

* Markets can be volatile, i.e. subject to sudden change.
* Movies are released to the global market at a staggering rate, multiple per day.
* Reviews of movies are written in their hundreds everyday. Websites like “Letterboxd”, a rapidly growing social media for reviewing films, is an example of how this practice is only becoming more common.

These three factors combined means that for the insights garnered to be relevant, this massive and growing dataset must be continually reprocessed. The quicker the turn over of the processing, the more relevant the insights.

**Section 3. Traditional Solution (2.5 pages)**

Language Used: Python 3

Steps:

1. Read in csv files to pandas dataframes for convenient manipulation

metadataDF = pd.read\_csv('TheMoviesDataset/movies\_metadata.csv', usecols=["id", "title", "budget", "genres", "popularity", "release\_date", "revenue", "runtime", "vote\_average", "vote\_count"])

2. Show correlation between how many movies actors are in and how well those movies are received

numFilmsCorrelation = cast\_stats.groupby('num\_films').agg({

    'vote\_average': 'mean', 'popularity': 'mean'}).reset\_index()

3.

Normally, in practice, before we develop the big data pipeline, we quickly prototype the processing logic on the same dataset (or its smaller version) first to test its feasibility and get its performance profile. The prototype should use no parallelism and can be any high-level programming language such as Shell, SQL, Python, Java, C++, etc. Please decompose your overall objective into several (roughly 3 to 6) small steps.

* [Note that each task should be directly translated to one or a few Shell or SQL statements or small code snippets in other programming languages.]
* Briefly introduce each step (2’)
* The key code [not all] of SQL or Shell or other single-threaded high-level programming language solutions should be presented here. (5’)
* The execution results / execution time / memory requirements should be presented here. (5’)

**Section 4. MapReduce Optimisation (2 pages)**

Please identify 1 or 2 most time-consuming steps in your Section 3 that can be optimised by big data programming paradigms: MapReduce. You are free to use either Hadoop MapReduce or Spark MapReduce (Spark Core API, NOT Spark SQL or Dataframe etc.)

* Explain why they can be optimised using MapReduce and present your expectations (e.g., reduce execution time by 2). (3’)
* Present MapReduce solution (3’) - Present MapReduce results. (3’) - Explain why the results match or deviate from your expectations. (3’ )