Case Study 3
MovieLens Analytics Pipeline

1. Aim:

To design and implement a data processing pipeline that analyses the MovieLens dataset using Apache Spark, providing insights through data integration, aggregation, and visualisation.

2. Objectives:

- Load and clean MovieLens datasets.
- Integrate data from multiple sources.
- Perform data transformations and aggregations.
- Calculate statistical metrics for movies and genres.
- Identify data anomalies and user activity patterns.
- Generate reports and visualisations for meaningful insights.

3. Datasets Involved:

- **Ratings Dataset:** Contains user ratings for movies.
- **Movies Dataset:** Includes movie details like title and genres.
- **Tags Dataset:** Contains user-generated tags for movies.
- **Links Dataset:** Provides external links for additional movie information.

4. Tools & Technologies:

- Apache Spark (Scala)
- Google Cloud Storage (GCS)
- Spark SQL
- Spark DataFrame API

5. Procedure:

Step 1: Environment Setup

- Configure Spark Session for local execution and Google Cloud Storage integration.

Step 2: Data Loading

- Load CSV datasets into DataFrames using Spark.

Step 3: Data Preparation

- Clean and transform data:
  - Format timestamps into dates.
  - Split genres into multiple rows.
  - Remove null values.

Step 4: Data Integration

- Join ratings, movies, tags, and links DataFrames based on movieId and userId.

Step 5: Statistical Computation

- Calculate movie statistics like average ratings, rating counts, and variance.

Step 6: Data Aggregation

- Aggregate data based on:
  - Genres: Average ratings and total rating counts.
  - Years: Yearly average ratings.
  - Users: Rating counts and averages per user.

Step 7: Anomaly Detection

- Identify movies with unusually high rating variance.

Step 8: Reporting and Visualisation

- Generate summary reports for:
  - Top-rated movies
  - Popular genres
  - Most active users
  - Yearly rating trends
- Plot and display:
  - Average ratings per year.
  - Rolling averages for ratings over time.

Step 9: Data Persistence

- Save processed and aggregated datasets to GCS in Parquet and JSON formats.

Here are the output images:

```
Loading Links dataset from: /Users/navadeep/Downloads/Work/Assignments/Datasets/movieDataset/link.csv
Displaying first 5 rows of Links dataset:
+-------+------+------+
|movieId|imdbId|tmdbId|
+-------+------+------+
|1      |114709|862   |
|2      |113497|8844  |
|3      |113228|15602 |
|4      |114885|31357 |
|5      |113041|11862 |
+-------+------+------+
only showing top 5 rows

Schema of Links dataset:
root
 |-- movieId: string (nullable = true)
 |-- imdbId: string (nullable = true)
 |-- tmdbId: string (nullable = true)

Links dataset count: 27278
```

```
Loading Tags dataset from: /Users/navadeep/Downloads/Work/Assignments/Datasets/movieDataset/tag.csv
Displaying first 5 rows of Tags dataset:
+------+-------+------------+-------------------+
|userId|movieId|tag         |timestamp          |
+------+-------+------------+-------------------+
|18    |4141   |Mark Waters |2009-04-24 18:19:40|
|65    |208    |dark hero   |2013-05-10 01:41:18|
|65    |353    |dark hero   |2013-05-10 01:41:19|
|65    |521    |noir thriller|2013-05-10 01:39:43|
|65    |592    |dark hero   |2013-05-10 01:41:18|
+------+-------+------------+-------------------+
only showing top 5 rows

Schema of Tags dataset:
root
 |-- userId: string (nullable = true)
 |-- movieId: string (nullable = true)
 |-- tag: string (nullable = true)
 |-- timestamp: string (nullable = true)

Tags dataset count: 465564
Cleaning tags data...
```

```
Loading Movies dataset from: /Users/navadeep/Downloads/Work/Assignments/Datasets/movieDataset/movie.csv
Displaying first 5 rows of Movies dataset:
+-------+---------------------------------+---------------------------------------+
|movieId|title                            |genres                                 |
+-------+---------------------------------+---------------------------------------+
|1      |Toy Story (1995)                 |Adventure|Animation|Children|Comedy|Fantasy|
|2      |Jumanji (1995)                   |Adventure|Children|Fantasy               |
|3      |Grumpier Old Men (1995)          |Comedy|Romance                         |
|4      |Waiting to Exhale (1995)         |Comedy|Drama|Romance                   |
|5      |Father of the Bride Part II (1995)|Comedy                                |
+-------+---------------------------------+---------------------------------------+
only showing top 5 rows

Schema of Movies dataset:
root
 |-- movieId: string (nullable = true)
 |-- title: string (nullable = true)
 |-- genres: string (nullable = true)

Movies dataset count: 27278
Transforming movies data...
```

```
Loading Ratings dataset from: /Users/navadeep/Downloads/Work/Assignments/Datasets/movieDataset/rating.csv
Displaying first 5 rows of Ratings dataset:
+------+-------+------+-------------------+
|userId|movieId|rating|timestamp          |
+------+-------+------+-------------------+
|1     |2      |3.5   |2005-04-02 23:53:47|
|1     |29     |3.5   |2005-04-02 23:31:16|
|1     |32     |3.5   |2005-04-02 23:33:39|
|1     |47     |3.5   |2005-04-02 23:32:07|
|1     |50     |3.5   |2005-04-02 23:29:40|
+------+-------+------+-------------------+
only showing top 5 rows

Schema of Ratings dataset:
root
 |-- userId: string (nullable = true)
 |-- movieId: string (nullable = true)
 |-- rating: string (nullable = true)
 |-- timestamp: string (nullable = true)

Ratings dataset count: 20000263
Preparing ratings data...
Raw Ratings:
+------+-------+------+-------------------+
|userId|movieId|rating|          timestamp|
+------+-------+------+-------------------+
|     1|      2|   3.5|2005-04-02 23:53:47|
|     1|     29|   3.5|2005-04-02 23:31:16|
|     1|     32|   3.5|2005-04-02 23:33:39|
|     1|     47|   3.5|2005-04-02 23:32:07|
|     1|     50|   3.5|2005-04-02 23:29:40|
+------+-------+------+-------------------+
only showing top 5 rows
```

```
Ratings:+------+-------+------+-----------------+-------------------+
|userId|movieId|rating|          timestamp|         ratingDate|
+------+-------+------+-----------------+-------------------+
|     1|      2|   3.5|2005-04-02 23:53:47|2005-04-02 23:53:47|
|     1|     29|   3.5|2005-04-02 23:31:16|2005-04-02 23:31:16|
|     1|     32|   3.5|2005-04-02 23:33:39|2005-04-02 23:33:39|
|     1|     47|   3.5|2005-04-02 23:32:07|2005-04-02 23:32:07|
|     1|     50|   3.5|2005-04-02 23:29:40|2005-04-02 23:29:40|
+------+-------+------+-----------------+-------------------+
only showing top 5 rows

+-------+---------------+--------------------+---------+
|movieId|          title|              genres|    genre|
+-------+---------------+--------------------+---------+
|      1|Toy Story (1995)|Adventure|Animati...|Adventure|
|      1|Toy Story (1995)|Adventure|Animati...|Animation|
|      1|Toy Story (1995)|Adventure|Animati...| Children|
|      1|Toy Story (1995)|Adventure|Animati...|   Comedy|
|      1|Toy Story (1995)|Adventure|Animati...|  Fantasy|
+-------+---------------+--------------------+---------+
only showing top 5 rows
```

```
+------+-------+-------------+-----------------+
|userId|movieId|          tag|        timestamp|
+------+-------+-------------+-----------------+
|    18|   4141|  Mark Waters|2009-04-24 18:19:40|
|    65|    208|    dark hero|2013-05-10 01:41:18|
|    65|    353|    dark hero|2013-05-10 01:41:19|
|    65|    521|noir thriller|2013-05-10 01:39:43|
|    65|    592|    dark hero|2013-05-10 01:41:18|
+------+-------+-------------+-----------------+
only showing top 5 rows

+-------+------+------+
|movieId|imdbId|tmdbId|
+-------+------+------+
|      1|114709|   862|
|      2|113497|  8844|
|      3|113228| 15602|
|      4|114885| 31357|
|      5|113041| 11862|
+-------+------+------+
only showing top 5 rows
```

Fig.ratings, movies, tags, links

```
Null check in Ratings - Column: movieId - Null Count: 0
Null check in Ratings - Column: userId - Null Count: 0
Null check in Movies - Column: movieId - Null Count: 0
Null check in Tags - Column: movieId - Null Count: 0
Null check in Tags - Column: userId - Null Count: 0
Null check in Links - Column: movieId - Null Count: 0
Integrating datasets...
Full Data:
+-------+------+------+-----------------+-----------------+-----------------+--------------------+--------+----+---------+------+------+
|movieId|userId|rating|        timestamp|       ratingDate|            title|              genres|   genre| tag|timestamp|imdbId|tmdbId|
+-------+------+------+-----------------+-----------------+-----------------+--------------------+--------+----+---------+------+------+
|   1035| 46736|     4|2008-10-03 23:14:18|2008-10-03 23:14:18|Sound of Music, T...|     Musical|Romance| Romance|null|     null| 59742| 15121|
|   1035| 46736|     4|2008-10-03 23:14:18|2008-10-03 23:14:18|Sound of Music, T...|     Musical|Romance| Musical|null|     null| 59742| 15121|
|   1563| 23312|     4|1999-10-16 11:42:02|1999-10-16 11:42:02|Dream With the Fi...|               Drama|   Drama|null|     null|119019| 47686|
|   2505| 11731|   4.5|2013-06-12 15:39:38|2013-06-12 15:39:38|         8MM (1999)|Drama|Mystery|Thr...|Thriller|null|     null|134273|  8224|
|   2505| 11731|   4.5|2013-06-12 15:39:38|2013-06-12 15:39:38|         8MM (1999)|Drama|Mystery|Thr...| Mystery|null|     null|134273|  8224|
+-------+------+------+-----------------+-----------------+-----------------+--------------------+--------+----+---------+------+------+
only showing top 5 rows

Saving data to: gs://task-dataset-bucket/FinalProject/CaseStudy3/movies_stats as parquet
```

Fig. Null checks and Integrated data

```
Generating reports and visualizations...
Aggregate bY Genre+-------+------+------+-----------------+-----------------+-----------------+--------------------+--------+----+---------+------+------+
|movieId|userId|rating|        timestamp|       ratingDate|            title|              genres|   genre| tag|timestamp|imdbId|tmdbId|
+-------+------+------+-----------------+-----------------+-----------------+--------------------+--------+----+---------+------+------+
|   1035| 46736|     4|2008-10-03 23:14:18|2008-10-03 23:14:18|Sound of Music, T...|     Musical|Romance| Romance|null|     null| 59742| 15121|
|   1035| 46736|     4|2008-10-03 23:14:18|2008-10-03 23:14:18|Sound of Music, T...|     Musical|Romance| Musical|null|     null| 59742| 15121|
|   1563| 23312|     4|1999-10-16 11:42:02|1999-10-16 11:42:02|Dream With the Fi...|               Drama|   Drama|null|     null|119019| 47686|
|   2505| 11731|   4.5|2013-06-12 15:39:38|2013-06-12 15:39:38|         8MM (1999)|Drama|Mystery|Thr...|Thriller|null|     null|134273|  8224|
|   2505| 11731|   4.5|2013-06-12 15:39:38|2013-06-12 15:39:38|         8MM (1999)|Drama|Mystery|Thr...| Mystery|null|     null|134273|  8224|
+-------+------+------+-----------------+-----------------+-----------------+--------------------+--------+----+---------+------+------+
only showing top 5 rows
```

Fig. Aggregate by Genre

Data visualisation

genre	avgGenreRating	totalRatingsByGenre
Crime	3.678323176438156	3344526
Romance	3.545343637579382	3847993
Thriller	3.511530456827378	5392978
Adventure	3.504920857792138	4433475
Drama	3.678117982649218	8995560
War	3.8107847596768254	1061841
Documentary	3.741678966053692	250690
Fantasy	3.510150631148478	2143315
Mystery	3.668876331848703	1586892
Musical	3.5615689542275617	879656
Animation	3.6215837663516655	1156610
Film-Noir	3.966126460007793	220718
(no genres listed)	3.030848329048843	389
IMAX	3.6580310830139138	511083
Horror	3.282691755347552	1507559
Western	3.573077723892207	427449
Comedy	3.4298469564432756	7580391
Children	3.4111114082632055	1682640
Action	3.447358900144718	5688312
Sci-Fi	3.4428975218952584	3206402

userId	totalRatings	averageRating
101205	166	3.8433734939759034
104603	195	3.6564102564102563
121556	193	3.139896373056995
131450	94	3.723404255319149
131682	233	4.180257510729613
133809	781	3.9622279129321383
136186	263	2.9163498098859315
1436	610	3.2508196721311475
22596	59	3.593220338983051
23318	376	3.7247340425531914
39641	167	3.8323353293413174
40740	263	3.3802281368821294
42688	1045	4.143540669856459
46870	70	3.3142857142857145
47880	852	3.841549295774648
49755	456	3.1359649122807016
60733	232	2.997844827586207
77930	752	3.9202127659574466
85022	187	3.9331550802139037
89517	86	3.5232558139534884

only showing top 20 rows

movieId	title	genre	averageRating	ratingCount	ratingVariance
318	Shawshank Redemption, The (1994)	Drama	4.448998801985281	64273	0.5150114835839518
318	Shawshank Redemption, The (1994)	Crime	4.448998801985281	64273	0.5150114835839518
858	Godfather, The (1972)	Drama	4.363334531942947	41715	0.7063544653259511
858	Godfather, The (1972)	Crime	4.363334531942947	41715	0.7063544653259511
50	Usual Suspects, The (1995)	Mystery	4.334675130851533	47573	0.5733776150707832
50	Usual Suspects, The (1995)	Thriller	4.334675130851533	47573	0.5733776150707832
50	Usual Suspects, The (1995)	Crime	4.334675130851533	47573	0.5733776150707832
527	Schindler's List (1993)	Drama	4.3104387441814405	50485	0.6829569523376989
527	Schindler's List (1993)	War	4.3104387441814405	50485	0.6829569523376989
77658	Cosmos (1980)	Documentary	4.279013539651838	1034	1.022511707464129

year	yearlyAverage
2003	3.485253490840837
2007	3.4785849349864235
2015	3.52402006158835
2006	3.468486949207258
2013	3.6607698061081626
1997	3.6043686609177428
2014	3.6209940750533343
2004	3.4380688007194675
1996	3.5614778965200102
1998	3.5261758778350414
2012	3.6308968552591345
2009	3.528494320853906
2001	3.5474717613541142
2005	3.436573723100307
2000	3.582719338481802
2010	3.562078686833833
2011	3.5897459515607797
2008	3.5474812625608765
1999	3.611680198668057
2002	3.501837202738953

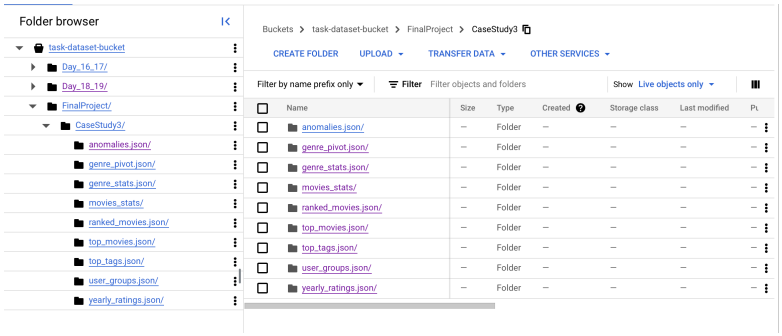
only showing top 20 rows

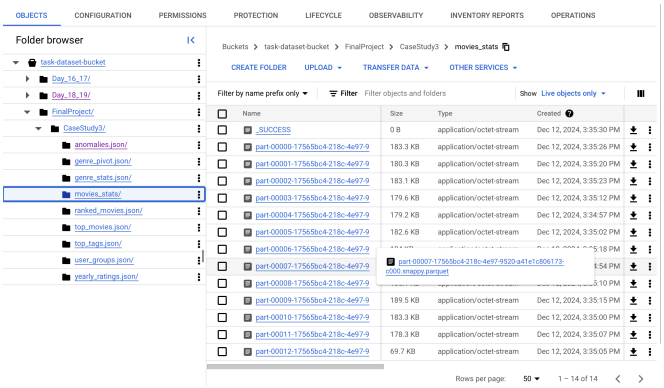
year	rollingAverage
1995	3.7
1996	3.5614778965200102
1997	3.6043686609177428
1998	3.5261758778350414
1999	3.611680198668057
2000	3.582719338481802
2001	3.5474717613541142
2002	3.501837202738953
2003	3.485253490840837
2004	3.4380688007194675
2005	3.436573723100307
2006	3.468486949207258
2007	3.4785849349864235
2008	3.5474812625608765
2009	3.528494320853906
2010	3.562078686833833
2011	3.5897459515607797
2012	3.6308968552591345
2013	3.6607698061081626
2014	3.6209940750533343

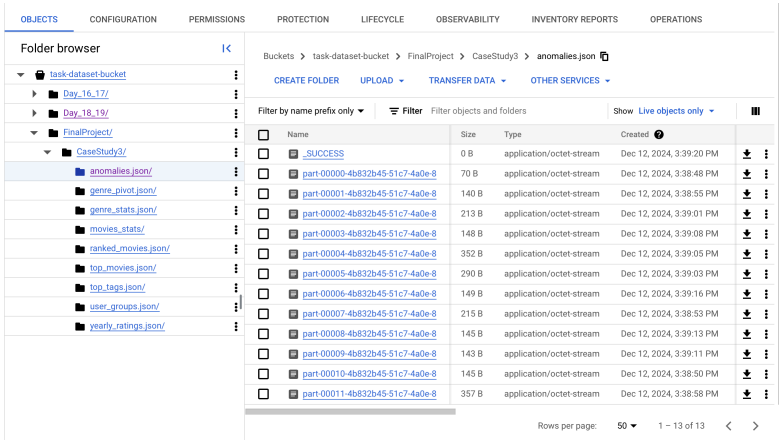
only showing top 20 rows

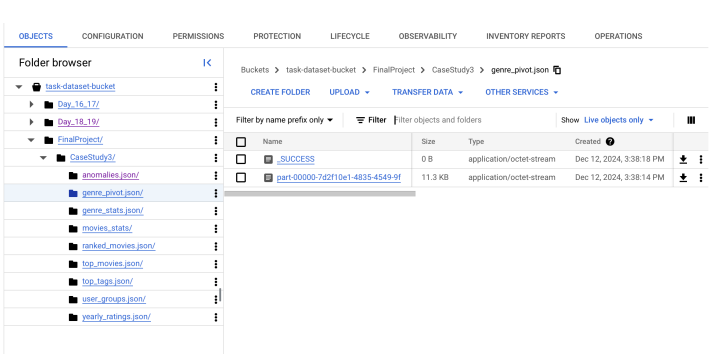
Reports and visualizations generated successfully!
Data Pipeline Execution Completed Successfully!

Google cloud storage









In the similar fashion, the hierarchy works.