A Parallel Recommender System Using A Collaborative Filtering Algorithm For Movie Recommender System

Projna Saha
School of Computer Science
Carleton University
projnasaha@cmail.carleton.ca

Types of Recommendation Systems

- **Content-Based Filtering:**

  Content-Based recommender system tries to guess the features or behavior of a user given the item's features, he/she reacts positively to [1].
- **Collaborative Filtering:**

  Collaborative does not need the features of the items to be given. Every user and item is described by a feature vector or embedding [2].

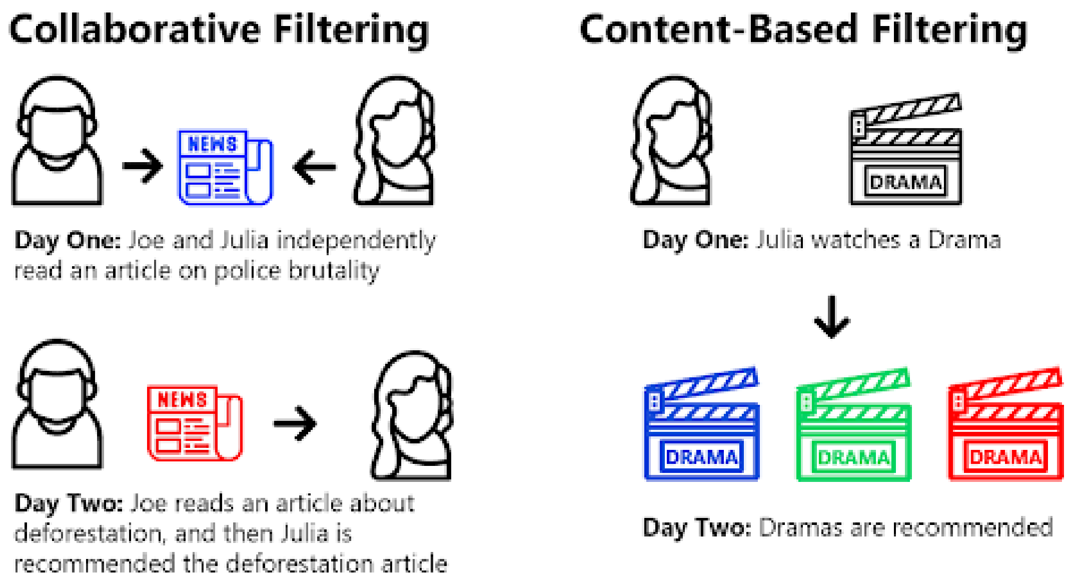


Image Source

Literature Review

Similarity Computation

- Cosine Vector (CV) Similarity [3]
- Pearson Correlation (PC) Similarity [4][5]
- Spearman Correlation (SC)[6]
- JacRA Similarity [7]

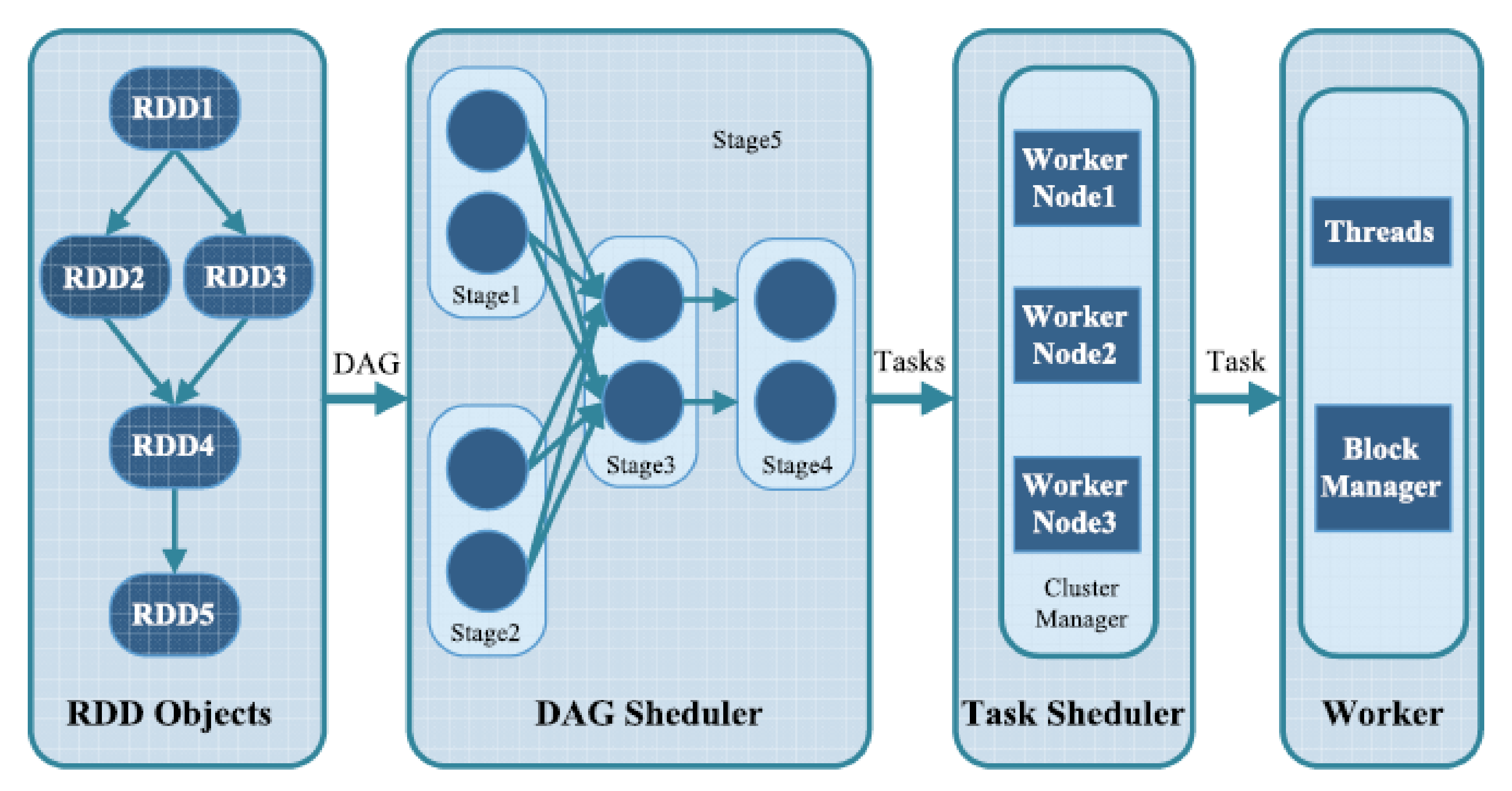
Rating Prediction

- Weighted Average (WA) [8]
- Mean-Centering (MC) [9][10]
- Z-Score (ZS) [11]

Hadoop vs Apache [12]

- Hadoop
  - Processing data using MapReduce in Hadoop is slow
  - Performs batch processing of data
  - Hadoop has more lines of code. Since it is written in Java, it takes more time to execute.
  - Hadoop supports Kerberos authentication, which is difficult to manage
- Apache
  - Spark processes data 100 times faster than MapReduce as it is dome in-memory
  - Performs both batch processing and real-time processing of data
  - Spark has fewer lines of code as it is implemented in Scala
  - Spark supports authentication via a shared secret. It can also run-on YARN leveraging the capability of Kerberos

The Task Scheduling Procedure In Spark [13]



Methodology

Algorithms That We Have Used

K-nearest neighbors (KNN) Algorithm

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems [14].

Alternating Least Square (ALS)

Alternating Least Square (ALS) is also a matrix factorization algorithm and it runs itself in a parallel fashion. ALS is implemented in Apache Spark ML and built for a larges-scale collaborative filtering problems [15].

Linear Regression Analysis

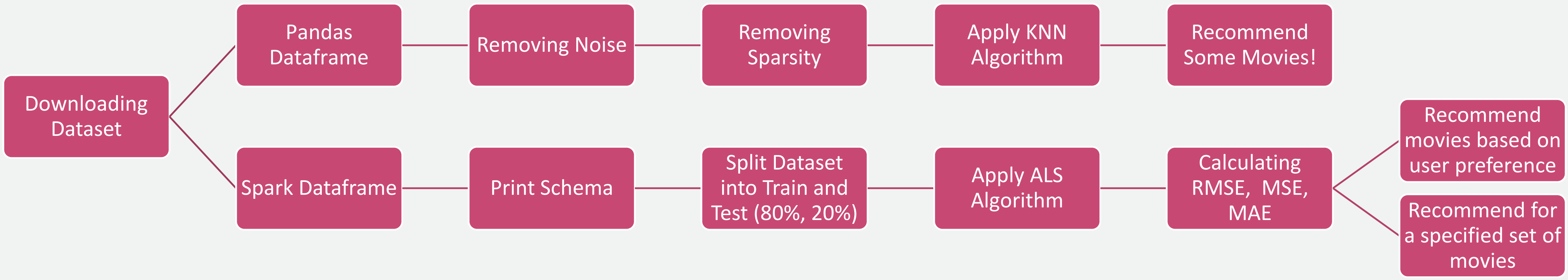
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable [16].

Our Dataset Statistics vs The Paper We Followed!

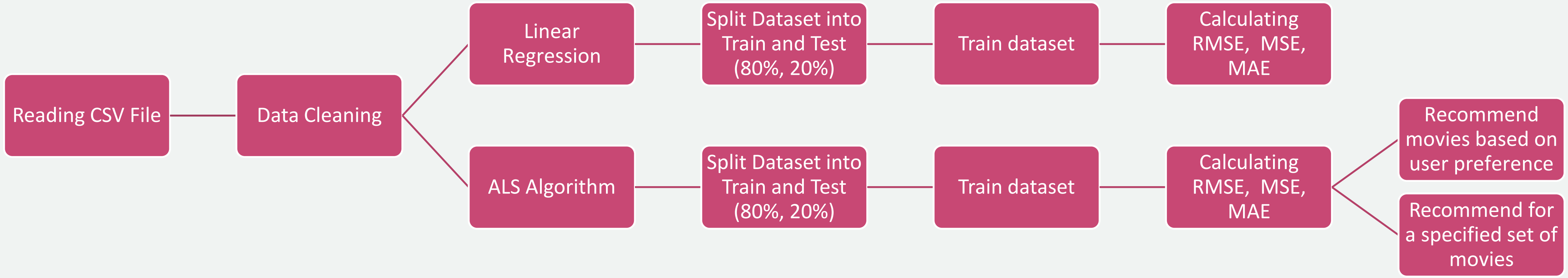
	Datasets	# of users	# of items	# of ratings	Sparsity
Our Dataset	MovieLens-100k	943	1682	100000	6.3%
	Netflix-6.2M	95325	412	6198103	15%

	Datasets	# of users	# of items	# of ratings	Sparsity
Dataset of The Paper We Followed [13]	WikiLens	326	5111	26937	1.6%
	MovieLens-100k	943	1682	100000	6.3%
	MovieLens-1M	6040	3900	1000209	4.2%

Methodology for MovieLens-100k Dataset



Methodology for Netflix-6.2M Dataset



Experimental Results

Movie Recommender System using KNN Algorithm :: MovieLens-100k

- Used KNN algorithm to compute similarity with Cosine Distance metric.
- We first check if the movie name input is in the database (CSV)
- If exists, then we use our recommendation system to find similar movies
- Sort them based on their similarity distance and output only the top 10 movies with their distances from the input movie
- All the movies in the top 10 are just like “Kolya” itself, therefore I think the result, in this case, is also good.

```
1  get_movie_recommendation('Kolya')
```

	Title	Distance
1	Fly Away Home	0.681671
2	Raise the Red Lantern	0.680329
3	Antonia's Line	0.679670
4	Like Water For Chocolate	0.673561
5	Angels and Insects	0.664926
6	L.A. Confidential	0.664229
7	Mrs. Brown	0.652900
8	Ulee's Gold	0.652874
9	Ridicule	0.652396
10	Lone Star	0.647755

Movie Recommender System using ALS Algorithm :: MovieLens-100k

Selecting the list of users for the movie which id is 17

user_id	title	ratings
475	From Dusk Till Dawn	3.6
78	From Dusk Till Dawn	3.380952380952381
248	From Dusk Till Dawn	3.5714285714285716
88	From Dusk Till Dawn	3.9523809523809526
797	From Dusk Till Dawn	2.6538461538461537
266	From Dusk Till Dawn	3.260869565217391
366	From Dusk Till Dawn	4.393939393939394
494	From Dusk Till Dawn	3.872340425531915
97	From Dusk Till Dawn	4.158730158730159
692	From Dusk Till Dawn	3.275

- Selecting the list of movies for the user whose id is 6

movie_id	title	ratings
1368	Mina Tannenbaum	3.6666666666666665
1463	Boys, Les	3.3333333333333335
1158	Fille seule, La	4.0
1202	Maybe, Maybe Not	3.5
207	Cyrano de Bergerac	3.8181818181818183
513	Third Man, The	4.333333333333333
1203	Top Hat	4.0476190476190474
1643	Angel Baby	3.75
652	Rosencrantz and G...	3.888888888888889
927	Flower of My Secr...	3.1666666666666665

Movie Recommender System using ALS Algorithm :: Netflix-6.2M

Selecting the list of users for the movie which id is 17

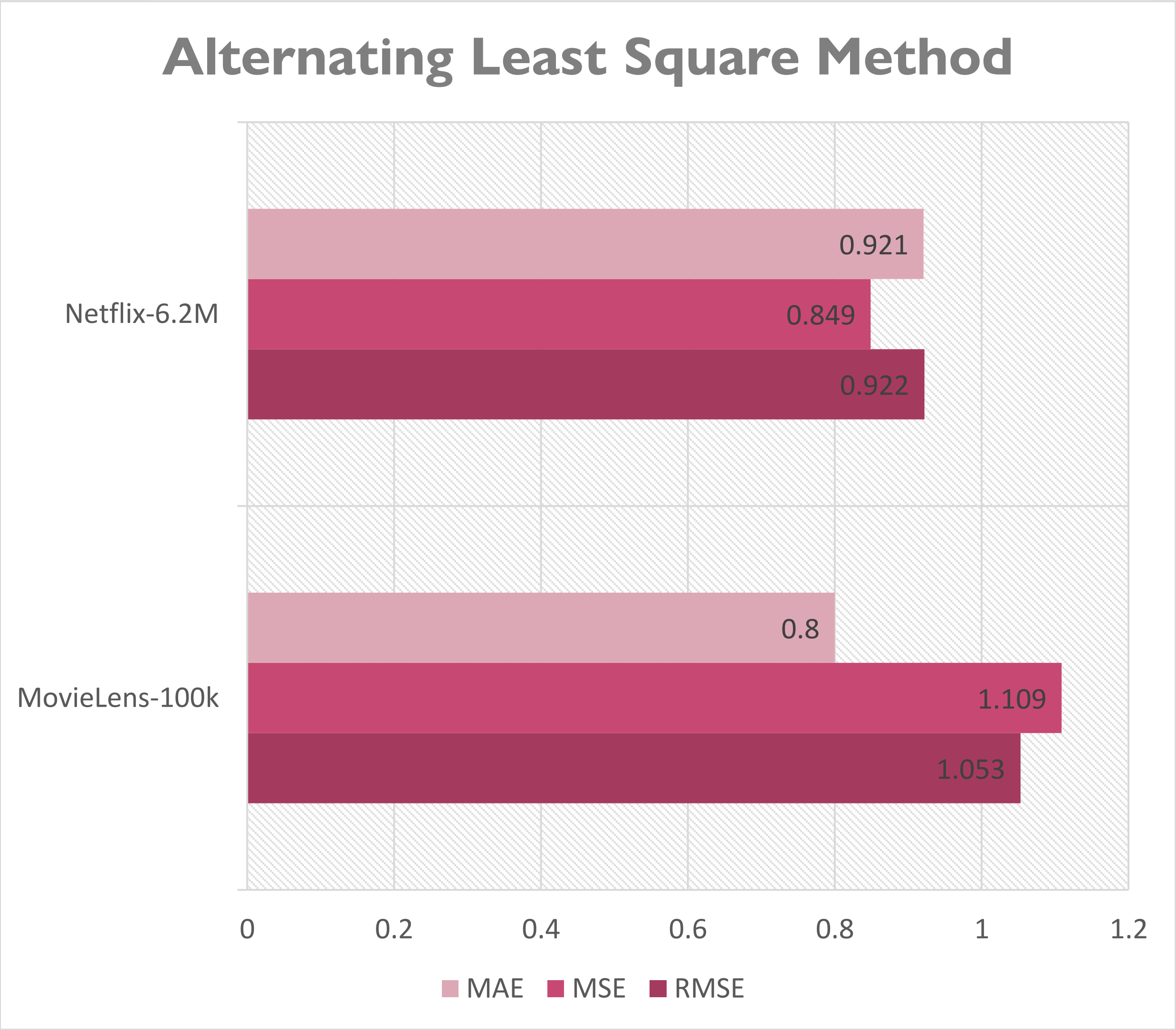
user_id	Name	ratings
1264514	National Lampoon'...	4.2745098039215685
364590	National Lampoon'...	4.866666666666666
1723350	National Lampoon'...	4.526315789473684
160876	National Lampoon'...	4.269230769230769
18764	National Lampoon'...	4.194444444444445
782679	National Lampoon'...	4.36734693877551
1180376	National Lampoon'...	3.8461538461538463
1038898	National Lampoon'...	4.177777777777778
1258697	National Lampoon'...	3.9166666666666665
1657241	National Lampoon'...	4.1568627450980395

- Selecting the list of movies for the user whose id is 6

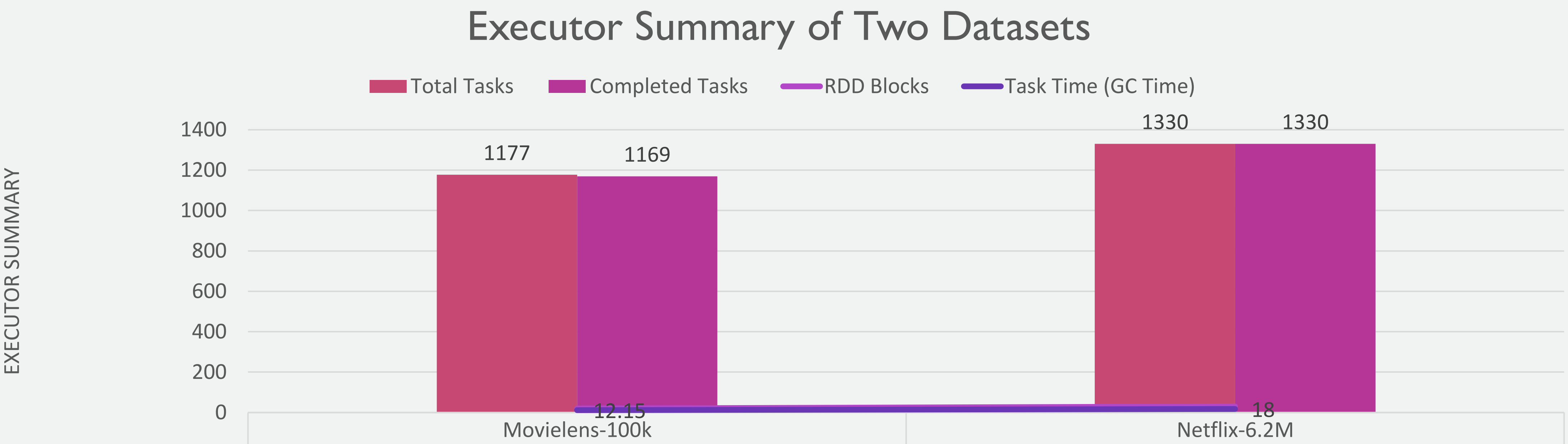
movie_id	Name	ratings
2862	The Silence of th...	4.304120879120879
3290	The Godfather	4.38083346646712
1692	Lonesome Dove	4.078693951248871
2782	Braveheart	4.260301246537396
3456	Lost: Season 1	4.65859938208033
3124	Titanic	3.760035682426405
2452	Lord of the Rings...	4.428412903907633
3391	Where the Red Fer...	3.9453125
1642	Casino: 10th Anni...	3.997315385487957
3153	Doctor Zhivago	3.949515316013959

Comparison Between Netflix-6.2M & ML-100K Datasets ALS Method

RMSE, MSE, MAE



Executor Summary of Two Datasets



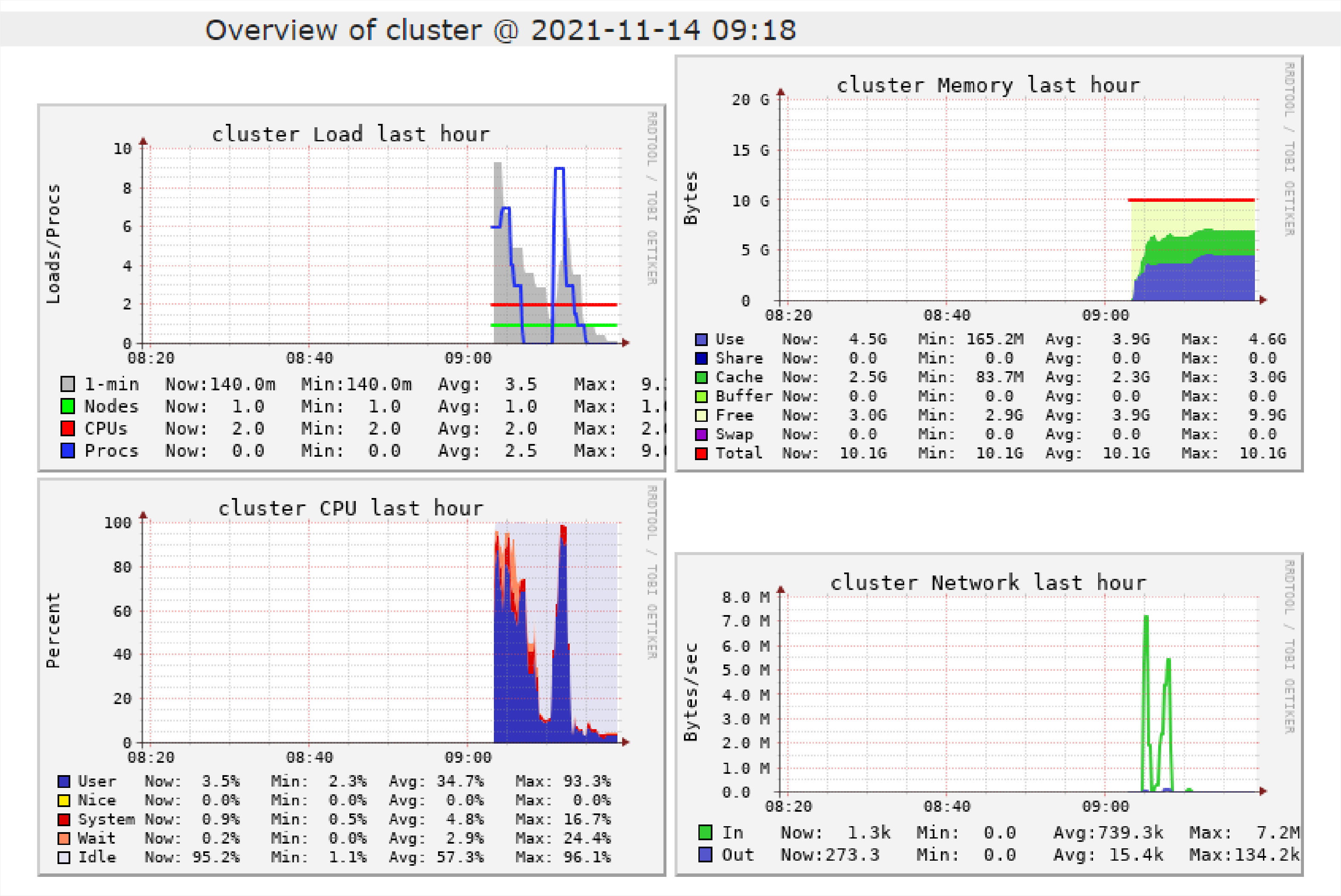
	Movielens-100k	Netflix-6.2M
Total Tasks	1177	1330
Completed Tasks	1169	1330
RDD Blocks	20	30
Task Time (GC Time)	12.15	18

DATASETS

Ganglia Cluster Report

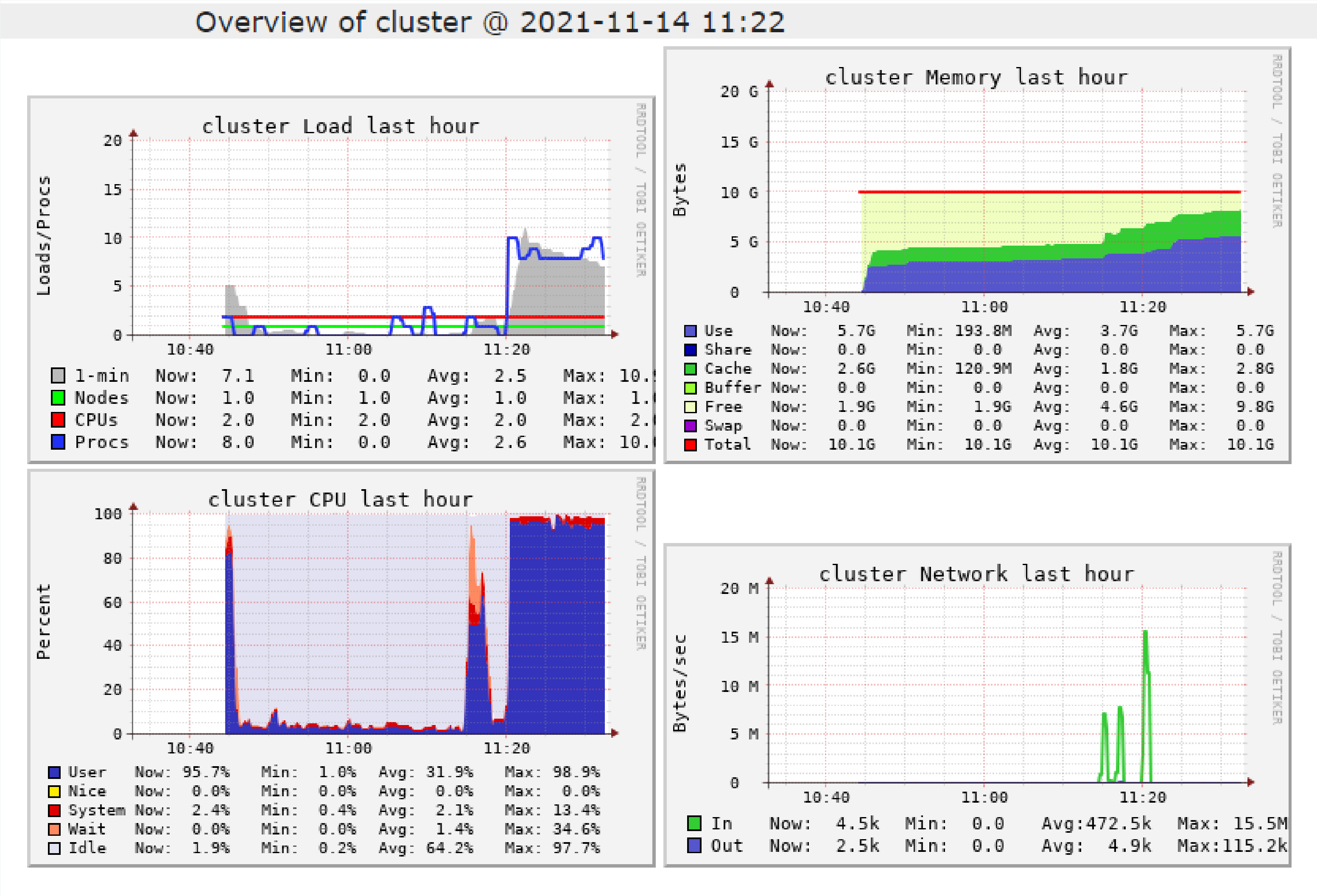
Ganglia Cluster Report :: MovieLens-100k (Host View)

Databricks Runtime Version:
10.1 (includes Apache Spark 3.2.0, Scala 2.12)

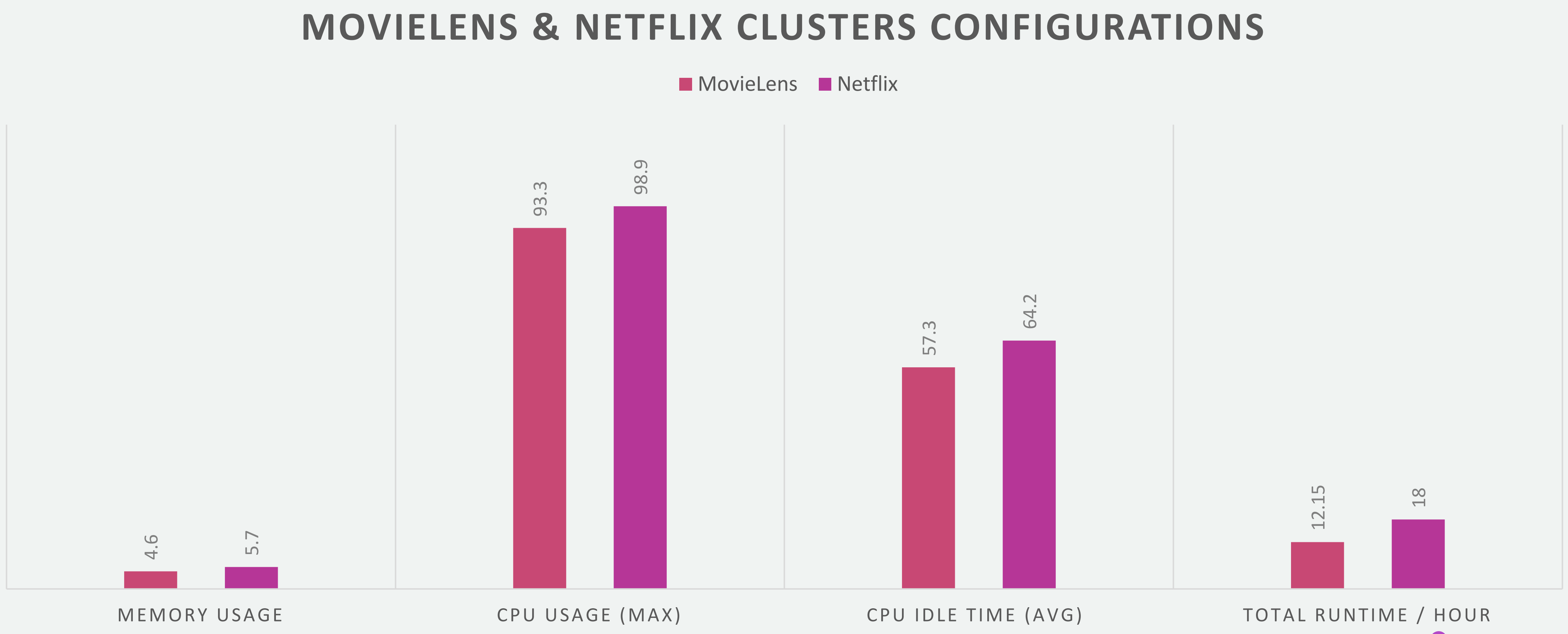


Ganglia Cluster Report :: Netflix-6.2M (Host View)

Databricks Runtime Version:
10.1 (includes Apache Spark 3.2.0, Scala 2.12)



MovieLens and Netflix Clusters Configurations - usage Overall



Limitations & Future Work

Unable to create multiple master / worker node in Microsoft Azure Cloud as a student

Databricks Community Edition only permits cluster with 8-core CPU with node 2

Future Work

Hybrid recommender systems

Reducing dataset size and user-item segmentation

References

[1] R. Ji, Y. Tian, and M. Ma, “Collaborative filtering recommendation algorithm basedon user characteristics,”2020 5th International Conference on Control, Robotics andCybernetics (CRC), p. 56–60, 2020.

[2] A. Pal, P. Parhi, and M. Aggarwal, “An improved content based collaborative filter-ing algorithm for movie recommendations,”2017 Tenth International Conference onContemporary Computing (IC3), p. 1–3, 2017.

[3] K. B. Fard, M. Nilashi, and N. Salim, “Recommender system based on semantic simi-larity,”International Journal of Electrical and Computer Engineering (IJECE), vol. 3,no. 6, 2013.

[4] M. Deshpande and G. Karypis, “Item-based top- n recommendation algorithms,”ACMTransactions on Information Systems, vol. 22, no. 1, p. 143–177, 2004.

[5] P. Ahlgren, B. Jarneving, and R. Rousseau, “Requirements for a cocitation similaritymeasure, with special reference to pearsons correlation coefficient,”Journal of theAmerican Society for Information Science and Technology, vol. 54, no. 6, p. 550–560,Apr 2003.

[6] X. Wu, Y. Huang, and S. Wang, “A new similarity computation method in collabo-rative filtering based recommendation system,”2017 IEEE 86th Vehicular TechnologyConference (VTC-Fall), 2017.

[7] J. Bobadilla, A. Hernando, F. Ortega, and A. Gutí errez, “Collaborative filtering basedon significances,”Information Sciences, vol. 185, no. 1, p. 1–17, 2012.

[8] A. Bonfietti and M. Lombardi, “The weighted average constraint,” vol. 7514, 10 2012,pp. 191–206.

[9] M. Hofer, “Mean centering,”The International Encyclopedia of Communication Re-search Methods, p. 1–3, 2017.

[10] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of pre-dictive algorithms for collaborative filtering,” Jan 1998. [Online]. Available:https://arxiv.org/abs/1301.7363

[11] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, “An algorithmic frameworkfor performing collaborative filtering,”Proceedings of the 22nd annual internationalACM SIGIR conference on Research and development in information retrieval - SIGIR99, 1999.

[12] A. Wakde, P. Shende, S. Waydande, S. Uttarwar, and G. Deshmukh, “Comparativeanalysis of hadoop tools and spark technology,” in2018 Fourth International Confer-ence on Computing Communication Control and Automation (ICCUBEA), 2018, pp.1–4

[13] J. Sun, Z. Wang, X. Luo, P. Shi, W. Wang, L. Wang, J.-H. Wang, and W. Zhao, “Aparallel recommender system using a collaborative filtering algorithm with correntropyfor social networks,”IEEE Transactions on Network Science and Engineering, vol. 7,no. 1, p. 91–103, 2020.

[14] P. Cunningham and S. Delany, “k-nearest neighbour classifiers,”Mult Classif Syst,vol. 54, 04 2007.

[15] S. Ghosh, N. Nahar, M. Wahab, M. Biswas, M. Hossain, and K. Andersson,Recom-mendation System for E-commerce Using Alternating Least Squares (ALS) on ApacheSpark, 02 2021, pp. 880–893.

[16] T. Jhalani, V. Kant, and P. Dwivedi, “A linear regression approach to multi-criteriarecommender system,” vol. 9714, 06 2016, pp. 235–243.

Discussion

How is Apache Spark different from MapReduce?

Why is Apache Spark faster than Apache Hadoop?

Is it possible to run Apache Spark without Hadoop?

What role does worker node play in Apache Spark Cluster? And what is the need to register a worker node with the driver program?

How can you trigger automatic clean-ups in Spark to handle accumulated metadata?

Thank You