Assignment (Marks 10)

Submit By: 3rd May 2022

Measuring Performance of PIG and HIVE

*** Only one submission per group is required**

Name of Group Member	Roll Number
Archana P Ajit	MT21ACS427
Christo Joseph	MT21ACS426
Jishnu Chandran	MT21ACS419
Rahul Nambiar	MT21ACS421

In this assignment, you will measure the performance of PIG and HIVE by testing them against various operations against varying sizes of data sets

The operations can be Arithmetic operation, Filter operation (filtered set 5%), Filter operation (filtered set 95%), Group of one column, Join. You can add other operations if you desire.

Sizes of the data set (increase 10X times): approx. 500 KB, 5 MB, 50 MB, 500MB, 5 GB, 50 GB

Put the code for all your operations for PIG in a single file named operationsPIG.txt and for HIVE, operationsHIVE.txt.

Hardware configurations:

Memory: 2GB
Processor: 11th Gen Intel® Core™ i5-1135G7 @ 2.40GHz
Disk Capacity: 100GB

LINK of the Data Set:

Size	Remark	Link
50KB	(Created From 500KB)	https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot
500KB	Melbourne housing snapshot	https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot
5MB	(Created from 500KB)	https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot
50MB	Riots/Protests in India,2016-2022 - 100k Datapoint	https://www.kaggle.com/datasets/shivkumarganesh/riots-in-india-19972022-acled-dataset-50k

500MB	Open Adresses US Northeast	https://www.kaggle.com/datasets/openaddresses/openaddresses-us-northeast?select=ny.csv
5GB	Large Car Dataset	https://www.kaggle.com/datasets/cisautomotiveapi/large-car-dataset?resource=download

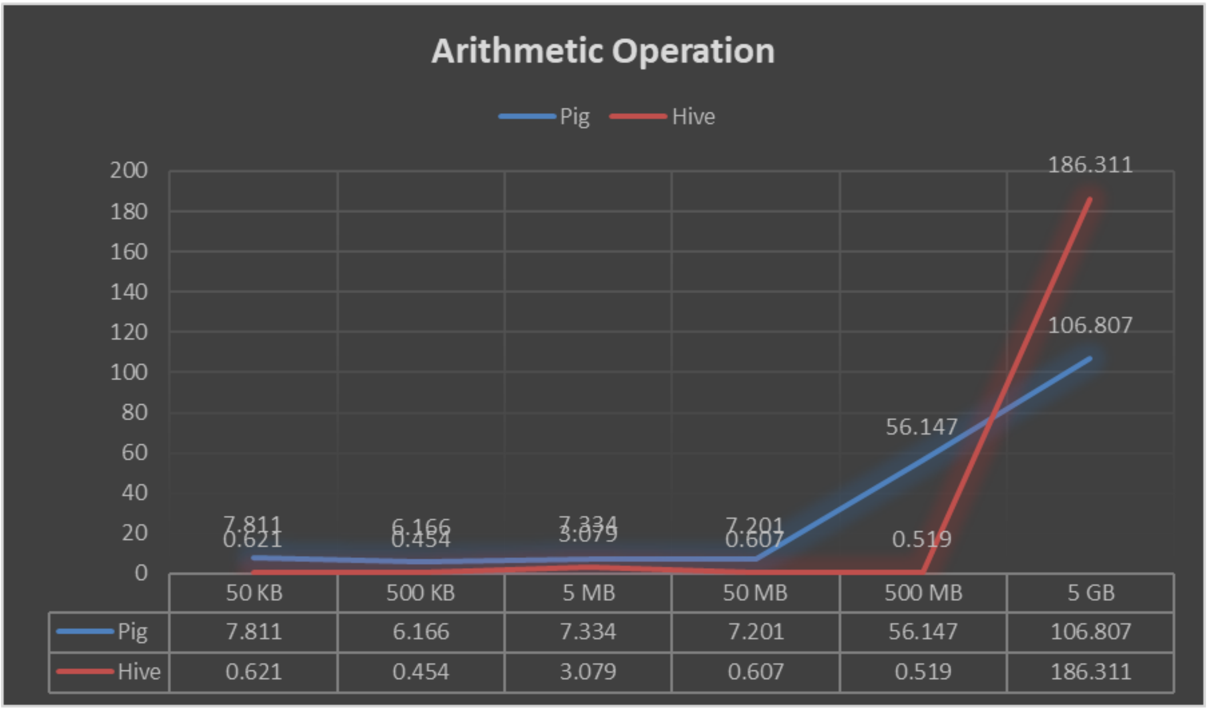
Fill the following table PIG:

	Size 50KB	Size 500KB	Size 5MB	Size 50MB	Size 500MB	Size 5GB
Arithmetic Operation	7.811s	6.166s	7.334s	7.201s	56.147s	106.807s
Filter 5%	6.645s	6.052s	6.341s	15.270s	51.969s	134.655
Filter 95%	6.343s	6.078s	6.588s	13.237s	53.091s	144.899s
Group By	7.685s	7.236s	9.508s	17.461s	105.311s	290.718s
Distinct	6.310s	6.638s	8.216s	17.454s	126.37s	410.899s
Join	6.907s	8.749s	94.017s	26.471s	192.569s	279.977s

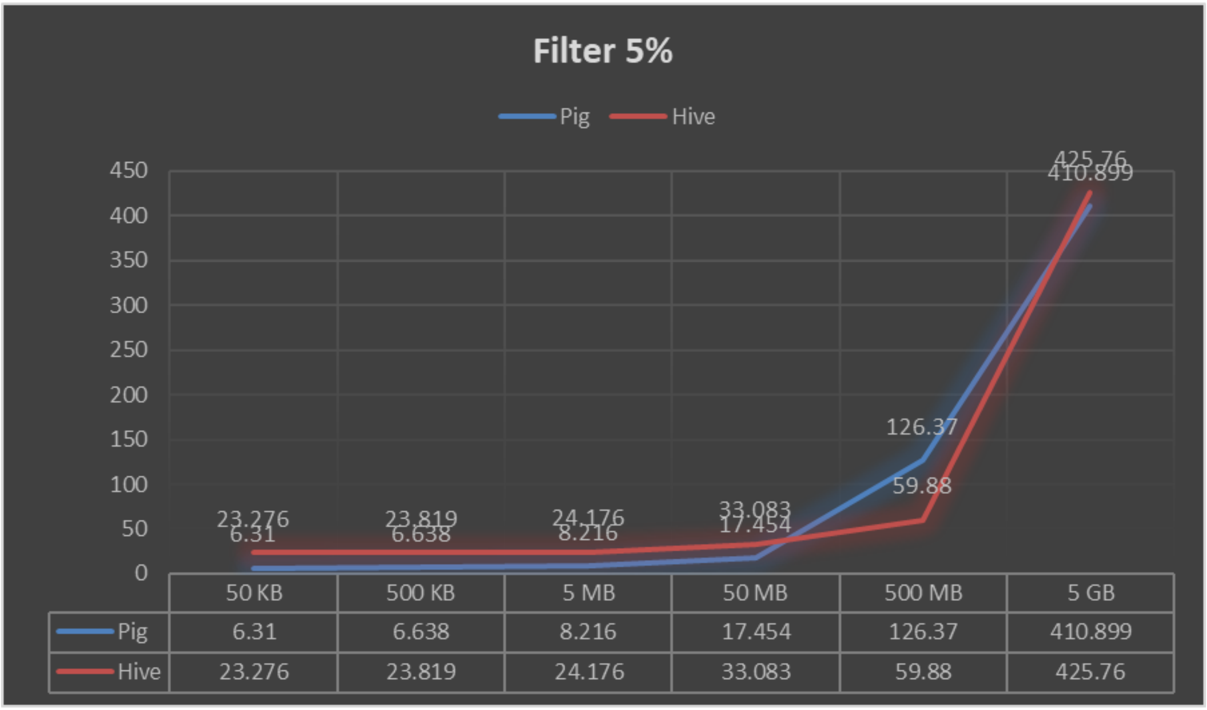
Fill the following table HIVE:

	Size 50KB	Size 500KB	Size 5MB	Size 50MB	Size 500MB	Size 5GB
Arithmetic Operation	0.621 s	0.454s	3.079s	0.607s	0.519s	186.311s
Filter 5%	0.629s	0.684s	0.544s	5.053s	0.708s	248.509s
Filter 95%	0.592s	0.578s	0.509s	0.798s	0.556s	210.602s
Group By	29.853s	26.863s	34.425s	45.27s	57.684 s	383.493s
Distinct	23.276s	23.819s	24.176s	33.083s	59.88s	425.76s
Join	21.47s	27.61s	174.328s	38.455s	113.668s	500.15s

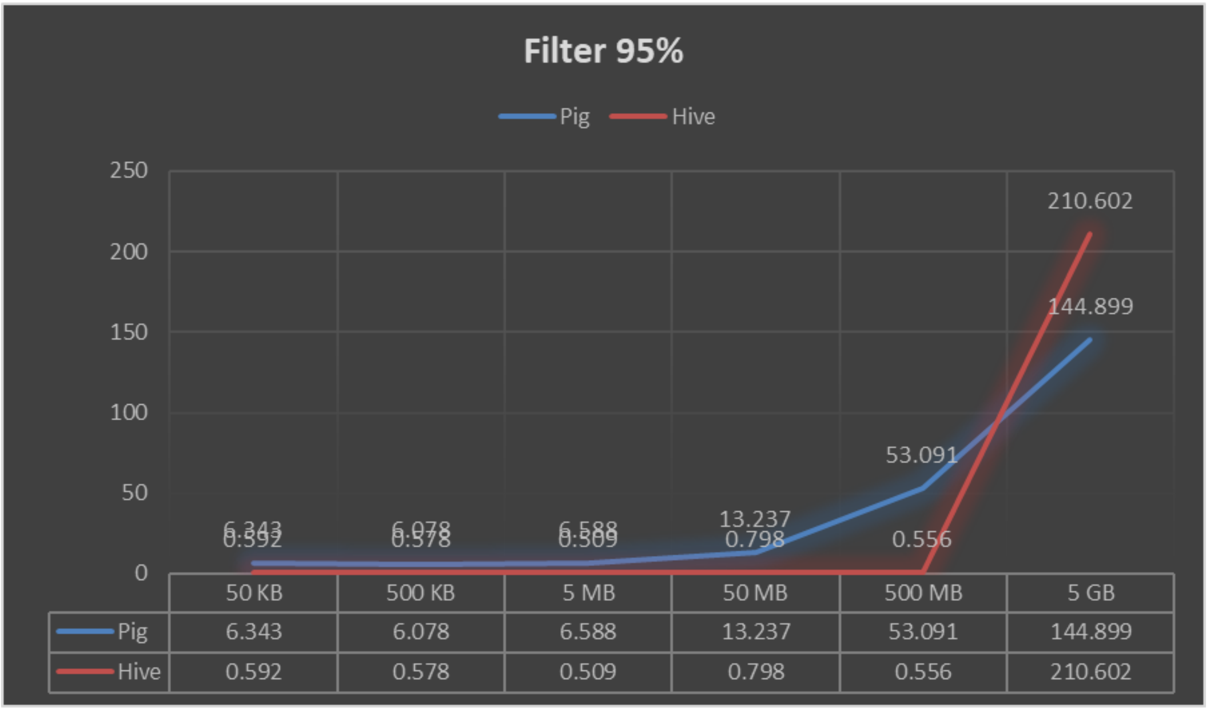
Graph:



	50 KB	500 KB	5 MB	50 MB	500 MB	5 GB
Pig	7.811	6.166	7.334	7.201	56.147	106.807
Hive	0.621	0.454	3.079	0.607	0.519	186.311

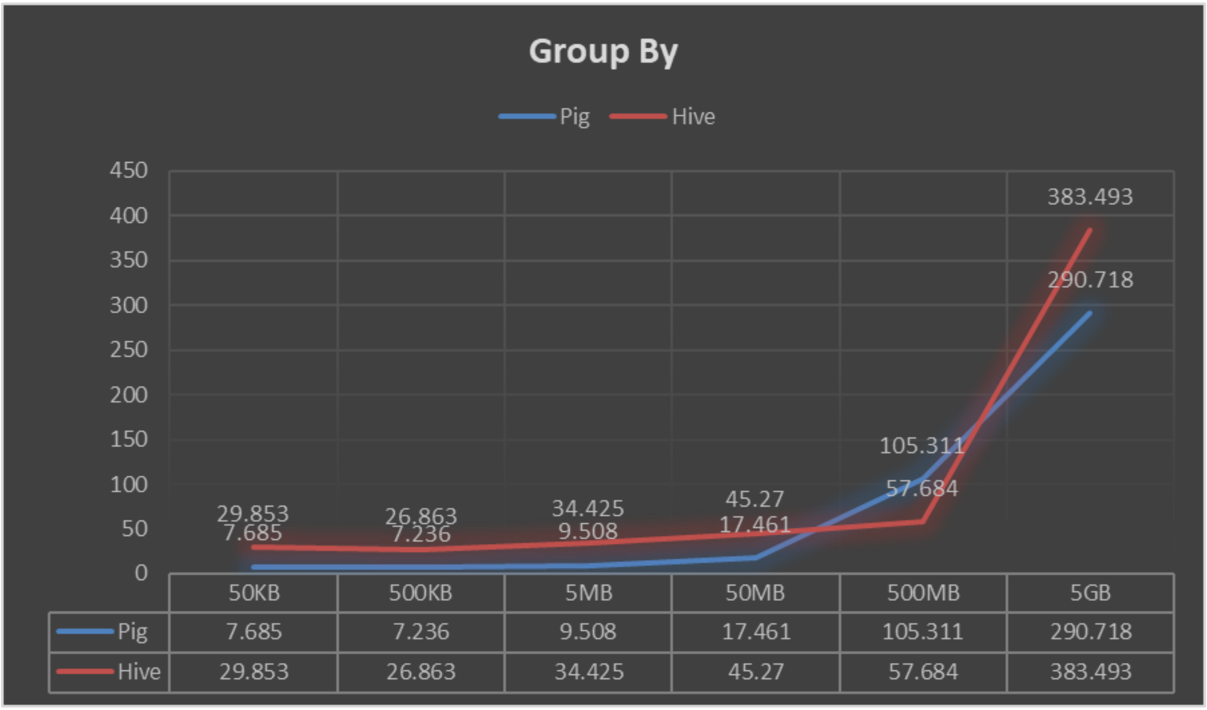


	50 KB	500 KB	5 MB	50 MB	500 MB	5 GB
Pig	6.31	6.638	8.216	17.454	126.37	410.899
Hive	23.276	23.819	24.176	33.083	59.88	425.76



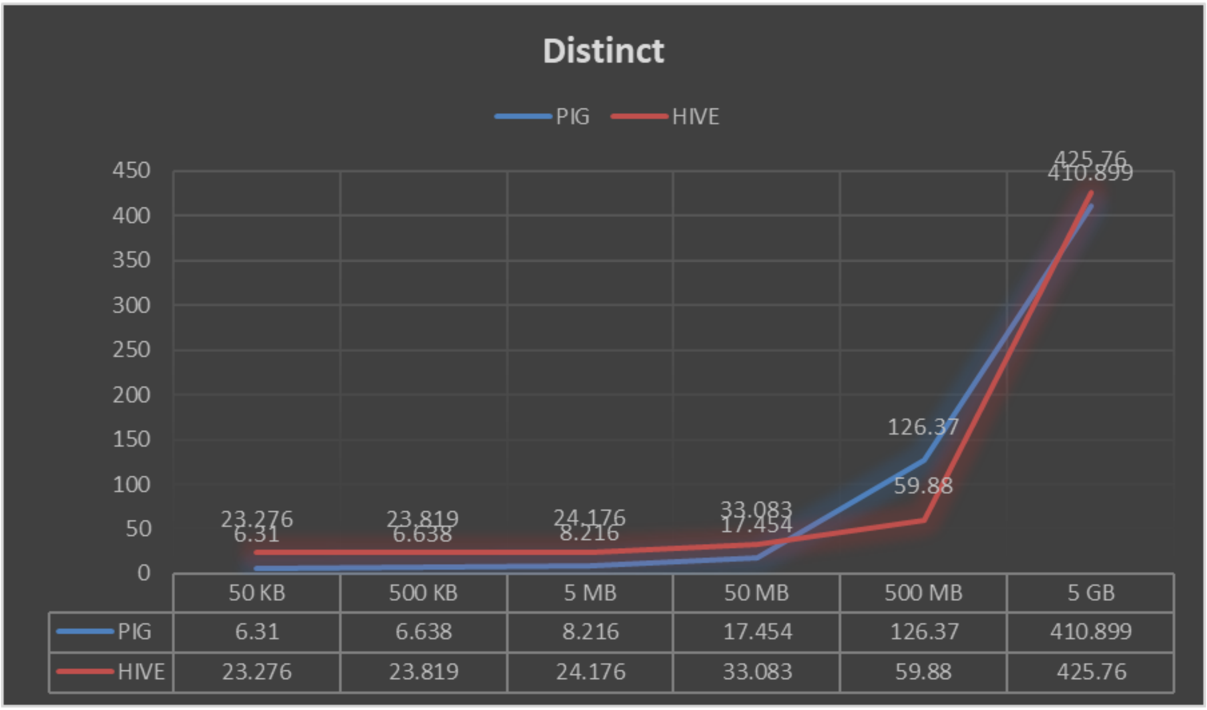
Filter 95%

	50 KB	500 KB	5 MB	50 MB	500 MB	5 GB
Pig	6.343	6.078	6.588	13.237	53.091	144.899
Hive	0.592	0.578	0.509	0.798	0.556	210.602

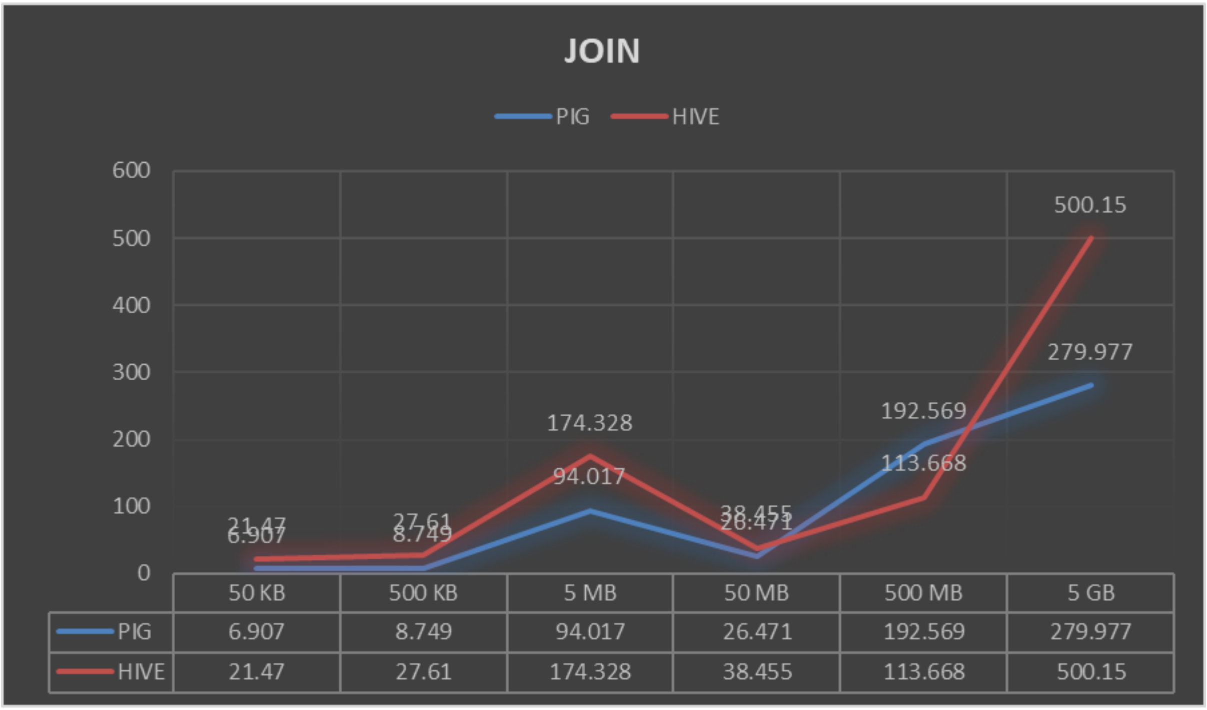


Group By

	50KB	500KB	5MB	50MB	500MB	5GB
Pig	7.685	7.236	9.508	17.461	105.311	290.718
Hive	29.853	26.863	34.425	45.27	57.684	383.493



	50 KB	500 KB	5 MB	50 MB	500 MB	5 GB
PIG	6.31	6.638	8.216	17.454	126.37	410.899
HIVE	23.276	23.819	24.176	33.083	59.88	425.76



	50 KB	500 KB	5 MB	50 MB	500 MB	5 GB
PIG	6.907	8.749	94.017	26.471	192.569	279.977
HIVE	21.47	27.61	174.328	38.455	113.668	500.15

Conclusion:

As per the above result we can conclude the following:

- Hive performed slightly better than Pig in case of arithmetic operations.
- Overall Pig performance is better than Hive for filtering 5% of the dataset.
- Overall Pig performance is better than Hive for filtering 95% of the dataset.
- Pig performance is dominating over Hive for Group By operation.
- Pig performance seems to be better than Hive for Distinct operation.
- Overall Pig performance is better than Hive for Join operation.

In conclusion, Pig performs better when compared to Hive in case of varying size of datasets.