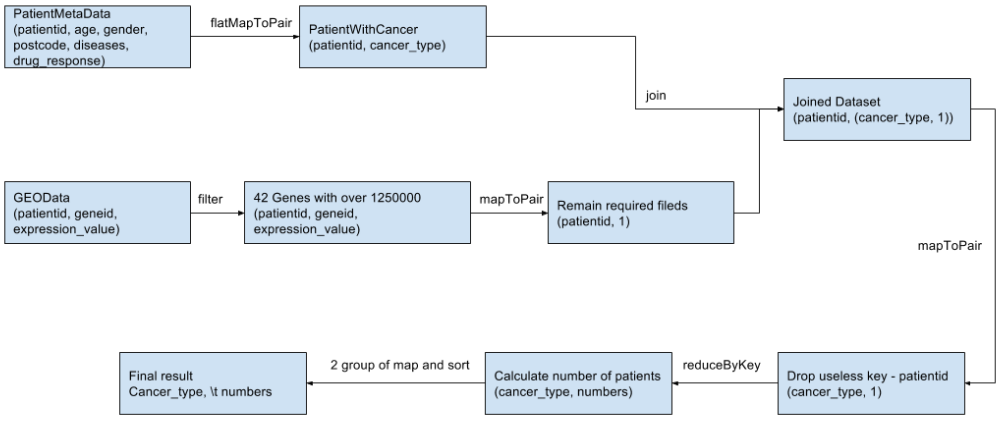
COMP5349 Assign3

Task1

In this task, the PatientMetaData file is read in Spark as PairRDD, followed by a flatMapToPair operation to filter the patients who get cancer. After that, the GEO file is read, followed by a filter operation to filter 42 gene with valid express value. In addition, a map function is used to transfer the result to (patientid, 1).
A join operation is executed upon two PairRDDs processed above based on the same key - patientid followed by a mapToPair transformation to adjust the joined result. Then reduceByKey operation is executed based on patientid and two sortByKey function is followed to get the required format.



Performance:

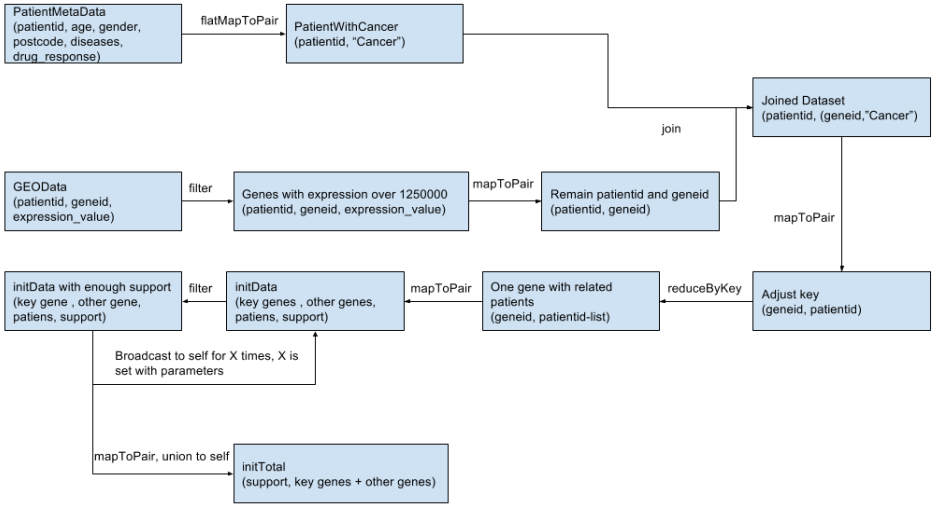
Dataset	Runtime	Run Command On Large Dataset
Small	12s	spark-submit \ --class assign3.Task1 \ --master yarn \ --num-executors 3 \ comp5349assign3-0.0.1.jar \ hdfs://soit-hdp-pro-1.ucc.usyd.edu.au:8020/share/genedata/large/ \ #input path hdfs://soit-hdp-pro-1.ucc.usyd.edu.au:8020/user/cshe6391/assign3/task1large #output path
Large	14s	

N.B. Remove comment after hash tag before execute command.

Task2

In this task, the PatientMetaData.txt is read and filtered based on the cancer list. On the other hand, the GEO.txt is read and filtered with valid expression value followed by a map to pair function to drop unrelated fields. A join operation is executed based on patientid.

After that, the PairRDD is grouped and reduced by geneid in that patients are collected together. InitData is the initial dataset of loop followed by a filter to drop some entry without enough support. In each loop, the result is broadcast in that they can be used in the next loop.
The result is collected by initTotal PairRDD meanwhile, the initData of every loop is styled and union to it.
FrequentItemMapper is a customised mapToPair function to help mine frequent items.
Another package called assign3P in the code is attached where a joined version is implemented. That version is designed for huge source input but in this project, the result of experiment is collected based on the code in assign3 package.



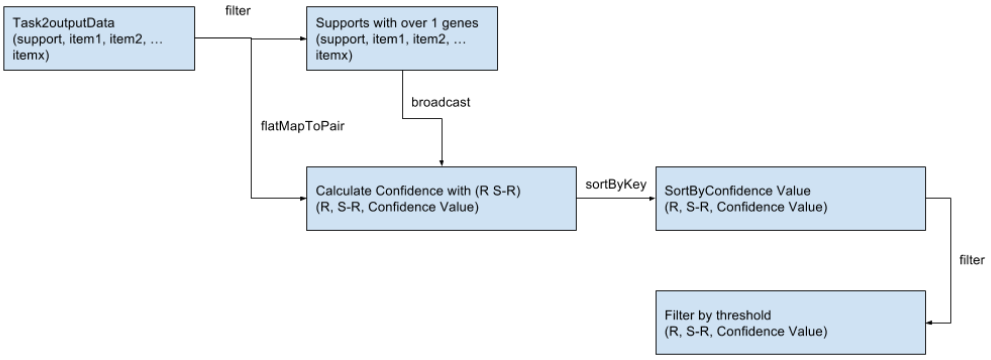
Performance:

Dataset	Runtime	Run Command On Large Dataset
Small	23s	spark-submit \ --class assign3.Task2 \ --master yarn \ --num-executors 10 \ #in this experiment 10 executors are used comp5349assign3-0.0.1.jar \ hdfs://soit-hdp-pro-1.ucc.usyd.edu.au:8020/share/genedata/large/ \ #input path hdfs://soit-hdp-pro-1.ucc.usyd.edu.au:8020/user/cshe6391/assign3/task2large \ #output path 0.3 \ #set support percentage 4 #set max iteration times
Large	147s	

N.B. Remove comment after hash tag before execute command.

Task3

In this task, the output of task 2 is loaded. A filter is applied to get a PairRDD of which every entry owns more than one genes. The result is broadcast to the original dataset to help generate the (R, S-R, Confidence Value) PairRDD. After that, the PairRDD is sorted by confidence value and is filtered according to its confidence value either. A customized getSet function is used to help find association rules.



Performance:

Dataset	Runtime	Run Command On Large Dataset
Small	13s	spark-submit \ --class assign3.Task3 \ --master yarn \ --num-executors 3 \ comp5349assign3-0.0.1.jar \ hdfs://soit-hdp-pro-1.ucc.usyd.edu.au:8020/user/cshe6391/assign3/task2 \ #input path hdfs://soit-hdp-pro-1.ucc.usyd.edu.au:8020/user/cshe6391/assign3/task3large \ #output path 0.6 #set confidence threshold
Large	14s	

N.B. Remove comment after hash tag before execute command.

Appendix

HDFS Location:

Item	Path
Task 1 small	/user/cshe6391/assign3/task1small
Task 1 large	/user/cshe6391/assign3/task1large
Task 2 small	/user/cshe6391/assign3/task2small
Task 2 large	/user/cshe6391/assign3/task2large
Task 3 small	/user/cshe6391/assign3/task3small
Task 3 large	/user/cshe6391/assign3/task3large