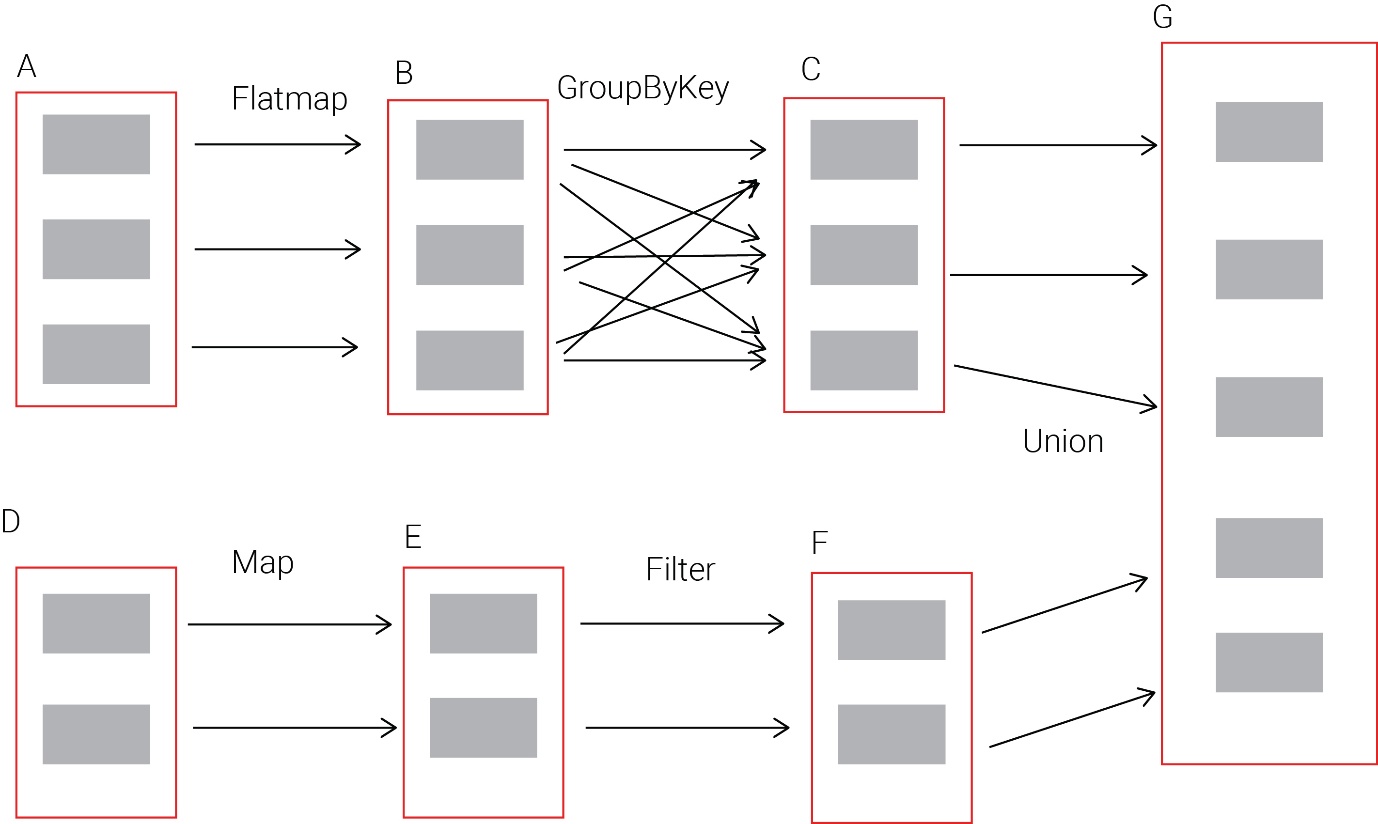
2023-CSE5BDC-T5-W - Big Data Management on the Cloud

Big Data Management on the Cloud Week 4. Quizzes or Notes



**Topic 1: Apache spark internals and programming**

Given the Apache Spark following execution graph, which set of RDDs belong to the same stage? (There are more than one correct answer.)

**A,B**

**D,E,F**

Given the following Apache Spark execution graph, which transformations are narrow dependences? (Note there are more than one correct answer)

D-> map -> E

* A -> Flatmap -> B
* C and F -> Union -> G

Suppose you are given an RDD with the following attributes (student name, subject, mark). You are asked to write a Spark program that returns the student's name with the highest mark (break ties arbitrarily). Which set of Spark operations would you use? (There is just one correct answer.)

sortBy, map

Suppose you are given an RDD with the following attributes (student name, subject, mark). You are asked to write a Spark program that returns the subject that gave the highest total mark (break ties arbitrarily). Which set of Spark operations would you use? (There is just one correct answer.)

map, reduceByKey, sortBy

**Topic 2: Introduction to SparkSQL**

Which statements about the SparkSQL’s query optimiser are true? (Choose two answers.)

-SparkSQL’s query optimiser can more optimally order query operations to result in faster query execution.

-SparkSQL’s query optimiser can potentially combine certain operations together to speed up query processing.

Which statements about the SparkSQL’s Tungsten engine are true? (Choose two answers.)

-The Tungsten engine consists of very optimised Java byte code for serialisation and deserialisation of objects.

-The Tungsten engine makes more optimal use of the CPU cache by storing data in more compact formats