title

wbg231

January 2023

1 Introduction

q1

- If no caching is involved, wide dependencies are just as fast as narrow dependencies
- this if false, regardless of if we are caching (ie) saving data to memory a wide dependencie has to use data from multiple paritions so it will have latency

question 3

- An RDD can depend on multiple parent and be reused by multiple descendents
- true
- The Dremel system was designed to efficiently process subsets of attributes over all records in a dataset.
- yep that is the point to look at cols not rows
- If a numerical column A has been compressed with run-length encoding, it must be decompressed to compute the average mean(A).
- run length encoding maps a vector to each of its unique values and the number of times they appeard so false
- descibe the roles of paritions in spark RDD's what do they do how do they effect disrivuted computing
- paritions are the amount of data that a worker node in RDD needs to work with
- we want the data in that parition to be as close as possible to the block the worker node is on

- raising the number of paritions beyond the number of workers means that after finishing one run, a worker will have to go back and do another (this increses comuncaiton costs)
- lowering the value of partions bellow the number of workers means that some worers are dealing with a lot of data while others are idle
-