**COMP9313 Assignment01**

**zID: z5224151**

**Name: ZANNING WANG**

**Q1:**

**1.**

To take replication factor is 3 as an example, due to different blocks have same replicas, we can get the following Venn diagram. From the graphs, we can see that if one circle represent one block lost, number 1 in the figure means the data have lost one replicas, number 2 in the figure means the data have lost two replicas, number 3 in the figure means the data have lost three replicas. So that we can get L1(k,N) under the replication factor = 3 as below:

图片包含 游戏机, 体育

描述已自动生成

This similar to the replication factor = 5, the L1(k,N) shows as below:

In order to calculate the L2(k,N), we should calculate the number of lost-replicas when the k-th block broken firstly. And due to the replication factor is 4, we know that the number of block who lose one replicas equal to 4L1(k-1,N), therefore we can get when the k(th) block lost, there are blocks lost 2 replicas, and we should plus when the k-1(th) block lost, there are blocks, which lost 2 replicas, however, there are some part been mis-calculated, when the k(th) block lost, there are some block have already lost two replicas, we should minus it: , therefore we can get the following formula:

The same as L3(k,N), L4(k,N), L5(k,N), the difference is coefficient is changed based on the number of replicas, therefore we can get the formulas like below:

**2.**

This problem can be treated as dynamic planning, which means if we want to calculate the L5(2,500), we should ues the L4(1,500) and L5(1,500), the similar for L4(2,500), L3(2,500), L2(2,500),like below:

L4(1,500), L5(1,500) => L5(2,500)

L3(1,500), L4(1,500) => L4(2,500)

L2(1,500), L3(1,500) => L4(2,500)

L1(1,500), L2(1,500) => L2(2,500)

And L1(1,500) = B, L2(1,500) = L3(1,500) = L4(1,500) = L5(1,500) = 0

According to the result we get from the last questions, L1(2,500) can be calculated from L5(2,500), L4(2,500), L3(2,500), L2(2,500)., like below:

L5(2,500), L4(2,500), L3(2,500), L2(2,500) =>L1(2,500)

In this way, we can write code to calculate L5(200,500) = 39736.77280169178

图片包含 游戏机

描述已自动生成

**Q2:**

**1.**

(1) In the rdd\_2, we get the name with score: [name,score];

(2) In the rdd\_3, we reduceByKey the rdd\_2 and get the student with the highest score he get. [name, highest\_score]

(3) In the rdd\_4, we reduceByKey the rdd\_2 and get the student with the lowest score he get. [name, lowest\_score]

(4) In the rdd\_5, we join rdd\_3 and rdd\_4 together. [name,highest\_score,lowest\_score]

(5) In the rdd\_6, add highest score and the lowest score of each student. [name, total\_score]

(6) Therefore we get the results as below:  
[(Tina,155),(Jimmy, 159),(Thomas,167),(Joseph,165)]

**2.**

The stage are divided according the transformation is narrow transformation or wide transformation. Only after a wide transformation, there will be a new stage. Therefore in this questions, we got three stage

图片包含 游戏机, 文字

描述已自动生成

**3.**

According the code snippet given, there are two shuffle between rdd\_3 and rdd\_5, rdd\_4 and rdd\_5, which will decrease the efficiency of the spark, we can group the key by student with all the score they get, and sum the highest and the lowest in one step, the code show below:

rdd\_1 = sc.parallelize(raw\_data)

rdd\_2 = sc.parallelize(raw\_data)

rdd\_3 = rdd\_2.groupByKey()

rdd\_4 = rdd\_3.mapValues(lambda e :min(e) + max(e))

rdd\_4.collect()

**Q3:**

**1.**

According to the assumption, the cos(θ(o, q)) ≥ 0.9, the θ must less than 25.842°

Pr[hi(o) = hi(q)] = 1- θ/π = 1-0.144 = 0.856

Therefore, we know pkq,o > 0.856

The probability of not finding any near duplicate is (1 - pkq,o)l

Therefore the equation is 1 – ( 1 - pkq,o)l ≥0.99, whereas l= 8.

In conclusion, at least 8 tables require to ensure the near duplicate with probability more than 99%.

**2.**

According to the assumption, the cos(θ(o, q)) < 0.8, the θ must more than 36.87°

Pr[hi(o) = hi(q)] = 1- θ/π = 1-0.205 = 0.795

Therefore, we know pkq,o < 0.795

The probability of not finding any near duplicate is (1 - pkq,o)l

Therefore the equation is 1 – ( 1 - pkq,o)l = 1-(1-0.795)­­­­10 = 0.9782

In conclusion, the maximum value of the probability of o to become a false positive of query q is 97.82%.