# COMP9313 Assignment1

Chen Wu
ID: z5244467

2020 年 8 月 2 日

## Question 1: HDFS

### Part (1)

1) When using Wayne to draw each part, it is very clear that different numbers of datanodes and files failded. According to the figure area, the following formula can be obtained

- $L_1(k,N) = k \cdot B - 2 \cdot L_2(k,N) - 3 \cdot L_3(k,N) - 4 \cdot L_4(k,N) - 5 \cdot L_5(k,N)$

2) $L_2(k,N)$ R represents the number of all replications

a. If k - 1 dataNodes failed and lost one replication: $4 \cdot L_1(k-1,N)/(N-k+1)$

b. If k - 1 dataNodes failed and have lost two replications: $L_2(k - 1,N)$

c. based on [b], if k-1 have lost tow replications, another replication lost: $3 \cdot L_2(k - 1,N)/(N - k + 1)$

- Therefore, the result can be acquired:

* $L_2(k,N) = 4 \cdot L_1(k - 1,N)/(N - k + 1) + L_2(k - 1,N) - 3 \cdot L_2(k - 1,N)/(N - k + 1)$

- similarly

* $L_3(k,N) = 3 \cdot L_2(k - 1,N)/(N - k + 1) + L_3(k - 1,N) - 2 \cdot L_3(k - 1,N)/(N - k + 1)$

* $L_4(k,N) = 2 \cdot L_3(k - 1,N)/(N - k + 1) + L_4(k - 1,N) - L_4(k - 1,N)/(N - k + 1)$

* $L_5(k,N) = L_4(k - 1,N)/(N - k + 1) + L_5(k - 1,N)$

### Part (2)

```
def update(l,k,table):
    # according to the rule which is found in question 1
        if l == 1:
            table[l][k] = k * B -2 * table[2][k] - 3 *
            table[3][k] - 4 * table[4][k] - 5 * table[5][k]
            return

        table[l][k] = ((6 - l) * table[l - 1][k - 1]) / (N - k + 1)
        + table[l][k - 1] - ((6 - l - 1) * table[l][k - 1]) / (N - k + 1)
        return
R,N,K= 20000000, 500, 200
B = R / N
table = [[0 for _ in range(201)] for _ in range(6)]
table[0][0]=B
tmp = [5,4,3,2,1]
for k in range(201):
        for l in tmp:
            update(l, k, table)
print(table[5][200])
```

Therefore, the final result is 39736.7728

## Question 2: Spark

### Part (1)

The step rdd_2 is to select name and score from raw_data;
rdd_3: Use the name as the key value to find the corresponding max score
rdd_4: Use the name as the key value to find the corresponding min score
rdd_5 is to merge the max and min score by same key
rdd_6 is to add the max and min score
The output: [(Tina,155),(Jimmy,159),(Thomas,167),(Joseph,165)]

### Part (2)

The number of stage is 3.
stage0: rdd_1 rdd_2
stage1: rdd_1 rdd_2

stage2: the rest of rdd
As we known, when a wide dependency is encountered, it is disconnected and divided into a stage; when a narrow dependency is encountered, the RDD is added to the stage. In this question, at first rdd_1 and rdd_2 was be created in stage0. Secondly, rdd_1 and rdd_2 was be created in stage1. Then reduceByKey belongs to wide dependency, which number is two. The remaining operations belong to another stage. So The number of stage should be three. Because rdd_3 and rdd_4 have same dataframe, join() here does not belong to wide dependency.

### Part (3)

It is clear that the times of using shuffle need to be reduce.

```
rdd_1 = sc.parallelize(raw_data)
rdd_2 = rdd_1.map(lambda x:(x[0], x[2]))
rdd_3 = rdd_2.combineByKey(lambda x : [x], lambda y, x : y + [x], lambda y1, y2 : y1 + y2)
rdd_4 = rdd_3.map(lambda x: (x[0], max(x[1]) + min(x[1])))
rdd_4.collect()
```

## Question 3: LSH

### Part (1)

From cos($\theta$(o,q))$\geq$0.9, we have that $\theta <$ arccos(0.9)$\approx$25.842°.
According to SimHash, we have that Pr[h$_i$(o)=h$_i$(q)].
Thus, P$_{q,o}$=Pr[h$_i$(o)=h$_i$(q)]1-$\theta/\phi$ > 0.856.
Simultaneously, the probability of find any near duplicate is 1 - (1-P$^k_{q,o}$)$^l$ Therefore, we can get 1-(1-0.856$^5$)$^l\geq$0.99, and the result is L$\geq$8

### Part (2)

As we known, false positive means that some Data should not become a candidate which is selected.
Since cos($\theta$(o,q))$\leq$0.8, this means that it is not a near duplicate. In order to become a false positive of query q, it is necessary that image o to be the candidate. Therefore, we can get $\theta$>36.8699° and P$_{q,o}$=Pr[h$_i$(o)=h$_i$(q)]1-$\theta/\phi$ < 0.796
Thus, we can get 1 - (1-P$^k_{q,o}$)$^l$ = 1-(1-0.795$^5$)$^l\leq$0.9782=97.82