How to find the median of a large dataset with Spark

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Anis Hamroun

Summary

What is a median ?

The **median** is the value separating the higher half from the lower half of a data sample.
wikipedia

Typical case

Let D such as D is included in R^N such as N isn’t considered as a “large number”. The typical algorithm is to sort the dataset D and return the $\frac{N}{2}$ ene number of the sorted dataset D.

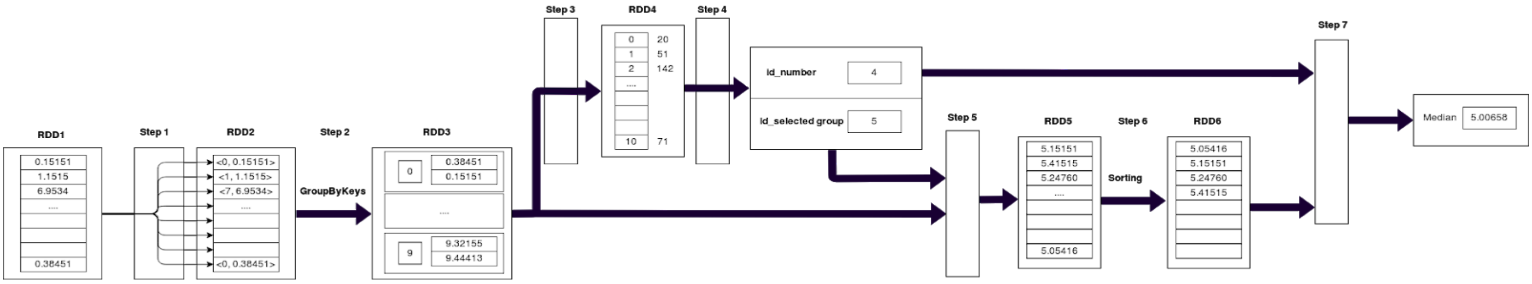
Pseudo Code

```
sorted_D = sort(D)

median = sorted_D[N/2]
```

The case of a Large Dataset

Let D such as D is included in R^N such as N is considered as a “large number”.
Assume D sorting takes too much time, then we have to find another manner to find the median.



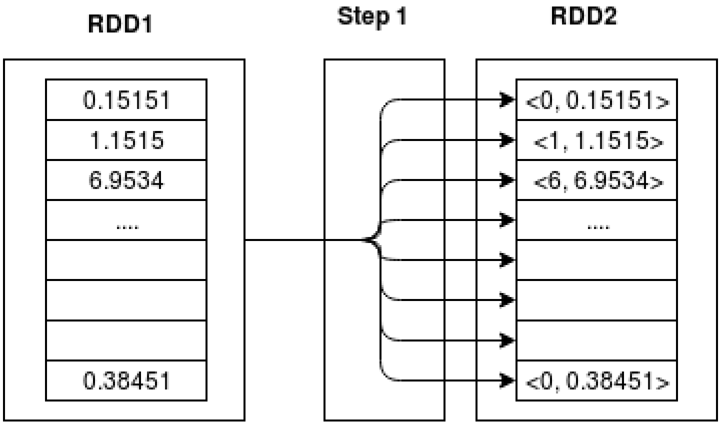
schema : Representation of median researching process

The main steps of the process :

- Assume a and b is included in D such as $\forall\ i \in [1;N]$, $a <= D[i] <= b$.
- Let Nb is the number of groups.
- We split the dataset D in several given the ascending order.
    - Let i an Integer such as Di is the iene group of D.
    - $\forall\ j \in [0;NB-1]$, $\forall\ u \in [0;\ len(Dj)\ -1\ ]$,
      $j\ = roundeddown(\ Dj[u]\ *(\frac{Nb}{b-a})\ -\ a\ *(\frac{Nb}{b-a}))$
- We have to find u such as $\sum_{j=0}^{u-1}$ len(Dj) < N/2 < $\sum_{j=0}^{u}$ len(Dj)
- that mean, the median is included in the group Du
- Then, we have to sort Du.
- Let Du’ = sort(Du)
- the median is Du’[$\frac{N}{2}\ -\ \sum_{j=0}^{u-1} len(Du)$]

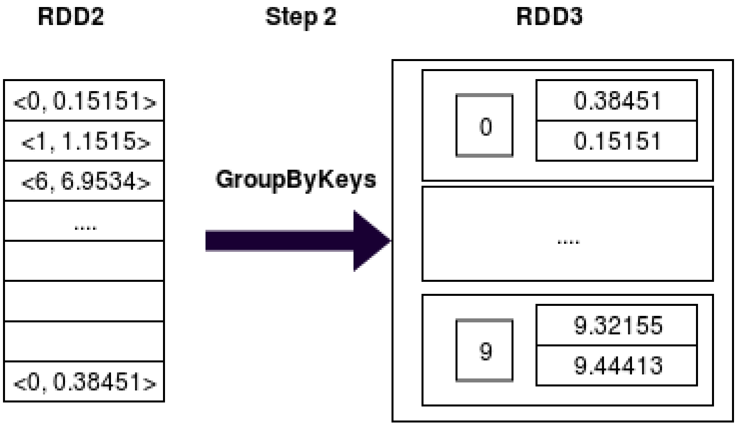
Detailed Processus

Step 01 :



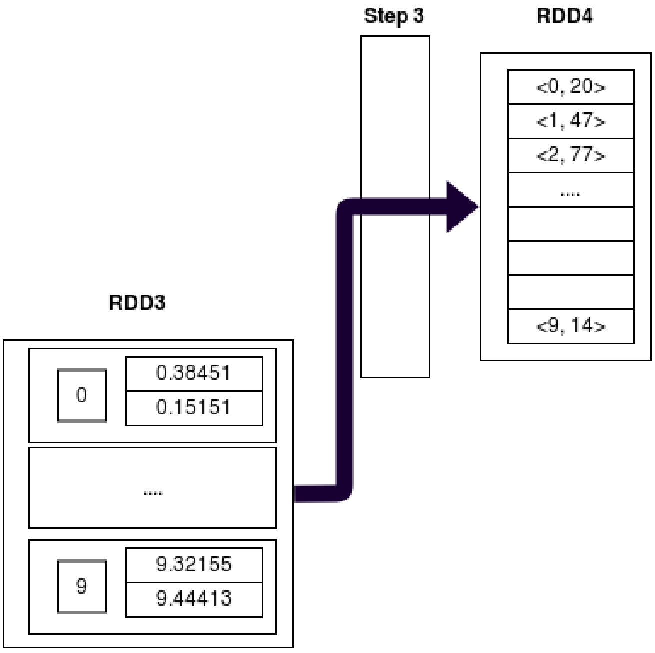
Map the RDD1 values in key-value pairs such as, for each key-value pair $key = roundeddown(Value * (\frac{Nb}{b-a}) - a * (\frac{Nb}{b-a}))$. With a and b respectively the minimum and the maximum of the RDD1.

Step 02 :



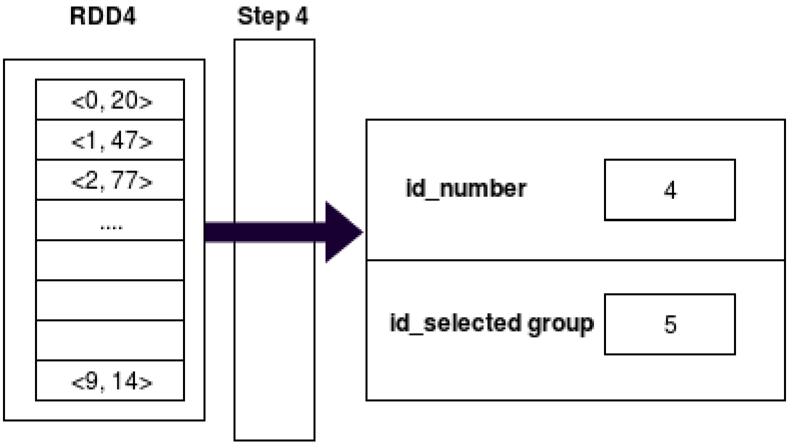
RDD2 is grouped by keys.

Step 03 :



During the Step 3, we create a RDD4 from RDD3 such as, let k a key such as, RDD3[k] is the list of values having for key k. So, RDD4[k] = len(RDD3[k])

Step 04 :

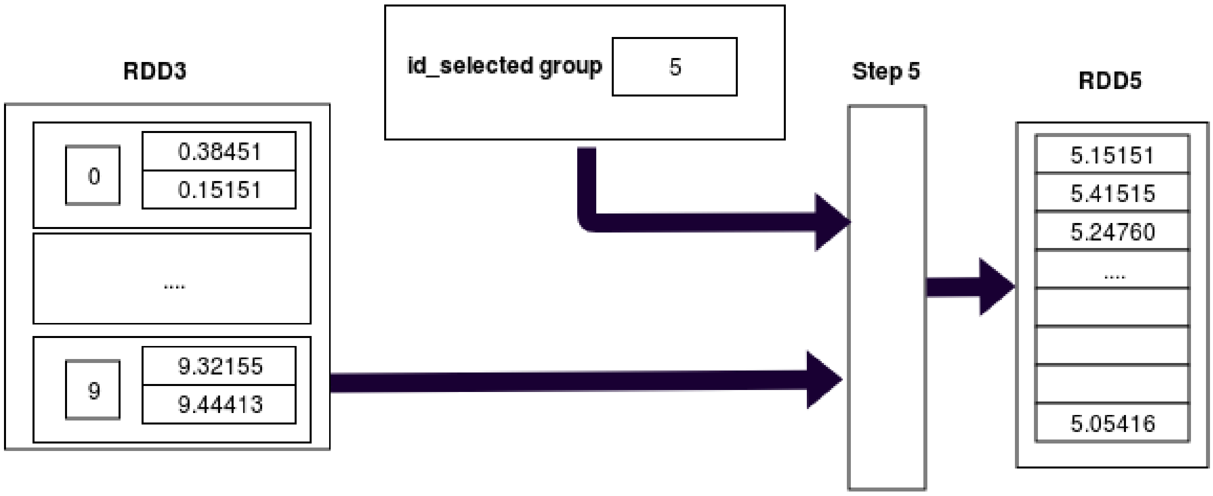


During the Step 04, we have to find u such as $\sum_{j=0}^{u-1}$ len(Dj) < N/2 < $\sum_{j=0}^{u}$ len(Dj).

u is the id_selected_group. That means, the median is included in the u-ene group.

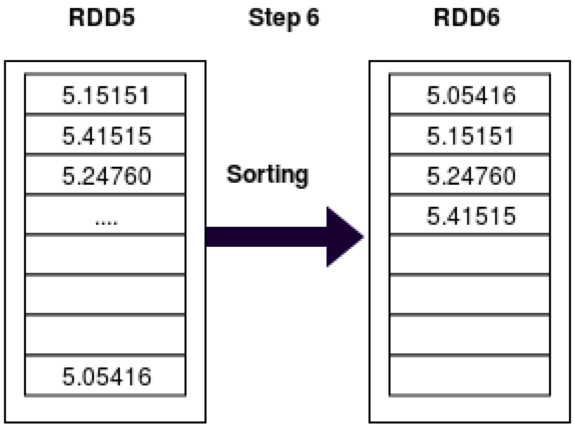
Additionnally, id_number = $\frac{N}{2} - \sum_{j=0}^{u-1} len(Du)$. id_number is the position of the median in Du', with Du' = sort(Du).

Step 05 :



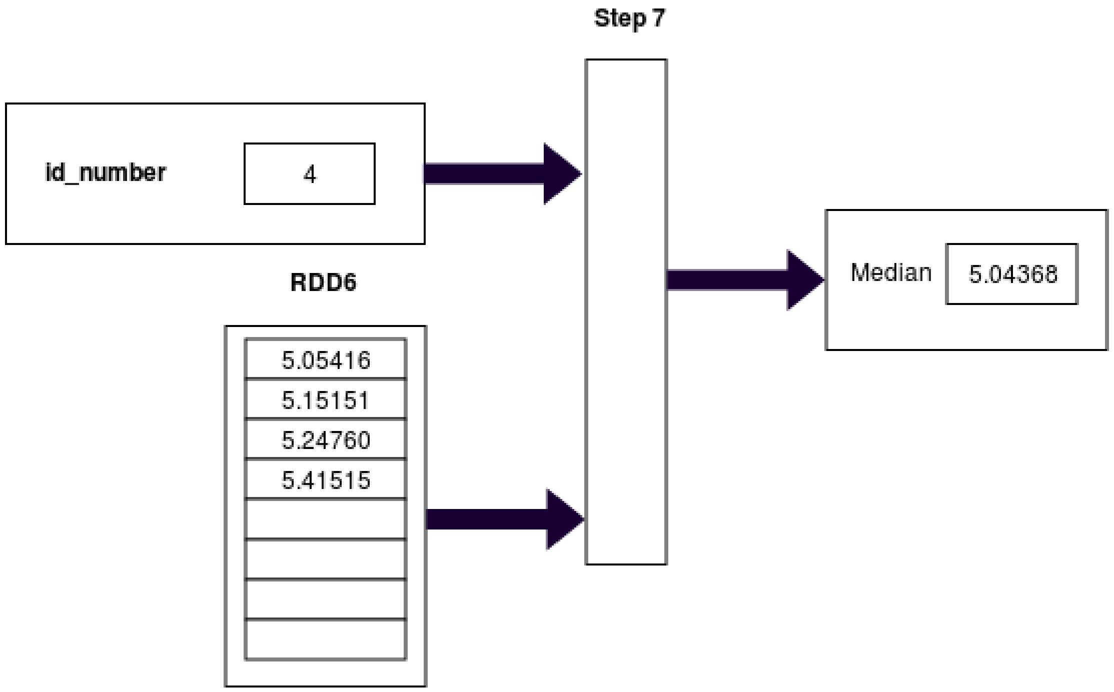
During this step, we choose the selected group.

Step 06 :



RDD6 equals to RDD5 after a sorting process.

Step 07 :



The Step 07 return the median which equals to Du'[$\frac{N}{2} - \sum_{j=0}^{u-1} len(Du)$]

Answers :

With Default configuration.

Data-1-sample.txt	Data-1.txt
median = 50.64663482	median = 50.00685338