The problem statement given to me was of a Call Data Record (CSV) File. I had to find out top 10 customers facing frequent call drops. So I used the following approach:

1. Load Data into HDFS
2. Create a key value paired Rdd where key is Phone\_no, value is Call Duration
3. Find per Key avg using Combine By Key() in Scala
4. Sort the data by values
5. Display top 10 records

The Complete code is given as:

1. val data = sc.textFile(“hdfs://localhost:9000/data/CDR.csv”)
2. val trandata = data.map(\_.split(“,”))
3. val keyvalue = trandata.map(x => (x(2),x(1).toInt))
4. val scorecom = keyvalue.combineByKey(
5. (v) => (v,1),
6. (acc: (Int,Int),v) => (acc.\_1 + v,acc.\_2 + 1),
7. (acc1: (Int,Int), acc2: (Int,Int)) => (acc1.\_1 + acc2.\_1,acc2.\_2 + acc2.\_2))
8. val scorecom1 = scorecom.map{ case(map,value) => (key,value.\_1 / value.\_2.toFloat) }
9. scorecom2 = scorecom1.sortBy(\_.\_2)
10. score2.take(10)

The last command displays top 10 customers facing frequent call drops.











