Q1​.Create an RDD (Resilient Distributed Dataset) named ​pagecounts​ from the input file.

Ans:val pagecounts = sc.textFile("/home/knoldus/Documents/spark\_session")

pagecounts: org.apache.spark.rdd.RDD[String] = /home/knoldus/Documents/spark\_session MapPartitionsRDD[4] at textFile at <console>:27

Q2​.Get the 10 records from the data and write the data which is getting printed/displayed.

Ans:pagecounts.take(10)

res12: Array[String] = Array(aa 112\_f.Kr 1 4606, aa 2,4-Dinitrophenylhydrazine/en/Brady%27s\_test 1 4680, aa 439\_f.Kr 1 4605, aa Celestino\_Marchant 1 4611, aa Eduard\_O%E2%80%99Rourke/de/Eduard\_O%27Rourke 1 4688, aa File:Douris\_Man\_with\_wax\_tablet.jpg 1 7848, aa File:UpdatedPlanets2006.jpg 1 8839, aa File:Wikipedia\_h2g2.jpg 1 8838, aa Islam\_in\_Tunisia/en/Shi%27a\_Islam\_in\_Tunisia 1 4675, aa Kk/

Special:Imagelist 1 4613)

Q3​.How many records in total are in the given data set ?

Ans:val pagecounts = sc.textFile("/home/knoldus/Documents/spark\_session")

pagecounts.count

res0: Long = 7598006

Q4​.The first field is the “project code” and contains information about the language of the pages. For example, the project code “en” indicates an English page. Derive an RDD containing only English pages from pagecounts​

Ans:val projectCode=pagecounts.map(line => line.split(" ")(0))

projectCode: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[7] at map at <console>:29

val enLanguageRDD=projectCode.filter(word=>word.contains("en"))

enLanguageRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[5] at filter at <console>:31

Q5​.How many records are there for English pages?

ans:enLanguageRDD.count

res4: Long = 2389787

Q6.Find the pages that were requested more than 200,000 times in total.

ans: