Question 2

1. **Create a function that given an RDD and a field (e.g. download_id), it computes an inverted index on the RDD for efficiently searching the records of the RDD using values of the field as keys?**
   Method:- We first create an RDD that is a collection of *case classes* that represent a line in the log file. Then we use HashSet to store the previous RDD as (key, HashSet). Here, key corresponds to the element of the given field and HashSet consists of all the entries with the corresponding key.

2. **Compute the number of different repositories accessed by the client 'ghtorrent-22' (without using the inverted index).**
   Answer:- 4601
   Method:- We create an RDD that is a collection of *case classes* that represent a line in the log file. Then filter those messages whose *download_id* is ghtorrent-22 and use MapReduce with the repository name (from *rest*) as the *key* and calculate the number of different repositories.

3. **Compute the number of different repositories accessed by the client 'ghtorrent-22' (with using the inverted index).**
   Answer:- 4601
   Method:- We use the Inverted Index computed in Q1 with the field as *download_id* and calculate the number of elements in the HashSet with ghtorrent-22 as key.