1. What is the purpose local or local[\*] in a scala code ?

Local sets the spark configuration to use the processers in the local machine. The [\*] tells spark to use all processers in the machine or you can define how many you would like spark to you use by entering a number.

2. What is the difference between map and flatMap ?

Map returns a new RDD by applying the functions passed through it to each element of the RDD. FlatMap does the same thing except the output it produces is more flattened. It can take one element and return 0, 1 or more elements back

3. What are two fundamental methods of creating a RDD in spark ?

You can reference a dataset in an external storage system such as HDFS and you can create a new RDD from an existing RDD by applying some sort of transformation and then calling an action.

4. What is the difference between map and mapValues ?

mapValues is only applicable for PairRDDs, meaning RDDs of the form RDD[(A, B)]. In that case, mapValues operates on the value only (the second part of the tuple), while map operates on the entire record (tuple of key and value).

5. What transformation helps us summarize data according to a key?

reducebyKey will assign key : value pairs to the data.