General Data Engineer Tech Test

Please submit the solution as a link to a GitHub repository.

Take Home Test

A directory in S3 contains files with two columns

1. The files contain headers, the column names are just random strings and they are not consistent across files
2. both columns are integer values
3. Some files are CSV some are TSV - so they are mixed, and you must handle this!
4. The empty string should represent 0
5. Henceforth the first column will be referred to as the key, and the second column the value
6. For any given key across the entire dataset (so all the files), there exists exactly one value that occurs an odd number of times. . E.g. you will see data like this:

```
// value 2 occurs odd number of times
1 -> 2
1 -> 3
1 -> 3

// value 4 occurs odd number of times
2 -> 4
2 -> 4
2 -> 4
```

   But never like this:

```
// three values occur odd number of times
1 -> 2
1 -> 3
1 -> 4

// no value for this key occurs odd number of times
2 -> 4
2 -> 4
```

Write an app in Scala that takes 3 parameters:

- An S3 path (input)
- An S3 path (output)
- An AWS `~/.aws/credentials` profile to initialise creds with, or param to indicate that creds should use default provider chain. Your app will assume these creds have full access to the S3 bucket.

Then in spark local mode the app should write file(s) to the output S3 path in TSV format with two columns such that

- The first column contains each key exactly once
- The second column contains the integer occurring an odd number of times for the key, as described by 6 above

NOTE: If your CV doesn't mention AWS you can assume local filesystem and not bother with the S3 stuff.

Bonus 1

This is a bonus and entirely optional. If you have already spent a long time on the above, it's recommended you skip this as we do not want to consume too much time of the candidates.

Provide more than one implementation of the algorithm, and as comments discuss the time & space complexity.

Bonus 2

This is a bonus and entirely optional. If you have already spent a long time on the above, it's recommended you skip this as we do not want to consume too much time of the candidates.

Instead of running the app in local mode, use the Java SDK for AWS `"com.amazonaws" % "aws-java-sdk" % "1.11.698"`, to run the app in an EMR cluster. Here the creds passed in are assumed to be admin creds for the AWS account.