1) Explain the differences between RDD and a traditional Relational Database System.

Resilient Distributed Dataset (RDD) vs Traditional Database.

RDD	**Traditional Database**
• RDD is an immutable distributed collection data. • RDD data will be portioned across nodes in cluster. • RDD is one of the use facing API in spark.	• The files are mutable • Traditional data is stored in fixed format or fields in a file • Can have you can have both fast reads and writes

Common Use cases of RDD-

RDD	**Traditional Database**
• Low level transformations, actions and control over data • For unstructured data like media files or stream of text • Not required a schema or columnar format	• For structured data

RDD is more like views in database and persistent RDDs are more like materialized views. RDMS are typically allow fine-grained read, write access to various entities.

2) Using pyspark create a word count application of all the words of the file **assignment_2_datafile.txt**(located in the files/datasets tab). Avoid counting trivial words such as vowels and pronouns.

In [32]:

```
from pyspark import SparkContext, SparkConf
import re
import sys

def normalizeWords(text):
    words = re.sub(r'[^A-Za-z0-9 ]','',text).lower().strip().split()
    return [w for w in words if len(w) > 3 and w !='']

if __name__ == '__main__':
    filename = "C:/Users/himan/Downloads/Ramp-Course/spark/assignment_2_datafile.txt"
    input = sc.textFile(filename)
    words = input.flatMap(normalizeWords)
    word_counts = words.map(lambda x: (x, 1)).reduceByKey(lambda x, y: (x + y))
    results = word_counts.collect()
    print(results)
```

```
[('1604', 1), ('tragedy', 3), ('prince', 5), ('denmark', 23), ('shakespeare', 1), ('dramatis', 1), ('king', 196), ('office
r', 3), ('nephew', 3), ('polonius', 36), ('chamberlain', 1), ('horatio', 47), ('courtier', 10), ('guildenstern', 30), ('gentl
eman', 16), ('soldier', 3), ('reynaldo', 6), ('players', 23), ('ambassadors', 6), ('getrude', 1), ('queen', 120), ('opheli
a', 29), ('daughter', 16), ('ghost', 34), ('hamlets', 10), ('father', 52), ('lords', 7), ('ladies', 4), ('officers', 2), ('so
ldiers', 7), ('messengers', 2), ('attendants', 9), ('scene', 24), ('elsinore', 23), ('platform', 4), ('before', 22), ('ente
r', 73), ('sentinelsfirst', 1), ('paces', 1), ('down', 29), ('approaches', 1), ('there', 77), ('fran', 8), ('stand', 15), ('u
nfold', 3), ('long', 17), ('live', 16), ('carefully', 1), ('upon', 55), ('twelve', 5), ('this', 300), ('thanks', 10), ('col
d', 6), ('heart', 29), ('have', 183), ('quiet', 5), ('guard', 4), ('mouse', 2), ('stirring', 1), ('good', 109), ('night', 3
6), ('meet', 9), ('rivals', 1), ('watch', 15), ('them', 74), ('make', 55), ('haste', 12), ('think', 47), ('hear', 32), ('grou
nd', 9), ('give', 60), ('farewell', 17), ('holla', 1), ('again', 34), ('seen', 22), ('says', 8), ('belief', 1), ('take', 3
6), ('touching', 3), ('dreaded', 1), ('entreated', 1), ('minutes', 1), ('apparition', 2), ('eyes', 23), ('speak', 63), ('appe
ar', 3), ('awhile', 10), ('once', 19), ('assail', 1), ('ears', 9), ('against', 24), ('story', 3), ('nights', 4), ('last', 1
4), ('when', 56), ('yond', 1), ('pole', 2), ('illume', 1), ('where', 54), ('burns', 2), ('myself', 12), ('beating', 3), ('pea
ce', 10), ('look', 37), ('figure', 5), ('like', 81), ('thou', 105), ('looks', 8), ('mark', 15), ('fear', 20), ('wonder', 3),
('would', 81), ('spoke', 2), ('question', 16), ('usurpst', 1), ('form', 13), ('sometimes', 3), ('charge', 8), ('away', 25),
('pale', 8), ('something', 14), ('more', 95), ('than', 50), ('believe', 19), ('sensible', 1), ('true', 22), ('avouch', 1),
('very', 66), ('parle', 1), ('polacks', 1), ('strange', 9), ('thus', 43), ('jump', 2), ('martial', 1), ('stalk', 1), ('partic
ular', 8), ('thought', 11), ('work', 10), ('know', 74), ('gross', 3), ('scope', 3), ('opinion', 1), ('state', 15), ('tell', 4
3), ('nightly', 1), ('subject', 3), ('cast', 6), ('cannon', 4), ('foreign', 1), ('mart', 1), ('implements', 1), ('impress',
```