**HDFS - Task 1**

**HDFS:**

* Hadoop distributed file system
* written in java
* store files across machine in a cluster

**Avro:**

Provide container files to store persistent data, remote procedure call, reach data structure,

It's compact, fast, binary data format.

**Dependencies Used in project:**

1. **Spark-core:**

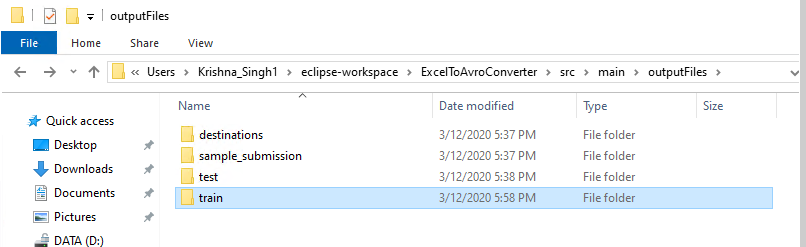
It is the fundamental unit of the whole Spark project. It provides all sort of functionalities like task **dispatching**, **scheduling**, and **input-output** operations etc. Spark makes use of Special data structure known as RDD (Resilient Distributed Dataset). It is the home for API that defines and manipulate the RDDs.

1. **Spark-sql:**

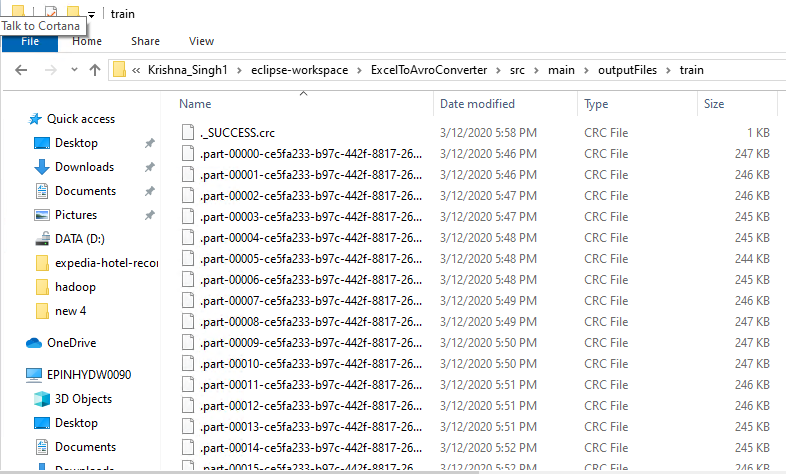
It is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It also provides powerful integration with the rest of the Spark ecosystem.

1. **Spark-Avro:** To support Avro file format.
2. **JUnit-Jupiter-API:** Used to write test cases for this project.
3. **Junit-Jupiter-Engine:** Used to write test cases for this project.

**Output Avro Files:**



**Files are automatically divided into sub-files:**



**Avro Files are encrypted:**

