# **Week\_2 Report**

**MongoDB to Pyspark connection:**

To connect Pyspark with MongoDB below jar file is needed.

* Downloaded the jar file from the link below and placed it in the python file path.

<https://repo1.maven.org/maven2/org/mongodb/spark/mongo-spark-connector_2.12/3.0.1/mongo-spark-connector_2.12-3.0.1.jar>

**MongoDB Installation**:

* Installed Community edition of MongoDB using link below:

<https://www.mongodb.com/try/download/community>

**Python Installation**:

* pip install pyspark

**Challenges Faced:**

* py4j.protocol.Py4JJavaError: An error occurred while calling o39.load.

: java.lang.ClassNotFoundException: Failed to find data source: mongo.

* py4j.protocol.Py4JJavaError:

An error occurred while calling o39.load. : java.lang.NoClassDefFoundError: com/mongodb/connection/DefaultClusterFactory

**Resolutions:**

Downloaded the below jar files.

Jars:

<https://repo1.maven.org/maven2/org/mongodb/mongo-java-driver/3.12.10/mongo-java-driver-3.12.10.jar>

<https://repo1.maven.org/maven2/org/mongodb/mongodb-driver-core/4.3.1/mongodb-driver-core-4.3.1.jar>

<https://repo1.maven.org/maven2/org/mongodb/bson/4.3.1/bson-4.3.1.jar>

**Week\_2 Output:**

1. Using Mongo spark connector, we have connected to the database in order to fetch the raw data.
2. Selected the columns which are required for the model prediction.
3. Removed the rows which has null values.







