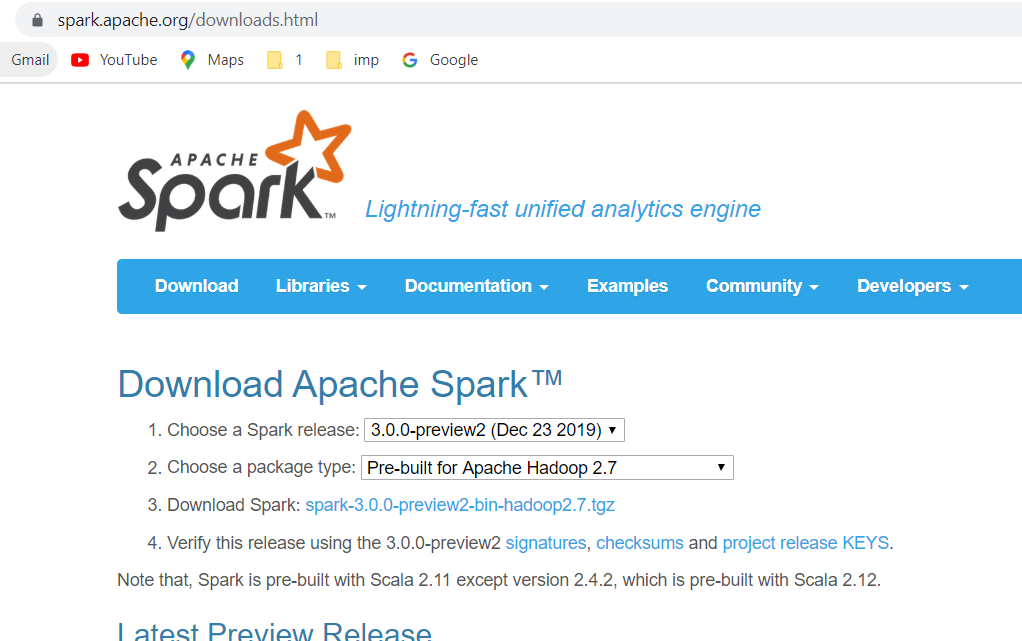
**Useful information Spark**

1. Download Spark from below url.

<https://spark.apache.org/downloads.html>



1. Run the spark-shell.cmd from bin folder to start the Spark.
2. Download winutils.exe from <http://public-repo-1.hortonworks.com/hdp-win-alpha/winutils.exe>.
3. SetUp your HADOOP\_HOME environment variable on the OS level or programmatically:

//System.setProperty("hadoop.home.dir", "full path to the folder with winutils");

HADOOP\_HOME=D:\Practice\spark-scala\{bin\winutil.exe}

Important points :

* Sc (spart Context ),spark(Spark Session)

Scala>:help

scala> :type sc

org.apache.spark.SparkContext

scala> :type spark

org.apache.spark.sql.SparkSession

scala> :history

1 jhj

2 :help

3 type sc

4 type:sc

5 :type sc

6 :type spark

7 :history

**RDD:**

1. RDD Immutable objects
2. It supports any language (scala,java,pythom)
3. Creation of RDD.
4. **Parallelize method**

* scala> val intArry = Array(41,56,85,9,85)

intArry: Array[Int] = Array(41, 56, 85, 9, 85)

scala> val intRDD = sc.**parallelize**(intArry)

intRDD: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[5] at parallelize at <console>:26

Or

scala> val intRDD = sc.parallelize(intArry,2)

intRDD: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[7] at parallelize at <console>:26

* **Reading data:**

scala> intRDD.first()

res6: Int = 41

scala> intRDD.take(3)

res7: Array[Int] = Array(41, 56, 85)

scala> intRDD.collect()

res10: Array[Int] = Array(41, 56, 85, 9, 85)

scala> intRDD.collect().foreach(println)

41

56

85

9

85

scala> intRDD.**partitions**.size

res14: Int = 8

1. **Using Text file**