**Big data tools and technologies:**

* ­Big data analytics requires higher computational power than the one available on our local systems. To avail this additional power, we can set up an AWS account and create an EC2 (Elastic Compute Cloud) instance
* For our learning purpose, we can use the 12-month free access to EC2 system (with limited RAM). This allows us to create a virtual environment from the distributed systems and can be used to run our analysis
* Apache Hadoop ecosystem allows us to distribute very large files across different machines. This is usually done through the HDFS i.e. Hadoop Distributed File System
* MapReduce is a programming model that can be used to perform large scale computations on a distributed data set. It is a way of splitting a computation task to a distributed set of files. It consists of a job tracker and multiple task trackers
* Apache Spark can be thought of as a better programming model than MapReduce since it does not specifically require files to be stored in HDFS (which is compulsory for MapReduce). Spark can also perform operations much faster than MapReduce since Spark keeps most of the data in the memory after transformation. Whereas MapReduce constantly writes data to disk after each map which eventually slows it down
* At the core of Apache Spark is the idea of a RDD i.e. Resilient Distributed Dataset (RDD). It has 4 main features – distributed collection of data, fault-tolerant, parallel operations and ability to use many data sources
* **Setup reference:** <https://medium.com/@josemarcialportilla/getting-spark-python-and-jupyter-notebook-running-on-amazon-ec2-dec599e1c297>

**Spark:**

* Actions are spark operations that produce a local object
  + Examples – first (), collect (), count (), take (), top (), takeSample (), sum (), mean (), stdev (), Reduce ()
* Transformations are spark operations that produce an RDD (i.e. a set of instructions or recipe that produce an RDD output. But you cannot view the output until you apply an action function to this RDD object)
  + Examples- filter (), map (), flatMap (), sample (), union (rdd), distinct (), sortBy (), ReduceByKey () (ReduceByKey is similar to GroupBy of pandas)
* Spark job is a sequence of transformations on data with a final action i.e. a combination of reading/creating a file, transforming and running an action on it
* Creating an RDD –
  + sc.parallelize(array)
  + sc.textFile(path/to/file)