**Create a document and write a few lines about what you understand about the 4 different modes of spark. Include advantages and disadvantages.**

The 4 different modes of spark are:

1. Spark Client Mode
2. Spark Cluster Mode
3. Spark Standalone mode
4. Spark Local Mode

**Spark Client Mode**

The spark client mode is a spark mode of deployment when the driver of a spark job runs on the local machine from which the job is submitted.

Advantage:

* If the machine that submitted the job is not remote, the network bandwidth between the Spark Infrastructure and where the Spark Job Driver runs is very high
* Good for debugging and testing

Disadvantage:

* If the machine submitting the spark job is remote from Spark Infrastructure, the network bandwidth between the Spark Infrastructure and the job driver will be low. In other word, it will result in high network latency

**Spark Cluster Mode**

The Spark Cluster Mode is another spark mode of deployment. It is defined by the behaviour of spark job driver and worker node in a spark cluster. The job driver runs on a worker node which is within a spark cluster.

Advantages:

* Since the spark job driver is running within the cluster, the rate of data transfer will be high. In other word, the network latency is low.
* There is low chance of network failure which may result in abrupt disruption in the driver operation
* If the local machine is shutdown, the cluster would not be affected since the driver is running in the cluster
* Good for running production job

Disadvantage:

* Not the best choice mode for debugging or testing

**Spark Standalone Mode**

Spark Standalone Mode is spark mode that has its own cluster. In simple term, Spark has a cluster manager. So, a cluster can be setup that spark would manage. The cluster would have master and worker nodes with configured amount of memory and CPU cores. Resources would be allocated base on the available cores.

Advantage:

* A simple way to run Spark application in a cluster environment

Disadvantage:

* Very rarely used because most organisation already have Hadoop and YARN installed for cluster management
* Storage is required but Spark has none

**Spark Local Mode**

In Spark Local Mode, the spark job is run in a local machine such as laptop or desktop. The Spark job driver and executor is created inside a single Java Virtual Machine (JVM) process.

Advantage:

* Spark-shell can be run using laptop or desktop
* Good for testing and debugging

Disadvantage:

* Low processing power
* Low security
* Storage limitation
* Horizontal scaling is not possible

**Shuffle operations in Spark.**

In Spark, large dataset can be partitioned across the spark cluster. The shuffle operation occours when particular data has to be shuffled across the cluster. For example, if there are 3 partitions across the cluster, each with its own executor, each partition might contain different group of data. So, if partition 1 has data grouped by (a, b, c), partion 2 data (c, b, a), partion 3 data (b, a,c). The shuffle operation will make sure that partition 1 will have only single group of data, either a, b or c, and the same operation will be applied to partition 2 and 3 across the cluster. What shuffle operation is doing is basically redistributing the data across different executor within the cluster

**What happens in terms of memory when you are using shuffling?**

In Spark, there is a term called RDD (Resilient Distributed Datasets). RDD is a collection of Read-Only or Immutable object which are partioned across the cluster. Spark can run the map task on all partitions and group the data in memory. If the data is bigger than the memory, spark would use the available disk space to store the rest of the data or map task. The shuffle operation on memory is usually faster. Another important point to note during shuffling is the caching operation. This is the process in which spark stores the shuffled data for reuse