# **ITCS 3190 Assignment 4**

Total 100 points

*Due: April 19, 2020. No late submission accepted*

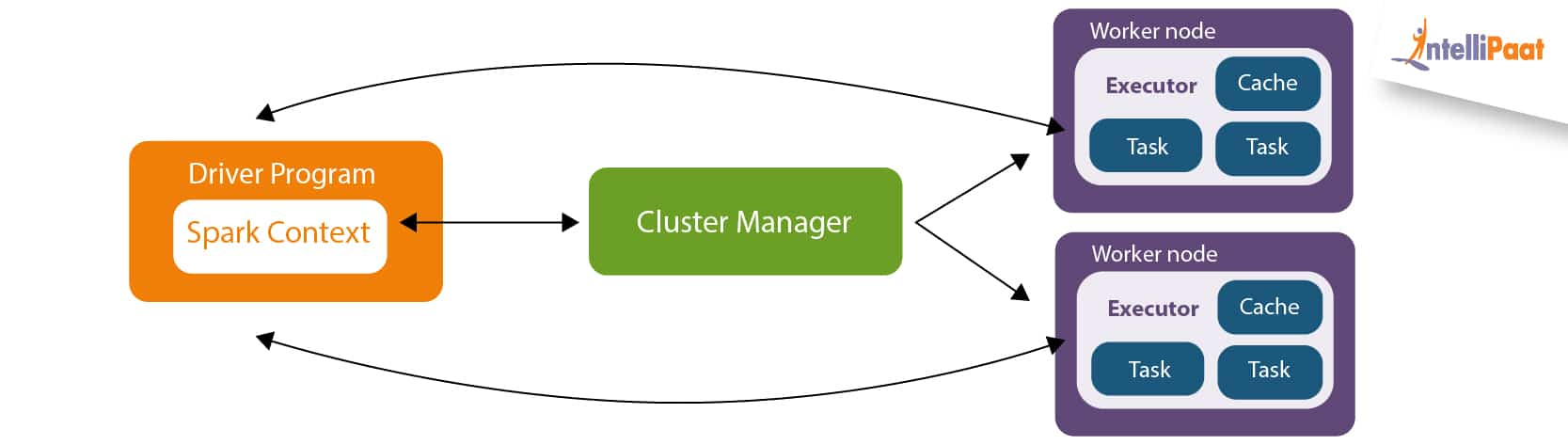
**Please refer to the slides and relative videos to finish this assignment**

1. (25 pts) Why data flow engine is used in clusters?

Data flow engines provide high level functions instead of message passing, are more common than MPI, and are scalable to vary large clusters

1. (25 pts) List two limitations of MapReduce.
2. Inefficient for multipass algorithms
3. Not efficient for data sharing

1. (25 pts) Draw the architecture of Spark and state how it works.



A driver program in Spark calls the main program of an application to create a Spark Context that consists of basic functionalities. The Spark Driver works with the Cluster Manager for various other jobs in addition to watching over the overall job execution with Spark Context. The Cluster Manager does resource allocation, and then splits the job into multiple smaller tasks to be distributed to worker nodes. These worker nodes execute tasks given by the Cluster Manager and return back to the Spark Context. Within each worker node, an executor is responsible for the execution of the task given to the worker.

1. (25 pts) What is Resilient Distributed Dataset (RDD) and how does it guarantee fault tolerance?

A resilient Distributed Dataset (RDD) is a read-only collection of elements that can be used on many devices at the same time. Each RDD can be divided into logical portions to be executed on different nodes of a cluster.

RDDs guarantee fault tolerance by keeping track of the history of the operations that have been used on the dataset to create it. If any partition of an RDD is lost, that partition can be recreated from the original dataset using the saved history of operations.