ASSESSMENT

NAME-SWARNA LATHA GUVVALA

DATE-22/12/2023

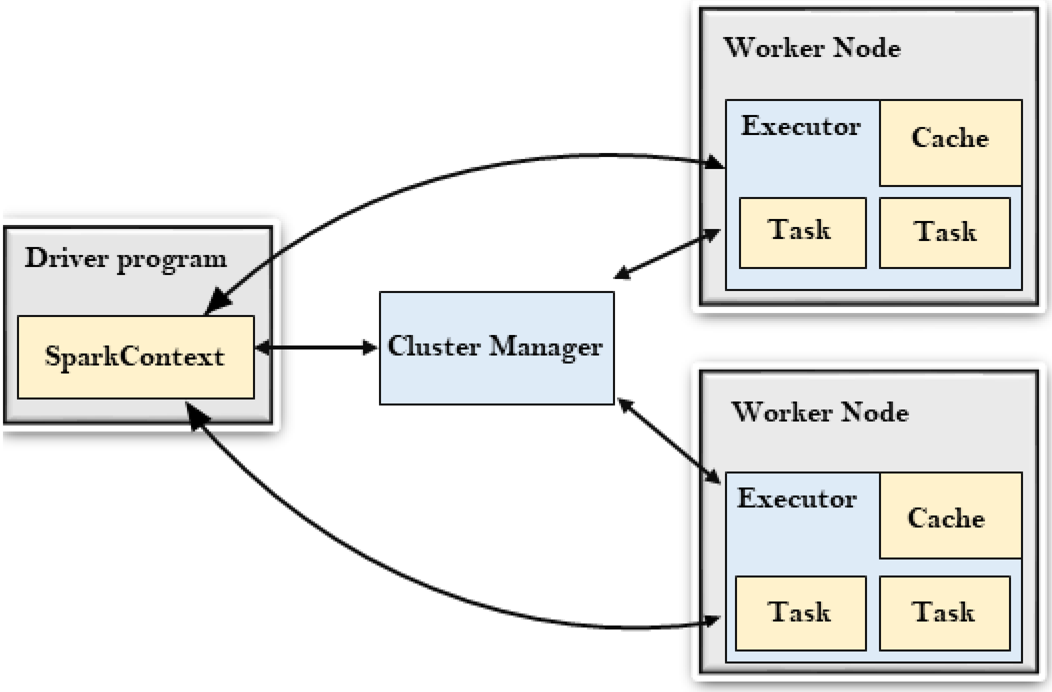
APACHE SPARK – INTRODUCTION:

Spark was introduced by Apache Software Foundation for speeding up the Hadoop computational computing software process.

Spark is not a modified version of Hadoop and is not, really, dependent on Hadoop because it has its own cluster management. Hadoop is just one of the ways to implement Spark.

Spark uses Hadoop in two ways – one is **storage** and second is **processing**. Since Spark has its own cluster management computation, it uses Hadoop for storage purpose only.

APACHE SPARK ARCHITECTURE:



DRIVER PROGRAM:

The Driver Program is a process that runs the main() function of the application and creates the **SparkContext** object. The purpose of **SparkContext** is to coordinate the spark applications, running as independent sets of processes on a cluster.

To run on a cluster, the **SparkContext** connects to a different type of cluster managers and then perform the following tasks: -

- It acquires executors on nodes in the cluster.
- Then, it sends your application code to the executors. Here, the application code can be defined by JAR or Python files passed to the SparkContext.
- At last, the SparkContext sends tasks to the executors to run.

CLUSTER MANAGER:

- The role of the cluster manager is to allocate resources across applications. The Spark is capable enough of running on a large number of clusters.
- It consists of various types of cluster managers such as Hadoop YARN, Apache Mesos and Standalone Scheduler.
- Here, the Standalone Scheduler is a standalone spark cluster manager that facilitates to install Spark on an empty set of machines.

WORKER NODE:

- The worker node is a slave node
- Its role is to run the application code in the cluster.

EXECUTOR:

- An executor is a process launched for an application on a worker node.
- It runs tasks and keeps data in memory or disk storage across them.
- It read and write data to the external sources.
- Every application contains its executor.

TASK:

- A unit of work that will be sent to one executor.

Following are the website links for the Spark API for each of these languages.

- **Scala API**
- **Java**
- **Python**

CLUSTER MANAGER TYPES:

The system currently supports several cluster managers:

- Standalone – a simple cluster manager included with Spark that makes it easy to set up a cluster.
- Apache Mesos – a general cluster manager that can also run Hadoop MapReduce and service applications. (Deprecated)
- Hadoop YARN – the resource manager in Hadoop 3.
- Kubernetes – an open-source system for automating deployment, scaling, and management of containerized applications.