### 2.3.3 Hadoop

**Apache Hadoop**

It is an open source framework for distributed batch processing (MapReduce) of big data.

**Hadoop YARN**

- It is the next generation architecture of Hadoop.
- The original processing engine of Hadoop (MapReduce) has been separated from the resource management component (which is now part of YARN).
- YARN is like operating system for Hadoop that supports different processing engines on a Hadoop cluster such as MapReduce for batch processing, Apache Tez for interactive queries, Apache Storm for stream processing.

**Resource Manager (RM)**

It manages the global assignment of compute resources (resources) to applications. RM consists of two main services:

- Scheduler: It is a pluggable service that manages and enforces the resource scheduling policy in the cluster.
- Applications Manager (AsM): AsM manages the running Application Masters in the cluster. AsM is responsible for starting application masters and for monitoring and restarting them on different nodes in case of failures.

**Application Master (AM)**

It manages the application’s life cycle. AM is responsible for negotiating resources from the RM and working with the NMs to execute and monitor the tasks.

**Node Manager (NM)**

A per-machine NM manages the user processes on that machine.

**Containers**

- Container is a bundle of resources allocated by RM (memory, CPU and network).
- It is a conceptual entity that grants an application the privilege to use a certain amount of resources on a given machine to run a task.
- Each node has an NM that spawns multiple containers based on the resource allocations made by the RM.

**Hadoop Schedulers**

- FIFO scheduler maintains a work queue in which the jobs are queued. The scheduler pulls jobs in first-in first-out manner (oldest job first) for scheduling. There is no concept of priority
- Fair Scheduler assigns resources to jobs such that each job gets an equal share of the available resources on average over time. Unlike the FIFO scheduler, which

forms a queue of jobs, the Fair Scheduler lets short jobs finish in reasonable time while not starving long jobs.

- In Capacity Scheduler, multiple named queues are defined, each with a configurable number of map and reduce slots. The Capacity Scheduler gives each queue its capacity when it contains jobs, and shares any unused capacity between the queues. Within each queue FIFO policy is used.