***What is Hadoop?***

Formally speaking, Hadoop is an open source framework for writing and running distributed

applications that process large amounts of data. Distributed computing is a wide and varied

field, but the key distinctions of Hadoop are that it is

 *Accessible*—Hadoop runs on large clusters of commodity machines or on cloud

computing services such as Amazon’s Elastic Compute Cloud (EC2).

 *Robust*—Because it is intended to run on commodity hardware, Hadoop is architected with the assumption of frequent hardware malfunctions. It can gracefully handle most

such failures.

 *Scalable*—Hadoop scales linearly to handle larger data by adding more nodes to the

cluster.

 *Simple*—Hadoop allows users to quickly write efficient parallel code.

Hadoop’s accessibility and simplicity give it an edge over writing and running large

distributed programs. Even college students can quickly and cheaply create their own Hadoop

cluster. On the other hand, its robustness and scalability make it suitable for even the most

demanding jobs at Yahoo! and Facebook. These features make Hadoop popular in both

academia and industry.

Figure 1.1 illustrates how one interacts with a Hadoop cluster. As you can see, a Hadoop

cluster is a set of commodity machines networked together in one location.**2**

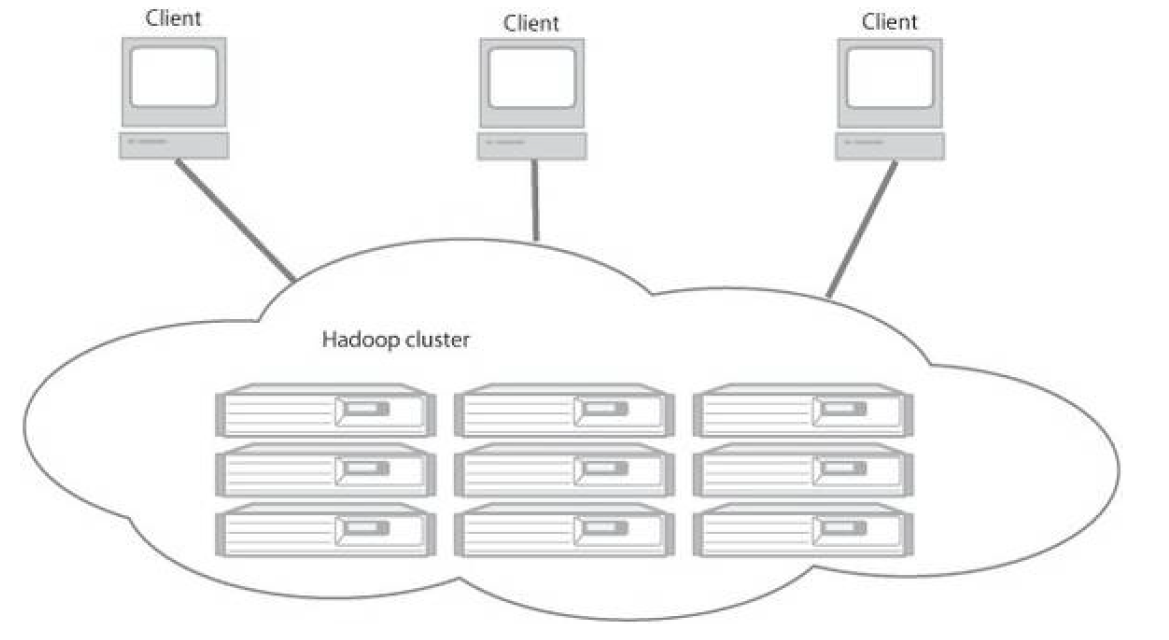


Figure 1.1 A Hadoop cluster has many parallel machines that store and process large data sets. Client

computers send jobs into this computer cloud and obtain results.

Data storage and processing all occur within this “cloud” of machines. Different users can

submit computing “jobs” to Hadoop from individual clients, which can be their own desktop

machines in remote locations from the Hadoop cluster.