About Hadoop

Table of contents

1. Overview

Hadoop is a framework for running applications on large clusters of commodity hardware. The Hadoop framework transparently provides applications both reliability and data motion. Hadoop implements a computational paradigm named map/reduce, where the application is divided into many small fragments of work, each of which may be executed or reexecuted on any node in the cluster. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both map/reduce and the distributed file system are designed so that node failures are automatically handled by the framework.

Hadoop has been demonstrated on clusters with 2000 nodes. The current design target is 10,000 node clusters.

Hadoop is a Lucene sub-project that contains the distributed computing platform that was formerly a part of Nutch. This includes the Hadoop Distributed Filesystem (HDFS) and an implementation of map/reduce.

For more information about Hadoop, please see the Hadoop wiki.