HadoopOverview

Overview of Hadoop

Hadoop is a collection of code libraries and programs useful for creating very large distributed systems. Much of the code was originally part of the Nutch search engine project.

Hadoop includes the following parts:

- conf, an assortment of classes for handling key-value pairs used in system configuration.
    - An HadoopMapReduce job is described with an XML Job Configuration File.
- DFS, the Hadoop Distributed Filesystem.
- io, an assortment of IO-related classes. Includes a compressed UTF8 string implementation, code for performing external sorts, and a "poor-man's B-Tree" implementation for looking up items in large key-value sets.
- ipc, a fast and easy remote procedure call system
- HadoopMapReduce, a distributed job allocation system built on top of DFS. It employs a MapReduce-like programming model