Seminar

Date: / /

Let's first understand what is ~~Hdfs~~ Hadoop.

Some operations run on a special file system called Google file systems that is highly optimized for this purpose, but these file systems are not open source. So a developer named Doug Cutting ~~with~~ and some ~~others~~ of his co-workers reverse engineered the GFS and called it Hadoop Distributed file System ~~or~~ Hadoop ~~This~~ is open source and distributed ~~by Apache~~ processing framework that manages data processing and storage for big data applicⁿ

~~HDFS is a file used in~~

This file system is used in ~~Hadoop based~~ distributed file system and it provides the base supports for the storage.

HDFS is a distributed file system that handles large data sets running on commodity hardwares.

~~It is used to scale a~~ It is the primary data storage system used by Hadoop applications which includes, stream processing, fraud detection and prevention, content management, financials sectors, healthcare sector and many more.

HDFS employs a NameNode and DataNode to implement distributed system that provides high-performance access to data across highly Hadoop clusters which is a collection of computers, So failure of at least one computers or server is inevitable. HDFS detects faults and automatically recover quickly.

HDFS have hundreds of nodes per cluster cluster to manage the huge datasets. and a requested task can be done efficiently, when computation takes place especially when there are huge datasets, it reduces the network traffic and increases the throughput.

HDFS is intended for more batch processing that interactive use, and it increases the throughput rate for high data which accomodate streaming access to data sets. also, It is designed to be portable across multiple hardware platforms and to be compatible with a variety of operating systems.

~~To~~ understand it ~~better we~~ can let's understand it better with an example,

Consider a file includes the phone numbers for everyone in a country,

The numbers for people with a last name starting with A might be stored on server 1, B on server 2 and so on

with Hadoop, piece of this ~~not~~ phonebook is stored across the cluster, and your program need the blocks from every server in the cluster to reconstruct the entire phonebook

If a server fails, HDFS replicates these smaller pieces onto two ~~smaller~~ additional servers by default., The redundancy can be increased or decrease as per file basis., This redundancy offers multiple benefits, the most obvious ~~being~~ will be high~~ly~~ availability.

Web-based companies like eBay, facebook, Linkedin and Twitter uses Hadoop file systems to manage pools of big data.