Hello, this is the topic video for the FIU 2017 Fall term Database class, my name is Huai, today I’d like to talk something about Hive

Story of Facebook data infrastructure team

Apache Hive is a data warehouse software project based on Apache Hadoop and HDFS for data query, summarization and analysis. Hive QL is a SQL-like declarative language to compile queries expression into Map-Reduce jobs executed on Hadoop. Four major part

Data warehouse

Data Warehouse is a large store of data accumulated from a wide range of sources within a company that support management decisions. Subject Oriented Integrated Non-Volatile Time Variant

Hadoop HDFS Mapreduce

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. Map-reduce

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

HiveQL no update delete individual records transactions limited index high latency much larger data size

Hive Architecture

Future Hbase spark tez