Summary of Hadoop/Hbase project

HBase (also known as Hadoop Database) is a distributed, scalable, and fault-tolerant data storage system. It is designed to handle large storage of structured data and is built on the Apache Hadoop platform. HBase is a NoSQL database.

HBase uses Apache Hadoop Distributed File System (HDFS) to store data. Data is distributed and replicated across nodes in the network to ensure high availability and reliability. HBase uses a column-family-oriented architecture and key-value data model.

Some characteristics of HBase:

* Horizontal Scalability: HBase is designed to scale horizontally easily, allowing additional servers to increase data storage and processing capacity.
* Flexible data structure: HBase uses a column-family-based database model.
* Consistency: HBase supports consistency in reading and writing data.
* Integration with Hadoop: HBase integrates tightly with Hadoop, uses Hadoop Distributed File System (HDFS) to store data and combines with Hadoop MapReduce to process data.

HBase architecture:

* HMaster: monitors all Regionservers in the cluster. All changes related to metadata are done through HMaster.
* Region Servers: managing and storing regions, HRegionServer is responsible for managing regions.
* ZooKeeper: stores metadata, region info.
* Hadoop Distributed File System (HDFS): HBase data is stored in HDFS, a distributed file system of the Apache Hadoop project.
* HBase Client: is the component responsible for interacting with HBase, including performing queries and writing data.

Hadoop is an open-source framework designed to process and store big data on a distributed computing cluster. It provides a distributed model for parallel data processing on multiple computers through splitting data and distributing them across nodes in the cluster.

Hadoop Distributed File System (HDFS) is the data storage system used by Hadoop. With the application of NameNode and DataNode architecture to implement a distributed file system. It provided high-performance access to data across Hadoop clusters.

* NameNode: manages file system metadata.
* DataNode: responsible for storing the actual data blocks on the hard disk.