Session17 Assignment 1- answers

1. Give a brief difference between HBASE and HDFS.

|  |  |
| --- | --- |
| HDFS is a distributed file system | HBase on the other hand is a distributed column oriented database. The filesystem of choice typically is HDFS owing to the tight integration between HBase and HDFS |
| 1. It is optimized for streaming access of large files. You would typically store files that are in the 100s of MB upwards on HDFS and access them through MapReduce to process them in batch mode. | 1. It gives you the ability to do random read/writes on your data which HDFS doesnt allow you to. |
| 2. HDFS is optimized for use cases where you write once and read many times like in the case of production logs. You can append to files in some of the recent versions but that is not a feature that is very commonly used. There is no concept of random writes. | 2. HBase stores data in the form of key value pairs in a columnar fashion. HBase provides a flexible data model. |
| 3. HDFS doesn’t do random reads very well. | 3. Fast scans across tables. |
|  | 4. Scale in terms of writes as well as total volume of data. |

1. List the main components of HBASE.

HBase Hmaster - Region assignment, DDL (create, delete tables) operations are handled by the HBase Master. responsible for: Coordinating the region servers and Admin functions

Regions and Region Server - HBase Tables are divided horizontally by row key range into “Regions.” A region contains all rows in the table between the region’s start key and end key. Regions are assigned to the nodes in the cluster, called “Region Servers,” and these serve data for reads and writes. A region server can serve about 1,000 regions.

ZooKeeper: The Coordinator - HBase uses ZooKeeper as a distributed coordination service to maintain server state in the cluster. Zookeeper maintains which servers are alive and available, and provides server failure notification. Zookeeper uses consensus to guarantee common shared state. Note that there should be three or five machines for consensus.

1. Does Hbase support sql?

Nope. HBase does not support SQL.

1. When should we use HBASE, list some of the scenarios for the same.

* If your application has a variable schema where each row is slightly different, then you should look at HBase. As an example, doing a modeling exercise using a standard relational schema; When you can’t add columns fast enough and most of them are NULL in each row, you should consider HBase.
* If you find that your data is stored in collections, for example some meta data, message data or binary data that is all keyed on the same value, then you should consider HBase.
* If you need key based access to data when storing or retrieving, then you should consider HBase

1. What are the different modes in which Hbase can be run?

HBase has two run modes: “Standalone HBase” and “Distributed”

1. Why is zookeeper needed in Hbase?

* HBase uses Zookeeper extensively for region assignment.
* Zookeeper has simple APIs written in java
* Directories and files in Zookeeper are called Znodes.
* Clients connect to Zookeeper via a session.
* Zookeeper is used to ensure only one Master is assigned.
* HBase cannot operate without Zookeeper.

1. Hbase is a schema less database, what does it mean?

It doesn't have the concept of fixed columns schema; defines only column families. The column names can be completely variable and the number of columns can vary by row – so you could have a table with billions of rows and could have rows with 5 or 5 million columns.

1. What is the minimum number of column family every Hbase table should have?

One

1. What is the benefit of using connection pool in Hbase?

For applications which require high-end multithreaded access (e.g., web-servers or application servers that may serve many application threads in a single JVM), you can pre-create an HConnection