**HBase: Word Count**

HBase is database over HDFS. HBase allows user to perform CRUD (create, read, update and delete) operations on big data, data as large as petabytes. It is column oriented database.

**Advantage of HBase:**

1: Stores key/value pairs in columnar fashion(columns are clubbed together as column families).

2: Provides low latency access to small amounts of data from within a large data set.

3: Provides flexible data model

**Table Structure in HBase:**

Tables are made of rows and columns. A cell is an array of bytes. Table is sorted based on row key. Data is accessed based on the primary key. The columns in the HBase table are grouped in column families. HBase is similar to RDBMS, rows are sorted based on key, and columns of the row belong to column family. Tables are partitioned horizontally into regions. Each region consists of subsets of rows. As the size of the table increases, the number of regions also increases. Regions are distributed over the HBase cluster wherein each node host subsets of the total regions.

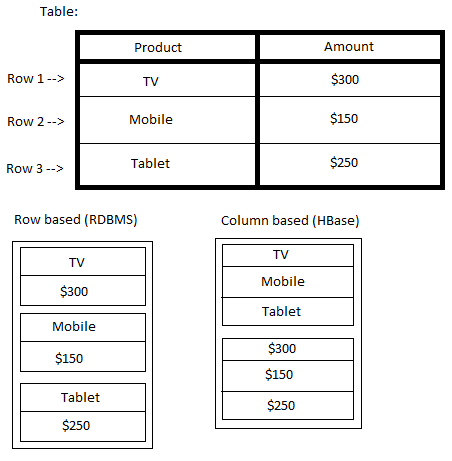


Fig: Storage on Disc for HBase

**Program Flow:**

WordCount is a program that counts the number of occurrences of a word in a given input dataset. In the previous project, the data is loaded from the HDFS. In this project, the data is loaded form the HBase. The dataset was ClueWeb09, which was created to support research on information retrieval and related human language technologies. It consists of about 1 billion webpages in ten languages that were collected from January 2009 to February 2009. The table schema is as follows:-

  
Fig: Data table schema for storing the ClueWeb09 dataset

**Mapper Function:**

A mapper function overrides the map function of the Mapper class. The input to the mapper is key and value wherein key is the rowkey of the HBase record related to the specified URL and value is the content that is stored in the text of the URL. The output of the map is key and value wherein key is the word and the value is frequency of the word in the context of the text.

**Reducer Function:**

Reducer combines the result of the mapper class. The reducer function sums up the occurrences of each word i.e. key and value wherein the key is the word and value is the frequency of that word. This output is then written back to the HBase table using put operation.