

HBase Introduction

Gwan-Hwan Hwang

2011/12/6

HBasics

- HBase is a distributed column-oriented database built on top of HDFS.
- HBase is the Hadoop application to use when you require real-time read/write random-access to very large datasets.

HBasics (Cont'd)

- The canonical HBase use case is the webtable, a table of crawled web pages and their attributes.
- Concurrently, the table is randomly accessed by crawlers running at various rates updating random rows while random web pages are served in read time as users click on website's cached-page feature.

Backdrop

- The HBase project was started toward the end of 2006.
- It was modeled after Google's "*Bigtable: A Distributed Storage System for Structured Data*"

Concepts

- Data model:
 - Labeled tables
 - Tables are made of rows and columns
 - Table cells
 - The intersection of row and column coordinates (are versioned)
 - The version is a timestamp auto-assigned by Hbase at the time of cell insertion.
 - A cell's content is an uninterpreted array of bytes.

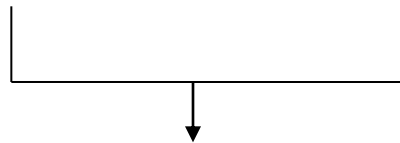
Concept (Cont'd)

- Table row keys are also byte arrays
- Table rows are sorted by row key, the table's primary key
- All table accesses are via the table primary key.
- Row columns are grouped into column families
 - All column family members have a common prefix
 - E.g. temperature:air, temperature:dew_point
 - The column family prefix must be composed of printable characters.
 - The qualifying tail, the column family qualifier, can be made of any arbitrary bytes.

Multidimensional Keys

- The map is indexed by a row key, column key, and timestamp; each value in the map is an uninterpreted array of bytes.

<row>, <family>:<qualifier>, <timestamp>



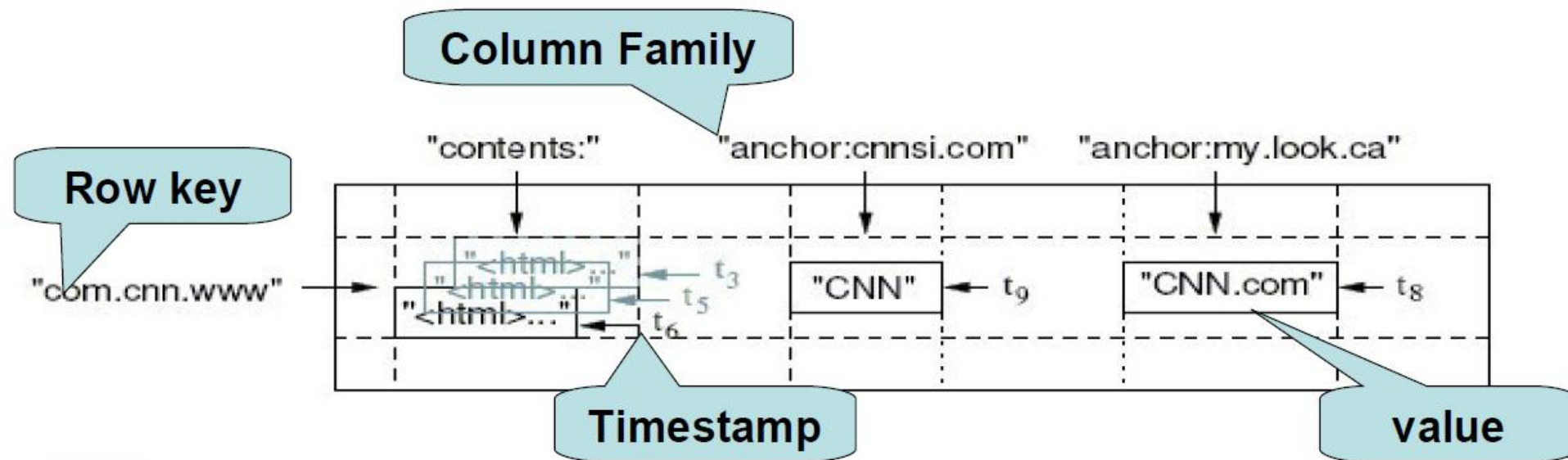
Column

- (row:string, column:string, time:int64) -> string
Qualifiers are additional level of indexing

Region	Row Keys	Column Family "Content"
Region 1	00000	...
	00001	...

	09999	...
Region 2	10000	...

	29999	...



Concept (Cont'd)

- Column families must be specified up front as part of the table schema definition.
 - New column family members can be added on demand.
- Physically, all column family members are stored together on the file system.

Concept (Cont'd)

- Generally speaking, HBase tables are like those in an RDBMS, only clls are versioned, rows are sorted, and columns can be added on the fly by the client as long as the column family they belong to preexists.

Regions

- Tables are automatically partitioned horizontally by HBase into regions.
- Each region comprises a subset of a table's rows.
- Initially, a table comprises a single region, but as the size of the region grows, after it crosses a configurable size threshold, it splits at a row boundary into two new regions of approximately equal size.

HBase shell

- Start Hbase shell
 - %hbase shell

- Related command

- create 'ghh','f1','f2'
- create 'ghh2',{NAME=>'f1'}, {NAME=>'f2'}
- scan 'ghh'
- put 'ghh','r1','f1:1','value1'
- put 'ghh','r1','f2:1','value2'
- put 'ghh','r1','f2:1','value3'
- put 'ghh','r1','f1:q2','value4'
- get 'ghh','r1'
- get 'ghh','r1',{COLUMN=>'f1:1'}
- put 'ghh','r1','f1','value1'
- delete 'ghh','r1','f1:'
- disable 'ghh'
- drop 'ghh'