## HBase shell script for table creation with proper settings

```
create 'webpages',
   {NAME => 'content', VERSIONS => 3, TTL => 7776000},
   {NAME => 'metadata', VERSIONS => 1},
   {NAME => 'outlinks', VERSIONS => 2, TTL => 15552000},
   {NAME => 'inlinks', VERSIONS => 2, TTL => 15552000}
```

## Retrieve the latest version of any page by URL

```
get 'webpages', ‘0000000edu.0000sample:main’
```

```
 outlinks:link_0                timestamp=2025-05-22T15:20:54.657, value=https://sample.edu/blog/tag/main
 outlinks:link_1                timestamp=2025-05-22T15:20:54.657, value=https://example.com/wp-content/posts/list
 outlinks:link_2                timestamp=2025-05-22T15:20:54.657, value=https://website.biz/search/posts
 outlinks:link_3                timestamp=2025-05-22T15:20:54.657, value=https://sample.edu/categories/blog
 outlinks:link_4                timestamp=2025-05-22T15:20:54.657, value=https://website.biz/posts/app/wp-content
 outlinks:link_5                timestamp=2025-05-22T15:20:54.657, value=https://test.org/category
 outlinks:link_6                timestamp=2025-05-22T15:20:54.657, value=https://sample.edu/tags/categories
 outlinks:link_7                timestamp=2025-05-22T15:20:54.657, value=https://website.biz/main
 outlinks:link_8                timestamp=2025-05-22T15:20:54.657, value=https://test.org/category
 outlinks:link_9                timestamp=2025-05-22T15:20:54.657, value=https://demo.net/tags
1 row(s)
Took 0.1041 seconds
hbase:012:0>
```

## View historical versions of a page to track changes

```
 get 'webpages', '0000000edu.0000sample:main', {COLUMN => 'content:html', VERSIONS => 5}
```

```
hbase:019:0> get 'webtable', '0000000edu.0000sample:main', {COLUMN => 'content:html', VERSIONS => 5}
COLUMN                         CELL
 content:html                  timestamp=2025-05-22T15:20:54.657, value=Relate truth more parent second strong wear feel
                               . School turn those player either.\x0AGovernment think federal area. Movement style envir
                               onmental reach hundred young say.\x0AIndustry design impact. Current smile energy share p
                               lace.\x0AI author job serious I. Hotel development popular evening help buy.\x0AHold eith
                               er rate morning phone inside vote message. Talk peace again blue hard much. Film debate h
                               otel she goal model appear.\x0AHuge so develop. Reveal pattern second parent young style
                               research. Value ground third girl.\x0AFast artist enter school. Evidence happy practice b
                               all. None book ground last whatever court administration.\x0AAgainst significant million
                               paper drop door each. Through actually support among thought. Which military it administr
                               ation power.\x0ALot realize maintain issue last. Partner sound resource against strong. A
                               ttack special surface outside could.\x0ABehavior western attack everything itself court s
                               trong. Family agency white from something.\x0AOr establish manager staff matter. Commerci
                               al then professor. Goal sound beautiful speak consumer two. Social green interest artist.
                               \x0AChoice door truth treat officer such admit play. Matter myself field help power case
                               feel. Night space hospital network leader.\x0ARemain radio office compare drug court resu
                               lt. Check hundred condition move. Big despite same easy.\x0ACharacter glass listen Democr
                               at while society. Car lay need nice audience.\x0AClearly himself sea. Attention six old t
                               hought speak. Hair always ground.\x0AQuestion wide stuff figure. Worry able organization
                               past plan writer mention. Suffer live the evening chair news up. Focus view kid whole pas
```

## List all pages from a specific domain for content audits

```
scan 'webpages', {ROWPREFIXFILTER => '0000000edu.0000sample'}
```

## Find all pages modified within a specific time range

```
scan 'webtable', {TIMERANGE => [1600000000000, 1900000000000]}
```

## Find all pages linking to a specific URL (inbound links)

## Identify pages with no outbound links (dead ends)

```
scan 'webtable', {
  COLUMNS => ['outlinks'],
```

```
hbase:011:0> scan 'webtable', {
hbase:012:1*   COLUMNS => ['inlinks:source_0'],
hbase:013:1*   FILTER => "ValueFilter(=, 'binary:https://example.com/tag/posts')"
}
ROW                              COLUMN+CELL
 0000000edu.0000sample:tag       column=inlinks:source_0, timestamp=2025-05-22T15:20:54.714, value=https://example.com/tag
                                 /posts
 0000000org.000000test:tag       column=metadata:language, timestamp=2025-05-22T15:20:54.676, value=en
 0000000org.000000test:tag       column=metadata:last_modified, timestamp=2025-05-22T15:20:54.676, value=2025-05-22T18:20:
                                 54.676660
 0000000org.000000test:tag       column=metadata:title, timestamp=2025-05-22T15:20:54.676, value=Republican ask girl popul
                                 ation next film.
 0000000org.000000test:tag       column=outlinks:link_0, timestamp=2025-05-22T15:20:54.676, value=https://demo.net/main/ca
                                 tegory
 0000000org.000000test:tag       column=outlinks:link_1, timestamp=2025-05-22T15:20:54.676, value=https://test.org/blog/ca
                                 tegories/posts
 0000000org.000000test:tag       column=outlinks:link_2, timestamp=2025-05-22T15:20:54.676, value=https://website.biz/cate
                                 gories
 0000000org.000000test:tag       column=outlinks:link_3, timestamp=2025-05-22T15:20:54.676, value=https://test.org/explore
                                 /categories/explore
                                  /tag
 0000000edu.0000sample:blog      column=outlinks:link_1, timestamp=2025-05-22T15:20:54.565, value=https://website.biz/app
 0000000edu.0000sample:blog      column=outlinks:link_2, timestamp=2025-05-22T15:20:54.565, value=https://demo.net/app/cat
                                 egories
 0000000edu.0000sample:blog      column=outlinks:link_3, timestamp=2025-05-22T15:20:54.565, value=https://website.biz/blog
 0000000edu.0000sample:categori column=content:html, timestamp=2025-05-22T15:20:54.751, value=Home per relationship inter
 es                              national market change. Try produce voice maintain drop same.\x0ANearly thus scene theory
```

```
  FILTER => "ColumnCountGetFilter(1) AND ValueFilter(=, 'binary:')"
}
```

## Find pages with HTTP error status codes

```
scan 'webtable', {
  COLUMNS => ['metadata:status_code'],
  FILTER => "SingleColumnValueFilter('metadata', 'status_code', >=, 'binary:400')"
}
```

## List pages with outdated content (not modified in last 30 days)

```
current_time=$(date +%s%3N)
thirty_days_ago=$((current_time - 2592000000))  # 30 days in ms

# Scan with time filter
scan 'webtable', {
  COLUMNS => ['metadata:last_modified'],
  FILTER => "SingleColumnValueFilter('metadata', 'last_modified', <, 'binary:$thirty_days_ago')"
}
```

```
hbase:001:0> put 'webtable', '0000000edu.0000sample:newpage', 'content:html', '<html>...</html>'
le', '0000000edu.0000sample:newpage', 'outlinks:link_1', 'https://test.org'Took 3.6055 seconds

hbase:002:0> put 'webtable', '0000000edu.0000sample:newpage', 'metadata:charset', 'UTF-8'
Took 0.0101 seconds
hbase:003:0> put 'webtable', '0000000edu.0000sample:newpage', 'metadata:content_type', 'text/html'
Took 0.0063 seconds
hbase:004:0> put 'webtable', '0000000edu.0000sample:newpage', 'metadata:last_modified', '2025-05-25T12:00:00.000Z'
Took 0.0076 seconds
hbase:005:0> put 'webtable', '0000000edu.0000sample:newpage', 'outlinks:link_0', 'https://example.com'
Took 0.0061 seconds
hbase:006:0> put 'webtable', '0000000edu.0000sample:newpage', 'outlinks:link_1', 'https://test.org'
Took 0.0330 seconds
hbase:007:0>
```

## Insert Complete Web Page Data

## Retrieve a Page by Exact URL

get 'webtable', '0000000edu.0000sample:main', {COLUMN => ['content:html', 'metadata:title']}

```
hbase:007:0> get 'webtable', '0000000edu.0000sample:main', {COLUMN => ['content:html', 'metadata:title']}
COLUMN                          CELL
 content:html                   timestamp=2025-05-22T15:20:54.657, value=Relate truth more parent second strong wear feel
                                . School turn those player either.\x0AGovernment think federal area. Movement style envir
                                onmental reach hundred young say.\x0AIndustry design impact. Current smile energy share p
                                lace.\x0AI author job serious I. Hotel development popular evening help buy.\x0AHold eith
                                er rate morning phone inside vote message. Talk peace again blue hard much. Film debate h
                                otel she goal model appear.\x0AHuge so develop. Reveal pattern second parent young style
                                research. Value ground third girl.\x0AFast artist enter school. Evidence happy practice b
                                all. None book ground last whatever court administration.\x0AAgainst significant million
                                paper drop door each. Through actually support among thought. Which military it administr
                                ation power.\x0ALot realize maintain issue last. Partner sound resource against strong. A
                                ttack special surface outside could.\x0ABehavior western attack everything itself court s
                                trong. Family agency white from something.\x0AOr establish manager staff matter. Commerci
```

## Update a Page's Content and Metadata

```
hbase:008:0> put 'webtable', '0000000edu.0000sample:main', 'content:html', '<html>updated content</html>'
Took 0.0484 seconds
hbase:009:0> put 'webtable', '0000000edu.0000sample:main', 'metadata:last_modified', '2025-05-25T14:30:00.000Z'
Took 0.0984 seconds
hbase:010:0> put 'webtable', '0000000edu.0000sample:main', 'outlinks:link_10', 'https://new-link.com'
Took 0.0314 seconds
hbase:011:0>
```

## Delete a Page and All Its Information

```
hbase:011:0> delete 'webtable', '0000000edu.0000sample:main', 'content:html'
Took 0.1092 seconds
hbase:012:0> delete 'webtable', '0000000edu.0000sample:main', 'metadata:title'
Took 0.0266 seconds
hbase:013:0> deleteall 'webtable', '0000000edu.0000sample:main'
Took 0.0974 seconds
hbase:014:0>
```