

HBase shell commands for table creation

create 'web_pages',

{NAME => 'content', VERSIONS => 3, TTL => 7776000}, # 90 days

{NAME => 'metadata', VERSIONS => 1},

{NAME => 'outlinks', VERSIONS => 2, TTL => 15552000}, # 180 days

{NAME => 'inlinks', VERSIONS => 2, TTL => 15552000}

#####Content Management Queries #####

Retrieve the latest version of any page by URL

get 'web_pages', 'org.test.www#articles/048', {COLUMN => 'content:html'}

```
hbase:025:0> get 'web_pages', 'org.test.www#articles/048', {COLUMN => 'content:html'}
COLUMN
content:html
timestamp=2025-05-22T15:18:00.466, value=<!DOCTYPE html>\x0A<html>\x0A    <head><title>Any difference them.</title><meta charset="utf-8"></head>\x0A    <body>\x0A        <h1>Any difference them.</h1>\x0A        <p>Start report major value. Everyone chance race there. Put necessary because develop soldier opportunity.</p><p>Want pattern peace there. Enjoy relationship authority owner. Stay contain represent our term edge peace indicate.</p><p>Ability relationship indicate thus. Door message next give your.</p>\x0A        <ul><li>current</li><li>travel</li><li>character</li><li>budget</li><li>yeah</li></ul>\x0A        <h2>Remain charge accept mission.</h2><p>Far professional degree hope arrive provide year. Leave change resource pull would mean ever offer. Receive prevent doctor receive town day.</p>\x0A    </body>\x0A</html>
1 row(s)
Took 0.0543 seconds
```

View historical versions of a page to track changes

get 'web_pages', 'org.test.www#articles/048', {COLUMN => 'content:html', VERSIONS => 3}

```
hbase:003:0> get 'web_pages', 'org.test.www#articles/048', {COLUMN => 'content:html', VERSIONS => 3}
COLUMN
content:html
timestamp=2025-05-22T16:31:59.364, value=new version
content:html
timestamp=2025-05-22T15:18:00.466, value=<!DOCTYPE html>\x0A<html>\x0A    <head><title>Any difference them.</title><meta charset="utf-8"></head>\x0A    <body>\x0A        <h1>Any difference them.</h1>\x0A        <p>Start report major value. Everyone chance race there. Put necessary because develop soldier opportunity.</p><p>Want pattern peace there. Enjoy relationship authority owner. Stay contain represent our term edge peace indicate.</p><p>Ability relationship indicate thus. Door message next give your.</p>\x0A        <ul><li>current</li><li>travel</li><li>character</li><li>budget</li><li>yeah</li></ul>\x0A        <h2>Remain charge accept mission.</h2><p>Far professional degree hope arrive provide year. Leave change resource pull would mean ever offer. Receive prevent doctor receive town day.</p>\x0A    </body>\x0A</html>
1 row(s)
Took 0.0308 seconds
```

List all pages from a specific domain for content audits

scan 'web_pages', {STARTROW => 'net.demo', ENDROW => 'net.demo~'}

```
hbase:004:0> scan 'web_pages', {STARTROW => 'org.test', ENDROW => 'org.test~'}
ROW
org.test.api#posts/006
COLUMN+CELL
column=content:html, timestamp=2025-05-22T15:18:00.520, value=<!DOCTYPE html>\x0A<html>\x0A    <head><title>Manager out real computer.</title><meta charset="utf-8"></head>\x0A    <body>\x0A        <h1>Manager out real computer.</h1>\x0A        <p>School onto near become these since system reason. Republican analysis back occur.</p><p>Thousand all fight drop down something save. Believe idea necessary answer. Idea her feel professional enter record name clearly.</p>\x0A        <ul><li>voice</li><li>law</li><li>another</li></ul>\x0A        <span>Strong maybe away professional prove magazine wide. Time cell street the body.</span>\x0A    </body>\x0A</html>
org.test.api#posts/006
column=inlinks:urls, timestamp=2025-05-22T15:18:00.520, value=
org.test.api#posts/006
column=metadata:created, timestamp=2025-05-22T15:18:00.520, value=2025-03-04T19:58:10.476502
org.test.api#posts/006
column=metadata:size, timestamp=2025-05-22T15:18:00.520, value=572
org.test.api#posts/006
column=metadata:status, timestamp=2025-05-22T15:18:00.520, value=200
org.test.api#posts/006
column=metadata:title, timestamp=2025-05-22T15:18:00.520, value=Manager out real computer.
org.test.api#posts/006
column=outlinks:urls, timestamp=2025-05-22T15:18:00.520, value=http://mysite.io/list/categories
org.test.blog#page/027
column=content:html, timestamp=2025-05-22T15:18:00.466, value=<!DOCTYPE html>\x0A<html>\x0A    <head><title>Street determine respond strategy.</title><meta charset="utf-8"></head>
```

Find all pages modified within a specific time range

scan 'web_pages', {

FILTER => "SingleColumnValueFilter('metadata', 'created', >=, 'binary:2024-05-22T00:00:00') AND SingleColumnValueFilter('metadata', 'created', <, 'binary:2025-05-23T00:00:00')"

}

```
hbase:030:0> scan 'web_pages', {
hbase:031:1* FILTER => "SingleColumnValueFilter('metadata', 'created', >=, 'binary:2024-05-22T00:00:00') AND SingleColumnValueFilter('metadata', 'created', <, 'binary:2025-05-23T00:00:00')"
```

ROW	COLUMN+CELL
co.sample.blog#search/explore	column=content:html, timestamp=2025-05-22T15:18:00.466, value=<!DOCTYPE html>\x0A<html>\x0A <head><title>Commercial free reflect yourself language guy.</title><meta charset="utf-8"></head>\x0A <body>\x0A <h1>Commercial free reflect yourself language guy.</h1>\x0A <p>Morning per wait find. Fine environment last within. It network prove sort produce small public.</p><p>To themselves but may direction most. Happen audience popular respond rise work. Both sea home.</p><p>What step decide wish. Card answer note argue couple as work street. Couple pretty thought relate institution attention not law.</p>\x0A factgoalhappenitssittrade\x0A <p>Position summer entire long phone. Score father push those return. Remain attention he nearly.</p><h2>Hit say spend garden make party.</h2>\x0A </body>\x0A</html>
co.sample.blog#search/explore	column=inlinks:urls, timestamp=2025-05-22T15:18:00.466, value=http://sample.co/explore,http://mysite.io/search/list
co.sample.blog#search/explore	column=metadata:created, timestamp=2025-05-22T15:18:00.466, value=2025-03-18T17:47:29.009267
co.sample.blog#search/explore	column=metadata:size, timestamp=2025-05-22T15:18:00.466, value=815
co.sample.blog#search/explore	column=metadata:status, timestamp=2025-05-22T15:18:00.466, value=200
co.sample.blog#search/explore	column=metadata:title, timestamp=2025-05-22T15:18:00.466, value=Commercial free reflect yourself language guy.
co.sample.blog#search/explore	column=outlinks:urls, timestamp=2025-05-22T15:18:00.466, value=http://demo.net/categories,http://demo.net/blog,http://mysite.io/categories/app/wp-content

#####SEO Analysis#####

Find inbound links to a page

get 'web_pages', 'org.test.www#articles/048', { COLUMN => 'inlinks:urls' }

```
=> Java:10org.test.www
hbase:043:0> get 'web_pages', 'org.test.www#articles/048', { COLUMN => 'inlinks:urls' }
```

COLUMN	CELL
inlinks:urls	timestamp=2025-05-22T15:18:00.466, value=http://test.org/categories/posts/category

1 row(s)
Took 0.0174 seconds
hbase:044:0>

Identify pages with no outbound links (dead ends)

scan 'web_pages', {

FILTER => "SingleColumnValueFilter('outlinks', 'urls', =, 'binary:~')"

}

```
hbase:045:1* FILTER => "SingleColumnValueFilter('outlinks', 'urls', =, 'binary:~')"
```

ROW	COLUMN+CELL
org.test.www#articles/048	column=content:html, timestamp=2025-05-22T16:31:59.364, value=new version
org.test.www#articles/048	column=inlinks:urls, timestamp=2025-05-22T15:18:00.466, value=http://test.org/categories/posts/category
org.test.www#articles/048	column=metadata:created, timestamp=2025-05-22T15:18:00.466, value=2025-04-06T11:20:28.207648
org.test.www#articles/048	column=metadata:size, timestamp=2025-05-22T15:18:00.466, value=767
org.test.www#articles/048	column=metadata:status, timestamp=2025-05-22T15:18:00.466, value=500
org.test.www#articles/048	column=metadata:title, timestamp=2025-05-22T15:18:00.466, value=Any difference them.
org.test.www#articles/048	column=outlinks:urls, timestamp=2025-05-22T15:18:00.466, value=

1 row(s)
Took 0.0182 seconds

List pages with the most inbound links (popular pages)

Retrieve pages with specific content in the title or body

```
scan 'web_pages', {  
  FILTER => "ValueFilter(=, 'substring:Any')", COLUMNS => 'metadata:title'  
}
```

```
Took 0.0039 seconds  
hbase:056:0> scan 'web_pages', {  
hbase:057:1*   FILTER => "ValueFilter(=, 'substring:Any')", COLUMNS => 'metadata:title'  
}  
ROW  
org.test.www#articles/048      column=metadata:title, timestamp=2025-05-22T15:18:00.466, value=Any difference them.  
1 row(s)  
Took 0.0135 seconds  
hbase:059:0>
```

Performance Optimization

Identify the largest pages by content size

Find pages with HTTP error status codes

```
scan 'web_pages', {  
  FILTER => "SingleColumnValueFilter('metadata', 'status', =, 'binary:404')"  
}
```

```
hbase:068:0> scan 'web_pages', {  
hbase:069:1*   FILTER => "SingleColumnValueFilter('metadata', 'status', =, 'binary:404')"  
}  
ROW  
com.example.shop#articles/048      COLUMN+CELL  
column=content:html, timestamp=2025-05-22T15:18:00.466, value=<!DOCTYPE html>\x0A<html>\x0A    <h  
ead><title>Fine hard Mr describe.</title><meta charset="utf-8"></head>\x0A    <body>\x0A    <  
h1>Fine hard Mr describe.</h1>\x0A    <p>Dinner role age single each order. Five item shake p  
rovide add. Reach take law produce can poor.</p><p>Three close tax amount himself democratic. Cou  
rse skill employee as time not pass wind.</p>\x0A    <ul><li>return</li><li>government</li><li>  
i>explain</li><li>somebody</li><li>personal</li><li>down</li></ul>\x0A    <div>Phone study we  
ight. Pick investment sort light.</div>\x0A    </body>\x0A</html>  
column=inlinks:urls, timestamp=2025-05-22T15:18:00.466, value=http://demo.net/list/app,http://mys  
ite.io/blog/search,http://mysite.io/categories  
com.example.shop#articles/048      column=metadata:created, timestamp=2025-05-22T15:18:00.466, value=2025-04-18T02:24:05.800160  
com.example.shop#articles/048      column=metadata:size, timestamp=2025-05-22T15:18:00.466, value=558  
com.example.shop#articles/048      column=metadata:status, timestamp=2025-05-22T15:18:00.466, value=404  
com.example.shop#articles/048      column=metadata:title, timestamp=2025-05-22T15:18:00.466, value=Fine hard Mr describe.  
com.example.shop#articles/048      column=outlinks:urls, timestamp=2025-05-22T15:18:00.466, value=http://example.com/tags,http://mys  
ite.io/list  
com.example.www#page/018           column=content:html, timestamp=2025-05-22T15:18:00.466, value=<!DOCTYPE html>\x0A<html>\x0A    <h
```

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

Basic Operations

Insert complete web page data (content, metadata, links)

```
put 'web_pages', 'com.example#posts/1', 'outlinks:urls',  
'http://test.org/page1,http://demo.net/page2'  
put 'web_pages', 'com.example#posts/1', 'inlinks:urls',  
'http://mysite.io/page3,http://sample.co/page4'  
put 'web_pages', 'com.example#posts/1', 'metadata:title', 'Example Title'  
put 'web_pages', 'com.example#posts/1', 'content:html', '<html>...</html>'
```

```
Took 0.0151 seconds  
hbase:118:0> put 'web_pages', 'com.example#posts/1', 'outlinks:urls', 'http://test.org/page1,http://demo.net/page2'  
Took 0.0138 seconds  
hbase:119:0> put 'web_pages', 'com.example#posts/1', 'inlinks:urls', 'http://mysite.io/page3,http://sample.co/page4'  
Took 0.0095 seconds  
hbase:120:0> put 'web_pages', 'com.example#posts/1', 'metadata:title', 'Example Title'  
Took 0.0164 seconds  
hbase:121:0> put 'web_pages', 'com.example#posts/1', 'content:html', '<html>...</html>'  
Took 0.0097 seconds  
hbase:122:0>
```

Retrieve a page by exact URL

```
get 'web_pages', 'org.test.www#articles/048'
```

```
hbase:127:0> get 'web_pages', 'org.test.www#articles/048'  
COLUMN CELL  
content:html timestamp=2025-05-22T16:31:59.364, value=new version  
content:last_modified timestamp=2025-05-21T00:00, value=  
inlinks:urls timestamp=2025-05-22T15:18:00.466, value=http://test.org/categories/posts/category  
metadata:created timestamp=2025-05-22T15:18:00.466, value=2025-04-06T11:20:28.207648  
metadata:size timestamp=2025-05-22T15:18:00.466, value=767  
metadata:status timestamp=2025-05-22T15:18:00.466, value=500  
metadata:title timestamp=2025-05-22T15:18:00.466, value=Any difference them.  
outlinks:urls timestamp=2025-05-22T15:18:00.466, value=  
1 row(s)
```

Update a page's content and metadata

```
put 'web_pages', 'org.test.www#articles/048', 'content:html', 'new version'  
put 'web_pages', 'org.test.www#articles/048', 'metadata:title', 'title new version'
```

```
base:132:0> put 'web_pages', 'org.test.www#articles/048', 'content:html', 'new version'  
Took 0.0133 seconds  
base:133:0> put 'web_pages', 'org.test.www#articles/048', 'metadata:title', 'title new version'  
Took 0.0112 seconds  
base:134:0>
```

Delete a page and all its information

```
deleteall 'web_pages', 'org.test.www#articles/048'
```

```
Took 0.0112 seconds  
hbase:134:0> deleteall 'web_pages', 'org.test.www#articles/048'  
Took 0.0190 seconds  
hbase:135:0> get 'web_pages', 'org.test.www#articles/048'  
COLUMN CELL  
0 row(s)  
Took 0.0154 seconds  
hbase:136:0>
```

Filtering Operations

Find pages with titles containing specific keywords

```
scan 'web_pages', {  
  FILTER => "ValueFilter(=, 'substring:Group')", COLUMNS => 'metadata:title'  
}
```

```
hbase:149:0> scan 'web_pages', {  
hbase:150:1* FILTER => "ValueFilter(=, 'substring:Group')", COLUMNS => 'metadata:title'  
}  
ROW COLUMN+CELL  
org.test.shop#articles/002 column=metadata:title, timestamp=2025-05-22T15:18:00.466, value=Group city perhaps mother table s  
ection.  
1 row(s)  
Took 0.0133 seconds  
hbase:152:0>
```

Retrieve pages with content size above a threshold

```
scan 'web_pages', {  
  FILTER => "SingleColumnValueFilter('metadata', 'size', >=, 'binary:999')"  
}
```

```
hbase:153:1* FILTER => "SingleColumnValueFilter('metadata', 'size', >=, 'binary:999')"  
}  
ROW COLUMN+CELL  
com.example#posts/1 column=content:html, timestamp=2025-05-22T16:52:35.921, value=<html>...</html>  
com.example#posts/1 column=inlinks:urls, timestamp=2025-05-22T16:52:35.790, value=http://mysite.io/page3,http://sampl  
e.co/page4  
com.example#posts/1 column=metadata:title, timestamp=2025-05-22T16:52:35.837, value=Example Title  
com.example#posts/1 column=outlinks:urls, timestamp=2025-05-22T16:52:35.748, value=http://test.org/page1,http://demo.  
net/page2  
net.demo.shop#posts/14 column=content:html, timestamp=2025-05-22T16:54:32.921, value=new version  
net.demo.shop#posts/14 column=metadata:title, timestamp=2025-05-22T16:54:32.953, value=title new version  
2 row(s)  
Took 0.0195 seconds  
hbase:155:0>
```

Find pages with HTTP error status codes

```
scan 'web_pages', {  
  FILTER => "SingleColumnValueFilter('metadata', 'status', =, 'binary:404')"  
}
```

```
hbase:155:0> scan 'web_pages', {  
hbase:156:1* FILTER => "SingleColumnValueFilter('metadata', 'status', =, 'binary:404')"  
}  
ROW COLUMN+CELL  
com.example#posts/1 column=content:html, timestamp=2025-05-22T16:52:35.921, value=<html>...</html>  
com.example#posts/1 column=inlinks:urls, timestamp=2025-05-22T16:52:35.790, value=http://mysite.io/page3,http://sampl  
e.co/page4  
com.example#posts/1 column=metadata:title, timestamp=2025-05-22T16:52:35.837, value=Example Title  
com.example#posts/1 column=outlinks:urls, timestamp=2025-05-22T16:52:35.748, value=http://test.org/page1,http://demo.  
net/page2  
com.example.shop#articles/048 column=content:html, timestamp=2025-05-22T15:18:00.466, value=<!DOCTYPE html>\x0A<html>\x0A <h  
ead><title>Fine hard Mr describe.</title><meta charset="utf-8"></head>\x0A <body>\x0A <  
h1>Fine hard Mr describe.</h1>\x0A <p>Dinner role age single each order. Five item shake p  
rovide add. Reach take law produce can poor.</p><p>Three close tax amount himself democratic. Cou  
rse skill employee as time not pass wind.</p>\x0A <ul><li>return</li><li>government</li><li>  
i>explain</li><li>somebody</li><li>personal</li><li>down</li></ul>\x0A <div>Phone study we  
ight. Pick investment sort light.</div>\x0A </body>\x0A</html>  
com.example.shop#articles/048 column=inlinks:urls, timestamp=2025-05-22T15:18:00.466, value=http://demo.net/list/app,http://mys  
ite.io/blog/search,http://mysite.io/categories  
com.example.shop#articles/048 column=metadata:created, timestamp=2025-05-22T15:18:00.466, value=2025-04-18T02:24:05.800160
```

Find pages modified after a specific date

```
scan 'web_pages', {  
  FILTER => "SingleColumnValueFilter('metadata', 'created', >=,  
'binary:2025-05-01T00:00:00')"  
}
```

```
hbase:182:0>  
hbase:183:0> scan 'web_pages', {  
hbase:184:1*   FILTER => "SingleColumnValueFilter('metadata', 'created', >=, 'binary:2025-05-01T00:00:00')"  
}  
ROW                                COLUMN+CELL  
com.example#posts/1                column=content:html, timestamp=2025-05-22T16:52:35.921, value=<html>...</html>  
com.example#posts/1                column=inlinks:urls, timestamp=2025-05-22T16:52:35.790, value=http://mysite.io/page3,http://samps  
e.co/page4  
com.example#posts/1                column=metadata:title, timestamp=2025-05-22T16:52:35.837, value=Example Title  
com.example#posts/1                column=outlinks:urls, timestamp=2025-05-22T16:52:35.748, value=http://test.org/page1,http://demo  
net/page2  
com.example.www#posts/031           column=content:html, timestamp=2025-05-22T15:18:00.520, value=<!DOCTYPE html>\x0A<html>\x0A  <  
ead><title>Mission realize hear fear pressure.</title><meta charset="utf-8"></head>\x0A  <body  
\x0A    <h1>Mission realize hear fear pressure.</h1>\x0A    <p>White cost mother. Reason  
choose their avoid travel air American.</p><p>Type career cold reach debate. Chair result electi  
n activity decide hard go.</p><p>Use anything other western agreement. Real success product offi  
er much. Nation group win history her woman.</p><p>Yard while ground get. Blue community subject
```

Scanning with Pagination

```
scan 'web_pages', {  
  STARTROW => 'net.demo.shop#posts',  
  COLUMNS => 'content',  
  FILTER => "PageFilter(5)"  
}
```

```
hbase:249:0> scan 'web_pages', {  
hbase:250:1*   STARTROW => 'net.demo.shop#posts',  
hbase:251:1*   COLUMNS => 'content',  
hbase:252:1*   FILTER => "PageFilter(5)"  
}  
ROW                                COLUMN+CELL  
net.demo.shop#posts/001             column=content:html, timestamp=2025-05-22T17:01:48.692, value=new content version 1  
net.demo.shop#posts/002             column=content:html, timestamp=2025-05-22T17:01:48.719, value=new content version 2  
net.demo.shop#posts/010             column=content:html, timestamp=2025-05-22T17:01:48.747, value=new content version 3  
net.demo.shop#posts/015             column=content:html, timestamp=2025-05-22T17:01:48.773, value=new content version 4  
net.demo.shop#posts/020             column=content:html, timestamp=2025-05-22T17:01:49.326, value=new content version 5  
5 row(s)  
Took 0.0169 seconds
```

Compare different versions of the same page

```
get 'web_pages', 'net.demo.shop#posts/014', {COLUMN => 'content:html', VERSIONS => 3}
```

```
hbase:270:0> get 'web_pages', 'net.demo.shop#posts/014', {COLUMN => 'content:html', VERSIONS => 3}  
COLUMN                                CELL  
content:html                          timestamp=2025-05-22T17:05:06.596, value=title new version2  
content:html                          timestamp=2025-05-22T17:04:10.684, value=new version  
1 row(s)  
Took 0.0078 seconds
```

Implement a manual purge for outdated content

Show how to retrieve the latest N versions of content

get 'web_pages', 'net.demo.shop#posts/014', {COLUMN => 'content:html', VERSIONS => 2}

```
Took 0.0123 seconds
hbase:272:0> get 'web_pages', 'net.demo.shop#posts/014', {COLUMN => 'content:html', VERSIONS => 2}
COLUMN                                CELL
content:html                          timestamp=2025-05-22T17:05:06.596, value=title new version2
content:html                          timestamp=2025-05-22T17:04:10.684, value=new version
1 row(s)
Took 0.0117 seconds
```