

Analytics with Hive

Regex SerDe and Views

`/^[Reg]ular[Ex]pression$/`

/^[Reg]ular[Ex]pression\$/

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200
2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php
HTTP/1.1" 200 2 "http://www.example.com/wordpress3/wp-admin/post-new.php"
"Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-US) AppleWebKit/534.3 (KHTML,
like Gecko) Chrome/6.0.472.25 Safari/534.3"



/^[Reg]ular[Ex]pression\$/

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200
      2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php
      HTTP/1.1" 200 2 "http://www.example.com/wordpress3/wp-admin/post-new.php"
      "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-US) AppleWebKit/534.3 (KHTML,
      like Gecko) Chrome/6.0.472.25 Safari/534.3"
```

123.65.150.10

-

-

[23/Aug/2010:03:50:59
+0000]

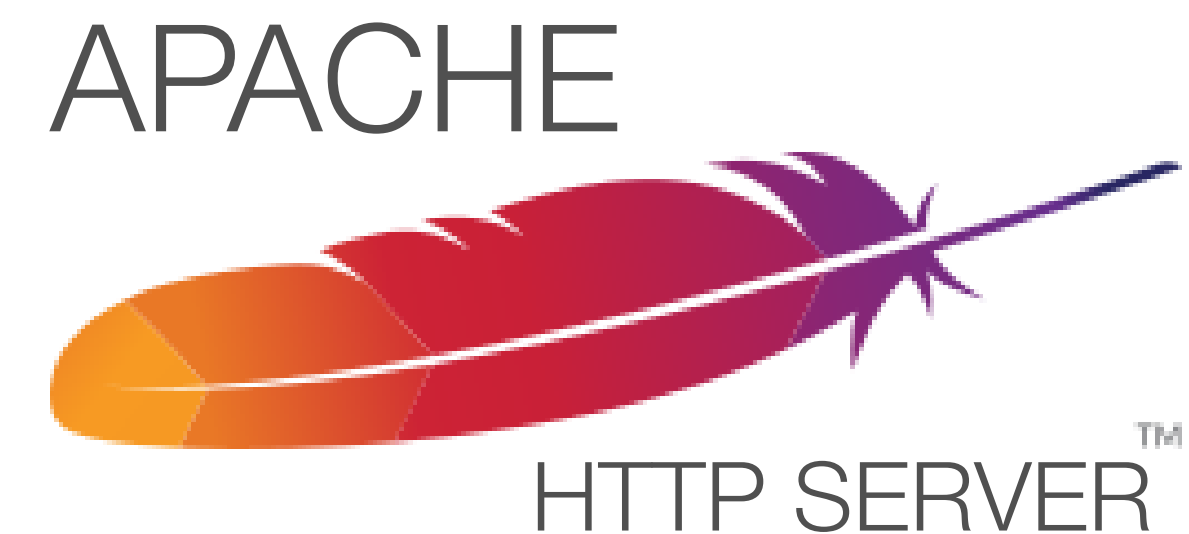
"POST

/wordpress3/wp-admin/admin-ajax.php

HTTP/1.1"


200

...



127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200
123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php
HTTP/1.1" 200

```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
LOCATION '/user/adral/local_log';
```




SERialisation + **DE**serialisation

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200
123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php
HTTP/1.1" 200

```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^(\\S*) (\\S*) (\\S*) \\[([^\]]*)\\] "([^"]*)" (\\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200
123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php
HTTP/1.1" 200

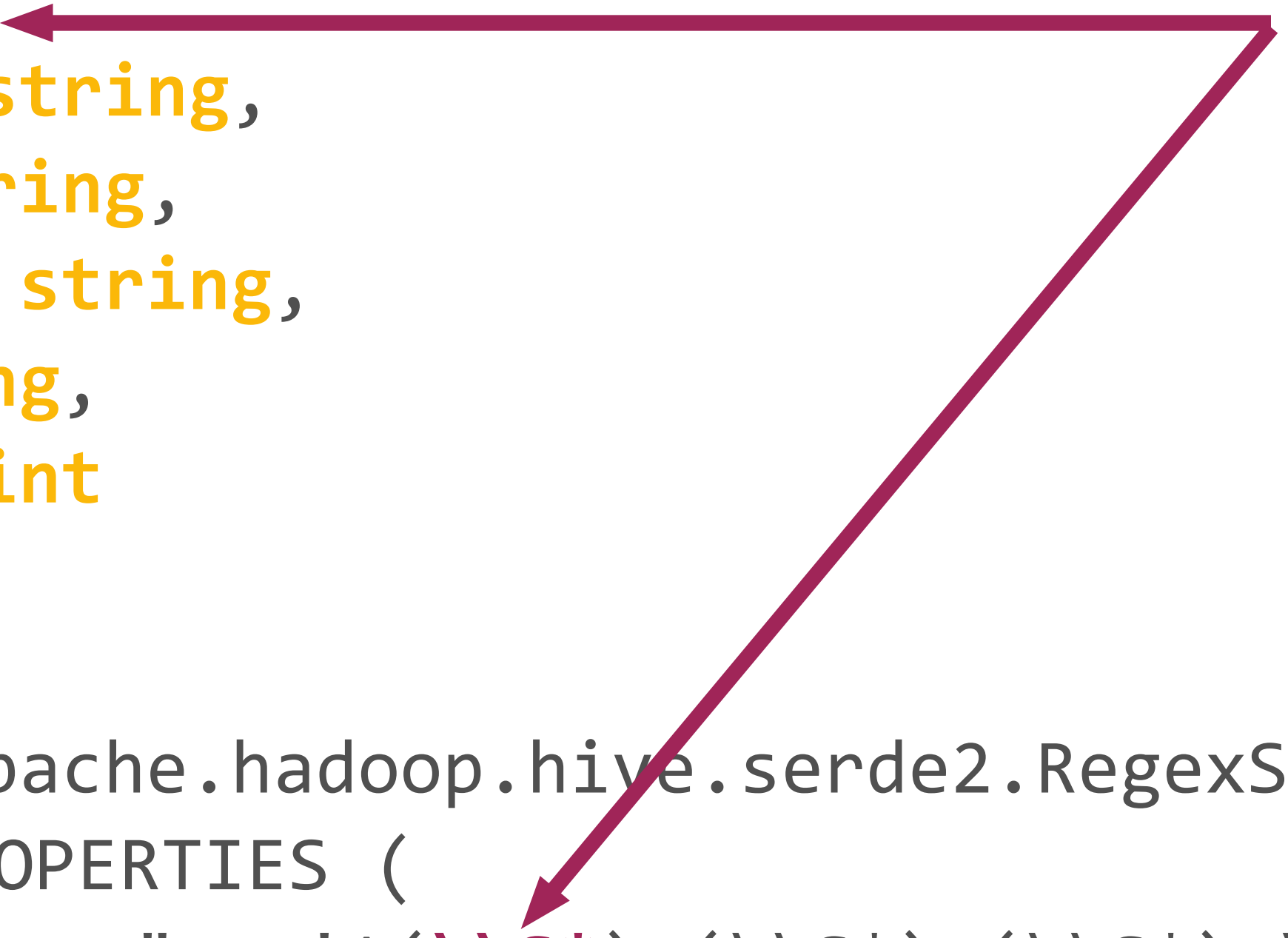
```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^((\\S*) (\\S*) (\\S*) \\[([^\\"\\]]*)\\] \"([^\"]*)\" (\\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```



127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200

123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php HTTP/1.1" 200

```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^(\S*) (\S*) (\S*) \[([^\]]*)\] "([^"]*)" (\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```



127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200

123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php
HTTP/1.1" 200

```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^(\S*) (\S*) (\S*) \\[([^\]]*)\] "([^"]*)" (\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```

The diagram illustrates the mapping between the column definitions in the `CREATE EXTERNAL TABLE` statement and the capture groups in the `input.regex` property. Three arrows originate from a common point above the `WITH SERDEPROPERTIES` block and point to the first three capture groups in the regex: `(\S*)`, `(\S*)`, and `(\S*)`. These capture groups correspond to the `ip`, `auth_unused`, and `auth_user` columns, respectively.

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200
123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php
HTTP/1.1" 200

```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^((\\S*) (\\S*) (\\S*) \\\\([\\^\\\\\\\\]*)\\\\\\\\' "([\\^"]*)" (\\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```



127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200

123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php HTTP/1.1" 200

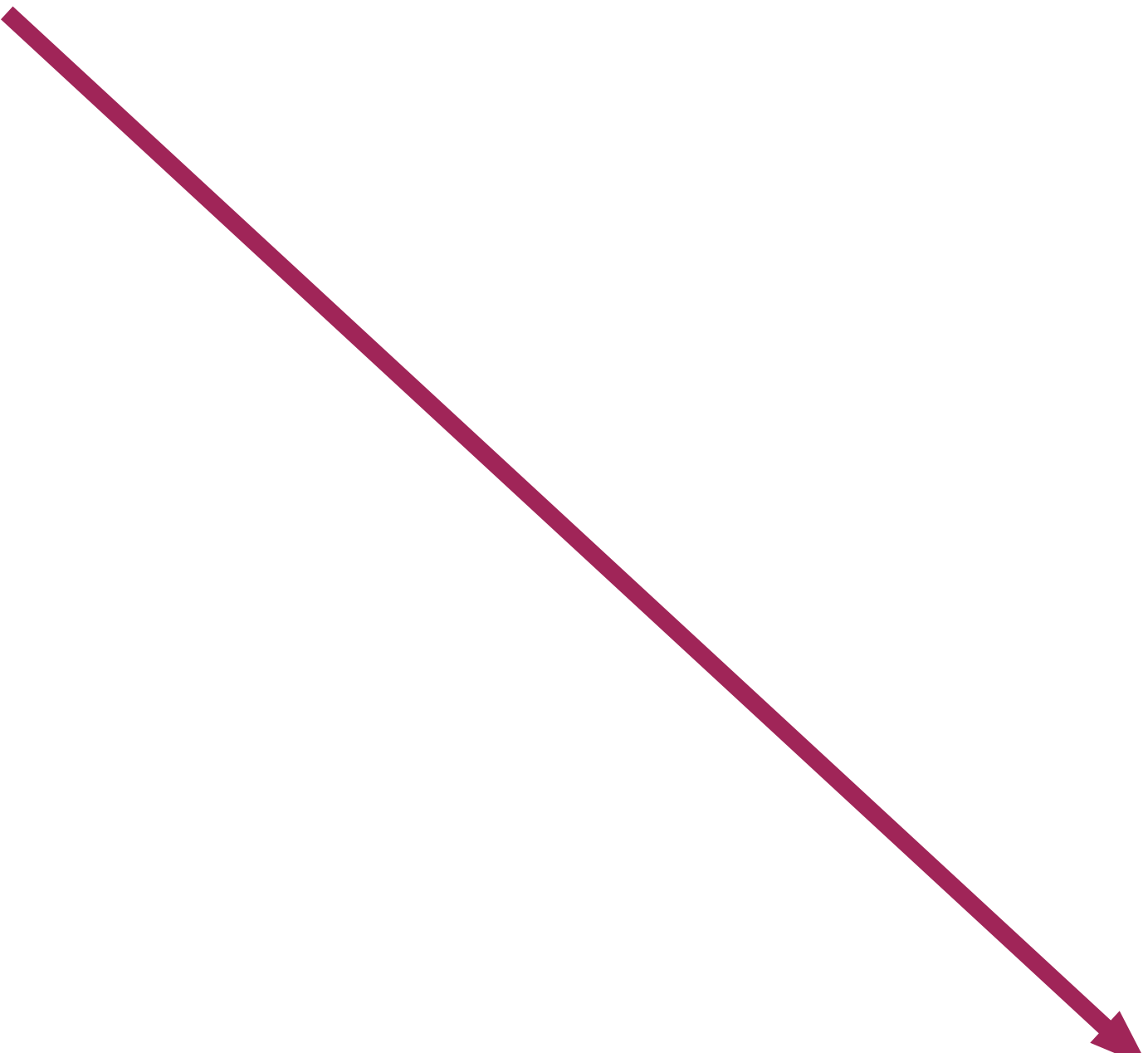
```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^(\\S*) (\\S*) (\\S*) \\.\\.([^[\\]]*)\\.\\. \"([^\"]*)\" (\\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```



127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200

123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php HTTP/1.1" 200

```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^(\\S*) (\\S*) (\\S*) \\[([^\]]*)\\] \"([^\"]*)\" (\\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```



127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200

123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php HTTP/1.1" 200


```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
    )                                org.apache.hadoop.hive.contrib.serde2.RegexSerDe  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^((\\S*) (\\S*) (\\S*) \\[([^\\"\\]]*)\\] \"([^\"]*)\" (\\S*) .*$'  
    )  
LOCATION '/user/adral/local_log';
```



127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" **200**

123.65.150.10 - - [23/Aug/2010:03:50:59 +0000] "POST /wordpress3/wp-admin/admin-ajax.php HTTP/1.1" **200**

```
CREATE EXTERNAL TABLE apache_log (  
    ip string,  
    auth_unused string,  
    auth_user string,  
    request_time string,  
    request string,  
    status_code int  
)  
ROW FORMAT  
SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = '^(\\S*) (\\S*) (\\S*) \\.\\.([^\"]*)\\.\\. \"([^\"]*)\" (\\S*) .*$'  
)  
LOCATION '/user/adral/local_log';
```



```
CREATE VIEW apache_log_view
```

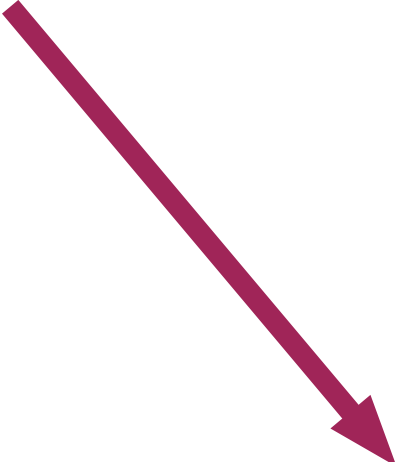
```
CREATE VIEW apache_log_view (  
    ip,  
    request_year,  
    request,  
    status_code  
)
```



```
CREATE VIEW apache_log_view (  
    ip,  
    request_year,  
    request,  
    status_code  
)  
AS SELECT  
    ip,  
    regexp_extract(request_time, "\\d+\\/\\w+\\/\\d+", 1),  
    request, status_code  
FROM apache_log;
```

```
CREATE VIEW apache_log_view (  
    ip,  
    request_year,  
    request,  
    status_code  
)  
AS SELECT  
    ip,  
    regexp_extract(request_time, "\\d+\\/\\w+\\/ (\\d+)", 1),  
    request, status_code  
FROM apache_log;
```

[10/Oct/2000:13:55:36-0700]



```
hive> SHOW TABLES;
```

```
OK
```

```
apache_log
```

```
apache_log_view
```

```
...
```

```
hive> SHOW TABLES;  
OK  
apache_log  
apache_log_view  
...
```

```
hive> SHOW CREATE TABLE apache_log_view;
```

```
OK
```

```
CREATE VIEW 'apache_log_view' AS SELECT 'ip' AS 'ip', '_c1' AS  
'request_year', 'request' AS 'request', 'status_code' AS 'status_code'  
FROM (SELECT  
    'apache_log'. 'ip',  
    regexp_extract('apache_log'. 'request_time', "\\d+\\V\\w+\\V(\\d+)", 1),  
    'apache_log'. 'request', 'apache_log'. 'status_code'  
FROM 'adral'. 'apache_log') 'adral.apache_log_view'
```

read: HDFS file split



parse into columns



casts + extracts



filter (WHERE)



...

Mapper

read: HDFS file split



parse into columns



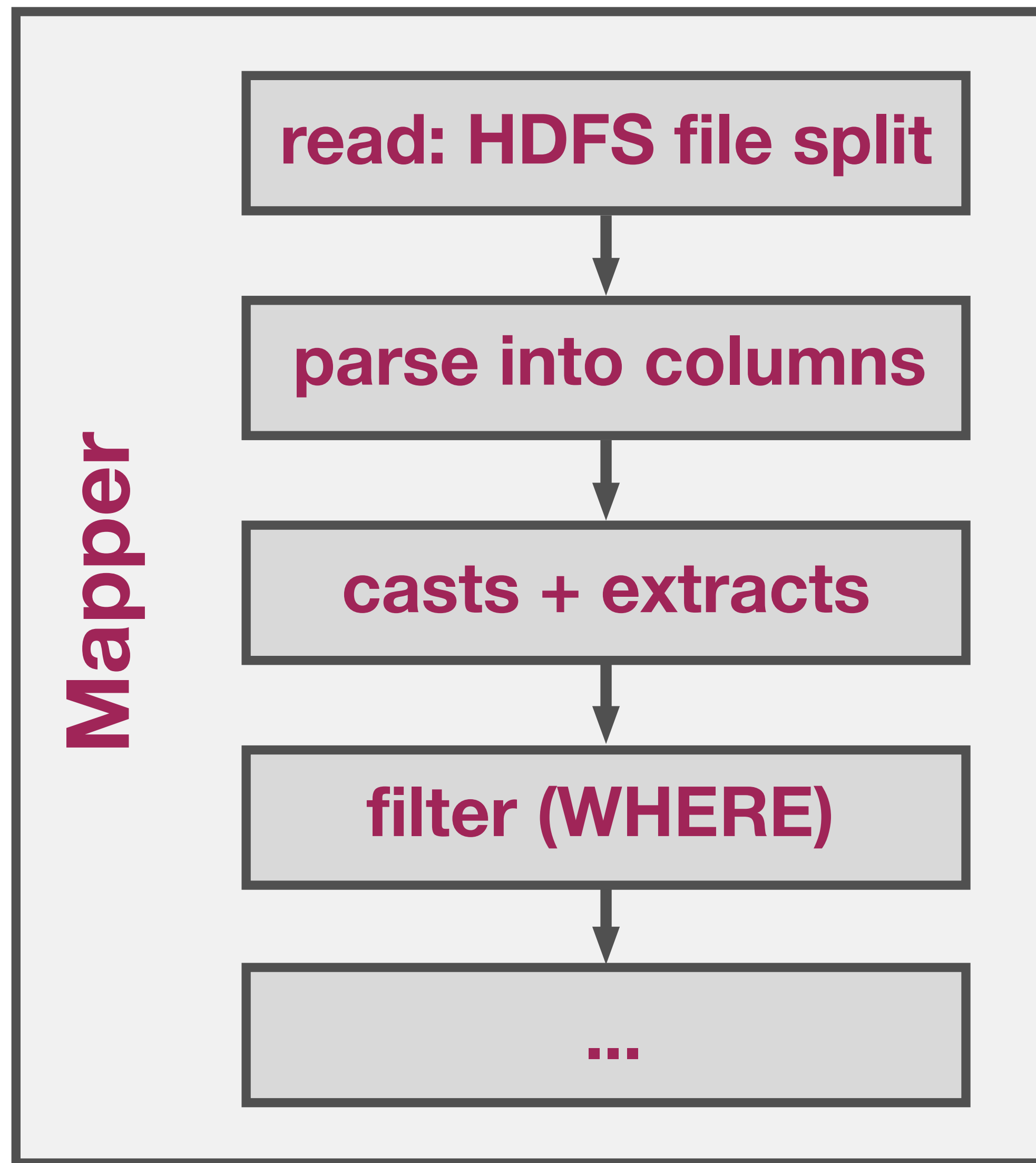
casts + extracts



filter (WHERE)

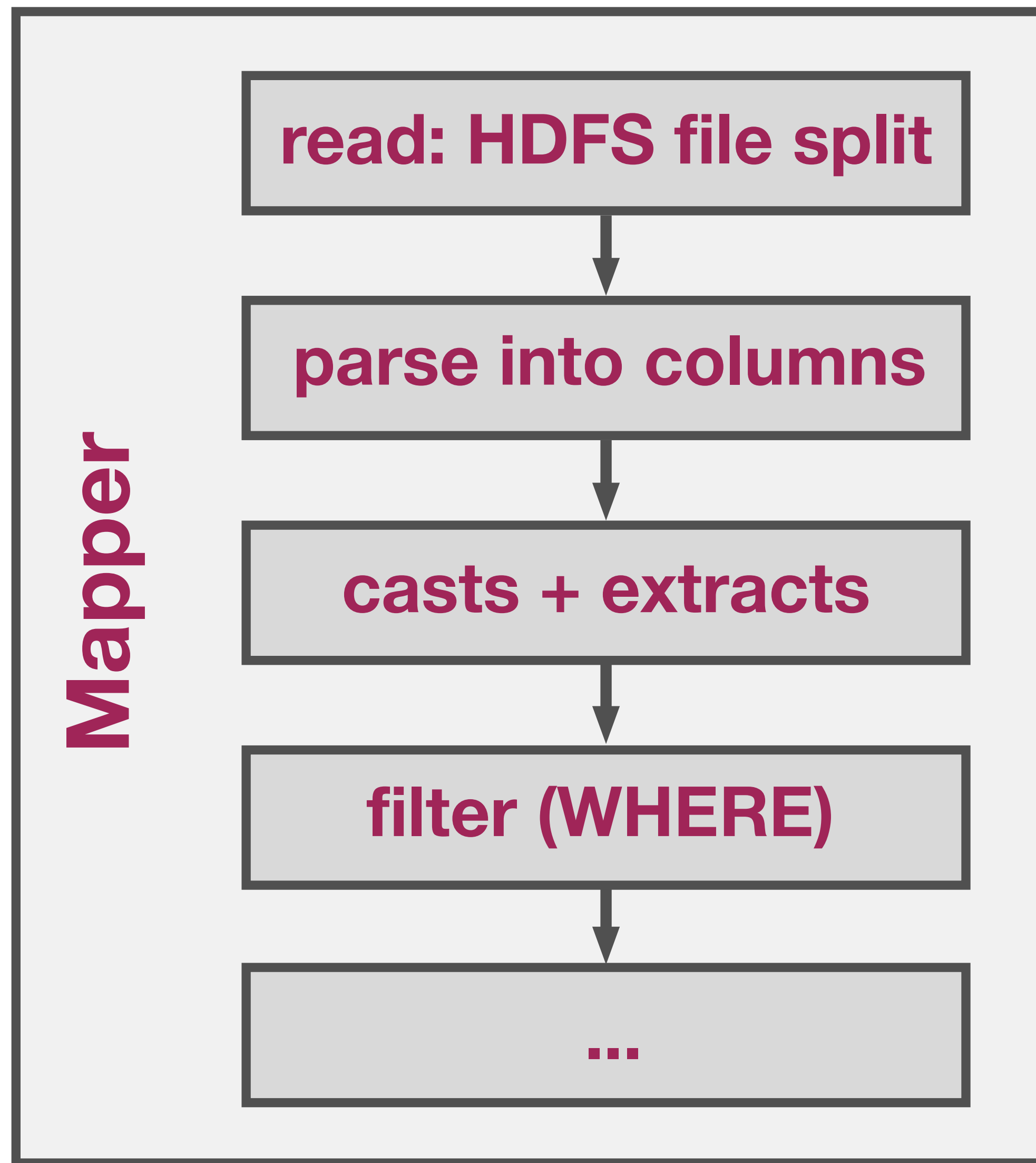


...



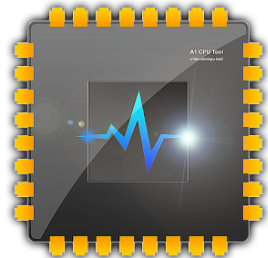
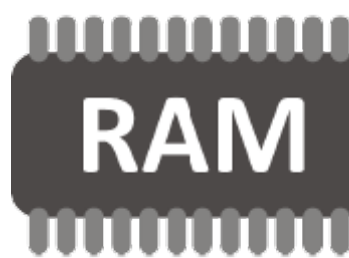
View Limitations

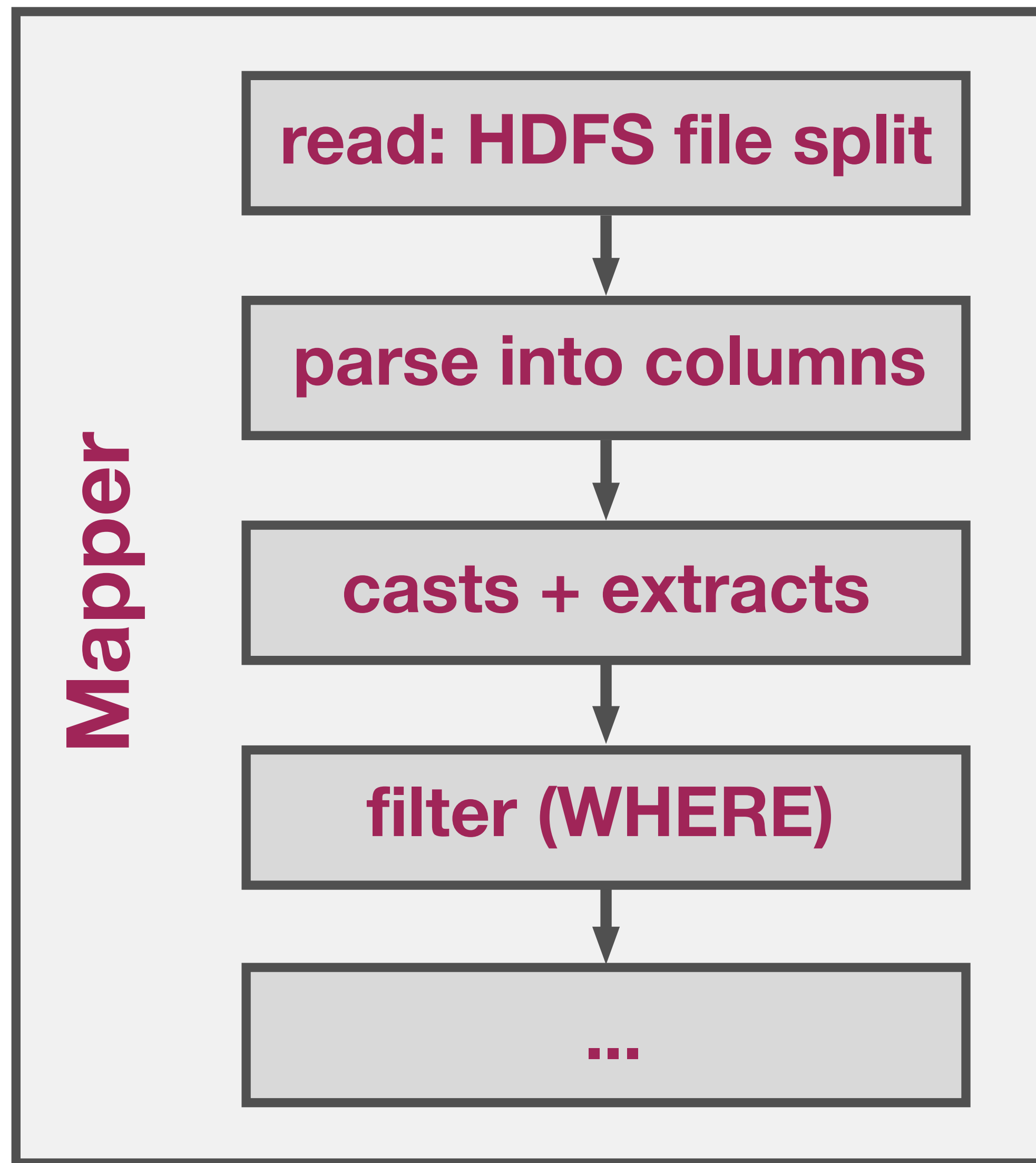
1. Read-only



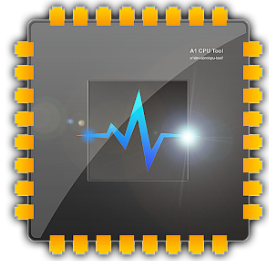
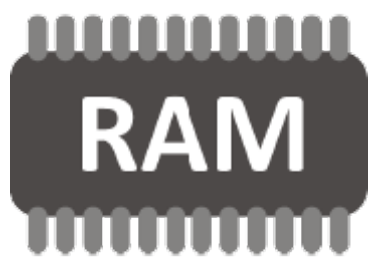
View Limitations

1. Read-only

2.  



View Limitations

1. Read-only
2.   RAM
3. Metainformation fluctuation

Summary

Summary

- You can **use** RegexSerde to create tables from complex data formats

see: <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF> (regex_extract)
see: <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#Language-ManualDDL-Create/Drop/AlterView> (create view)

Summary

- You can **use** RegexSerde to create tables from complex data formats
- You can **list** pros and cons of Hive views

see: <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF> (regex_extract)

see: <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#Language-ManualDDL-Create/Drop/AlterView> (create view)