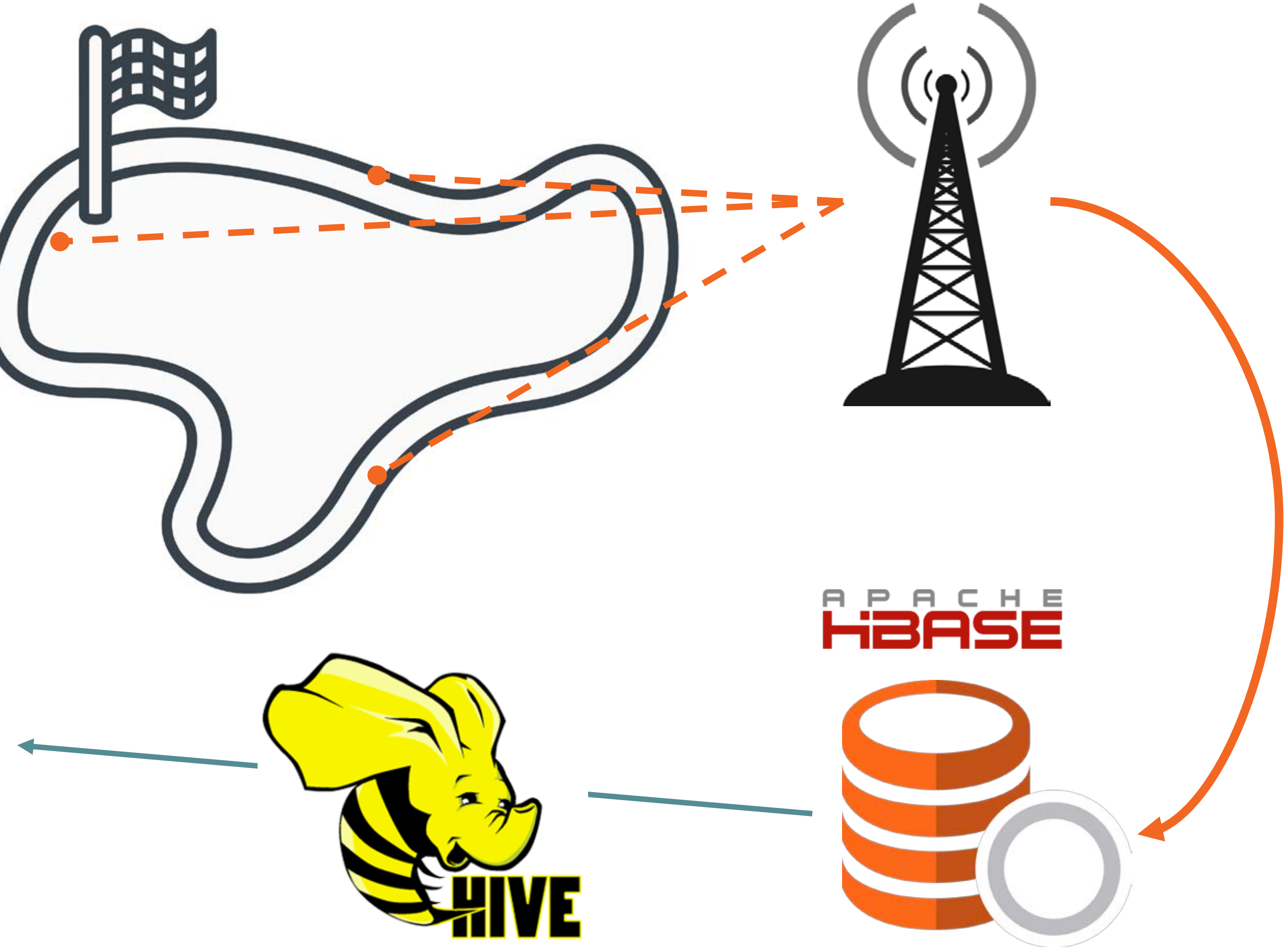
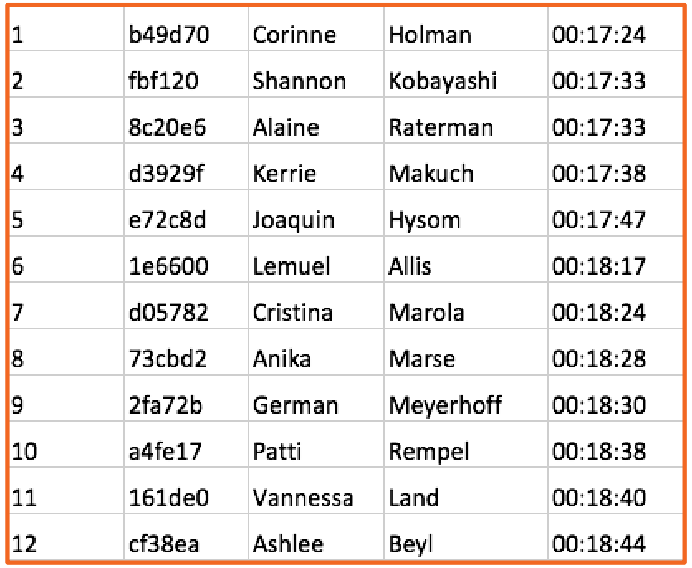
Querying Race Data with Hive

Elton Stoneman

@EltonStoneman | blog.sixeyed.com





1	b49d70	Corinne	Holman	00:17:24
2	fbf120	Shannon	Kobayashi	00:17:33
3	8c20e6	Alaine	Raterman	00:17:33
4	d3929f	Kerrie	Makuch	00:17:38
5	e72c8d	Joaquin	Hysom	00:17:47
6	1e6600	Lemuel	Allis	00:18:17
7	d05782	Cristina	Marola	00:18:24
8	73cbd2	Anika	Marse	00:18:28
9	2fa72b	German	Meyerhoff	00:18:30
10	a4fe17	Patti	Rempel	00:18:38
11	161de0	Vannessa	Land	00:18:40
12	cf38ea	Ashlee	Beyl	00:18:44

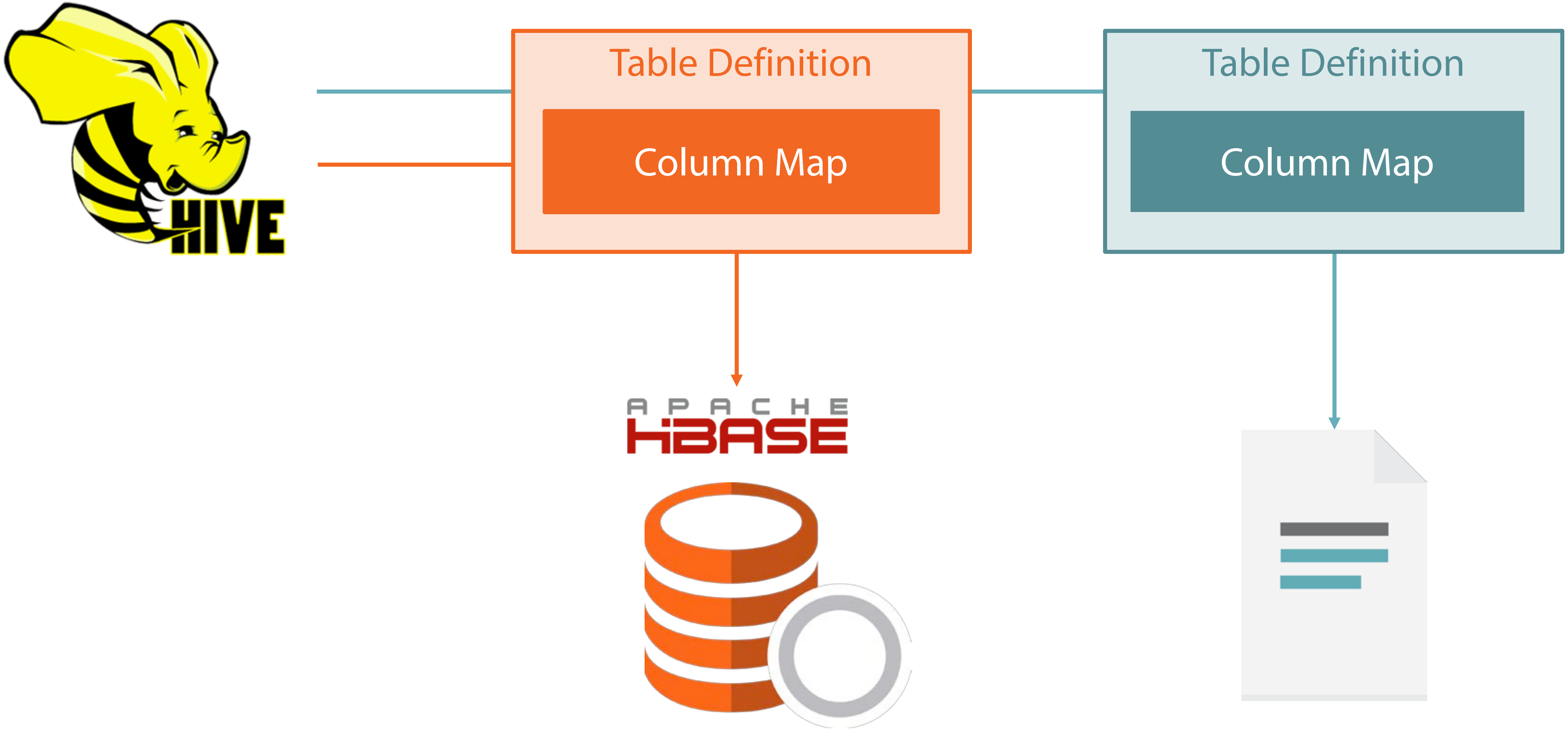


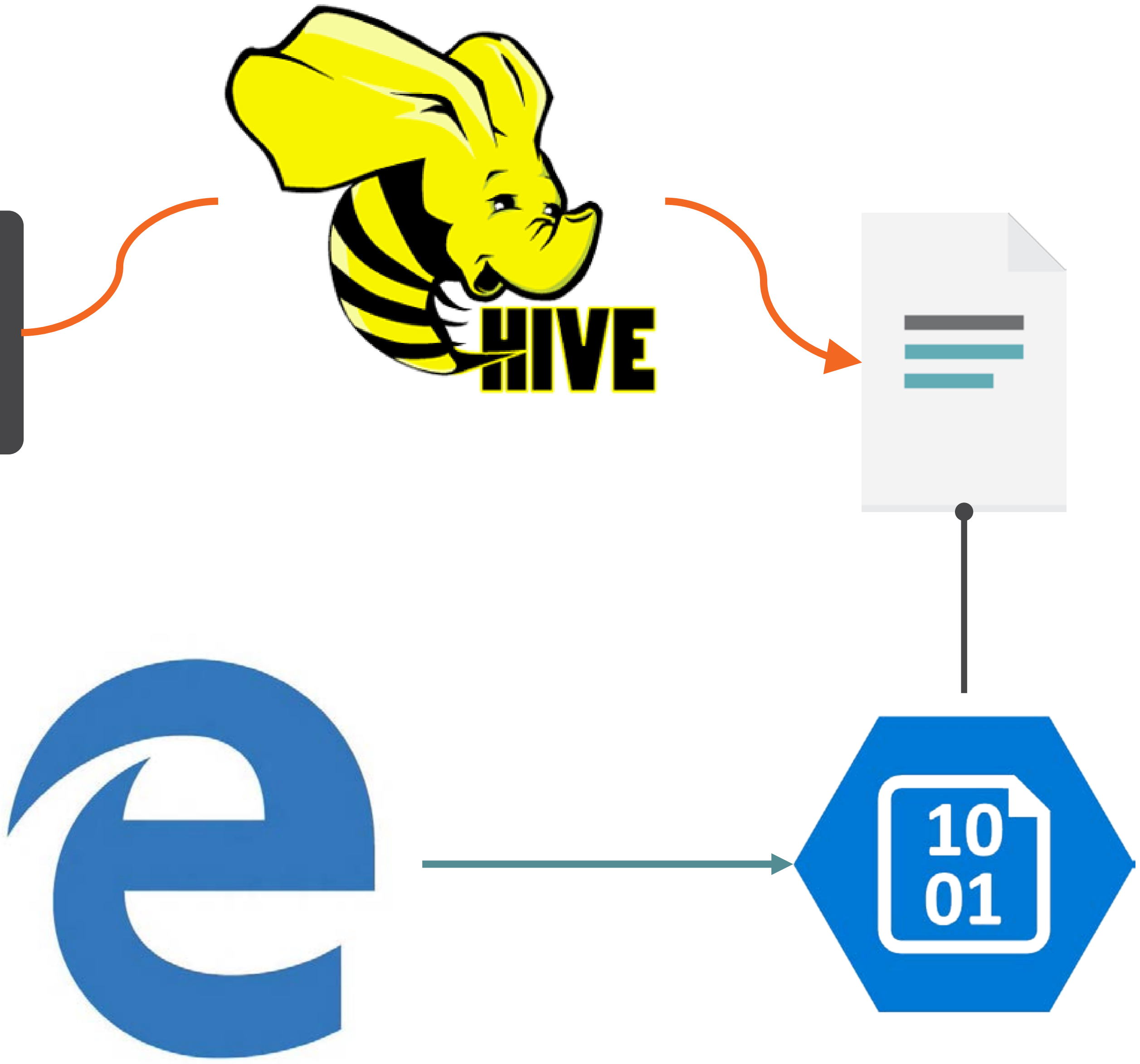
Table Definition

Column Map

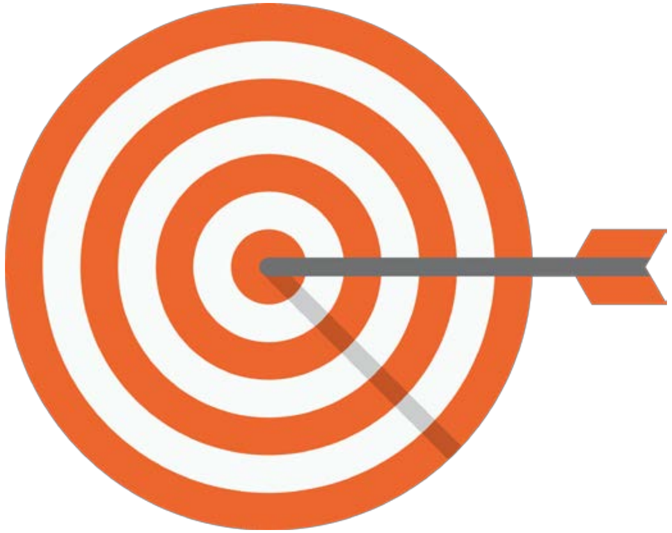
Table Definition

Column Map

```
INSERT INTO TABLE race_results
SELECT...
```



Module Goals



Generate Race Results

Map HBase Tables in Hive

Query Data with HiveQL

Map CSV Files in Hive

Write Output to Azure

```
SELECT MIN(lastWrite), MAX(lastWrite), COUNT(*)
FROM races
WHERE process = 'RaceTiming-1009-1442-5-1444398328';
```

HiveQL

Querying HBase

```
CREATE EXTERNAL TABLE races(lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```

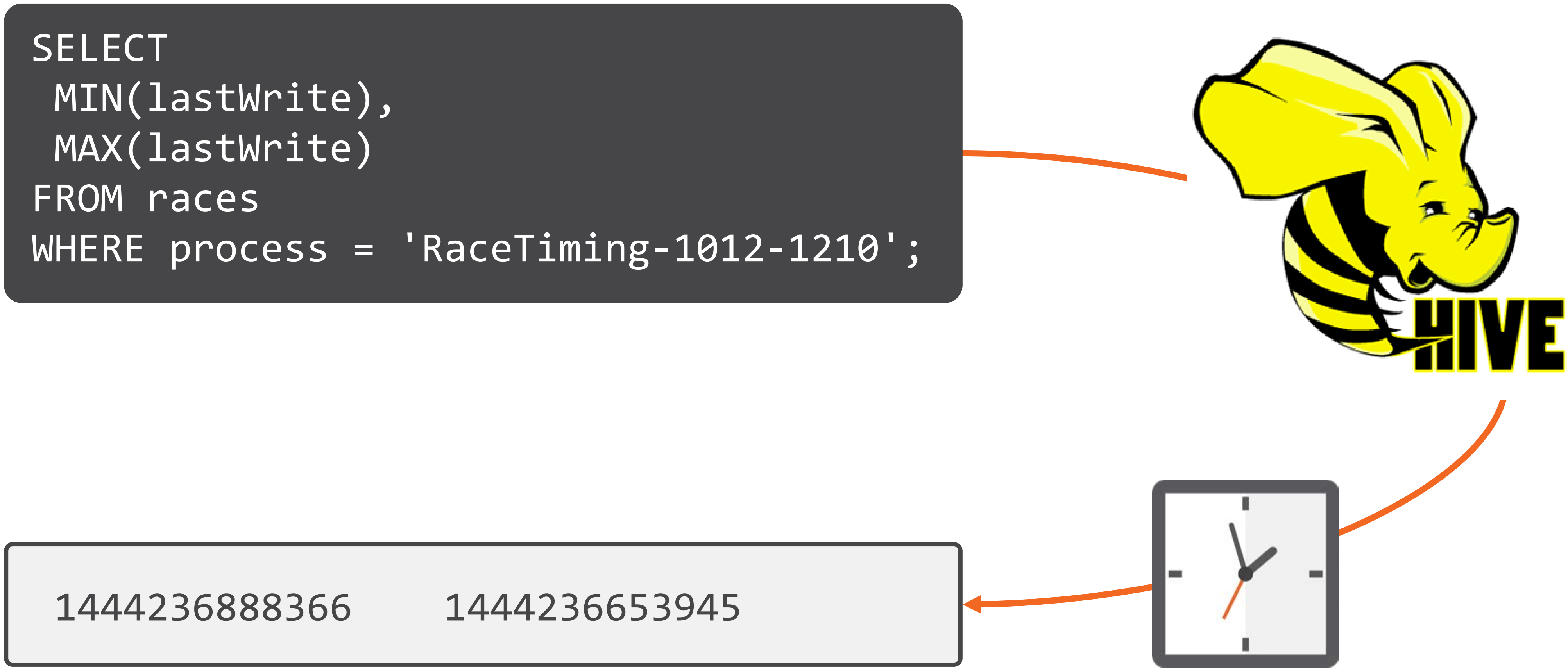
Create Table

Mapping external data

```
SELECT
 MIN(lastWrite),
 MAX(lastWrite)
FROM races
WHERE process = 'RaceTiming-1012-1210';
```

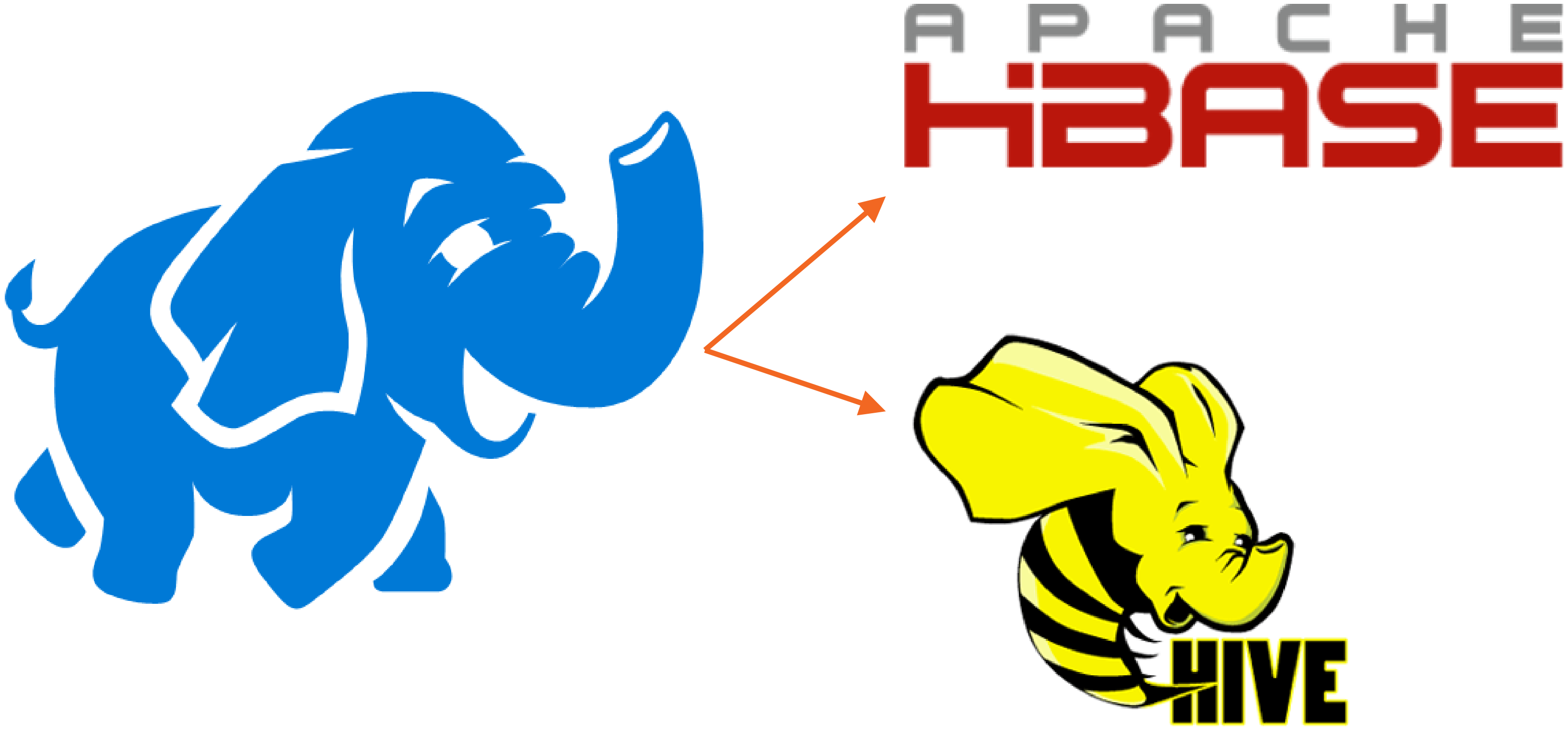
races	au

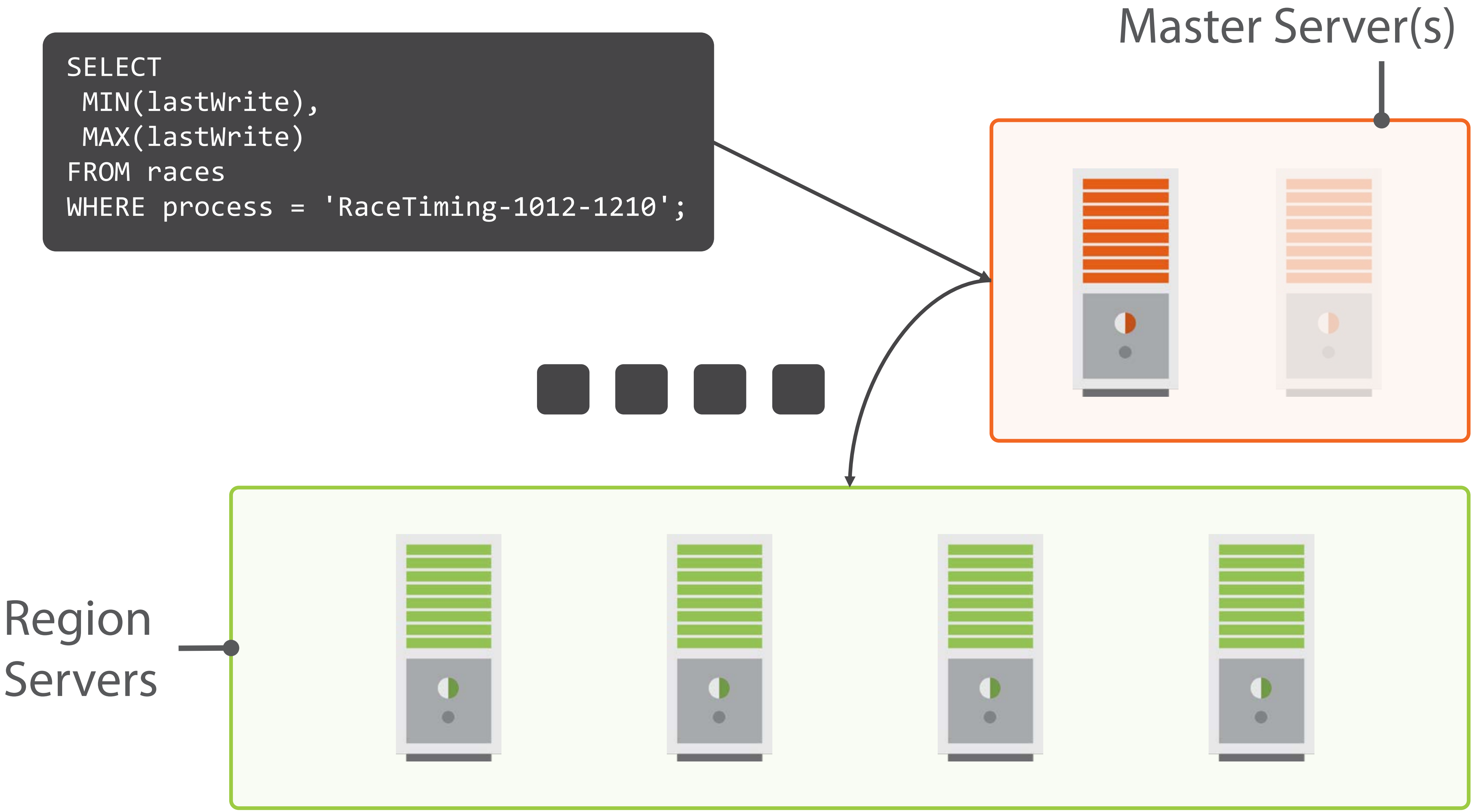
RowKey	au:w	au:p
a6545da436	1444641212016	RaceTiming-1012-1210



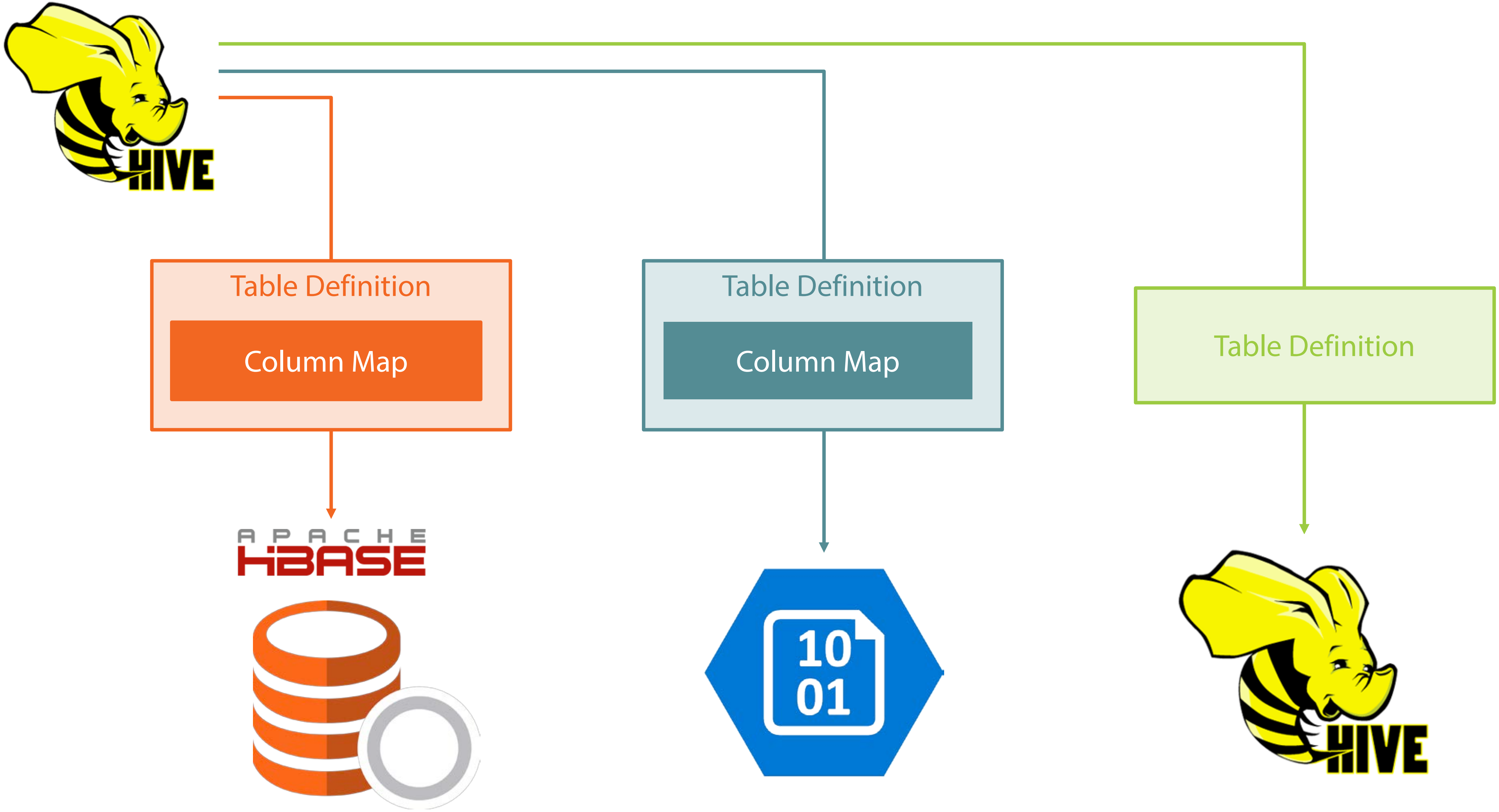
```
SELECT
 MIN(lastWrite),
 MAX(lastWrite)
FROM races
WHERE process = 'RaceTiming-1012-1210';
```

```
1444236888366    1444236653945
```


APACHE
HBASE
HIVE



```
SELECT
 MIN(lastWrite),
 MAX(lastWrite)
FROM races
WHERE process = 'RaceTiming-1012-1210';
```


HIVE
Table Definition
Column Map
Table Definition
Column Map
Table Definition
APACHE
HBASE
10
01
HIVE

External Tables

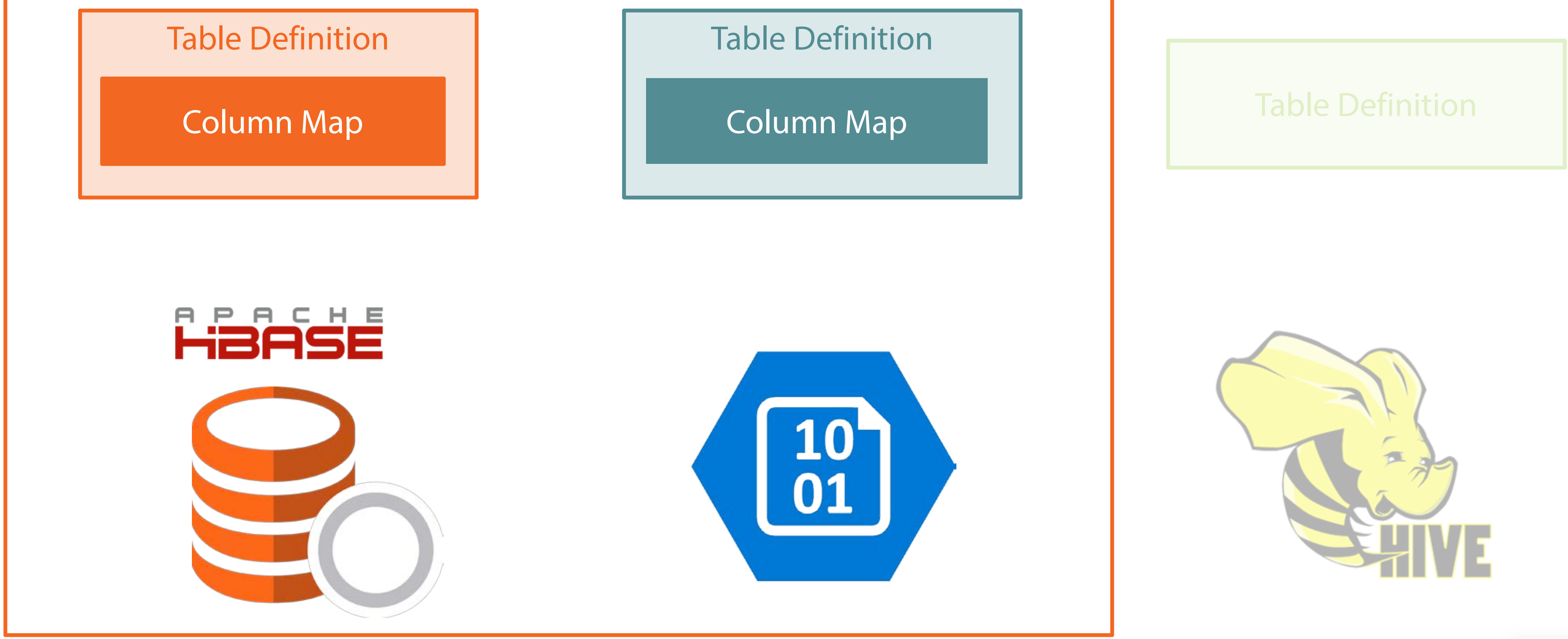


Table Definition

Column Map

Table Definition

Column Map

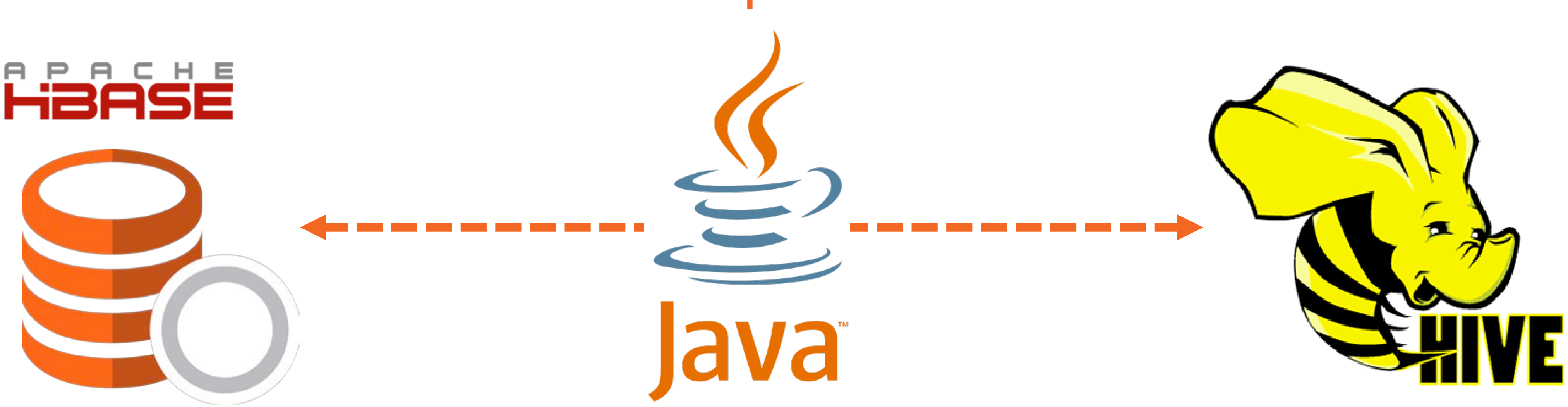
Table Definition

```
CREATE EXTERNAL TABLE races(rowkey STRING, lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```

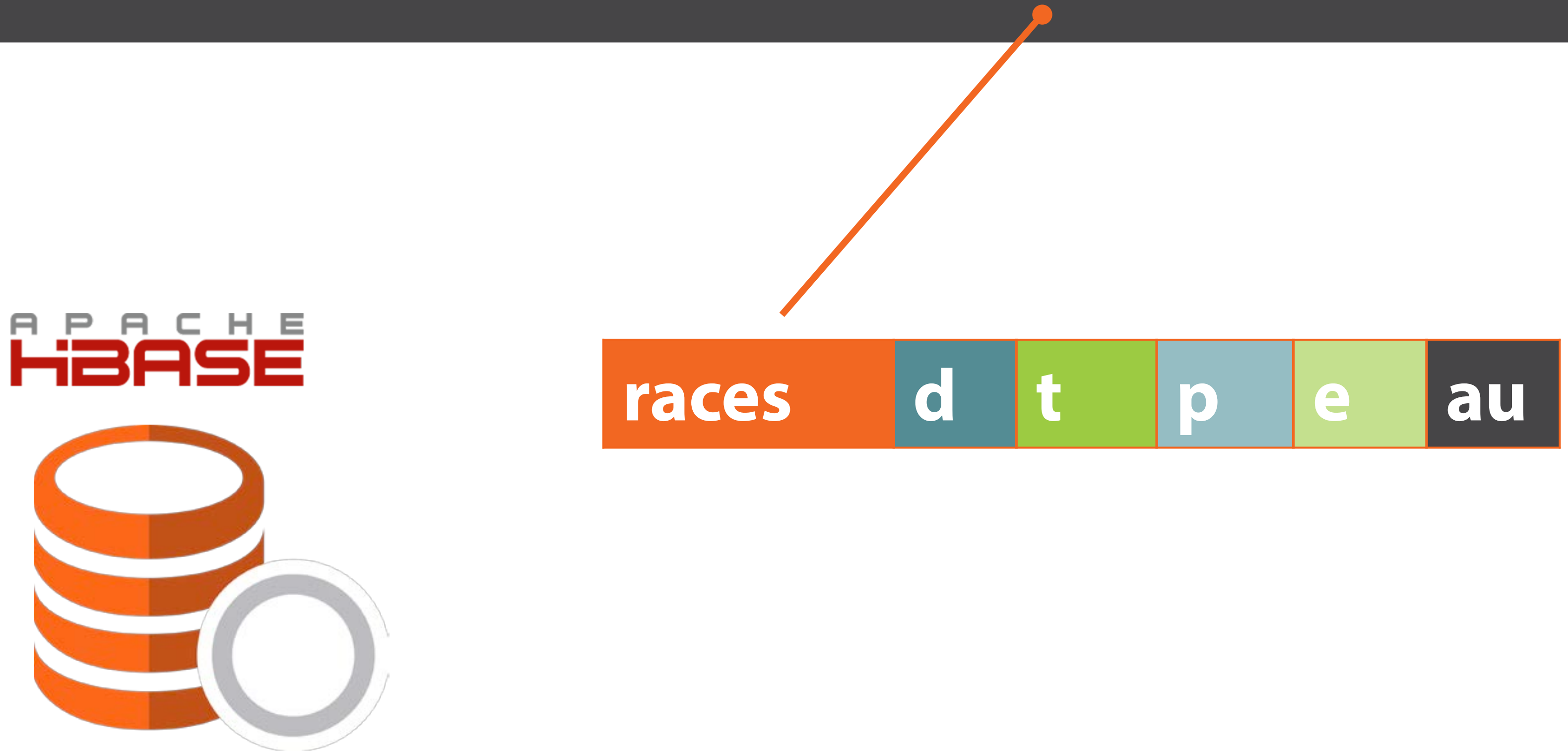
races

lastWrite	process
"1444236888366"	"RaceTiming-1012-1210"
"1444236888366"	"RaceTiming-1012-1210"
"1444236888366"	"RaceTiming-1012-1210"
NULL	"RaceTiming-1012-1530"

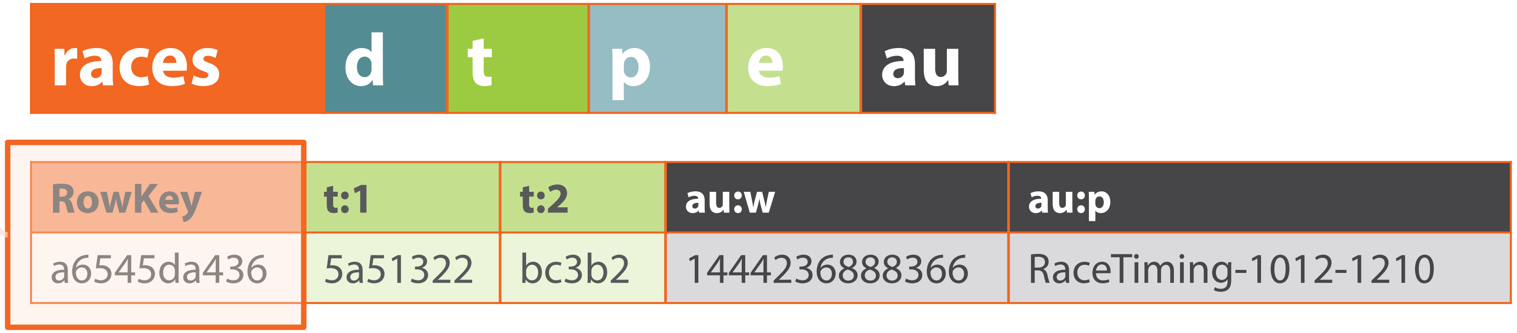
```
CREATE EXTERNAL TABLE races(rowkey STRING, lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```



```
CREATE EXTERNAL TABLE races(rowkey STRING, lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```

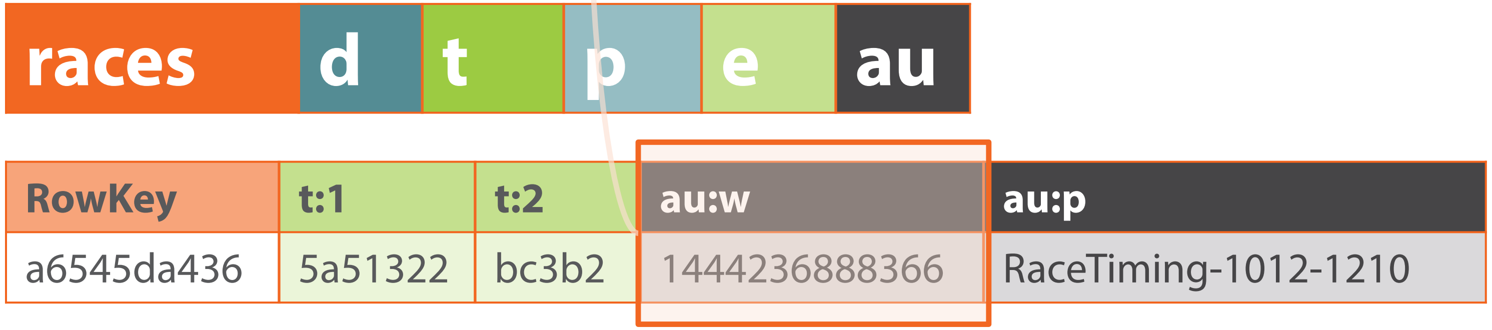


```
CREATE EXTERNAL TABLE races(rowkey STRING, lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = ':key,au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```



RowKey	t:1	t:2	au:w	au:p
a6545da436	5a51322	bc3b2	1444236888366	RaceTiming-1012-1210

```
CREATE EXTERNAL TABLE races(rowkey STRING, lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = ':key,au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```



races	d	t	p	e	au

RowKey	t:1	t:2	au:w	au:p
a6545da436	5a51322	bc3b2	1444236888366	RaceTiming-1012-1210

```
CREATE EXTERNAL TABLE races(rowkey STRING, lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```

races	d	t	p	e	au

RowKey	t:1	t:2	au:w	au:p
a6545da436	5a51322	bc3b2	1444236888366	RaceTiming-1012-1210

```
CREATE EXTERNAL TABLE races(rowkey STRING, lastWrite STRING, process STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,au:w,au:p')
TBLPROPERTIES ('hbase.table.name' = 'races');
```

races	d	t	p	e	au

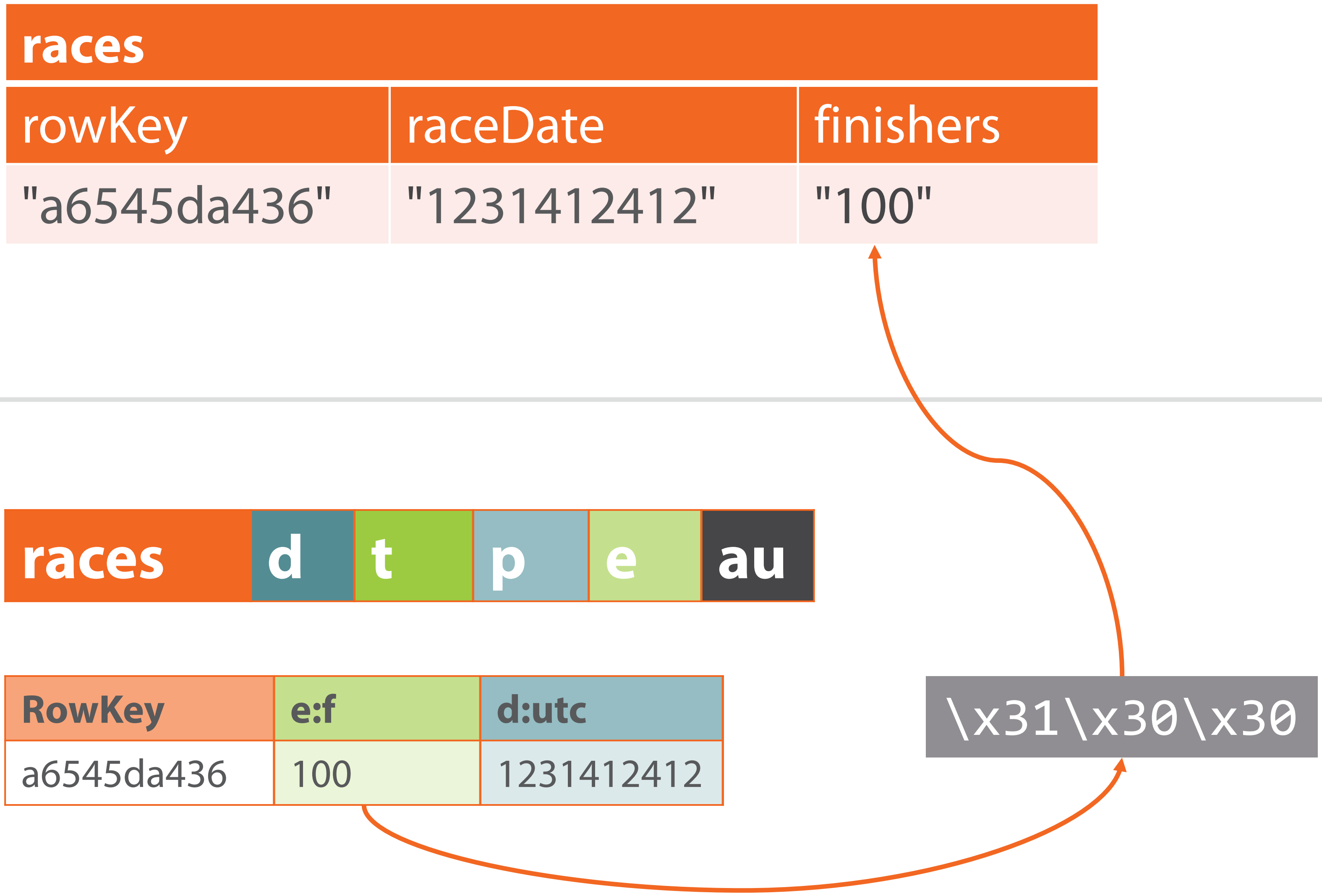
RowKey	t:1	t:2	au:w	au:p
a6545da436	5a51322	bc3b2	1444236888366	RaceTiming-1012-1210

Demo: HBase Data Types

Byte Arrays

Type Encoding

.NET & Java Interop



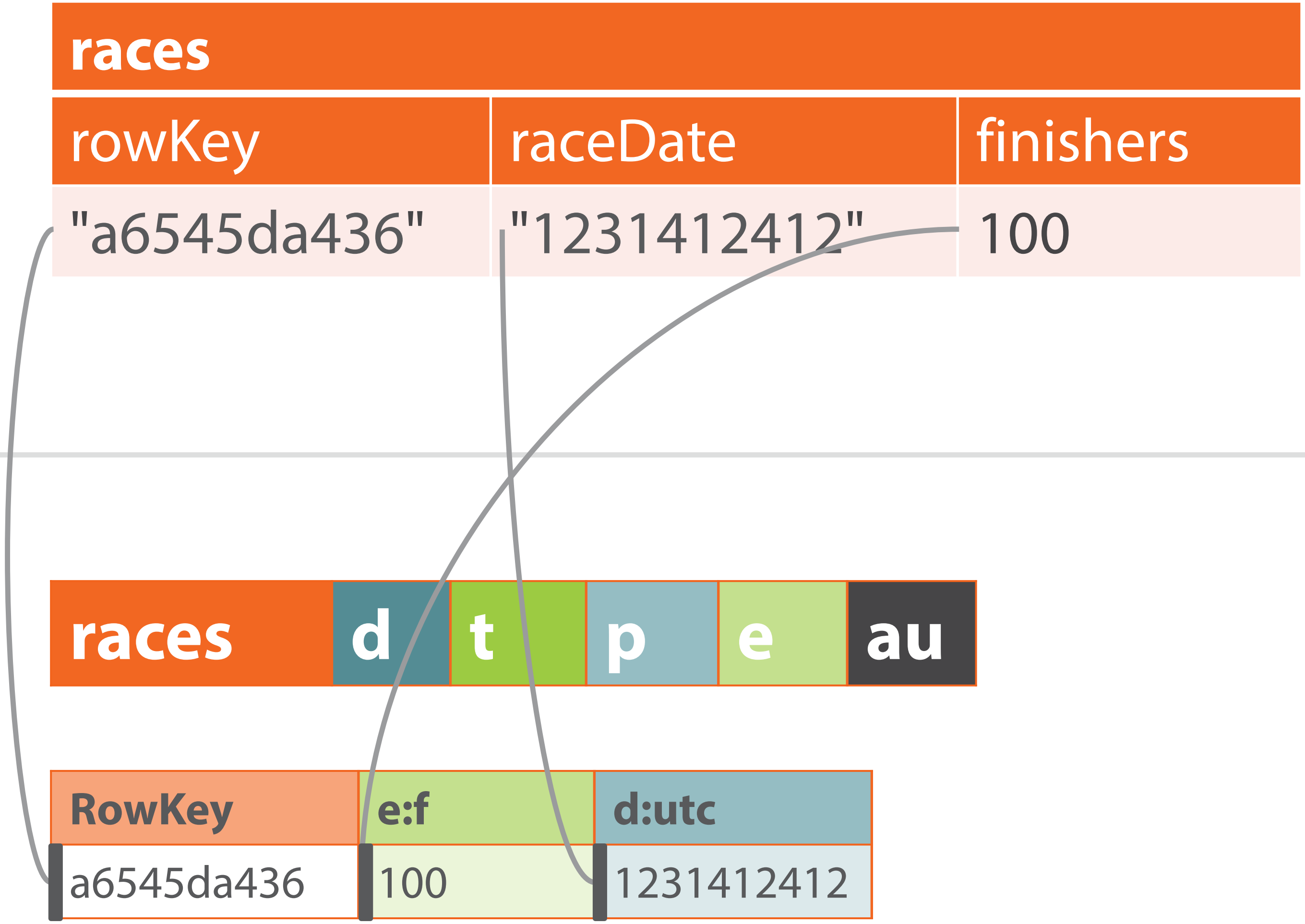
races

rowKey	raceDate	finishers
"a6545da436"	"1231412412"	"100"

races	**d**	**t**	**p**	**e**	**au**

RowKey	**e:f**	**d:utc**
a6545da436	100	1231412412

HIVE



races		
rowKey	raceDate	finishers
"a6545da436"	"1231412412"	100

APACHE HBASE

races	d	t	p	e	au

RowKey	e:f	d:utc
a6545da436	100	1231412412

races			
rowKey	raceDate	finishers	positions
"a6545da436"	"1231412412"	100	?

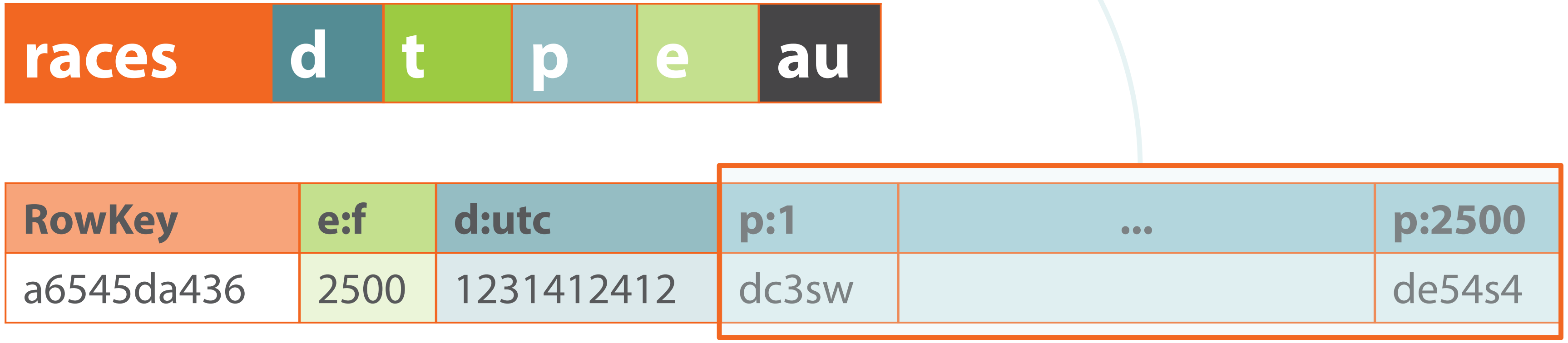
races	d	t	p	e	au

RowKey	e:f	d:utc	p:1	...	p:2500
a6545da436	2500	1231412412	dc3sw		de54s4

```
CREATE EXTERNAL TABLE races(rowkey STRING, positions MAP<STRING,STRING>)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,p:')
TBLPROPERTIES ('hbase.table.name' = 'races');
```

```
CREATE EXTERNAL TABLE races(rowkey STRING, positions MAP<STRING,STRING>)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,p:')
TBLPROPERTIES ('hbase.table.name' = 'races');
```

APACHE HBASE



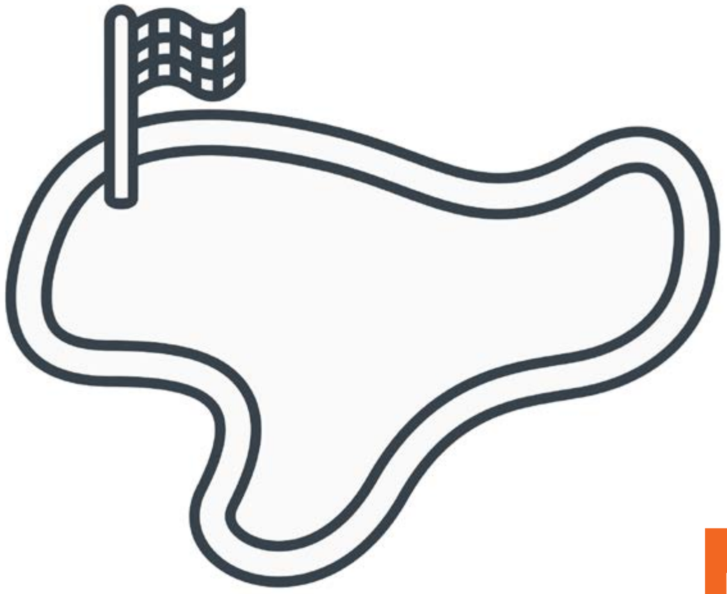
races	d	t	p	e	au

RowKey	e:f	d:utc	p:1	...	p:2500
a6545da436	2500	1231412412	dc3sw		de54s4

races			
rowKey	raceDate	finishers	positions
"a6545da436"	"1231412412"	100	{ ... }

races	d	t	p	e	au

RowKey	e:f	d:utc	p:1	...	p:2500
a6545da436	2500	1231412412	dc3sw		de54s4



Position	Racer ID	First Name	Last Name	Time
1	b49d70	Corinne	Holman	00:17:24
2	fbf120	Shannon	Kobayashi	00:17:33
3	8c20e6	Alaine	Raterman	00:17:33
4	d3929f	Kerrie	Makuch	00:17:38
5	e72c8d	Joaquin	Hysom	00:17:47

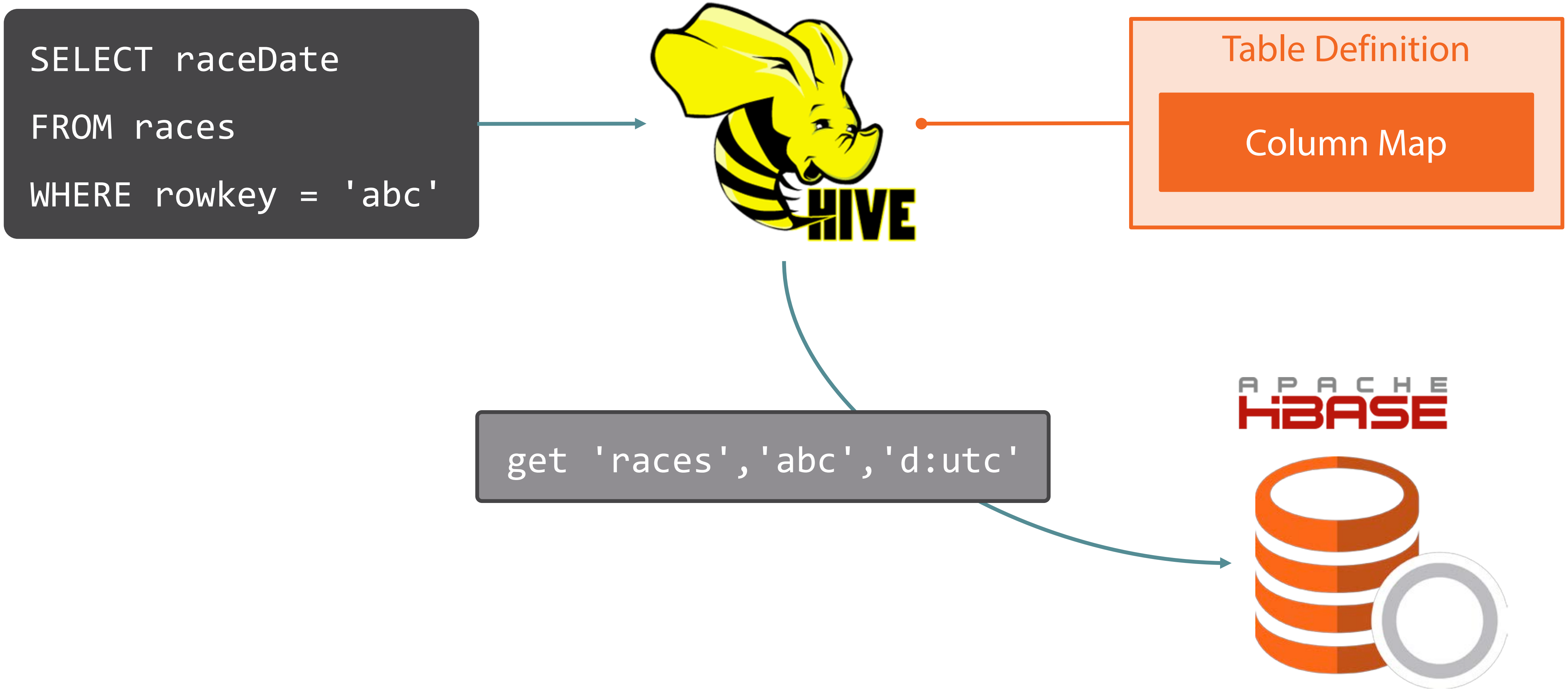
sector-times	te	d

RowKey	t:a31c4	t:bc3b2	d:1	d:t
e4324|a6545da436	1231412412	1231483602	71190	71190

Row key = **{racerId} | {raceId}**

Duration column family = **d**

Total duration column = **d:t**



```
SELECT raceDate
FROM races
WHERE rowkey = 'abc'
```

Table Definition

Column Map

```
get 'races','abc','d:utc'
```

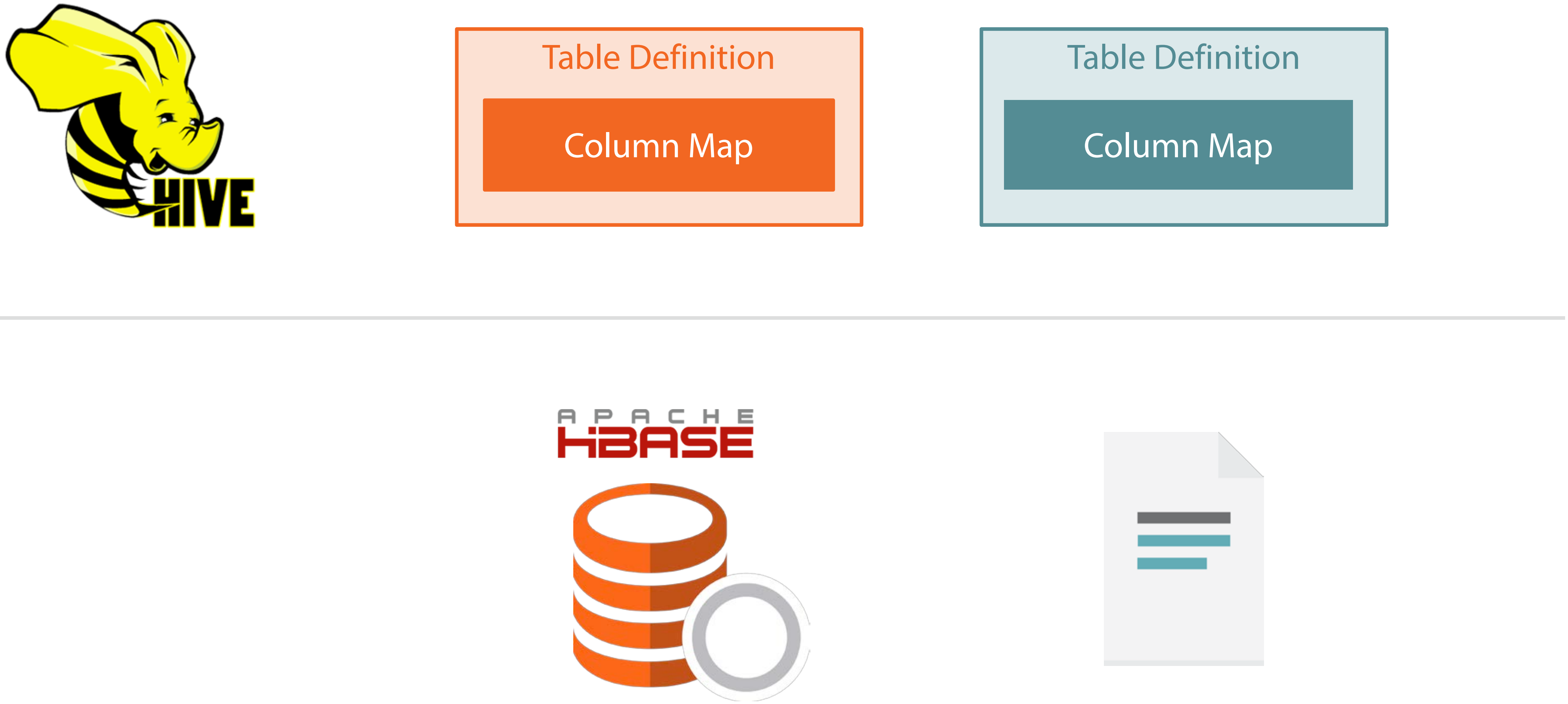


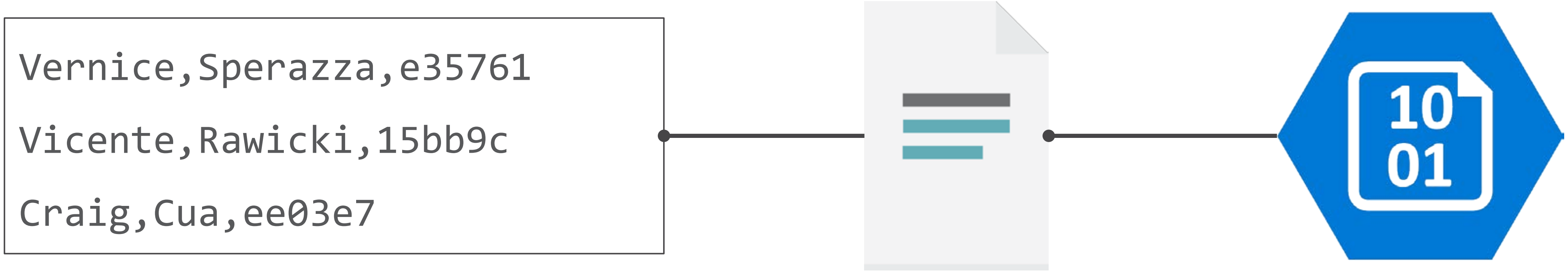
Table Definition

Column Map

Table Definition

Column Map

```
CREATE EXTERNAL TABLE racer_details
 (firstName STRING, lastName STRING, racerId STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/race-timing/racers';
```



```
Vernice,Sperazza,e35761
Vicente,Rawicki,15bb9c
Craig,Cua,ee03e7
```

Demo: Fetching Results

Join HBase to CSV

Filter CSV Source

APACHE HBASE

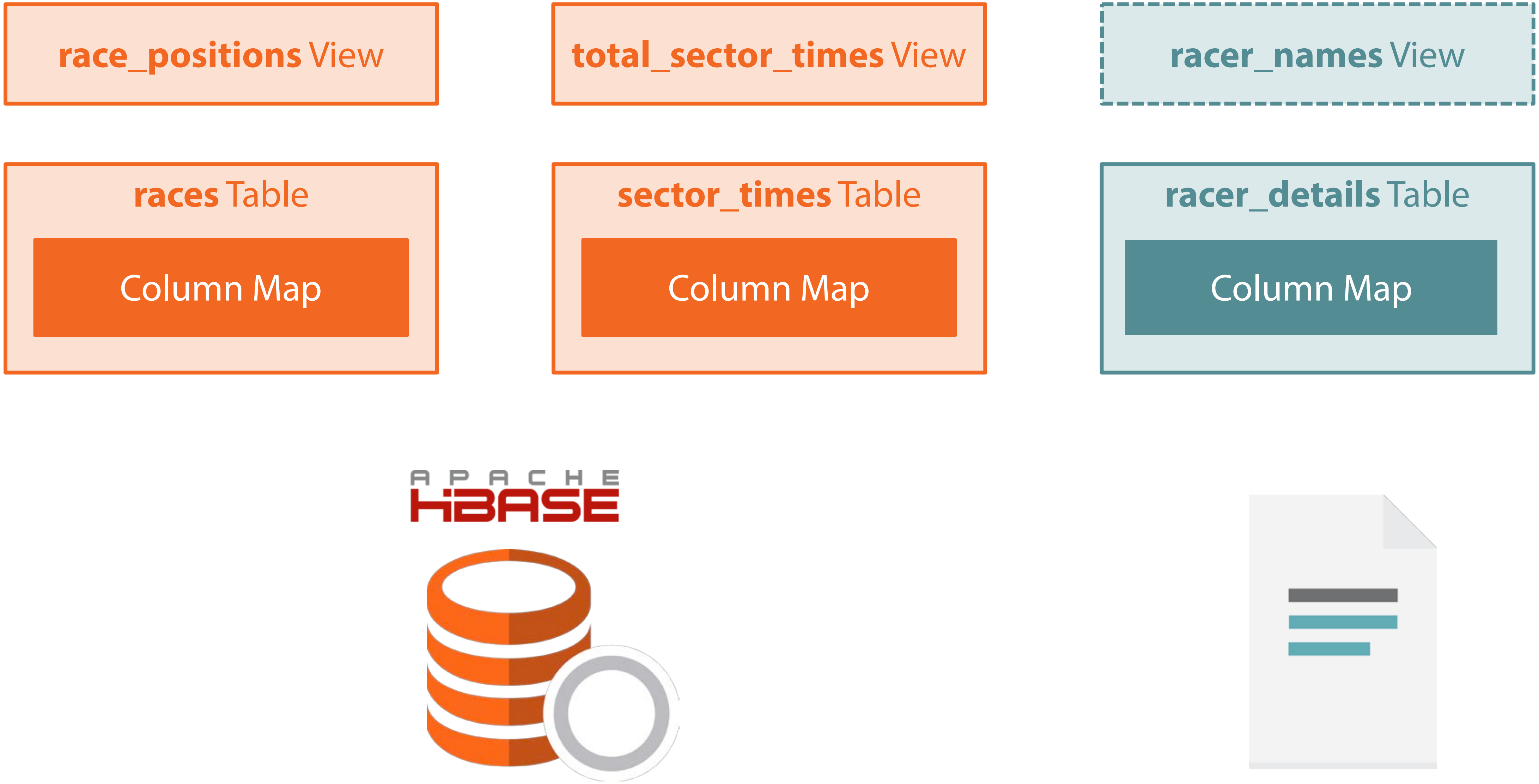
races	d	t	p	e	au

RowKey	e:f	d:utc	p:1	...	p:10
a6545da436	10	1231412412	e4324		rce2t132

sector-times	te	d

RowKey	t:a31c4	t:bc3b2	d:1	d:t
e4324|a6545da436	1231412412	1231483602	71190	71190

```
Naomi,Lavezzo, e4324
```



race_positions View

total_sector_times View

racer_names View

races Table

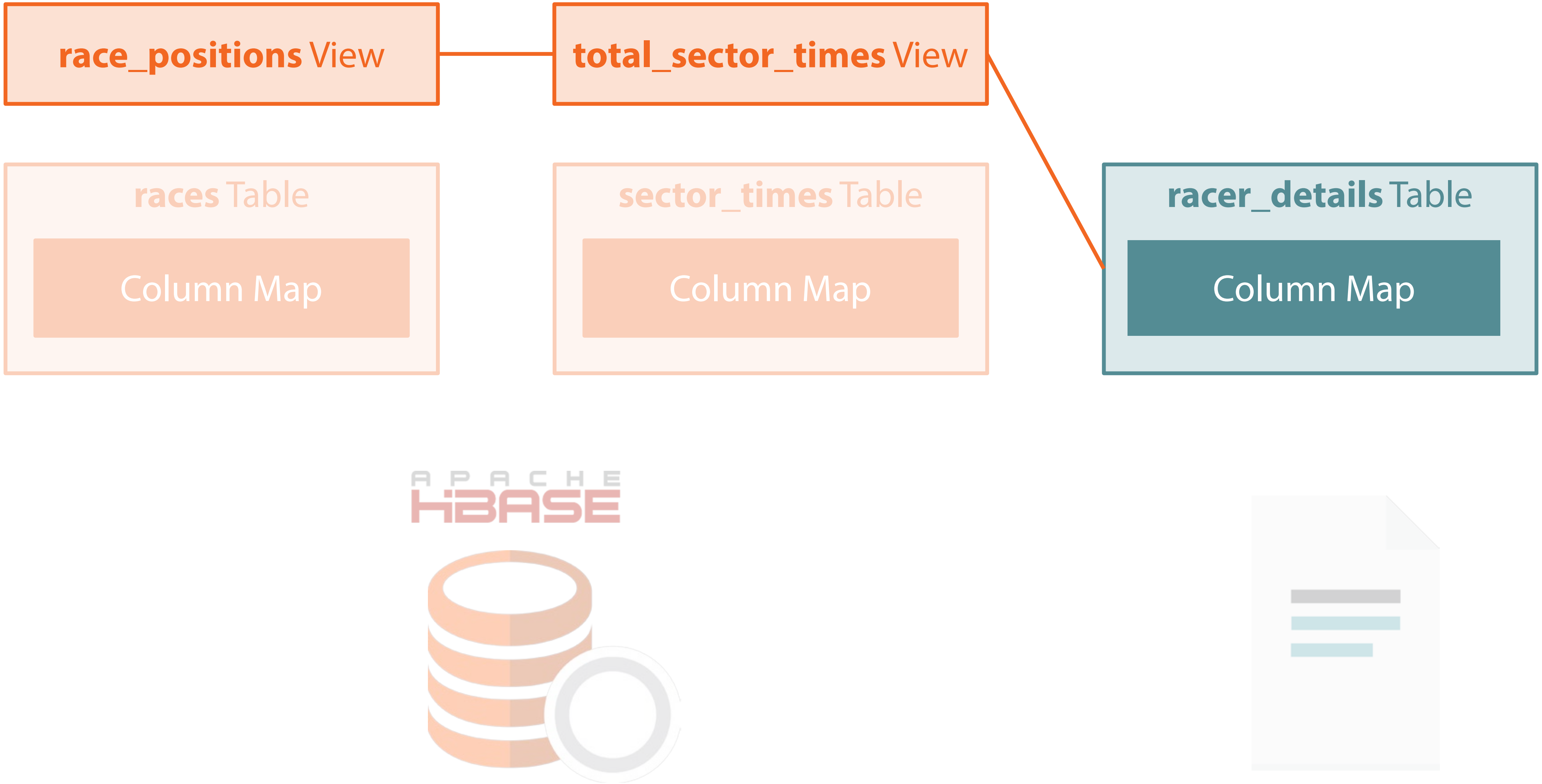
Column Map

sector_times Table

Column Map

racer_details Table

Column Map



race_positions View

total_sector_times View

races Table

Column Map

sector_times Table

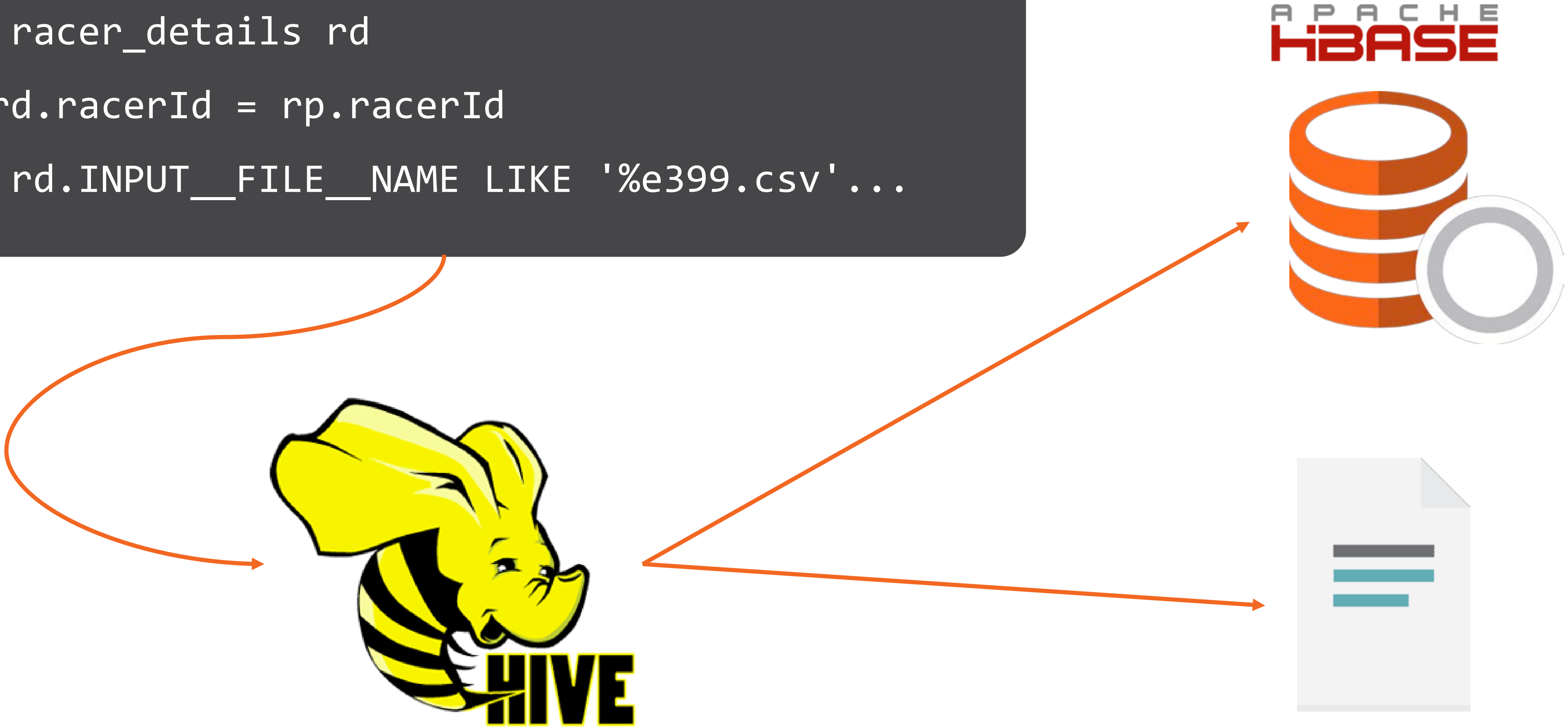
Column Map

racer_details Table

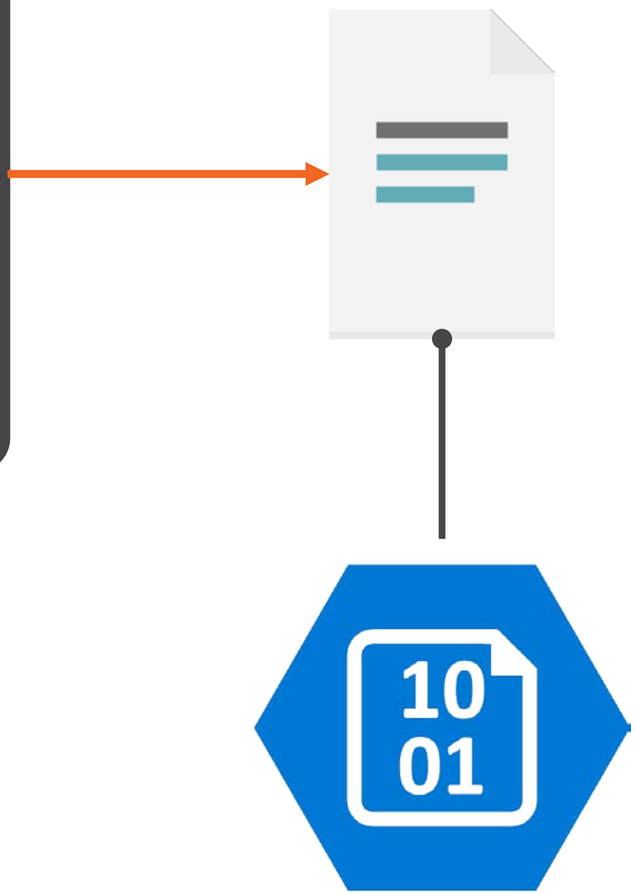
Column Map

APACHE HBASE

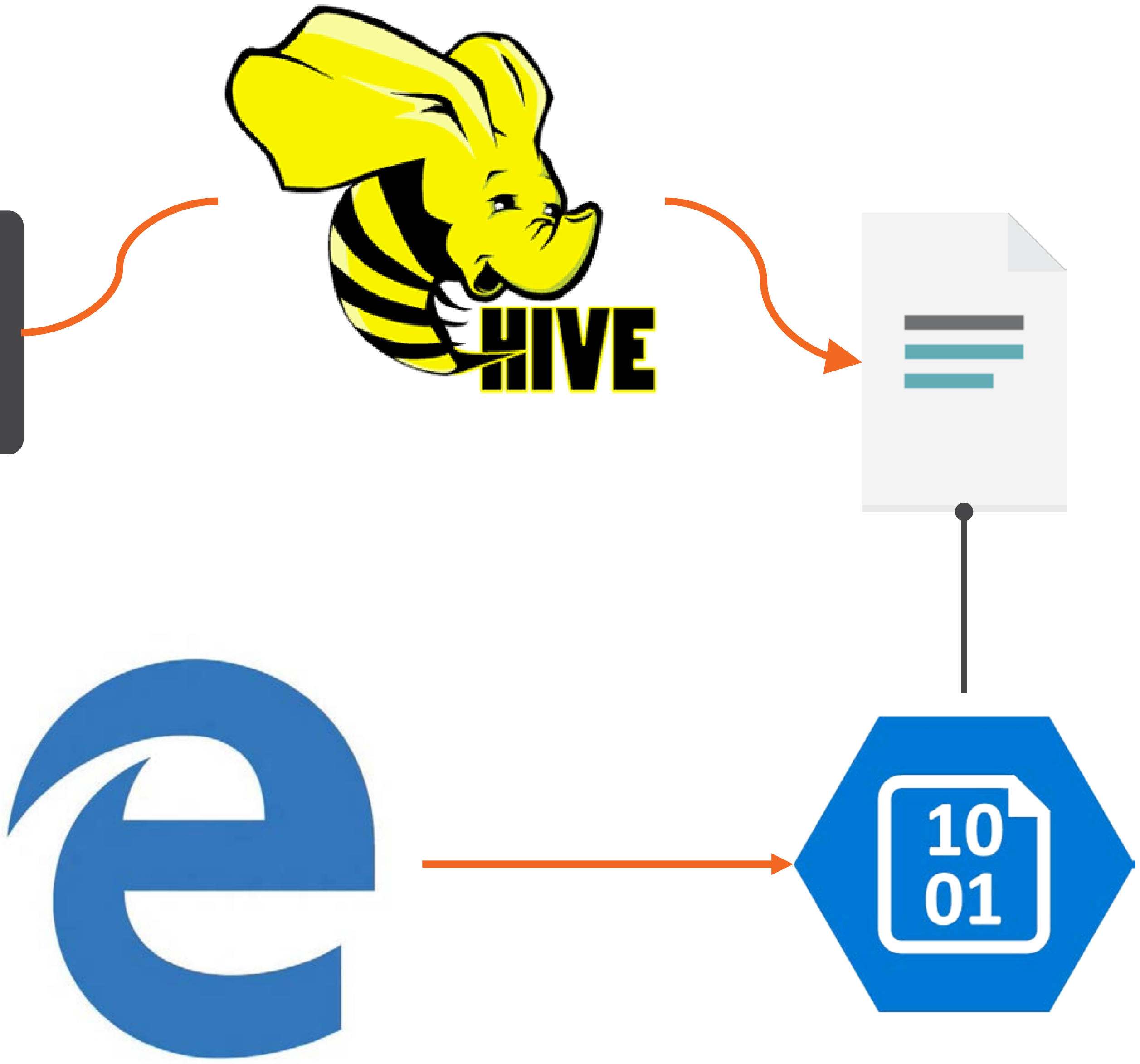
```
SELECT rp.position, rp.racerId, rd.firstName...
FROM race_positions rp
JOIN racer_details rd
 ON rd.racerId = rp.racerId
 AND rd.INPUT__FILE__NAME LIKE '%e399.csv'...
```



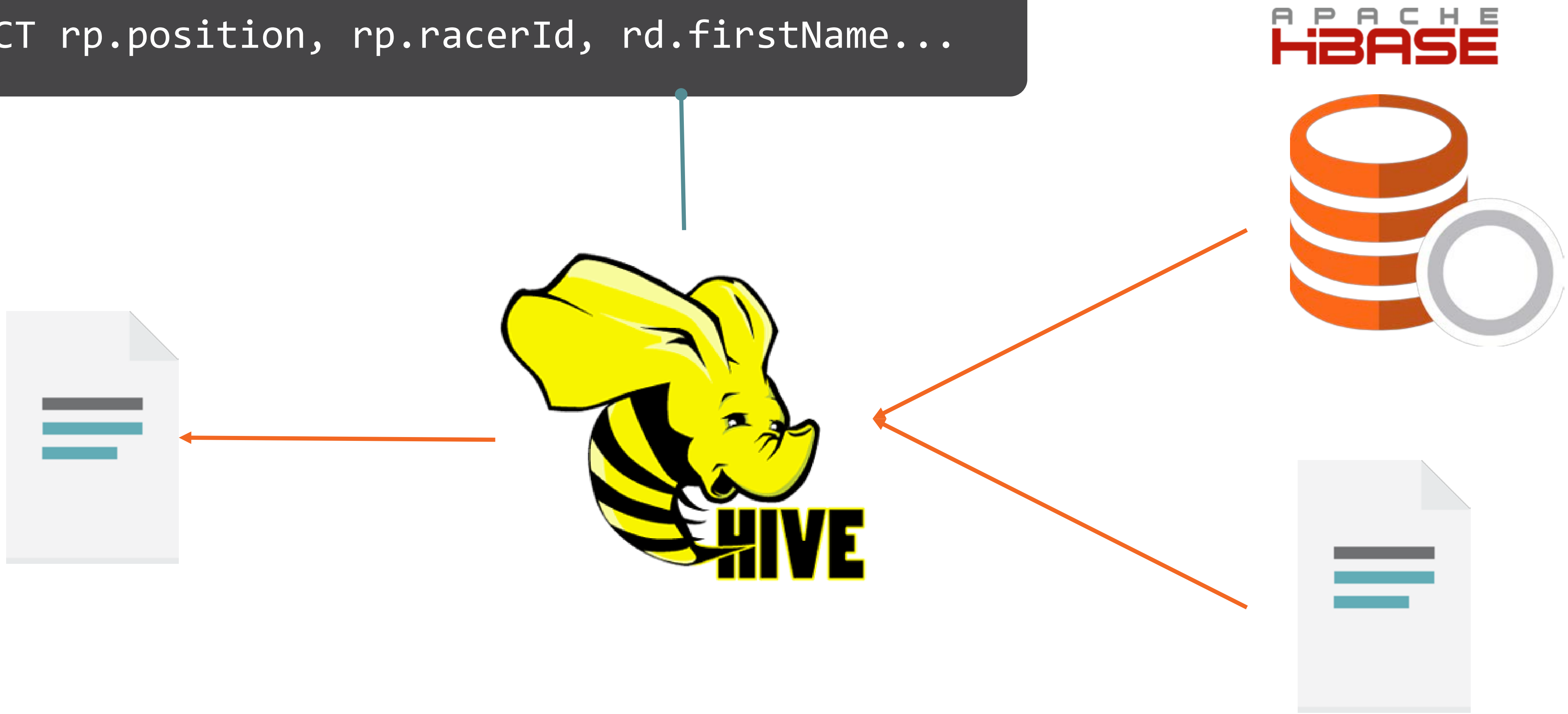
```
SELECT rp.position, rp.racerId, rd.firstName...
FROM race_positions rp
JOIN racer_details rd
 ON rd.racerId = rp.racerId
 AND rd.INPUT__FILE__NAME LIKE '%e399.csv'...
```



```
INSERT INTO TABLE race_results
SELECT...
```

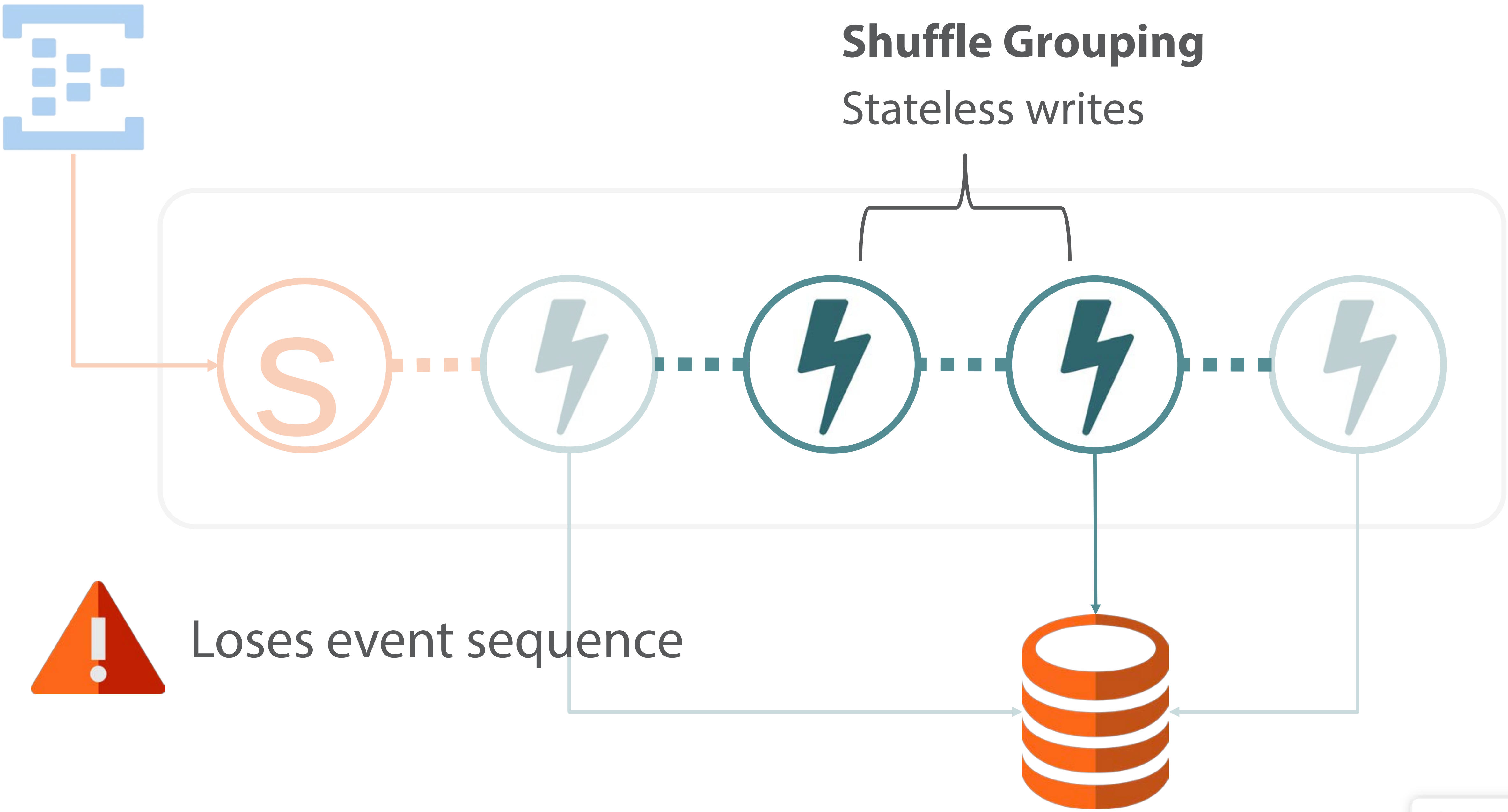


```
INSERT INTO TABLE race_results
PARTITION (raceId='e399')
SELECT rp.position, rp.racerId, rd.firstName...
```

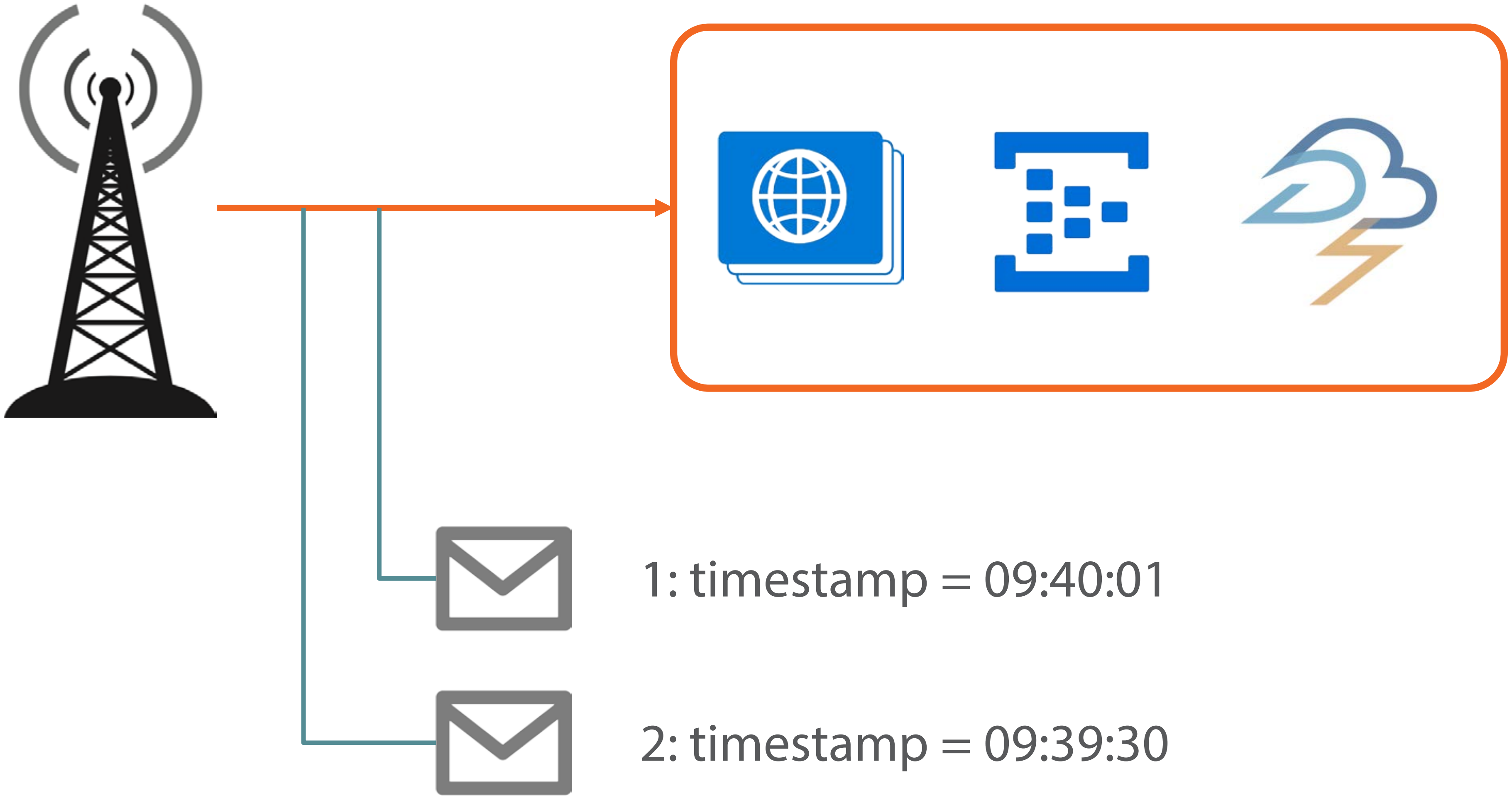


Shuffle Grouping

Stateless writes

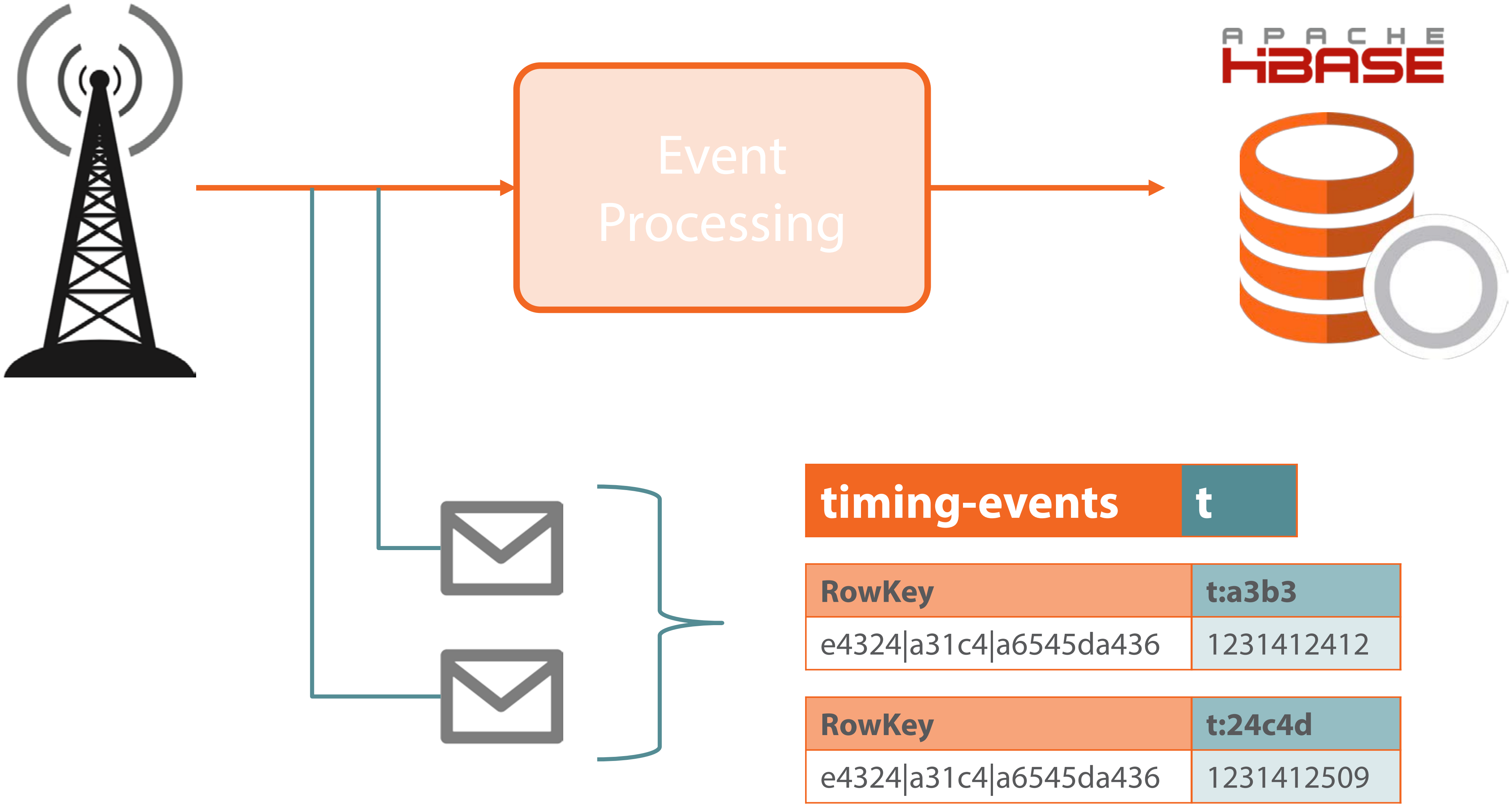


Loses event sequence



1: timestamp = 09:40:01

2: timestamp = 09:39:30



timing-events	t

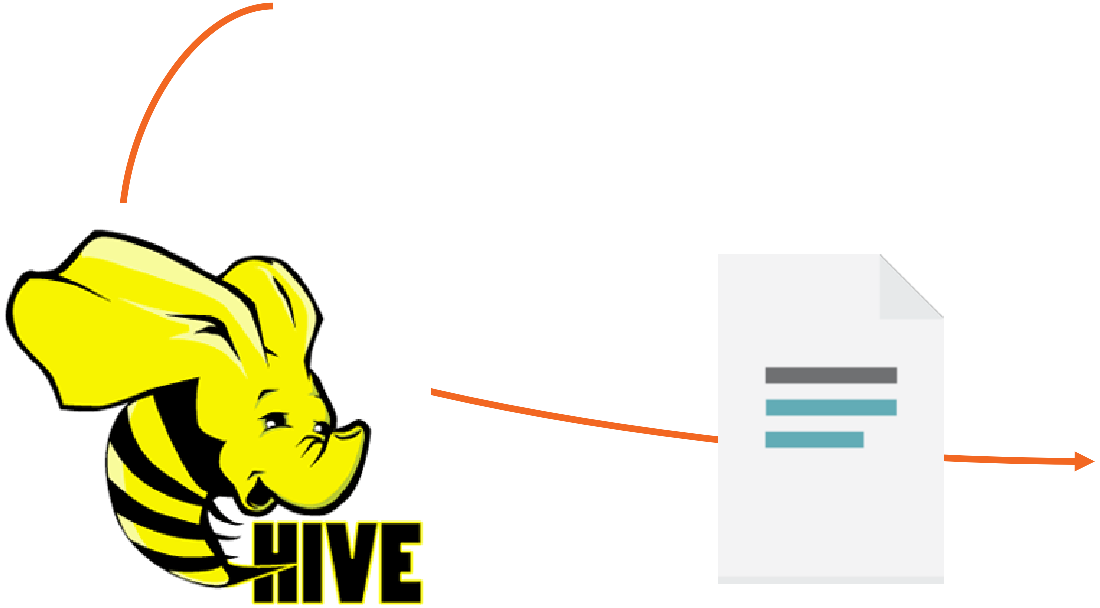
RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

RowKey	t:24c4d
e4324|a31c4|a6545da436	1231412509

timing-events	t

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

RowKey	t:24c4d
e4324|a31c4|a6545da436	1231412509

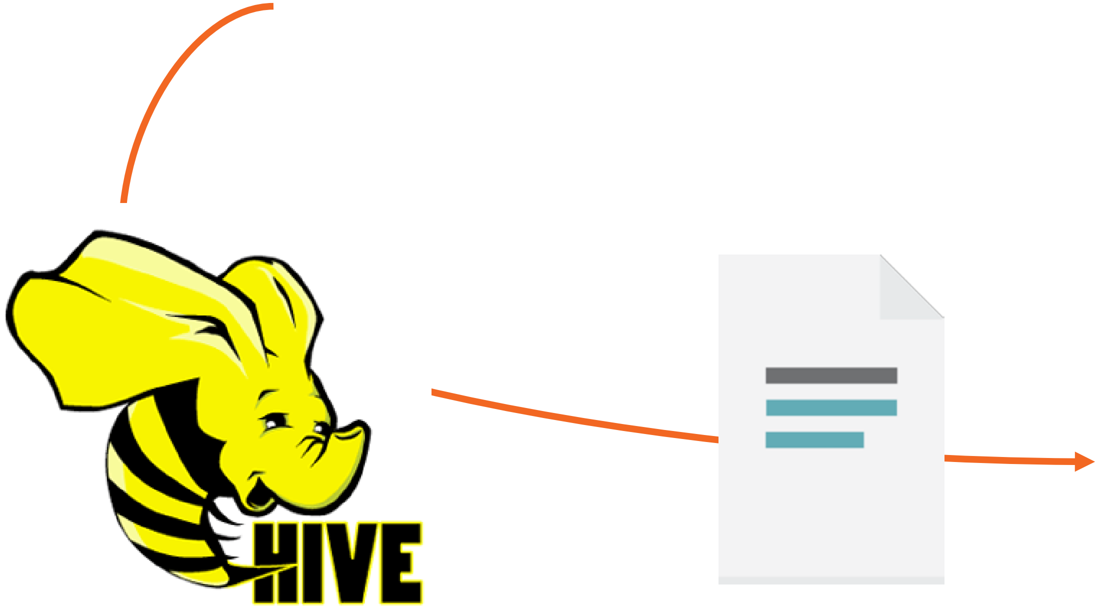


1	b49d70	Corinne	Holman	00:17:24
2	fbf120	Shannon	Kobayashi	00:17:33
3	8c20e6	Alaine	Raterman	00:17:33
4	d3929f	Kerrie	Makuch	00:17:38
5	e72c8d	Joaquin	Hysom	00:17:47
6	1e6600	Lemuel	Allis	00:18:17
7	d05782	Cristina	Marola	00:18:24
8	73cbd2	Anika	Marse	00:18:28
9	2fa72b	German	Meyerhoff	00:18:30
10	a4fe17	Patti	Rempel	00:18:38
11	161de0	Vannessa	Land	00:18:40
12	cf38ea	Ashlee	Beyl	00:18:44

timing-events	t

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

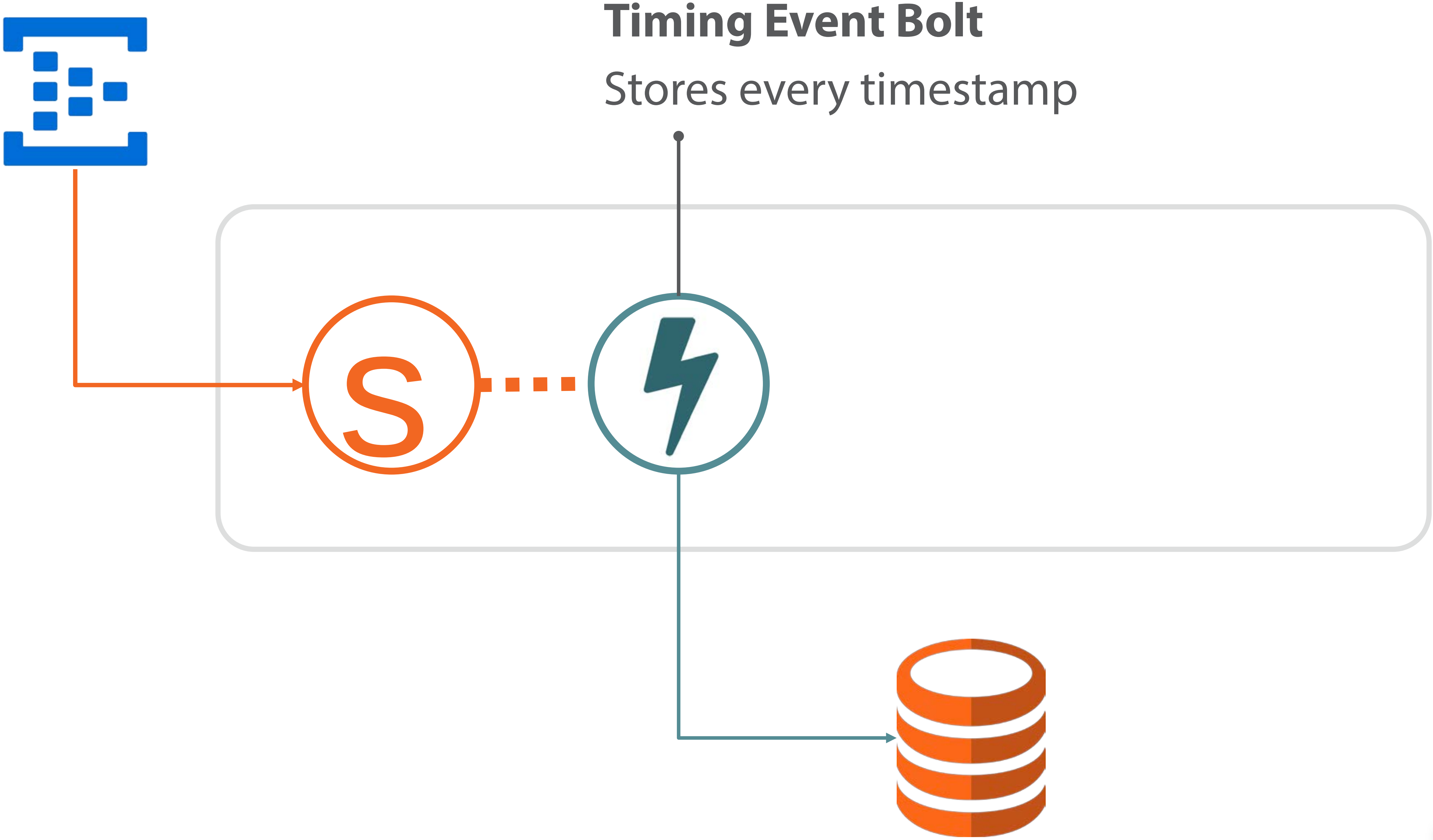
RowKey	t:24c4d
e4324|a31c4|a6545da436	1231412509

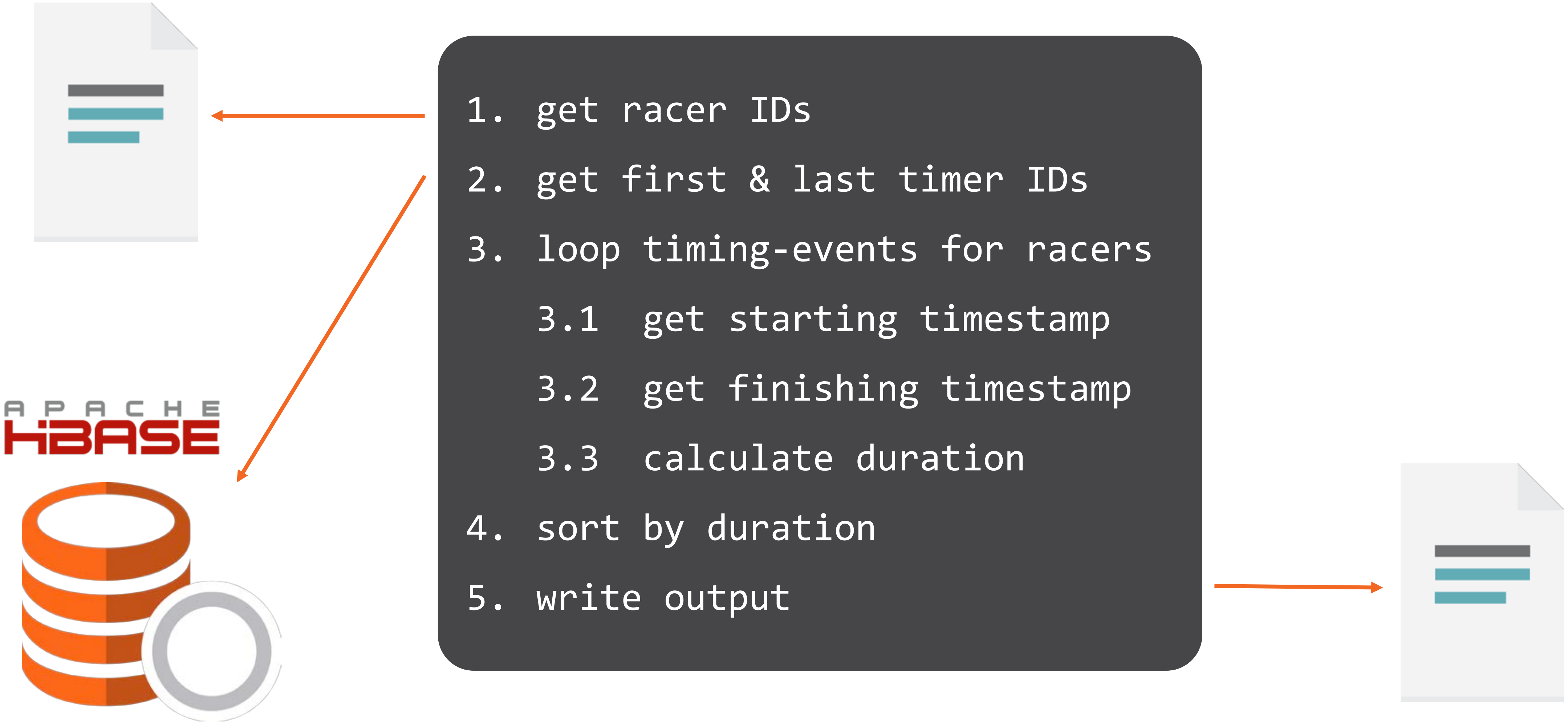


96	e08b91	Marylin	Cornelison	00:30:35
97	8c78ca	Jenise	Dulle	00:30:38
98	ee03e7	Craig	Cua	00:31:00
99	15bb9c	Vicente	Rawicki	00:31:02
100	e35761	Vernice	Sperazza	00:31:26
	30bf87	Joana	Reinecke	-
	746b04	Horace	Knapp	-
	bcc3a6	Val	Cellucci	-
	526d0f	Marta	Brace	-
	185650	Velda	Mancilla	-
	d0485c	Bettina	Kham	-
	f06a05	Anisha	Shulick	-

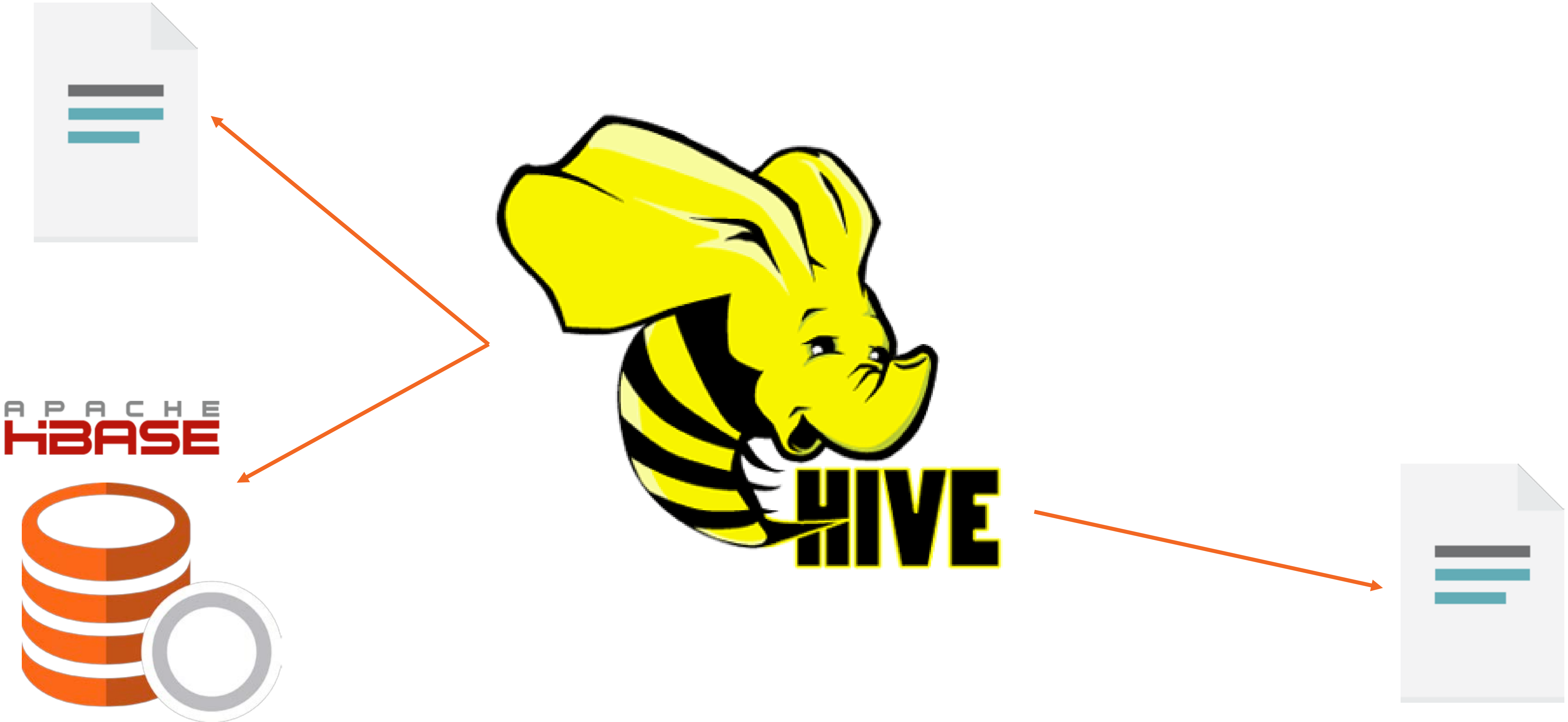
Timing Event Bolt

Stores every timestamp





```
1.  get racer IDs
2.  get first & last timer IDs
3.  loop timing-events for racers
    3.1  get starting timestamp
    3.2  get finishing timestamp
    3.3  calculate duration
4.  sort by duration
5.  write output
```


APACHE
HBASE
HIVE

Demo: Calculating Results

Earliest Event Timestamps

First & Last Timer IDs

Collection Functions

timing-events	**t**

RowKey	t:a3b3	t:24c4d
e4324|a31c4|a6545da436	1231412412	1231412509

t = timestamp **column family**

Name = random string

Value = timestamp

Demo: Calculating Results

Joins & Cross Joins

Ranking Results

Performance

APACHE HBASE

timing-events	**t**

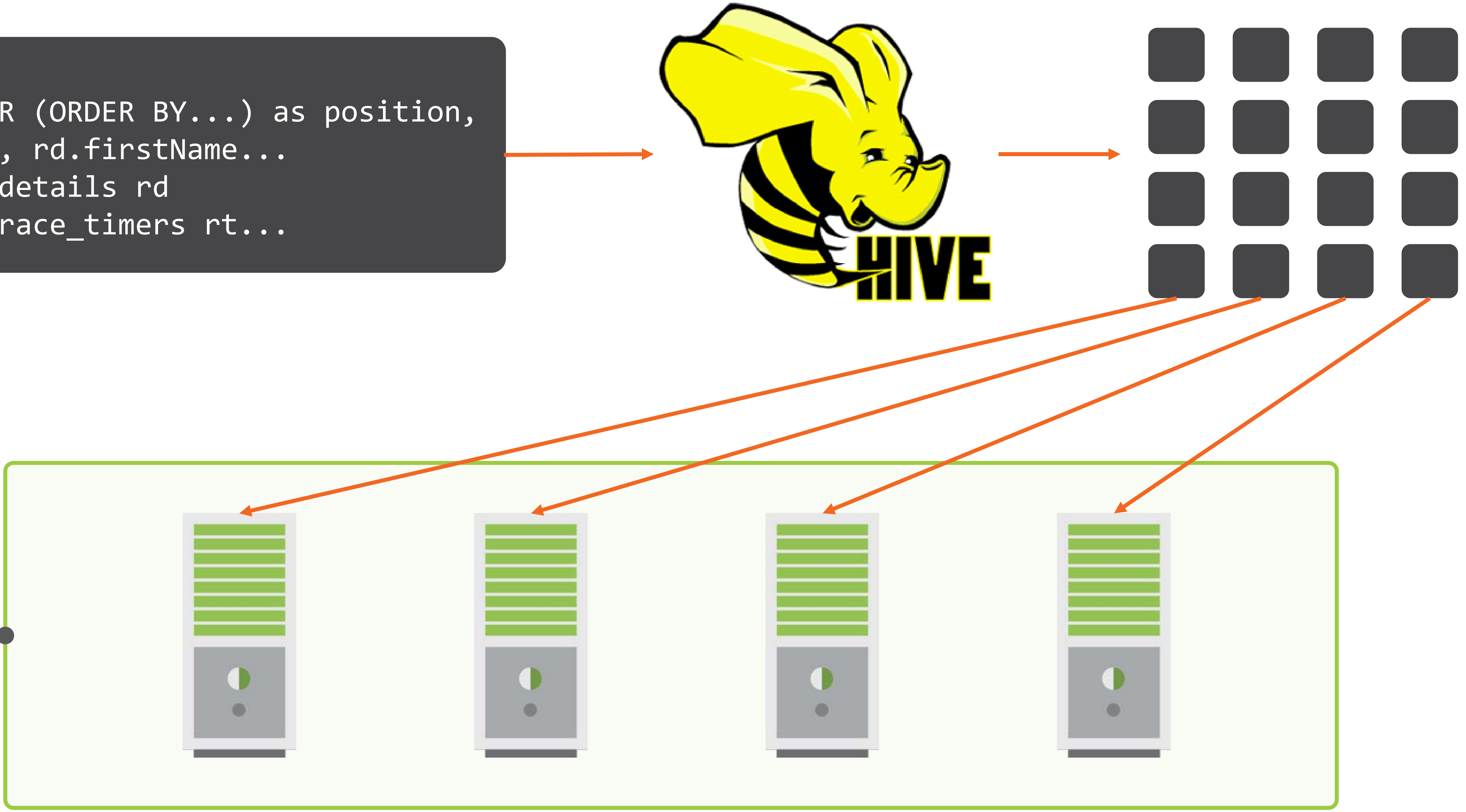
RowKey	**t:a3b3**	**t:24c4d**
e4324|a31c4|a6545da436	1231412412	1231412509

300K rows

Table scan

200+ seconds

```
SELECT
 rank() OVER (ORDER BY...) as position,
 rd.racerid, rd.firstName...
FROM racer_details rd
CROSS JOIN race_timers rt...
```



HBase
Region
Servers

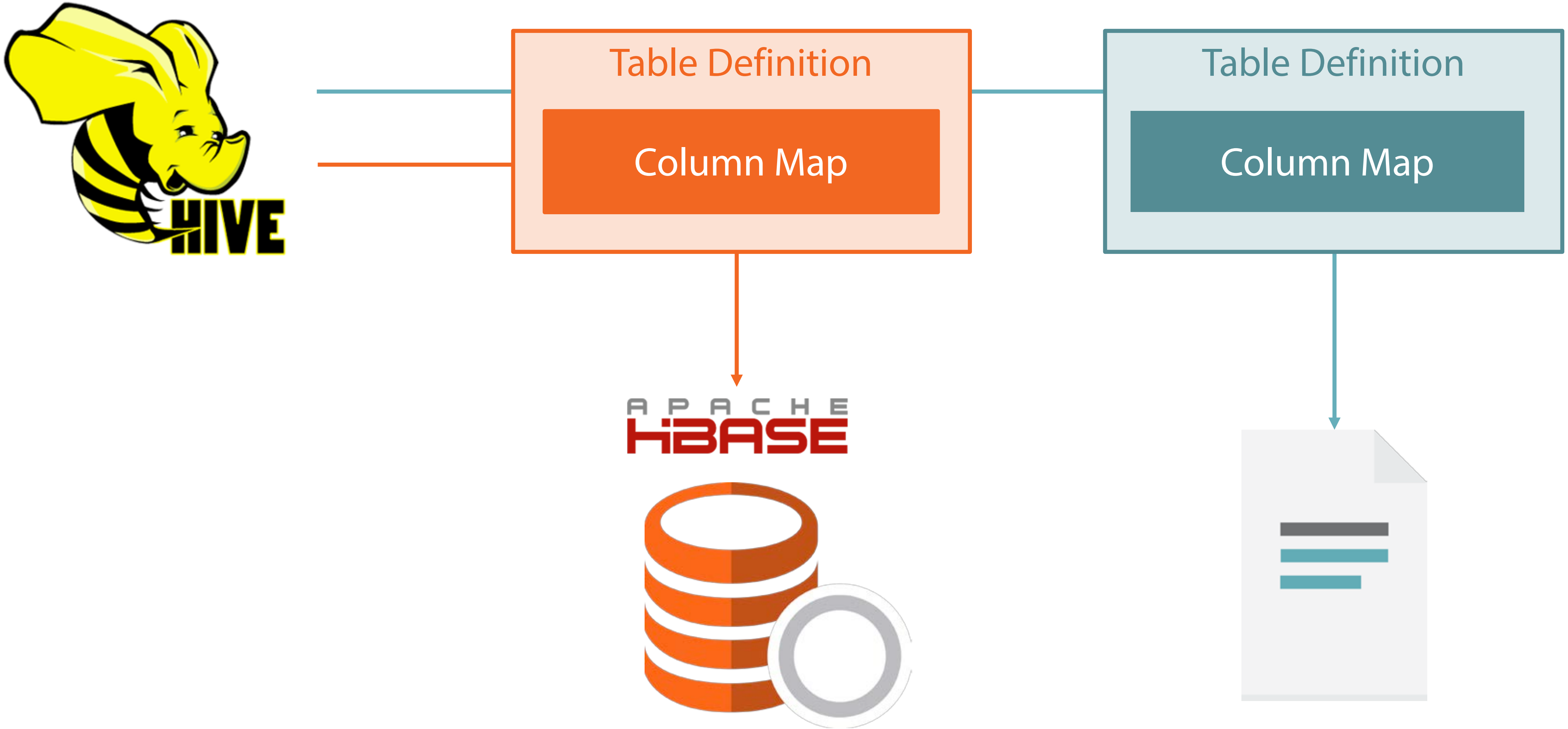


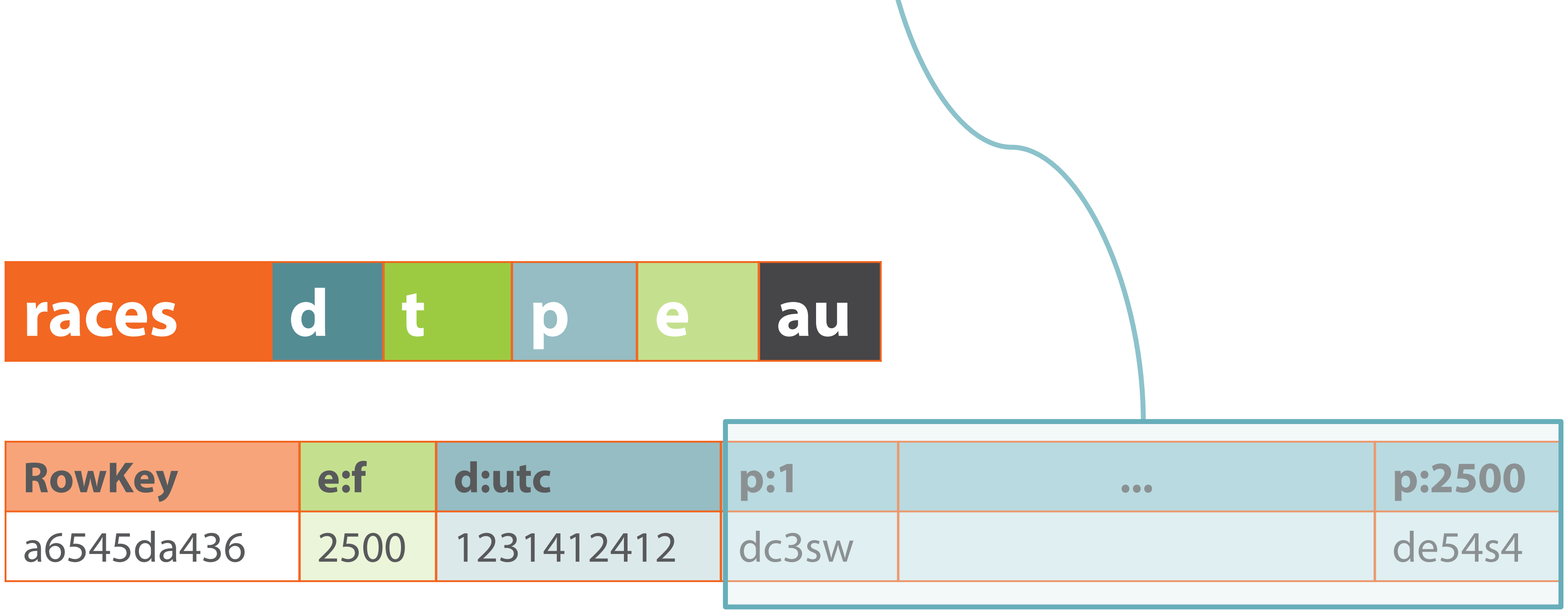
Table Definition

Column Map

Table Definition

Column Map

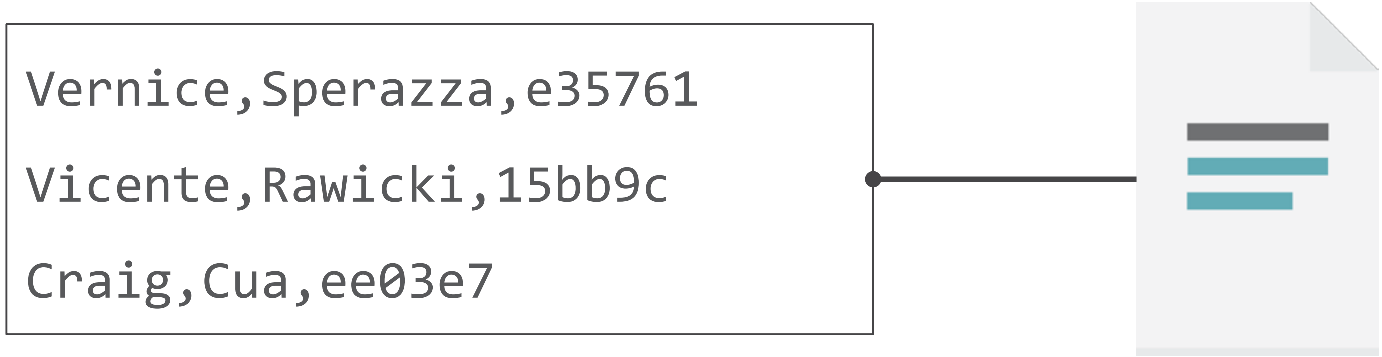
```
CREATE EXTERNAL TABLE races(rowkey STRING, positions MAP<STRING,STRING>)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ('hbase.columns.mapping' = :key,p:')
TBLPROPERTIES ('hbase.table.name' = 'races');
```



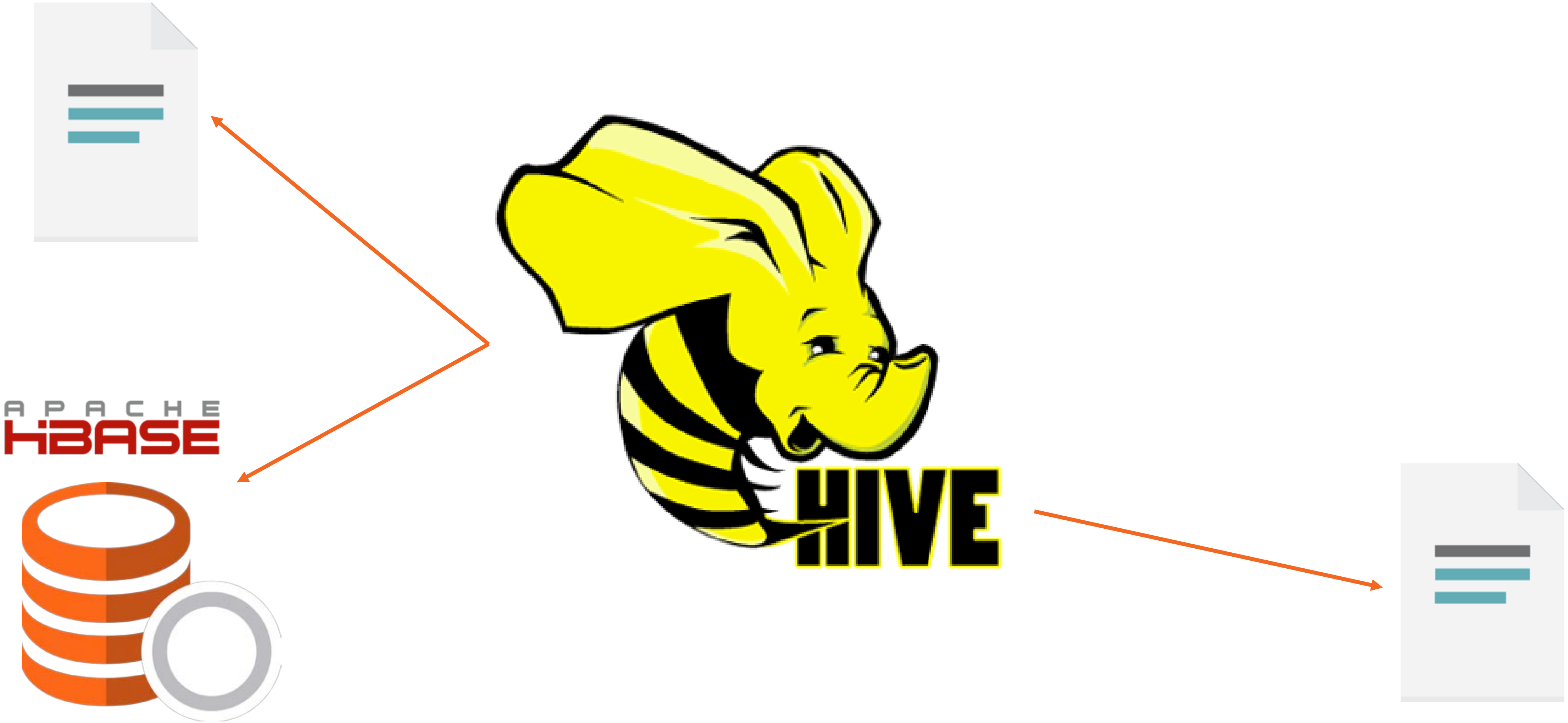
races	d	t	p	e	au

RowKey	e:f	d:utc	p:1	...	p:2500
a6545da436	2500	1231412412	dc3sw		de54s4

```
CREATE EXTERNAL TABLE racer_details
 (firstName STRING, lastName STRING, racerId STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/race-timing/racers';
```



```
Vernice,Sperazza,e35761
Vicente,Rawicki,15bb9c
Craig,Cua,ee03e7
```


APACHE HBASE
HIVE

```
CREATE VIEW earliest_timing_events(racerId, timerId, raceId, timestamp)
AS SELECT split(ROWKEY, '\\|')[0], split(ROWKEY, '\\|')[1] ...
          CAST(sort_array(map_values(timestamps))[0] AS BIGINT)
FROM timing_events;
```

APACHE HBASE

timing-events	t

RowKey	t:a3b3	t:24c4d
e4324|a31c4|a6545da436	1231412412	1231412509

```
SELECT
rank() OVER (ORDER BY t2.earliestTimestamp-t1.earliestTimestamp)
 AS position,
rd.racerid, rd.firstName, rd.LastName ...
```

APACHE HBASE

timing-events | **t**

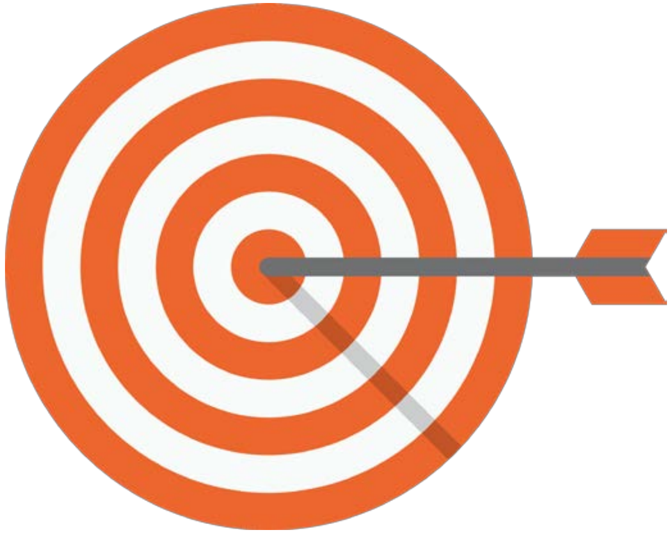
RowKey	t:a3b3	t:24c4d
e4324|a31c4|a6545da436	1231412412	1231412509

2.5M rows

606 seconds

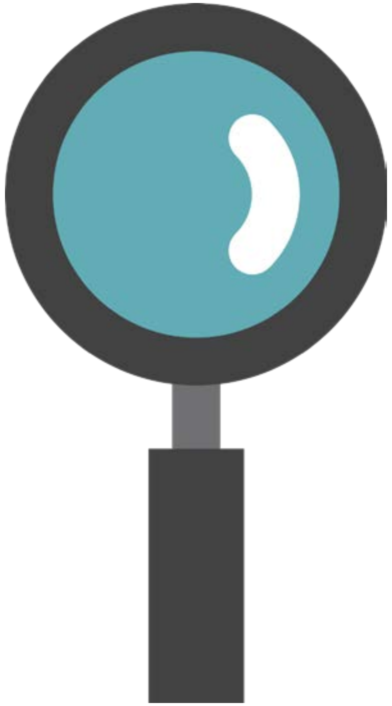
= **4.2K** rows/sec

Module Goals



- Generate Race Results ✓
- Map HBase Tables in Hive ✓
- Query Data with HiveQL ✓
- Map CSV Files in Hive ✓
- Write Output to Azure ✓

Coming Next



Hive Architecture

Custom UDFs

Additional Azure Storage

Hive & PowerShell