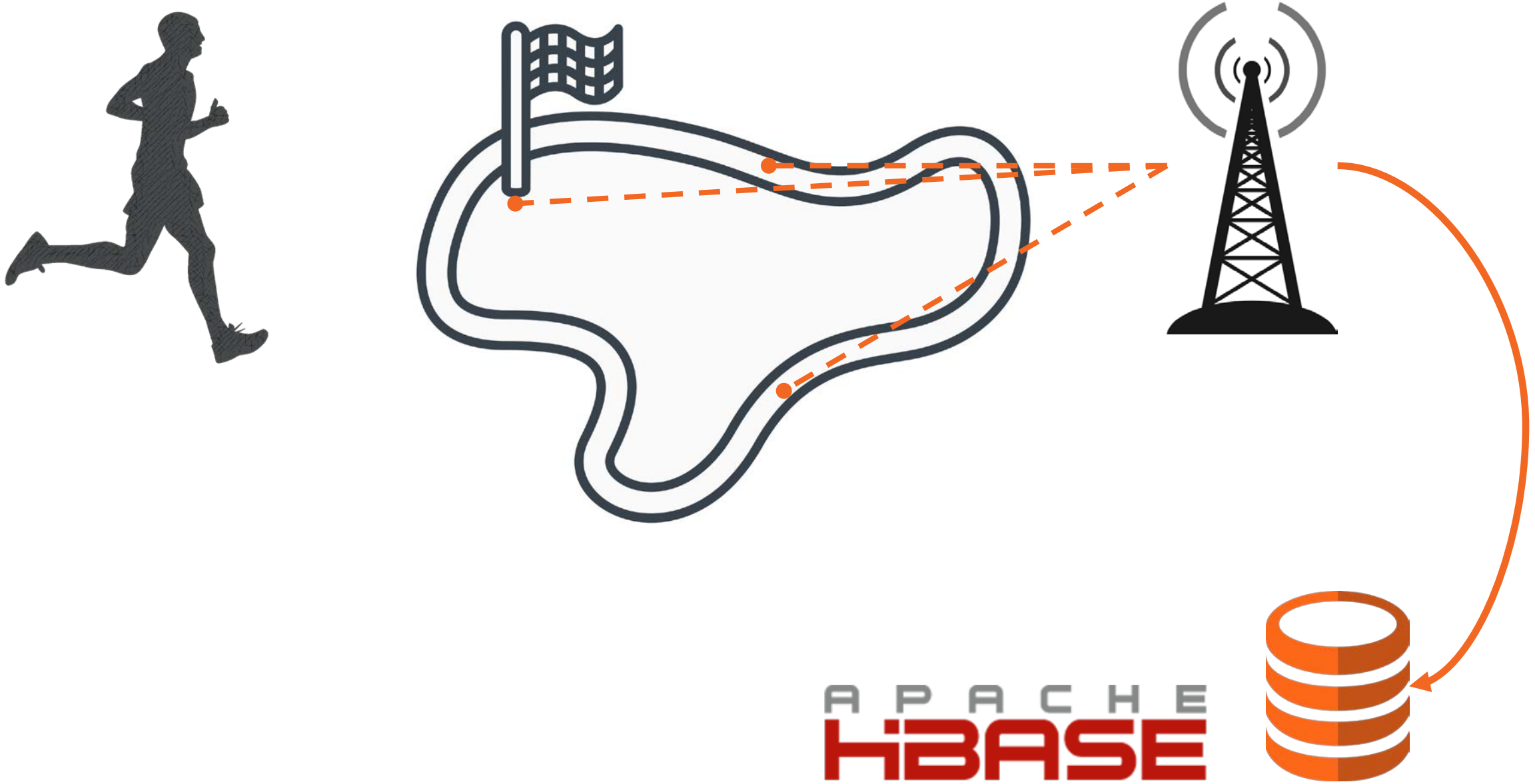
Storing Race Data in HBase

Elton Stoneman

@EltonStoneman | blog.sixeyed.com



APACHE HBASE

Races	t	p

RowKey	t:1	t:2	p:1	p:2
a6545da436	34b2f131	5a51322	bc3b2	cf2434

RowKey	t:1	t:2	t:3	p:1	p:2
b3242b44d	ee142dd	234fe1a	876abd7	234a2bf	457ba

RowKey	t:1	t:2
c1312d312e	24b4f131	66a1e21

Microsoft® SQL Server®

Races

RaceId	Date	Entries
123123	2015-09-04	100

RaceTimers

RaceId	TimerId
123123	245335

Timers

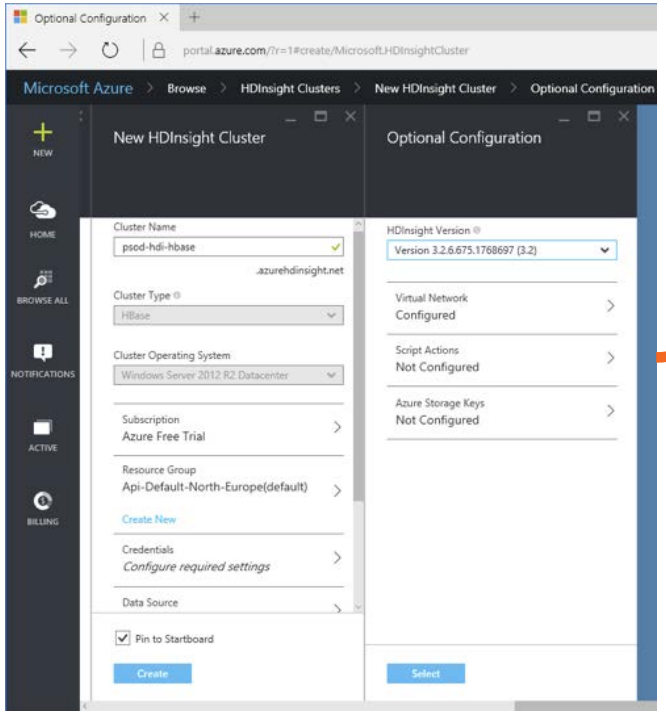
TimerId	AssetTag	PurchaseDate	LastServiceDate
245335	TM-T01-034	2014-12-02	2015-06-23

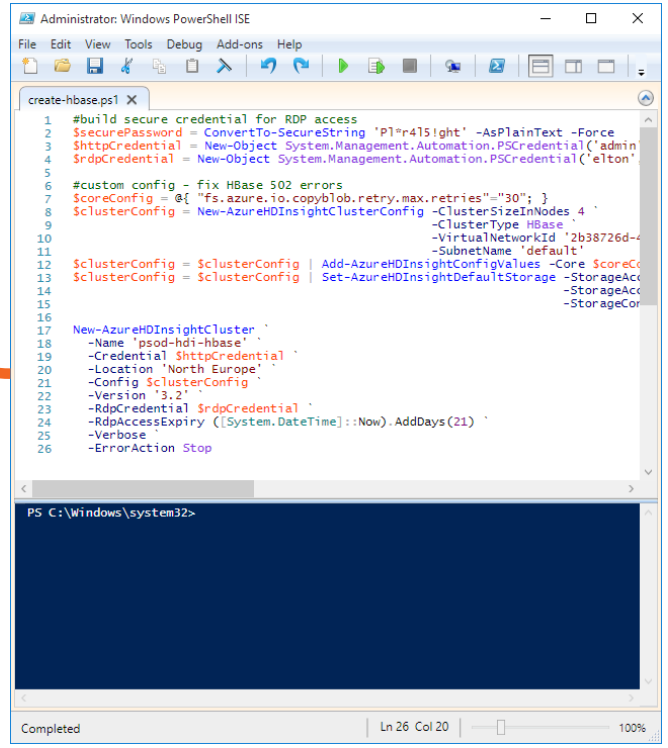
mongoDB

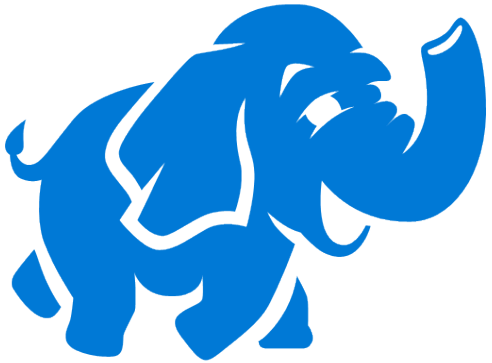
Race	
Id	bqwd65fb6
Date	2015-09-04
Entries	100
Timers	

Timer	
Id	2fba5fb6
AssetTag	TM-T01-034

Timer	
Id	124ab131c
AssetTag	TM-T02-006

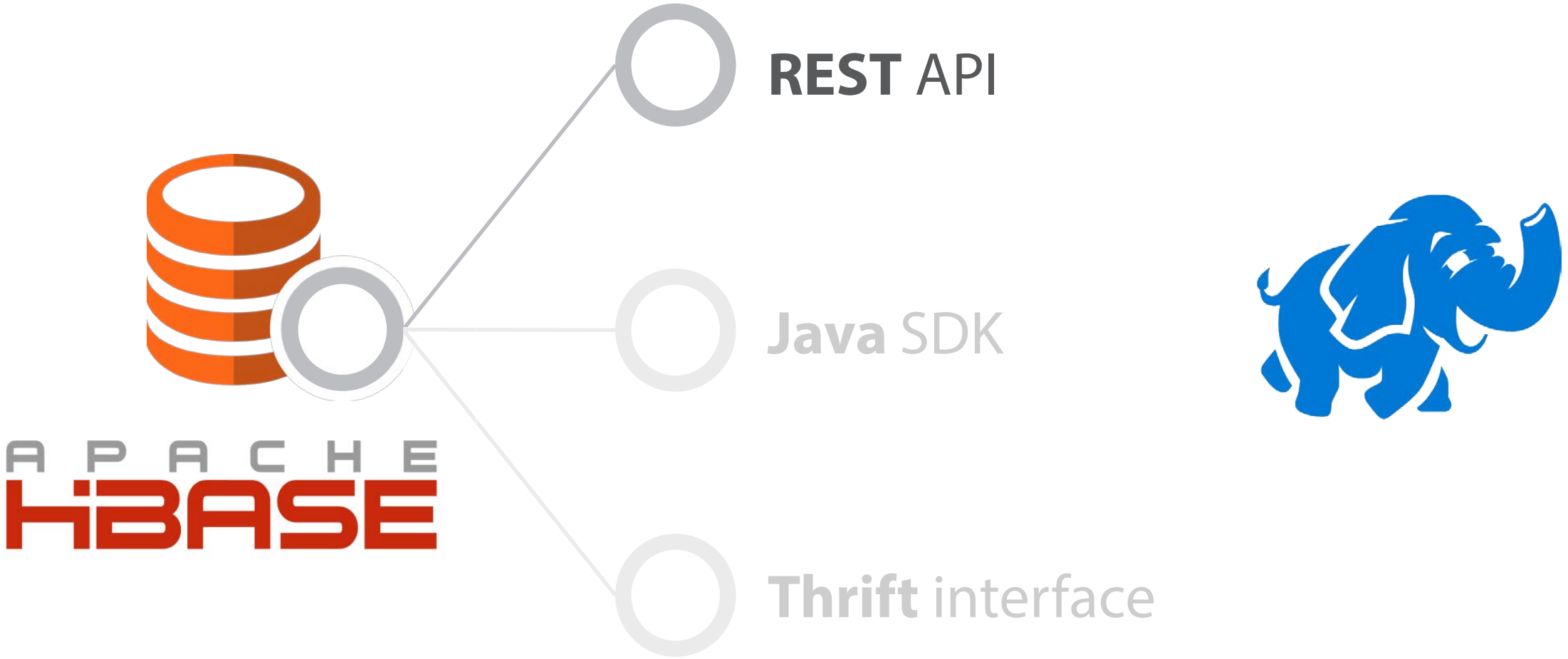


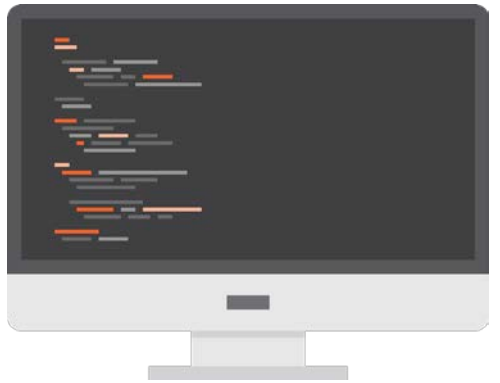


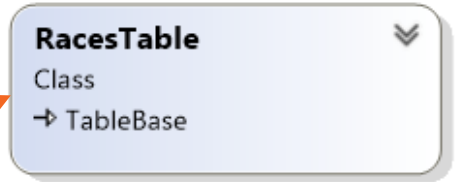


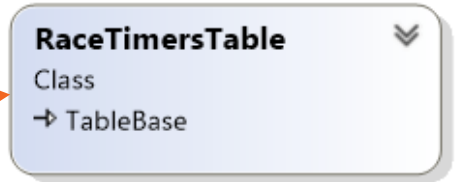
APACHE HBASE

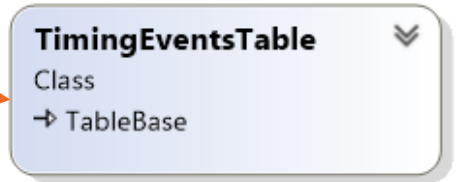
```
> list

> create 'races','t', 'p'

> describe 'races'

> put 'races', 'r1', 't:1', '14b3b4'

> count 'races'

> get 'races', 'r1'
```



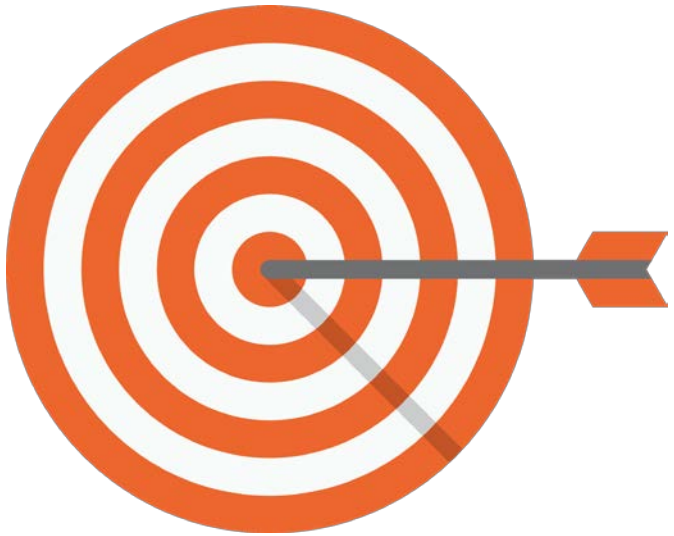








Module Goals



- Structuring data in HBase
- Data access patterns
- Accessing HBase from .NET
- Running HBase on HDInsight

APACHE HBASE

races | **d** | **t** | **p** | **e**

RowKey	d:utc	t:1	t:2	p:1	e:f
a6545da436	1231412412	5a51322	bc3b2	e4324	1

RowKey	d:utc	t:1
a6545da436	1231412412	5a51322

sector-times | **te** | **d**

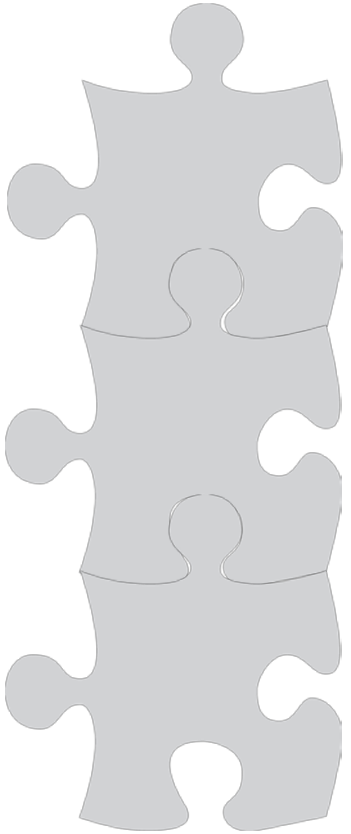
RowKey	t:a31c4	t:bc3b2	d:1	d:t
e4324|a6545da436	1231412412	1231483602	71190	71190

RowKey	t:a31c4
ba23d|a6545da436	1231460234

Table

Row

Column



races | **d** | **t** | **p** | **e**

RowKey	d:utc	t:1	t:2	p:1	e:f
a6545da436	1231412412	5a51322	bc3b2	e4324	1

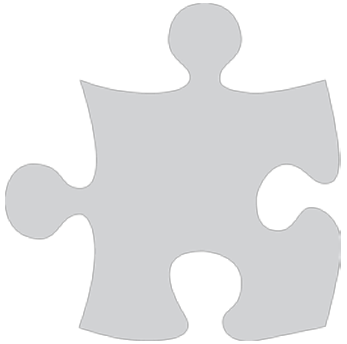
RowKey	d:utc	t:1
a6545da436	1231412412	5a51322

sector-times | **te** | **d**

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

RowKey	t:a31c4
ba23d|a6545da436	1231460234

Table



races	d	t	p	e

RowKey	d:utc	t:1	t:2	p:1	e:f
a6545da436	1231412412	5a51322	bc3b2	e4324	1

RowKey	d:utc	t:1
a6545da436	1231412412	5a51322

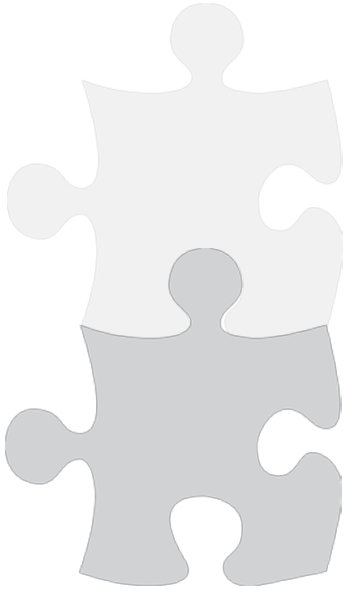
sector-times	te	d

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

RowKey	t:a31c4
ba23d|a6545da436	1231460234

Table

Row



races	d	t	p	e

RowKey	d:utc	t:1	t:2	p:1	e:f
a6545da436	1231412412	5a51322	bc3b2	e4324	1

RowKey	d:utc	t:1
a6545da436	1231412412	5a51322

sector-times	te	d

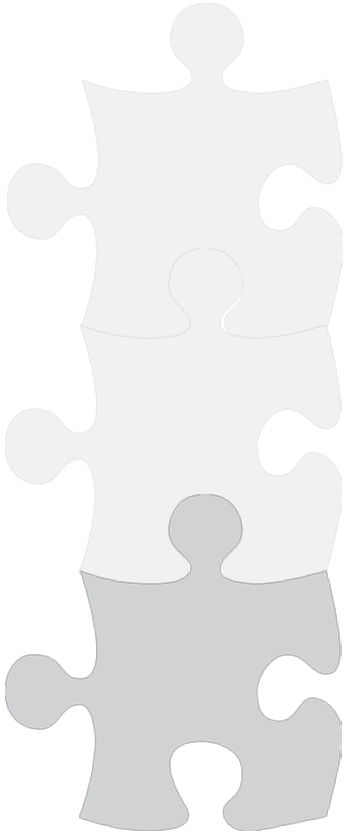
RowKey	t:a31c4	t:bc3b2	d:1	d:t
e4324|a6545da436	1231412412	1231483602	71190	71190

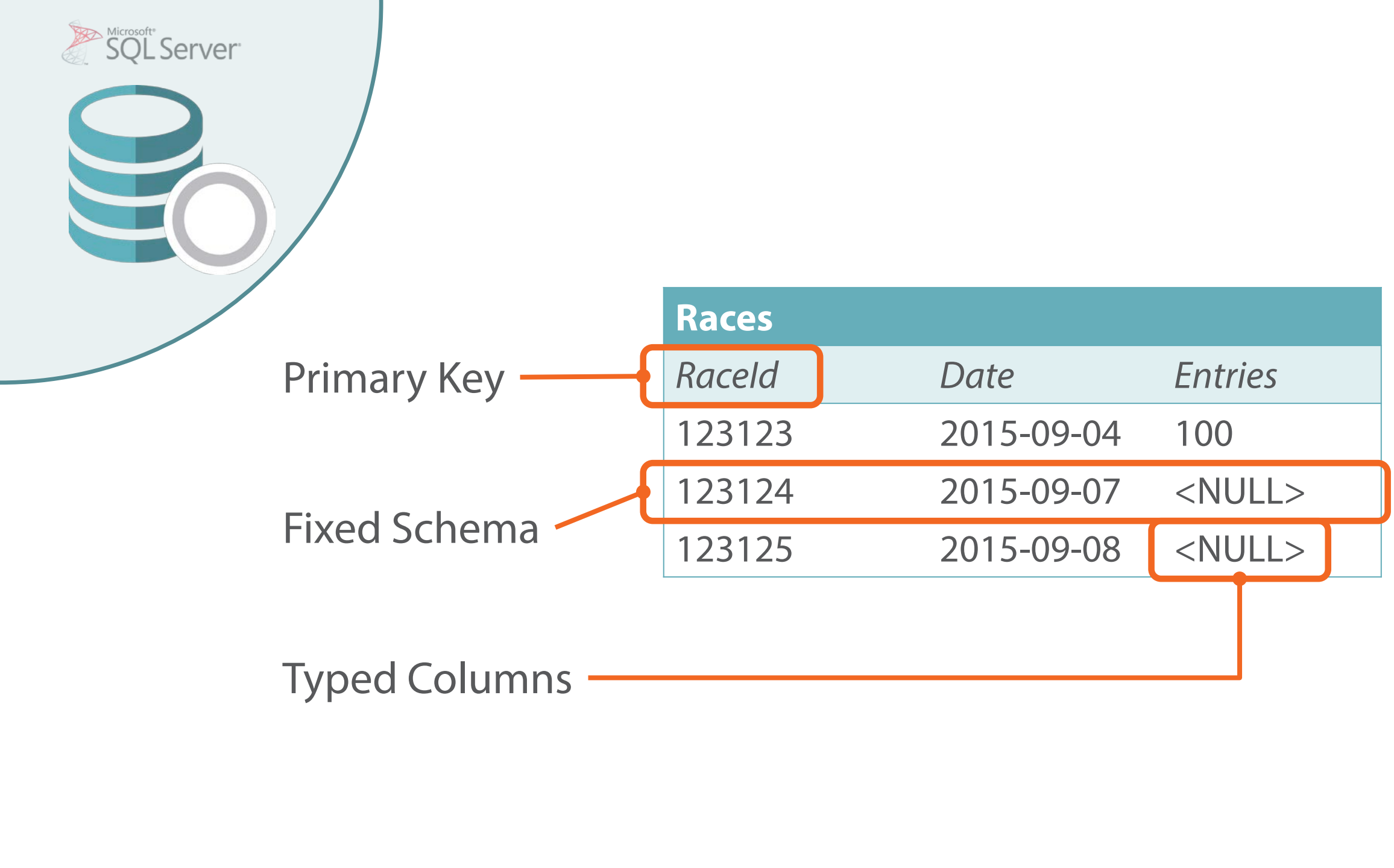
RowKey	t:a31c4
ba23d|a6545da436	1231460234

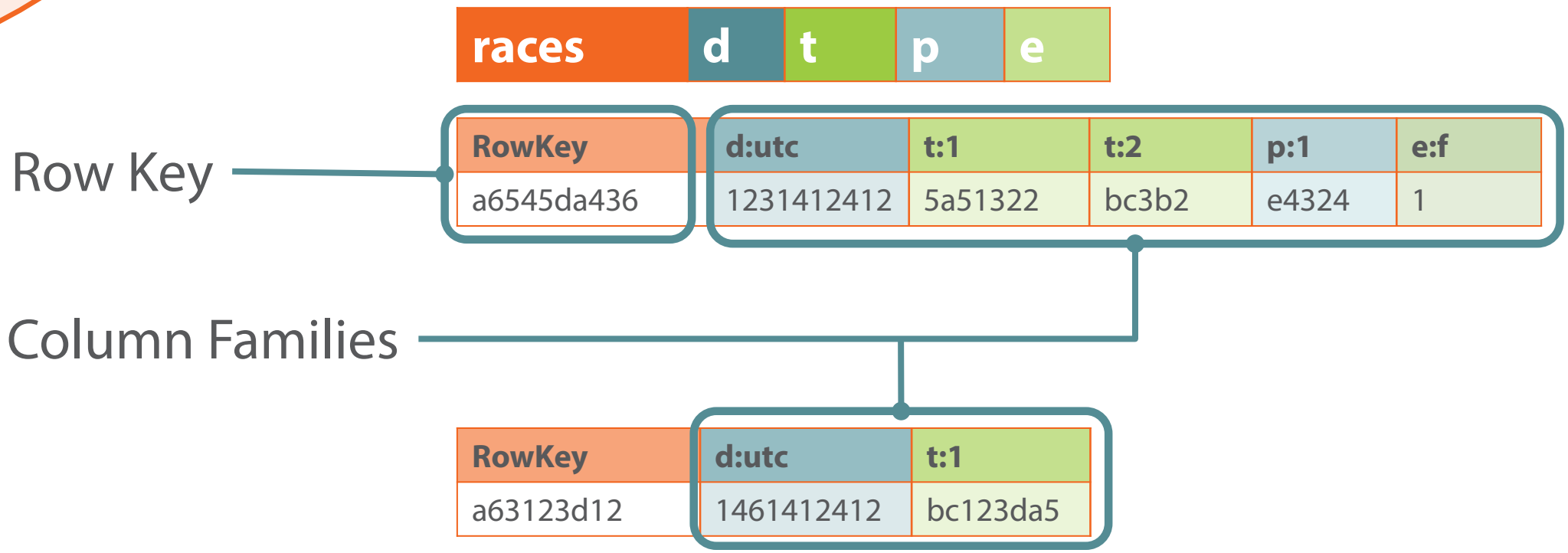
Table

Row

Column







APACHE HBASE

races	d	t	p	e

RowKey	d:utc	t:1	t:2	p:1	e:f
a6545da436	1231412412	5a51322	bc3b2	e4324	1

RowKey	d:utc	t:1
a623123d12	1461412412	bc123da5

```
get 'races', 'a623123d12'
```

```
scan 'races', { STARTROW=>'a6'}
```

APACHE HBASE

races	d	t	p	e

RowKey	d:utc	t:1	t:2	p:1	e:f
a6545da436	1231412412	5a51322	bc3b2	e4324	1

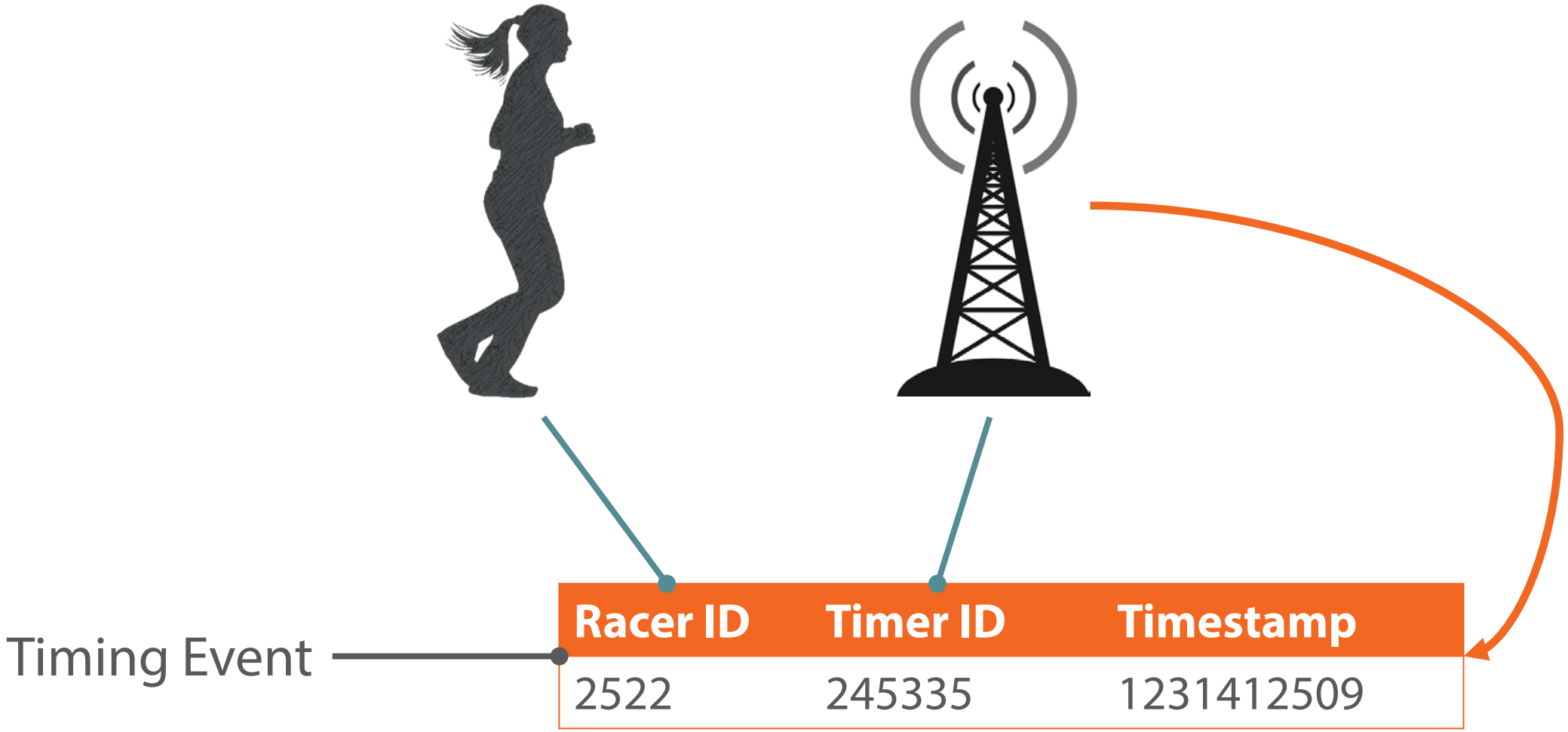
RowKey	d:utc	t:1
a623123d12	1461412412	bc123da5

Primary access by row key

```
get 'races', 'a623123d12'
```

```
scan 'races', { STARTROW=>'a6'}
```

```
scan 'races', { 'e:f'=>'1'}
```



Races

RaceId	Date	Entries
123123	2015-09-04	100

RaceTimers

RaceId	TimerId
123123	245335

Timers

TimerId	AssetTag	PurchaseDate	LastServiceDate
245335	TM-T01-034	2014-12-02	2015-06-23

```
SELECT r.RaceId
FROM Races r
JOIN RaceTimers rt
ON rt.RaceId = r.RaceId
WHERE rt.TimerId = {TIMER-ID}
AND r.Date = {EVENT-DATE}
```

Races

RaceId	Date	Entries
123123	2015-09-04	100

RaceTimers

RaceId	TimerId
123123	245335

Races

RaceId	Date	Entries
123123	2015-09-04	100

RaceTimers

RaceId	TimerId
123123	245335

TimingEvents

RaceId	TimerId	RacerId	Timestamp
123123	245335	2522	1231412509

Timers

TimerId	AssetTag	PurchaseDate	LastServiceDate
245335	TM-T01-034	2014-12-02	2015-06-23

Races

RaceId	Date	Entries
123123	2015-09-04	100

RaceTimers

RaceId	TimerId
123123	245335

TimingEvents

RaceId	TimerId	RacerId	Timestamp
123123	245335	2522	1231412509

Timers

TimerId	AssetTag	PurchaseDate	LastServiceDate
245335	TM-T01-034	2014-12-02	2015-06-23

Races

RaceId	Date	Entries
123123	2015-09-04	100

RaceTimers

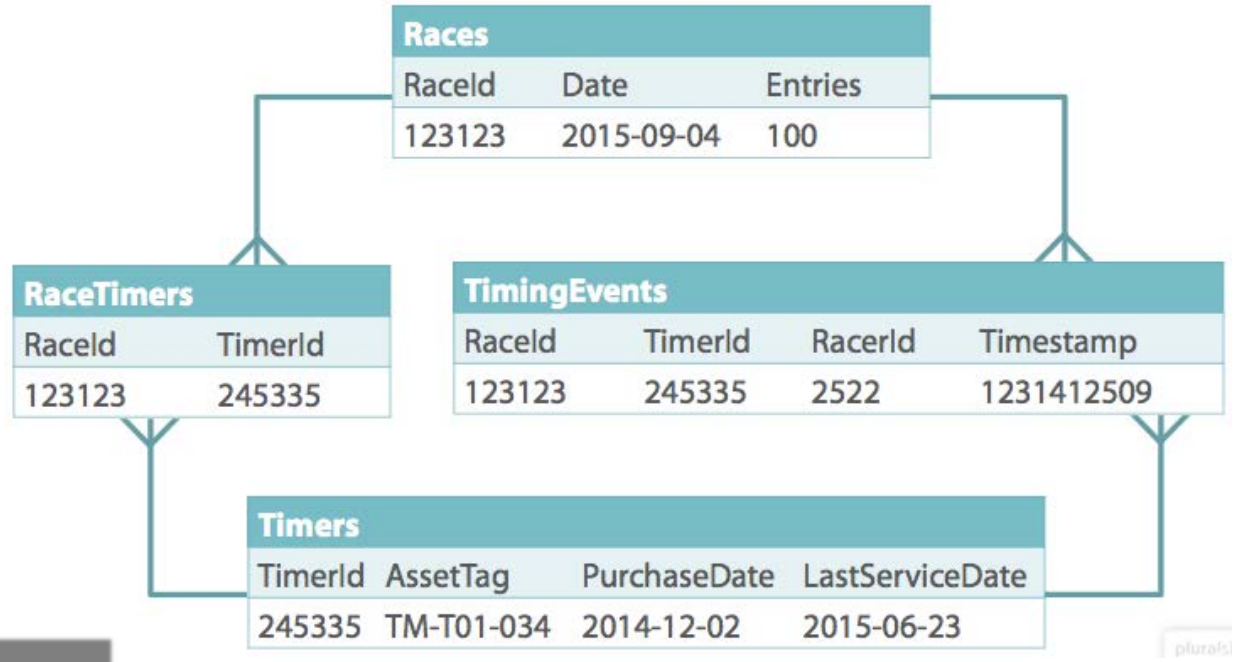
RaceId	TimerId
123123	245335

TimingEvents

RaceId	TimerId	RacerId	Timestamp
123123	245335	2522	1231412509

Timers

TimerId	AssetTag	PurchaseDate	LastServiceDate
245335	TM-T01-034	2014-12-02	2015-06-23



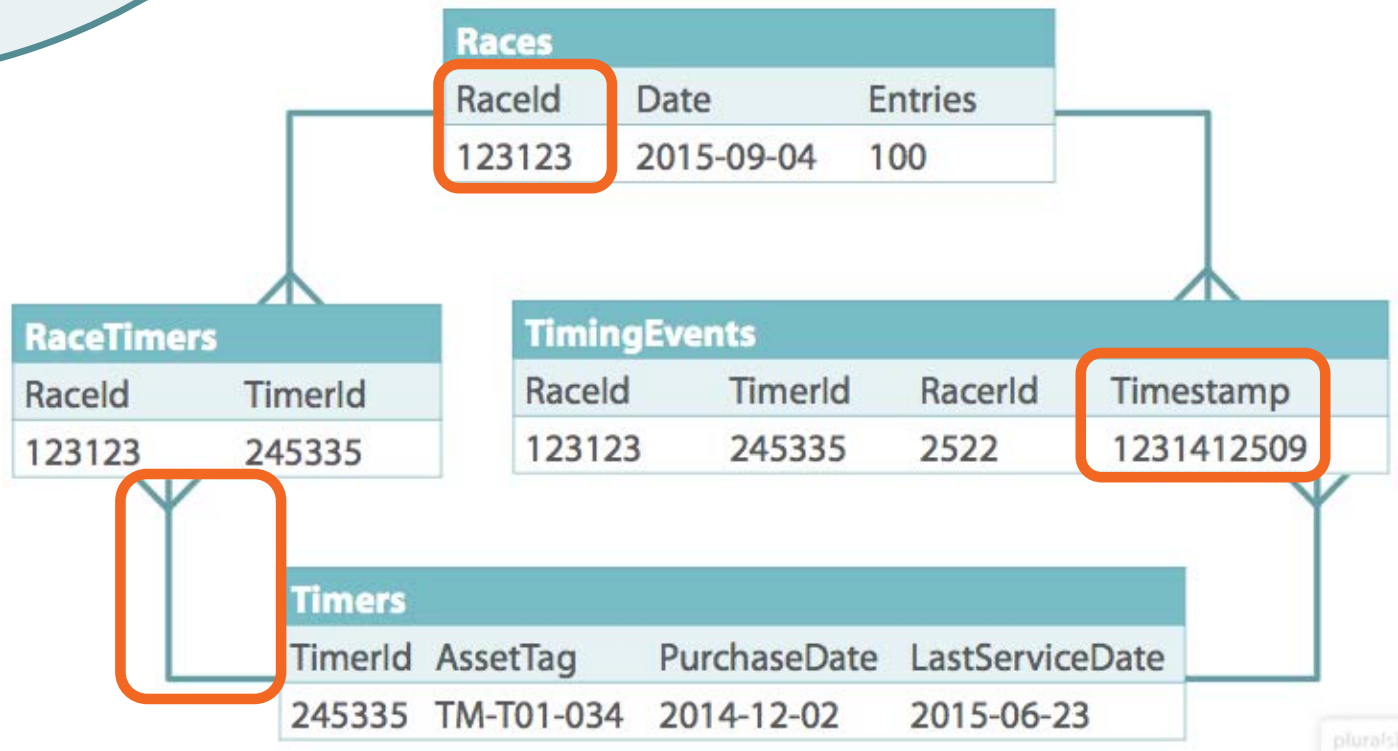

```
SELECT
  te.RacerId,
  (max(te.Timestamp) -
   min(te.Timestamp)) as Duration
FROM TimingEvents te
GROUP BY te.RacerId
WHERE te.RaceId = {RACE-ID}
ORDER BY Duration
```

Microsoft SQL Server

DB-generated IDs

Explicit relationships

High-level **querying**



races

RowKey
a6545da436

RowKey
b3242b44d

Client-generated IDs

Not sequential

GUID-like **string**

races	t

RowKey	t:1	t:2
a6545da436	5a51322	bc3b2

RowKey	t:1	..	t:12
b3242b44d	5a51322		b12d125

t = timers **column family**

Name = timer index

Value = timer ID

races	d	t	p	e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

RowKey	t:1	d:utc
b3242b44d	5a51322	1231412412

d = dates **column family**

d:utc = race date

e = entrants **column family**

e:f = finishers

Timing Event

Racer ID	Timer ID	Timestamp
e4324	bc3b2	1231412509

Race ID ?

races	d	t	p	e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

RowKey	t:1	d:utc
b3242b44d	5a51322	1231412412

```
scan 't1', {COLUMNS => ['t'] }
```

race-timers

RowKey
5a51322|20150907

Rowkey=

{timerId}|{raceDate}

race-timers	r

RowKey	r:a6545da436
5a51322|20150907	1

RowKey	r:a6545da436
bc3b2|20150907	2

r = race **column family**

Name = race ID

Value = timer index

races	d	t	p	e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

race-timers	r

RowKey	r:a6545da436
5a51322|20150907	1

RowKey	r:a6545da436
bc3b2|20150907	2

timing-events

RowKey
a6545da436|e4324|a31c4

Rowkey=

{raceId}|{timerId}|{racerID}

timing-events

RowKey
a6545da436|e4324|a31c4

RowKey
a6545da436|e4324|cf2434

RowKey
a6545da436|bc3b2|a31c4

Rowkey=

{raceId}|{timerId}|{racerID}

```
scan 'timing-events', { STARTROW=>'a6545da436|'}
```

timing-events

RowKey
a6545da436|e4324|a31c4

RowKey
a6545da436|e4324|cf2434

RowKey
a6545da436|bc3b2|a31c4

Rowkey=

{raceId}|{timerId}|{racerID}

```
scan 'timing-events', { STARTROW=>'a6545da436|e4324|'}
```

timing-events

RowKey
a6545da436|e4324|a31c4

RowKey
a6545da436|e4324|cf2434

RowKey
a6545da436|bc3b2|a31c4

Rowkey=

{raceId}|{timerId}|{racerID}

```
get 'timing-events', 'a6545da436|e4324|a31c4'
```

timing-events

RowKey
a6545da436|e4324|a31c4

RowKey
a6545da436|e4324|cf2434

Rowkey as **byte[]**

```
97 54 53 52 53 100 97 52 51 54 124 101 …

97 54 53 52 53 100 97 52 51 54 124 101 …
```

First 11 bytes match

APACHE HBASE

timing-events

RowKey
a6545da436|e4324|a31c4

RowKey
a6545da436|e4324|cf2434

Rowkey=

{raceId}|{timerId}|{racerID}

timing-events

RowKey
e4324|a31c4|a6545da436

RowKey
cf2434|a31c4|a6545da436

Rowkey=

{**racerId**}|{**timerId**}|{**raceID**}

timing-events

RowKey
e4324|a31c4|a6545da436

RowKey
cf2434|a31c4|a6545da436

Rowkey as **byte[]**

```
101 52 51 50 52 124 97 51 49 99 52 124 …
```

```
99 102 50 52 51 52 124 97 51 49 99 52 …
```

No bytes match

timing-events

RowKey
e4324|a31c4|a6545da436

RowKey
cf2434|a31c4|a6545da436

RowKey
e4324|bc3b2|a6545da436

Rowkey=

{**racerId**}|{**timerId**}|{**raceID**}

```
scan 'timing-events', { STARTROW=>'a6545da436|'}
```

timing-events

RowKey
e4324|a31c4|a6545da436

RowKey
cf2434|a31c4|a6545da436

RowKey
e4324|bc3b2|a6545da436

Rowkey=

{**racerId**}|{**timerId**}|{**raceID**}

```
scan 'timing-events', { STARTROW=>'a6545da436|a31c4'}
```

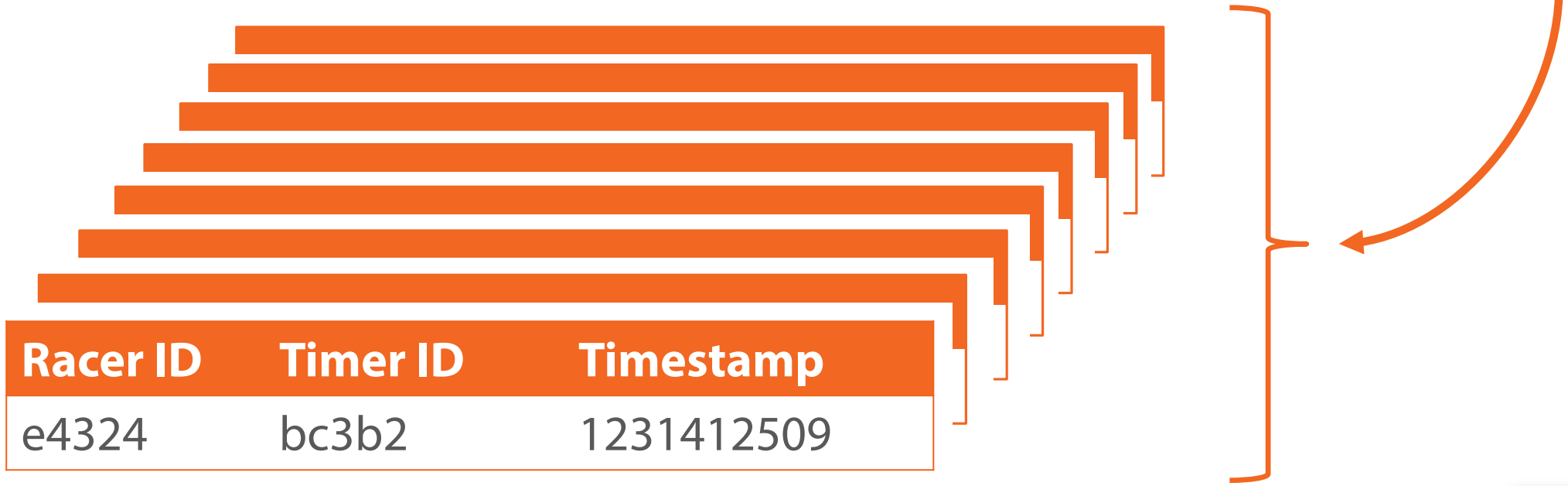
timing-events	t

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

t = timestamp **column family**

Name = random string

Value = timestamp



Racer ID	Timer ID	Timestamp
e4324	bc3b2	1231412509

APACHE HBASE

timing-events	t

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

APACHE HBASE

timing-events	t

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

APACHE HBASE

timing-events	t

RowKey	t:a3b3	t:24c4d	t:3b54d
e4324|a31c4|a6545da436	1231412412	1231412509	1231412510

timing-events	t

RowKey	t:a3b3	t:24c4d	t:3b54d	…
e4324|a31c4|a6545da436	1231412412	1231412509	1231412510	

timing-events

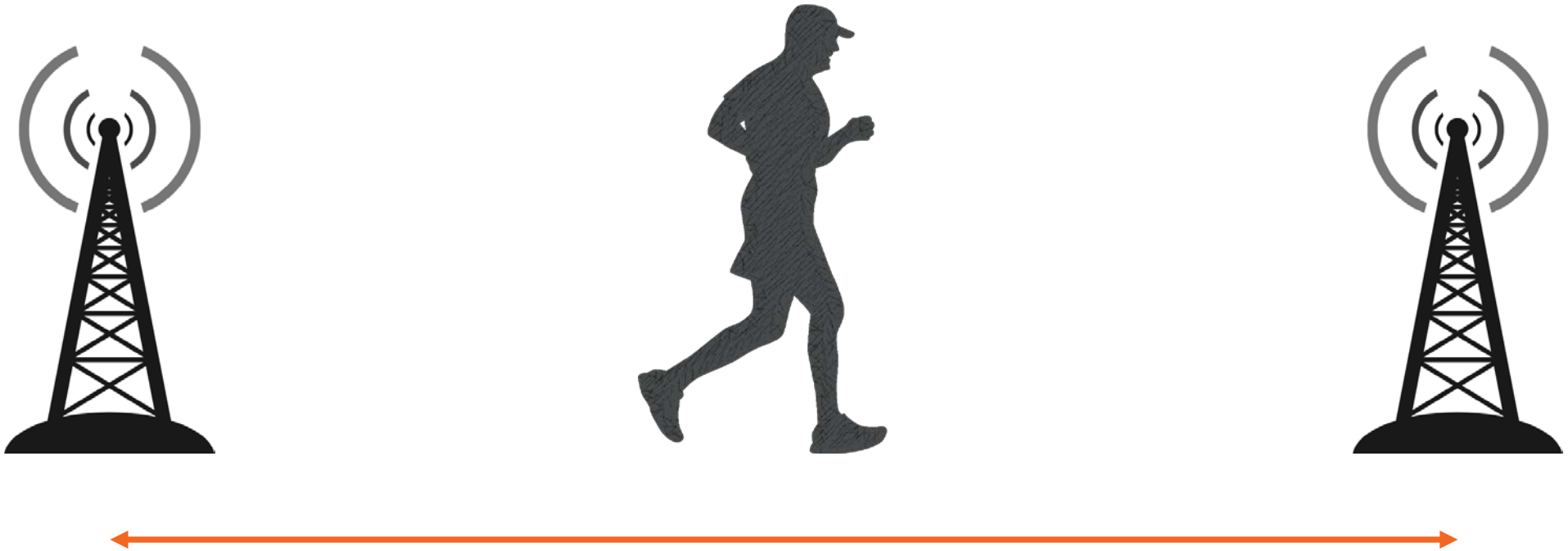
RowKey	t:a3b3	t:24c4d
e4324|a31c4|a6545da436	1231412412	1231412509

Duration ?

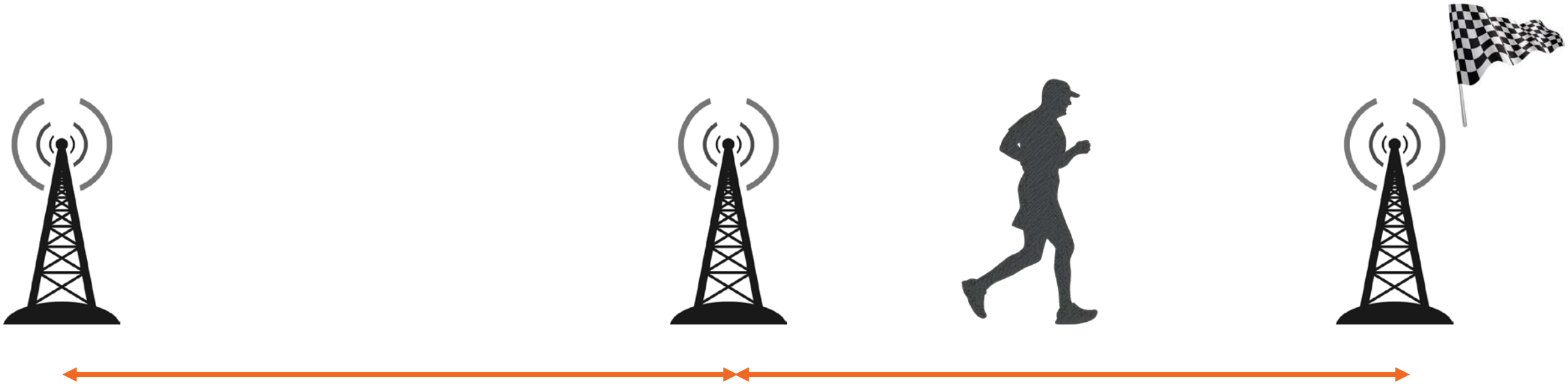
Order ?

RowKey	t:a3b3
cf2434|a31c4|a6545da436	1231452512

RowKey	t:25cda	t:d13ca4	t:32a1e
e4324|bc3b2|a6545da436	1231412412	1231412506	1231412642



Racer ID	Sector	Duration (ms)
e4324	1	367,412



Racer ID	Sector	Duration (ms)
e4324	1	367,412
e4324	2	386,818
	Total	754,230

Racer ID	Total Duration (ms)	Position
e4324	754,230	1
c31dc	756,606	2

sector-times

RowKey
e4324|a6545da436

Rowkey= {**racerId**}|{**raceID**}

RowKey
e4324|a6545da436

te = timing events

d = sector durations

sector-times	te	d

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

te = **column family**

Name = timer ID

Value = first timestamp

APACHE HBASE

sector-times	te	d

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

d = **column family**

Name = sector number

Value = duration (ms)

APACHE HBASE

sector-times	te	d

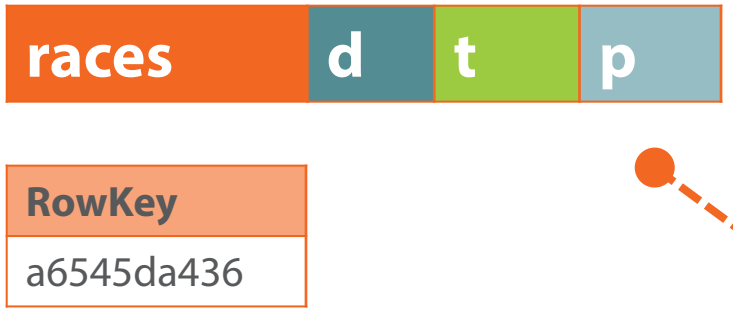
RowKey	t:a31c4	t:bc3b2	d:1	d:t
e4324|a6545da436	1231412412	1231483602	71190	71190

sector-times | **te** | **d**

RowKey	t:a31c4	t:bc3b2	d:1	d:t
e4324|a6545da436	123 412412	1231483602	71190	71190

timing-events | **t**

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



RowKey
a6545da436

d = race dates

t = timers

p = result positions

APACHE HBASE

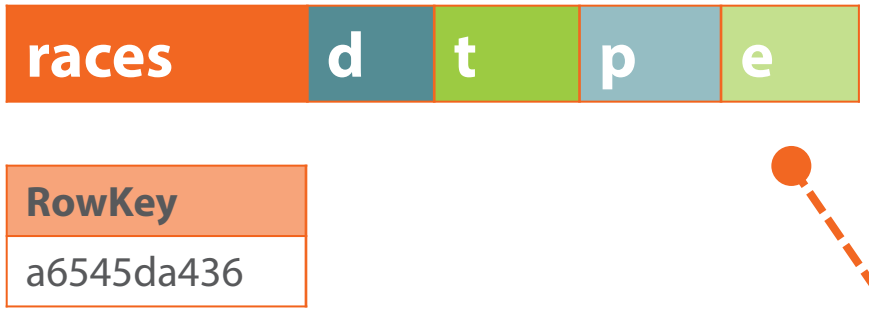
races	d	t	p

RowKey	t:1	t:2	d:utc	p:1
a6545da436	5a51322	bc3b2	1231412412	e4324

p = **column family**

Name = position number

Value = racer ID



d = race dates

t = timers

p = result positions

e = entries

races	d	t	p	e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

e:f = finished entrants

Value = number of finishers

races	d	t	p	e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

```
put 'races', 'p:1', 'e4324', 'e:f', '1'
```

races	d	t	p	e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

race-timers	r

RowKey	r:a6545da436
a31c4|20150907	1

timing-events	t

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

sector-times	te	d

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

races | d | t | p | e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

race-timers | r

RowKey	r:a6545da436
a31c4|20150907	1

timing-events | t

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

sector-times | te | d

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

APACHE HBASE

races	d	t	p	e

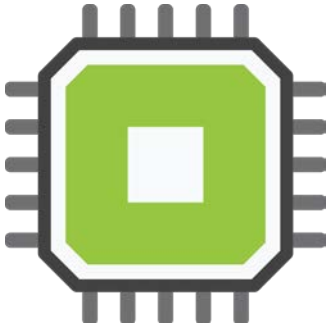
RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

race-timers	r

RowKey	r:a6545da436
a31c4|20150907	1

timing-events	t

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



sector-times	te	d

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

APACHE HBASE

races	d	t	p	e

RowKey	t:1	t:2	d:utc	p:1	e:f
a6545da436	5a51322	bc3b2	1231412412	e4324	1

race-timers	r

RowKey	r:a6545da436
a31c4|20150907	1

timing-events	t

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

sector-times	te	d

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

APACHE HBASE

v4

races	d	t	p	e
raceID	dates	timers	positions	entries

race-timers	r
raceID | date	race

timing-events	t
racerID | timerID | raceID	timestamps

sector-times	te	d
racerID | raceID	timing events	durations

v3

races	d	t	p	e
raceID	dates	timers	positions	entries

race-timing-points	r
raceID | date	race

timing-events	t
raceID | timerID | racerID	timestamps

sector-times	te	d
racerID | raceID	timing events	durations

v2

races	d	t	p	e
raceID	dates	timers	positions	entries

race-timing-points	r
raceID | date	race

timing-events	t
raceID | timerID | racerID	timestamps

sector-times	t	tp
raceID | racerID | sectorNumber	timing events	timing points

v1

races	d	t
raceID	dates	timers

race-timing-points	r
raceID | date	race

timing-events	t
raceID | timerID | racerID	timestamps

sector-times	t	tp
raceID | racerID | sectorNumber	timing events	timing points

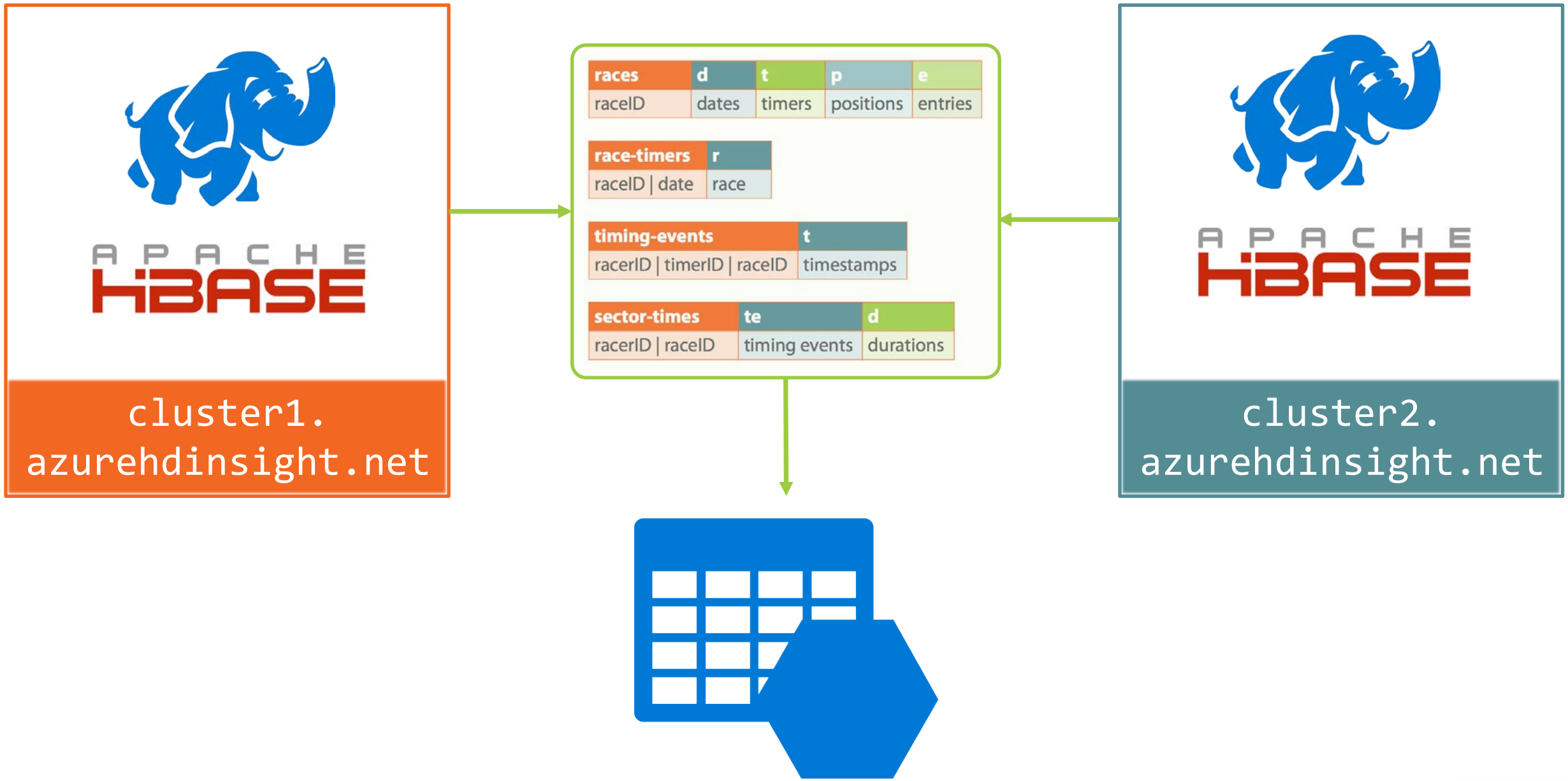
race-results	r
raceID | position	racer

Demo: HBase on HDInsight

Azure Portal

PowerShell

Custom Configuration



Demo: HBase Shell

Cluster Administration

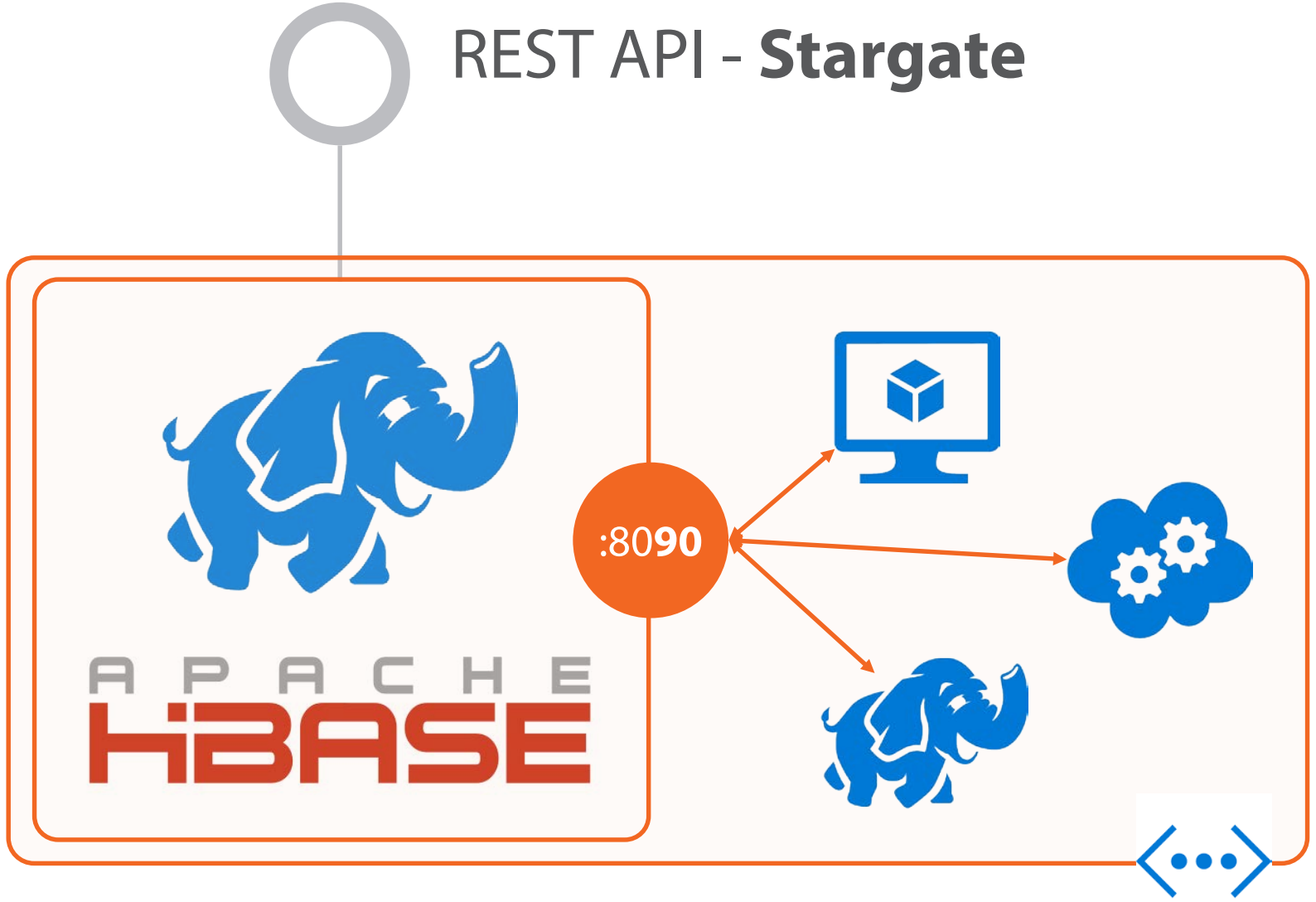
Data Definition

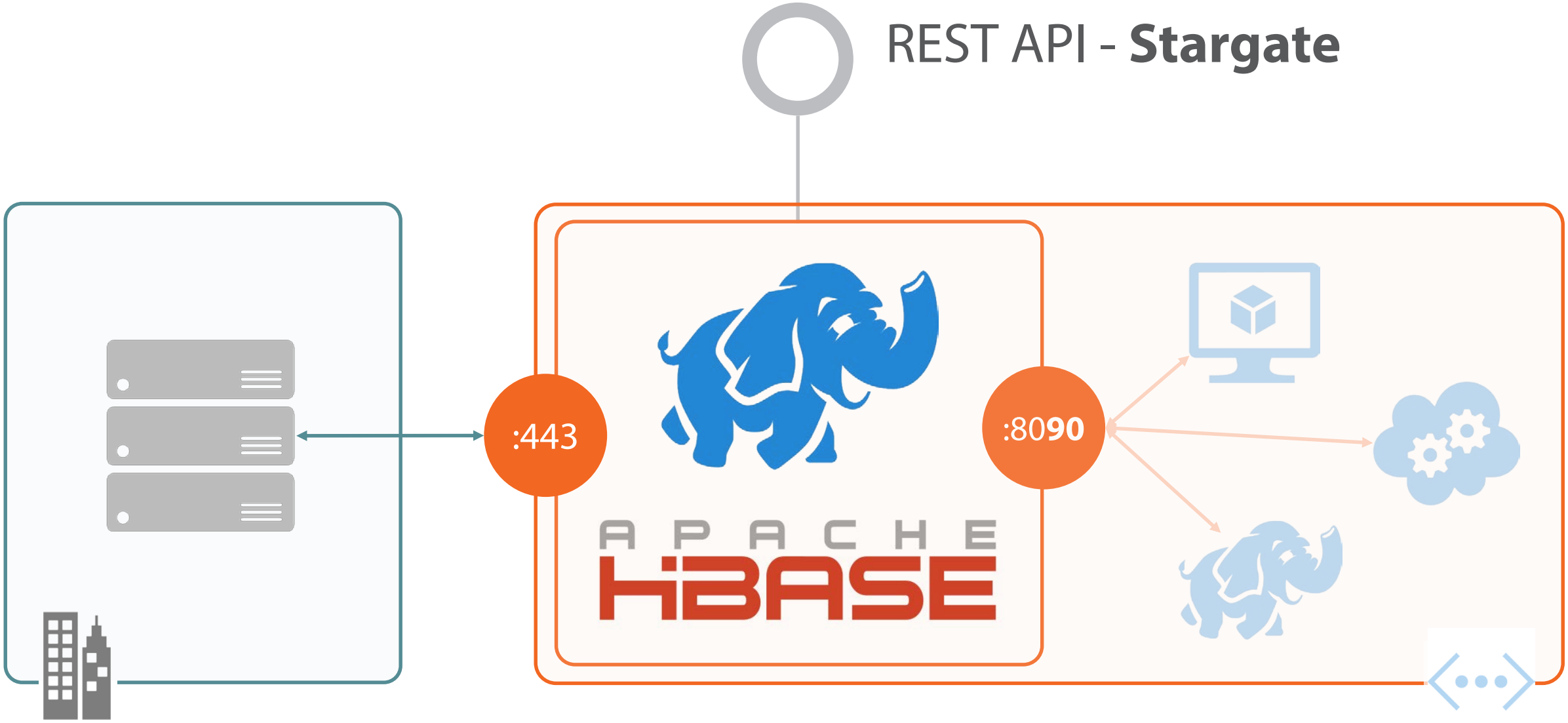
Data Modification

REST API - **Stargate**

REST API - **Stargate**

REST API - **Stargate**

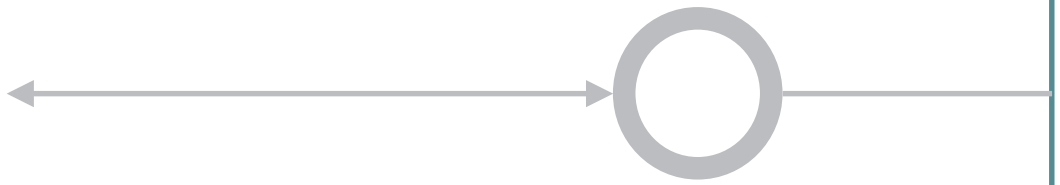


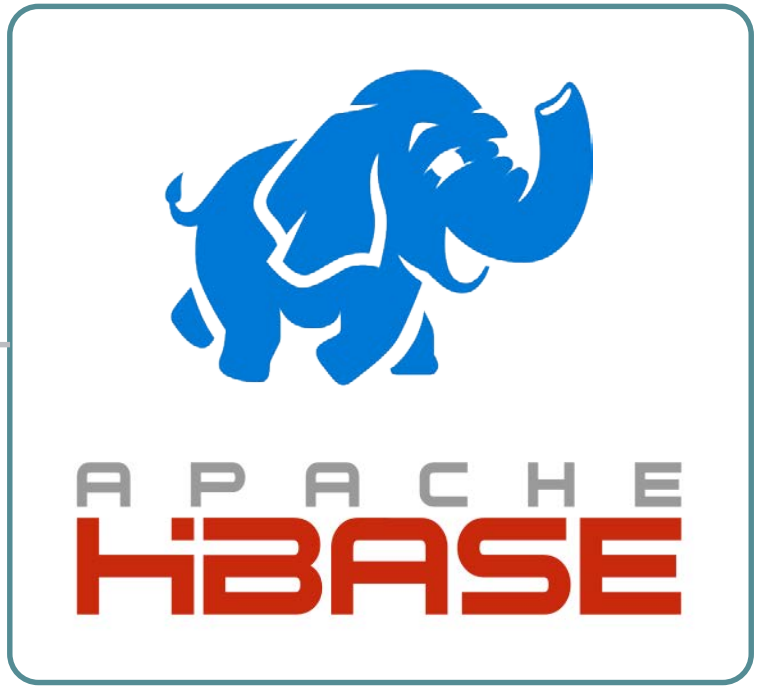

REST API - Stargate
:443
APACHE
HBASE
:8090

Demo: Stargate

Explore REST API

Azure Virtual Network





Microsoft HBase REST Client Library for .NET By: maxluk

Last Published: 2015-03-13 | Latest Version: 0.2.2

.NET Client Library for Azure HDInsight HBase clusters.

1,598 total downloads | Tags Microsoft HDInsight HBase

C# HBase binding By: vchekan

Last Published: 2011-08-29 | Latest Version: 1.0.0

C# binding for HBase. This is a "dumb" client which uses Thrift interface and is not aware of HBase regoins.

1,570 total downloads | Tags hbase

HBase.Stargate.Client.Autofac By: jbatte47

Last Published: 2014-06-24 | Latest Version: 1.1.1

Autofac support for HBase.Stargate.Client

678 total downloads | Licenses BSD-2-Clause

HBase.Stargate.Client By: jbatte47

Last Published: 2014-06-24 | Latest Version: 1.1.1

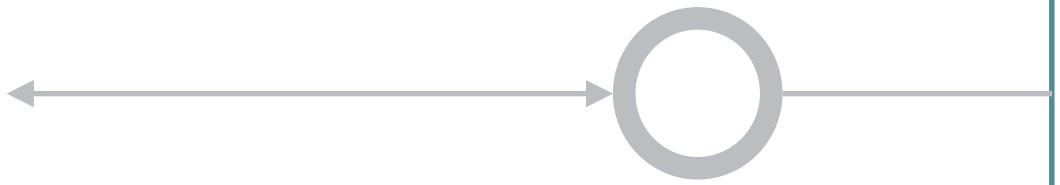
A simple .NET client library for HBase Stargate

991 total downloads | Licenses BSD-2-Clause

Real World Big Data in Azure

Elton Stoneman

@EltonStoneman | blog.sixeyed.com



Microsoft HBase REST Client Library for .NET By: maxluk

Last Published: 2015-03-13 | Latest Version: 0.2.2

.NET Client Library for Azure HDInsight HBase clusters.

1,598 total downloads | Tags Microsoft HDInsight HBase

C# HBase binding By: vchekan

Last Published: 2011-08-29 | Latest Version: 1.0.0

C# binding for HBase. This is a "dumb" client which uses Thrift interface and is not aware of HBase regoins.

1,570 total downloads | Tags hbase

HBase.Stargate.Client.Autofac By: jbatte47

Last Published: 2014-06-24 | Latest Version: 1.1.1

Autofac support for HBase.Stargate.Client

678 total downloads | Licenses BSD-2-Clause

HBase.Stargate.Client By: jbatte47

Last Published: 2014-06-24 | Latest Version: 1.1.1

A simple .NET client library for HBase Stargate

991 total downloads | Licenses BSD-2-Clause

Demo: HBase with .NET

HBase.Stargate.Client package

LINQPad

Demo: Storing Timing Events

HBase Data Access Layer

TimingEventsTable class

Append-only approach

timing-events	t

RowKey	t:1
e4324|a31c4|a6545da436	1231412412

APACHE HBASE

timing-events	t

RowKey	t:1	t:2
e4324|a31c4|a6545da436	1231412412	1231412509

APACHE HBASE

timing-events	t

RowKey	t:1	t:2	t:3
e4324|a31c4|a6545da436	1231412412	1231412509	1231412510

timing-events | **t**

RowKey	t:1	t:2	t:3
e4324|a31c4|a6545da436	1231412412	1231412509	1231412510

```
get 'timing-events', 'e4324|a31c4|a6545da436'
```

```
put 'timing-events', 'e4324|a31c4|a6545da436',
't:4', '12314126194'
```

APACHE HBASE

timing-events	t

RowKey	t:a3b3	t:24c4d	t:3b54d
e4324|a31c4|a6545da436	1231412412	1231412509	1231412510

timing-events | **t**

RowKey	t:a3b3	t:24c4d	t:3b54d
e4324|a31c4|a6545da436	1231412412	1231412509	1231412510

```
get 'timing-events', 'e4324|a31c4|a6545da436'
```

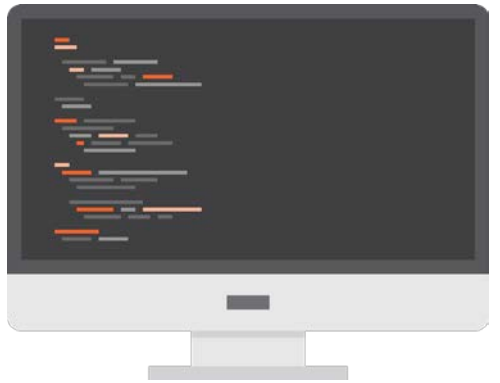
```
put 'timing-events', 'e4324|a31c4|a6545da436',
't:bb245a', '12314126194'
```

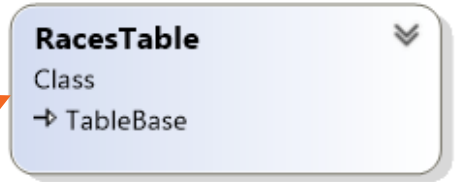
Demo: Storing Sector Times

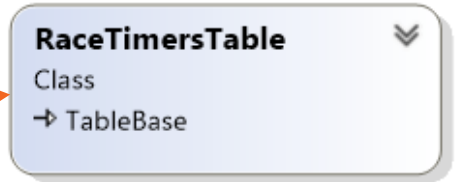
Refactored TableBase

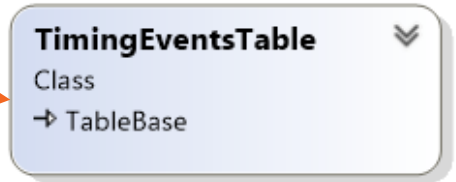
SectorTimesTable class

Reading & writing

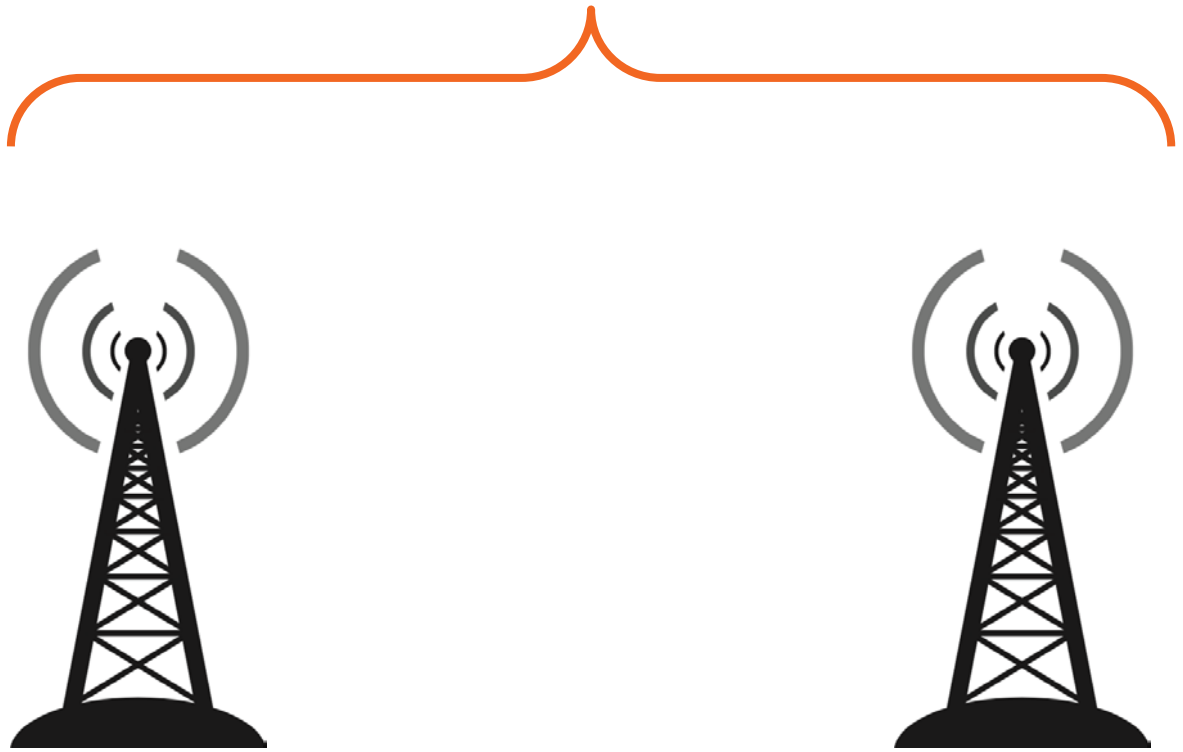




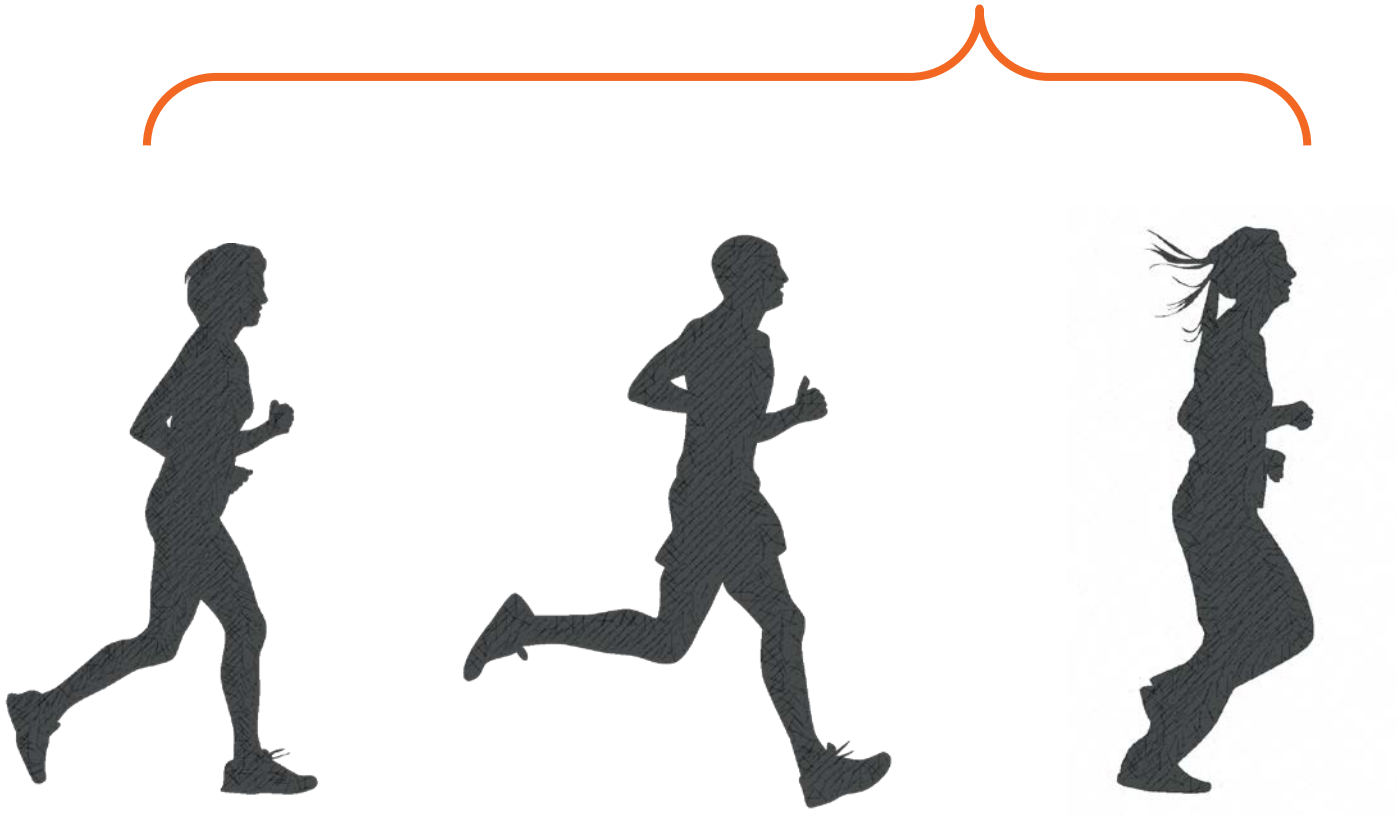




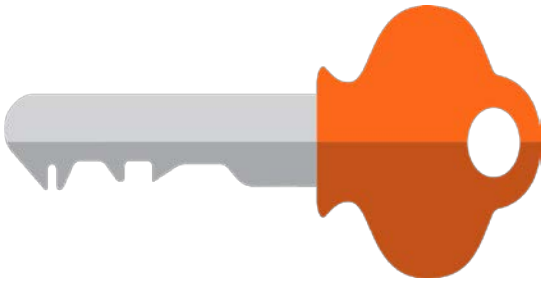
races	d	t	p	e
raceID	dates	timers	positions	entries

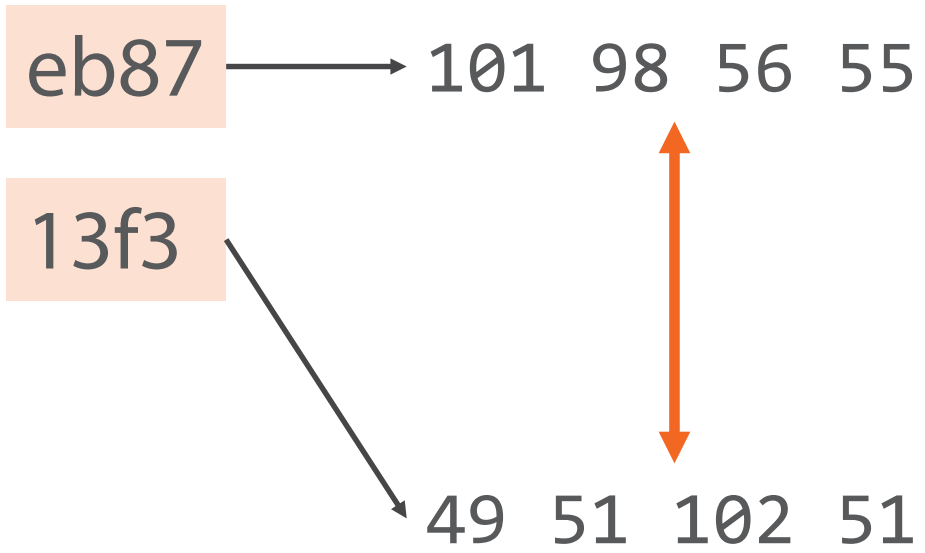


races	d	t	p	e
raceID	dates	timers	positions	entries



races	d	t	p	e
raceID	dates	timers	positions	entries





races	d	t	p	e

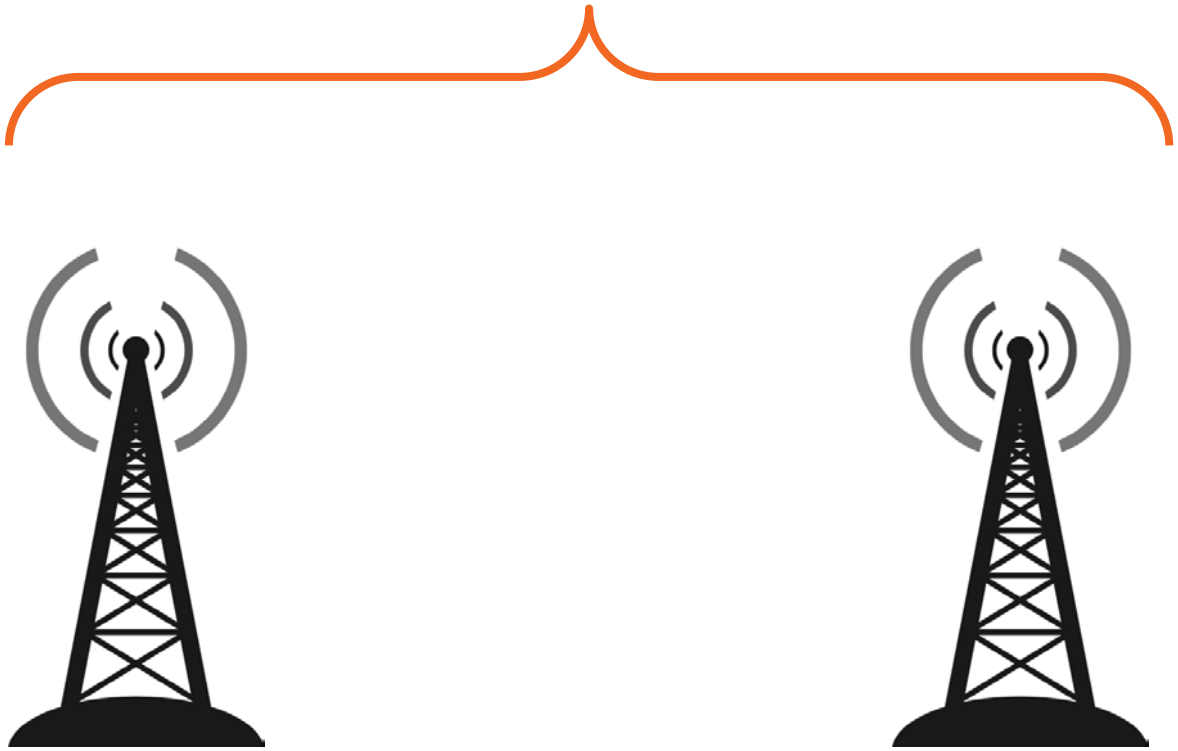
RowKey	d:utc	t:1	t:2	p:1	e:f
e399	1441880447000	a3b	f19	b49d70	1

Transactional data

Navigation **by ID**

Atomic updates

race-timers	r
raceID | date	race



race-timers	r

RowKey	r:e399
a3b|20150910	1

Reference data

Access **by key**

Fast lookup

timing-events	t
racerID | timerID | raceID	timestamps

timing-events	t
racerID | timerID | raceID	timestamps

RowKey	t:4530a	t:8f9f2	t:24c4d
887971|a3b|e399	1441880506000	1441880506412	1441880506603

RowKey	t:8201e	t:3f6d7
d3929f|a3b|e399	1441880487000	1441880487014

timing-events	t
racerID | timerID | raceID	timestamps

200 rows

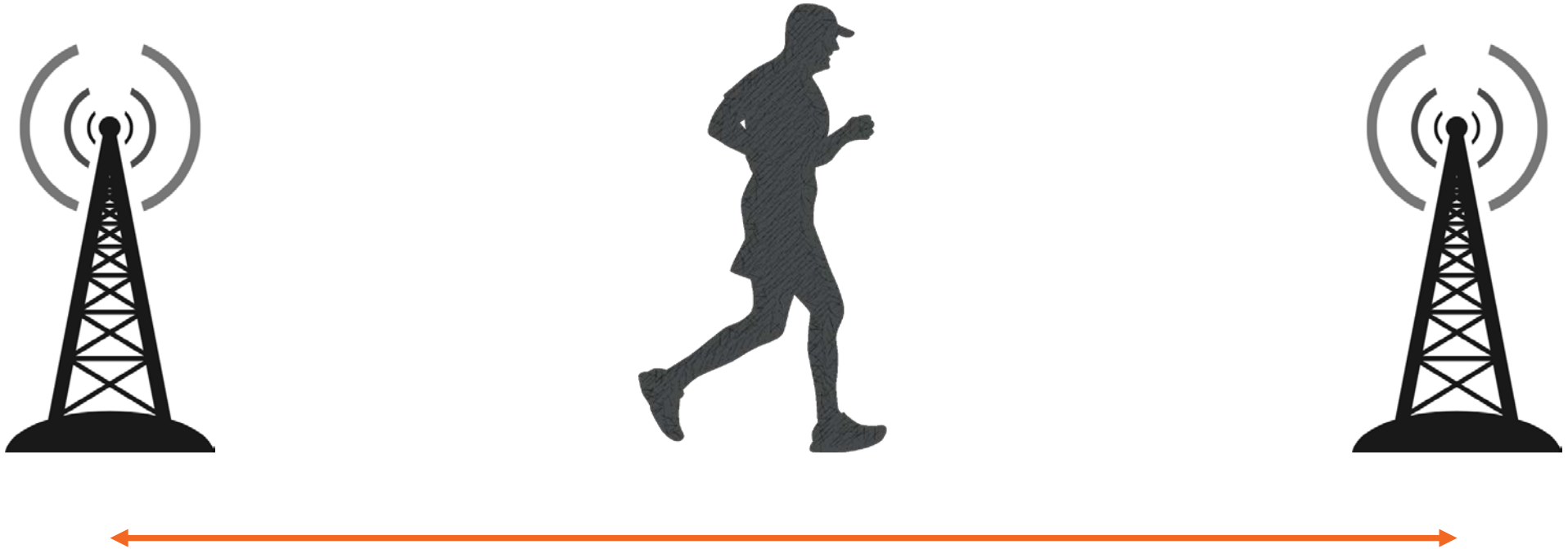
50 columns per row

timing-events	t
racerID | timerID | raceID	timestamps

5.1M rows

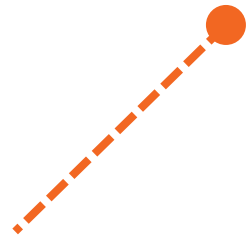
100 columns per row

sector-times	te	d
racerID | raceID	timing events	durations



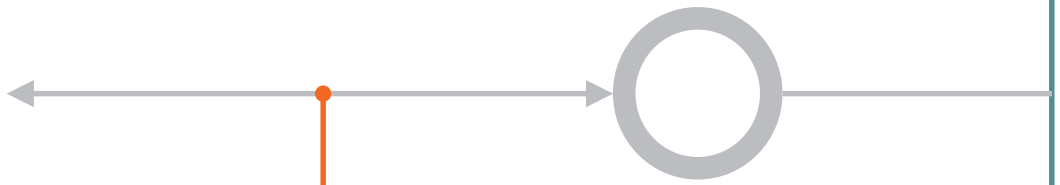
sector-times	te	d
racerID | raceID	timing events	durations

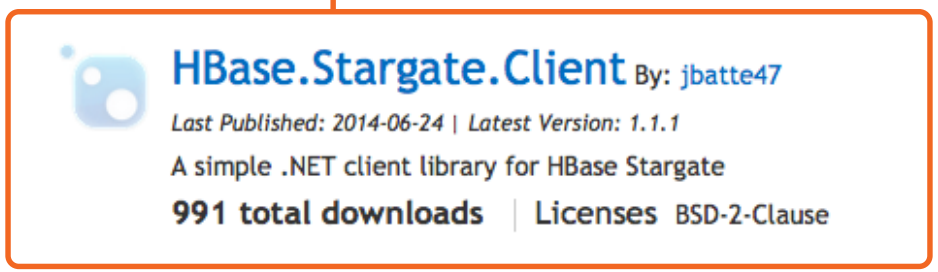
RowKey	te:a3b	te:f19	d:1	d:t
503f19|e399	1441880503000	1441882003000	1500000	1500000

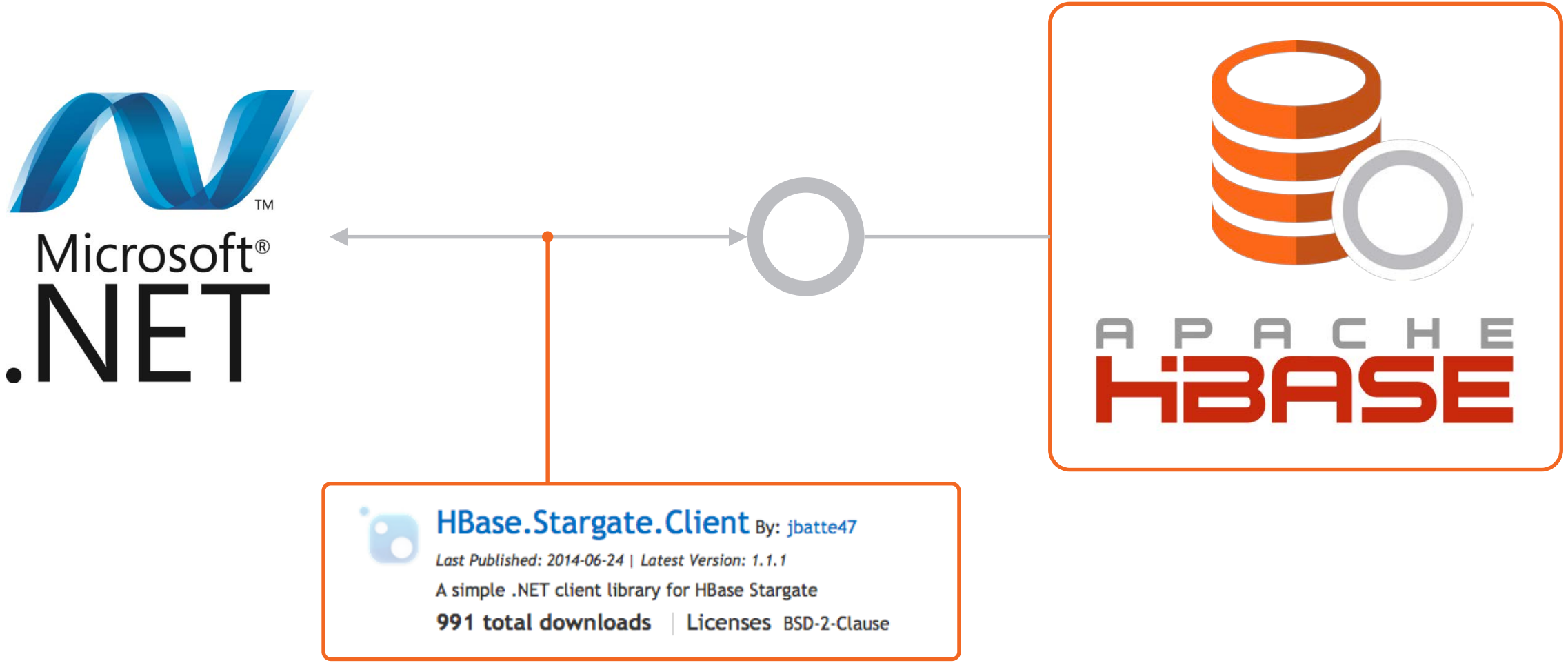


Computed data

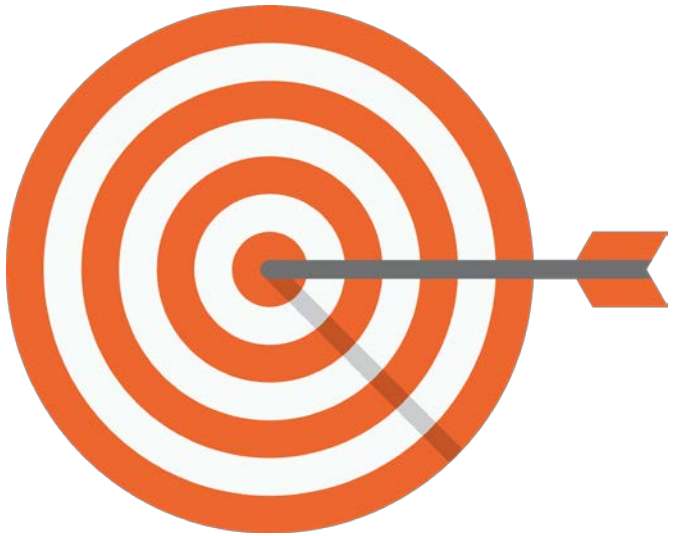
Real-time results





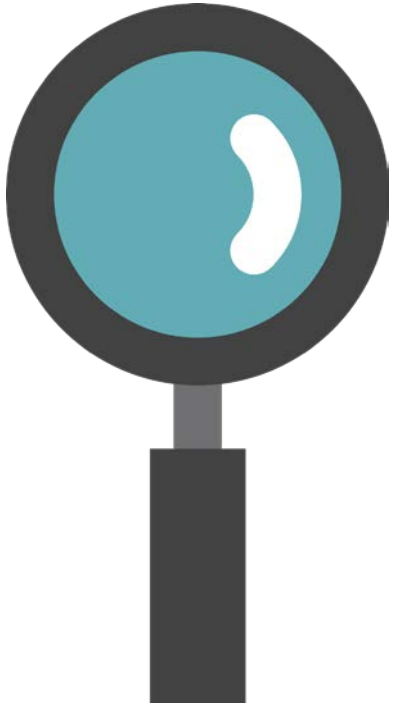


Module Goals



- Structuring data in HBase
- Data access patterns
- Accessing HBase from .NET
- Running HBase on HDInsight

Coming Next



HFiles and HDFS

Table setup

Modelling HBase in .NET

Administration & performance