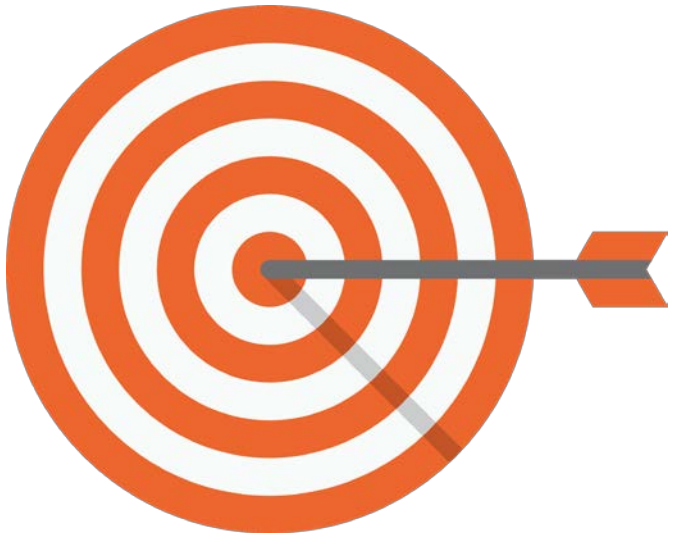
Hive Deep Dive

Elton Stoneman

@EltonStoneman | blog.sixeyed.com

pluralsight

Module Goals



Hive Processing Model

Monitoring Queries in YARN

pluralsight

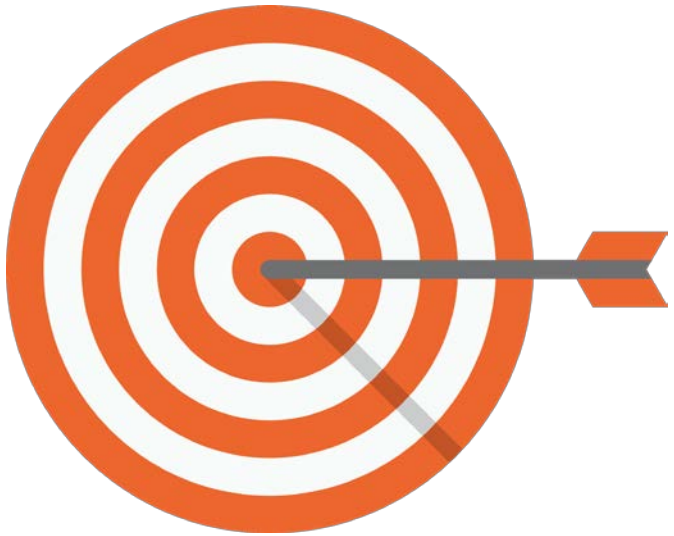
timing-events	t

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

x41M rows

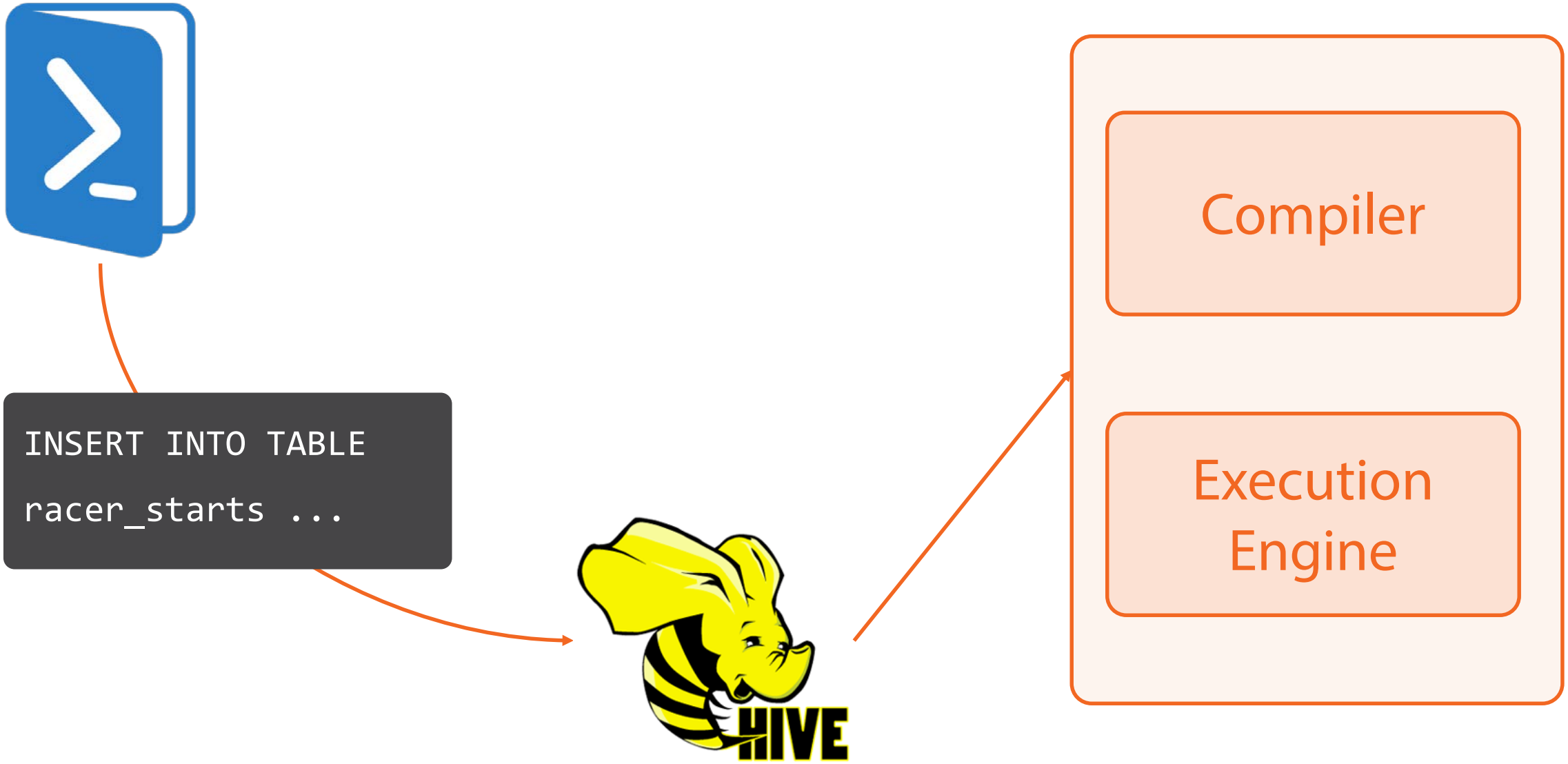
pluralsight

Module Goals

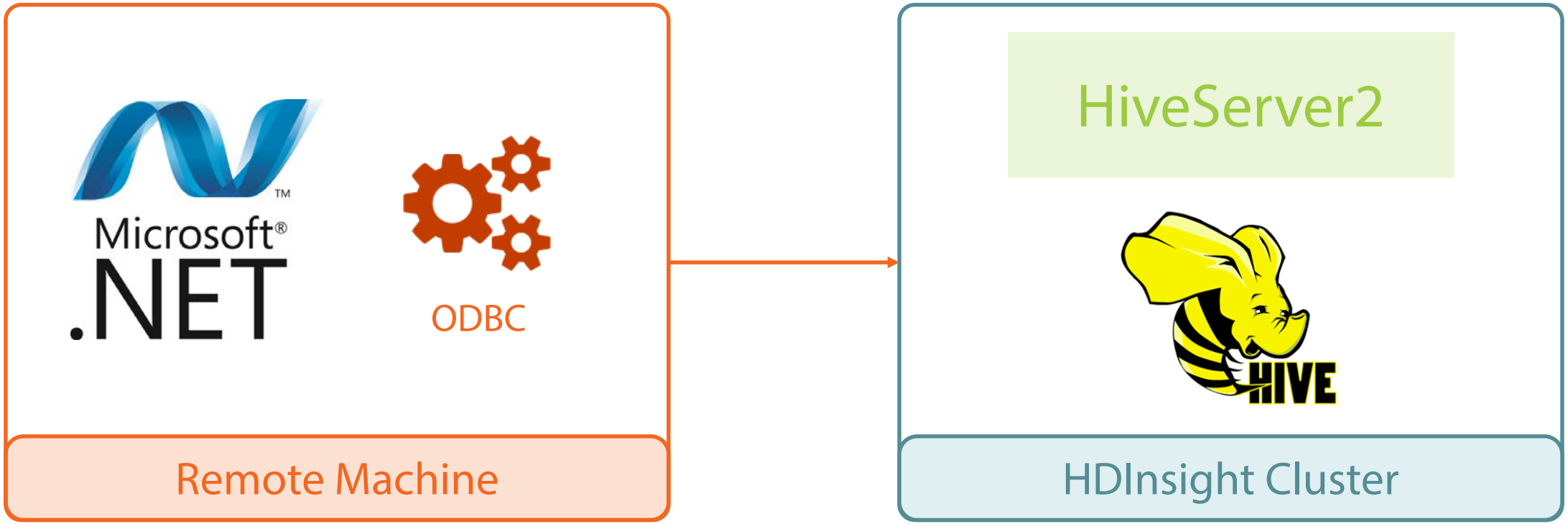


- Hive Processing Model
- Monitoring Queries in YARN
- Query Plans & Optimization

pluralsight

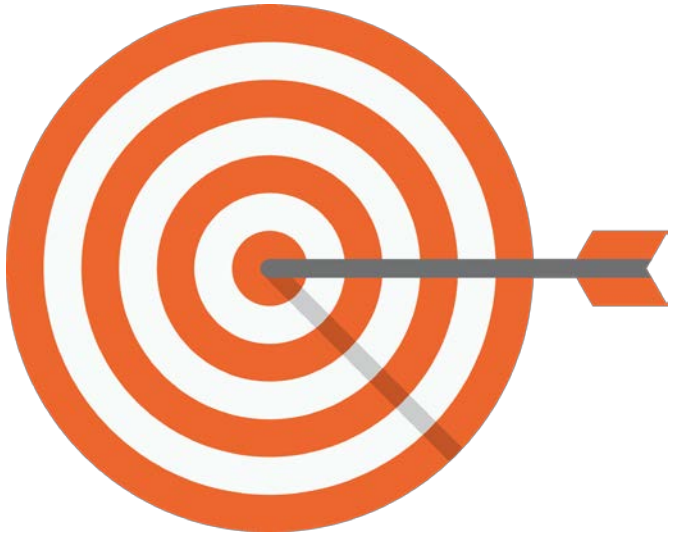


pluralsight



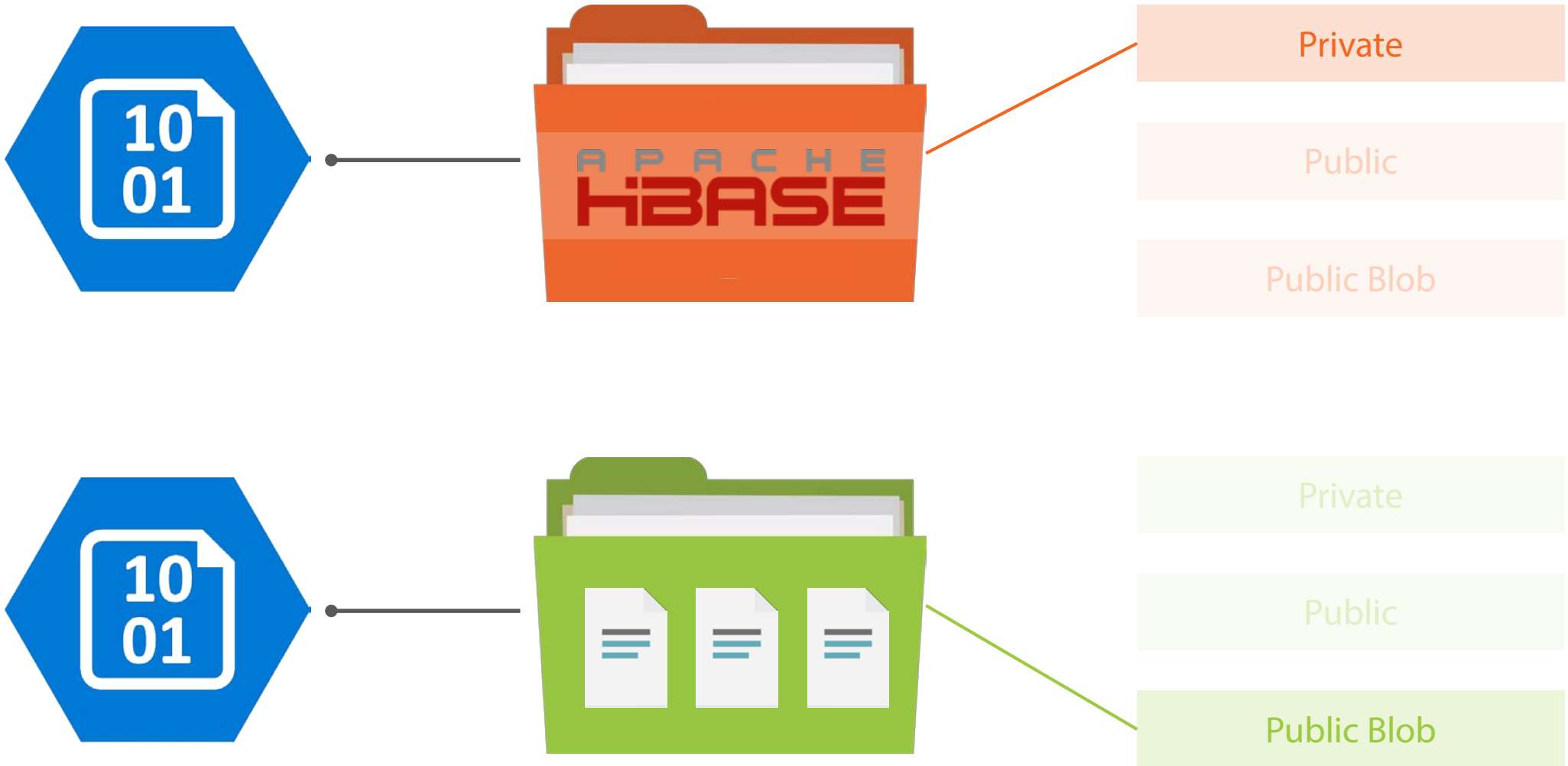
pluralsight

Module Goals



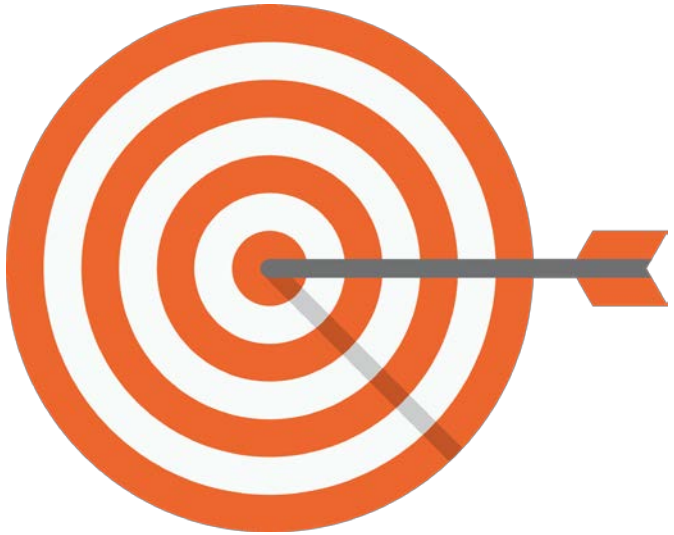
- Hive Processing Model
- Monitoring Queries in YARN
- Query Plans & Optimization
- Hive with PowerShell & .NET

pluralsight



pluralsight

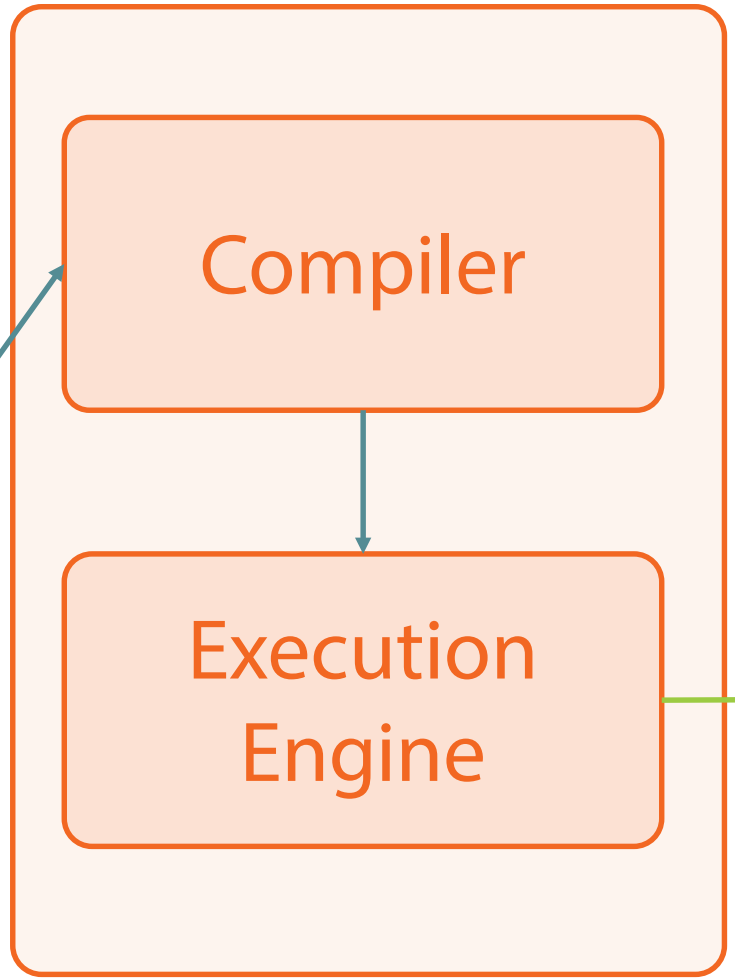
Module Goals

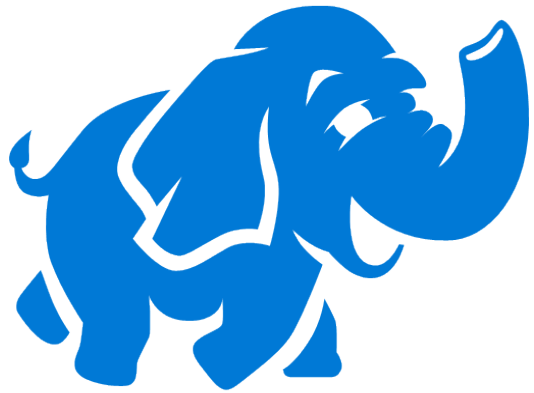


- Hive Processing Model
- Monitoring Queries in YARN
- Query Plans & Optimization
- Hive with PowerShell & .NET
- Initialising Hive in HDInsight

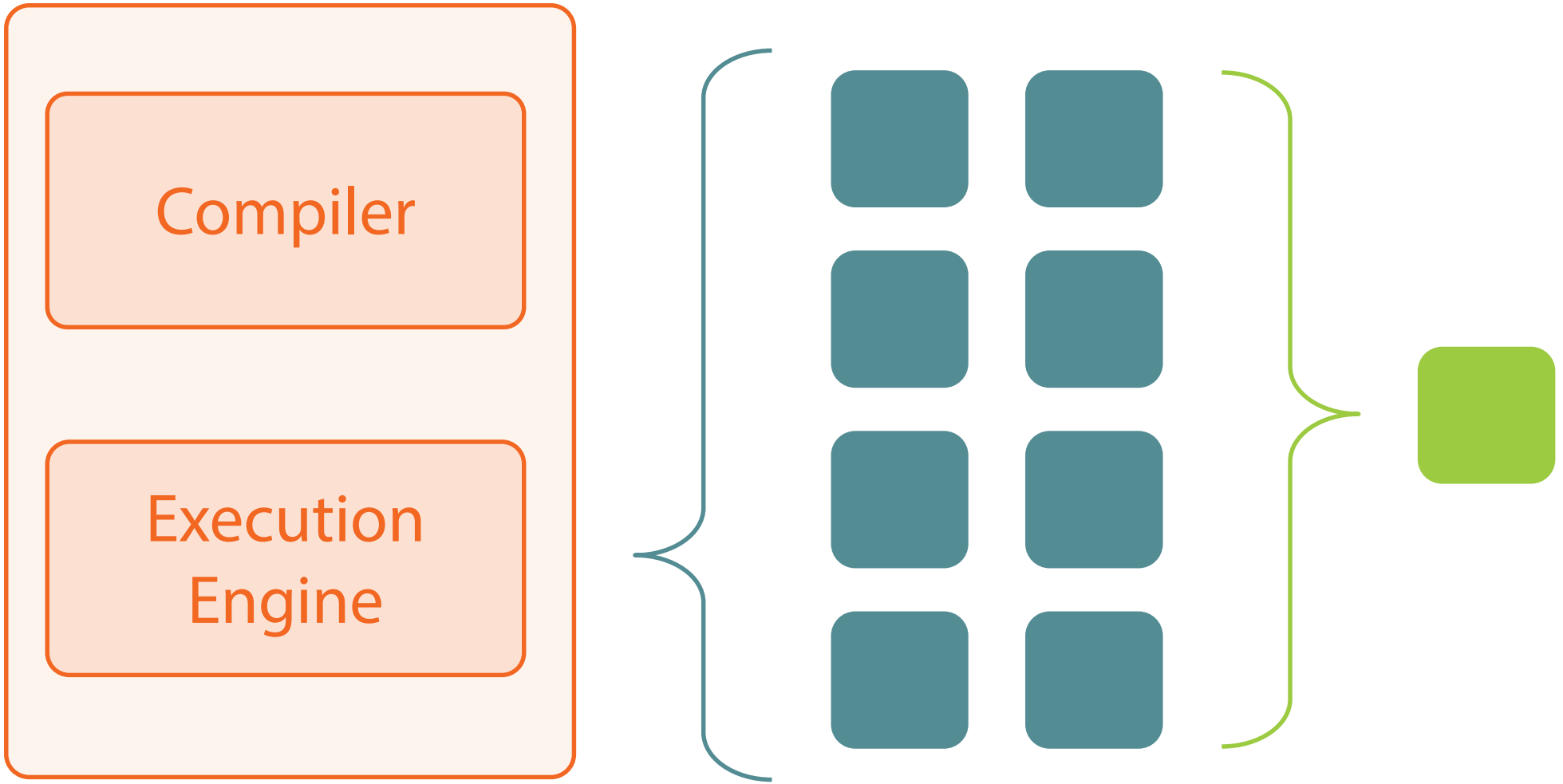
pluralsight

```
SELECT COUNT(*)
FROM timing_events;
```

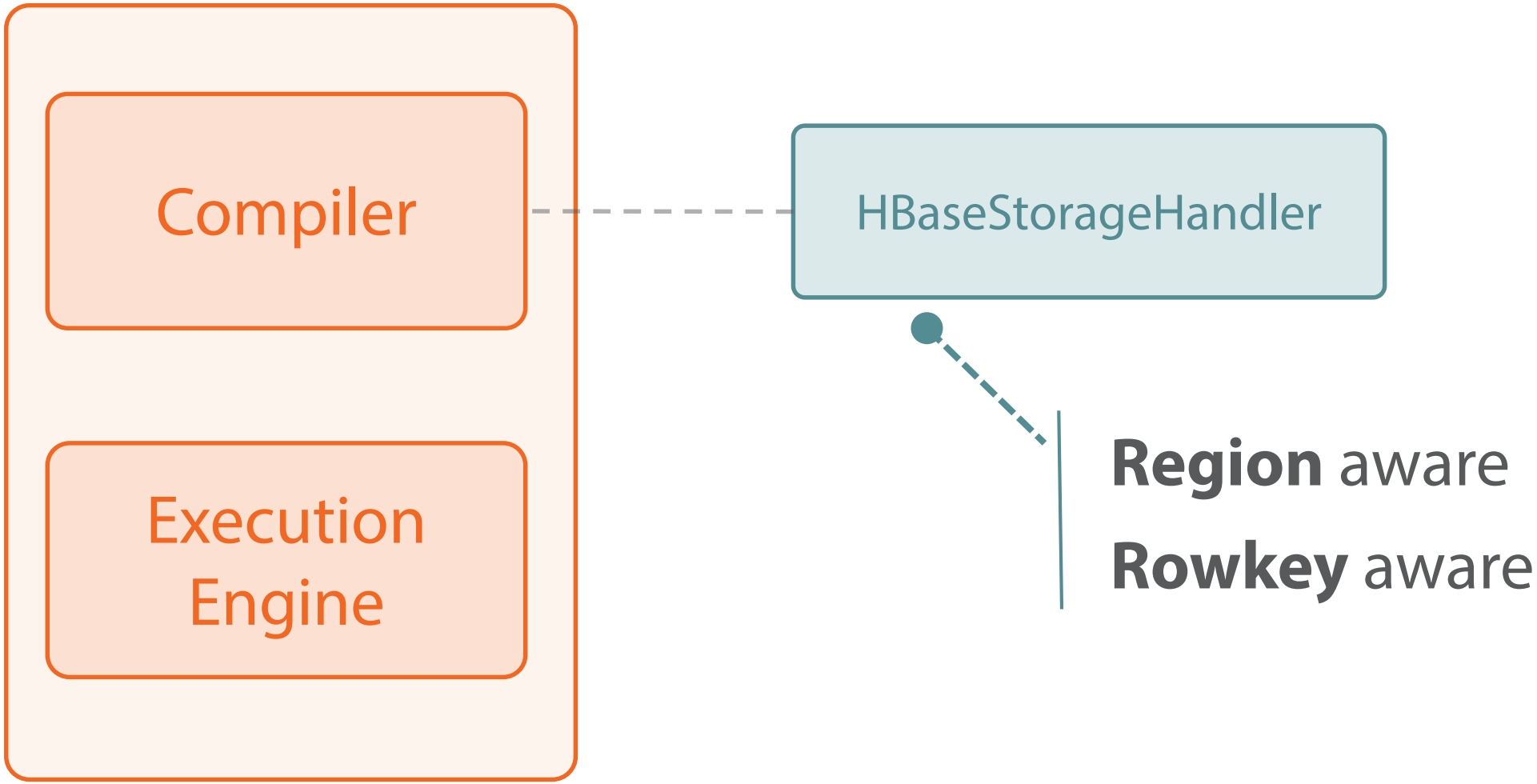




pluralsight



pluralsight



pluralsight

```
SELECT COUNT(*)
FROM timing-events;
```

APACHE HBASE

pluralsight

```
SELECT COUNT(*)
FROM timing-events;
```

count rows
timing-events
partition **1**

count rows
timing-events
partition **2**

count rows
timing-events
partition **3**

• • •

count rows
timing-events
partition **256**

pluralsight

count rows
timing-events
partition **1**
=165K

count rows
timing-events
partition **2**
=170K

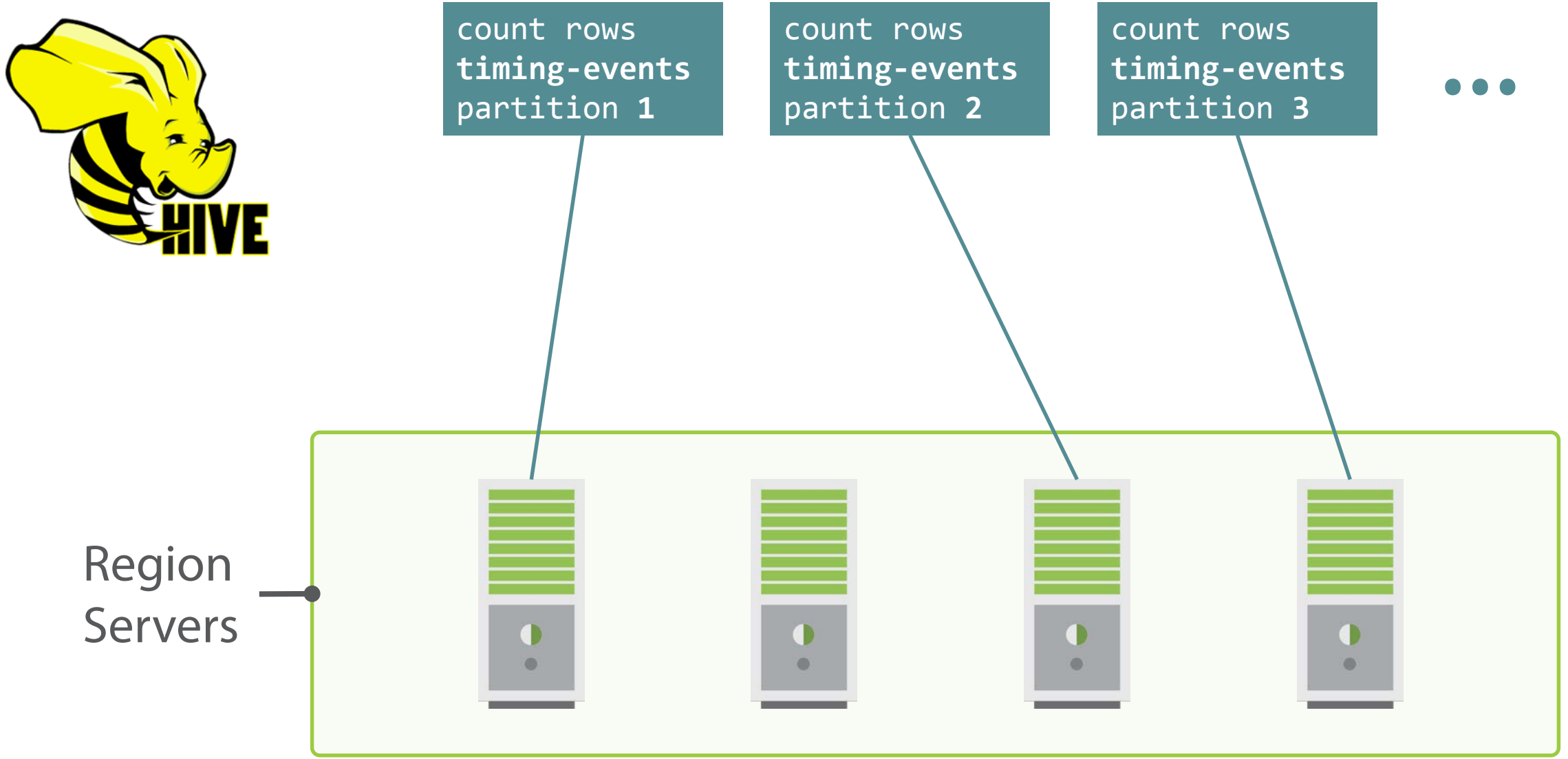
count rows
timing-events
partition **3**
=160K

• • •

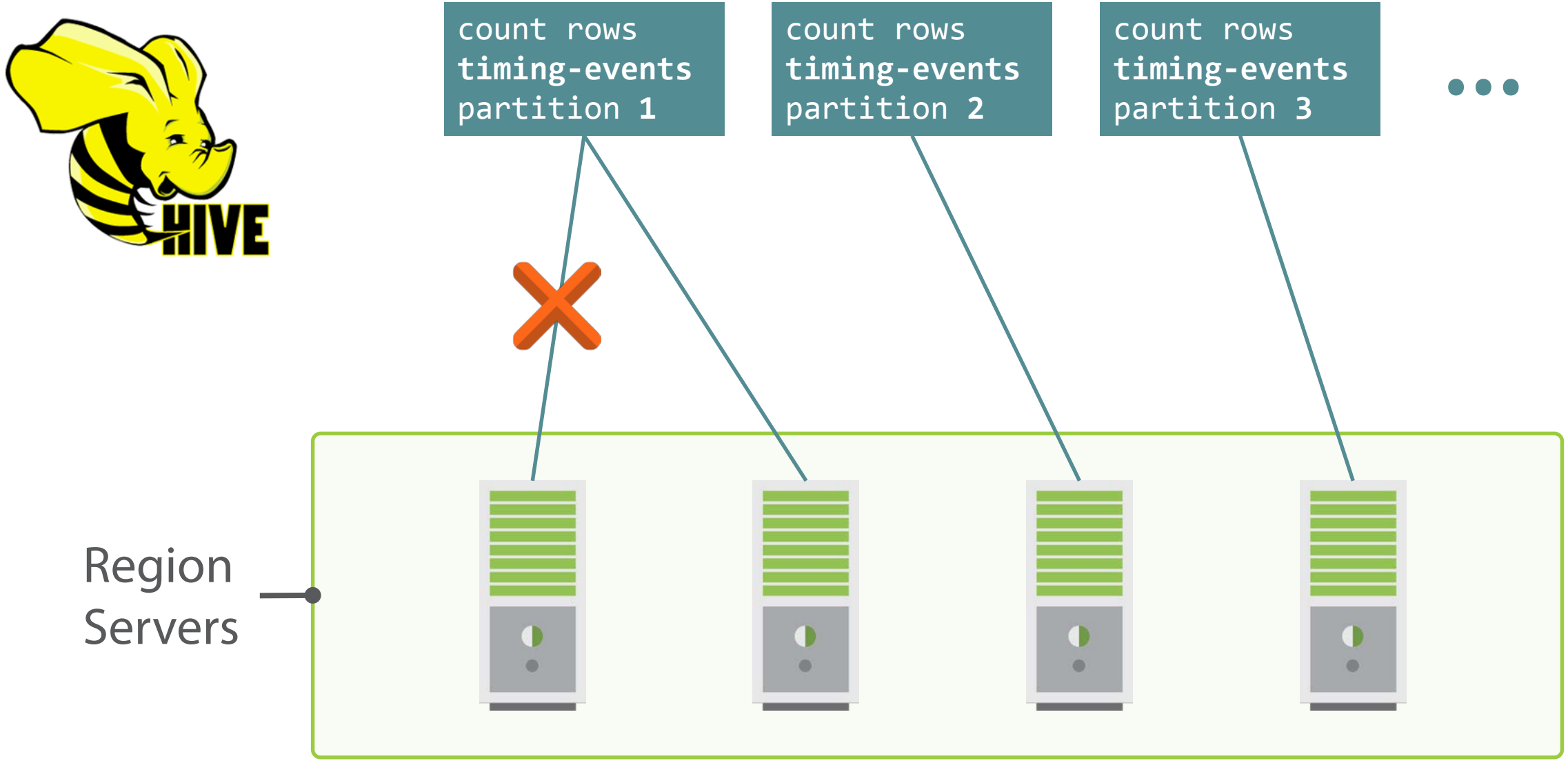
count rows
timing-events
partition **256**
=165K

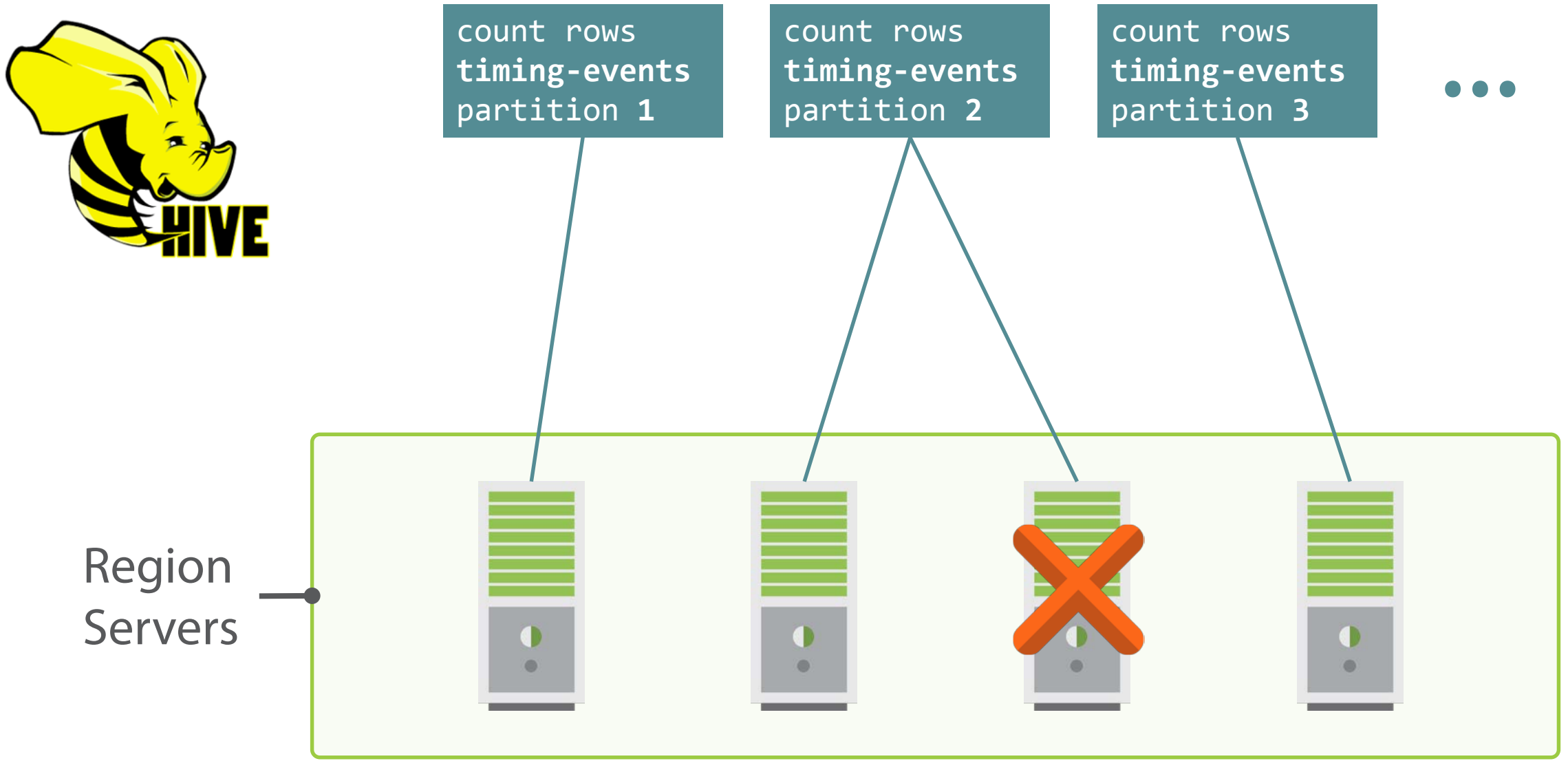
sum map
totals
= 41M

pluralsight

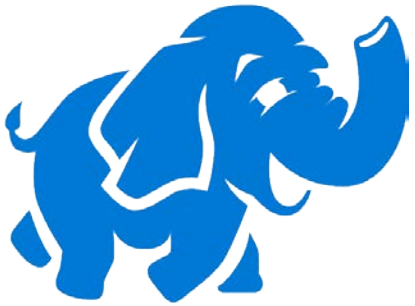


pluralsight


count rows
timing-events
partition 1
count rows
timing-events
partition 2
count rows
timing-events
partition 3
Region
Servers



pluralsight

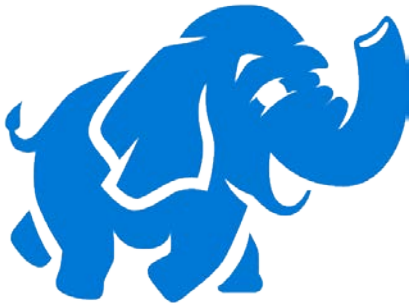


YARN

Yet Another Resource Negotiator

Job scheduling

pluralsight



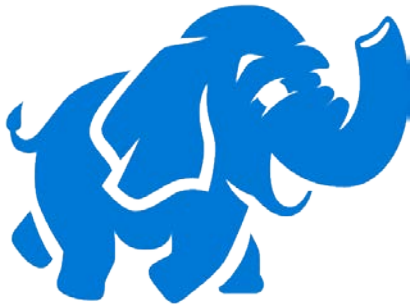
YARN

Yet Another Resource Negotiator

HDFS

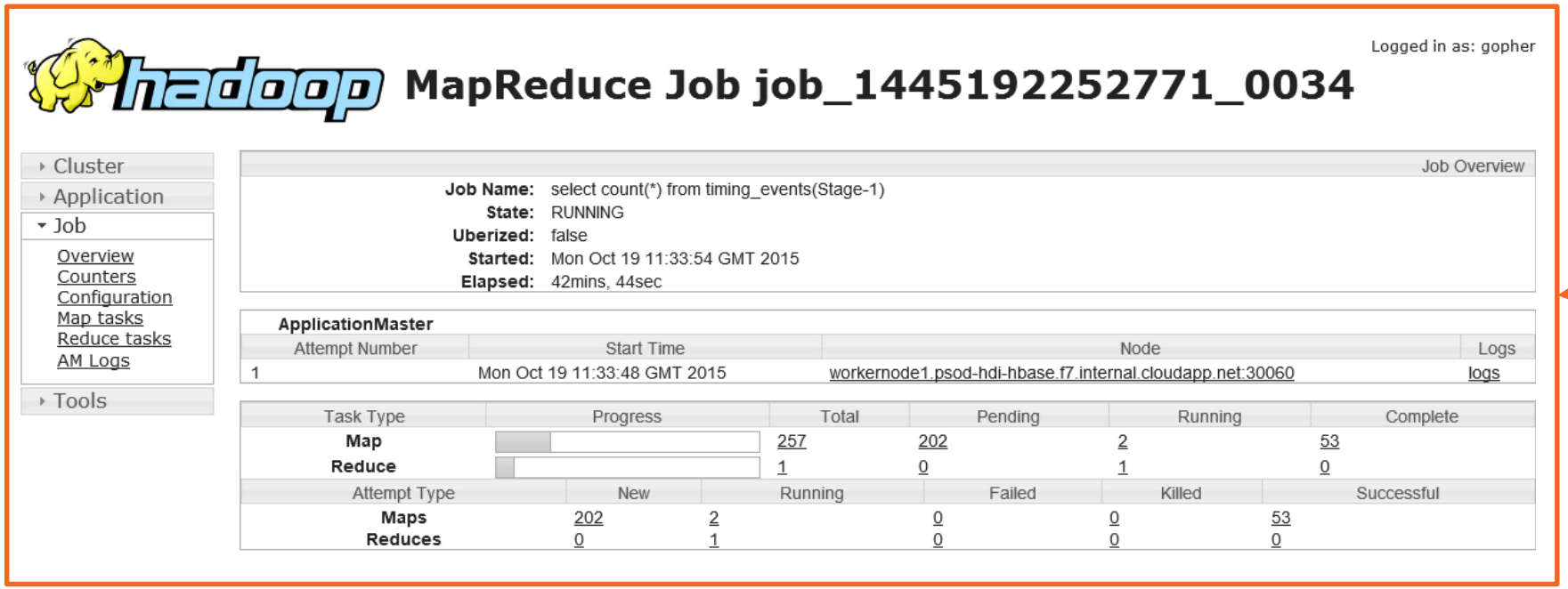
Hadoop Distributed File System

pluralsight



YARN

Yet Another Resource Negotiator



pluralsight

Demo: YARN UI

Hive Query

HBase Regions

YARN Jobs

pluralsight

Demo: Execution Plan

Hive EXPLAIN

HBase row keys

Filter push-down

pluralsight

```
EXPLAIN
SELECT COUNT(*)
FROM timing_events;
```

Compiler

Execution Engine

pluralsight

```
...
JOIN earliest_timing_events t1
 ON t1.timerid = rt.firstTimerId
 AND t1.raceid = rt.raceid
 AND t1.racerid = rd.racerid
```

timing-events	t

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

pluralsight

```
...
JOIN earliest_timing_events t1
 ON t1.rowkey = rd.racerid + '|' + rt.firstTimerId + '|' + rt.raceid
```

Rowkey filter

Table scan

timing-events	t

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

Filter push-down only on **constant values**

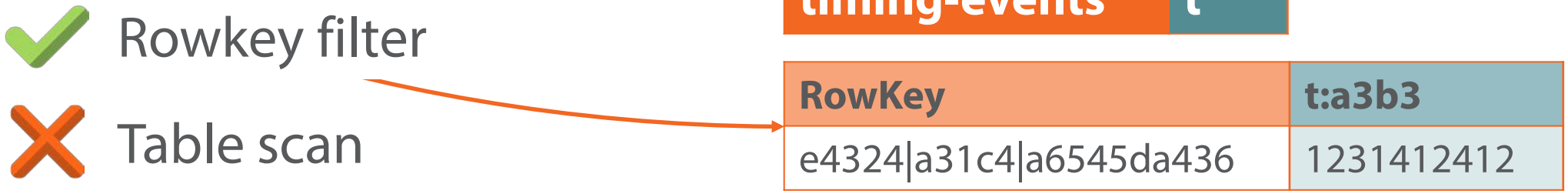
pluralsight

```
...
JOIN earliest_timing_events t1
...
WHERE t1.rowkey BETWEEN '08e54e|dc1c|04b4dd' AND 'fe2994|dc1c|04b4dd'
```

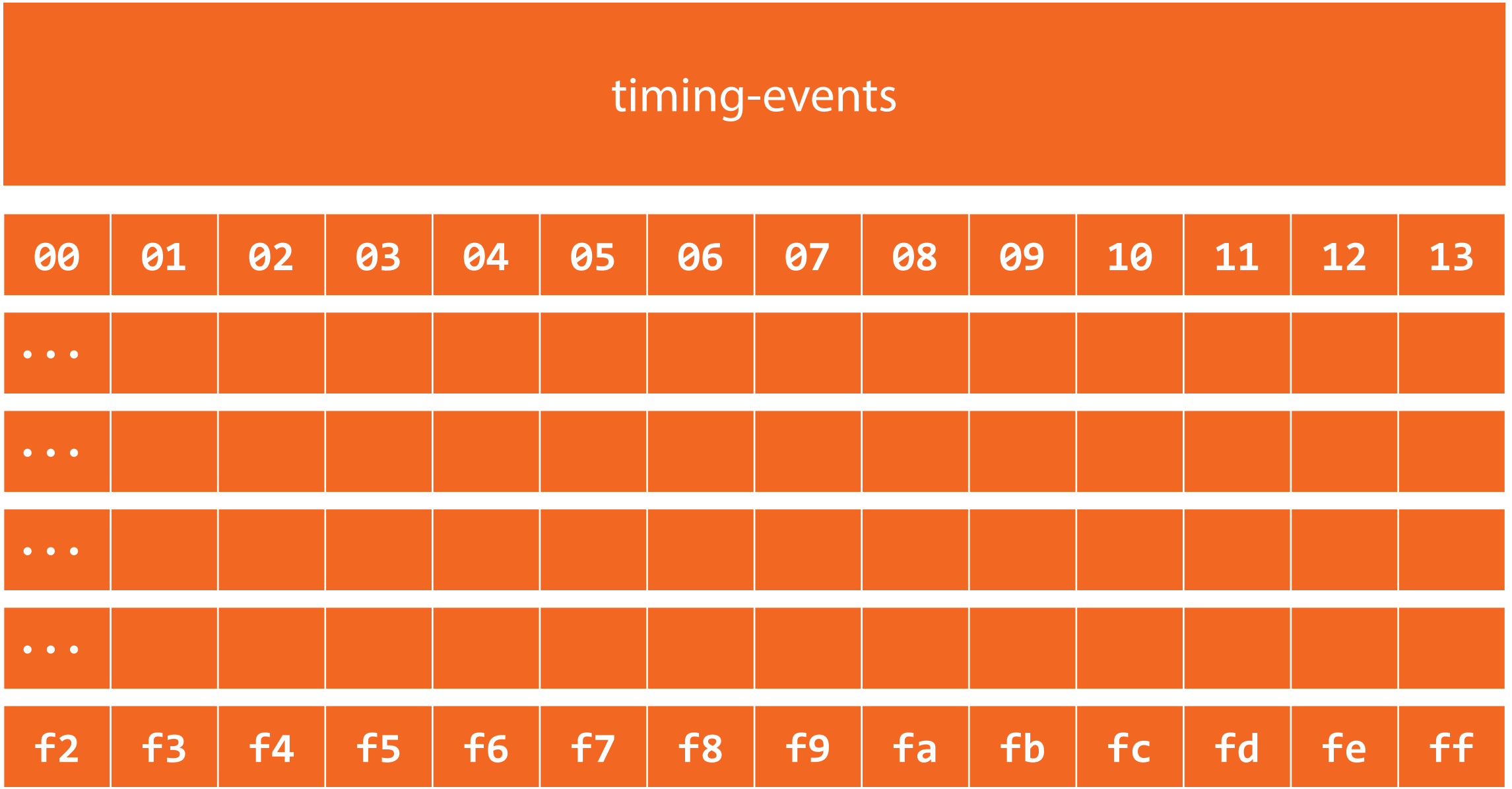
First Name	*Last Name*	*Racer ID*
Dallas	Miyashiro	08e54e
Janna	Przybyl	0a3ec3
Olene	Skubis	0e7ee4
...		
Nell	Pao	fbdd74
Arlyne	Dellinger	fe073a
Chana	Perera	fe2994

pluralsight

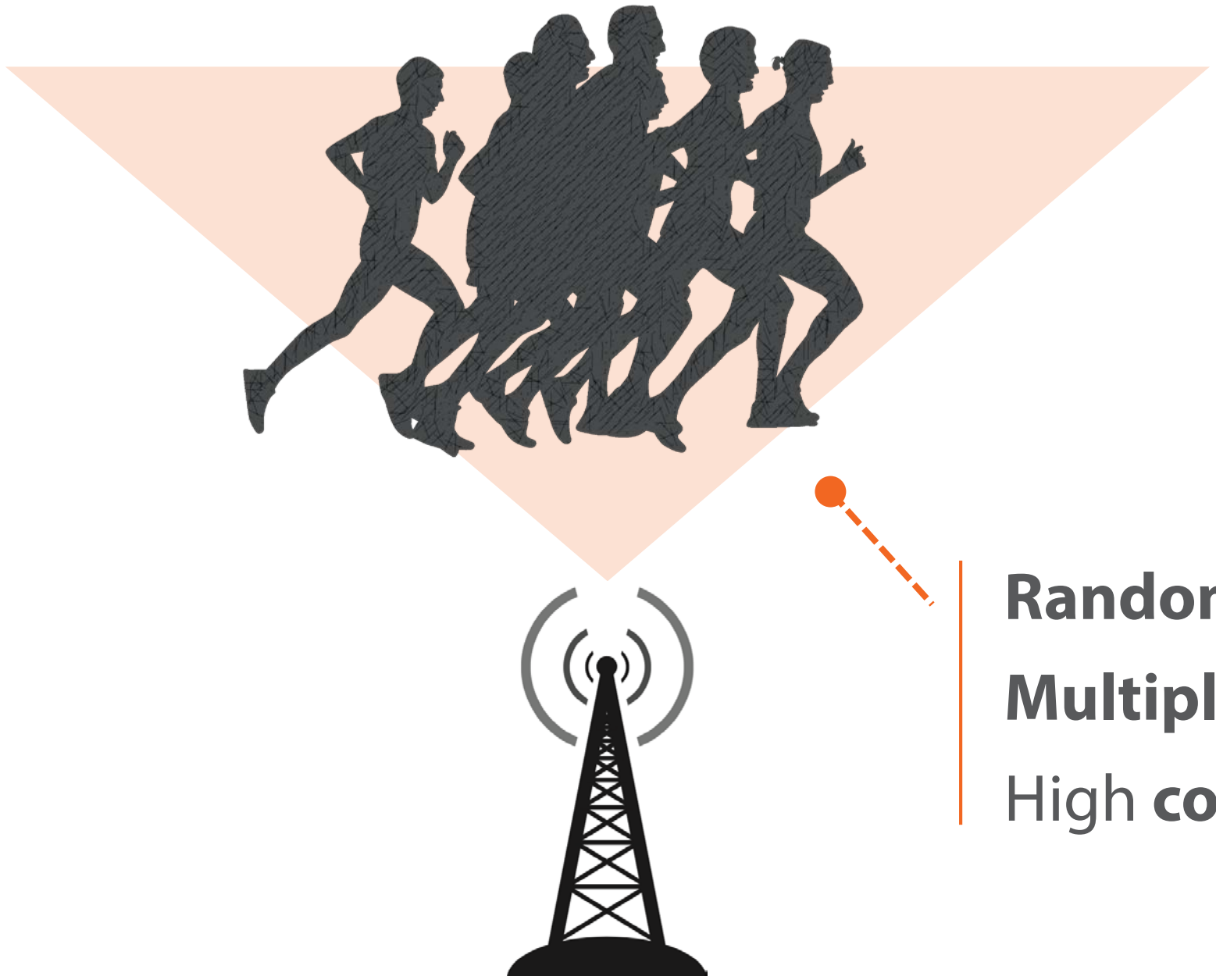
```
...
JOIN earliest_timing_events t1
...
WHERE t1.rowkey BETWEEN '08e54e|dc1c|04b4dd' AND 'fe2994|dc1c|04b4dd'
```



pluralsight



pluralsight

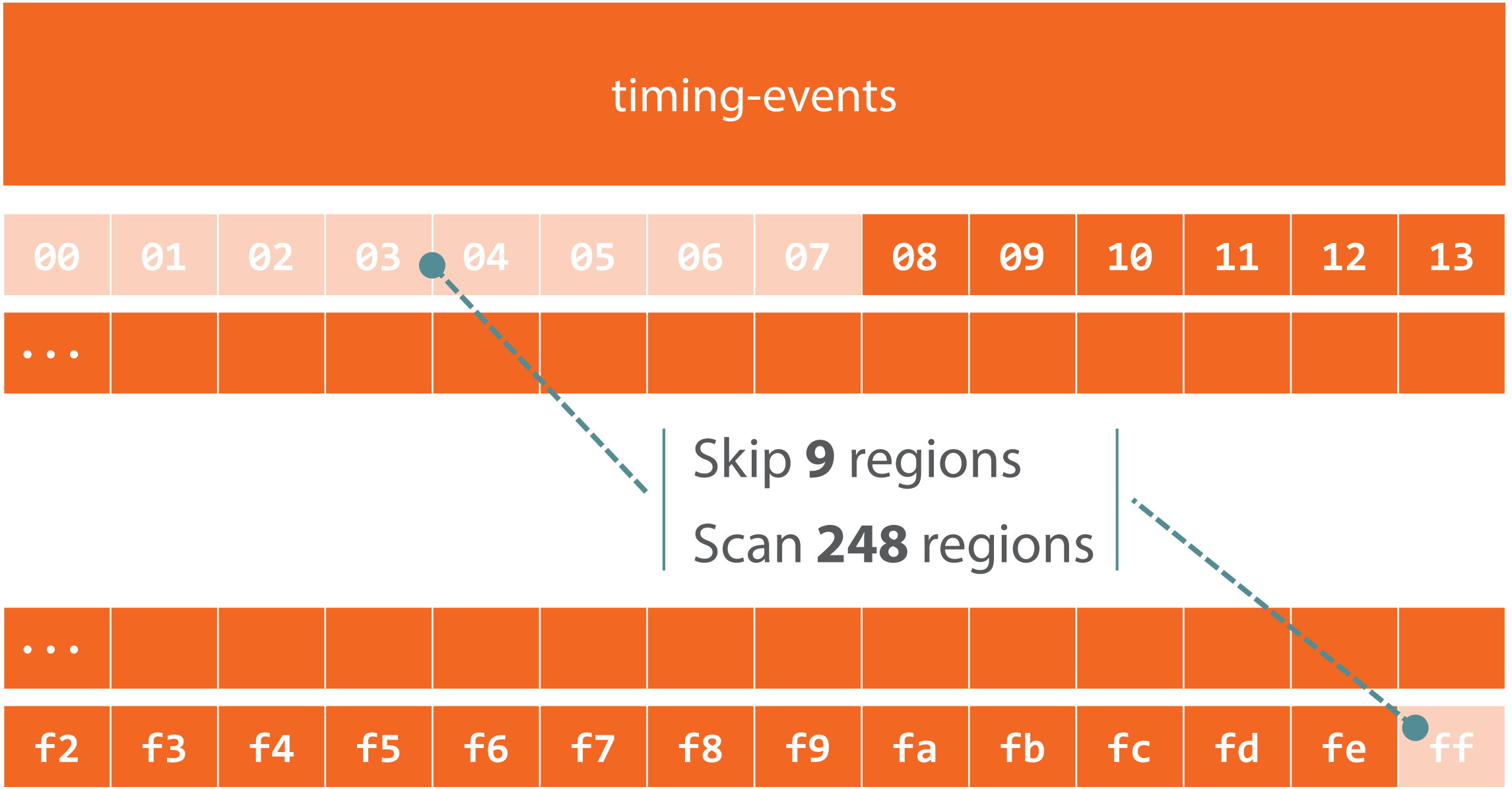


Random IDs

Multiple regions

High **concurrency**

pluralsight



pluralsight

```
INSERT INTO #race_rowkeys
SELECT racerId + '|dc1c|04b4dd'
FROM racer_details...
```

```
FOR rowkey IN (
 SELECT * FROM #race_rowkeys
)
LOOP
 --process racer
END LOOP;
```

First Name	Last Name	Racer ID
Dallas	Miyashiro	08e54e
Janna	Przybyl	0a3ec3
...		
Arlyne	Dellinger	fe073a
Chana	Perera	fe2994

timing-events | **t**

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

pluralsight

Demo: Hive and PowerShell

Submitting Query Jobs

Parameterising Queries

Reading Results

pluralsight

60+ seconds

- Compiler
- Execution Engine

```
INSERT INTO TABLE
racer_starts ...
```

pluralsight

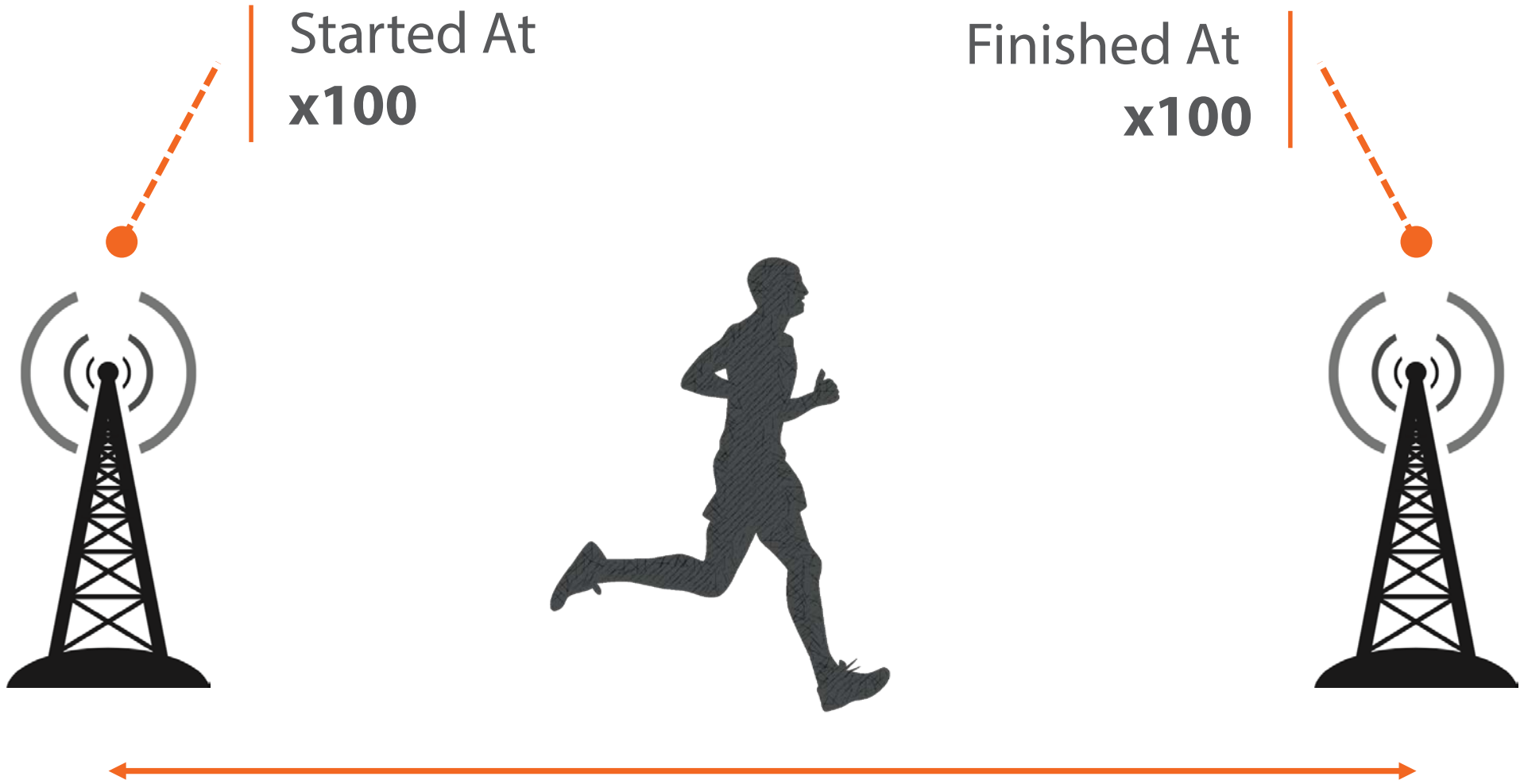
Demo: Multiple Hive Jobs

Bulk INSERTs

Multiple Queries

Running in Parallel

pluralsight



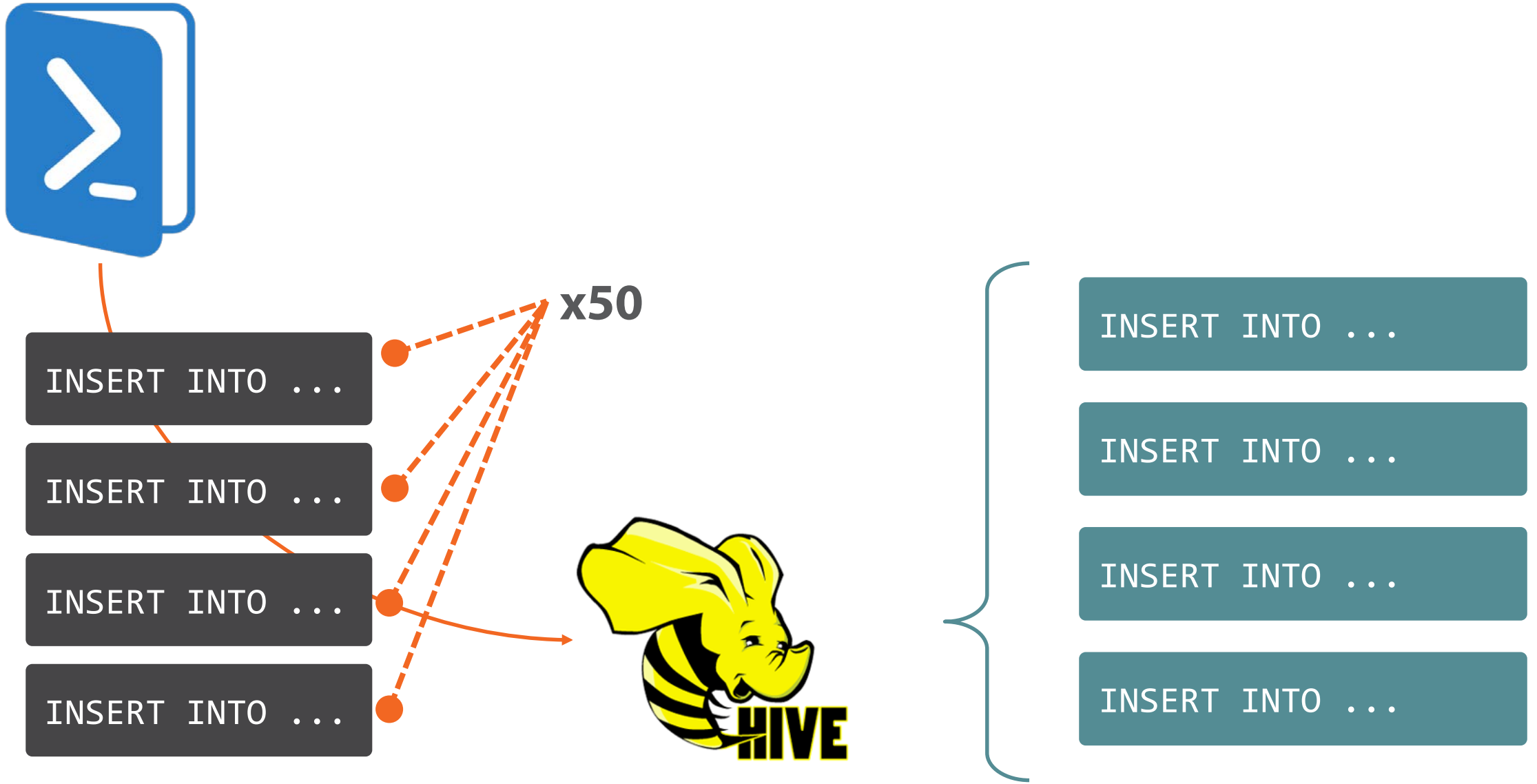
pluralsight

x200

```
INSERT INTO TABLE
racer_starts ...
```

```
INSERT INTO ...
INSERT INTO ...
INSERT INTO ...
INSERT INTO ...
...
INSERT INTO ...
```

pluralsight

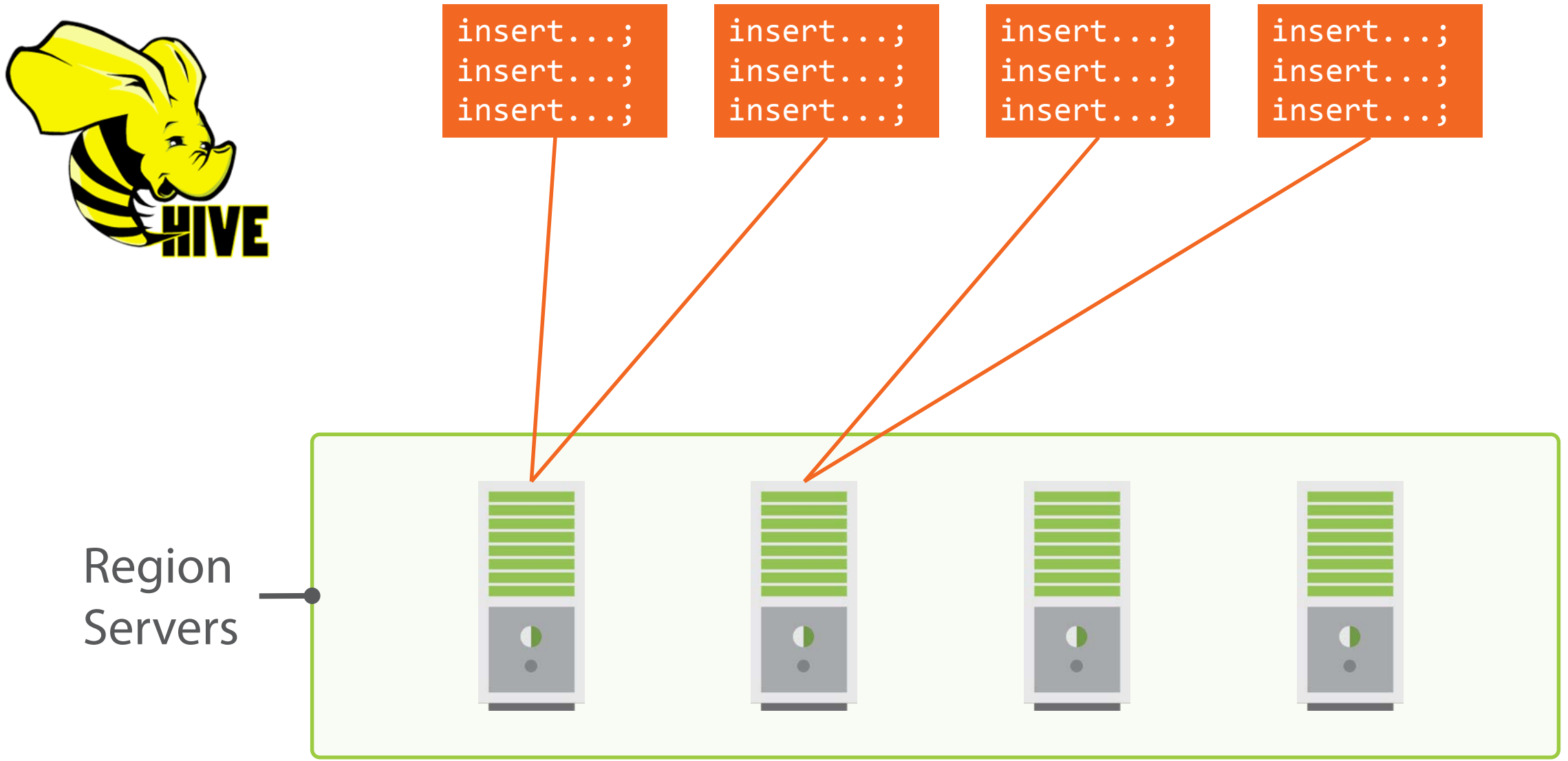


pluralsight

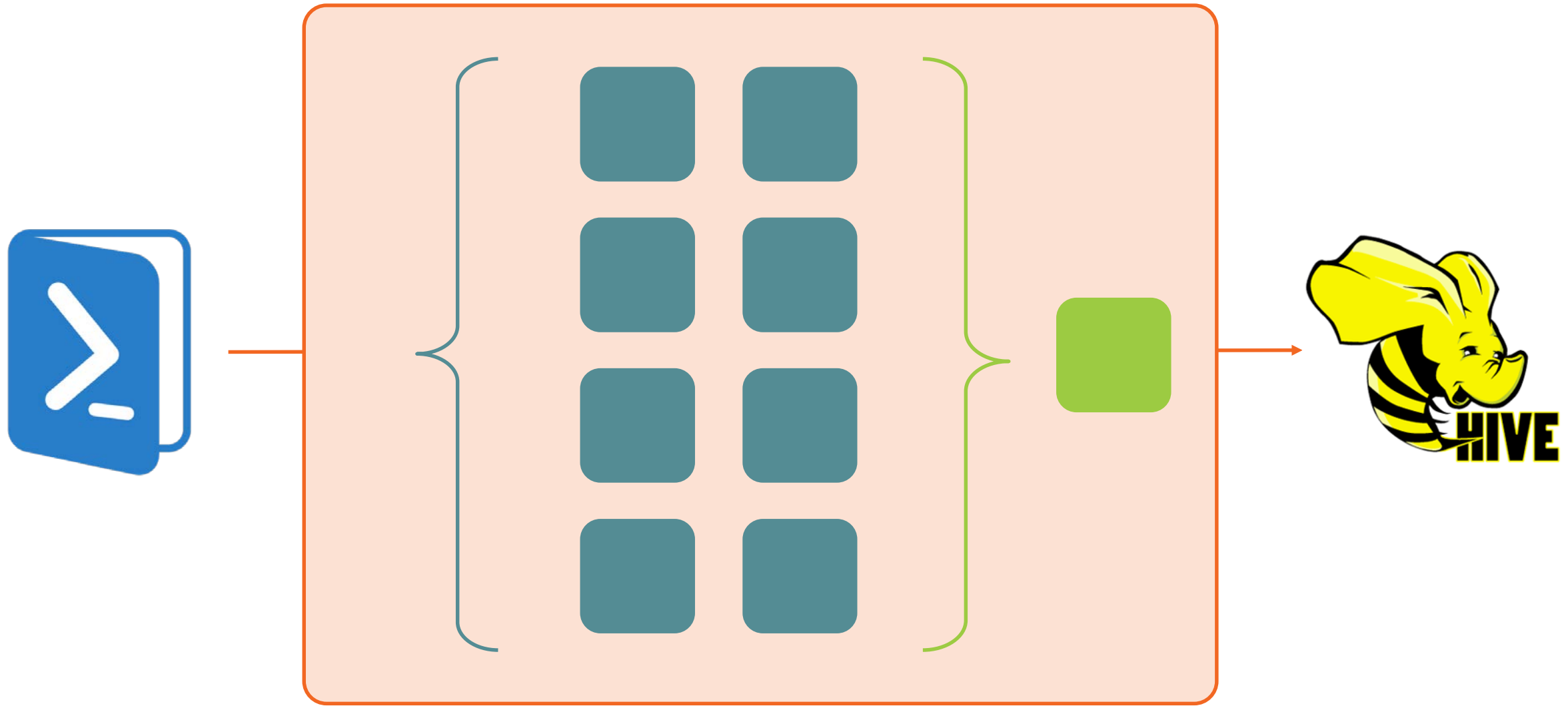
```
INSERT racer_starts ...
INSERT racer_starts ...
INSERT racer_finishes ...
INSERT racer_finishes ...
```

x50

pluralsight



pluralsight



pluralsight

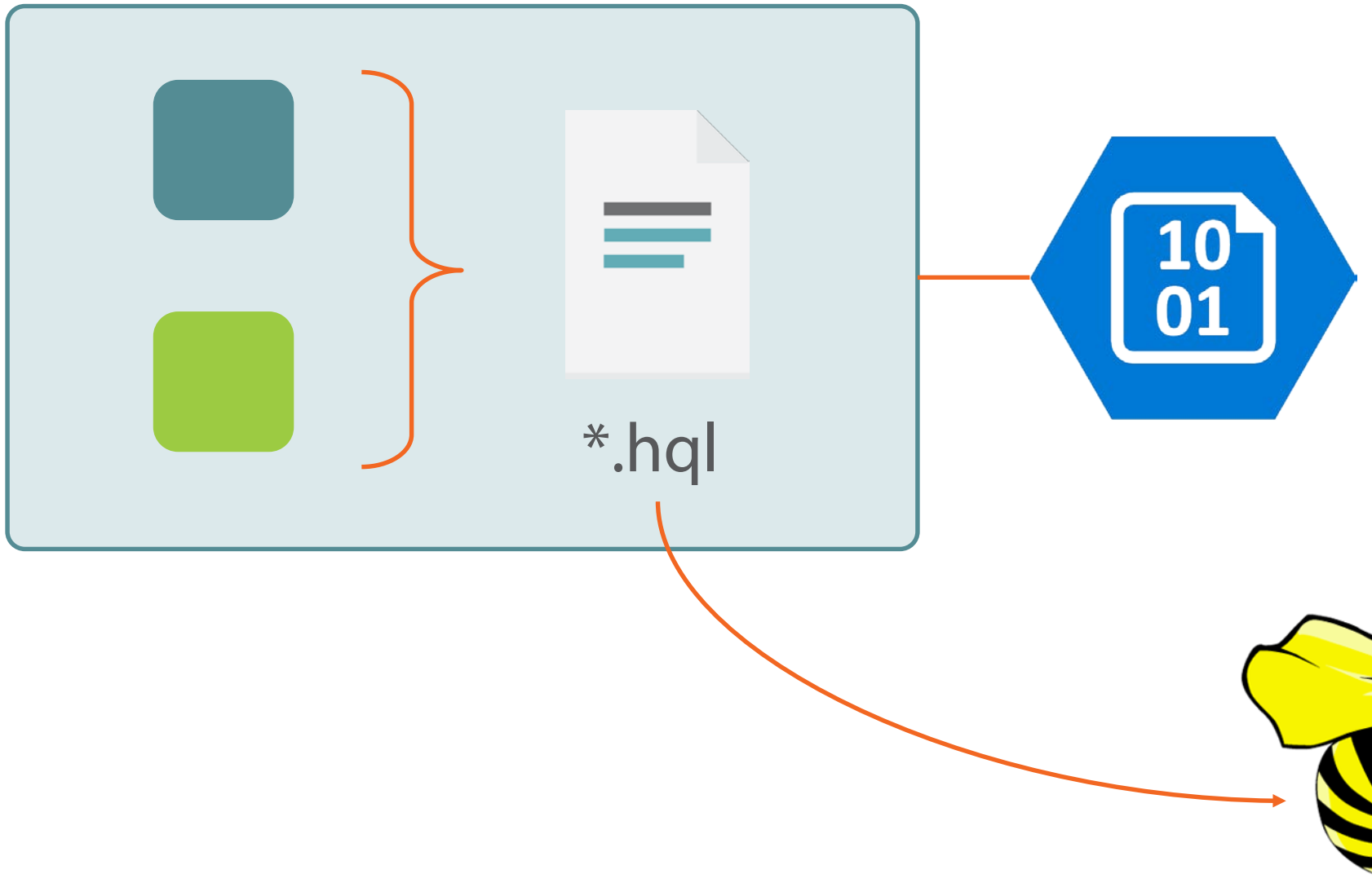
```
INSERT INTO TABLE racer_starts PARTITION (raceId = '{1}')
 SELECT racerId, earliestTimestamp
 FROM earliest_timing_events
  WHERE rowkey = '{2}'
```

- ✔ Rowkey filter
- ✘ Table scan

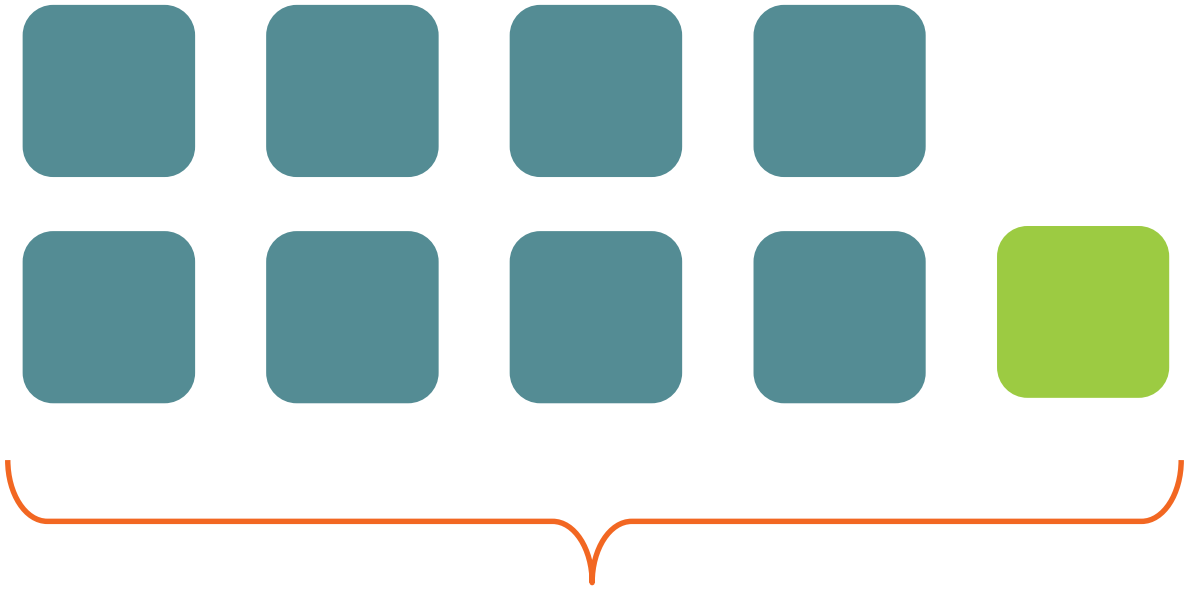
timing-events	t

RowKey	t:a3b3
e4324|a31c4|a6545da436	1231412412

pluralsight

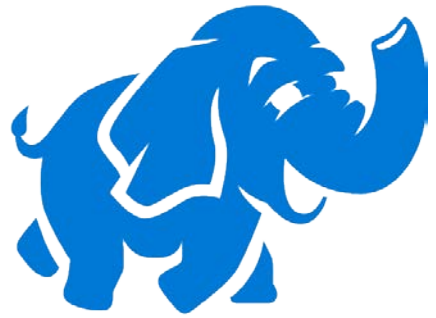


pluralsight

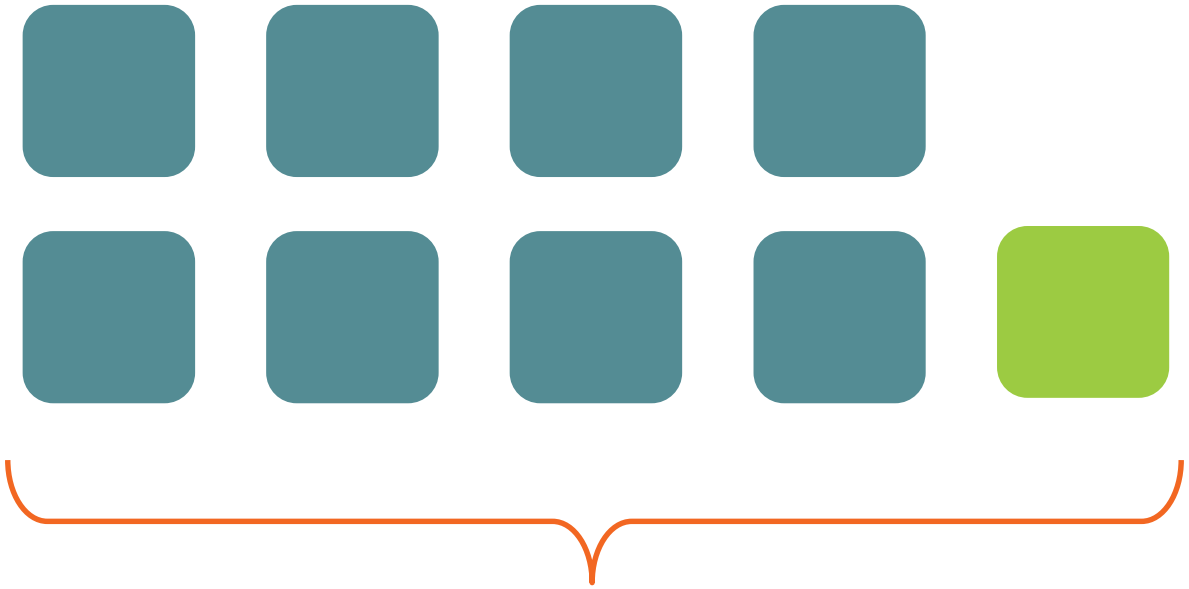


YARN

Yet Another Resource Negotiator



pluralsight

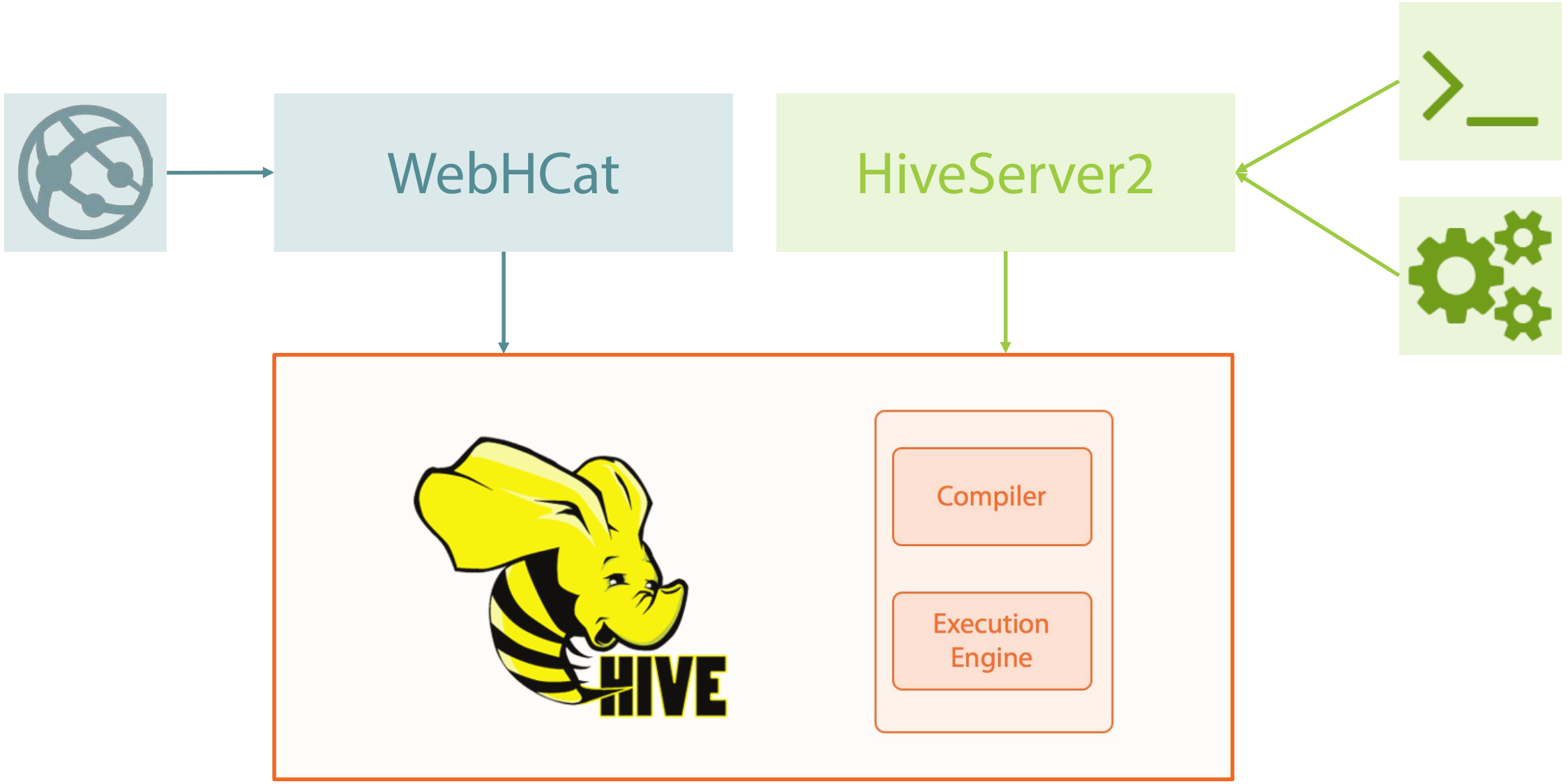


27+ minutes

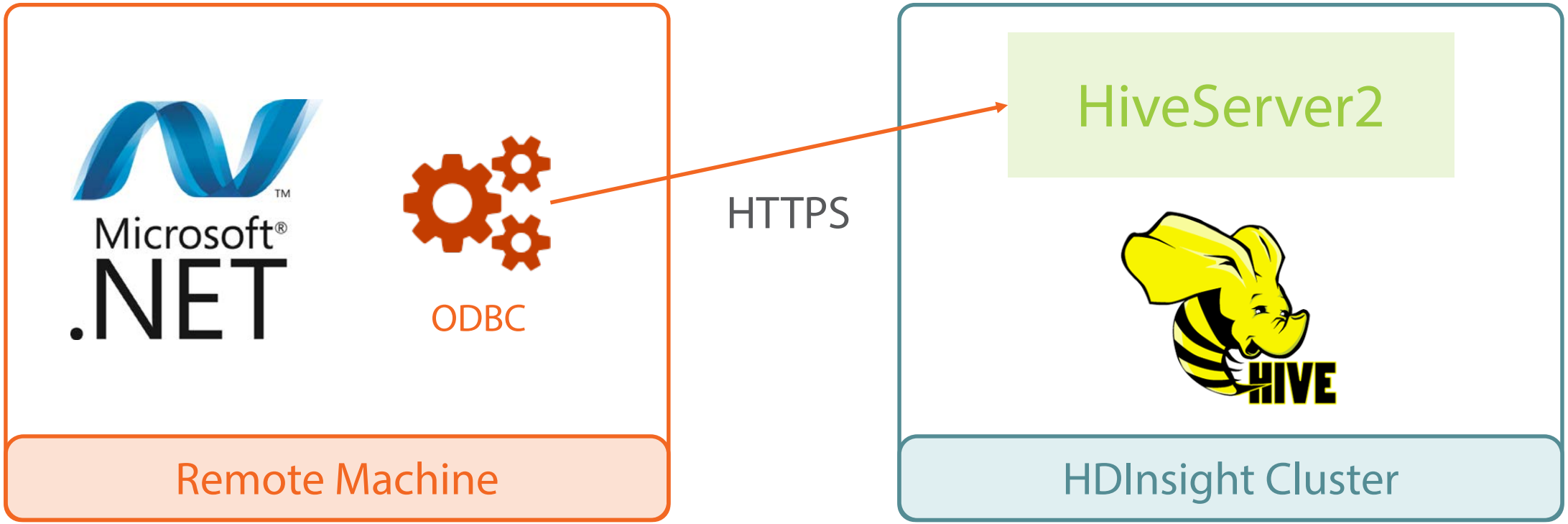
YARN

Yet Another Resource Negotiator

pluralsight



pluralsight



pluralsight

Demo: User Defined Functions

Hadoop Streaming Interface

.NET Console App

HiveQL Integration

pluralsight

Demo: Cluster Setup

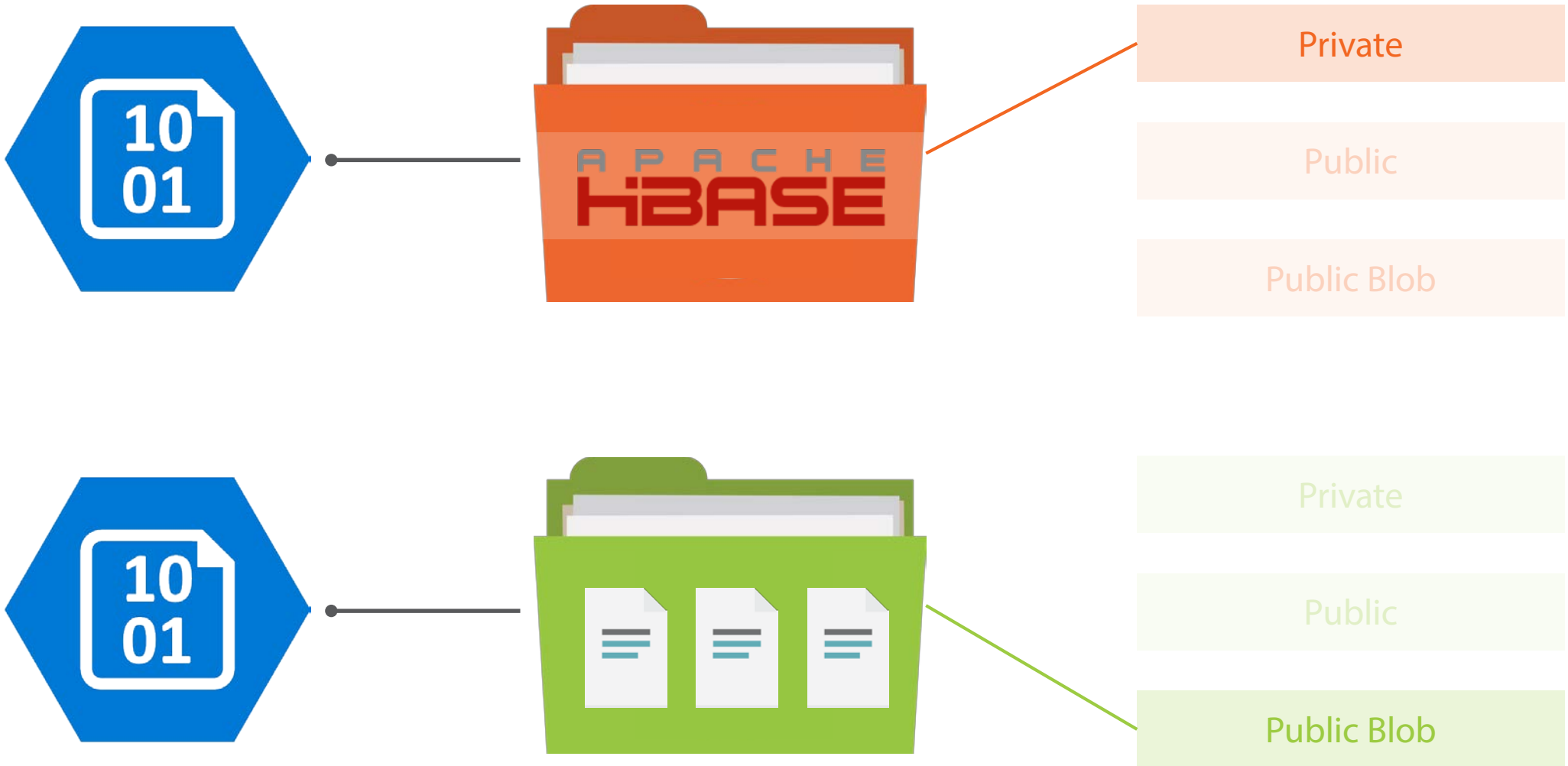
Additional Storage Accounts

Hive Metastore Setup

PowerShell Script

pluralsight

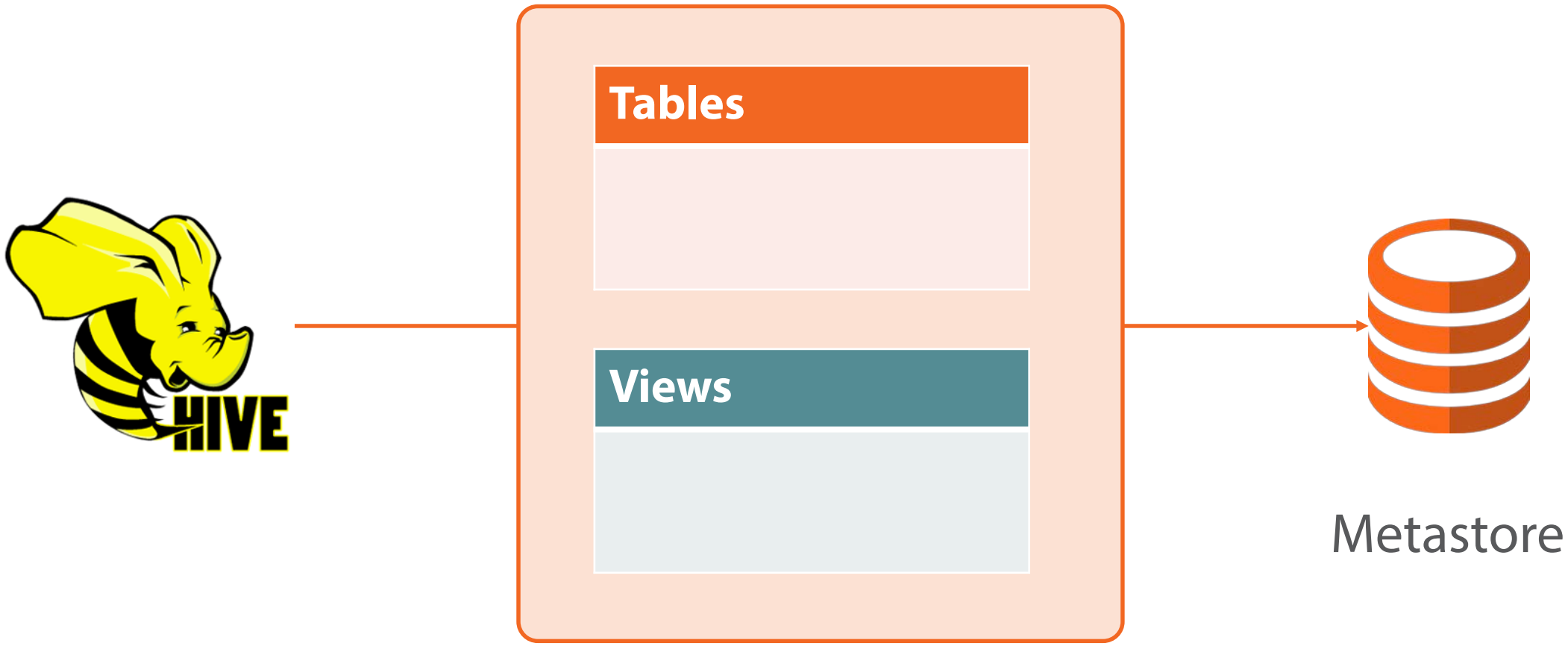
pluralsight



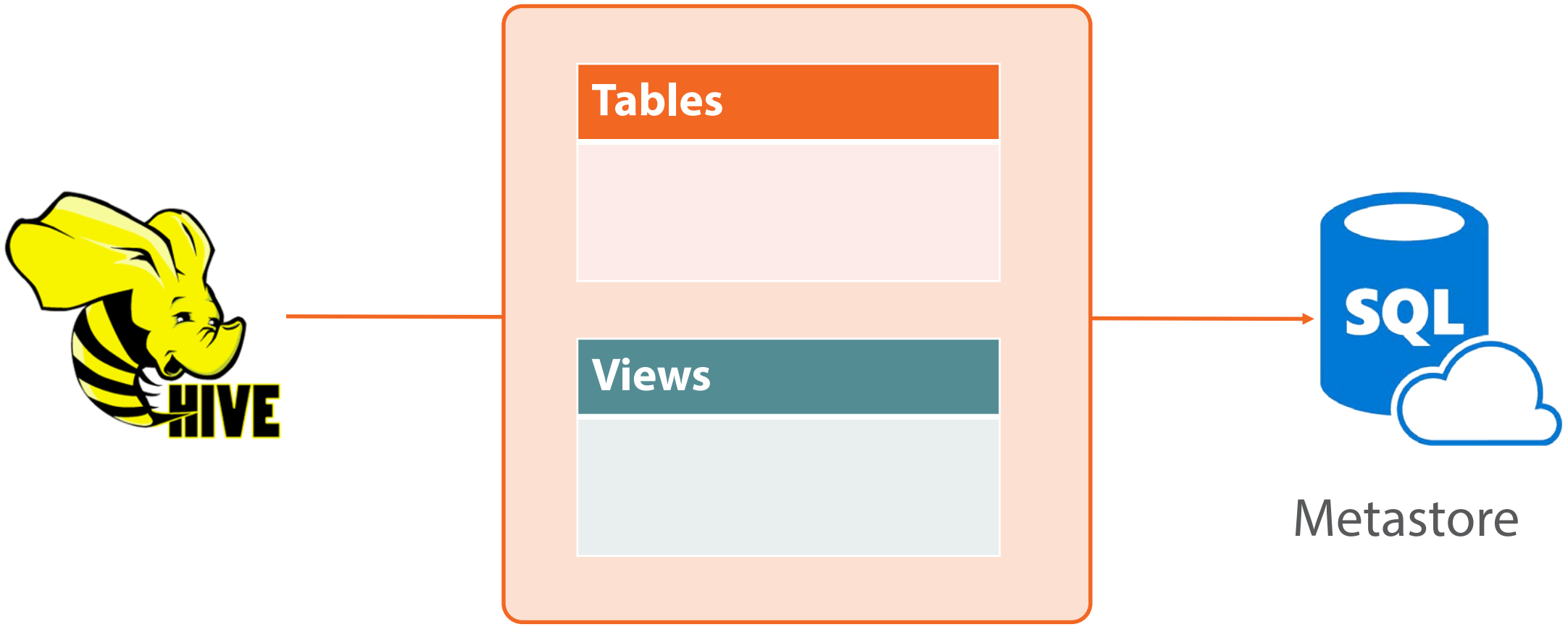
pluralsight

http://**timing.sixeyed.com**/**public**/race_results/**raceId=e399**

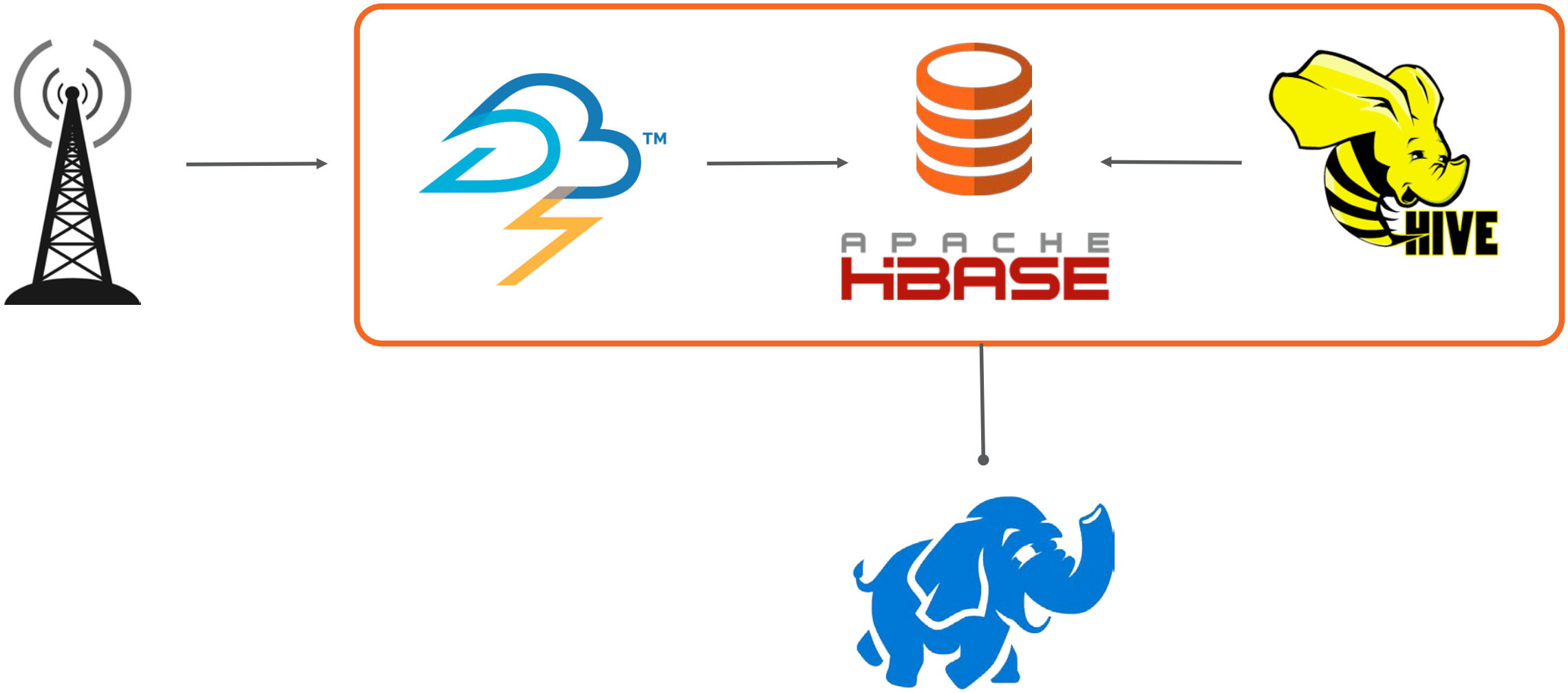
pluralsight

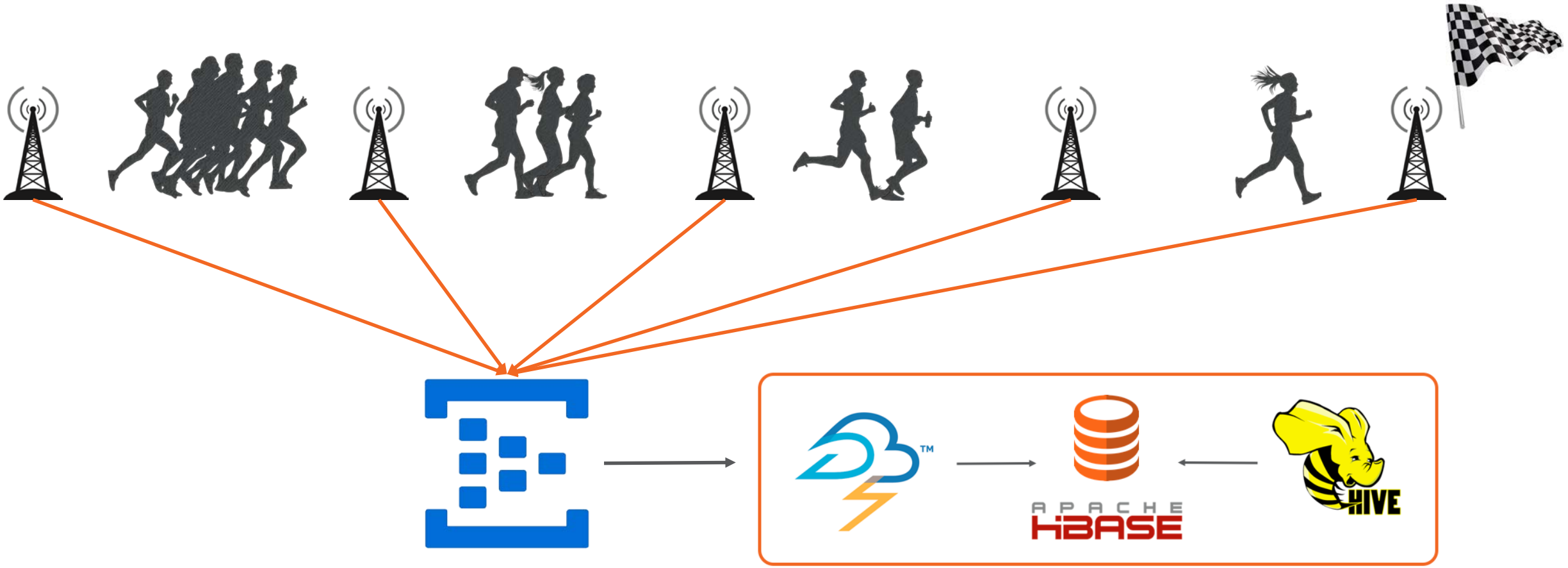


pluralsight

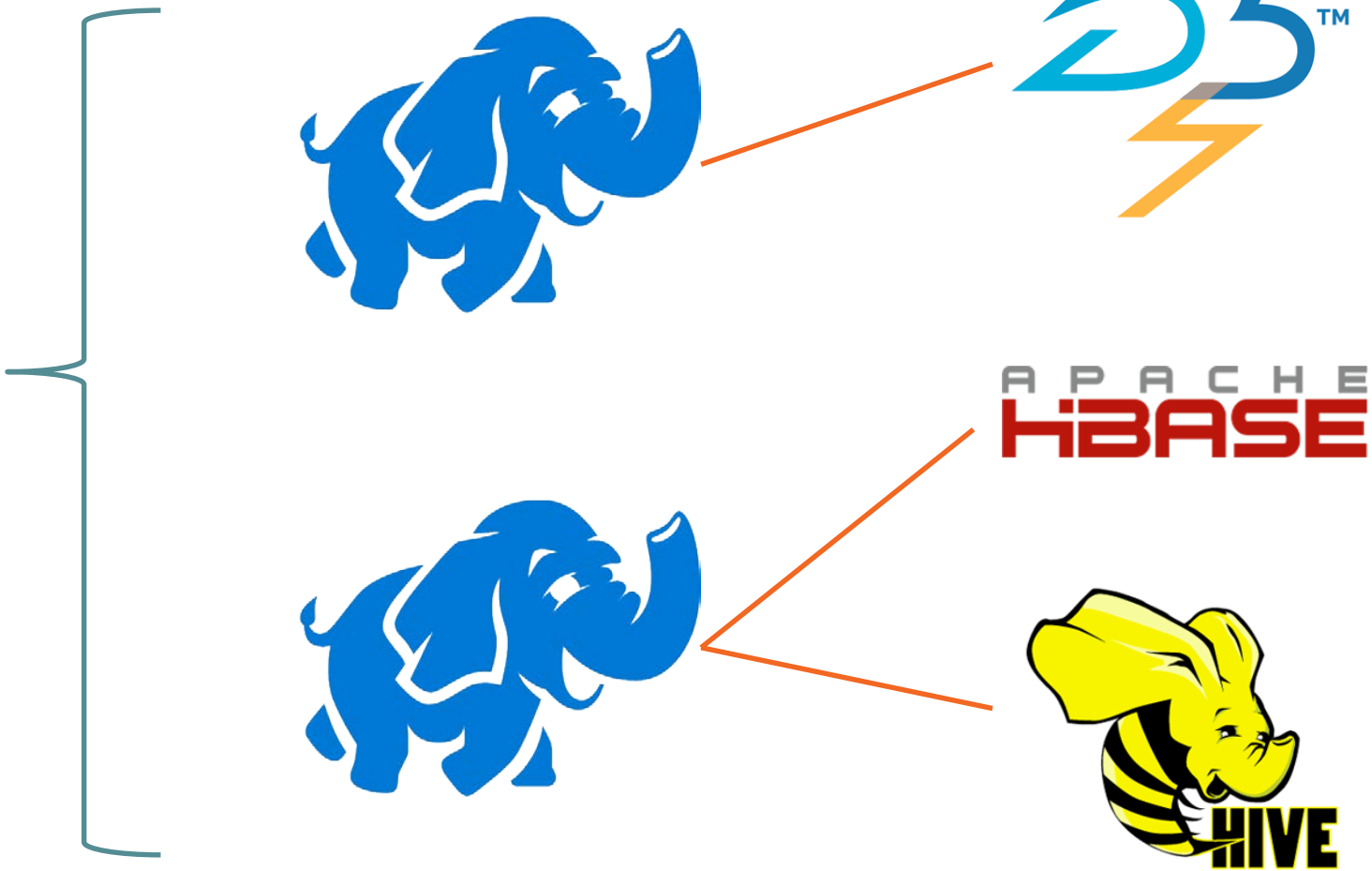


pluralsight

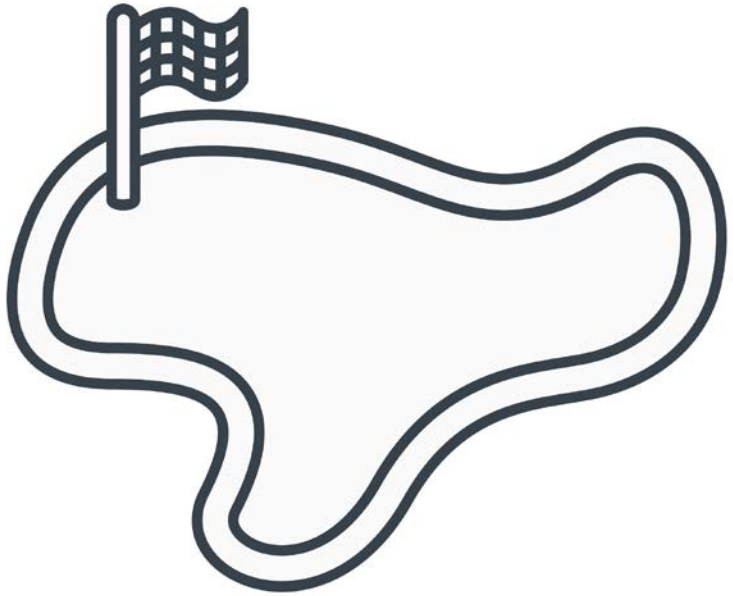

APACHE
HBASE
HIVE

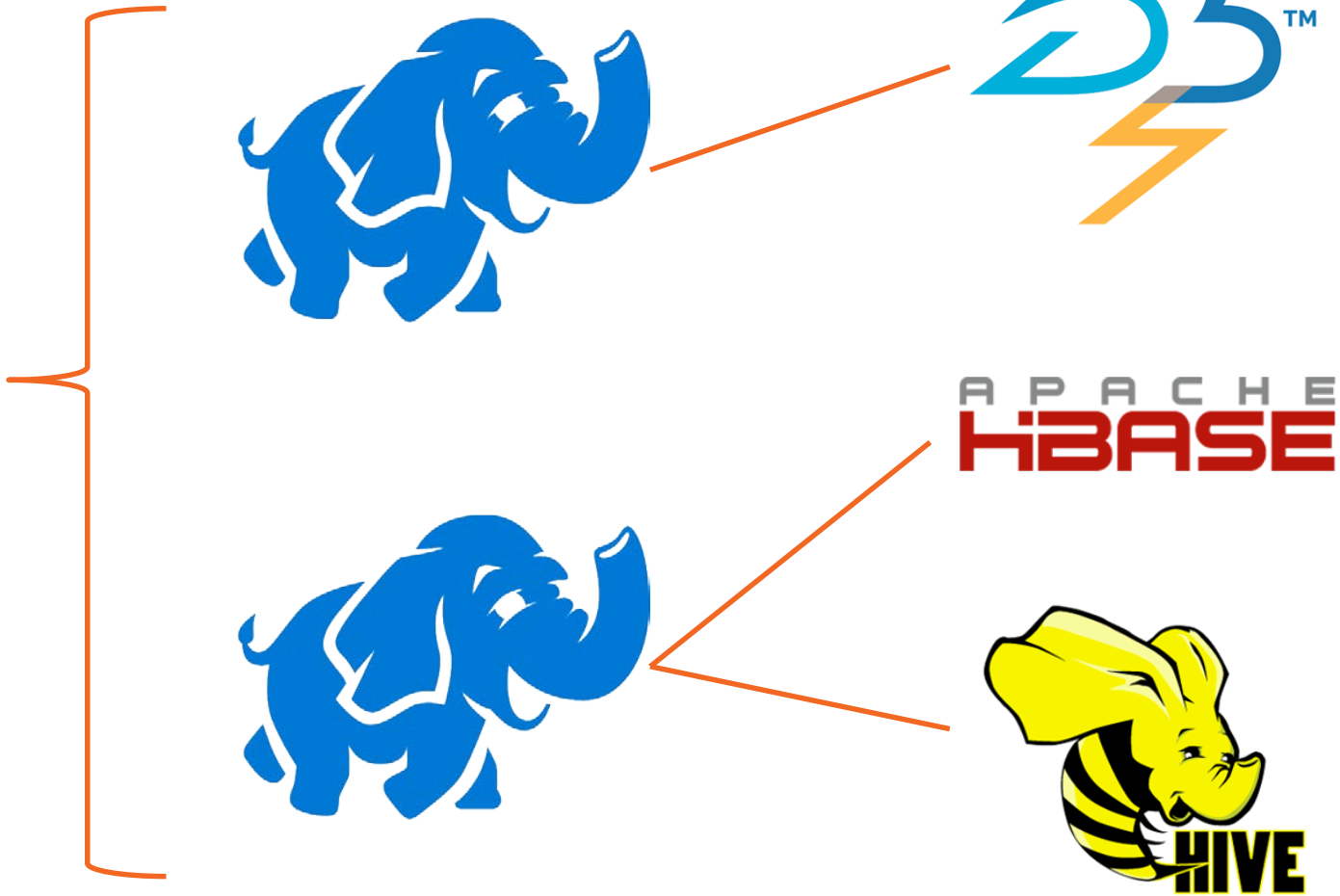


pluralsight

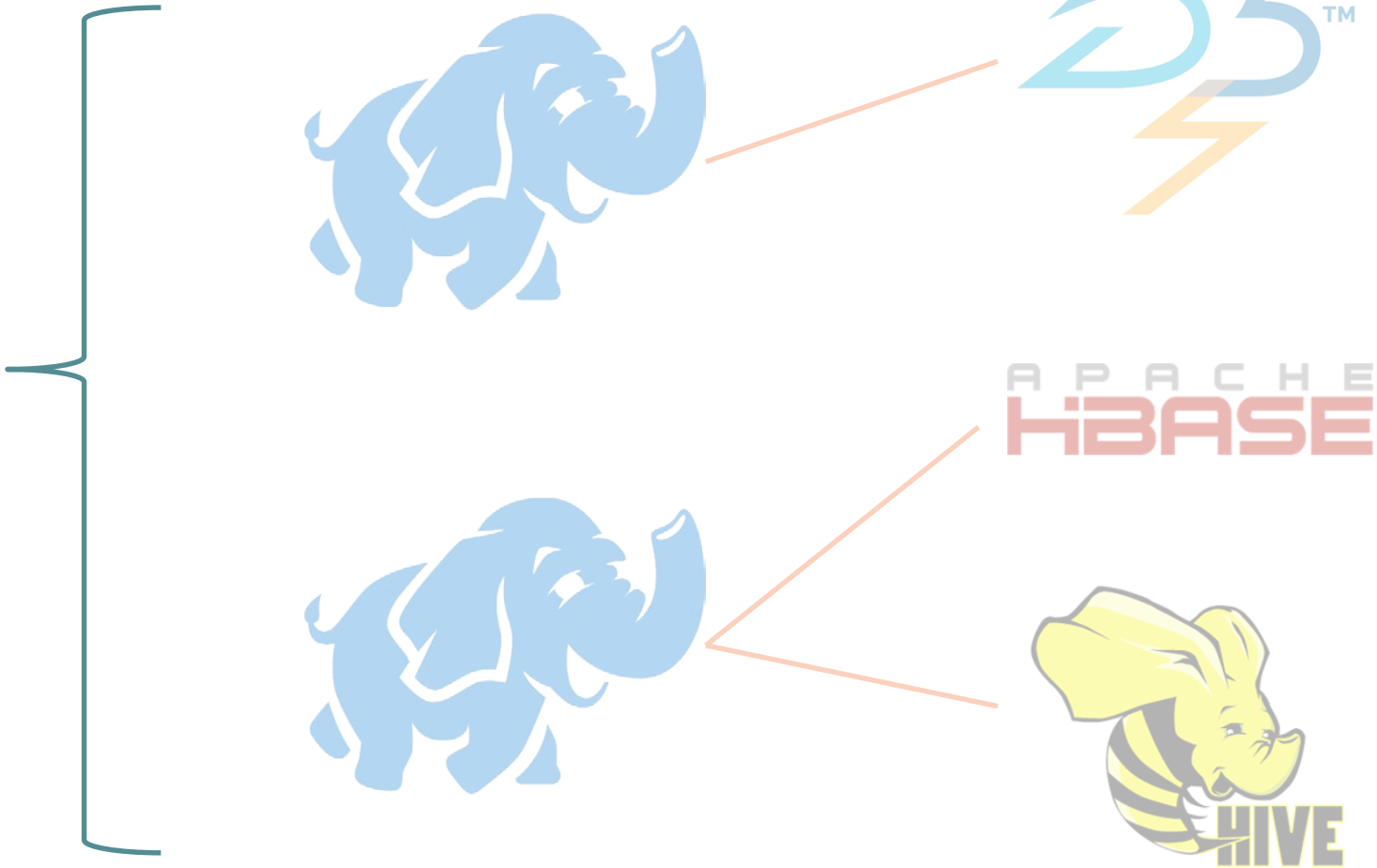


pluralsight

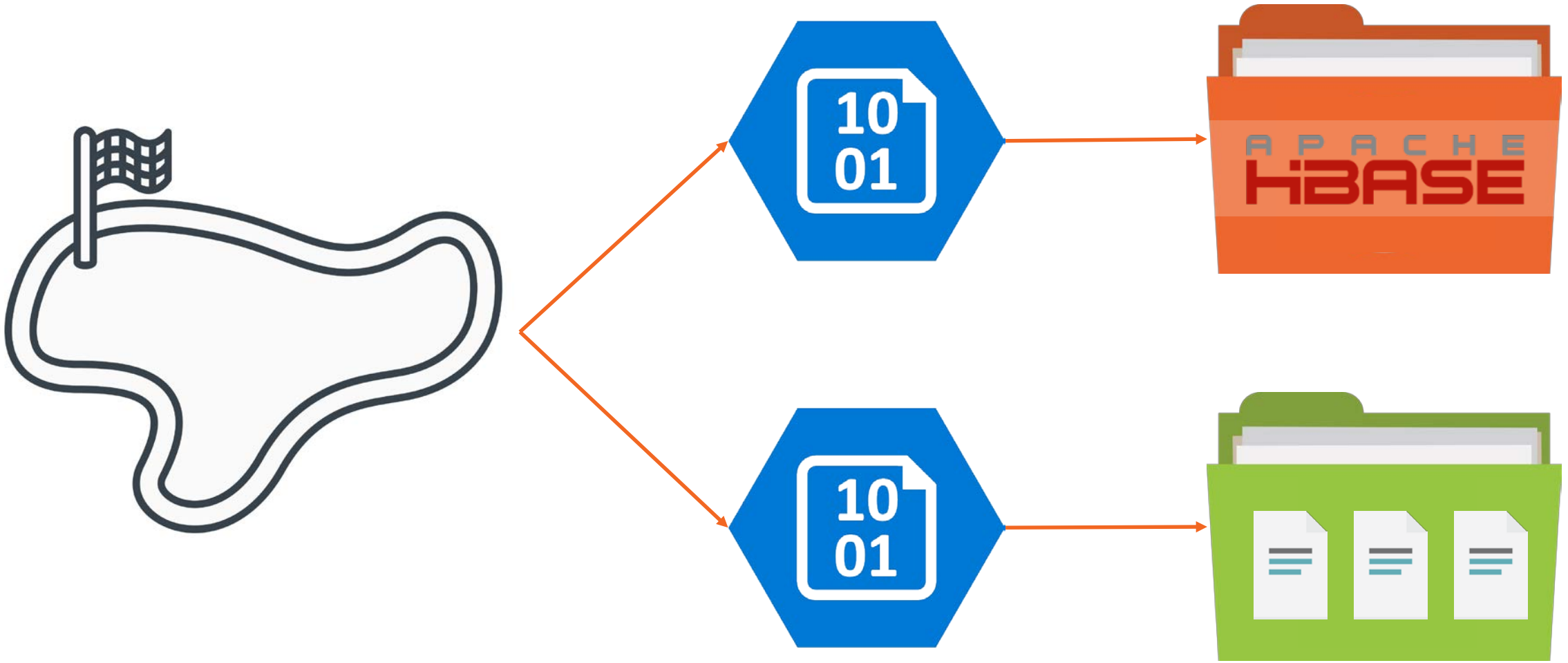




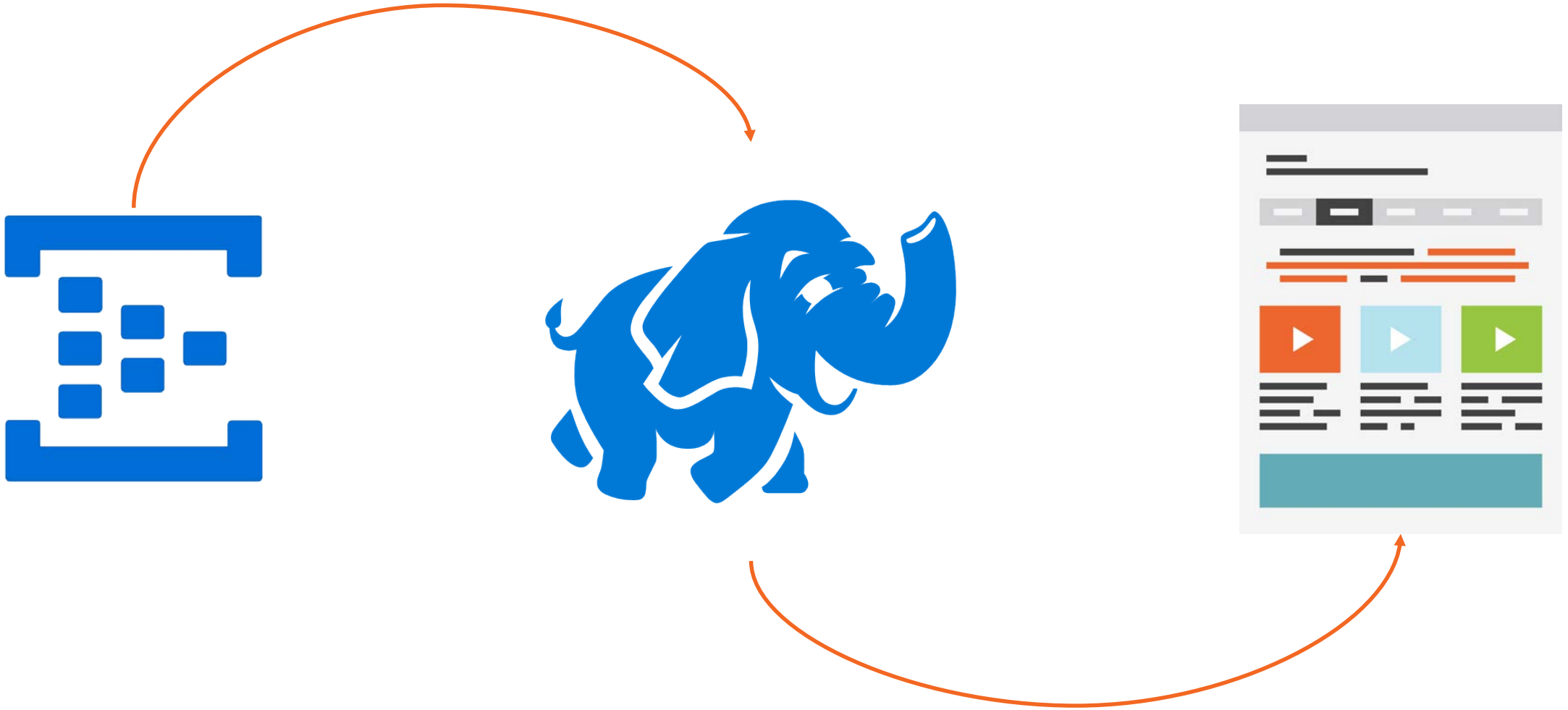
pluralsight



pluralsight

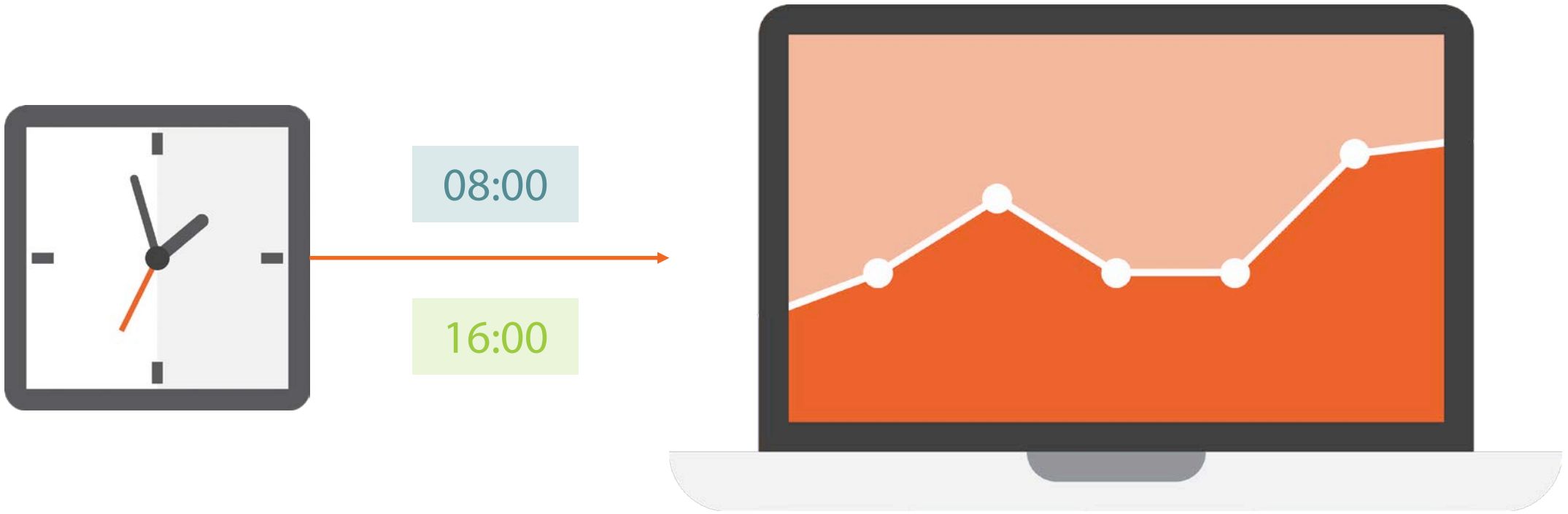


pluralsight

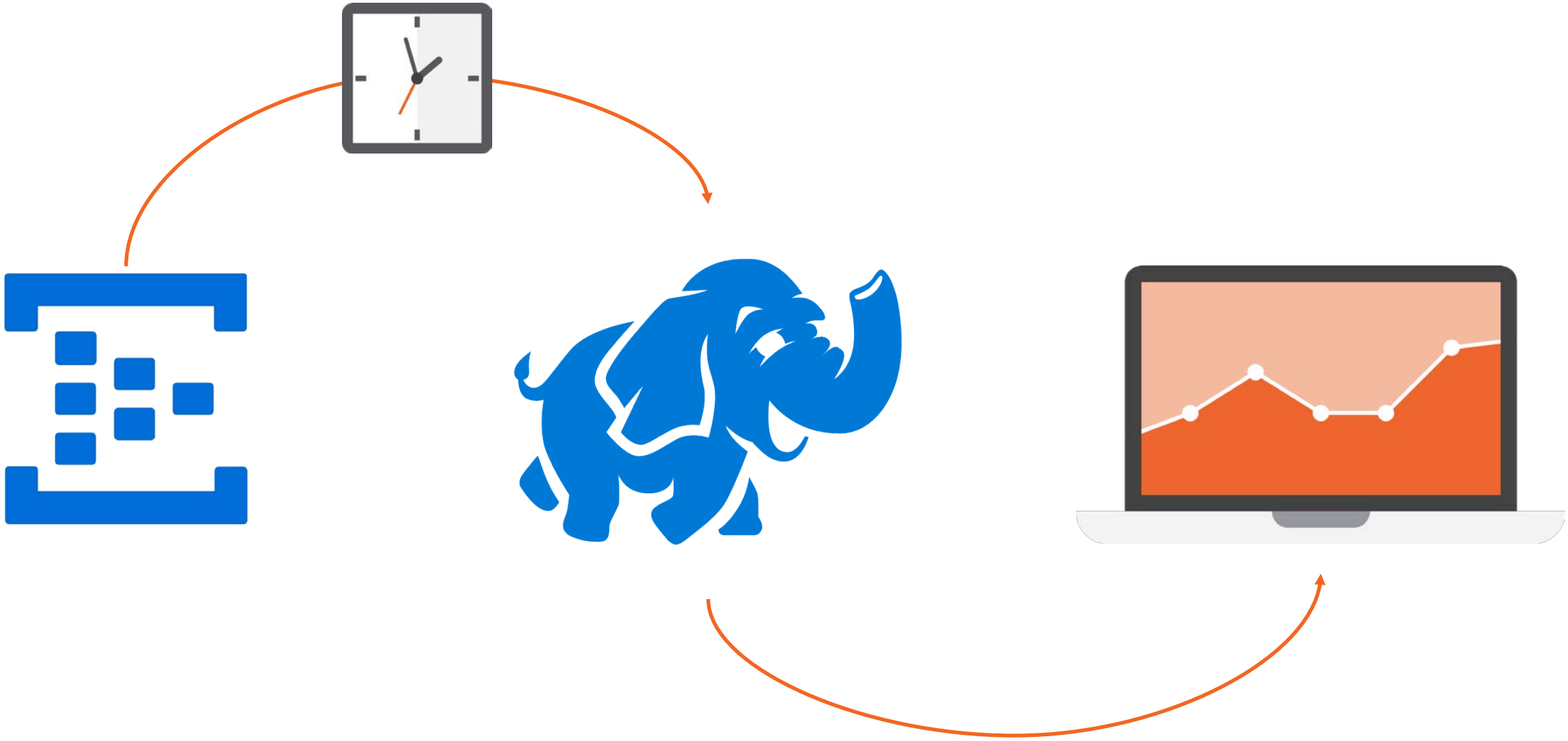


pluralsight

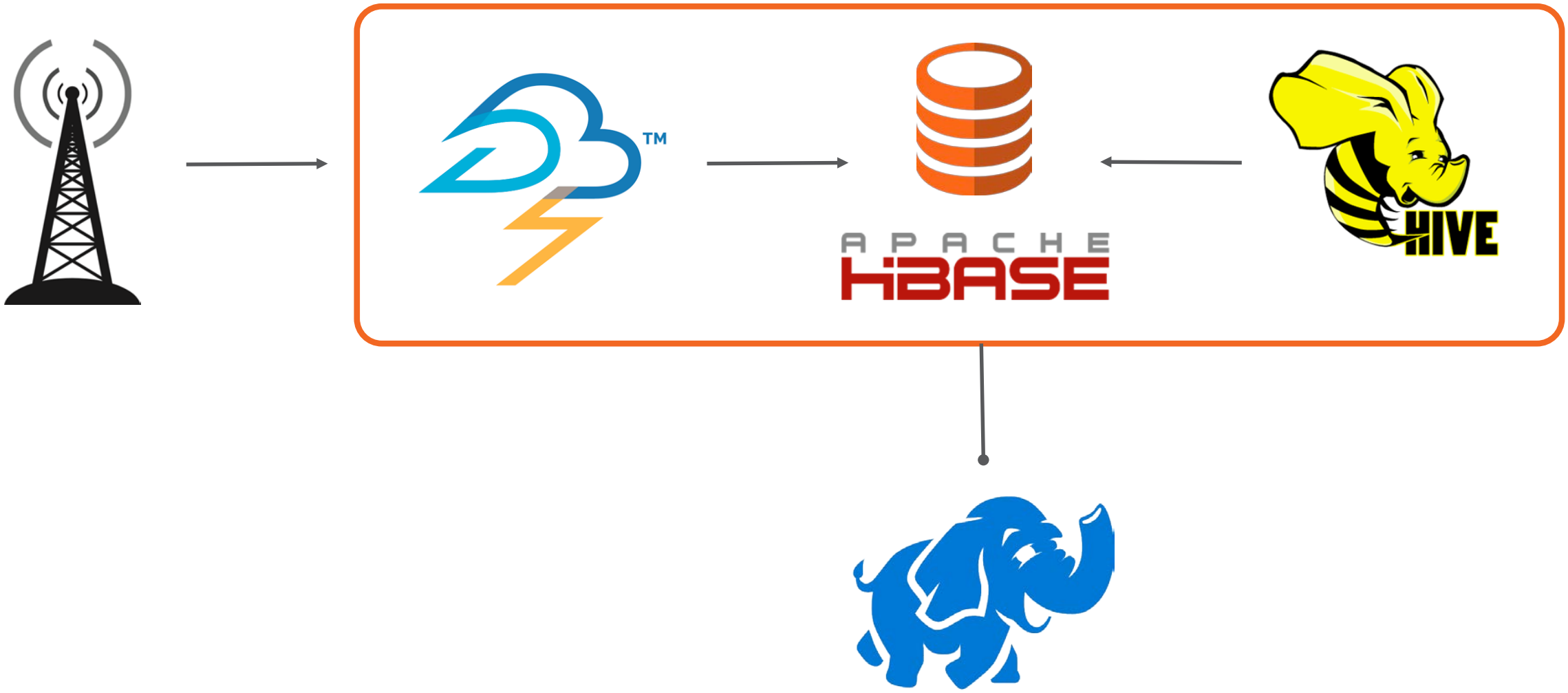
pluralsight



pluralsight



pluralsight


APACHE
HBASE
HIVE