Big data: An architecture for the real-time analysis

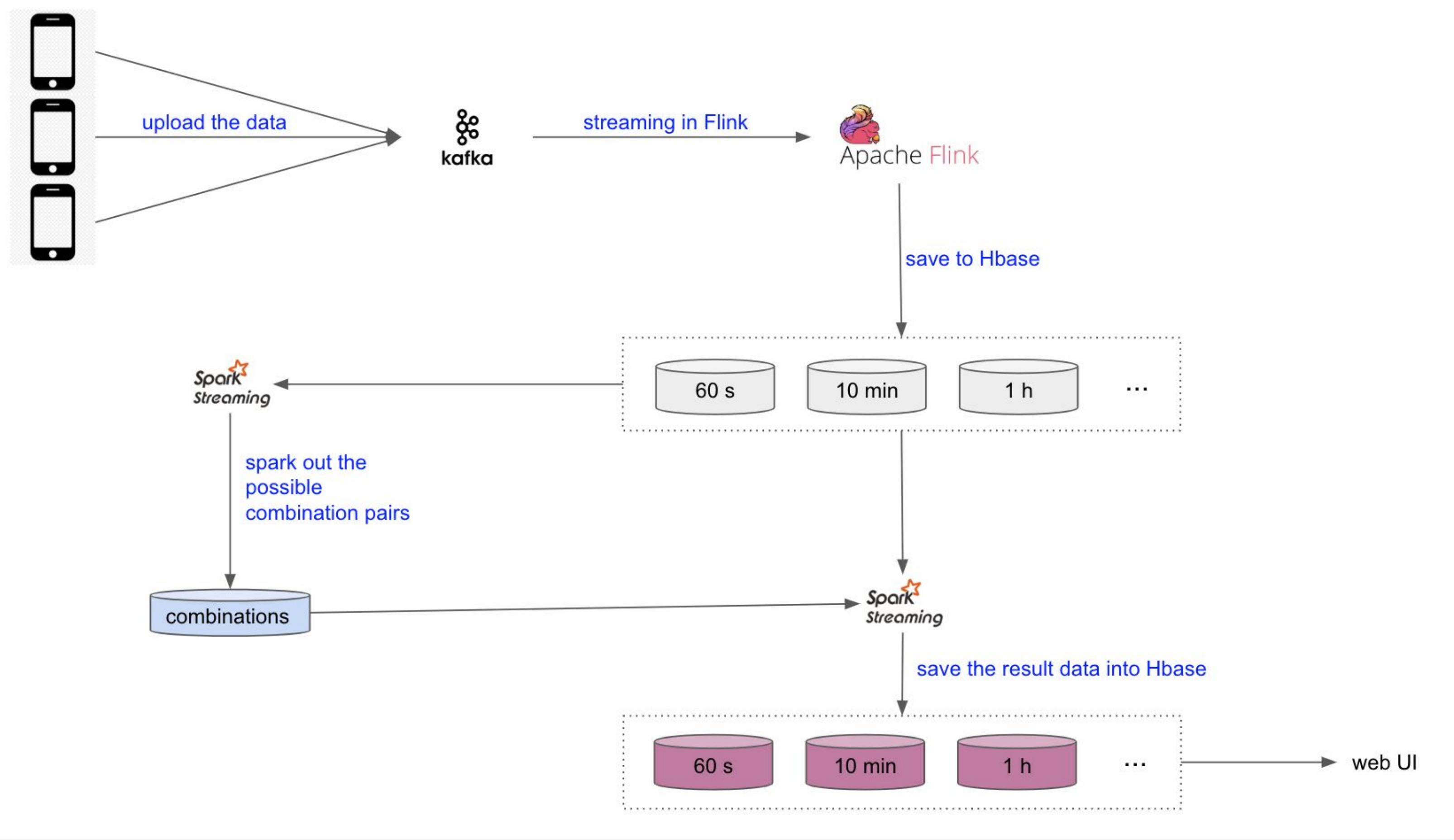
wshen24@asu.edu

Purpose

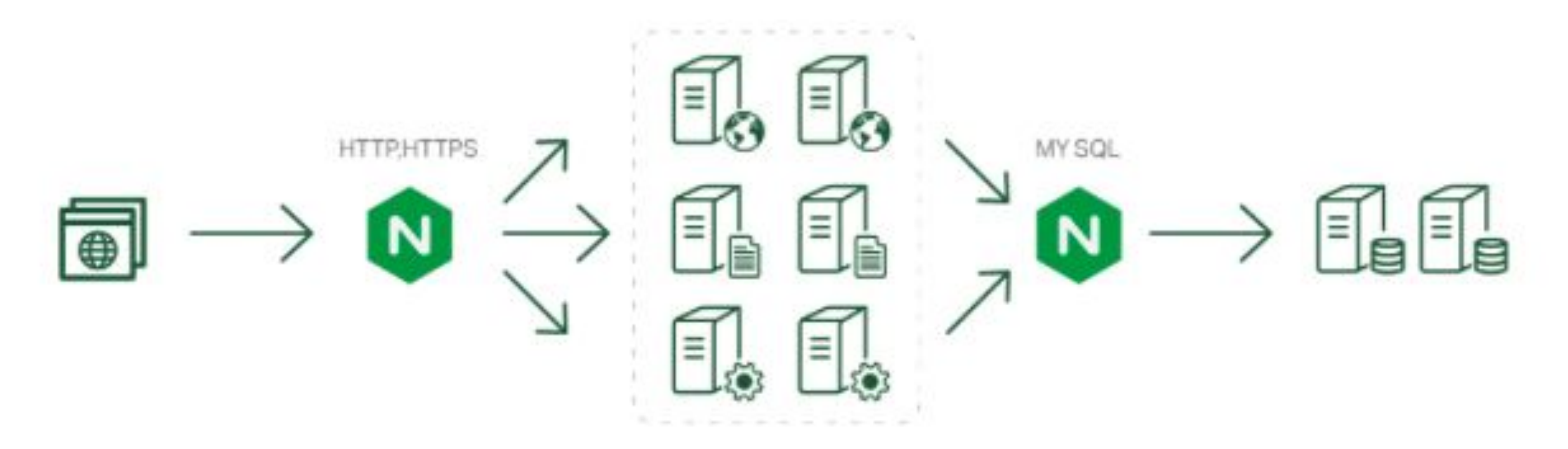
Get the number of particular users (location, mobile or age etc) in a reasonable time

easy?

Overall architecture: Hbase cluster, Kafka cluster, Flink Cluster, Spark cluster, Python server ...



Old-fashioned architecture



However, what happens if the page review speed at 1 million/s?

What happens if we want to know more detailed information, when, where and how etc.

What can we know those information in reasonable real time?

How can we get the page views during a period of time?

```
SELECT COUNT(*) FROM table.logs WHERE time = ‘2019-01-20’
```

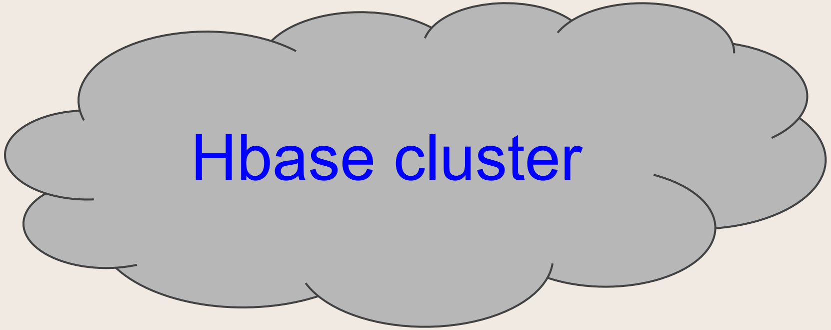
If more details, visitor from AZ and using iPhone?

```
SELECT COUNT(*) FROM table.logs WHERE time = ‘2019-01-20’ AND location = ‘LA’ AND device = ‘iPhone’
```

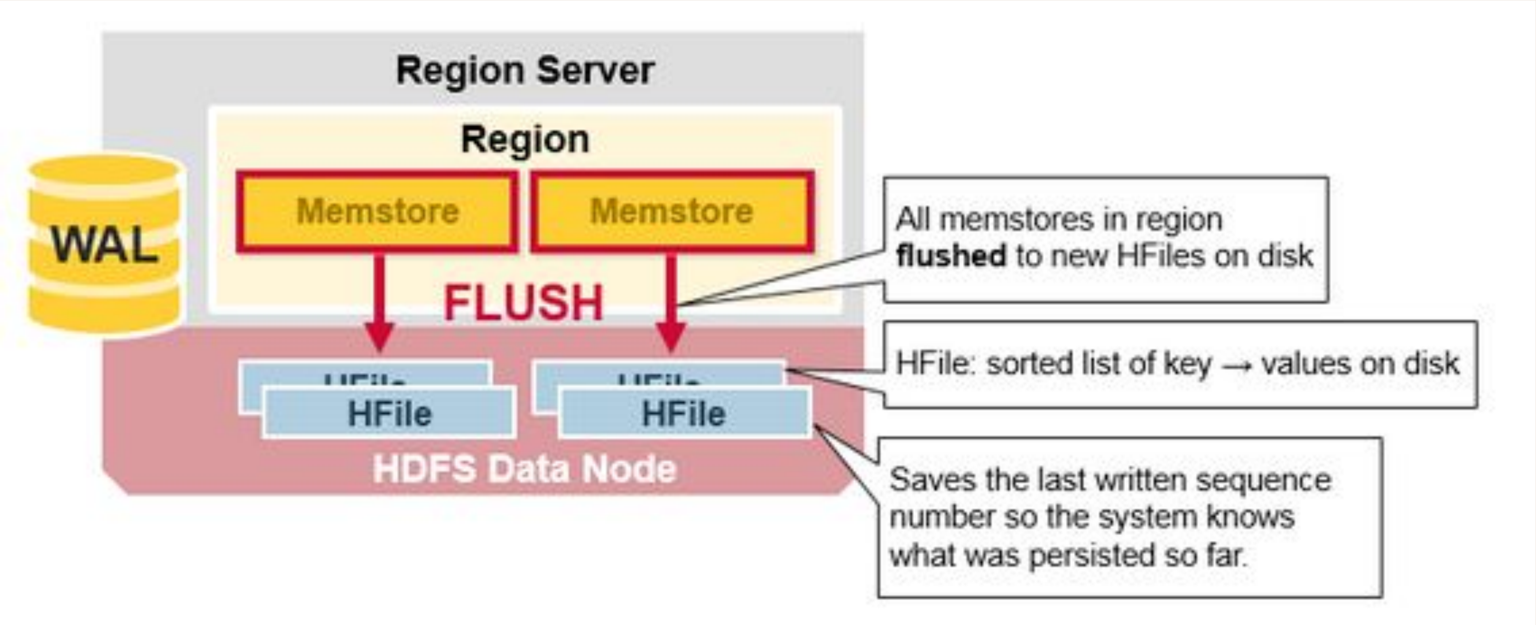
It is not always the case when data is over 1 billions even 1, 000 billions

Storage

1,000, 000, 000 new messages every day !!!



WAL: Write-Ahead-Log



All memstores in region **flushed** to new HFiles on disk

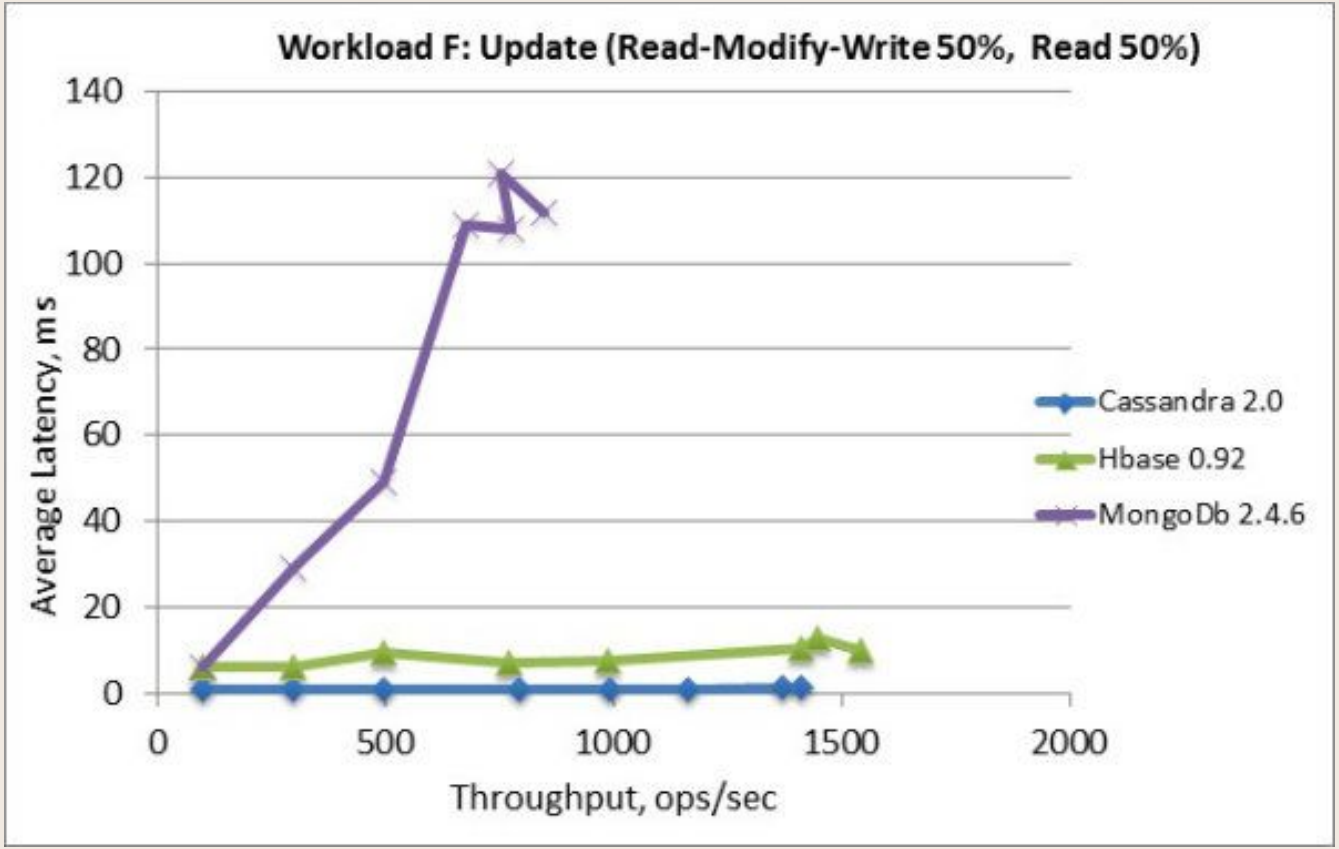
HFile: sorted list of key → values on disk

Saves the last written sequence number so the system knows what was persisted so far.

Great Performance on Writing with huge data !!!

Reference: http://www.larsgeorge.com/2010/01/hbase-architecture-101-write-ahead-log.html

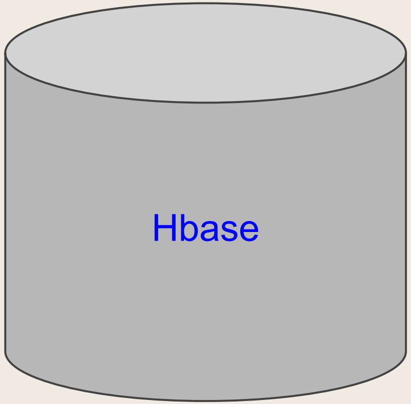
Read performance



Key-Value !!!
Calibrate keys

The design of key

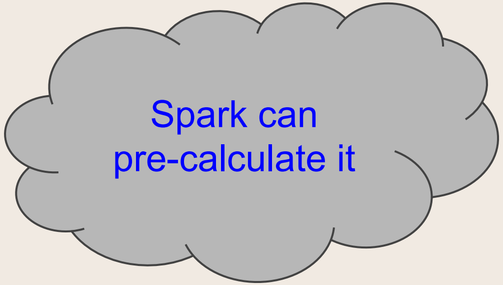
SELECT COUNT(*) FROM table.logs WHERE time = ‘2019-01-20’ AND location = ‘LA’ AND device = ‘iPhone’



Key:
{cluster_id}#{date}#{app_id}#{location->LA}#{device->iPhone}

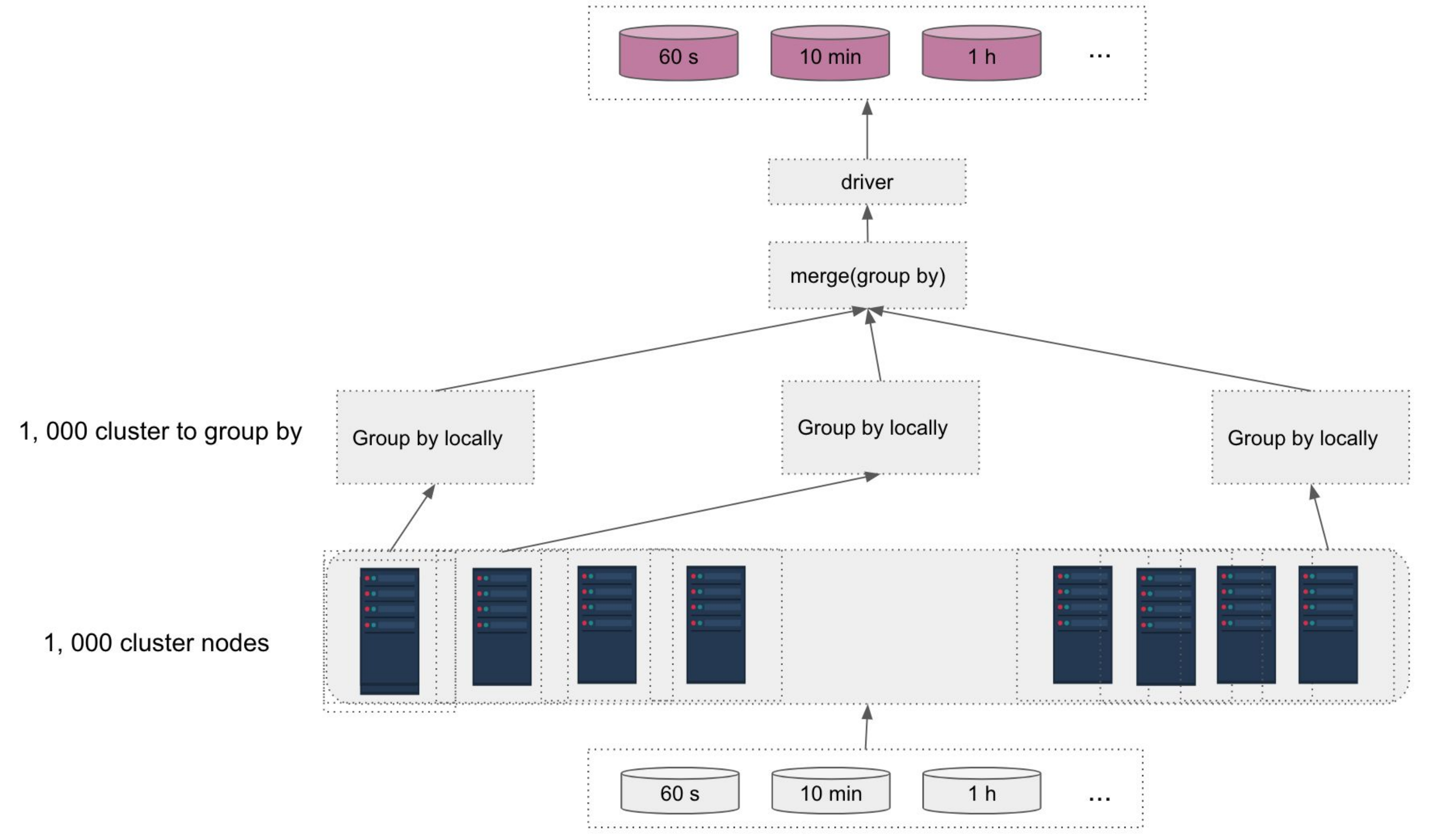
Value:
Integer

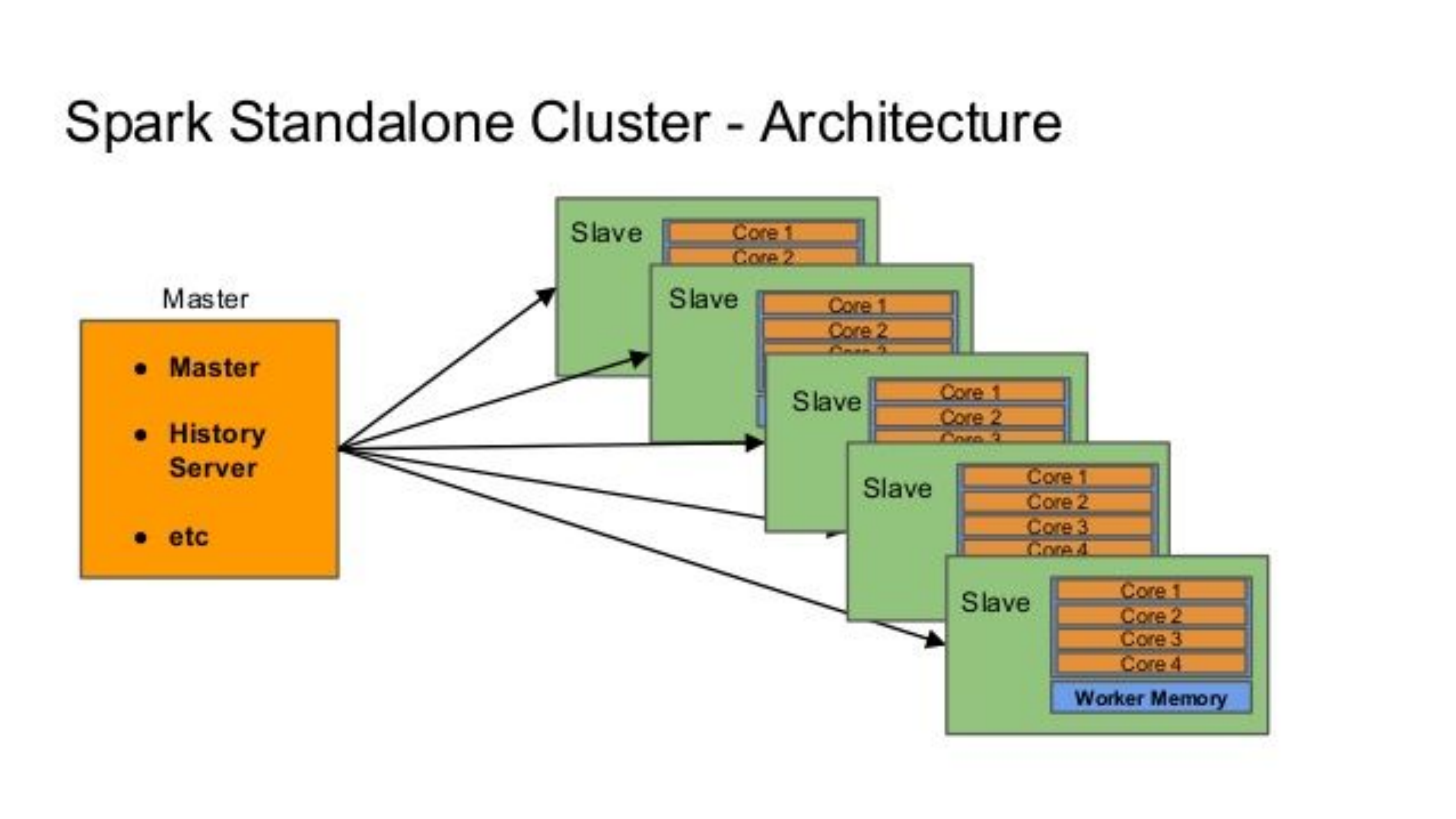
How to get that Key and Value?


Spark can
pre-calculate it

Reference: http://hbase.apache.org/book.html#rowkey.design

Spark flow


60 s
10 min
1 h
...
driver
merge(group by)
1, 000 cluster to group by
Group by locally
Group by locally
Group by locally
1, 000 cluster nodes
60 s
10 min
1 h
...



Spark Standalone Cluster - Architecture

Congratulations! It’s a patent now!

End