BEAM

Apache Beam at scale using **Apache Spark**

David Moravek
Lead Software Engineer @ Seznam.cz

BERLIN 2019

Internet Obrázky Videa Články Firmy Mapy Zboží Slovník

SEZNAM.CZ

berlin

Berlín – Wikipedie

https://cs.wikipedia.org/wiki/**Berlín**

Hlavním městem Německa se stal roku 1991 a od sjednocení Německa (a tím i obou částí města) **Berlín** patří k největším městům v Evropě a je druhým největším městem Evropské unie.

Všeobecný přehled · Dějiny města · Obyvatelstvo · Politické zřízení

Berlin › Obrázky.cz

Další obrázky ›

Berlin › Články.seznam.cz

Freshlabels míří na západ. Nový obchod s batohy v Berlíně navrhla Lenka Míková | PROČ NE?!

Před 4 dny

procne.ihned.cz

V Berlíně vyhlásily mimořádný stav. Příčinou je intezivní déšť | EuroZprávy.cz

Před 4 dny

eurozpravy.cz

Více článků ›

Berlín

Hlavní město

Webová stránka | Navigovat | Tipy v okolí

Berlín je hlavní město a zároveň i spolková země Spolkové republiky Německo. Hlavním městem Německa se stal roku 1991 a od sjednocení Německa… Číst dále

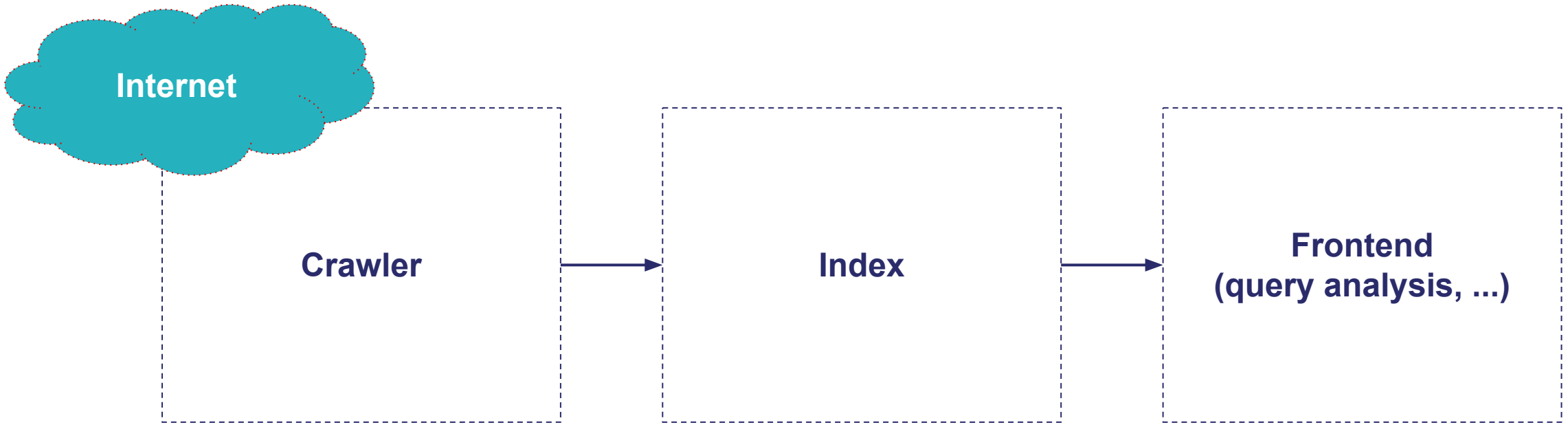
Německo

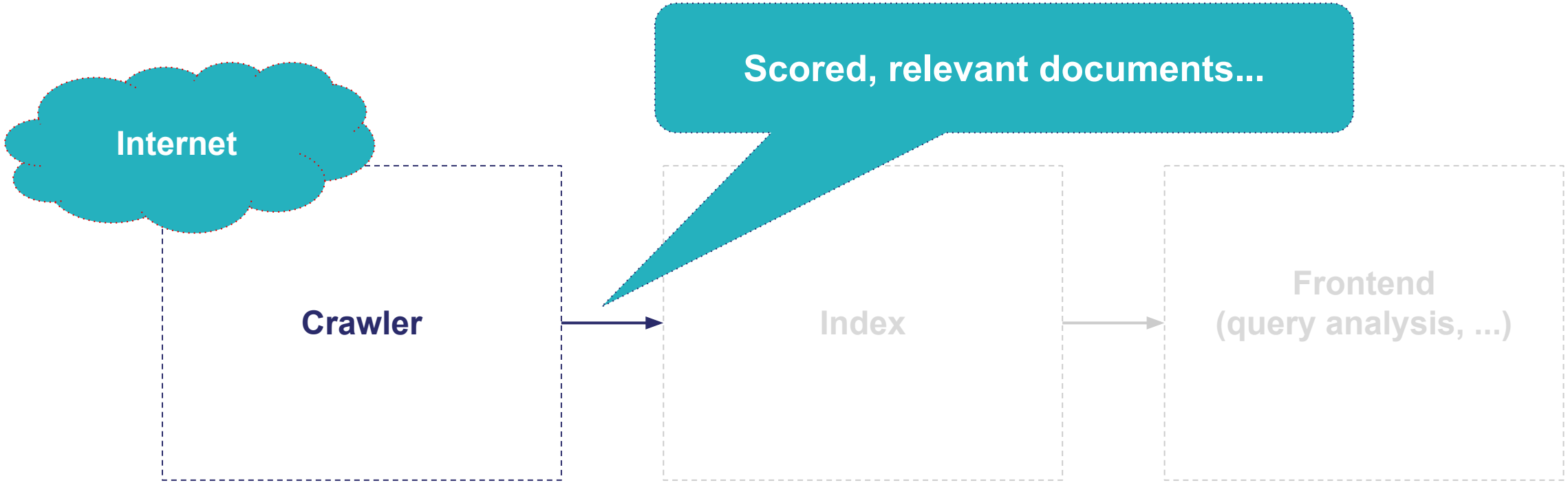
Více informací na Mapy.cz

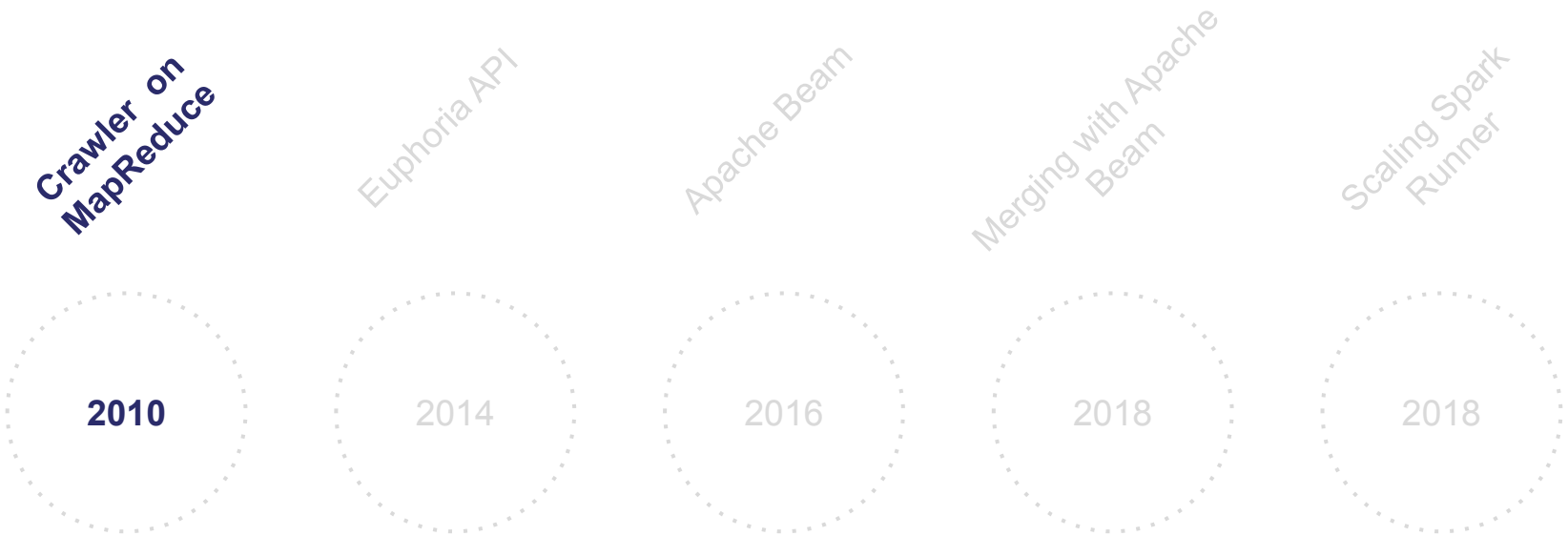
Počet obyvatel:	**3,61 mil.**
Nadmořská výška:	**34 m n. m.**
Rozloha:	**891,12 km²**
Nejvyšší představitel:	Michael Müller

Může se hodit:

cs.wikipedia.org


Internet
Crawler
Index
Frontend
(query analysis, ...)


Scored, relevant documents...
Internet
Crawler
Index
Frontend
(query analysis, ...)


Crawler on MapReduce
2010
Euphoria API
2014
Apache Beam
2016
Merging with Apache Beam
2018
Scaling Spark Runner
2018

BEAM

40,000,000,000 rows in DB

BEAM

40,000,000,000 rows in DB

300,000,000 downloads a day

BEAM

> **1,100** bare-metal servers

BEAM

> **1,100** bare-metal servers

> **13 PB** storage

BEAM

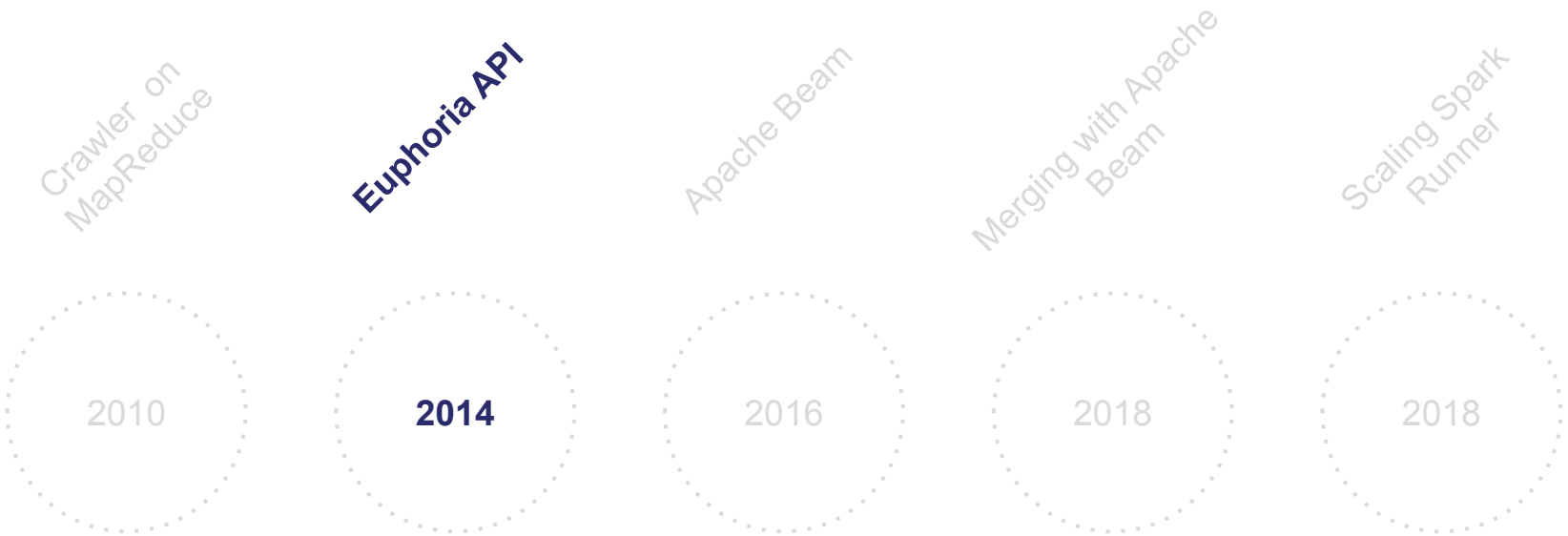
> **1,100** bare-metal servers

> **13 PB** storage

> **50 TB** memory

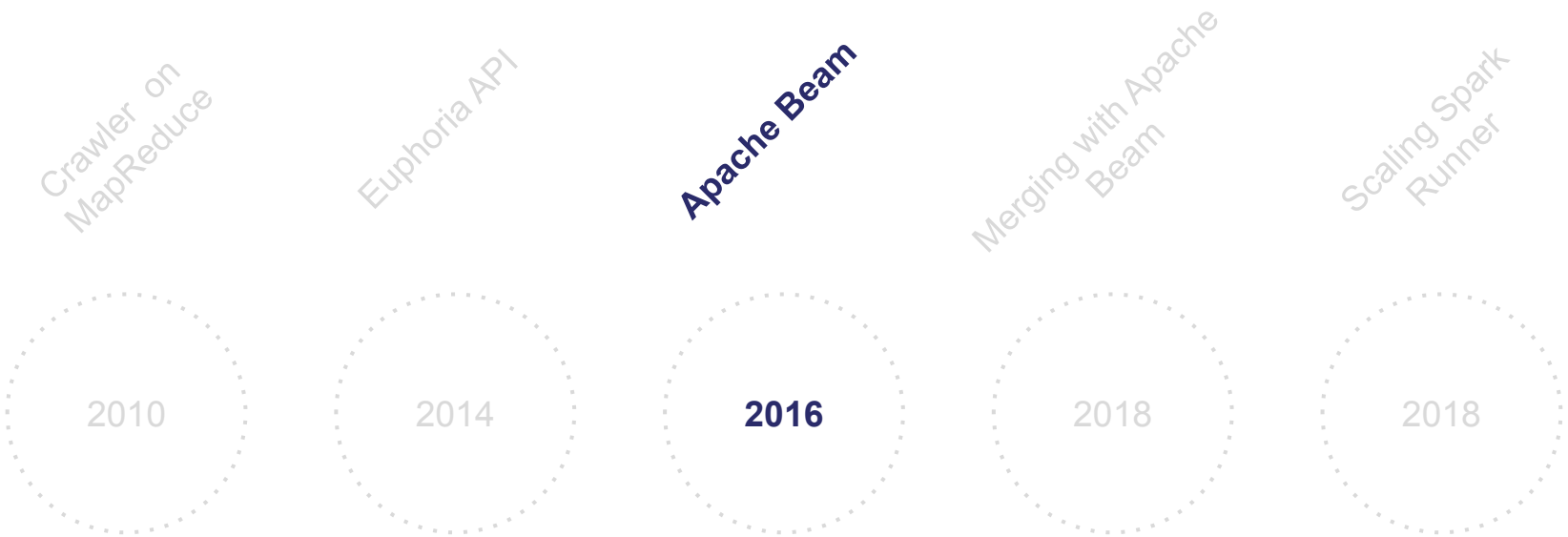
BEAM

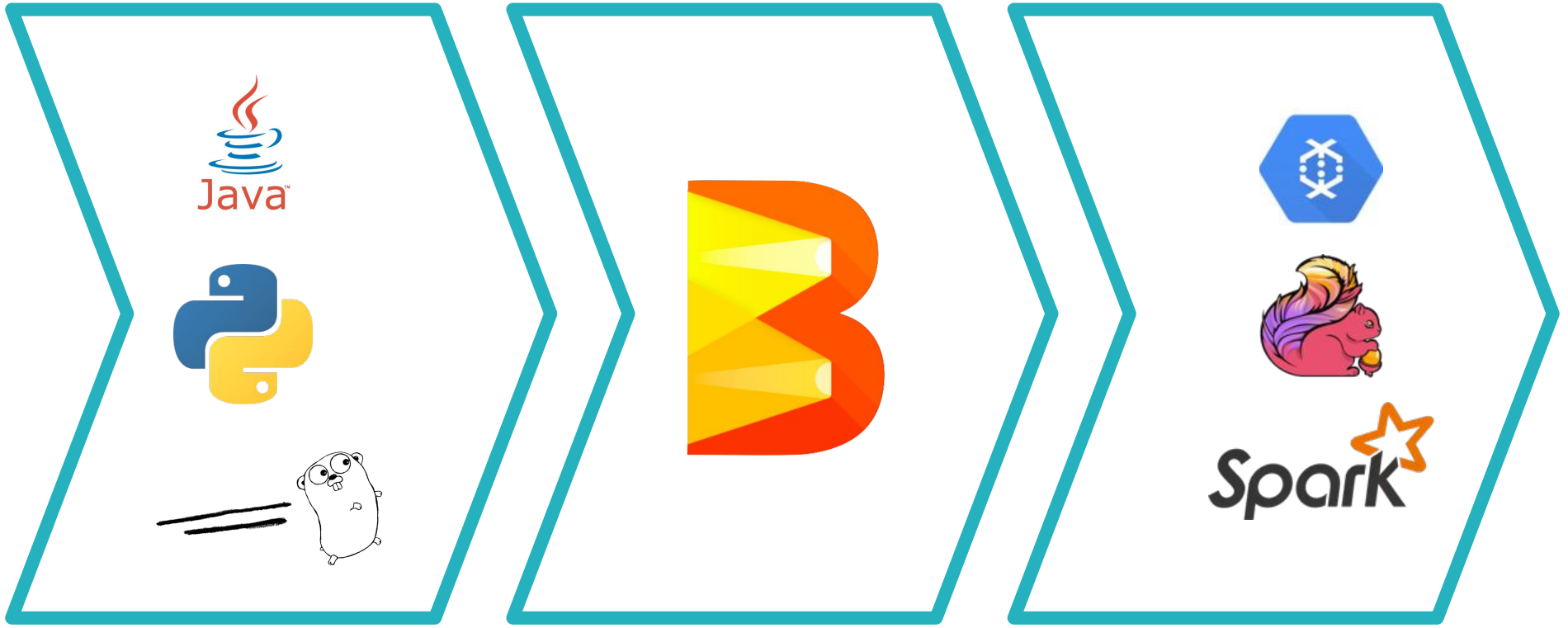
MapReduce is slow and expensive


Crawler on MapReduce
2010
Euphoria API
2014
Apache Beam
2016
Merging with Apache Beam
2018
Scaling Spark Runner
2018

An engine independent programming model which can express both **batch** and **streaming pipelines**.

Java SDK **Runner**


Crawler on MapReduce
Euphoria API
Apache Beam
Merging with Apache Beam
Scaling Spark Runner
2010
2014
2016
2018
2018



SDK

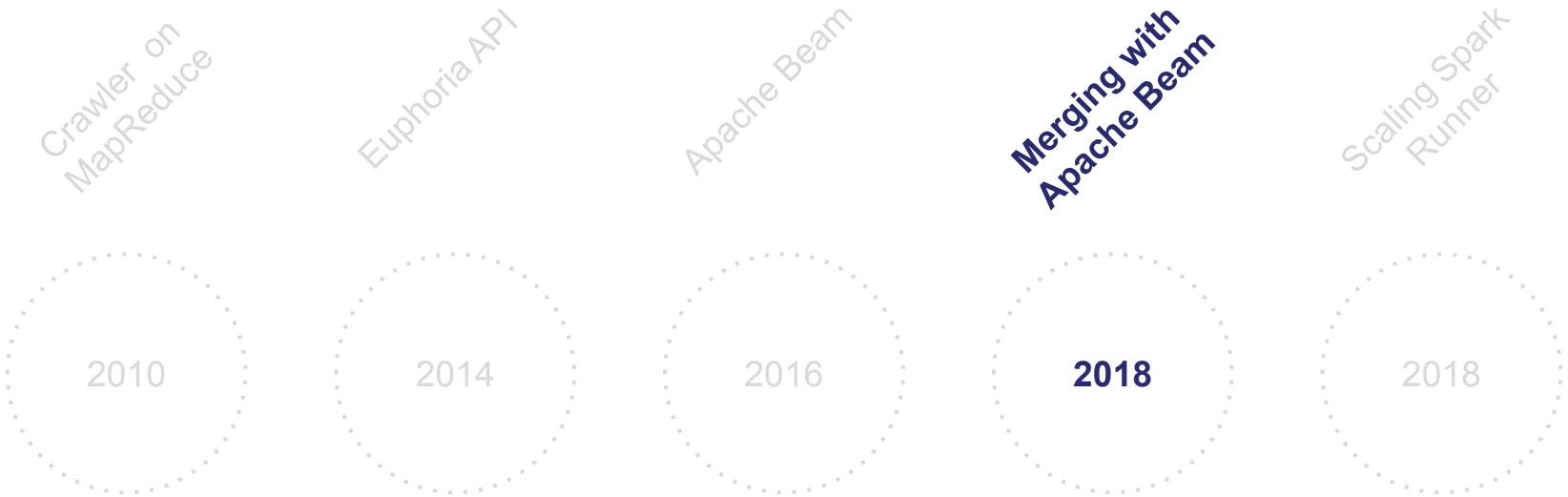
Model

Runner

Python, Java, or Go: It's Your Choice with **Apache Beam**

Tomorrow from **11:00** to **11:40**

Maximilian Michels, Ismaël Mejía


Crawler on MapReduce
Euphoria API
Apache Beam
Merging with Apache Beam
Scaling Spark Runner
2010
2014
2016
2018
2018

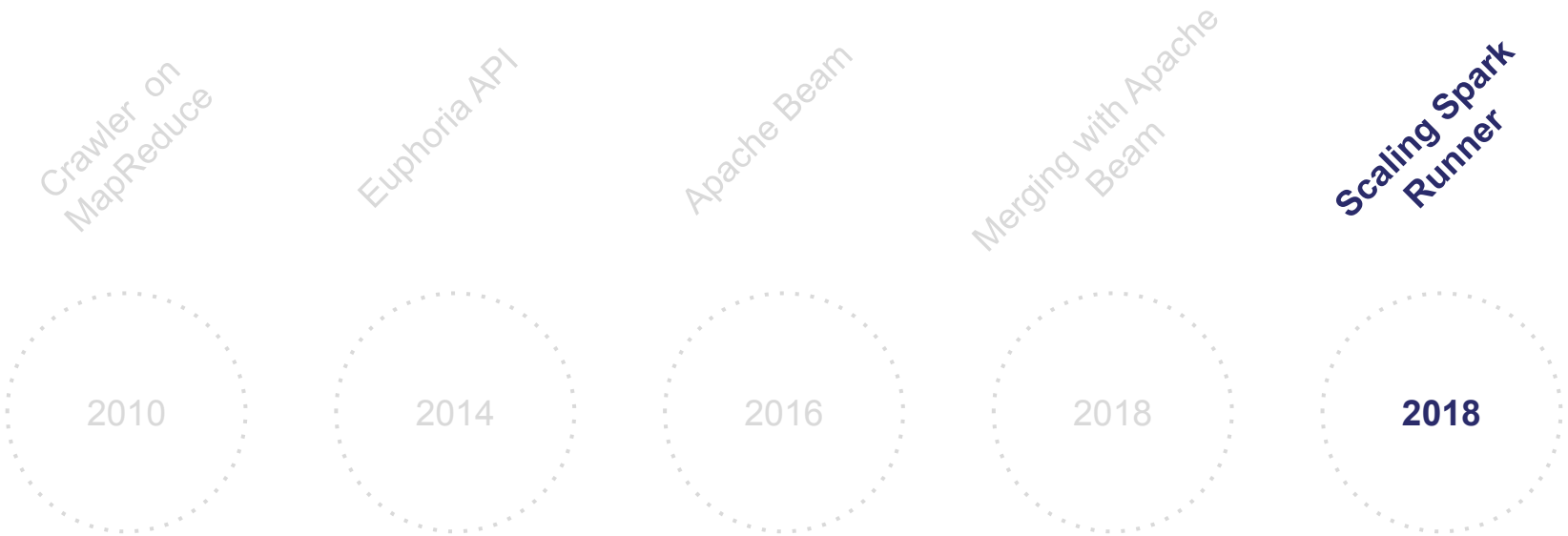
BEAM-3900

beam

BEAM-3900

beam

https://beam.apache.org/documentation/sdks/java/euphoria/


Crawler on MapReduce
Euphoria API
Apache Beam
Merging with Apache Beam
Scaling Spark Runner
2010
2014
2016
2018
2018

BEAM

Turning the **internet upside down**

BEAM

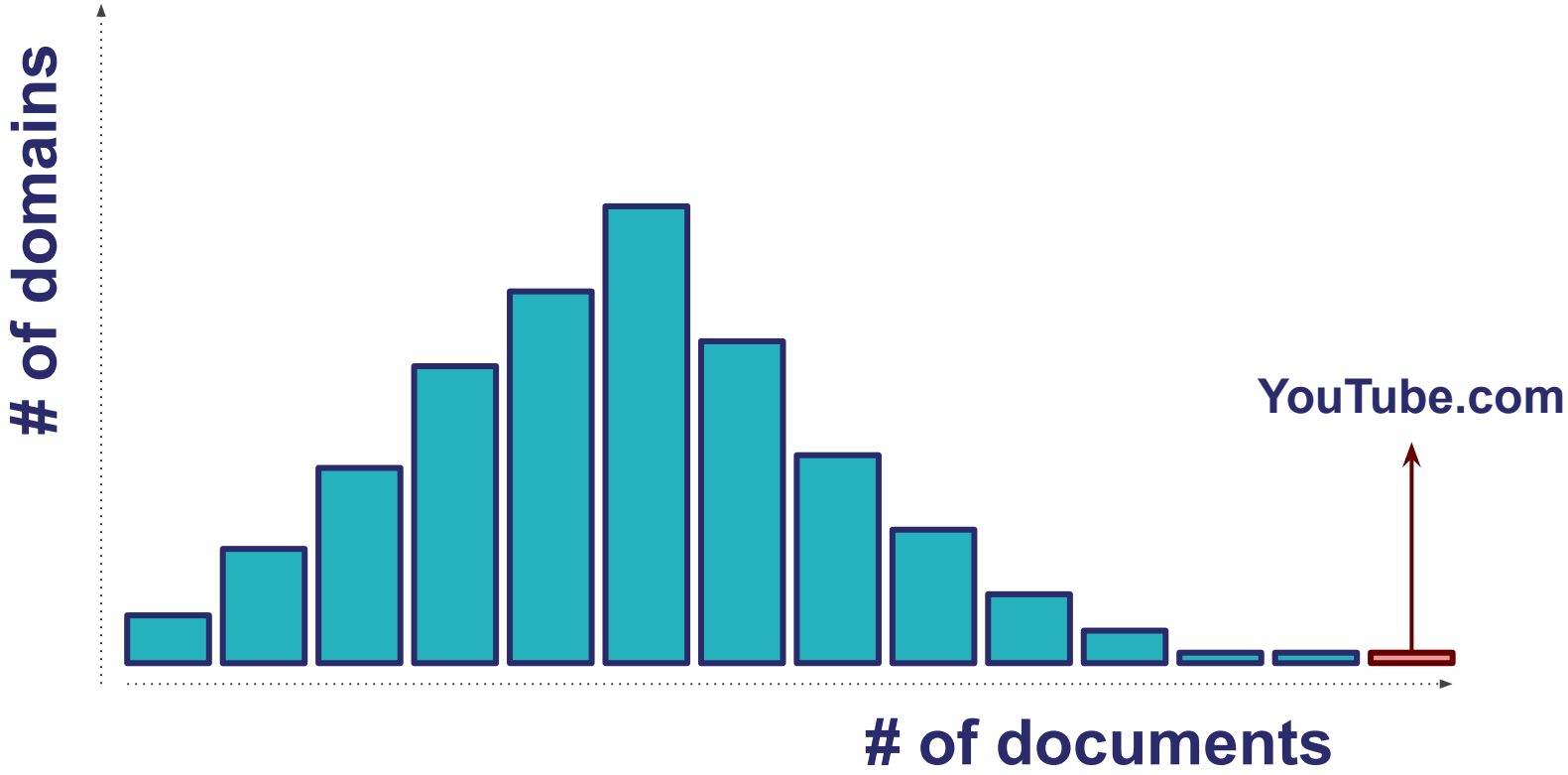
Exponential data skew

EXAMPLE

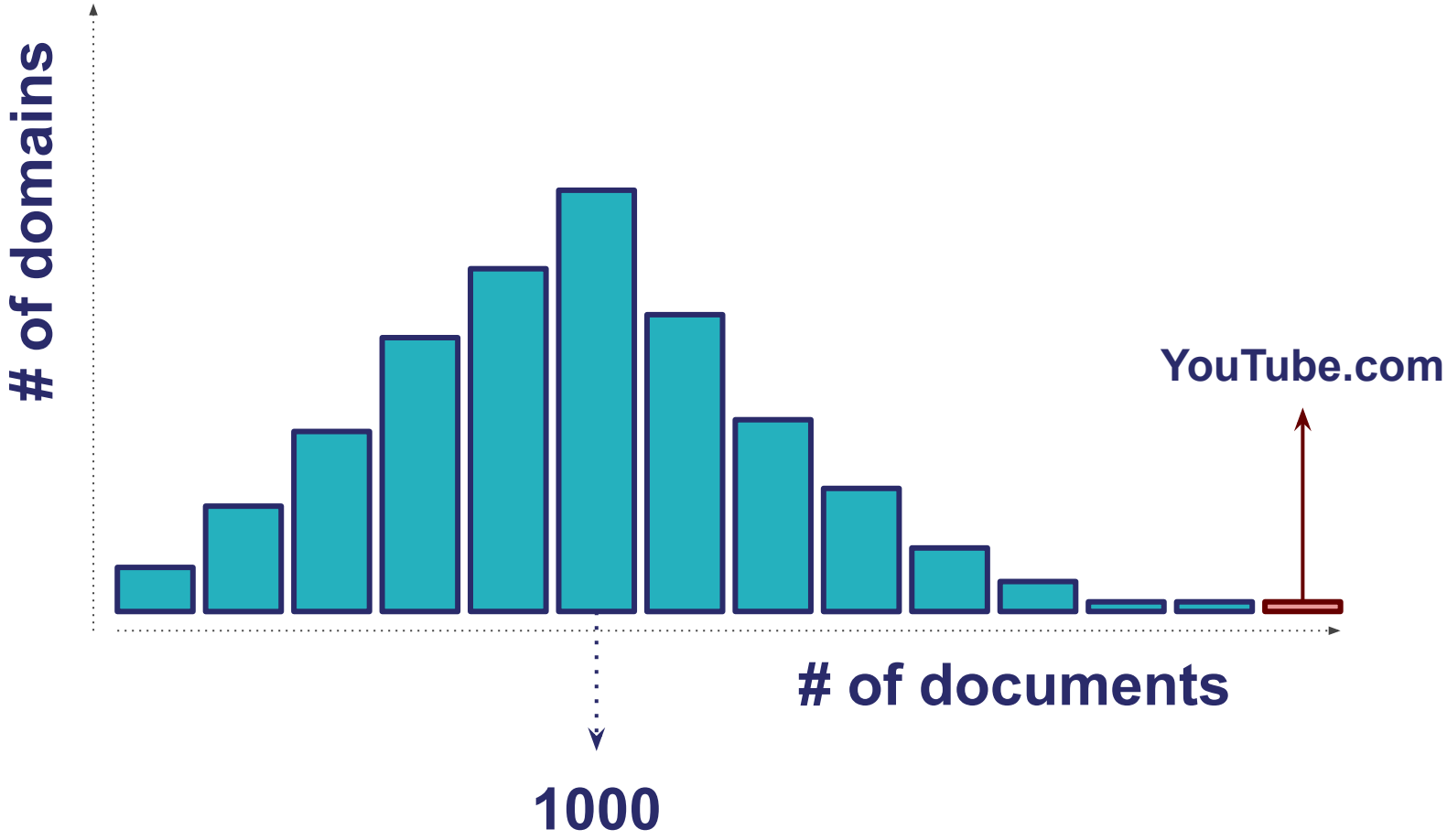
BEAM

joining **documents** with **domains**

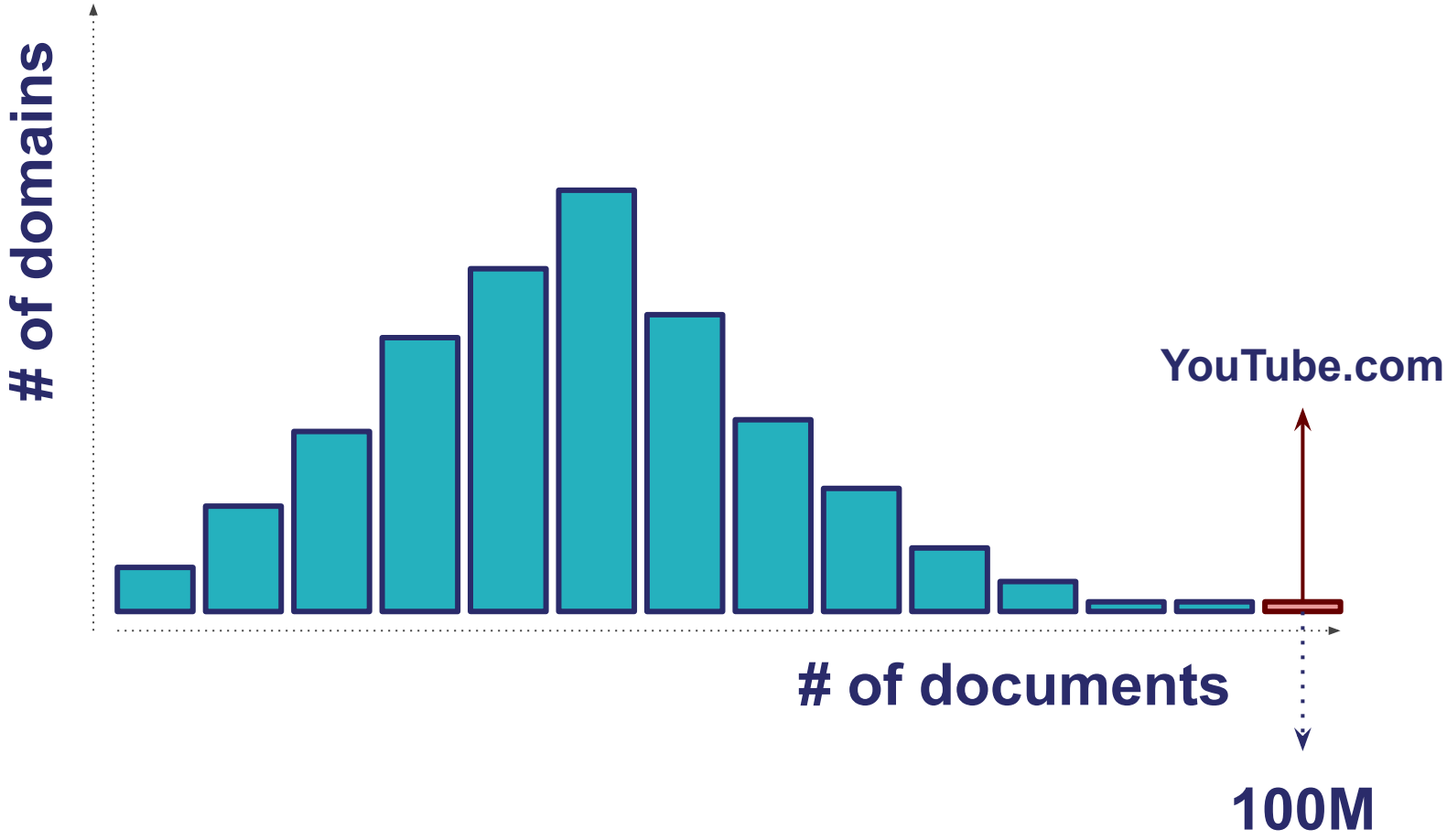
BEAM



BEAM



BEAM



Summary Metrics for 3130 Completed Tasks

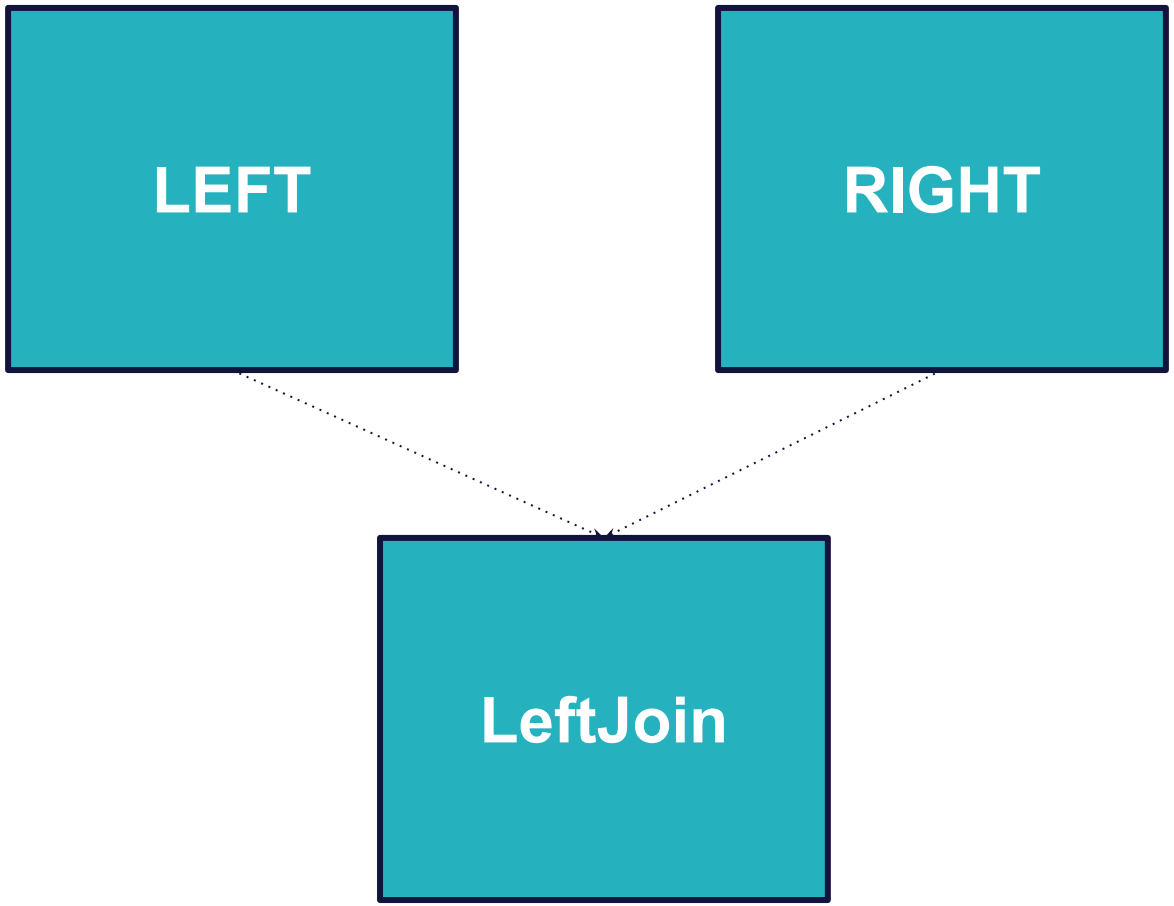
Metric	Min	25th percentile	Median	75th percentile	Max
Duration	48 s	2.5 min	3.5 min	4.8 min	28 min
GC Time	0.5 s	3 s	4 s	6 s	33 s
Output Size / Records	15.0 MB / 296710	62.2 MB / 1135248	88.9 MB / 1554320	121.3 MB / 2070174	561.7 MB / 10055260
Shuffle Read Size / Records	49.6 MB / 105710	142.8 MB / 231421	184.9 MB / 298787	236.4 MB / 390047	1643.6 MB / 6241261
Shuffle spill (memory)	0.0 B	0.0 B	0.0 B	0.0 B	4.7 GB
Shuffle spill (disk)	0.0 B	0.0 B	0.0 B	0.0 B	1594.5 MB

BEAM

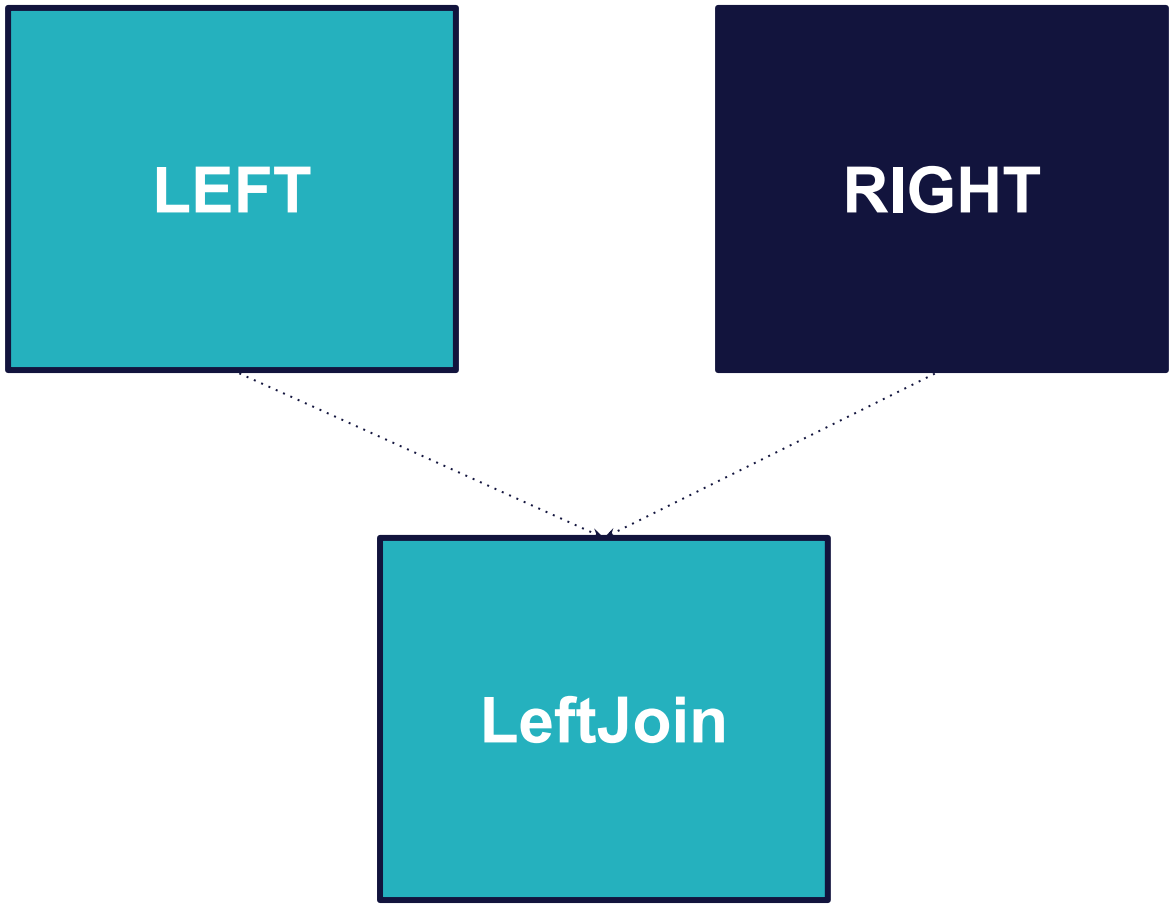
Summary Metrics for 3130 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	48 s	2.5 min	3.5 min	4.8 min	28 min
GC Time	0.5 s	3 s	4 s	6 s	33 s
Output Size / Records	15.0 MB / 296710	62.2 MB / 1135248	88.9 MB / 1554320	121.3 MB / 2070174	561.7 MB / 10055260
Shuffle Read Size / Records	49.6 MB / 105710	142.8 MB / 231421	184.9 MB / 298787	236.4 MB / 390047	1643.6 MB / 6241261
Shuffle spill (memory)	0.0 B	0.0 B	0.0 B	0.0 B	4.7 GB
Shuffle spill (disk)	0.0 B	0.0 B	0.0 B	0.0 B	1594.5 MB

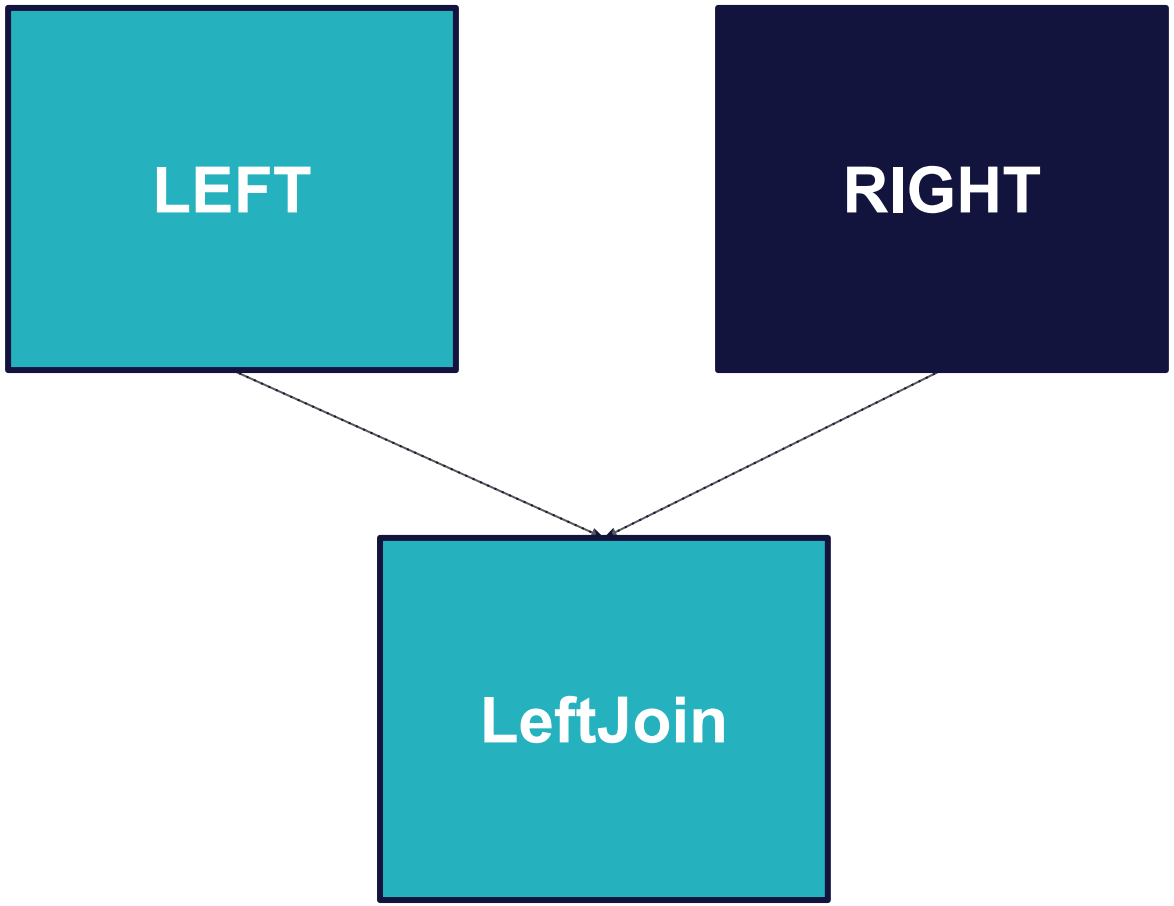
Solution A
Map-side Join


LEFT
RIGHT
LeftJoin

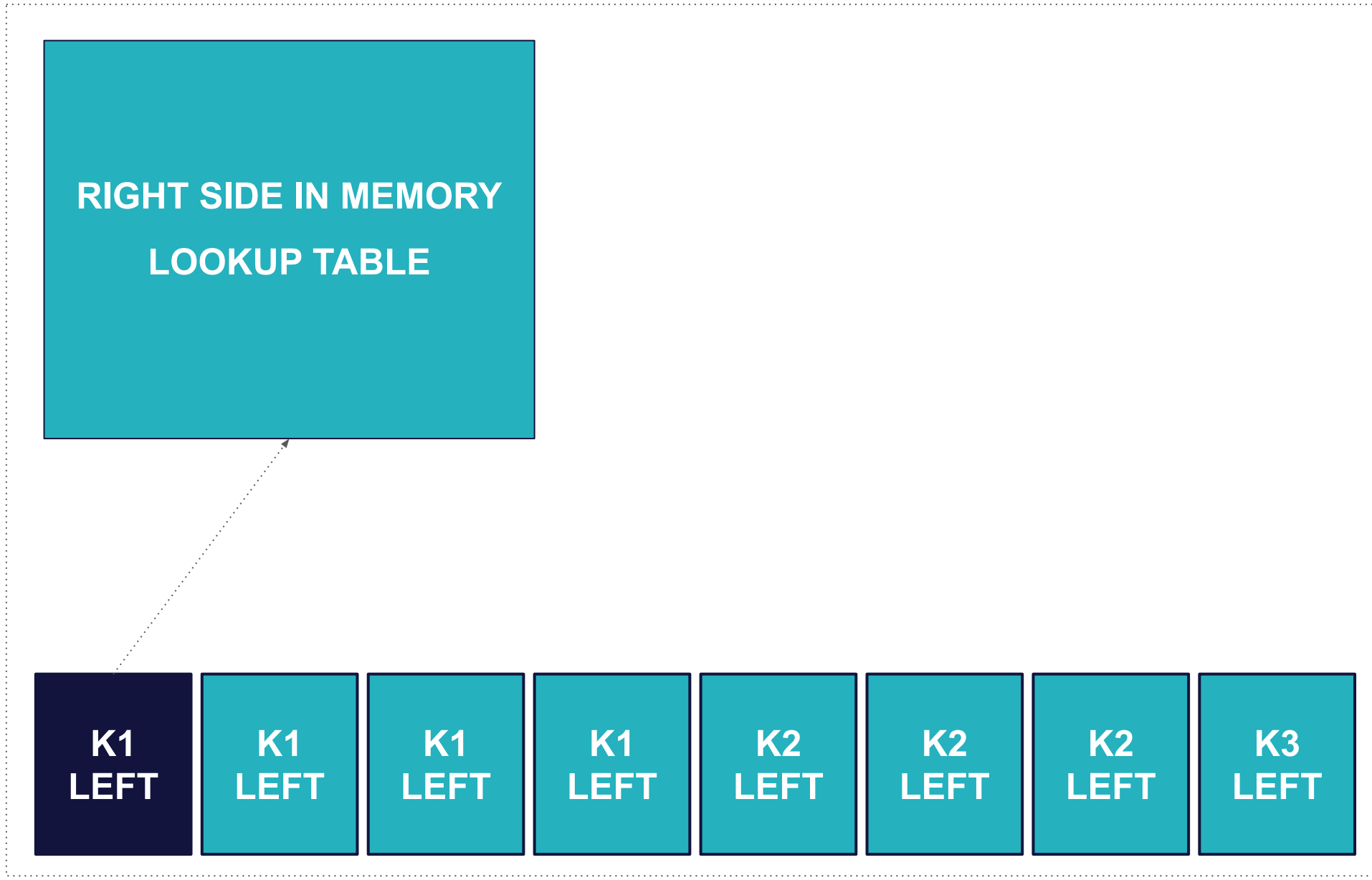
FITS IN MEMORY



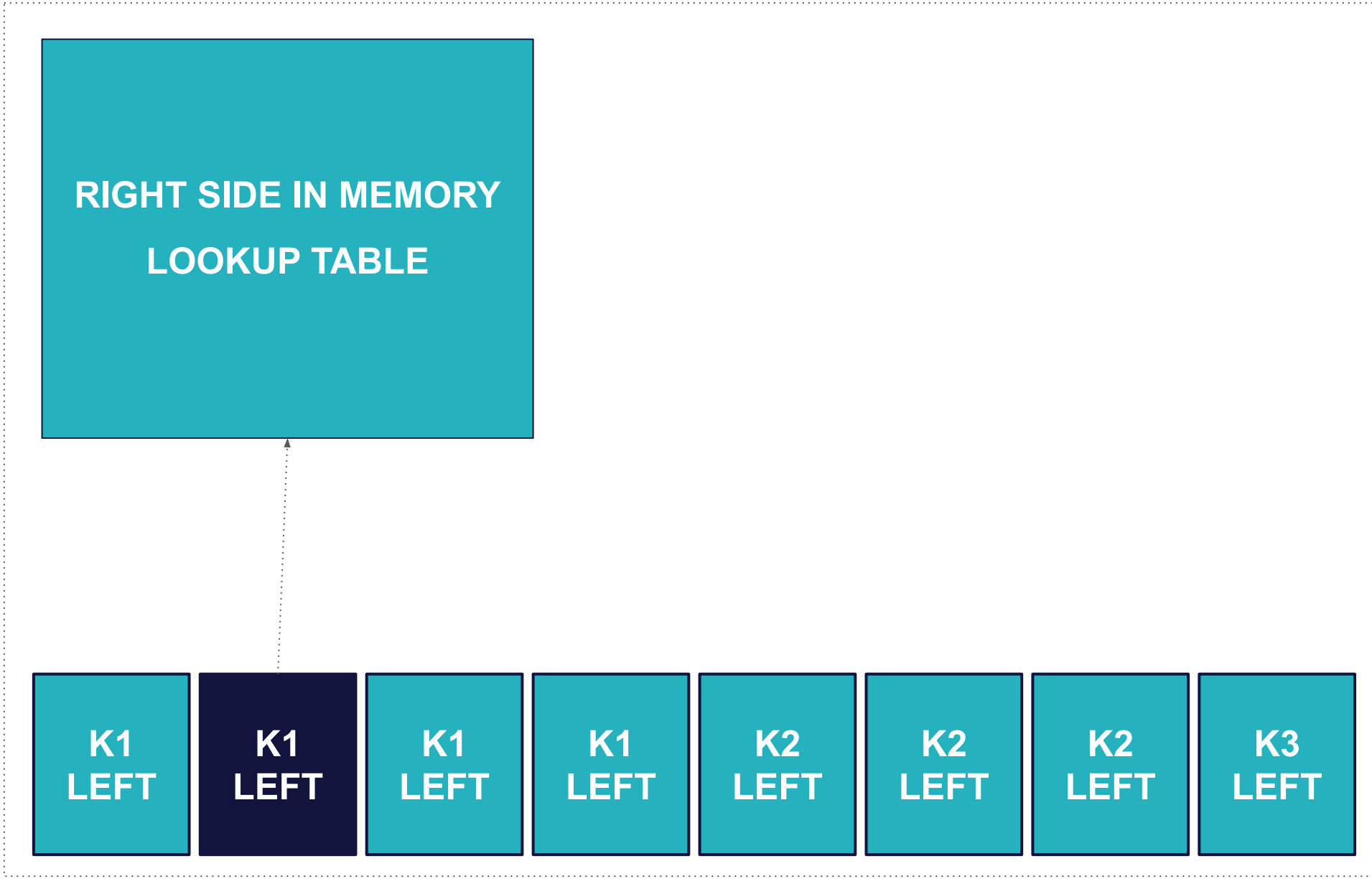
FITS IN MEMORY

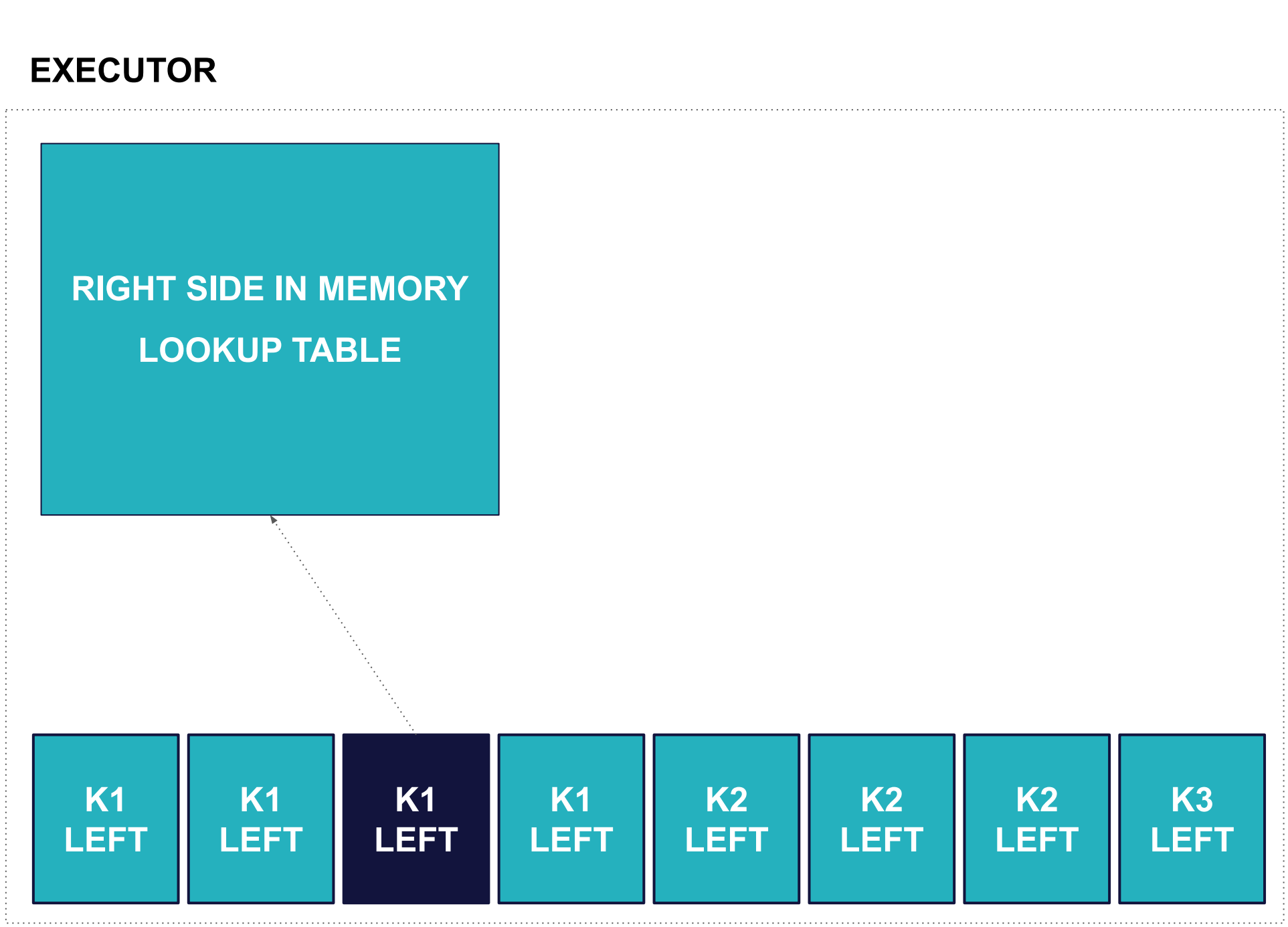


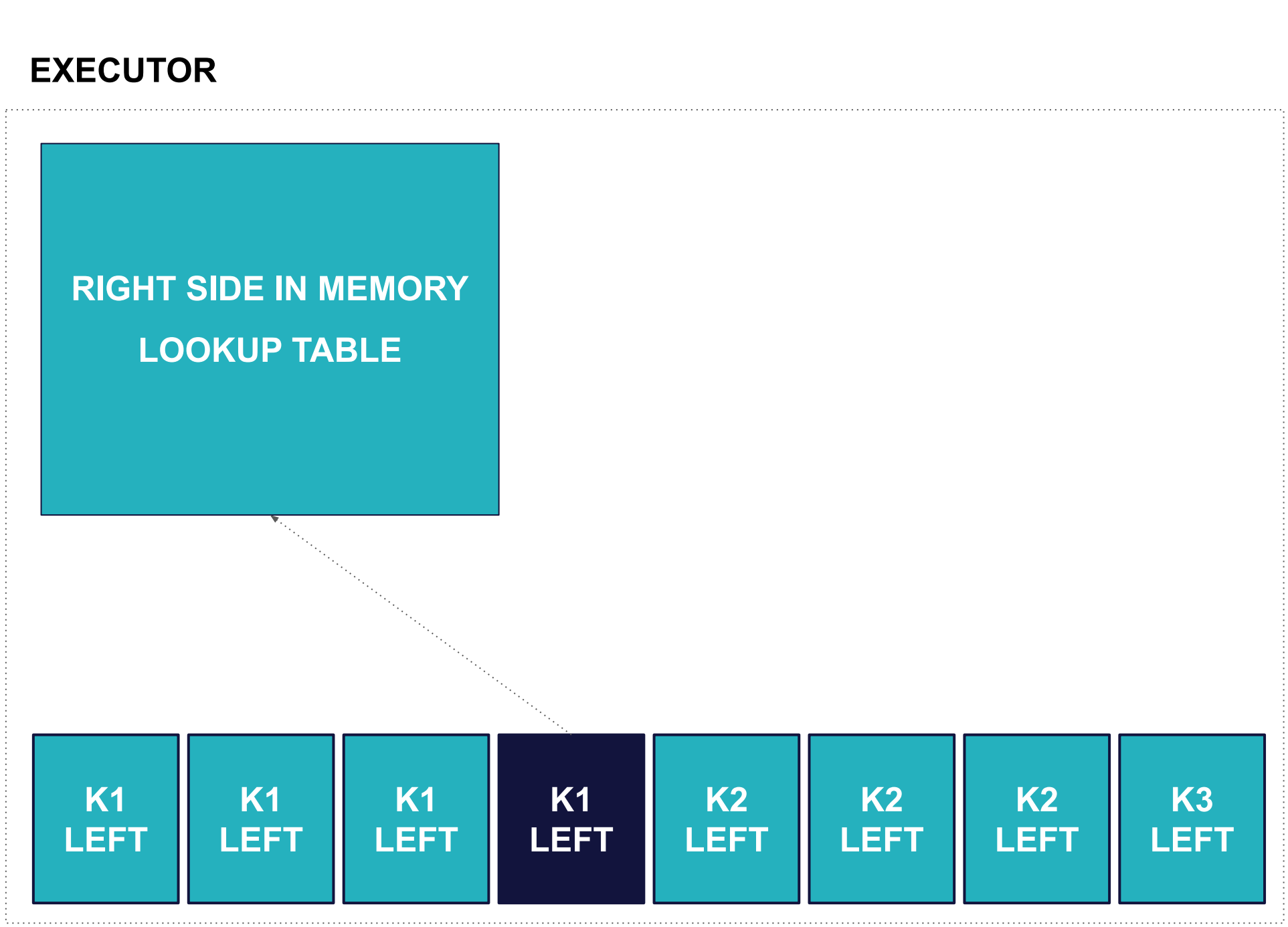
Collect results and broadcast them to all executors.


EXECUTOR
RIGHT SIDE IN MEMORY
LOOKUP TABLE
K1 LEFT
K1 LEFT
K1 LEFT
K1 LEFT
K2 LEFT
K2 LEFT
K2 LEFT
K3 LEFT

EXECUTOR


RIGHT SIDE IN MEMORY
LOOKUP TABLE
K1 LEFT
K1 LEFT
K1 LEFT
K1 LEFT
K2 LEFT
K2 LEFT
K2 LEFT
K3 LEFT

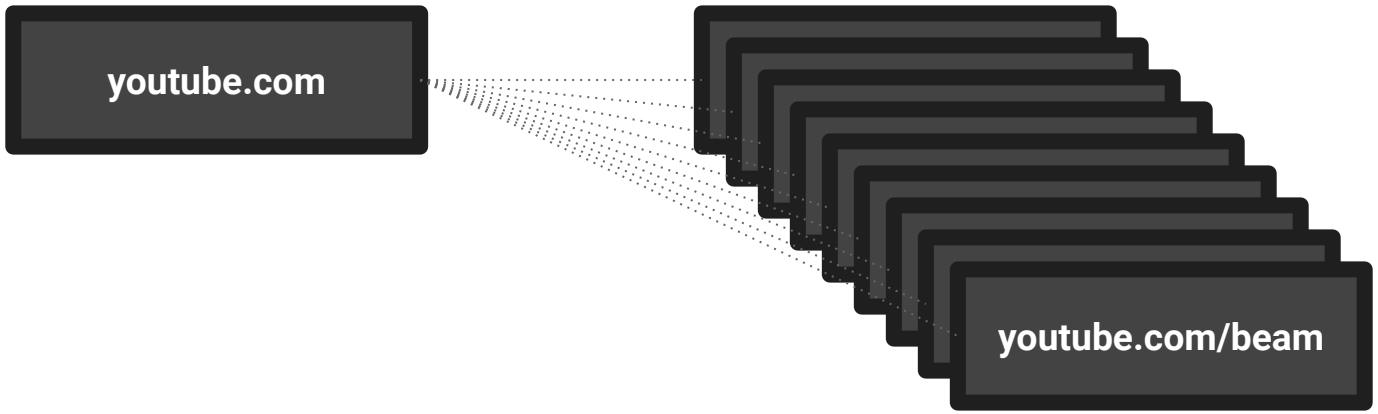

EXECUTOR
RIGHT SIDE IN MEMORY
LOOKUP TABLE
K1 LEFT
K1 LEFT
K1 LEFT
K1 LEFT
K2 LEFT
K2 LEFT
K2 LEFT
K3 LEFT


EXECUTOR
RIGHT SIDE IN MEMORY
LOOKUP TABLE
K1 LEFT
K1 LEFT
K1 LEFT
K1 LEFT
K2 LEFT
K2 LEFT
K2 LEFT
K3 LEFT

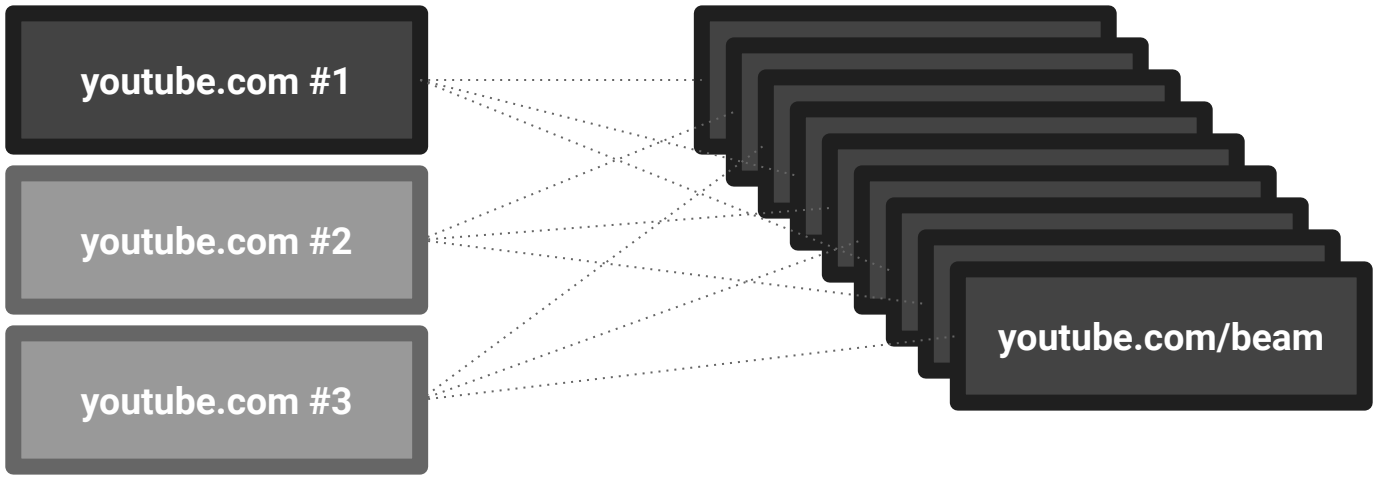
BEAM

Solution B
Splitting large keys

BEAM



BEAM



BEAM

All values for a single key must fit in-memory

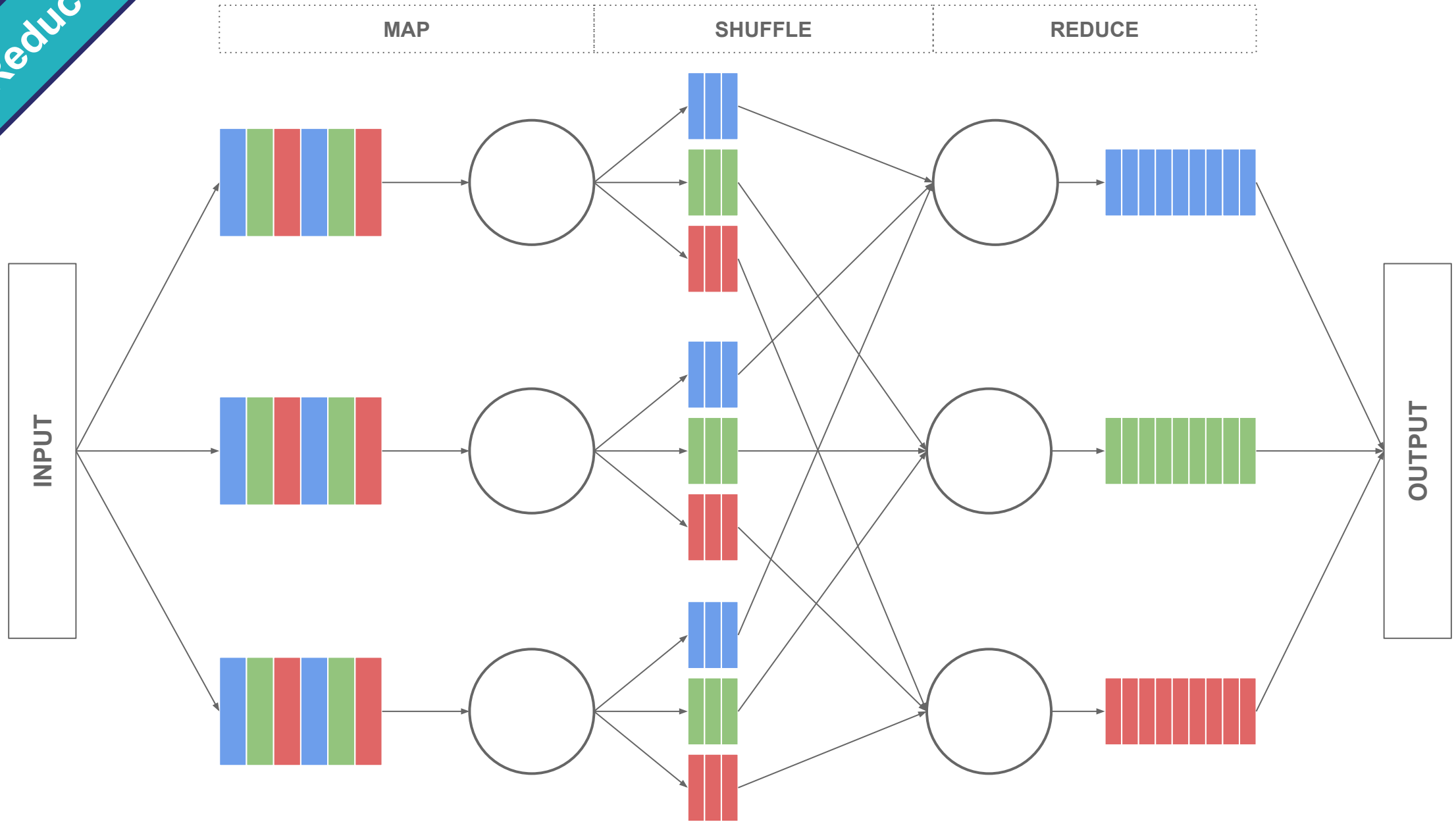
BEAM

groupByKey vs reduceByKey

BEAM

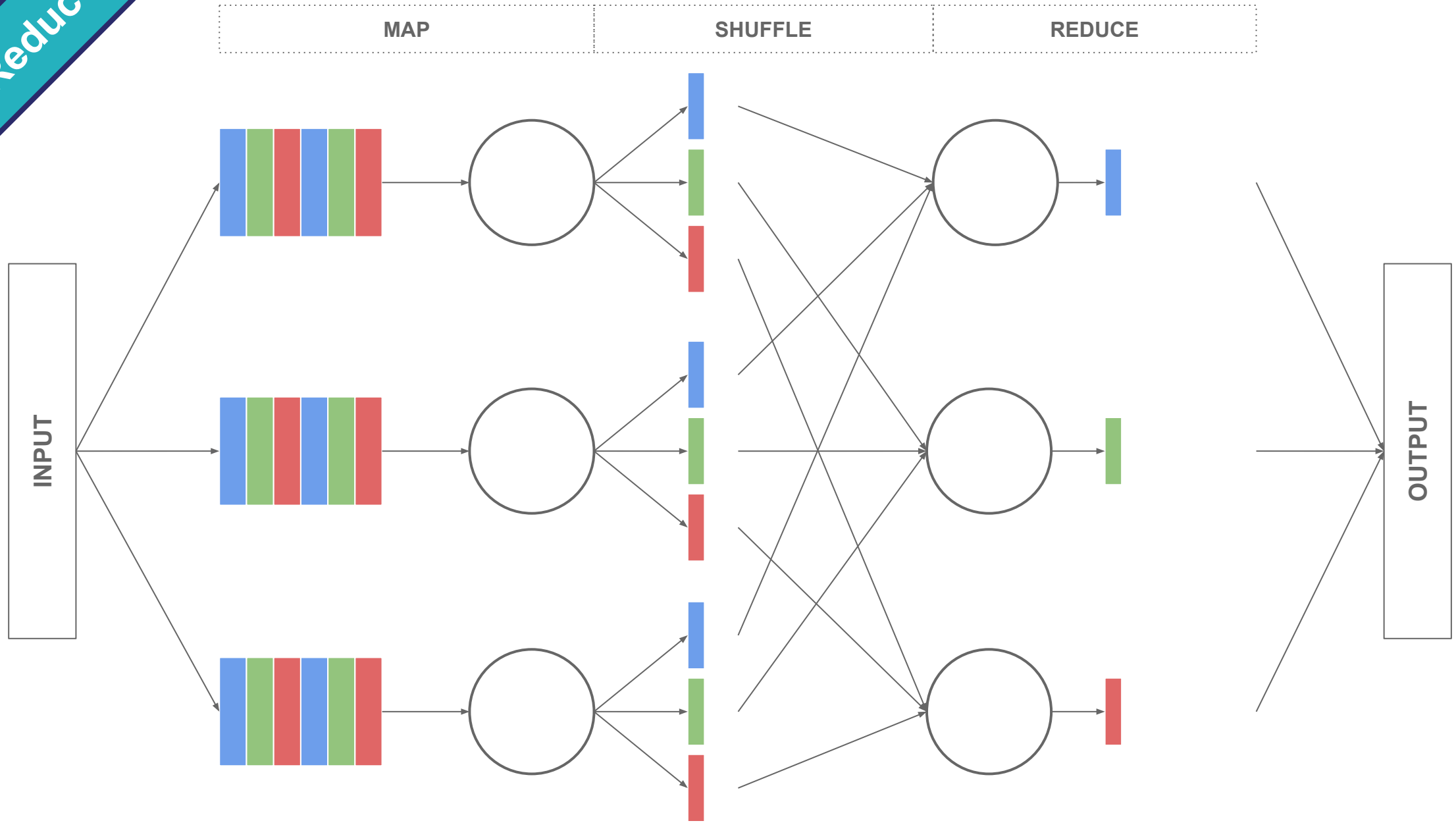
GroupByKey.create VS Combine.perKey

MapReduce



BERLIN 2019

MapReduce



BERLIN 2019

QUIZ

BEAM

```
final JavaPairRDD<byte[], byte[]> myRDD = ...;
grouped = myRDD.groupByKey();
```

QUIZ

BERLIN 2019

BEAM

```
final JavaPairRDD<byte[], byte[]> myRDD = ...;
grouped = myRDD.groupByKey();
```

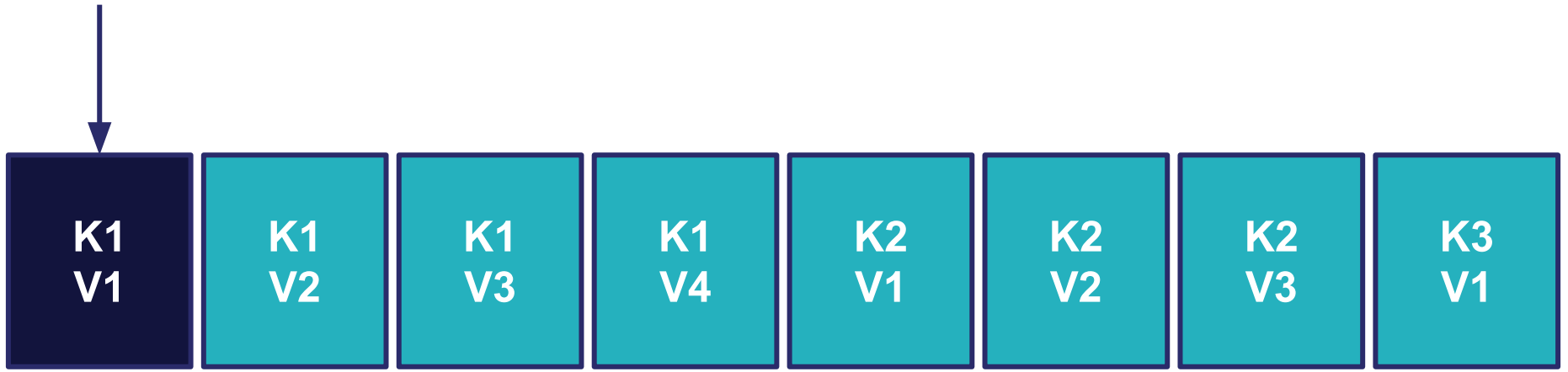
BERLIN 2019

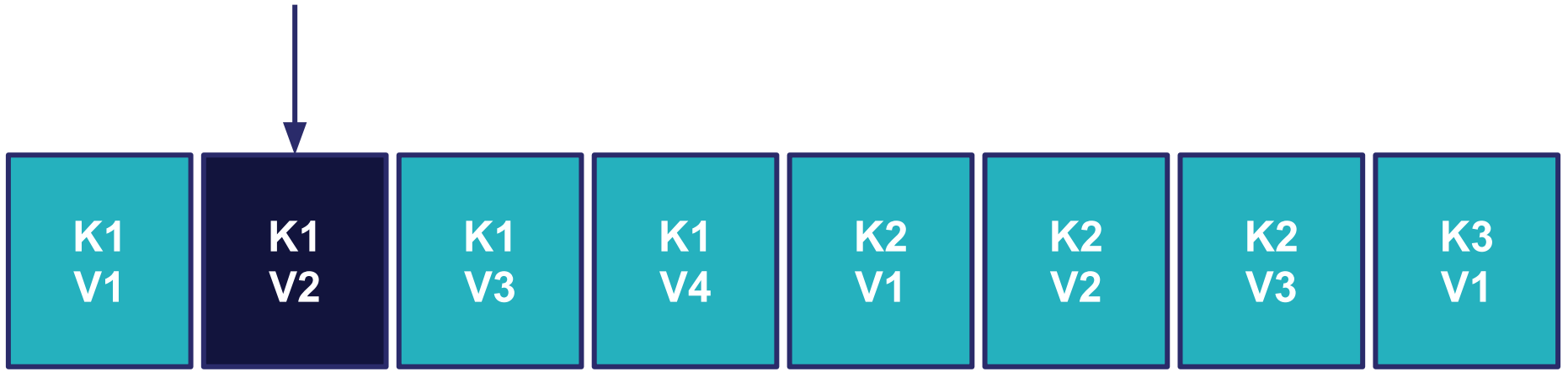
BEAM

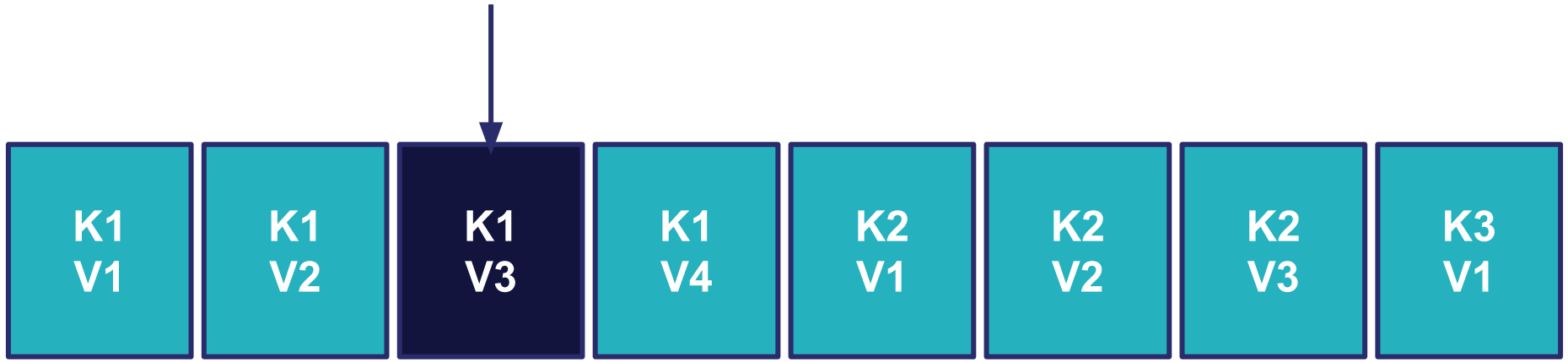
Memory efficient GroupByKey

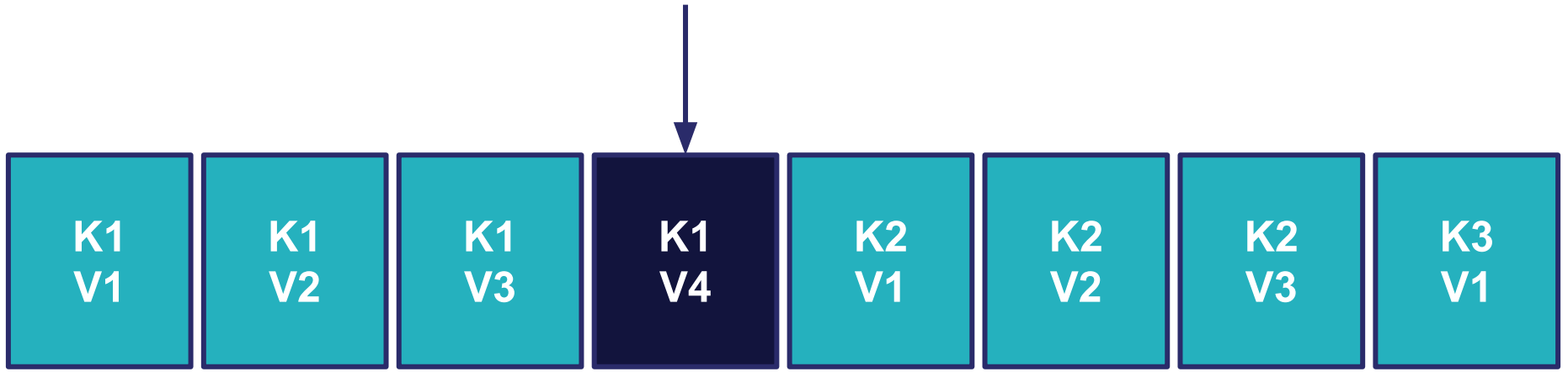
repartitionAndSortWithinPartitions(*partitioner*, *comparator*)


K1
V1
K1
V2
K1
V3
K1
V4
K2
V1
K2
V2
K2
V3
K3
V1


K1
V1
K1
V2
K1
V3
K1
V4
K2
V1
K2
V2
K2
V3
K3
V1


K1
V1
K1
V2
K1
V3
K1
V4
K2
V1
K2
V2
K2
V3
K3
V1


K1
V1
K1
V2
K1
V3
K1
V4
K2
V1
K2
V2
K2
V3
K3
V1


K1
V1
K1
V2
K1
V3
K1
V4
K2
V1
K2
V2
K2
V3
K3
V1

BEAM

Now everything should work, right?

FetchFailedException

Summary Metrics for 32651 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	4.2 min	6.7 min	7.6 min	9.0 min	57 min
Scheduler Delay	0 ms	18 ms	21 ms	26 ms	21 s
Task Deserialization Time	4 ms	9 ms	11 ms	14 ms	2.1 min
GC Time	4 s	21 s	24 s	28 s	11 min
Result Serialization Time	0 ms	0 ms	0 ms	0 ms	1 s
Getting Result Time	0 ms	0 ms	0 ms	0 ms	0 ms
Peak Execution Memory	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B
Shuffle Read Blocked Time	1 s	60 s	1.4 min	2.0 min	20 min
Shuffle Read Size / Records	1450.4 MB / 3487300	1668.1 MB / 4025965	1710.8 MB / 4133493	1755.4 MB / 4244198	1971.6 MB / 4807860
Shuffle Remote Reads	1450.4 MB	1668.1 MB	1710.8 MB	1755.4 MB	1971.6 MB
Shuffle Write Size / Records	69.0 B / 1	71.0 B / 1	71.0 B / 1	71.0 B / 1	71.0 B / 1
Shuffle spill (memory)	1757.8 MB	5.1 GB	5.3 GB	6.8 GB	7.6 GB
Shuffle spill (disk)	407.5 MB	1238.7 MB	1277.2 MB	1632.5 MB	1791.2 MB

Summary Metrics for 32651 Completed Tasks

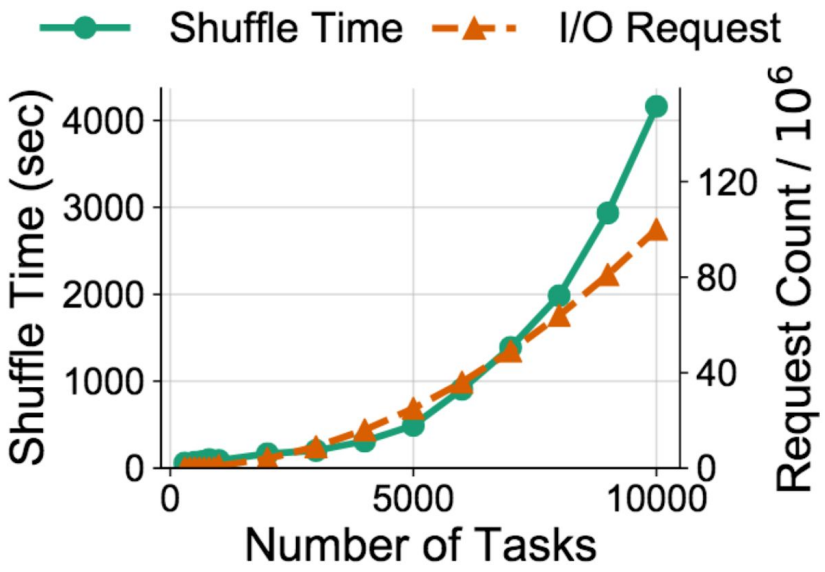
Metric	Min	25th percentile	Median	75th percentile	Max
Duration	4.2 min	6.7 min	7.6 min	9.0 min	57 min
Scheduler Delay	0 ms	18 ms	21 ms	26 ms	21 s
Task Deserialization Time	4 ms	9 ms	11 ms	14 ms	2.1 min
GC Time	4 s	21 s	24 s	28 s	11 min
Result Serialization Time	0 ms	0 ms	0 ms	0 ms	1 s
Getting Result Time	0 ms	0 ms	0 ms	0 ms	0 ms
Peak Execution Memory	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B
Shuffle Read Blocked Time	1 s	60 s	1.4 min	2.0 min	20 min
Shuffle Read Size / Records	1450.4 MB / 3487300	1668.1 MB / 4025965	1710.8 MB / 4133493	1755.4 MB / 4244198	1971.6 MB / 4807860
Shuffle Remote Reads	1450.4 MB	1668.1 MB	1710.8 MB	1755.4 MB	1971.6 MB
Shuffle Write Size / Records	69.0 B / 1	71.0 B / 1	71.0 B / 1	71.0 B / 1	71.0 B / 1
Shuffle spill (memory)	1757.8 MB	5.1 GB	5.3 GB	6.8 GB	7.6 GB
Shuffle spill (disk)	407.5 MB	1238.7 MB	1277.2 MB	1632.5 MB	1791.2 MB

Summary Metrics for 32651 Completed Tasks

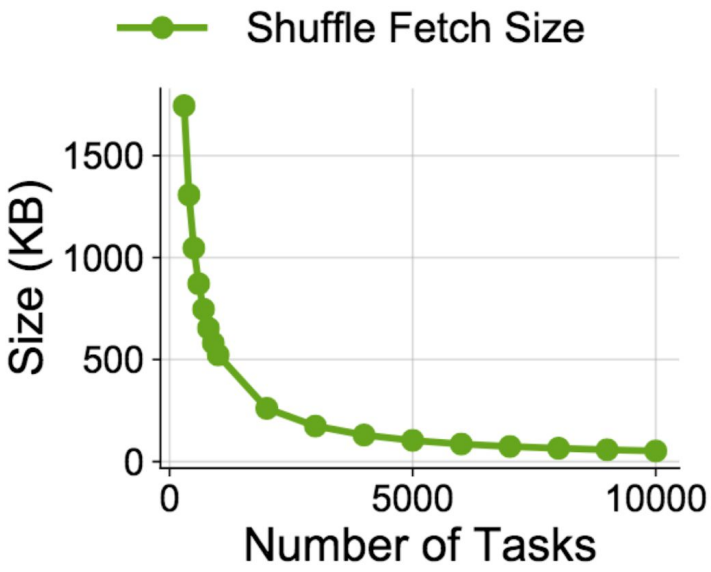
Metric	Min	25th percentile	Median	75th percentile	Max
Duration	4.2 min	6.7 min	7.6 min	9.0 min	57 min
Scheduler Delay	0 ms	18 ms	21 ms	26 ms	21 s
Task Deserialization Time	4 ms	9 ms	11 ms	14 ms	2.1 min
GC Time	4 s	21 s	24 s	28 s	11 min
Result Serialization Time	0 ms	0 ms	0 ms	0 ms	1 s
Getting Result Time	0 ms	0 ms	0 ms	0 ms	0 ms
Peak Execution Memory	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B
Shuffle Read Blocked Time	1 s	60 s	1.4 min	2.0 min	20 min
Shuffle Read Size / Records	1450.4 MB / 3487300	1668.1 MB / 4025965	1710.8 MB / 4133493	1755.4 MB / 4244198	1971.6 MB / 4807860
Shuffle Remote Reads	1450.4 MB	1668.1 MB	1710.8 MB	1755.4 MB	1971.6 MB
Shuffle Write Size / Records	69.0 B / 1	71.0 B / 1	71.0 B / 1	71.0 B / 1	71.0 B / 1
Shuffle spill (memory)	1757.8 MB	5.1 GB	5.3 GB	6.8 GB	7.6 GB
Shuffle spill (disk)	407.5 MB	1238.7 MB	1277.2 MB	1632.5 MB	1791.2 MB

BEAM

Understanding Spark shuffle



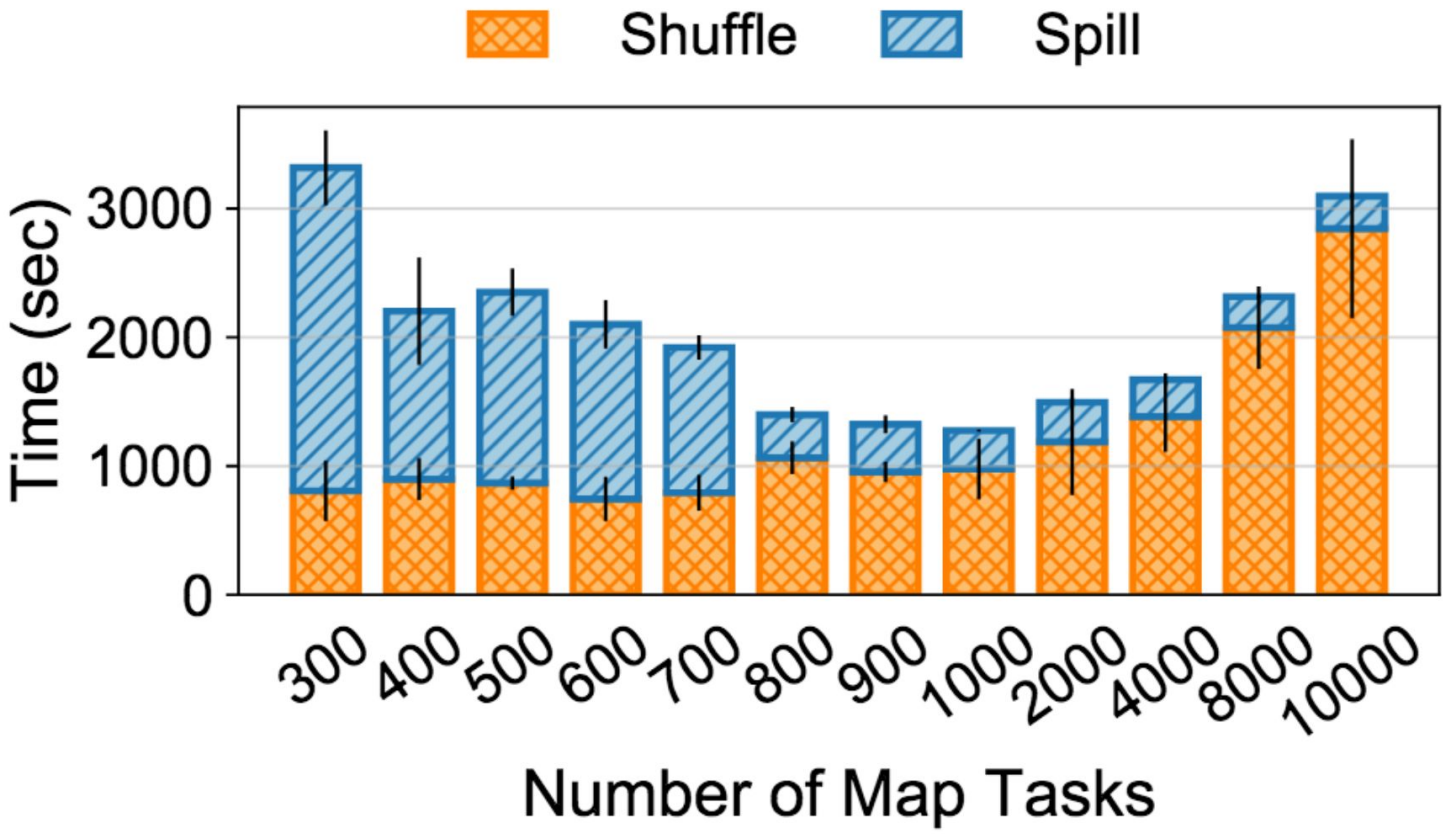
(a) Shuffle time and I/O requests.



(b) Average I/O request size.

Riffle: Optimized Shuffle Service for Large-Scale Data Analytics

https://haoyuzhang.org/publications/riffle-eurosys18.pdf



Riffle: Optimized Shuffle Service for Large-Scale Data Analytics

https://haoyuzhang.org/publications/riffle-eurosys18.pdf

Performance

BERLIN 2019

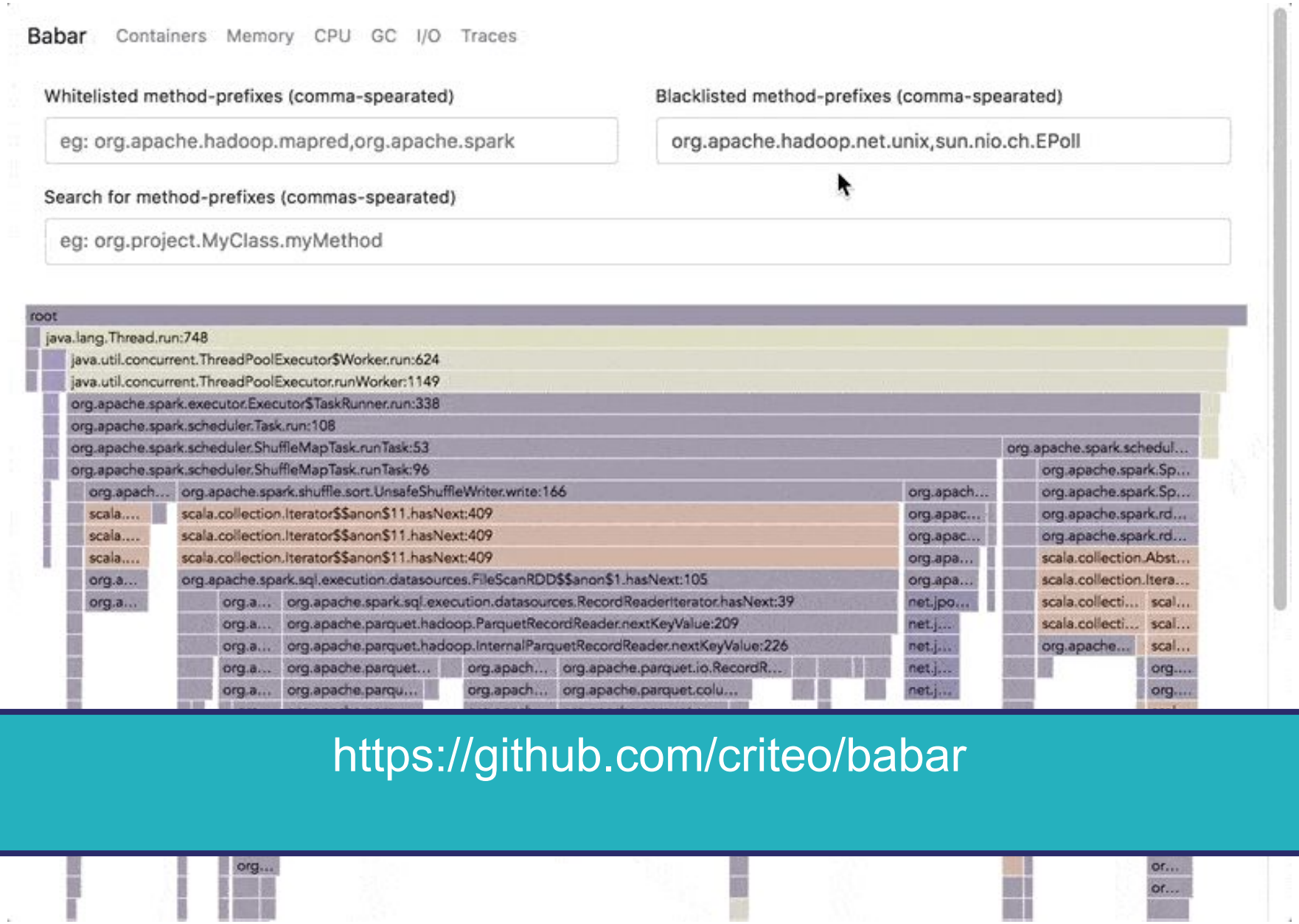
BEAM

byte-based shuffle

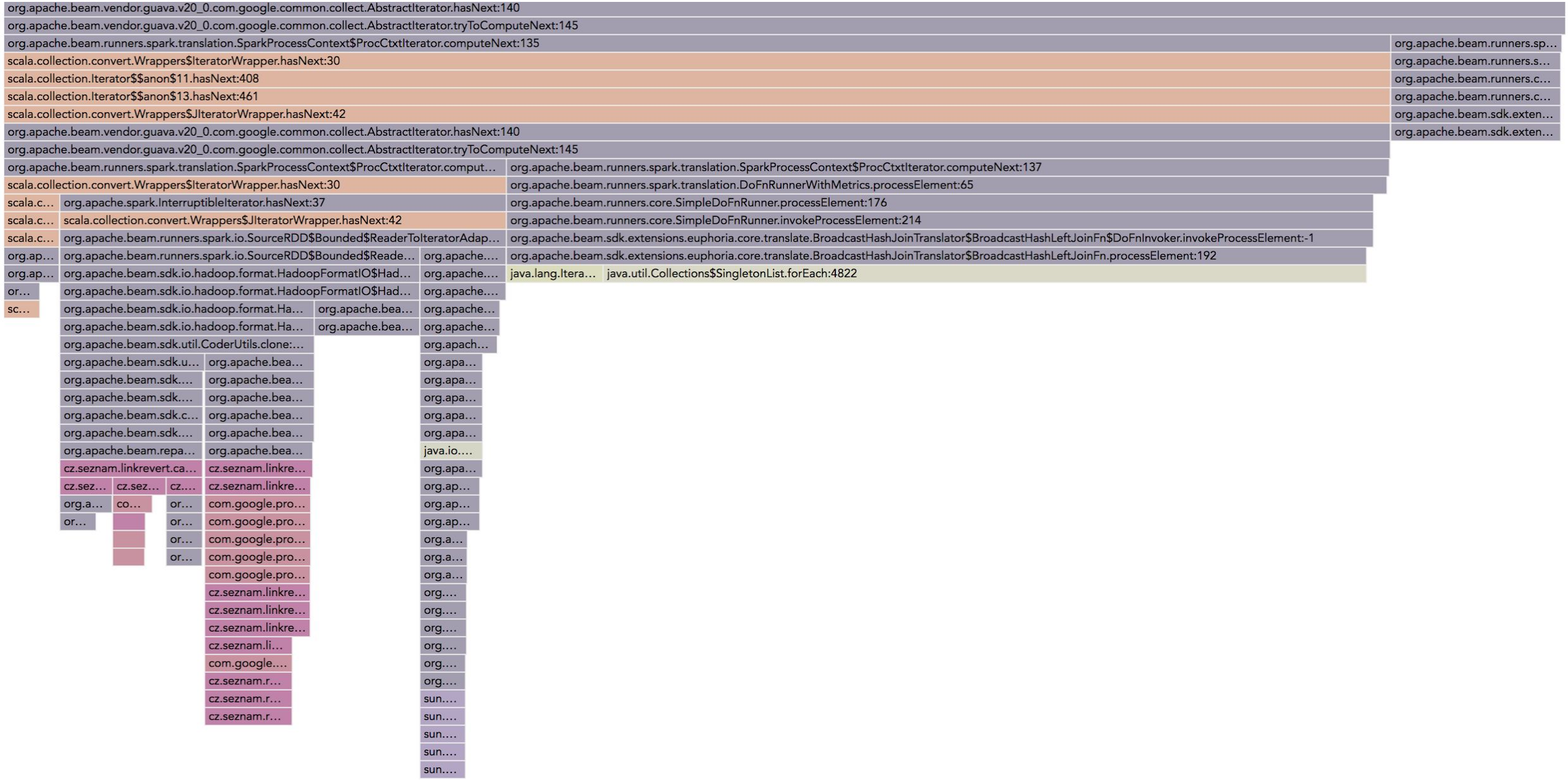
BERLIN 2019

BEAM

debugging spark pipelines



https://github.com/criteo/babar


org.apache.beam.vendor.guava.v20_0.com.google.common.collect.AbstractIterator.hasNext:140
org.apache.beam.vendor.guava.v20_0.com.google.common.collect.AbstractIterator.tryToComputeNext:145
org.apache.beam.runners.spark.translation.SparkProcessContext$ProcCtxtIterator.computeNext:135
org.apache.beam.runners.sp…
scala.collection.convert.Wrappers$IteratorWrapper.hasNext:30
org.apache.beam.runners.s…
scala.collection.Iterator$$anon$11.hasNext:408
org.apache.beam.runners.c…
scala.collection.Iterator$$anon$13.hasNext:461
org.apache.beam.runners.c…
scala.collection.convert.Wrappers$JIteratorWrapper.hasNext:42
org.apache.beam.sdk.exten…
org.apache.beam.vendor.guava.v20_0.com.google.common.collect.AbstractIterator.hasNext:140
org.apache.beam.sdk.exten…
org.apache.beam.vendor.guava.v20_0.com.google.common.collect.AbstractIterator.tryToComputeNext:145
org.apache.beam.runners.spark.translation.SparkProcessContext$ProcCtxtIterator.comput…
org.apache.beam.runners.spark.translation.SparkProcessContext$ProcCtxtIterator.computeNext:137
scala.collection.convert.Wrappers$IteratorWrapper.hasNext:30
org.apache.beam.runners.spark.translation.DoFnRunnerWithMetrics.processElement:65
scala.c…
org.apache.spark.InterruptibleIterator.hasNext:37
org.apache.beam.runners.core.SimpleDoFnRunner.processElement:176
scala.c…
scala.collection.convert.Wrappers$JIteratorWrapper.hasNext:42
org.apache.beam.runners.core.SimpleDoFnRunner.invokeProcessElement:214
scala.c…
org.apache.beam.runners.spark.io.SourceRDD$Bounded$ReaderToIteratorAdap…
org.apache.beam.sdk.extensions.euphoria.core.translate.BroadcastHashJoinTranslator$BroadcastHashLeftJoinFn$DoFnInvoker.invokeProcessElement:-1
org.ap…
org.apache.beam.runners.spark.io.SourceRDD$Bounded$Reade…
org.apache.…
org.apache.beam.sdk.extensions.euphoria.core.translate.BroadcastHashJoinTranslator$BroadcastHashLeftJoinFn.processElement:192
org.ap…
org.apache.beam.sdk.io.hadoop.format.HadoopFormatIO$Had…
org.apache.…
java.lang.Itera…
java.util.Collections$SingletonList.forEach:4822
or…
org.apache.beam.sdk.io.hadoop.format.HadoopFormatIO$Had…
org.apache.…
sc…
org.apache.beam.sdk.io.hadoop.format.Ha…
org.apache.bea…
org.apache…
org.apache.beam.sdk.io.hadoop.format.Ha…
org.apache.bea…
org.apache…
org.apache.beam.sdk.util.CoderUtils.clone:…
org.apach…
org.apache.beam.sdk.u…
org.apache.bea…
org.apa…
org.apache.beam.sdk.…
org.apache.bea…
org.apa…
org.apache.beam.sdk.…
org.apache.bea…
org.apa…
org.apache.beam.sdk.c…
org.apache.bea…
org.apa…
org.apache.beam.sdk.…
org.apache.bea…
org.apa…
org.apache.beam.repa…
org.apache.bea…
java.io.…
cz.seznam.linkrevert.ca…
cz.seznam.linkre…
org.apa…
cz.sez…
cz.sez…
cz.…
cz.seznam.linkre…
org.ap…
org.a…
co…
or…
com.google.pro…
org.ap…
or…
or…
com.google.pro…
org.ap…
or…
com.google.pro…
org.a…
or…
com.google.pro…
org.a…
com.google.pro…
org.a…
cz.seznam.linkre…
org.…
cz.seznam.linkre…
org.…
cz.seznam.linkre…
org.…
cz.seznam.li…
org.…
com.google.…
org.…
cz.seznam.r…
org.…
cz.seznam.r…
sun.…
cz.seznam.r…
sun.…
sun.…
sun.…
sun.…

BEAM-6332

BEAM

Combine performance

BEAM

heap dumps :)

Production ready for batch!

BEAM

Streaming your shared ride

Today from **15:20** to **16:00**

Thomas Weise

BEAM

Beam Summit Europe 2019

June 19-20, 2019

https://beamsummit.org/

BERLIN 2019

@davidmoravek
@ApacheBeam