Storm

Distributed and fault-tolerant realtime computation

Nathan Marz
Twitter

Basic info

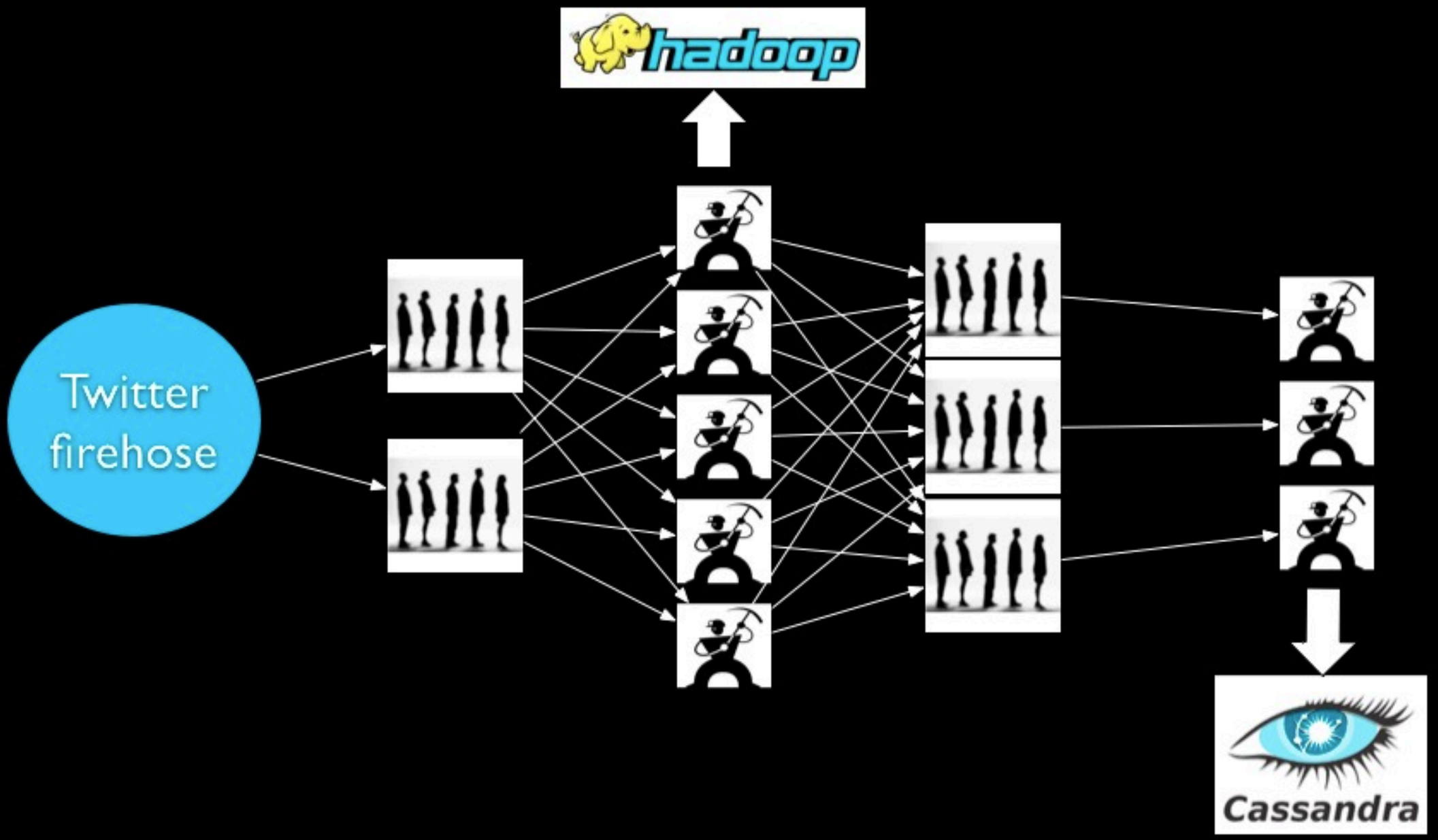
- Open sourced September 19th
- Implementation is 15,000 lines of code
- Used by over 25 companies
- >2900 watchers on Github (most watched JVM project)
- Very active mailing list
  - >3000 messages
  - >800 members

Before Storm

Queues

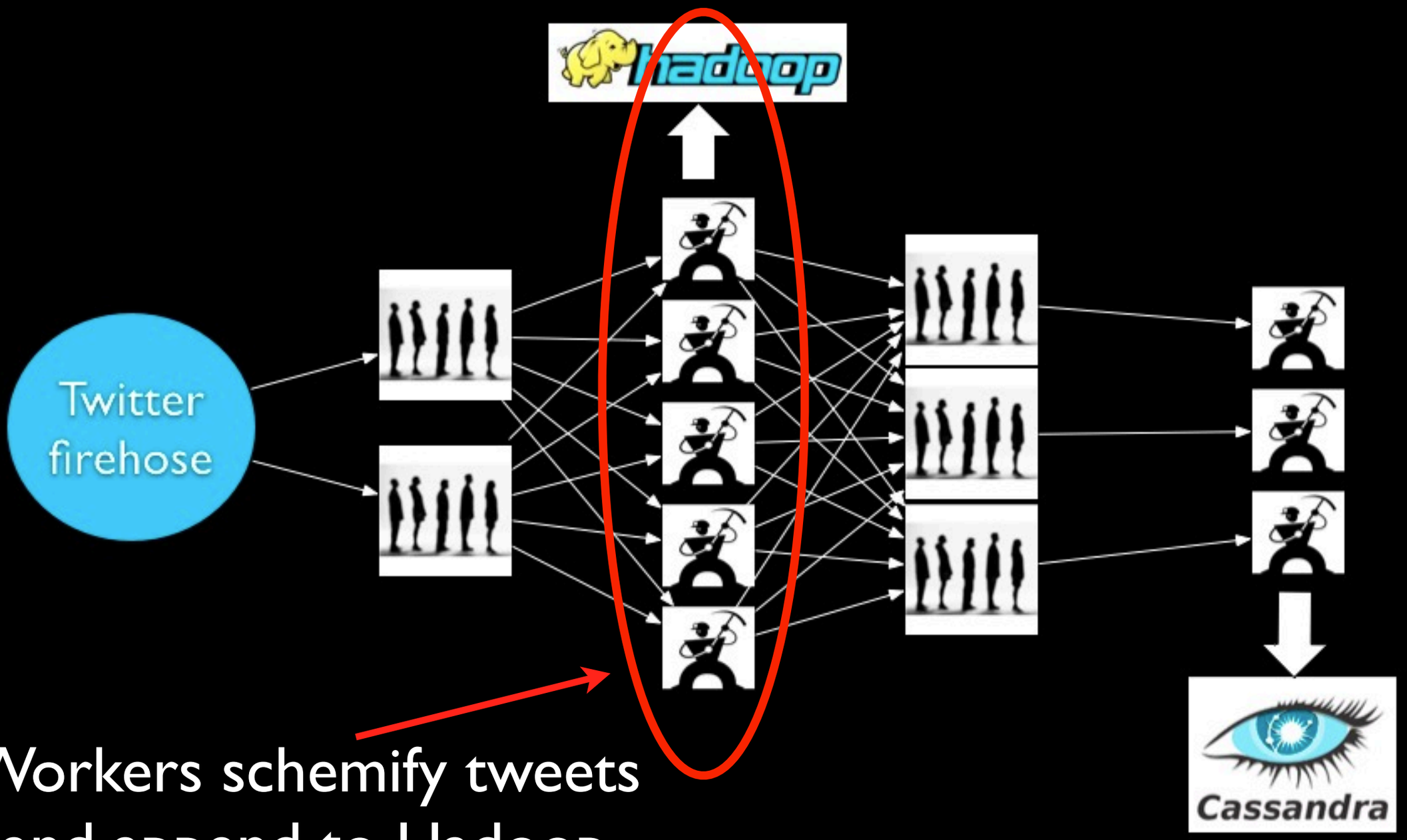
Workers

Example



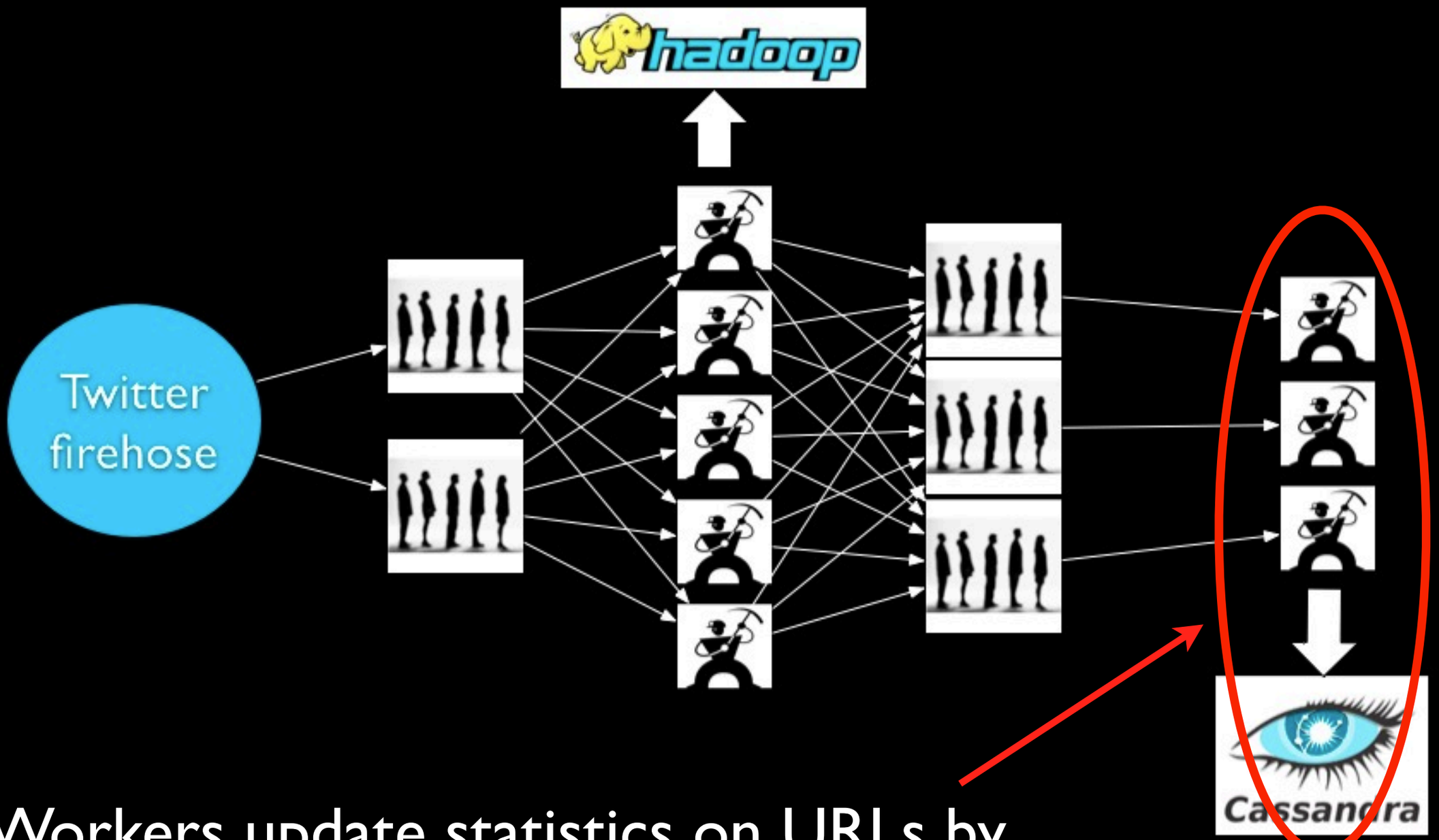
(simplified)

Example



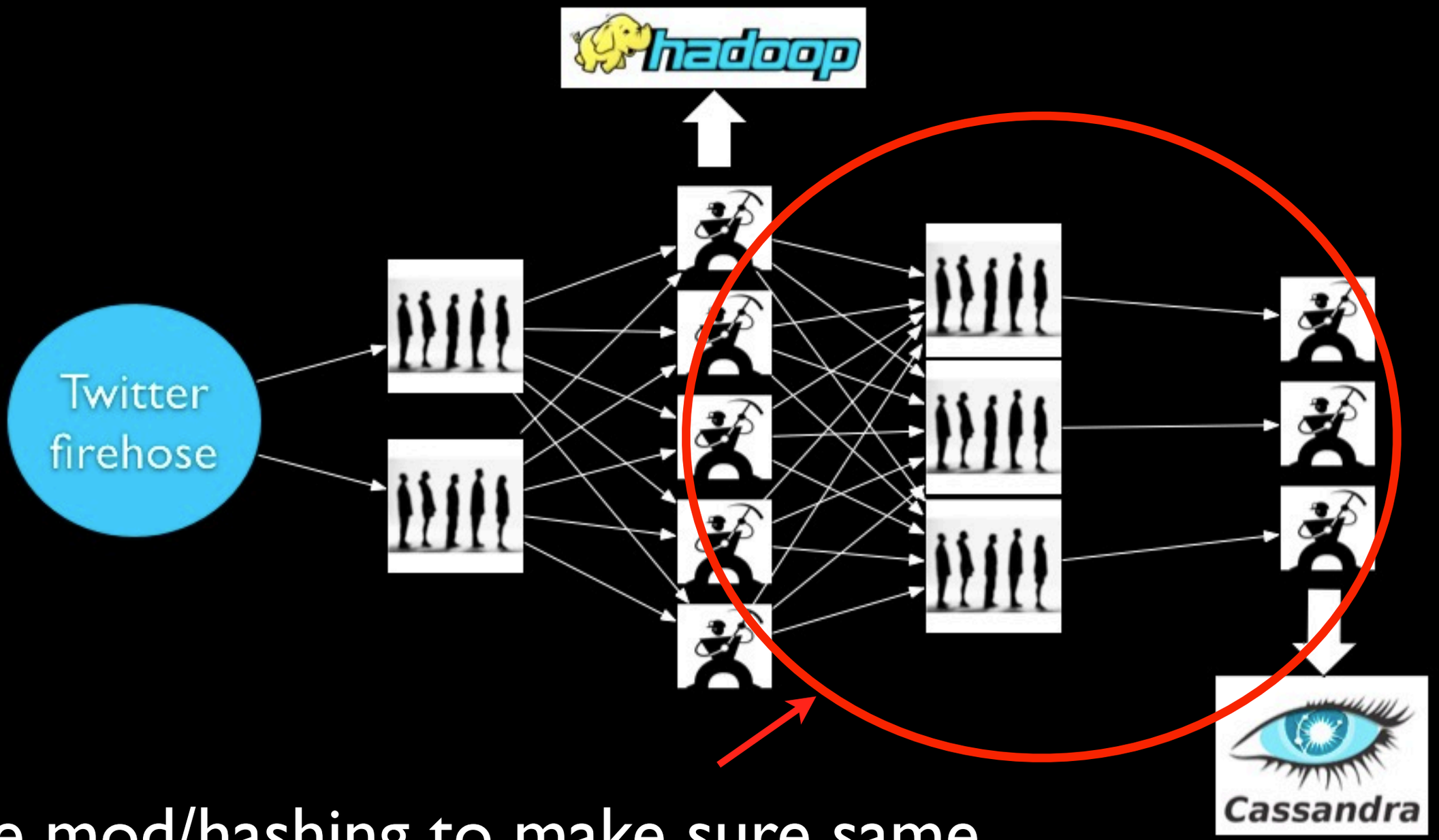
Workers schemify tweets and append to Hadoop

Example



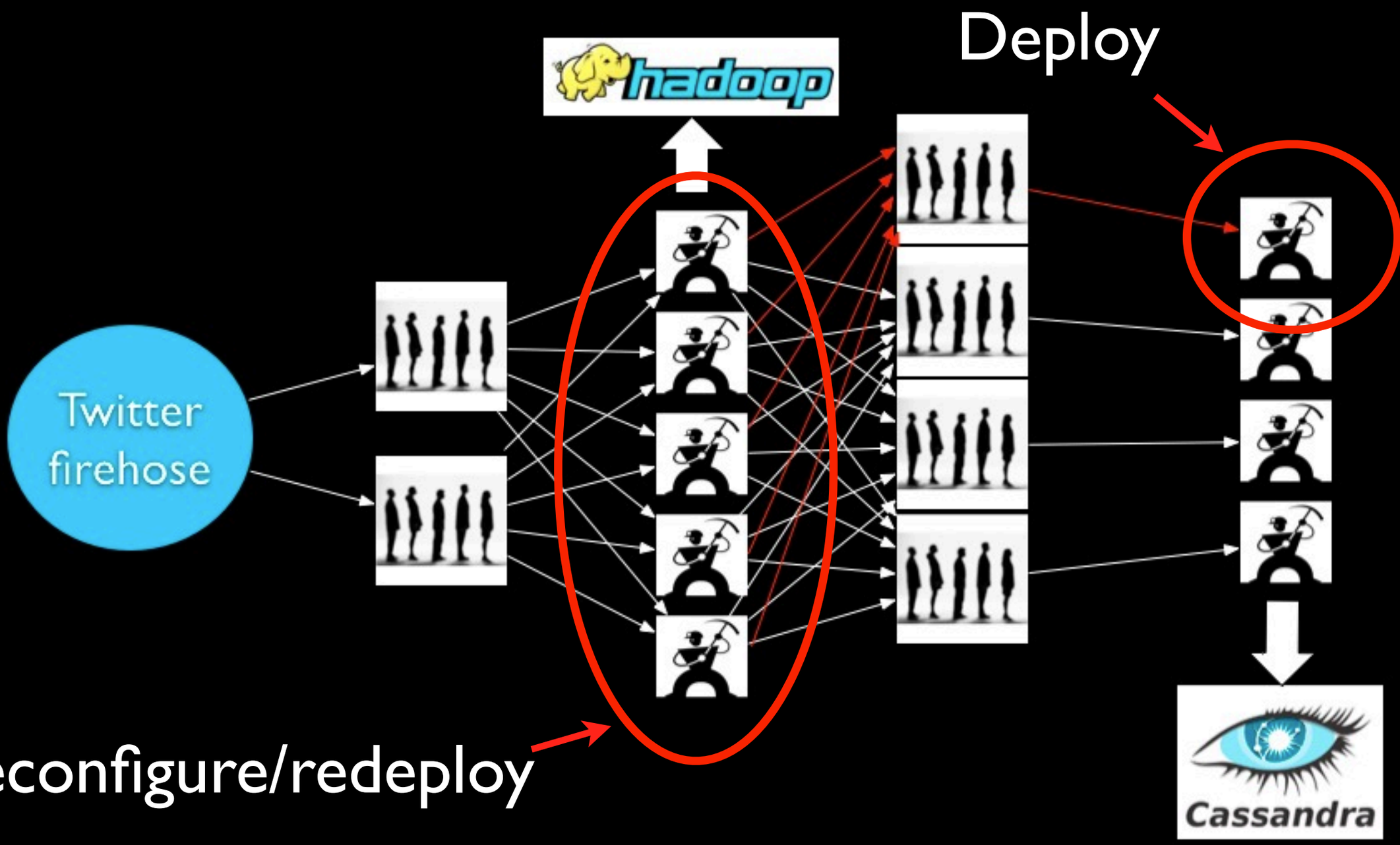
Workers update statistics on URLs by incrementing counters in Cassandra

Example



Use mod/hashing to make sure same URL always goes to same worker

Scaling


Deploy
hadoop
Twitter firehose
Reconfigure/redeploy
Cassandra

Problems

- Scaling is painful
- Poor fault-tolerance
- Coding is tedious

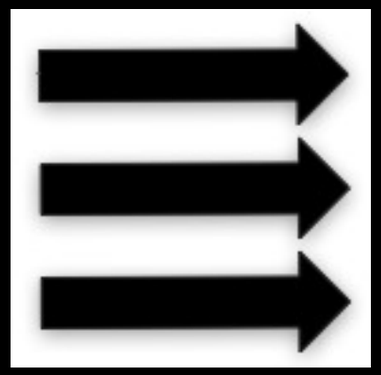
What we want

- Guaranteed data processing
- Horizontal scalability
- Fault-tolerance
- No intermediate message brokers!
- Higher level abstraction than message passing
- “Just works”

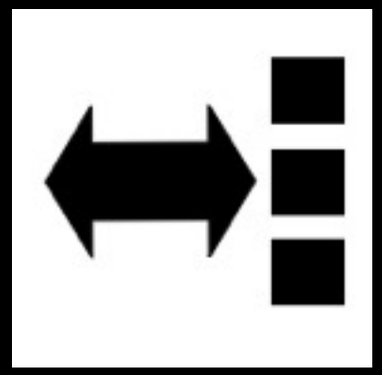
Storm

- ✔ Guaranteed data processing
- ✔ Horizontal scalability
- ✔ Fault-tolerance
- ✔ No intermediate message brokers!
- ✔ Higher level abstraction than message passing
- ✔ “Just works”

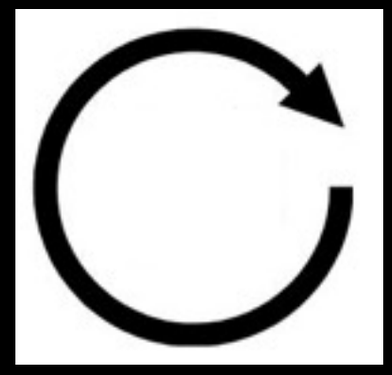
Use cases



Stream processing

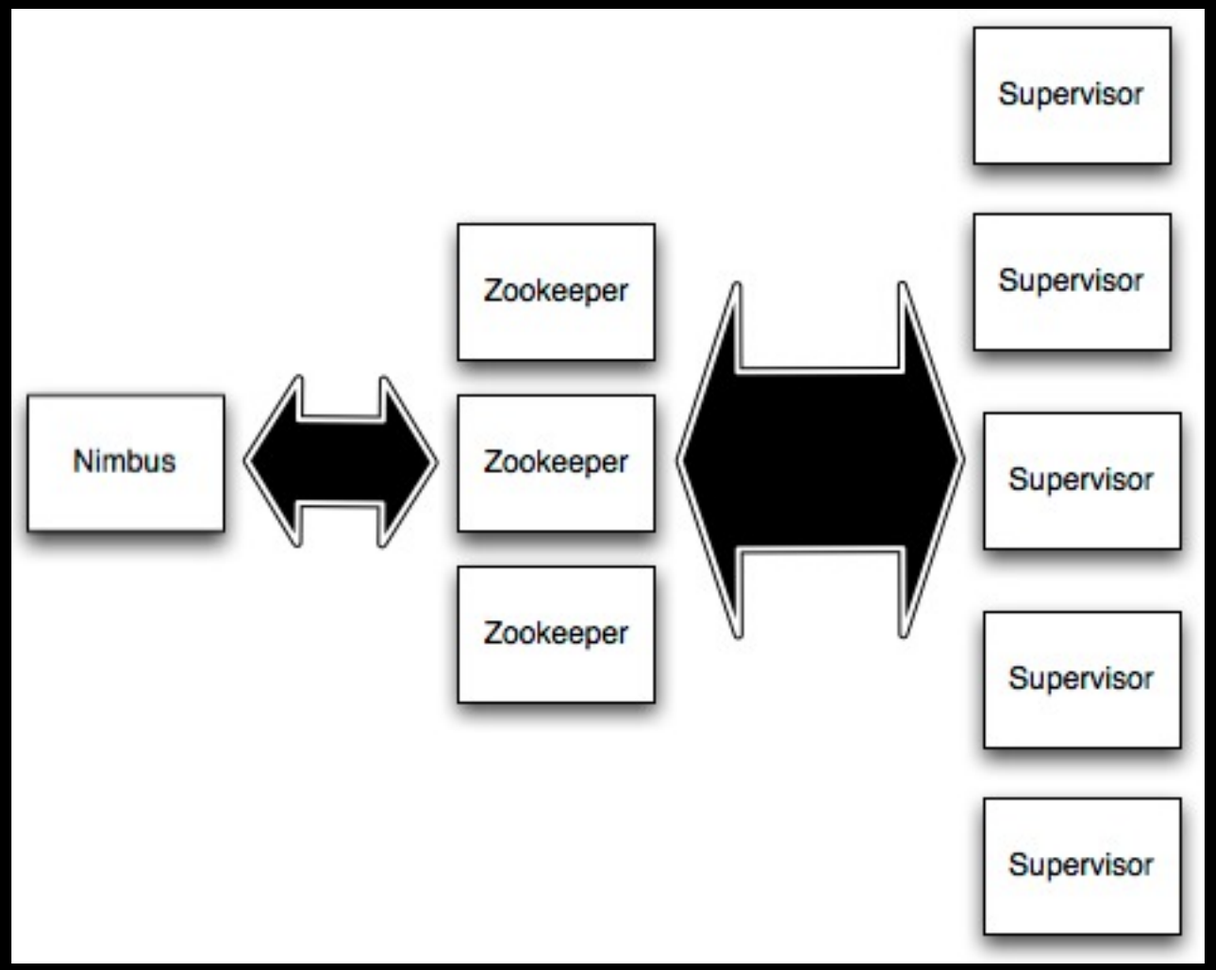


Distributed RPC

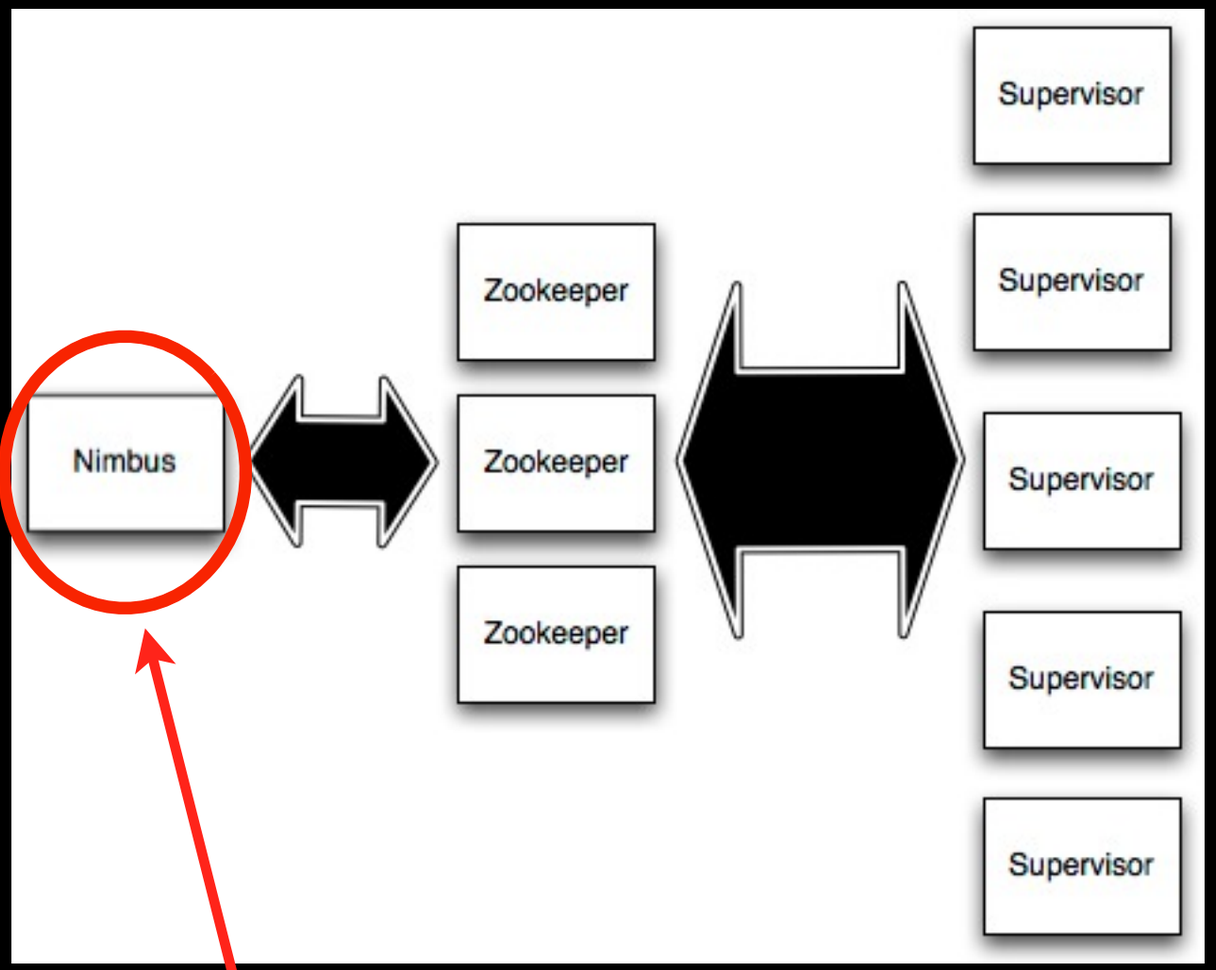


Continuous computation

Storm Cluster

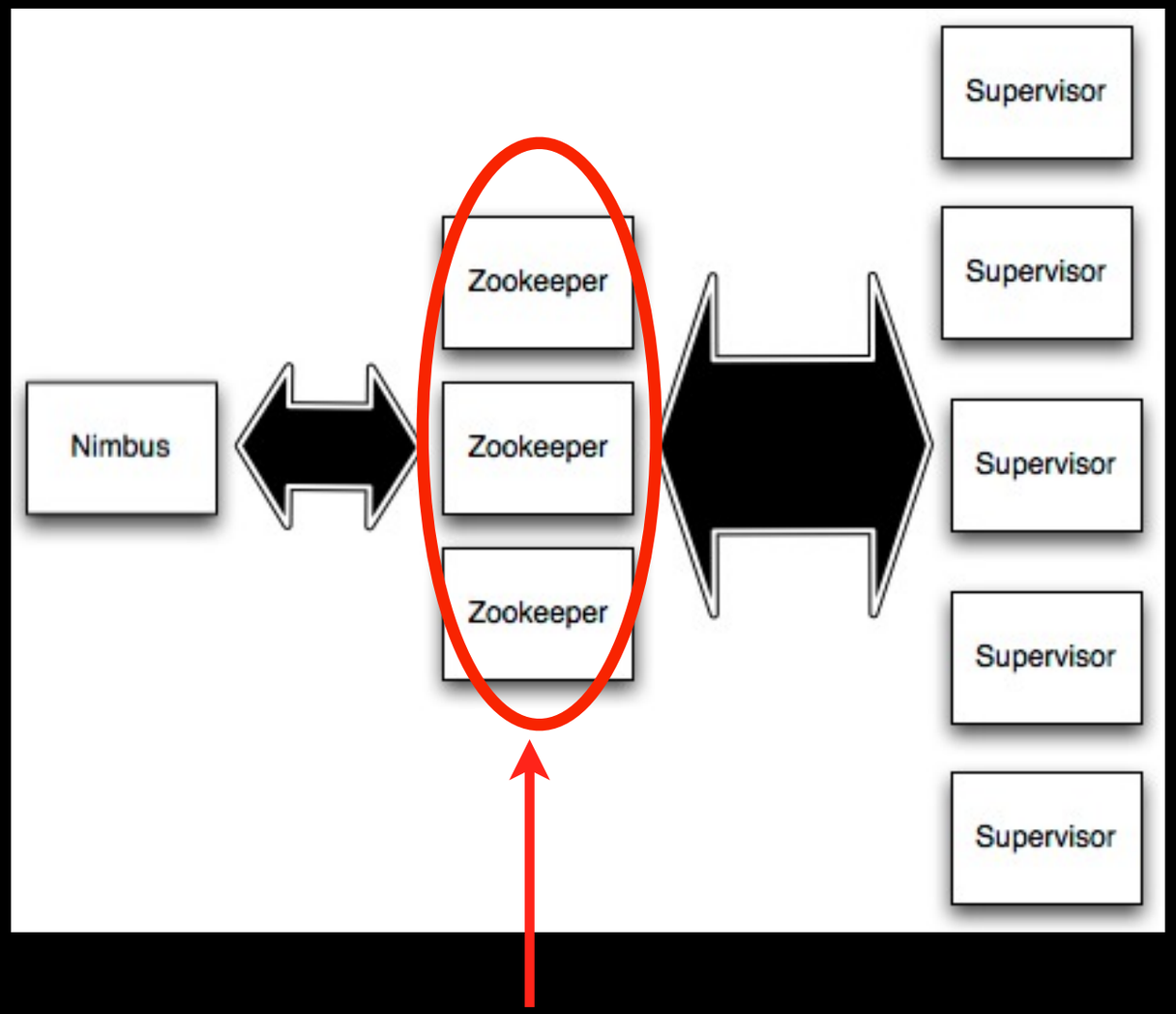


Storm Cluster



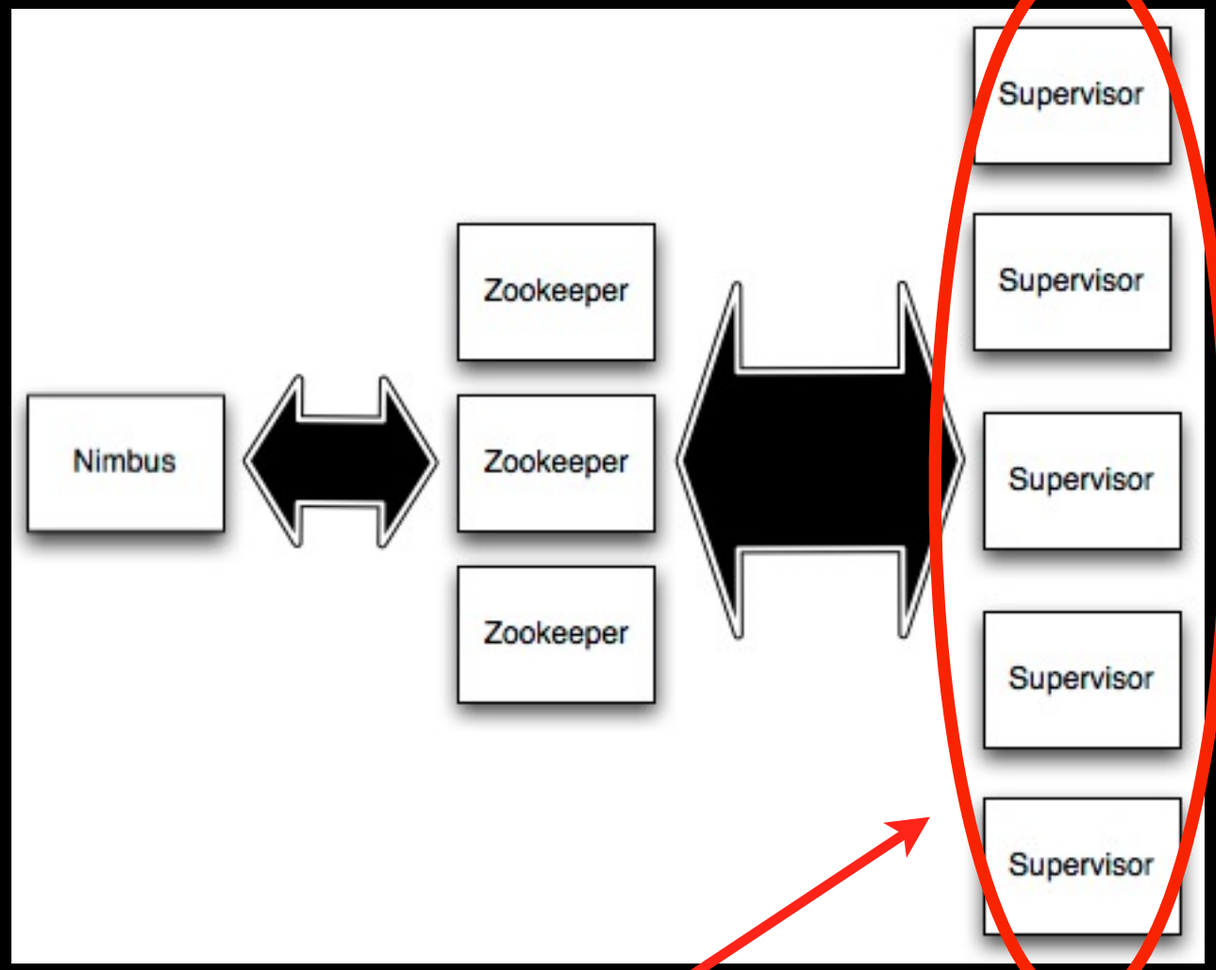
Master node (similar to Hadoop JobTracker)

Storm Cluster



Used for cluster coordination

Storm Cluster



Run worker processes

Starting a topology

```
storm jar mycode.jar twitter.storm.MyTopology demo
```

Killing a topology

```
storm kill demo
```

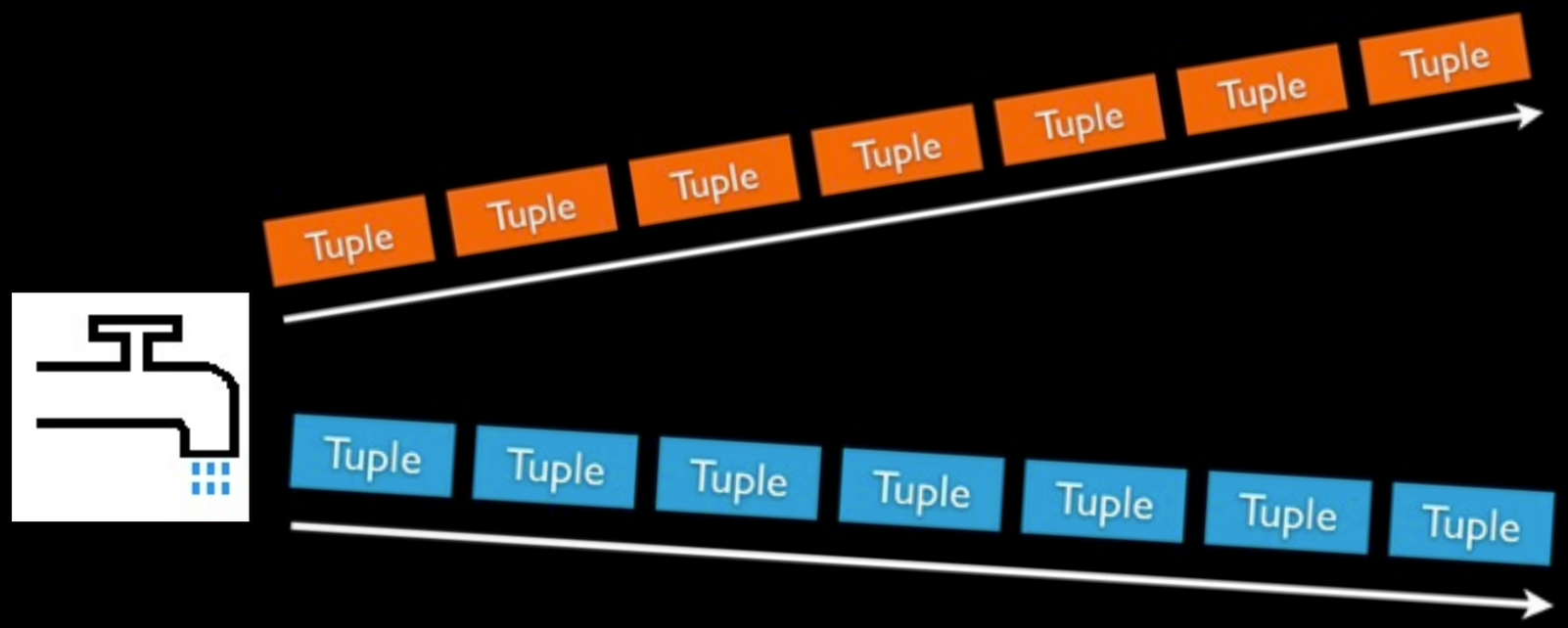
Concepts

- Streams
- Spouts
- Bolts
- Topologies

Streams

Unbounded sequence of tuples

Spouts

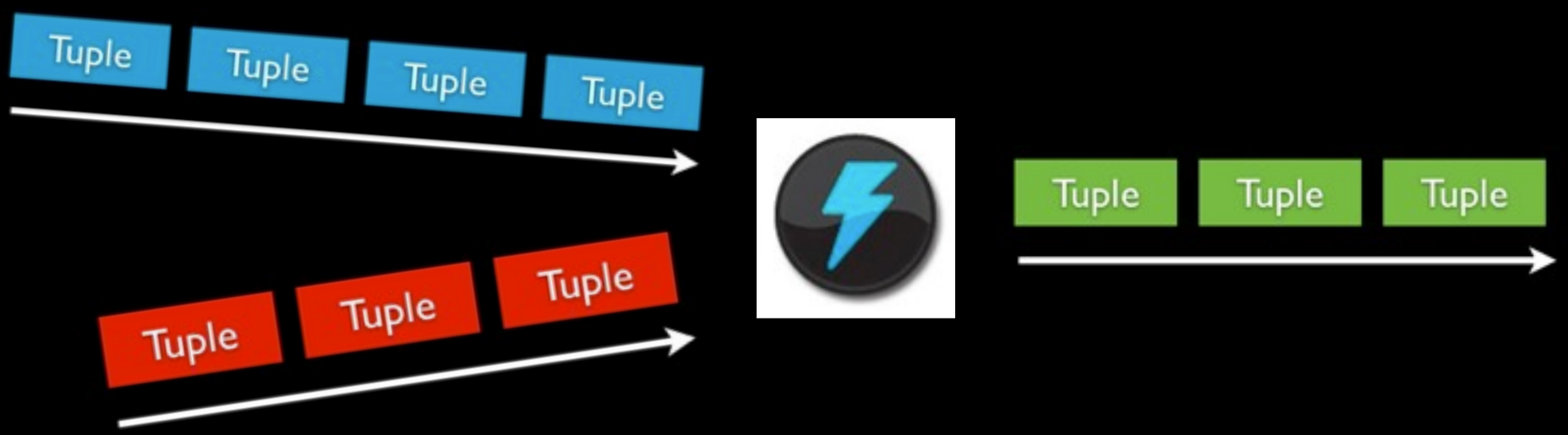


Source of streams

Spout examples

- Read from Kestrel queue
- Read from Twitter streaming API

Bolts

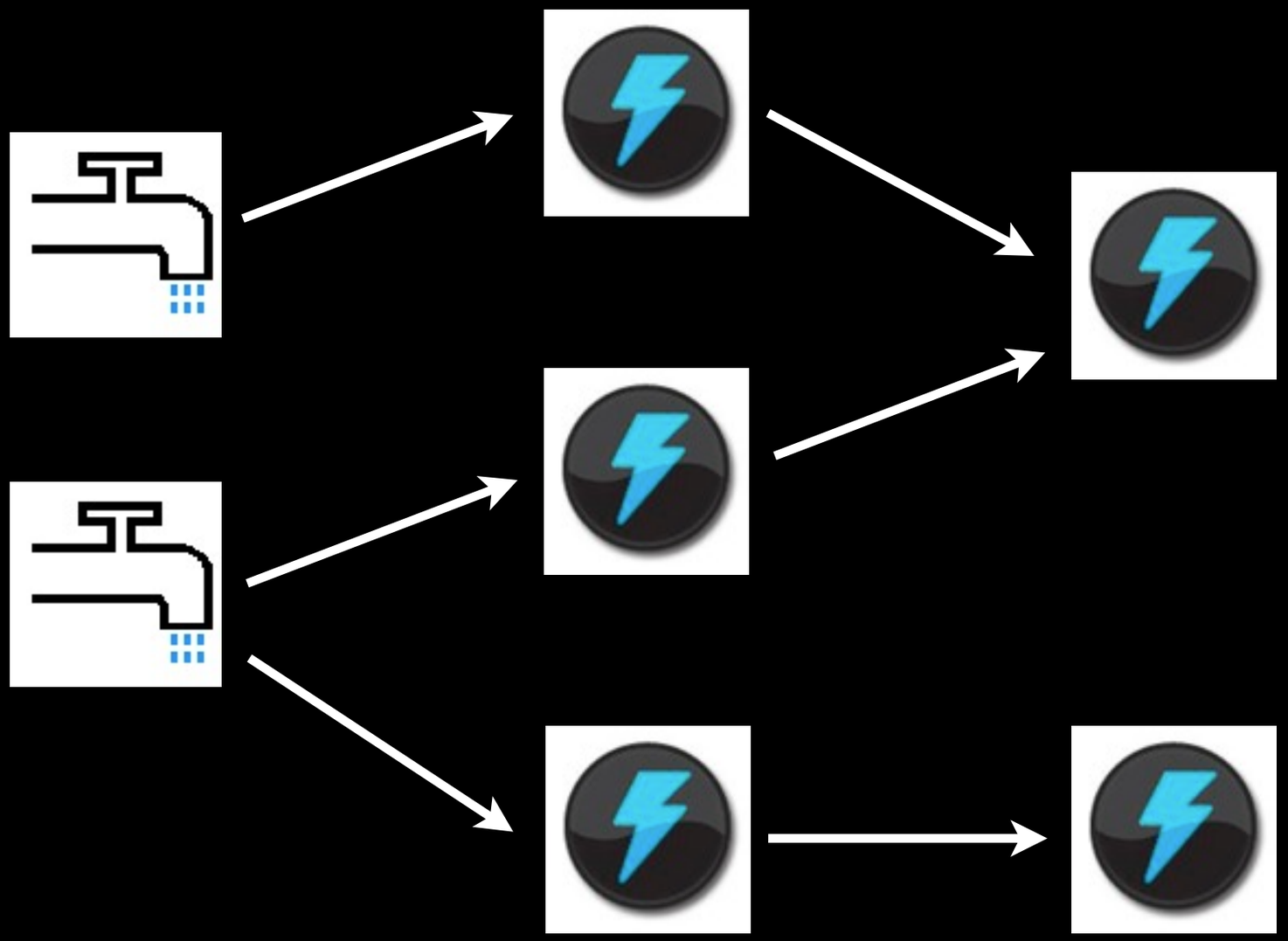


Processes input streams and produces new streams

Bolts

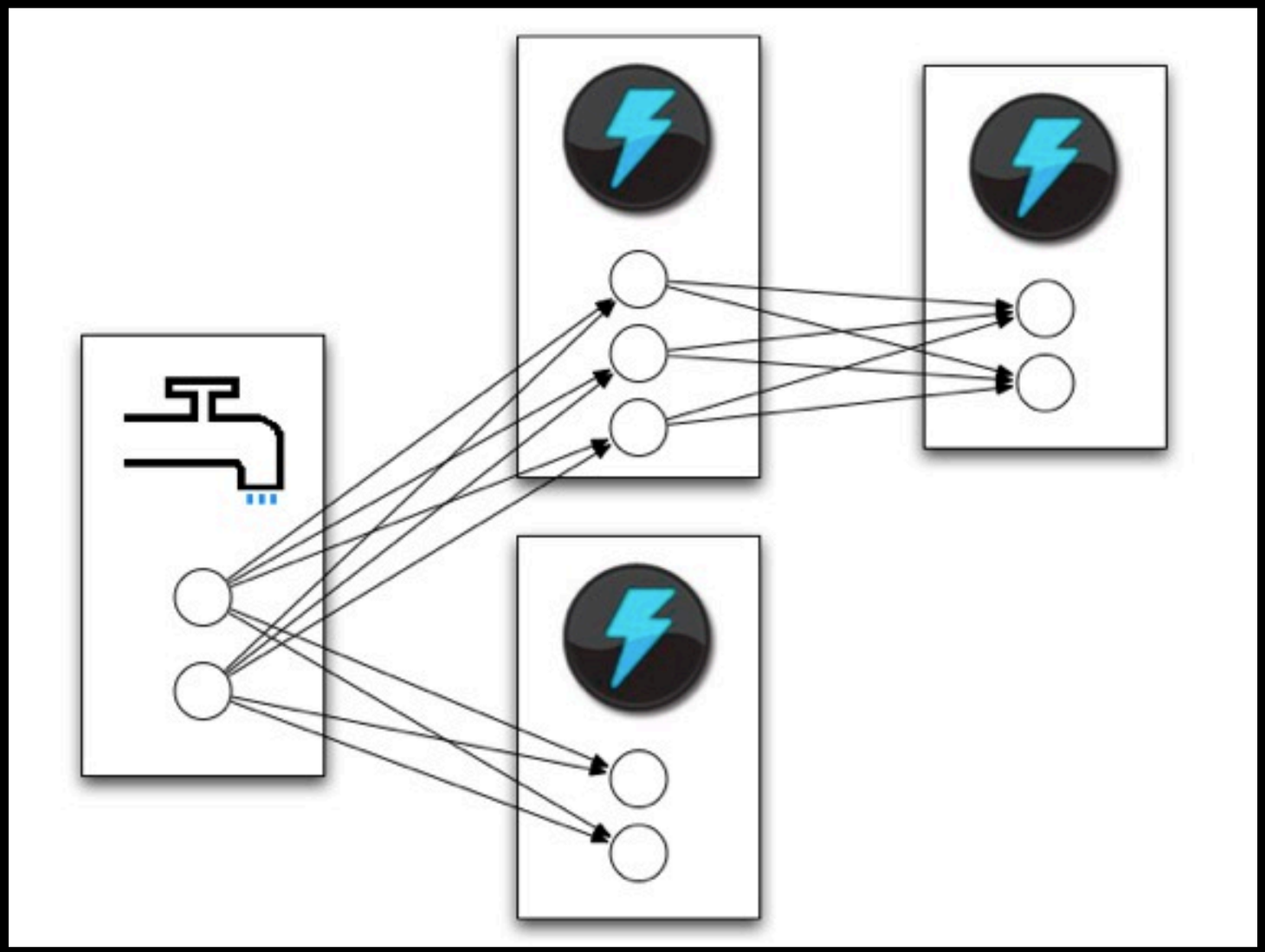
- Functions
- Filters
- Aggregation
- Joins
- Talk to databases

Topology



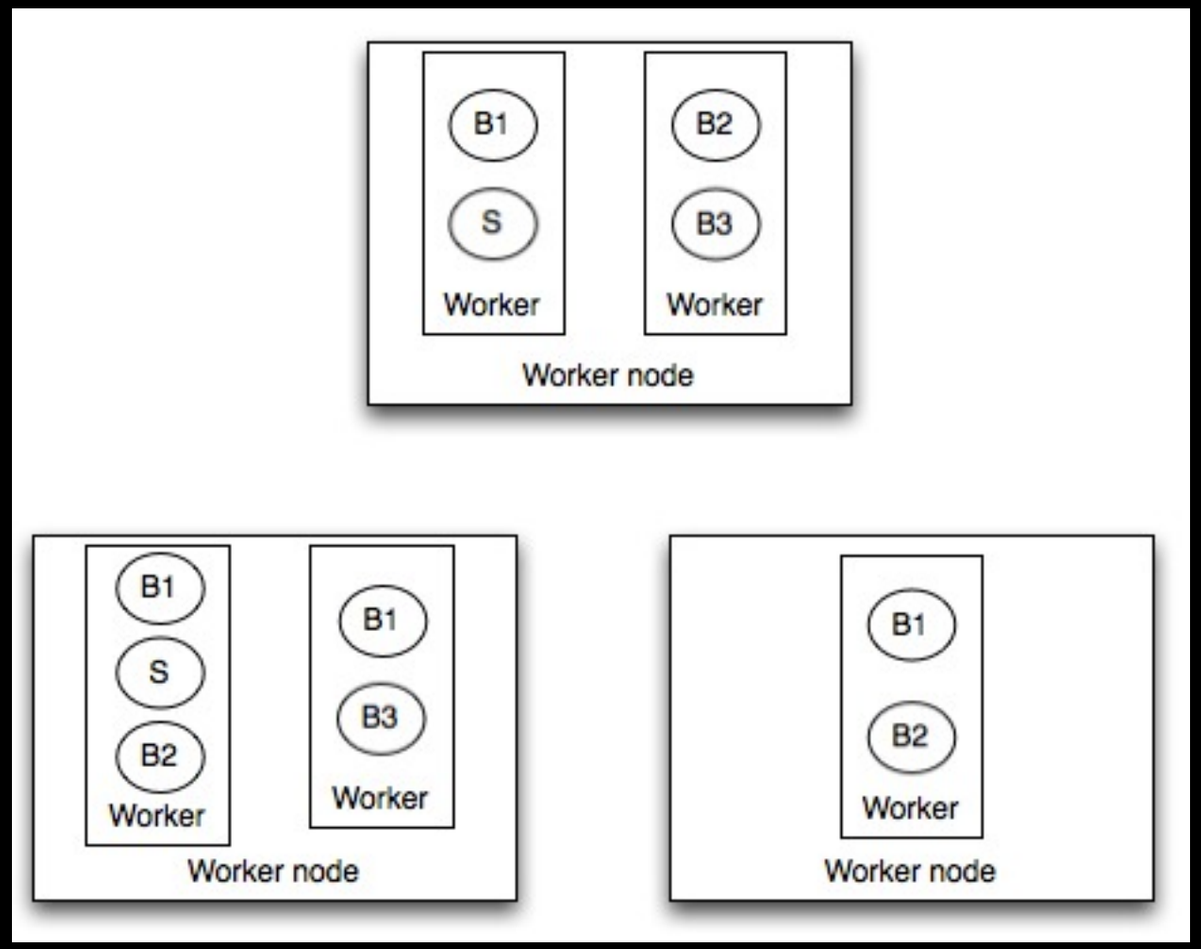
Network of spouts and bolts

Tasks



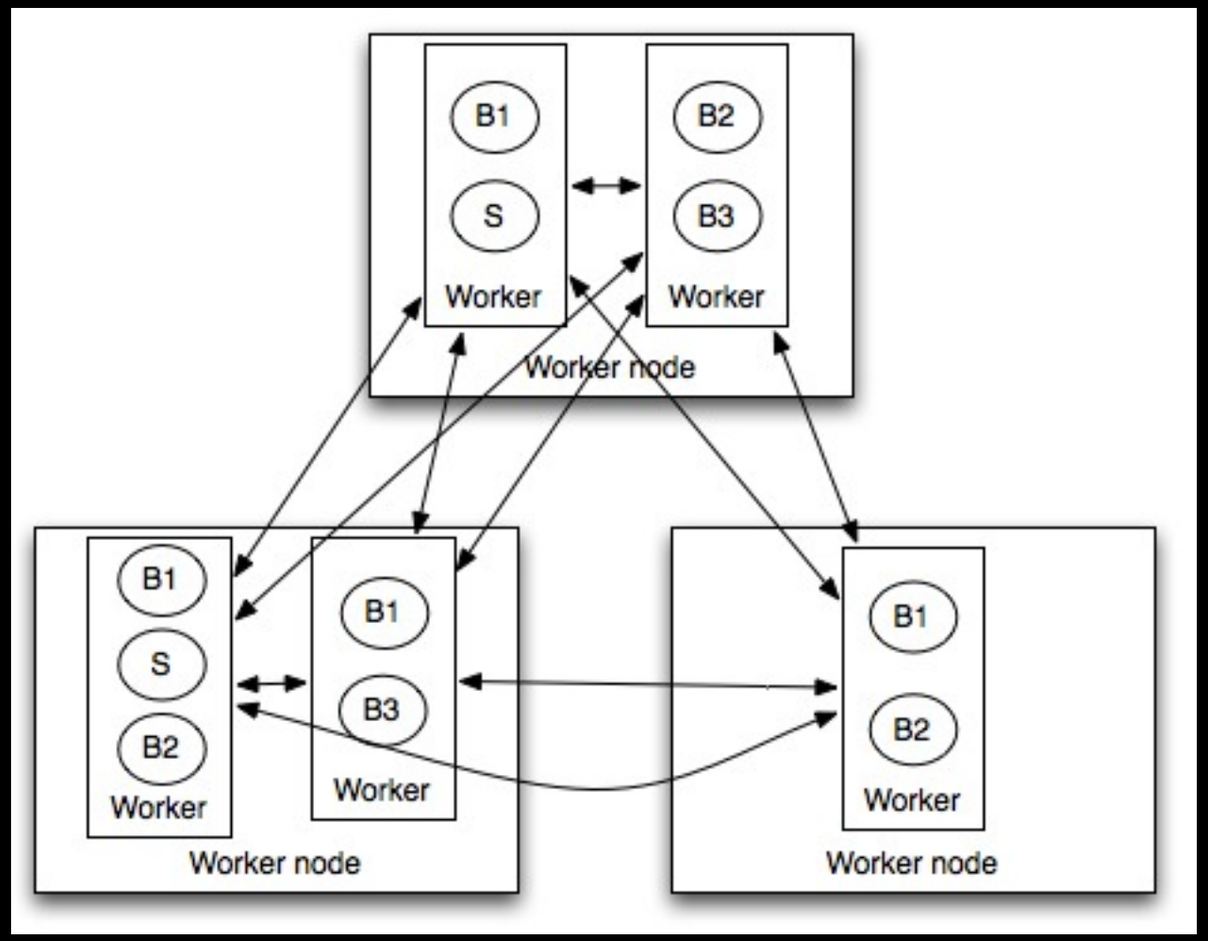
Spouts and bolts execute as
many tasks across the cluster

Task execution



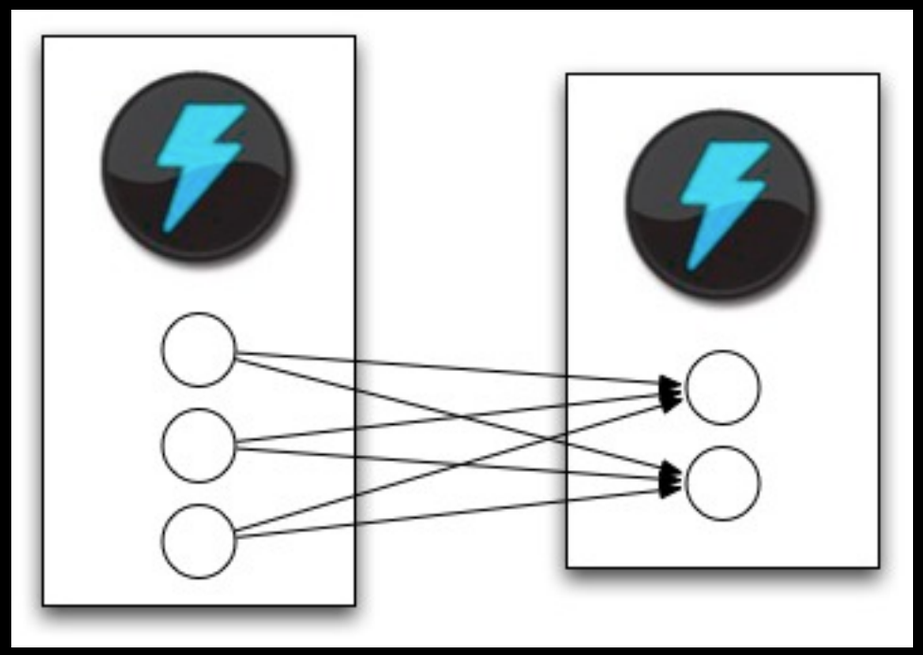
Tasks are spread across the cluster

Task execution



Tasks are spread across the cluster

Stream grouping

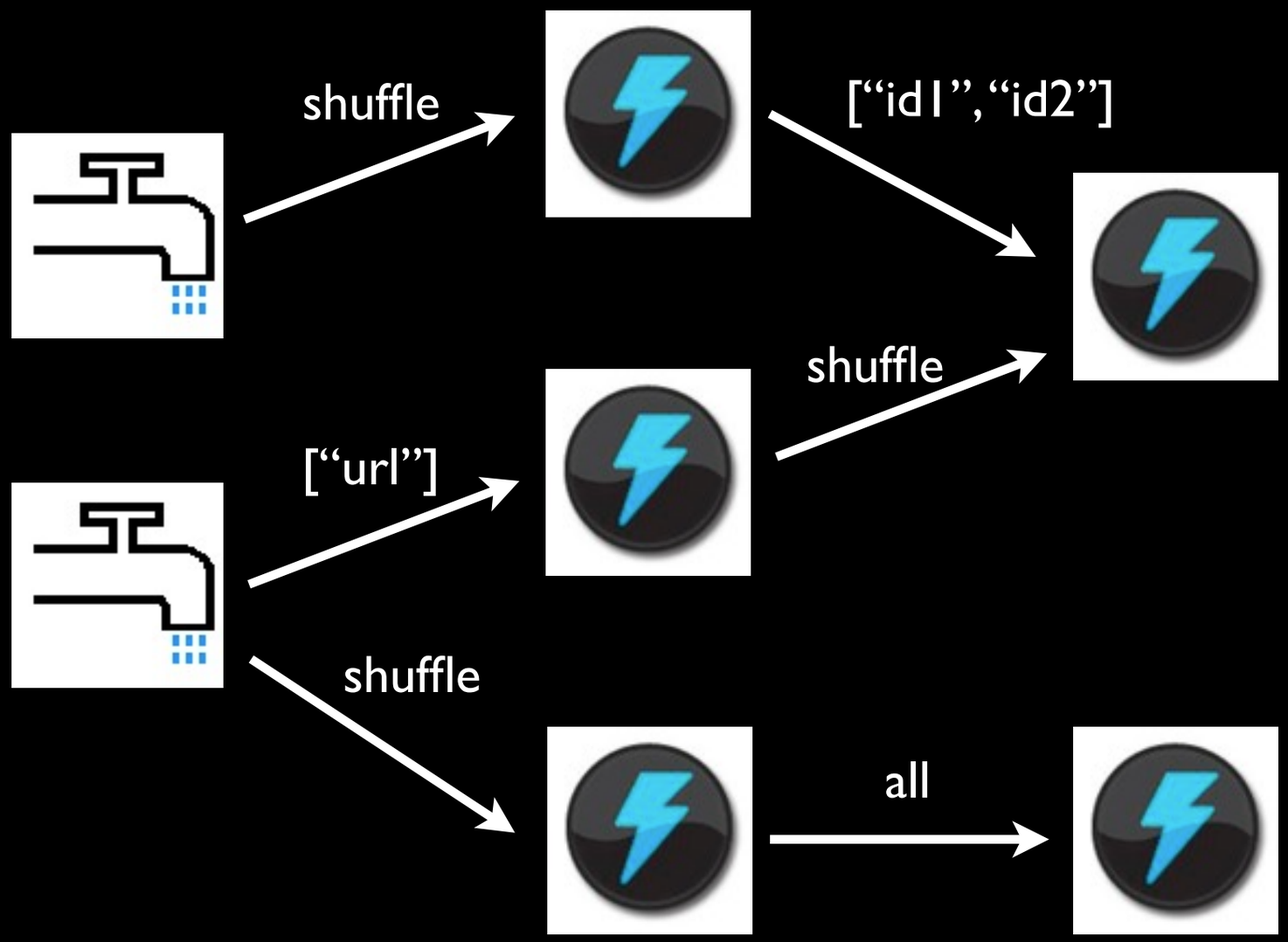


When a tuple is emitted, which task does it go to?

Stream grouping

- **Shuffle grouping:** pick a random task
- **Fields grouping:** mod hashing on a subset of tuple fields
- **All grouping:** send to all tasks
- **Global grouping:** pick task with lowest id

Topology



Streaming word count

```
TopologyBuilder builder = new TopologyBuilder();
```

TopologyBuilder is used to construct topologies in Java

Streaming word count

```
builder.setSpout("spout",
                 new KestrelSpout(
                     "kestrel.twitter.com"
                     22133,
                     "sentence_queue"),
                 5);
```

Define a spout in the topology with parallelism of 5 tasks

Streaming word count

```
builder.setBolt("split", new SplitSentence(), 8)
        .shuffleGrouping("spout");
```

Split sentences into words with parallelism of 8 tasks

Streaming word count

```
builder.setBolt("split", new SplitSentence(), 8)
        .shuffleGrouping("spout");
```

Consumer decides what data it receives and how it gets grouped

Split sentences into words with parallelism of 8 tasks

Streaming word count

```
builder.setBolt("count", new WordCount(), 12)
        .fieldsGrouping("split", new Fields("word"));
```

Create a word count stream

Streaming word count

```
public static class SplitSentence extends ShellBolt implements IRichBolt {
    public SplitSentence() {
        super("python", "splitsentence.py");
    }

    public void declareOutputFields(OutputFieldsDeclarer declarer) {
        declarer.declare(new Fields("word"));
    }
}
```

```
import storm

class SplitSentenceBolt(storm.BasicBolt):
    def process(self, tup):
        words = tup.values[0].split(" ")
        for word in words:
          storm.emit([word])
```

splitsentence.py

Streaming word count

```
public static class WordCount implements IBasicBolt {
    Map<String, Integer> counts = new HashMap<String, Integer>();

    public void prepare(Map conf, TopologyContext context) {
    }

    public void execute(Tuple tuple, BasicOutputCollector collector) {
        String word = tuple.getString(0);
        Integer count = counts.get(word);
        if(count==null) count = 0;
        count++;
        counts.put(word, count);
        collector.emit(new Values(word, count));
    }

    public void cleanup() {
    }

    public void declareOutputFields(OutputFieldsDeclarer declarer) {
        declarer.declare(new Fields("word", "count"));
    }
}
```

Streaming word count

```
Map conf = new HashMap();
conf.put(Config.TOPOLOGY_WORKERS, 10);

StormSubmitter.submitTopology("word-count", conf, builder.createTopology());
```

Submitting topology to a cluster

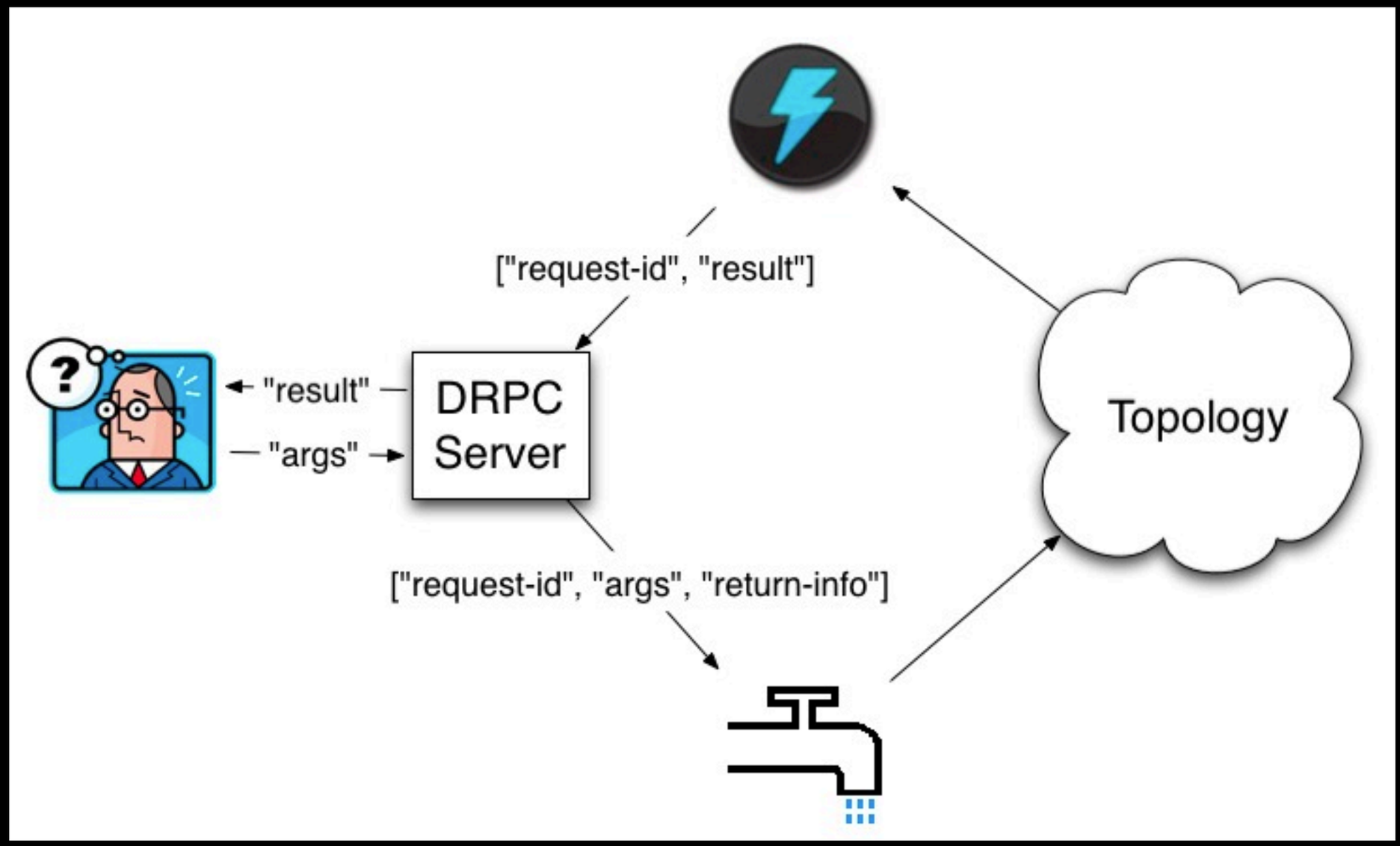
Streaming word count

```
LocalCluster cluster = new LocalCluster();

Map conf = new HashMap();
conf.put(Config.TOPOLOGY_DEBUG, true);

cluster.submitTopology("demo", conf, builder.createTopology());
```

Running topology in local mode


Demo

Distributed RPC



Data flow for Distributed RPC

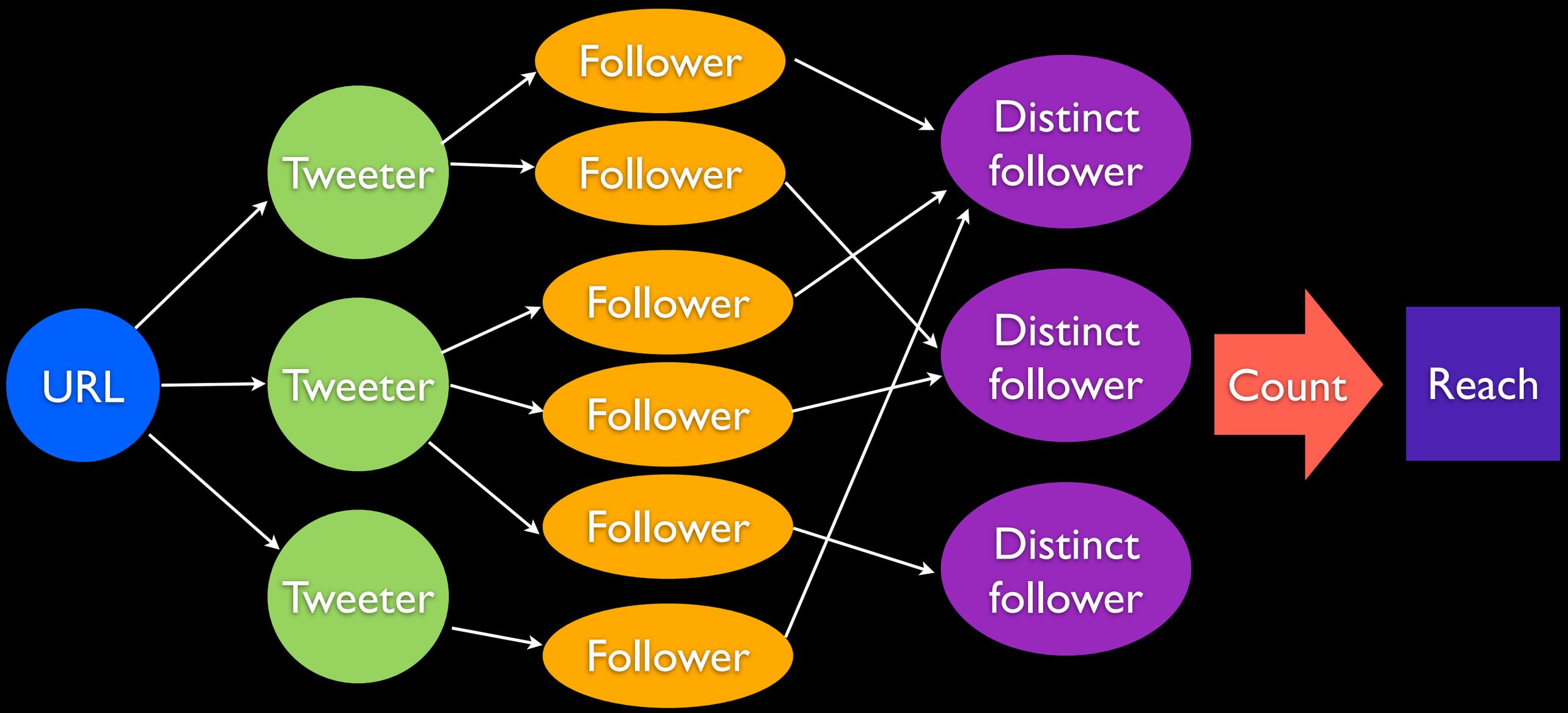
DRPC Example

Computing "reach" of a URL on the fly

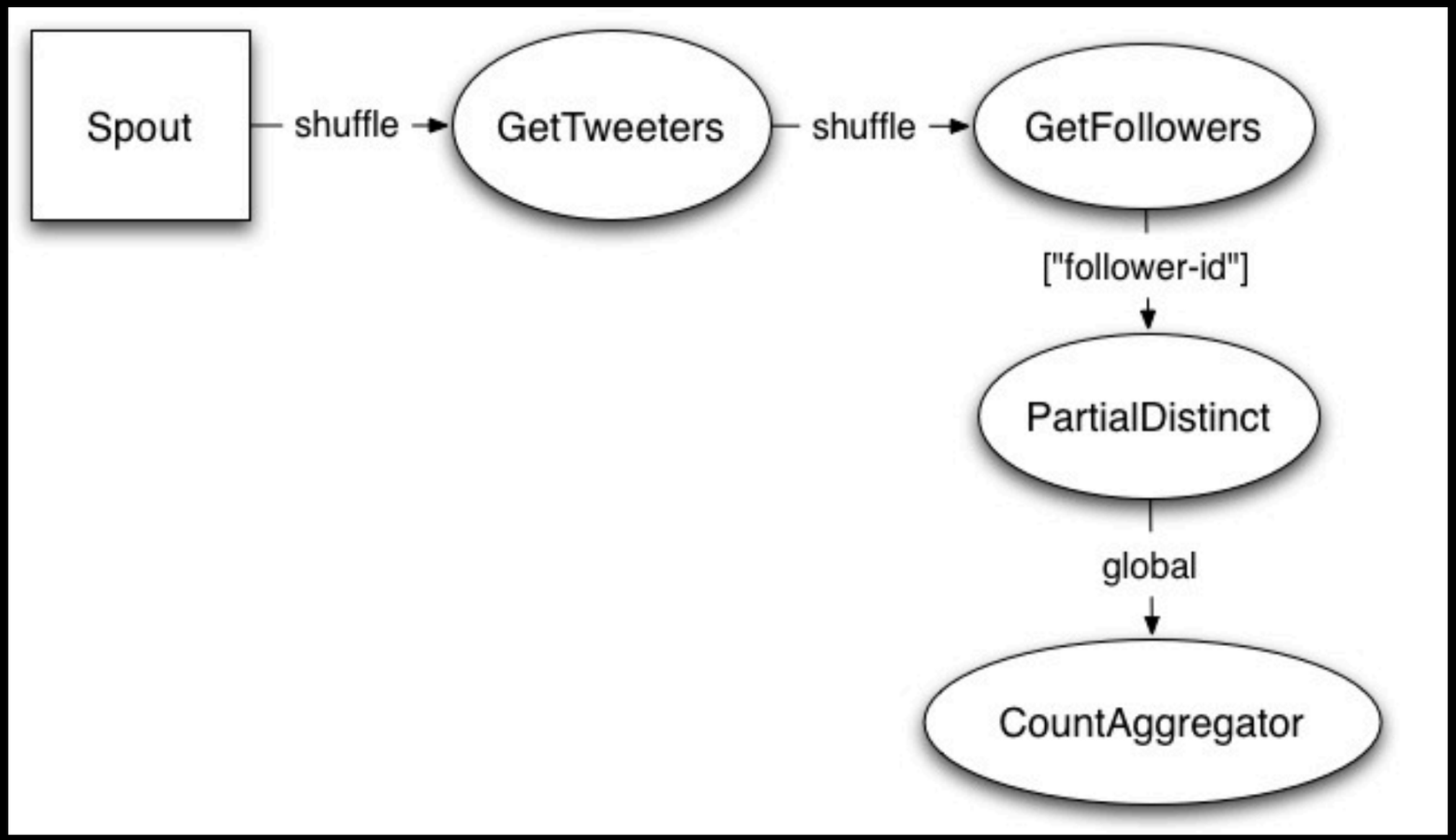
Reach

Reach is the number of unique people exposed to a URL on Twitter

Computing reach



Reach topology



Reach topology

```
LinearDRPCTopologyBuilder builder = new LinearDRPCTopologyBuilder("reach");
builder.addBolt(new GetTweeters(), 3);
builder.addBolt(new GetFollowers(), 12)
        .shuffleGrouping();
builder.addBolt(new PartialUniquer(), 6)
        .fieldsGrouping(new Fields("id", "follower"));
builder.addBolt(new CountAggregator(), 2)
        .fieldsGrouping(new Fields("id"));
```

Reach topology

```
public static class PartialUniquer extends BaseBatchBolt {
    BatchOutputCollector _collector;
    Object _id;
    Set<String> _followers = new HashSet<String>();

    public void prepare(Map conf, TopologyContext context,
                        BatchOutputCollector collector, Object id) {
        _collector = collector;
        _id = id;
    }

    public void execute(Tuple tuple) {
        _followers.add(tuple.getString(1));
    }

    public void finishBatch() {
        _collector.emit(new Values(_id, _followers.size()));
    }
```

Reach topology

```
public static class PartialUniquer extends BaseBatchBolt {
    BatchOutputCollector _collector;
    Object _id;
    Set<String> _followers = new HashSet<String>();

    public void prepare(Map conf, TopologyContext context,
                        BatchOutputCollector collector, Object id) {
        _collector = collector;
        _id = id;
    }

    public void execute(Tuple tuple) {
        _followers.add(tuple.getString(1));
    }

    public void finishBatch() {
        _collector.emit(new Values(_id, _followers.size()));
    }
```

Keep set of followers for each request id in memory

Reach topology

```
public static class PartialUniquer extends BaseBatchBolt {
    BatchOutputCollector _collector;
    Object _id;
    Set<String> _followers = new HashSet<String>();

    public void prepare(Map conf, TopologyContext context,
                        BatchOutputCollector collector, Object id) {
        _collector = collector;
        _id = id;
    }

    public void execute(Tuple tuple) {
        _followers.add(tuple.getString(1));
    }

    public void finishBatch() {
        _collector.emit(new Values(_id, _followers.size()));
    }
```

Update followers set when receive a new follower

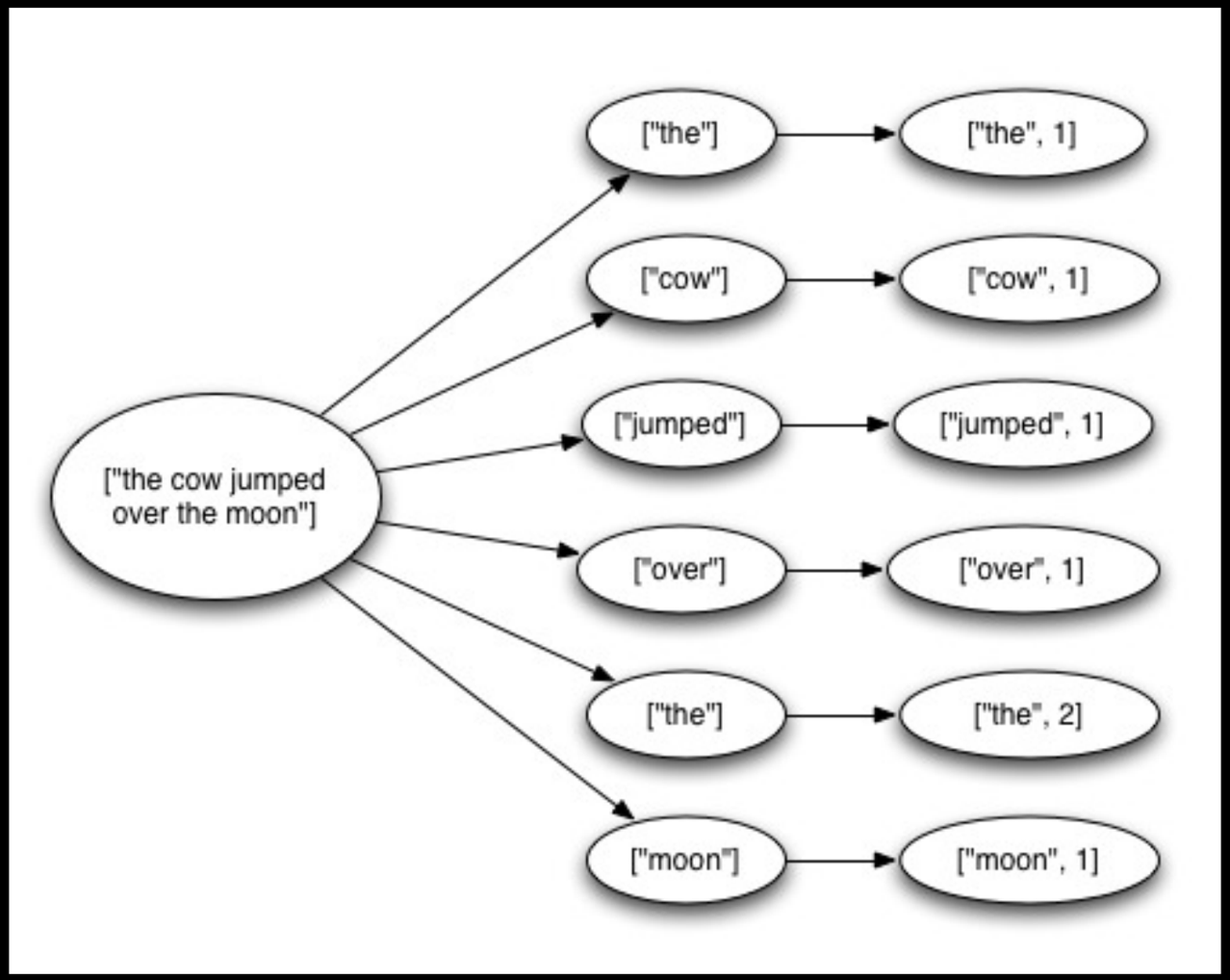
Reach topology

```
public static class PartialUniquer extends BaseBatchBolt {
    BatchOutputCollector _collector;
    Object _id;
    Set<String> _followers = new HashSet<String>();

    public void prepare(Map conf, TopologyContext context,
                        BatchOutputColle
        _collector = collector;
        _id = id;
    }

    public void execute(Tuple tuple) {
        _followers.add(tuple.getString(1));
    }

    public void finishBatch() {
        _collector.emit(new Values(_id, _followers.size()));
    }
```

Emit partial count after receiving all followers for a request id


Demo

Guaranteeing message processing



"Tuple tree"

Guaranteeing message processing

- A spout tuple is not fully processed until all tuples in the tree have been completed

Guaranteeing message processing

- If the tuple tree is not completed within a specified timeout, the spout tuple is replayed

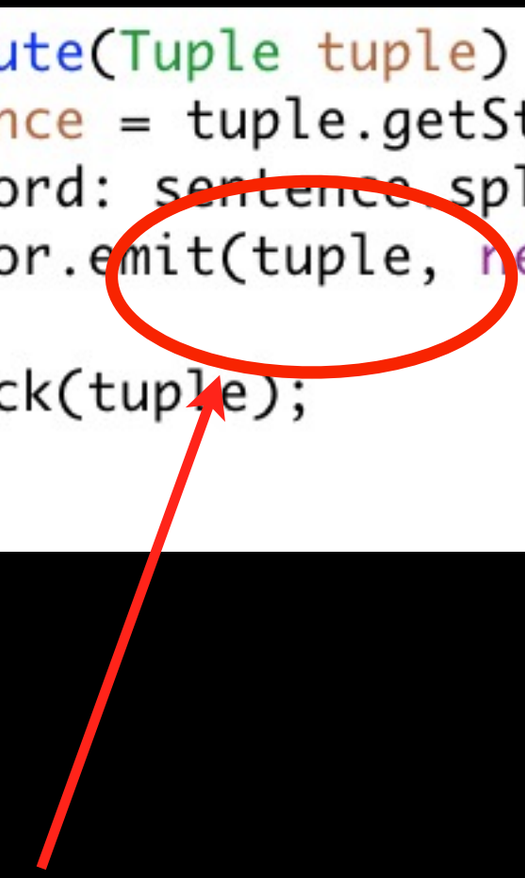
Guaranteeing message processing

```
public void execute(Tuple tuple) {
    String sentence = tuple.getString(0);
    for(String word: sentence.split(" ")) {
        _collector.emit(tuple, new Values(word));
    }
    _collector.ack(tuple);
}
```

Reliability API

Guaranteeing message processing

```
public void execute(Tuple tuple) {
    String sentence = tuple.getString(0);
    for(String word: sentence.split(" ")) {
        _collector.emit(tuple, new Values(word));
    }
    _collector.ack(tuple);
}
```



"Anchoring" creates a new edge in the tuple tree

Guaranteeing message processing

```
public void execute(Tuple tuple) {
    String sentence = tuple.getString(0);
    for(String word: sentence.split(" ")) {
        _collector.emit(tuple, new Values(word));
    }
    _collector.ack(tuple);
}
```

Marks a single node in the tree as complete

Guaranteeing message processing

- Storm tracks tuple trees for you in an extremely efficient way

Transactional topologies

How do you do idempotent counting with an at least once delivery guarantee?

Transactional topologies

Won’t you overcount?

Transactional topologies

Transactional topologies solve this problem

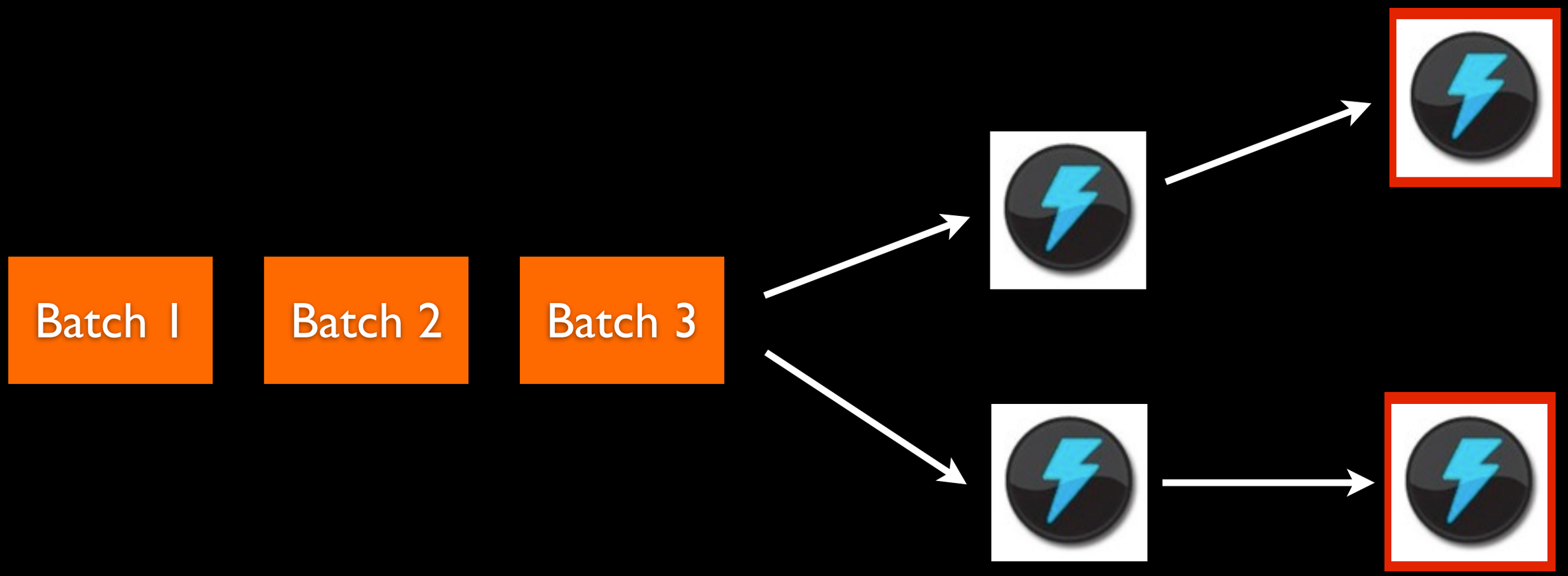
Transactional topologies

Built completely on top of Storm's primitives
of streams, spouts, and bolts

Transactional topologies

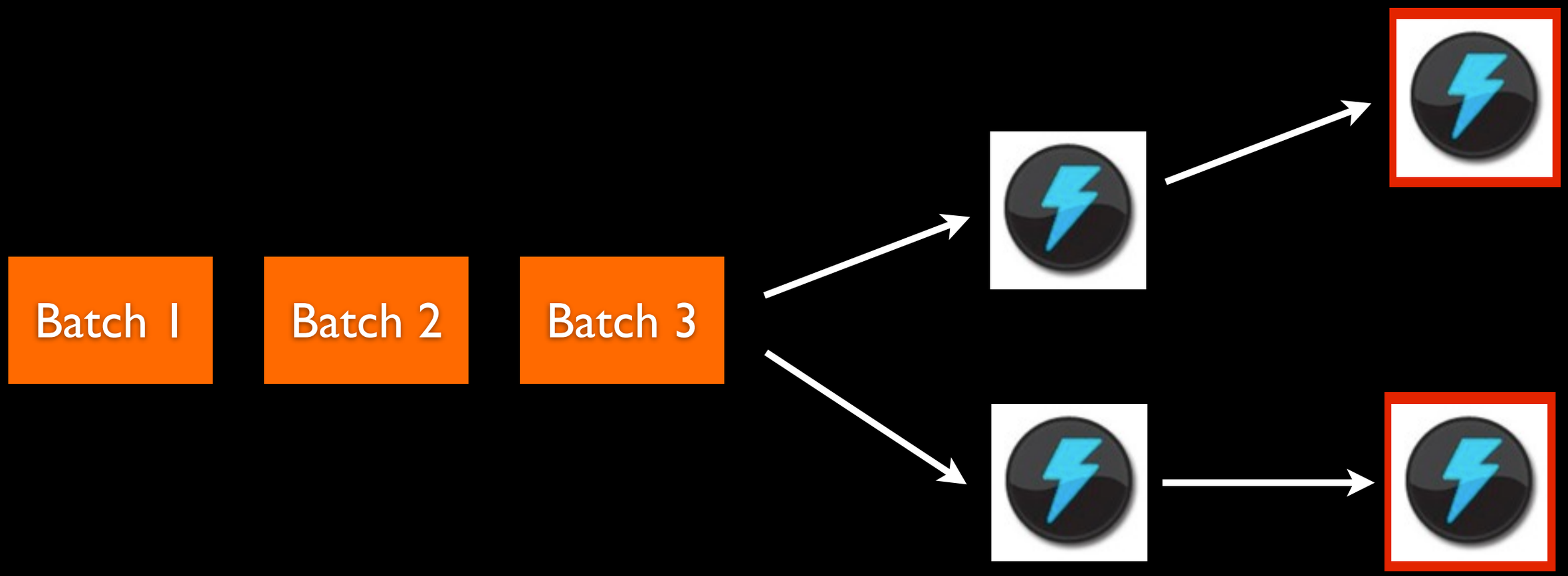
Enables fault-tolerant, exactly-once messaging semantics

Transactional topologies



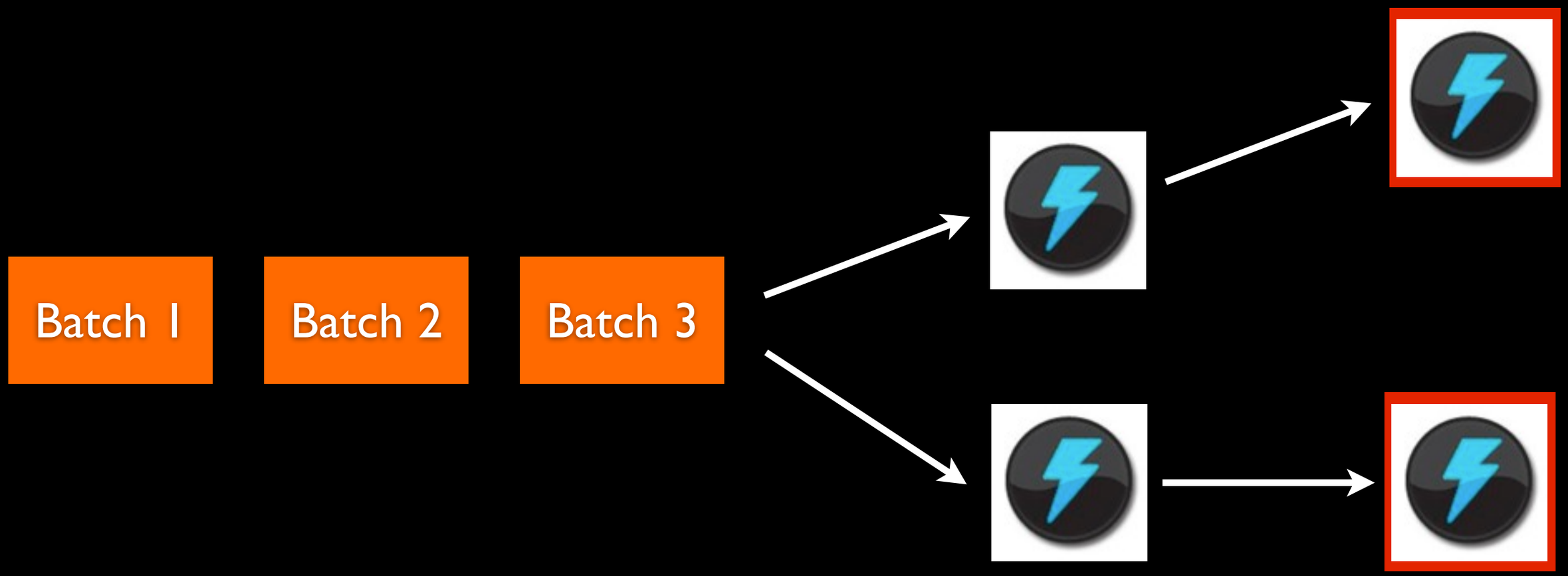
Process small batches of tuples

Transactional topologies



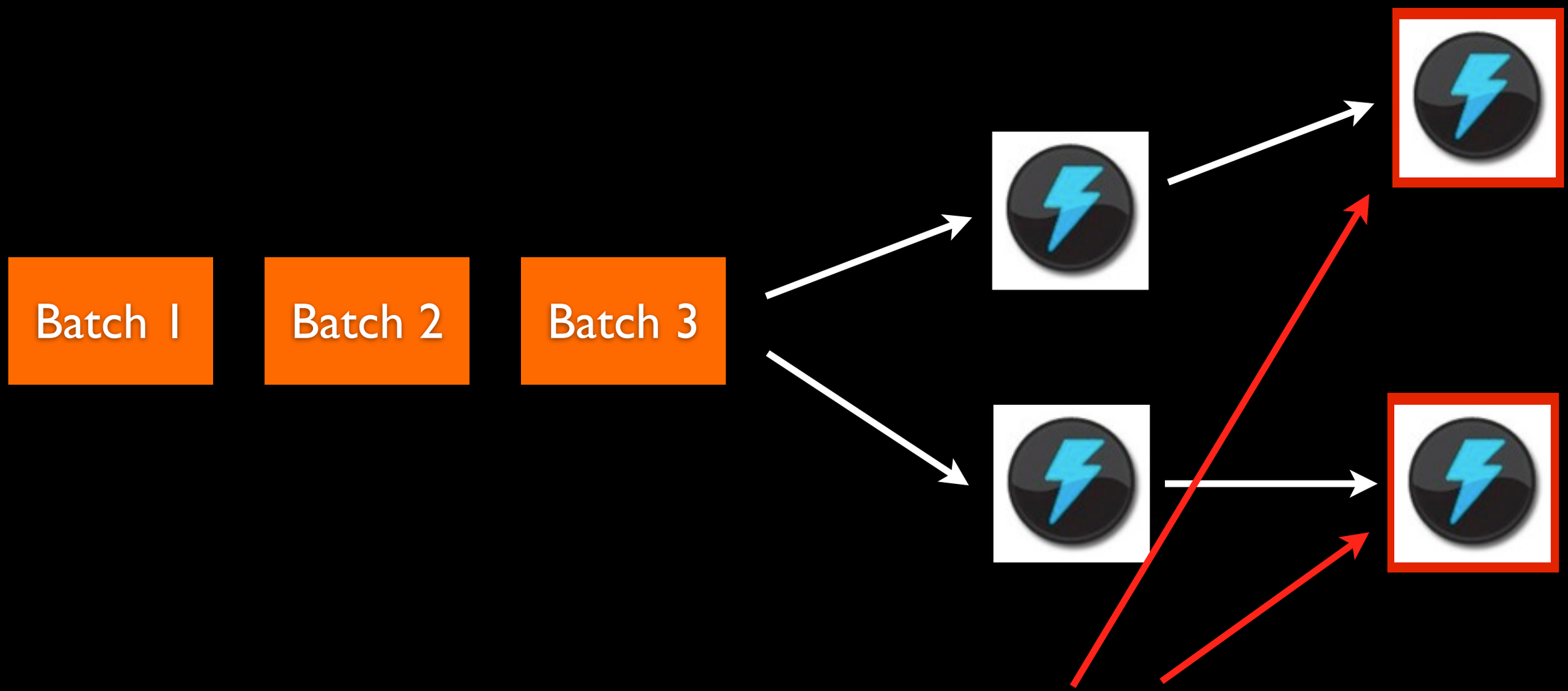
If a batch fails, replay the whole batch

Transactional topologies



Once a batch is completed, commit the batch

Transactional topologies



Bolts can optionally be “committers”

Transactional topologies

Commit 1 | Commit 1 | Commit 2 | Commit 3 | Commit 4 | Commit 4

Commits are ordered. If there's a failure during commit, the whole batch + commit is retried

Example

```
public class IdempotentCountingBolt extends BaseTransactionalBolt
                                    implements ICommitter {
    TransactionAttempt _attempt;
    BatchOutputCollector _collector;
    int _count = 0;

    public void prepare(Map conf, TopologyContext context,
            BatchOutputCollector collector,
            TransactionAttempt attempt) {
        _collector = collector;
        _attempt = attempt;
    }

    public void execute(Tuple tuple) {
        _count += 1;
    }

    public void finishBatch() {
        CurrentValue current = getCurrentValue();
        if(current.txid!=_attempt.getTransactionId()) {
            setCurrentValue(current.count + _count, _attempt.getTransactionId());
        }
    }
```

Example

```
public class IdempotentCountingBolt extends BaseTransactionalBolt
                                    implements ICommitter {
    TransactionAttempt _attempt;
    BatchOutputCollector _collector;
    int _count = 0;

    public void prepare(Map conf, TopologyContext co
            BatchOutputCollector collector,
            TransactionAttempt attempt) {
        _collector = collector;
        _attempt = attempt;
    }

    public void execute(Tuple tuple) {
        _count += 1;
    }

    public void finishBatch() {
        CurrentValue current = getCurrentValue();
        if(current.txid!=_attempt.getTransactionId()) {
            setCurrentValue(current.count + _count, _attempt.getTransactionId());
        }
    }
```

New instance of this object for every transaction attempt

Example

```
public class IdempotentCountingBolt extends BaseTransactionalBolt
                                    implements ICommitter {
    TransactionAttempt _attempt;
    BatchOutputCollector _collector;
    int _count = 0;

    public void prepare(Map conf, TopologyContext context,
            BatchOutputCollector collector,
            TransactionAttempt attempt) {
        _collector = collector;
        _attempt = attempt;
    }

    public void execute(Tuple tuple) {
        _count += 1;
    }

    public void finishBatch() {
        CurrentValue current = getCurrentValue();
        if(current.txid!=_attempt.getTransactionId()) {
            setCurrentValue(current.count + _count, _attempt.getTransactionId());
        }
    }
```

Aggregate the count for this batch

Example

```
public class IdempotentCountingBolt extends BaseTransactionalBolt
                                    implements ICommitter {
    TransactionAttempt _attempt;
    BatchOutputCollector _collector;
    int _count = 0;

    public void prepare(Map conf, TopologyContext context,
            BatchOutputCollector collector,
            TransactionAttempt attempt) {
        _collector = collector;
        _attempt = attempt;
    }

    public void execute(Tuple tuple) {
        _count += 1;
    }

    public void finishBatch() {
        CurrentValue current = getCurrentValue();
        if(current.txid!=_attempt.getTransactionId()) {
            setCurrentValue(current.count + _count, _attempt.getTransactionId());
        }
    }
```

Only update database if transaction ids differ

Example

```
public class IdempotentCountingBolt extends BaseTransactionalBolt
                                    implements ICommitter {
    TransactionAttempt _attempt;
    BatchOutputCollector _collector;
    int _count = 0;

    public void prepare(Map conf, TopologyContext context,
            BatchOutputCollector collector,
            TransactionAttempt attempt) {
        _collector = collector;
        _attempt = attempt;
    }

    public void execute(Tuple tuple) {
        _count += 1;
    }

    public void finishBatch() {
        CurrentValue current = getCurrentValue();
        if(current.txid!=_attempt.getTransactionId()) {
            setCurrentValue(current.count + _count, _attempt.getTransactionId());
        }
    }
```

This enables idempotency since commits are ordered

Example

```
public class IdempotentCountingBolt extends BaseTransactionalBolt
                                    implements ICommitter {
    TransactionAttempt _attempt;
    BatchOutputCollector _collector;
    int _count = 0;

    public void prepare(Map conf, TopologyContext context,
            BatchOutputCollector collector,
            TransactionAttempt attempt) {
        _collector = collector;
        _attempt = attempt;
    }

    public void execute(Tuple tuple) {
        _count += 1;
    }

    public void finishBatch() {
        CurrentValue current = getCurrentValue();
        if(current.txid!=_attempt.getTransactionId()) {
            setCurrentValue(current.count + _count, _attempt.getTransactionId());
        }
    }
```

(Credit goes to Kafka devs for this trick)

Transactional topologies

Multiple batches can be processed in parallel,
but commits are guaranteed to be ordered

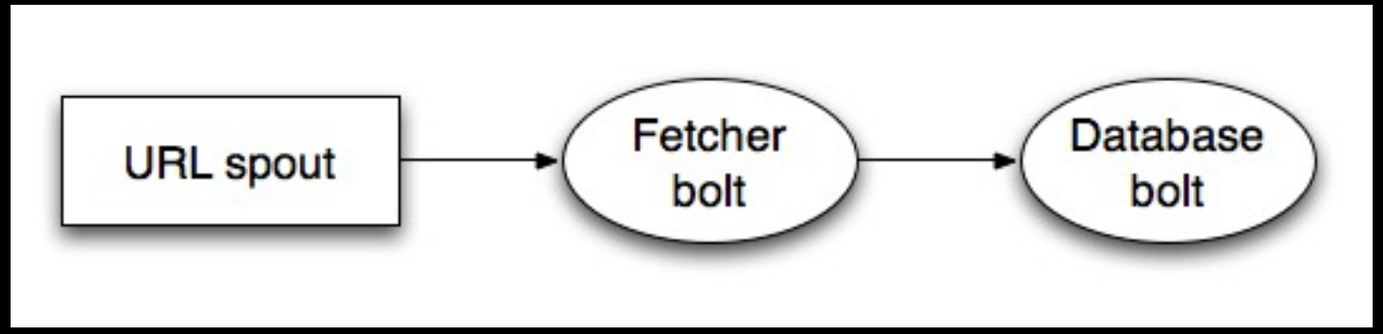
Transactional topologies

- Requires a more sophisticated source queue than Kestrel or RabbitMQ
- storm-contrib has a transactional spout implementation for Kafka

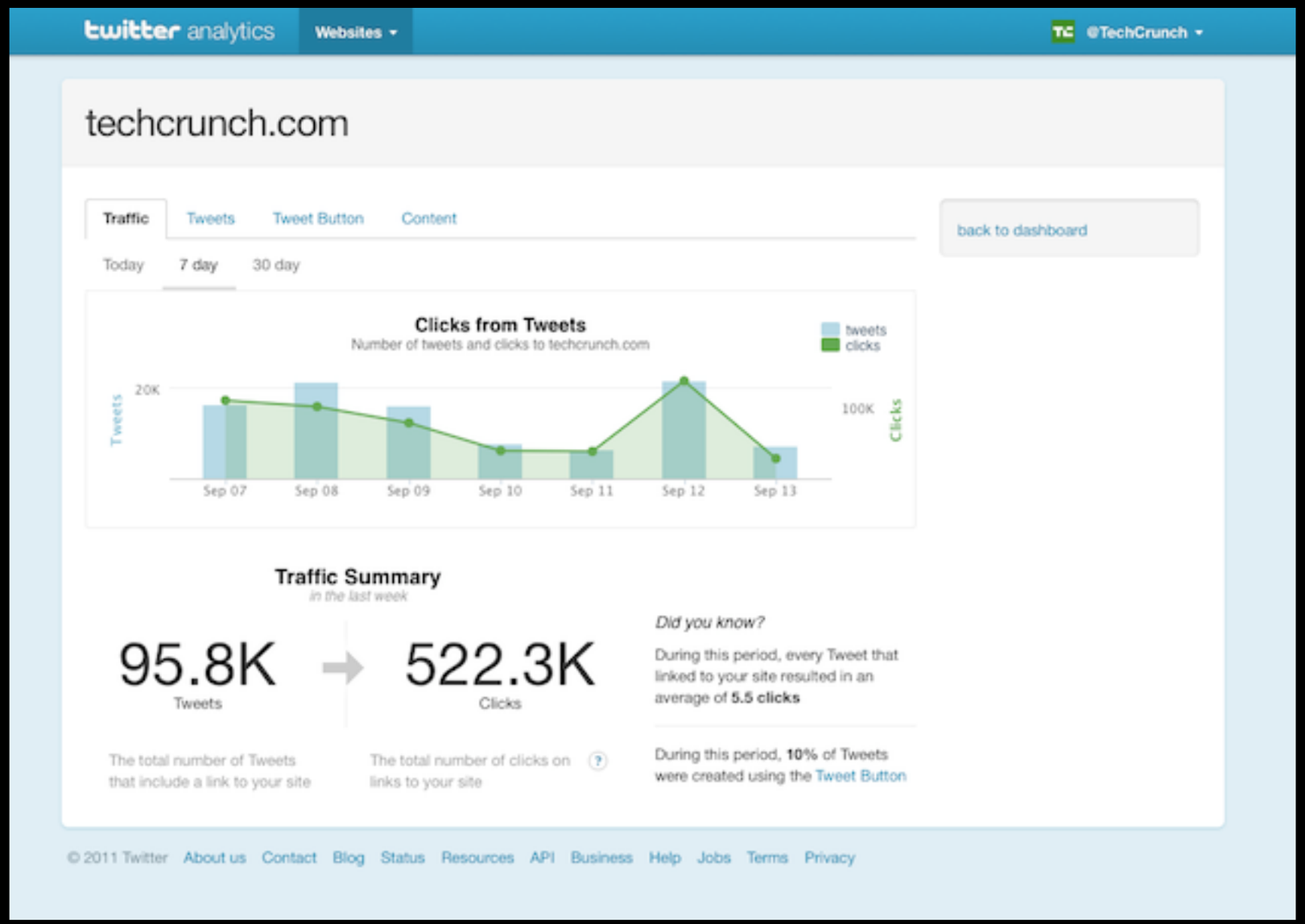
Example #1

Crawler

Example #1

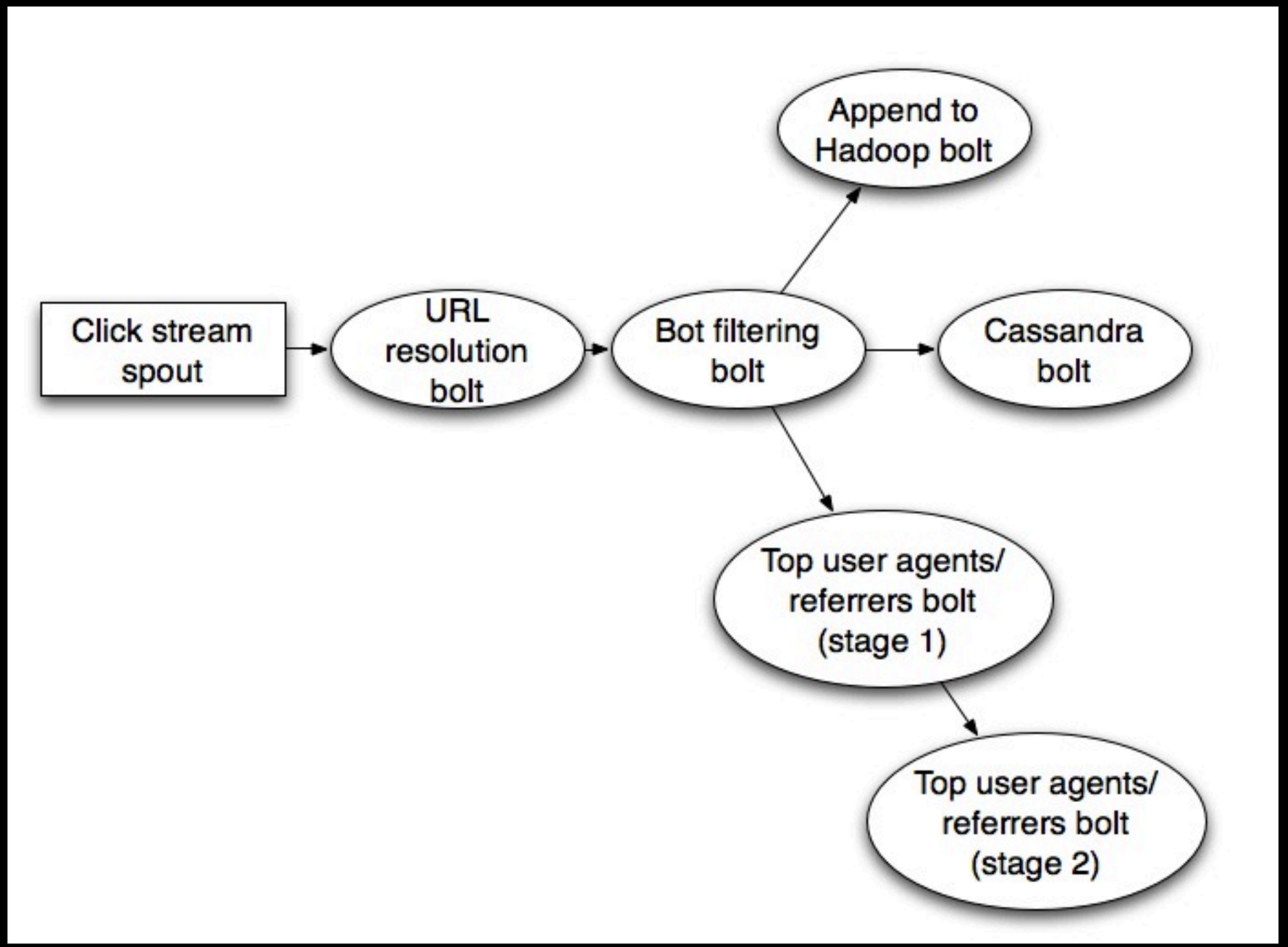


Example #2



Twitter Web Analytics

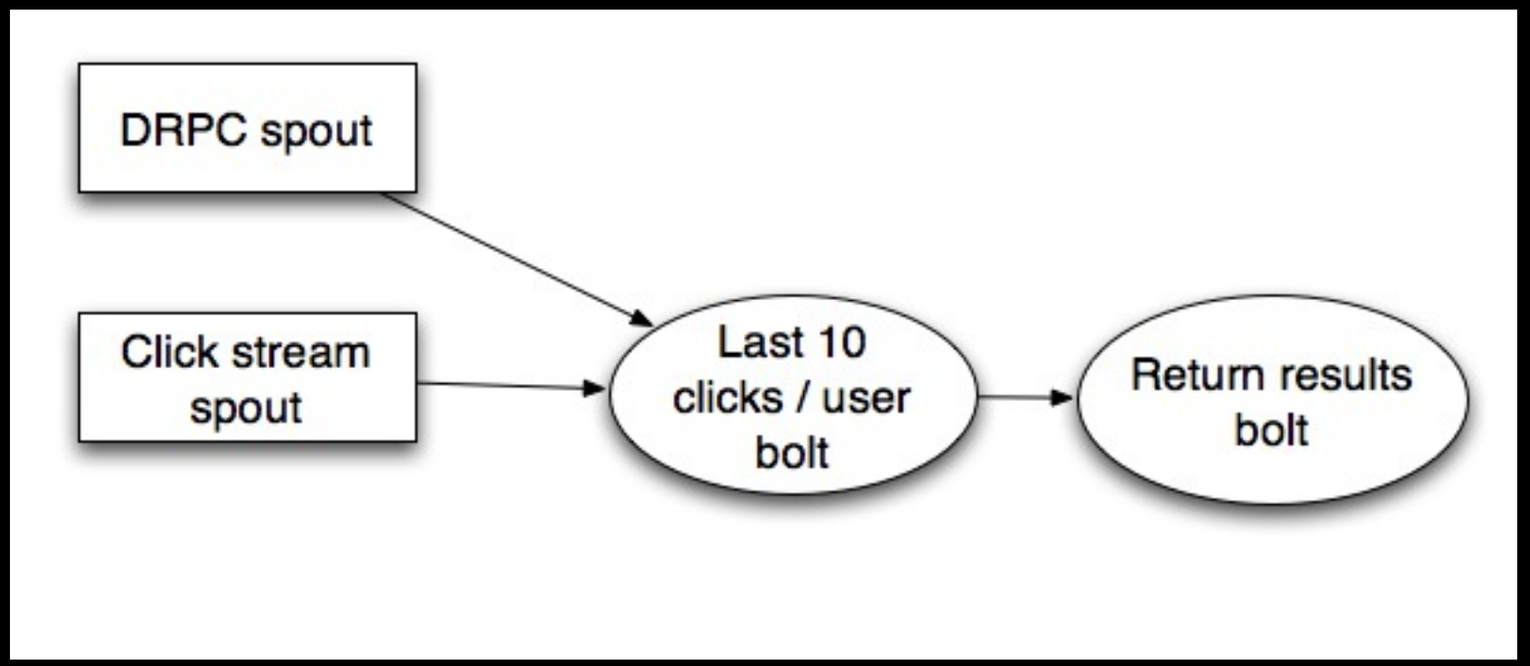
Example #2


Append to Hadoop bolt
Click stream spout
URL resolution bolt
Bot filtering bolt
Cassandra bolt
Top user agents/ referrers bolt (stage 1)
Top user agents/ referrers bolt (stage 2)

Example #3

Storm as a “database”

Example #3


DRPC spout
Click stream spout
Last 10 clicks / user bolt
Return results bolt

Storm UI

Storm UI

Topology summary

Name	Id	Uptime	Num workers	Num tasks
poseidon	poseidon-1-1314658150	23h 17m 0s	80	765

Topology stats

Window	Emitted	Transferred	Complete latency (ms)	Acked	Failed
10m 0s	24786020	24786000	4131.688	2338940	0
3h 0m 0s	621695800	621694600	4463.830	59353840	0
1d 0h 0m 0s	4447725560	4447716960	4278.459	438710100	0
All time	4447725560	4447716960	4278.459	438710100	0

Spouts (All time)

Id	Parallelism	Emitted	Transferred	Complete latency (ms)	Acked	Failed	Last error
1	160	877453060	877453060	4278.459	438710100	0	

Bolts (All time)

Id	Parallelism	Emitted	Transferred	Process latency (ms)	Acked	Failed	Last error
-1	4	438716440	438716440	0.009	2223890060	0	
2	160	877451720	877451720	0.320	438725980	0	
3	160	1264258160	1264258160	5.438	438724980	0	
4	18	55946080	55946080	0.215	55946040	0	
5	18	55947280	55947280	0.121	55947280	0	
6	18	55945660	55945660	0.229	55945660	0	
7	18	55946480	55946480	0.145	55946580	0	
8	18	81512620	81512620	0.209	81512620	0	
9	30	438710060	438710060	4205.639	438710140	0	
10	80	163024580	163024580	0.194	81512200	0	

Storm on EC2

https://github.com/nathanmarz/storm-deploy

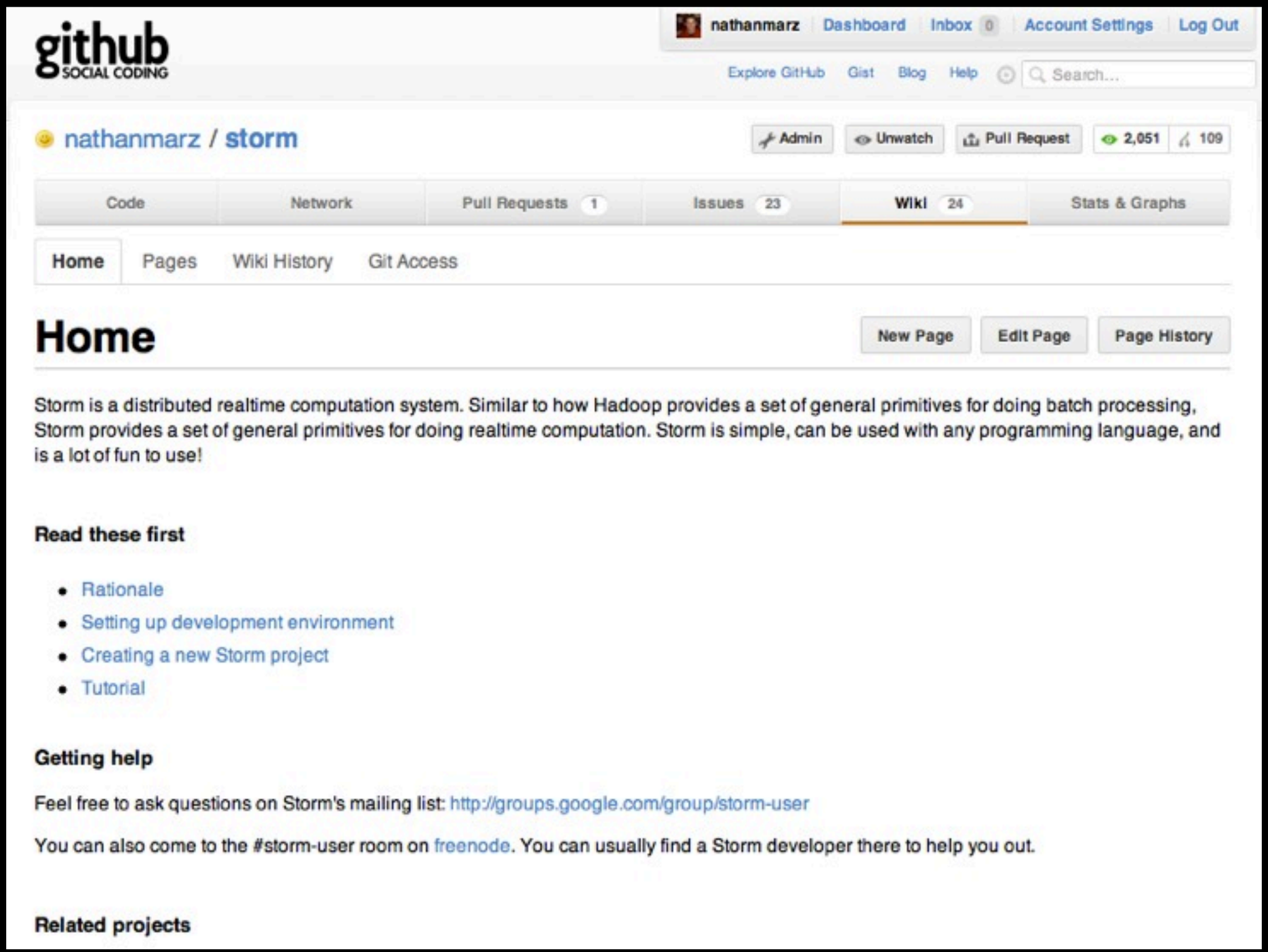
One-click deploy tool

Starter code

https://github.com/nathanmarz/storm-starter

Example topologies

Documentation



Ecosystem

- Scala, JRuby, and Clojure DSL's
- Kestrel, Redis, AMQP, JMS, and other spout adapters
- Multilang adapters
- Cassandra, MongoDB integration

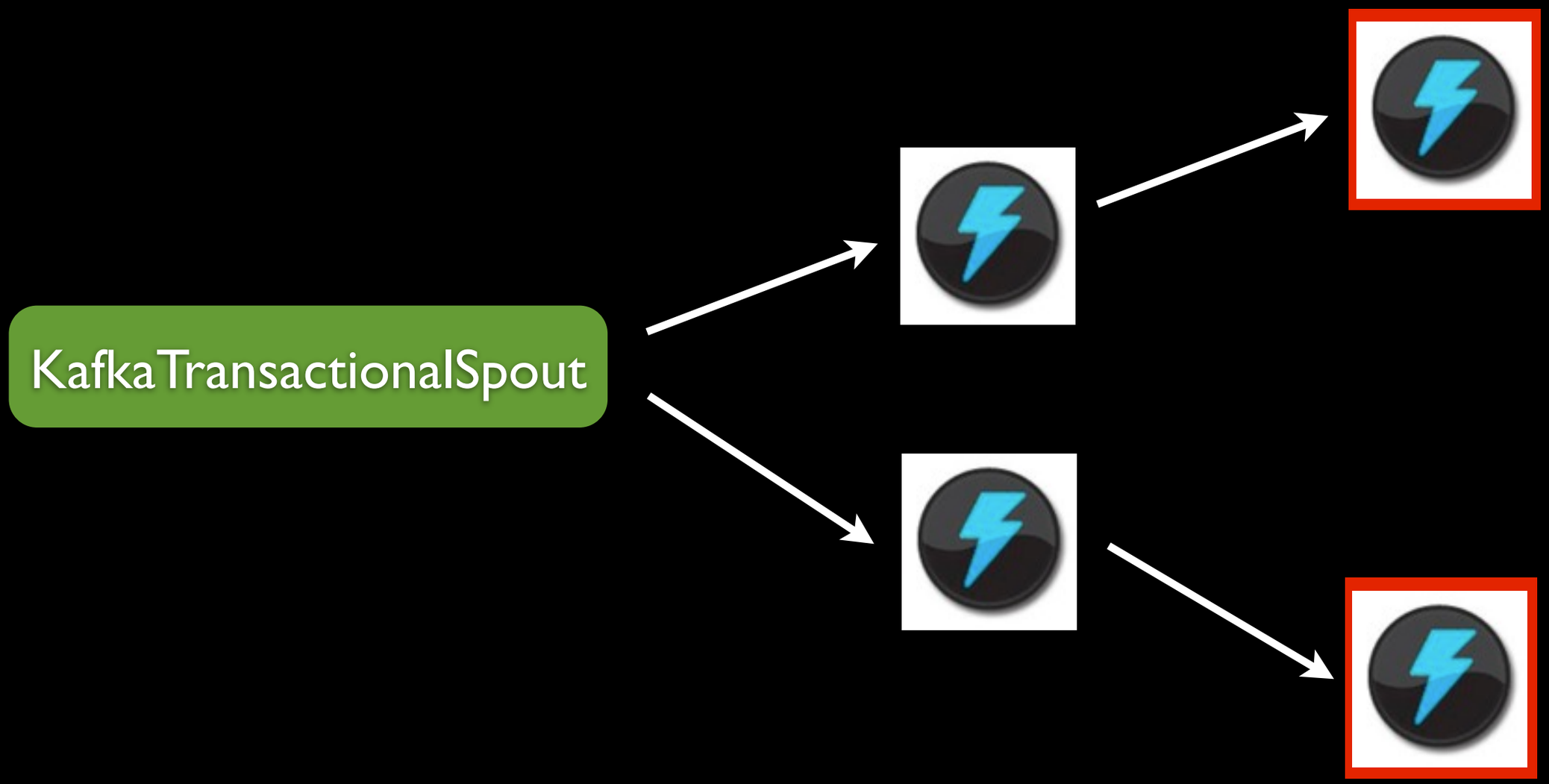
Questions?

http://github.com/nathanmarz/storm

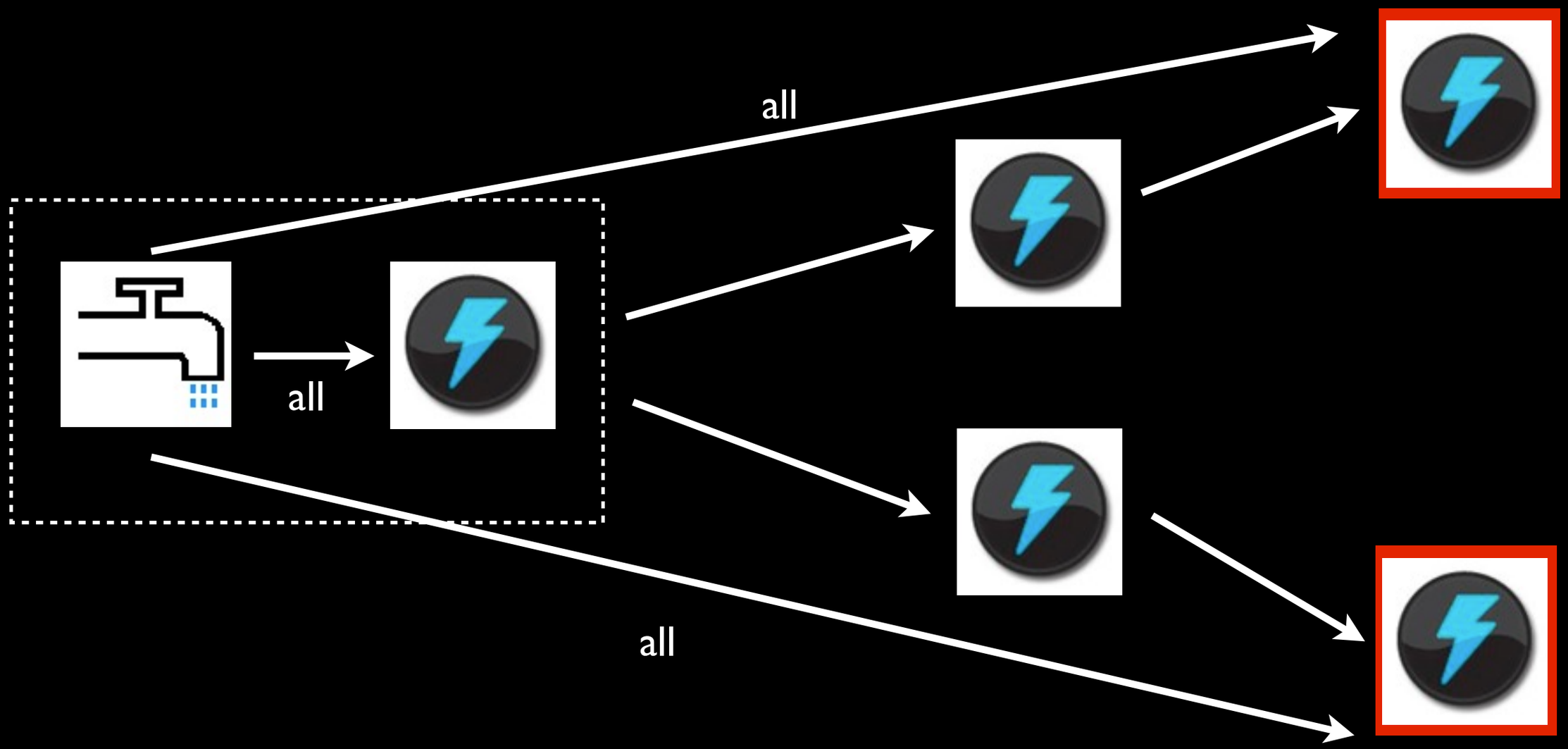
Future work

- State spout
- Storm on Mesos
- “Swapping”
- Auto-scaling
- Higher level abstractions

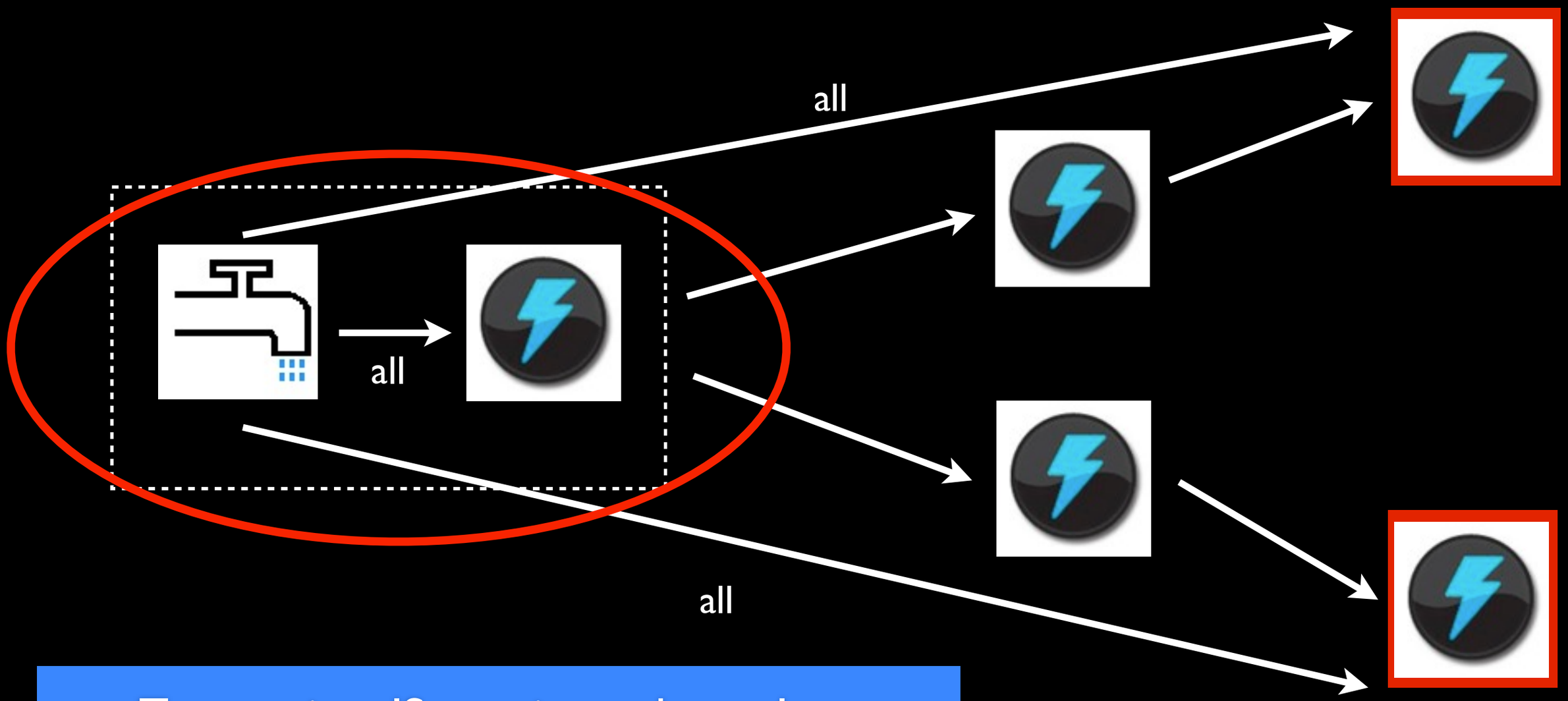
Implementation



Implementation

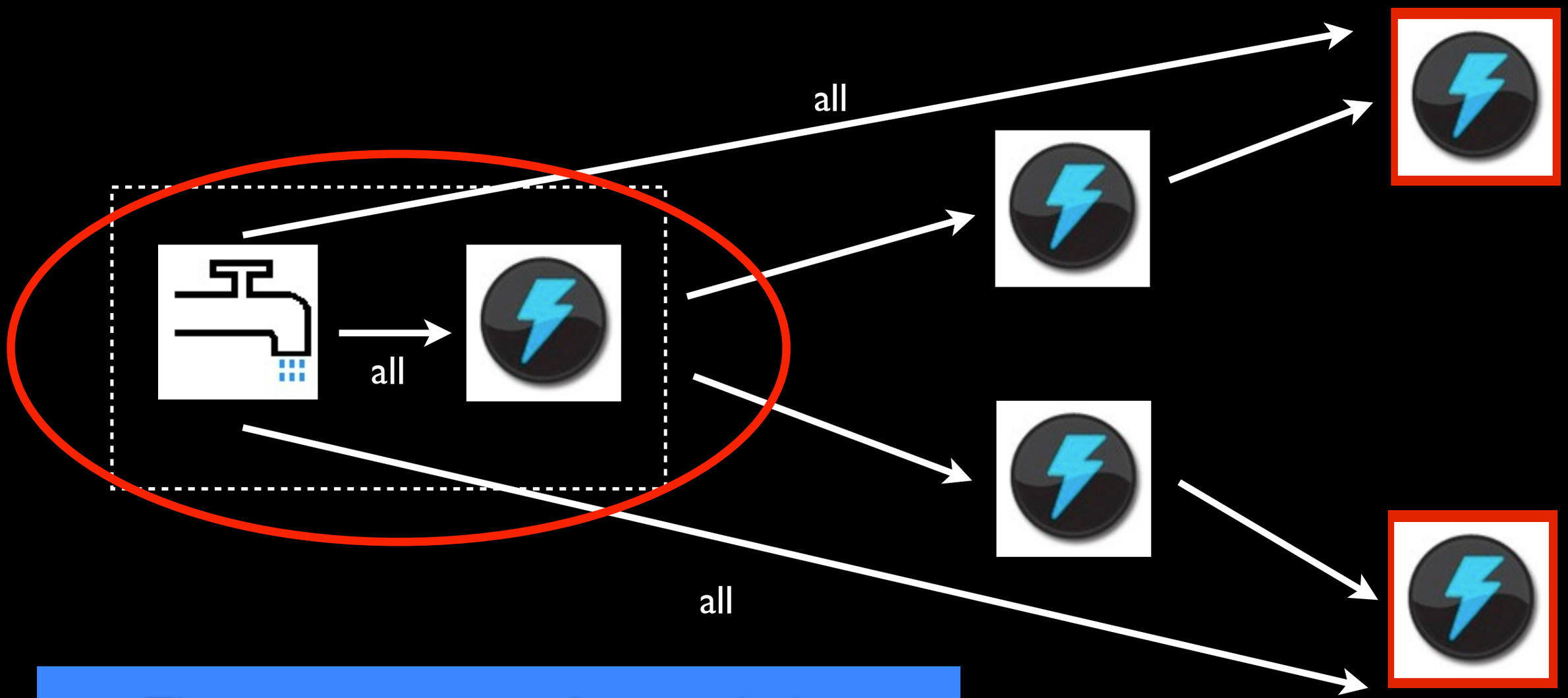


Implementation



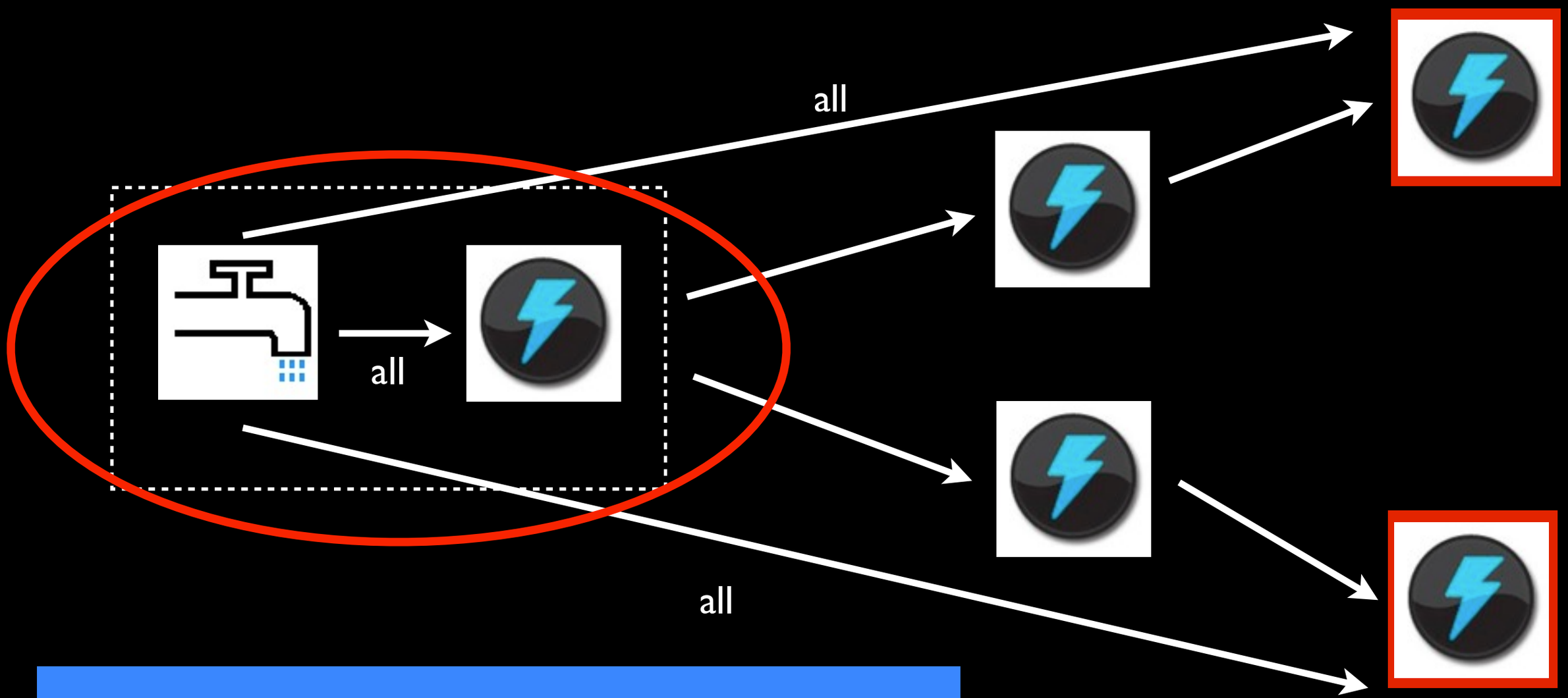
TransactionalSpout is a subtopology consisting of a spout and a bolt

Implementation



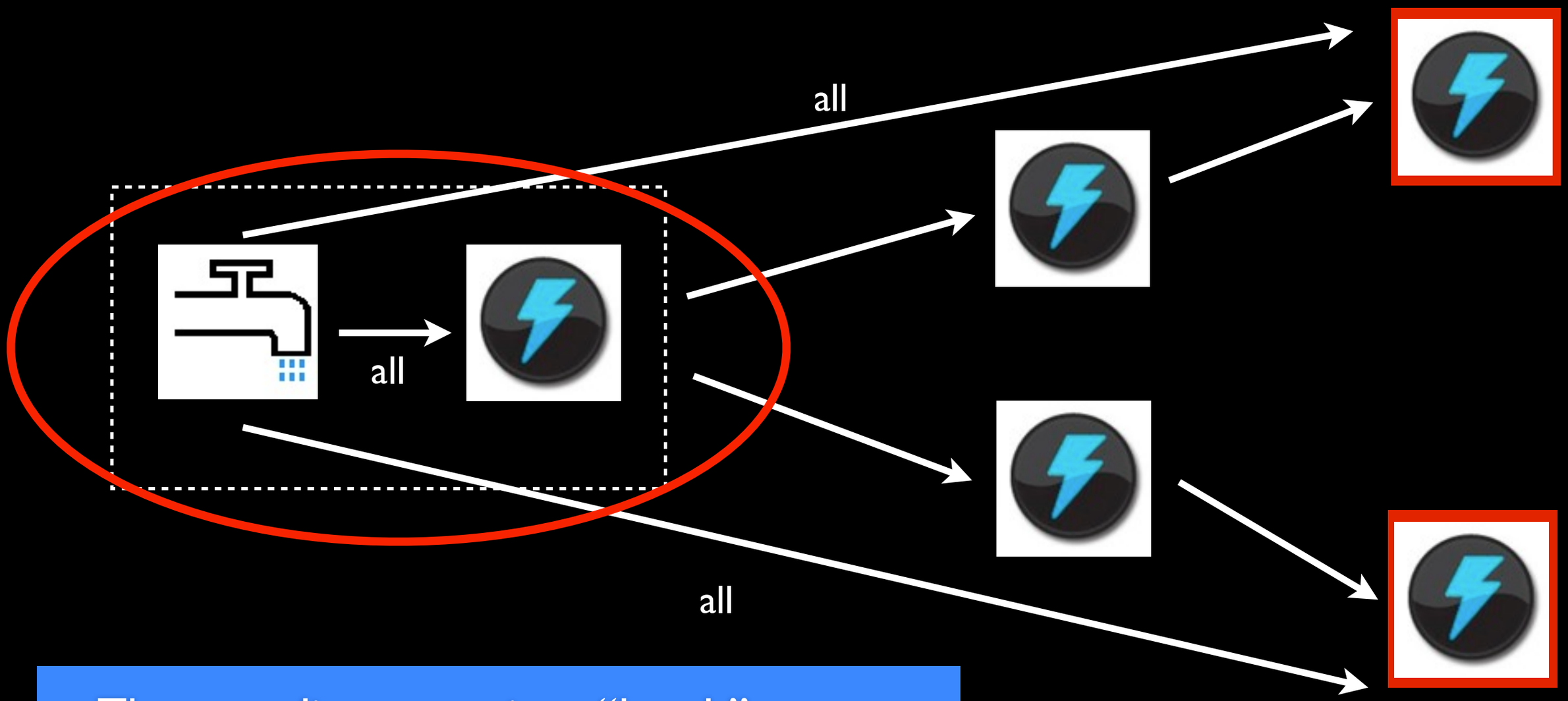
The spout consists of one task that coordinates the transactions

Implementation



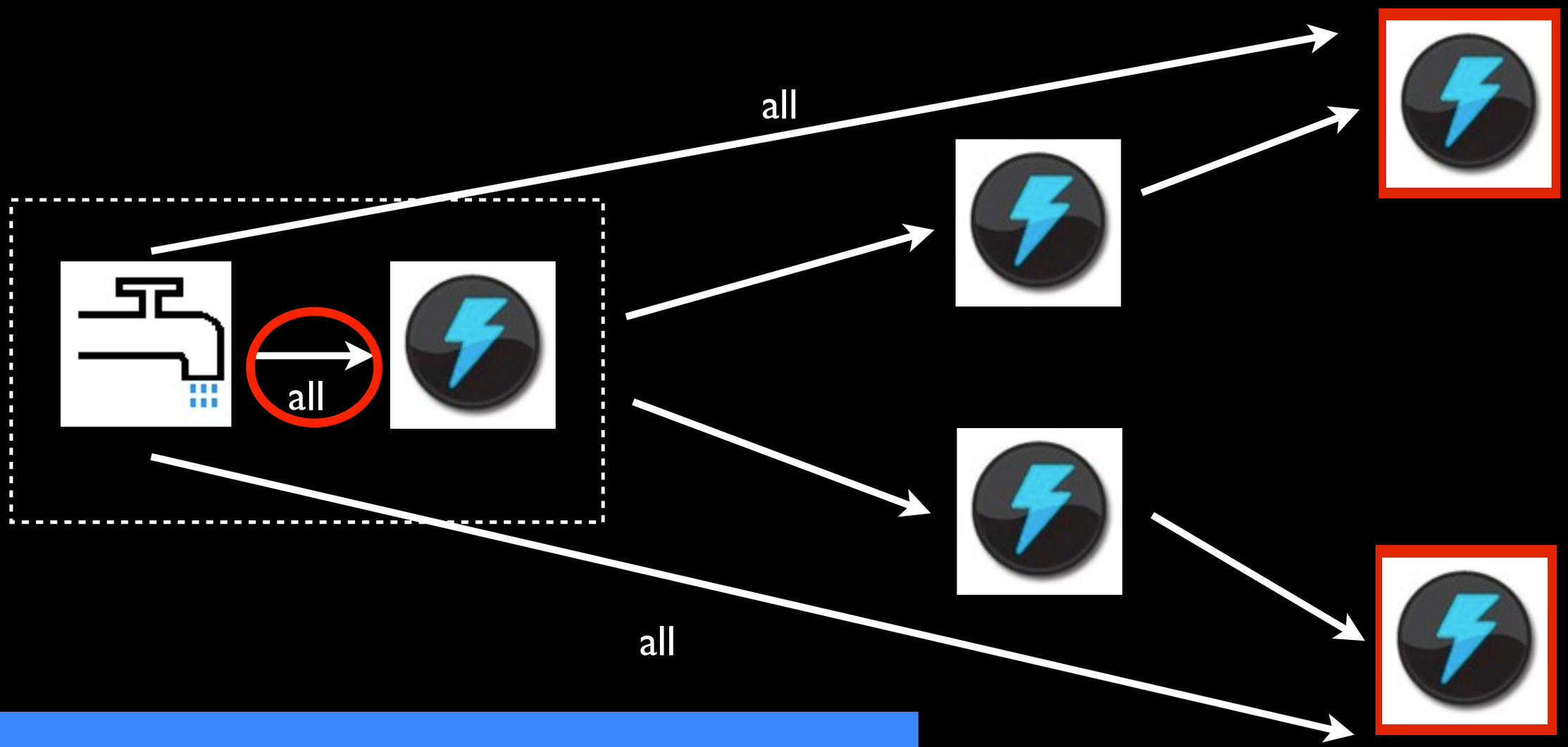
The bolt emits the batches of tuples

Implementation



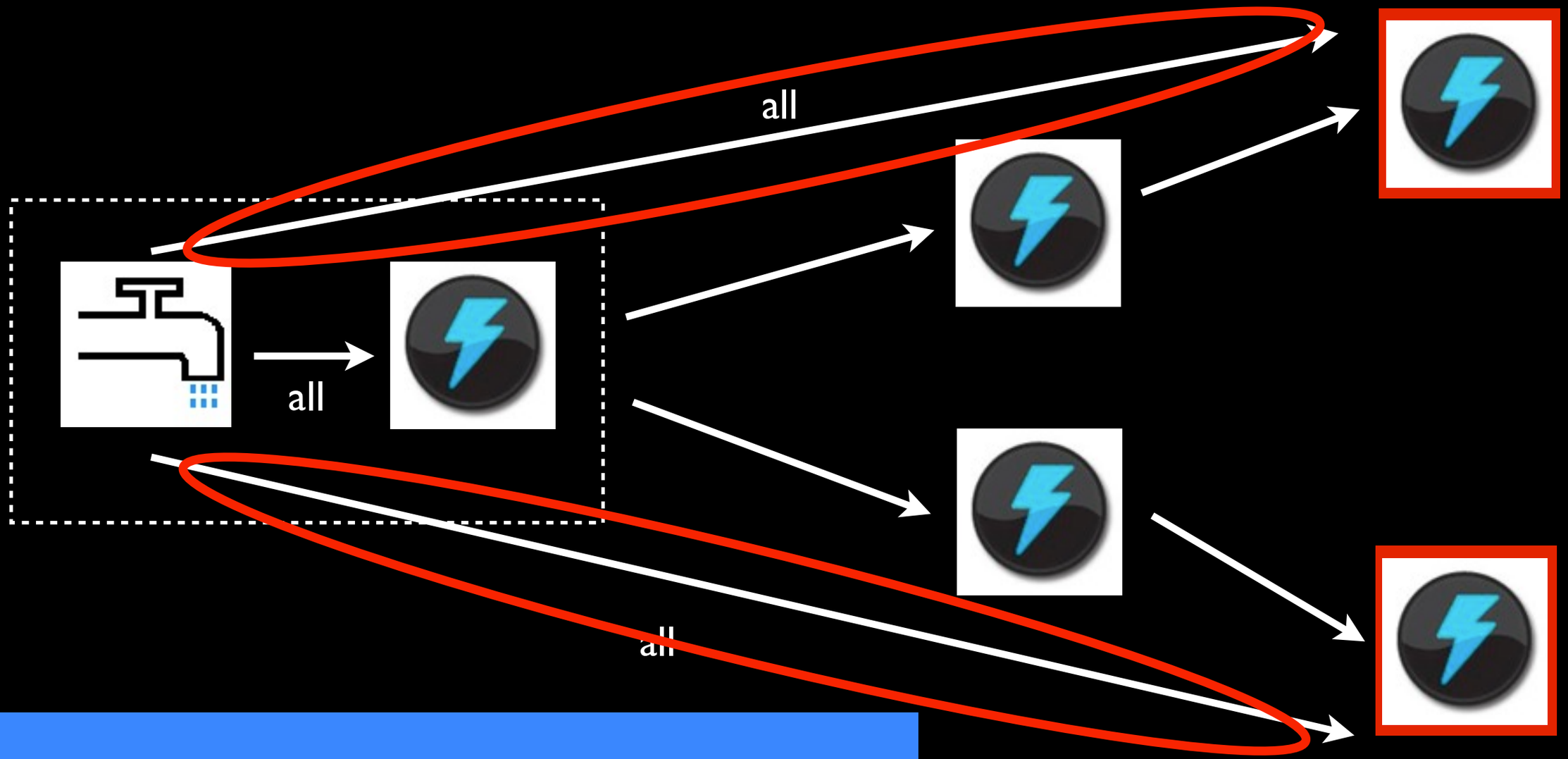
The coordinator emits a “batch” stream and a “commit stream”

Implementation



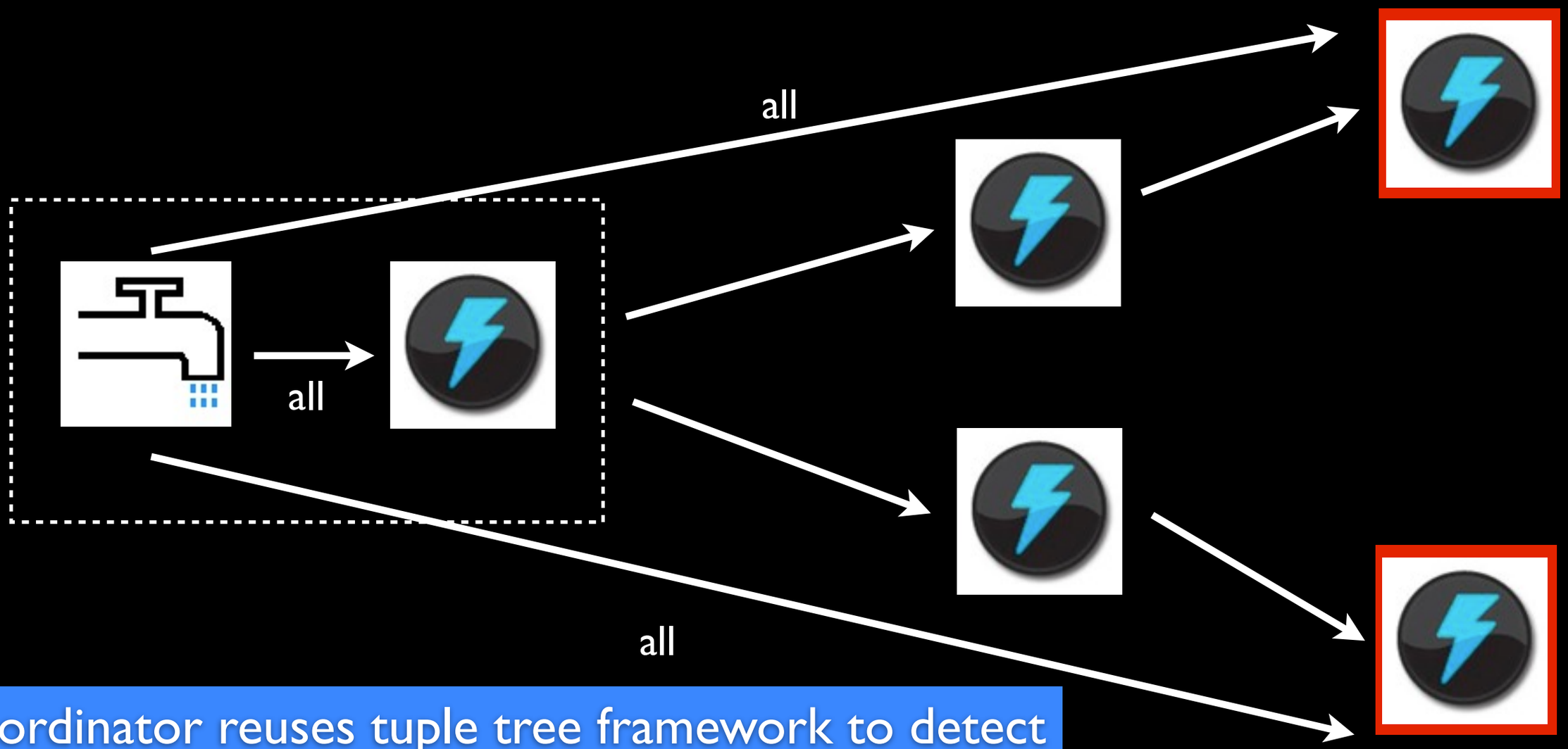
Batch stream

Implementation



Commit stream

Implementation



Coordinator reuses tuple tree framework to detect success or failure of batches or commits and replays appropriately