

LEVERAGE ON OPEN SOURCE TO BUILD UP A DATA PRODUCT

Darren Chen @ DataCon.TW

2017.09.30



About Me

- Darren Chen
- Sr. Software Engineer @ Trend Micro
- Enthusiast in
 - Big data technologies
 - Docker
- Speaker in
 - DockerCon 2016
 - iThome Big Data Innovation Summit 2016



5,239,978

5,239,978

Lines of code

One engineer

100

Lines of code/ pre day

143 years



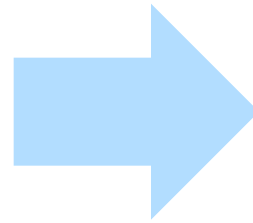
**KEEP
CALM
AND
DON'T REINVENT
THE WHEEL**

0 → 1

1. Frame the Problem
2. Get the Data
3. Explore the Data
4. Prepare the Data
5. Short-List Promising Models
6. Fine-Tune the System
7. Online

ML Project Checklist

1. Frame the Problem
2. Get the Data
3. Explore the Data
4. Prepare the Data
5. Short-List Promising Models
6. Fine-Tune the System
7. Online



Data Product Checklist

1. Frame the Problem
2. Get the Data
3. Explore the Data
4. Prepare Requirements
5. Design Architecture
6. Improve the System
7. Online

ML Project Checklist

1. Frame the Problem
2. Get the Data
3. Explore the Data
4. Prepare the Data
5. Short-List Promising Models
6. Fine-Tune the System
7. Online

Data Product Checklist

1. Frame the Problem
2. Get the Data
3. Explore the Data
4. Prepare Requirements
5. Design Architecture
6. Improve the System
7. Online

TIME

1. Frame the Problem

- You need to build up an on-premises product which can detect threats among many kinds of logs

1. Frame the Problem

- You need to build up an on-premises product which can detect threats among many kinds of logs

1. Frame the Problem

- You need to build up an on-premises product which can detect threats among many kinds of logs



2. Get the Data

- Q: How to import data to our system ? [\[Data Import\]](#)
- A: 1. For streaming data, such as TCP/UDP, syslog ...
2. For batch data, such as files

3. Explore the Data

- Q1: How many kinds of log types ? [Data Variety]

A1: More than 6 types

- Q2: What is the size? [Data Volume]

A2: >1,000,000,000 records per day/per type

- Q3: What else ? [Data Characteristics]

A3: Time series

4. Prepare Requirements

Data Import

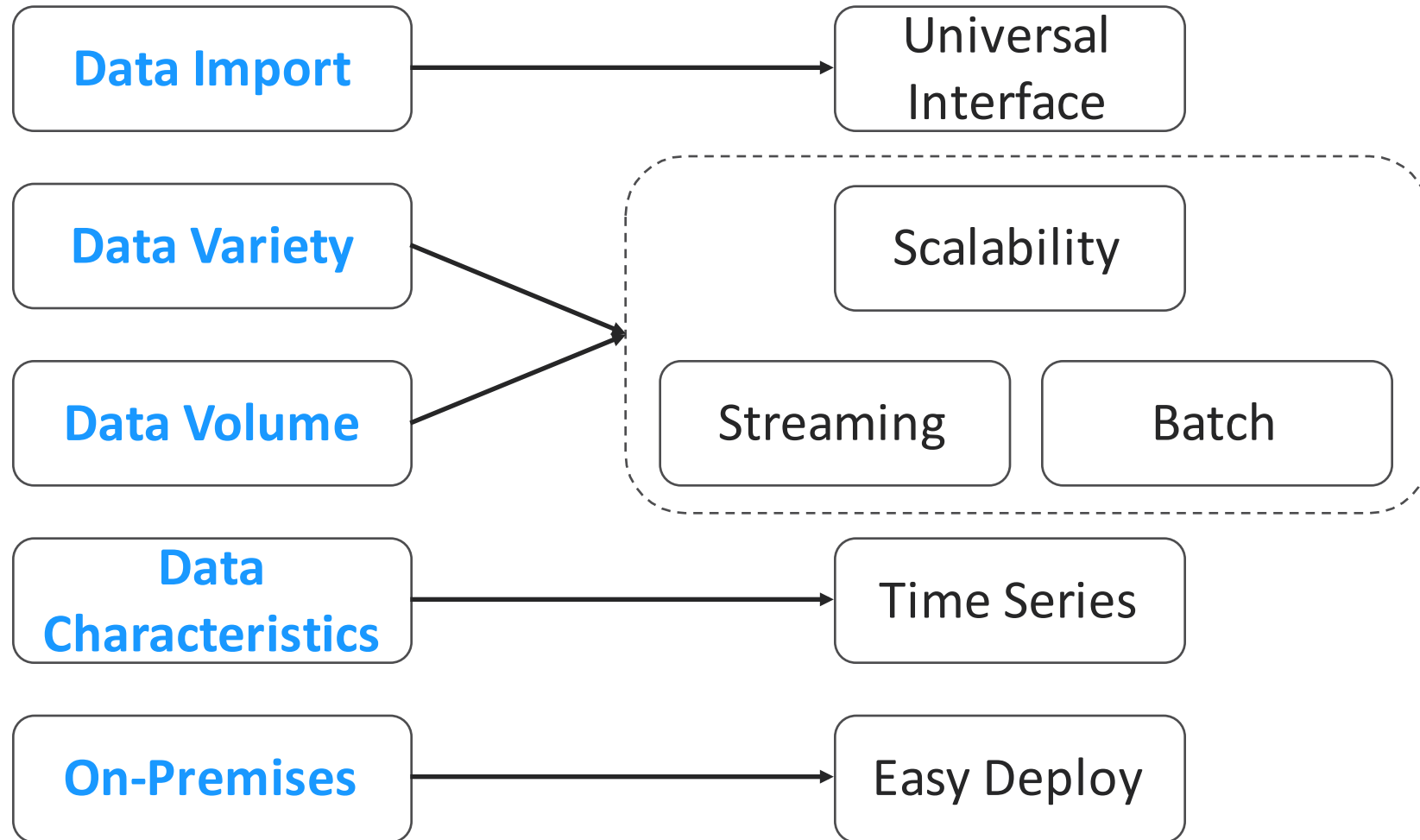
Data Variety

Data Volume

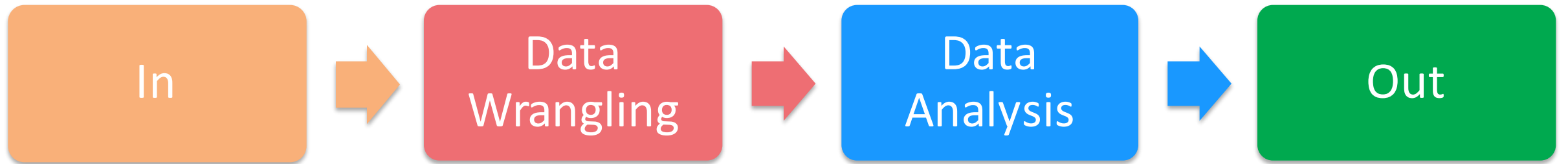
**Data
Characteristics**

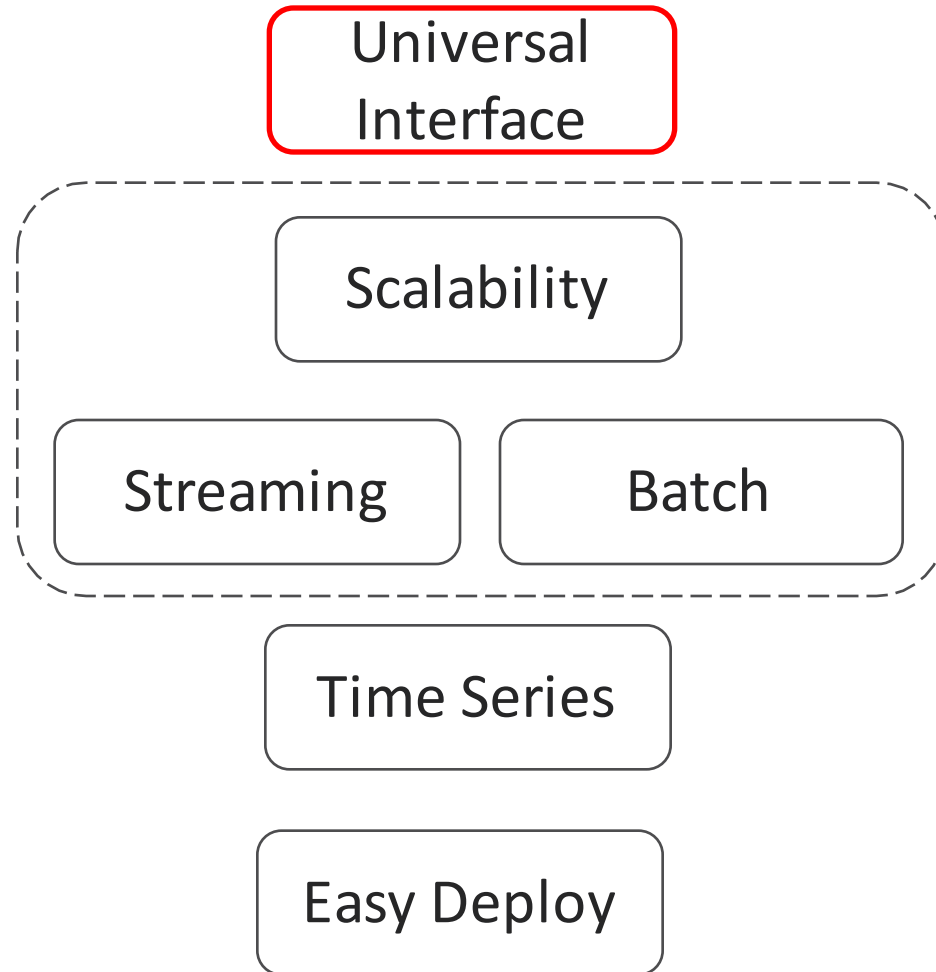
On-Premises

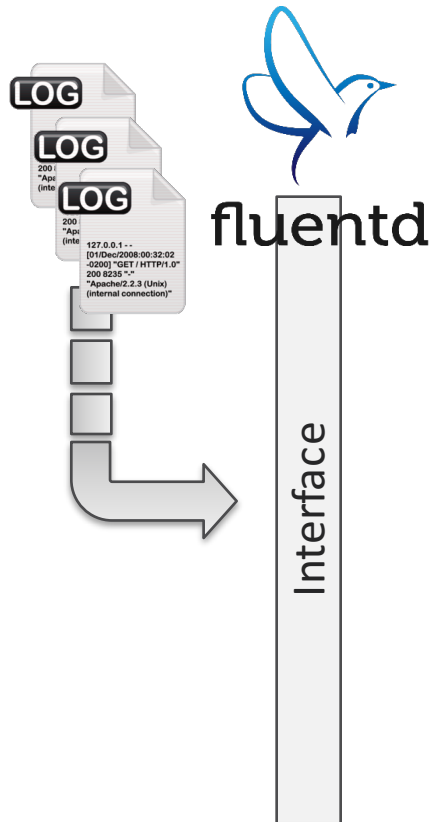
4. Prepare Requirements

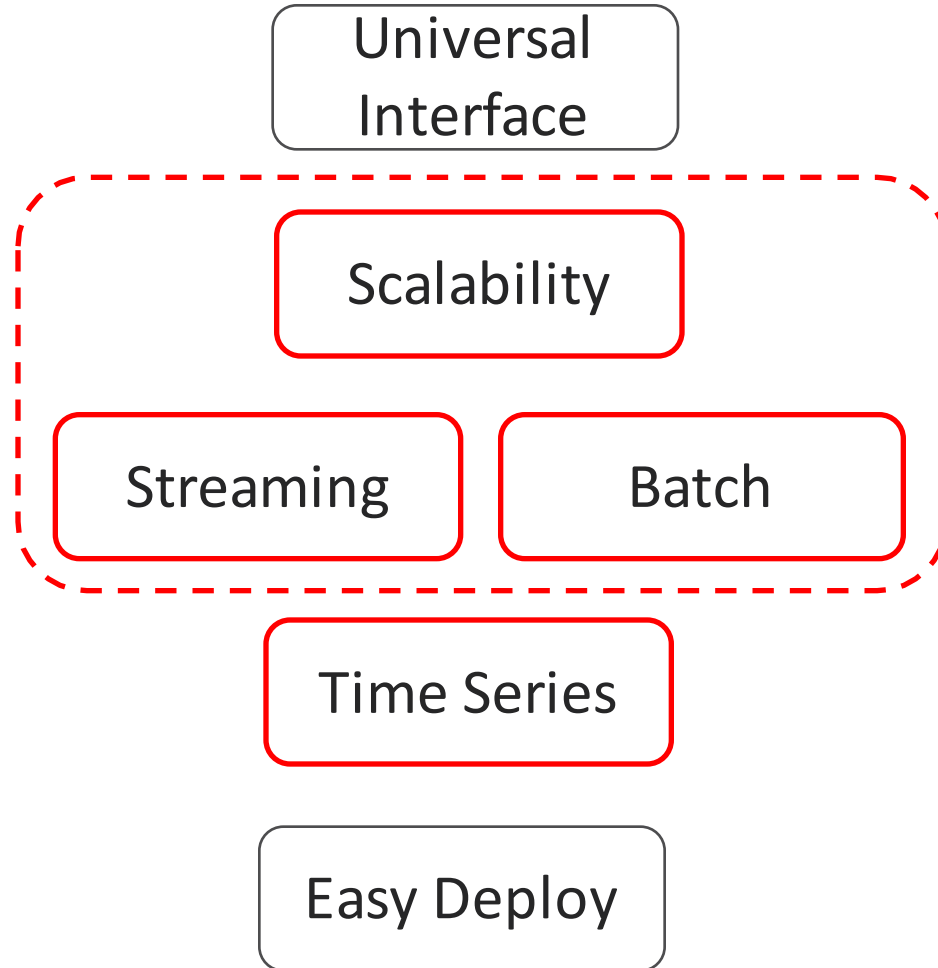


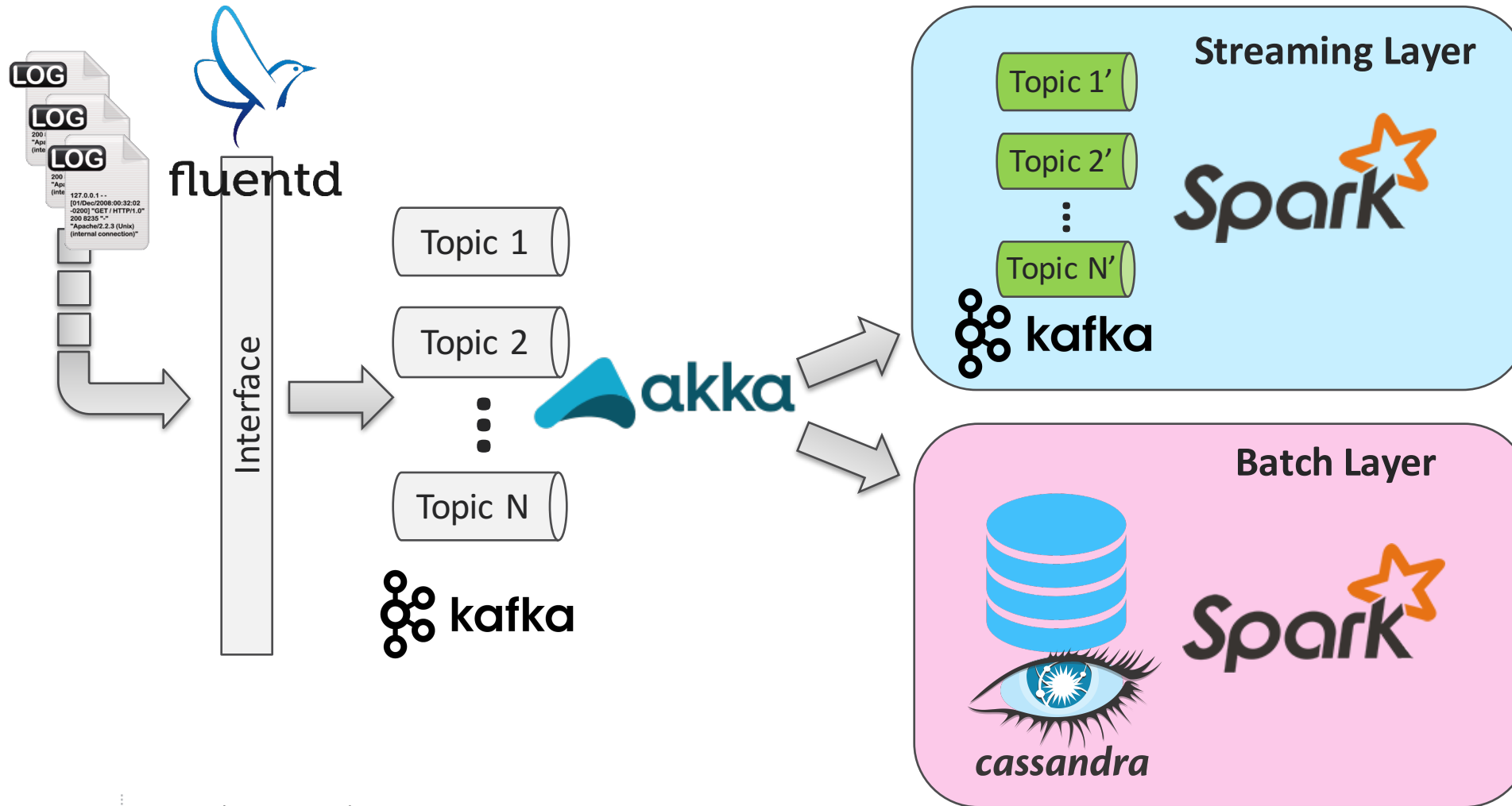
5. Design Architecture

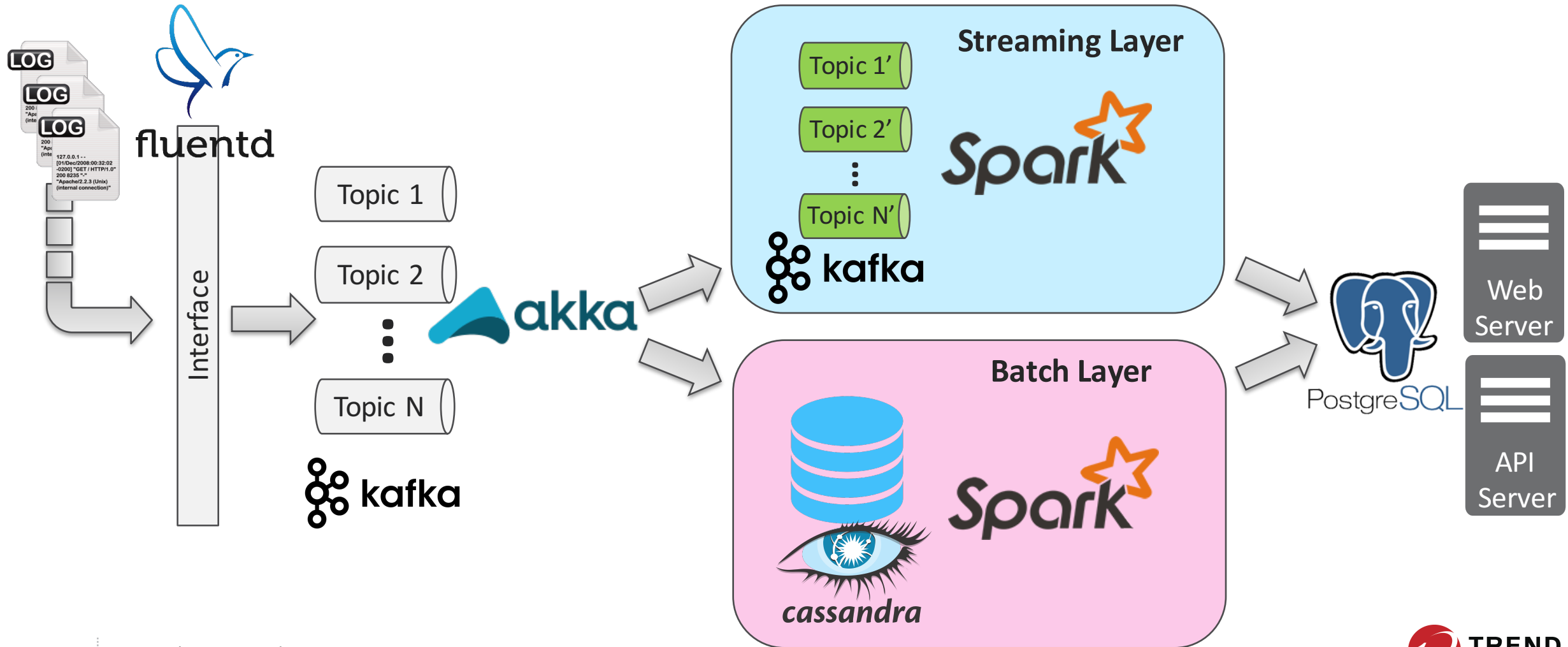


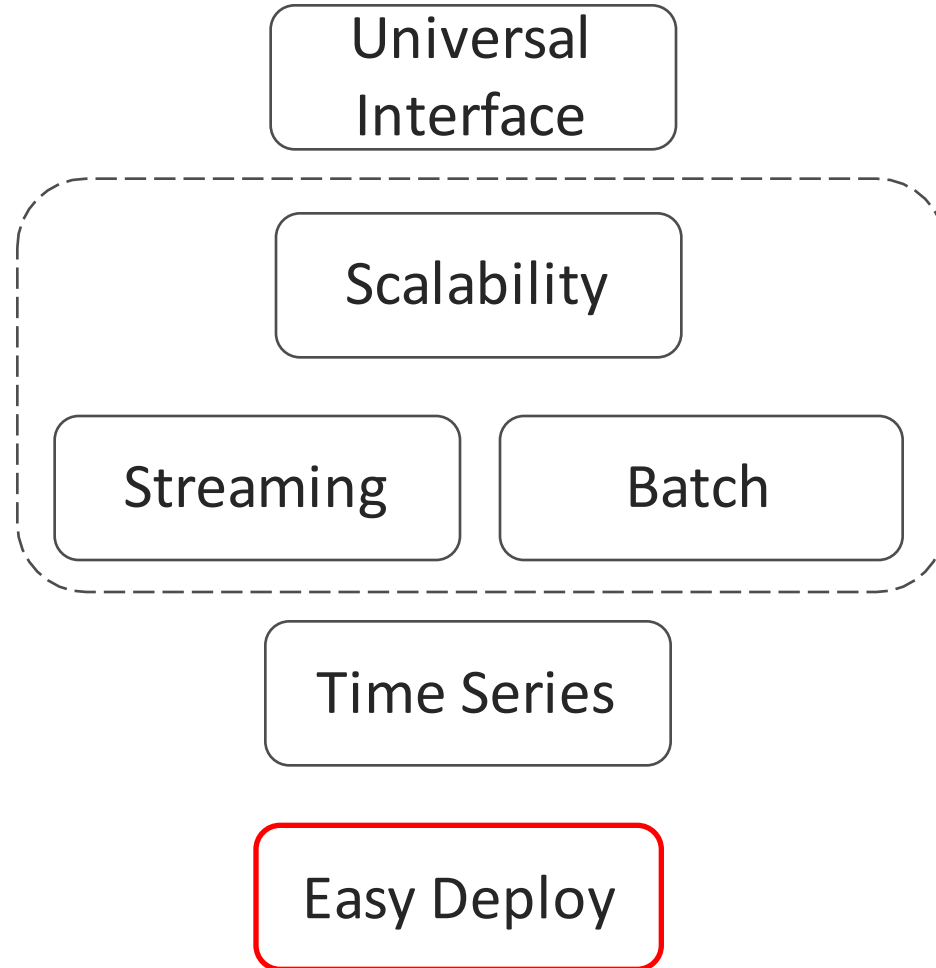




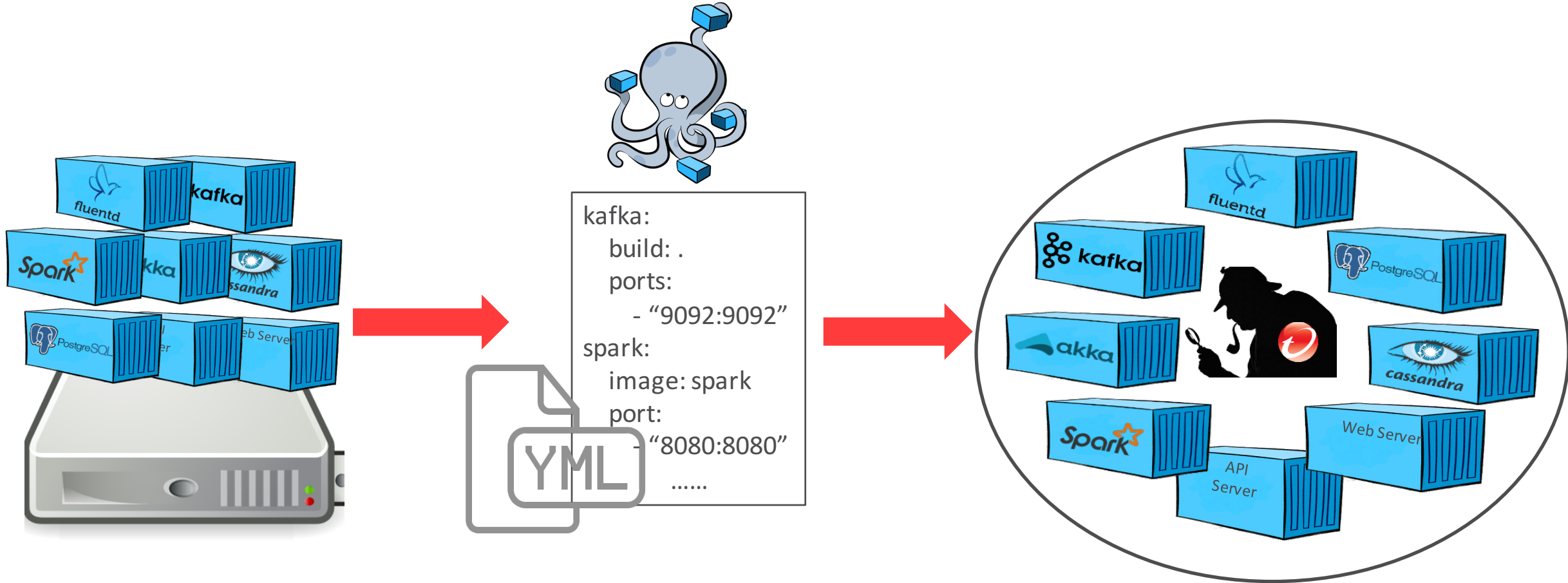




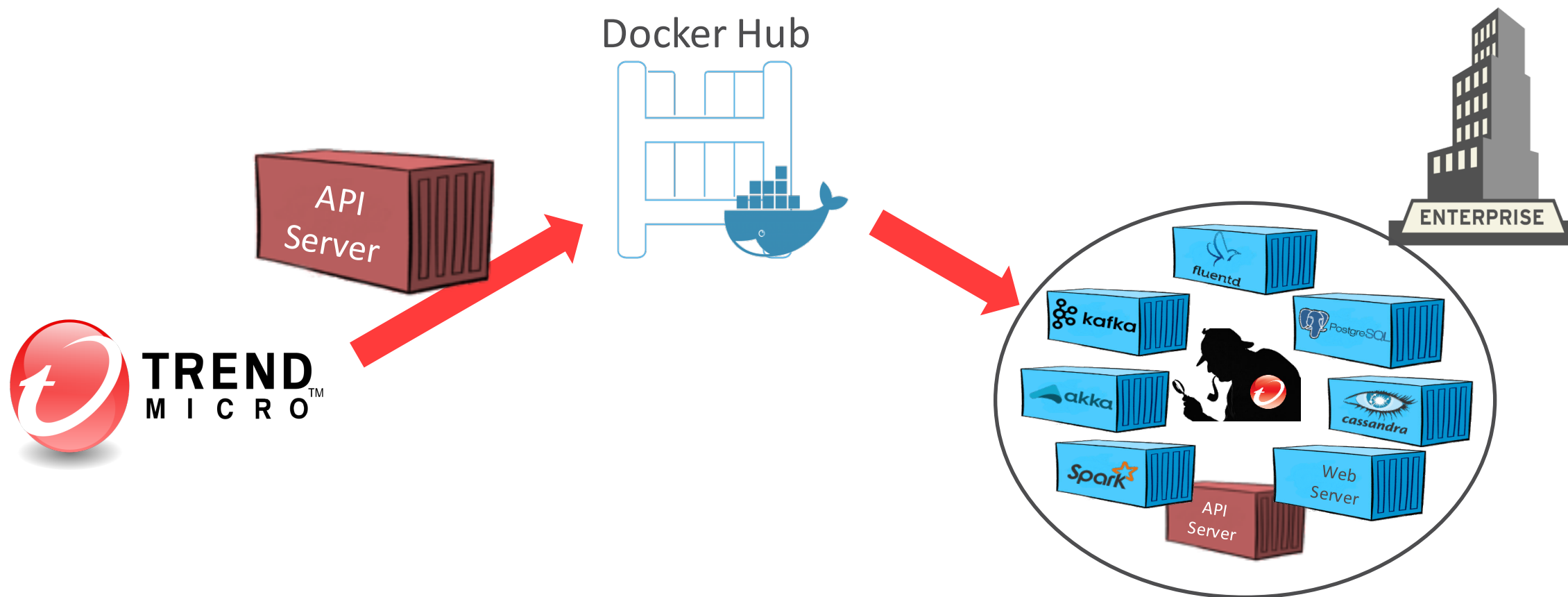




Dockerize Product

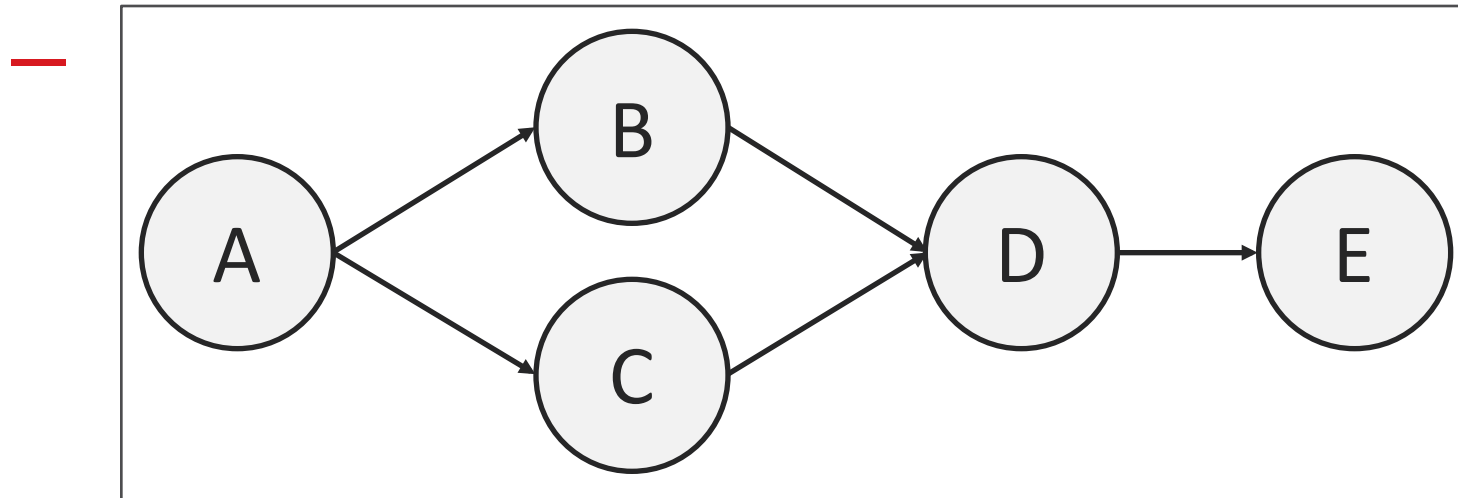


Deploy Dockerized Product



6. Improve the System

- Profiling
- Scheduling

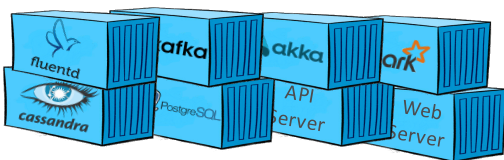


Profiling - Monitor Stack



cAdvisor

CPU, Memory, Network

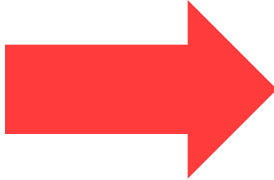
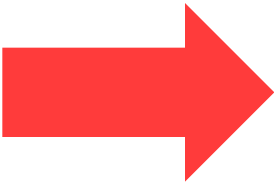




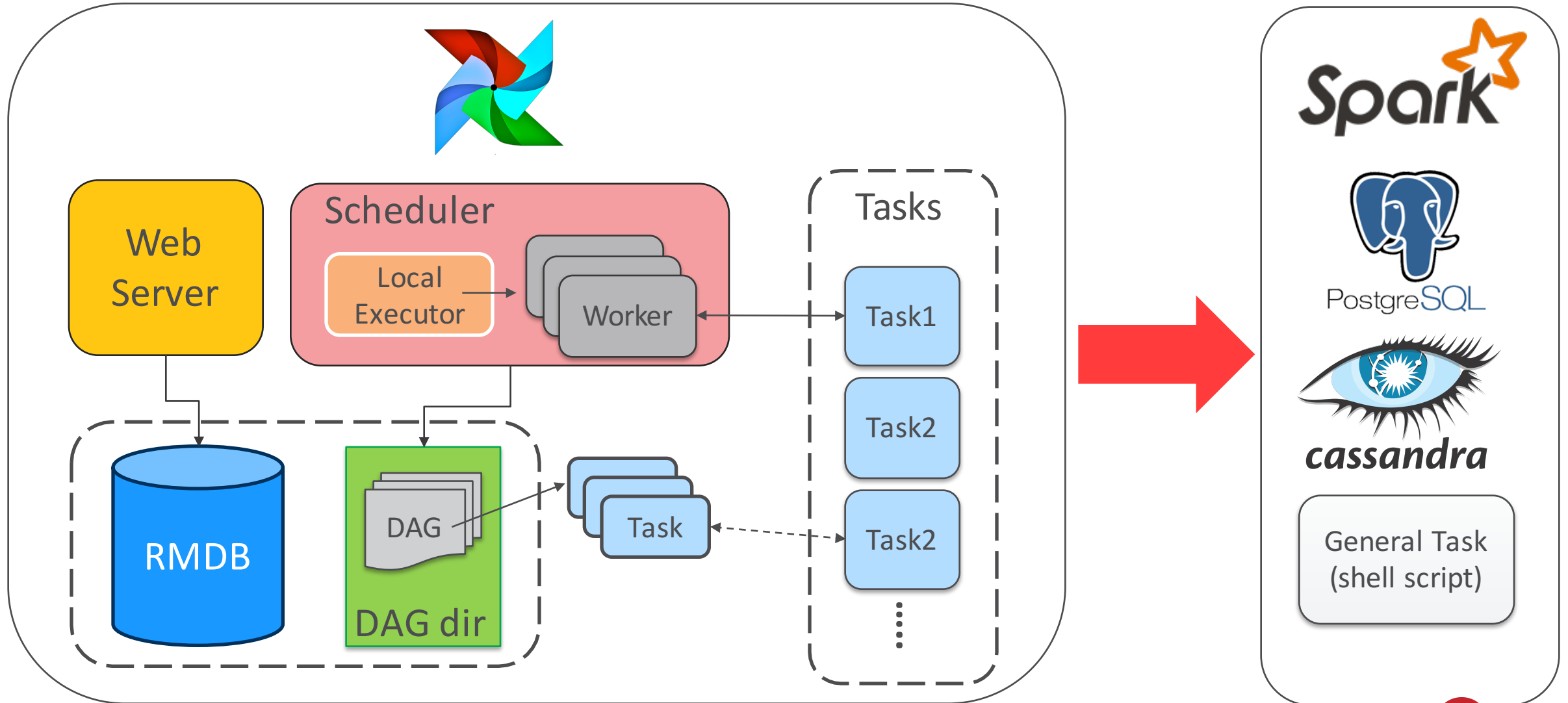
Spark

kafka

APP
Metrics



Scheduling - Airflow



Data Product Checklist

1. Frame the Problem
2. Get the Data
3. Explore the Data
4. Prepare Requirements
5. Design Architecture
6. Improve the System



Online !

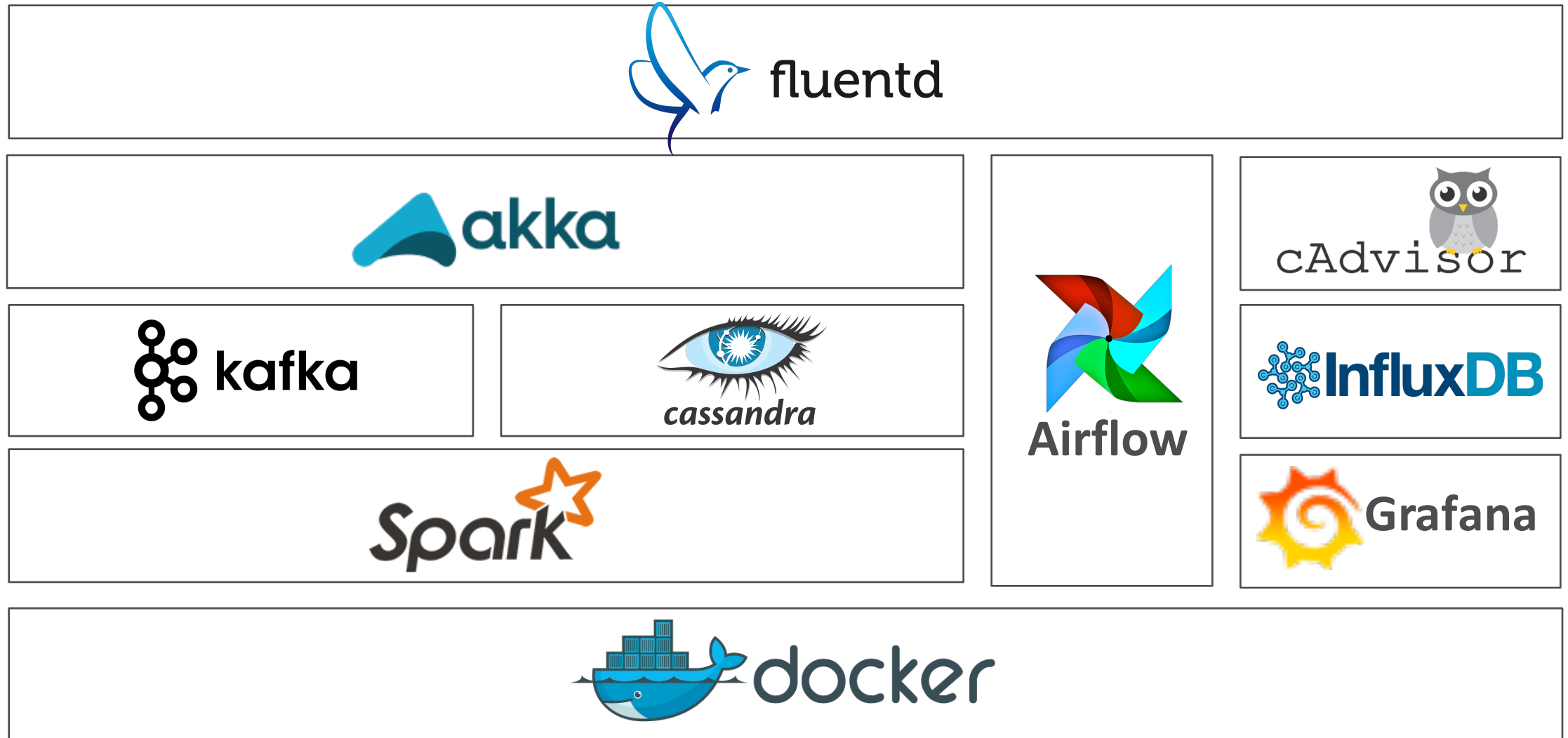
1 → 1.1

The Startup Curve

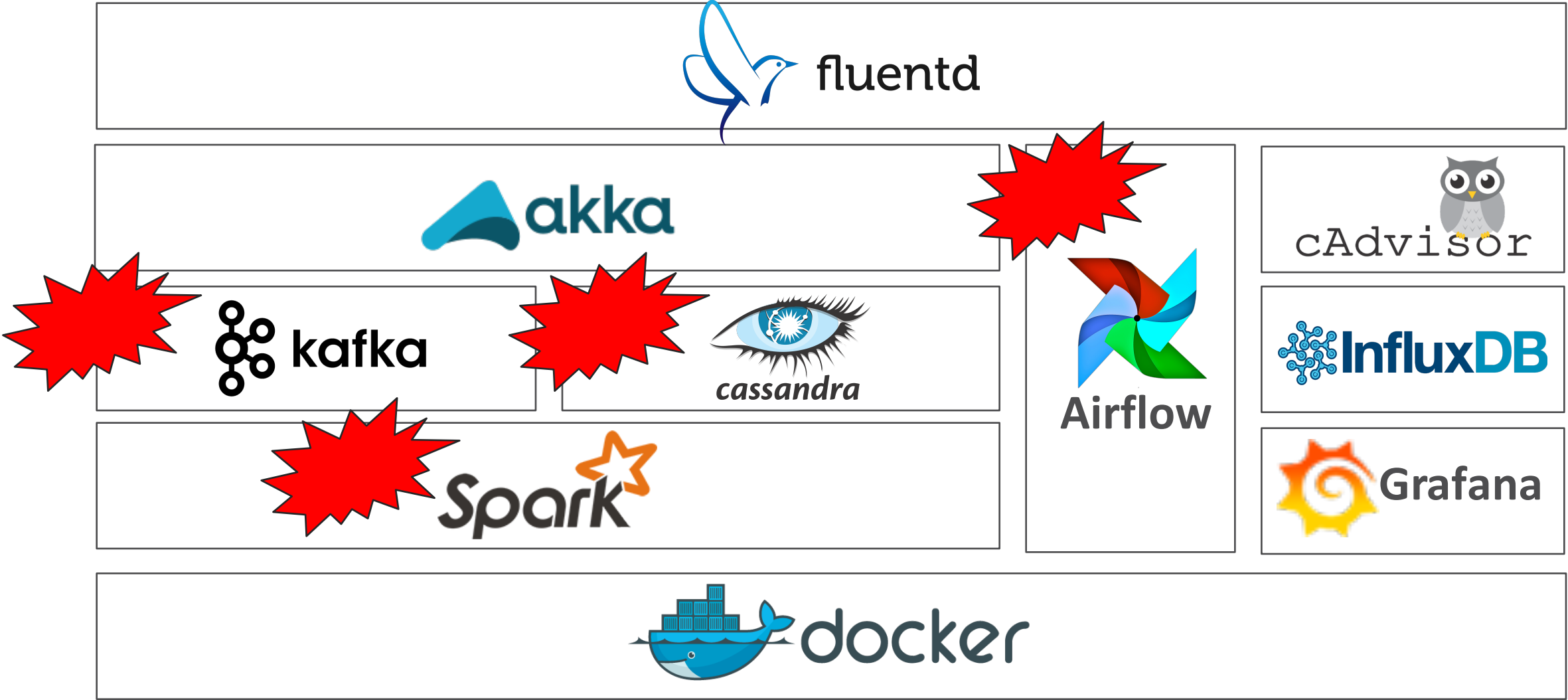


Source: Paul Graham; ayc.com

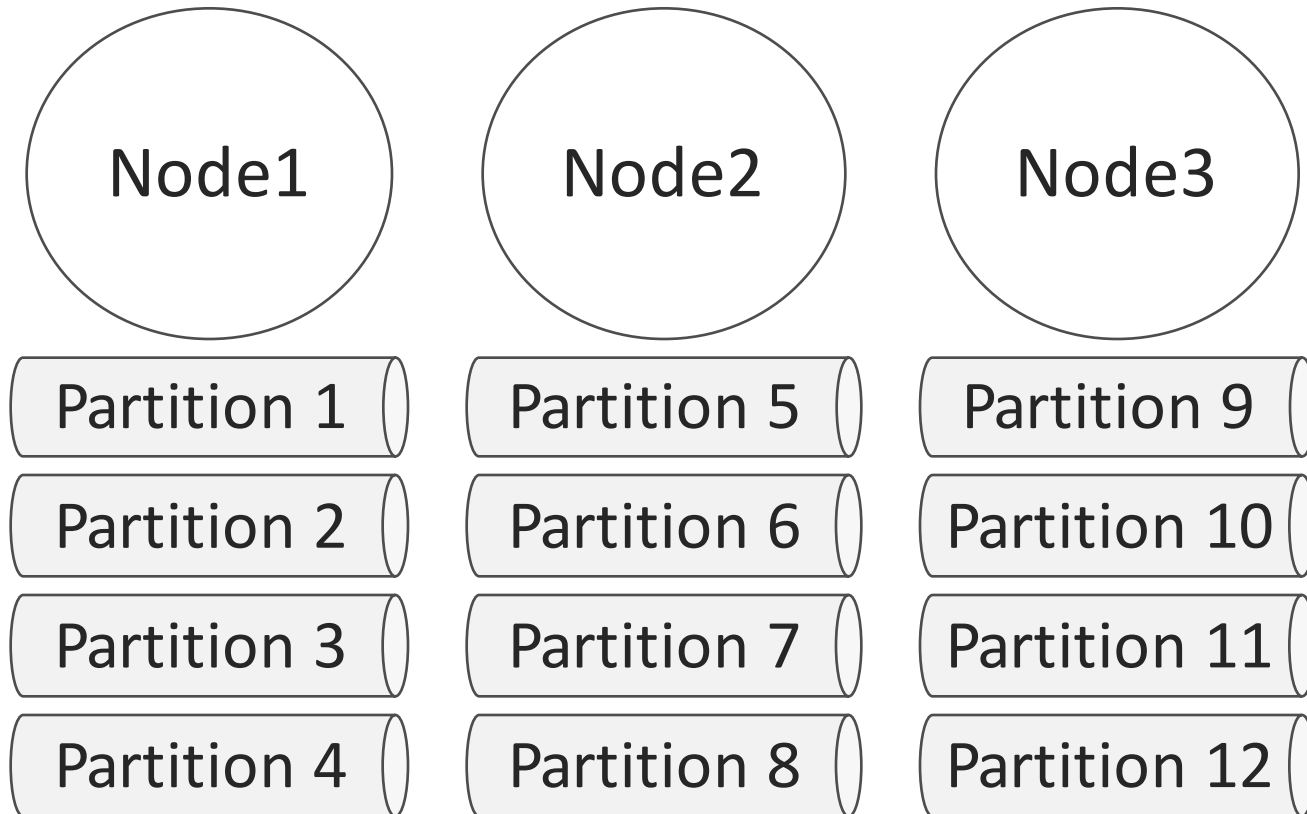
Data Product Stack



Data Product Stack



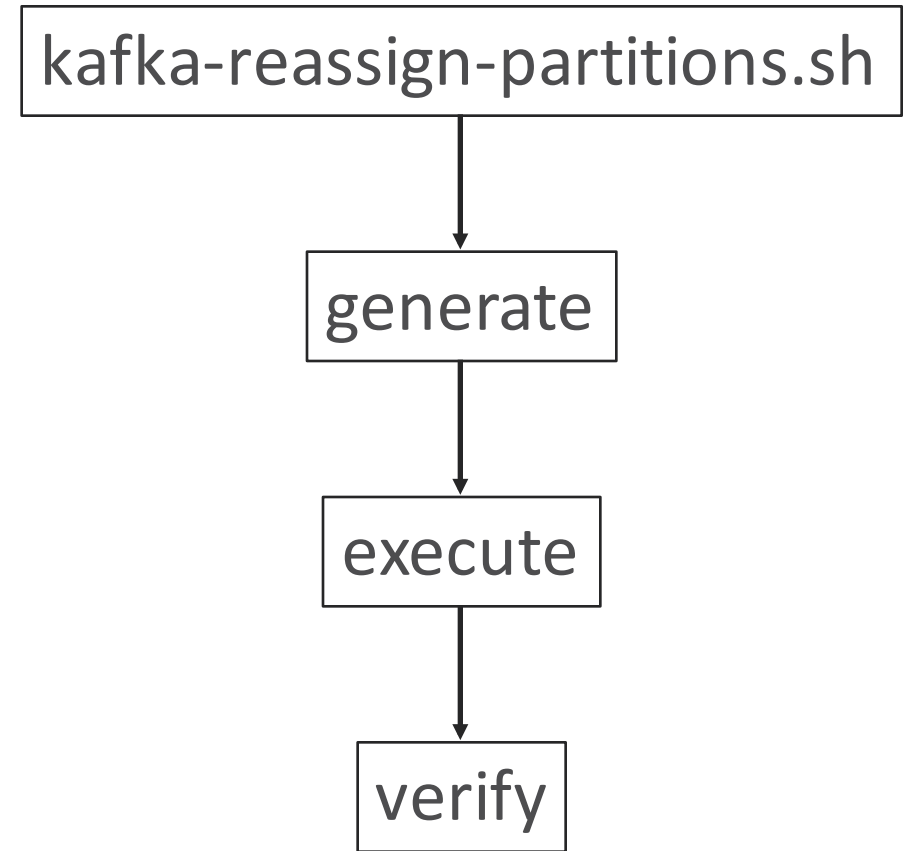
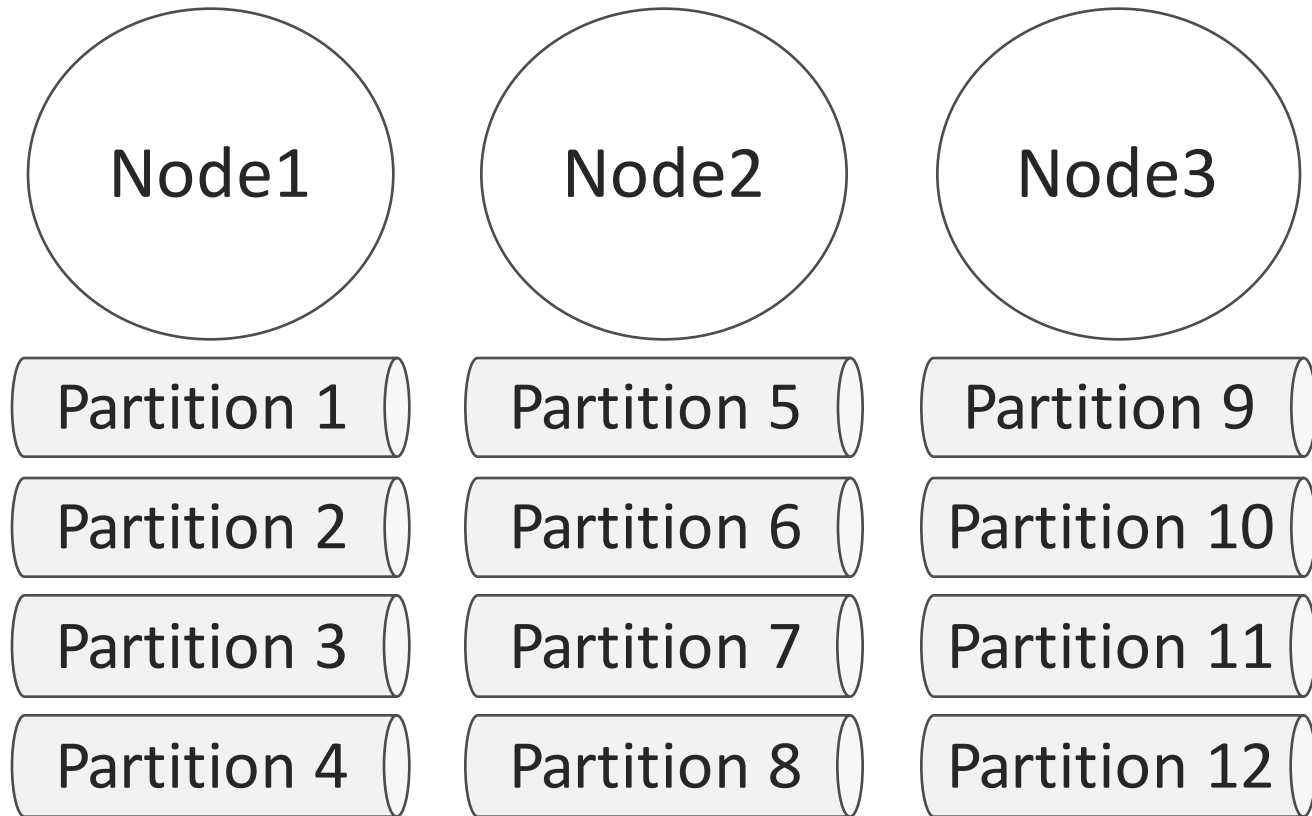
4-1 Kafka Scale Out



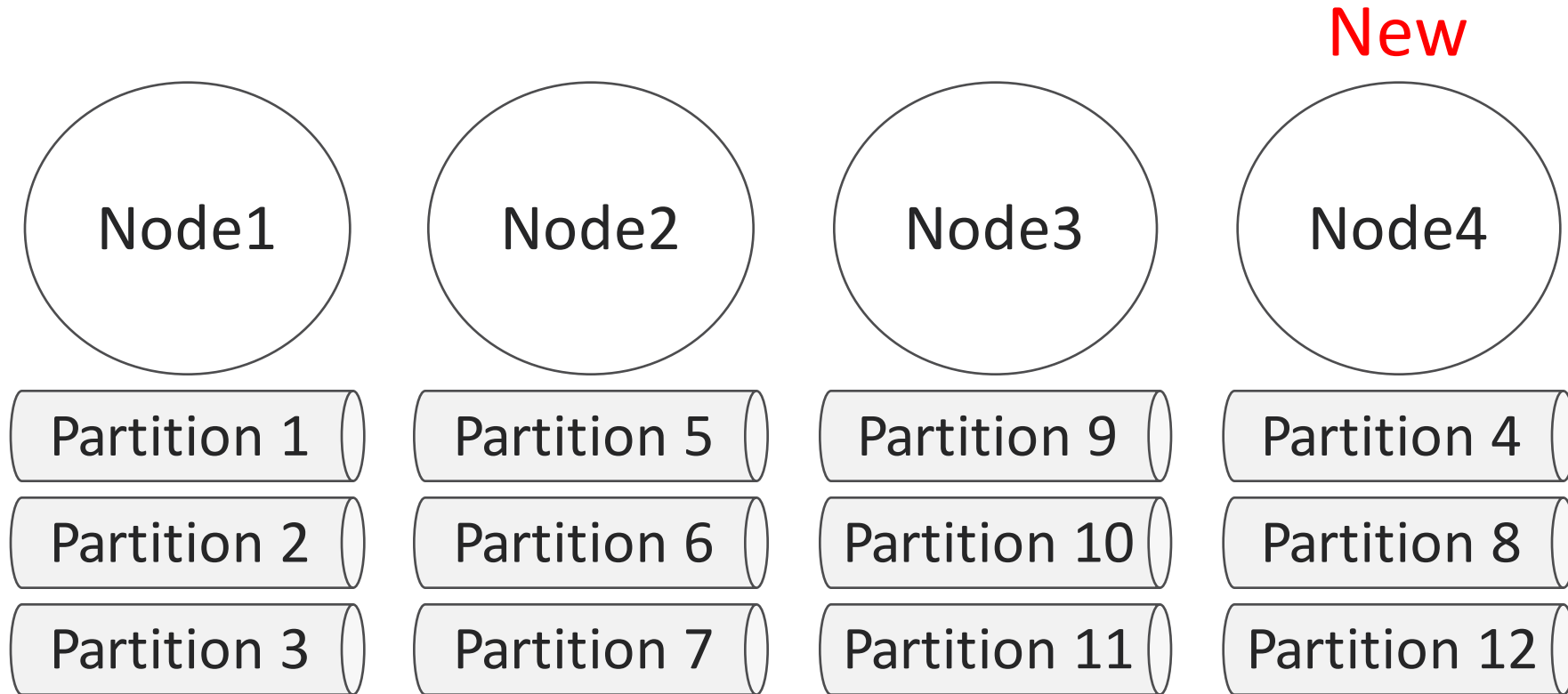
Example

```
bin/kafka-topics.sh --create  
    --zookeeper localhost:2181  
    --replication-factor 1  
    --partitions 12  
    --topic test
```

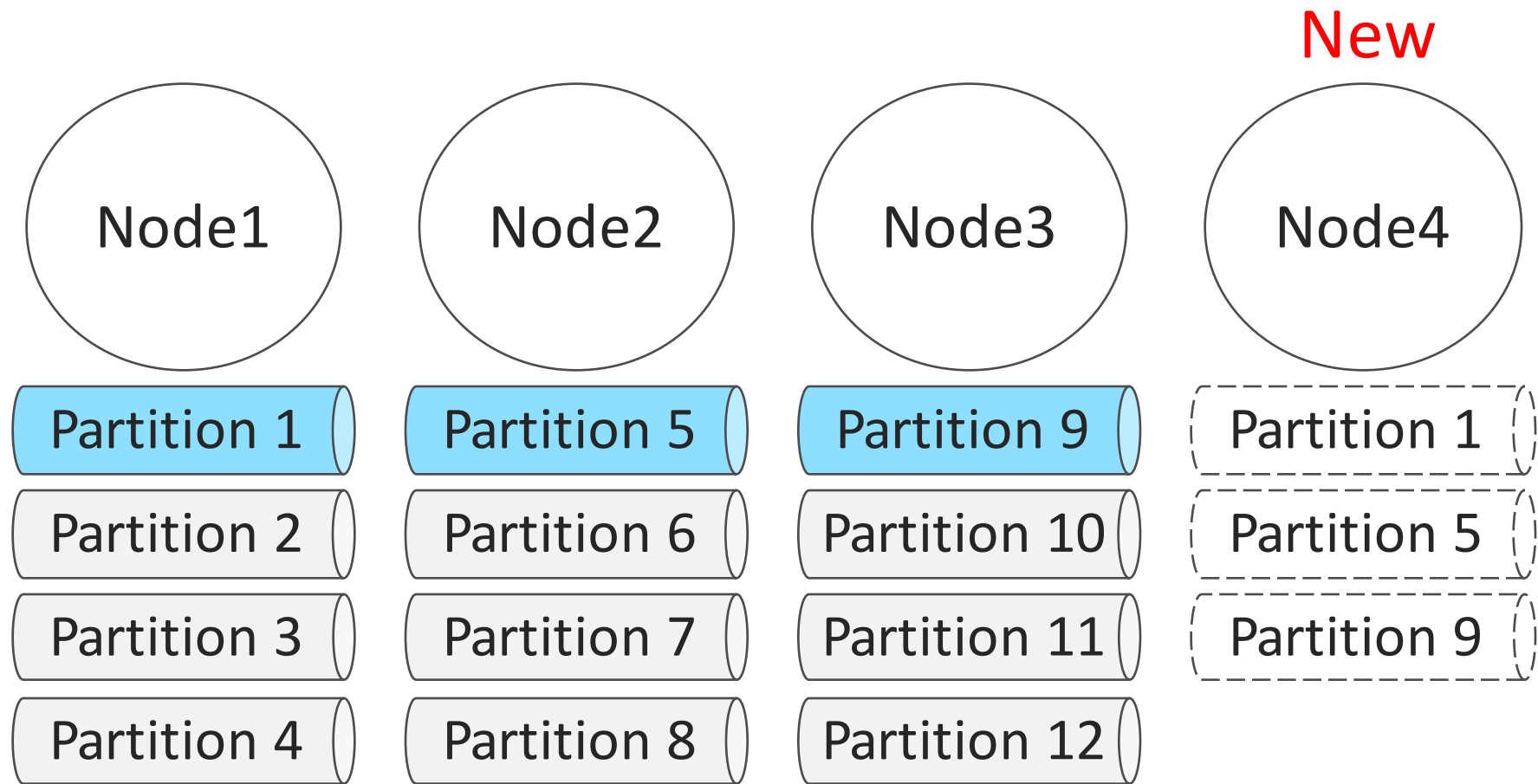
Kafka Scale Out



Kafka Scale Out



Kafka Scale Out



Kafka Scale Out

Example

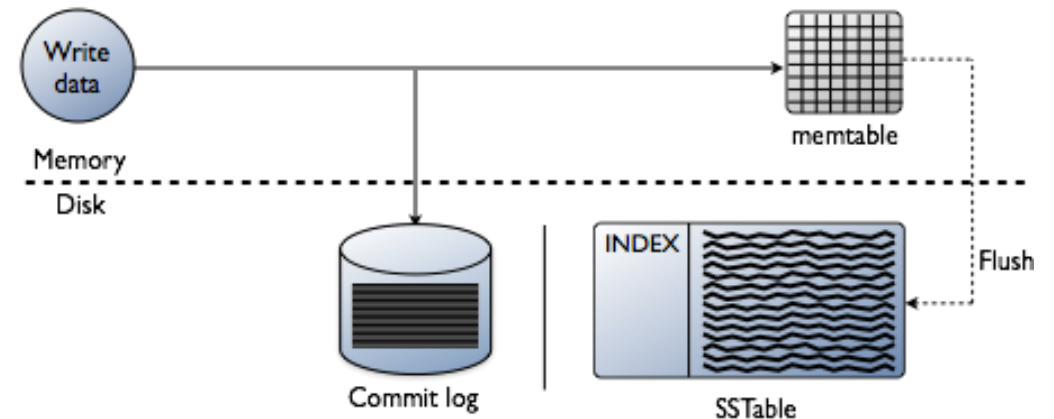
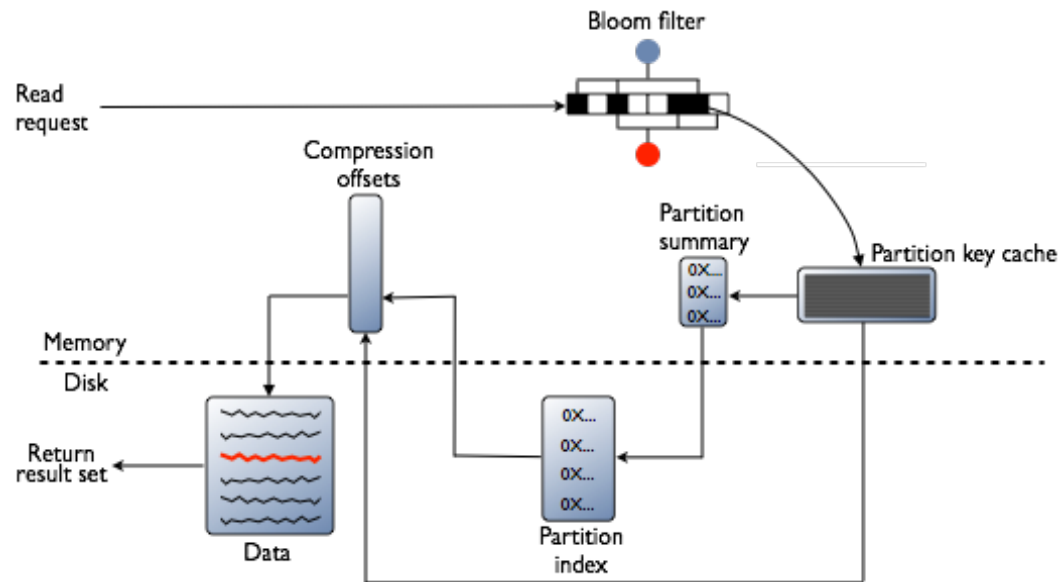
```
bin/kafka-topics.sh --create
  --zookeeper localhost:2181
  --replica-assignment 1,2,3
  --topic test
```

```
--replica-assignment
  <broker_id_for_part1_replica1 :
  broker_id_for_part1_replica2 ,
  broker_id_for_part2_replica1 :
  broker_id_for_part2_replica2 , ...>
```

A list of manual partition-to-broker assignments for the topic being created or altered.

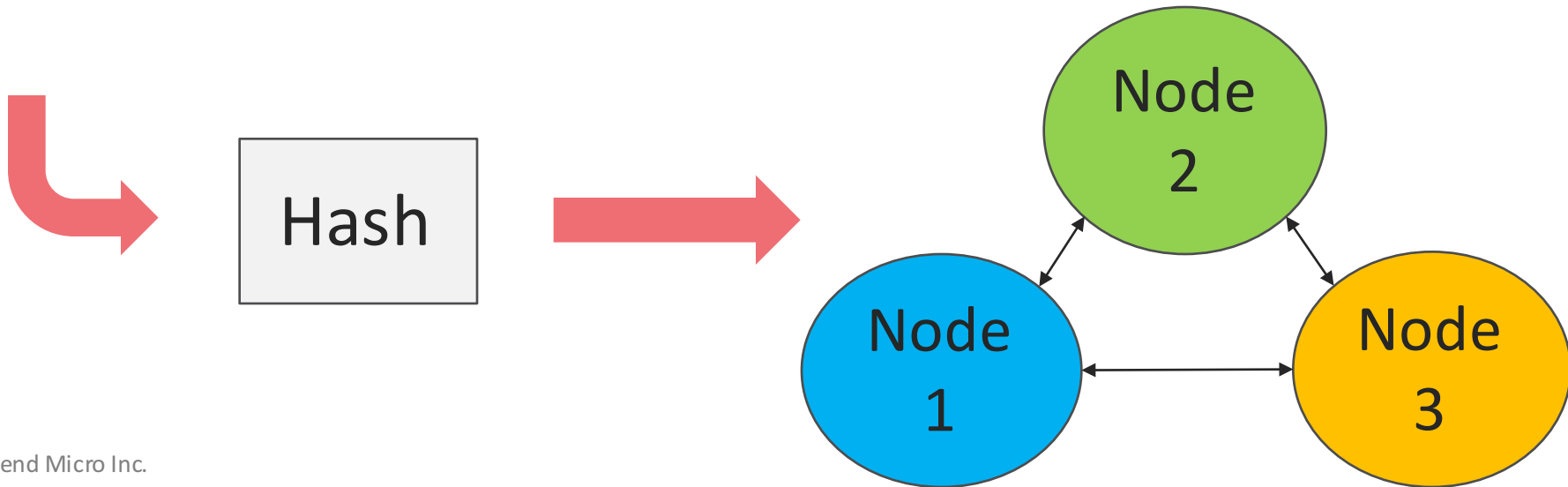
Cassandra Read Path

- Read path is more complicated than write path
- Row cache disable by default



Cassandra Partition

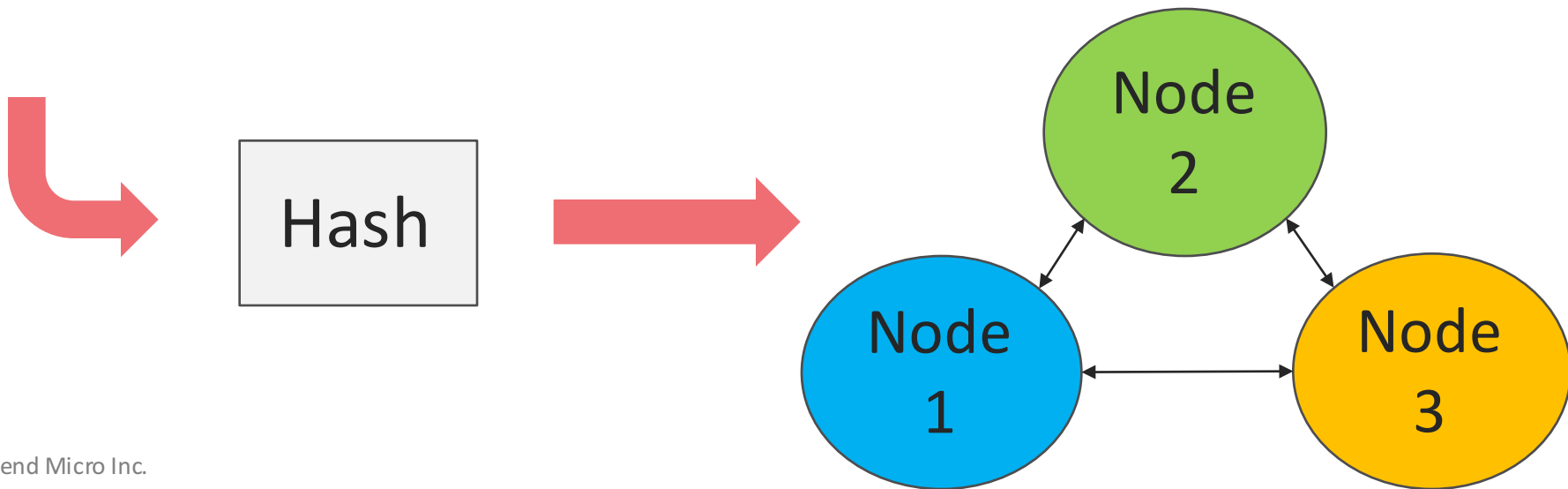
Partition Key	Clustering Key		Column1	Column2
YYMMDDHH	name	addr	...	...
2016120609	Darren	addr1	xxx	xxx
2016120609	David	addr2	ooo	ooo
2016120609	...	...	...	...



Cassandra Partition

Partition Key	Clustering Key		Column1	Column2
(shardId , YYMMDDHH)	name	addr	...	...

(1 , 2016120609)	Darren	addr1	xxx	xxx
(2 , 2016120609)	David	addr2	ooo	ooo
(3 , 2016120609)	...	...	...	...



Spark Pushdown – Cassandra

- Optimize filter operators
- Translate Spark SQL to your data source engine

Spark Pushdown – Cassandra

Partition Key	Clustering Key		Column1	Column2
(shardId, YMMDDHH)	name	addr	...	...

Example

```
select("shardid", "yymmddhh", "name")  
  .where("shardid = 0 AND yymmddhh IN ('2016-07-23-07')")
```

```
C* Filters: [EqualTo(shardid,0), In(yymmddhh,[Ljava.lang.Object;@4b7740e3)]  
Spark Filters []
```

Spark Pushdown – Cassandra

Partition Key	Clustering Key		Column1	Column2
(YYMMDDHH, shardId)	name	addr	...	...

Example

```
select("shardid", "yymmddhh", "name")  
  .where("shardid = 0 AND yymmddhh IN ('2016-07-23-07')")
```

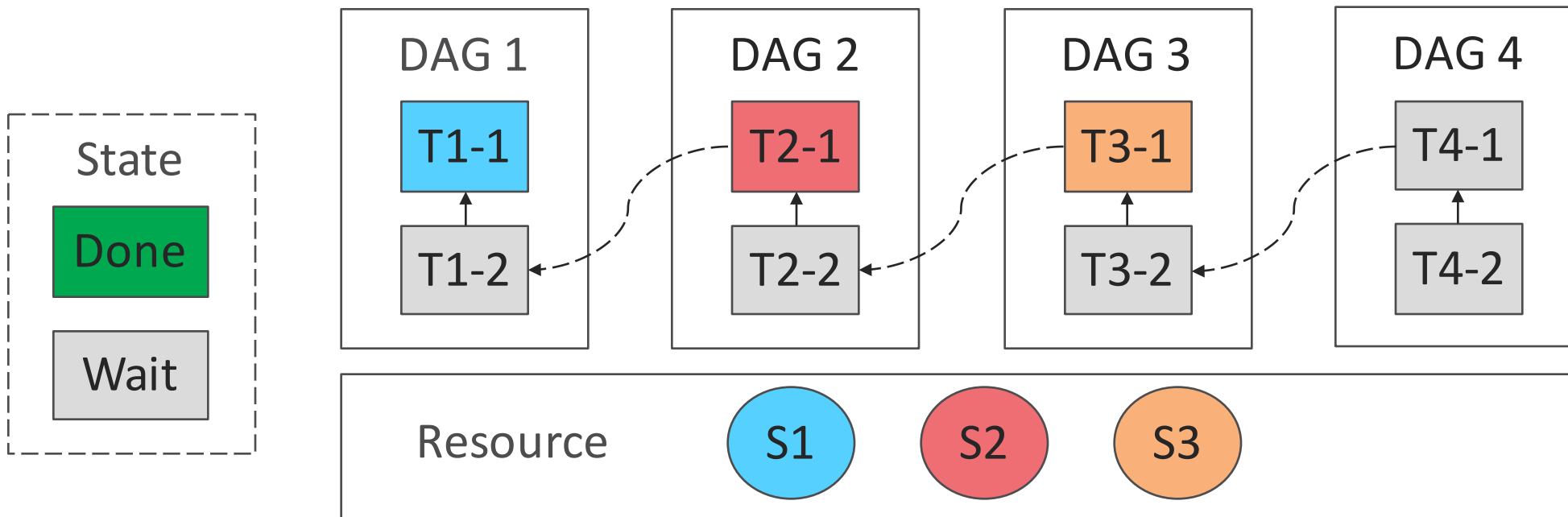
C* Filters: ☐

Spark Filters [In(yymmddhh,[Ljava.lang.Object;@513ad59c), EqualTo(shardid,0)]

Only the last part of the partition key can be an IN predicate

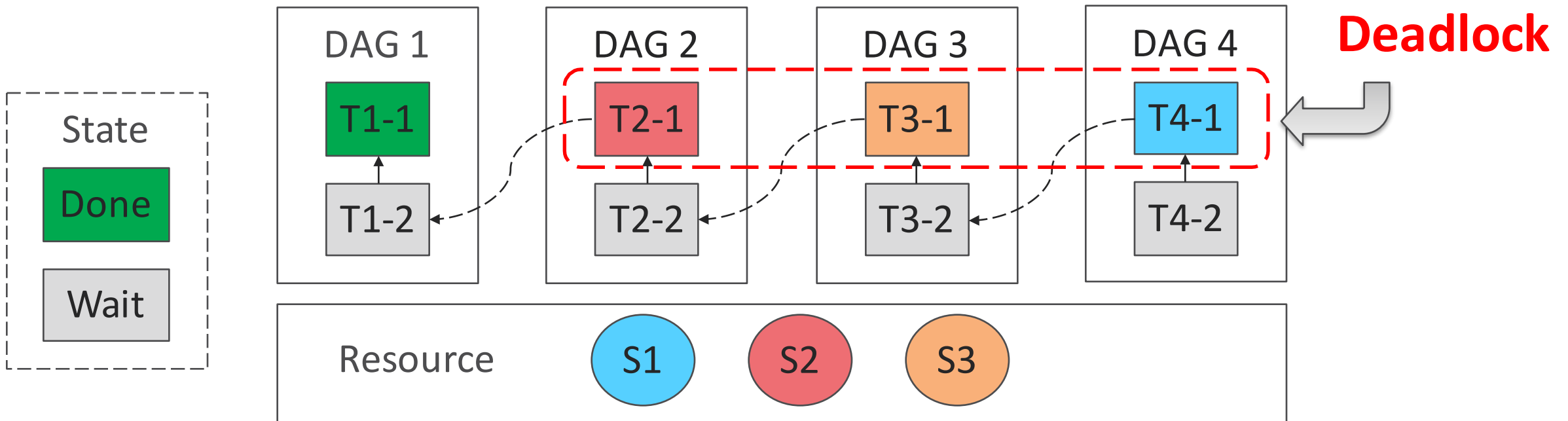
Airflow

- Use “**ExternalTaskSensors**” carefully



Airflow

- Use “ExternalTaskSensors” carefully



Conclusions

Conclusions

- Align open source release cycle
- Understand open source background could help you walk on the right path
- Use state-of-the-art technologies wisely
- Be open mind on open source
- Everything fulfill your needs, it's good enough



**KEEP
CALM
AND
DON'T REINVENT
THE WHEEL**

But you should know
“how to use it”



Q&A

References

- DockerCon 2016

- Video: <https://www.youtube.com/watch?v=j54VzJEwnck&list=PLkA60AVN3hh9gnrYwNO6zTb9U3i1Y9FMY&index=37>
- Slides: <http://www.slideshare.net/Docker/using-the-sdack-architecture-on-security-event-inspection-by-yulun-chen-and-evans-ye>

- iThome Big Data Innovation Summit 2016

- Slides: http://s.itho.me/bigdata/2016/day2/12061310%20iThome_SDACK_Darren_Chen.pdf

- Spark push down

- <https://github.com/datastax/spark-cassandra-connector/blob/master/spark-cassandra-connector/src/main/scala/org/apache/spark/sql/cassandra/BasicCassandraPredicatePushDown.scala>