Autonomous workload rebalancing in Kafka

Indrajeet Kumar

Site Reliability Engineer - LinkedIn

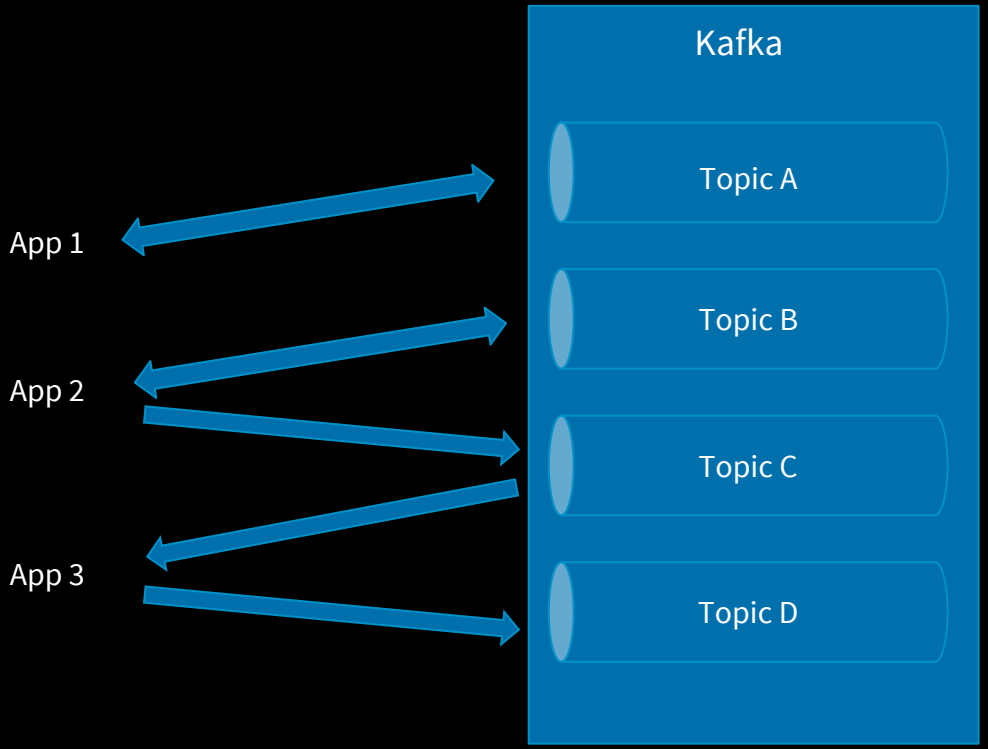
Agenda

- Workload distribution problem
- Manual - Built-in tools
- Semi-automated - Kafka-assigner
- Autonomous - Cruise Control

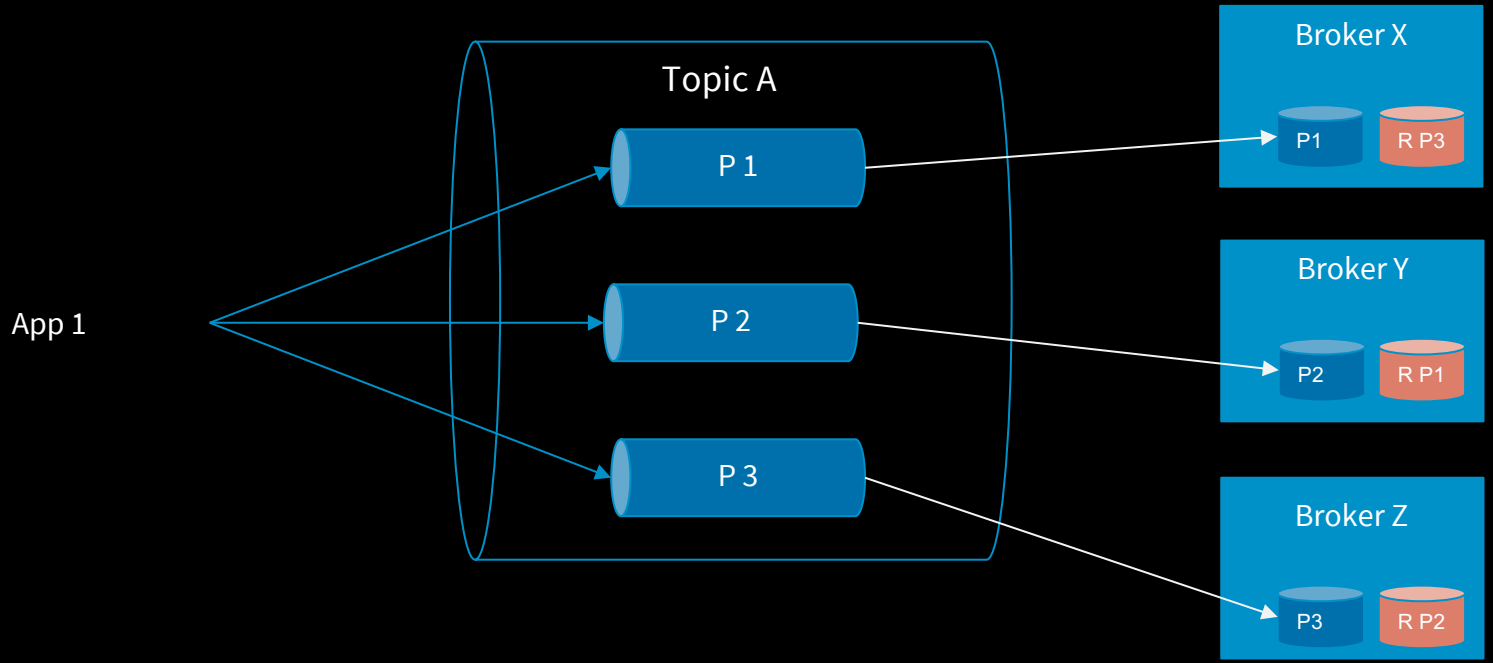
Workload distribution problem

- Important for Distributed Systems
- Harder to work around with Stateful systems

Kafka Overview

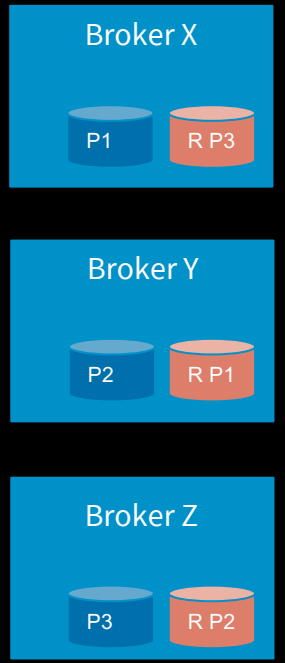


Kafka Overview



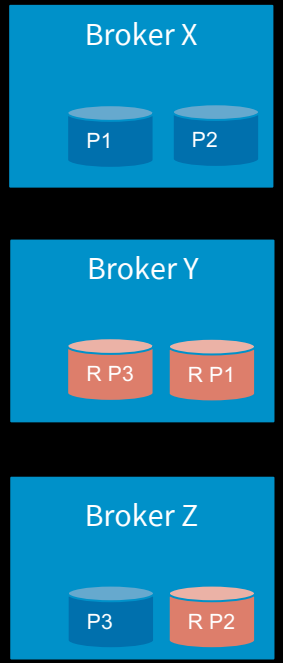
Workload in Kafka

- Leader Partitions
- Total Partitions
- Partition Sizes



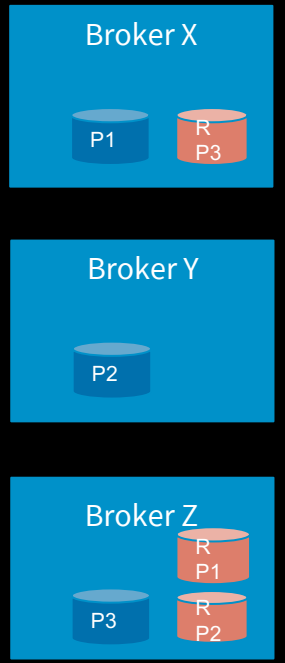
Workload in Kafka

- ☐ **Leader Partitions**
- ☐ Total Partitions
- ☐ Partition Sizes



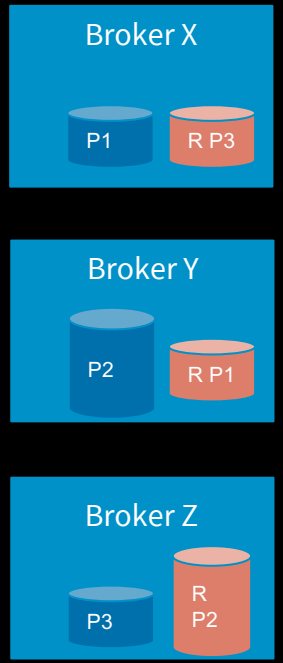
Workload in Kafka

- Leader Partitions
- **Total Partitions**
- Partition Sizes



Workload in Kafka

- Leader Partitions
- Total Partitions
- **Partition Sizes**



Workload distribution problem - Some causes

- Major factors which affect workload balance are:
  - Bad partition distribution
  - Hard host failures
  - Soft host failures
  - Traffic patterns

Kafka workload distribution - Solution

- Rebalance the partitions!
  - Disk usage
  - Network usage
  - Number of partitions
  - Partition leadership count

Usual operations in Kafka

- ☐ Preferred Leader Election
- ☐ Partition rebalance
- ☐ Bump Partition counts
- ☐ Add/Remove brokers

Kafka at LinkedIn

Kafka at LinkedIn

- 4.5 Trillion messages a day
- 2500+ kafka brokers
- 1 PB In
- 3.9 PB Out

Kafka admin utilities

- Out of the box tools:
  - bin/kafka-reassign-partitions.sh
  - bin/kafka-preferred-replica-election.sh

Example run of built-in tools

- Rebalancing Partitions:

```
$ cat topics-to-move.json
{"topics":
    [{"topic": "foo1"},{"topic": "foo2"}],
    "version":1
}

$ ./bin/kafka-reassign-partitions.sh --topics-to-move-json-file topics-to-move.json --broker-list "5,6,7" --generate

$ cat partitions-to-move.json
{"partitions":
            [{"topic": "foo",
              "partition": 1,
              "replicas": [1,2,4] }],
  "version":1
}

$ ./bin/kafka-reassign-partitions.sh --reassignment-json-file partitions-to-move.json --execute
```

Problems with stock tools

- Manual
- Less optimal
- Slow

Kafka Assigner

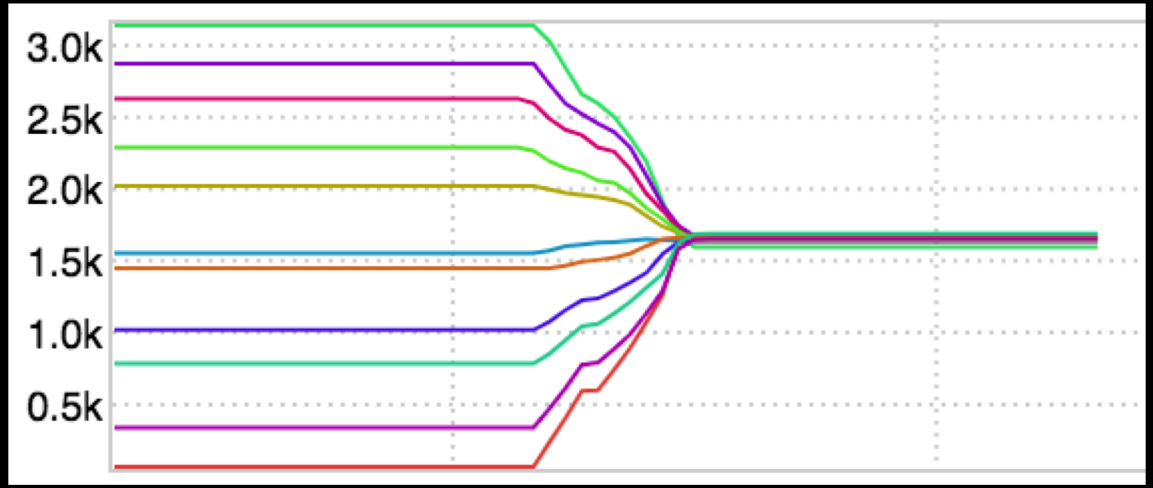
Kafka assigner

- High level administrative commands
- Under the hood, it uses the ‘kafka-utils/bin/’ scripts
- It also allows to do complex rebalances with multiple goals

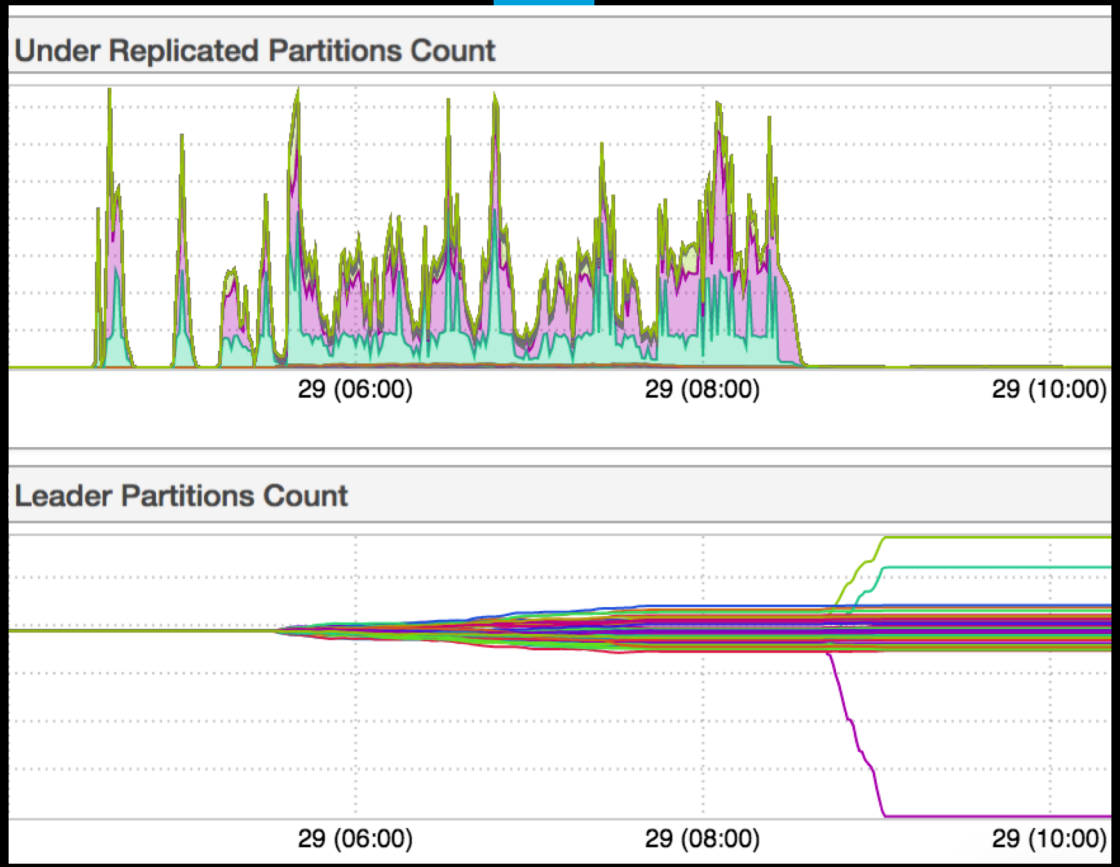
Kafka assigner

```
reorder                 Reelect partition leaders using replica reordering
balance                 Rebalance partitions across the cluster
elect                   Reelect partition leaders using preferred replica
                        election
trim                    Remove partitions from some brokers (reducing RF)
remove                  Move partitions from one broker to one or more other
                        brokers (maintaining RF)
set-replication-factor
                        Increase the replication factor of the specified
                        topics
clone                   Copy partitions from some brokers to a new broker
                        (increasing RF)
```

Preferred Leader election



Case of URPs


Under Replicated Partitions Count
29 (06:00)
29 (08:00)
29 (10:00)
Leader Partitions Count
29 (06:00)
29 (08:00)
29 (10:00)

Kafka assigner

- Pros:
  - High level admin commands
  - Simple to use
  - Allows chaining rebalance goals
  - Easy to remove all partitions from a broker

Kafka assigner

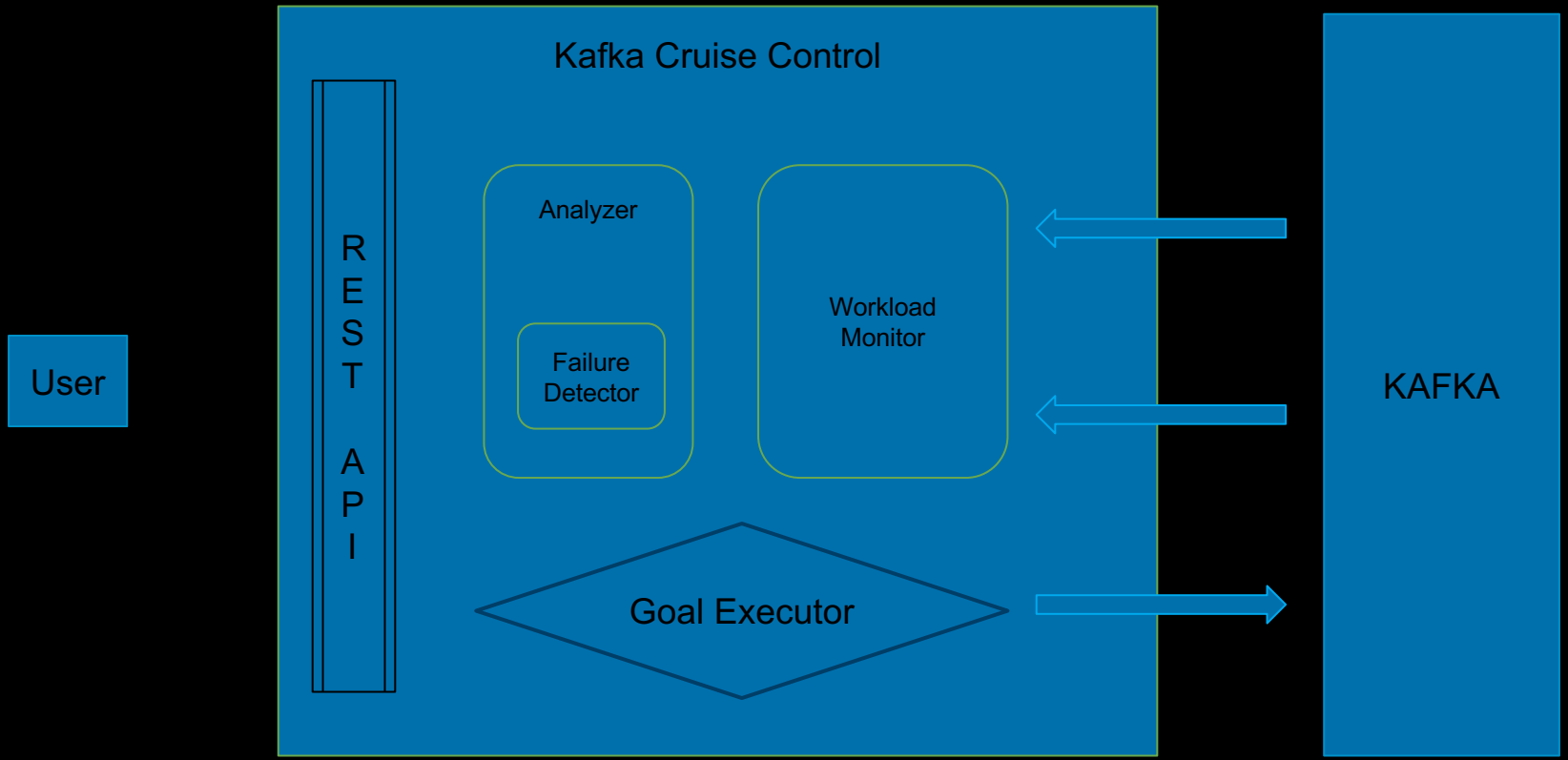
- Cons:
  - Where did you run it?
  - In-optimal balances in certain cases
  - Needs manual invocation and supervision

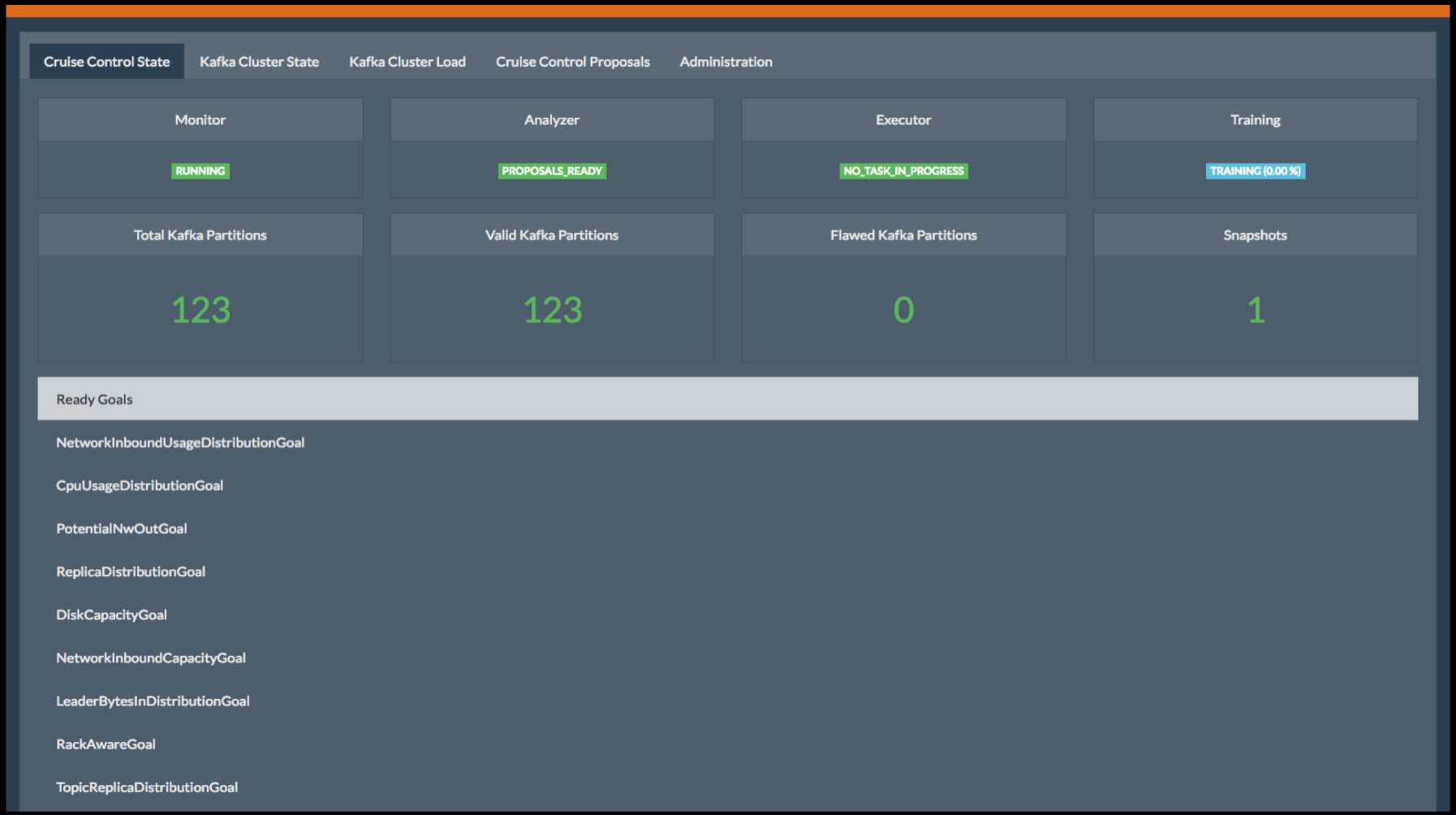
Cruise Control

Cruise Control

- Central System
- Complete live health of the cluster
- Manual/Automatic management of workload

Design


Kafka Cruise Control
REST API
Analyzer
Failure Detector
Workload Monitor
Goal Executor
User
KAFKA


Cruise Control State
Kafka Cluster State
Kafka Cluster Load
Cruise Control Proposals
Administration
Monitor
RUNNING
Analyzer
PROPOSALS_READY
Executor
NO_TASK_IN_PROGRESS
Training
TRAINING (0.00 %)
Total Kafka Partitions
123
Valid Kafka Partitions
123
Flawed Kafka Partitions
0
Snapshots
1
Ready Goals
NetworkInboundUsageDistributionGoal
CpuUsageDistributionGoal
PotentialNwOutGoal
ReplicaDistributionGoal
DiskCapacityGoal
NetworkInboundCapacityGoal
LeaderBytesInDistributionGoal
RackAwareGoal
TopicReplicaDistributionGoal

Cruise Control State | Kafka Cluster State | Kafka Cluster Load | Cruise Control Proposals | **Administration**

Administrative Section

ALERT: Any Actions that you do in this section will have consequences on your Kafka Cluster. Please think twice before executing these actions.

1. Add broker to kafka cluster
2. Remove broker from kafka cluster
3. Demote broker from kafka cluster
4. Rebalance kafka cluster
5. Stop Execution

Broker Administration

Replicas	Host	Broker	Status	Disk	CPU	Leader In	Follower In	Out	Potential Out	[]
115	nareshv-mn1	1	ALIVE	24 KB	0 %	48 Bps	26 Bps	97 Bps	150 Bps	☐
104	nareshv-mn1	2	ALIVE	65 KB	0 %	107 Bps	34 Bps	149 Bps	218 Bps	☐
18	nareshv-mn1	3	ALIVE	10 KB	0 %	9 Bps	24 Bps	18 Bps	66 Bps	☑

Flags: ☑ Dryrun ☑ Kafka Assigner Mode
Remove Broker(s) url: http://localhost:8080/kafkacruisecontrol/remove_broker?kafka_assigner=true&dryrun=true&brokerid=3&json=true
Demote Broker(s) url: http://localhost:8080/kafkacruisecontrol/demote_broker?kafka_assigner=true&dryrun=true&brokerid=3&json=true

Remove 1 Broker | Demote 1 Broker

Cruise Control

Balancing Performance

Racks: 10

Brokers: 40

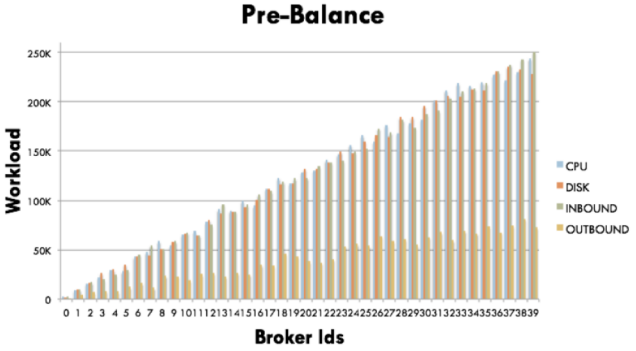
Entities: 50K

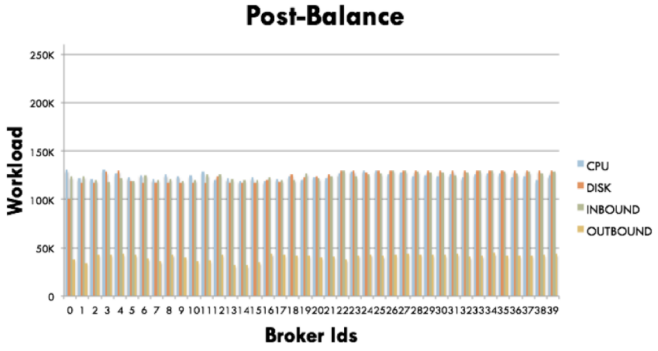
Topics: 3K

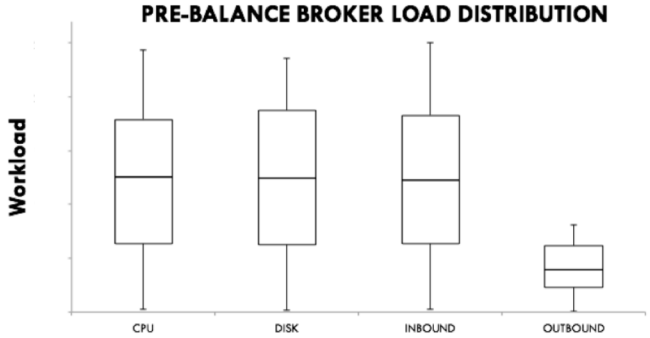
Replication Factor: 3

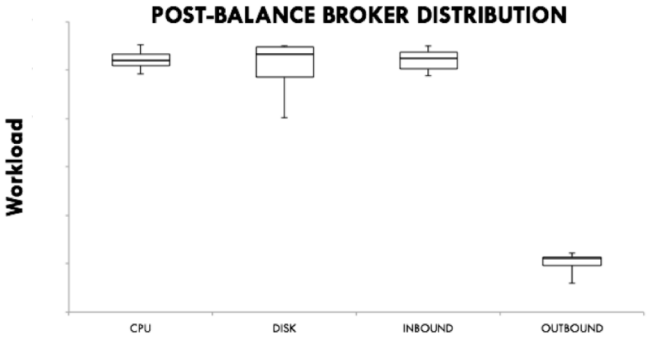
Entity distribution: Exponential

Balance percentage (for all resources): 1.05









CC setup requirements

- Kafka > 0.11.0.0
- Drop in jar

Features already built-in

- ☐ Resource utilization tracking
- ☐ Multi-goal rebalance
- ☐ Anomaly detection
- ☐ Admin operations

How is CC doing?

- Save SRE's time to debug/fix kafka workload issues
- Very fast operations
- Central place to look at for globally distributed teams
- Self-heal !!

Resources

Kafka shipped admin-tools:
https://github.com/apache/kafka/tree/trunk/bin

Kafka Assigner:
https://github.com/linkedin/kafka-tools/wiki/Kafka-Assigner

Cruise Control:
https://github.com/linkedin/cruise-control

Connect with me: https://www.linkedin.com/in/indrajeetkm/

Questions