Kafka partition 搬移成本估計

演講者 : 孫祥鈞

October 2022 @ JCConf

Link to the Astraea Project

https://github.com/skiptests/astraea

關於我

- 演講者
  - 孫祥鈞
- 成功大學的研究生
  - 研究 Kafka 的負載平衡問題
  - https://hackmd.io/@XiangJunSun/XiangJunSun
- 維護 Astraea 專案
  - 一系列 Kafka 維運工具
  - https://github.com/skiptests/astraea

- 核心開發者
  - 孫祥鈞, 方竫泓, 李宜桓, 李政憲, 王懿宸, 蔡嘉平, 蕭宏章, 鄧智懋, 魏連興, 陳嘉晟, 李兆恆
- 特別感謝
  - 成功大學
  - 原昌工業
  - 教育部
  - 來自其他公司的工程師的貢獻

今天的演講

大綱

1. Kafka Topic & Partitions
2. reassign partition & 情境
3. partition搬移成本估計
4. 結論

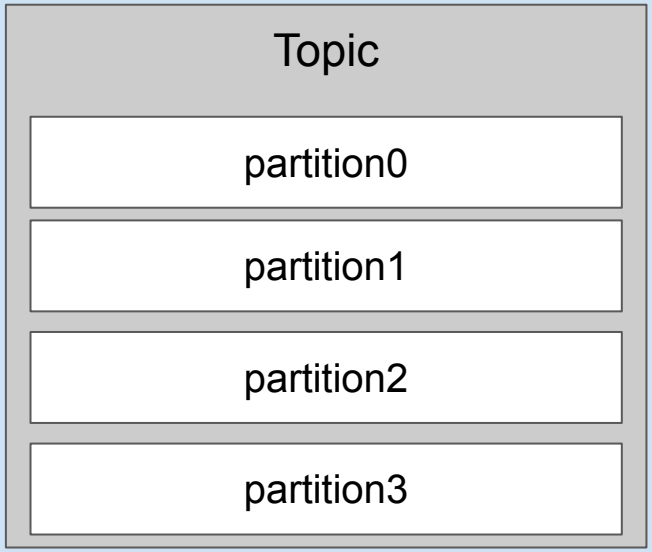
Kafka Topic & Partitions

reassign partition & 情境

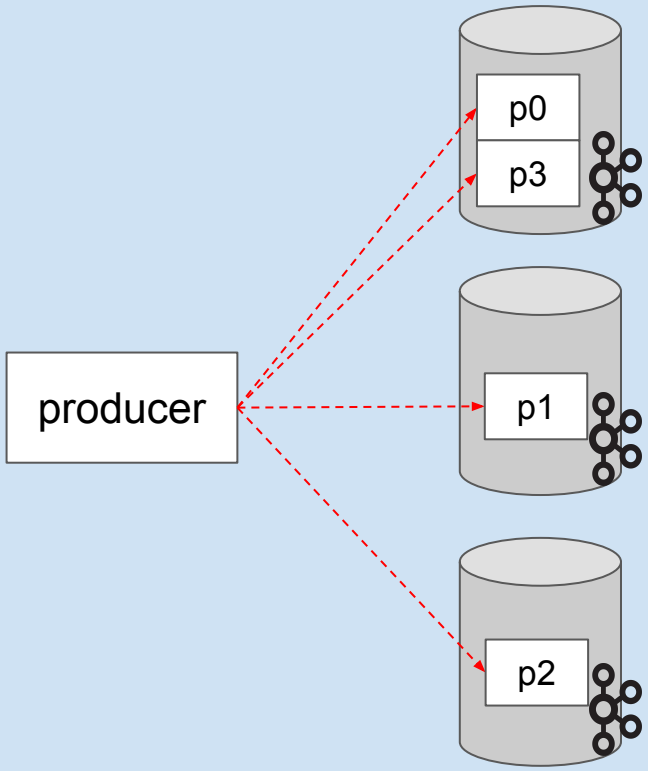
partition搬移成本估計

結論

Topic & Partitions



- Topic
  - Topic 是一個邏輯上的概念，通常producer打資料會以Topic為單位去打
- Partitions
  - 由Topic 切分出來，Partition 可以分散式的存放在不同機器中，以防止單台機器故障



Kafka Topic & Partitions

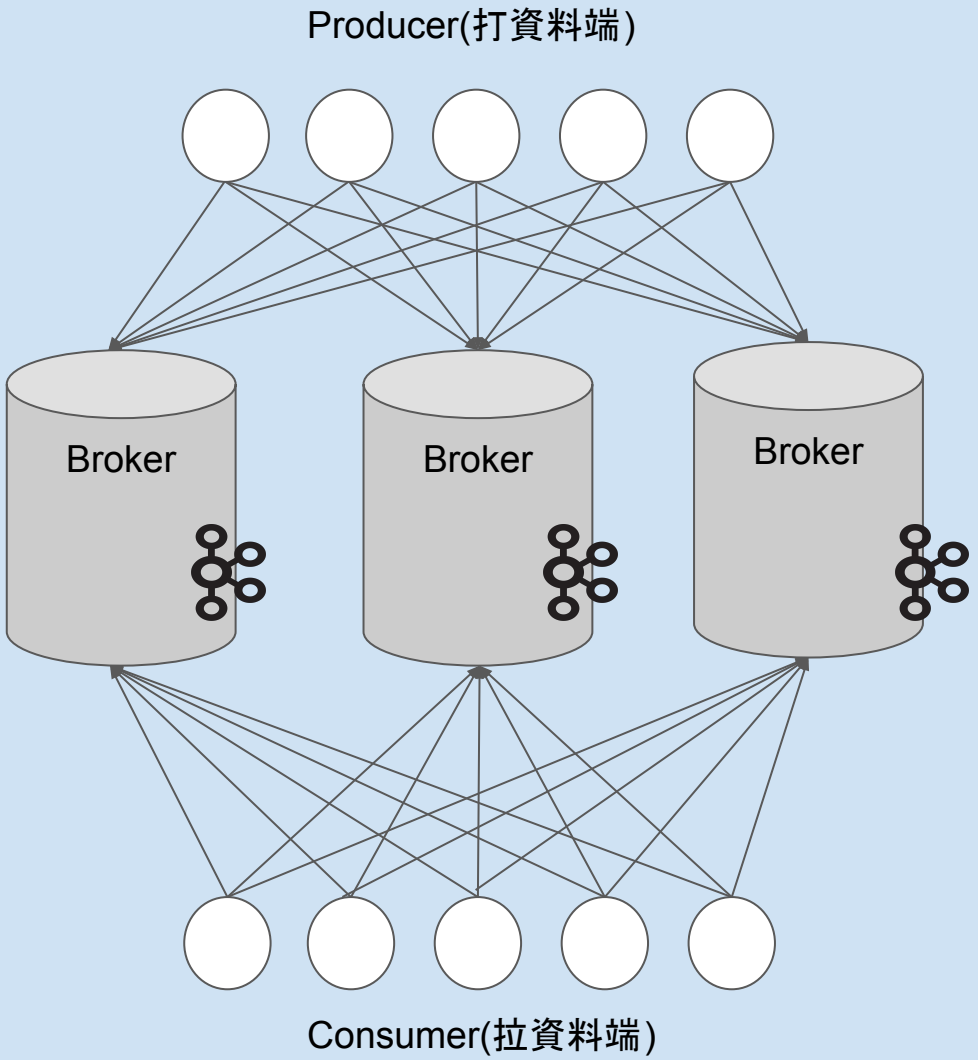
reassign partition & 情境

partition搬移成本估計

結論

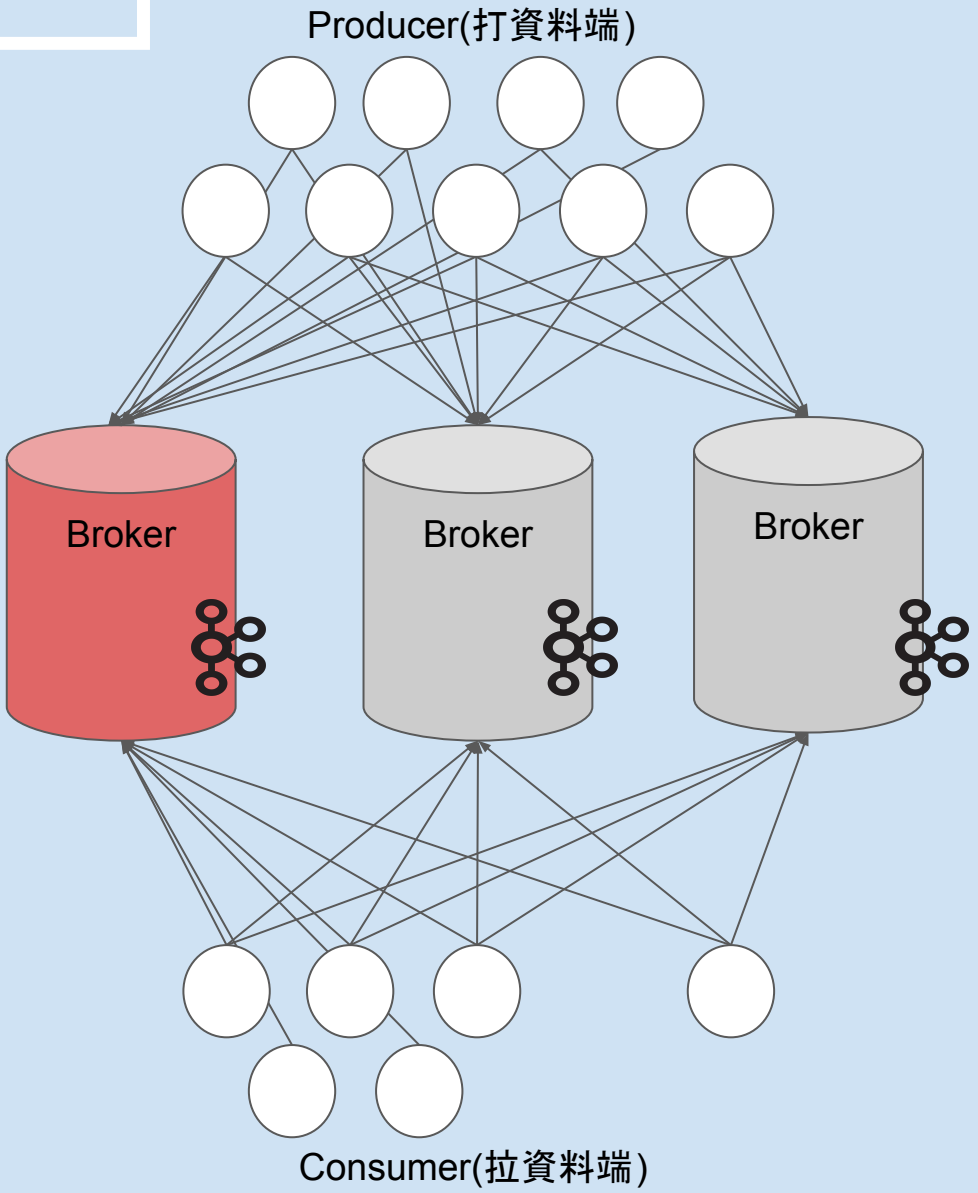
Kafka 叢集負載平衡議題(1/2)

- 新叢集
  - **表現正常**



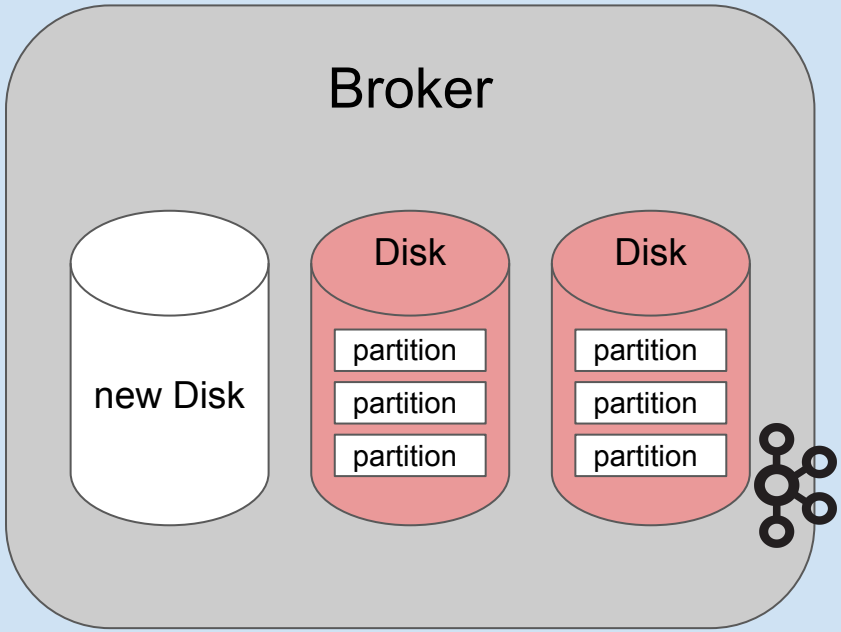
Kafka 叢集負載平衡議題(2/2)

- 新叢集
  - 表現正常
- 叢集使用
  - **上層業務需求變化**



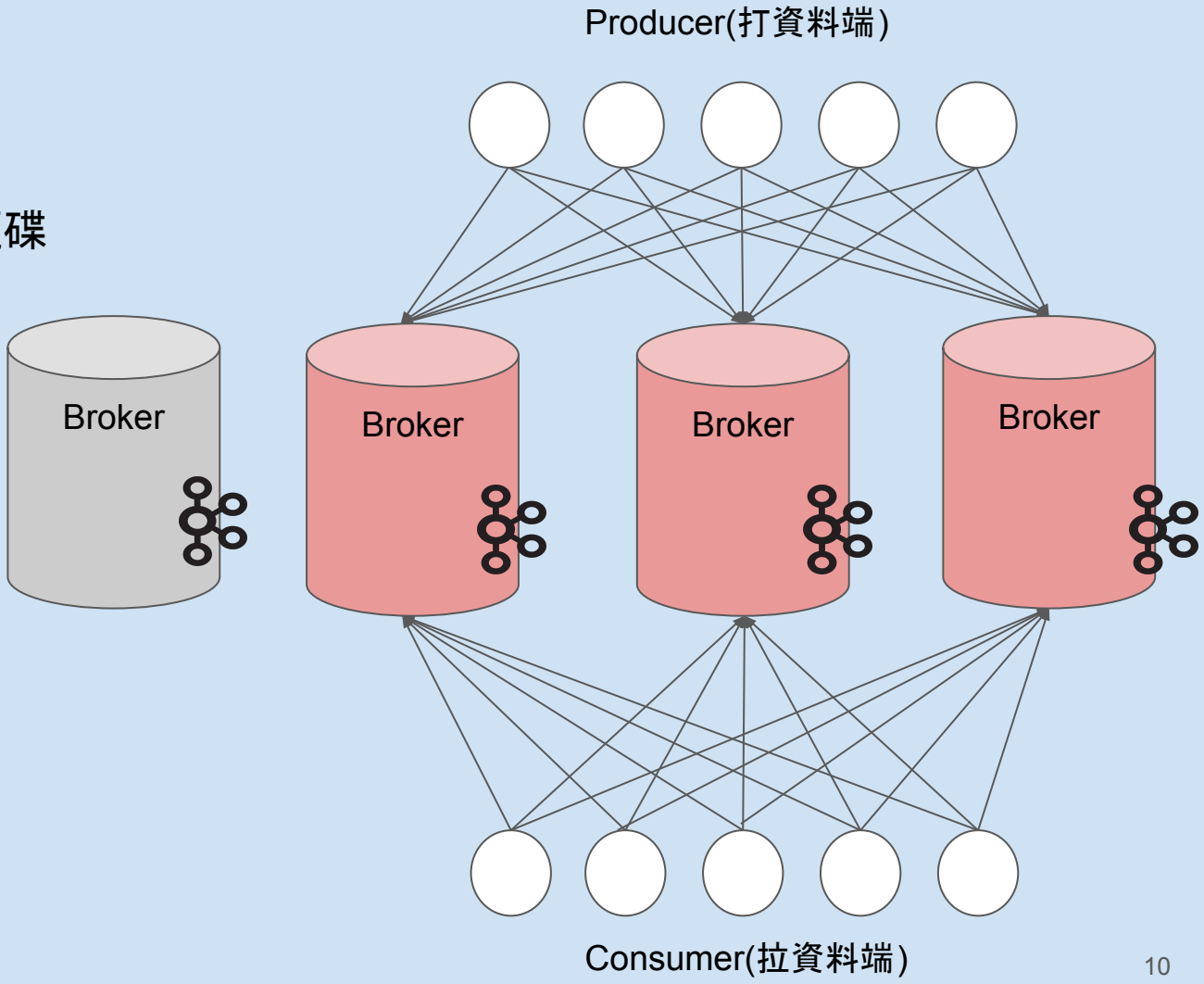
擴展叢集(1/2)

- 叢集使用一段時間後
  - **Broker新增/更換硬碟**



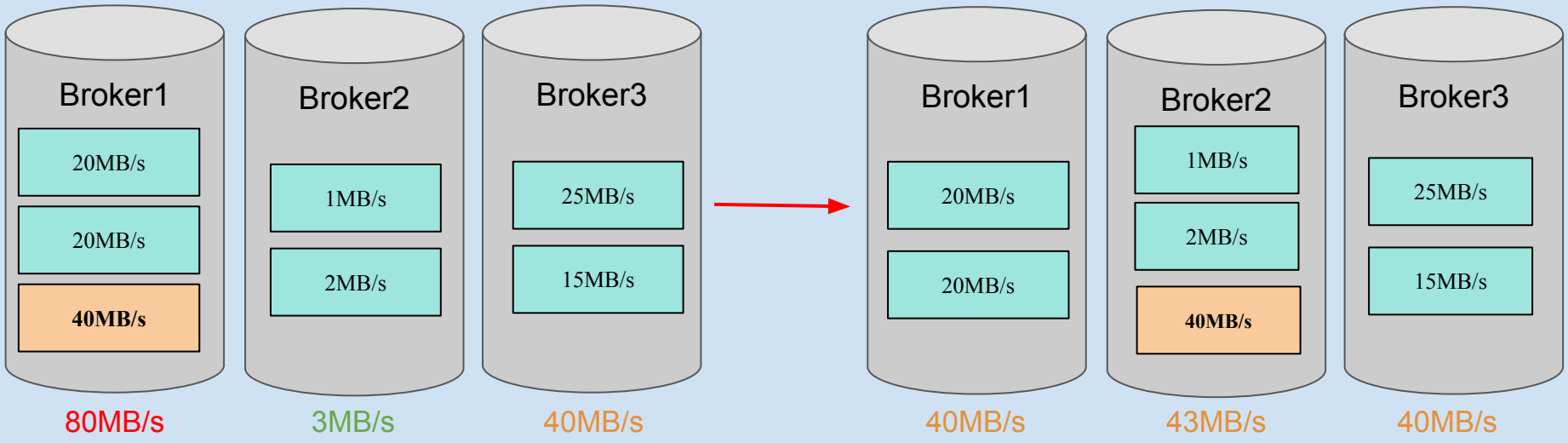
擴展叢集(2/2)

- 叢集使用一段時間後
  - Broker新增/更換硬碟
  - **增加/取代節點**



kafka reassign API

- Apache Kafka 提供了一個API可以調整叢集的partition分佈
  - 只需要簡單使用API就可以了嗎?



Kafka Topic & Partitions

reassign partition & 情境

partition搬移成本估計

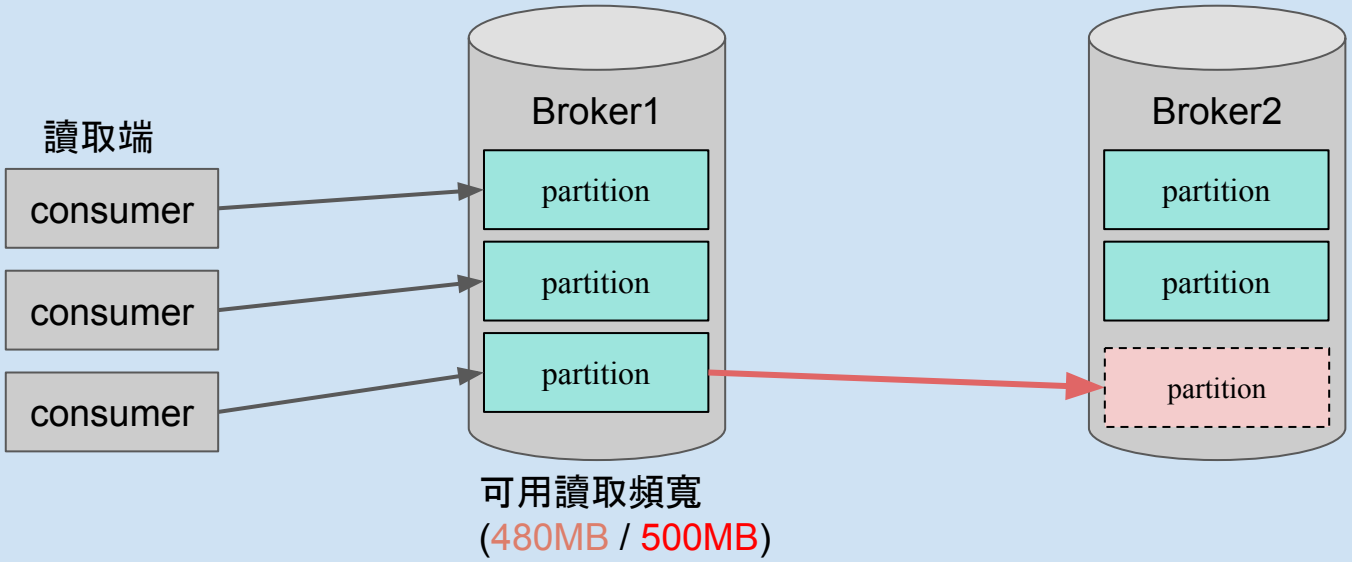
結論

為何要有成本估計(1/2)

- 在搬移partition過程中會占用硬體資源
  - 網路讀取/寫入流量
    - 讀取負責接收資料的partition
    - 寫入搬移目的的partition
  - 磁碟使用率
    - 讀取負責接收資料的partition
    - 寫入搬移目的的partition
  - 磁碟空間
    - partition會佔用儲存空間

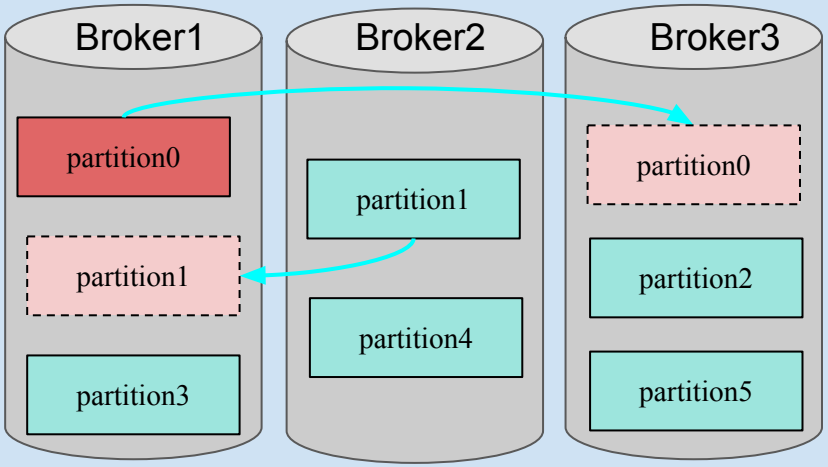
為何要有成本估計(2/2)

- 不顧一切的隨意搬移可能會發生什麼事?
  - **上層的應用受到影響**



為何要有成本估計(2/2)

- 不顧一切的隨意搬移可能會發生什麼事?
  - 上層的應用受到影響
  - **搬移過程中磁碟空間佔用太多**

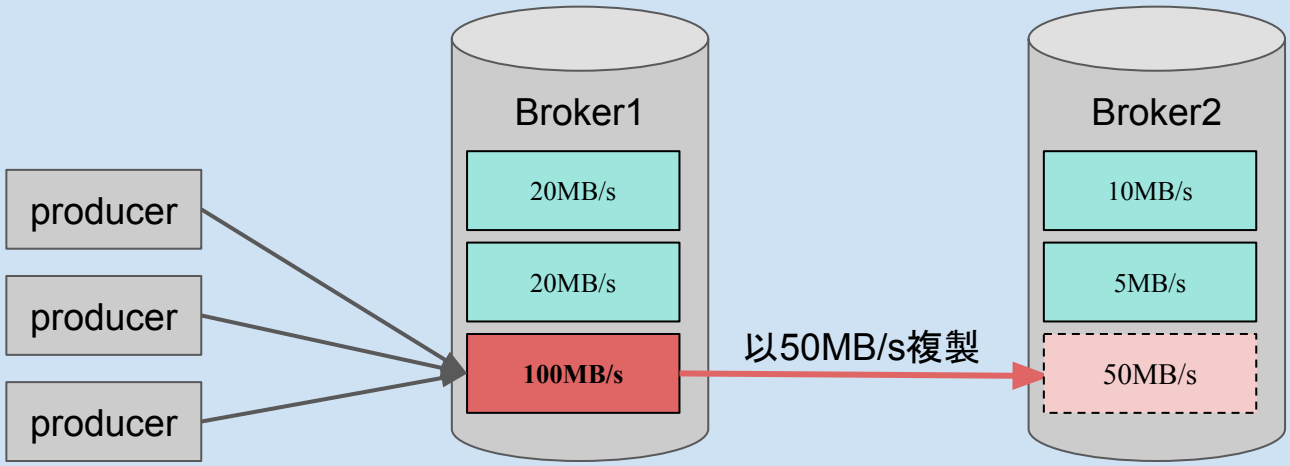


搬移計畫

	from	to
partition0	Broker1	Broker3
partition1	Broker2	Broker1

為何要有成本估計(2/2)

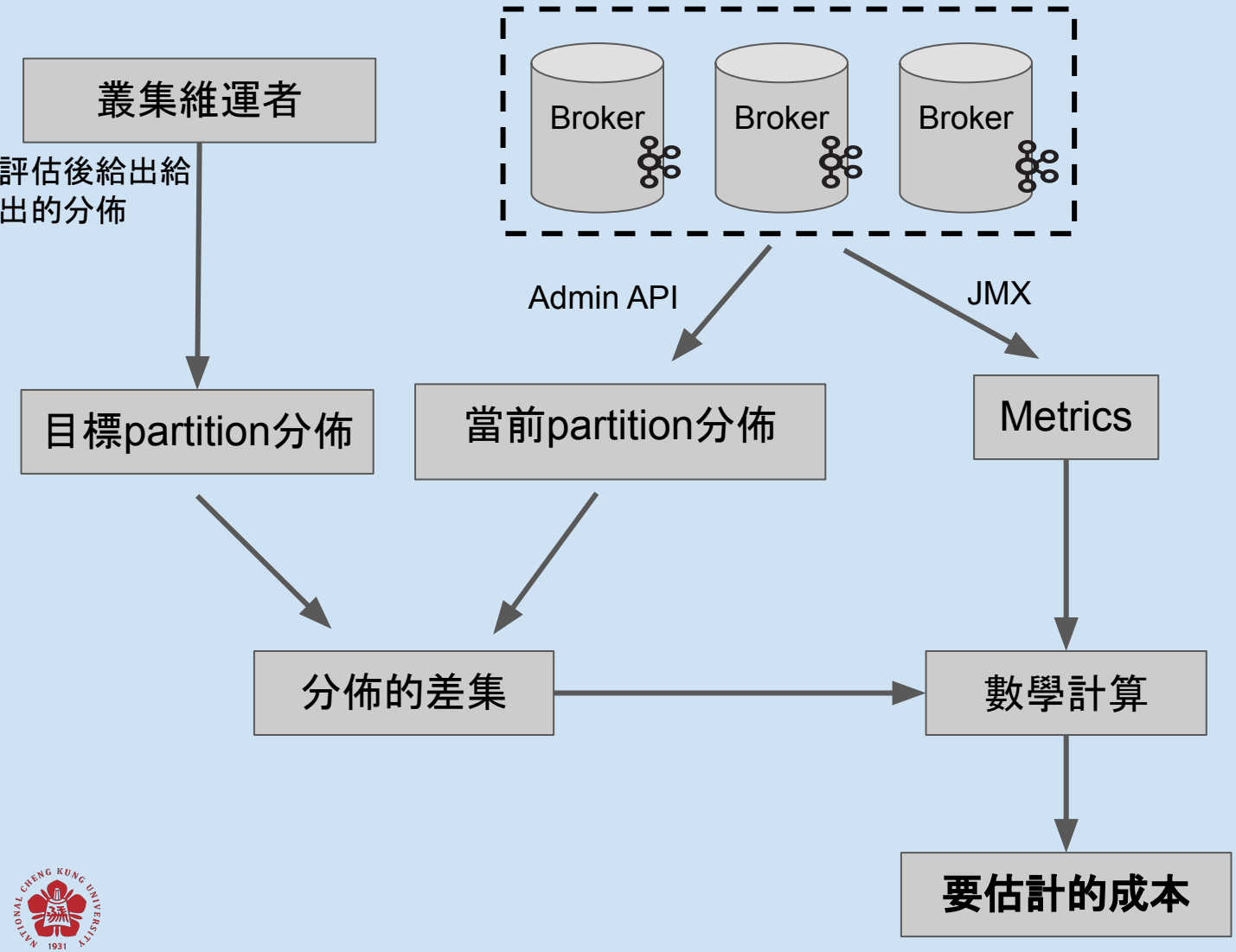
- 不顧一切的隨意搬移可能會發生什麼事?
  - 上層的應用受到影響
  - 搬移過程中磁碟空間佔用太多
  - **partition搬移時間太長**



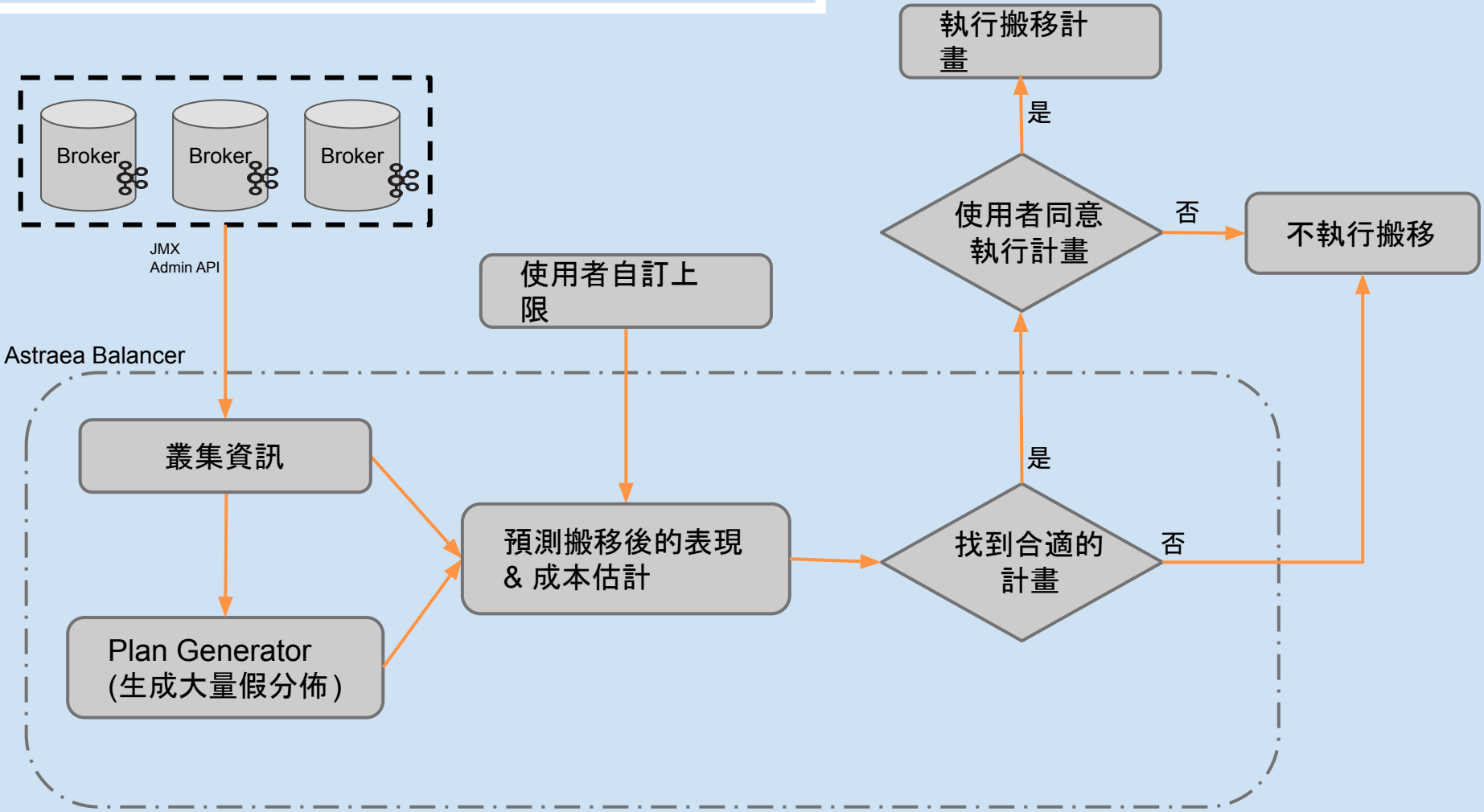
如何估計成本(1/2)

- 取得當前叢集資訊
  - 透過Kafka Admin API取得partition分佈
  - 透過JMX取得partition log size, broker 讀取寫入流量等資訊
- 使用者輸入自訂搬移資源上限
  - 搬移的總partition數量
  - 搬移的總partition size
  - 搬移預計要占用網路頻寬
  - 搬移預計的搬移時間

如何估計成本(2/2)


叢集維運者
評估後給出給出的分佈
Broker
Broker
Broker
Admin API
JMX
目標partition分佈
當前partition分佈
Metrics
分佈的差集
數學計算
要估計的成本

套用成本估計到Balancer(1/2)



套用成本估計到Balancer(2/2)

- 執行搬移計畫前


```
"changes":[
      {
         "topic":"test-3",
         "partition":1,
         "before":[
            {
               "brokerId":1001,
               "directory":"/tmp/log-folder-2",
               "size":18012068
            },
            {
               "brokerId":1002,
               "directory":"/tmp/log-folder-2",
               "size":18012068
            }
         ],
         "after":[
            {
               "brokerId":1001,
               "directory":"/tmp/log-folder-2"
            },
            {
               "brokerId":1003,
               "directory":"/tmp/log-folder-1"
            }
         ]
      }
```

```
"migrationCosts":[
      {
         "function":"size",
         "totalCost":93634045,
         "cost":[
            {
               "brokerId":1001,
               "cost":-11997701
            },
            {
               "brokerId":1002,
               "cost":14824599
            },
            {
               "brokerId":1003,
               "cost":-2826898
            }
         ],
         "unit":"byte"
      }
   ]
```

Kafka Topic & Partitions

reassign partition & 情境

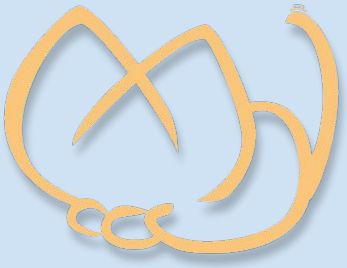
partition搬移成本估計

結論

結論

- 成本估計在搬移partition有一定的幫助
  - 確保搬移不會佔用太多資源
  - 減少搬移花費時間
  - 保證前端應用可以正常運行

感謝聆聽



Link to the Astraea Project

https://github.com/skiptests/astraea