Apache Kafkaでの大量データ処理がKubenetesで簡単にできて嬉しかった話

Fluentd,Spark Streaming

株式会社マイクロアド 大澤 昂太

自己紹介など

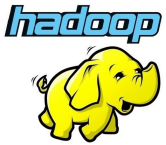
大澤 昂太

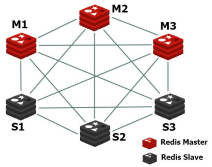
サーバサイドエンジニアをしています。

分散システム関連の開発が多いです。

K8sは何回か勉強したのですがあまり理解できなかったです。

Dockerはよく使います





株式会社マイクロアド

インターネット広告の会社です。

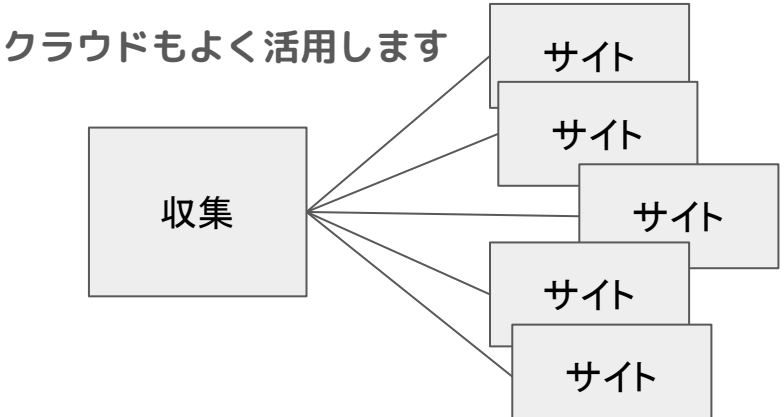
設計に起因する苦しい時代を経験しているので設計へのこだわりが強いです。

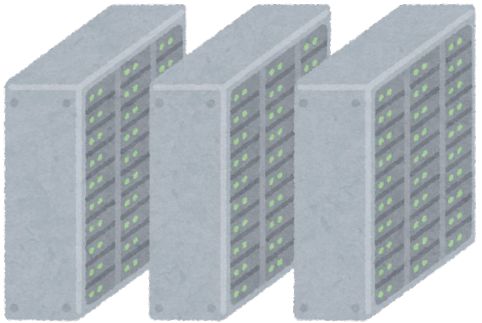
関数型言語のScalaやサーバサイドKotlinなどJVMの活用が多いです。

広告業界とデータセンターについて

- 広告業界では自社で取り扱ってないサイトのトラフィックを取り扱う
- このトラフィックは一日TB単位になることがある

- **休みなく**24時間動き続ける数千台単位のサーバが欲しくなるのでオンプレのデータセンターのほうがコスパが良い機能が多い（と思われる）

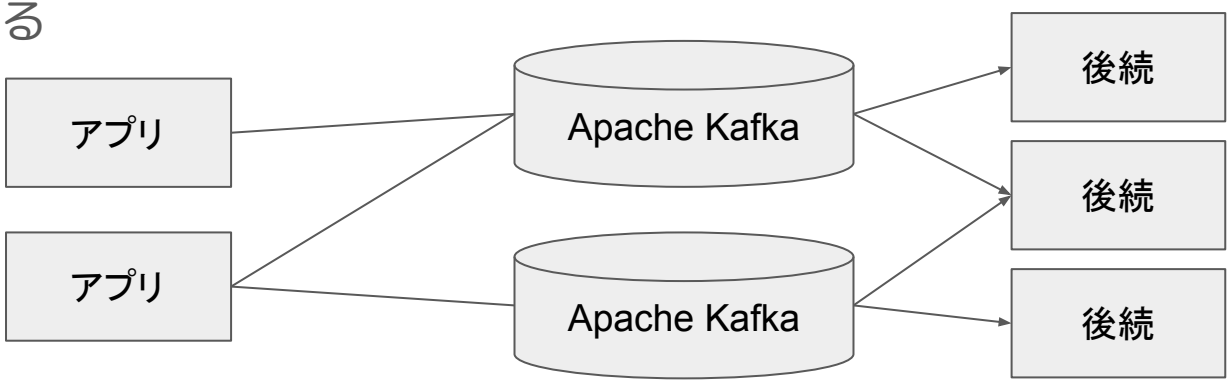




Apache Kafkaの活用について

- Kafkaとはリアルタイムキューでリアルタイムにログを流したりできるデータベースに近いもの
- 一般的なPub/Subとの違いとしてクラスタリングができ、大規模なデータでも転送できる

- 弊社ではトラフィックを解析する機能が多いのでApache Kafkaを多数活用している

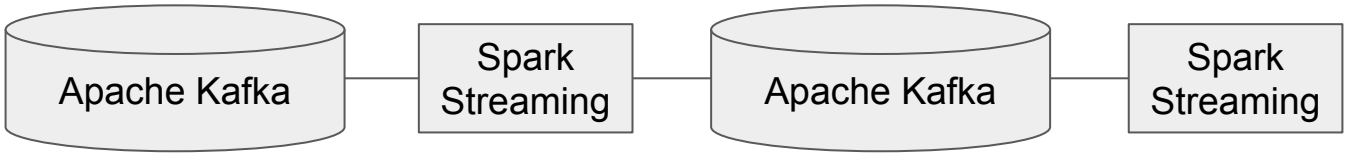


後続が大量に増えてもスケールできる！

Spark Streamingの活用と問題について

- 弊社ではApache Kafkaのデータ処理をする際には主にSparkストリーミングを活用している

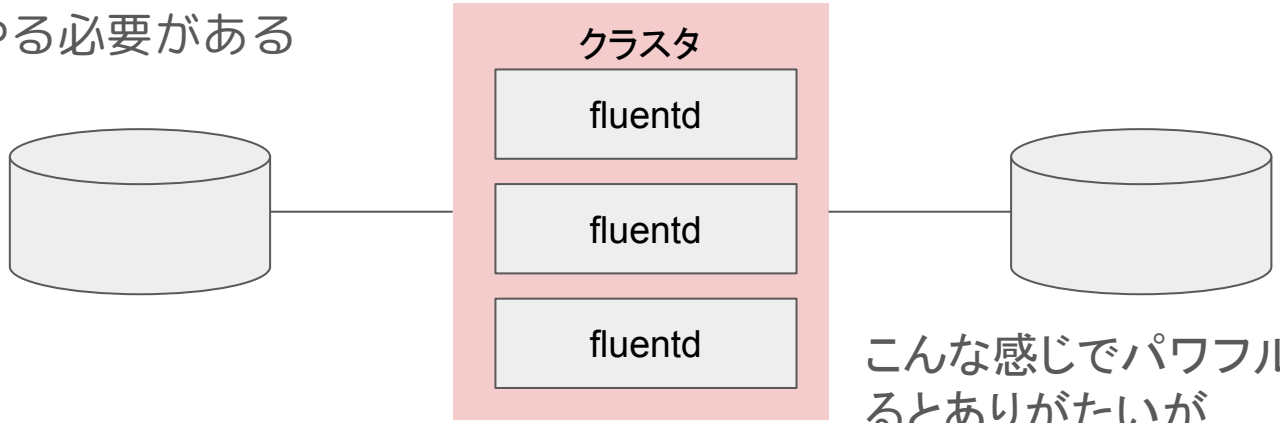
- 問題点として軽い機能でもそれなりの規模の開発が必要
- （弊社Hadoopの問題で）オーバーコミットが安定せずCPUコアの消費量が多い
- Hadoopクラスタが老朽化していてどうにかしたい



Fluentdの活用と弱点

- Fluentdはストリーム処理を行うことができるミドルウェア
- 設定ファイルを少し書くだけでデータ転送等を行うことができる
- 弊社でもよく使われるが利用用途は限られていた

- 複数サーバやマルチコアCPUを活用することが難しい
- 対Kafkaだと擬似的にクラスタリングできるが、デプロイ周りを上手く作ってやる必要がある

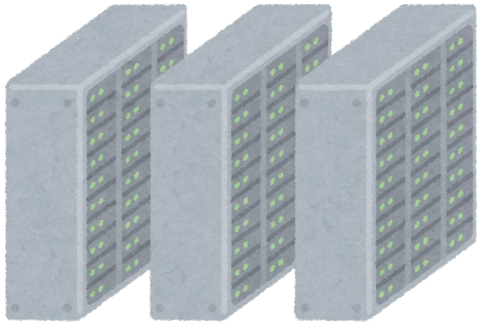


こんな感じでパワフルに動いてくれるとありがたいが…

会社でのKubernetesの導入

- マイクロアドでは数年くらいK8sの準備をインフラチームの方で進めている
- それ以外のサーバはDockerデーモンを直で使っていることが多い

- ある時、リアルタイム処理の開発が必要になったが大した処理ではないのでFluentdとK8sを組み合わせて使ってみることにした



RANCHER

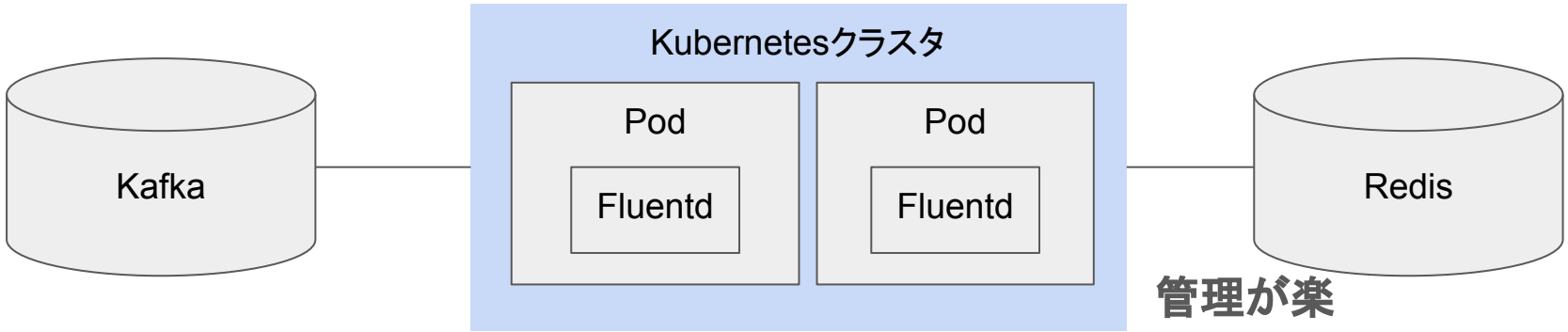
kubernetes

K8sでFluentdの弱点が解消できた

- Kafka consumer groupによって複数のFluentdのプロセスは一つのKafkaトピックから重複することなく一つのデータを受け取ることができる

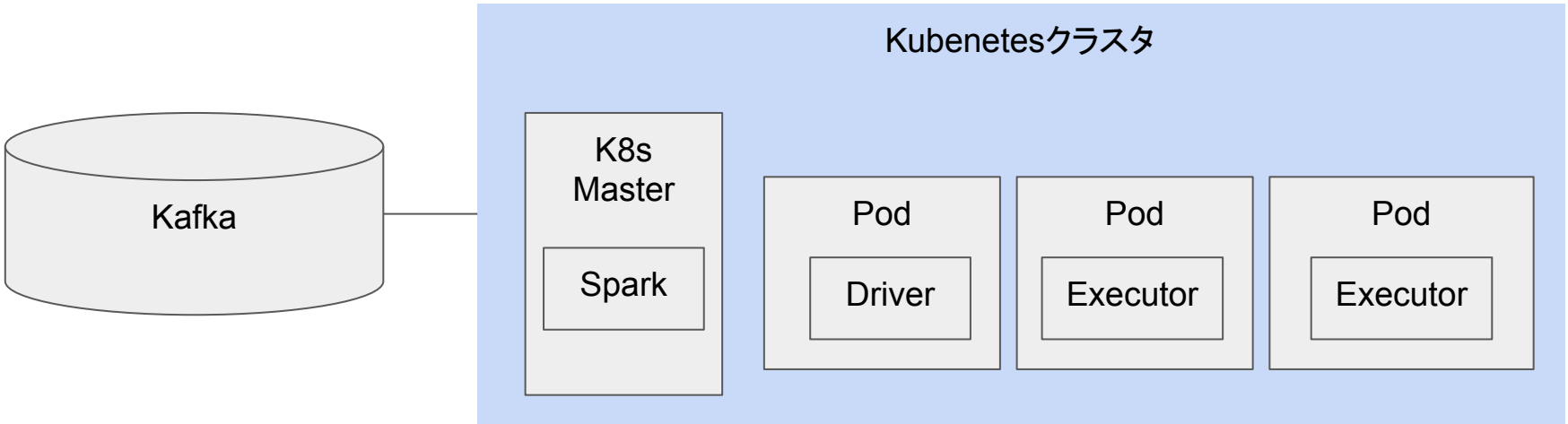
これにより、そこまで難しいこと考えることなくk8sにデプロイして

- k8sでFluentdクラスタを作ることができた
- マルチコアのCPUを有効活用をできた
- 複数サーバに展開してHA構成にすることができた



K8sでSpark Streamingの移行先の目処が付きそう

- Spark Streamingを動かしているHadoopの老朽化問題があったが
- K8sでのSpark検証が進んでいて移行することができそうな状況



まとめ

- KubernetesでFluentdの弱点を大きく改善することができた
- ノード管理やHA化なども難しいこと考えることなく実現できた

- KubernetesでSpark Streamingを活用することができた
- 弊社のリアルタイム処理周りはK8sでほとんどカバーすることができそう

- k8sの威力を思い知り勉強を再開しようと思った

MicroAd
Redesigning the Future Life

ご清聴ありがとうございました