広告配信システムにおけるデータ基盤移行の事例紹介

株式会社マイクロアド

高橋 唐樹

自己紹介 / 高橋 唐樹

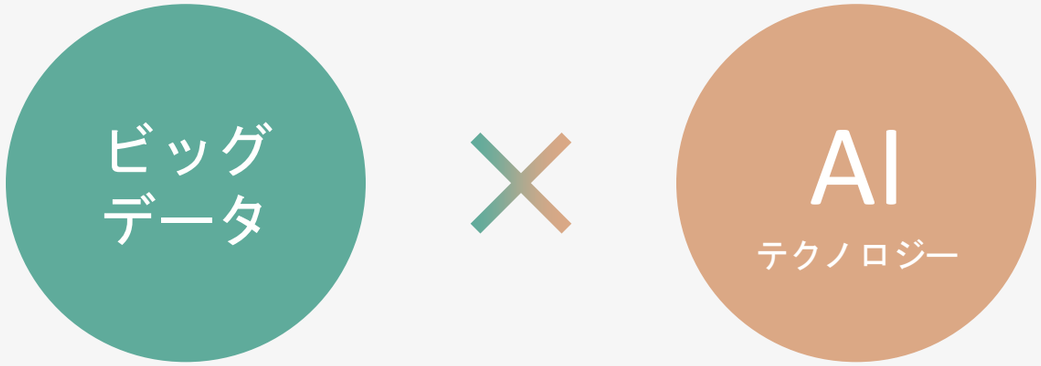
- 株式会社マイクロアド システム開発部
  マーケティングプロダクト開発ユニット
- 仕事: 大規模データ基盤関連開発
- 好きな言語: Python
- 経歴
  - 2022/3: 東京農工大学大学院 修士課程修了
  - 2022/4: 株式会社マイクロアドに新卒入社

目次

目次

マイクロアドについて / VISION

Redesigning the Future Life



データとテクノロジーの力で
“未来を予測する”

マイクロアドについて / 事業内容

- 自社製品を提供するデータプロダクト事業
  - データプラットフォーム UNIVERSE
  - 広告主向けプラットフォーム UNIVERSE Ads
  - 媒体社向けプラットフォーム MicroAd COMPASS
  - 業種特化マーケティングプロダクト
  - その他 広告サービス
- 他社製品を扱うコンサルティング事業
  - 海外コンサルティングサービス
  - メディア向けコンサルティングサービス

マイクロアドについて / 事業内容

- 自社製品を提供するデータプロダクト事業
  - データプラットフォーム UNIVERSE
  - 広告主向けプラットフォーム UNIVERSE Ads
  - 媒体社向けプラットフォーム MicroAd COMPASS
  - 業種特化マーケティングプロダクト
  - その他 広告サービス
- 他社製品を扱うコンサルティング事業
  - 海外コンサルテイングサービス
  - メデイア向けコンサルティングサービス

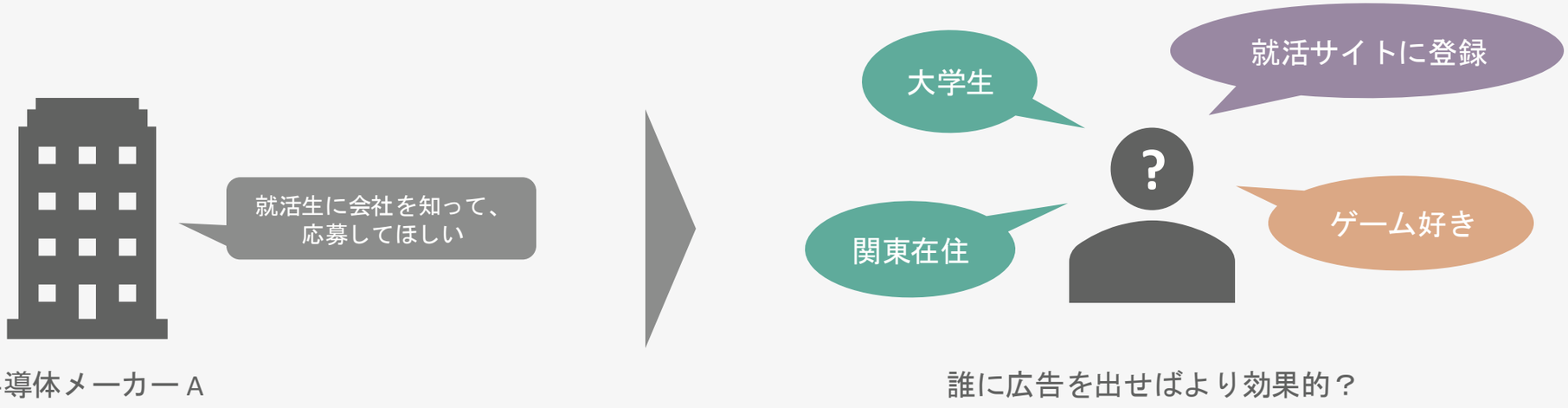
インターネット広告

インターネット広告 / プログラマティック広告

- インターネット広告
  - Webサイトやスマートフォンアプリなどを対象とした広告の形態
  - 配信する広告を動的に変更することが可能
- プログラマティック広告
  - 現在のインターネット広告の主流
  - 広告効果を最大化するため運用し続けていく広告
    - 収益の増加
    - 認知の拡大

インターネット広告 / ターゲティング

- ユーザー個々人の属性・趣味嗜好・行動などにあわせた広告を配信する手法
  - 広告効果を高めるのに有効
- データのリアルタイム性や分析しながらの運用が重要
  - 本当に届けたい人に広告が届いているのか？

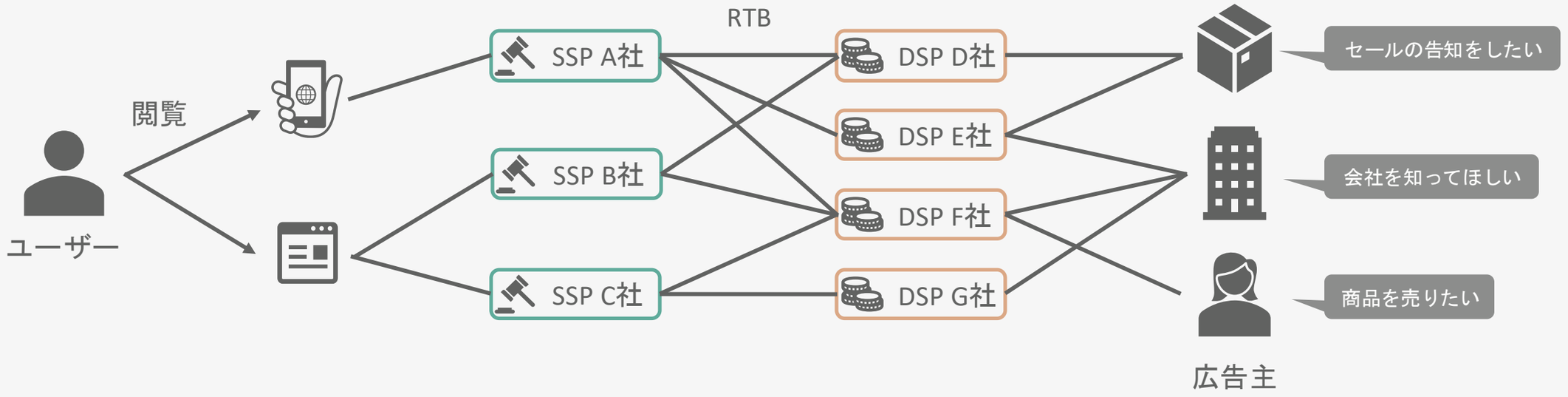


半導体メーカー A

誰に広告を出せばより効果的？

インターネット広告 / RTB (Real-Time Bidding)

- 表示する広告を決めるためのオークションの仕組み
- SSP (Supply-Side Platform)
  - メディアの利益を最大化するためオークションを開催するシステム
- DSP (Demand-Side Platform)
  - 広告主の利益を最大化するためオークションに参加するシステム



インターネット広告 / 大規模データ

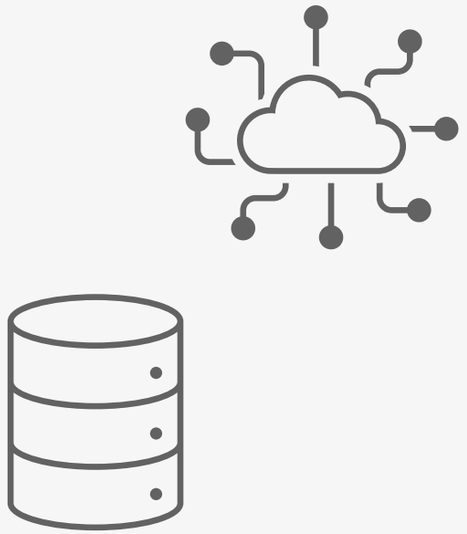
- ターゲティング, RTBの例だけでも様々なデータが関わってくる
  - ターゲティング用のユーザー情報: 170億件
    - 精度向上のため他社から連携されるデータが大量にある
  - RTB結果のログ: 60億件/日
    - 入札処理毎にログを記録しているため膨大な量になる
  - 広告配信結果分析用のログ
    - 多角的に分析するための集計処理が多数ある
- マイクロアドの場合、処理するデータ量は10TB/日～にも及ぶ
  - リアルタイム性が求められるデータも

分散データベース・分散ストレージ・分散処理システムなどが必要となる

目次

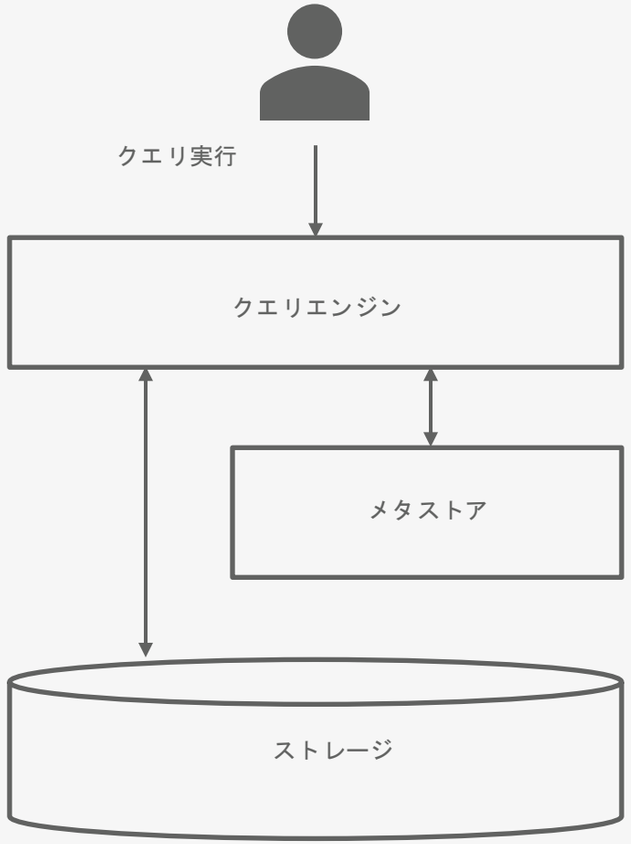
大規模データ基盤 / 前提

- データの大部分をオンプレミスのサーバーで保持している
  - データ量が非常に多いのでクラウドだとコストが掛かりすぎる
- ハードウェアの調達や費用の都合で移行が必要に
  - 2024/4〜 移行開始
  - 年内移行完了を目指して現在も進行中
- 大規模データのETL処理が250〜程度存在している
  - ETL: データの抽出・変換・蓄積



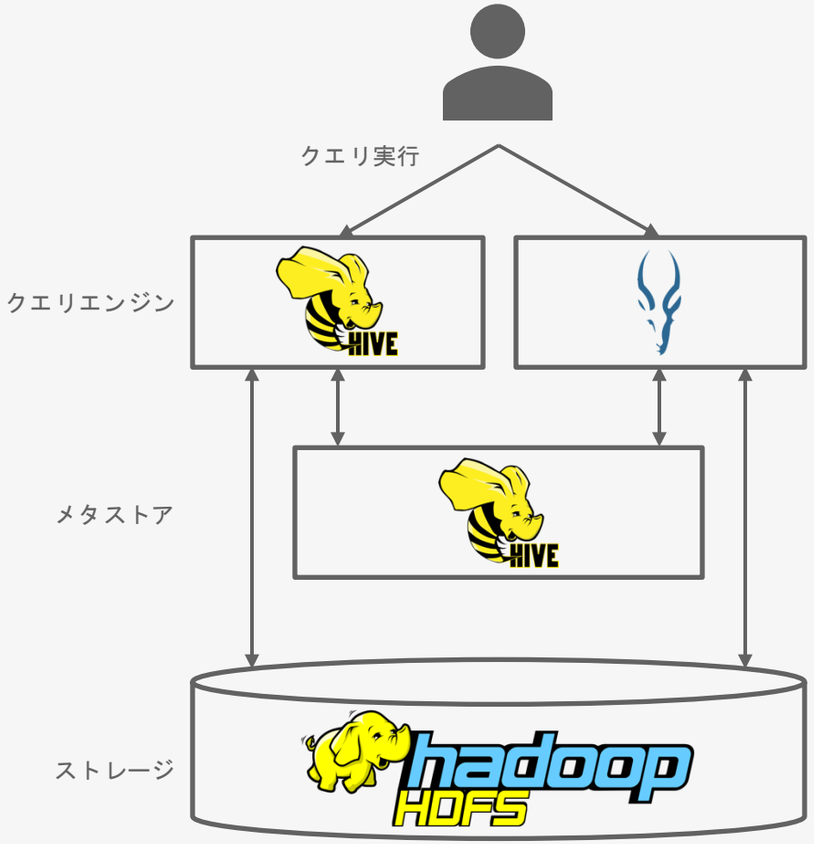
大規模データ基盤 / 要件

- クエリエンジン: クエリを実行するためのシステム
- メタストア: テーブル定義などのメタデータを管理するシステム
- ストレージ: テーブルの実データの保存場所
- 求めること
  - 大規模データ処理に適していること
  - スケーラビリティ
  - 耐障害性



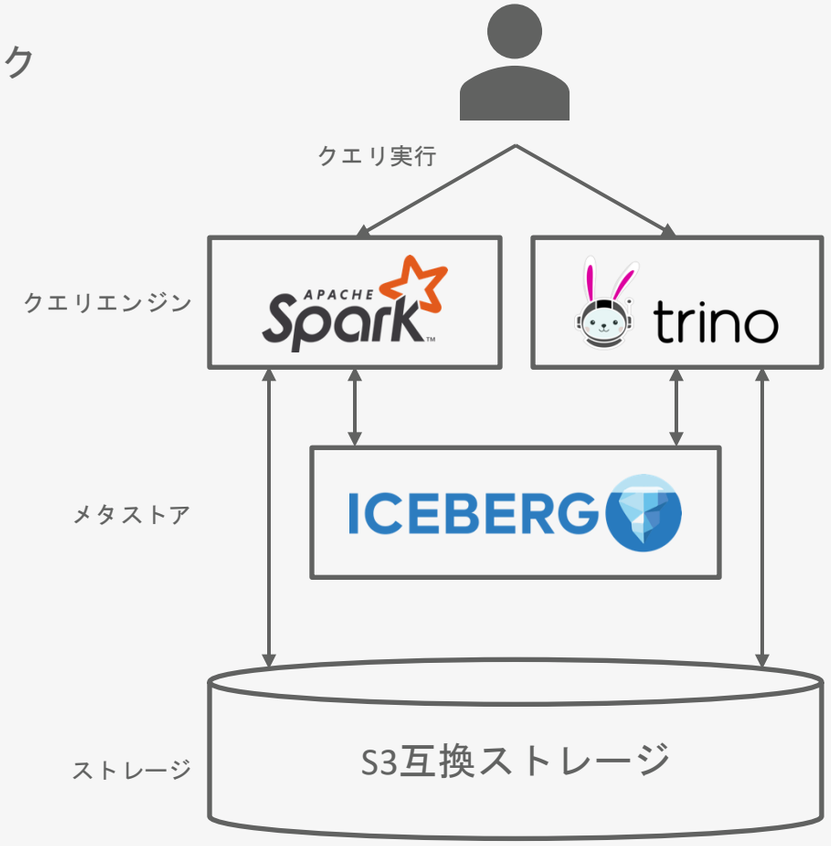
大規模データ基盤 / 移行前の構成

- クエリエンジン: Hive, Impala
  - Hadoop上のデータをSQL風に操作するためのシステム
  - Hive: MapReduceというHadoopの仕組みを利用
    - 耐障害性に優れているためETL処理で主に使用
  - Impala: 独自の分散処理方法を実装
    - 実行速度に優れているため分析で主に使用
- メタストア: Hive
- ストレージ: HDFS
  - 大規模データ処理向けの分散ファイルシステム
  - レプリケーション (データの複製) による耐障害性
  - Hadoop向け分散ファイルシステム



大規模データ基盤 / 移行後の構成

- クエリエンジン: Spark, Trino
  - Spark: SQLやデータフレームで操作可能な分散処理フレームワーク
    - 耐障害性に優れているためETL処理で主に使用
    - 特にSpark connectという技術を主軸として移行
  - Trino: データソースを横断して操作できる分散SQLエンジン
    - 実行速度に優れているため分析で主に使用
- メタストア: Iceberg
  - Open Table Formatと呼ばれる技術
  - 詳細は後述
- ストレージ: S3互換ストレージ
  - S3プロトコルでアクセス可能なオブジェクトストレージ
  - 消失訂正符号などによる耐障害性



大規模データ基盤 / 構成変更によるメリット

- ストレージとコンピュートを分離できた
  - どちらもHadoopに依存していた
    - 「コンピュートのみスケールさせる」みたいなことが柔軟にできない
  - ストレージ用・コンピュート用のサーバーが独立した状態に
- 今後Icebergから移行する際も改修が容易
  - Sparkが対応しているデータソースであれば切り替えられる
- SQLベースだけでなくデータフレームベースでの操作が可能
  - SQLだと表現しにくいビジネスロジックもあった
  - 可読性やテスト容易性を考慮して選択可能

移行の事例紹介 / Icebergの嬉しい機能

- 行レベルでの更新・削除
  - Hiveでは1行でも変更があると全データを書き直していた
  - 差分だけを書き込むことで更新・削除のI/Oを抑えられる
    - ターゲティング用のユーザー情報は頻繁に更新されるので便利
- Time travel / Rollback
  - 過去の特定の時点でのテーブルの状態を参照・復元できる
  - 再現性のあるクエリ実行ができる
    - 請求用の集計などで過去の時点と同じ結果を出せる
  - 障害発生時の切り戻しが容易
- 活かしきれていないが他にも色々
  - まだ移行しきってないので運用は手探りですが…

目次

移行の事例紹介 / 移行による利点

- 必要な機能なのか・適切な仕様なのかを検討できた
  - 誰が使っているのかわからない機能
  - 今のユースケースにあっていない集計テーブル
- 古いコードをリファクタできた
  - メンテナンス困難になっていた機能
  - 今の実装とは方針が違うコード
  - 古いバージョンのPythonで動いていた機能
- データ基盤の技術スタックが”今風”に
  - コミュニティが活発

移行の事例紹介 / 苦労していること: 技術面

- 運用に関する情報がまだまだ少ない
  - 大規模データで長期間運用したような記事は少ない
  - ベストプラクティスがわからない
- 「そもそもできるのか」から調査が始まる
  - 同じことをやろうとしている記事すら無い
  - 技術調査の結論を出すのが難しい
    - 先入観で試さずに結論を出すのも危険
- ライブラリ側にバグがあることも
  - Spark connectはまだ正式版じゃない
    - 利用者も多くはないので報告して修正してもらったり

移行の事例紹介 / 苦労していること: その他

- リプレイスなので優先度が下がりがち
  - 売上に直結する機能開発ではない
  - 重要性を説明して優先度を上げていくしかない
- 作業者のモチベーションの維持
  - 単調作業になる部分も多い
  - 新しいものを作る喜びが少ない
- 移行対象が多くて進捗が見えにくい
  - 細かくマイルストーンを設定する
  - 進捗を計測できる状態を整える

移行の事例紹介 / やってよかった・やったほうがよかったこと

- (どれも当たり前かもですが)
- 期限・ゴール・意思決定者を明確にする
  - 計画を立てるにもまずはここから
- 進捗を計測できる状態をキープする
  - 必要なのは最初の気合
- ドキュメントを残す
  - 今後苦労する誰かのために
- 優先度が下がらないようアクションし続ける
  - 先延ばしにしないという覚悟

まとめ

- 広告配信システムの大規模データ基盤を移行中
  - HiveからSpark x Icebergの基盤に移行
  - 既存の機能の見直しや実装の改善などの機会になる
  - 長期間のプロジェクトになるのでマネジメントが大変
  - ゴールに関しての合意を形成してから動き始めましょう
  - 社内向けのドキュメントはちゃんと残しましょう

We are hiring!

- 3/3, 3/4のオンサイト会場ではブース出展します！
  - スポンサーブース G14
- 技術ブログ、X (Twitter)で情報発信してます！
  - X (Twitter): @microad_dev

株式会社マイクロアド採用サイト

MicroAd Developers Blog

株式会社 マイクロアド
〒150-0031 東京都渋谷区桜丘町20-1
渋谷インフォスタワー 13F

www.microad.co.jp