GMOペパボ

サービスに寄り添うログ基盤

- ログ収集のその先に -

三宅悠介 / GMO PEPABO inc.

2016.07.02 はてな・ペパボ技術大会〜インフラ技術基盤〜@京都

GMOペパボ

三宅悠介　@monochromegane

プリンシパルエンジニア

minne事業部

http://blog.monochromegane.com/

minne

https://minne.com

目次

- Webサービスと行動ログ
- Bigfoot
- サービスに寄り添うログ基盤

Webサービスと行動ログ

ログはいいぞ

行動ログ

行動ログ

アプリケーション層で出力するログ

いつ、だれが、なにをやったかが特定できる

最終的な行動結果だけでなく、途中の**どこであきらめたか、どう迷ったか**がわかる

行動ログには
サービス改善のヒントがつまっている

行動ログの活用段階

行動ログの活用段階

収集: 行動ログが出力され、取りまとめられている状態

分析: 取りまとめた行動ログを視覚化、分析できる状態

活用: 分析した行動ログをもとに継続的なサービス改善が行えている状態

ログ基盤

ログ基盤に大切なこと

“ログの活用”

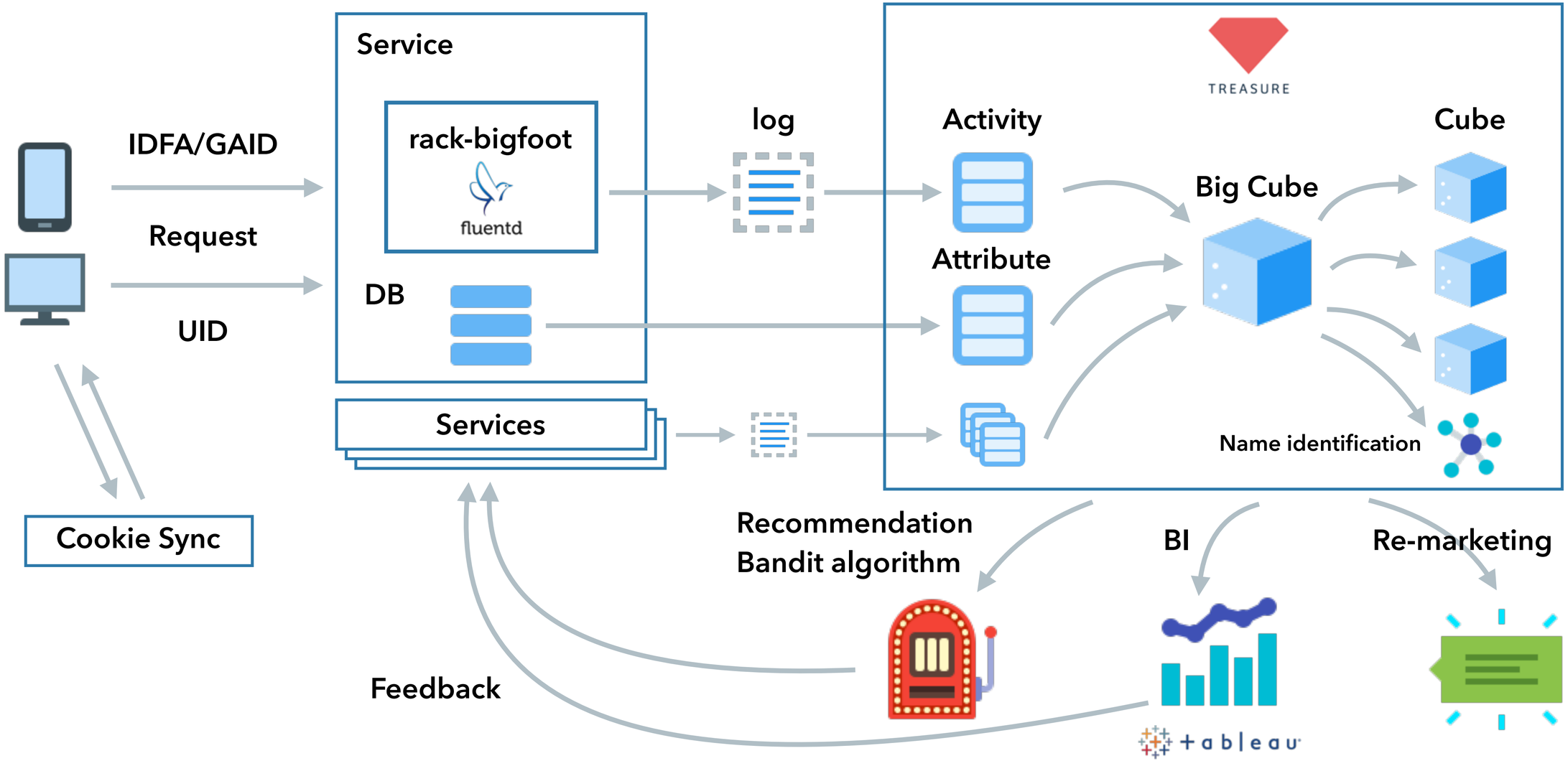
ログ “活用” 基盤

Bigfoot

Bigfoot 👣

- **ペパボの次世代ログ “活用” 基盤**
- 行動ログの収集、分析、活用の各段階において、全社で利用できる汎用性と具体的な活用方法を提供
- 国内最大級ハンドメイドマーケットminneを支えるログ基盤

Bigfoot 👣



https://icons8.com

Bigfootを支える技術

収集・分析

ログを送る

rack-bigfoot

- RailsアプリケーションとFluentdをつなぐRackミドルウェア
- Bigfootに必要な共通パラメタをリクエスト・レスポンスヘッダから取得
- サービス固有のパラメタを付与することも可能

```
Rails.application.config.app_middleware.insert_after ActionDispatch::Callbacks,
Rack::Bigfoot do |config|
  config.service        = 'minne'
  config.environment    = Rails.env
  config.enable_fluent = Rails.env.production? || Rails.env.staging?

  config.ignore_path_patterns << %r(\A/healthcheck)
  config.headers << 'HTTP_X_CLIENT_VERSION'
end
```

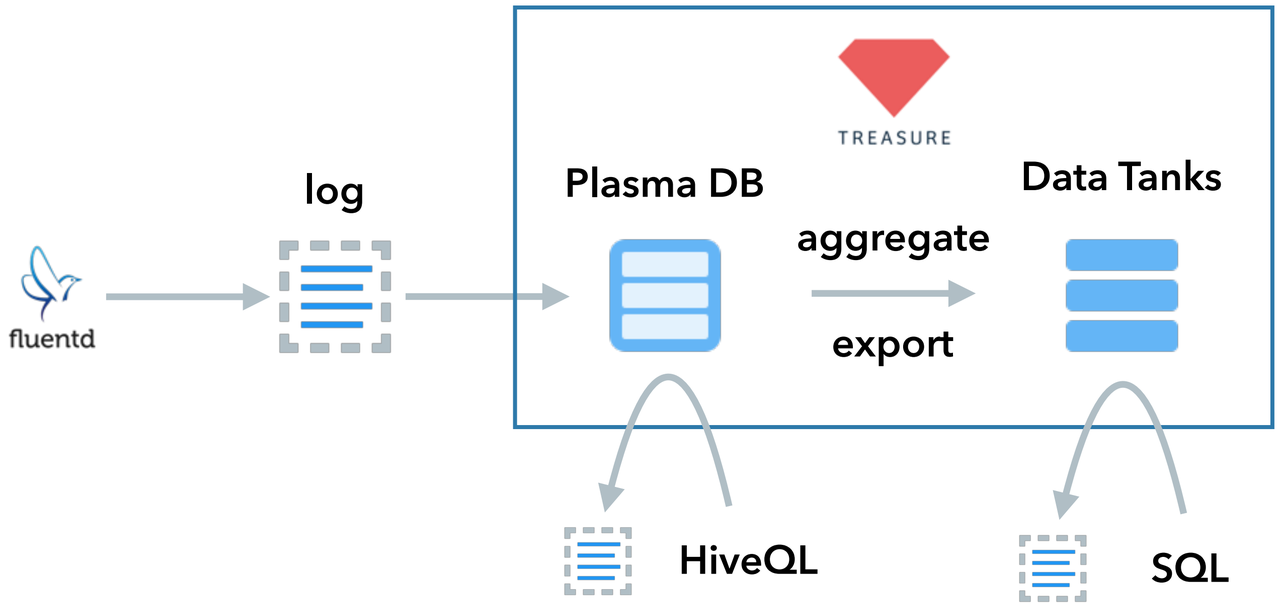
ログをためる

Treasure Data

・クラウド型データマネジメントサービス

・https://www.treasuredata.com/

・大容量のログ保存、分散処理による高速なログ操作



ログを扱う

Hive QL

- Treasure Data上の行動ログをSQLライクに扱う
  - http://hive.apache.org/
  - https://docs.treasuredata.com/articles/hive

```
SELECT
  TD_TIME_FORMAT(time, 'yyyy-MM-dd HH:mm:ss', 'JST') AS timestamp,
  response_time,
  request_method,
  path_info
FROM
  activity
WHERE
  TD_TIME_RANGE(time, '2016-07-01 10:00:00', '2016-07-01 12:00:00', 'JST');
```

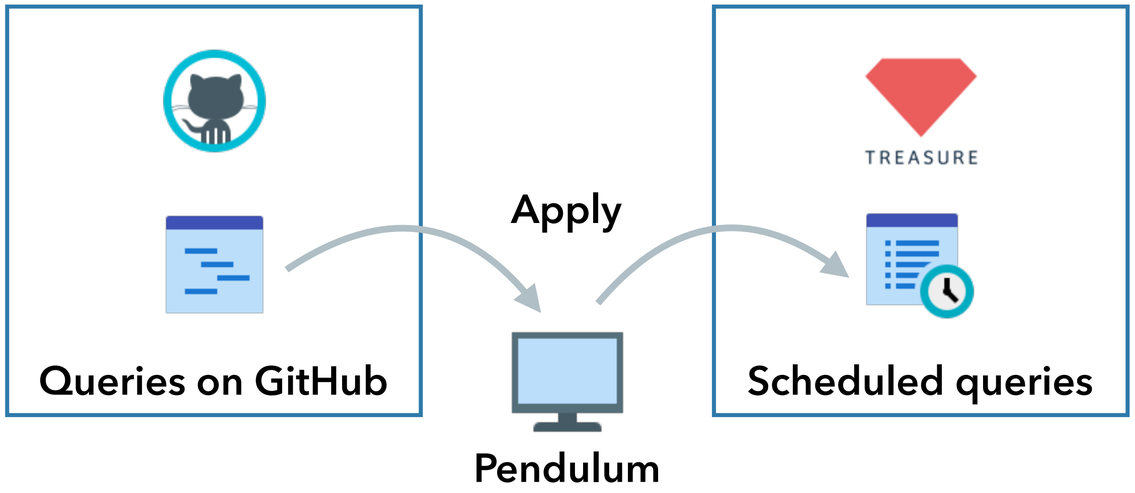
ワークフロー

・TreasureDataのスケジュールクエリを利用

・クエリのコード管理用に Pendulumを開発

・https://github.com/monochromegane/pendulum

・DSLによってスケジュールクエリを記述し、コード管理



Pendulum

Schedfile

```
schedule 'test-scheduled-job' do
  database     'db_name'
  query        'select time from access;'
  retry_limit 0
  priority     :normal
  cron         '30 0 * * *'
  timezone     'Asia/Tokyo'
  delay        0
  result_url   'td://@/db_name/table_name'
end
```

Apply

```
$ pendulum --apikey='...' -a --dry-run
$ pendulum --apikey='...' -a
```

Digdag移行中

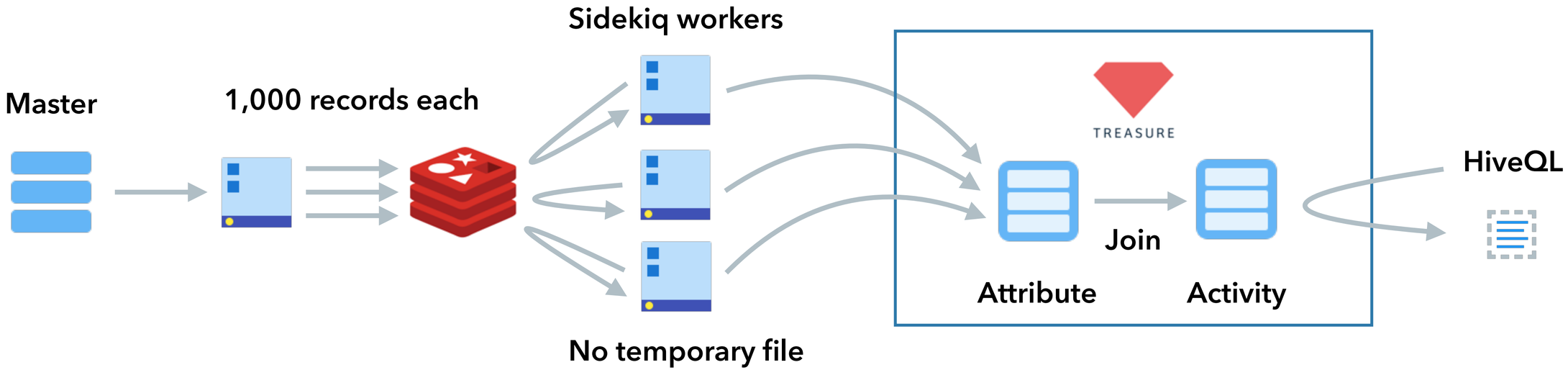
https://github.com/treasure-data/digdag

ログを便利にする

属性情報

・行動ログと属性情報を組み合わせることで分析時の幅が広がる

```
def perform(*args)
  User.order(:id).select(:id).find_in_batches do |users|
    UserAttributesUploadJob.perform_later(users.first.id, users.last.id)
  end
end
```

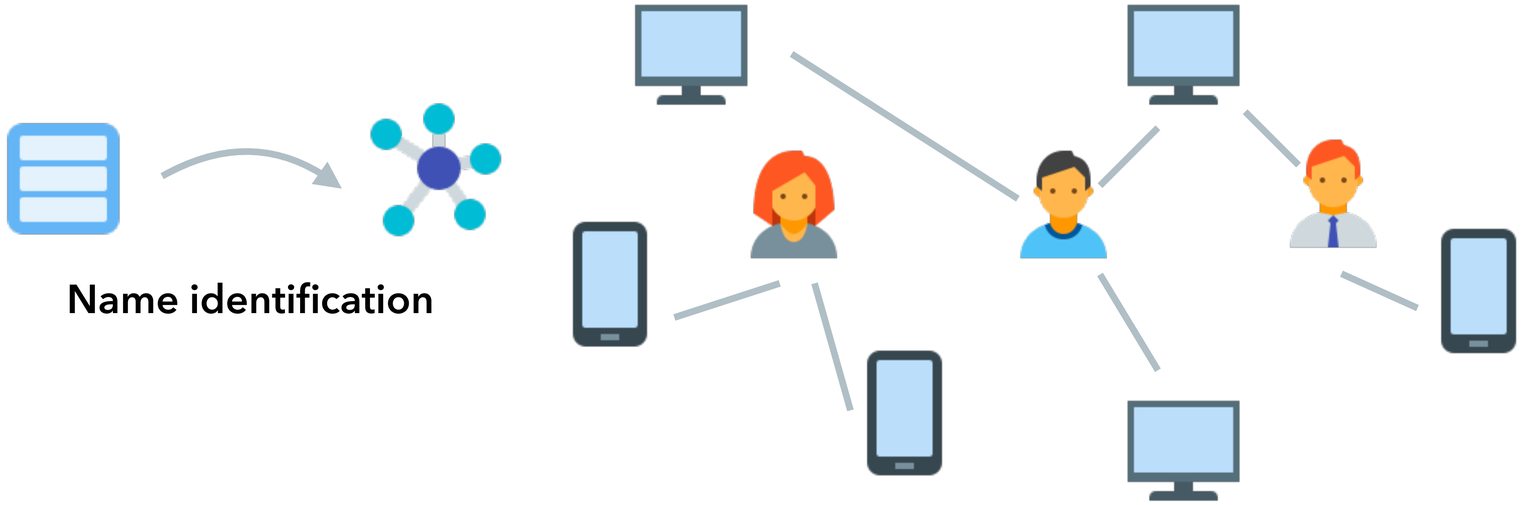


名寄せ

・サービスのアカウントと各クライアントをマッピング

・未ログイン状態のアカウントも名寄せ後に過去に遡って紐付け

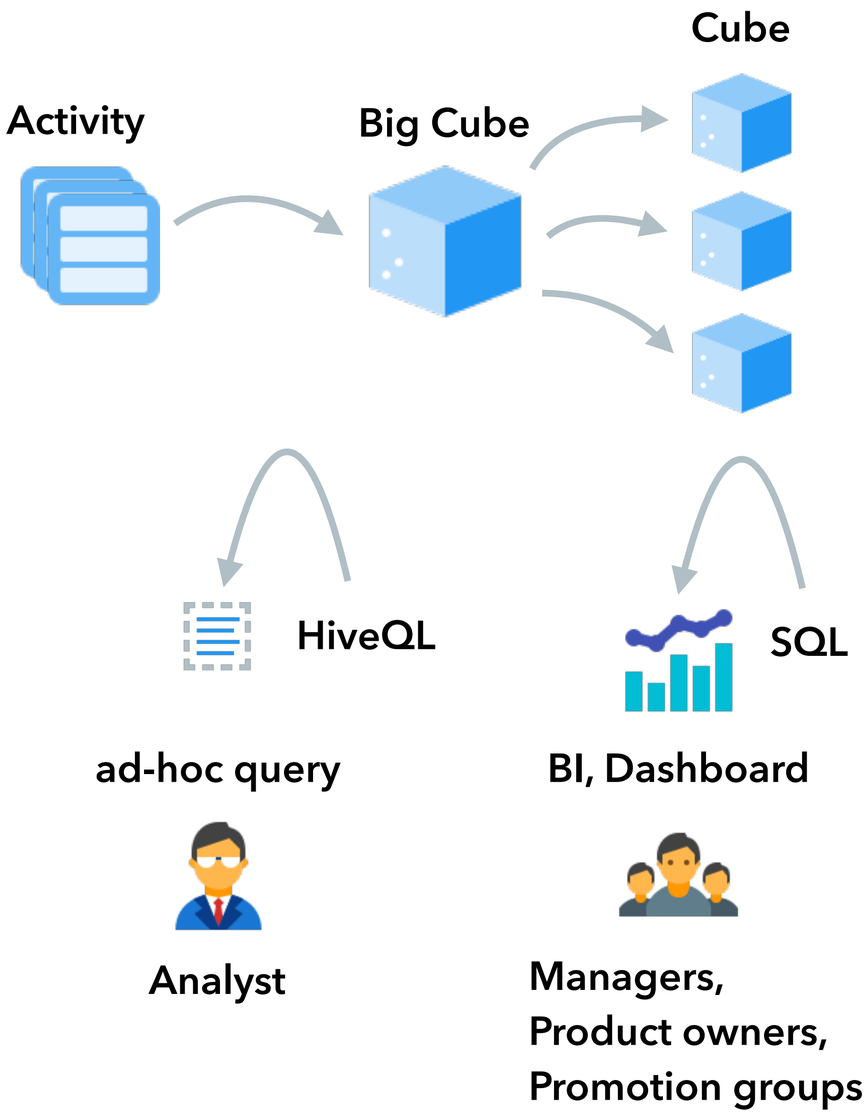
・Cookie Syncと組み合わせてサービスをまたいだマッピングも可能



ログを分析する

Big CubeとCube

- 全サービスの行動ログを集約したBig Cube
- 切口が確定したものは1メジャーカラム、1ディメンションカラムの単位でCubeに切り出し
  - メジャー: 定量化可能なカラム
  - ディメンション: 集計の切り口となるカラム
  - 例: 時間ごとの売上、都道府県ごとの作品数
- Cubeはデータマートに置き、高速に参照できるようにする



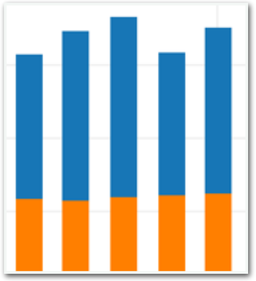
視覚化と分析

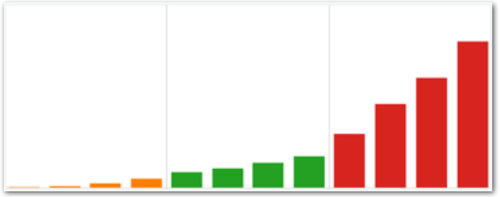
・視覚化と分析には Tableau社の Tableau Desktopを利用

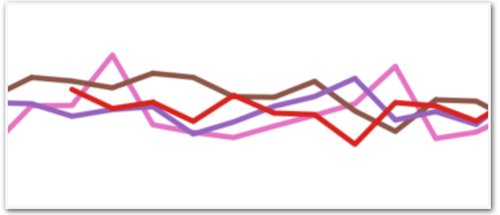
・http://www.tableau.com/

・データソースとしてTreasure Dataを選択可

- **ダッシュボード例**
- 流通額、キャンセル金額、注文金額、ユーザー単価
- 累積会員数、注文単価、新規登録ユーザー、DAUC 新規、DAUC 既存
- 受注作品数、受注率、受注作品価格、受注可能作品数
- 総在庫数、在庫単価、在庫総額
- 受注可能作家数、販売中作品数、開店中作品数、総作品数







活用

活用

・分析した結果をもとに仮説を立ててシステムの改修を行う

・画面デザインの変更、ステップの見直し

・A/Bテスト

→ **静的なフィードバック**

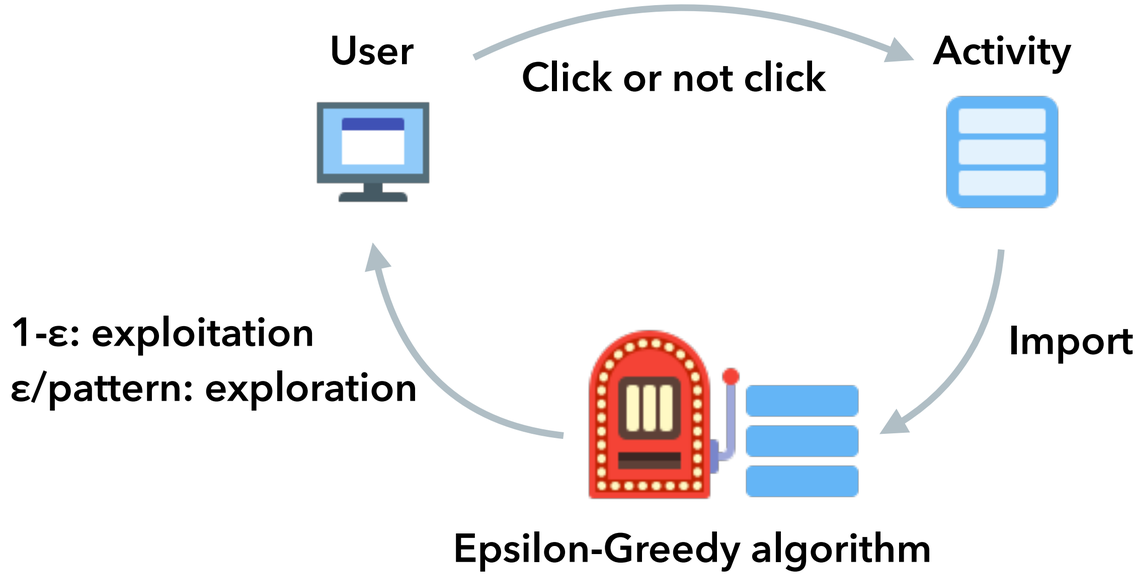
動的なフィードバック

バンディットアルゴリズム

・探求と活用の割合を更新し続けることでA/Bテストの機会損失を減らす

・https://www.oreilly.co.jp/books/9784873116273/

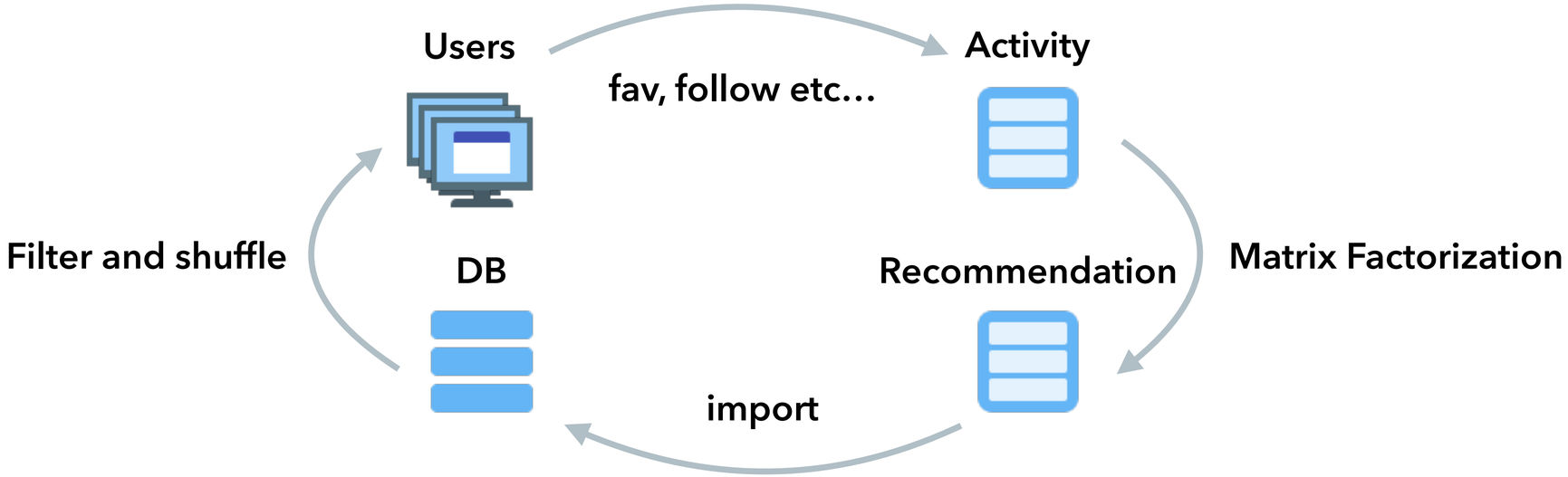
・例えば、ある機能のCTRを改善するために8割は最善の手法（活用）、残り2割で複数の手法を試す（探求）



レコメンド

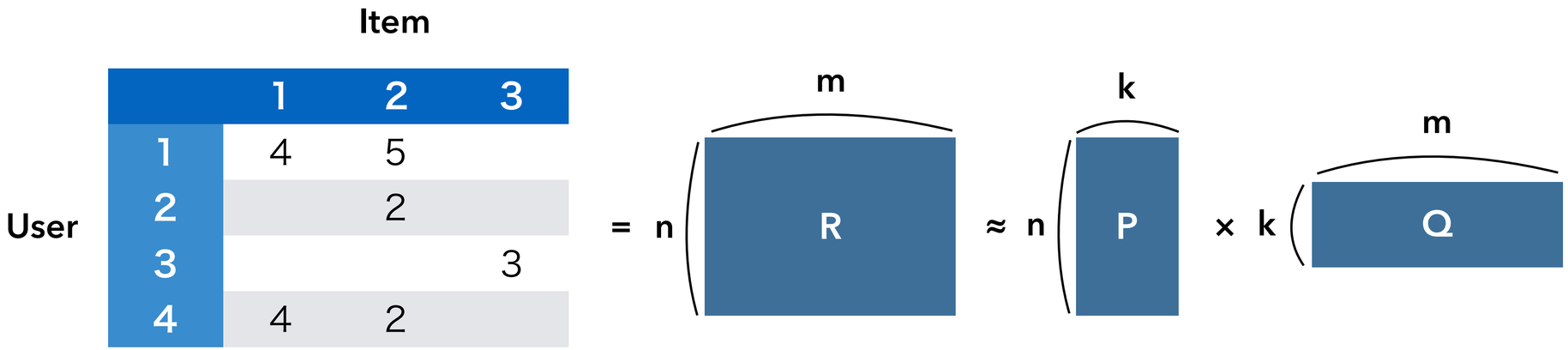
・minne「あなたにおすすめの作家」

・ユーザーの行動を基に作家をレーティング



レコメンド - Matrix Factorization

- 協調フィルタリング - ユーザの嗜好情報を蓄積し、あるユーザと嗜好の類似した他のユーザの情報を用いて推論を行う
- **Matrix Factorization**
  - 次元削減
  - ユーザーや作品ごとの評価の偏りがあり、疎なデータに対する評価予測



レコメンド - Matrix Factorization

予測

$$R'_{ui} = \mu + B_u + B_i + \vec{P}_u^T \vec{Q}_i$$

平均　バイアス

学習

$$\min_{P,Q,B} \sum_{(u,i) \in R} (R_{ui} - R'_{ui})^2 + \lambda(\|B_u\|^2 + \|B_i\|^2 + \|\vec{P}_u\|^2 + \|\vec{Q}_i\|^2)$$

誤差　　正則化項

Hivemall

レコメンド - Matrix Factorization

```
SELECT
  idx,
  array_avg(u_rank) as Pu,
  array_avg(i_rank) as Qi,
  avg(u_bias) as Bu,
  avg(i_bias) as Bi,
  min(mu) as mu
FROM (
  SELECT train_mf_sgd(account_id, creator_id, rating,
'-factor 20 -iter 50 -update_mu') AS (idx, u_rank,
i_rank, u_bias, i_bias, mu)
  FROM training
) t
GROUP BY idx;
```

などなど

browse, cart abandonment 

・いわゆる、閲覧放棄、カート放棄の作品を行動ログから抽出

・特定の条件で呼び戻しの通知を行う

広告連携 

・行動ログから関連性の高い広告を出す

・リマーケティング

・広告対象のセグメント化（絞込、除外）

サービスに寄り添うログ基盤

サービスに寄り添うログ基盤

・単にログを集めるだけにせず、分析、活用の段階を補助する

・静的なフィードバックから動的なフィードバックへ

・行動ログの循環により、**なめらかな世界へ**

ログはいいぞ

おわり

君もペパボで働かないか？

最新の採用情報をチェック→ @pb_recruit