2013.10.21
第13回 関西DB勉強会

データビジネスの最前線
データクリーンルームとは？
Snowflakeで実現する新たなデータ流通の仕組みを解説

エクスチュア株式会社 喜田紘介

自己紹介

喜田紘介 @kkkida_twtr

- エクスチュア株式会社
- 日本PostgreSQLユーザ会 理事長
- Snowflake公式サブコミュニティ
  Team Data Clean Room発起人

最近の仕事

約半年間、Snowflakeのデータクリーンルーム構築案件、社内データ基盤の設計、社内の情報整備やメンバー育成に従事。

データベースが好き、ITが好き、コミュニティが好きを活かしてコミュニティ盛り上げの一助になれれば幸いです。

Team Data Clean Room

データクリーンルームの構築ノウハウ、ユースケースを蓄積・発信

特に日本ではデータを**買いたい人・売りたい人**どちらも手探りのなか、先行企業の事例やアイデアが集まる数少ない場です。
これからの**データ流通の新しい形！**
売り手も買い手も一緒に盛り上げていきましょう！

▲DCRってなに？勉強会の動画 公開中

本セッションでは

データの流通？
データクリーンルームって？

Snowflake DCR

アーキテクチャから見る
注目ポイント

結局なにがおきるの？
利用イメージをもっと膨らませて
「相手探し」の第一歩を踏み出そう！

SoR から SoE、SoIの時代へ

蓄積されたデータを分析し新たな洞察を得るためのシステム

- ✓ ML/AI
- ✓ パブリックデータ

顧客とのエンゲージ（約束・結びつき）を主眼に置いたCXのためのシステム

- ✓ コールセンター
- ✓ リコメンド
- ✓ CDP

事実を正確に記録し、間違わない
失わない、流出しないためのシステム

- ✓ 会計
- ✓ 契約管理
- ✓ 在庫管理

SoE、SoIの時代 - どのようなデータに価値があるか -

SoI
System of Insights

SoE
System of Engagement

SoR
System of Records

事業全体の改革につながる新たな知見を得たい

- ✓ 年単位の鮮度
- ✓ 超大規模データから特徴抽出（ML/AI）
- ✓ パブリックデータとの突合

その道の専門機関が収集したデータ

過去10年の売上データ × 同期間の天候データ

顧客の指向を理解し、要望にフィットする対応したい

- ✓ 個人に紐づく履歴
- ✓ 界隈のトレンドも加味
- ✓ 日単位の鮮度

多くの属性をとらえる

mail_address	実績	A	B	C	D	E
aaa@mail	**成約**	✓		✓	✓	✓
bbb@mail	トライアル	✓	✓			
ccc@mail	失注	✓				✓
ddd@mail	**成約**	✓		✓	✓	✓

データを買う：個人に紐づく属性を充実させてデータドリブンな施策へ

購入歴のある人に
DMを送ろう

経済系の記事に
広告だそうよ

旅行代理店とコラボ
キャンペーンやろうよ

会員ID	購入歴あり			経済	IT	エンタメ	スポーツ	旅行
aaa@mail	✓			✓	✓		✓	✓
bbb@mail	✓			✓				
ccc@mail	✓			✓				✓
eee@mail	✓			✓	✓		✓	✓
fff@mail	✓							
hhh@mail	✓			✓	✓	✓	✓	✓

会員IDに紐づく属性が充実している状態

何らかの集計・分析を経て施策へ

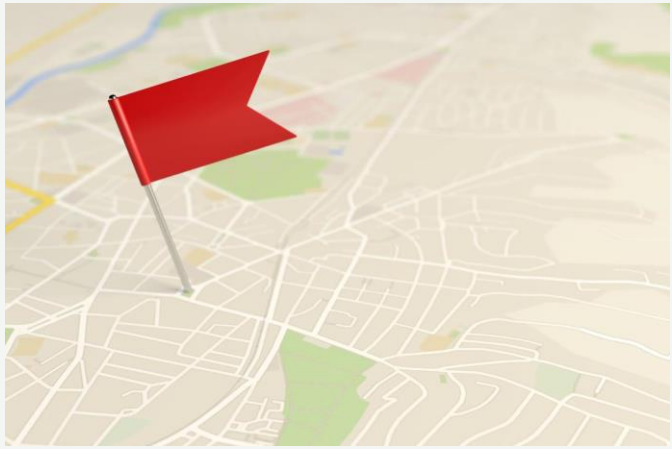
分析：会員全体のうち該当者が多い属性は？

キーワード	会員数	該当者	割合
経済	20000	16000	**80%**
IT	20000	4000	20%
エンタメ	20000	100	2%
スポーツ	20000	2000	10%
旅行	20000	12000	**60%**

データを売る：魅力あるデータは魅力あるコンテンツや仕組みに集まる

メディア

興味・関心にフィットする
豊富なコンテンツ



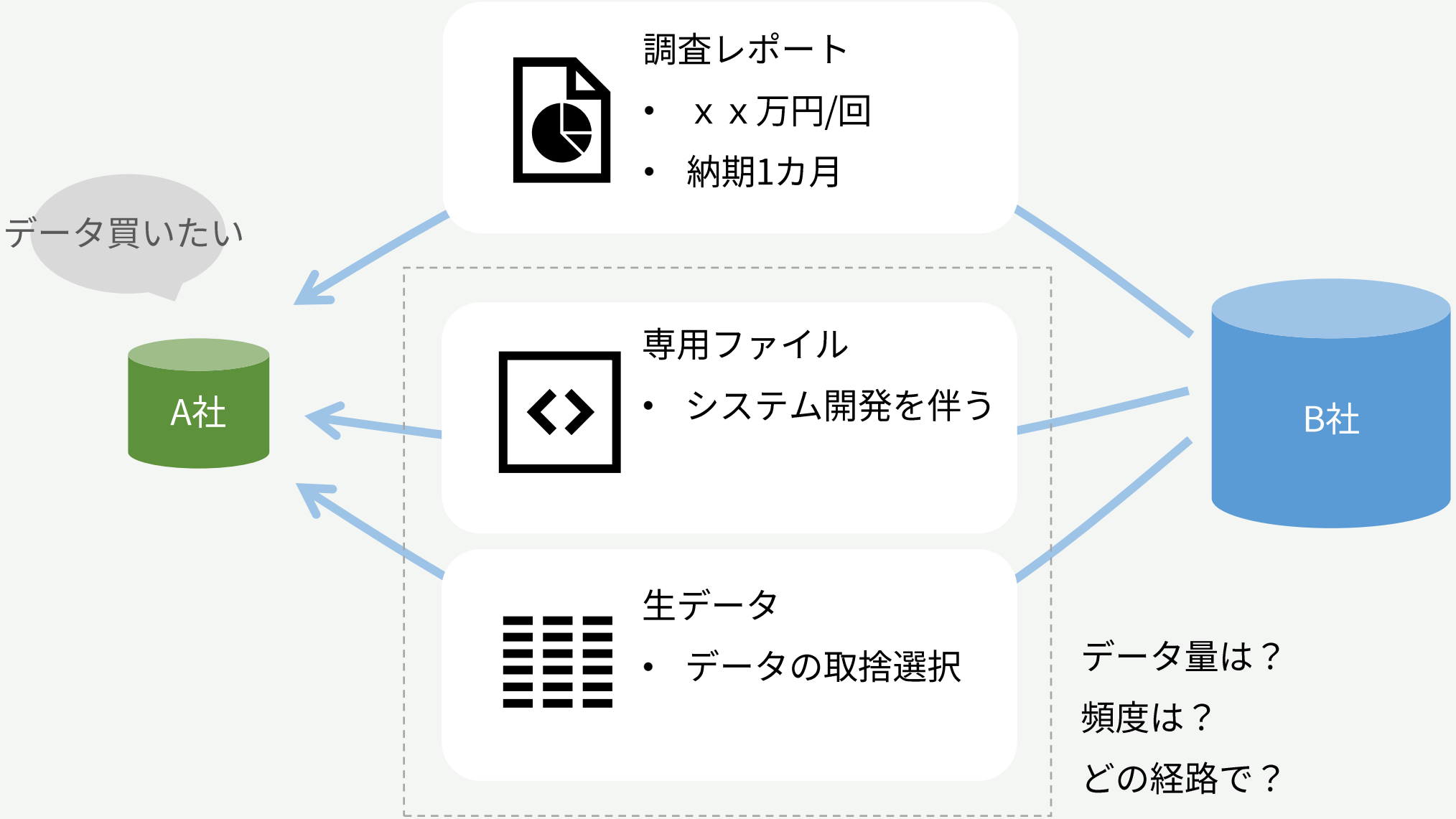
IoT、センサー、位置情報

現在地が取得できている
ことでの利便性

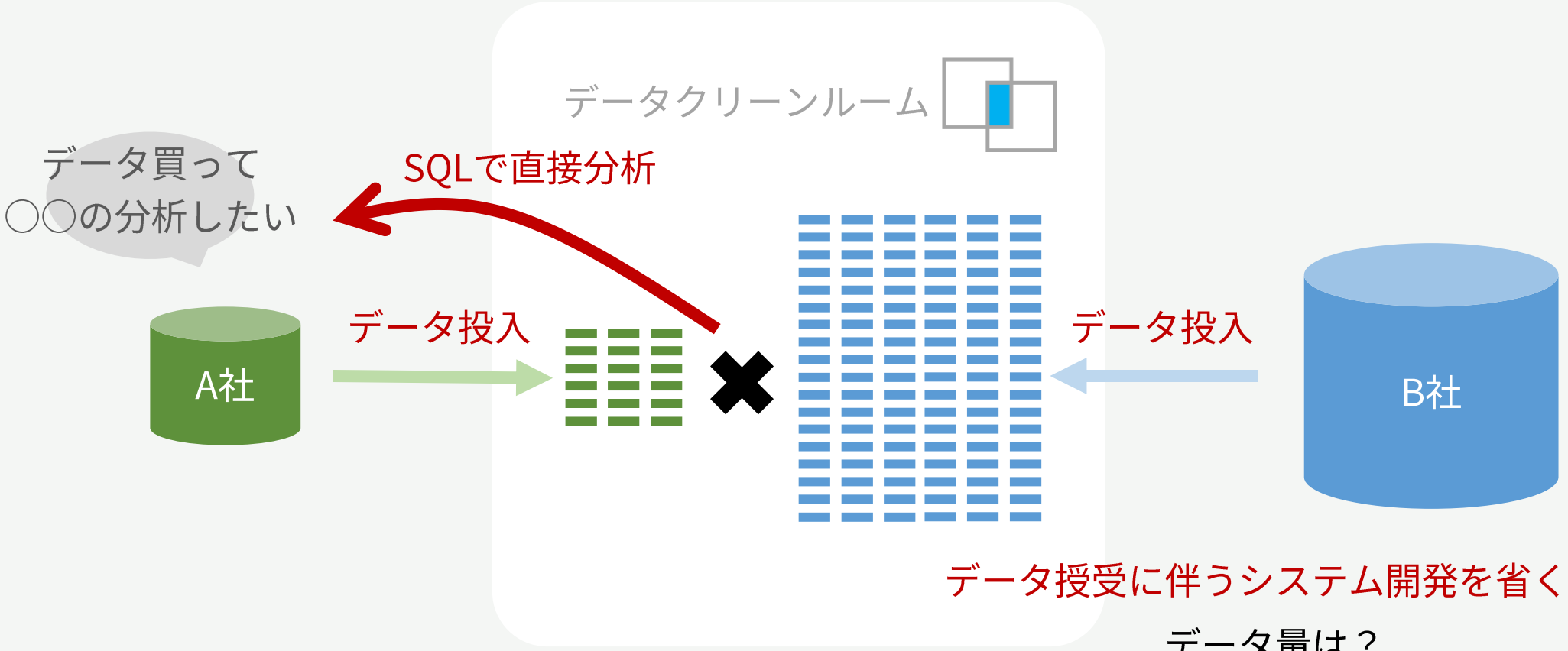
決済

お得感や利便性
店舗そのものの価値

データを授受するために



データクリーンルーム：データを授受せずに「分析」を達成



データ授受に伴うシステム開発を省く

データ量は？

頻度は？

どの経路で？

データを買う：個人に紐づく属性を充実させてデータドリブンな施策へ

購入歴のある人に
DMを送ろう

経済系の記事に
広告だそうよ

旅行代理店とコラボ
キャンペーンやろうよ

データ買って
○○の分析したい

会員ID	購入歴あり			経済	IT	エンタメ	スポーツ	旅行
aaa@mail	✓			✓	✓		✓	✓
bbb@mail	✓			✓				
ccc@mail	✓			✓				✓
eee@mail	✓			✓	✓		✓	✓
fff@mail	✓							
hhh@mail	✓			✓	✓	✓	✓	✓

会員IDに紐づく属性が充実している状態

何らかの集計・分析を経て施策へ

分析：会員全体のうち該当者が多い属性は？

キーワード	会員数	該当者	割合
経済	20000	16000	**80%**
IT	20000	4000	20%
エンタメ	20000	100	2%
スポーツ	20000	2000	10%
旅行	20000	12000	**60%**

データクリーンルームで実行するクエリのイメージ（NGな例）

会員に紐づく属性が直接見えてしまう！

会員ID				経済	IT	エン	スポ	旅行
aaa@mail	✓			✓	✓		✓	✓
bbb@mail	✓			✓				
ccc@mail	✓							

```
SELECT * FROM
   list_b
JOIN
   list_a
ON 会員ID＝mail_address
```

A社

会員ID	購入歴あり		
aaa@mail	✓		
bbb@mail	✓		
yyy@mail	✓		
hhh@mail	✓		

B社

mail_address	経済	コ	エンタメ	スポーツ	旅行
aaa@mail	✓	✓		✓	✓
bbb@mail					
ccc@mail		✓			✓
ddd@mail	✓	✓			✓
eee@mail	✓	✓			
jjj@mail				✓	
kkk@mail			✓	✓	

データ流通といってもプライバシーへの配慮は最重視しなければならない（法律、仕組み、ユーザ感情など）

GROUP BYしやすい形に整える

mail_address	経済	IT	エンタメ	スポーツ	旅行
aaa@mail	✓	✓		✓	✓
bbb@mail	✓				✓
ccc@mail		✓			✓
ddd@mail	✓	✓			✓
eee@mail	✓	✓			
fff@mail					
ggg@mail		✓			
hhh@mail	✓	✓	✓	✓	✓
iii@mail		✓	✓	✓	✓
jjj@mail				✓	
kkk@mail			✓	✓	

UNPIVOT

mail_address	keyword	flag
aaa@mail	経済	1
aaa@mail	IT	1
aaa@mail	エンタメ	
aaa@mail	スポーツ	1
aaa@mail	旅行	1
bbb@mail	経済	1
bbb@mail	IT	1
bbb@mail	エンタメ	
bbb@mail	スポーツ	
bbb@mail	旅行	1
ccc@mail	経済	1
:		

```
SELECT
   keyword,count_if(flag)
FROM
   list_b
GROUP BY
   keyword
```

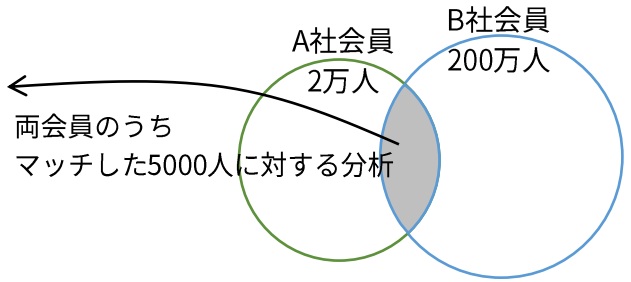
keyword	Countif
経済	200000
IT	100000
エンタメ	300000
旅行	150000

B社会員
200万人

データクリーンルームで実行しやすいSQL

A社のリストを考慮したキーワード別レポートが取得できた

キーワード	マッチした数	該当者	割合
経済	5000	4000	**80%**
IT	5000	1000	20%
旅行	5000	3000	**60%**



マッチ数や割合は別のクエリで算出

A社

会員ID	購入歴あり		
aaa@mail	✓		
bbb@mail	✓		
ccc@mail	✓		
eee@mail	✓		
fff@mail	✓		
xxx@mail	✓		
yyy@mail	✓		
hhh@mail	✓		

```
SELECT
   keyword,count_if(flag)
FROM
   list_b
JOIN
   list_a
ON
   会員ID＝mail_address
GROUP BY keyword
```

B社

mail_address	keyword	flag
aaa@mail	経済	1
aaa@mail	IT	1
aaa@mail	エンタメ	
aaa@mail	スポーツ	1
aaa@mail	旅行	1
bbb@mail	経済	1
bbb@mail	IT	1
bbb@mail	エンタメ	
bbb@mail	スポーツ	
bbb@mail	旅行	1
ccc@mail	経済	1
:		

keyword	Countif
経済	200000
IT	100000
エンタメ	300000
旅行	150000

```
SELECT
   keyword
   count_if(flag)
FROM
   list_b
GROUP BY
   keyword
```

本セッションでは

データの流通？
データクリーンルームって？

Snowflake DCR

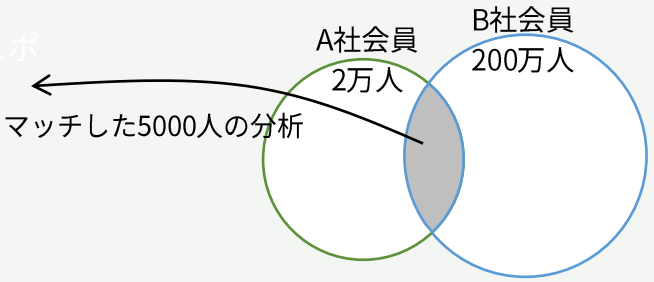
アーキテクチャから見る
注目ポイント

結局なにがおきるの？
利用イメージをもっと膨らませて
「相手探し」の第一歩を踏み出そう！

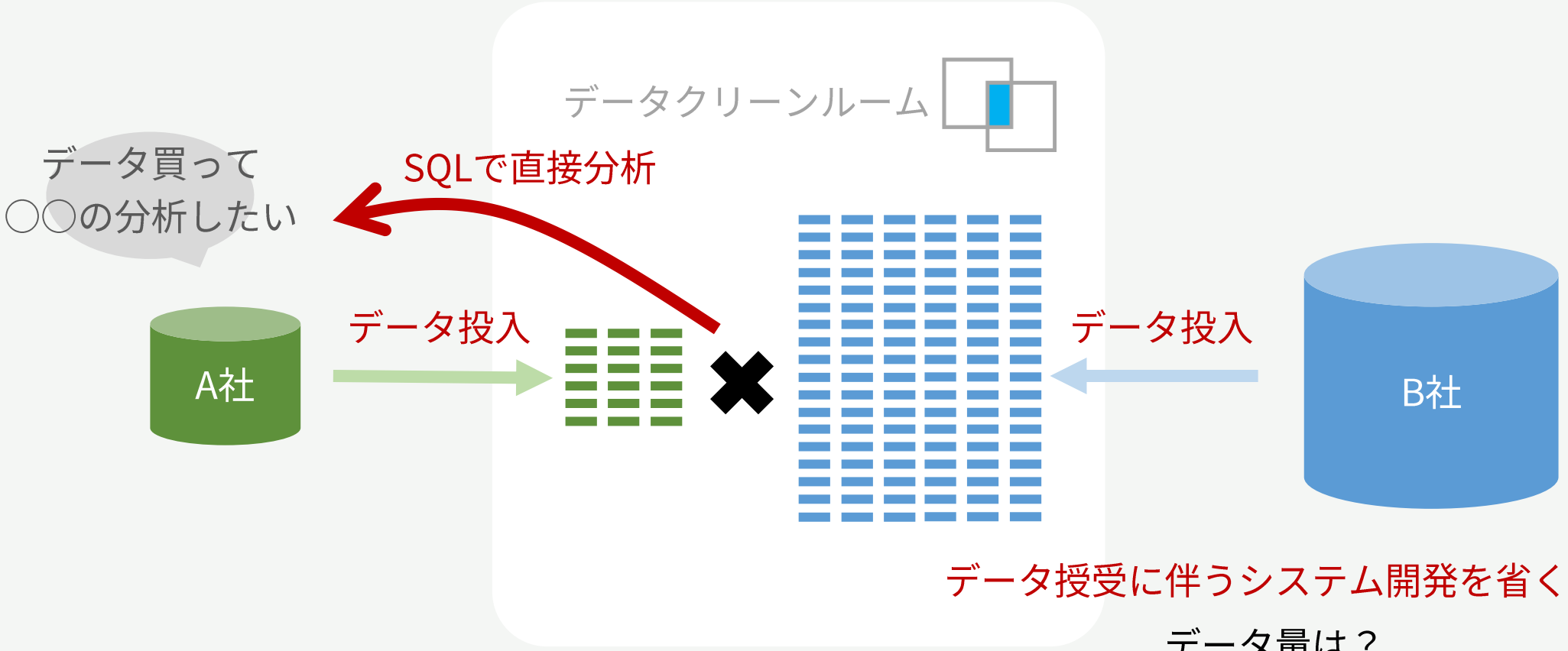
データクリーンルーム = 分析のためのデータ流通の1つの形

・データ授受のためのシステム開発を省力化
・プライバシーへの対応

キーワード	マッチした数	該当者	割合
経済	5000	4000	**80%**
IT	5000	1000	20%
旅行	5000	3000	**60%**



データクリーンルーム：データを授受せずに「分析」を達成



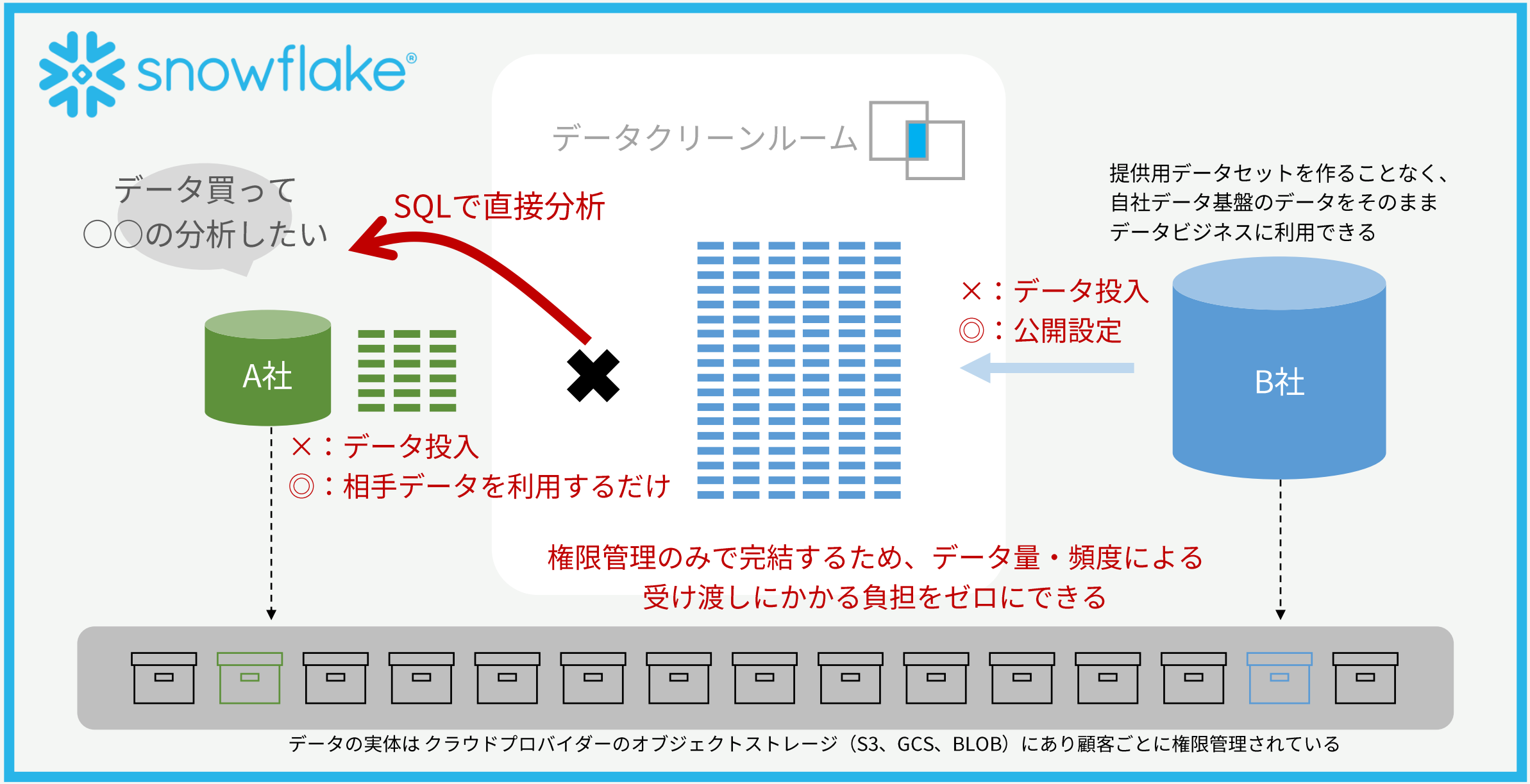
データ授受に伴うシステム開発を省く

データ量は？

頻度は？

どの経路で？

Snowflake：超巨大なマルチテナントのDWHサービス



Snowflake DCRでプライバシーを重視したSQLを強制

rap_mapping_table

company	account	allowed_query
A	AA12345	クエリから生成したハッシュ値
B	BB23456	
C	CC34567	

```
SELECT sha2($$SELECT 商品名,count(*)
FROM lineitem AS remote
JOIN my_schema.category_list AS local
ON remote.商品名 = local.商品名
WHERE local.valid_status = true
GROUP BY 商品名;$$);
```

JOINやGROUP BYを含むSQLを事前登録

行アクセスポリシーで**許可済みクエリ**のみ実行を許可

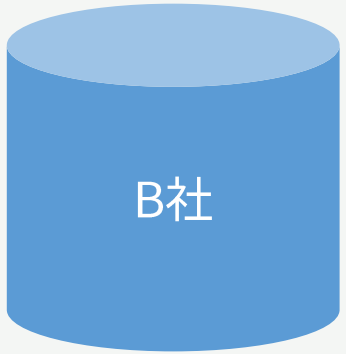
```
CREATE OR REPLACE ROW ACCESS POLICY security.dcr_rap_by_supplier
AS (filter_value integer) RETURNS BOOLEAN ->
  'dummy' = current_account()
      OR
  EXISTS (SELECT 1 FROM security.rap_mapping_table
          WHERE account = current_account()
          AND allowed_query = sha2(current_statement())
          AND company = filter_value
          )
;
```

filter_value = A

RAPをアタッチした列「会社名」に対して
filter_valueとの一致をチェック

許可済みSQLで直接分析

公開設定



Snowflakeでのデータクリーンルーム構築

本セッションでは

データの流通？
データクリーンルームって？

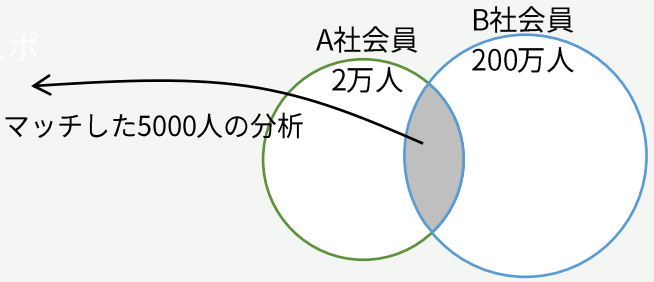
Snowflake DCR

アーキテクチャから見る
注目ポイント

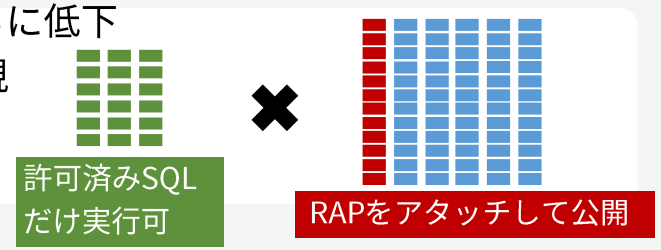
結局なにがおきるの？
利用イメージをもっと膨らませて
「相手探し」の第一歩を踏み出そう！

データクリーンルーム＝分析のためのデータ流通の1つの形
・データ授受のためのシステム開発を省力化
・プライバシーへの対応

キーワード	マッチした数	該当者	割合
経済	5000	4000	**80%**
IT	5000	1000	20%
旅行	5000	3000	**60%**



Snowflakeのデータクリーンルーム
・マルチテナントゆえ公開までのハードルがさらに低下
・JOINやGROUP BYを強制、プライバシーを重視



データを買う：自社の顧客リストだけでデータドリブンを始められる

経済系の記事に
広告だそうよ

旅行代理店とコラボ
キャンペーンやろうよ

会員ID	購入歴あり			経済	IT	エンタメ	スポーツ	旅行
aaa@mail	✓			✓	✓		✓	✓
bbb@mail	✓			✓				
ccc@mail	✓			✓				✓
eee@mail	✓			✓	✓		✓	✓
fff@mail	✓							
hhh@mail	✓			✓	✓	✓	✓	✓

会員IDに紐づく属性が充実している状態

何らかの集計・分析を経て施策へ

分析：会員全体のうち該当者が多い属性は？

キーワード	会員数	該当者	割合
経済	20000	16000	**80%**
IT	20000	4000	20%
エンタメ	20000	100	2%
スポーツ	20000	2000	10%
旅行	20000	12000	**60%**

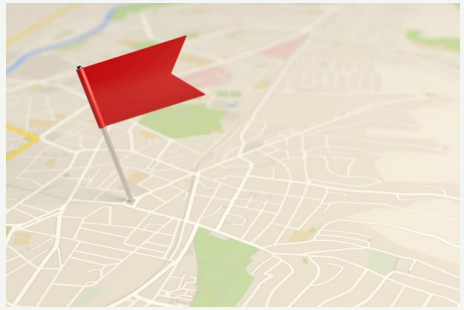
データを売る：データビジネスの参入障壁は下がっている

これまで　膨大なユーザーベース × 幅広いコンテンツや設備がないと収益化できない → データビジネスの専業化

これから　参入を阻むシステム化のコストは大幅に下がり、データそのものの価値が収益につながる

メディア

興味・関心にフィットする
豊富なコンテンツ



IoT、センサー、位置情報

現在地が取得できている
ことでの利便性

決済

お得感や利便性
店舗そのものの価値

本セッションでは

データの流通？
データクリーンルームって？

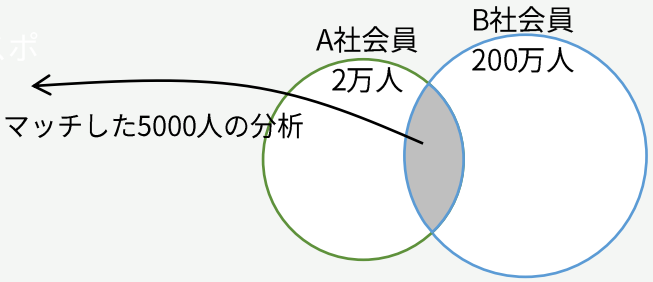
Snowflake DCR

アーキテクチャから見る
注目ポイント

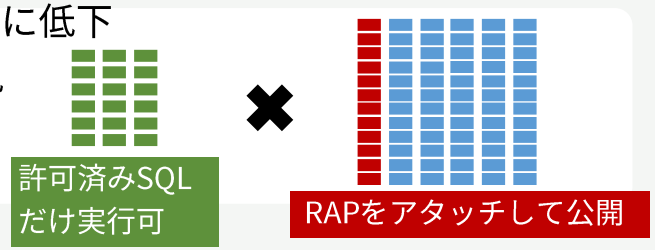
結局なにがおきるの？
利用イメージをもっと膨らませて
「相手探し」の第一歩を踏み出そう！

データクリーンルーム＝分析のためのデータ流通の1つの形
・データ授受のためのシステム開発を省力化
・プライバシーへの対応

キーワード	マッチした数	該当者	割合
経済	5000	4000	**80%**
IT	5000	1000	20%
旅行	5000	3000	**60%**



Snowflakeのデータクリーンルーム
・マルチテナントゆえ公開までのハードルがさらに低下
・JOINやGROUP BYを強制、プライバシーを重視



これからのデータ流通
・参入障壁が下がって有益なデータを売り買い
・外部データ利用が当然、分析したものが勝者に

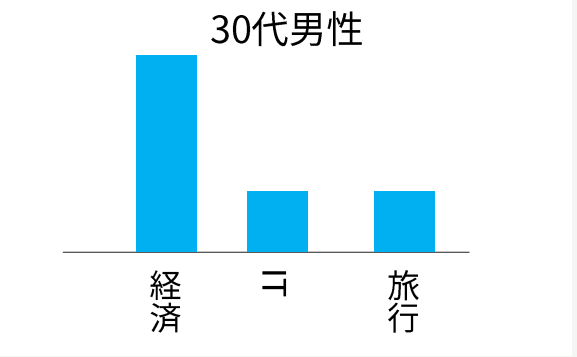
これから来る課題

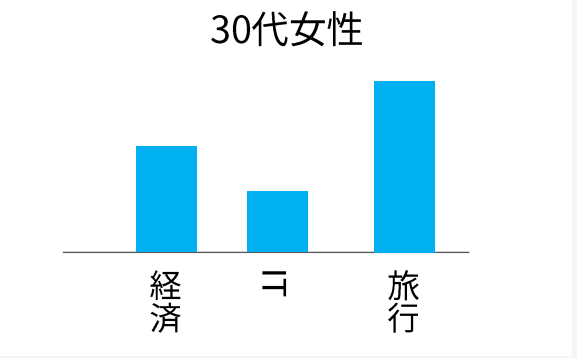
結合に使えるユーザー一覧があれば
すぐに始められる

会員ID	購入済	年齢層	性別
aaa@mail	✓	20代	男
bbb@mail	✓	30代	女
ccc@mail	✓	30代	男
eee@mail	✓		
fff@mail	✓		
xxx@mail	✓		
yyy@mail	✓		
hhh@mail	✓		

自社顧客を捉えるためのキーの充足
→メアド登録が
消費者のメリットになる施策

自社でさらに分析を深めるための
属性を収集する





Data to the People

すべての人にデータを