

ガバナンスとアジリティ データ民主化を実現させる LookML とは？

Google Cloud カスタマー エンジニア
蓮池 拓哉

本日のアジェンダ

データ民主化の現状と課題	01
--------------	----

データ民主化を支援する LookML とは？	02
------------------------	----

効率的なデータサイエンスワークフロー	03
--------------------	----

01

データ民主化の現 状と課題

データ民主化に向けた課題

ガバナンス

指標を算出する際のデータの取得元、計算手法、条件などが人によってバラバラであるため、同じ指標でも値が異なったりすることから定義の確認や数字のずれなどによるコミュニケーションコストが増加したり、判断を誤ってしまう

公開範囲の設定

社内外に蓄積されたデータを十分に活用することでビジネスのあらゆるところでその価値を得たいと考えるが、セキュリティの問題から限定的なデータの開示に留まるため、データ活用の価値も限定的にしか得ることができず、取り組みが持続されない

人材・スキル不足

データ分析をする前提としてデータの所在やデータ定義への理解、取り出すための技術が必要になるため、データ活用を始められるまでのリードタイムが長くなり、ひいてはデータ活用自体がされない

データ民主化を実現する方法

共通言語化

取得元や計算方法も含め社内の指標定義を統一化し共有していくことで関係者間の認識の相違などを排除し、コミュニケーションコストを削減し、“人によって異なる”状況がなくなることでデータを社内の共通言語化する

柔軟かつ適切な公開範囲

機密度や必要な情報レベルに合わせたセキュアなデータ利用環境提供し幅広いデータをタイムリーに利用できることで、データから得る価値を最大化

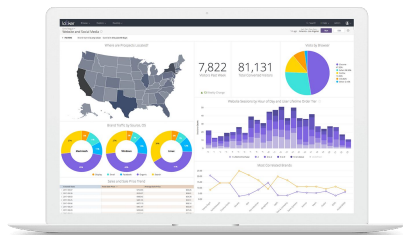
分析ナレッジの資産化

利用できるデータにどのようなものがあるか、それがどのような意味を持つのかを1箇所にまとめ確認できることで調べ、見つけ、使う、という流れをスムーズに実現し、データを身近に

可視化・分析を行う
データアナリティクスツール

より強力なデータ活用

BI だけでなく、
データ活用が広がる機能



データを
アクションに繋げる

slack zapier Airtable

Segment twilio Marketo

メールでレポート配信

動的なアラート発報

機械学習/ AI のワークフローに
データを送信

自社アプリに組み込んで活用
データの収益化

他社向けサービスに埋め込みBIとして活用

iFrame で自社ポータルサイトなどへ配置

SFA など他の SaaS ツール画面への表示

API 連携による柔軟な開発

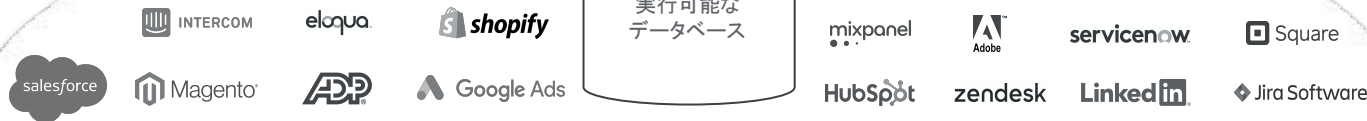
堅牢なプラットフォーム

データと指標を
統合する

Looker

共通の指標 | 権限の管理 | Git でのバージョンの管理 | セキュリティ | イン データベース | Cloud

企業が持つ様々なデータ



02

データ民主化を支援する LookML とは？

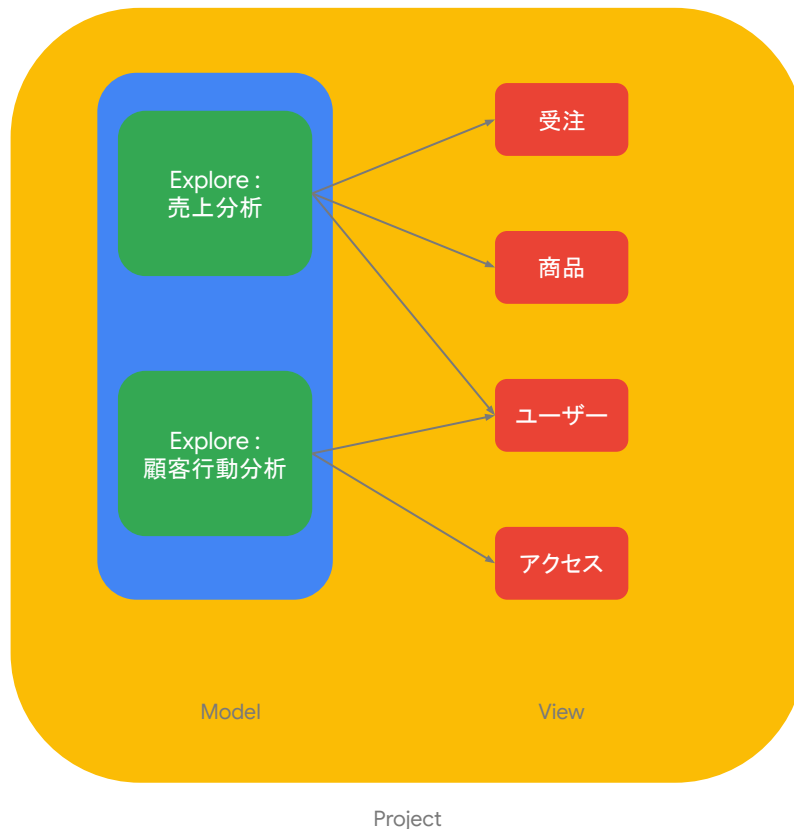
LookML で実現する 定義の一元管理

- SQL を抽象化し Looker に
情報を提供
- データベースとユーザーの間にモ
デルレイヤーを作成
- Git による構成管理で定義の
氾濫を排除

```
dimension_group: created {  
  type: time  
  timeframes: [raw, time, date, week, month, quarter, year]  
  sql: ${TABLE}.created_at ;;  
}  
  
dimension: status {  
  type: string  
  sql: ${TABLE}.status ;;  
}  
  
dimension: traffic_source {  
  type: string  
  sql: ${TABLE}.traffic_source ;;  
}  
  
dimension: user_id {  
  type: number  
  sql: ${TABLE}.user_id ;;  
}  
  
measure: count {  
  type: count  
  drill_fields: [id, users.id, users.first_name, users.last_name]  
}  
  
measure: count_html {  
  type: count  
  drill_fields: [id, users.id, users.first_name, users.last_name]  
  html:  
    {% if value > 400 %}  
    <b><p style="color:white; background-color:darkgreen; margin: 0"> {{value}} </p>  
    {% elsif value > 380 %}  
    <b><p style="color:white; background-color:goldenrod; margin: 0"> {{value}} </p>  
    {% else %}  
    <b><p style="color:white; background-color:darkred; margin: 0"> {{value}} </p>  
    {% endif %}  
    ;;  
}
```

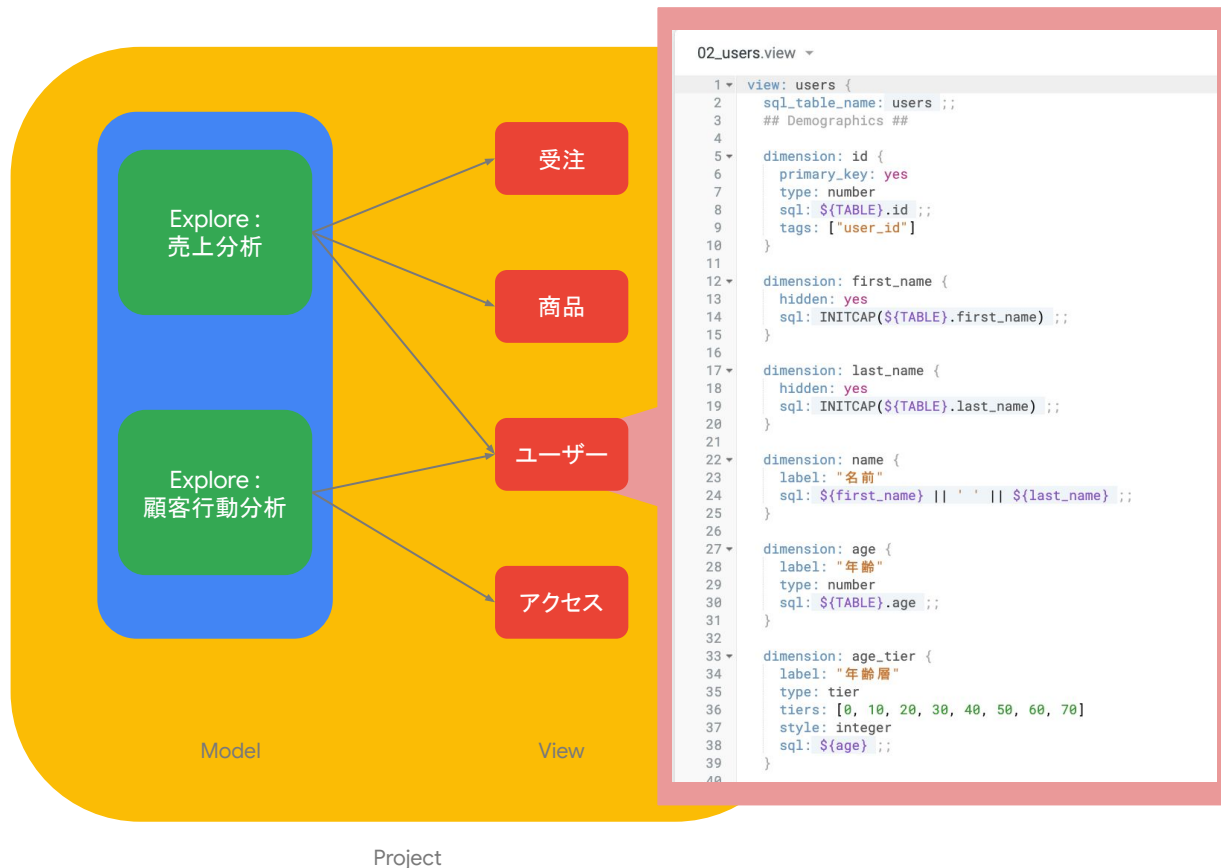

再利用可能な定義情報

- View ファイル
Looker で取り扱う論理的なテーブル定義
- Model ファイル
テーブル間の結合ロジック
- Project
ファイルを管理する単位



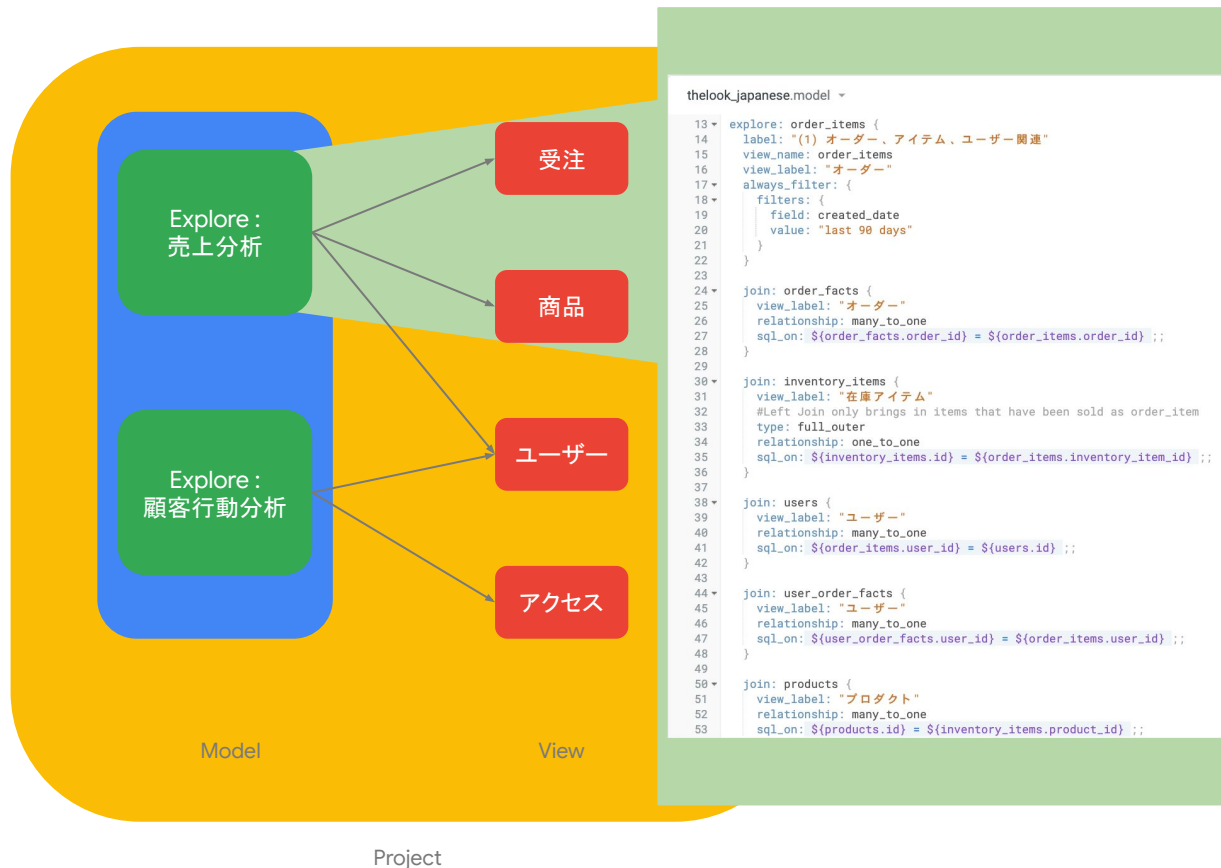
再利用可能な定義情報

- View ファイル
Looker で取り扱う論理的なテーブル定義
- Model ファイル
テーブル間の結合ロジック
- Project
ファイルを管理する単位



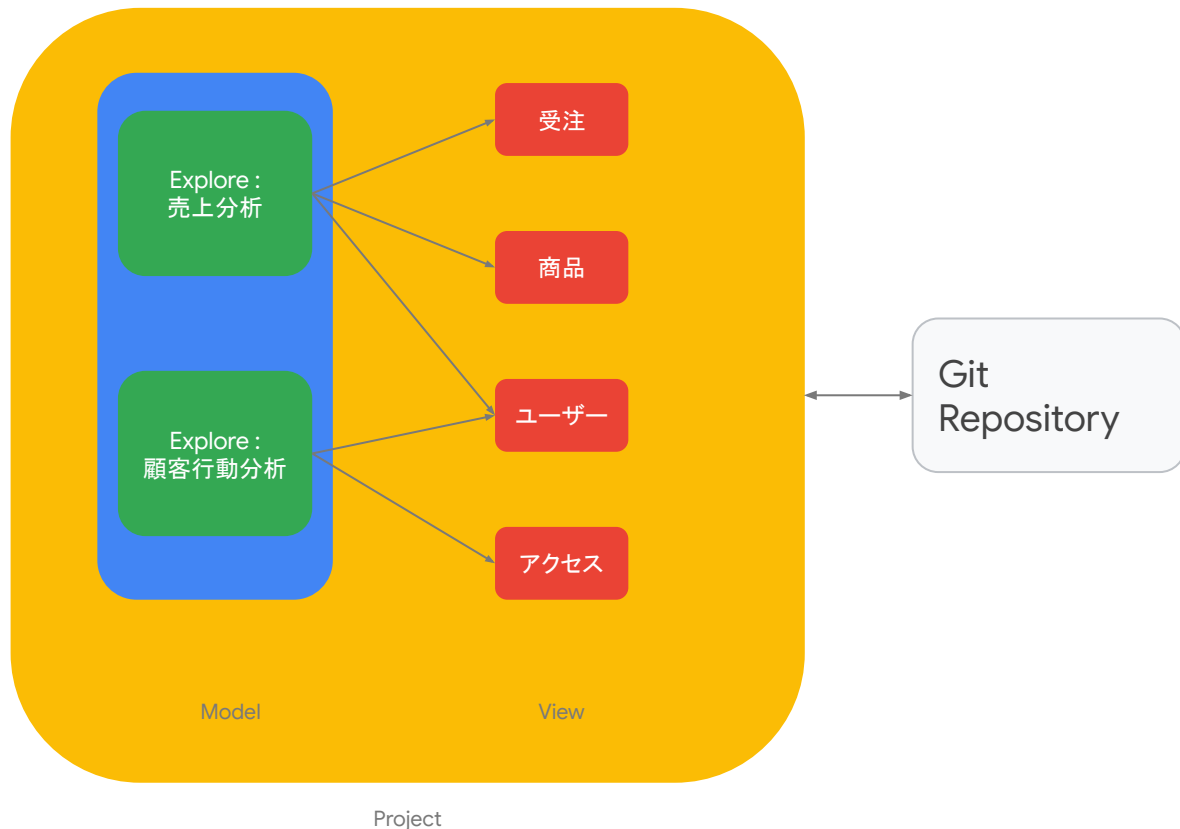
再利用可能な定義情報

- View ファイル
Looker で取り扱う論理的なテーブル定義
- Model ファイル
テーブル間の結合ロジック
- Project
ファイルを管理する単位



再利用可能な定義情報

- View ファイル
Looker で取り扱う論理的なテーブル定義
- Model ファイル
テーブル間の結合ロジック
- Project
ファイルを管理する単位



LookML 編集デモ

Looker で実現する柔軟かつ容易なアクセス制御

行レベル フィルター

ユーザーの所属部署やエリアなどに合わせて、フィルターを操作することなく、いつでも自分に関係しているデータだけが表示されることで、データを探す手間と管理側でマートを分割する手間を削減

列レベルのアクセス権限

職位や職責に合わせて機密度の異なるデータへのアクセス権を正しく設定することで、公開できるデータの種類を増やし、データ利活用の幅を拡大

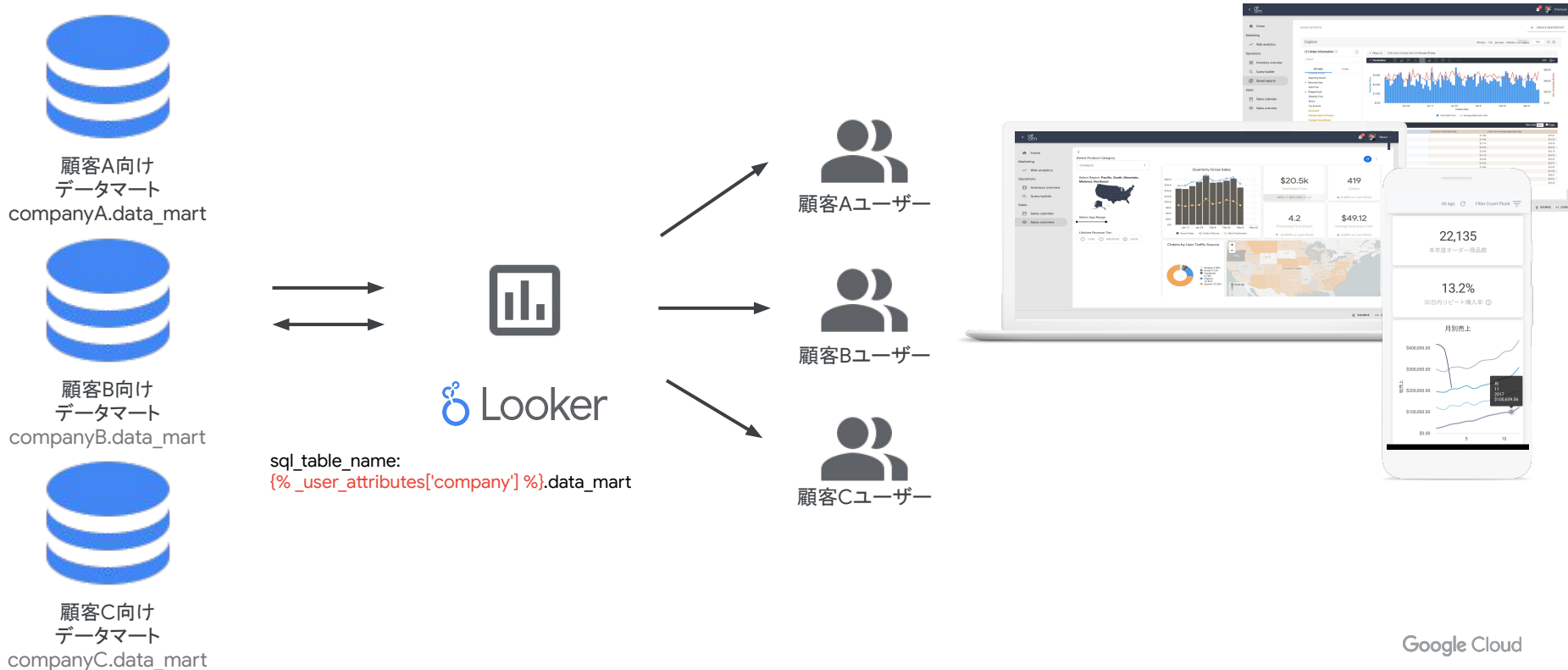
属性に合わせた取得元の変更

アクセス可能なユーザーが異なる同一フォーマットのデータも属性状況に合わせて取得元を動的に変更できることにより運用工数を削減し、データの外部公開を実現

属性に合わせた公開範囲変更のデモ

パラメーターを利用した接続データの動的な出しわけ

アクセスユーザーに合わせて取得元を変更したいような場合でも単一ダッシュボード、単一コネクションで対応



データ ディクショナリー

ユーザーが利用可能なデータ項目に
どんなものがあるのか、データの種類は？ど
んな定義で集計されているのか？などを辞書
形式に簡単に検索、確認が可能

Looker

閲覧Explore開発管理

検索

Data Dictionary

Select a Model
ECサイトデータ

Explores
Search Model

(1) オーダー、アイテム、ユーザー関連
(2) ページイベントデータ関連
(3) ウェブセッションデータ関連
(4) バスケット解析
(5) 顧客内シェア解析
(6) カスタマージャーニーマップ
(7) 在庫分析
(8) 在庫スナップショット分析
(9) 期間比較分析

ECサイトデータ

Select a field for more information.

Filter fields in this Explore

Has Description
Yes No
Fields
Dimension Measure
Has Tags
Yes No
Has Comments
Yes No
Type
String Date Number Yesno Tier Location Zipcode Count Sum Median Average Percent Max
Min

オーダー

Field Label	Description	LookML Name	Type	SQL
ID		order_items.id	Number	<code>\${TABLE}.id</code>
オーダーID		order_items. order_id	Number	<code>\${TABLE}.order_id</code>
オーダー原価		order_facts. order_cost	Number	<code>\${TABLE}.order_cost</code>
オーダー粗利益		order_facts. order_gross_margin	Number	<code>order_facts.order_gross_margin</code>
オーダー順		order_facts. order_sequence_number	Number	<code>\${TABLE}.order_sequence_number</code>
オーダー額		order_facts. order_amount	Number	<code>\${TABLE}.order_amount</code>
コホート期間		order_items. cohort_time_period	String	<code>CASE WHEN (\${parameter time_period %}) = 'Date' THEN \${created_date}::varchar WHEN (\${par...</code>
...		...	...	<code>CASE WHEN \${TABLE}.status =</code>

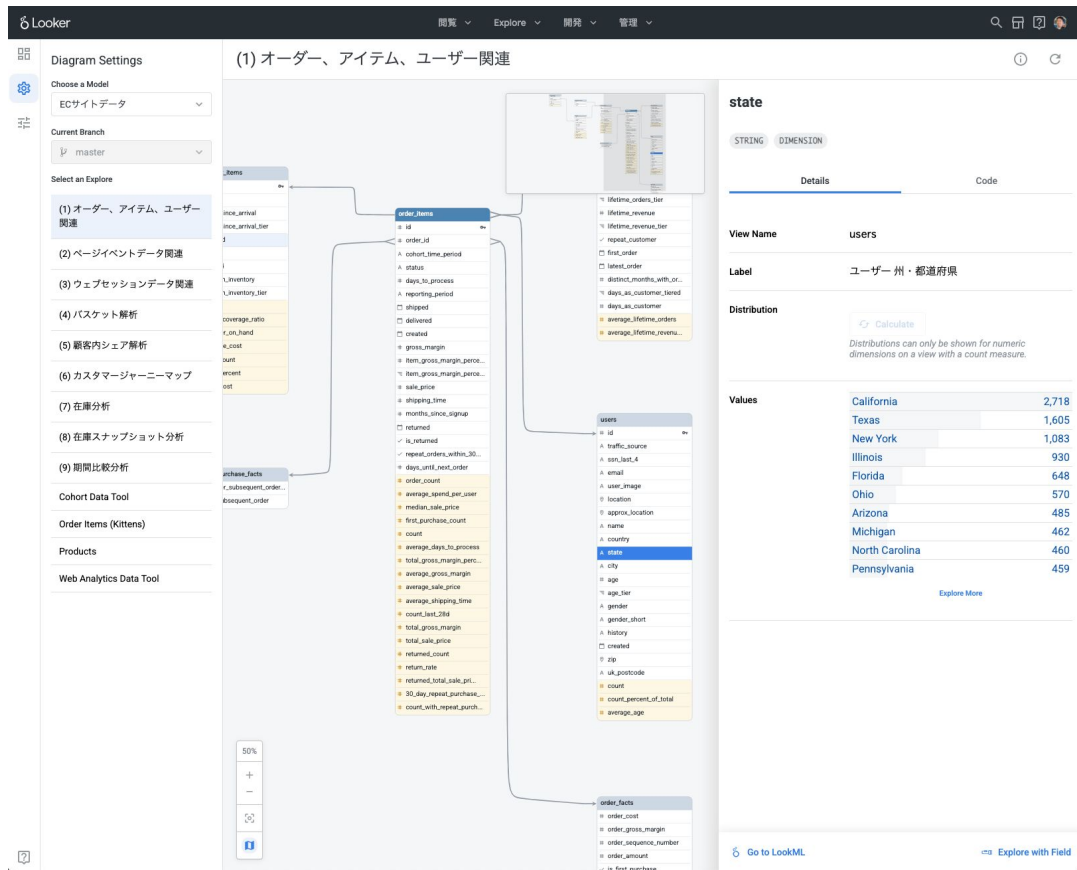
LookML ダイアグラム

SQL やテーブル構造に詳しい

ユーザー向けにテーブル同士の

関係性や値の分布状況などを表示し、

データセットの構造を提供



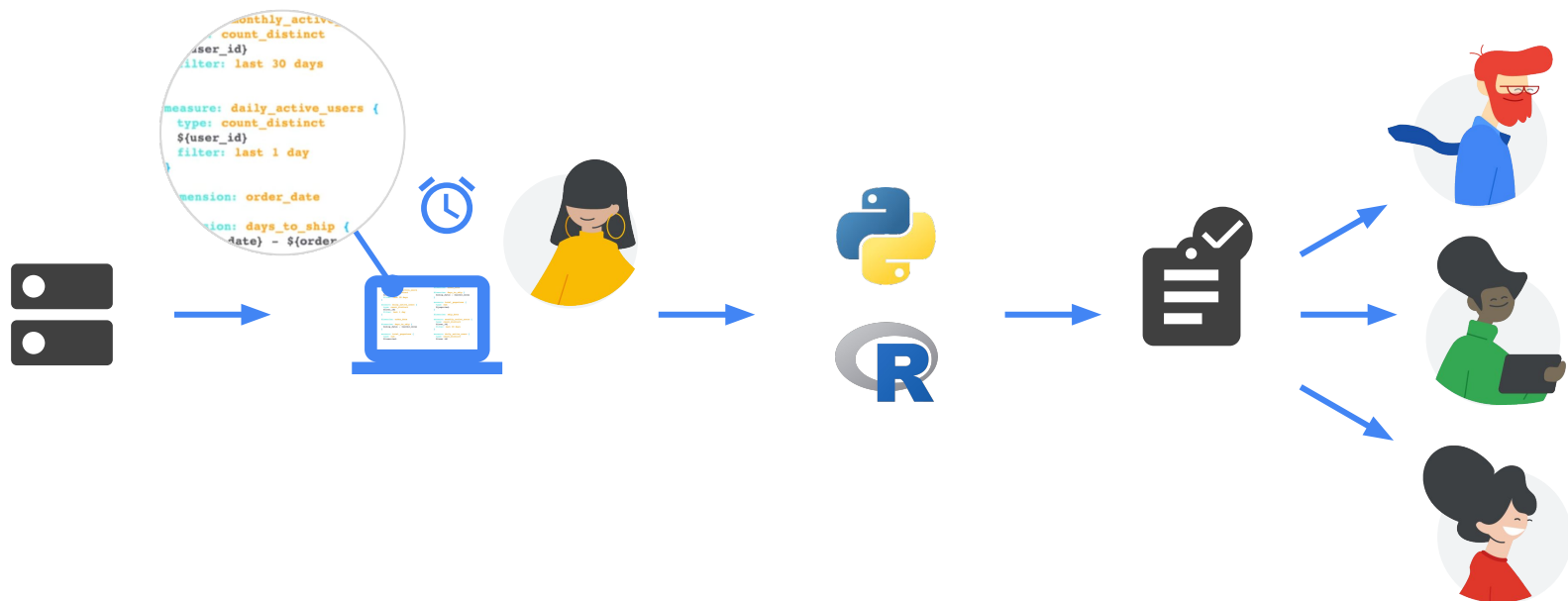
ツールだけでなくナレッジも見える形で提供し、
データ文化を醸成”

03

効率的なデータサイエンス ワークフロー

Looker + BigQuery

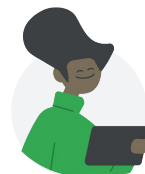
従来のデータサイエンスワークフロー



Looker と Google BigQuery ML を組み合わせた ワークフロー



BigQuery



BigQuery ML

```
CREATE OR REPLACE MODEL
  bqml_tutorial.london_station_clusters OPTIONS(model_type='kmeans',
    num_clusters=4) AS
WITH
  hs AS (
    SELECT
      h.start_station_name AS station_name,
    IF
      (EXTRACT(DAYOFWEEK
        FROM
          h.start_date) = 1
      OR EXTRACT(DAYOFWEEK
        FROM
          h.start_date) = 7,
      "weekend",
      "weekday") AS isweekday,
    h.duration,
    ST_DISTANCE(ST_GEOGPOINT(s.longitude,
      s.latitude),
      ST_GEOGPOINT(-0.1,
        51.5))/1000 AS distance_from_city_center
```

```
FROM
  `bigquery-public-data.london_bicycles.cycle_hire` AS h
JOIN
  `bigquery-public-data.london_bicycles.cycle_stations` AS s
ON
  h.start_station_id = s.id
WHERE
  h.start_date BETWEEN CAST('2015-01-01 00:00:00' AS TIMESTAMP)
    AND CAST('2016-01-01 00:00:00' AS TIMESTAMP) ),
stationstats AS (
  SELECT
    station_name,
    isweekday,
    AVG(duration) AS duration,
    COUNT(duration) AS num_trips,
    MAX(distance_from_city_center) AS distance_from_city_center
  FROM
    hs
  GROUP BY
    station_name, isweekday)
SELECT
  * EXCEPT(station_name, isweekday)
FROM
  stationstats
```

LookML + BigQuery ML

```
view: trip_count_input {  
  derived_table: {  
    explore_source: trip {  
      column: start_date {}  
      column: humidity { field: daily_weather.humidity }  
      column: temperature { field: daily_weather.temperature }  
      column: trip_count {}  
    }  
  }  
}
```

```
view: trip_count_regression {  
  # 2019-08-21 Bruce - Set max iterations down to 50 from 100. Getting error saying max allowable is 50  
  derived_table: {  
    datagroup_trigger: sweet_datagroup  
    sql_create:  
      CREATE OR REPLACE MODEL ${SQL_TABLE_NAME}  
      OPTIONS(model_type='linear_reg'  
        , labels=['trip_count']  
        , min_rel_progress = 0.05  
        , max_iteration = 50  
      ) AS  
      SELECT  
        * EXCEPT(start_date)  
      FROM ${trip_count_input.SQL_TABLE_NAME};  
  }  
}
```


モデルパフォーマンス

FILTERS

Error Matrix Threshold ≥ 0.65 Model Name is any value

85.81%

Accuracy (static)

82.64%

Recall (static)

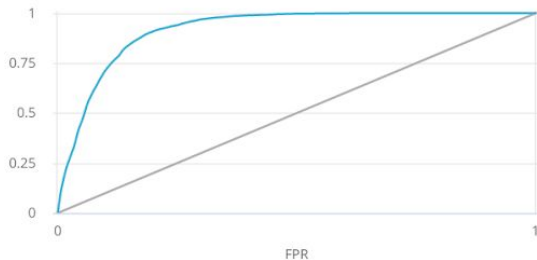
17.660%

F1 Score (static)

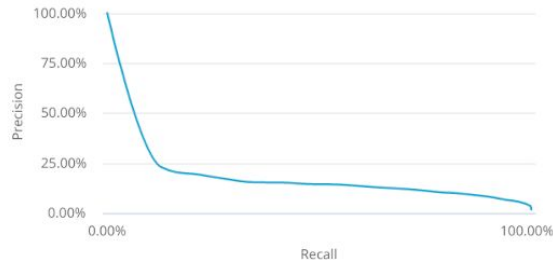
9.89%

Precision (static)

ROC Curve



Precision-Recall Curve



10,520

True Positives

70,859

False Positive

6,102

False Negatives

815,372

True Negatives

91.48%

Accuracy (threshold)

63.29%

Recall (threshold)

21.469%

F1 (threshold)

12.93%

Precision (threshold)

ROC Curve Details (use thresholds for filter values)

Threshold ^	False Negatives	False Positives	True Negatives	True Positives	Precision	Recall	Threshold Accuracy
0.0007	0	886,231	0	16,622	1.84%	100.00%	1.84%
0.0066	0	877,579	8,652	16,622	1.86%	100.00%	2.80%
0.0109	0	868,201	18,030	16,622	1.88%	100.00%	3.84%
0.0124	0	860,356	25,875	16,622	1.90%	100.00%	4.71%

Looker で実現する データ民主化

- データ定義を一元管理、
統一し”共通言語化”することで
”データ分析を均質化”
- 適切な公開範囲を設定し
あらゆるデータを開示することで、デー
タから得る”価値を最大化”
- データ利用を身近にすることで
”データの民主化を実現”



Thank you.