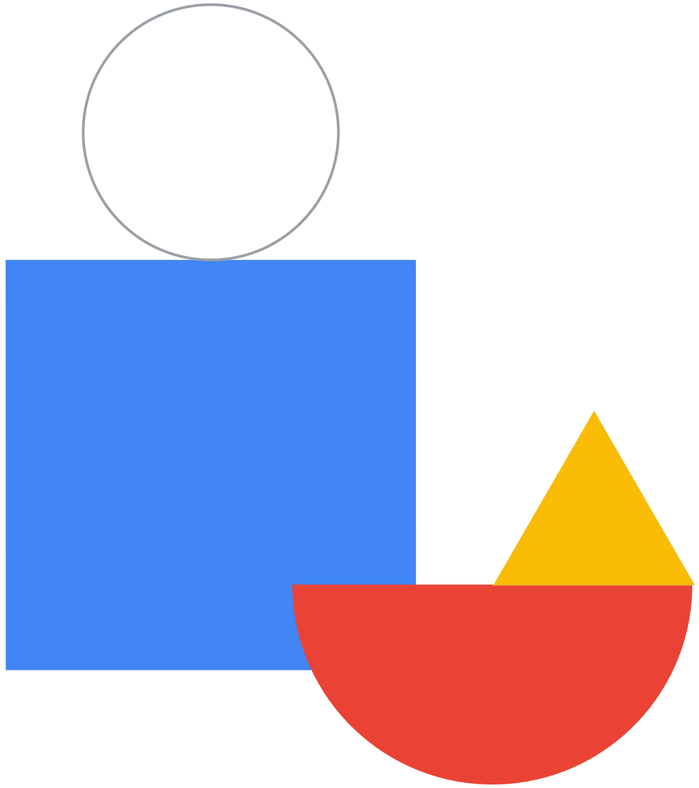
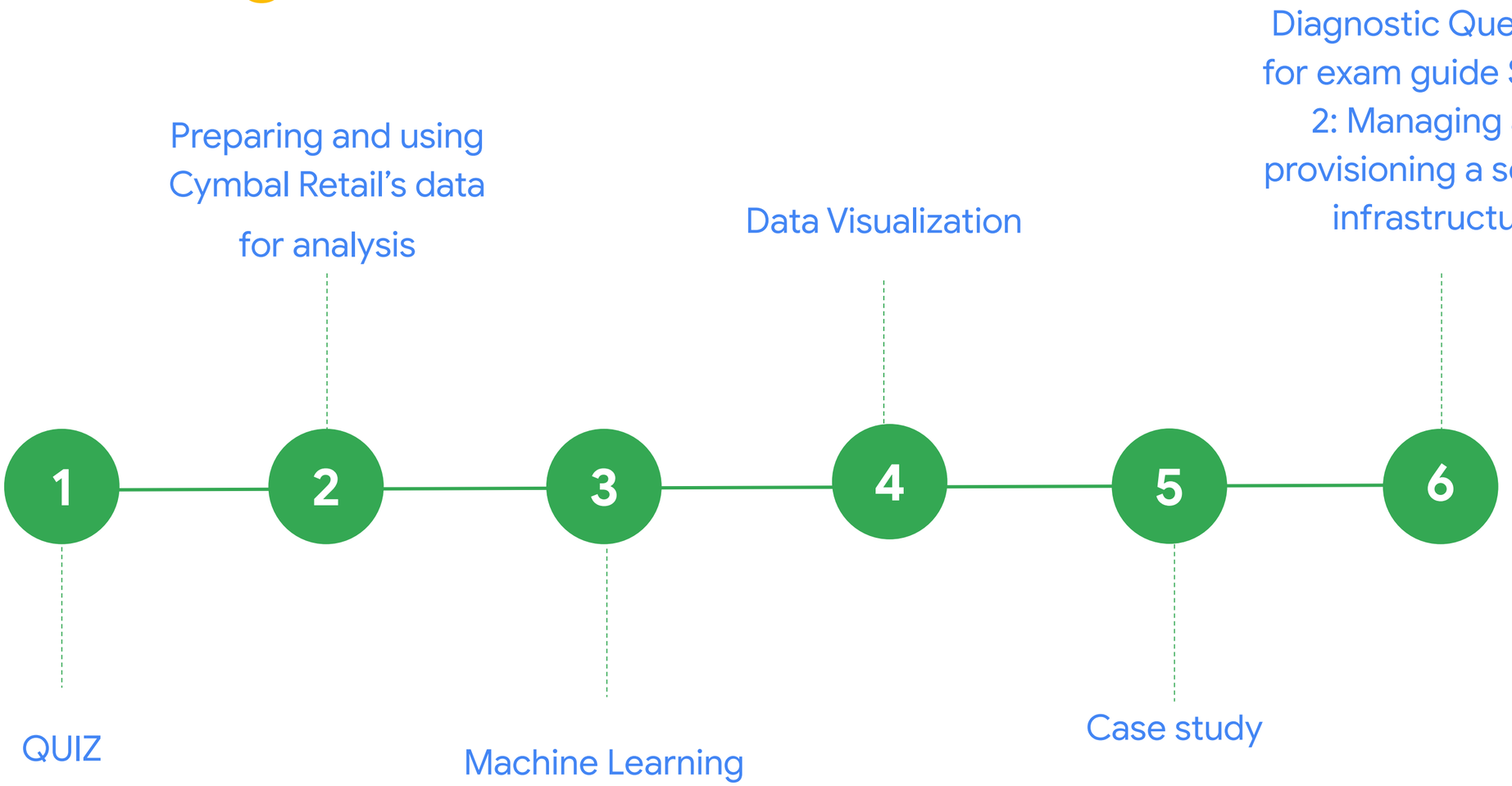
Google Cloud



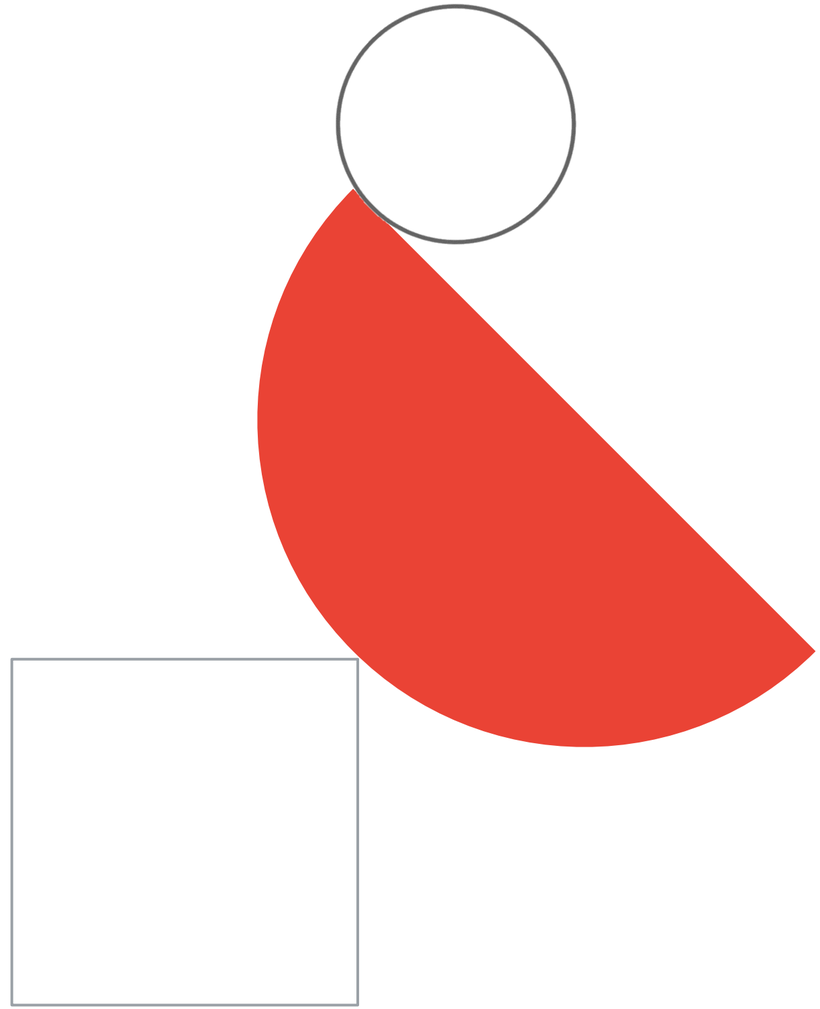
Preparing for Your Professional Data Engineer journey

Module 4: Preparing and Using Data for Analysis

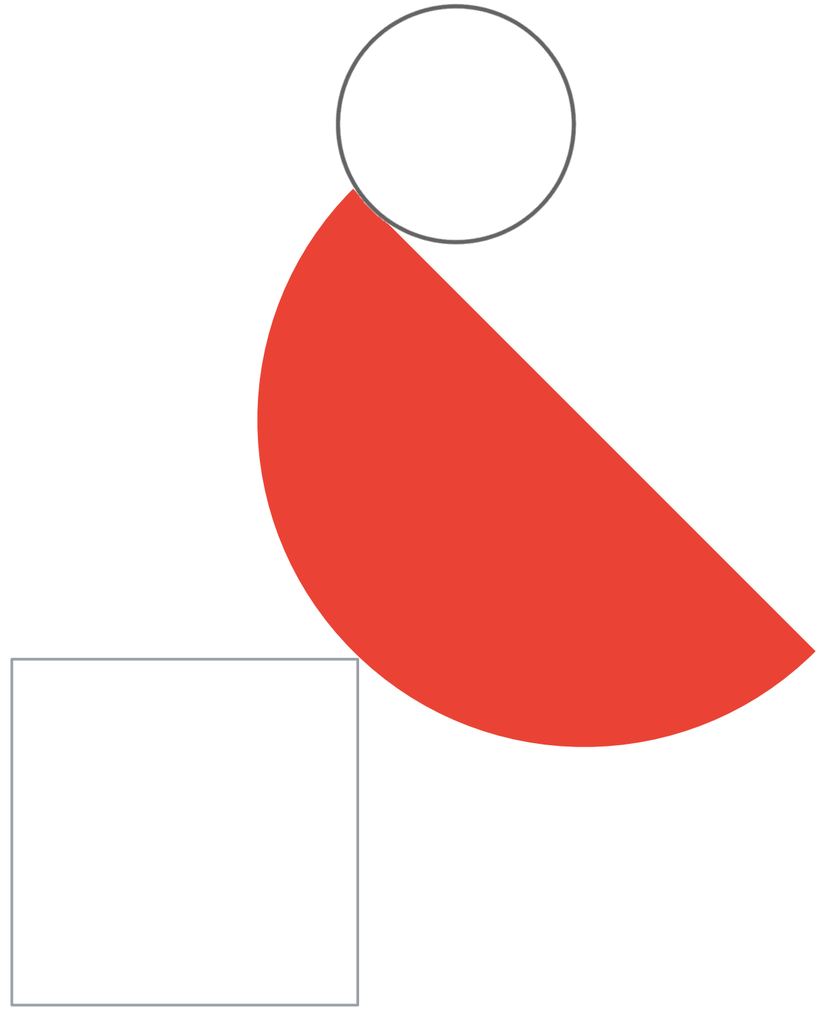
Week 5 agenda



QUIZ time!

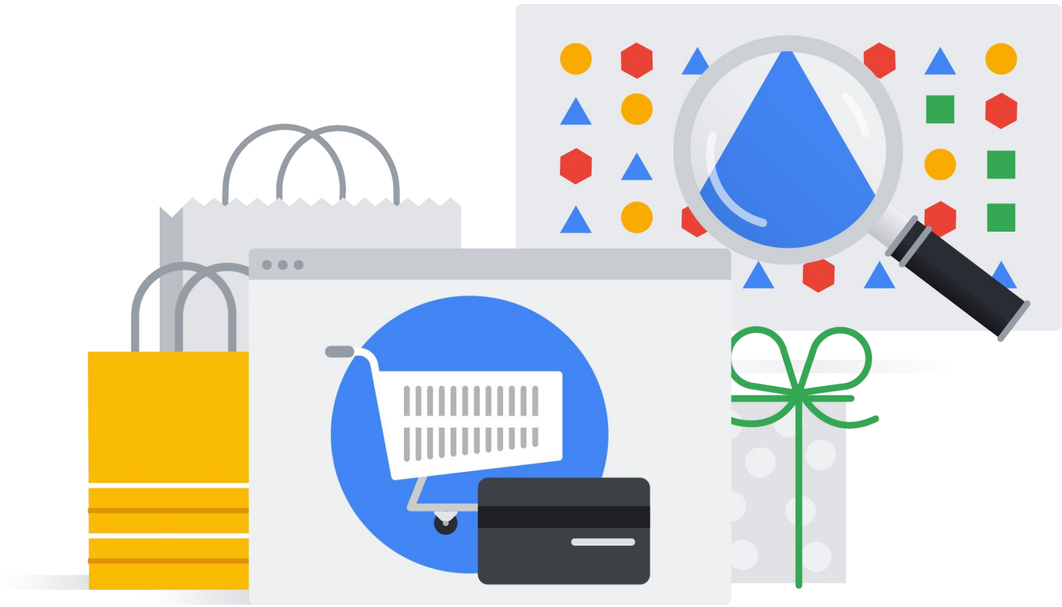


Preparing and using Cymbal Retail's data for analysis



Helping Cymbal Retail use data to make decisions

- Cymbal Retail wants to use data to make informed decisions.
- Systems need to be flexible, agile, and customizable.
- You need to help support the business in using data.

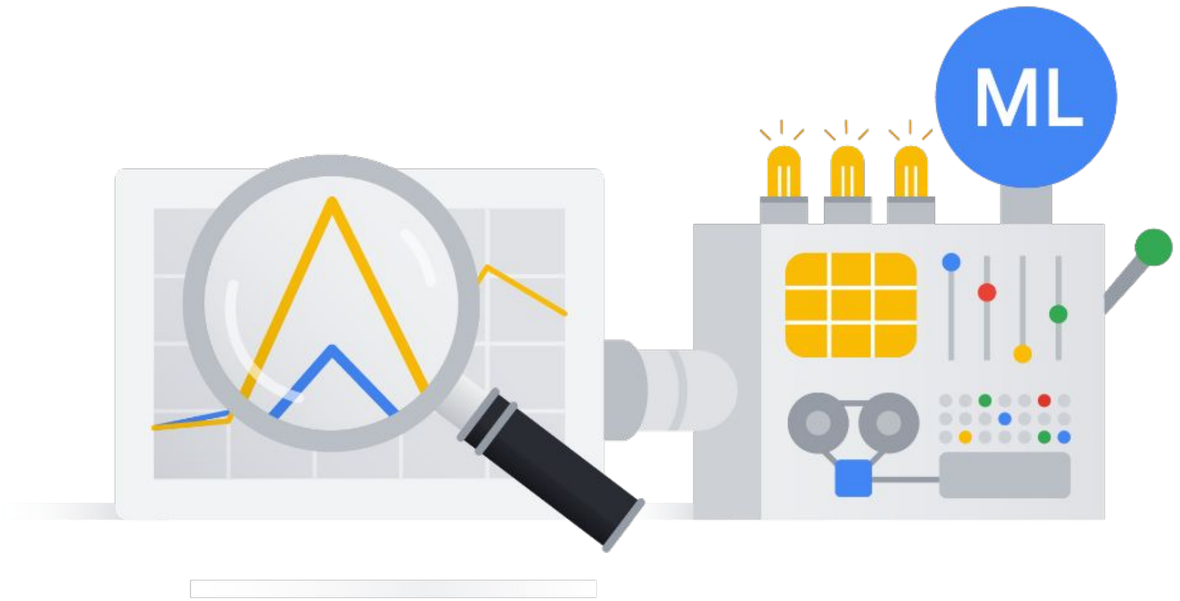


Sharing data and reporting

- Partner-specific reports and accounts
- Limited data access for industry and government bodies

Supporting machine learning advancements

- Cymbal Retail is using fine-tuned predictions to estimate demand and individual customer behavior.
- You need to help process data to make it usable for building machine learning models.

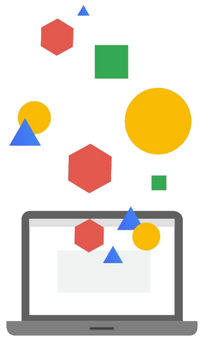


Machine Learning

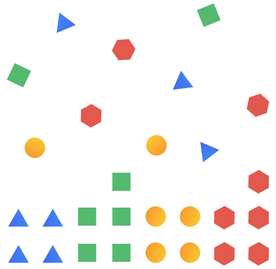
Examples are based on fictional data.

In reality, machine learning is...

1. Collect data



2. Organize data



3. Create model

4. Use machines to flesh out the model from data

5. Deploy fleshed out model

TIP

Which step or steps is the question asking about?

ML models

Supervised learning

Task-driven and identifies a **goal**

Past data to predict future trends

Classification

Predicts a categorical variable

Use an image to tell the difference between a cat and a dog

Regression

Predicts a continuous number

Use past sales of an item to predict a future trend

Unsupervised learning

Data-driven and identifies a **pattern**

Group customers together

Clustering

Groups data points together

Use customer demographics to determine customer segmentation

Association

Identifies underlying relationships

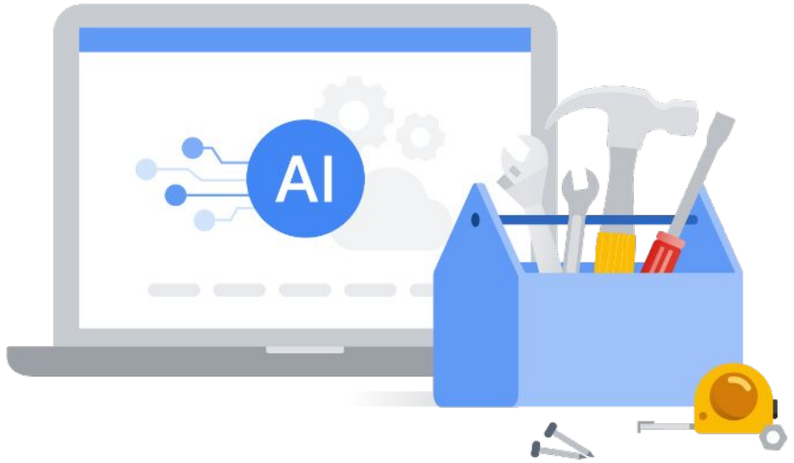
Correlation between two products to place them closer in a grocery store

Dimensionality reduction

Reduces the number of dimensions

Combining characteristics to create a quote

ML options on Google Cloud



Preconfigured

Low-code / No-code

DIY

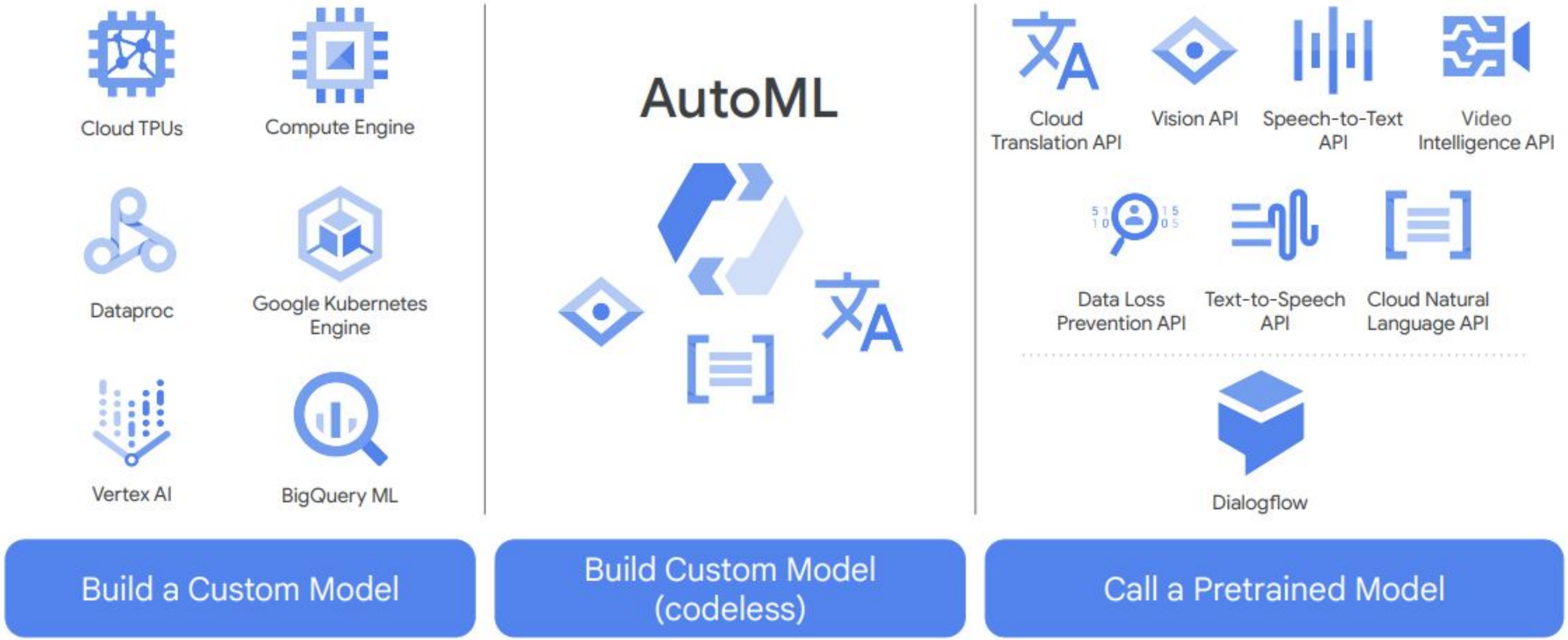
Pre-trained APIs

BigQuery ML

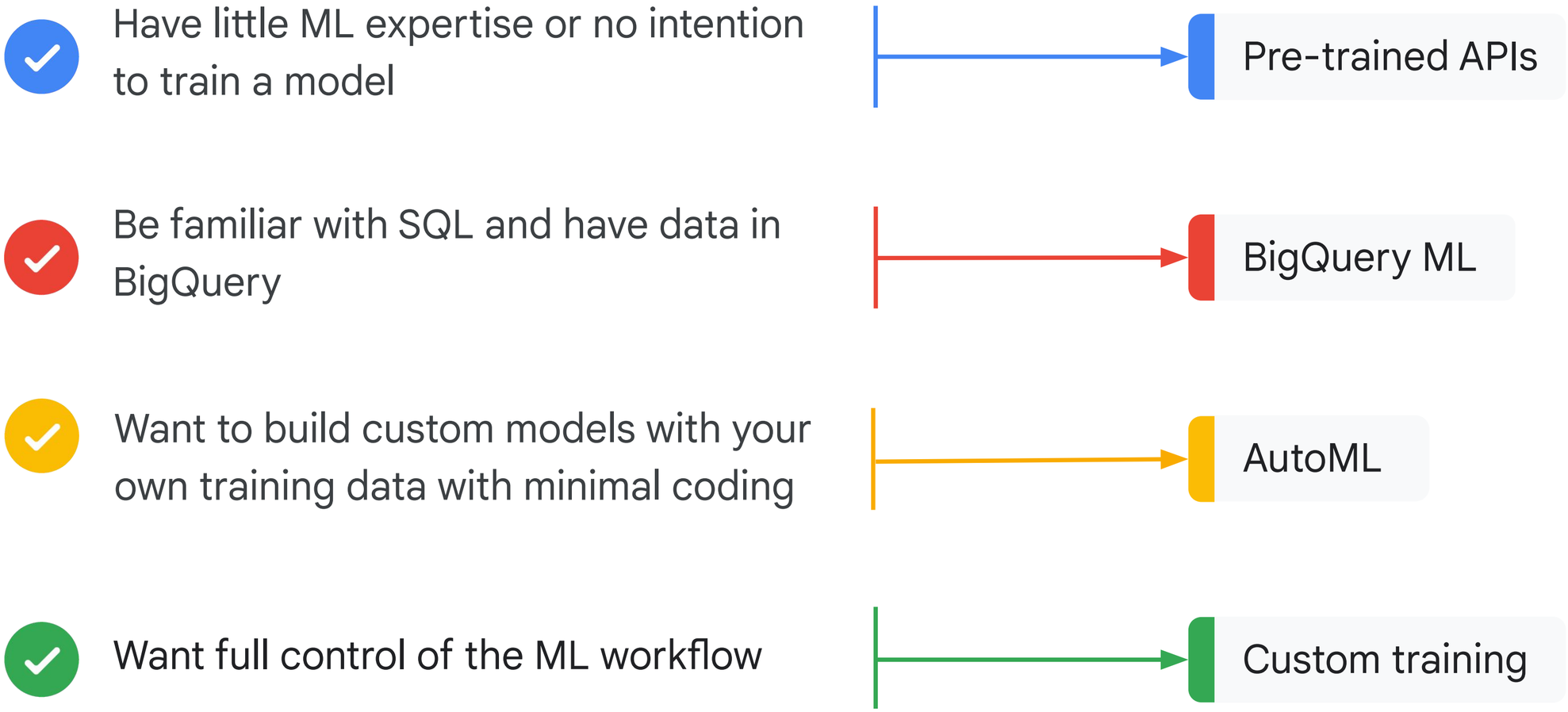
AutoML

Custom training

Machine Learning Solutions



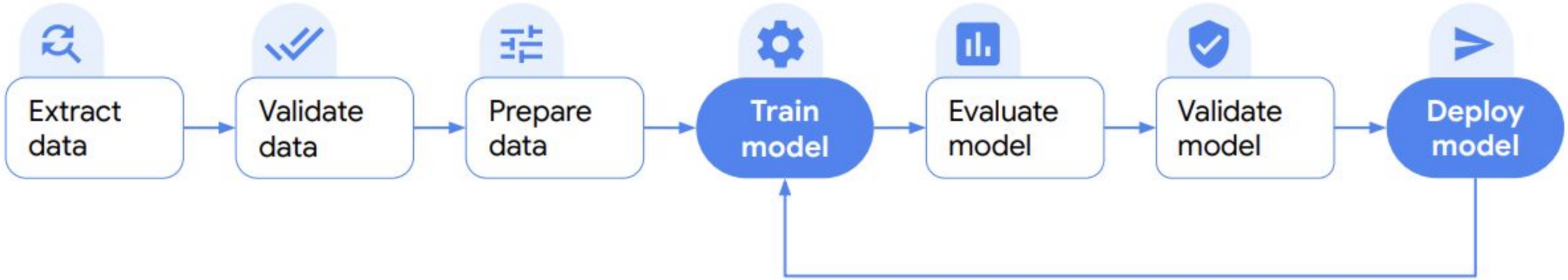
Decide the right ML option: business needs, ML expertise, budget



Pre-built API examples

Speech-to-Text API	Converts audio to text for data processing.
Cloud Natural Language API	Recognizes parts of speech called entities and sentiment.
Cloud Translation API	Converts text from one language to another.
Text-to-Speech API	Converts text into high-quality voice audio.
Vision API	Works with and recognizes content in static images.
Video Intelligence API	Recognizes motion and action in video.

ML Pipelines



BigQuery ML

1. **Execute** ML initiatives without moving data from BigQuery
2. **Iterate** on models in SQL in BigQuery to increase development speed
3. **Automate** common ML tasks and hyperparameter tuning

https://cloud.google.com/bigquery-ml/docs/tutorials

Working with BigQuery ML

01

Dataset

02

Create/ train

```
CREATE MODEL `bqml_tutorial.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
```

03

Evaluate

```
FROM
  ML.EVALUATE(MODEL
`bqml_tutorial.sample_model`,
    TABLE eval_table)
```

04

Predict/ classify

```
FROM
    ML.PREDICT(MODEL
`bqml_tutorial.sample_model`,
      table game_to_predict) ) AS
predict
```

BigQuery ML supported models and features

Classification

- Logistic regression
- DNN classifier (TensorFlow)
- XGBoost
- AutoML Tables
- Wide and Deep NNs[Preview, GA H1'21]

Regression

- Linear regression
- DNN regressor (TensorFlow)
- XGBoost
- AutoML Tables
- Wide and Deep NNs[Preview, GA H1'21]

Other Models

- k-means clustering
- Time series forecasting
- Recommendation: Matrix factorization
- Time series anomaly detection[Preview Q2'21, GA H2'21]

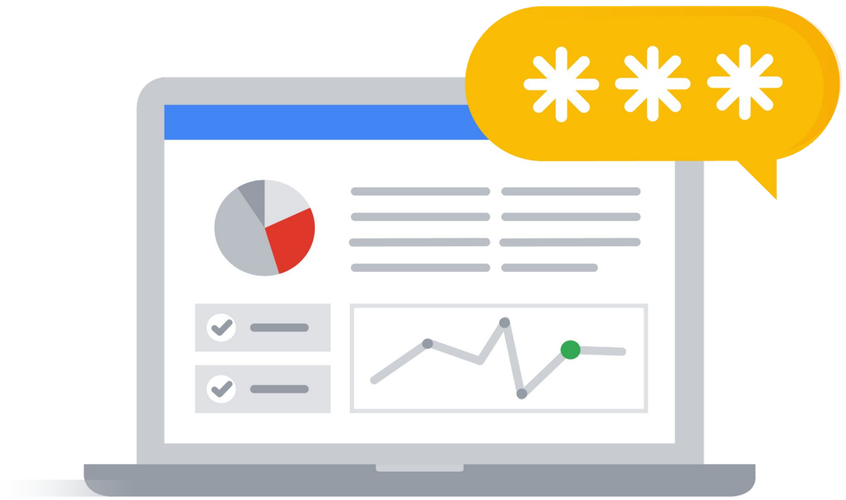
Model ops and explainability

- Import/export TensorFlow models for batch and online prediction
- **hyperparameter tuning using Cloud AI Vizier[Preview H1'21, GA H2'21]**
- **Model explainability using Cloud AI[Preview H1'21, GA H2'21]**
- **Managed Kubernetes and TFX pipelines[Preview H2'21, GA 2022]**
- **List models for comparison and online deployment in Cloud AI[Preview H2'21, GA 2022]**
- Model versioning, continuous monitoring[future]

Data Visualization

Examples are based on fictional data.

What is Looker?



Looker can help you:

- See the data your company collects.
- Answer questions as you have them.
- Stay up to date with the status of your business.
- Use data for daily decisions, instead of waiting for reports.

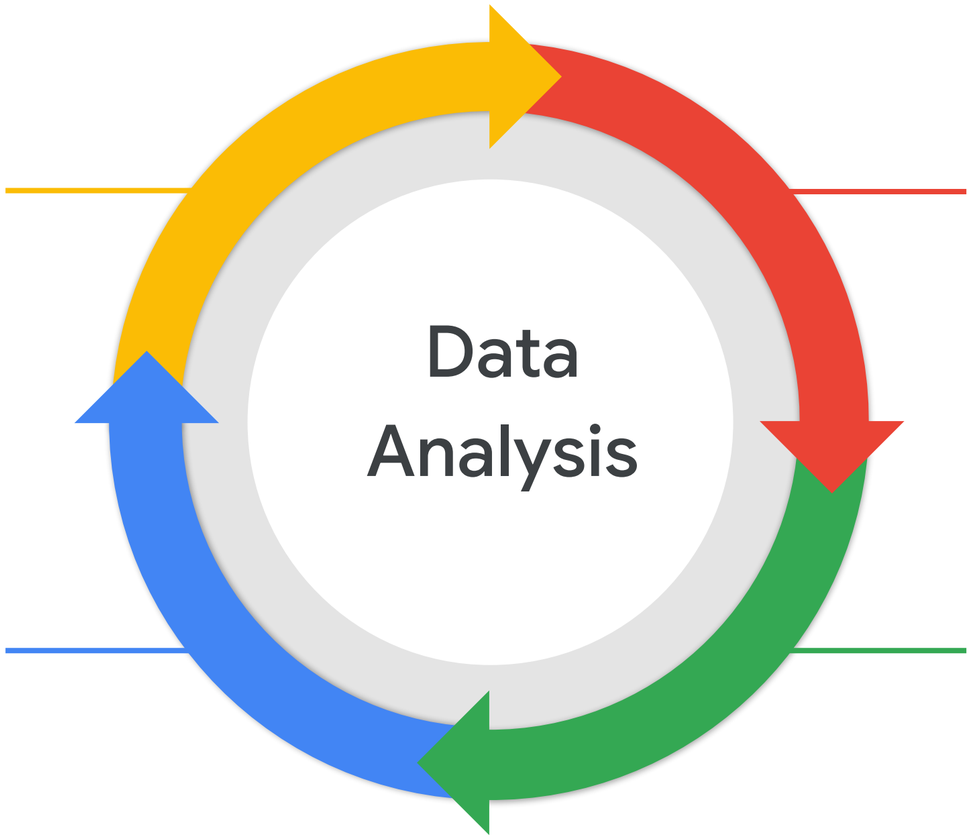
The role of Looker in the data analysis process

Define questions, facilitated by data exploration in Looker

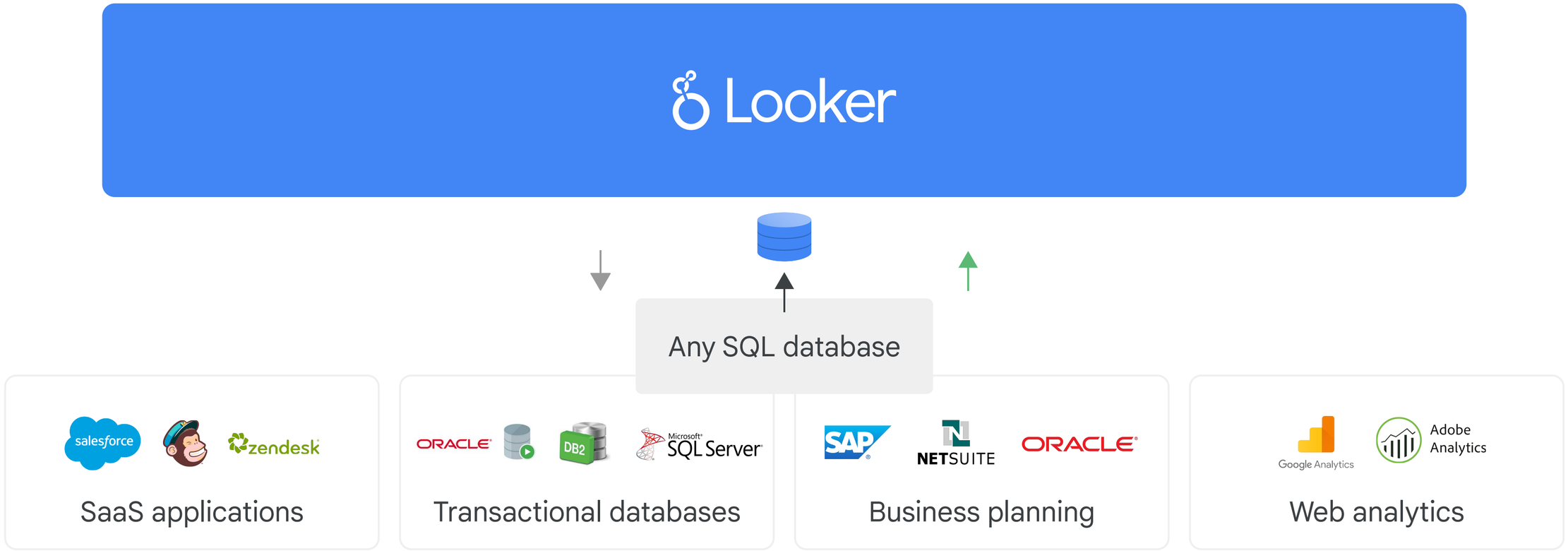
Identify required data by reviewing available data in Looker

Analyze data using Looker

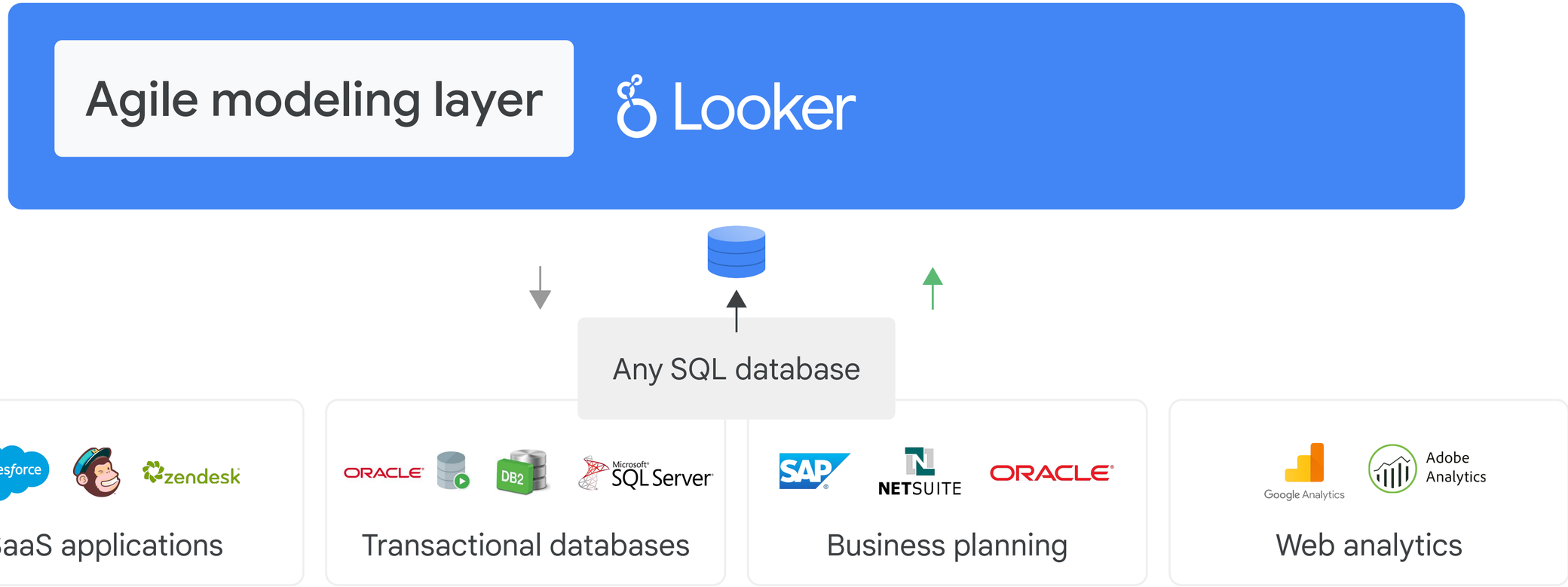
Interpret results from shareable visualizations and dashboards in Looker



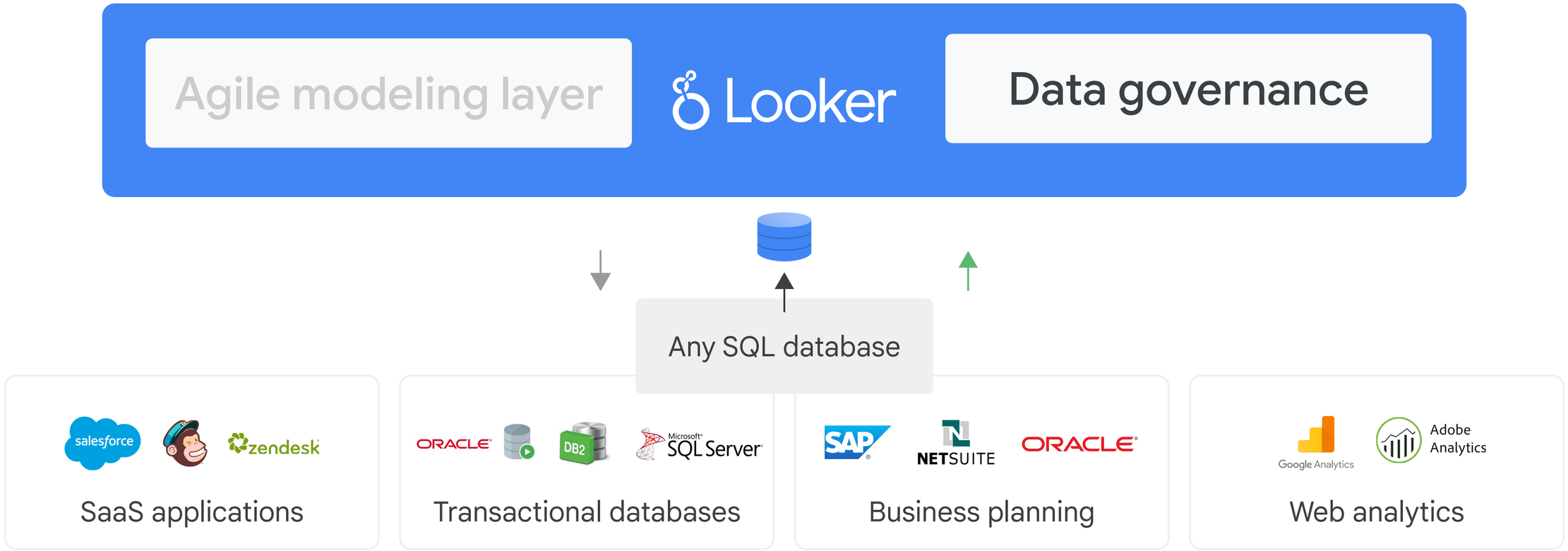
Architecture: Looker in the data stack



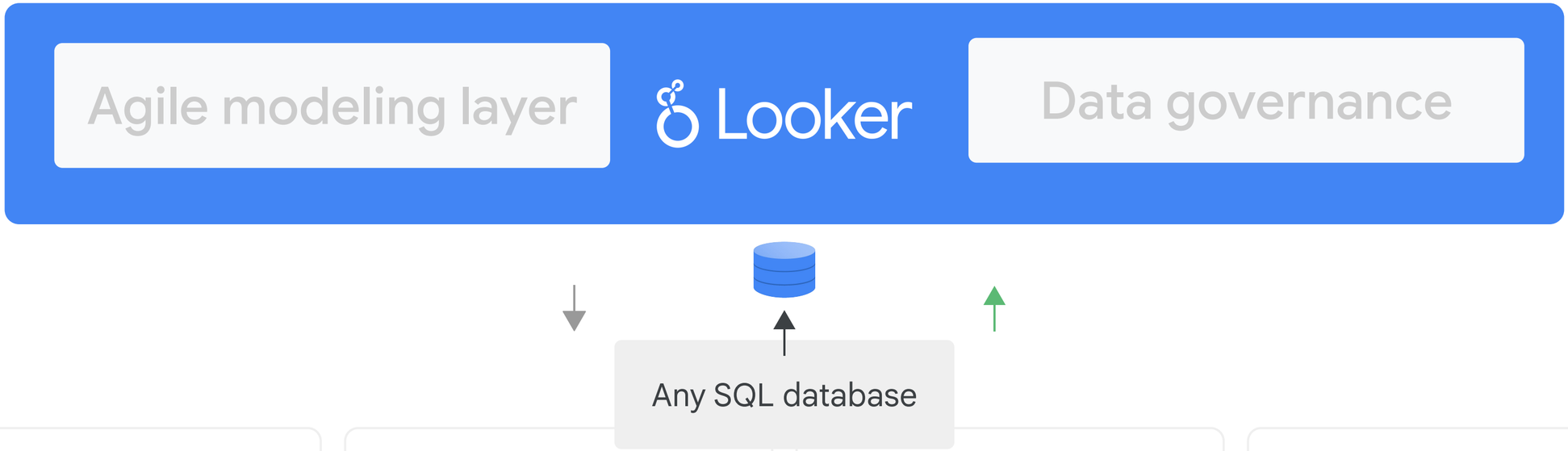
Architecture: Looker's agile modeling layer



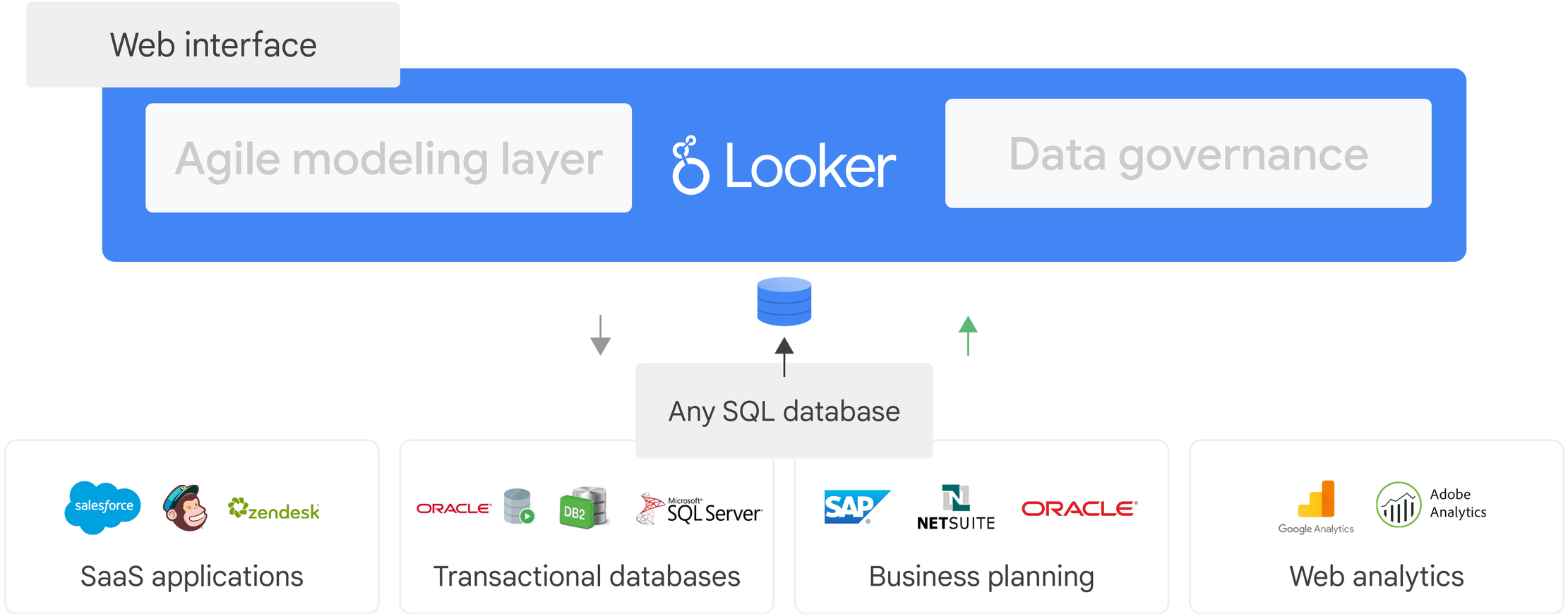
Architecture: Looker data governance



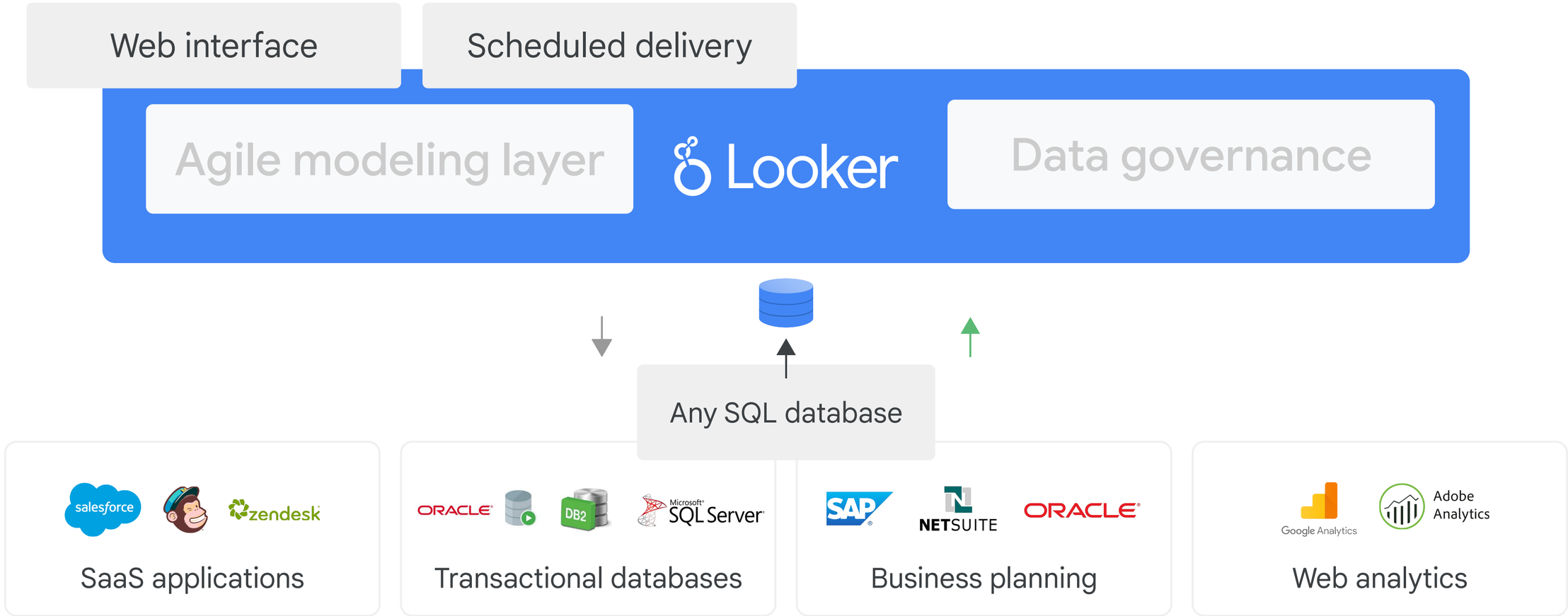
Architecture: Looker data democratization



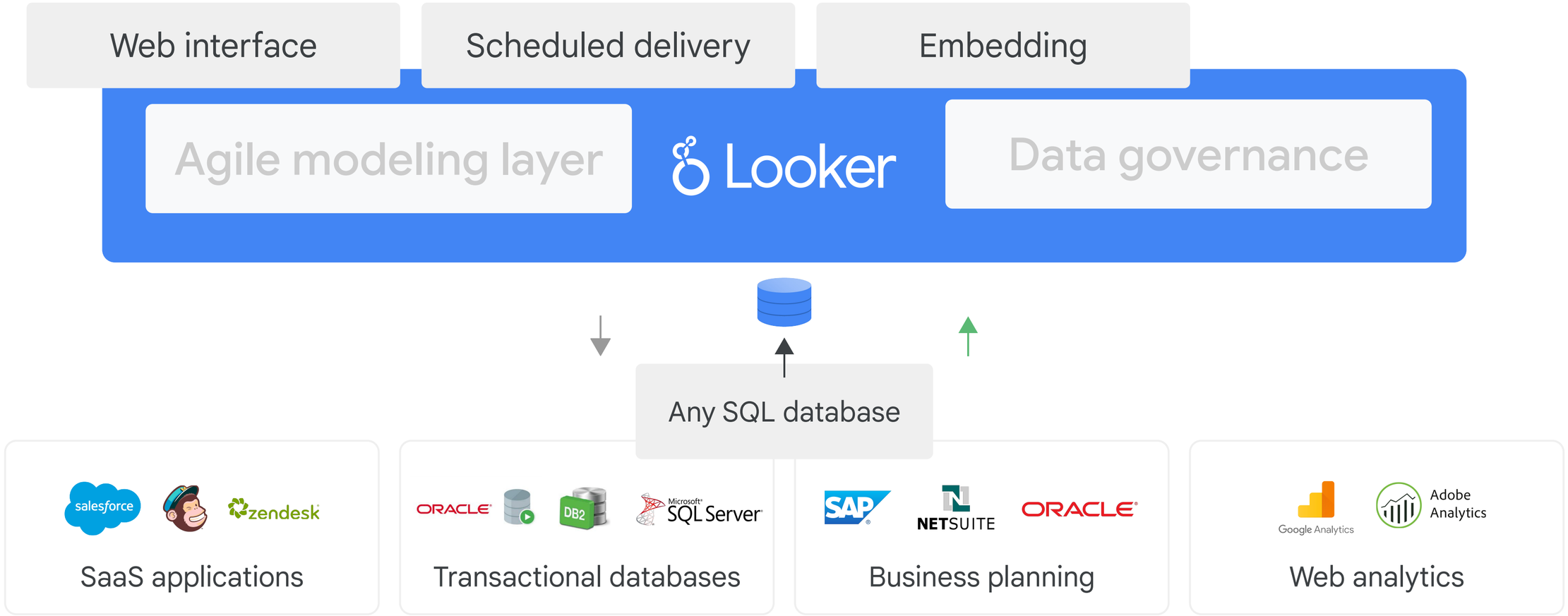
Architecture Looker's user interface



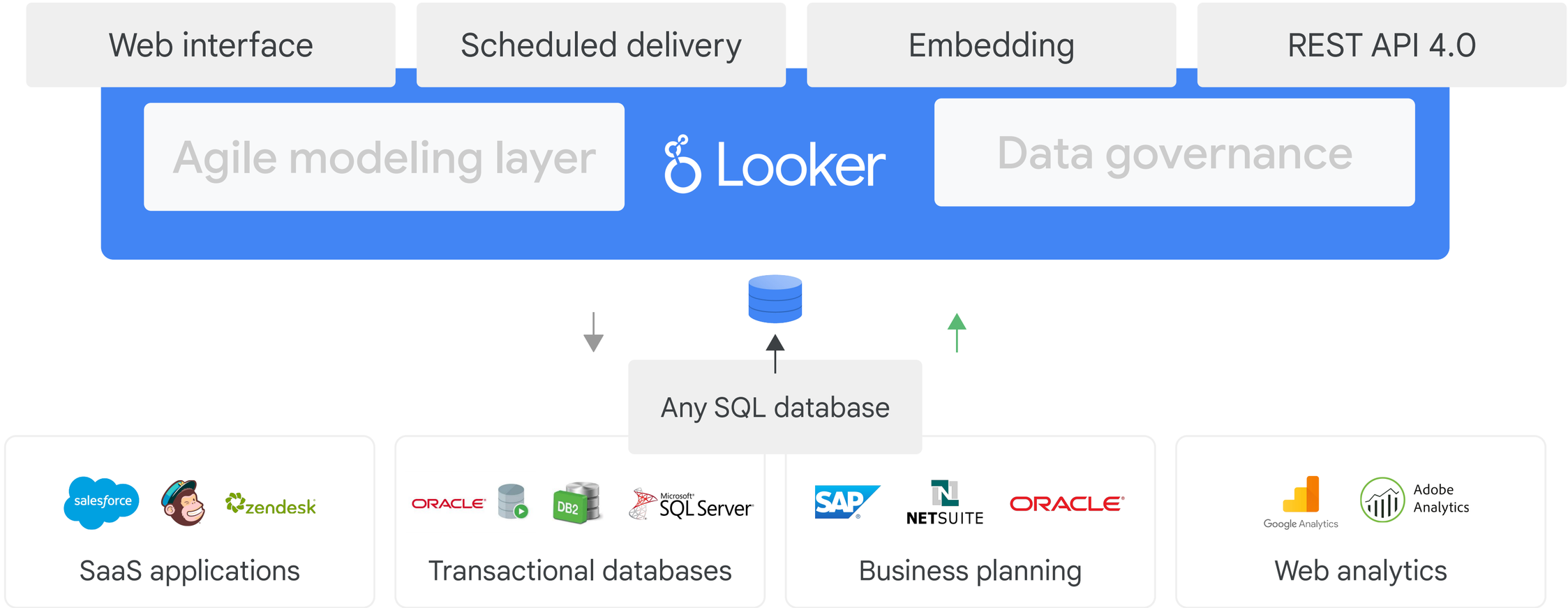
Architecture: Looker's data scheduling and delivery



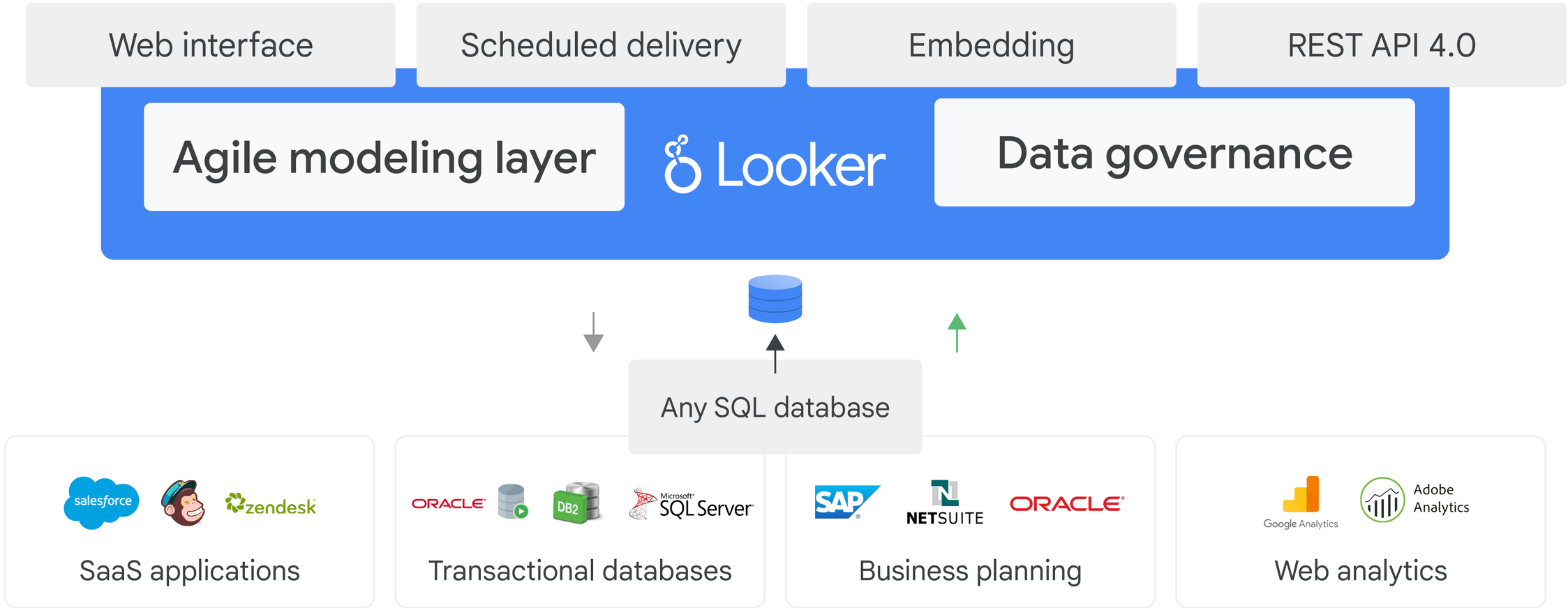
Architecture: Embed Looker content in your applications



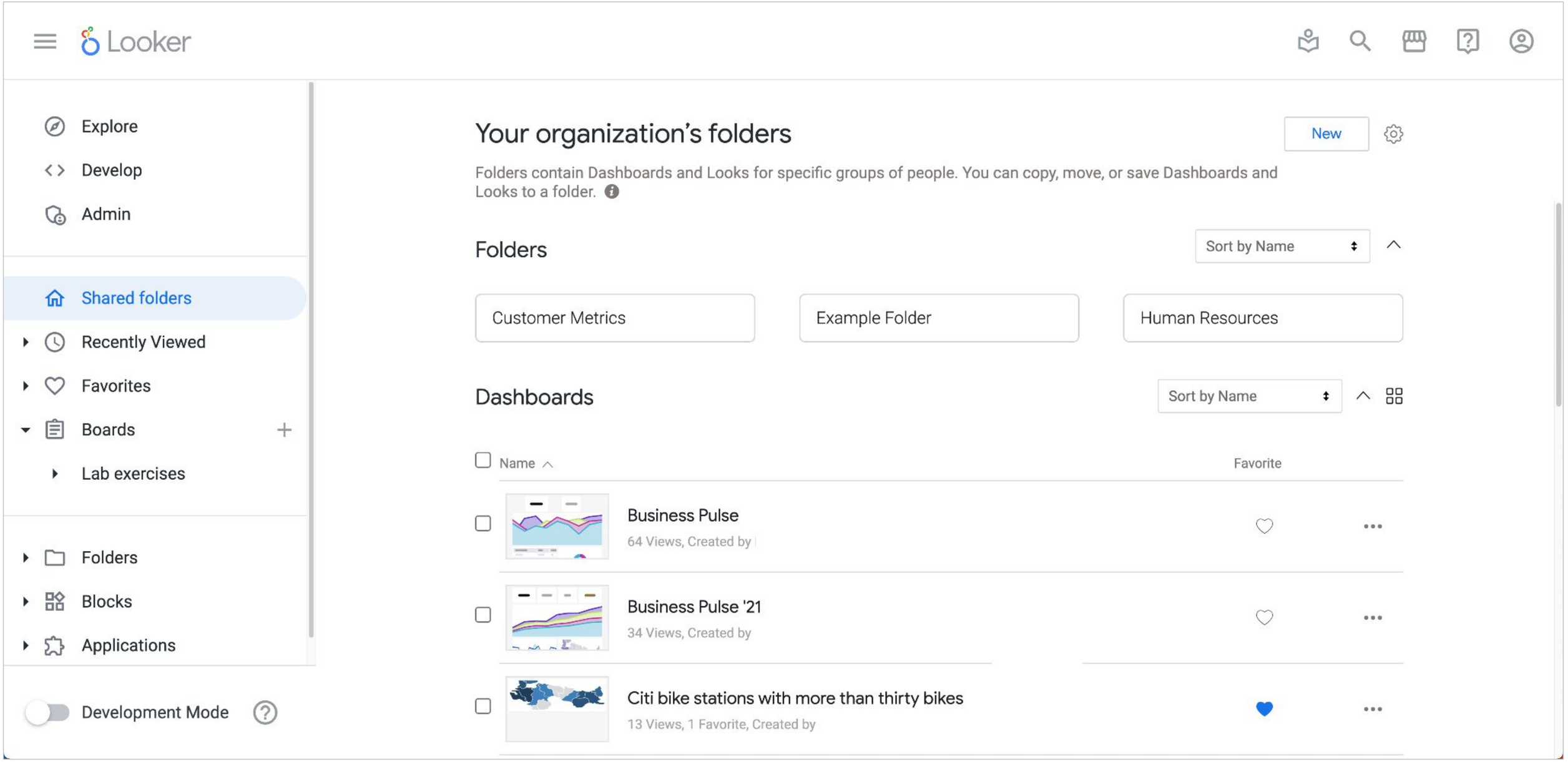
Architecture: The Looker API



Architecture: Looker's rich development framework

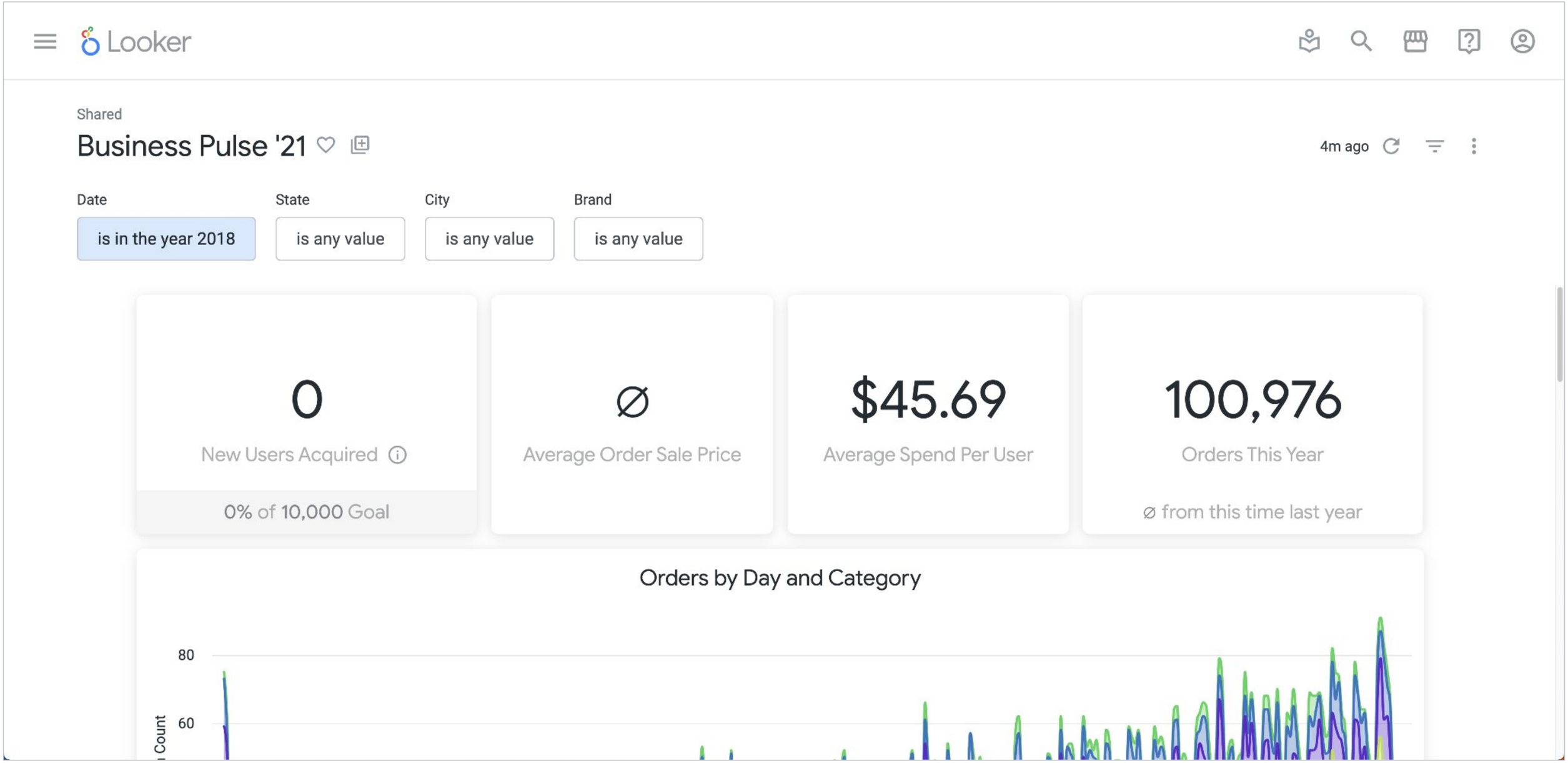


Default home screen



Examples are based on fictional data.

Dashboards



Examples are based on fictional data.

Many visualization types

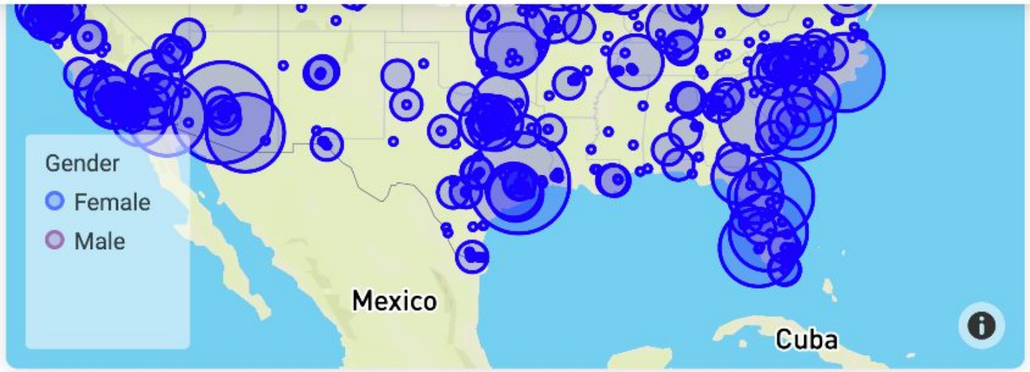
Looker

Shared

Business Pulse '21

2m ago

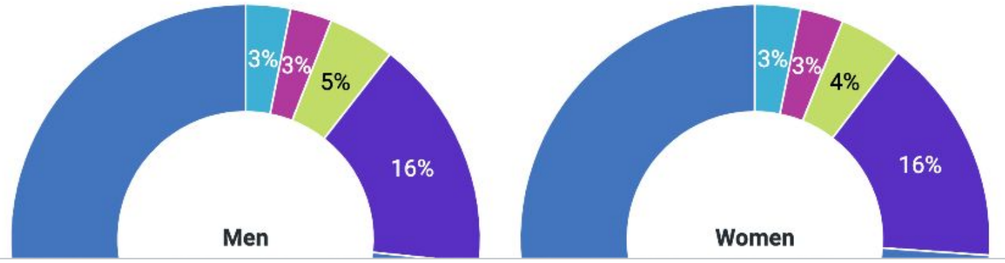
Date	State	City	Brand
is in the year 2018	is any value	is any value	is any value



Brand Sales

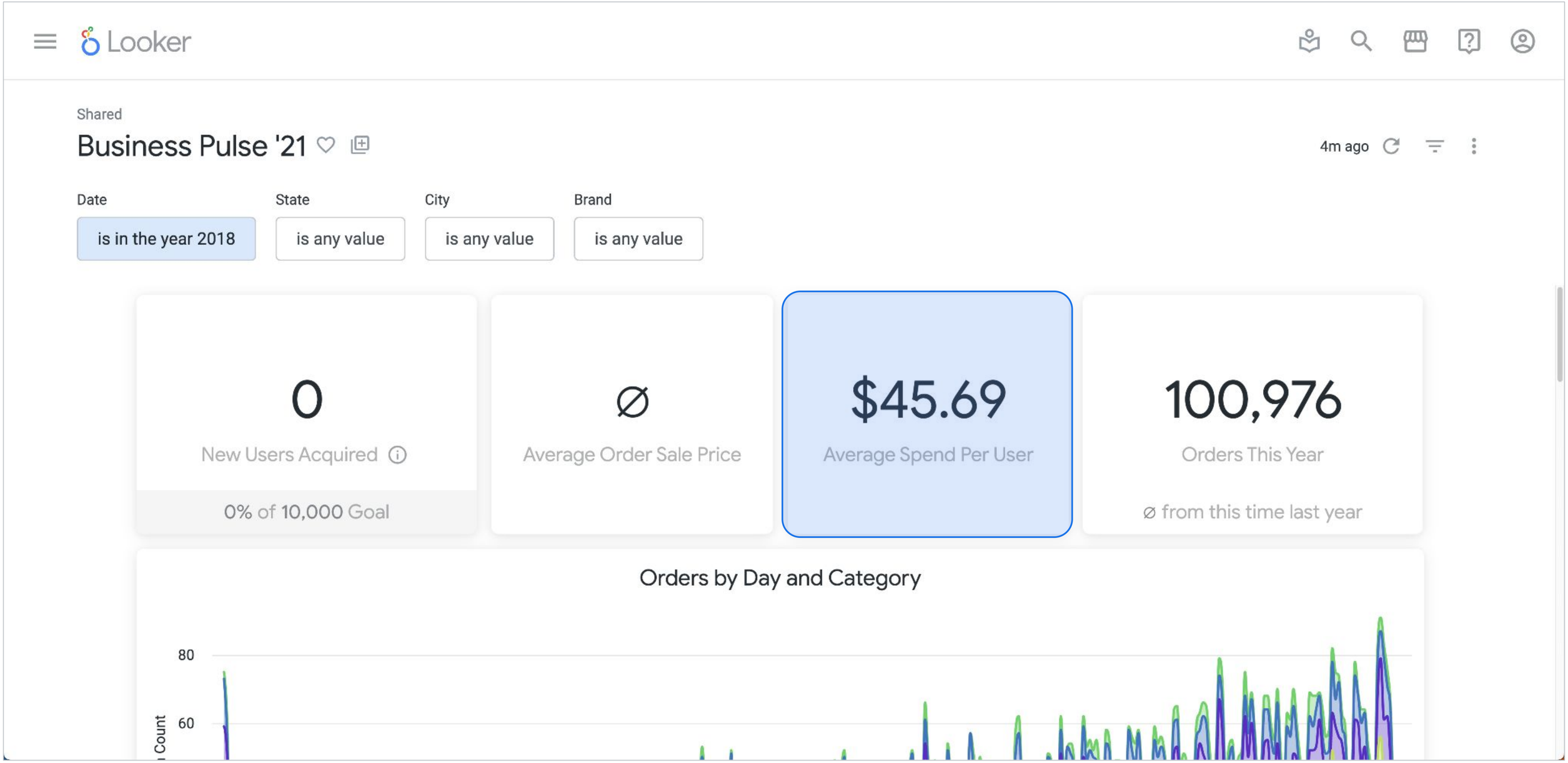
	Brand	Order Count	Average Spend per User	Users
1	Example Brand 1	1,413	$32.13	1,369
2	Example Brand 2	1,317	$32.33	1,288
3	Example Brand 92	971	$62.96	939
4	Example Brand 3	741	$44.05	730
5	Example Brand 4	682	$44.02	674

Marketing Channel by User Demographic



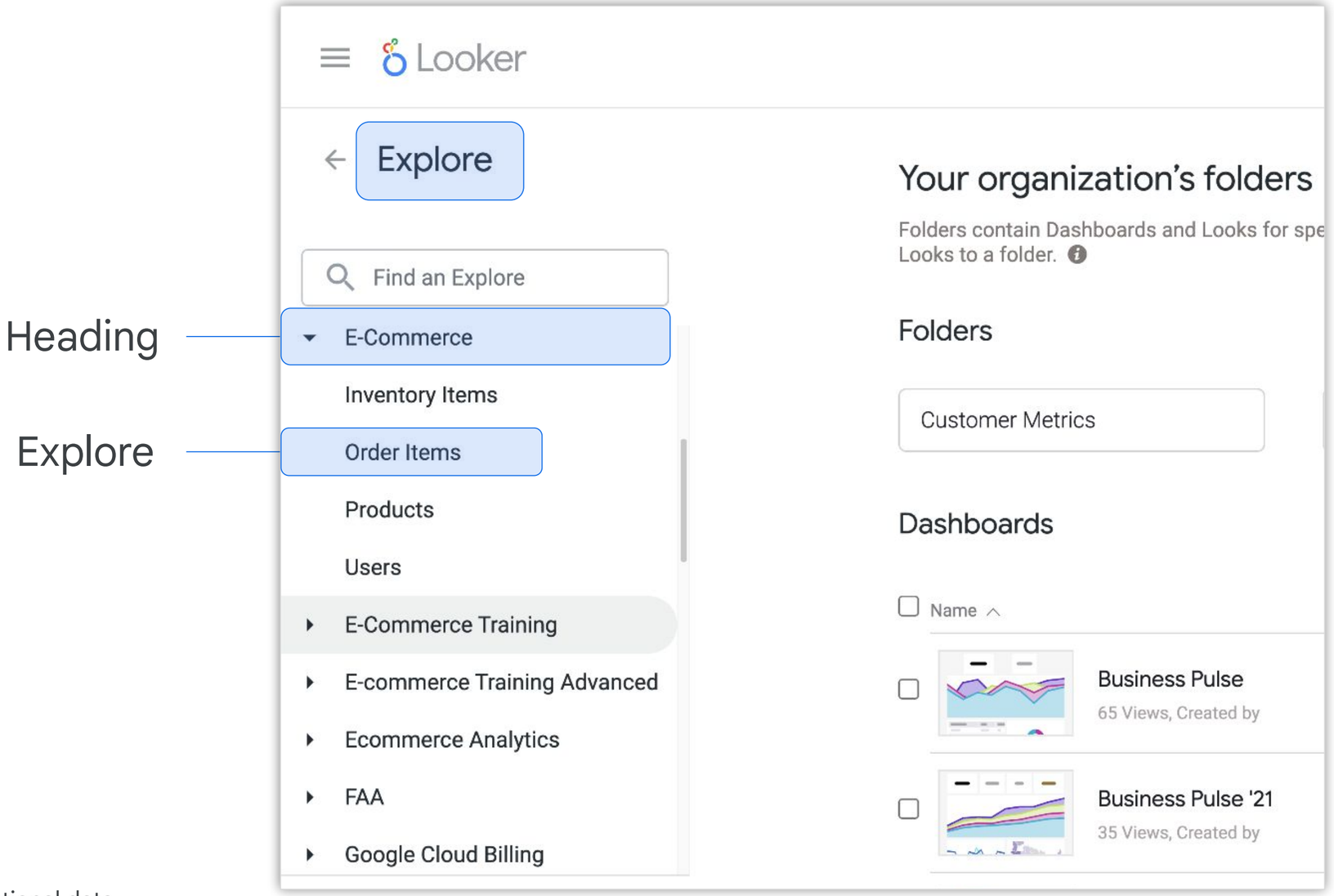
Examples are based on fictional data.

Drilling into dashboards



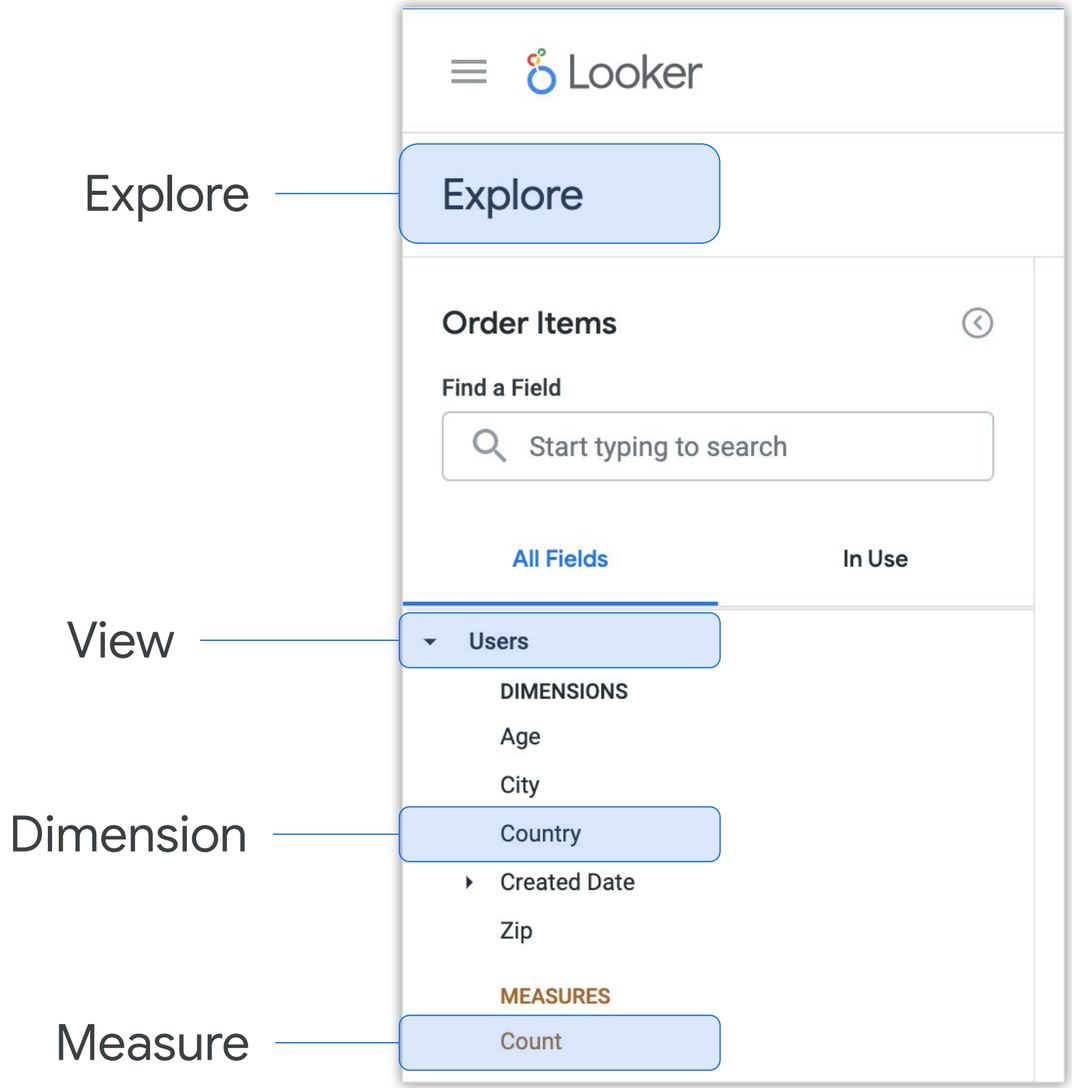
Examples are based on fictional data.

Key vocabulary



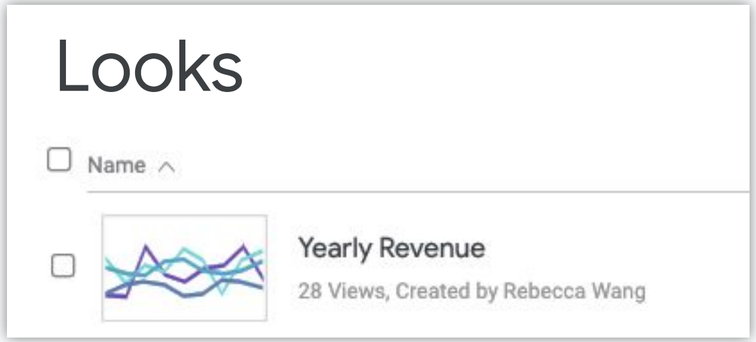
Examples are based on fictional data.

Key vocabulary

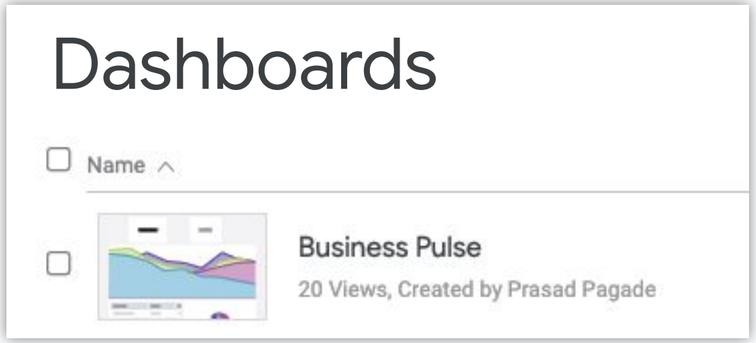


Examples are based on fictional data.

Key vocabulary



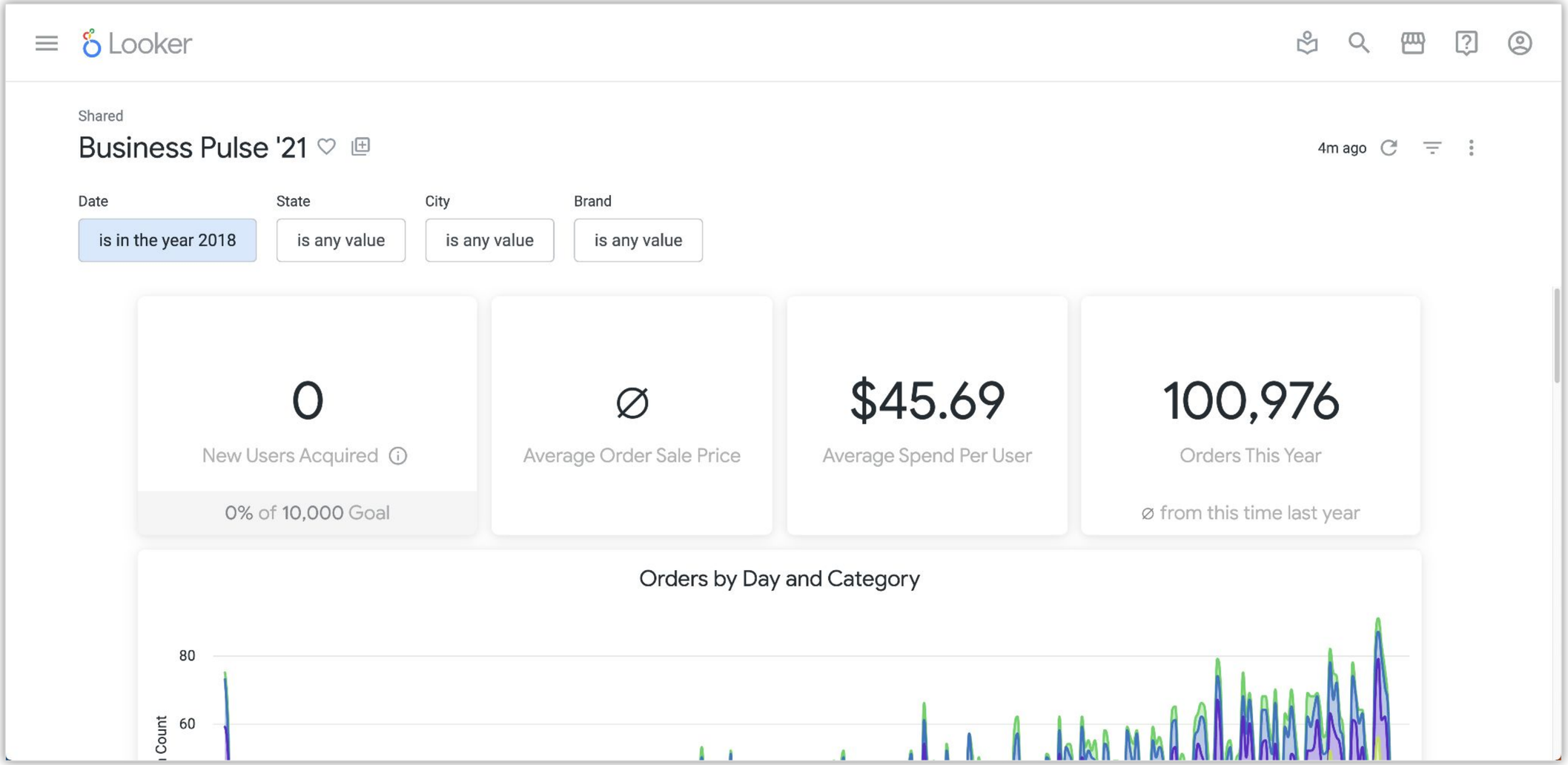
Looks are standalone reports or visualizations.



Dashboards can contain multiple visualizations referred to as tiles.

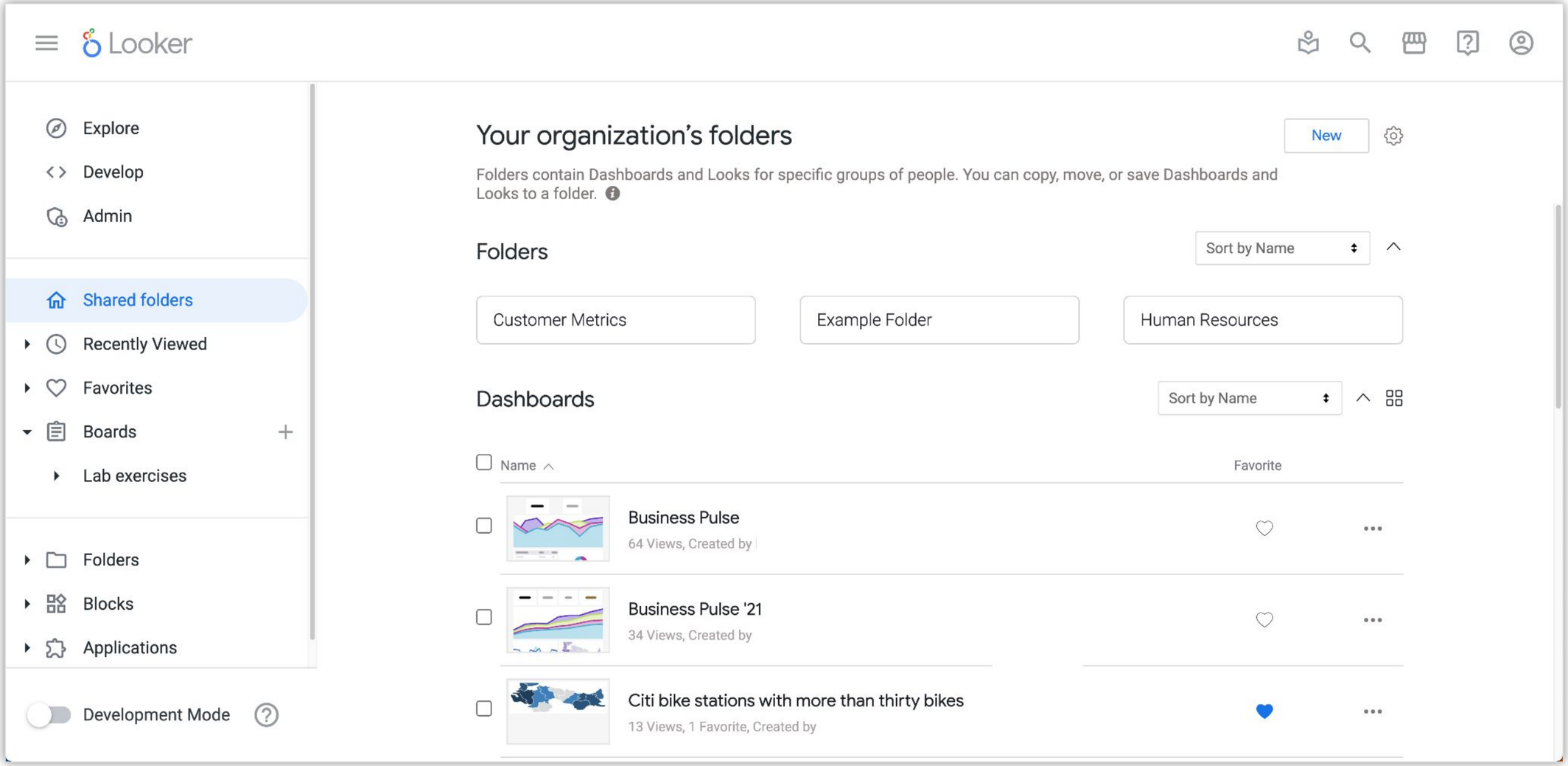
Examples are based on fictional data.

Dashboards



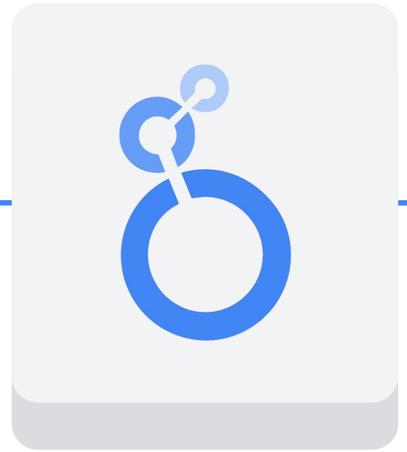
Examples are based on fictional data.

Shared folders



Examples are based on fictional data.

Looker versus Looker Studio



Looker

Platform for BI, data applications, and embedded analytics

Looker Studio

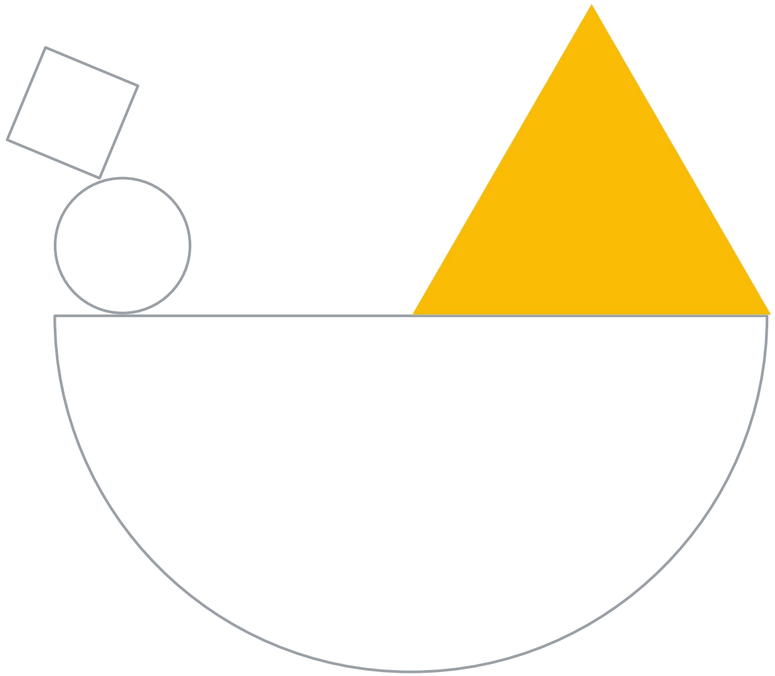
Interactive data suite for dashboarding, reporting, and analytics

Looker Studio Pro

- A paid version of Looker Studio
- Manage access to content at scale
- Team workspaces
- Set up organizational ownership of Looker Studio assets
- Multiple delivery schedules
- Get help with Cloud Customer Care

Looker Studio

Case Study



Case Study - II

A customer had this interesting business requirement...

- Capture data reading and updates events to know who, what, when, and where.
- Separation of who manages the data and who can read the data.
- Allocate costs appropriately; costs to read/process vs. costs to store.
- Prevent exfiltration of data to other Google Cloud projects and to external systems.

Data engineer case study 02:

We mapped that to technical requirements like this...

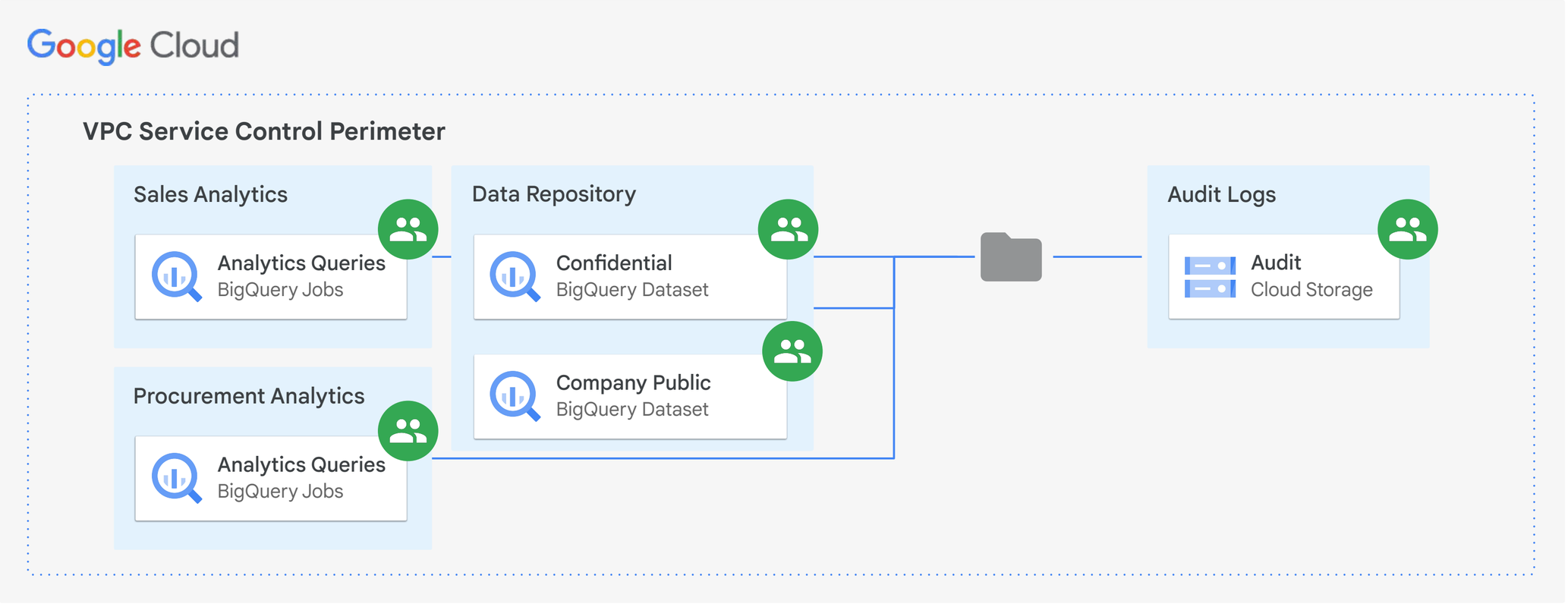
All access to data should be captured in audit logs.

All access to data should be managed via IAM.

Configure service perimeters with VPC service controls.

Data engineer case study 02:

And this is how we implemented that technical requirement



Case Study - III

Case study III

A customer had this interesting business requirement...

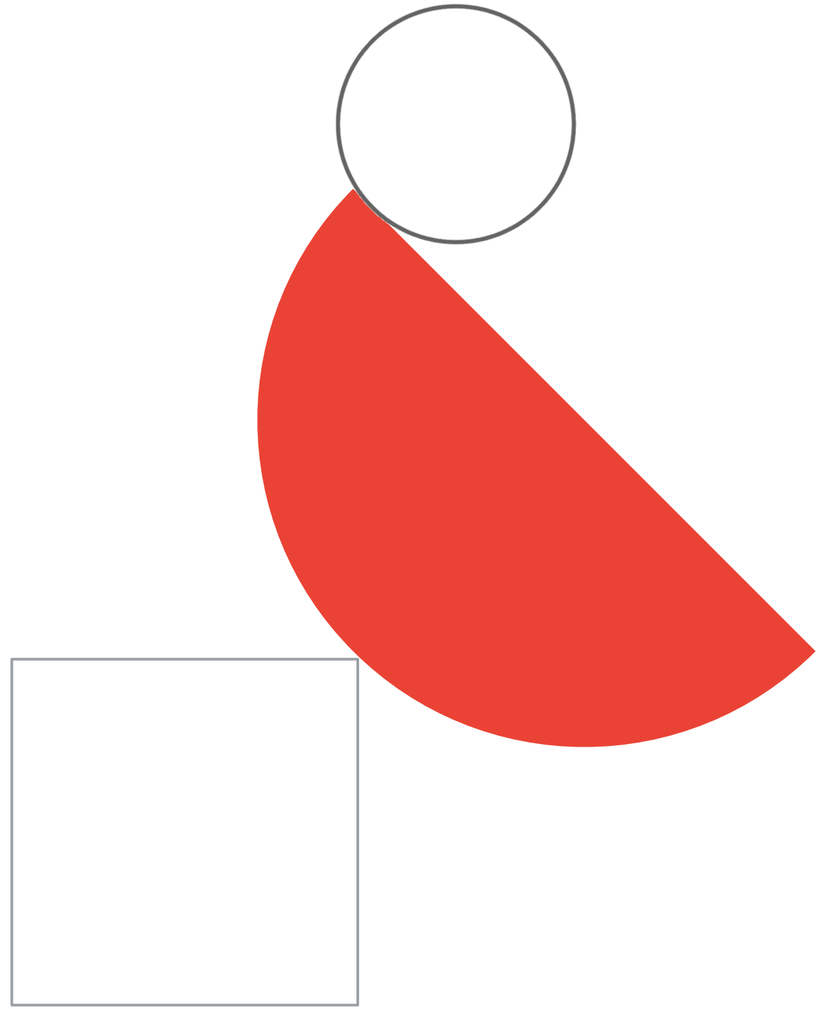
The overall business requirement was to migrate to Cloud a on-premise reporting solution aimed to produce daily reports to regulators.

The on-premises solution was coded using a ‘SQL-like language’ and it was run on a Hadoop cluster using Spark/MapReduce (leveraging proprietary third-party software).

The client wanted to optimize the target cloud solution to:

- Minimize changes to their processes and codebase.
- Leverage PaaS and native cloud solution (i.e., avoid third-party software).
- Significantly improve performance (from hours to minutes).

[optional] Links to useful materials



Optional materials 1

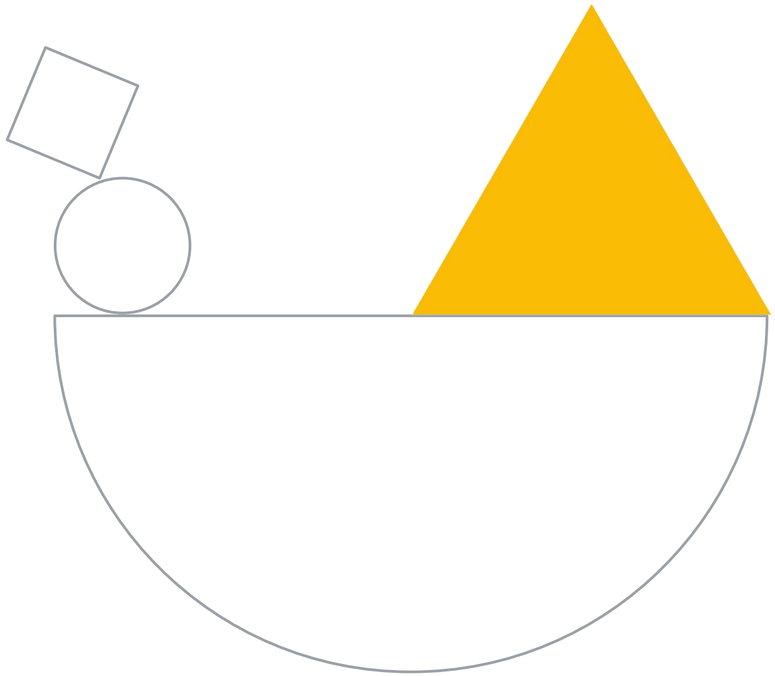
[READING]

- Analyze data with Looker Studio
- Materialized views
- Introduction to Analytics Hub and manage data exchanges

[VIDEOS]

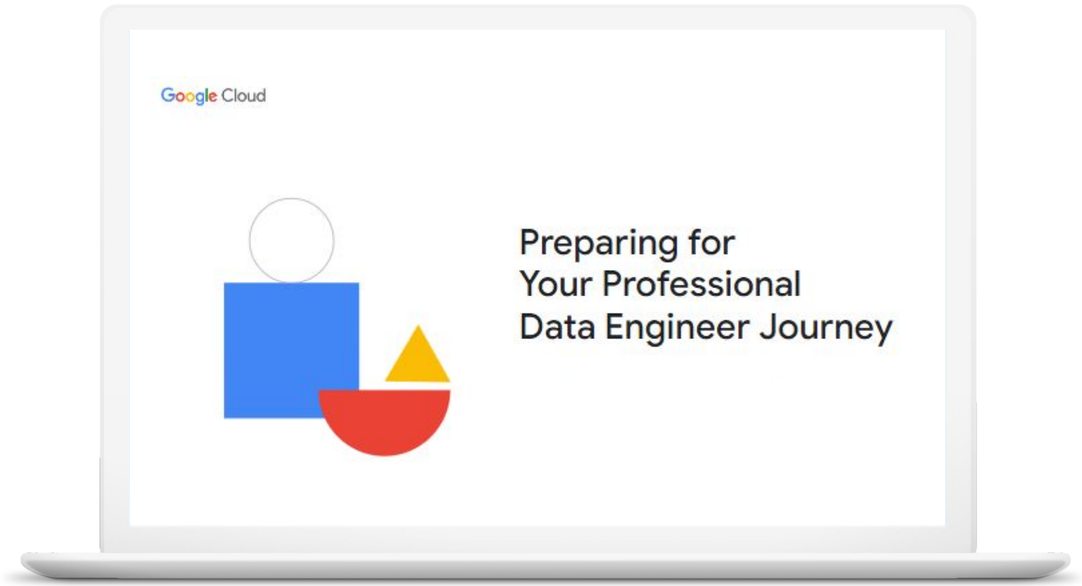
- Visualizing query results
- Scaling Interactive and Insightful Dashboards with Data Studio and BigQuery
- Analytics Hub
- Secure data exchanges and data sharing with Analytics Hub

Diagnostic questions



Your study plan:

Preparing and using data for analysis



4.1 Preparing data for visualization

4.2 Sharing data

4.3 Exploring and analyzing data

4.1 Preparing data for visualization

Considerations include:

- Connecting to tools
- Precalculating fields
- BigQuery materialized views (view logic)
- Determining granularity of time data
- Troubleshooting poor performing queries
- Identity and Access Management (IAM) and Cloud Data Loss Prevention (DLP)

4.1 Diagnostic Question 01 Discussion

Your company uses Google Workspace and your leadership team is familiar with its business apps and collaboration tools. They want a cost-effective solution that uses their existing knowledge to evaluate, analyze, filter, and visualize data that is stored in BigQuery.

What should you do to create a solution for the leadership team?

A. Create models in Looker.
B. Configure Connected Sheets.
C. Configure Tableau.
D. Configure Looker Studio.

4.1 Diagnostic Question 01 Discussion

Your company uses **Google Workspace** and your leadership team is familiar with its business apps and collaboration tools. They want a cost-effective solution that uses their **existing knowledge** to evaluate, analyze, filter, and visualize data that is stored in BigQuery.

What should you do to create a solution for the leadership team?

A. Create models in Looker.
B. Configure Connected Sheets.
C. Configure Tableau.
D. Configure Looker Studio.

4.1 Diagnostic Question 01 Discussion

Your company uses **Google Workspace** and your leadership team is familiar with its business apps and collaboration tools. They want a cost-effective solution that uses their **existing knowledge** to evaluate, analyze, filter, and visualize data that is stored in BigQuery.

What should you do to create a solution for the leadership team?

A. Create models in Looker.

B. Configure Connected Sheets.

C. Configure Tableau.

D. Configure Looker Studio.

4.1 Diagnostic Question 02 Discussion

You have data in PostgreSQL that was designed to reduce redundancy. You are transferring this data to BigQuery for analytics. The source data is hierarchical and frequently queried together. You need to design a BigQuery schema that is performant.

What should you do?

A. Use nested and repeated fields.

B. Retain the data in normalized form always.

C. Copy the primary tables and use federated queries for secondary tables.

D. Copy the normalized data into partitions.

4.1 Diagnostic Question 02 Discussion

You have data in PostgreSQL that was designed to reduce redundancy. You are transferring this data to **BigQuery** for analytics. The source data is **hierarchical and frequently queried together**. You need to design a BigQuery schema that is performant.

What should you do?

A. Use nested and repeated fields.

B. Retain the data in normalized form always.

C. Copy the primary tables and use federated queries for secondary tables.

D. Copy the normalized data into partitions.

4.1 Diagnostic Question 02 Discussion

You have data in PostgreSQL that was designed to reduce redundancy. You are transferring this data to BigQuery for analytics. The source data is hierarchical and frequently queried together. You need to design a BigQuery schema that is performant.

What should you do?

A. Use nested and repeated fields.

B. Retain the data in normalized form always.

C. Copy the primary tables and use federated queries for secondary tables.

D. Copy the normalized data into partitions.

4.1 Diagnostic Question 03 Discussion

You repeatedly run the same queries by joining multiple tables. The original tables change about ten times per day. You want an optimized querying approach.

Which feature should you use in Bigquery?

A. Views
B. Materialized views
C. Federated queries
D. Partitions

4.1 Diagnostic Question 03 Discussion

You repeatedly run the same queries by joining multiple tables. The original tables change about ten times per day. You want an **optimized querying** approach.

Which feature should you use in Bigquery?

A. Views
B. Materialized views
C. Federated queries
D. Partitions

4.1 Diagnostic Question 03 Discussion

You repeatedly run the same queries by joining multiple tables. The original tables change about ten times per day. You want an optimized querying approach.

A. Views
B. Materialized views
C. Federated queries
D. Partitions

Which feature should you use in Bigquery?

4.1 Diagnostic Question 04 Discussion

You have analytics data stored in BigQuery. You need an efficient way to compute values across a group of rows and return a single result for each row.

What should you do?

A. Use an aggregate function.
B. Use a UDF (user-defined function).
C. Use BigQuery ML.
D. Use a window function with an OVER clause.

4.1 Diagnostic Question 04 Discussion

You have analytics data stored in **BigQuery.** You need an efficient way to **compute values across a group of rows and return a single result for each row.**

What should you do?

A. Use an aggregate function.
B. Use a UDF (user-defined function).
C. Use BigQuery ML.
D. Use a window function with an OVER clause.

4.1 Diagnostic Question 04 Discussion

You have analytics data stored in **BigQuery**. You need an efficient way to **compute values across a group of rows and return a single result for each row.**

What should you do?

A. Use an aggregate function.
B. Use a UDF (user-defined function).
C. Use BigQuery ML.
D. Use a window function with an OVER clause.

4.1 Diagnostic Question 05 Discussion

You need to optimize the performance of queries in BigQuery. Your tables are not partitioned or clustered.

What optimization technique can you use?

A. Batch your updates and inserts.
B. Use the LIMIT clause to reduce the data read.
C. Filter data as late as possible.
D. Perform self-joins on data.

4.1 Diagnostic Question 05 Discussion

You need to **optimize the performance of queries** in BigQuery. Your tables are n**ot partitioned or clustered**.

What optimization technique can you use?

A. Batch your updates and inserts.
B. Use the LIMIT clause to reduce the data read.
C. Filter data as late as possible.
D. Perform self-joins on data.

4.1 Diagnostic Question 05 Discussion

You need to **optimize the performance of queries** in BigQuery. Your tables are **not partitioned or clustered**.

What optimization technique can you use?

A. Batch your updates and inserts.

B. Use the LIMIT clause to reduce the data read.

C. Filter data as late as possible.

D. Perform self-joins on data.

4.1 Preparing data for visualization

Courses

Google Cloud Big Data and Machine Learning Fundamentals

- Data Engineering for streaming data

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a data warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- Dataflow streaming features
- Advanced BigQuery functionality and performance

Serverless Data Processing with Dataflow: Develop Pipelines

- Windows, watermarks, and triggers

Skill Badges

Perform Foundational Data, ML, and AI Tasks in Google Cloud

Engineer Data with Google Cloud

Documentation

Introduction to analysis and business intelligence tools

Use nested and repeated fields

Introduction to materialized views

Window function calls

Optimize query computation

Optimize query computation

4.2 Sharing data

Considerations include:

- Defining rules to share data
- Publishing datasets
- Publishing reports and visualizations
- Analytics Hub

4.2 Diagnostic Question 06 Discussion

Your data in BigQuery has some columns that are extremely sensitive. You need to enable only some users to see certain columns.

What should you do?

A. Create a new dataset with the column's data.

B. Create a new table with the column's data.

C. Use policy tags.

D. Use Identity and Access Management (IAM) permissions.

4.2 Diagnostic Question 06 Discussion

Your data in BigQuery has some **columns** that are **extremely sensitive.** You need to enable **only some users** to see certain columns.

What should you do?

A. Create a new dataset with the column's data.

B. Create a new table with the column's data.

C. Use policy tags.

D. Use Identity and Access Management (IAM) permissions.

4.2 Diagnostic Question 06 Discussion

Your data in BigQuery has some **columns** that are **extremely sensitive.** You need to enable **only some users** to see certain columns.

What should you do?

A. Create a new dataset with the column's data.

B. Create a new table with the column's data.

C. Use policy tags.

D. Use Identity and Access Management (IAM) permissions.

4.2 Diagnostic Question 07 Discussion

Your business has collected industry-relevant data over many years. The processed data is useful for your partners and they are willing to pay for its usage. You need to ensure proper access control over the data.

What should you do?

A. Export the data to zip files and share it through Cloud Storage.

B. Host the data on Analytics Hub.

C. Export the data to persistent disks and share it through an FTP endpoint.

D. Host the data on Cloud SQL.

4.2 Diagnostic Question 07 Discussion

Your business has collected industry-relevant data over many years. The processed data is useful for your partners and they are **willing to pay** for its usage. You need to ensure **proper access control** over the data.

What should you do?

A. Export the data to zip files and share it through Cloud Storage.

B. Host the data on Analytics Hub.

C. Export the data to persistent disks and share it through an FTP endpoint.

D. Host the data on Cloud SQL.

4.2 Diagnostic Question 07 Discussion

Your business has collected industry-relevant data over many years. The processed data is useful for your partners and they are **willing to pay** for its usage. You need to ensure **proper access control** over the data.

What should you do?

A. Export the data to zip files and share it through Cloud Storage.

B. Host the data on Analytics Hub.

C. Export the data to persistent disks and share it through an FTP endpoint.

D. Host the data on Cloud SQL.

4.2 Diagnostic Question 08 Discussion

You have a complex set of data that comes from multiple sources. The analysts in your team need to analyze the data, visualize it, and publish reports to internal and external stakeholders. You need to make it easier for the analysts to work with the data by abstracting the multiple data sources.

A. Looker Studio
B. Connected Sheets
C. D3.js library
D. Looker

What tool do you recommend?

4.2 Diagnostic Question 08 Discussion

You have a complex set of data that comes from multiple sources. The analysts in your team need to **analyze the data, visualize it, and publish reports to internal and external stakeholders**. You need to make it **easier for the analysts to work with the data by abstracting the multiple data sources.**

What tool do you recommend?

A. Looker Studio
B. Connected Sheets
C. D3.js library
D. Looker

4.2 Diagnostic Question 08 Discussion

You have a complex set of data that comes from multiple sources. The analysts in your team need to **analyze the data, visualize it, and publish reports to internal and external stakeholders**. You need to make it **easier for the analysts to work with the data by abstracting the multiple data sources.**

What tool do you recommend?

A. Looker Studio
B. Connected Sheets
C. D3.js library
D. Looker

4.2 Sharing data

Courses

Google Cloud Big Data and Machine Learning Fundamentals

- Data Engineering for Streaming Data

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Introduction to Data Engineering

Building Batch Data Pipelines on Google Cloud

- Introduction to Building Batch Data Pipelines

Skill Badges

Data Catalog Fundamentals

Documentation

Introduction to column-level access control

Analytics Hub | Data Exchange and Data Sharing | Google Cloud

Introduction to Analytics Hub | BigQuery

Secure data exchanges and data sharing with Analytics Hub

Looker business intelligence platform embedded analytics

4.3 Exploring and analyzing data

Considerations include:

- Preparing data for feature engineering (training and serving machine learning models
- Conducting data discovery

4.3 Diagnostic Question 09 Discussion

You built machine learning (ML) models based on your own data. In production, the ML models are not giving satisfactory results. When you examine the data, it appears that the existing data is not sufficiently representing the business goals. You need to create a more accurate machine learning model.

What should you do?

A. Train the model with more of similar data.

B. Perform L2 regularization.

C. Perform feature engineering, and use domain knowledge to enhance the column data.

D. Train the model with the same data, but use more epochs.

4.3 Diagnostic Question 09 Discussion

You built machine learning (ML) models based on your own data. In production, the ML models are not giving satisfactory results. When you examine the data, it appears that the **existing data is not sufficiently representing the business goals.** You need to create a more accurate machine learning model.

What should you do?

A. Train the model with more of similar data.

B. Perform L2 regularization.

C. Perform feature engineering, and use domain knowledge to enhance the column data.

D. Train the model with the same data, but use more epochs.

4.3 Diagnostic Question 09 Discussion

You built machine learning (ML) models based on your own data. In production, the ML models are not giving satisfactory results. When you examine the data, it appears that the **existing data is not sufficiently representing the business goals.** You need to create a more accurate machine learning model.

What should you do?

A. Train the model with more of similar data.

B. Perform L2 regularization.

C. Perform feature engineering, and use domain knowledge to enhance the column data.

D. Train the model with the same data, but use more epochs.

4.3 Diagnostic Question 10 Discussion

You used Dataplex to create lakes and zones for your business data. However, some files are not being discovered.

What could be the issue?

A. You have an exclude pattern that matches the files.

B. You have scheduled discovery to run every hour.

C. The files are in ORC format.

D. The files are in Parquet format.

4.3 Diagnostic Question 10 Discussion

You used Dataplex to create lakes and zones for your business data. However, **some files are not being discovered.**

What could be the issue?

A. You have an exclude pattern that matches the files.
B. You have scheduled discovery to run every hour.
C. The files are in ORC format.
D. The files are in Parquet format.

4.3 Diagnostic Question 10 Discussion

You used Dataplex to create lakes and zones for your business data. However, **some files are not being discovered.**

What could be the issue?

A. You have an exclude pattern that matches the files.
B. You have scheduled discovery to run every hour.
C. The files are in ORC format.
D. The files are in Parquet format.

4.3 Exploring and analyzing data

Courses

[Google Cloud Big Data and Machine Learning Fundamentals]

- Big Data with BigQuery
- The machine learning workflow with Vertex A

[Modernizing Data Lakes and Data Warehouses on Google Cloud]

- Introduction to Data Engineering

[Building Batch Data Pipelines on Google Cloud]

- Introduction to building batch data pipelinesi

[Smart Analytics, Machine Learning, and AI on Google Cloud]

- Custom model building with SQL in BigQuery ML

Skill Badges

[Engineer Data with Google Cloud]

Documentation

[Use the BigQuery ML TRANSFORM clause for feature engineering | Google Cloud]

[Feature preprocessing overview | BigQuery | Google Cloud]

[Discover data | Dataplex | Google Cloud]

Knowledge check 1

You need to share inventory data from Cymbal Retail with a partner company that uses BigQuery to store and analyze its data. What tool can you use to securely and efficiently share the data?

A. Cloud Storage

B. Data Loss Prevention (DLP)

C. Data Catalog

D. Analytics Hub

Knowledge check 1

You need to share inventory data from Cymbal Retail with a partner company that uses BigQuery to store and analyze its data. What tool can you use to securely and efficiently share the data?

A. Cloud Storage

B. Data Loss Prevention (DLP)

C. Data Catalog

D. Analytics Hub

Knowledge check 2

Cymbal Retail has a team of ML engineers that builds and maintains machine learning models. As a Professional Data Engineer, how will you support this team?

A. Identify what type of data is required to build ML models.

B. Process and prepare existing data to enable feature engineering.

C. Finalize the type of machine learning model to use.

D. Keep on improving the machine learning model after initial deployment.

Knowledge check 2

Cymbal Retail has a team of ML engineers that builds and maintains machine learning models. As a Professional Data Engineer, how will you support this team?

A. Identify what type of data is required to build ML models.

B. Process and prepare existing data to enable feature engineering.

C. Finalize the type of machine learning model to use.

D. Keep on improving the machine learning model after initial deployment.