databricks

Data Analysis with Databricks SQL

databricks

Getting Started in the Course

Introductions

- Name
- SQL/Databricks Experience
- Professional Responsibilities
- Fun Personal Interest/Fact
- Expectations for the Course

Lesson goals

1. Get to know the instructor and other students
2. Describe the course objectives
3. Describe course structure
4. Give an overview of the technical environment

Course Objectives

- Import data and persist it in Databricks SQL as schemas, tables, and views
- Query data in Databricks SQL
- Use Databricks SQL to create visualizations and dashboards

Agenda

Module Name	Duration
Getting Started with Databricks SQL	1 hr 40 min
Basic SQL on Databricks SQL	2 hrs
Presenting Data Visually	2 hrs 20 min

- We will take 10 minute breaks at the beginning of every hour
- We will take a one hour break for lunch

Getting Started with Databricks SQL

Lesson Name	Duration
Getting Started with Databricks SQL	10 min
Navigating Databricks SQL Demo	20 min
Unity Catalog on Databricks SQL	10 min
Schemas, Tables, and Views on Databricks SQL Demo	20 min
Schemas, Tables, and Views on Databricks SQL Lab	20 min

Basic SQL on Databricks SQL

Lesson Name	Duration
Ingesting Data for Databricks SQL	10 min
Ingesting Data Demo	20 min
Ingesting Data Lab	20 min
Joins	10 min
Delta Commands in Databricks SQL Demo	20 min
Optional: Basic SQL Demo	30 min
Basic SQL Lab	30 min

Presenting Data Visually

Lesson Name	Duration
Data Visualizations	10 min
Data Visualizations and Dashboards Demo	30 min
Data Visualizations and Dashboards Lab	20 min
Notifying Stakeholders Demo	20 min
Notifying Stakeholders Lab	20 min
Final Lab Assignment Overview	10 min
Final Lab Assignment	30 min

Technical Environment Overview

The Databricks SQL workspace

- Everyone is in the same workspace
- Everyone has their own schema
- Everyone is using the same SQL warehouse
- Only the instructor has administrator privileges in the workspace
  - Only a select few tasks in this course require admin privileges
  - You will see these tasks in the slides in order to provide context
  - The demos and labs do not require admin privileges

Technical Environment Overview, cont.

Working through the course

- Slides
  - Some content is provided using slide decks
  - Take notes, if you wish
  - The slide decks will be provided to you
- Demonstrations
  - Work through these together with the instructor
- Labs
  - Reinforce your learning by accomplishing tasks on your own
  - Ask questions, as needed

Databricks Certified Data Analyst Associate

Certification helps you gain industry recognition, competitive differentiation, greater productivity, and results.

- This course helps you prepare for the Databricks **Certified Data Analyst Associate exam**
- Please see the Databricks Academy for additional prep materials

For more information visit:
databricks.com/learn/certification

Databricks Certified Data Analyst Associate

Become Certified

- Five certification pathways
  - Data Analyst
  - Business Leader
  - Data Engineer
  - Machine Learning Practitioner
  - Platform Administrator
- Recommended Self-Paced Courses
  - How to Ingest Data for Databricks SQL
  - How to Integrate BI Tools with Databricks SQL

databricks

Getting Started with Databricks SQL

Lesson goals

1. Describe what Databricks is.
2. Describe what Databricks SQL is.
3. Describe the benefits of Databricks SQL.
4. Warehouse Configuration

What is Databricks?

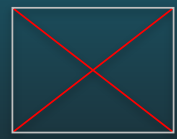
Databricks' vision is to enable data-driven innovation to all enterprises

databricks Lakehouse Platform

SIMPLE ⬡ OPEN ⬡ COLLABORATIVE

Data Engineering	BI & SQL Analytics	Real-time Data Applications	Data Science & Machine Learning

Data Management & Governance



Open Data Lake

aws Microsoft Azure Google Cloud

Visual ETL & Data Ingestion

Azure Data Factory, Fivetran, CONFLUENT, Informatica, talend, Qlik, StreamSets, precisely, Infoworks, wanDISCO, TRIFACTA, dbt, mongoDB., MATILLION

Business Intelligence

Power BI, Azure Synapse, Google BigQuery, TIBCO Spotfire, Amazon Redshift

Machine Learning

Amazon SageMaker, Azure Machine Learning, MathWorks, Google AI Platform, R Studio, H2O.ai

Data Providers

dun & bradstreet, SAFEGRAPH, experian, yipitDATA, arm INSIGHT

databricks

Lakehouse Platform

Microsoft Azure, aws, Google Cloud

Centralized Governance

AWS Glue, Informatica, Alation, PRIVACERA, IMMUTA, collibra

Top Consulting & SI Partners

Insight, NEUDESIC, PARIVEDA, accenture, avanade, wipro, Infosys, Tech Mahindra, GO DATA DRIVEN, ZEAL, AGILISIUM, BLUEGRANITE, Hitachi Solutions, ONICA, Capgemini, Cognizant, Deloitte., ELASTACLOUD, OPENVALUE, b.telligent, kumulus, BlueShift, deep_dive, Mindtree, slalom, TATA CONSULTANCY SERVICES, JUST ANALYTICS, Intellify, Versor

Open

Unify your data ecosystem with open source, standards, and formats

Partners

450+

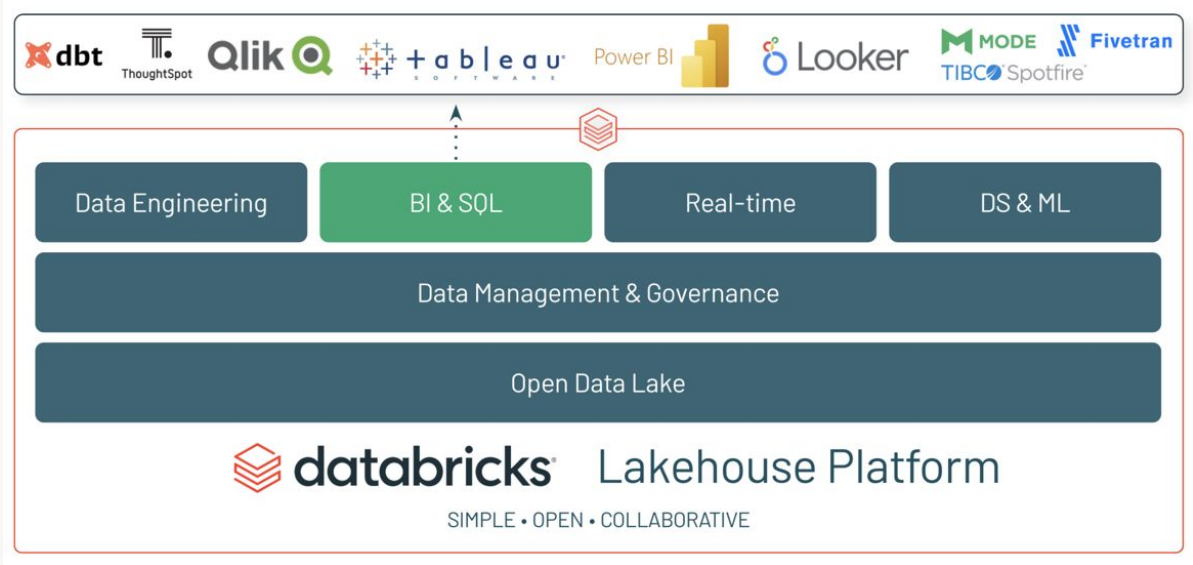
Across the data landscape

Databricks SQL

Databricks SQL

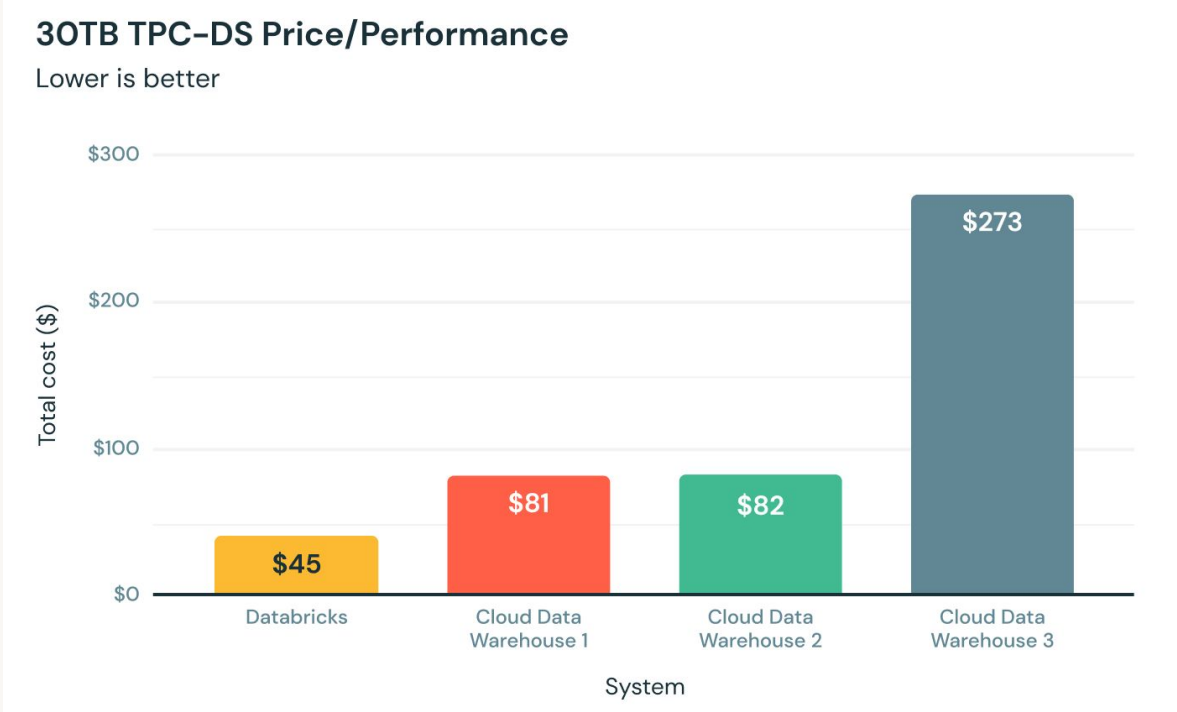
Delivering analytics on the freshest data with data warehouse performance and data lake economics

- Better price / performance than other cloud data warehouses
- Simplify discovery and sharing of new insights
- Connect to familiar BI tools, like Tableau or Power BI
- Simplified administration and governance



Better price / performance

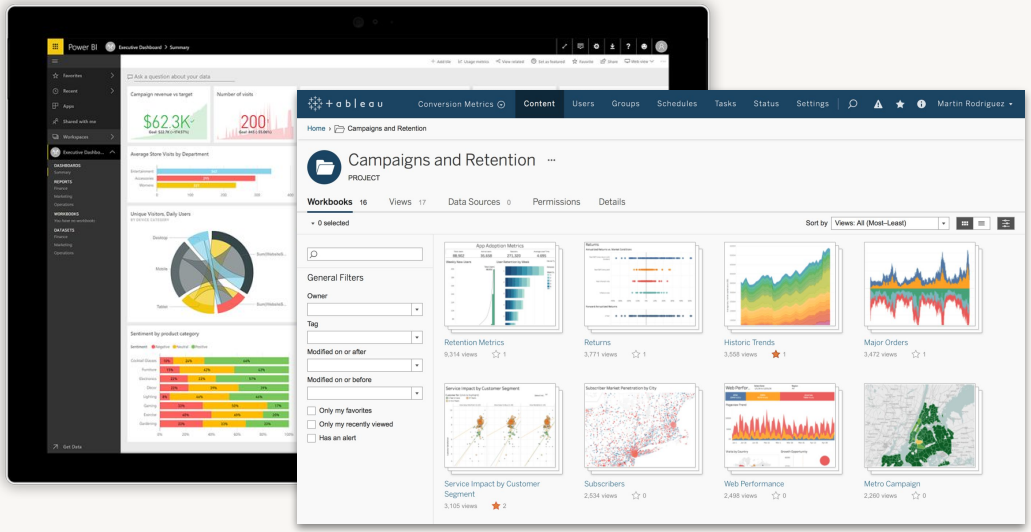
Run SQL queries on your lakehouse and analyze your freshest data with **up to 6x better price/performance** than traditional cloud data warehouses.



Source: Performance Benchmark with Barcelona Supercomputing Center

Better together | Broad integration with BI tools

Connect your preferred BI tools with optimized connectors that provide fast performance, low latency, and high user conconcurrency to your data lake for your existing BI tools.

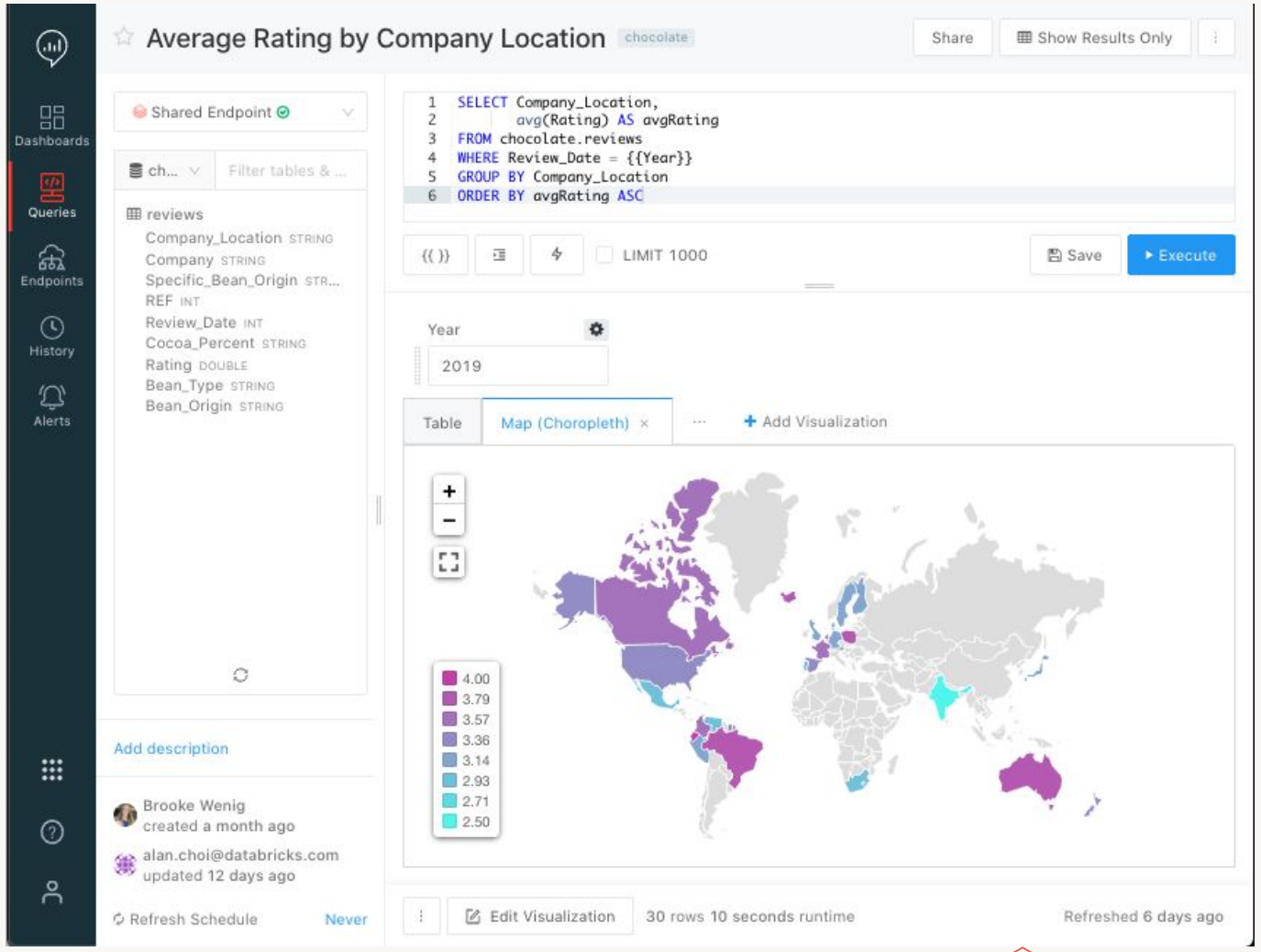


Coming soon:

Why use Databricks SQL?

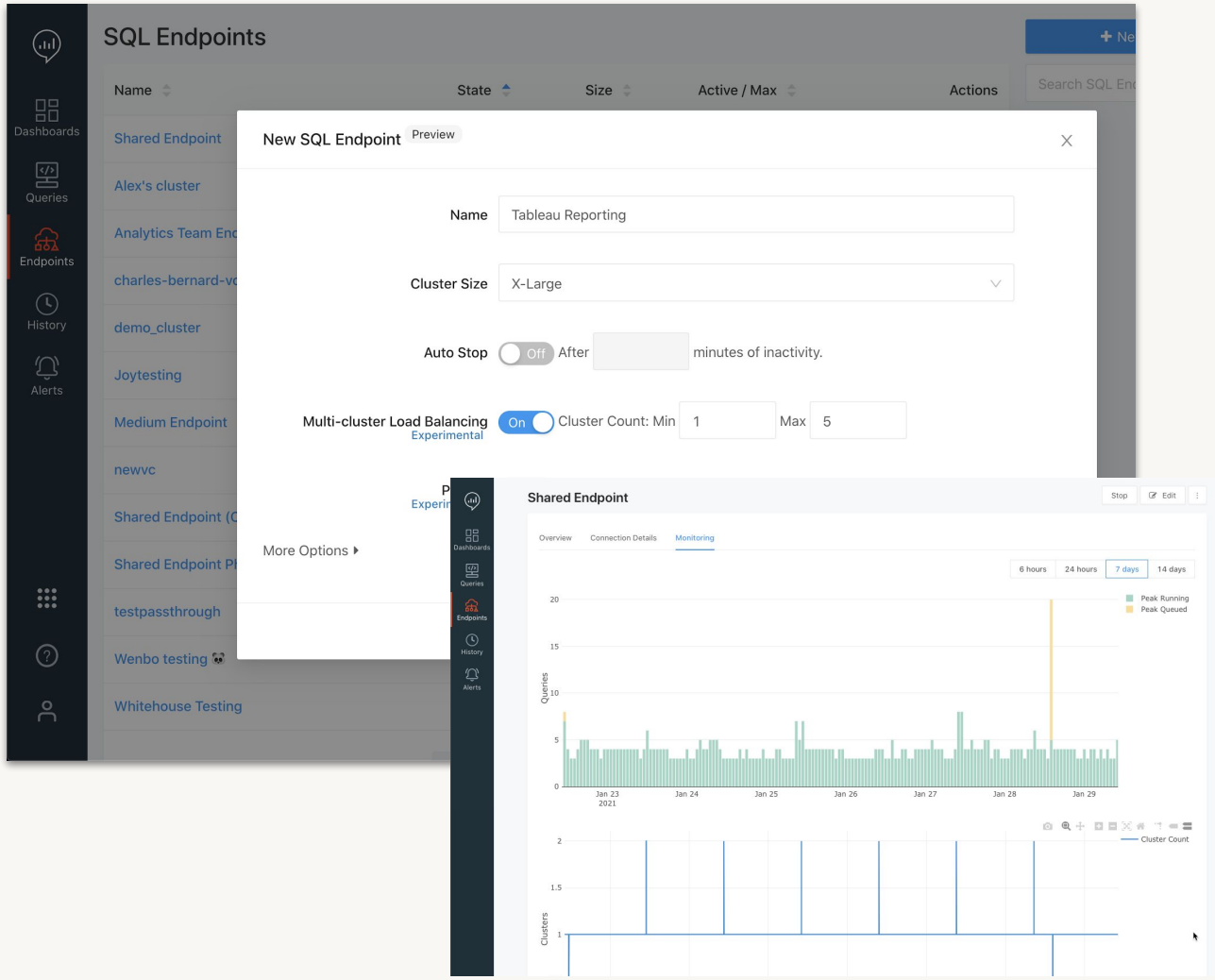
A new home for data analysts

Enable data analysts to quickly **perform ad-hoc and exploratory data analysis**, with a new SQL query editor, visualizations and dashboards. Automatic alerts can be triggered for critical changes, allowing to respond to business needs faster.



Simple administration and governance

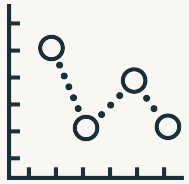
Quickly setup SQL / BI optimized compute with SQL warehouses. Databricks automatically determines instance types and configuration for the best price/performance. Then, easily manage usage, perform quick auditing, and troubleshooting with query history.



Use Cases

Connect existing BI tools to one source of truth for all your data

Maximize existing investments by connecting your preferred BI tools to your data lake with Databricks SQL Warehouses. Re-engineered and optimized connectors ensure fast performance, low latency, and high user concurrency to your data lake. Now analysts can use the best tool for the job on one single source of truth for your data while minimizing more ETL and data silos.



Collaboratively explore the latest and freshest data

Respond to business needs faster with a self-served experience designed for every analysts in your organization. Databricks SQL Analytics provides a simple and secure access to data, ability to create or reuse SQL queries to analyze the data that sits directly on your data lake, and quickly mock-up and iterate on visualizations and dashboards that fit best the business.

Build data-enhanced applications

Build rich and custom data enhanced applications for your own organization or your customers. Benefit from the ease of connectivity, management, and better price / performance of Databricks SQL Analytics to simplify development of data-enhanced applications at scale, all served from your data lake.

Warehouse Configuration

Warehouse Configuration

AWS

Cluster size	Driver size	Worker count
2X-Small	i3.2xlarge	1
X-Small	i3.2xlarge	2
Small	i3.4xlarge	4
Medium	i3.8xlarge	8
Large	i3.8xlarge	16
X-Large	i3.16xlarge	32
2X-Large	i3.16xlarge	64
3X-Large	i3.16xlarge	128
4X-Large	i3.16xlarge	256

Azure

Cluster size	Driver size	Worker count
2X-Small	Standard_E8ds_v4	1
X-Small	Standard_E8ds_v4	2
Small	Standard_E16ds_v4	4
Medium	Standard_E32ds_v4	8
Large	Standard_E32ds_v4	16
X-Large	Standard_E64ds_v4	32
2X-Large	Standard_E64ds_v4	64
3X-Large	Standard_E64ds_v4	128
4X-Large	Standard_E64ds_v4	256

Warehouse Configuration

In the course, SQL Warehouses have the following settings

- Cluster size – 2X-Small
- Scaling – Min: 1, Max 1
- Auto-stop – After two hours

01-2 – DEMO: NAVIGATING DATABRICKS SQL

databricks

Unity Catalog on Databricks SQL

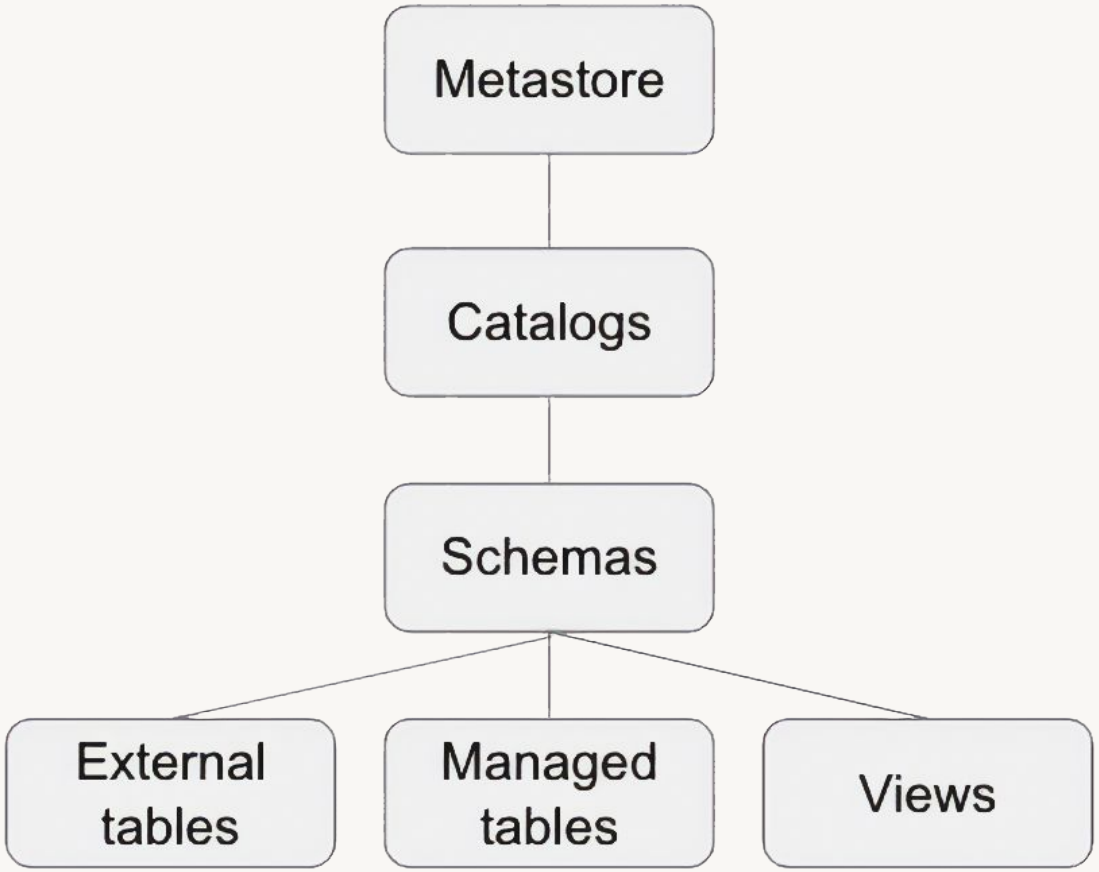
Lesson goals

1. Describe the object model in Unity Catalog.

2. Write queries using three-level namespace notation.

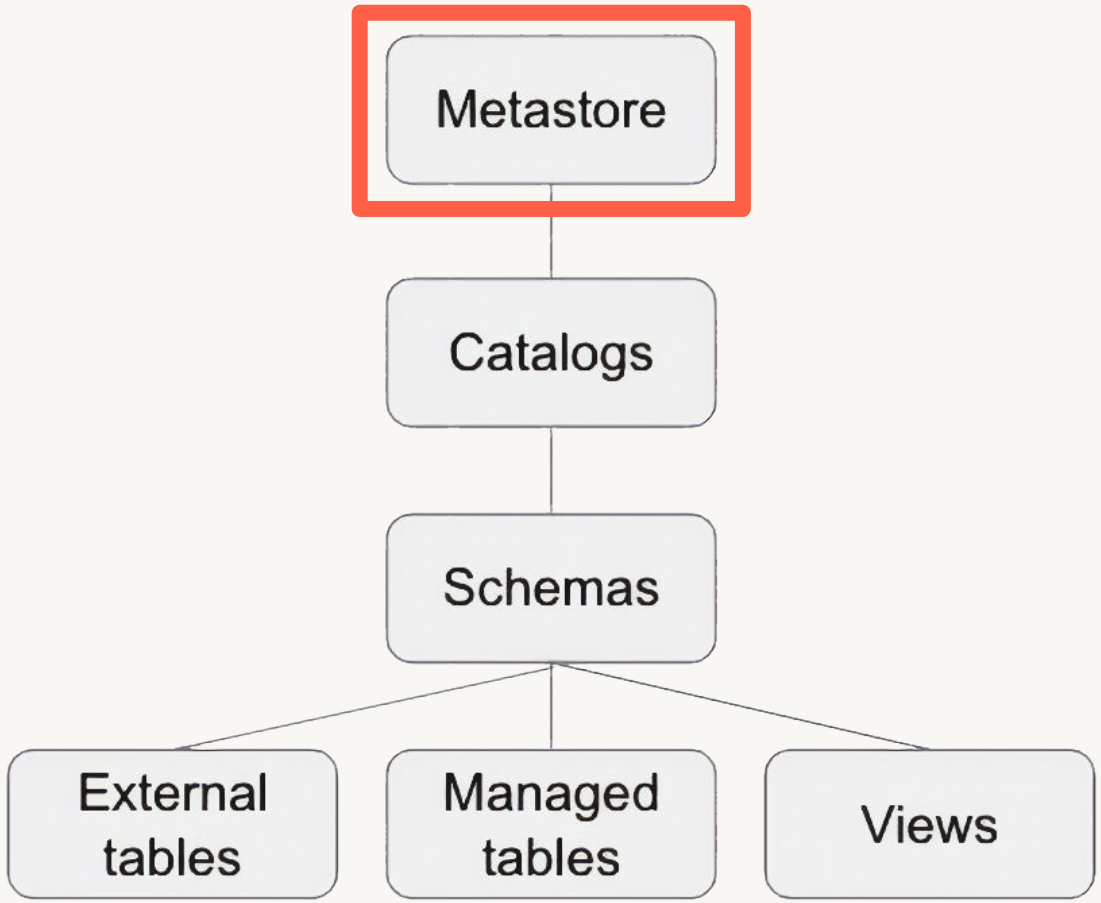
Unity Catalog Object Model

Object Model



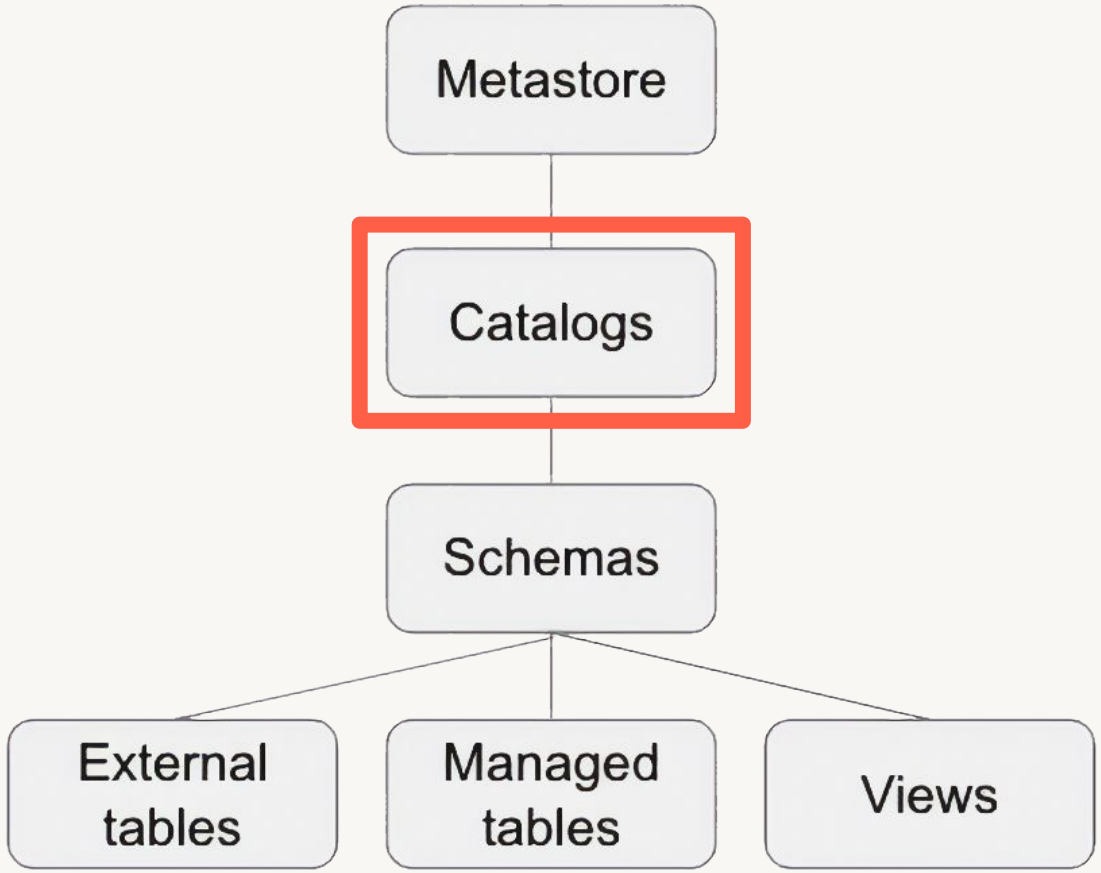
Metastore

- Stores data assets
- Permissions
- Created with default storage location (external object store)
- Metastore Admin



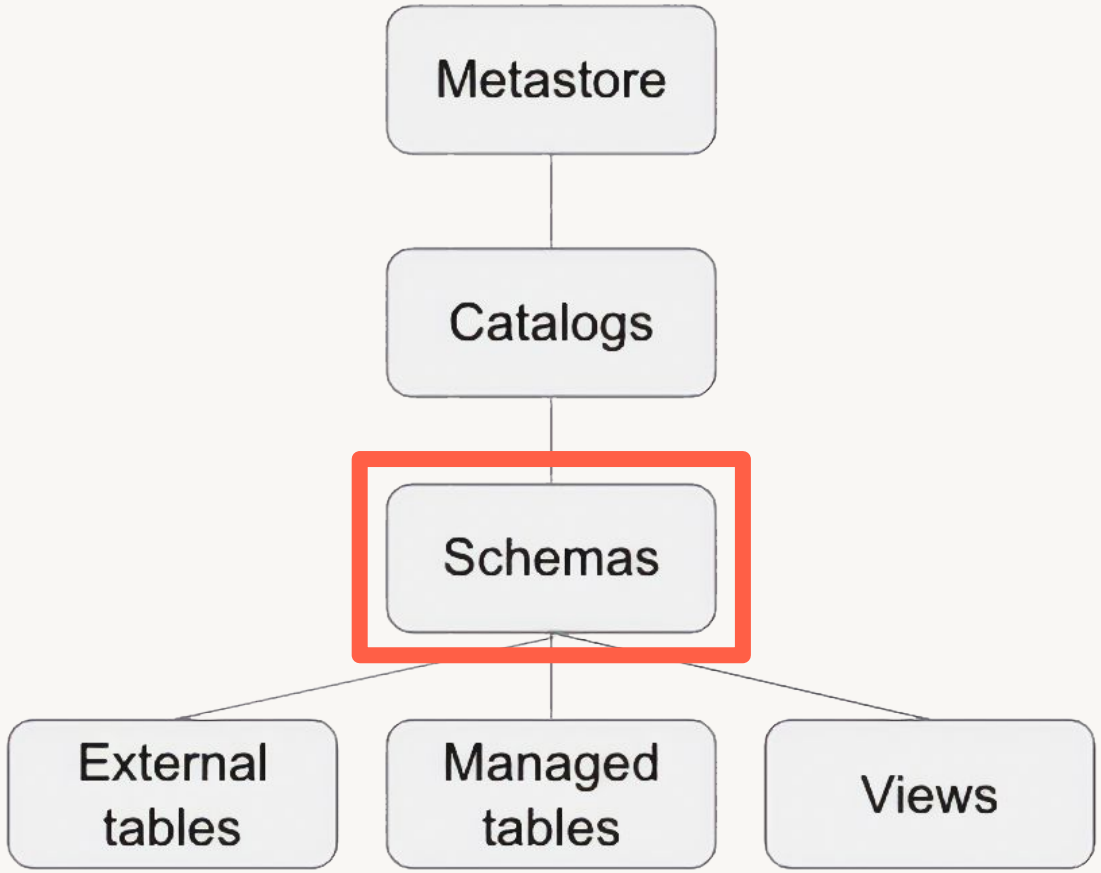
Catalog

- First level of organization
- Users can see all catalogs where `USAGE` is granted



Schema

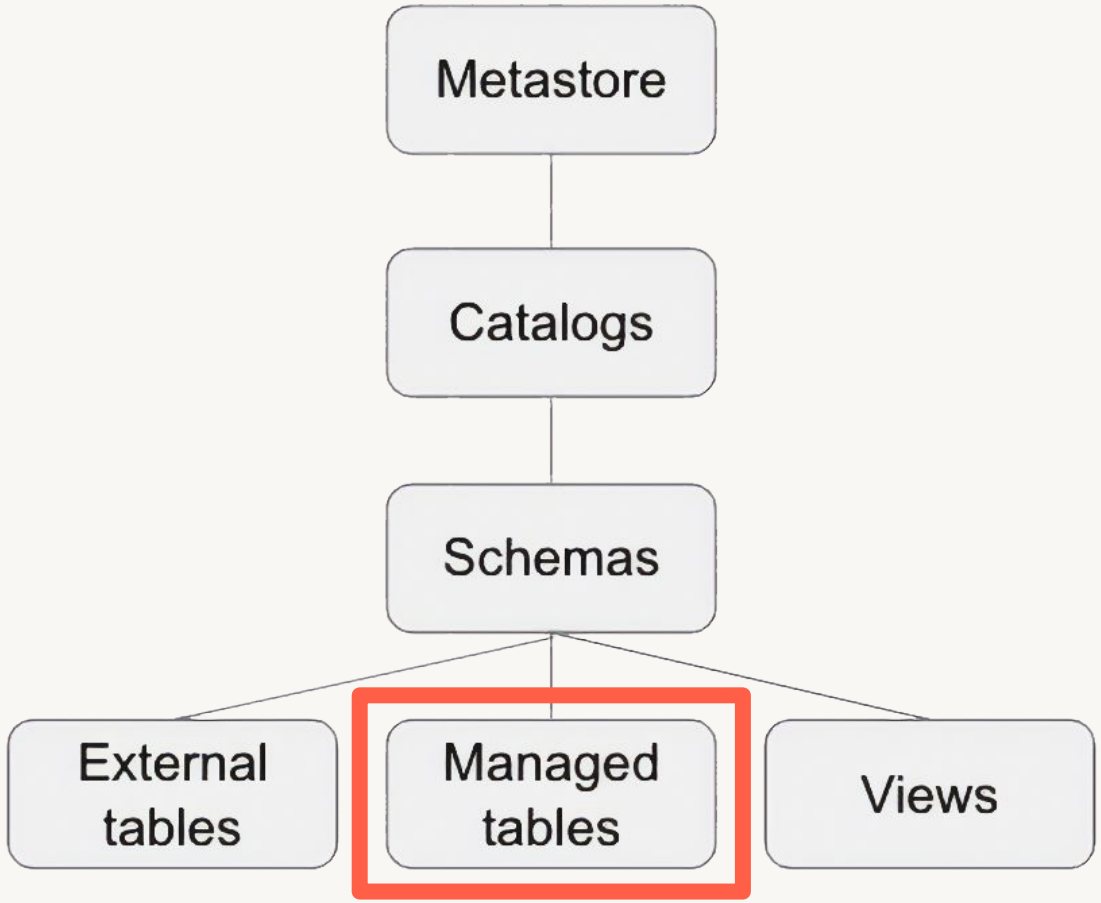
- aka, Database
- Second level of organization
- Users can see all schemas where `USAGE` is granted on both the schema and the catalog



Managed Table

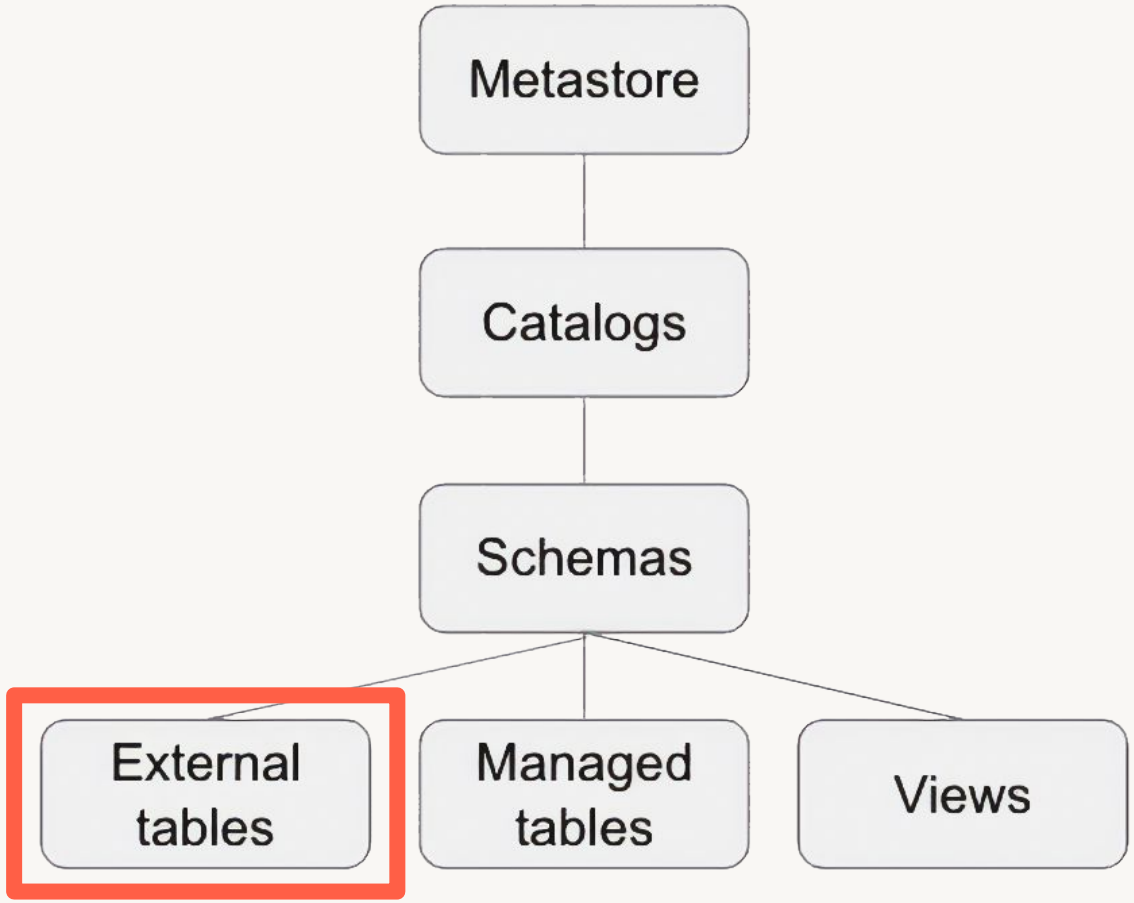
- Third level of organization
- Supported format: Delta
- Data is written to a new directory in the metastore's default location
- Created using `CREATE TABLE` statement with no `LOCATION` clause
- Example:

```
CREATE TABLE table1 …
```



External Table

- Third level of organization
- Data stored in a location outside the managed storage location
- `DROP TABLE` does not delete data
- Can easily clone a table to a new schema or table name without moving data
- Supported formats:
  - Delta, csv, json, avro, parquet, orc, text



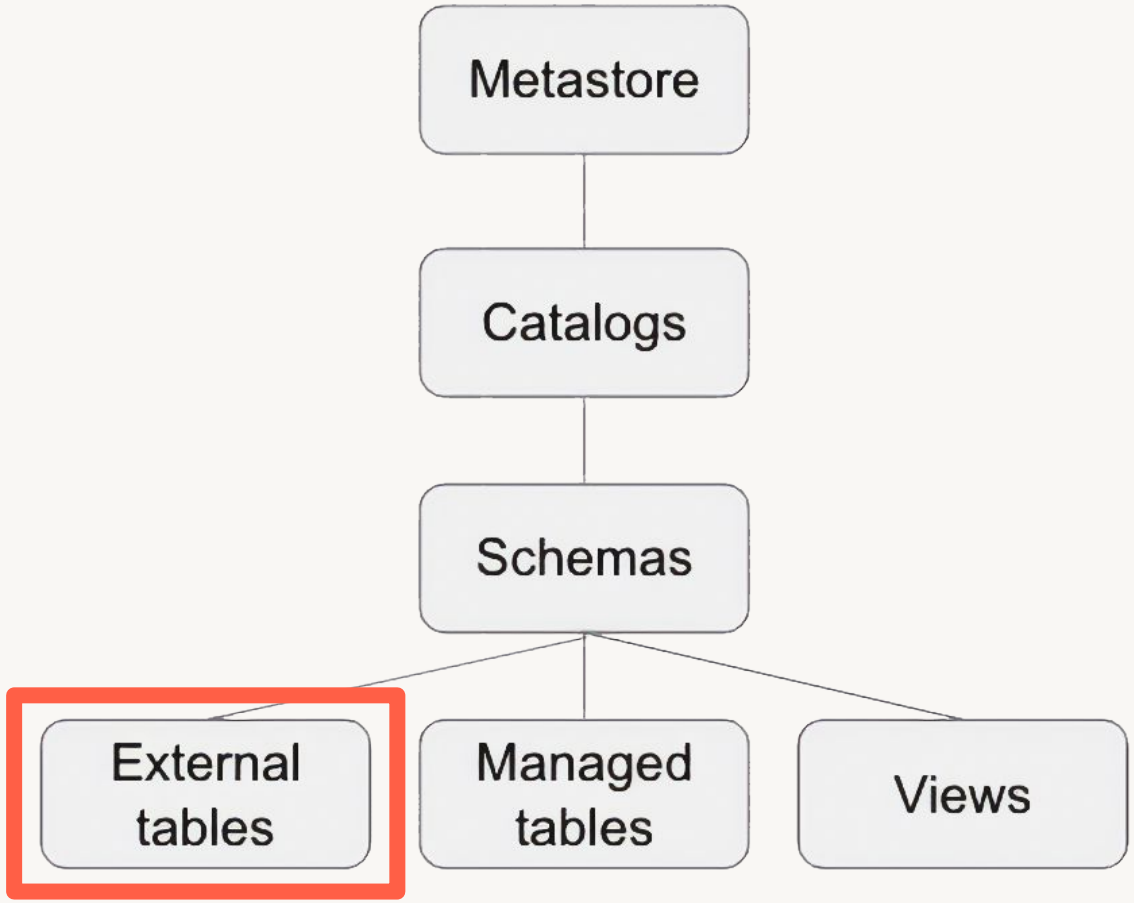
Creating External Tables

- Two credential types:
  - Storage Credential or External Location
- Use the LOCATION clause
- Example using External Location only

```
CREATE TABLE table2
  LOCATION 's3://<bucket_path>/<table_directory>'
...
```

- Example using Storage Credential

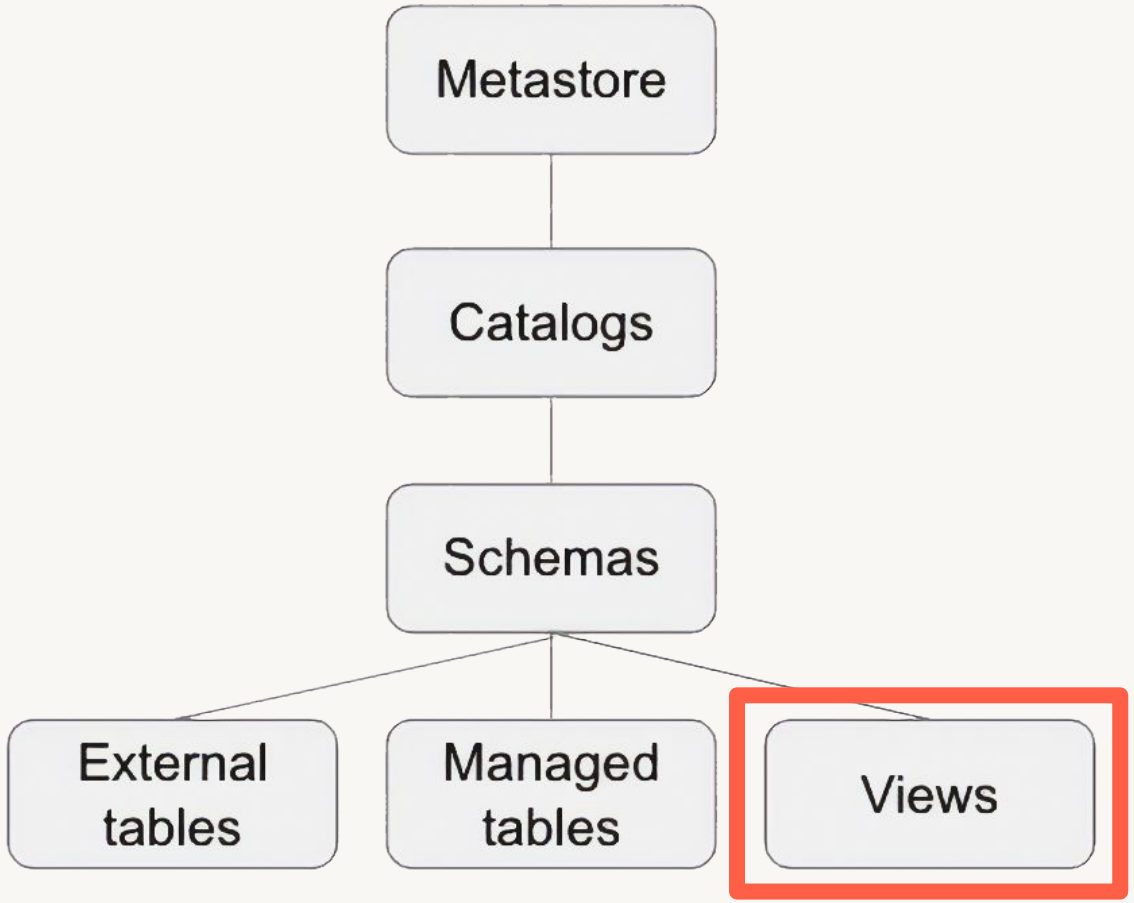
```
CREATE TABLE table2
    LOCATION 's3://<bucket_path>/<table_directory>'
    ...
    WITH CREDENTIAL <credential-name>;
```



View

- Third level of organization
- Can be composed from tables and views in multiple schemas or catalogs
- Created using `CREATE VIEW`:

```
CREATE VIEW view1 AS
  SELECT column1, column2
    FROM table1 ...
```



Three-Level Namespace Notation

- Data objects must be specified with three elements, depending on granularity required: Catalog, Schema, and Table
- Example:

```
CREATE TABLE main.default.department
 (
    deptcode   INT,
    deptname  STRING,
    location  STRING
 );
```

- Or, with a `USE` statement:

```
USE main.default;
SELECT * FROM department;
```

01-4 – DEMO: SCHEMAS, TABLES, AND VIEWS ON DATABRICKS SQL

01-5 – LAB: SCHEMAS, TABLES, AND VIEWS ON DATABRICKS SQL

databricks

Ingesting Data for Databricks SQL

Lesson goals

1. Describe how to connect Databricks SQL to an object store
2. Explain how Partner Connect can be used to ingest data
3. Provide proper data access privileges to users

Ingesting Existing Data

- Databricks SQL can ingest Parquet, JSON, CSV, Delta, and more
  - Individual file
  - Full directory of files of a single type
- Example (Azure Databricks):

```
CREATE TABLE table1 LOCATION

    'wasbs://[account].blob.core.windows.net/[container]/[path/]'
```

Partner Connect

- Connect to Databricks partners
- Data ingestion, preparation, BI, and visualization tools
- Data Ingestion:
  - Fivetran
  - Rivery
- Click Partner Connect in the sidebar menu to get started
- More detail in Databricks Academy course:
  - How to Ingest Data for Databricks SQL

GRANT and REVOKE

- Databricks SQL supports standard `GRANT` and `REVOKE` statements in SQL
- Permission types include `CREATE`, `MODIFY`, `SELECT`, `USAGE`, and more.
- Permissions can be granted to users, groups, or both
- Can also grant all permissions
- Example:

```
GRANT ALL PRIVILEGES ON TABLE table1 TO finance;
```

- Revoke privileges in the same way

Data Explorer

- A UI tool for working with database entities
- Grant and revoke permissions, view schema details, preview sample data, and see table details and properties
- Click “Data” in the sidebar menu to access the Data Explorer

02-1 – DEMO: INGESTING DATA

02-3 – LAB: INGESTING DATA

databricks

Joins

Join

- Combine rows from two relations based on a criteria
- Relations are tables, views, and more
- Many join types: `INNER`, `LEFT`, `RIGHT`, `FULL`, `SEMI`, `ANTI`, and `CROSS`
- The criteria is a boolean expression that specifies how the relations will be joined

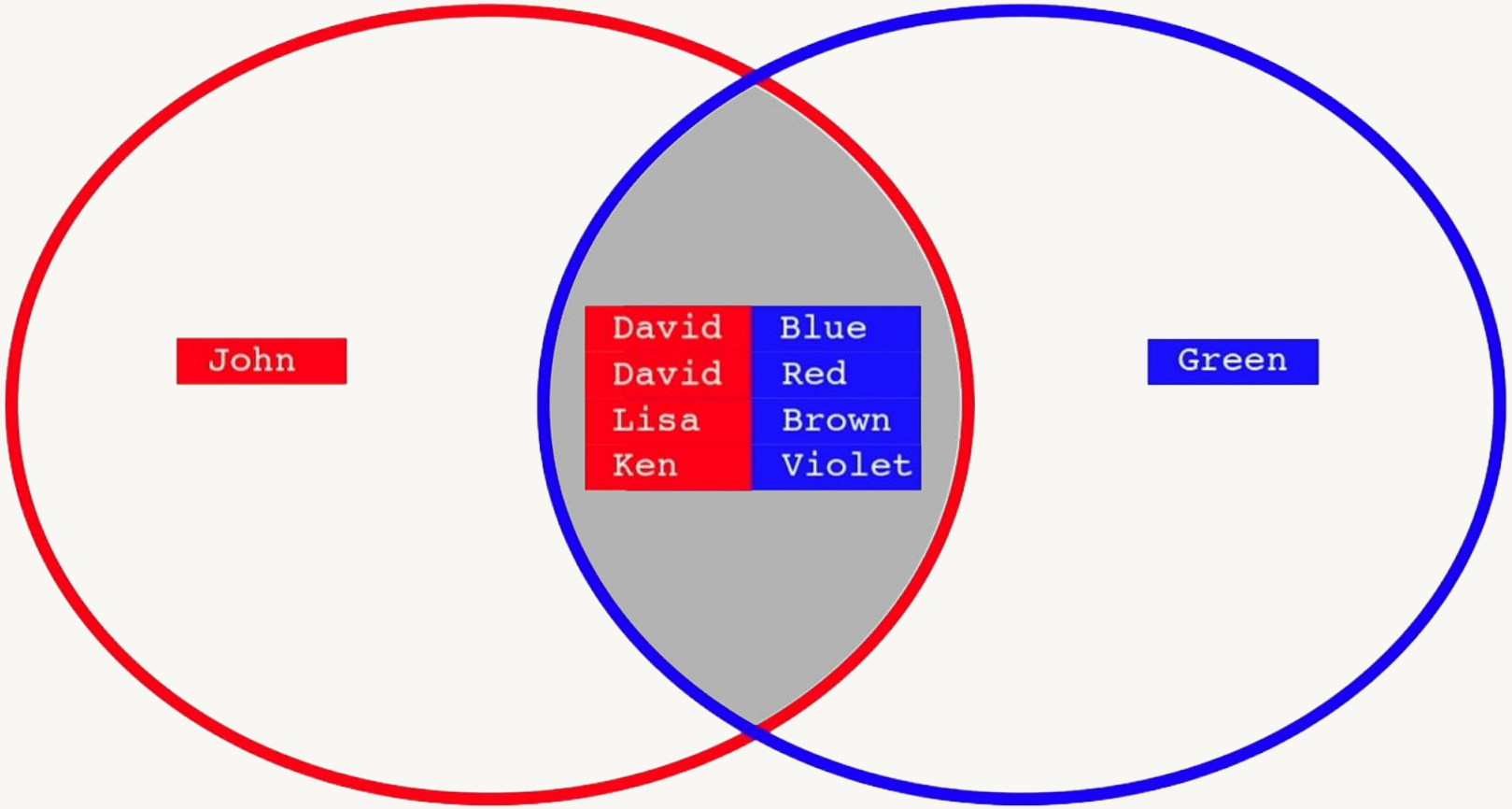
Join

- Example:

```
SELECT id, name, deptname
  FROM employee
  INNER JOIN department ON employee.deptno =
    department.deptno;
```

```
SELECT name, f_color FROM table1 INNER JOIN table2 ON table1.id = table2.pid;
```

id	name
1	David
2	John
3	Lisa
4	Ken



pid	f_color
1	Blue
1	Red
3	Brown
4	Violet
5	Green

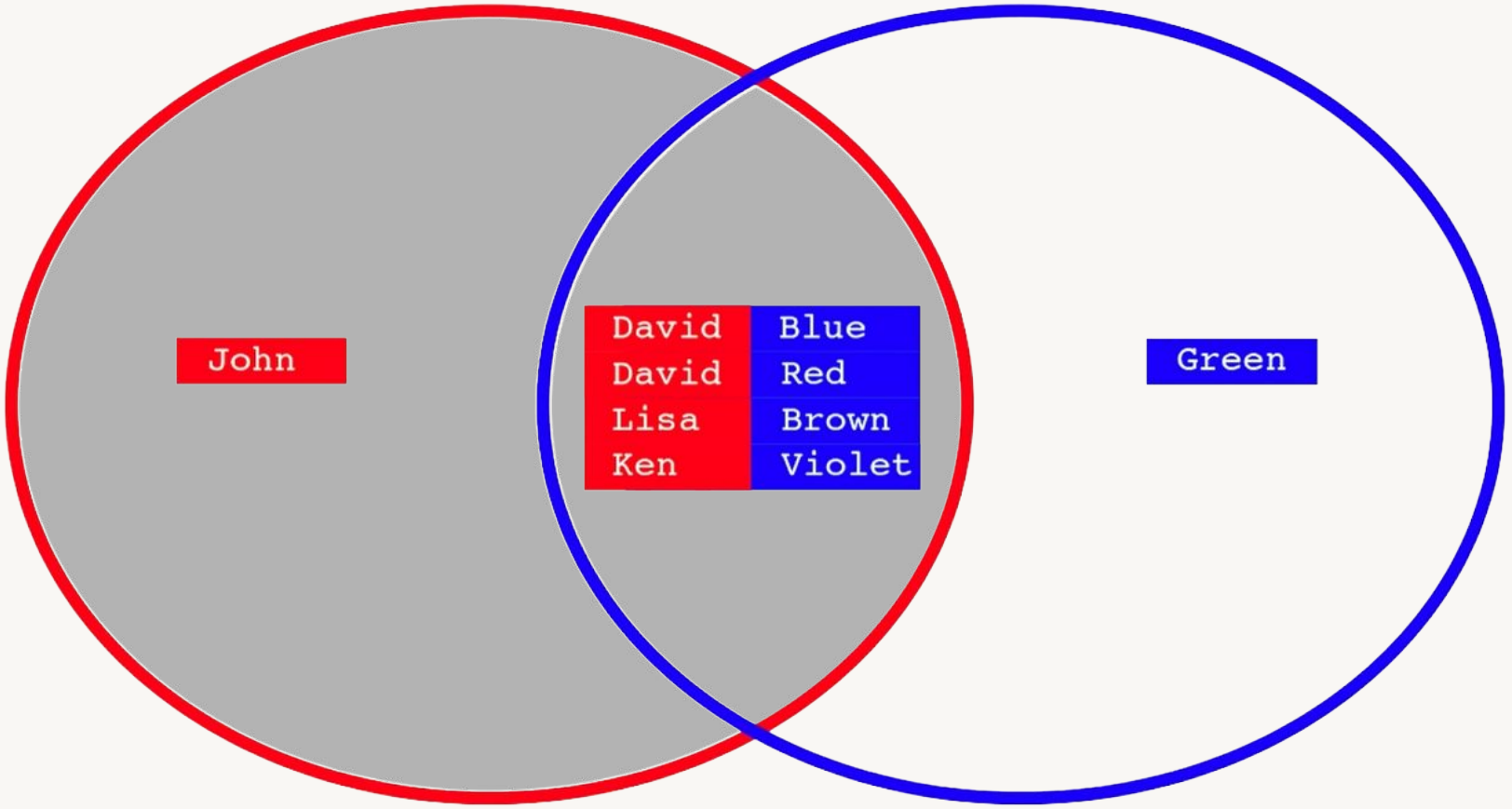
INNER JOIN

Output:

David	Blue
David	Red
Lisa	Brown
Ken	Violet

```
SELECT name, f_color FROM table1 LEFT OUTER JOIN table2 ON table1.id = table2.pid;
```

id	name
1	David
2	John
3	Lisa
4	Ken



pid	f_color
1	Blue
1	Red
3	Brown
4	Violet
5	Green

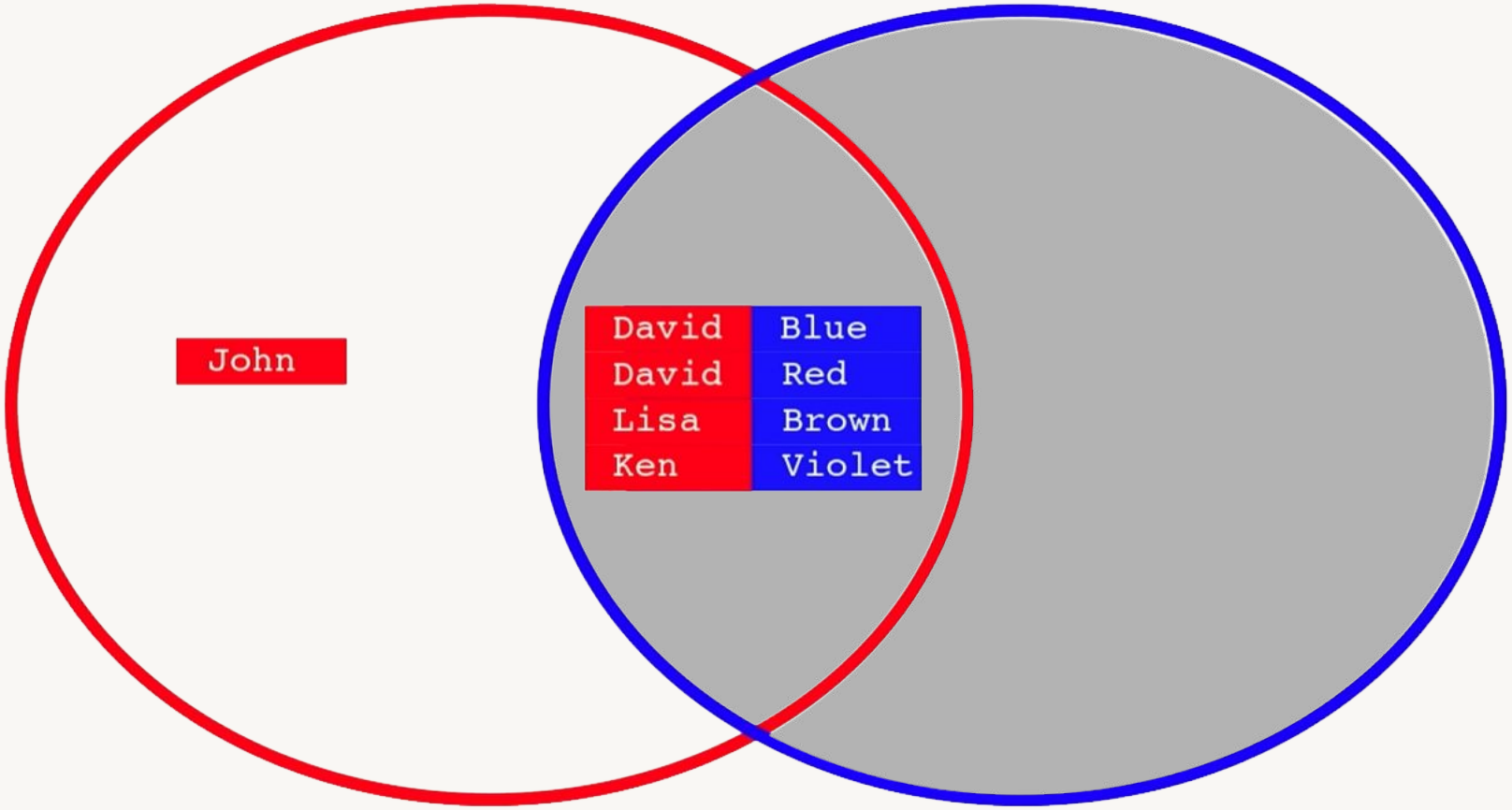
LEFT JOIN

Output:

David	Blue
David	Red
Lisa	Brown
Ken	Violet
John	NULL

```
SELECT name, f_color FROM table1 RIGHT OUTER JOIN table2 ON table1.id = table2.pid;
```

id	name
1	David
2	John
3	Lisa
4	Ken



pid	f_color
1	Blue
1	Red
3	Brown
4	Violet
5	Green

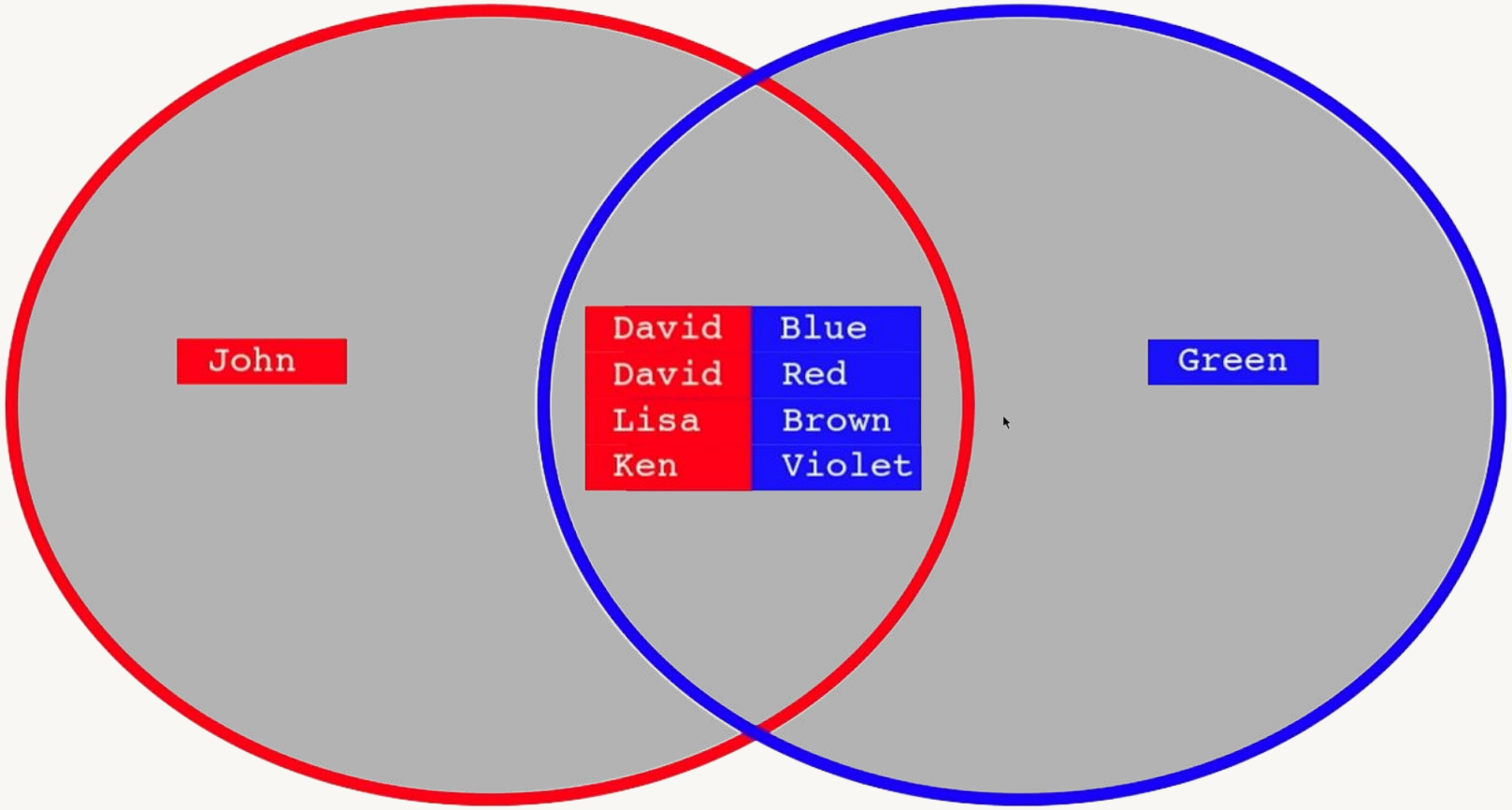
RIGHT JOIN

Output:

David	Blue
David	Red
Lisa	Brown
Ken	Violet
NULL	Green

```
SELECT name, f_color FROM table1 FULL OUTER JOIN table2 ON table1.id = table2.pid;
```

id	name
1	David
2	John
3	Lisa
4	Ken



pid	f_color
1	Blue
1	Red
3	Brown
4	Violet
5	Green

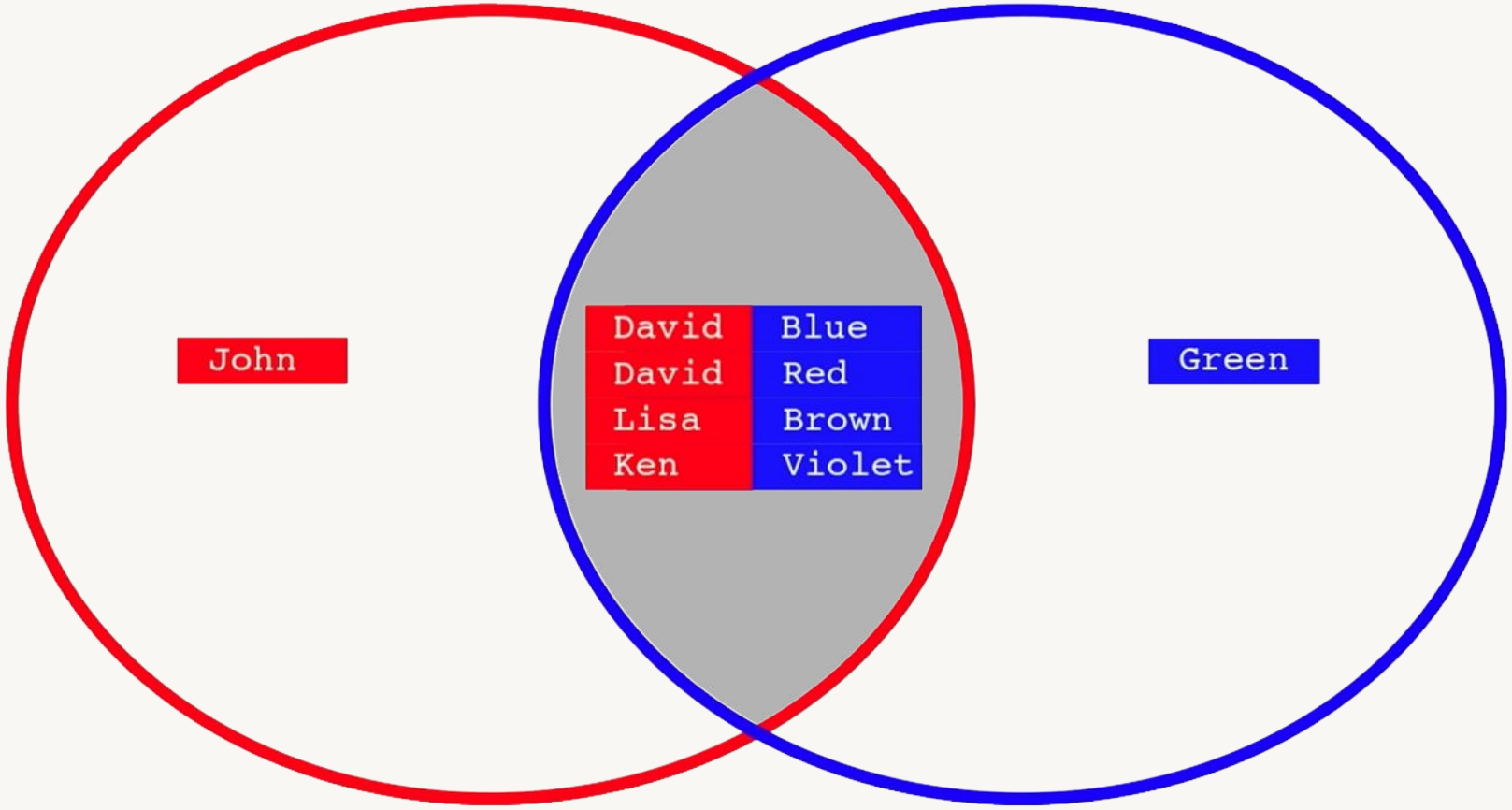
FULL JOIN

Output:

David	Blue
David	Red
Lisa	Brown
Ken	Violet
John	NULL
NULL	Green

```
SELECT name FROM table1 LEFT SEMI JOIN table2 ON table1.id = table2.pid;
```

id	name
1	David
2	John
3	Lisa
4	Ken



pid	f_color
1	Blue
1	Red
3	Brown
4	Violet
5	Green

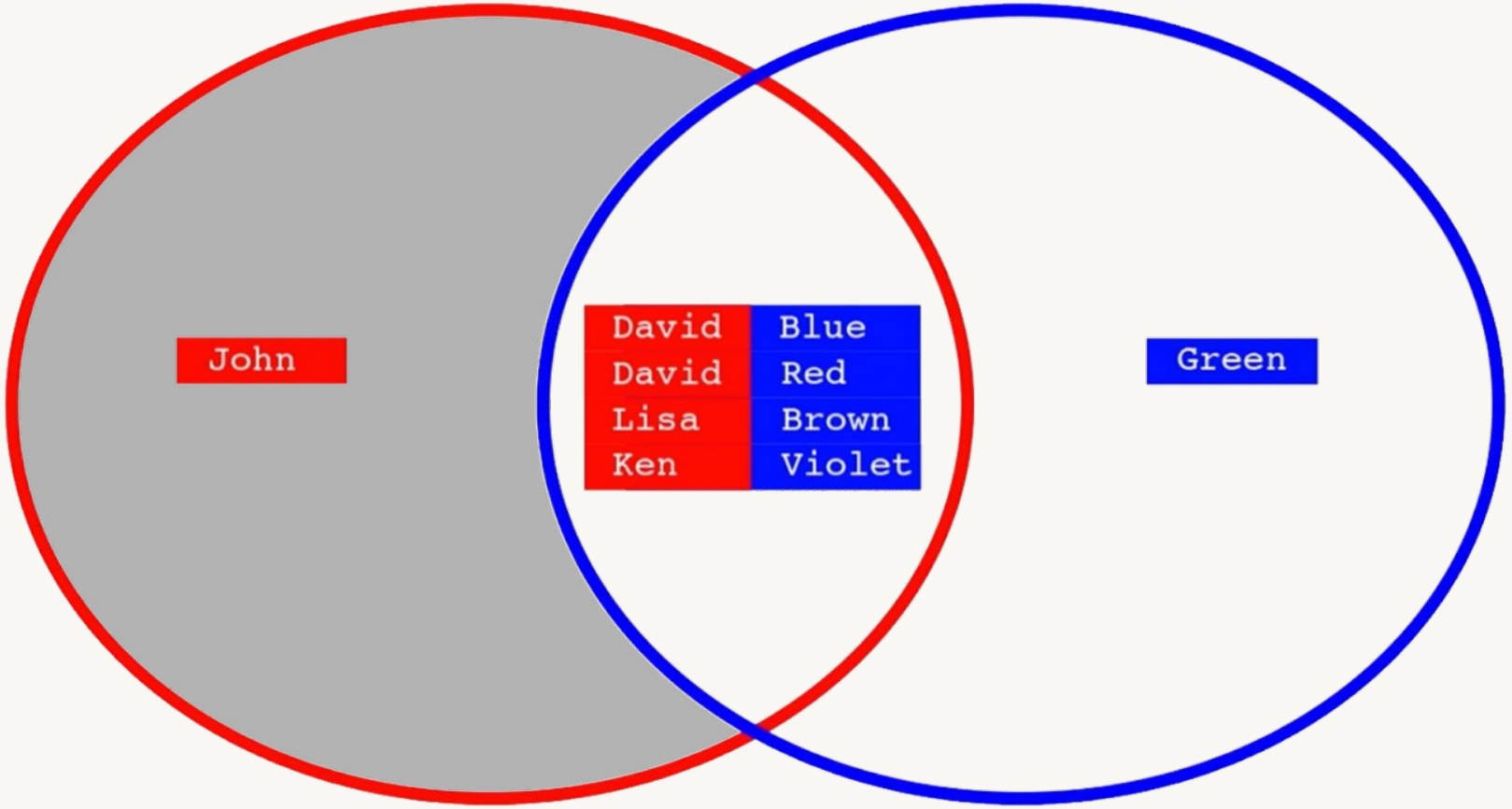
SEMI JOIN

Output:

David
Lisa
Ken

```
SELECT name FROM table1 LEFT ANTI JOIN table2 ON table1.id = table2.pid;
```

id	name
1	David
2	John
3	Lisa
4	Ken



pid	f_color
1	Blue
1	Red
3	Brown
4	Violet
5	Green

ANTI JOIN

Output: John

```
SELECT name, f_color FROM table1 CROSS JOIN table2 ON table1.id = table2.pid;
```

id	name
1	David
2	John
3	Lisa
4	Ken

pid	f_color
1	Blue
1	Red
3	Brown
4	Violet
5	Green

Output:

David	Blue
David	Red
David	Brown
David	Violet
David	Green
John	Blue
John	Red
John	Brown
John	Violet
John	Green
Lisa	Blue
Lisa	Red
Lisa	Brown
Lisa	Violet
Lisa	Green
Ken	Blue
Ken	Red
Ken	Brown
Ken	Violet
Ken	Green

CROSS JOIN

02-5-1 – DEMO: DELTA COMMANDS IN DATABRICKS SQL

02-5-2 – DEMO:
OPTIONAL:
BASIC SQL

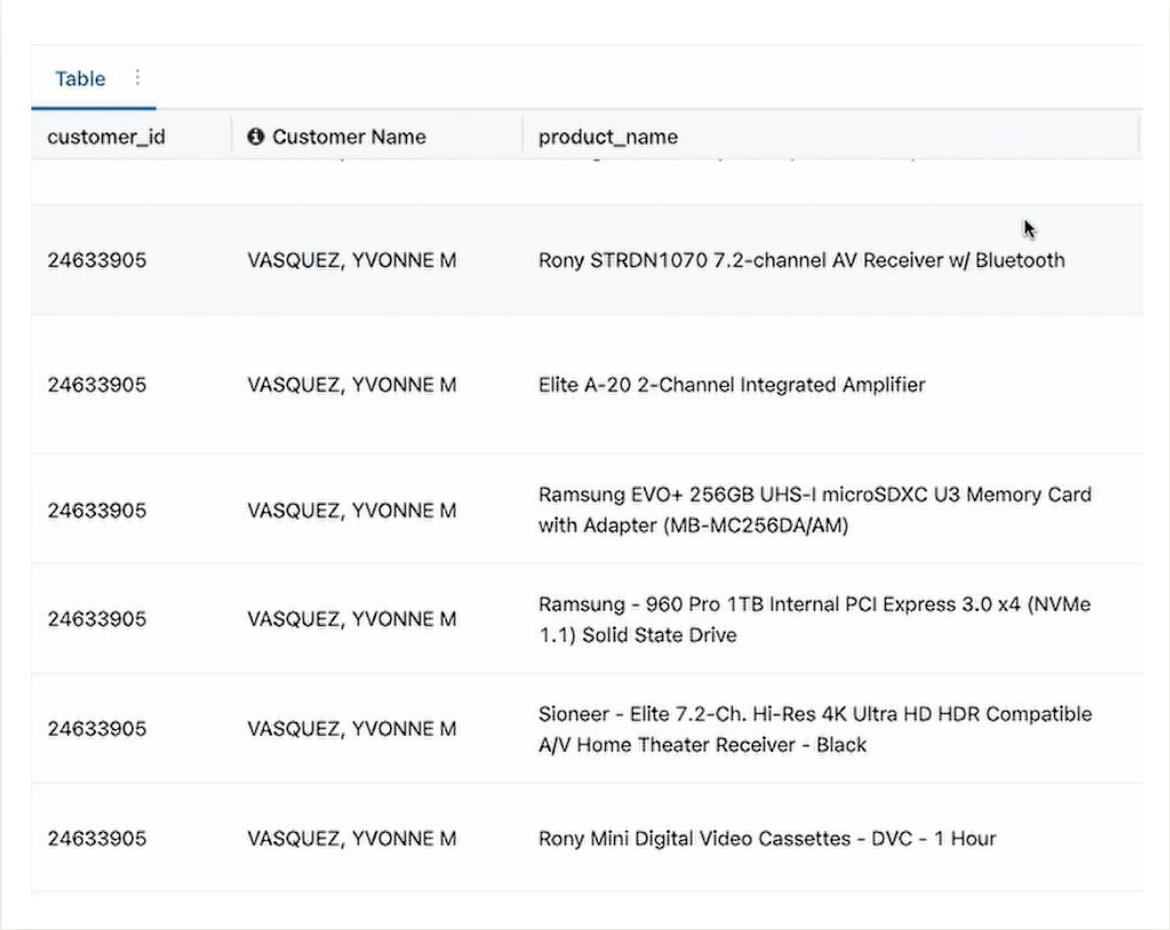
02-6 – LAB: BASIC SQL

databricks

Data Visualization

Table

- Default visualization
- Customizable columns
- Change heading
- Add description
- Change font color
- Conditional font color
  - Based on each data value



Details, Counter, Pivot

Details

customer_id	29717491
customer_name	MATTHEWS, CHRIS E
product_name	Opple NakBook - 12 - Core
order_date	2019-08-12
product_category	Opple
product	{"curr":"USD","id":"AVpf-2 English"","price":3553,"qty
total_price	35530

Counter

2,757,026 Dollars (3,000,000 Dollars)

Sales Goal

Pivot

Table

Count

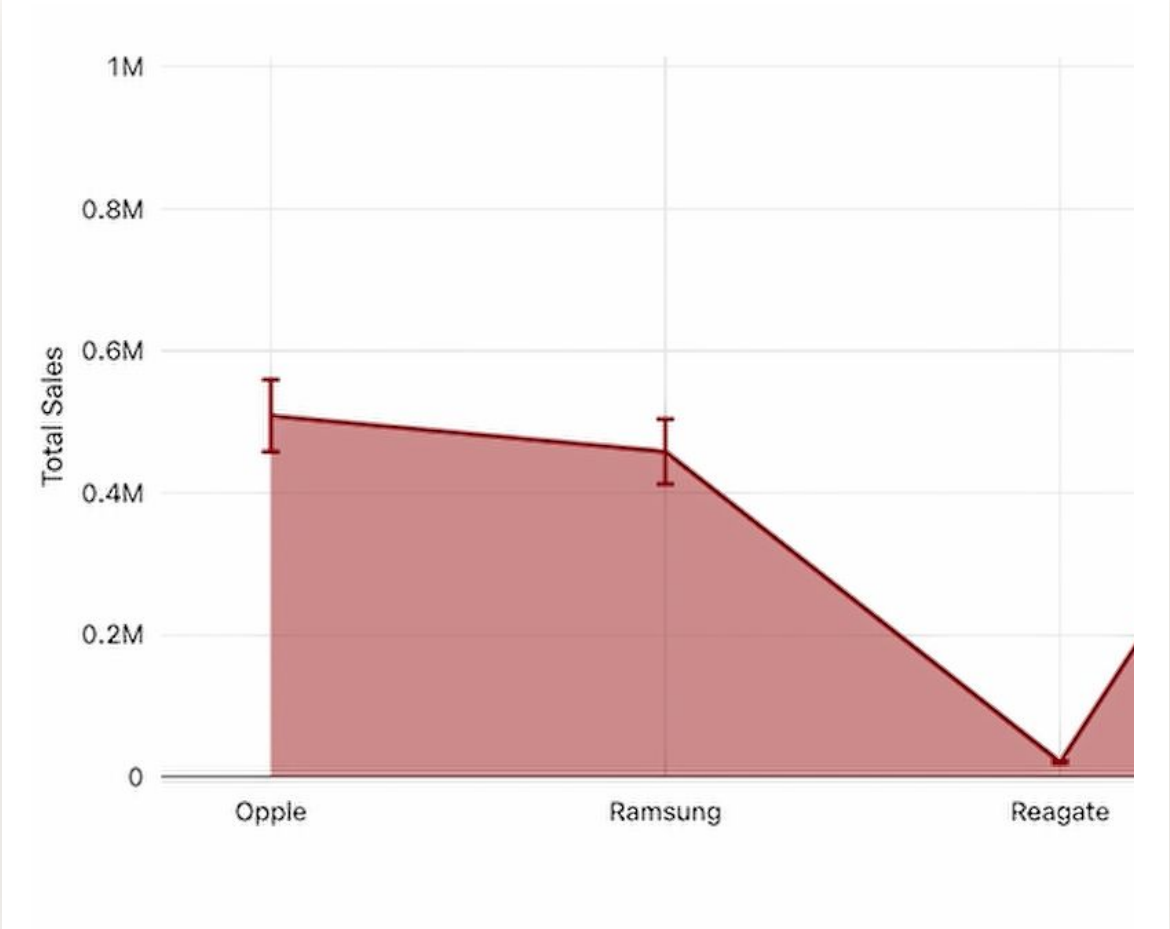
Month

Category

Category	Month 8	9	10
Opple	22	18	19
Ramsung	19	15	25
Reagate	2	2	1
Rony	32	38	34
Sioneer	21	33	22
Zamaha	4	6	3
Totals	100	112	104

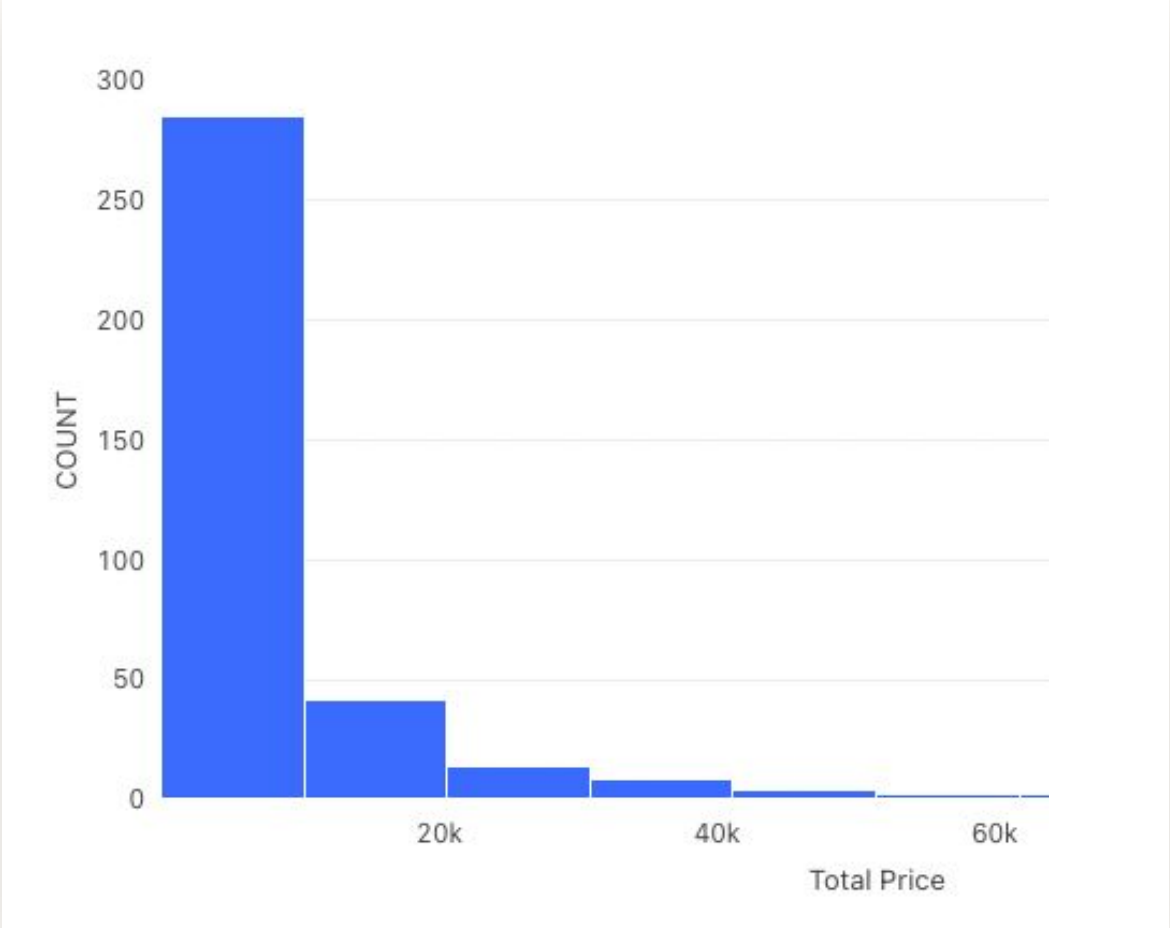
Charts

- Chart types: Line, Bar, Area, Pie, Scatter, Bubble, Heatmap, and Box
- Grouping
- Stacking
- Error Bars



Histogram

- Display a count
- Control of number of buckets

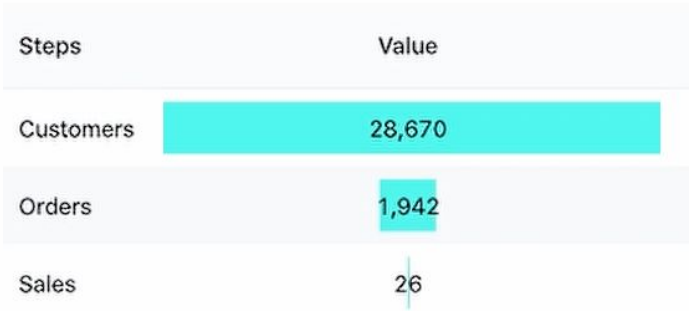


Cohort, Funnel, Word Cloud

Cohort

Time	Customers	1	2	3	4	5
January 2018	82	0.00%	0.00%	4.88%	0.00%	0.00%
February 2018	0	-	-	-	-	-
March 2018	72	0.00%	2.78%	0.00%	0.00%	0.00%
April 2018	78	0.00%	1.28%	0.00%	0.00%	0.00%
May 2018	70	8.57%	0.00%	0.00%	0.00%	0.00%
June 2018	87	0.00%	0.00%	2.30%	0.00%	0.00%
July 2018	86	4.65%	0.00%	0.00%	0.00%	0.00%
August 2018	80	0.00%	65.00%	61.25%	0.00%	0.00%
September 2018	72	4.17%	0.00%	0.00%	0.00%	-
October 2018	88	0.00%	4.55%	0.00%	-	-
November 2018	78	5.13%	0.00%	-	-	-
December 2018	77	0.00%	-	-	-	-
January 2019	91	-	-	-	-	-
February 2019	93	-	-	-	-	-
March 2019	84	-	-	-	-	-
April 2019	69	-	-	-	-	-

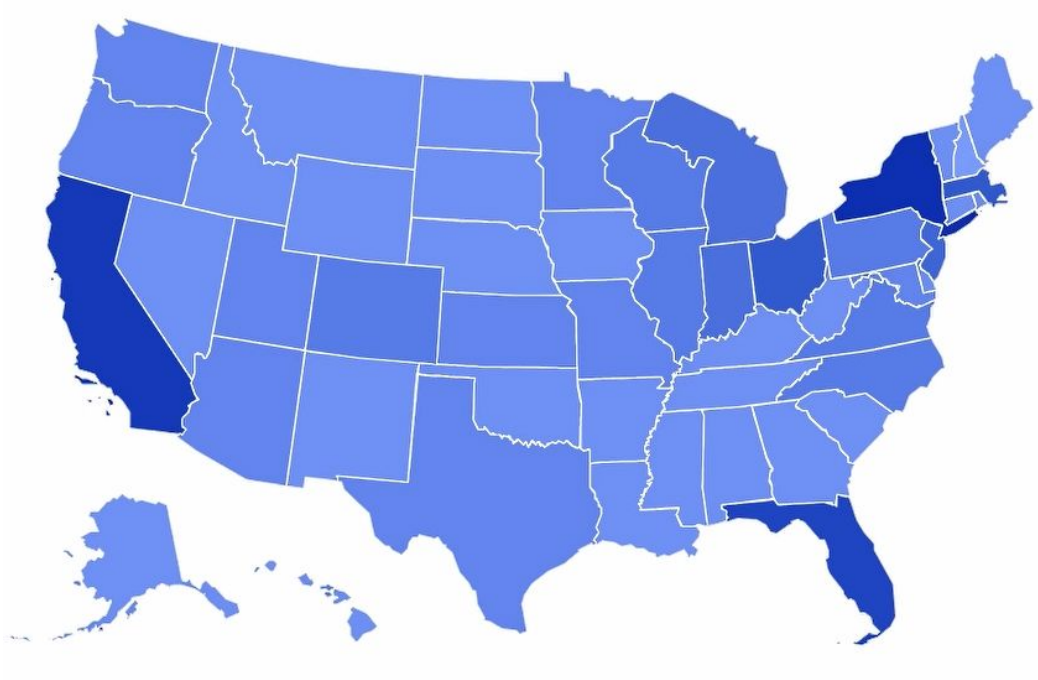
Funnel



Word Cloud

Maps

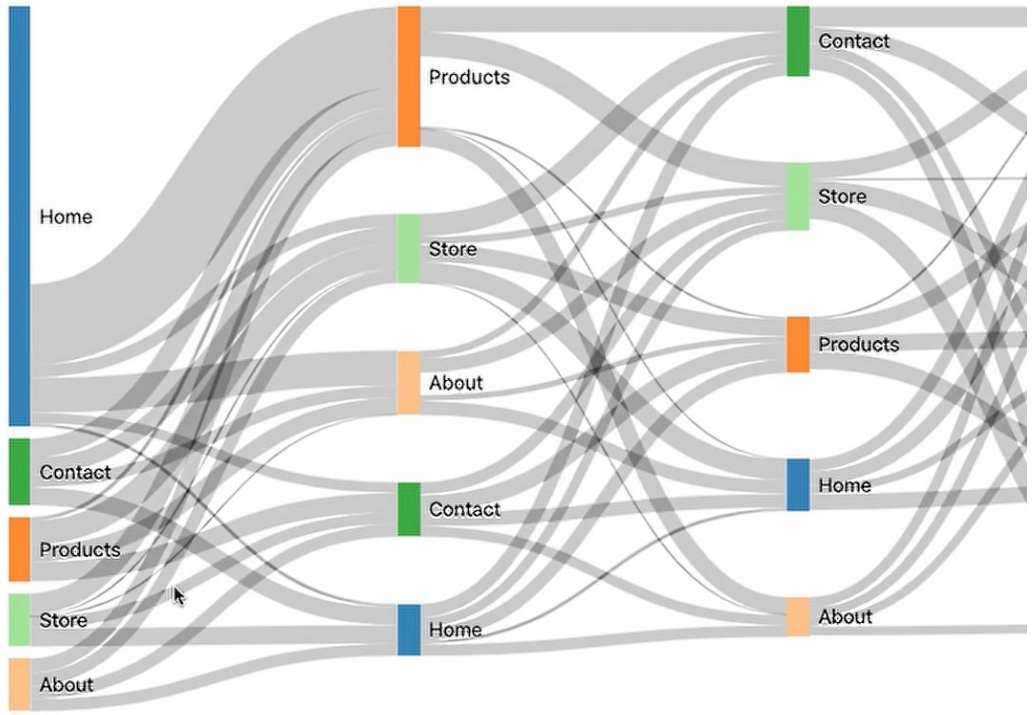
Choropleth Map



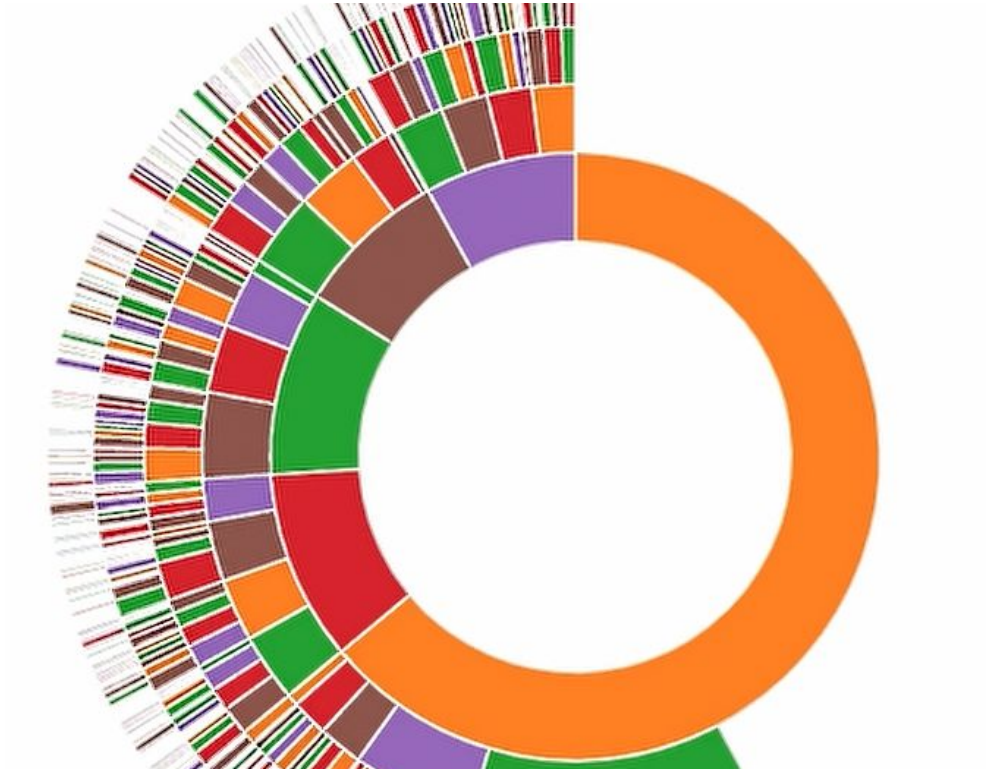
Marker Map

Sankey and Sunburst

Sankey



Sunburst



Databricks Academy

- We aren't going to cover all visualizations in this course.
- More detail can be found in the Databricks Academy course:
  - Data Visualization with Databricks SQL
- If you are using Tableau or PowerBI, you can connect both to Databricks SQL
- More detail in the Databricks Academy course:
  - How to Integrate BI Tools with Databricks SQL

03-2 – DEMO: DATA VISUALIZATIONS AND DASHBOARDS

03-3 – LAB: DATA VISUALIZATIONS AND DASHBOARDS

03-4 – DEMO: NOTIFYING STAKEHOLDERS

03-5 – LAB: NOTIFYING STAKEHOLDERS

03-6 – LAB:
FINAL LAB ASSIGNMENT