Introducing a Modern SQL Experience in Apache Beam

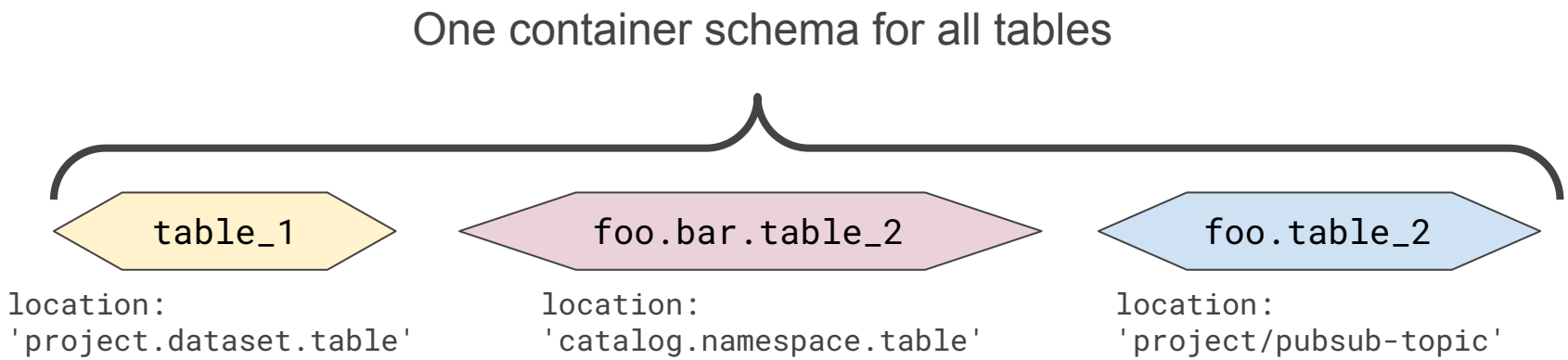
Ahmed Abualsaud

Old Frictions

Flat schema

Old Frictions

Flat schema



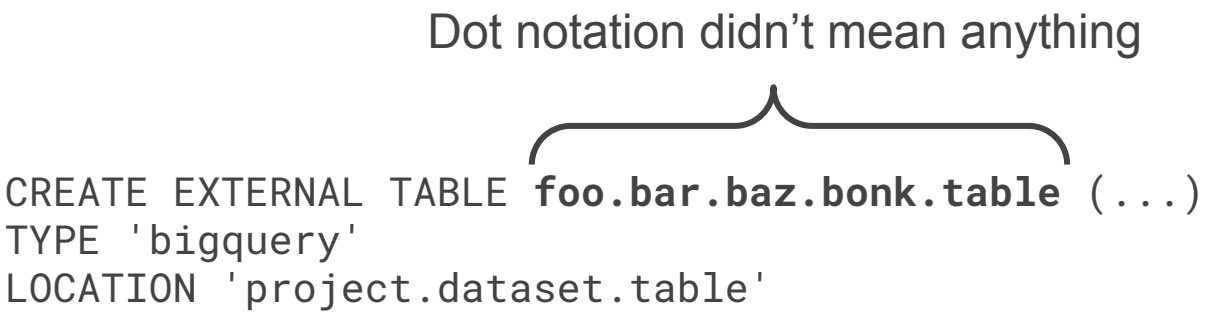
Old Frictions

Flat schema

```
CREATE EXTERNAL TABLE foo.bar.baz.bonk.table (...)
TYPE 'bigquery'
LOCATION 'project.dataset.table'
```

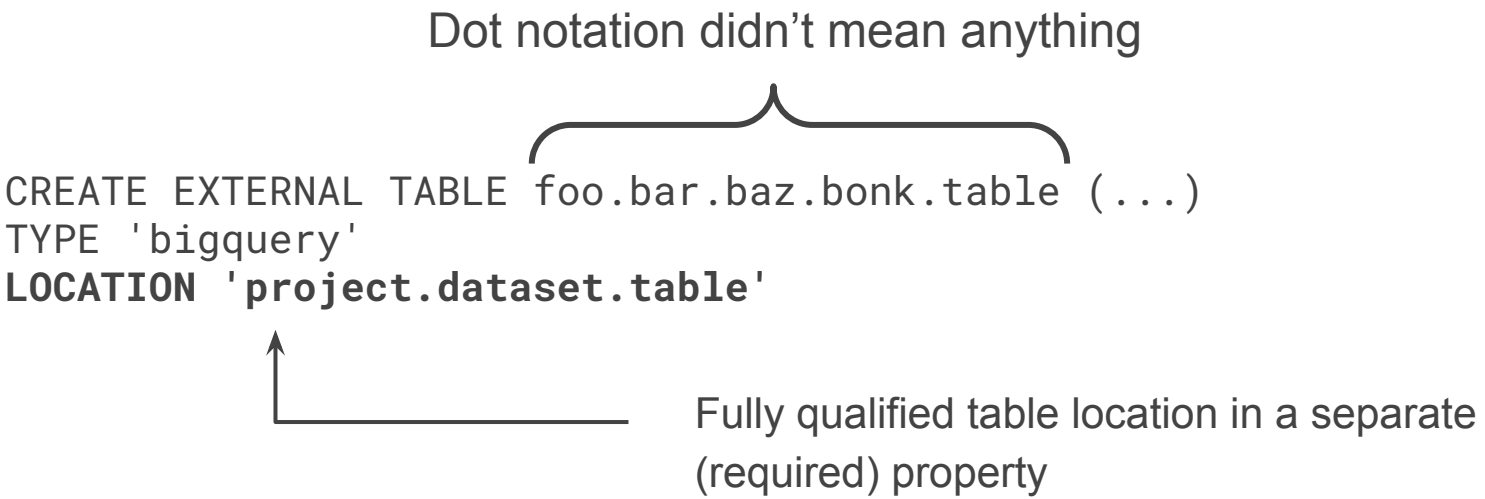
Old Frictions

Flat schema



Old Frictions

Flat schema



Old Frictions

Flat schema

No Catalog or Database concepts

Old Frictions

Flat schema

No Catalog or Database concepts

No discovery

- Always needed to run a CREATE TABLE before operating on the table

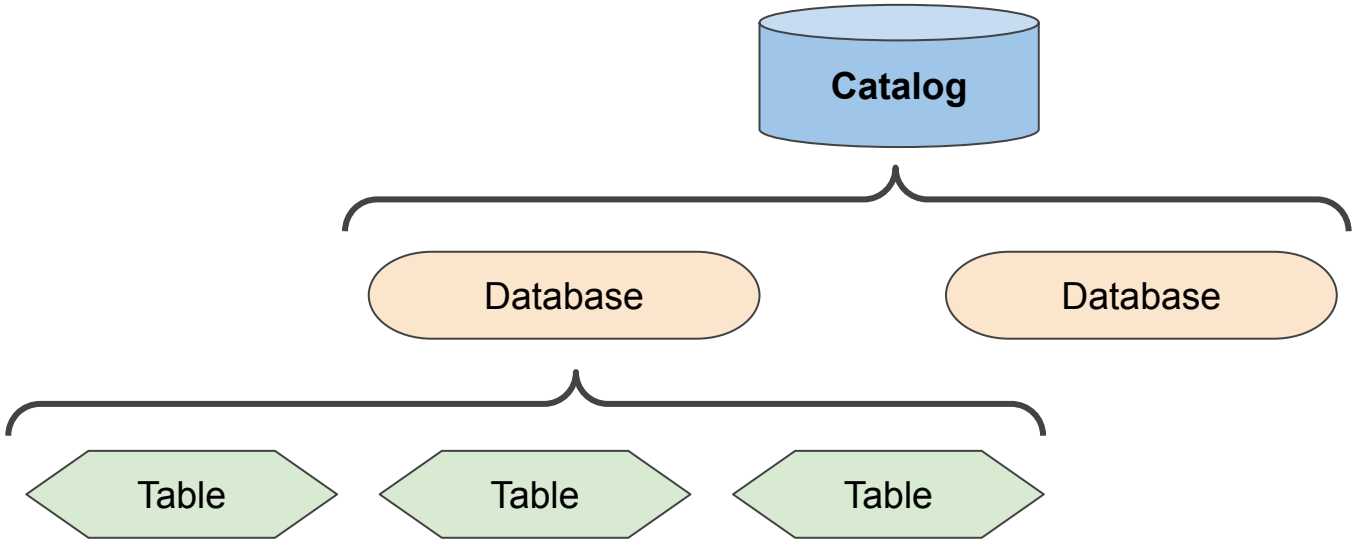
Iceberg Exposed the Gap

SQL considered a first-class language.

Open table format that can be used independently of any platform.

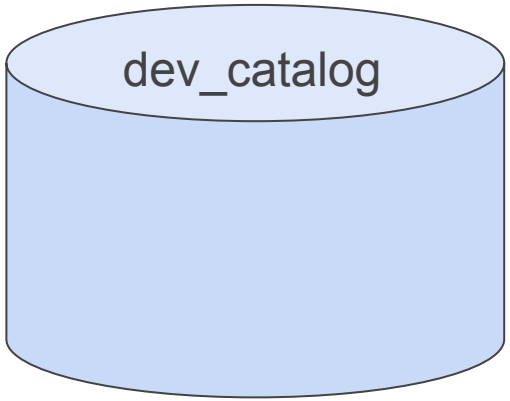
Clear concepts for catalogs, namespaces, tables.

Introducing the Standard Hierarchy



Why Catalogs Matter

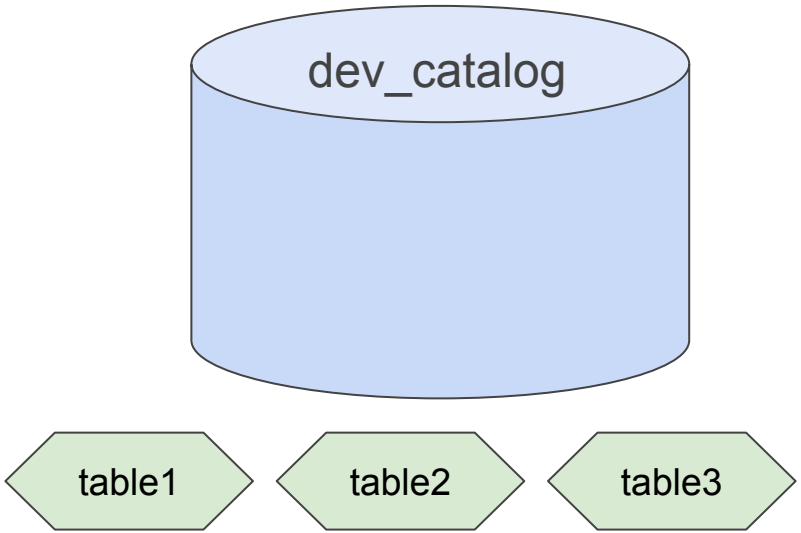
Register an environment once



Why Catalogs Matter

Register an environment once

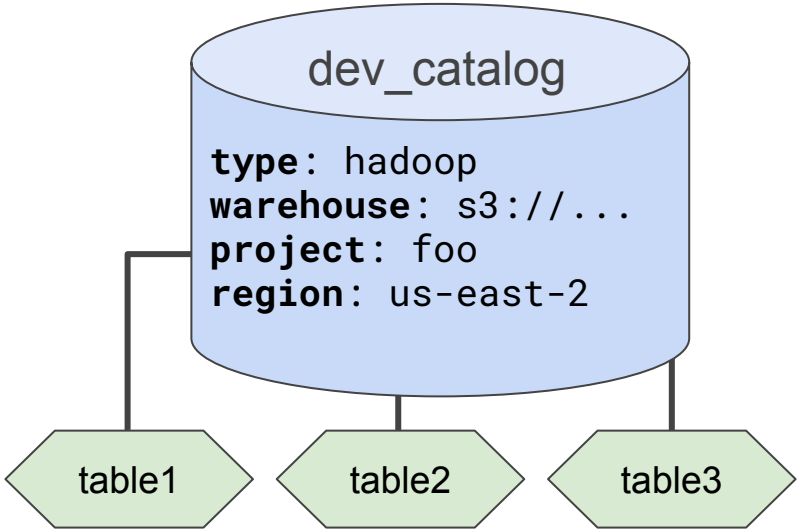
- Use it to **discover databases and tables**



Why Catalogs Matter

Register an environment once

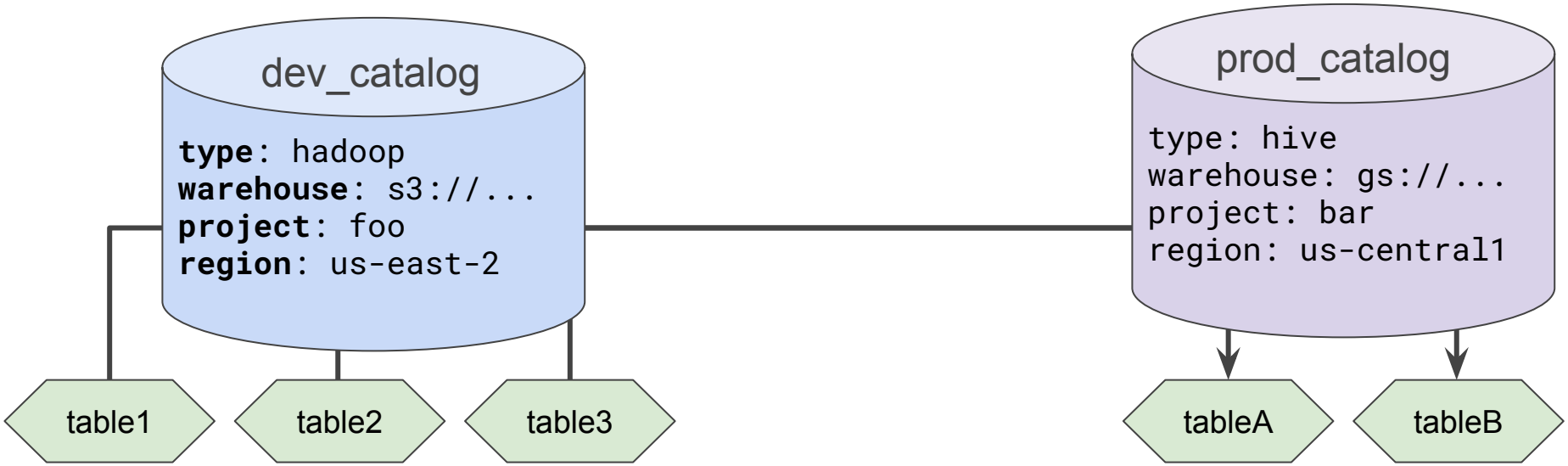
- Use it to discover databases and tables
- Use the same configuration to connect to different tables



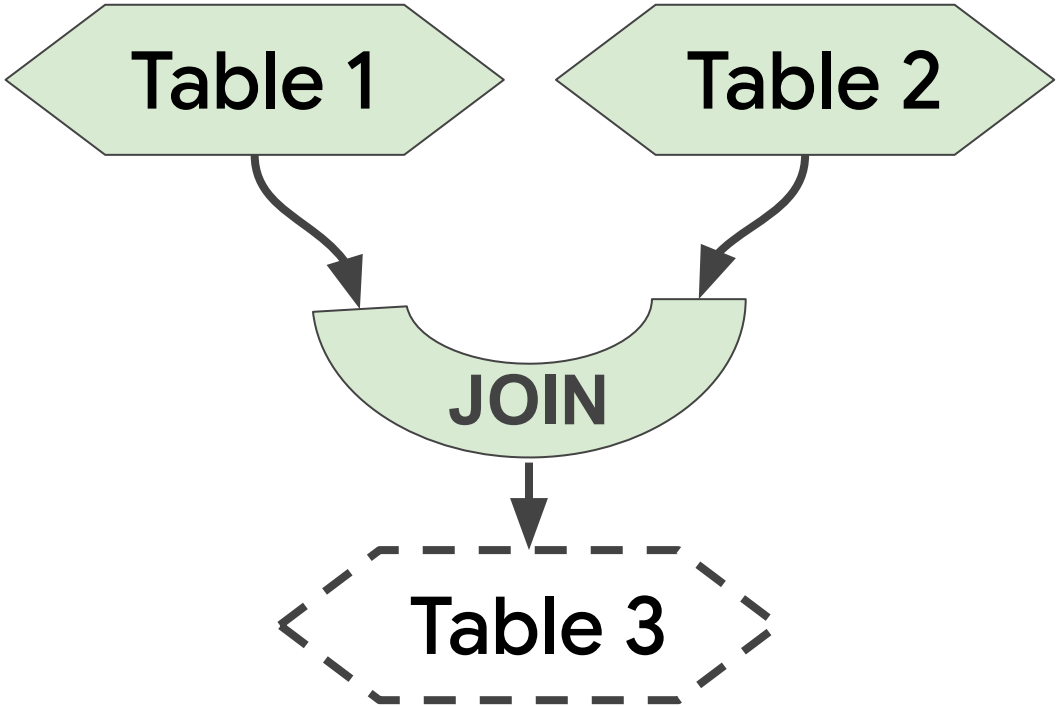
Why Catalogs Matter

Register an environment once

- Use it to discover databases and tables
- Use the same configuration to connect to different tables



Let’s take a JOIN example



Before…

Setup

```
CREATE EXTERNAL TABLE names (id INT64, name STRING)
TYPE 'iceberg'
LOCATION 'namespace.names'
TBLPROPERTIES '{
    "catalog_properties.type": "hadoop",
    "catalog_properties.warehouse": "s3://bucket/dir",
    "catalog_properties.io-impl": "org.apache.iceberg.aws.s3.S3FileIO"
}';

CREATE EXTERNAL TABLE ages (id INT64, age INT32)
TYPE 'iceberg'
LOCATION 'namespace.ages'
TBLPROPERTIES '{
    "catalog_properties.type": "hadoop",
    "catalog_properties.warehouse": "s3://bucket/dir",
    "catalog_properties.io-impl": "org.apache.iceberg.aws.s3.S3FileIO"
}';

CREATE EXTERNAL TABLE people (INT64, name STRING, age INT32)
TYPE 'iceberg'
LOCATION 'namespace.people'
TBLPROPERTIES '{
    "catalog_properties.type": "hadoop",
    "catalog_properties.warehouse": "s3://bucket/dir",
    "catalog_properties.io-impl": "org.apache.iceberg.aws.s3.S3FileIO"
}';
```

Query

```
INSERT INTO people (id, name, age)
SELECT
    names.id,
    names.name,
    ages.age
FROM names
INNER JOIN ages
    ON names.id = ages.id;
```

… and After

Setup

```
CREATE CATALOG my_catalog
TYPE 'iceberg'
PROPERTIES (
    'type' = 'hadoop',
    'warehouse' = 's3://bucket/dir',
    'io-impl' = 'org.apache.iceberg.aws.s3.S3FileIO'
);

USE DATABASE my_catalog.namespace;

CREATE EXTERNAL TABLE people (INT64, name STRING, age INT32)
TYPE 'iceberg';
```

Query

```
INSERT INTO people (id, name, age)
SELECT
    names.id,
    names.name,
    ages.age
FROM names
INNER JOIN ages
    ON names.id = ages.id;
```

Cross-Catalog Queries

```
CREATE CATALOG sales_america
TYPE 'bigquery'
PROPERTIES (
    'project' = 'company_sales'
    'location' = 'us-central1',
);

CREATE CATALOG sales_europe
TYPE 'iceberg'
PROPERTIES (
    'type' = 'rest',
    'uri'= 'https://uri.com/restcatalog'
    'warehouse' = 'gs://europe/sales',
);
```

```
SELECT
  a.customer_id,
  a.total_spend + e.total_spend AS total_spend
FROM sales_america.analytics.sales AS a
JOIN sales_europe.analytics.sales AS e
  ON a.customer_id = e.customer_id;
```

Introduced DDL for Catalogs, Databases, and Tables

	SHOW	USE	CREATE	ALTER	DROP
Catalog					
Database				N/A	
Table		N/A			

https://beam.apache.org/documentation/dsls/sql/ddl/

Demo time

https://github.com/ahmedabu98/beam-sql-demo

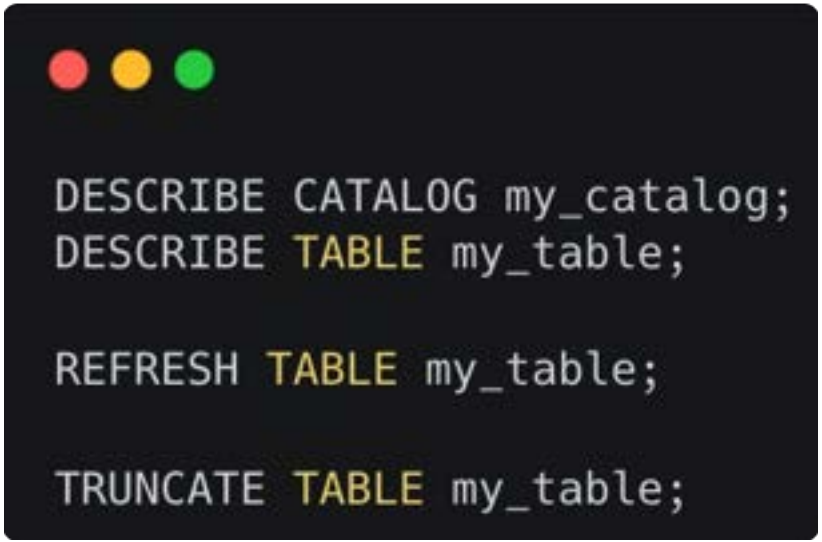
What to Unlock Next

Table procedures (maintenance, admin functions)

```
CALL prod_iceberg.system.rewrite_data_files(table => 'sales.orders');
CALL prod_iceberg.system.rewrite_manifests(table => 'sales.orders');
CALL prod_iceberg.system.expire_snapshots(table => 'sales.orders');
CALL prod_iceberg.system.remove_orphan_files(table => 'sales.orders');
```

What to Unlock Next

More DDL



What to Unlock Next

Catalog implementations for more data sources

```
CREATE CATALOG pg_catalog
TYPE 'jdbc'
PROPERTIES (
    'driver'    = 'org.postgresql.Driver',
    'base-url' = 'jdbc:postgresql://localhost:5432',
    'username' = '...',
    'password' = '...',
    'default_database' = 'public'
);
```

```
CREATE CATALOG kafka_catalog
TYPE 'kafka'
PROPERTIES (
  'bootstrap.servers' = 'broker:9092',
  'schema.registry.url' = 'https://schema-registry',
  'format' = 'avro',
  'watermark.type' = 'PROCESSINGTIME',
);
```

Ahmed Abualsaud

QUESTIONS?

Contact Info:

linkedin.com/in/ahmedabu98

github.com/ahmedabu98

a.abualsaud98@gmail.com

BEAM SUMMIT NYC 2026