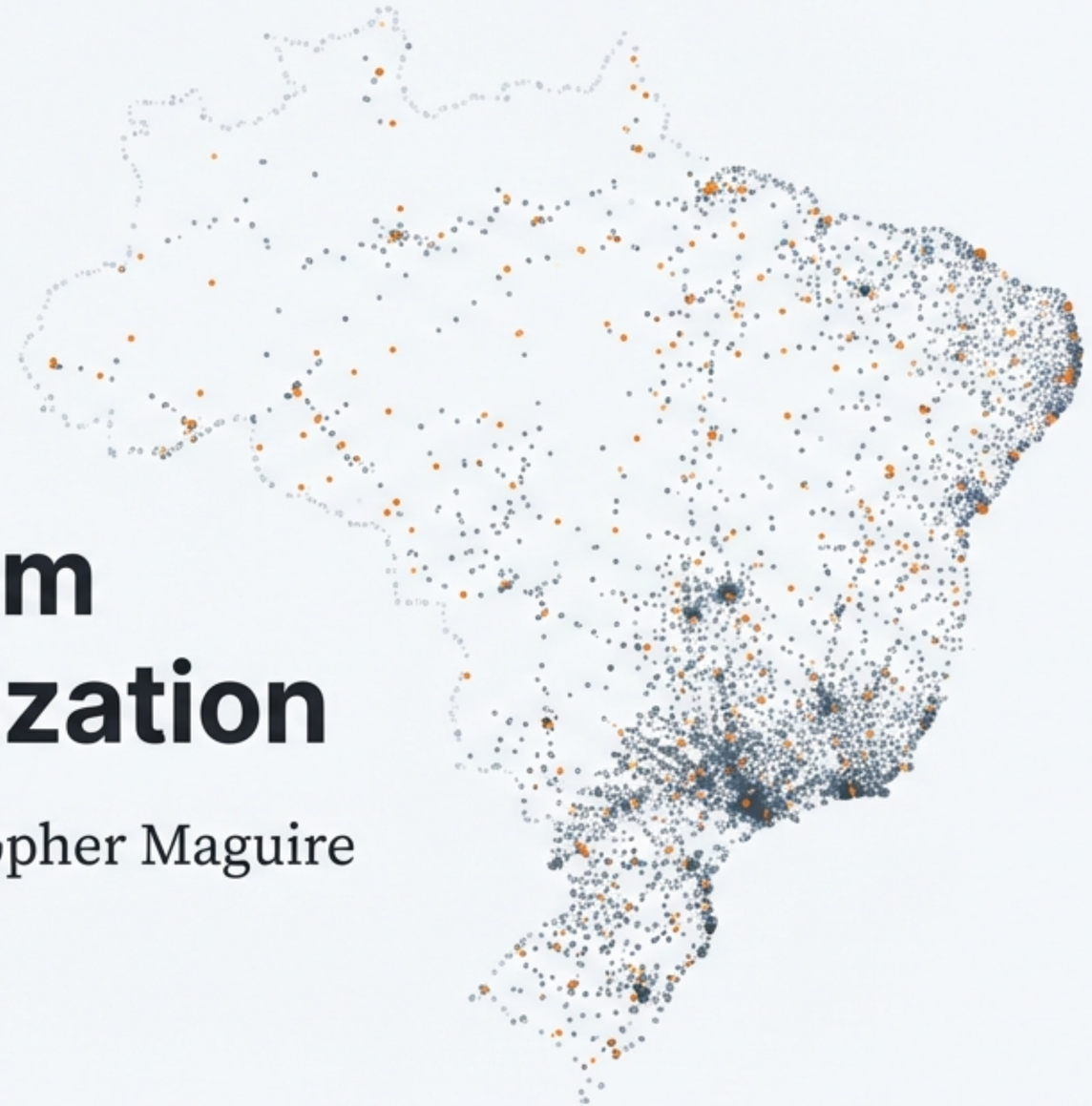
# Analyzing Brazilian E-commerce: A Data Analyst's Journey from Exploration to Optimization

A case study on the Olist dataset by Christopher Maguire



# The Mission: To Learn SQL by Analyzing a Real-World Dataset

- **Objective:** To demonstrate core data analyst skills using the **Olist Brazilian E-commerce** public dataset from Kaggle.
- **Method:** A self-directed learning approach, writing SQL code “struggling through each line to learn as I go.”
- **Dataset Scope:** A rich collection of tables including `customers`, `orders`, `order_items`, `payments`, `products`, and `sellers`.

# Translating Business Curiosity into Actionable Queries

## What foundational metrics can we uncover about the Olist e-commerce platform?

### Key Questions Guiding the Initial Analysis

1. How many customers and orders exist in the dataset?
2. What is the Average Order Value (AOV)?
3. How often do customers place multiple orders?
4. How do payments correspond to orders and customers?

# The Initial Toolkit: Python, Pandas, and SQL in Google Colab

## Core Components

**Environment:** Google Colab / Jupyter Notebook for interactive analysis.

**Language:** Python with the Pandas library for data manipulation.

**Analysis Engine:** SQL queries executed directly within the Python environment.

```
-- Calculating Average Order Value
SELECT
    SUM(payment_value) / COUNT(DISTINCT order_id) AS AOV
FROM
    payments;
```

Connects technical execution directly to a core business metric.

# Uncovering the Core Business Metrics

## Total Customers

Tens of thousands of customers

## Total Orders

A large volume of transactions

## Average Order Value (AOV)

**160.99 BRL**

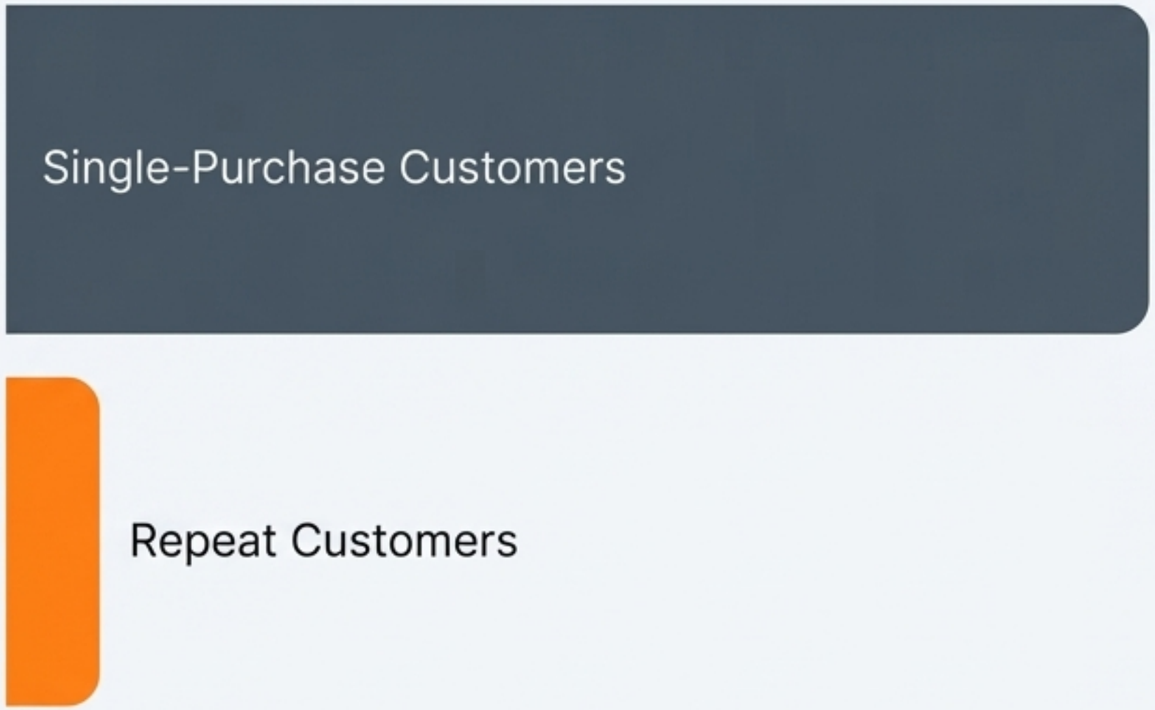
## Payment Integrity

Sum of payments aligns with total orders, indicating data quality.

Initial analysis reveals consistent purchasing behavior and a solid data foundation.

# The Key Insight: A Significant Opportunity to Improve Customer Retention

- **A significant portion of customers place only one order.**
- **Implication:** While repeat customers exist, their proportion is low compared to first-time buyers.
- **Business opportunity:** This points to a clear opportunity for customer engagement strategies designed to increase repeat purchases and lifetime value.



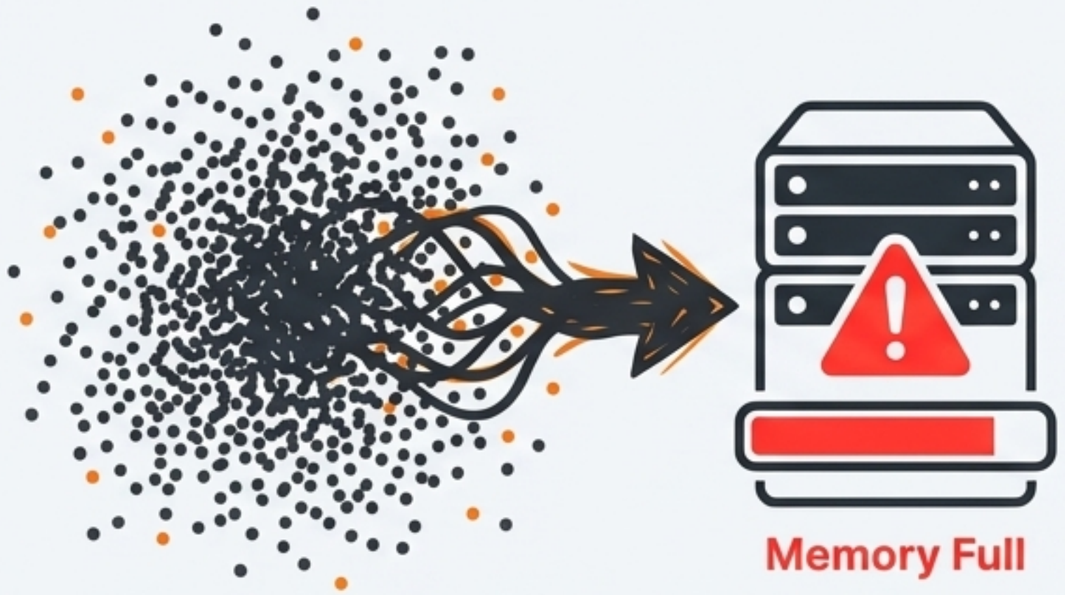
# Pushing Further: Visualizing the Geographic Distribution of Customers

- **New Objective:** To create an interactive heat map in Tableau Public showing where the most customers are located across Brazil.
- **Datasets Used:**
  - olist_customers_dataset.csv
  - olist_geolocation_dataset.csv
- **The Goal:** Highlight the largest customer bases while ensuring the visualization remains performant and usable.

# The Challenge: Tableau Public’s Memory Limits vs. High-Cardinality Geospatial Data

**The Core Problem:** Attempting to plot tens of thousands of individual customer locations caused major performance issues.

- **Issue 1: Memory Overload:** High-cardinality data from `customer_city` and `customer_unique_id` quickly exceeded Tableau Public's memory limits.
- **Issue 2: Performance Bottleneck:** Using joins at the raw row level threatened to "explode the number of rows," making the visualization unresponsive.

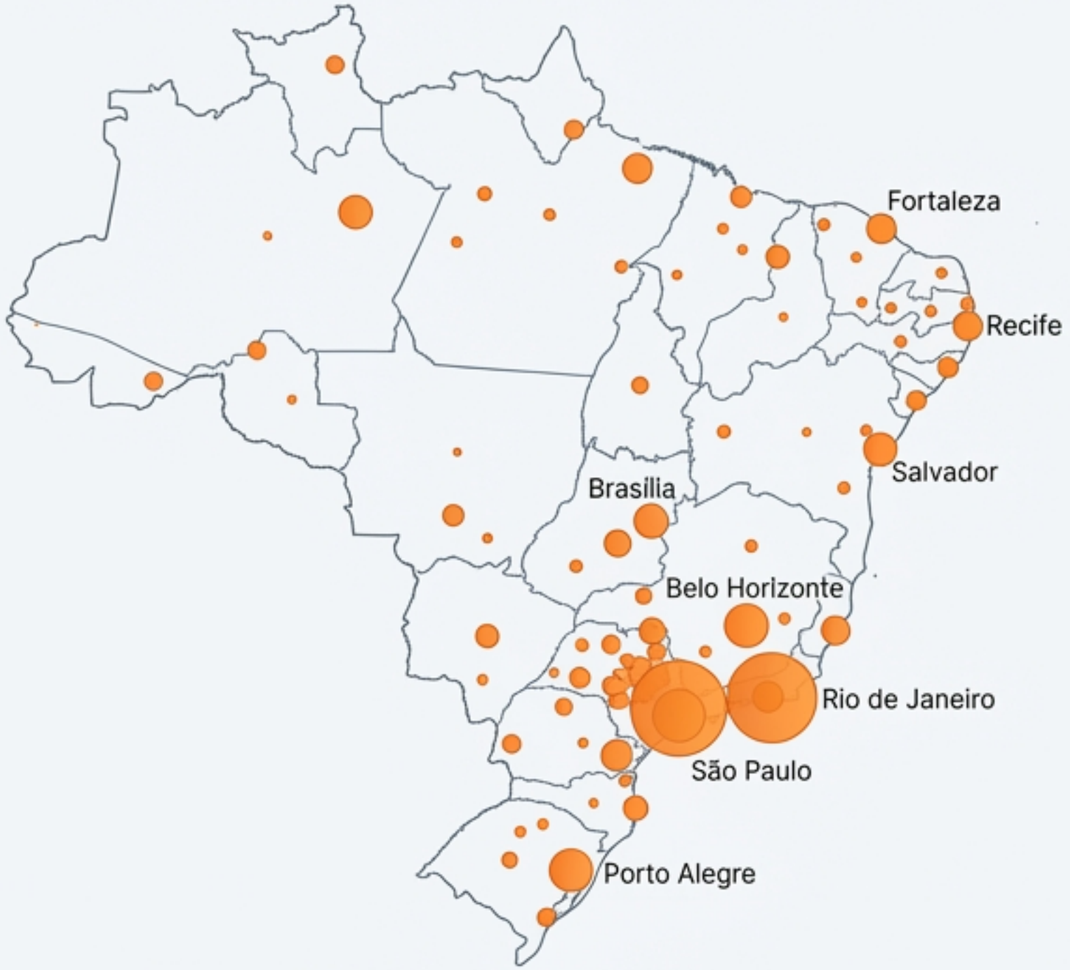



Raw Geospatial Data

# The Solution: A Multi-Pronged Optimization Strategy

1. **Aggregate Before Visualizing:** Created a calculated field `COUNT([customer_unique_id])` to aggregate customers at the city level, dramatically reducing the number of marks to be rendered.
2. **Filter Strategically with Top N:** Applied a "Top 200 by `COUNT(customer_unique_id)`" filter on `` `customer_city` `` to focus on the most significant locations.
3. **Leverage Context Filters:** Used `` `customer_state` `` as a context filter to further reduce memory usage before other filters were applied.
4. **Use Tableau's Native Engine:** Utilized Tableau's generated Latitude/Longitude fields instead of raw CSV columns for optimized geographic processing.

# The Result: An Interactive and Performant Customer Heat Map




- High-cardinality data can be managed effectively with strategic filtering and aggregation.
- The visualization successfully highlights key customer hubs without sacrificing performance.

View on Tableau Public:
https://public.tableau.com/shared/7RBSX3Y69

# A New Bottleneck Emerges: The Limitations of the Development Environment

## The Problem

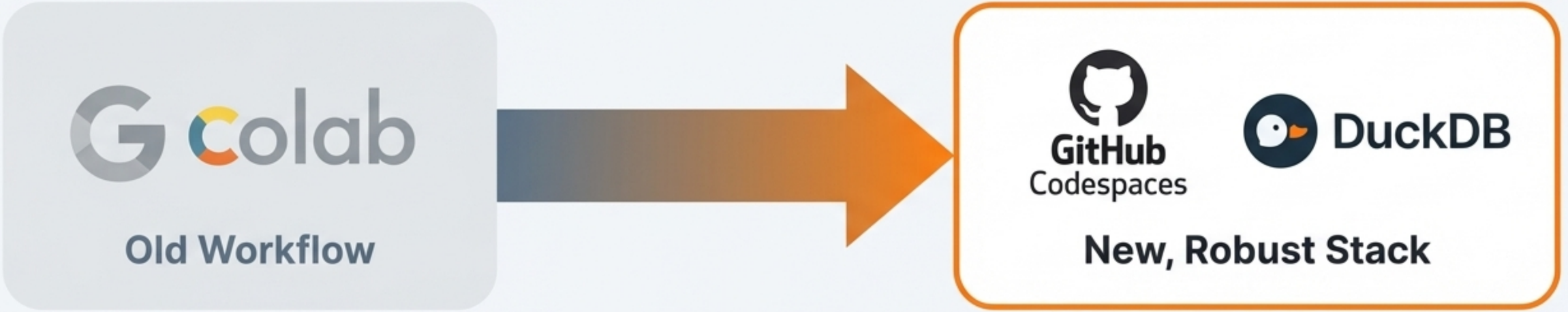
The Google Colab workflow was proving to be fragile and inefficient for a scalable project.

- **Instability**: "Google colab notebook keep crashing due to file storage issues."
- **Inefficiency**: "Colab does not save the Kaggle CSVs permanently so I had to keep uploading them. (very ineffective)"


## The Realization

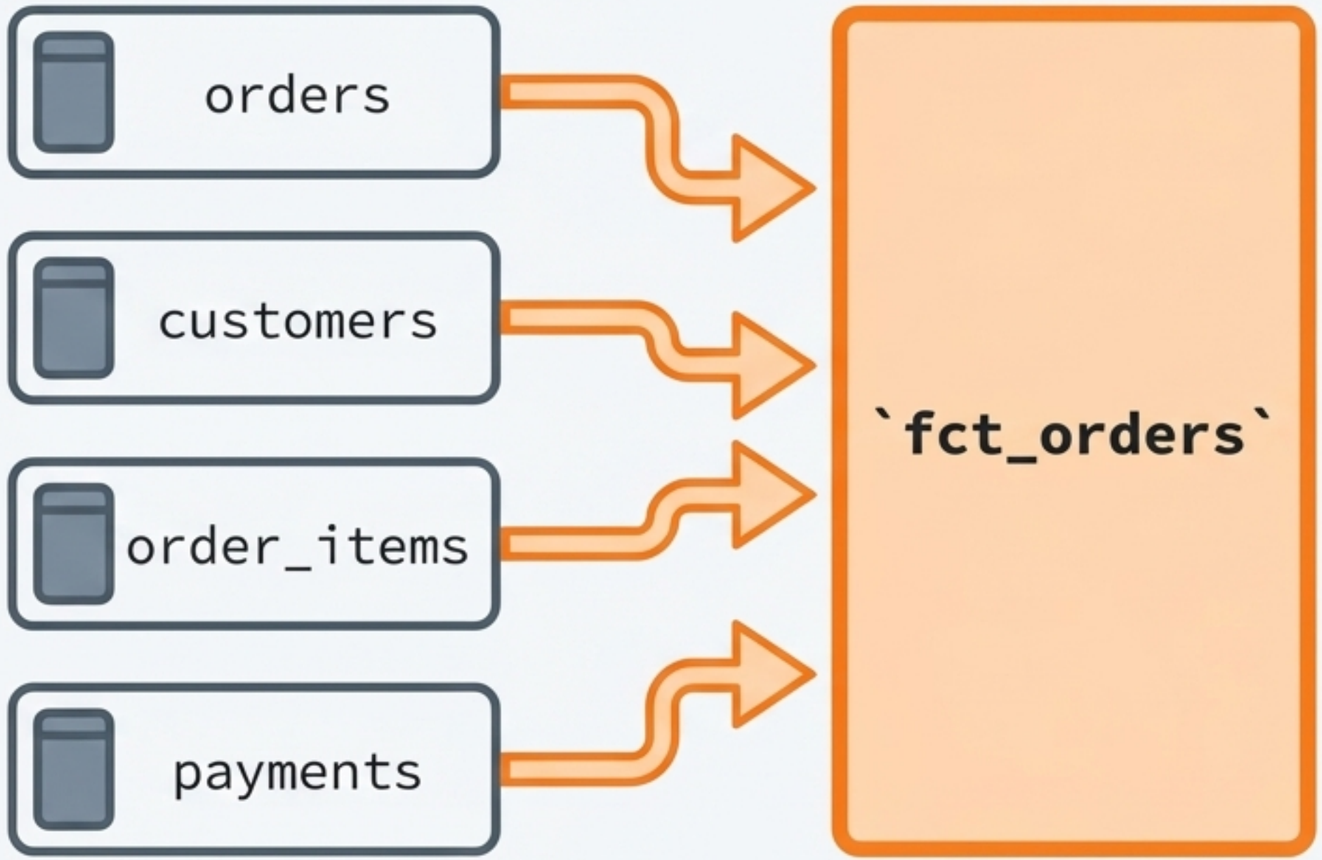
A more robust and permanent solution was needed for the data storage and processing pipeline.

# The Pivot: Building a Professional Data Pipeline with Codespaces and DuckDB




- **GitHub Codespaces**: Provided a persistent, cloud-based development environment where data could be stored directly in the repository.
- **DuckDB**: An in-process SQL database that allowed for running complex SQL queries directly on CSV files without the overhead of a traditional database server.
- **The Benefit**: This created a self-contained, reproducible, and efficient workflow from data ingestion to analysis.

# Creating a Single Source of Truth: The `fct_orders` Table

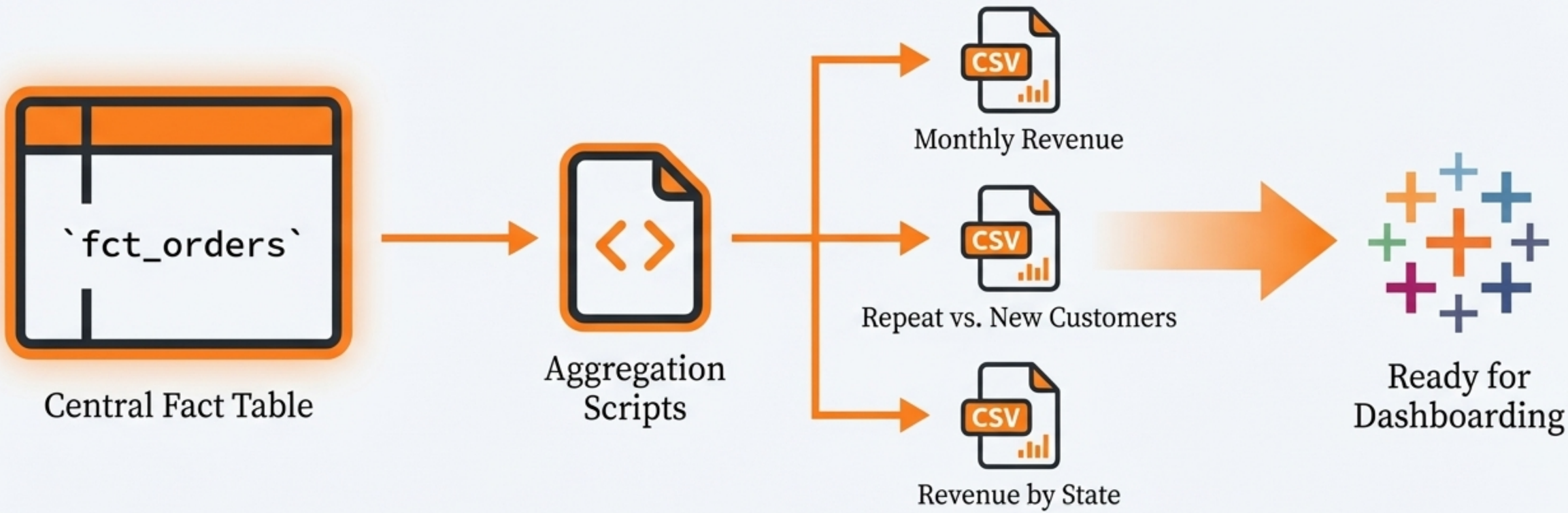



## Key Transformations

- Calculated total revenue for each order.
- Joined in relevant customer information.
- Aggregated and added payment details.
- Created a new flag to identify repeat customers.

Combined multiple raw tables into one comprehensive fact table to serve as the foundation for all analysis.

# From a Centralized Table to Ready-for-Visualization Exports




The result is a repeatable and automated process for preparing data, ready to be fed directly into Tableau for dashboard creation.

# A Journey of Growth: From Foundational Analysis to a Scalable Pipeline

## Foundational SQL Analysis

- Translating business needs into SQL.
- Calculating core metrics (AOV).
- Deriving initial business insights.

## Visualization & Optimization

- Tackling high-cardinality data in Tableau.
- Strategic filtering and aggregation.
- Creating performant, interactive maps.

## Data Engineering & Workflow

- Migrating from Colab to Codespaces.
- Using DuckDB for efficient SQL on files.
- Building a central `fct_orders` table.