

Data Stack: 0 to 100 in under a week

Greg Ratner CTO @ Troops

Twitter: @GregRatner



TR O O P S

# Motivation

- Transparency in customer behavior
- Signed contract with Looker
- Resource-constrained team
- Less than 1 week to go-live

# Goals

- Blueprint for data stack from scratch
- Understand how different pieces fit together

# Requirements



Salesforce

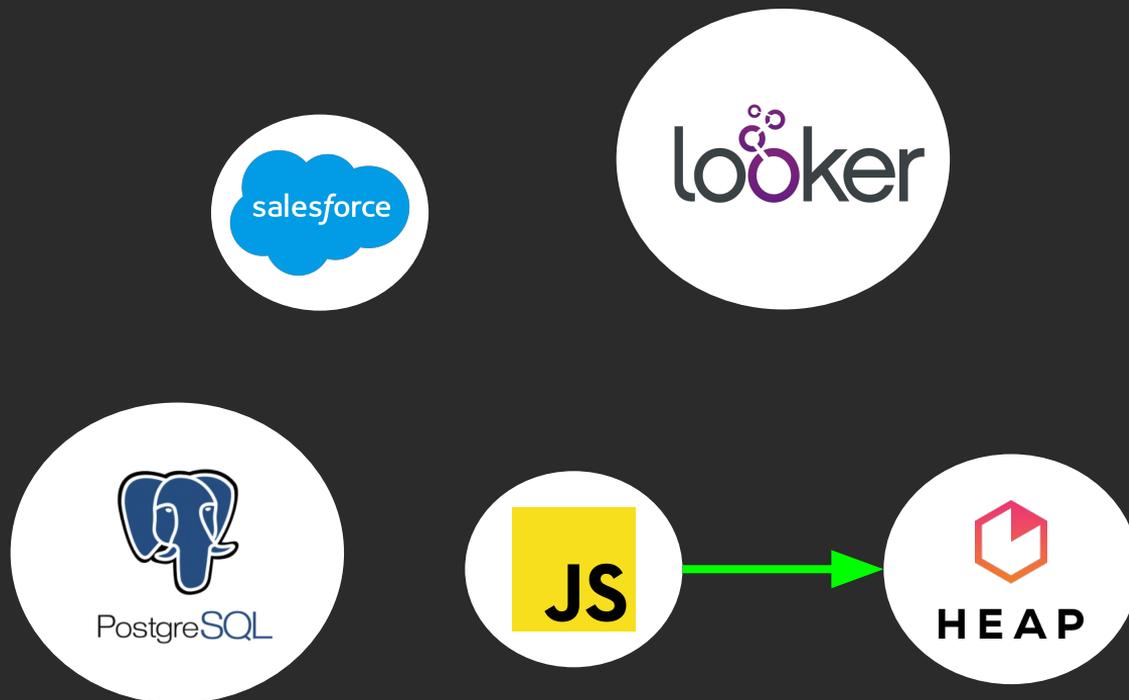
Website  
Click Events

Product DB  
Event Data



TROOPS  
troops.ai

# Starting State: Disjoint Datasets



How do we connect the dots?

Step 1: single source of truth

# Data Warehouse

- So why not just SQL on RDS?
- Best Bets: Amazon RedShift or Google BigQuery



Amazon Redshift



Google BigQuery

Let's set up Redshift

**Redshift dashboard**

- Clusters
- Snapshots
- Security
- Parameter groups
- Workload management
- Reserved nodes
- Events
- Connect client

## Launch cluster

Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse solution that makes it simple and cost-effective to efficiently analyze all your data using your existing business intelligence tools.

[Launch cluster](#)

Note: Your cluster will launch in the US East (N. Virginia) region

## Resources



You are using the following Amazon Redshift resources in the US East (N. Virginia) region (used):

**Clusters (1)**[Increase cluster limit](#)**Snapshots (4)**[Manual \(1\)](#)[Automated \(3\)](#)**Security**[Subnet groups \(1\)](#)[HSM connections \(0\)](#)[HSM certificates \(0\)](#)**Parameter groups (1)**[Reserved nodes \(1\)](#)[Events \(4\)](#)[Event subscriptions \(0\)](#)

**CLUSTER DETAILS****NODE CONFIGURATION**

Provide the details of your cluster. Fields marked with \* are required.

**Cluster identifier\*****Database name****Database port\*****Master user name\*****Master user password\*****Confirm password\*****TROOPS**

troops.ai

Choose a number of nodes and node type below. Number of Compute Nodes is required for m



The ds2 and dc2 node types replace the ds1 and dc1 node types, respectively. The newer ds2 and dc2 node types provide higher performance than ds1 and dc1 at no extra cost. [Learn more.](#)

Node type

CPU 7 EC2 Compute Units (2 virtual cores) per node

Memory 15.25 GiB per node

Storage 160GB SSD storage per node

I/O performance Moderate

Cluster type

Number of compute nodes\*

Maximum 1

Minimum 1



aws Services Resource Groups

Redshift dashboard

Clusters

Snapshots

Security

Parameter groups

Workload management

Reserved nodes

Events

Connect client

CLUSTER DETAILS NODE CONFIGURATION **ADDITIONAL CONFIGURATION** REVIEW

Provide the optional additional configuration details below.

**Cluster parameter group**  Parameter group to associate with this cluster.

**Encrypt database**  None  KMS  HSM [Learn more about database encryption](#)

Configure networking options:

**Choose a VPC**  The identifier of the VPC in which you want to create your cluster

**Cluster subnet group**  Selected Cluster Subnet Group may limit the choice of Availability Zones

**Publicly accessible**  Yes  No Select Yes if you want the cluster to be accessible from the public internet. Select No if you want it to be accessible only from within your private VPC network

**Choose a public IP address**  Yes  No Select Yes if you want the cluster to have a public IP address that can be accessed from the public Internet, select No if you want the cluster to have a private IP address that can only be accessed from within the VPC.

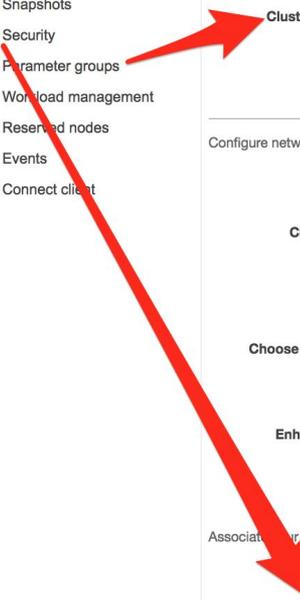
**Enhanced VPC Routing**  Yes  No Select Yes if you want to enable Enhanced VPC Routing. [Learn more](#)

**Availability zone**  The EC2 Availability Zone that the cluster will be created in.

Associate your cluster with one or more security groups.

**Cluster security group** A default security group will be created when this cluster is launched.

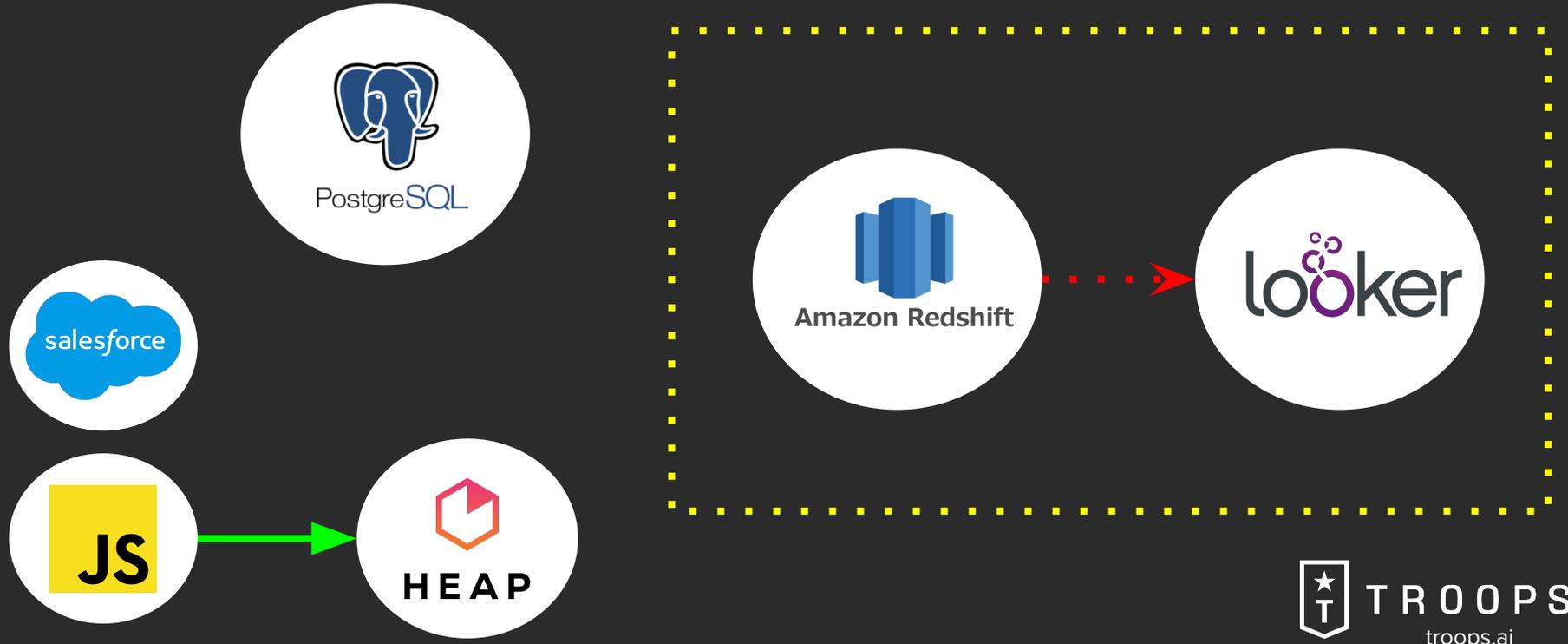
**VPC security groups**  List of VPC security groups to associate with this cluster.



We got a Data Warehouse!



## Step 2: Now let's connect Redshift to Looker



## New Connection

**General**

Settings

Labs

Legacy Features

Support Access

**On****Users**

Users

Groups

Roles

Content Access

User Attributes

**Name \***

Troops Data Warehouse



The name you will use to refer to this connection in your model.

**Dialect \***

Amazon Redshift

**Host:Port \***

troops-dw.gk3nf0sf.us-east-1.redshift.amazonaws.c



5439

**Database \***

dwdb

**Username \***

readonly

**Password**

.....

**Schema**

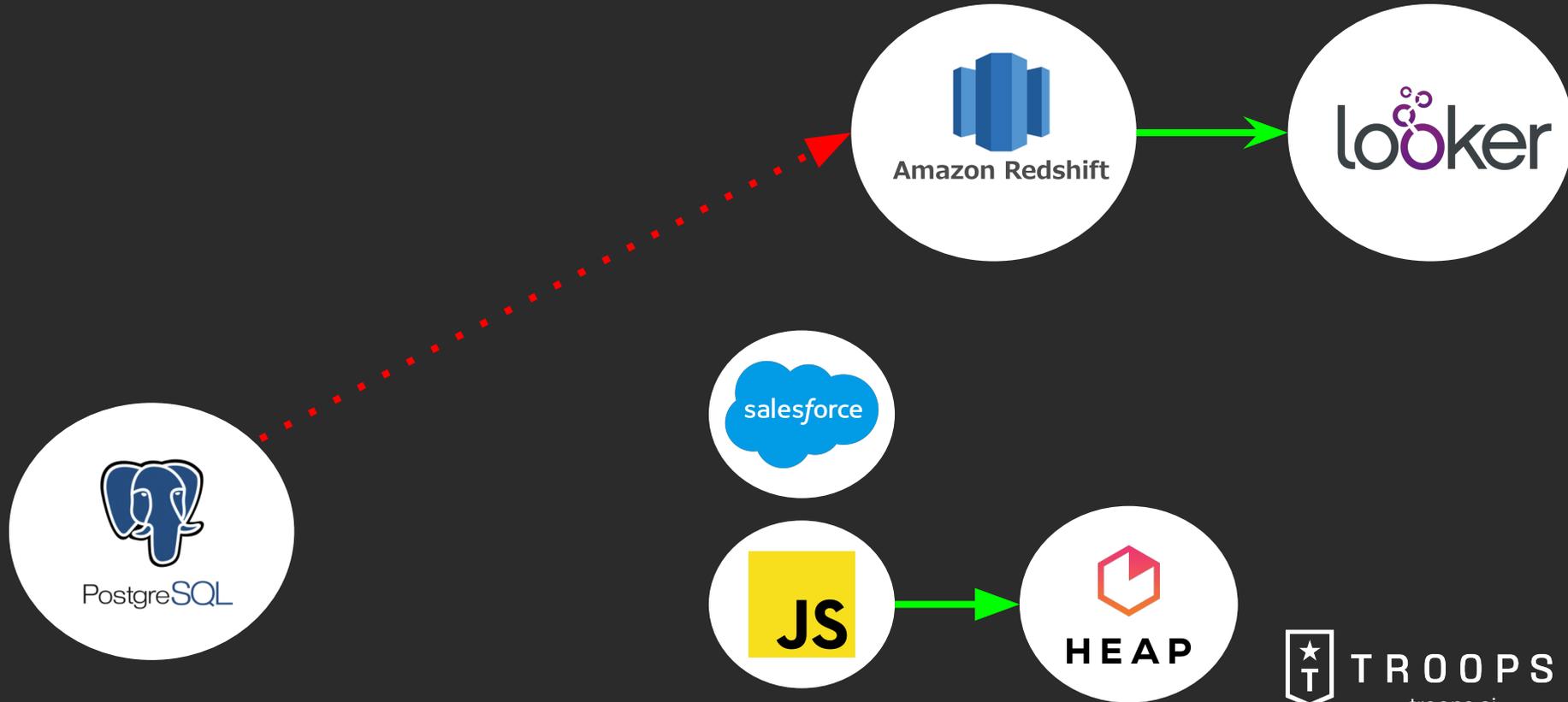
The default schema that Looker will examine in SQL Runner and during LookML project generation.



Looker connected. That was easy!



# How do we get from PSQL to Redshift?



Not so easy



# A few options



alooma



Fivetran



Stitch



xplenty

# Why we chose FiveTran

- Cost efficiency
- Ease of Setup
- Replication performance



Step 3: connect FiveTran to Redshift

SSH Tunnel?  > Database location  > Allow network access

Host

IP (1.2.3.4) or host (your.server.com)

Port

5439

AWS Services Edit

Amazon Redshift

**Clusters**

Snapshots

Security

Parameter Groups

Reserved Nodes

Events

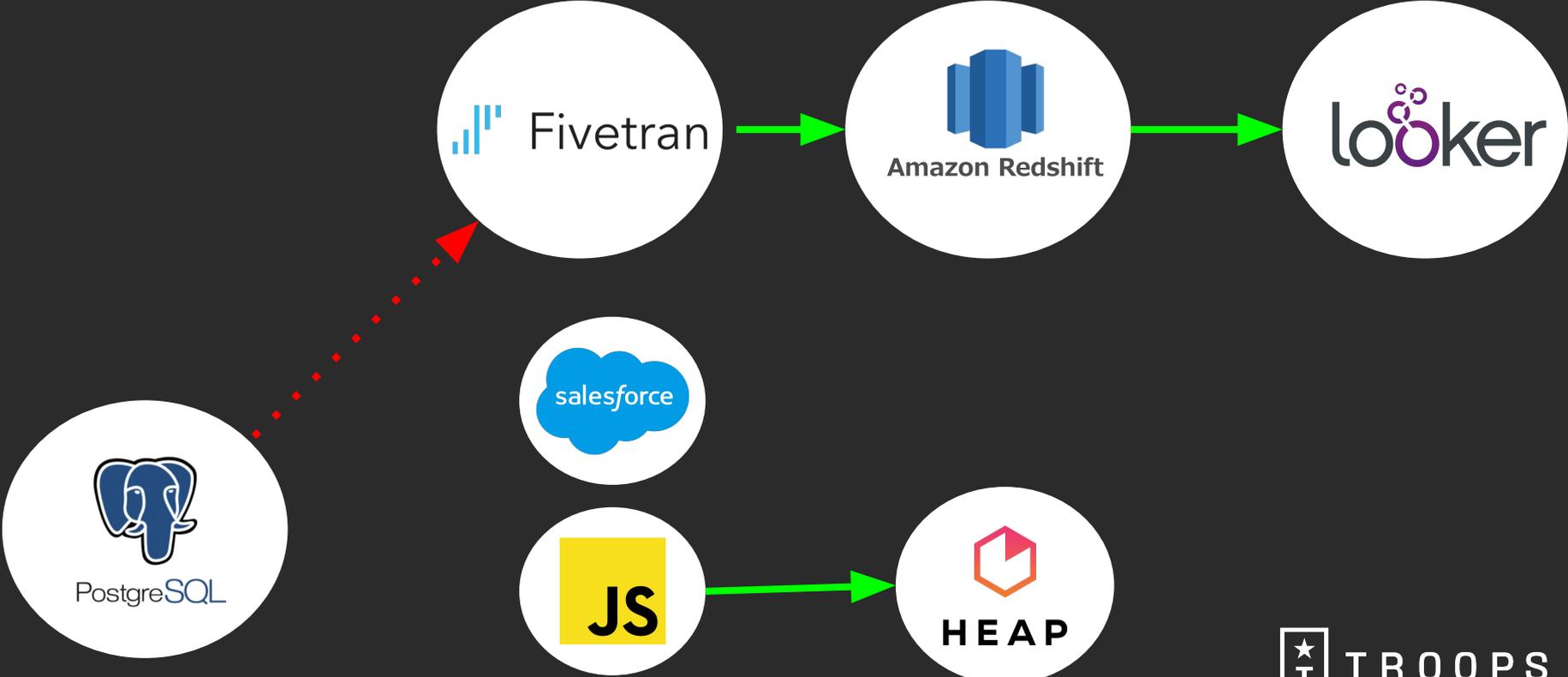
Launch Cluster Manage Tags

Cluster	Cluster Status
production	available

Bam!



# Now PSQL to FiveTran



# Not so fast ...

Troops Warehouse

Support Users Greg Ratner tech@troops.ai

SSH Tunnel?  > Allow network access

Host  
IP (1.2.3.4) or host (your.server.com)

Port  
5432

## Create Read Replica

Creating a read replica is necessary because it allows Fivetran to integrate your data without putting unnecessary load on or interrupting the queries running on your Master server. If you already have a read replica, you can skip to the next section "Enable access".

In your RDS Dashboard, select the Postgres instance you would like to integrate:

AWS Services Edit

RDS Dashboard

Instances

Reserved Purchases

Snapshots

Security Groups

Parameter Groups

Option Groups

Launch DB Instance Show Monitoring Instance Actions

Filter: All Instances Search DB Instances...

	DB Instance	VPC	Multi-AZ	Class	Status
<input checked="" type="checkbox"/>	master-5-6-22	vpc-aff75ac4	Yes	db.t2.micro	available

Step 4: create a read replica



AWS

Services

Edit

## RDS Dashboard

Instances

Reserved Purchases

Snapshots

Security Groups

Parameter Groups

Option Groups

Subnet Groups

Events

Event Subscriptions

Launch DB Instance

Show Monitoring

Instance Actions

Filter: All Instances

Search DB



DB Instance

VPC

Status



master-5-6-22

vpc-aff75ac4

available

Modify

Reboot

Delete

Create Read Replica

Promote Read Replica

Take DB Snapshot

Restore to Point in Time

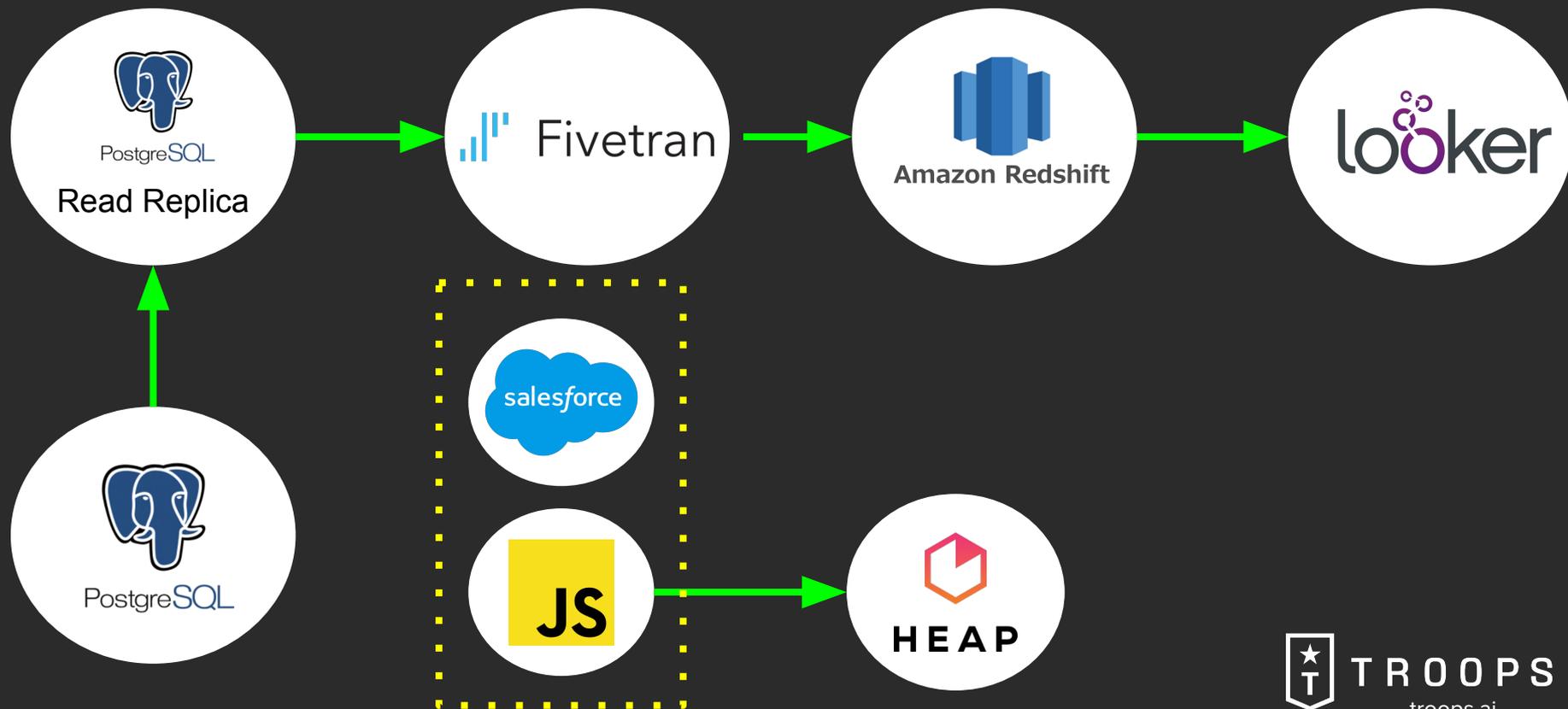
See Details



TROOPS

troops.ai

# What about other data sources?



# Segment to the Rescue!



Step 5: connect Redshift to Segment



## Destinations Catalog

🔍 Type to search...

### FEATURED

Most Installed

New & Noteworthy

### CATEGORIES

All

A/B Testing

Advertising

### Warehouses



Redshift



Postgres



BigQuery





FEATURED

Most Installed

New & Noteworthy

CATEGORIES

All

A/B Testing

Advertising

## Step 2: Enter your credentials

You'll want to copy and paste these credentials from your warehouse provider. If you're not sure where to go, you can follow this guide or you can invite someone else to help you.

**Username**

**Password**

**Host**

**Port**

**Database name**

Connect



Q Type to search... View All Add Destination

### Warehouses

Name ^	Status	Sources	
 redshift	● Enabled	  +2	...

Success!



Step 6: connect Salesforce & Heap

Troops Team Plan

Sources Destinations

### Sources Catalog

CATEGORIES

- All
- Advertising
- Analytics
- CRM**
- Email
- Enrichment
- Helpdesk

### CRM



**Salesforce** >



**HubSpot** >

Sources Destinations

### Destinations Catalog

Type to search...

[View All](#)



**Mixpanel** >  
Analytics



**Amplitude** >  
Analytics



**Heap** >  
Analytics



**Keen IO** >  
Analytics



**KISSmetrics** >  
Analytics



**Amazon S3** >  
Analytics



**HubSpot** >  
Analytics

< CANCEL

## Add Salesforce

NICKNAME

Salesforce

Pick a name to help you identify this source in Segment.

SCHEMA NAME

sfdc\_prod

This is how you will query this source when using SQL.

1

```
SELECT * FROM sfdc_prod.accounts
```

CONNECT

# Simple wizard interface



TROOPS  
troops.ai

Step 7: replace Heap JS with Segment JS

## Source setup

### Name \*

Identifies this source within your workspace, and typically includes the product area and environment. E.g. Website Prod or App Dev.

### Warehouses

Choose to sync this data to your warehouse below. The schema name will be **website\_events\_js** in your Warehouses ([change](#)). You will have the option to connect to additional destinations later.

redshift

### Website URL

The full URL where you will install analytics.js.

Add Source



Step 8: profit!



### Sources

Add Source



#### Javascript (3)

Sending to 12 Destinations



#### Salesforce

Sending to 1 Destination

[View Sources](#)

### Destinations

Add Destination



#### Redshift

Receiving from 4 Sources



#### Intercom (3)

Receiving from 3 Sources

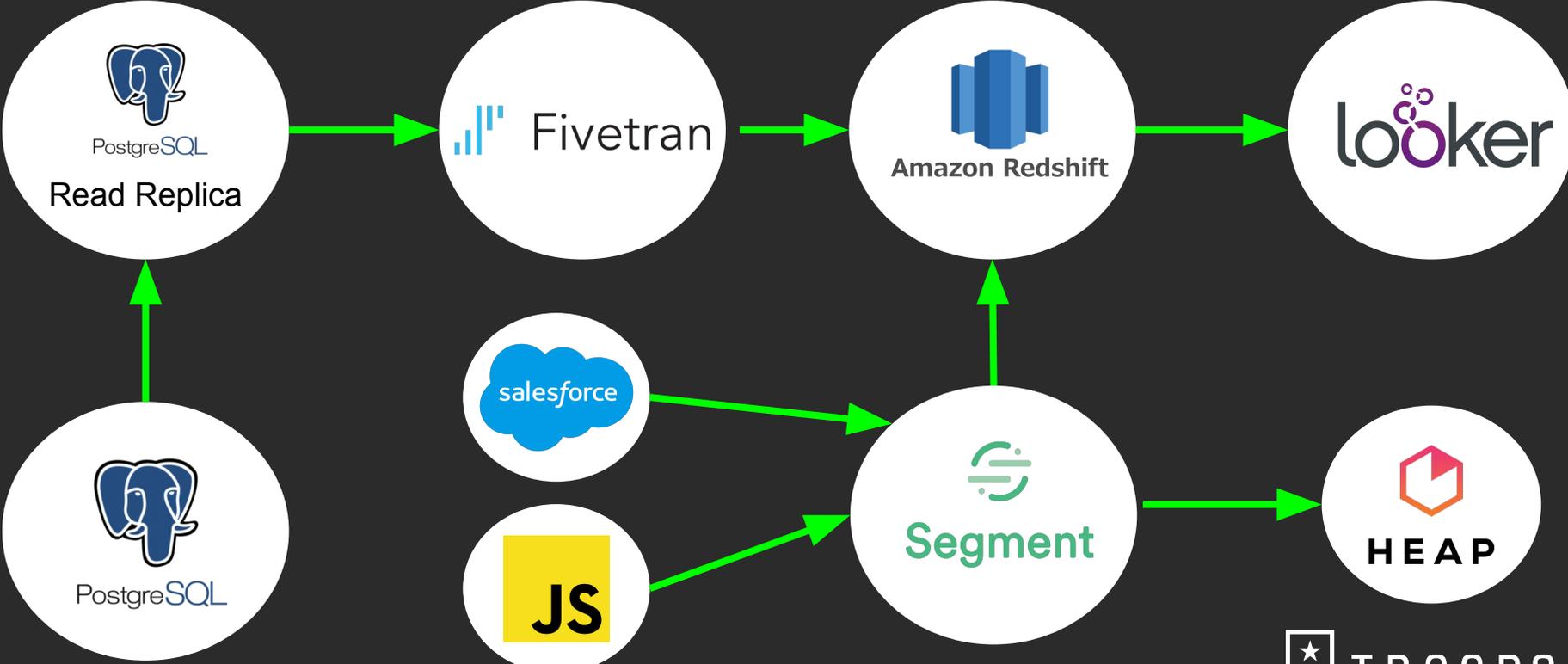


#### Heap (3)

Receiving from 3 Sources



# Final State



# Stats

- 115 Looks created
- 23 Dashboards
- 70% of the company created looks

# Results

- Measurable quarterly OKRs
- Transparency in feature performance
- Transparency in sales performance
- Faster investor updates and board decks
- Increased ownership and accountability

# Thanks!

Greg Ratner CTO @ Troops

Twitter: @GregRatner

