Reddit R/Place Insights

By: Dakota Brown

What is the problem to solve?

Reddit's R/Place had only four guidelines:

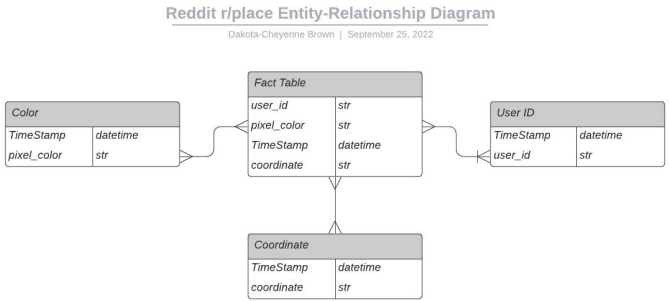
- There is an empty canvas.
- You may place a tile upon it, but you must wait to place another.
- Individually you can create something.
- Together you can create something more.

With over 30 GBs of data extracted from Reddit, I created a data pipeline to transform data from their 2017 and 2022 to see if there were any insights to be gained between the two years such as growth in the platform and whether or not the top users were bots.

Cleaning/Transformations

- Looking at the data during Exploratory Data Analysis, the columns between the two years were were similar but with some discrepancies.
- Another thought was trying figure out the best way to model data with very few columns.

I thought the best way to do so was to use a star schema.



Cleaning/Transformations cont.

I decided to use PySpark on an EMR cluster to fix the 2017 data to match the 2022 data.

With that, I had to:

- Change the formatting of the colors
- Combine the coordinates into one pair

Due to the amount of data I needed to make sure I used extra large clusters with enough computing power and memory to handle 30 GBs of data and cache the data frames in place during transformations.

Testing

When testing the code, I needed to be sure that I pulled the data correctly from reddit and that it was stored in S3 without any problems.

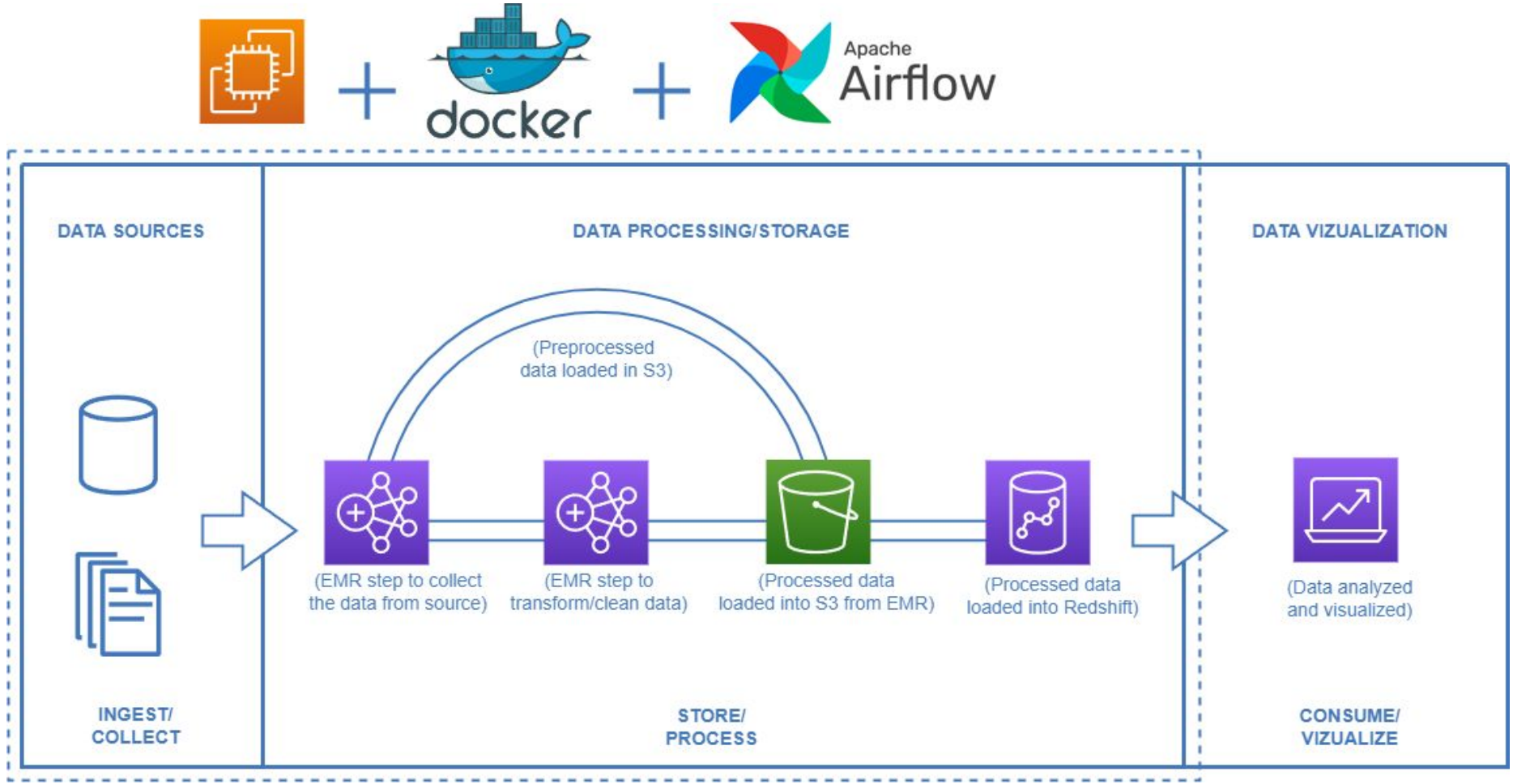
- With the 2017 data it was stored in one pdf
- With the 2022 data it was compressed and split into almost 80 different pdfs
- Extract the data from either a zip or gzip

Through my testing I was able to assure that all of the data was pulled correctly, extracted from zips and gzips, and stored correctly into S3 before and after transformations

Architecture

When choosing the architecture for my project, I wanted it to be reusable, redeployable, and able to be ran by other engineers easily. I also needed to be sure the instances I used were correctly sized.

- I chose to use Apache Airflow for it use of DAGs and so it could be ran on a schedule
- From there an EMR instance was created that ran ETL code from S3
- The data was extracted into an S3, transformed in PySpark, and loaded back into S3
- All of this was containerized in Docker and ran on an EC2 instance


docker
Apache
Airflow
DATA SOURCES
DATA PROCESSING/STORAGE
DATA VIZUALIZATION
(Preprocessed
data loaded in S3)
(EMR step to collect
the data from source)
(EMR step to
transform/clean data)
(Processed data
loaded into S3 from EMR)
(Processed data
loaded into Redshift)
(Data analyzed
and visualized)
INGEST/
COLLECT
STORE/
PROCESS
CONSUME/
VIZUALIZE

Metrics

Some of the things I wanted to track in this project were:

- The growth (if any) between the two years
- Top users and their most used colors in 2017 and 2022
- The top 5 colors used in 2017 and 2022
- The bottom 5 colors in 2017 and 2022

The biggest metric would be top users and their colors to see if there was any bot activity.

Insights

Since the usernames were hashed, it’s harder to see if there was a difference between bot activity or a concerted effort by a fandom of individuals wanting to make their art look well.

At the end of the day, there was a ten time increase in both number of users as well as number of pixels placed on the canvas from 2017 to 2022.

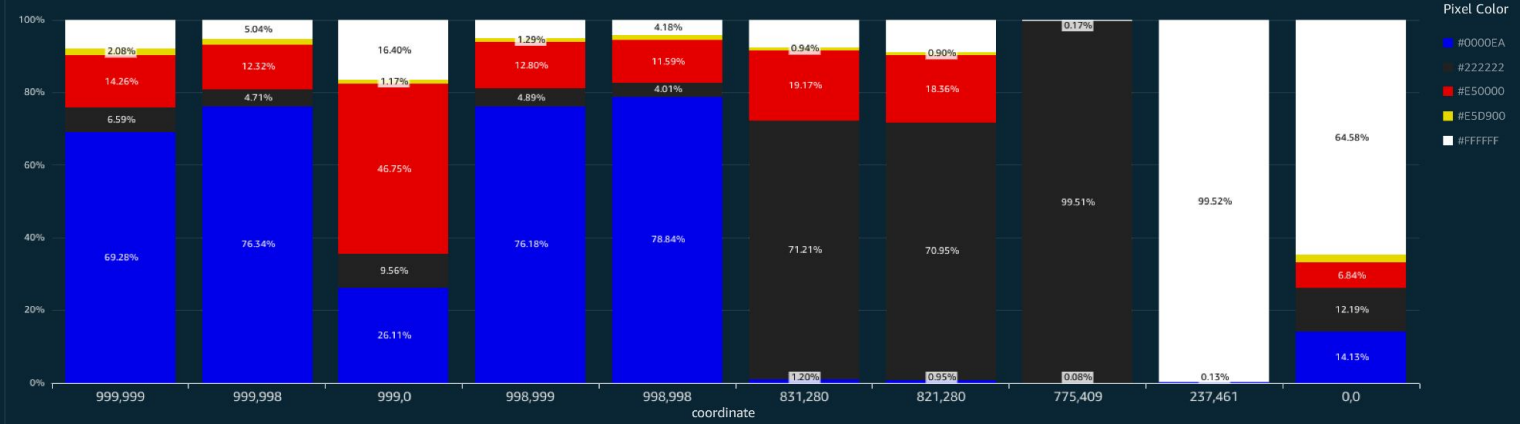
Total Number of Pixels Placed in 2017	Total Number of Pixels Placed in 2022
16,567,567	160,353,104
Distinct Number of Colors Used in 2017	**Distinct Number of Colors Used in 2022**
16	32
Distinct Number of Users in 2017	**Distinct Number of Users in 2022**
1,166,940	10,381,163

August 18, 2022 2:43 AM (GMT)

Top 10 Contested Coordinates and Its Colors in 2017

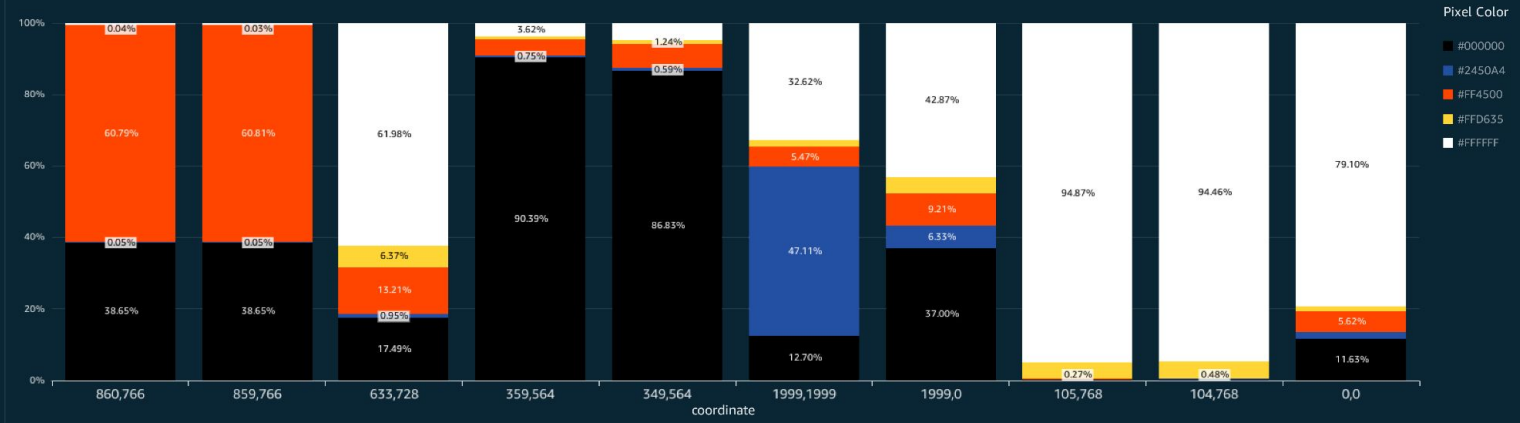
(Including #222222 and #FFFFFF)

SHOWING TOP 10 IN COORDINATE AND TOP 5 IN PIXEL_COLOR



Top 10 Contested Coordinates and Its Colors in 2022

(Including #000000 and #FFFFFF)



August 18, 2022 2:44 AM (GMT)

Powered by QuickSight