Peter Meleney
DATA 516
November 30, 2019

Project Title

Abstract

Fractal is a company dedicated to protecting photographers' intellectual property online. To that end Fractal requires that their customers only upload photos that are their own original work. Fractal has developed several technical solutions to ensure that photos that are uploaded are unique, and not derivatives or copies of work already on the site, one of these features involves using Locality-Sensitive Hashing (LSH) to identify photos that have a high likelihood of being copies of work already on the site. Fractal maintains a database of 1000 hashes of every photo on the site, this database must be searched every time a new photo is uploaded. The speed of the search is paramount because this feature will determine the length of time a customer is waiting before an uploaded image is accepted onto the platform. This paper investigates the speed and scalability of two leading cloud-based Relational Database Managements Systems (RDBMS): Spark and Snowflake in searching for duplicate LSH hashes in databases of variable size.

1 Introduction

Fractal is a company dedicated to protecting photographers' intellectual property online. Fractal runs a website which guarantees that once an image is uploaded, no other copy or derivative of that image may be uploaded for as long as the image is maintained. Fractal achieves this in two ways: (1) images are digitally watermarked and if the watermark, or any part of it, is detected in an uploaded image, that image is rejected by the website's back-end, and (2) Fractal stores 1000 LSH hashes of each uploaded image, if an image with an identical LSH is uploaded to our servers, we subject that image to further screening to determine if it is copy or a derivative of a previously uploaded image.

LSH is used for similarity searching at scale by large technology companies like Uber**??** and Pinterest**??**. LSH is particularly useful because by tuning the parameters b and r, where b is the number of bands, and r is the number of rows per band, we can finely control the proportion of false positives. These parameters are important to data curation because they control the number of hashed values that must be stored in the database.

2 Data

For our experiment we will use photos from the 2017 Common Objects in Context (COCO) image dataset. Created by Microsoft, this repository is free to access online at http://cocodataset.org/. Since we are not creating a computer vision model in this application, I combined the Train, Test, Val, and Unlabeled repositories, for a total of 287,360 unique images of common objects. The images are in color, are of variable size, and together are over 46.8GB on disk.