

Final Project Report

Shunchang Wang | Big Data Remote Course

Professor Fan Zhang| 10/19/2018 – 11/19/2018

Content

[1. Introduction 2](#_Toc1308988)

[2. Basic Skills Learning 2](#_Toc1308989)

[2.1 Python 2](#_Toc1308990)

[2.2 Github 2](#_Toc1308991)

[2.3 Linux Command 3](#_Toc1308992)

[2.4 Introduction of mnist 3](#_Toc1308993)

[3. Service 4](#_Toc1308994)

[3.1 Docker 4](#_Toc1308995)

[3.2 Flask 4](#_Toc1308996)

[3.3 RESTful web service 4](#_Toc1308997)

[4. Big data storage 4](#_Toc1308998)

[4.1 NoSQL 4](#_Toc1308999)

[4.2 Cassandra 5](#_Toc1309000)

[5. Data Processing 5](#_Toc1309001)

[5.1 SPARK 5](#_Toc1309002)

[5.2 IMPORTANT tools for visualize data 6](#_Toc1309003)

[5.3 OVERVIEW of the Cloud Native 6](#_Toc1309004)

[6. PROJECT – Mnist 6](#_Toc1309005)

[Introspection 7](#_Toc1309006)

# Introduction

Big data is a term used to refer to data sets that are too large or too complex for traditional data-processing application software to adequately deal with. Data with many cases offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. Other concepts later attributed with big data are veracity and value. Through the program, we learned more about the service, data storing and processing to finish the project. In the program, we learn to build a service recognize user’s handwrite input and turn to a digital data to store in Cassandra. And we can use Cassandra to check the data we stored.

# Basic Skills Learning

In this part, I will give a brief introduction of the basic things we need to have.

## Python

Python is an interpreted, high-level, general-purpose programming language. In Python, everything is an object, and data type is a property of the object. Lists, Tuples and Dictionaries are the data types. Also, booleans, if statement and for loops were explained in the program.

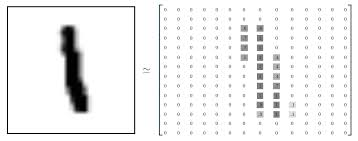
## Github

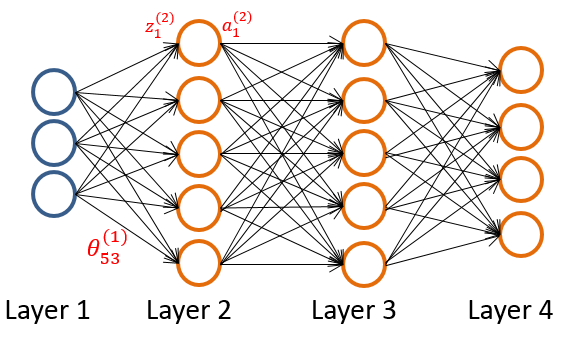
GitHub is Inc. is a web-based hosting service for version control using Git. It is mostly used for computer code. It offers all the distributed version control and source code management functionality of Git as well as adding its own features. In the program, we understand the difference between central server and remoter server. In addition, we get familiar with the way to use branching.

## Linux Command

Linux is a family of free and open-source software operating systems based on the Linux kernel. In the course, the professor teaches us the common commands for the Linux system.

## Introduction of mnist





Mnist based on the neural network, the input will go through many layers of the network and train to make predicate output more and more close to the output we want. After hundreds of training or more, it will get accurate results.

# Service

## Docker

Docker is a computer program that performs operating-system-level virtualization, also known as "containerization". It is an open platform for developers and sysadmins to build, ship and run distributed applications. Its features include light-weight, portable, self-sufficient. In the program, we learn to use docker as container to build a service for the project. More details on <https://docs.docker.com/get-started/part2/>.

## Flask

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. We can use flask to build a service for other people to use.

## RESTful web service

The REST architecture was originally designed to fit the HTTP protocol that the world wide web uses. For instance, the Uber use the Google map service to locate passages and drivers. More information we can get in <https://blog.miguelgrinberg.com/post/designing-a-restful-api-with-python-and-flask>

# Big data storage

Big data system application involving three V’s, volume, velocity and variety, we need the system be availability in multi region and can response very fast and reliable, with no single point of failure.

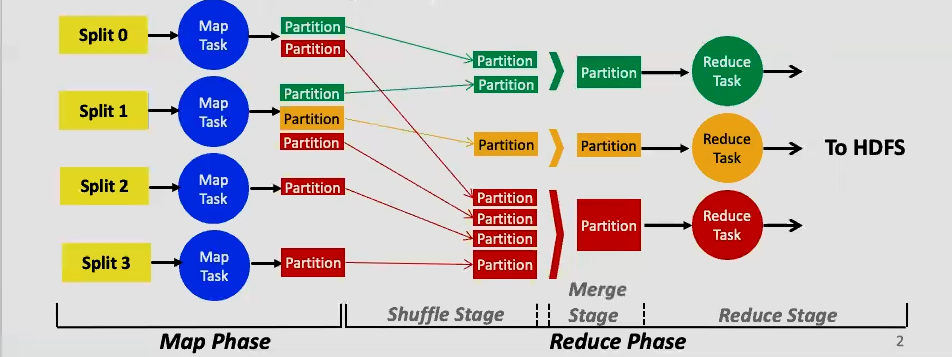
## NoSQL

There were four broad classes of non-relational database: Graph, Key-Value, Document and Column Family. Its key factors include consistency, availability and Partition Tolerance to navigate its landscape.

## Cassandra

Cassandra’s database can massively linearly scalable NoSQL database with highly performant. And it has no single point of failure due to horizontal scaling. Cassandra is particularly useful for eBay, Spotify and Instagram. Its data is modeled using Cassandra Query Language(CQL), and we learn more about it in the following website: <http://cassandra.apache.org/doc/latest/cql/index.html>

* 1. Hadoop MapReduce



MapReduce is a pervasive data processing framework on the cloud. Hadoop is an open source implementation of MapReduce. Hadoop MapReduce incorporate two phases, encompass multiple Map and Reduce tasks.

# Data Processing

Data processing’s goals are low latency queries on historical data/ low latency queries on live data / sophisticated data processing.

## 5.1 SPARK

Spark is In-memory framework for interactive and iterative computations. It is easy for spark to combine batch, streaming and interactive computations and develop sophisticated algorithms. It uses RDDS (resilient distributed datasets) to retain fault tolerance, data locality and scalability.

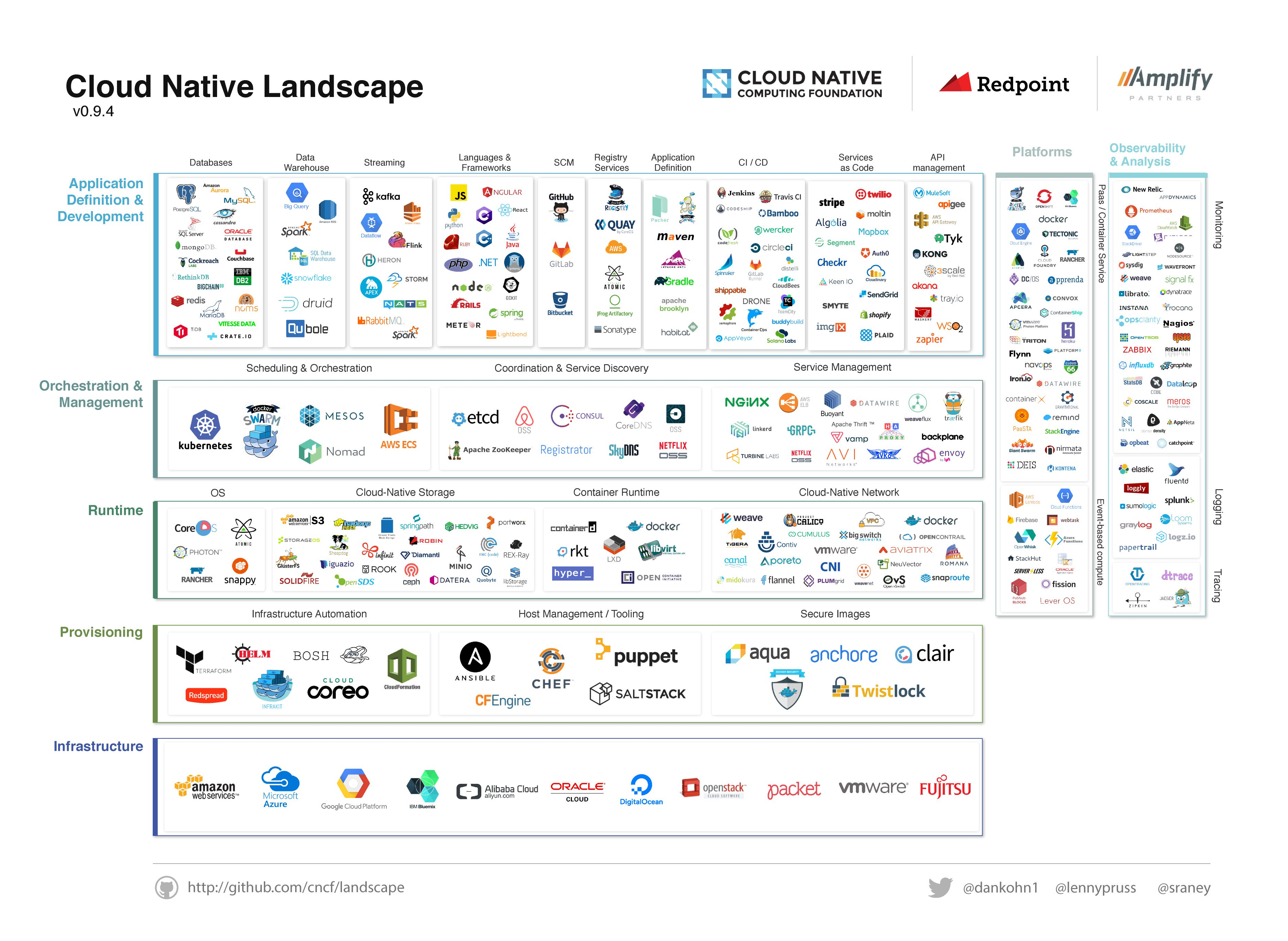
## 5.2 IMPORTANT tools for visualize data

AMCharts a website used for data visualization.

Data-Driven Documents another website to visualize the data

PlayGround Tensflow a website is helpful for us to visualize the project.

## 5.3 OVERVIEW of the Cloud Native

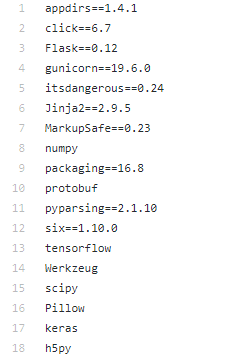
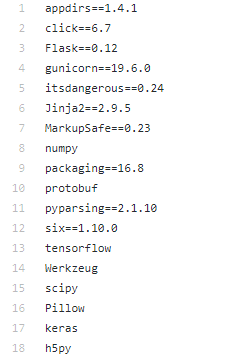


In this picture, it shows us the all the tools and applications for the Big Data.

# PROJECT – Mnist

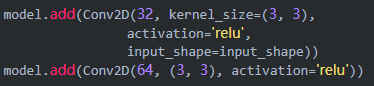
Goals: We use the mouse to draw the number in the webpage generate by the docker and python file and the webpage will give the result of the handwrite. Also, it will store the output in the Cassandra by using the CQL to make the table for the input.

Start: we need to install the required package and make sure they are latest.



As we can see, we use Keras to train the model, Keras is an open source neural network library written in Python. Thus, by using Keras, we build the model step by step.

model = Sequential()

and the core part is the conv2D layer, which can build like this, 

And then we will add maxpooling to prevent overfitting



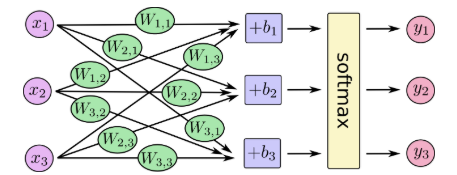
Next steps will enter the dense layer, firstly we need to build a flatten layer



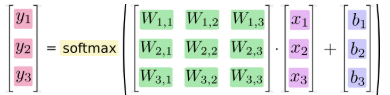
And then we will enter the dense layer and build a dropout layer to prevent overfit

In the last, we can have fully connected layers of softmax function

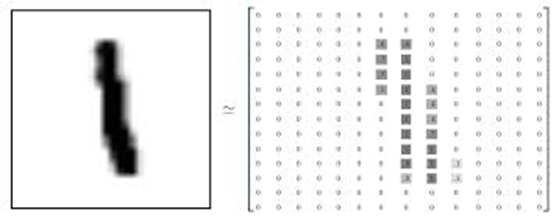
And SoftMax regression function can be explained in the following pictures



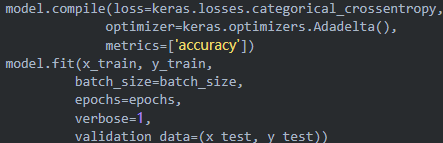
We can write SoftMax into the equation



Therefore we can see how the SoftMax works for recognize the handwrite digits.



Thus, we can build a train for the model by the following codes

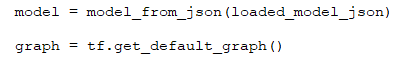


This model will be stored in another python file we named load.py

We use json (a lightweight data-interchange format) to store and load the model

Store 

Load

Therefore, we can generate the model and graph for the next steps

In the last python file, we can directly use the model and graph. But first we need to connect to the Cassandra database, create container by the docker:

**docker run --name some-cassandra -p 9042:9042 -d cassandra:latest**

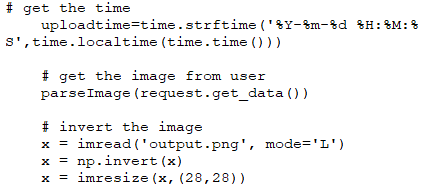
**docker run --name some-cassandra -p 9042:9042 -d cassandra:latest**

simply change “some” to the container name will works.

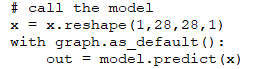
And we need write the code in the python file to connect the Cassandra database,



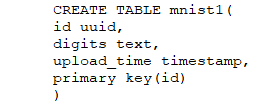
We can get the input by the following code



At this time, we can call the model and graph to predict the result

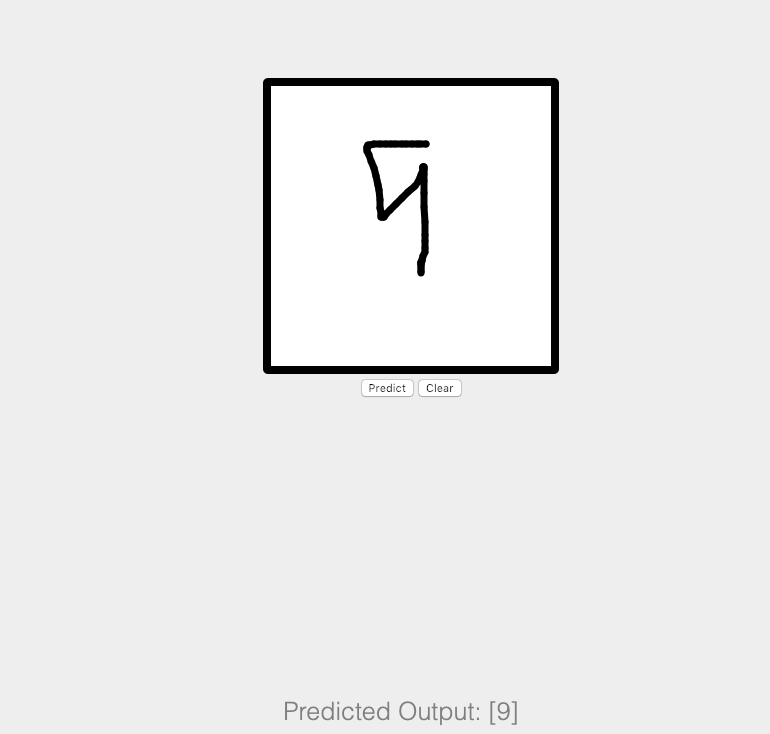
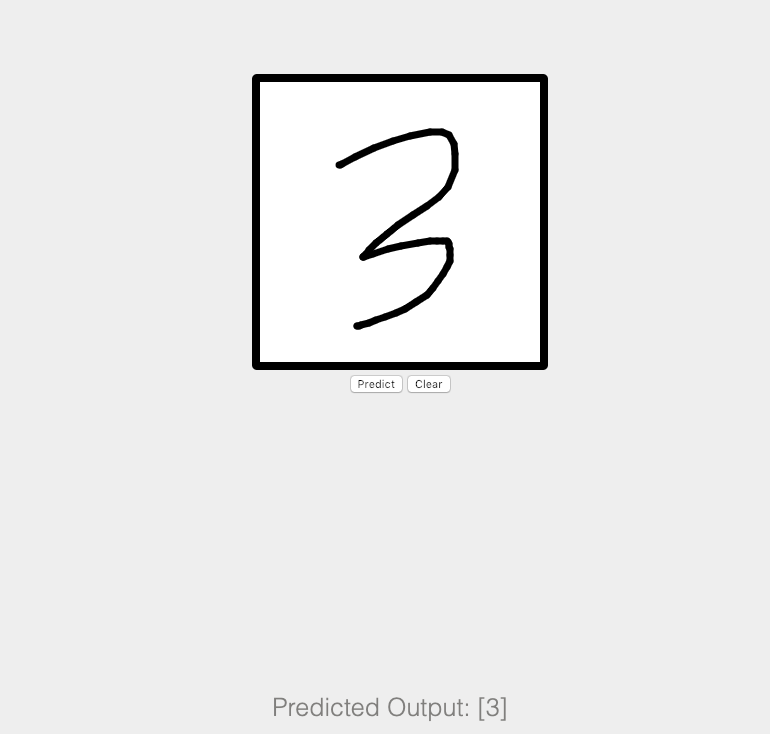


We can build a database after we get the time, input and output



As a result, we can store and read the database by Cassandra.

Output:

Like in these pictures, I draw the digits and it shows the predicted output.

## Introspection

Through the program, I learned the importance of big data and why people develop many tools to handle the problems of Big Data. Before taking the class, big data for me is just a term. I do not even know it is really close to us, like Amazon, eBay and Google, these companies ‘website rely on the big data. After the class I feel the professor give us the way to explore the big data area, the tools and platform he recommends I feel is helpful. If I have other chances, I would like to take his courses again.

Besides, I see the limitation of my coding. I do not know much about python code, even though it has some similarities between Java code, it is a tough thing for me to learn and handle. However, the experience to use the code is not bad, I can understand this code and I get some ideas to write the python code in the future. Docker, Flask and Cassandra, I feel I will use them sometimes in the future. Big Data is an area filled with challenges and potentials, thus I really appreciate the professor Zhang can be the guide for me to get in touch with the big data. Therefore, even with heavy courses in the college, I will sometimes learn the big data in my spare time.