CONCRETE COMPRESSIVE STRENGTH PREDICTION

ARCHITECTURE

REVISION No. : 1.0
LAST DATE OF REVISION : MAY 2ND,2022

DOCUMENT VERSION CONTROL

DATE	VERSION	DESCRIPTION	AUTHOR
02-05-2022	1	PROJECT ARCHITECTURE	MOHD. USAMA

TABLE OF CONTENTS

1 Introduction

1.1 What is Architecture Design Document?

Any software needs the architectural design to represent the design of the software. It defines architectural design as "the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system. Machine Learning architecture is defined as the subject that has evolved from the concept of fantasy to the proof of reality.

2 Scope

Machine learning solutions are used to solve a wide variety of problems, but in nearly all cases the core components are the same. Whether you simply want to understand the skeleton of machine learning solutions better or are embarking on building your own, understanding these components - and how they interact - can help. Whenever we try to create architecture, we try to create Development, UAT(User Acceptance Testing Environment), and Production architecture.

We have to look for couple of things in term of architecture as follows:

- **Latency:** Latency is the delay or time taken between a user's action and a web application's response to that action.
- **Frequency:** Based on frequency of incoming data it could be Batch, Minibatch or Streaming Data.
- **Scalability:** Scalability is a characteristic of a software system or organization that describes its capability to perform well under an increased workload. A system that scales well can maintain its level of efficiency during increased operational demand.
- **Security:** The security of a system is a crucial task. It is a process of ensuring the confidentiality and integrity of the data.

Architecture

Based on Latency, Frequency, Scalability and Security, we select couple of things like :

- **Database :** Data is only useful if it's accessible, so it needs to be stored – ideally in a consistent structure and conveniently in one place such as in databases like Cassandra, MongoDB, MySQL, etc.

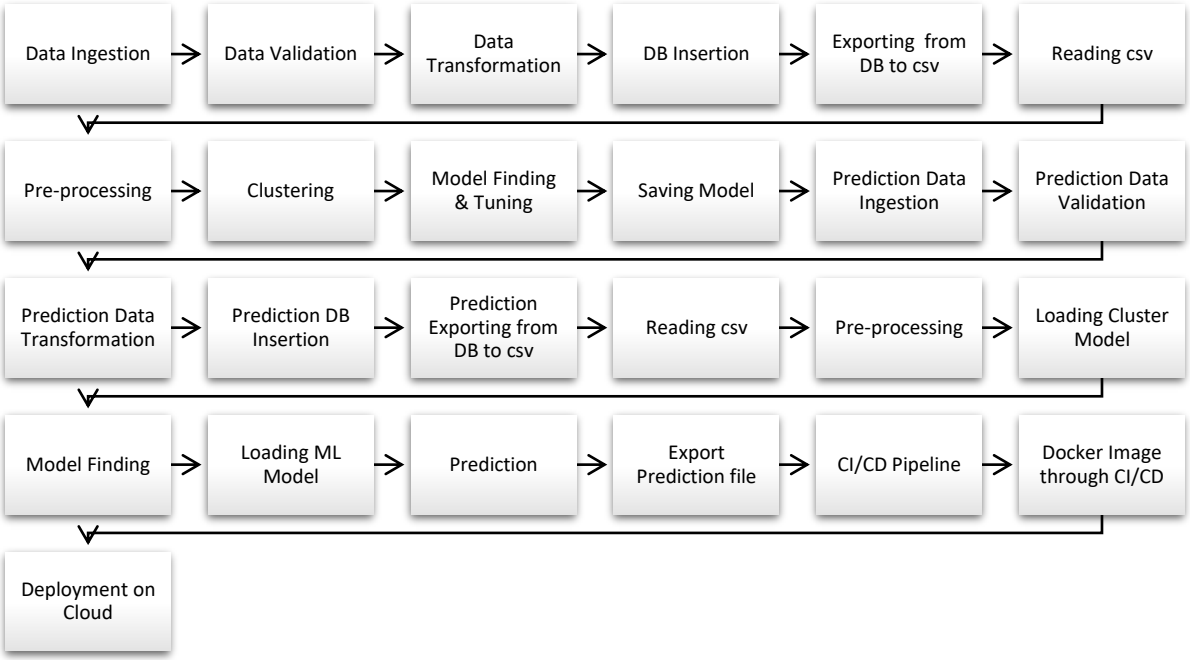
- **API :** We need some way to interact with our model and give it problems to solve. Usually this takes the form of an API, a user interface, or a command-line interface.

- **Cloud Platform :** Feature engineering, training, and prediction all need to be scheduled on our compute infrastructure (such as AWS or Azure) – usually with non-trivial interdependence. So we need to reliably orchestrate our tasks. Orchestration is the configuration of multiple tasks (some may be automated) into one complete end-to-end process or job.

- **Monitoring Tool :** We need to regularly check our model's performance. This usually involves periodically generating a report or showing performance history in a dashboard.

- **Security Tool :** Authentication keeps our models secure and makes sure only those who have permission can use them.

3 Architecture



3.1 Architecture Description

- **Data Ingestion :** Dataset is taken from Kaggle link provided by I-Neuron
- **Data Validation :** Validation checks have been made like Name and Number of Columns, File name, Missing values in columns, etc.
- **Data Transformation :** Columns have been renamed to shorter name for easy access. No other necessary transformation required in the dataset.
- **Database Insertion :** After validation of the data and doing necessary transformations, the data is pushed to Cassandra Database. This is the last step in Training Raw Data Validation process.
- **Exporting from DB to csv :** Validated data that was pushed to Cassandran DB is retrieved into csv file. This csv file is used to train our model after necessary steps.
- **Reading csv :** The csv file that we received is now accessed using pandas
- **Data Pre-processing :** Dataset is checked for missing values, duplicate values, skewed data, outliers, scales, distributions, etc . We have found only duplicate rows are present which will not contribute anything for prediction, so have removed those rows. There are no other pre-processing steps required in this dataset.

- **Clustering :** Clustering of model is done to make the clusters of data having similarity between them.
- **Model Finding :** Model is trained based on each of the cluster dataset using appropriate ML algorithms and out of these models the best one is selected based on the R2 score for each of the clusters.
- **Model Tuning :** After finding out the best model for each of the clusters, we do hyperparameters tuning of those models to get better score. And then based on the best parameters, the model is trained.
- **Saving Model :** After training we save our model which can be used for prediction of unknown data.
- **Prediction Steps :** When we receive the new data from client for which we have to do predictions, we do all the Data Validation, Data Ingestion, Data Transformation, DB Operations, and Pre-processing steps, and then we will load our cluster model that he have created during Clustering step in Training and then using that model, it will make the clusters of prediction dataset and then based on that cluster number we use the cluster specific trained model to do the predictions of cluster and then we append those results into a single file which we have to save for Client.
- **CI/CD Pipeline :** Circle CI is used for the purpose of CI/CD operations which is used to create docker image for DockerHub and to deploy on cloud automatically just by committing our code to GitHub Repository.
- **Deployment :** The project is deployed to AWS and Heroku cloud platforms.