Big Data Processing and Data Analytics

Sheik Mohammed Shaw S, Navin Tharmaraj E , Dr.R.Karthikeyan

Students-Department of CSE-PSNA,Dindigul

Head of Department-Department of CSE-PSNA, Dindigul

Abstract— All applications in current trends need to use Machine Learning and Big Data in this huge data world.Machine learning is a type of artificial intelligence that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.Machine learning use historical data as input to predict new output values.Small Scale and even large developers cannot maintain Machine Learning Algorithms and other big data processing techniques all by themselves.To harvest, process and analyze the data we need a data processing engine.Our project is to provide them with a machine learning engine which runs on.a specified port on their servers which collects the data sets in real time.This which can be queried using a simple Query Language and analyze the database on the query.The purpose of our project is to create a Data Processing Engine for Big Data Analytics. Machine learning and Big Data plays a major role in the current huge data world. Our project is to maintain a dedicated machine learning and data processing engine, so that even small developers without the knowledge of machine learning and big data analysis can produce expected results based on Machine Learning to make their application smooth. All applications in current trends need to use Machine Learning and Big Data in this huge data world. Machine learning is a subset of artificial intelligence that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning uses historical data as input to predict new output values. Small Scale and even large developers cannot maintain Machine Learning Algorithms and other big data processing techniques all by themselves. To harvest,

process and analyze the data we need a data processing engine. This project is to provide them a data processing engine which runs on a specified port on their servers which collects the data sets in real time. This can be queried using a simple Query Language and analyze the data and produce respected results using ML based on the query

I. INTRODUCTION

The history of databases begins with the 2 earliest computerized examples. Charles Bach man designed the primary computerized info within the early Sixties. This 1st info was called the Integrated information Store, or IDS. This was shortly followed by the data Management System, an info created by IBM.

Database process was originally utilized in major firms and enormous organizations as the basis of enormous transaction-processing systems. Later, as microcomputers gained quality, info technology migrated to micros and was used for single-user, personal info applications. Next, as micros were connected along in work teams, info technology rapped to the workgroup setting. Finally, database square measure is getting used nowadays for web and computer network applications. As indicated antecedent, a management system (DBMS) may be a cluster of programs used as an associate degree interface between an info associate degree and an applications program. DBMS's square measure is classified by the kind of info model they support. A relative package would follow the relative model, for instance. The functions of a package embrace information storage and retrieval, info modifications, information manipulation, and report generation. A information definition language (DDL) may be an assortment of directions and commands accustomed to outline and describe information and data relationships during a specific info. File descriptions, space descriptions, record descriptions, and set descriptions square measure terms the DDL defines and uses.

Machine learning was 1st formed from the mathematical modeling of neural networks. A paper by logistical Bruno Walter Pitts and neurobiologist Warren McCulloch, printed in 1943, tried to mathematically project thought processes and deciding on human psychological features. From now on, "intelligent" machine learning algorithms and pc programs began to seem, doing everything from coming up with travel routes for salespeople, to taking part in board games with humans like checkers and tit-tat-toe. Intelligent machines went on to try everything from mistreatment speech recognition to learning to pronounce words the method a baby would learn to defeating a world chess champion at his own game. The infographic below shows the history of machine learning and the way it grew from mathematical models to stylish technology.

II.FUNCTIONAL ARCHITECTURE

Django was created within the fall of 2003, once the net programmers at the Lawrence Journal-World newspaper, Adrian Holiday and Simon Willis on, began exploitation Python to make applications. Jacob Kaplan-Moss was employed early in Django's development shortly before Simon Willis on's billet finished.[16] it had been free in public beneath a BSD license in July 2005. The framework was named once player Django Reinhardt.[17] Adrian Holiday could be an itinerant jazz player and a giant fan of Django Reinhardt

In June 2008, it had been proclaimed that a freshly shaped Django package Foundation (DSF) would maintain Django within the future.

Despite having its own language, like naming the due objects generating the protocol responses "views",[7] the core Django framework will be seen as associate degree MVC design.[8] It consists of associate degree object-relational plotter (ORM) that mediates between information models (defined as Python classes) and a on-line database ("Model"), a system for process protocol requests with an on-line emulating system ("View"), and a regular-expression-based uniform resource locator dispatcher ("Controller").Also enclosed within the core framework are:

- a lightweight and standalone net server for development and testing
- a kind serialization and validation system which will translate between markup language forms and values appropriate for storage within the information
- an example system that utilizes the thought of inheritance borrowed from object-oriented programming
- a caching framework which will use any of many cache strategies
- support for middleware categories which will intervene at numerous stages of request process and do custom functions
- an internal dispatcher system that permits elements of associate degree application to speak events to every alternative via pre-defined signals

- an internationalization system, as well as translations of Django's own elements into a range of languages
- a serialization system which will turn out and skim XML and/or JSON representations of Django model instances
- a system for extending the capabilities of the example engine
- an interface to Python's intrinsic unit check framework

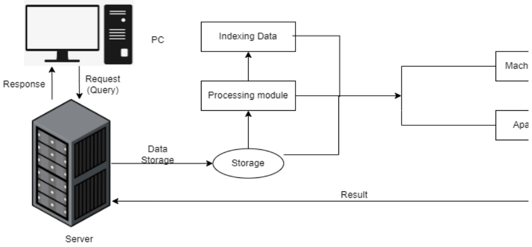


FIGURE 1. High Level Diagram

TensorFlow is an Associate in Nursing ASCII text file library for numerical computation and large-scale machine learning that facilitate Google Brain TensorFlow, the method of exploiting knowledge, coaching models, serving predictions, and purifying future results.Tensorflow bundles along Machine Learning and Deep Learning models and algorithms. It uses Python as a convenient front-end and runs it with efficiency in optimized C++.TensorFlow permits developers to form a graph of computations to perform. Every node within the graph represents a calculation and every affiliation represents knowledge. Hence, rather than managing low-details like working out correct ways in which to hitch the output of 1 perform to the input of another, the developer will specialize in the logic of the application.The deep learning computer science analysis team at Google, Google Brain, within the year 2015 developed TensorFlow for Google's internal use. This ASCII text file code library is employed by the analysis team to perform many vital tasks.TensorFlow is nowadays the foremost common code library. Their square measures many real-world applications of deep learning that TensorFlow commonly uses. Being Associate in Nursing ASCII text file library for deep learning and machine learning, TensorFlow finds a job to play in text-based applications, image recognition, voice search, and plenty of additional. Deepfake, Facebook's image recognition system, uses TensorFlow for image recognition. It's utilized by Apple's Siri for voice recognition. Each Google app that you simply use has created sensible use of TensorFlow to form your expertise higher.

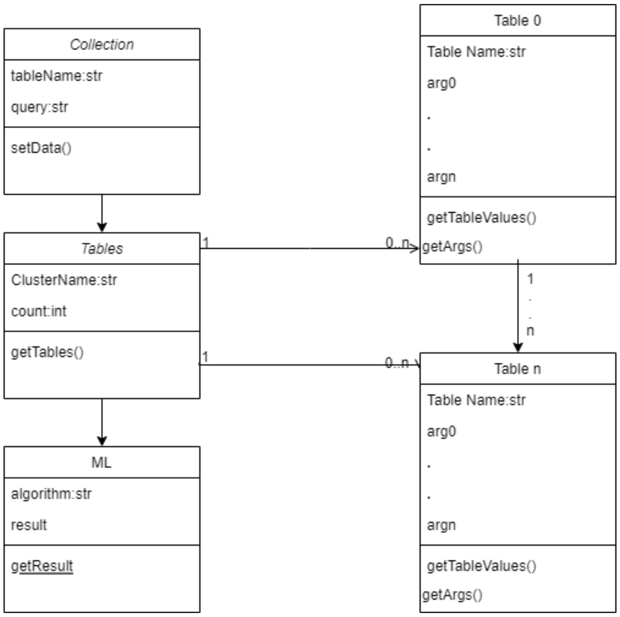
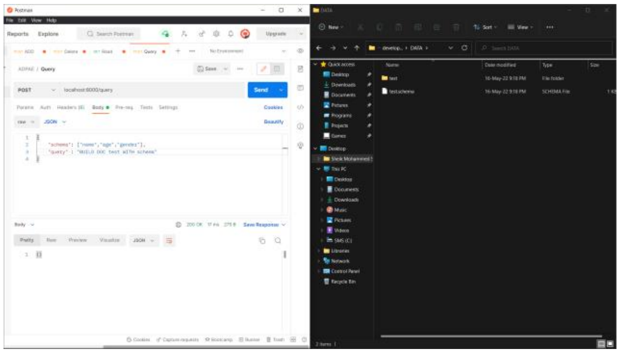


Figure 2.Low Level Diagram

We use the postman api currently.The deliveryman API permits you to programmatically access information held in a deliveryman account with ease

- The easiest way to start with the API is to click the Run in deliveryman button at the highest of the documentation page and use the deliveryman App to send requests.

- You need a legitimate API Key to send requests to the API endpoints. You'll get your key from the integrations' dashboard.
-
- The API has AN access rate limit applied thereto.
- The deliveryman API can solely reply to secured communication done over HTTPS. HTTP requests are sent a 301 direct to corresponding HTTPS resources.
- Response to each request is shipped in JSON format. Just in case the API request leads to miscalculation, it's drawn by AN "error": key within the JSON response.
- The request technique (verb) determines the character of action you propose to perform. Missed invitation created victimization. The GET technique implies that you simply need to fetch one thing from the delivery man, and POST implies you would like to avoid wasting one thing on the new deliveryman.
- The API calls can respond with acceptable HTTP standing codes for all requests. Among deliveryman consumers, once a response is received, the standing code is highlighted and is in the middle of a facilitated text that indicates the potential which means of the response code. A two hundred OKs indicates all went well, whereas 4XX or 5XX response codes indicate miscalculation from the requesting consumer or our API servers severally.



POST localhost:6000/query

```
{
  "schema": ["name","age","gender"],
  "query" : "BUILD DOC test WITH schema"
}
```

200 OK

```
[]
```

Figure 3. Creating a database using Postman

- Individual resources in your deliveryman Account are accessible victimization with its distinctive ID (uid). The uid could be an easy concatenation of the resource owner's user-id and also the resource-id. For instance, a collection's uid is }-}.

III. LITERATURE SURVEY

1. “Big Data Analytics: A Literature Review Paper ”

In the info era, huge amounts of knowledge became accessible to be had to call manufacturers.Big information refers to datasets that don't seem to be solely massive, however conjointly high in selection and speed, that makes them tough to handle victimization ancient tools and techniques.Due to the ascent of such information, solutions have to be compelled to be studied and provided so as to handle and extract price and information from these datasets.Furthermore, call manufacturers have to be compelled to be able to gain valuable insights from such varied and quickly dynamical information, starting from daily transactions to client interactions and social network information. Such a price may be provided by victimization of massive information analytics, that is the application of advanced analytics techniques on massive information. This paper aims to research a number of the various analytics strategies and tools which might be applied to massive information, yet because the opportunities provided by the appliance of huge information analytics in numerous call domains

IV. WORKING AND IMPLEMENTATION

HARDWARE REQUIREMENTS

- Processor : Intel Core i5/i7
- RAM : Minimum 4 GB
- Hard Disk : Minimum 40 GB
- Cache Memory : Minimum 512 KB

SOFTWARE REQUIREMENTS :

- Operating System : Windows XP ,7,8,10,11,Linux,Mac OS
- Programming Language : Python 3.10.4
- Backend : Django
- Tools : Postman,Tensor flow

In our Data Processing Engine For Big Data Analytics, there are 2 modules which consist of Data Collections, Data Processing.In the Data Collection Module we use REST api to collect the data as Strings. In the Data Processing Module we split the Strings into multiple smaller json files according to the given schema.

When the raw text file hits a certain count(1000) or a PROCESS query has been given to the back-end then it performs data processing.The raw text file contents are split into many smaller json files which restores the key-value pairs.While searching for a particular record ,the data has been either pulled from these smaller json files or unprocessed raw text files.These smaller json files forms the input data frame for our machine learning data analysis module.

We use the K-means cluster algorithm to partition data into clusters so that we can classify the cluster type and use it for further referencesThe Json and raw text files are converted into csv according to the Machine Learning framework .

Modules:-

MODULE 0 : Data Collection

We create a back end which uses REST api with various end points supporting crud operations.REST api collects data as key-value pairs. The key-values pairs are stored in a raw text file.

MODULE 1 : Data Processing

When the raw text file hits a certain count(1000) or a PROCESS query has been given to the back-end then it performs data processing. The raw text file contents are split into many smaller json files which restores the key-value pairs. While searching for a particular record ,the data has been either pulled from these smaller json files or unprocessed raw text files. These smaller json files form the input data frame for our machine learning data analysis module.

Module 2: Data Analysis

The Json and raw text files are converted into csv according to the Machine Learning framework .Initially we support K-means clustering algorithm which partitions data into clusters so that we can classify the cluster type and use it for further references.

V. TESTING

System Testing is an invaluable process in any software development life cycle. The main cause of testing is to find existing errors.. It is the process of testing software with the intent of ensuring that the product fulfills its requirements and user expectations and does not fail. There are several kinds of tests. Each type of test addresses a specific testing requirement.

The most important part in software development life cycle is converting the design specs into a flawless source code.

The following three are the main characteristics of a great test:

- A good test is essential.
- A good test should not be "subpar".
- A good test should be nigh perfect.

White Box Testing:

White box testing, typically referred to as glass-box testing, may be a legal action style technique that uses the management structure of the procedural style to derive a look at cases. victimization white box testing strategies, the applied scientist will derive take a look at cases that

- This guarantees that every endpoint inside the module has been tested at least once.
- Test all logical decisions on their true and false sides.
- Test internal data structure to ensure their integrity.

For example in this project white box testing is performed against the data processing module . Without creating the document using a query, if we try to store data then it responds with the message “Build the document” else the data should be inserted.

Black Box Testing:

This method treats the coded data processing module as a black box. The module runs with processes that are more likely to cause errors. Then the error is checked in the output to see if it has any. This method unfortunately cannot be used to test all errors, because some errors may depend on the code or algorithm used to implement this module.

VI. CONCLUSION

This system can be used by developers who need to use machine learning to classify and analyze their data. It is simple, robust, efficient and easily accessible. The design of our system is to maintain an efficient relationship between Machine Learning framework and developers, who have less knowledge about machine learning. Therefore, the proposed goals have been achieved successfully.

VII. REFERENCE

1. https://www.researchgate.net/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper

   Big Data Analytics Paper
2. https://itnext.io/processing-engines-for-big-data-5827bfad6b02
3. K-Means Clustering Algorithm

   https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning
4. Django. https://docs.djangoproject.com/en/4.0/