International Journal of Engineering & Technology, 7 (2.33) (2018) 688-690

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Digital unstructured data leverage using mongo DB and python

P. T. Mary Theresa Rani [1] *, Dr. S. Prasanna [2]

[1] Research Scholar, Department of Computer Science, VISTAS, Chennai, India
[2] Professor, Department of Computer Science, VISTAS, Chennai, India
**Corresponding author E-mail: Pt_mary@yahoo.com*

Abstract

Every human being creates massive unstructured data every day with the technology growth. Investors to gain digital business should use these data. They will need this data for reaching the global market. In this paper, we will analyze the category of unstructured data and the way it is processed using MongoDB and Python. We will discuss using case study and derive the mechanism of faster and flexible data accessing methodology to support the Organization.

Keywords: *Digital Universe; Mongo DB, Python; Pymongo; Unstructured Data.*

1. Introduction

Unstructured data for not only storage and analyzation. The next evaluation in computing which is machine learning is mainly depending on data and training. The predictive models of data should be accurate and readily available for accessing. This data will train the machine and enables the machine to predict the testing data [1].
Unstructured data will exists 80% in the business world [2]. Figure21 represents the sources.

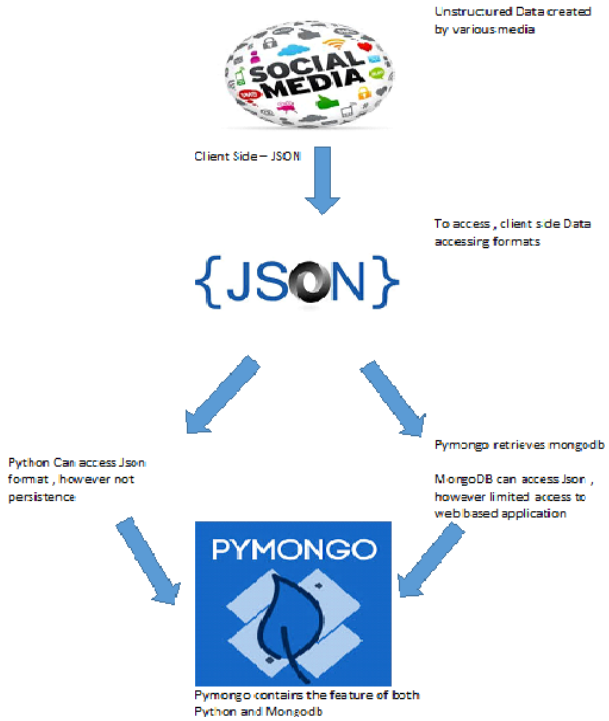

Fig. 2: Pymonogo Unstructured Data.

2. NO SQL unstructured data

The mass usage of unstructured data demands the need of an architecture to perform storage and querying. Not Only SQL (NOSQL) becomes popular in addressing this need. According to the data model and storage, NoSQL databases categorized as key-value, document column and graphs stores, which will support all sources of unstructured data in the digital universe [3].
Though we mention NoSQL data models do not follow any structure, it supports CAP (Consistency, Availability, Partition tolerance) model and ensures data reliability.
The semi-structured data formats CSV; XML JSON can also be stored in NoSQL database. The below are document store type NOSQL model used for accessing any semi or unstructured data [4].
B.1 Comma Separated Values (CSV).
Generally used to transfer data between applications. For example when organization need to transfer database data into spreadsheet, they can convert into CSV.
B.2 JavaScript Object Notation (JSON).
JSON is a format, which is readable by man as well as machine. It is widely used for data interchange. It supports all datatypes and arrays
B.3 Binary JSON (BSON)
BSON is extended version of JSON, which provides additional data type. It also uses for data encore and decode in various languages
B.4 XML (eXtensible Markup Language)
XML is a very flexible text format. Initially it is designed to access large electronic data.
Document NoSQL supports any data files, electronic data and large objects. Their Schema is dynamic and extended based on the growth of data.
While NoSQL deals unorganized data, the support of cloud computing service model, database as service (DBaas) changed the trend of data accessing methodologies [3].

3. Mono DB

Mongodb is the open source, schema less NoSQL database. There are many document NoSQL databases are available. However, MongoDb is popular due to its unique features.
Unlike other NOSQL databases, MongoDB has features like indexing, balancing load and faster querying. It will also act as file system during fault tolerance[5].
The following are some of the best features of MongoDB that will support optimal execution of unstructured data [6].
Sharding: The support of sharing process allows the data to be stored in different machines helps MongoDb to be reliable when the data growth is massive and during downtime.
B. Advanced Text Searching.
Advanced Text Searching:
The full text search feature helps in searching the whole database by using the keyword or phrase given by end user. Irrespective of grammar and spelling, this will return sounds like text as output. In social network, if the user attempt for a search using a word 'photo', free text search option will search all posts and returns objects.
Map Reduce: Map reduce is used in faster retrival of data. When the volume of data is large. This will first queries the collection list and maps the result document and returns key value pairs. This will be further divided into key and values.
ng the cross reference number (Example: As pointed in (1) the…).

4. Use case of unstructured data in mongo DB

We will discuss some of the implementation of document type data using MongoDB with the below case study.
ABC product company need to collect e-feedback form, which contains product id, product name, company address, consumer name and consumer address. In SQL structure, we may need to use two tables and use join queries. The same will be accessed as document object in MongoDB.
JSON using MongoDB:
Consider he below JSON data format

```
{
'_id': one,
'Name': {'ABC smart watch'},
'Type': ['round', 'square'],
'Consumer': [
{
'ID': 'Mathur',
'Address': 'India',
'Remarks': 'Excellent'
},
{
ID': 'Joyson',
'Address': 'UK',
'Remarks': 'Good'
}
]
}
```

Step one – Import the data collecting mongoimport db test collection ABCProduct
Step 2 Querying using Find method use test db.ABCProduct.find ()
The above result set will contain all the documents in JSON
To retrieve customer who has given Excellent Feedback,
Db. ABCProduct. find ("Consumer.remarks', 'Excellent')

4.1. CSV and mongo DB

Consider the same scenario and the file format is in text or CSV. Let us take file name as company.csv
We need to import the CSV file in Mongodb as database. Once we have imported and for querying, we will be similar to earlier case
Glimpse of Company.CSV file format.

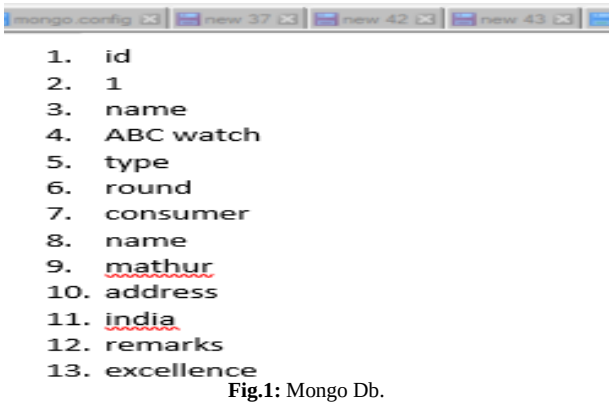

Fig.1: Mongo Db.

Mongo import -d ABCDbase -c ABCProduct –-type CSV – -file Company.csv –headerline
ABCDbase is the datbase name
ABCProduct is the collection name
XML & BSON: Mongodb will not support XML document. We can convert XML document into JSON, BSON or CSV format and import document [7].
PYTHON AND DATA ACCESSING: Python is the common purpose high level programming language. With fewer lines of coding, the language will meet the requirement of user. It supports multiple programing methods like object oriented, functional, procedural or dynamic. Python is portable, open source, simple and fast [8].
Libraries: Python supports enormous set of library files that are used in various field. Some of the libraries mainly designed for improving portability
Python Tuples: Tuples are immutable list in python. The format of tuples supports unstructured data. It generally used to group set of related data
For example

```
tup1 = ("Joyson","UK","Good");
tup2 = ("Mathur","India","Excellent" );
Print (tup1)
```

Networking & Webscarping: Using Urllib we can easily access the HTTP network. URLlib will retrieve the required webpage and handles all header details

```
Import urllib.
Myfile=urllib.open
(www.ABCproduct.com/product/customerfeedback.csv)
```

Once we retrieve the page, Webscarping used to retrieve data based on few patterns

```
fhand = open('feedback.txt', 'w')
Size = 0
While True:
Info = Myfile.read(100000)
If len (info) < 1: break
Size = size + Len (info)
Handwrite (info)
fhand.close ()
```

The above program parses the URL and retrieve information. The content of customerfeedback.csv file will be stored 100000 character at a time into local file named 'output.txt' using webscarping method. This file can be further used for the purpose of search and analyzation
JSON: Python parses JSON data format and parsed JSON Object will be represented as Python Object and Structure.

```
Import json
Input = '" [{
"Id": "001",
"Name": "Joyson",
}} ]'"
Info = json.loads (input)
Print 'Usercount:' len(info)
```

Spidering using Database: Python supports querying both SQL and NoSQL databases. We need import the relevant Spiders are used to provide pages for search engines.
Python uses data modeling for breaking the data into multiple tables and build relationships among them.
Thus, Python is widely used for accessing structured and unstructured data from various sources. The networking and webscraping technique of Python is used to parse the universal data and help in analyzing the same.

5. Mongo DB and pythan

We have discussed how MongoDB database and Python are supporting unstructured data. While MongoDB is supporting migration of larger data into open easy to access format, Python uses simple querying mechanism to retrieve the web data. [Table 1]
We can use the combination of MongoDB and Python for better performance [9]. There are many GUI tools to add MongoDB database in python. We can also use systematic procedure to implement the connectivity. pymongo is the driver for including Mongodb in Python

```
Import pymongo
Conn = pymongo.Connection ()
Db. =conn.ABCproduct
```

Here ABCproduct is the collection name
We can also use pip command to install the package

```
Pip install mongodb
```

1) from pymongo import MongoClient
2) Creating connection Con=MongoClient ()Alternatively, we can use the complete url Con= ("mongodb://mongodb0.ABCComp.net:1028")
3) Access data base Db=Con.test
4) Access the collection using data base Db.ABCProduct

Once we establish this connection, we can use JSON or BSON formats to perform data operations.
Let us consider our use case. We have called ABCProduct collection, which is stored in MongoDB
This collection can be used in Python programming

```
Product=Db.ABCProduct
Product1 = {‘_id’: [1], ‘name’: {‘ABC smart watch’},
 ‘Type’: [‘round’, 'square'], ‘Consumer’: [{‘ID’: 'Rani',
 ‘Address’: 'India', ‘remarks’: 'Good'
 },
```

We can add or modify the content by using

```
Newfb=Product.insert_many(Product1)
```

Python is being high-level scalable web programming language; the integration of MongoDB with Python will be a good combination in monitoring and accessing unstructured data.

6. Monogo DB cloud

Cloud computing offers optimum utilization of services. In our case study, the organization wants to access the feedback of customer. The data is in unstructured format and we have used MongoDB and Python for storing and retrieval.
Cloud data store, Database as a Service (DBaaS) [3] helps in transforming this operational workload into vendor based. Experimenting the sharding technique of MongoDB in cloud and retrieving using Python will solve the purpose of organization

7. Conclusion

In this article, we have seen the huge amount of data accumulated by new age technologies. We need less expensive, highly efficient, more flexible database technologies and analysis tools to handle this situation. This data is getting accessed day-to-day, understanding and using it in optimized way will help the global business growth for organization.
We have analyzed the performance of Python and MongoDB to handle this unstructured data. Python language supports easy web accessing packages. The integrity of Mongodb with Python programming language will provide best service for the industry. [Figure 2]. Further Mongodb can be accessed using Cloud service to reduce the cost and provide rental service to the organization. The future study will be using real case studies using Mongodb and Python in cloud platform.

References

[1] William Vorhies, "How NoSQL Fundamentally Changed Machine Learning",Blog,April 27, 2015.
[2] Mark Burdon, “https: / /www.ibm.com /blogs /watson/ 2018/01/top-5-challenges-cmos-face-in-2018-and-how-ai-can-solve-them/”, Blog, Jan 18, 2018.
[3] Suman Tiwari,M.Akkalakshmi, KrishnaKasyap Bhagavatula, "Analysis of NoSQL databases:MongoDB,HBASe, Neo4J", International Journal of Engineering Trends and Technology (IJETT),April 2017.
[4] Jozef Jarosciak ,"Introduction to NoSQL & Document Data Store",Blog,January 29, 2017.
[5] Vidushi Jain,Aviral Upadhyay, "Mongodb and Nosql",International Journal of Computer Applications (0975 – 8887),Volume 167 – No.10, June 2017.
[6] I am programmer blog, http://www.improgrammer.net/most-popular-nosql-database,Oct 2017.
[7] Michael Good,”Converting XML to JSON, Raw Use in MongoDB, and Spring Batch”, BLOG, Dec. 03, 17.
[8] Masoud Nosrati,"Python: An appropriate language for real world programming", World Applied Programming, Vol (1), No (2), June 2011. 110-117.
[9] Sarkarsinha H. Rajput,Anand S. Jain,Priyanka B. Patil, Mahesh D.Patil,"Mongo DB GUI Operation Using Python",International Journal of Advanced Research in Computer and Communication Engineering,Vol. 4, Issue 3, March 2015.
[10] International Journal of Computer Science and Mobile Computing, A Monthly Journal of Computer Science and Information Technology,ISSN 2320–088X,IJCSMC, Vol. 4, Issue. 2, February 2015, pg.73 – 79.
[11] Nayak, A., Poriya, A., & Poojary, D., “Type of NOSQL databases and its comparison with relational databases”, International Journal of Applied Information Systems, 5(4), 16-19.
[12] Padhy, R. P., Patra, M. R., & Satapathy, S. C, “RDBMS to NoSQL: reviewing some next-generation non-relational database”, International Journal of Advanced Engineering Science and Technologies, 11(1), 15-30.
[13] Apache CouchDB http://wiki.apache.org/couchdb.
[14] Clarence GOH, “How Singapore investors can profit from unstructured data”, Singapore Management University, Sep 2017.