© 2018 IJSRST | Volume 4 | Issue 9 | Print ISSN : 2395-6011 | Online ISSN : 2395-602X
Themed Section: Science and Technology

Large Scale Monitoring System in IT Industries using Big Data

[1]Pallavi Baruah Guha, [2]Dr. Bhairab Sarma

[1]HOD and Assistant Professor, Computer Science Department, Asian Institute of Management and Technology, Guwahati, India

[2]Associate Professor, Computer Science Department, University of Science and Technology, Meghalaya, India

ABSTRACT

This research paper is based on modelling technique and building a prediction model using Python programming language PANDA to predict data set on large-scale monitoring system using Big Data Analytics in IT Industries. In this research paper, the researcher developed a programming modelling technique which would be identify the customer behaviours patterns using large scale of data. The programming language Python to perform the full life-cycle of any data set. It includes reading, analysing, visualizing and finally making predictions. The Researcher focused on the modelling techniques how attributes / data of applicants or customers are providing a significant role to make a specific decision or generate a new information about their candidatures towards predictions on specific real life problems.

Keywords : Big Data, ETL, PANDAS

I. INTRODUCTION

The term Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety) (Laney, 2001). Big Data analytics – the process of analysing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyse and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools.

Big data analytics refers to the strategy of analysing large volumes of data, or big data. This big data is gathered from a wide variety of sources, including social networks, videos, digital images, sensors, and sales transaction records. The aim in analysing all this data is to uncover patterns and connections that might otherwise be invisible, and that might provide valuable insights about the users who created it. Through this insight, businesses may be able to gain an edge over their rivals and make superior business decisions.

The technological advances in storage, processing, and analysis of Big Data include (a) the rapidly decreasing cost of storage and CPU power in recent years; (b) the flexibility and cost-effectiveness of datacentres and cloud computing for elastic computation and storage; and (c) the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing. These advances have created several differences between traditional analytics and Big Data analytics.

1. Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends.
2. Big Data tools such as the Hadoop ecosystem and No-SQL databases provide the technology to increase the processing speed of complex queries and analytics. 3. Extract, Transform, and Load (ETL) in traditional data warehouses is rigid because users have to define schemas ahead of time. As a result, after a data warehouse has been deployed, incorporating a new schema might be difficult. With Big Data tools, users do not have to use predefined formats. They can load structured and unstructured data in a variety of formats and can choose how best to use the data.

Big Data technologies can be divided into two groups: batch processing, which are analytics on data at rest, and stream processing, which are analytics on data in motion. Real-time processing does not always need to reside in memory, and new interactive analyses of large-scale data sets through new technologies like Drill and Dremel provide new paradigms for data analysis.

The complexity of the Internet has rapidly increased, making it more important and challenging to design scalable network monitoring tools. Network monitoring typically requires rolling data analysis, i.e., continuously and incrementally updating (rolling-over) various reports and statistics over high volume data streams. In this paper, we describe DB-Stream, which is an SQL-based system that explicitly supports incremental queries for rolling data analysis. We also present a performance comparison of DB-Stream with a parallel data processing engine (Spark), showing that, in some scenarios, a single DB-Stream node can outperform a cluster of ten Spark nodes on rolling network monitoring workloads. Although our performance evaluation is based on network monitoring data, our results can be generalized to other Big Data problems with high volume and velocity (Arian Bär and Alessandro Finamore (2014)).

II. LITERATURE REVIEW

Jun Liu and Feng Liu (2014) analysed the network traffic monitoring and analysis is of theoretical and practical significance for optimizing network resource and improving user experience. However, existing solutions, which usually rely on a high-performance server with large storage capacity, are not scalable for detailed analysis of a large volume of traffic data. In this article, we present a traffic monitoring and analysis system for large-scale networks based on Hadoop, an open-source distributed computing platform for big data processing on commodity hardware.

Pedro Domingos(2018)stated the machine learning algorithms can figure out how to perform important tasks by generalizing from examples. This is often feasible and cost-effective where manual programming is not. As more data becomes available, more ambitious problems can be tackled. As a result, machine learning is widely used in computer science and other fields. However, developing successful machine learning applications requires a substantial amount of “black art” that is hard to find in textbooks. This article summarizes twelve key lessons that machine learning researchers and practitioners have learned. These include pitfalls to avoid, important issues to focus on, and answers to common questions

Nada Elgendy and Ahmed Elragal(2014) researcher stated that in the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to

handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains.

E. F. CODD (1970), stated that future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information. Existing no inferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on n-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced.

mRajeev Gupta and Himanshu Gupt (2012) focused on many industries, such as telecom, health care, retail, pharmaceutical, financial services, etc., generate large amounts of data. Gaining critical business insights by querying and analyzing such massive amounts of data is becoming the need of the hour. The warehouses and solutions built around them are unable to provide reasonable response times in handling expanding data volumes. One can either perform analytics on big volume once in days or one can perform transactions on small amounts of data in seconds. With the new requirements, one needs to ensure the real-time or near real-time response for huge amount of data.

K. Leahy, K. Bruton and D. T. J. O'Sulliva (2015) has stated that in recent years, many initiatives and groups have been formed to advance smart manufacturing, with the most prominent being the Smart Manufacturing Leadership Coalition (SMLC), Industry 4.0, and the Industrial Internet Consortium. These initiatives comprise industry, academic and government partners, and contribute to the development of strategic policies, guidelines, and roadmaps relating to smart manufacturing adoption. In turn, many of these recommendations may be implemented using data-centric technologies, such as Big Data, Machine Learning, Simulation, Internet of Things and Cyber Physical Systems, to realise smart operations in the factory. Given the importance of machine uptime and availability in smart manufacturing, this research centres on the application of data-driven analytics to industrial equipment maintenance. The main contributions of this research are a set of data and system requirements for implementing equipment maintenance applications in industrial environments, and an information system model that provides a scalable and fault tolerant big data pipeline for integrating, processing and analysing industrial equipment data.

CliffEngle and Antonio Lupher(2012) emphasized on Shark is a research data analysis system built on a novel coarse-grained distributed shared-memory abstraction. Shark marries query processing with deep data analysis, providing a unified system for easy data manipulation using SQL and pushing sophisticated analysis closer to data. It scales to thousands of nodes in a fault-tolerant manner. Shark

can answer queries 40X faster than Apache Hive and run machine learning programs 25X faster than MapReduce programs in Apache Hadoop on large datasets. Modern data analysis employs statistical methods that go well beyond the roll-up and drill-down capabilities provided by traditional enterprise data warehouse (EDW) solutions. Data scientists appreciate the ability to use SQL for simple data manipulation but rely on other systems for machine learning on these data. What is needed is a system that consolidates both. For sophisticated data analysis at scale, it is important to exploit in-memory computation.

Badrish Chandramouli and Jonathan Goldstein (2013) focused on analytics over the increasing quantity of data stored in the Cloud has become very expensive, particularly due to the pay-as-you-go Cloud computation model. Data scientists typically manually extract samples of increasing data size (progressive samples) using domain-specific sampling strategies for exploratory querying. This provides them with user-control, repeatable semantics, and result provenance. However, such solutions result in tedious workflows that preclude the reuse of work across samples. On the other hand, existing approximate query processing systems report early results, but do not offer the above benefits for complex ad-hoc queries.

D. P. Acharjya and Kauser Ahmed P (2016) focused on a huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data challenges, open research issues, and various tools associated with it. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.

Matei Zaharia and Mosharaf Chowdhury (2018) focused on Map Reduce and its variants have been highly successful in implementing large-scale data-intensive applications on commodity clusters. However, most of these systems are built around an acyclic data flow model that is not suitable for other popular applications. This paper focuses on one such class of applications: those that reuse a working set of data across multiple parallel operations. This includes many iterative machine learning algorithms, as well as interactive data analysis tools. We propose a new framework called Spark that supports these applications while retaining the scalability and fault tolerance of Map Reduce.

G. Sabarmathi and Dr. R. Chinnaiyan (2016) focused on a collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications. Big data is not just about size. Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data. It aims to answer questions that were previously unanswered. Big Data constantly facing significant challenges like outsized, heterogeneity, noisy labels, non-stationary distribution. Capturing, storing, searching, sharing & analysing. The four dimensions (V's) of Big Data It is important to recognize the full potential of Big Data by addressing these technical challenges with new ways of thinking and transformative solutions. If these challenges are resolved on time, there will be a plenteous opportunities to provide major advancement in science, medicine and business. While there is clearly an important research space examining the fundamental methods and technologies for big data analytics, it is vital to acknowledge that it is also necessary to fund domain targeted research that allows specialized solutions to

be developed for specific applications. Healthcare, in general, deserves to be a natural candidate for this kind of evaluation.

III. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

To analysis data on large-scale monitoring system using big data analytics in IT industries is one of the significant research problem in the real world, to identify the natures of attributes, their behaviours patterns, their originality or validation which are directly predicted to the decision making process in IT Industries. The researcher developed a modelling technique for prediction model perform the full life-cycle of any data set, which includes reading, analysing, visualizing and finally making predictions. The researcher stated the following research objectives which are significant with building prediction model.

1. To identify the data source
2. Analysis the data using Python Programing Modelling Technique- PANDAS
3. To Generate the Report
4. To develop a Modelling Techniques / Prediction Model to derive new information from large data set about applicants candidatures.

In this research paper, the researcher stated the current research problem to identify the customer behaviours patterns using data analysis techniques with the help of Python programming Technique-PANDAS. The researcher used this technical tools to analyse the data which comes from any sources of data from the external sources.

IV. CONCEPTUAL FRAMEWORK OF THE RESEARCH STUDY

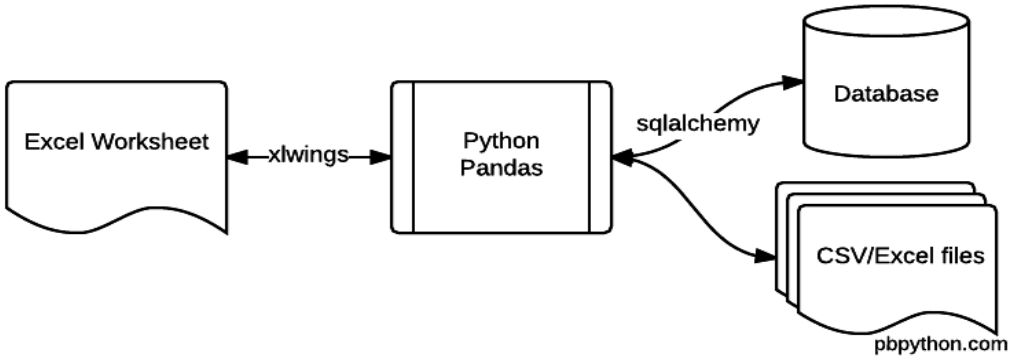

Figure 1. Data Analysis Using python Pandas

The conceptual framework of the research study is based on large amount of data which has be analysed using python programming PANDAS data analysis model. It direct access data from the different sources of external data system, stored it during processing and generating the report towards the decision making process.

V. METHODOLOGY ADAPTED FOR DATA ANALYSIS

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric Python packages. Pandas is one of those packages, and makes importing and analysing data much easier. Pandas builds on packages like

NUMPY and MATPLOTLIB to give you a single, convenient for data analysis and visualization work.
IMPORTING DATA WITH PANDAS

The first step we'll take is to read the data in. The data is stored as a comma-separated values, or csv, file, where each row is separated by a new line, and each column by a comma (,). Here are the first few rows of the ign.csv file:

There's also a leading column that contains row index values. We can safely ignore this column, but we'll dive into what index values are later on. In order to be able to work with the data in Python, we'll need to read the csv file into a Pandas Data Frame. A Data Frame is a way to represent and work with tabular data. Tabular data has rows and columns, just like our csv file.

In order to read in the data, we'll need to use the pandas..read_csv function. This function will take in a csv file and return a Data Frame. The below code will:

1. Import the panda's library. We rename it to pd so it's faster to type out.
2. Read ign.csv into a Data Frame, and assign the result to reviews.

Once we read in a Data Frame, Pandas gives us two methods that make it fast to print out the data. These functions are:

1. Pandas. Data Frame .head -- prints the first N rows of a Data Frame. By default.
2. Pandas. Data Frame. tail -- prints the last N rows of a Data Frame. By default

Indexing Data Frames with Pandas

Earlier, we used the head method to print the first 5 rows of reviews. We could accomplish the same thing using the pandas.DataFrame.iloc method. The iloc method allows us to retrieve rows and columns by position. In order to do that, we'll need to specify the positions of the rows that we want, and the positions of the columns that we want as well.

VI. RESULTS AND DISCUSSION

pbp_proj.py: programming Module

```
import pandas as pd
from sqlalchemy import create_engine
from xlwings import Workbook, Range
import os
def summarize_sales():
 """
Retrieve the account number and date ranges from the Excel sheet
Read in the data from the sqlite database, then manipulate and return it to exce
 """
# Make a connection to the calling Excel file
wb = Workbook.caller()
```

```
# Connect to sqlite db
db_file = os.path.join(os.path.dirname(wb.fullname), 'pbp_proj.db')
engine = create_engine(r"sqlite:///{}".format(db_file))

# Retrieve the account number from the excel sheet as an int
account = Range('B2').options(numbers=int).value

# Get our dates - in real life would need to do some error checking to ensure
# the correct format
start_date = Range('D2').value
end_date = Range('F2').value

# Clear existing data
Range('A5:F100').clear_contents()

# Create SQL query
sql = 'SELECT * from sales WHERE account="{}" AND date BETWEEN "{}" AND
"{}"'.format(account, start_date, end_date)

# Read query directly into a dataframe
sales_data = pd.read_sql(sql, engine)

# Analyze the data however we want
summary = sales_data.groupby(["sku"])["quantity", "ext-price"].sum()

total_sales = sales_data["ext-price"].sum()

# Output the results
if summary.empty:
```

```
    Range('A5').value = "No Data for account {}".format(account)
else:
    Range('A5').options(index=True).value = summary
    Range('E5').value = "Total Sales"
    Range('F5').value = total_sales
```

Here is a sample result:

A	B	C	D	E	F	G	H	I
Account Number	740150	**Start Date**	1/1/2014	**End Date**	3/1/2014		Retrieve Sales	
Sales History								
sku	quantity	ext-price		Total Sales	18395.6			
B1-20000	67	5659.83						
B1-33087	35	3263.75						
B1-33364	6	147.84						
B1-38851	20	1619.21						
B1-50809	8	156.8						
B1-53102	1	68.06						
B1-86481	20	608.2						
S1-06532	10	658.7						
S1-30248	21	295.05						
S1-47412	27	975.24						
S1-93683	21	217.14						
S2-10342	47	4543.96						
S2-16558	2	181.82						

Figure 2. Output of a Sample Data Analysis

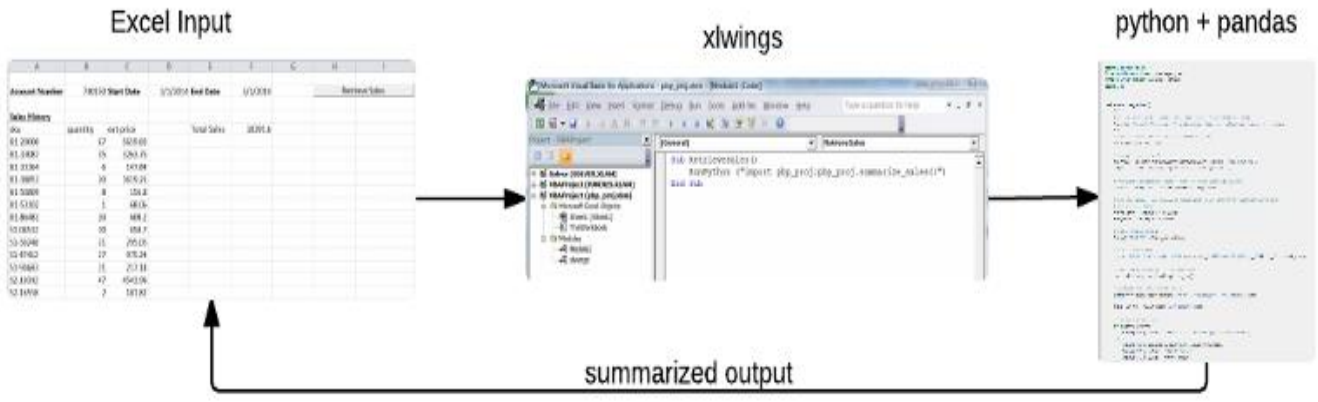

Figure 3. Summarized of a Sample Data Analysis

VII. CONCLUSION

In this research paper the researcher stated the current research problem and proposed a modelling technique to analysis the large amount of data using

Python programming PANDAS tools which takes the data from the different sources from the external system and generate a report towards the customer behaviours patterns. The researcher also stated a proposed model and programming algorithms to identify the problem and analysis to generate the report towards decision making process in IT industries.

VIII. REFERENCES

[1]. Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Stamford, CT: META Group.

[2]. Jun Liu and Feng Liu (2014),' Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop', Published in: IEEE Network (Volume: 28, Issue: 4, July-August 2014) Page(s): 32 - 39 Date of Publication: 24 July 2014 Print ISSN: 0890-8044.

[3]. Pedro Domingo (2018),' A Few Useful Things to Know about Machine Learning', Department of Computer Science and Engineering University of Washington Seattle, WA 98195-2350, U.S.A.

[4]. Nada Elgendy and Ahmed Elragal(2014),' Big Data Analytics: A Literature Review Paper', Conference Paper in Lecture Notes in Computer Science · August 2014, P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, 2014. © Springer International Publishing Switzerland 2014.

[5]. E. F. CODD (1970),' A Relational Model of Data for Large Shared Data Banks', Information Retrieval, Volume 13 / Number 6 / June, 1970, ACM.

[6]. Rajeev Gupta and Himanshu Gupt (2012),' Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?', BDA 2012, LNCS 7678, pp. 42–61, 2012. © Springer-Verlag Berlin Heidelberg 2012

[7]. K. Leahy, K. Bruton and D. T. J. O'Sulliva (2015),'An industrial big data pipeline for data- driven analytics maintenance applications in large- scale smart manufacturing facilities', Journal of Big Data, Springer, O'Donovan et al. Journal of Big Data (2015) 2:25

[8]. CliffEngle and Antonio Lupher(2012),' Shark: Fast Data Analysis Using Coarse-grained Distributed Memory', SIGMOD'12,May20-24,2012,Scottsdale,Arizona,USA. Copyright2012ACM978-1-4503-1247-9/12/05.

[9]. Badrish Chandramouli and Jonathan Goldstein (2013),' Scalable Progressive Analytics on Big Data in the Cloud', August 26th - 30th 2013, Riva del Garda, Trento, Italy. Proceedings of the VLDB Endowment, Vol. 6, No. 14 Copyright 2013 VLDB Endowment 2150-8097/13/14.

[10]. D. P. Acharjya and Kauser Ahmed P (2016),' A Survey on Big Data Analytics: Challenges, Open Research Issues and Tool', (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016

[11]. Matei Zaharia and Mosharaf Chowdhury (2018),' Spark: Cluster Computing with Working Sets', University of California, Berkeley.

[12]. G. Sabarmathi and Dr. R. Chinnaiyan (2016),' Big Data Analytics Research Opportunities and Challenges- A Review', International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 10, October 2016, ISSN: 2277 128X

[13]. Arian Bar and Alessandro Finamore (2014),' Large-scale network traffic monitoring with DB-Stream, a system for rolling big data analyses, published in: 2014 IEEE International Conference on Big Data (Big Data).