International Journal of Engineering Research and General Science Volume 4, Issue 1, January-February, 2016
ISSN 2091-2730

In-memory Data Analytics On Top Of Hadoop Environment For Detecting Fraudulent Transactions And Analysing Customer Behaviour

Rushikesh C. Nere[1], Prof. Pravin S. Game[2]
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
rushikeshcnere@gmail.com [1]

Abstract— Data is being generated by everything around us at all times. Every digital resource and social media operations produces it. Mobile devices, systems and sensors carry it. Data Analytics is an emerging technology in every area where large amount of data is generated. It helps business organizations to forecast the generated data (structured or unstructured) and to make proper strategies and decisions to prevent any kind of loss to them. Nowadays, online stores, banks continuously generate huge amount of data. Processing of this data takes significant amount of time. Hence, what online stores and banks need is to acquire the capability to process their big data in lesser time. This is easily possible, with the help of high-performance analytics solutions. This paper aims to implement high performance analytics solution using in-memory analytics and hadoop technology which detects fraudulent transactions and does customer behaviour analysis. Use of in-memory analytics reduces the processing and response time and generates faster results. On the other hand, hadoop is to be used for large scale data processing in parallel. Also, Hive is used on the top of hadoop for extracting structured data and analyzing the big data by processing queries and using map-reduce functions internally.

Keywords—Data Analytics; Big Data; In-memory Analytics; Hadoop; Hive; Map-reduce

INTRODUCTION

Big data is arriving from various sources at an alarming velocity, volume and variety, veracity and value. To extract meaningful value from big data, one need optimal processing power, analytics capabilities and skills. In-memory analytics is a technique for querying data when it is stored in a computer's main memory i.e. random access memory (RAM), instead of querying data that resides on secondary storage.

This results in tremendously shortened query processing and response times, allowing business intelligence (BI) and analytic applications to support speedy decision making. If the cost of RAM is neglected, in-memory analytics is becoming feasible for many business organizations. BI and other analytics applications have large supported data in RAM, but earlier 32-bit operating systems provided only 4 GB of addressable memory. Newer 64-bit operating systems, with up to 1 terabyte (TB) addressable memory (and perhaps
more in the future), have made it possible to store and process large volumes of data; potentially an entire data warehouse or data mart in a computer's RAM.

II. Motivation

Traditional analytics approaches such as Business Intelligence, Operation Research and Data Mining are no longer enough to harvest value from big data and automate decision making or provide a feedback loop into systems for future improvement and self-tuning. For data scientists and decision makers, there is often need to analyze the data to visualize it or numerically analyze the data in order to get an insight.

Advanced analytics however, does not stop at this point. The ultimate aim is to exploit and extract complex relationships between data and quantitatively measure the amounts and quality of data. This usually involves machine learning and statistical analysis to make solid mathematical models and abstractions that could be used to predict future behaviour or even optimize and improve the performance for multiple goals algorithmically.

The rate of data generated on digital universe is increasing exponentially. Current tools and technologies are not up to the mark to store and process huge amount of data in lesser time. They are also unable to extract value from these data which is most important. When an enterprise can leverage all the information available with large data rather than just a part of its data then it has a superior advantage over the market competitors. Big Data can help to review insights and make effective decisions. In order to handle big data modified paradigms are required.

In-memory analytics can reduce or minimize the need for data indexing and storing aggregated data in OLAP cubes or aggregate

tables. This reduces IT costs and offers quicker implementation of BI and analytics applications. It is expected that as BI and analytic applications implement in-memory analytics, traditional data warehouses may eventually be used only for data that is not queried frequently.

LITERATURE SURVEY

Analytics can be executed using various analytical approaches. Different predictive models, decision model and descriptive models are used for the analyzing data. A combinational approach for analyzing big data by integrating predictive analytics is helpful to automate the decision making process in business. *Big data* analytics need greater use of predictive analytics to discover hidden patterns and their relationships to visualize and explore data [1].

Descriptive Models	**Decision Models**	**Predictive Models**
Find clusters of data elements with similar characteristics. Focus is on as many variables as possible.	Find optimal and most certain outcome for a specific decision. Focus is on specific decision.	Find causality relationships and patterns between explanatory variables and dependent variables. Focus is on specific variables.
Examples: customer segmentation based on socio-demographic characteristics, life cycle, profitability, product preferences	Examples: critical path, network planning, scheduling, resource optimization, simulation, stochastic modeling	Examples: next customer preference, fraud, credit worthiness, system failure

Table 1. Comparison between various analytics model

For different types of big data, different frameworks are required to run analytics. There are many frameworks for processing big data. Some of them includes Apache Hadoop, Apache Drill and Project Storm. Hadoop runs Map-Reduce function whereas Storm clusters maintain topologies. Apache Drill is a query processing framework which is basically used to scan entire tables [2].

Features	Apache Hadoop	Storm	Apache Drill
Owner	Community	Community	Community
Workload	Batch Processing	Real Time Computation / Stream Analysis	Interactive and Ad-hoc Analysis
Source Code	Open	Open	Open
Low Latency	No	Yes	Yes
Complexity	Less	Less	High

Table 2. Comparison between different analytics frameworks

One study has shown that how to investigate customers' dissatisfaction behavior and how to find types of customers' dissatisfaction behavior for productively responding to that. Management of customers' dissatisfaction behaviours is related to their satisfaction. To predict the behaviour of customers data must be analyzed using qualitative methods. The result drawn from analyzing data by these methods showed that the dissatisfaction of service or product quality and disappointment have different effects on behaviors. Customers who feel greater dissatisfaction with service quality become aggressive and respond in various ways such as replacement, cancellation, refund. These results of customers' dissatisfaction behaviors shows that organization has to manage both service quality and the customer's experience dimension [3]. The time taken by the process to investigate dissatisfaction type may result in loss of customers.

Analysis of customer behavior is another crucial part in order to retain existing customers. In personalization applications, it is necessary to know that how essential the contextual information is when building customer behavioral model. The contextual information is also important to predict customer behavior. The degree of contextual information has to be calculated so that the number of behavioral characteristics can be decided. Three main questions has to be addressed to improve customer behavior's predictability: 1) Does context matter when building models of customers' behavior? 2) To which degree is it possible to extract the contextual information from the data? and 3) How do we use the contextual information for predicting customers' behavior? [4].

Another study describes that sensitivity analysis is another technique to predict customer behavior in precise manner. In this technique, some models for predicting behavior of customers are formed prior to analysis of behavior type. When the outcome of behavioral analysis is produced, it is then matched with the prediction models formed earlier. Then, depending upon sensitivity of outcome i.e. the outcome which is similar to one of the formed models is mapped to that respective model and then the prediction is carried out according to that model. In this approach, only matching of outcome is to be done with given models instead of processing and executing each model for analyzing human behavior [5].

The accuracy ratio in above study has to be calculated precisely and results showed that accuracy ratio is less while matching with predictive models.

SYSTEM ARCHITECTURE

The system is combination of in-memory analytics which helps improve the performance by minimizing overhead of bring data into main memory each time from secondary storage and hadoop framework. Hadoop helps in handling big data by storing smaller chunks and in Hadoop Distributed File System (HDFS).

Below is the architectural block diagram of proposed system:

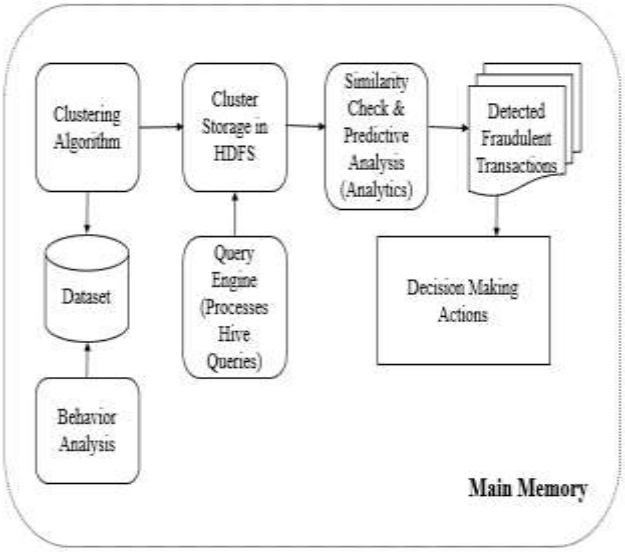


Figure 1. Proposed System Architecture

Mathematical Model

Let S be the system such that,
S = {s, e, X, Y, fmain, DD, NDD, ffriend | ϕ}
Where,
s- initial state
e- end state
X- input of the system
Y- output
Fmain- main algorithm resulting into outcome
DD- deterministic data
NDD- non-deterministic data
ϕ- constraint

In this problem, consider X be the unstructured data as input.
X = D
And D can be defined as D = {d_1, d_2, …… , d_n},
Where d_i is dataset and $d_i \in D$

Y can be treated as output.
Y = {T} ^ {C}
And T can be defined as T = {t_1, t_2, ……. , t_m}
Where t_i is detected fraudulent transaction and $t_i \in T$
And C can be defined as C = {c_1, c_2, ……. , c_m}
Where c_i is the type of dissatisfaction by customer and $c_i \in C$

DD: Given data from which comprises of different data sets i.e. d_1, d_2,…., d_n. DD → D

NDD: It is the data which is to be determined. NDD → {T} ^ {C}

Functions:
fmain → $d(x_i , x_j)$ and it is given by

$d(x_i, x_j) = [(x_{i1} - x_{j1})^2 + \ldots. + (x_{in} - x_{jn})^2]^{1/2}$

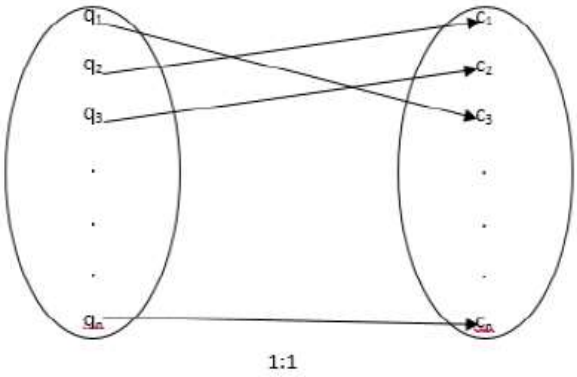
$f_{friend} \rightarrow \forall\ d_i \exists\ G \mid d(g_i, g_j) < d(g_{max}, g_{min})$
where, $G = \{G_1, G_2, \ldots\ldots\ldots, G_n\}$

ϕ- Di ∈ D | D is structured data.

Success: Set of fraudulent transactions and set of behavior type is obtained precisely.

Failure: Set of fraudulent transactions and set of behavior type is not obtained.

Mapping:



Query set Cluster set

Figure 2. Mapping between query set and cluster set

Executional Steps

1. Start
2. Bring entire data into main memory
3. Check if data is structured, if No go to step 4, else step 5
4. Convert unstructured data into structured data
5. Store clusters on HDFS
6. Perform similarity check i.e. calculate intra and inter cluster distance by Euclidean Distance formula
7. Extract odd entry by analyzing results obtained by queries
8. Group odd entries
9. Stop

Algorithm Used

DBSCAN Algorithm for clustering:

1. Create a graph of points which will be the points to be clustered
2. For each core-point p create an edge from p to every point z in the $_{\varepsilon}$-neighborhood of p
3. Assign M to the points of the graph;
4. If N does not contain any core points stop
5. Select a core point p in M
6. Let R be the set of points that can be reached from p by going forward;

i. create a cluster containing R ∪ {p}

ii. M=M/(R ∪ {p})

7. Continue with step 4

CONCLUSION

This paper aims to provide an optimized high performance data analytic technique for detecting fraudulent transactions and analyzing customer behavior. Also, proposed scheme uses in-memory analytic technique which drastically enhances system performance by minimizing overhead of bringing data in main memory from secondary storage. By utilizing hive on top of hadoop and map-reduce framework, obtaining queried results becomes an easy task. Thus, in the proposed scheme using in-memory analytics and hadoop framework we can achieve faster data processing and greater parallelization degree.

FUTURE WORK

Current approach uses offline data set for clustering records, process and analyze data. That means real time data capturing is not done. To address this issue real time data analysis will be the future work for this system. Also, risk analysis can be added as an extra objective which may analyze customer and counterparty risks.

ACKNOWLEDGMENT

I take this opportunity to express my deep sense of gratitude towards my esteemed guide Prof. P. S. Game, Pune Institute Of Computer Technology, Pune for giving me this splendid opportunity to select and present this project topic. I thank him, for his indispensable support, priceless suggestions and for most valuable time lent as and when required. I wish to express my thanks to Prof. G.P Potdar, HoD (Computer Engineering), Pune Institute of Computer Technology, Pune for encouragement and providing me with best facilities for my project work. I am also thankful to Dr. P. T. Kulkarni, Principal, Pune Institute of Computer Technology, Pune for his encouragement and support.

REFERENCES:

[1] Prof. M.S. Prasada S. Hanumanth Sastry,"Big Data and Predictive Analytics in ERP Systems for Automating Decision Making Process"

[2] Parth Chandarana, M. Vijayalakshmi, "Big Data Analytics Frameworks" in 2014 International Conference on Circuits, Systems, Communication and Information TechnologyApplications (CSCITA).

[3] Hangil Sun,"Research on the Customers Dissatisfaction behavior Types After Product Purchase from the Internet Shopping Mall: Case Analysis for Korea Post Office Shopping" in PICMET 2009 Proceedings, August 2-6, Portland, Oregon USA 2009 PICMET

[4] Cosimo Palmisano, Alexander Tuzhilin, Michele Gorgoglione, "Using Context to Improve Predictive Modeling of Customers in Personalization Applications" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 11, NOVEMBER 2008

[5] Alinda Kokkinou, David A. Cranage, "Modeling Human Behavior In Customer Based Processes: The Use Of Scenario-Based Surveys" in Proceedings of the 2011 Winter Simulation Conference