ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Net Banking Transactional Warehouse Migration towards Hadoop Distributed File System

B.Rajalakshmi , Dr.N.Saravanan

M.E. Student, Department Of CSE, MNM Jain Engineering College,Chennai, India

Professor, Department of CSE, MNM Jain Engineering College,Chennai, India

ABSTRACT: Visa/master are acclaimed in the payments cards industry, Any card transactions happens via them irrespective of host provider, current data flow edges to a Traditional RDBMS and warehoused concurrently. Re-platforming the legacy setup ingestion towards Big- data Solutions for economical benefits and scalability by customizing the Industrial Big Data ingestion and analysis Platform (IBDP) with business requirement in addition to data analysis add-on is integrated to the system for evident analytics layer.

KEYWORDS: Hadoop Distributed File System, Relational Database Management System, sqoop, Hive, Pentaho, Data Analysis.

I. INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, and updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.“There is little doubt that the quantities of data now available are indeed large, but that’s not the most relevant characteristic of this new data ecosystem.

II. RELATED METHODS

Traditional Internet industry and society with new trends and promising technologies. For industrial information with high amount and renewal speed characteristics, resulting in difficult data ingestion and analysis, this paper presented an Industrial Big Data ingestion and analysis Platform (IBDP). In the platform, we integrated HDFS, Spark, Hive, HBase, Flume, Sqoop, OpenStack etc. It works well for industrial data ingestion and analysis. In addition, we report some case studies on industrial big data processing flows respect to different data types.Facing different data types and applications, enterprises need to use a variety of large data tools and methods to carry out data ingestion and analysis in the production process. When they process the data, they need to think about the fusion of business and data flow. Also, the enterprises need to support the business decisions from the mass online data out of the enterprises. When analyzing different data in the manufacturing process, it is obviously inefficient to make the analysis respectively. So we need to build an industrial support framework which can collect and analyze different types of data. It can also provide data and analysis results to practical applications.[1]

Recently Cloud based Hadoop has gained a lot of interest that offer ready to use Hadoop cluster environment for processing of Big Data, eliminating the operational challenges of on-site hardware investment, IT support, and installing,

DOI: 10.15680/IJIRCCE.2017. 0503241

configuring of Hadoop components such as HDFS and MapReduce. On demand Hadoop as a service helps the industries to focus on business growth and based on pay peruse model for Big Data processing with auto-scaling of Hadoopcluster feature. In this paper implementation of various MapReduce jobs like Pi, TeraSort, WordCount has been done on cloud based Hadoop deployment by using Microsoft AzureCloud services. Performance of MapReduce jobs has been evaluated with respect to CPU execution time with varying size of Hadoop cluster. From the experimental result, it is found that CPU execution time to finish the jobs decrease as the number of Data Nodes in HDInsight cluster increases and indicates the good response time with increase in performance as well as more customer satisfaction. This large amount of data is having valuable information, and data without effective analysis mechanism is just a noise. So, in order to extract meaningful information from these large datasets, Hadoop is a popular open source framework for processing of large data sets on distributed commodity of hardware. But, in today'sworld for addressing this large amount of data is achallenging task that requires expensive hardware,dedicated storage, and complex software, making adaptationof Big Data technology prohibitive for small Enterprises. Toavoid this bottleneck , cloud computing provides computing resources such as storage, servers, computing power on payper use model rather than building yours own expensivesoftware and hardware infrastructure. The use of cloud computing helps the organizations to focus on their businessgoals and profits without worrying about issues such asavailability of resources, infrastructure, IT experts. With thecloud, enterprises can scale up or down to the desired levelof processing power and storage space easily and quickly. So in order to allow for future data needs, there is need to analyze the Hadoop framework on cloud computing environment.[2]

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedicalsciences. This article presents a HACE theorem that characterizes the features of the Big Datarevolution, and proposes a Big Data processing model, from the data mining perspective. This data- drivenmodel involves demand-driven aggregation of information sources, mining and analysis, user interestmodeling, and security and privacy considerations. We analyze the challenging issues in the data-drivenmodel and also in the Big Data revolution.Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Dataapplications. Being autonomous, each data sources is able to generate and collect information withoutinvolving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) settingwhere each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data related applications, such as Google, Flicker, Facebook, andWalmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technicaldesigns, but also the results of the legislation and the regulation rules in different countries/regions. Forexample, Asian markets of Walmart are inherently different from its North American markets in terms ofseasonal promotions, top sell items, and customer behaviors. More specifically, the local governmentregulations also impact on the wholesale management process and eventually result in data representations and data warehouses for local markets.[3]

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of "Big Data," including the recent announcement from the WhiteHouse about new funding initiatives across different agencies, that target research for Big Data. While the promise of Big Data is real –for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 – there is no clear consensus on what is Big Data. In fact, there have been many controversial statements about Big Data, such as "Size is the only thing that matters." In this panel we will try to explore the controversies and debunk the myths surrounding Big Data.[4]

Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Finance.These applications place very different demands on Bigtable, both in terms of data size (from URLs to web pages to satellite imagery) and latency requirements (from backend bulk processing to real-time data serving).Despite these varied demands, Bigtable has successfully provided a flexible, high-performance solution for all of these Google products. In this paper we describe the simple data model provided by Bigtable, which gives clients dynamic control over data layout and format, and we describe the design and implementation of Bigtable.[5]

Today, we‟re surrounded by data like the air. The exponential growth of data first presented challenges to cutting-edge businesses such as whatsapp, Google, Yahoo, Amazon, Microsoft, Facebook, Twitter etc. Data volumes to be processed by cloud applications are growing much faster than computing power. This growth demands new strategies for processing and analyzing information. Hadoop- MapReduce has become a powerful Computation Model addresses to these problems. Hadoop HDFS became more popular amongst all the Big Data tools as it is open source with flexible scalability, less total cost of ownership & allows data stores of any form without the need to have data types or schemas defined. Hadoop MapReduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. In this paper I have provided an overview, architecture and components of Hadoop, HCFS (Hadoop Cluster File System) and MapReduce programming model, its various applications and implementations in Cloud Environments.[6]

As usage of cloud computing increases, customers are mainly concerned about choosing cloud infrastructure with sufficient security. Concerns are greater in the multitenant environment on a public cloud. This paper addresses the security assessment of OpenStack open source cloud solution and virtual machine instances with different operating systems hosted in the cloud. The methodology and realized experiments target vulnerabilities from both inside and outside the cloud. We tested four different platforms and analyzed the security assessment. The main conclusions of the realized experiments show that multi-tenant environment raises new security challenges, there are more vulnerabilities from inside than outside and that Linux based Ubuntu, CentOS and Fedora are less vulnerable than Windows. We discuss details about these vulnerabilities and show how they can be solved by appropriate patches and other solutions.[7]

III. PROPOSED SYSTEM

The proposed system is based on the following prevention and authentication for the data security.

1. Big data Ingestion framework is implemented into the existing flow and the data is stored over HDFS.
2. Hive provides the data warehouse infra-structure over the HDFS.
3. Reporting layer is add-on over the HDFS data model for GUI representation of data.

A. ARICHITECTURE

In this paper we propose a Big data Ingestion framework is implemented into the existing flow and the data is stored over HDFS, Hive provides the data warehouse infra-structure over the HDFS. Reporting layer (Pentaho) is add-on over the HDFS data model for GUI representation of data.

DESIGN STRUCTURE

The Design structure is divided into 3 Micro service layers for the big data solution

- Data Ingestion Layer.
- Data Analysis Layer.
- Data Storage Layer.

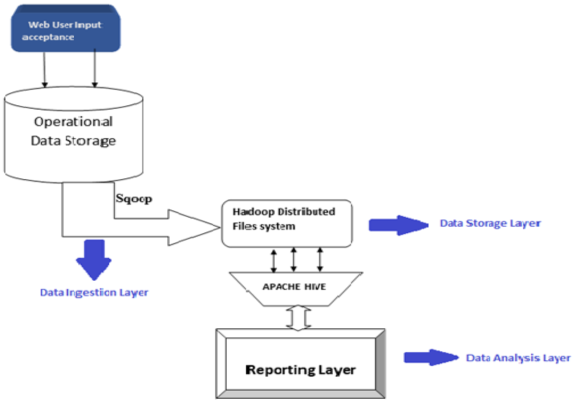

Figure 1

B.FRONT END MODULE

Front end module, projects as the User Interface web-application which gets transactional inputs from the end-user. Authentications of Individual user happens in the UI post user is allowed to their customized dashboards.Payment transactions like NEFT(money transfer) and bill payment happens in the front UI and Transactional status appears as a popups to alert user.Front end Module is the gateway for the end-customer irrespective of the hosts.Transactional Modules are mapped into MYSQL database

C. OPERATIONAL DATABASE MODULE MYSQL acts the operational database and it handles the transactional data in the entity table.Five entity tables gets updated on every transactions.Transactional details, payment details, payment history are handled in each entity tables and get updated instantly.Separate table maintained on each transactional inputs.Customer Critical Information like cardno,customerid, account number, holder name and card details are masked.

D. BIG DATA INGESTION

Apache Sqoop is used efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.Sqoop tool
'import' is used to import table data from the table to the Hadoop file system as a text file or a binary file.Sqoop scripting is handled inside a shell script which undergoes a batch process periodically Hadoop ingesting.

E. HADOOP DATA MODELLING Relational data ingestion using sqoop script gateways the data towards hdfs instituting an semi
structure tedious for analytical layer.Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.Hive lays out schema in a database and processed data into HDFS.It provides SQL type language for querying called HQL.

F. REPORTING LAYER

Pentaho reporting tool is layered over the hdfs for data analytics.Reporting tools are widely used to support decision making and to measure organizational and team performance. Companies use them for financial consolidation, for evaluation of strategies and policies and often just for plain reporting.Reporting tools allow companies to create attractive reports easily. In tabular or graphical format. With data from Excel, a data warehouse or the organization's ERP system

IV. CONCLUSION

Visa/master transactional data model has been Re- plat formed and the legacy setup ingestion towards Big-data Solutions for economic benefits and scalability by customizing the Industrial Big Data ingestion and analysis Platform (IBDP) with business requirement in addition to data analysis add-on is integrated to the system for evident analytics layer. Cost efficient and reduced SLA provides better business solutions for business betterment.

REFERENCES

[1] Cun Ji, Shijun Liu, Chenglei Yang, Lei Wu, Li Pan,"IBDP: An Industrial Big Data Ingestion and Analysis Platform and Case Studies," International Conference on Identification, Information, and Knowledge in the Internet of Things, 2015.

[2]V. Santosh Karthikeyan, Dr. N. Muthu, "Analysing Big Data with Hadoop Cluster in HDinsight Azure cloud," Research Scholar, Department of Management Studies St.Peter's University, Vol 7 issue 2 July 2016.

[3]Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data,"Department of Computer Science & Engineering, University of Vermont, USA, 2012.

[4]Alexandros Labrinidis, H. V. Jagadish, " Challenges and Opportunities with Big Data," The 38th International Conference on Very Large Data Bases, August 27th,2012.

[5] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C.Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber , "Bigtable: A Distributed Storage System for Structured Data,"Department of Computer Sciences, 7th USENIX Symposium on Operating Systems Design and Implementation, OSDI '06.

[6] Lalit Malik et al, "Mapreduce Algorithms Optimizes the Potential of Big Data," International Journal of Computer Science and Mobile Computing, Vol.4, Issue.6, June- 2015, pg. 663-674.

[7] Sasko Ristov, Marjan Gusev and Aleksandar Donevski, , "OpenStack Cloud Security Vulnerabilities from Inside and Outside," Faculty of Information Sciences and Computer Engineering, Skopje, Macedonia, The Fourth International Conference on Cloud Computing,GRIDs, and Virtualization,2013.