ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

A Detailed Survey on Big Data Application in Global Banking Data Management & Decision Making

[1] Kavyashree. J, [2] Gouri Jambure, [3] Vasudeva. R
[1] [2] VII Semester, [3] Asst. Professor
B.E Department of Computer Science Engineering Engineering, C. Byregowda Institute of Technology (CBIT)
Kolar (Karnataka), India.

Abstract: **- In today's world of Investment Banking and other financial domain areas, the requirement and demand for the automation in data processing is very high. The data is accumulated from different data sources with an increase in the rules and regulations, but that should also come with a plan of cost reduction without compromising in quality and scalability. The underlining technologies that handle big data with should guarantee of optimization and also keep global financial institutions interest in it. So this paper or case study covers the Big Data architecture and design that would help banking institutions make key decisions. We have used Hadoop map-reduce and no-SQL flexibility also maintaining the quality, banking rules and standards. The data that is proposed to be consumed or used for this analysis is from different sources and techniques, techniques that are followed in regular banking practices. That would include "front end" or "backend data processing". Business process modelling would require data transmission OR orchestration from different sources that are required to make key and important financial decisions.**

I. INTRODUCTION

Companies these days rely heavily on data.Maintaning and appending information, bank requies these incremental data to generate Historical reports.This paper's effort is to depict the cuurent data gathering, how to use in reducted format

II. RELATED WORK

Apart from just decision making in banking sector big data plays a huge role in other areas as well, e.g Retail banking. Individual banks are also data businesses. There are several key points that can help us understand how the retail banking is benefitted by this bigdata/datasets. There are several instance where the retail bankers need to propagate the secure banking information from one stream to another, may be it is the banking transaction data. This is one of the key secure data in the banking sector, may be it is the customer information, accounts data, clearing data etc.

Banking sector or retail banking transactions are one of the first domain that accepted the electronic channels for transaction and storing the static data. A team of dedicated resources were appointed to work on processing this data, so it can be used for further analysis and reporting by the back office system. Apart from being critical information, volume of the data or data sets is really huge. So even for retail transactions, several applications were developed in the initial phases and current era which have intelligence of making calculated predictions and decision makings based on the data sets One more key factor in managing these data sets is the updated records which would also mean that the application should be able to handle real time data for their analysis. This has advantage of less human intervention and automated decisions upto certain extent.

II. DATA SETS

The reason for big data to come in existence was the data sets growing leaps and bounds, which was very complex. The software applications or tools were No longer sufficient to handle this large process data. One of the prime reasons of big data sets was also collecting the data sets from different sources which could help in decision making and faster in analyzing. Following are some characteristics factors that can be very crucial while making decisions related to handling big data ets:[3]

Pricing

In order to make pricing decisions the technology underneath should be able to handle the strategies, and can gather and manage large client data sets, while can be distrubuted in chunks based on the clients willingness to spend on the core product or the services that are offered. Banks can do the anlaysis and set te pricing based on their individual segments. Some of them can be automated and get maximum ROI.

Decision Trees

Purpose of decision tree is to classify data into certain target field. Heavy decision trees can have field that might be of numeric values, typical example would be price in case of any product in the market. Decision trees are used to analyze which are the most important fields for a particular data set.And which of them can be used in other algorithms.So random Forests is a technique used to increase the efficiency of decision tree models by creating an ensemble of slightly varying decision trees that model the same target. By having slight variances between the decision trees acts as a safety net between possibleerrors and noise of an individual decision tree. Each tree in the ensemble then votes a target field, where the winning target field is then assigned to the data.

III .MAP REDUCED DATA

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data(multi-terabytes data-sets) in parallel on large clusters(thousands of nodes) of commodity hardware in a reliable, fault tolerant manner A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the output of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are sorted in a file system.The framework takes care of scheduling tasks,, monitoring them and re-executes the failed tasks. In this above fig. the input is given in various statements. This input is divided into small segments and these segments are called as input splits and the process is called splitting. Next these splits are separated into words and along with the words the count is kept track and this process is mapping. Further, these words are grouped into similar ones separately. This is the shuffling process. The reducer eliminates the duplicated entries from each segment and the total number is kept track of. Finally, the over all output is given in terms of words and the total number times each word is repeated. This is the process of MapReduce. [1]

Outcome

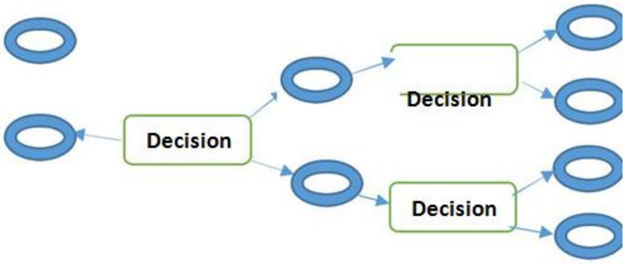

Fig 1. Decisions Trees

IV. CLUSTERING

What is clustering, it is nothing but searching for meaningful patterns within a Data Sets. It is very important to understand which patterns are most useful within the data. Cluster processing helps finding these common patterns of data and form a group Typica banking example would be customers maintaining an average balance in their accounts for a particular quarter, are grouped together. Cluster basically breaks a complex problem into smaller chunks, where a manul intervention can identify the pattern. Cluters are mainly the object of study.

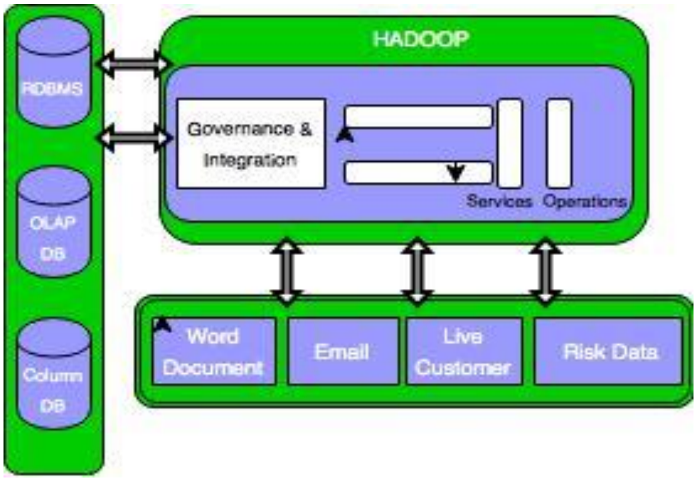

Fig 2. Architecture

The above diagram depicts how Different data sources are connected and the data is accumulated in one Data Repository and further processing with Hadoop. Key feature here is handling the big data sets or Chunks. Which can be cost effective, efficient and well managed infrastructure. Hadoop is responsible for managing and searching different patterns against the Data Sets. Different Repositories e.g the OLAP and Column Databases are interacting with the system where the accumulation happens. The security and operations strategies are applied. Different industry security standards are applied. Map & Reduce programming framework helps chunking the input data that is gathered from different sources, then the action of grouping data is done. Failure and load imbalances are handled in correct way and back up plans and environments are used then. Hadoop's distributed file system is a way where data is stored in files and directories.Moving computation rather than the actual data is always economical.

V .COMPARISON

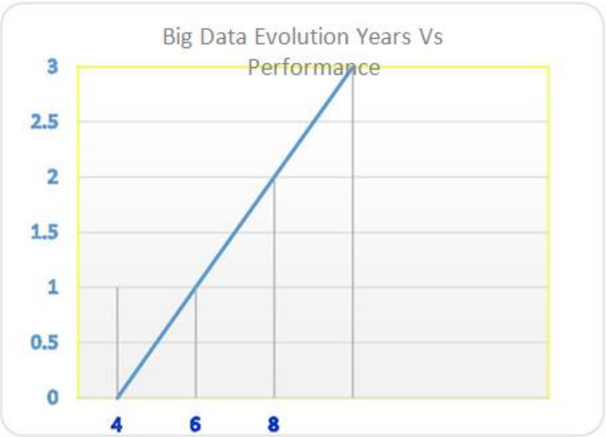

Fig 3. Comparison

Over the past several years companies all over the world have been able to get the benefits of big data and the analytics report around it. So what did the companies do to achieve this? Creating a data warehouse, then creating software programs to perform the analytics. So companies found that the cost was less and more improvements over the years.This was across all the industry domains. Finanace, Insurance, retail Banking etc. Survey results indicated that to produce these significant returns, companies need to invest substantially in data-analytics talent and in big data IT capabilities. Time is also a big factor when it comes to analysis and is very important, because performance improvements results and not that quick.In all terms the big data performance is measured in different domains – operations, customer-facing functions, and strategic and business intelligence. In terms of performance and average increase in profits from big data investments was in range 3-6 %, then over the period it increased from 7-9% in span of 5 years. One of the reason that the performance growth was seen due to the hiring of skilled employees in the organization's. From technology stand point Hadoop clusters can be one of the reasons to see good performances, along with Java and hardware tunings. A council is dedicatedly working towards providing the benchmarks basis on which the tuning parameters can be set. Also there are other tuning that need to be done, Server, Network..etc. An end to end pipeline benchmark not only measures the performance of individual stages in the data pipeline, but also takes into account performance of data exchange and possible transformation between different stages.

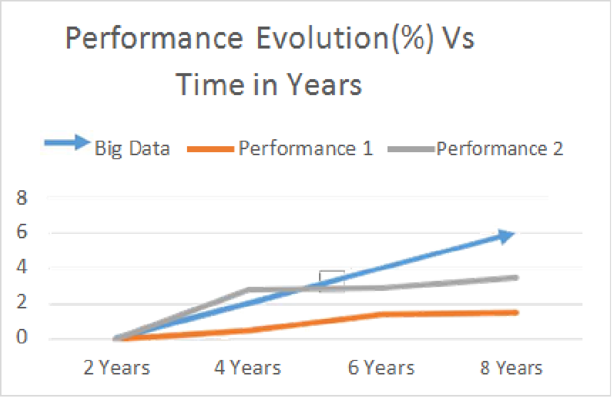

Fig 4. Comparison

REFERENCES

[1] Kasi Periyasamy, and Vinoth Perkinian, "Dynamically Reconfigurable User Interface - A Case Study from Health Care Application," Dept. of Computer Science University of Wisconsin-La Crosse La Crosse, WI, U.S.A.

[2] Yelena Yesha2 ∗, 2Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore Iterative Unified Clustering in Big Data County (UMBC, Santa Clara, CA, 2015

[3] S. Arora and I. Chana, "A survey of clustering techniques for big data analysis," Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference -, Noida, 2014, pp. 59-65. [4] J. Handy, The Cache Memory Book, Morgan Kaufmann, 1998.

[4] A. Munar, E. Chiner, I. A Big Data Financial Information Management Architecture for Global Banking Sales GFT Group Valencia, Spain