Big Data Storage Analytics

Subash Thota
Data Architect

Abstract

The amount of data being generated that enterprises acquire every day is increasing exponentially. The size of the databases has been growing at exponential rates in today's businesses. It is viable to store massive amounts of information on low-cost platforms such as Hadoop. Ironically storage industry has come a long way from providing just a storage system to customers based on customer's request to storage vendors anticipating customer needs and providing timely storage provisioning and sizing advice for business continuity.

One of the significant challenges that the storage industry faces is how to process and analyze these large volumes of data and aligning their products and services according to the capacity and performance requirements of customers. This whitepaper talks about how storage industry can leverage the Big Data ecosystem and data analytics to address the very problem of proactively identifying customer needs.

Key Words and Phrases

Big Data Analytics, Social Analytics, Storage Analytics, Data Management, Information Quality, Data Mitigation, Metadata, Data Profiling

Introduction

The explosion of data that happened during the past decade was so phenomenal, such that storage vendors had to build data storage systems with higher and higher capacities to satisfy customer needs. According to a study done by International Data Corporation (IDC), Storage is a crucial piece of the infrastructure component, growing at a compound annual growth rate (CAGR) of 53% between 2011 and 2016. The amount of data produced, handled, and stored by most organizations will continue to grow aggressively for the foreseeable future. So, Storage vendors are competing with one another and pushing their limits to meet the ever-growing needs of their customers.

The big data surge has energized the selection of Hadoop and other big data batch processing engines, but it is also pushing beyond batch and into a more analytic approach. Fortunately, there are open source tools that do pretty good job.

- Storm is a distributed computation system. It works the same way as Hadoop providing batch processing for performing analyses. The storm is easy to use, and it works with any programming language. It is very scalable andfault-tolerant.

- ClouderaofferstheCloudera Enterprisetoolsthatallow,interactiveanalyticalqueriesofthedatawhichis stored in HBase or HDFS. It is an essential part of ClouderaImpala.

- Grid Gain is an enterprise open source grid computing made for Java. It is compatible with Hadoop DFS, and it offers a substitute to Hadoop's Map Reduce. Grid Gain offers a distributed, in-memory, real-time and scalable data grid, which is the link between data sources and different applications.

- SpaceCurvecandiscoverunderlyingpatternsinmultidimensionalgeodata.Geodataisdifferentdatathan standard data as mobile devices create new datasets fast and not in the form traditional databases were used to untilrecently.

- Aerospike is a flash-optimized, in-memory open source NoSQL key-value database. It handles very high-volumestreamsofdata.Enterpriseswithexistingtransactionapplicationswouldneedtomigratethemto Aerospike for real-time analyticsintegration.

- IBM BLU is based on the standard IBM DB2 OLTP 10.5 offering. BLU Acceleration capabilities are designed mainly for "read-mostly" inline analytics. It can leverage SIMD (single instruction, multiple data) on IBM Power7orPower8chipstoimproveperformance.APIsinDB2canpotentiallyeasemigrationfromOracle.

- Microsoft SQL Server now has an In-Memory OLTP extension. Integration with SQL Server means you can havebothmemory-optimizedtablesanddiskbasedtablesinthedatabase,andqueryacrossbothtypes of tables, Microsoftasserts.

- Oracle’s database now offers an in-memory option for performing analytic queries in parallel. Itintegrates with high-availability options such as ORACLE RAC and DataGuard.

- HANAstandsforHigh-PerformanceAnalyticAppliance.SAPHANA’sgreatestvalueisinproviding specific operational reports within minutes – rather than days.

Storage Terminologies
DAS - Direct-attached storage (DAS) is a digital storage system directly attached to a server or workstation, without a storage network in between. It is a retronym, mainly used to differentiate non-networked storage from the concepts of storage area network (SAN) and network-attached storage (NAS).

SAN - A storage area network (SAN) that are primarily used to enhance storage devices, such as disk arrays, which are accessible to servers so that the devices appear like locally attached devices to the operating system.

NAS - Network-attached storage (NAS) is file-level computer information storage associated with a network providing data access to a heterogeneous group of clients. NAS frameworks are organized machines which contain at least one hard drives, regularly orchestrated into intelligent, excess stockpiling compartments or RAID. System joined capacity expels the obligation of record serving from different servers on the system. They normally give access to records utilizing system document sharing conventions, for example, NFS, SMB/CIFS, or AFP.

Data Storage is the entity that provides the physical space for all the applications to store and access all their data. This data will be used by different types of applications including, standalone desktop applications, web applications, mobile applications, etc. There are numerous ways or techniques to store, manage and access this data, which keeps evolving with the pace of new inventions and breakthroughs in technology.

There is a whole industry built around this which acts as Data Providers for applications of different forms and sizes. The storage Industry constantly adapts to the growing needs of companies to store more data in limited spaces, by using techniques like de-duplication, compression, etc. The storage industry uses cutting-edge technologies to provide seem less access to data, also provides security, integrity, and reliability of data.

The first approach that we will discuss in this article uses Greenplum database on a Hadoop environment and Tableau and the other approach in this article user R with Hadoop.

The first solution is a web portal that hosts various dashboards from tableau server. This Big Data portal is secured by LDAP authentication and provides various customized views or dashboards to users based on their roles. Data is stored in Greenplum database hosted in a Hadoop environment, Analytics dashboards and sheets are developed in Tableau desktop, which is published in tableau server. A web portal developed using Microsoft .NET technology is used to provide users with a way to view and interact with different data sets. This application is hosted on IIS on a Windows server.

Performance logs can be collected from the storage arrays on a predefined schedule. This statistical data contained in these log files gives the customer’s as well as vendors a very good insight into the configuration and performance statistics of the storage system. These log files are uploaded to the vendor's Big data ecosystem for analysis.
For the specific solution that we’re discussing in this whitepaper, the log files are processed using specializedsoftware tools that traverse the logs and extract relevant performance and configuration data into the Greenplumdatabase. The database is designed to hold data for different storage components like LUNs, Disks, Storage Processors, etc. on separate tables.

As we discussed earlier, the Database used here is Greenplum hosted on a Hadoop platform. The

database is designed in such a way that separate tables are maintained for different storage components and their related

performance/configuration data. The storage system’s serial number in conjunction with the log collection time is used to uniquely identify the data of a storage system at a period. This design helps us to pull out the performance statistics of a customer’s array at any point in time.

For data analytics, we are using Tableau desktop and Tableau server software. These software’s are developed and licensed by "Tableau Software." There are several other licensed, and opensource analytics software’s on the market, Tableau is the one we found that fits our requirement. Even though we're talking about Tableau in this whitepaper, the overall process of data analytics can be replicated on another analytics software’s as well.

Tableau Desktop is a per-user licensed software that can be installed on Windows or Mac Operating systems. Tableau has the provision to connect to various databases and provide a visual representation of the data. Database drivers are also required by Tableau to connect to different data sources. These drivers can also be downloaded from the Tableau support website. A Tableau project or workbook can comprise of several worksheets and dashboards.

Severalworksheetscanbegroupedunderadashboard.Forexample, wecandevelopdifferent worksheetstoshow the performance of different components in a storage system and then group all the worksheets into a dashboard. This dashboard can then be published onto a Tableau server fordistribution.

Tableau Server is a licensed software and needs to be installed on a separate server. It runs as a service, and it hosts the worksheets and dashboards for viewing on web browsers and embedding onto websites. In addition to the built- in user system, for user authentication and group association, Tableau supports Microsoft Active Directory. Tableau Server leverages fast databases through live data connections or can extract and refresh your data in–memory.

Tableau Worksheets hold your data views. You can save individual worksheets as bookmarks. Each worksheet can be connected to only one data source. However, different worksheets in a workbook can be connected to different data sources. Workbooks hold one or more worksheets and dashboards. By saving a workbook, you can save all open sheets in one file that can then be easily shared. 6.4 Tableau Dashboards A dashboard is a collection of several worksheets shown in a single location where you can compare and monitor a variety of data simultaneously. Figures 3 & 4 shows a single dashboard that has two worksheets, Opportunity Dashboard & Sales Dashboard arranged as tabs. These are sample dashboards pulled from Tableau website and are not part of the solution that we're discussing. This is only for the reader's understanding.

The views in a dashboard are connected to the worksheets they represent. That means when you make changes to the worksheet, the dashboard is updated, and subsequently, any changes you make to the dashboard affects the worksheet. This interaction is important to remember when you are annotating, formatting, and resizing the views in your dashboard. While dashboards are an easy way to summarize and monitor at a glance, you can go back and edit the original view by jumping to a selected worksheet. Additionally, you can duplicate worksheets directly from the dashboard to perform in-depth analysis without affecting the dashboard. Finally, you can hide worksheets that are used in dashboards, so they are not shown in the filmstrip, sheet sorter, or in the tabs along the bottom of the workbook.

The statistics and configuration data of the storage arrays used by customers world over is extracted and dumped into the Greenplum database. The Database schema is designed in such a way that the data about different components like LUN, Disk, RAID, etc. are arranged in different tables with the storage system's serial number as the primary key. So, using this key, we can get the data about a full system from these tables. Data analytics is cooked up using this data and the Tableau software.

The following dashboards were developed using Tableau Desktop and Greenplum database and then published onto the Tableau Server.

System Statistics

This dashboard is developed using the data from System specific tables like throughput(IOPS), bandwidth (MB/S), capacity (GB), storage system’s geography. This dashboard is comprised of four worksheets that show a bar graph or a table showing the split-up of the above-mentioned statistics data based on different models of the storage systems. For example, the user can understand which model of the array gives maximum throughput on the field or what kind of bandwidth requirements does the existing customers have. Filters are provided to view the statistics coming from the user base in a country or city. Filters are also provided to filter storage systems based on model name or throughput or bandwidth. This dashboard gives the user a very good overview of how their storage systems are performing in real-world situations and this will give a very good input on developing new storage arrays and technologies to better address customer needs.

Customer Statistics

As the name suggests, this dashboard is used to display customer specific statistics like Customer's geography, Model distribution, customer's industry like finance, Insurance, etc. This dashboard is comprised of different worksheets.

The worksheet that shows customer's geography gives a clear picture of the user base distribution. The filter available will enable the user to drill down to a country or city level and understand the distribution of storage arrays at that level. The model distribution and Industry distribution are two worksheets that give the user a good insight into the penetration of different models of storage systems in various industries. This dashboard will also help the user in understanding their customer base and come up with more innovations and technologies to address their larger customer base.

Application Statistics

Customers using huge datacenters and storage arrays use them for running various applications. This information is captured as part of the data extraction and is available in the application specific tables. The application-specific dashboard has different workbooks and filters to present a graphical view of the application distribution worldwide, type of storage system models used for running a different class of applications, the performance that these applications and users are extracting from the storage systems and much more. This representation helps the storage array manufacturer to better understand how customers are using their storage products to run different applications. This also helps the manufacturer to assist users or provide suggestions to improve their performance.

For example, a storage array configured with a RAID level will be best for running OLTP applications like exchange, but may not be the best solution for running SQL Server, so in such cases, the manufacturer can help them to get better performance by asking them to use a different RAID level or type of Disk.

Health Statistics

The health of the storage systems deployed by customers is the most critical data to be monitored. Any issues on the arrays can result in degenerated performance to even Data Loss or Data Unavailable situations (DU/DL). DU/DL can have a very adverse effect on the customer's Business and market. To be competitive in today's market customers expect 100% reliability and connectivity to their data and applications. So, any disruptions to data are not something storage vendors can take lightly or put up with. To ensure near to 100% up-times and data connectivity it's important for Storage vendors to monitor the health of their storage arrays on a periodic basis, identify or predict any part failures and provide pro-active solutions to prevent any catastrophic failures. The Health check statistics is comprised of various storage array related worksheets and a Health check dashboard. These worksheets provide a graphical representation of systems that are continuously being over-utilized regarding processor usage, disk saturations, bandwidth and several other properties. Tableau can show all the storage systems that are showing degraded performance or exceeding a certain threshold of utilization. This helps the storage vendors to get in touch with their customers and proactively propose some storage solutions to prevent any DU/DL from happening. Again, filters are provided to drill down to country, customer or even model level.

Marketing Statistics

Any business exists to grow their revenue and increase their customer base. In that perspective, marketing their products effectively towards a targeted group of customers will help them increase their revenue and profits significantly. In this ever-growing competitive Storage Industry, staying ahead of the curve is important. Developing a good product is one thing and selling it is another. So, the marketing statistics that tableau provides helps the storage vendors to identify marketing opportunities among their existing customers. Marketing statistics also has a group oftableau worksheets and a Tableau dashboard. This dashboard helps the business user to understand what percentage of their entire product line/features are used by different customers. Analyzing this in combination with various other dashboards that we discussed above, helps the company to target specific customers and showcase some performance or cost benefits they will be able to achieve if they use some of the new products/features provided by this storage company. Thereby helping the company as well as the customer to perform better and grow their business.

Analytics Website

Awebsitewasdevelopedtovariousdisplayedtypesofdashboardsdiscussedabove.Thewebsiteaccesswasprovided using Active Directory authentication, to allow only the active directory users from that domain to view the Analyticspages.ThiswebsitewasdevelopedinASP.NET,HTML,CSS,andjQuery.IIS7.0wasusedtohostthe website.

The second approach uses R with Hadoop, R is an open source software to perform statistical analysis on data. R is a programming language used by data scientists, statisticians and others who need to make a statistical analysis of data and glean key insights from data using mechanisms, such as regression, clustering, classification, and text analysis.

R gives a wide assortment of factual, machine learning (straight and nonlinear demonstrating, established measurable tests, time-arrangement examination, grouping, bunching) and graphical systems, and is exceedingly extensible. R has different worked in and also expanded capacities for factual, machine learning, and perception assignments, for example:

- Dataloading•Datatransformation•Dataextraction•Datacleaning•Statisticalanalysis•Predictivemodeling •Data visualization

It is one of the most popular open source statistical analysis packages available on the market today. With its developing rundown of bundles, R would now be able to associate with other information stores, for example, MySQL, SQLite, Mongo DB, and Hadoop for information stockpiling exercises. There are over 3,000 R packages, and the list is growing day by day. R bundles are independent units of R usefulness that can be conjured as capacities. A comprehensive list of these packages can be found at http://cran.r-project.org/ called Comprehensive R Archive Network (CRAN).

R enables a wide range of operations. Statistical operations, such as probability, mean, min, max, distribution, and regression, linear regression, classification, logistic regression, and clustering. Data processing operations are as follows:

- Data exploration: Explore all the possible values ofdatasets.
- Data cleaning: Clean massivedatasets.
- Dataanalysis:Performanalyticsondatawithdescriptiveandpredictiveanalyticsdatavisualization.

If we consider a consolidated RHadoop framework, R will deal with information investigation operations with the preparatory capacities, for example, information stacking, investigation, examination, and perception, and Hadoop will deal with parallel information stockpiling and additionally calculation control against appropriated information.

The integration of data-driven tools and technologies can build a powerful, scalable system that has features of R and Hadoop. There are three ways to link R and Hadoop are as follows:

- RHIPE
- RHadoop
- Hadoopstreaming

RHIPE

RHIPE stands for R and Hadoop Integrated Programming Environment. The RHIPE package uses the Divide and Recombine technique to perform data analytics over Big Data. Since RHIPE is a Java package, it acts as a Java bridge between R and Hadoop. There are some Hadoop components that will be used for data analytics operations with R and Hadoop. The components of RHIPE are as follows:

RClient is an R application that calls the JobTracker to execute the activity with a sign of a few MapReduce work assets, for example, Mapper, Reducer, input design, yield arrange, input record, yield document, and other a few parameters that can deal with the MapReduce occupations with RClient.

A JobTracker is the ace hub of the Hadoop MapReduce operations for introducing and observing the MapReduce employments over the Hadoop group.TaskTracker is a slave hub in the Hadoop group. It executes the MapReduce employments according to the requests were given by JobTracker, recovers the information lumps, and run R-particular Mapper and Reducer over it. At long last, the yield will be composed to the HDFS registry.

HDFS is a file system distributed over Hadoop clusters with several data nodes. It provides data services for various data operations.

RHadoop

RHadoop is an incredible open source programming system of R for performing information examination on the Hadoop stage through R capacities. RHadoop is a gathering of three R bundles for giving expansive information operations an R situation. RHadoop is accessible with three fundamental R bundles: rhdfs, rmr, and rhbase. Each of them offers diverse Hadoop highlights.

rhdfs is an R interface for giving the HDFS ease of use from the R reassure. As Hadoop MapReduce programs compose their yield on HDFS, it is anything but difficult to get to them by calling the rhdfs techniques. The R software engineer can without much of a stretch perform read and compose operations on dispersed information records. Fundamentally, rhdfs bundle calls the HDFS API in the backend to work information sources put away on HDFS.

rmr is an R interface for giving Hadoop MapReduce office inside the R condition. In this way, the R developer needs to simply isolate their application rationale into the guide and diminish stages and submit it with the rmr strategies. From that point onward, rmr calls the Hadoop spilling MapReduce API with a few employment parameters as information catalog, yield index, mapper, reducer, et cetera, to play out the R MapReduce work over Hadoop cluster.
rhbase is an R interface for working the Hadoop HBase information source put away in the appropriated arrange using a Thrift server. The rhbase bundle is outlined with a few techniques for reinstatement and read/compose and table control operations.Hadoop MapReduce and HDFS will be utilized inside the R support with the assistance of RHadoop rhdfs and rmr bundles. These bundles are sufficient to run Hadoop MapReduce from R. Fundamentally rhdfs gives HDFS information operations while rmr gives MapReduce execution operations.

RHadoop likewise incorporates another bundle called speedy check, which is intended for troubleshooting the created MapReduce work characterized by the rmr bundle.

Hadoop Streaming with R
Hadoop streaming is a Hadoop utility for running the Hadoop MapReduce work with executable contents, for example, Mapper and Reducer. This is like the pipe operation in Linux. With this, the content information document is imprinted on stream (stdin), which is given as a contribution to Mapper and the yield (stdout) of Mapper is given as a contribution to Reducer; at last, Reducer composes the yield to the HDFS catalog.

The principle favorable position of the Hadoop gushing utility is that it permits Java, and

non-Java, modified MapReduce occupations to be executed over Hadoop bunches. Likewise, it deals with the advance of running MapReduce jobs.

The Hadoop gushing backings the Perl, Python, PHP, R, and C++ programming dialects. To run an application writtenin other programming dialects, the engineer simply needs to interpret the application rationale into the Mapper and Reducer areas with the key and esteem yield components.

Understanding Data Analytics Life Cycle
The characterized information examination procedures of a venture life cycle ought to be trailed by groupings for successfully accomplishing the objective utilizing input datasets. This information investigation process may incorporate recognizing the information examination issues, planning, gathering datasets, information examination, and information perception.

Identifying the problem: Today, business examination patterns change by performing information investigation over web datasets for developing business. Since their information estimate is expanding step by step by step, their diagnostic application should be versatile for gathering bits of knowledge from their datasets.

With the assistance of web examination, we can take care of the business investigation issues. How about we accept that we have a huge web-based business site, and we need to know how to expand the business. We can distinguish the vital pages of our site by sorting them according to fame into high, medium, and low. Given these prevalent pages, their sorts, their activity sources, and their substance, we will have the capacity to choose the guide to enhance business by enhancing web movement, and content.

Designing data requirement: To play out the information investigation for aissue, it needs datasets from related spaces. Given the area and issue determination, the information source can be chosen and considering the issue definition; the information properties of these datasets can be opposed.

Forinstance,ifwewillperformweb-basedsocialnetworkinginvestigation(issuedetail), weutilizetheinformation sourceasFacebookorTwitter.Fordistinguishingthe clientqualities,werequireclientprofiledata,likes,an dposts as dataattributes.

Pre-processing data: In data analytics, we do not use the same data attributes, data sources, data

tools, and algorithms as all of them will not store data in the same format. This leads to the activities, such as data transformation, data aggregation, data profiling, data augmentation, data formatting, and data sorting, to provide the data in a structured and supported format to be used for data analytics.

In simple terms, pre-handling is utilized to perform information operation to interpret information into a field information arrange before giving information to calculations or instruments. The information investigation process will then be started with this organized information as the information. If there should be an occurrence of Big Data, the datasets should be designed and transferred to Hadoop Distributed File System (HDFS) and utilized further by different hubs with Mappers and Reducers in Hadoop groups.

Performing examination over information: After information is accessible in the required configuration for information investigation calculations, information investigation operations will be performed. The information examination operations are performed for finding important data from information to take better choices towards business with information mining ideas. It might either utilize unmistakable or prescient examination for business insight.

Analytics can be performed with various machine learning as well as custom algorithmic concepts, such as regression, classification, clustering, and model-based recommendation. For Big Data, similar calculations can be meant Map Reduce calculations for running them on Hadoop groups by interpreting their information investigation rationale to the MapReduce work which is to keep running on Hadoop clusters. These models should be additionally assessed and enhanced by different assessment phases of machine learning ideas. Enhanced or advanced calculations can give better bits of knowledge.

Visualizing data: Information perception is utilized for showing the yield of information examination. Perception is an intuitive approach to speak to the information experiences. This should be possible with different informationrepresentation programming and R bundles. R has an assortment of bundles for the perception of datasets. (a) ggplot2 like Plots for facet scales (b) rCharts like Dashboard charts

Conclusion

The advent of Big Data and Analytics is like a double-edged sword. This can be used for good as well as bad purposes. So, this must be used with good ethics and moral values. Some benefits include understanding customers better and providing them the exact solutions that they need even before they ask for it. Also, it can be used to point out some of the shortcomings that they have with their existing infrastructure and help them to increase their productivity. Some ill effects include selling customer information/statistics to other marketing firms, which can result in spying, scams, fraud, etc. So, next time you permit a mobile app or any website/store to collect your personal information, make sure to think twice, investigate the integrity of the company requesting the data and then take a call.

The important thing we can achieve using Hadoop is we are bringing the computation towards data rather than approaching data towards computation platform. Big Data analysis never involves throwing a lot of data into a computer and waiting for the output to pop out. The core R engine can process and work on the very limited amount of data. As Hadoop is very popular for Big Data processing, R can be made scalable by using a platform as Hadoop. Utilizing R with Hadoop will give a flexible information examination stage that will scale contingent upon the span of the dataset to be broke down. Experienced software engineers would then be able to compose Map/Reduce modules in R and run it utilizing Hadoop's parallel handling Map/Reduce component to distinguish designs in the dataset.

References

1) Big Data Analytics with R and Hadoop Vignesh Prajapati, Packt Publishing, 1st edition,2013.
2) Thota, S., 2017. Big Data Quality. Encyclopedia of Big Data, pp.1-5. https://link.springer.com/referenceworkentry/10.1007/978-3-319-32001-4_240-1
3) MachineLearningwithRBrettLantz,PacktPublishing,1stedition,October2013.
4) Hadoop For Dummies Dirk deRoos, Paul C. Zikopoulos, Bruce Brown, Rafael Coss, and Roman B. Melnyk, John Wiley & Sons, Inc., 1st edition 2014.
5) Hadoop Beginner's Guide Garry Turkington, Packt Publishing, 2013.

About the Author

Subash Thota works as Data Architect and specializes in Big Data, Cloud, Data Integration and Data Analytics with significantexperienceinProjectManagement,Agile,andDataGovernance.Subashhaswrittenseveralpapersinthe field of Big Data, the Cloud,andAnalytics.