Real Time Analytics for Non-Web Based Applications

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Arvind Jindal
(Roll No. 851232001)

Under the supervision of:
Dr. Inderveer Chana
Associate Professor
Computer Science and Engineering Department

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

July 2015

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "**Real time Analytics for Non-Web based applications**", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of ***Dr. Inderveer Chana*** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Arvind Jindal

(Arvind Jindal)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Inderveer Chana)
Associate Professor
CSED, Thapar University
Patiala

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. S. Bhatiaa)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

First of all, I am thankful to God for his blessings and showing me the right direction. With His mercy, it has been made possible for me to reach so far.

I would like to express my deep gratitude to my supervisor Dr. Inderveer Chana, Associate Professor, Computer Science & Engineering Department for her invaluable guidance, support and encouragement during this thesis work. She always provided a motivating and enthusiastic atmosphere to work with; it was a great pleasure to do this thesis under her supervision.

I am also thankful to Dr. Deepak Garg, Head, Computer Science and Engineering Department and Mr. Ashutosh Mishra, P.G. coordinator for providing us adequate environment, facility for carrying out thesis work. I express my gratitude to all the staff members of Computer Science and Engineering Department for providing seminars and encouraging towards research work.

I extend my thanks to Mr. Kedar Shiralkar, Saju Divakaran and Neeraj Garg for sharing their expertise and time to help me accomplish this work.

I want to express my appreciation to every person who contributed with either inspirational or actual work to this thesis. Last but not the least I am highly grateful to all my family members for their inspiration and ever encouraging moral support, which enables me to pursue my studies.

Arvind Jindal

Arvind Jindal

Abstract

Enormous software applications, both web and non-web based are built by organizations as a solution to solve problems. However, in today's competitive world providing a solution is not sufficient. One also needs to keep a track of the usage of the applications, activities being performed by the users on the application and look for the areas which if improved act as customer delight. Also, it is becoming clear that such data analytics needs to be carried out in real time for the business success. Although it is easier to track and analyze web based applications, but the same is not true for non-web based applications.

In this thesis, a framework is proposed which would help one perform such real time data analytics for non-web based applications possible over cloud. The performance and accuracy of the framework is evaluated with the help of a case study. The experimental results demonstrate that this framework effectively helps provide real time data insight for non-web based applications.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

This chapter familiarizes progression of Cloud computing and discovering its features offered and several deployment models. It will also deliver an insight about the web and non-web based applications and the necessity for accomplishing real / near real time analytics on top of these applications shadowed by the organization of thesis.

1.1 Cloud Computing Evolution with Time

Cloud computing is a term coined to describe the set of services being provided to the customers over a network on demand. One can also say that cloud computing is the internet based computing, whereby several IT services i.e. shared resources, useful data and applications are provided to users on demand. Cloud computing is not something which appeared overnight. It has evolved through various phases as below:

- **Mainframe Computing:** In the past, the information infrastructure of various organizations was powered from mainframe. There was one physical location where a large, powerful computer managed and ran various software applications. Although multiple applications could be easily supported over one mainframe, but managing and maintaining this large hardware was quite inefficient and expensive.
- **Distributed Computing:** With low cost computing becoming more available, the solution to the above problem of mainframe computing was to have multiple cheaper computers, instead of an expensive mainframe. Each computer can run independently leading to higher reliability. However, communication cost was high and coordination between these independent machines was difficult.
- **Cloud Computing:** Today cloud computing is in place. The cloud can be compared to a shared network of resources on which the people and organizations can run the software and store data. Rather than providing products to the clients as in the earlier phases the cloud providers provide services for which the client

need to pay as per the usage. In this the client is allowed to use the services of cloud but they don't own any part of it.

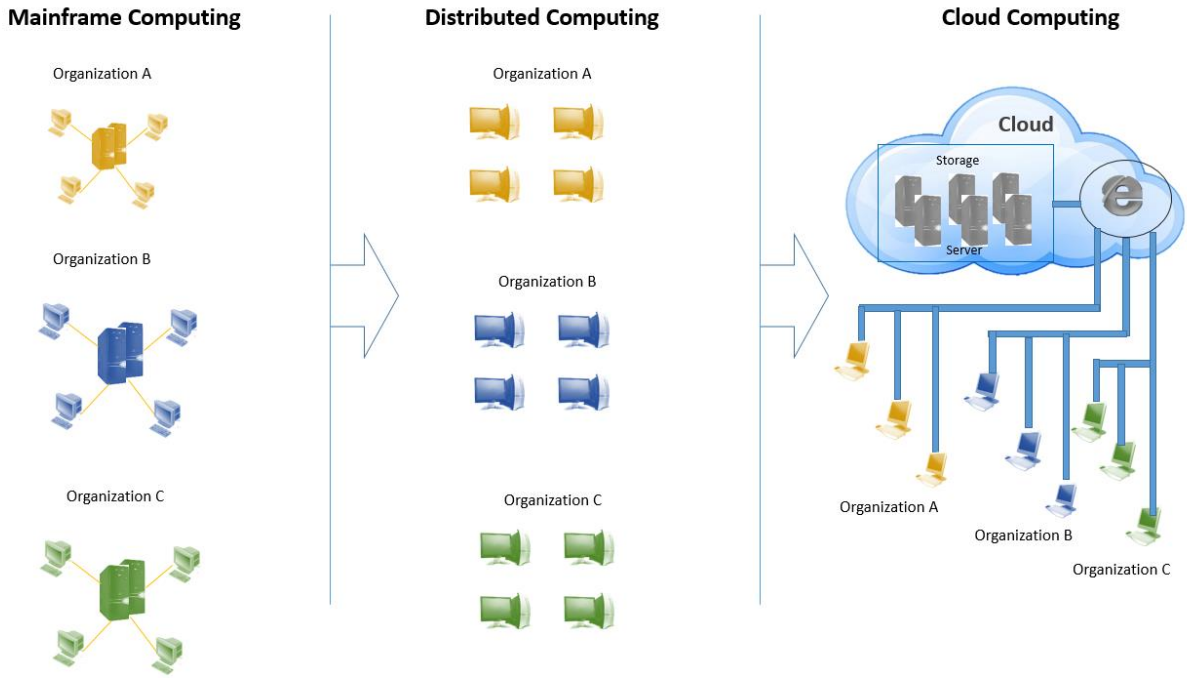


Figure1.1 Evolution of Cloud Computing

There are various characterizations for cloud computing, but all of them harmonize in that it is a new exemplar grounded on a pay-per-use model to compliantly access hardware and software resources through Internet, permitting companies to upsurge performance and reduce budgets.

For better understanding of the Cloud Computing model one needs to comprehend three broad areas which outline the core of Cloud Computing and distinguish them from the prevailing computing models which include:

1. Essential Characteristics
2. Service Models
3. Deployment Models

The critical features of which the cloud model is made of are as follows:

- **On-demand** – One can compare it with pay-per-use as is the instance with electricity for which person pay for the quantity of electricity that have been utilized. Under this prototypical model, users pay solitary for what they need and for what they have used according to the essential capacity. The organizations

have to pay for the software and hardware, even though these resources not required at their full volume.

- **Network access** – The resources that are essential for retrieving and altering are situated outside the company and therefore need certain mechanisms to solve the purpose.
- **Pooling of Resources** – The cloud workers own a set of servers and other storage devices to aid its patrons. Henceforth a multi occupant model is used to assist numerous customers with dissimilar physical and virtual resources being allocated and reallocated according to the need and claim of the customer. It generally provides the customer a sense of position individuality, i.e. having no information or regulation over the exact position of the provided resources.
- **Rapid provision** – Resources can be loosely provisioned and released, even routinely, to meet the demand of the client .This typical meets one of the major concerns of Business Intelligence. Here one needs to gauge the data barn once it reaches its supreme storage and performing capabilities.
- **Measured service** –It provides transparency for both the consumer and provider of the service by inbuilt capabilities for supervising, directing and recording of the resource usage.

Cloud can be classified on the basis of service models. These are as follows:

- **Infrastructure as a service (IaaS)** – It involves providing hardware related services to the provider using the principles of cloud computing. These could include some kind of virtual servers or storage services, i.e. disk or database storage. Leading vendors in IaaS are Amazon S3, Amazon EC2, Flexiscale, etc.
- **Platform as a Service (PaaS)** – It includes complete development platform offering over cloud. Leading players providing Platform as a Service are Microsoft Azure, Google App Engine, force.com of Sales force, etc.
- **Software as a Service (SaaS)** – It involves offering complete software on the cloud. The software application being hosted by the cloud vendor can be used by the usesr on the pay-per-use basis. Some of the major market leaders in this field are Gmail, Google docs, hotmail and Business Productivity Online Suite (BPOS), etc.

Nowadays when one talks about applying intelligence to Business with Cloud Computing two more services can be thought of offering over cloud which are:

- Data as a Service (Data management as a Service)
- Analytics as a Service [6]

Further, now a days since the organizations are interested in real time data. So, analytics as a service can be further extended to Real time analytics as a service.

As defined by Gartner, one can classify cloud computing model on the basis of the target audience as [8]:

- **Public Cloud Computing Model** – It is a style of computing where scalable and elastic IT-enabled capabilities are delivered as a service to the external customers using Internet technologies.
- **Private Cloud Computing Model** – It is a style of computing where scalable and elastic IT-enabled capabilities are delivered as a service to the internal customers using Internet technologies.

Further two other types of cloud computing models have been described by NIST on the basis of deployment as below [4]:

- **Community Cloud Computing** – In this type of cloud computing the infrastructure is provisioned for exclusive use by a specific community of users from organizations having common concerns.
- **Hybrid Cloud Computing** – Here the cloud infrastructure is a composition of two or more distinct cloud infrastructures (public, private, or community)

1.2 Software applications

All the available software applications in general can be fragmented into two basic types which are as follows:

1. Web Based Applications
2. Non-Web Based Applications.

An application in general is any piece of algorithm / software that delivers commands and directions to the computer. This furthermore enables it to accomplish a particular task precisely.

Web based applications are intended and anticipated for accessing information via a web browser or any client that aids as a kind of interface.

Non-Web based apps are envisioned to be used without any access via any browser over interned. These are in fact proposed for offline usage only.

1.2.1 Web based applications

All the web based applications have one thing common, i.e. they require a mandatory web browser service to be run. Such programs which necessarily run on web are produced in a browser-compatible programming platform like CSS, JavaScript, HTML or a combination of these. Also any kind of web based application is fully relied on a web browser to solidify the usability.

Web applications are prevalent and dominant because of their ubiquity of web browsers. Also the ease and the convenience provided by using a web browser as a client escalates its popularity all the more. The skill to apprise and preserve web applications without allocating and connecting software on possibly millions of customer computers is a key factor for their universal acceptance. The essential support for cross-platform compatibility makes it very admirable in the world of internet. Most commonly used web applications include shopping websites, Wikipedia, webmail services and many other functions.

One of the most prevalent disadvantage of such kind of applications are security fissures and splits. These are of major alarm as they encompass both enterprise evidence and confidential client data. Defending these significant assets is a noteworthy fragment of any browser reliant application and there are some crucial functional areas that must be involved in the expansion process. This embraces procedures for authentication, authorization, input, and logging and examining. Constructing security into the applications from the foundation can be more operative and less troublesome.

Web applications are software as a service (SaaS) in cloud computing model web. There are business submissions provided for SaaS for enterprises for usage reliant fee. Other web applications are often obtainable charge free, producing revenue from billboards and commercials revealed in web application interface.

Many businesses are permitted by open source applications such as retail or e-commerce businesses that simplifies effortlessly creating an online marketing store. Originations in all characteristics of web applications are delivering remarkable economic value by aggregating competition by dropping fences to let the new companies enter the market.
Some of the most common examples of browser applications are presentation tools, spreadsheets etc. There can also be more advanced applications that are web reliant such as video editing, project management and computer-aided design etc.

1.2.2 Non Web based applications

A non-web based or a desktop application refers to any application which can be installed on computer (laptop or a desktop) and used for specific tasks only. These applications can be used by multiple users in a networked environment.
While working on a non-Web-based application, access is frequently achieved by running an executable program on the user's computer. Unlike a Web app, the users are restricted to the discrete sitting at the keyboard and watching the screen. Additionally, an application intended for offline use depends on its own procedures for both program implementation and the user interface.

1.2.3 Architecture diagram for Software applications

The software applications are frequently fragmented into rational logical masses which are called "tiers". Basically every tier is allotted a role.
Old-style applications entail only of 1 tier. This tier resides on the client machine, but web applications offer themselves to an n-tiered methodology by nature. Although many discrepancies are conceivable and plausible, the utmost shared assembly is a three-tiered architecture.
The three-tier architecture model, which is the central framework for the logical design model, fragments an application's modules into three tiers of services. These tiers do not essentially resemble to physical locations on several computers on a network, but noticeably to logical layers of the application. How the parts of an application are allocated in a physical topology can alter, subject to the system requirements.
Below are the concise reports of the services allotted to each tier:

1. The **presentation tier**, or user services layer, hands a user access to the application. This layer offers data to the user and optionally allows data manipulation and data entry. The two main groupings of user interface for this layer are the Non-Web based application and the Web-based application. Web-based applications frequently constitute most of the data manipulation structures that traditional applications work on. This is achieved buy the use of Dynamic HTML and client-side data sources and data cursors.
2. The **middle tier**, or business services layer, comprises of business and data rules. Also denoted as the business logic tier, the middle tier is where developers can resolve mission-critical business difficulties and attain major efficiency advantages. The modules that comprise this layer can occur on a server machine, to support in resource sharing. These modules can be used to impose business rules, for example business algorithms and governmental or legal rules, and data rules, which are planned to keep the data structures reliable within specific or multiple databases. As these middle-tier modules are not bind to a specific client, they can be consumed by all applications and can be shifted to different locations, as response time and other rules demand. For example, simple edits can be positioned on the client side to reduce network round-trips, or data rules can be positioned in stored procedures.
3. The **data tier**, or data services layer, networks with persistent data typically stored in a database or in permanent storage. This is the real DBMS access layer. It can be retrieved through the business services layer and on event by the user services layer. This layer comprises of data access modules (rather than raw DBMS connections) to help in resource sharing and to permit clients to be designed without installing the DBMS libraries and ODBC drivers on single client.

Throughout an application's life cycle, the three-tier methodology delivers benefits such as flexibility, maintainability, manageability, reusability, and scalability. One can reuse and share the modules and services build and can allocate them across a network of computers as required.

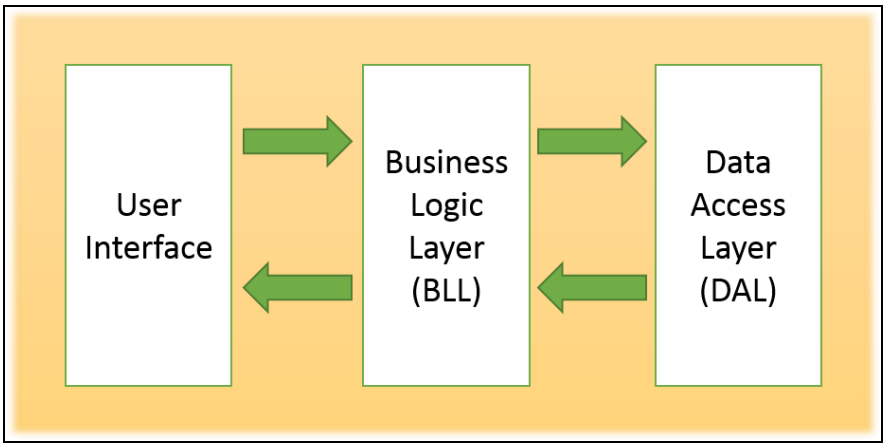


Figure 1.2 3-Tier Architecture of Software Applications

1.3 Analytics

Analytics is the study of data via various quantitative methods such as statistics, simulation, optimization along with descriptive and predictive data mining to yield insights which are unlikely to be discovered using the usual methodologies of business intelligence (BI) (for example query and reporting).

Analytics can be broadly categorized into 4 types:

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

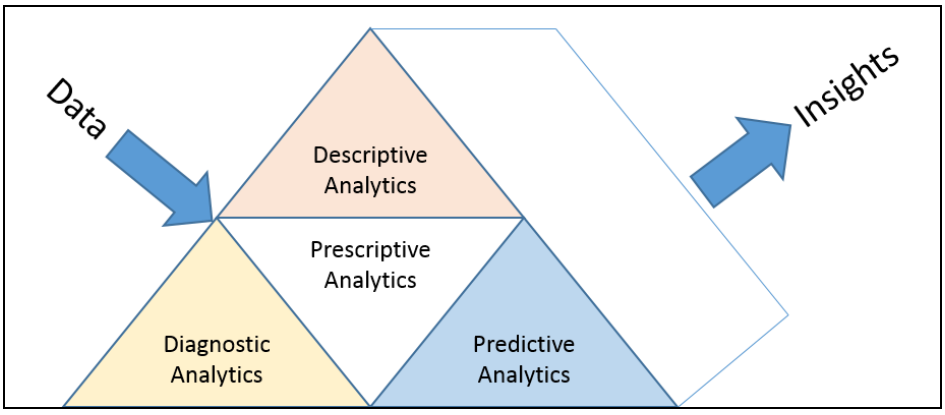


Figure 1.3 Type of Analytics

1.3.1 Descriptive Analytics

Descriptive Analytics is used to answer the “What” part of the business, i.e. "What happened and what is happening?" This uses various data mining and data aggregation techniques to provide the insight about the past and the present business situation. Some of the common examples are the reports from which one can get the insights of the company’s production, operation, inventory, etc.

1.3.2 Diagnostic Analytics

Diagnostics Analytics is used to answer the “Why” part of the business, i.e. "Why something happened in the business?" One cannot fix something is he does not know why that thing happened. So, one can use the diagnostic analytics to drill down to the lowest level to get the insight of why that thing happened. Various visualizations can be used to uncover the correlations and patterns and provide the reasons for why the revenues of the company is down or why the sales is high, etc. One can spot the essential factors which directly or indirectly affect the business.

1.3.3 Predictive Analytics

Predictive Analytics is used to predict the future, i.e. "What is expected to happen next?" This is studying the historical and real-time data, identifying the patterns and hence predict the future. One can use predictive analytics to identify the upcoming risks and opportunities and hence help grow the business.

1.3.4 Prescriptive Analytics

Prescriptive Analytics helps to advice the business users on the possible outcomes, i.e. help users “prescribe” various possible actions and further guide them towards a solution. On top of “what would happen” it also provide information on “why it would happen” and recommendations to what should be done next.

1.4 Research Motivation

Based on the Gartner’s Priority Matrix for Emerging technologies it has been found that analytics is one the most emerging technologies. The organizations are in dire need of getting the real time data analytics so that they can get the insight of the data and better plan and get maximum benefit from that.

Organizations are providing enormous software solutions (both web and non-web based) to solve real world problems. Although it is easier to track and analyze the web based applications but the same is not true for non-web based application. Aim of this thesis is to provide a framework which would help in getting the real time insight of how the non-web based applications are being used across the globe.

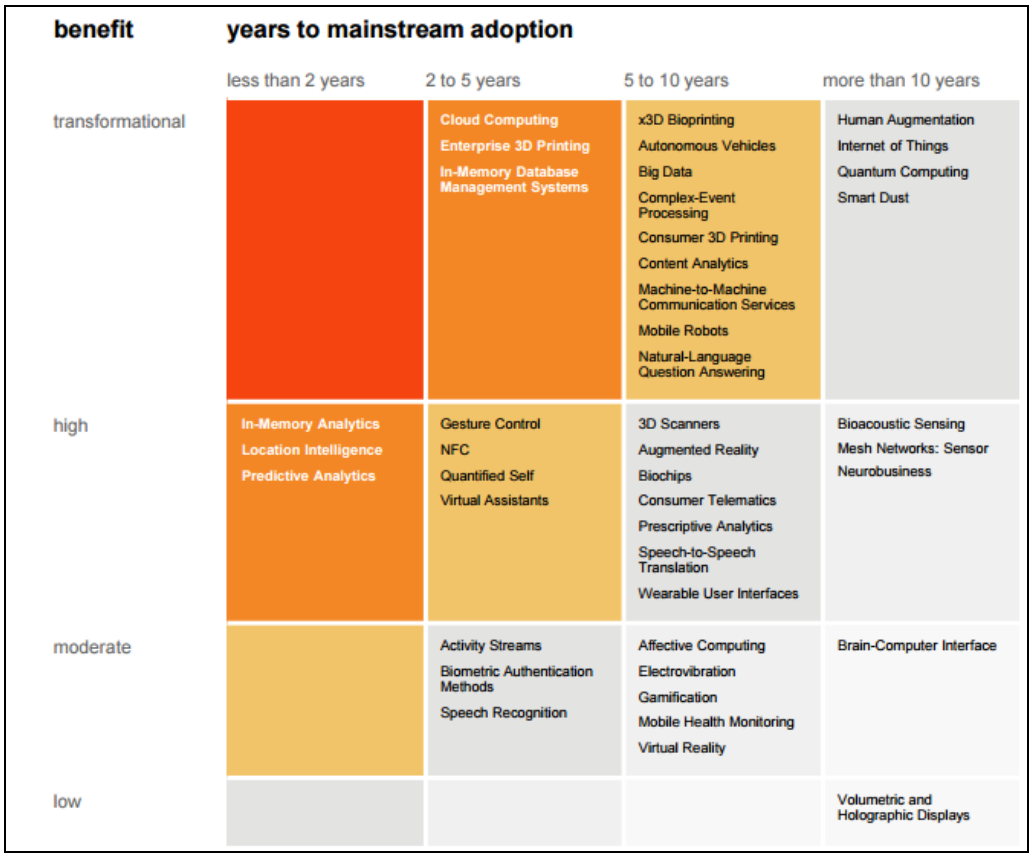


Figure 1.4 Gartner's Priority Matrix for Emerging Technologies [23]

1.5 Thesis Contribution

Below is the thesis contribution:

1. In this thesis various existing real time analytics techniques have been analyzed and compared.
2. A framework has been designed which can be used to perform real / near-real time analytics for non-web based applications.
3. The proposed framework is implemented using Google analytics and SQL server components and further the results are compared with R.

1.6 Organization of Thesis

The rest of the thesis is organized as follows:

Chapter 2: This chapter contains exhaustive description of literature survey done to study the importance of real time analytics for the organizations and the existing implementation of real time analytics over cloud for the software applications.

Chapter 3: This chapter presents the problem statement along with the objectives if this research work.

Chapter 4: This chapter describes the framework that have been built to solve the stated problem.

Chapter 5: This chapter focuses on related concepts followed by implementation details and experimental results.

Chapter 6: In this chapter conclusion, followed by possible future research work is discussed.

Chapter 2

Literature Survey

This chapter examines about the real time analytics for different software applications and the current implementation to gather the need. It also constitutes the tabular evaluation between the current implementations.

2.1 Real Time Analytics

Real-time analytics is an expression used to recommend to the analytics which are able to be recovered as they arise into a system. In different words, the word analytics is utilized to define data patterns that offer meaning to a business or other unit, where analysts gather valuable information by classifying through and analyzing that data.

While the term real-time analytics suggests practically prompt access and use of analytical data, some specialists offer a more concrete time frame for what institutes real-time analytics, such as implying that real-time analytics includes data used in a time frame of one minute of it being entered into the system. A usual example of real-time analytics is a system where managers or others can remotely analyze order information that's revised as soon as an order is made or handled. By staying associated to an IT architecture, these users will be adept to see the orders embodied as they happen, hence tracing orders in real time.

Few more examples of real-time analytics would be any repeatedly updating or refreshed results regarding user events by customer, for example page views, shopping cart use, website navigation, or any other kind of online or digital activity. Such kinds of data can be exceedingly important to businesses that need to conduct dynamic analysis and reporting in order to rapidly answer to trends in user behavior.

2.1.1 Traditional Analytics Approach

Conventionally, major part of business intelligence and analytics solutions have been planned and constructed around the insight of batch operations that transfer data between different persevered data stores like analytics cubes etc. Then the users will have to issue

a request against the data at rest for provisioning situations like dashboards, ad-hoc analysis etc. Although the method has confirmed to be exceptionally positive and progressive over a couple of years, and still continues to be exceedingly appropriate solution to various business processes in today's world, the absolute capacity and rapidity of data being produced by today's applications or campaigns is putting a lot of stress on this prevailing pattern.

To avoid some of the experiments linked to appropriateness of information being made obtainable to conclusion support systems, corporations have often observed to enhance their prevailing architectures using methods like quickening data loading time tables, or physically scaling out data alteration developments.

In some occurrences this has demonstrated to be a positive remediation for producing data to be available quicker, although data availability is commonly still restricted to minutes at best. There is a limit as to recurrently that an excerpt/ load development can be instantiated and accomplished in contradiction of a source system prior to being resumed.

2.1.2 Analytic Techniques

Various Data Mining techniques are used to perform analytics on the set of data and discover the hidden information. Data mining is the combination of statistical analysis and Artificial Intelligence to discover the "hidden" information from the data. Once the hidden information is revealed important business decisions can be taken. Data mining can be broadly categorized into 3 categories:

- Classification/Clustering
- Association Rules
- Sequence Analysis

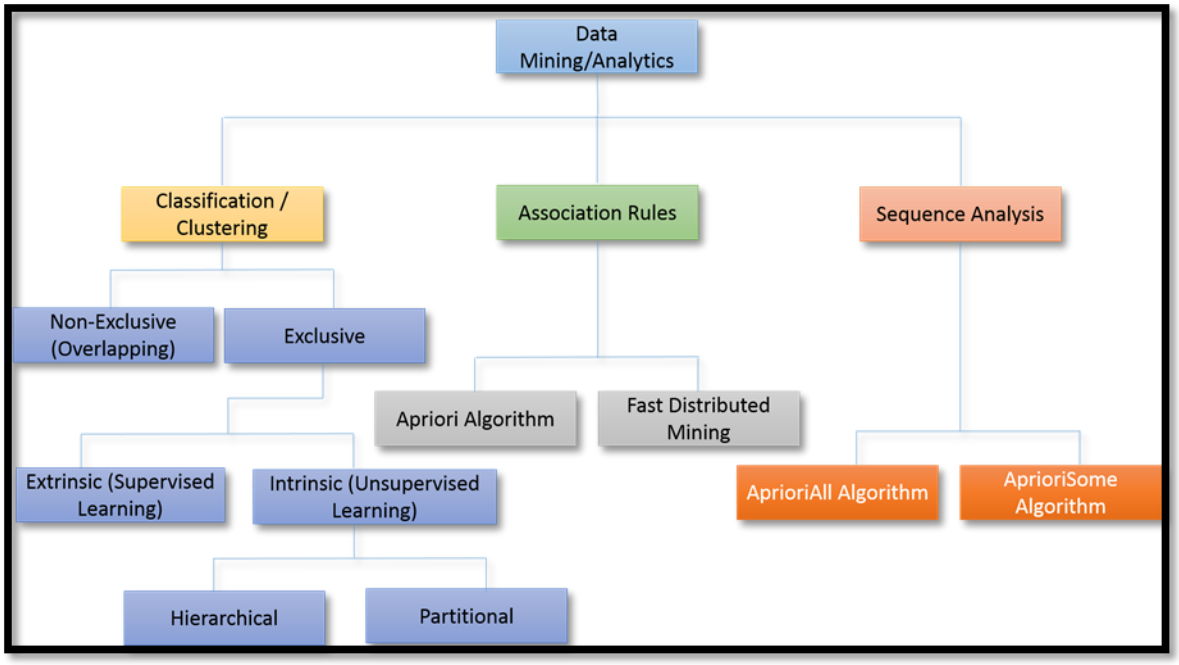


Figure 2.1 Data Mining Techniques

2.1.2.1 Classification/Clustering

Clustering refers to combining or grouping similar objects together. The data set is analyzed and grouped on the basis of certain rules to categorize upcoming information. The significant problem in grouping is to determine the rules which help partition the data into categories.

Nonexclusive vs Exclusive

In Nonexclusive classification, an object can be classified to various classes, whereas in exclusive classification and object can be categorized under only one cluster. E.g. grouping of a set of people according to age is exclusive categorization while grouping a group on the basis of disease is nonexclusive categorization.

Extrinsic vs Intrinsic

Using both the proximity matrix and the category label to classify the objects is categorized under extrinsic or supervised learning, while using only proximity matrix for classification is Intrinsic or unsupervised learning.

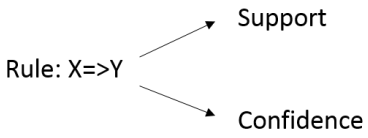
Hierarchical vs Partitional

Depending on the type of structure imposed on data intrinsic classification is further categorized into Hierarchical and Partitional. Division of objects into non-overlapping

clusters is known as partitional classification. When one further sub divide the clusters in sub-clusters it is known as Hierarchical clustering.

2.1.2.2 Association Rules

A set of various if-then statements used to discover the hidden relationship between unrelated datasets, relational databases, etc. are known as Association rules. These set up relationships between the objects which are commonly used together. For e.g. if a person buys a bread than he would also buy a butter, etc. The association rules follow basic confidence-support criteria, i.e. the association rules usually satisfy the user supplied minimum support and confidence at the same time.



Apriori Algorithm

Level wise search strategy is used in this algorithm, in which k itemsets are used to explore k+1 itemsets. This process continues as long as the datasets appear often in the database. The method used to implement Apriori algorithm is BFS (Breadth First Search) and the structure used is hash tree type.

Fast Distributed Mining (FDM)

It is the distributed data mining algorithm which follows the same principle as the Apriori algorithm. However, in this technique an association between the global data sets and the local data sets is determined which helps generate less number of candidate sets which further reduces the messages to be passed.

2.1.2.3 Sequence Analysis

The process of extracting various sequential patterns with the support of these patterns exceeding the minimal threshold support is known as Sequential pattern mining. It helps to extract those sequences which reflect the most frequent behavior of the sequence database. The data being considered for extracting the sequence is the one that appears in separate transactions.

Apriori-all vs Apriori-some

Apriori-all and apriori-some algorithm uses the concept of count-all/some algorithm in which either a person include maximal and non-maximal sequences or only the maximal sequence. Apriori-all is count-all algorithm in which both the maximal and non-maximal sequence are included, while apriori-some is a count-some algorithm in which only the maximal sequences are considered. However, one need to consider that not too many longer sequences are considered which don't have minimal support.

2.1.2.4 Commonly used analytic algorithms

The k-means algorithm

This is an iterative algorithm to partition the given data set into k (user specified) clusters. Initially the k points are chosen as the centroids from the input data set having n data points. Then the below two steps iterates till convergence:

- The entire data set is partitioned into n clusters by assigning each data point to its closest centroid. The closeness of the data point is determined by the Euclidean distance between the data point and the centroid.
- Each centroid is relocated to the center of all the data points of that cluster.

The algorithm converges once no relocation of centroid occurs.

The Apriori algorithm

It is a level-wise complete search algorithm used to find item sets from the given transaction data and derive various association rules. Initially Apriori scans the entire dataset and search for the frequent item sets of size 1 which satisfy the minimum support criteria. Then the algorithm iterates on the below three steps and determines all the frequent item sets.

- Generate the candidate frequent item sets of size k+1 using the frequent item set of size k.
- Scan the dataset to determine the support of all the frequent item sets.
- Add the item sets which satisfy the minimum support criteria to the set (F_{k+1})

PageRank Algorithm

It is a search ranking algorithm which utilizes the hyperlinks on web. This algorithm uses the link structure of the web to determine the overall page quality. The rank prestige is used to determine the pagerank algorithm:

- Depending upon the number of in links to a page the prestige is determined. The more the in links, more is the prestige value of that page *p*.
- The pages pointing to page p have their own prestige value also. So, a page with higher prestige vale pointing to page *p* is more important than the page with low prestige value pointing to page *p*.

DBSCAN Algorithm

DBSCAN is a density based clustering algorithm which help separate regions of low density from another regions of high density. Below is the algorithmic details for implementing DBSCAN:

- Mark the data points as core, border and noise points.
- Once the points are identified, eliminate the points which are noise points.
- Mark the edge between all the core points which are within the specified distance.
- Segregate all the groups of core points into separate clusters.
- Each border point is assigned to one of the clusters of the core points.

2.1.3 Frequent Scenarios for Real-Time Analytics

Real-time incident based analytics and decision engines can be functional to a range of industry scenarios. The below sections provides a subset of examples of where a real time data analytics solution can append value for business.

Inventory Management [11]: Optimized inventory control can assist businesses make sure that retail store shelves keep up a repetitive level of stock to assist random purchasing habits. Understocking items can cause missed sales opportunities, and further worsen, the likely loss of a customer to a player. Customers who are incapable to buy one item, mainly at the beginning of the shopping process, more tending to not buy any items and in its place move straight to a similar retail store.

By implementing a real-time analytics system to associate present stock on the shelf beside items being checked-out at the inventory, stock clerks can be clued-up in real-time regarding declining shelf volumes of any specific product line. This situation becomes

important during periods of wavering demand, such as seasonal shopping where a particular product (for example, Christmas caps) may see considerably higher demand from users, following in low shelf stock scenarios.

Demand Based Price Elasticity: For industries that keep highly elastic pricing models depending on demand, the beginning of real-time analytics and decision making is an essential solution for sustaining scenarios of rapid demand growth where price adjustment need to be automated.

Think of a hotel chain or an airline. Both of these industries continually regulate their product pricing according to seasonal demand. Such shifts in demand, for example a large conference in a city, can effect travel reservations for a particular destination. In this example, real-time analytics can assist discover a burst in the rate of bookings being done. By leveraging sequential queries, such as the count of bookings in a range of past 30 minutes, recalculated every 1 minute going ahead, the system can instantly detect an amplify in the rate of bookings over time as it is eventual, and consequently either change pricing customarily or inform an operator that a surge of bookings is presently happening.

Infrastructure Monitoring [11]: Data centers continually grow in reaction to the surge demand for compute and storage resources. Big enterprises can house thousands of servers, while cloud hosting providers can hold tens or hundreds of thousands of servers. For operations to sustain this scale of infrastructure in a consistent and proficient manner, a traditional reactive technique will not scale suitably to support these operations in a cost effective way.

Infrastructure happenings such as above average CPU consumption or expanded temperature rise can be found and surfaced with prioritization with operational dashboards for consideration. Furthermore, machine learning models can be functional and invoked to find out shifting capacity necessities or potential equipment failures former to the value of service being ruined.

Device Telemetry [11]: The count of connected devices is rising at an alarming rate. While few connected devices for example security badge readers are prime in their acceptance, others such as vehicles or smart buildings are very much in their maturity. By allowing real-time analytics for joint vehicles, manufactures allow new levels of visibility into the current running performance of a particular vehicle, rather than waiting to

download telemetry data scheduled vehicle service. This can allow further examples such as predictive maintenance where the health of the vehicle is constantly validated beside proven machine learning models to notice anomalies in the vehicle.

Big commercial and industrial buildings depend on a number of systems to assist day to day operations. One instance is the air handlers that assist climate control necessities within the grounds. By relating real-time airflow data from the air handlers with the alteration of temperature in the building, operators can find in real-time if the equipment is performing at optimum levels.

Website Analytics / Content Management [11]: Websites can create a vast amount of data on the basis of users' interactions. By knowing metrics for example time being spent on a page or click paths within the site, web designers can keep on refining layouts and satisfy to suit the user. Real-time web metrics can be advantageous to know the result of advertising venture planned to move people to a site. Moreover, by leveraging real-time telemetry and analyzing site actions, developers can execute highly interactive A/B testing of site layouts to repeatedly iterate changes and optimizations. This can be specifically valuable in sites that familiarize heavy usage.

Fraud Detection [11]: The cost of fake transactions for large scale financial institutions can rush into the millions of dollars per year. Some scenarios comprise identity theft and stolen credit card particulars. Solving for these scenarios is a specific time sensitive matter as the amount of time in use to find out a fake transaction and automatically cancel a credit card can straight force the bottom line of a card issuing company.

Real-time analytics and decision engines are specifically well-suited to support this scenario as each payment can be analyzed as the dealings occur. This may comprise processes such as counting the rate of transactions that are happening for a single credit card, or relating the use of the same card through different geographical distances. It would definitely look doubtful for the same card to be used at a retail outlet in Boston and Melbourne at the same time!

Additionally, by leveraging a machine learning model based on historical spending patterns for the card, a decision engine can find if a seemingly legitimate purchase has a superior probability of being fake for that given credit card. This enables customized insights rather than a more general rules based approach.

2.2 Comparative Analysis of Existing Solutions

Various prominent research groups both in academe and organizations are performing in the area of real time analytics for numerous software applications. In this segment, have discussed recently developed solutions followed by a tabular comparison of the existing techniques.

- **Gianluca Privitera, Stephen P. Emmons and Giacomo Ghidini** [1] proposed a solution for soft real time analysis of GPRS traffic using Apache Spark. The suggested solution acquires GPRS traffic, manages it, and decorates it with facts about the appliances, networks, and Machine to machine applications. It then calculates a complete array of information that is presented in maps and graphs on a live Web console, or provided to other applications for further processing.
- **Shuang Chen, Yanzhi Wang, Mahboobeh Ghorbani, Massoud Pedram, Paul Bogdan** [2] have proposed fractal modeling technique to perform analysis on trace based statistics and carry out predictions about the job inter-arrival time and accumulated resource demand. The allocations of essential parameters comprising job interval time and resource request per task in terms of CPU, storage and memory are mined from the dataset and extrapolated using alpha stable distribution.
- **Charles Feddersen** [3] has provided a framework using Microsoft Azure to perform event based real time analytics on streaming data rather than performing it on stored data. Consumers can advantage from the execution and expenditure elasticity provided by cloud based PaaS to fulfill elastic, competent framework that can ascend based on industry needs and varying market dynamics.
- **Alina M. Chircu, Eldar Sultanow, Flavius C. Chircu** [4] have come up with a solution performing track and trace based analytics for supply chain of pharmaceutical sector using SAP HANA. They have secured maximum benefit by utilizing in memory computing capability to extract and transform the large amount of data for efficient medication tracking and tracing.
- **Alex Guazzelli, Kostantinos Stathatos, Michael Zeller** [5] using open standards and cloud computing framework have come up with a solution which allows easy

exchange between various analytical applications. A scoring engine platform based on predictive analytics has been described which leverages the above elements to provide a useful deployment process for various analytical models using Predictive Model Markup language.

- **Shiming Zhang, Yin Yang, Wei Fan, Marianne Winslett** [6] have designed and implemented a real time analytical solution which works for data generated from web browsing done on mobile and call logs of a mobile operator using OceanRT. OceanRT extensively uses the RDMA operations, which lessens networking charges without and need for particular hardware. Also, it utilizes the capabilities of parallel computing of every node along with the storage schemes in the cloud via a novel architecture which consists of Access-Query Engines to further optimize the queries having multiple dimensions and joins.
- **Qiming Chen, Meichun Hsu, Hans Zeller** [7] proposes a continuous query model supporting both dynamic streaming and static relations data. This allows a long SQL query instance to execute in a cycle via cut-and-rewind mechanism. A table-ring and label switching mechanism has been developed to have the continuously generated analytics results staged efficiently. The proposed solution has been built by extending the capabilities of PostgreSQL engine.
- **Joao Cardoso** [8] has developed a solution which helps analyze and visualize the activities performed on web based application by any user. The technologies stack used to incorporate the above solution are Ruby, Rack, the Ruby on Rails and MongoDB. One needs to append a tracking code in the header or footer section of the web page for getting the real time usage of the web site in terms of number of visits to the site, unique users accessing the site, etc.
- **Georgia Fotaki** [9] has presented two frameworks, which is a first step towards an effectual online customer dissection approach efficient for assisting an online promotion policy. First framework help select the segregation type of the customer based on Clustering, classification, association, etc. and the second framework helps to apply the relevant algorithm for analysis based on the segmentation type of the customer.

- **E. J. Lingerfelt, S. S. Desai, O. E. B. Messer, E. J. Lentz, and C. A. Holt** [10] have developed a software application, Bellerophon, to support a High performance computing solution called CHIMERA. This application helps the geographically spread team of associates to carry out near real time data analysis. It also helps perform regression testing on a variety of supercomputing platform.

Sr. No.	Technique	Framework/Technology Stack Used	Merits	Demerits
1	Real-Time Big Data Analytics for the Enterprise	SAP HANA, Intel, Apache Hadoop	- Provide a common authentication and authorization framework for financial, government, healthcare and e-commerce environment. - End to End security - Fast, transparent data encryption	The applications should be connected via internet to enable real time analysis.
2	Soft Real-Time GPRS Traffic Analytics for Commercial M2M Communications Using Spark [1]	Spark, Spark SQL and Spark Streaming	- Provides soft real time analysis for applications running over wireless sensor network or using M2M communication.	- Better algorithms can be used for packet dissection, transaction reconstruction, and context reconstruction. - The framework can be used only for applications running over wireless sensor networks.
3	Trace-Based Analysis and Prediction of Cloud Computing User Behavior Using the Fractal Modeling Technique [2]	Fractal Modeling technique	- Helps in efficient resource management based on the predictions from data captured from job execution duration, inter arrival time, etc. using Fractal modeling technique.	- Usage limited for cloud based applications.
4	Real Time event processing with Microsoft Azure stream analytics. [3]	Microsoft Azure	- Serves as a blueprint for designing and deploying real time event processing solutions with Microsoft Azure.	- Solution can only be used with Microsoft Azure. Cannot be leveraged with some other cloud provider.

5	Cloud Computing for Big Data Entrepreneurship in the Supply Chain: Using SAP HANA for Pharmaceutical Track-and-Trace Analytics [4]	SAP HANA	- Track and trace analytics can be performed for Pharmaceutical firms to monitor safety and security in supply chain. - In memory computing approach has been used to process huge amount of data.	- Focused only on pharmaceutical sector, Radio Frequency Identification tags are essential for tracking and performing analytics.
6	Efficient Deployment of Predictive Analytics through Open Standards and Cloud Computing [5]	Predictive Model Markup Language (PMML)	- Predictive Model Markup Language capabilities have been used to allow models to be easily deployed between analytical applications. - These predictive models can be accessed from anywhere in enterprise by web services.	- The existing models needs to be represented in PMML to use this deployment approach.
7	Design and Implementation of a Real-Time Interactive Analytics System for Large Spatio-Temporal Data [6]	Optimized OceanRT framework	- Real time analytics can be performed on Spatio-Temporal data using OceanRT. - Uses a three system design of RDMA links, Acess query engines and improved storage scheme for better performance over existing system.	- Limiting to only Spatio-Temporal data set and queries such as mobile web browsing data and call logs.
8	Experience in Continuous analytics as a Service (CaaaS) [7]	Cycle-Based Framework, PostgreSQL	- Works on the framework which eradicates the limitation of store first and analyze later approach by querying data stream chunk by chunk rather than querying static table.	- Connection management, privacy, security related issues are not taken into consideration.
9	Real-time Web Analytics [8]	Ruby, Rack, Ruby on Rails web framework, MongoDB	- Helps to track every minute happening over internet in real time.	- The solution helps to track only the web based applications and has no technique to track non web based

				applications.
10	Exploring Big Data Opportunities for Online Customer Segmentation [9]	Apache Hadoop, R programming Language, Apache Solr, Micro strategy	- Combines the three important fields: Online advertising, Customer segregation, and Big Data Analytics in one framework to help real time analytics for extremely huge customer datasets.	- Efficient OCEM (Online Customer engagement management) tools can be used to capture the attributes of the online customers.
11	Near Real-time Data Analysis of Core-Collapse Supernova Simulations With Bellerophon [10]	Bellerophon	- Helps to provide near real-time analysis for supernova events at petabyte level to CHIMERA's geographically dispersed group of people.	- This application has been built to support and perform near real time analysis for only one application, CHIMERA. - Not enabled for Mobile phones.

Table 2.1 Comparative Analysis of Existing Solutions

The above table shows the comparison of the existing solutions which helps in performing real / near real time analytics for different software applications. These techniques used different large scale datasets to perform analysis. Also some of the techniques are applied only to a limited set of applications.

2.3 Conclusion

In this chapter literature review and comparative analysis of various framework / techniques have been done which are used for performing real time analysis. These techniques are compared and their merits and demerits are presented in tabular form. The next chapter provides the difference between the above techniques/ frameworks and the problem statement along with the objectives of this thesis work.

Chapter 3

Problem Statement

In the previous chapter various frameworks / techniques available for performing real time analysis for software applications were discussed. This chapter focusses on the problem statement taken up in the thesis.

3.1 Gap Analysis

In today's competitive world, exploring and analyzing data to envisage market trends and to improve organization performance is a vital business activity. Enormous software applications, both web and non-web based, are built as a part of solution to various problems. However, providing a solution is not enough. The organizations also need to keep a track of whether those applications are being used or not.

Based on the literature survey, there are various techniques and frameworks built to perform such real time analysis for web based solutions. But, these techniques cannot be used for performing analysis for non-web based applications in real / near real time. So, there is a need to design a framework which can help perform real / near real time analysis for non-web based software applications.

Sr. No.	Technique	Framework/Technology Stack Used	Limitations
1	Real-Time Big Data Analytics for the Enterprise	SAP HANA, Intel, Apache Hadoop	The applications should be connected via internet to enable real time analysis.
2	Soft Real-Time GPRS Traffic Analytics for Commercial M2M Communications Using Spark [1]	Spark, Spark SQL and Spark Streaming	- Better algorithms can be used for packet dissection, transaction reconstruction, and context reconstruction. - The configuration parameters of underlying cluster can be further improved. - The framework can be used only for applications running over wireless sensor networks.

3	Trace-Based Analysis and Prediction of Cloud Computing User Behavior Using the Fractal Modeling Technique [2]	Fractal Modeling technique	- Usage limited for cloud based applications.
4	Design and Implementation of a Real-Time Interactive Analytics System for Large Spatio-Temporal Data [6]	Optimized OceanRT framework	- Limiting to only Spatio-Temporal data set and queries such as mobile web browsing data and call logs.
5	Realtime Web Analytics [8]	Ruby, Ruby on Rails web framework, MongoDB	- The solution helps to track only the web based applications and has no technique to track non web based applications.
6	Exploring Big Data opportunities for Online Customer Segmentation [9]	Apache Hadoop, R programming Language, Apache Solr, Microstrategy	- Deals with Customer data captured over internet for Segmenting them and further performing analytics on that data.
7	Near Real-time Data Analysis of Core-Collapse Supernova Simulations With Bellerophon [10]	Bellerophon	- This application has been built to support and perform near real time analysis for only one application, CHIMERA. - Not enabled for Mobile phones.

Table 3.1 Limitations of Existing Solutions

3.2 Problem Statement

In today's world it is becoming clear that to be successful in business, it is essential that one not only provide solutions to the client but also analysis the real time usage of those applications, so that such analysis can to used further to optimize user experience and help in real time decision making.

Problem is to determine a framework which would help in real time analysis and tracking for non-web based applications. In current techniques and framework it is not possible for business to track the usage of non-web based in real / near-real time and further perform analysis on that data.

The purpose of thesis is to offer a framework which would help the organizations to perform real / near real time analytics for non-web based software applications and help optimize the user experience and perform audit trails for these applications.

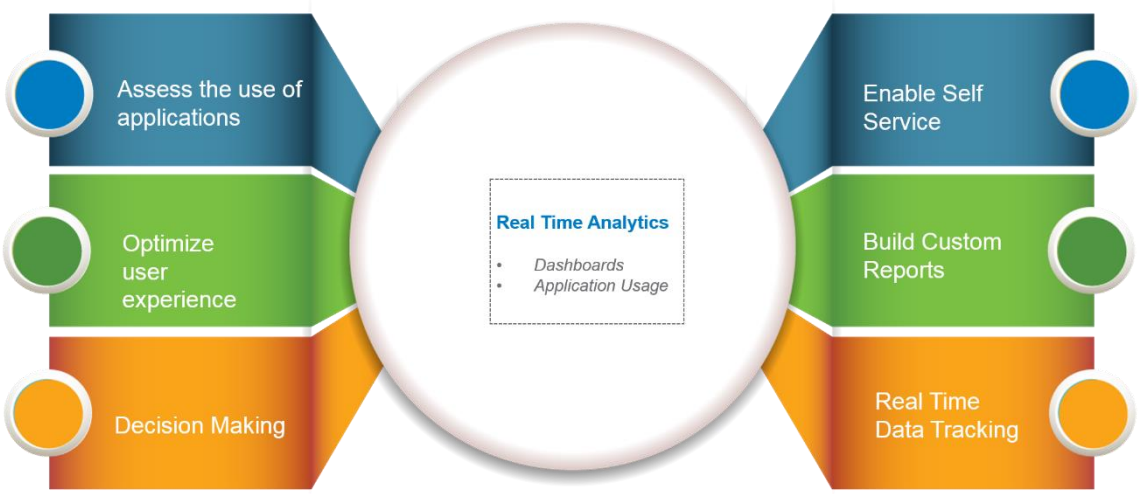


Figure 3.1 Expectation from Analytic Applications

3.3 Objectives

The purpose of thesis is as below:

i. To study the existing frameworks / techniques available for performing real time analytics for various software applications.

ii. To propose and design a framework which would help perform real time analytics for non-web based applications on Cloud.

iii. To implement, validate and compare the proposed framework on Google analytics and R.

3.4Conclusion

The gap analysis along with the problem statement has been presented in this chapter. Also the objectives of the work to be done are described. The next chapter gives the solution design for the problem statement stated above.

Chapter 4

Proposed Framework

In this chapter, the solution to the problem stated in the previous chapter is provided with the help of the layered diagram and data flow diagram.

4.1Design of Solution

The solution to the problem (real / near real time analytics for non-web based software applications) has been depicted through proposed framework and Data Flow Diagram. Following section presents the framework of the solution using architecture and data flow diagram.

4.1.1 Framework of Proposed Solution

In the proposed framework shown in figure 4.1 the user details along with the action performed are captured from the database trace utility through which the application is connected. As soon as the details are captured from the database trace, the user sensitive information is encrypted using encryption algorithm and custom dimension and variables are created. This information is pushed to the cloud where further processing takes place and the end users / application owner views the application access details like the number of users accessing the application, average hits on the application, etc. in real / near real time.

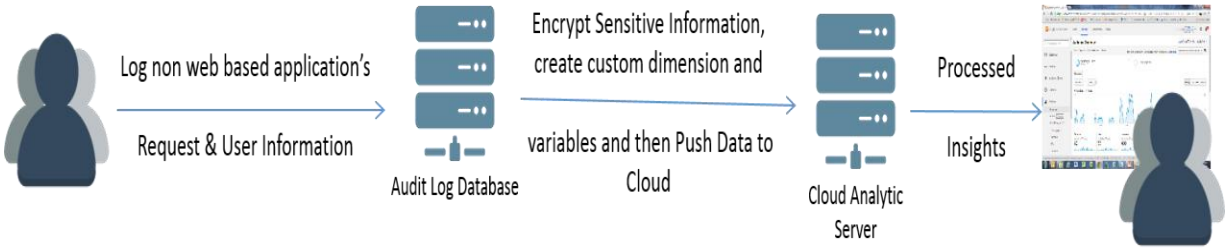


Figure 4.1 Proposed Framework

Figure 4.1 shows the main stages of the proposed framework that works in three stages:

Stage 1: Log the information of the user along with the request details via trace utility.

Stage 2: Capture the required information and apply encryption algorithm for user sensitive information and push that to cloud.

Stage 3: The information pushed to cloud database is used for further insights to view the application access details, etc. by building custom reports.

4.1.2 Detailed execution Stages:

The detailed explanation of how the proposed framework works is given below:

Stage 1: Log Details

Once the user access an application, it fires a request in the form of query to the database (e.g. SQL server, Oracle, OLAP cubes, etc.) to which that application is connected. The log details, i.e. user details along with the query fired on the database, can be captured from the trace utility of the database.

Stage 2: Capturing the required information and pushing it to cloud

This stage takes input from stage 1. Once log details are captured from the trace utility of the database, the required information, e.g. user details, application name, query execution time, etc. can be extracted using set of database components. The sensitive information like the user details, etc. are further encrypted using a mechanism so that the privacy of the user is maintained. Using this information the custom dimension and variables are derived which are pushed to cloud.

Stage 3: Processing Insight

In this stage the custom dimensions and variables which are pushed to cloud in stage 2 are used. On top of these custom variable a custom report is built which gives insight to the application owners and other authorized stakeholders about how many users are accessing the application, % number of new visits, etc.

4.1.3 Algorithm Details

In this section the algorithm details are provided which are proposed to encrypt the data and push the data to cloud.

Encryption Algorithm: The encryption algorithm takes the user sensitive data as input and gives the encrypted data as output. First the algorithm captures the data which is to be encrypted and converts the data into the ASCII format. This ASCII output is further

converted into hexadecimal number and summed up with the ASCII output to get the encrypted data.

Algo: Encryption Algorithm
1 **Input**: User sensitive information (UData) and **Output**: Encrypted data (EData) 2 Foreach User sensitive information Do 3 Set counter = 0 4 While counter <= Length(UData) Do 5 Set EData = EData + ASCII(Substring(UData,counter,1) 6 Set Counter=Counter + 1 7 End While 8 Set EData = EData + INTTOHEX(EData) 9 Return EData

Data Push Algorithm: In this algorithm, the required data is converted into custom variables and dimensions, which are further collected and then pushed on to the cloud.

Algo: Data Push Algorithm
1 Input: Log information (LData) 2 Foreach Ldata Do 3 Derive Custom Variables and Dimensions 4 Collect Custom Variables and Dimensions 5 Push to Cloud 9 End

4.1.4 Flowchart of proposed framework

Flowchart of the proposed framework for real time analytics for non-web based applications has been designed in figure 4.2.

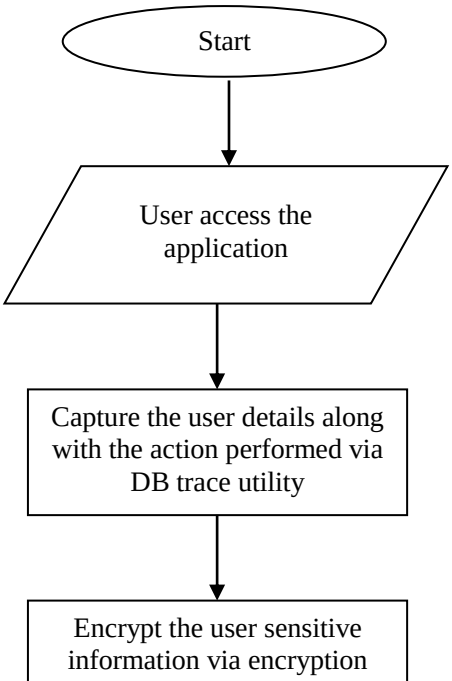


Figure 4.2 Flowchart of Proposed Framework

4.2 Conclusion

In this chapter the design of the proposed framework was presented. The detailed explanation of the proposed solution was given along its execution diagram and flow chart. In the next chapter the implementation of the proposed approach and results of the experiment would be gathered using a case study.

Chapter 5

Implementation and Experimental Results

This chapter of thesis work will focus on tools and technologies for setting up the framework, implementation of the proposed framework and implementation results using a particular case study.

5.1 Tools for setting up Environment

Cloud applications have different set of composition, configuration and deployment requirements. Various tools used to implement the proposed framework for performing real / near-real time analytics for non-web based applications are given below:

5.1.1 Google analytics

Google Analytics is an analytics tool being presented by Google, which provides a foresight into the websites on-going activities, offering users with information which allows them to make decisions on the performance of the website, conversion and design. Websites and web activity can be precisely calculated to offer detailed insight about the site. It helps provide answers to the users on questions as below:

- Is my site being used or not?
- How many users are accessing the site?
- How many users are returning back to visit the site?

Some of the key features being offered by Google Analytics include:

- **Map Overlay** – this helps one comprehend how to target movements by geography in the best possible way.
- **Internal Search** – this also allows one to follow how individuals utilize the search box on the site. This data can be sourced to put search alternatives on the website.
- Funnel Visualization – this feature can be used to optimize the checkout and click-paths.

However in the proposed solution these functionalities are further extended for non-web based applications as well, so that maximize benefit can be leveraged using these functionalities.

5.1.2 R

R is an integrated solution which provides services for data management, computation and graphical display. Various advanced features being supported by R are:

- an operational data management and storage provision
- a huge, comprehensive, integrated group of intermediate components for data analysis
- graphical services for data analysis and demonstration directly at the computer
- a group of operators for computation of arrays

5.1.3 SQL Server Integration Services

SQL Server Integration Services (SSIS) is the Microsoft component of SQL server database which is used to apply data integration operations to files and databases of the various business applications. It is just not limited to the simple ETL process. This facilitates application developers and DBA's to design, execute, and control complex, HPC ETL applications.

Below are some of the tasks that can be performed using SSIS:

- One can select data from various heterogeneous sources and apply any operation or functions which is required to apply business logic like standardization, merging, cleansing, etc.
- One can also automate various SQL server administrative functions, Cube building, etc. using certain set of procedures provided by SSIS.

5.2 Description of Case Study

Client uses certain set of reports built in Excel, a non-web based application, with source being OLAP cubes to meet certain requirements, i.e. to understand the shipment, distribution of products, etc. Now, the requirement of client is to perform the audit trails for these non-web based reporting applications and understand the usage of reports in real time so that user experience can be further optimized.

5.3 Implementation of Proposed Framework

The proposed framework has been implemented and tested on two analytical tools: Google Analytics and R

5.3.1 Implementation of framework using Google Analytics

To implement the requirement described in the above section the below framework is developed.

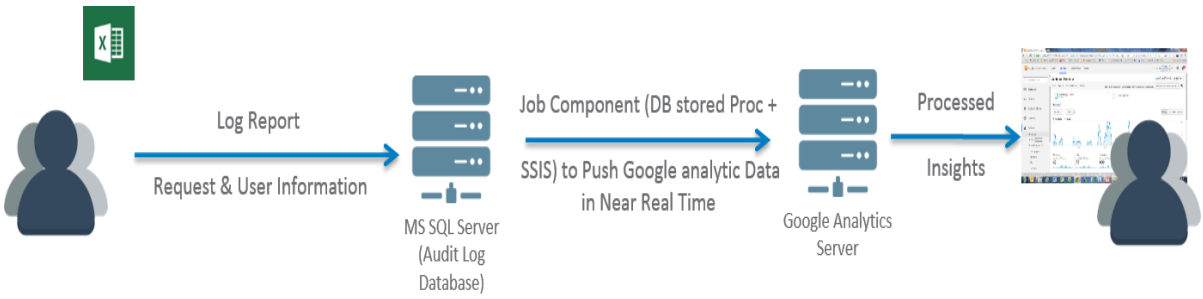


Figure 5.1 Framework to implement proposed technique

The detailed description of the above steps are described below:

Step 1: The user details along with the request information is logged in the MS SQL server Audit Log Database using the ASTrace utility.

Step 2: Job components (DB stored Procedures and SSIS package) are built to capture the latest user information and further derive the required custom variables to be pushed to Google analytics

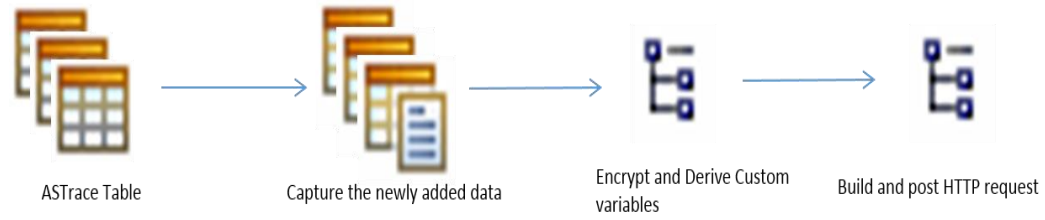


Figure 5.2 Job Component Overview

- **ASTrace Table**: This is a SQL server table which holds the information logged by the ASTrace service.

	RowNumber	EventClass	EventSubclass	TextData	ConnectionID	NTUserName	ApplicationName	StartTime	CurrentTime	Duration
1	0	65528	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	1	9	0	SELECT (AddCalculatedMembers({[Trade Channel Stru...	4	X	Shipment Details	2015-06-12 08:31:04.000	2015-06-12 08:31:04.000	NULL
3	2	10	0	SELECT (AddCalculatedMembers({[Trade Channel Stru...	4	Z	Distribution Details	2015-06-12 08:31:04.000	2015-06-12 08:31:04.000	20
4	3	9	0	WITH MEMBER [Measures].cChildren As 'AddCalculat...	4	X	Distribution Details	2015-06-12 08:31:07.000	2015-06-12 08:31:07.000	NULL
5	4	10	0	WITH MEMBER [Measures].cChildren As 'AddCalculat...	4	Y	Shipment Details	2015-06-12 08:31:07.000	2015-06-12 08:31:07.000	39
6	5	9	0	WITH MEMBER [Measures].cChildren As 'AddCalculat...	4	Y	Distribution Details	2015-06-12 08:31:09.000	2015-06-12 08:31:09.000	NULL
7	6	10	0	WITH MEMBER [Measures].cChildren As 'AddCalculat...	4	X	Shipment Details	2015-06-12 08:31:09.000	2015-06-12 08:31:09.000	18

Figure 5.3 ASTrace table

- View is built on top of ASTrace Table to capture the details about the recently accessed reports (i.e. Report Name, User Name, etc.)
- The captured details from the view are further encrypted using encryption algorithm for converting sensitive information to an encrypted string. The encryption algorithm is as below:
    - Set counter =0
    - While counter <= Length (UserName)
    - Set UserNameCode = UserNameCode+ASCII(SUBSTRING(UserName,counter,1)
    - Set Counter = Counter + 1
    - End
    - Set UserNameCode = UserNameCode + INTTOHEX(UserNameCode)
- Once the required custom variables are derived, a HTTP URL is created by placing the custom variables in the URL at the required location which is further posted to Google analytics using OLE automation procedures of Microsoft SQL server. Below are the set of OLE automation procedures used to post the derived URL to Google Analytics:
    - Execute sp_OACreate 'MSXML2.ServerXMLHttp', OLEobj OUT

- Execute sp_OAMethod OLEobj,'open',NULL, 'GET', strURL, false
- Execute sp_OAMethod OLEobj, 'setRequestHeader', NULL, 'Content-Type', 'application/x-www-form-urlencoded'
- Execute sp_OAMethod OLEobj,send,NULL,''
- Execute sp_OAGetProperty OLEobj,'status', objstatus OUT
- Execute sp_OADestroy OLEobj

Once the above OLE automation procedures are executed in sequence the information is pushed to Google Analytics.

With the data being pushed to google analytics further insights can be performed on the data in the form of number of users accessing the reports, number of active users, percentage of new users, etc.

The users across the globe accessed the reports and the below dashboards were generated in real time as and when the users were accessing the reports.

Figure 5.4 shows the number of active users at that point of time who are accessing the reports along with the location details from where the reports are being accessed. The geographical location of the active users, the application section which the user is accessing, etc. can be captured from this section.

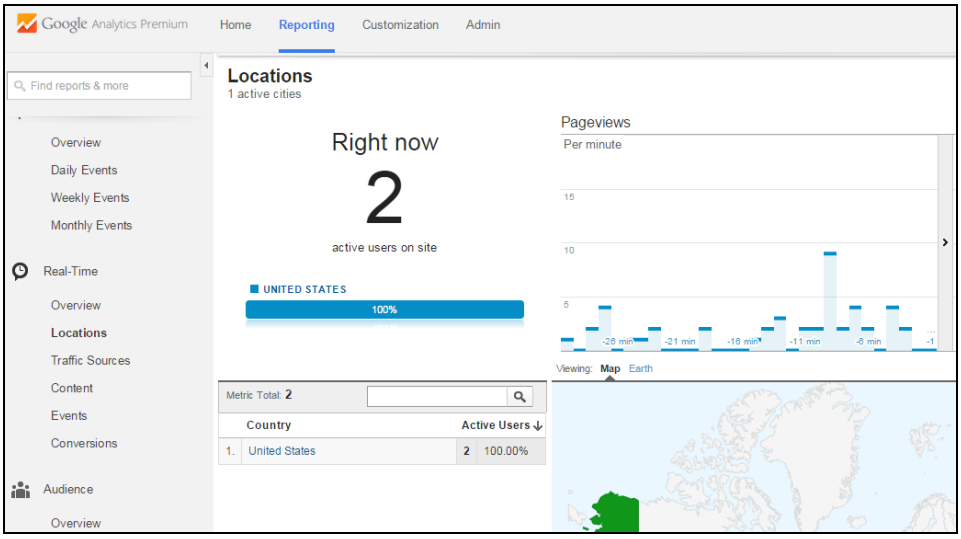


Figure 5.4 Number of Active Users

Figure 5.5 shows the comparison between the new vs returning users. From here one can identify how many new users are getting linked with the application and how many users returned back to use the application. It gives a clear picture about the visitors visiting the application.

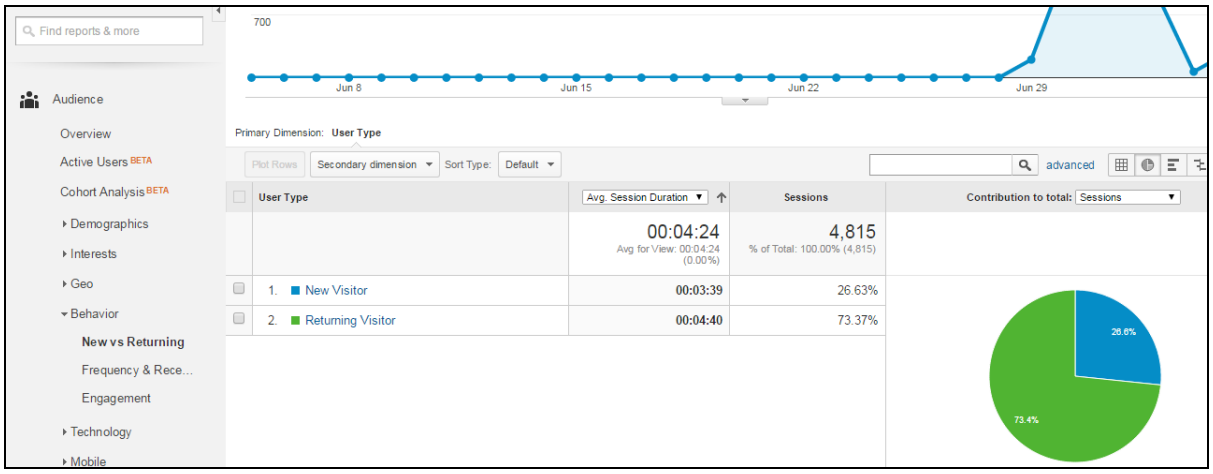


Figure 5.5 New vs returning Users

Figure 5.6 helps provide the user flow i.e. the click path of the group of users in order to access the user actions. This provides an interesting insight about how the users are finding the application, which path the users are taking and at which point the users are dropping off which provides an opportunity for further optimization of application.

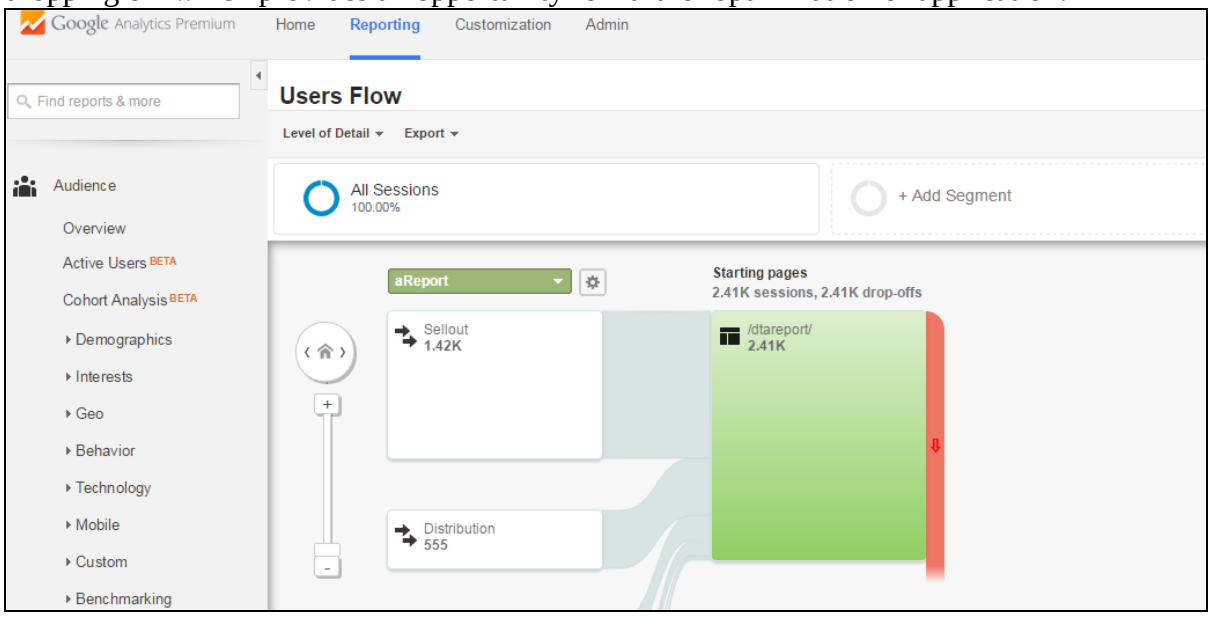


Figure 5.6 Users Flow

Figure 5.7 provides the insight about the average timing the report took to display the result to the users. From this view one can identify the performance of the application, i.e. how the application is performing, how long does the application takes to give the results to the user, etc.

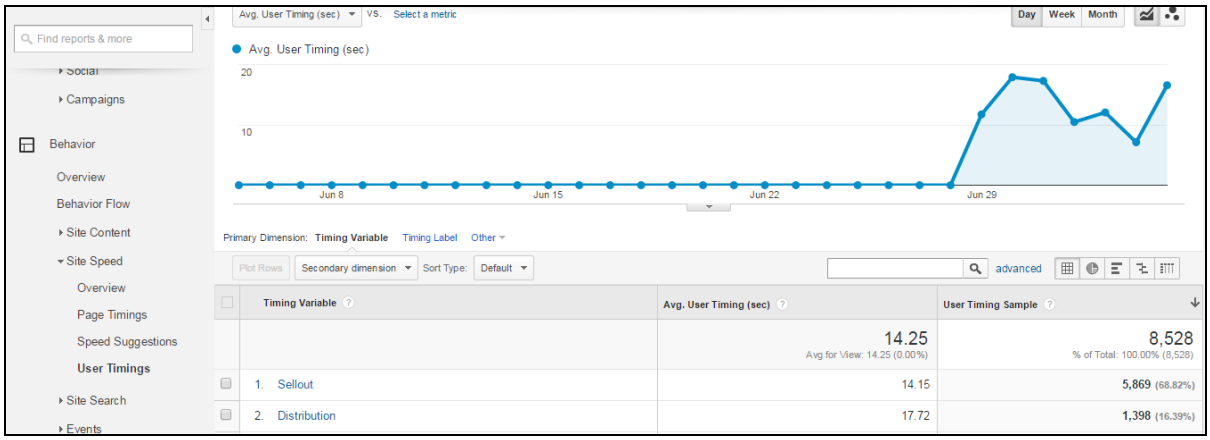


Figure 5.7 Report Speed – User Timings

Figure 5.8 displays the user count across time dimension at day level which can further be viewed at week, month and year level. In the selected time range one gets to know how many users are using the application. If any enhancement is made in the application one can track if that enhancement has helped improved the usage or not.

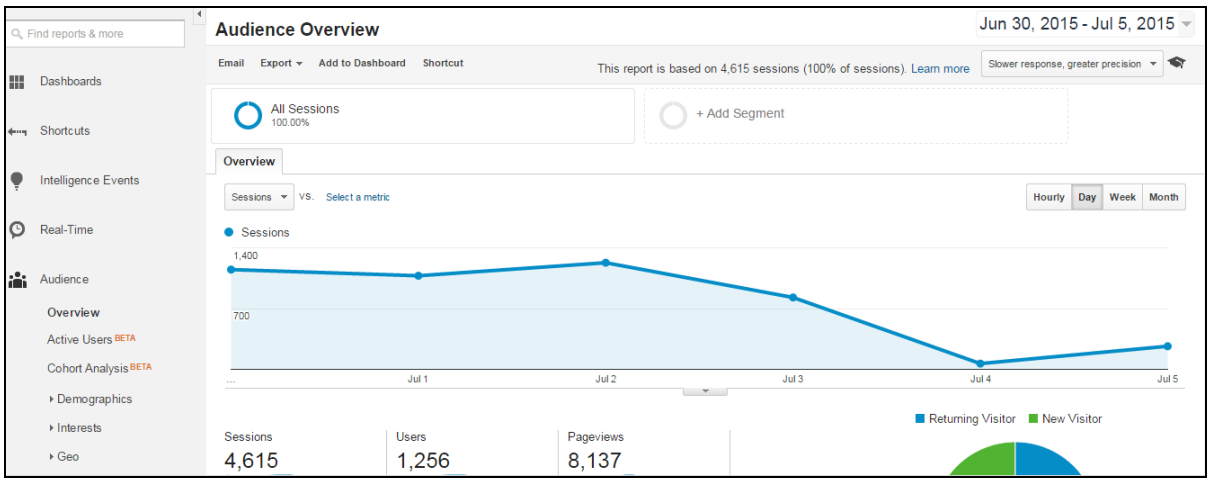


Figure 5.8 Comparing and contrasting Day wise report usage

Figure 5.9 display the custom report which can be built by the application owners to view the data accordingly and take appropriate actions. This is more of feature for those who want to build custom reports according to their need view the report variables across the required set of dimensions.

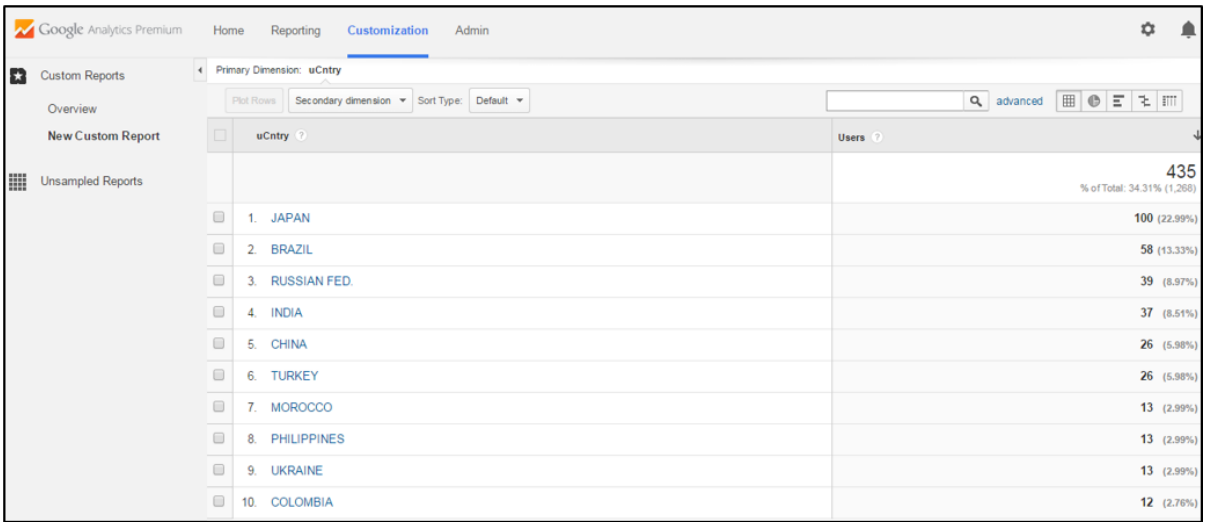


Figure 5.9 Country Wise view of report access

5.3.2 Implementation of framework using R

As an alternative, the above framework has also been implemented over another analytic tool known as R to get the insight about the non-web based applications. Below are the various steps that are used to perform analytics for non-web based applications:

Step 1: The requested information of the user accessing the report is logged in the MS SQL server Audit Log Database using the ASTrace utility.

Step 2: Using various Job components (DB stored Procedures and SSIS package) the latest user information along with the required custom variables are pushed to server for performing analytics.

Step 3: Analytics is performed in R on top of the data being pushed in step 2.

Figure 5.10 display the count of active users that are currently using the applications in R. This helps one know the current usage of application.

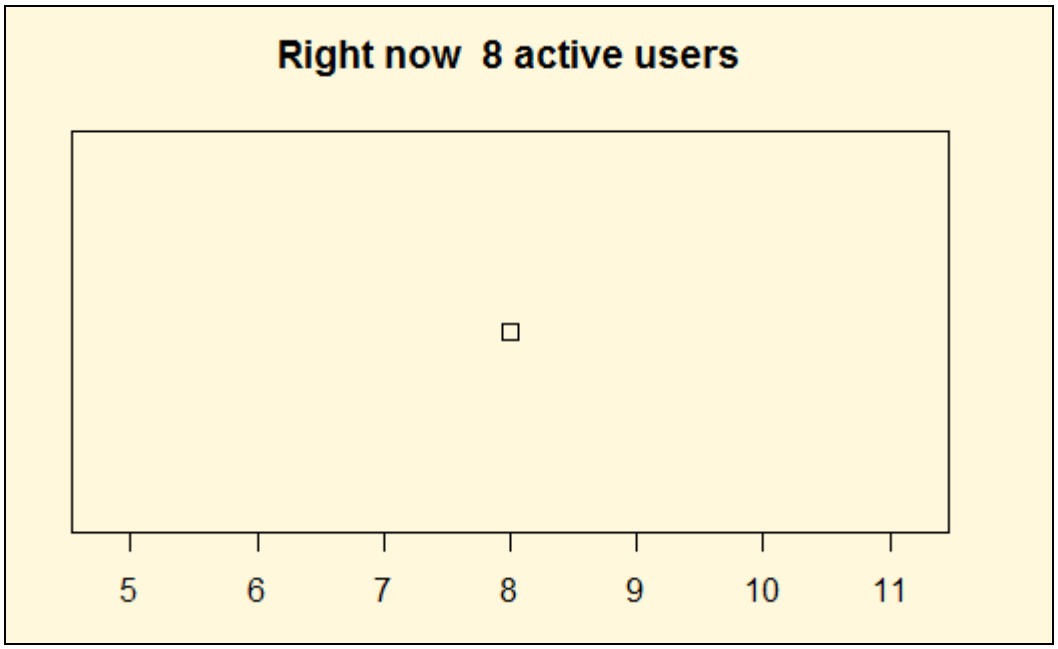


Figure 5.10 Active user details

Figure 5.11 display the count of new users who are currently accessing the application. This helps one know the details of those users who are accessing the application for the first time.

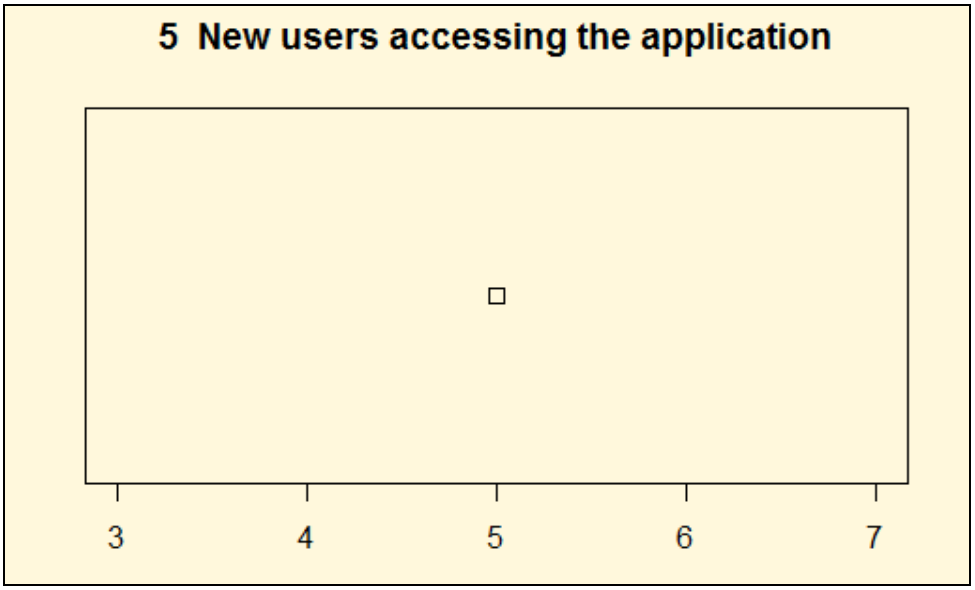


Figure 5.11 New users accessing the application

Figure 5.12 display the count of those users who are returning again to access the application at that moment of time.

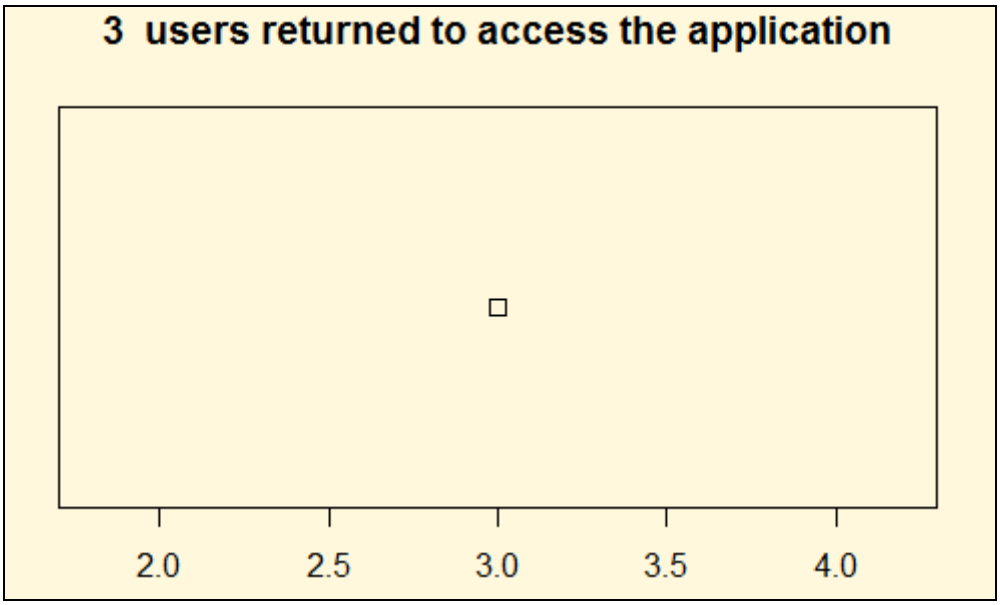


Figure 5.12 Returning users accessing the application

Figure 5.13 display the spilt of users on the basis of new vs the ones who are returning back to access the application.

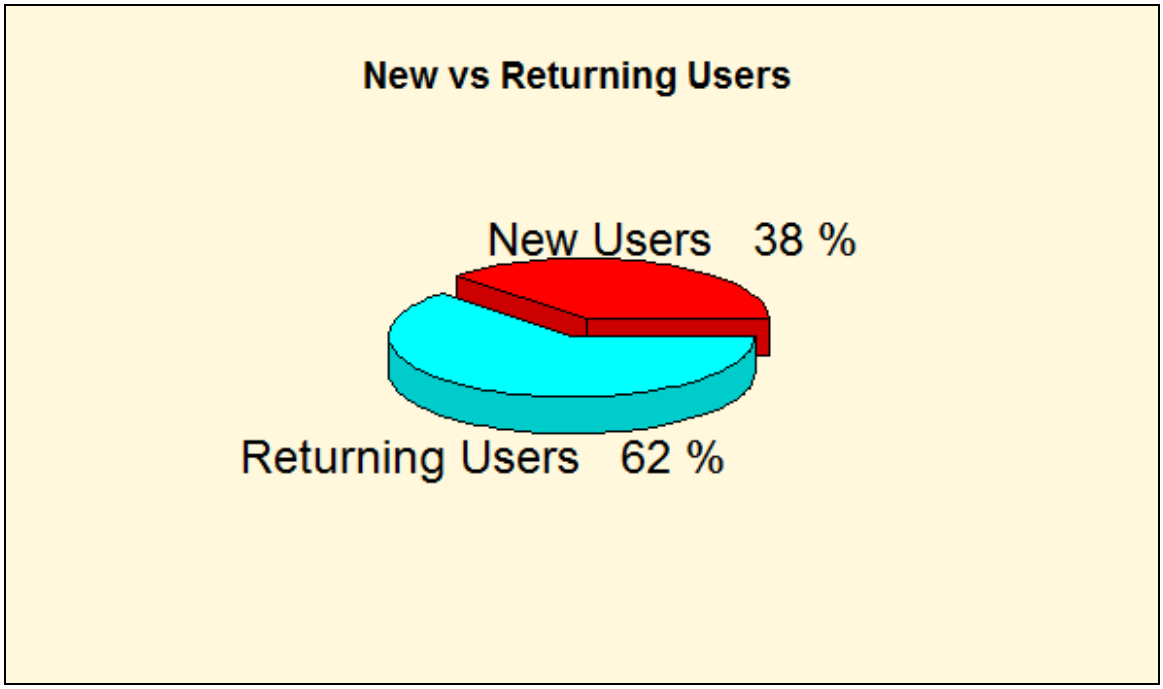


Figure 5.13 New vs Returning users

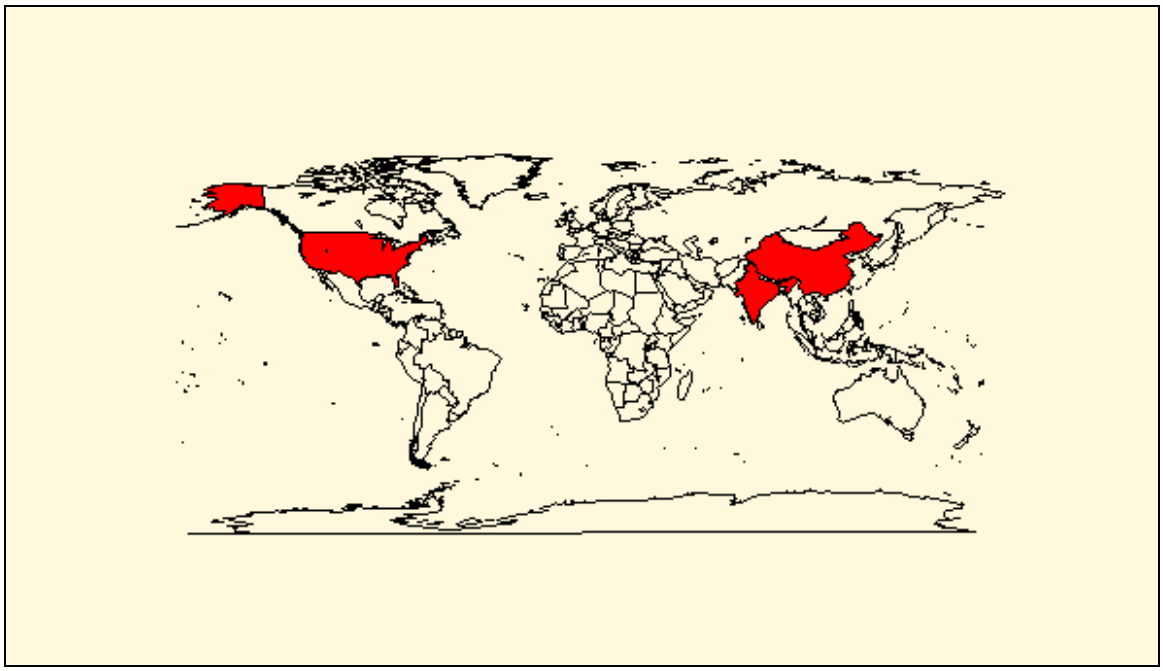


Figure 5.14 Application access details across globe

5.4 Experimental Results

The experiments performed are on the basis of 2 parameters, performance (upload time) and accuracy. The analytic features being used for the case study are further compared for both google analytics and R in this section.

5.4.1 Test for Performance (upload time)

The above setup has been tested for various reports across locations for a time span of one month and below are the average timings in which the data is uploaded to google analytics and R and the applications owners are able to view the user details:

Region	Average time taken in Google Analytics (in sec)	Average time taken in R (in sec)
India	1.34	1.7
Europe	1.23	2.1
China	1.4	1.8
Asean	1.35	1.5
Latin America	1.2	1.7

Table 5.1: Upload time statistics

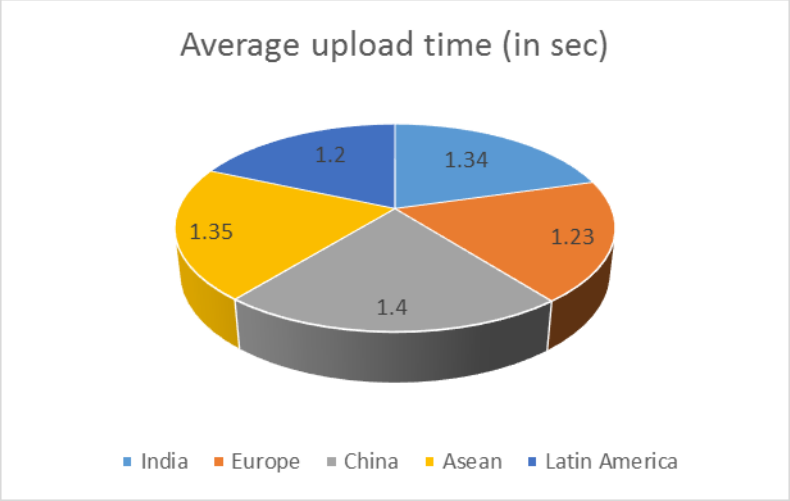


Figure 5.15 Average Upload Time per region in Google Analytics

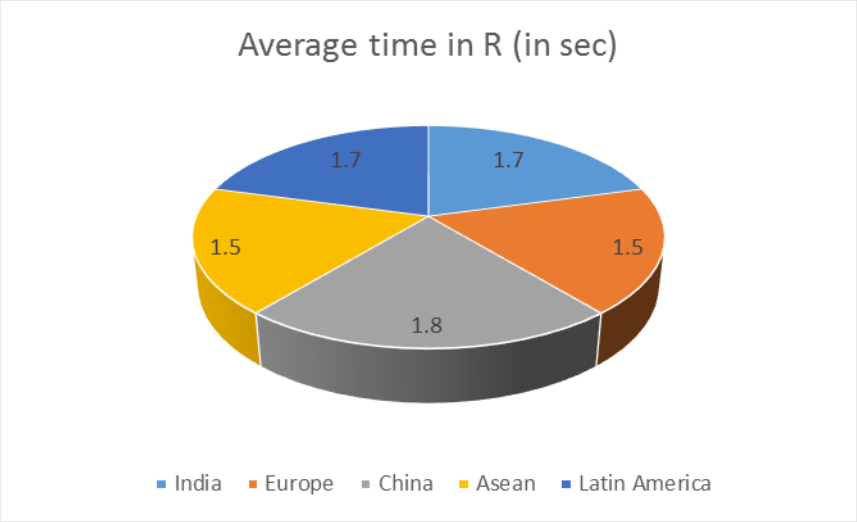


Figure 5.16 Average Upload Time per region in R

Hence, it is clear from the above statistics that the data upload time to both Google analytics and R for non-web based applications is in the range of 1-2 sec from the time

the report is accessed by the user. Also, R takes slightly longer time to display the data in real time than Google Analytics.

5.4.2 Test for Accuracy

Second round of test was performed the entire month to test if the data uploaded on Google analytics was correctly visualized or not and it found that if the custom variables are correctly placed in the specified location in the HTTP URL then the results are promising and the data was correctly displayed.

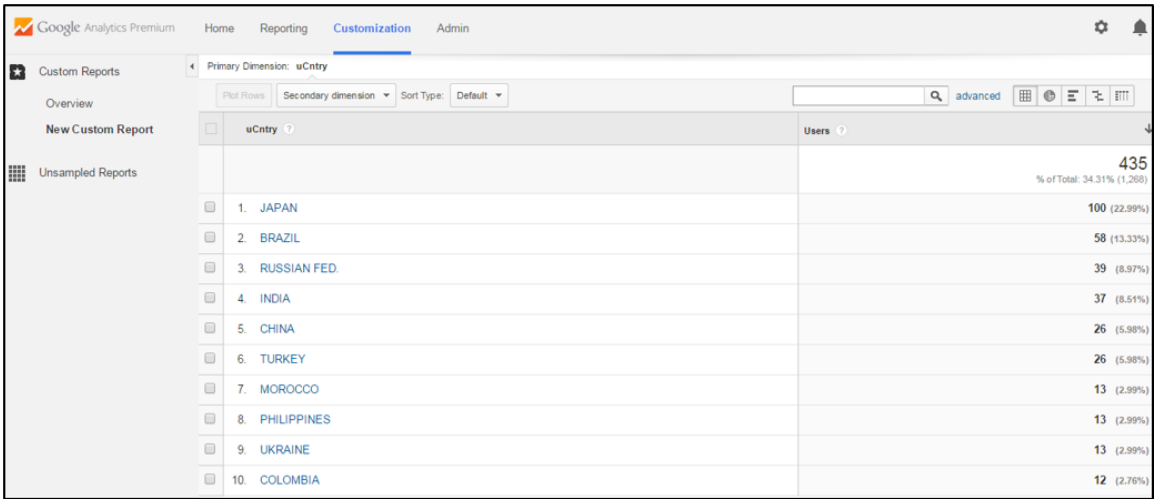


Figure 5.17 Data Accuracy

5.4.3 Comparative analysis between Google Analytics and R

In this section comparative analysis is done between the features supported by Google Analytics and the features implemented in R.

Sr. No.	Features	Implementation by Google Analytics	Implementation by R
1	Real Time User Access	Yes	Yes
2	Returning User	Yes	Yes
3	New User	Yes	Yes
4	User Flow	Yes	No
5	User Timings	Yes	Yes
6	Historical User Access	Yes	Yes
7	Run Time Custom Reports	Yes	No
8	Analytics via Graph	Yes	Yes

Table 5.2 Analysis between Google Analytics and R

5.5 Conclusion

This chapter gives the details of the implementation along with the experimental results. In the next chapter the conclusion and future directions of this work is presented with thesis contribution.

Chapter 6

Conclusion and Future Scope

This chapter discusses the conclusion of work done in thesis and ends with a vision of future direction which can be taken further.

6.1Conclusion

This thesis gives an introduction to cloud computing and software applications and a background about the various analytics techniques. In this work various existing real time analytics techniques have been discussed, followed by comparative analysis of these techniques. A framework has been built to perform real time analytics for non-web based applications. The framework has been implemented using a case study. It is developed using SQL server components and Google analytics. The experimental results have been gathered which shows us that the data is ready for analysis in a time frame of 1-2 sec. It has been tested from multiple locations across the globe.

6.2Thesis Contribution

Below is the thesis contribution:

- In this thesis various existing real time analytics techniques have been analyzed and compared.
- A framework has been designed which can be used to perform real / near-real time analytics for non-web based applications.
- The proposed framework is implemented using Google analytics and SQL server components and further the results are compared with R.
- Experimental results depict that one gets the insight about the non-web based applications in the time span of 1-2 sec.

6.3Future Scope

- In the current approach the capabilities of Google analytics are extended to capture the data of non-web based applications. Also, analytics features of R are used to provide insight to the users about the non-web based applications.
- However, in future further enhancements can be made to analytic algorithms to get better insight about the usage of non-web based applications which would help the organizations in better audit trailing, decision making and performing perspective analysis. Also, further algorithms can be developed to implement the missing features of R.

References

[1] Privitera, G.; Ghidini, G.; Emmons, S.P.; Levine, D.; Bellavista, P.; Smith, J.O., "Soft real-time GPRS traffic analytics for commercial M2M communications using spark," Smart Computing (SMARTCOMP), 2014 International Conference on , vol., no., pp.13,20, 3-5 Nov. 2014 doi: 10.1109/SMARTCOMP.2014.7043833

[2] Shuang Chen; Ghorbani, M.; Yanzhi Wang; Bogdan, P.; Pedram, M., "Trace-based analysis and prediction of cloud computing user behavior using the fractal modeling technique," Big Data (BigData Congress), 2014 IEEE International Congress on , vol., no., pp.733,739, June 27 2014-July 2 2014

[3] C. Feddersen, "Real-time event processing with Microsoft Azure Stream Analytics", Microsoft, vol. 1, 2015.

[4] Chircu, A.M.; Sultanow, E.; Chircu, F.C., "Cloud computing for big data entrepreneurship in the supply chain: using SAP HANA for pharmaceutical track-and-trace analytics," Services (SERVICES), 2014 IEEE World Congress on , vol., no., pp.450,451, June 27 2014-July 2 2014.

[5] Alex Guazzelli, Kostantinos Stathatos, Michael Zeller, "Efficient deployment of predictive analytics through open standards and cloud computing", ACM SIGKDD Explorations, June 2009

[6] Shiming Zhang, Yin Yang, Wei Fan, Marianne Winslett, "Design and Implementation of a Real-Time Interactive Analytics System for Large Spatio-Temporal Data", Proceedings of the VLDB Endowment, Volume 7 Issue 13, August 2014

[7] Qiming Chen, Meichun Hsu, Hans Zeller, "Experience in continuous analytics as a service (CaaaS)", ACM New York, NY, USA ©2011, Mar 2011

[8] João Cardoso, 'Realtime Web Analytics', HAAGA-HELLA University of Applied Science, 2011

[9] Georgia Fotaki, “Exploring Big Data Opportunities for Online Customer Segmentation”, IGI Publishing Hershey, PA, USA, Vol. 5, issue 3, July 2014.

[10] E. J. Lingerfelt, S. S. Desai, O. E. B. Messer, E. J. Lentz, and C. A. Holt, “Near real-time data analysis of core-collapse supernova simulations with bellerophon”, Procedia computer science, Vol. 29, pp. 1504,1514, 2014

[11] C. Feddersen, “Real-time event processing with Microsoft Azure Stream Analytics”, Microsoft, vol. 1, 2015.

[12] Gartner. Gartner Identifies the Top 10 Strategic Technology Trends for 2013. October, 2012. http://www.gartner.com/newsroom/id/2209615.

[13] Gartner. Gartner Identifies the Top 10 Strategic Technology Trends for 2014. October, 2013 http://www.gartner.com/newsroom/id/2603623.

[14] Doug Washburn, Lauren E. Nelson with James Staten, Christopher Mines, Eric Chi. Cloud Computing Helps Accelerate Green IT. June, 2011. http://www.forrester.com/Cloud+Computing+Helps+Accelerate+Green+IT/fulltext/-/E-RES58938.

[15] Peter Mell and Tim Grance, The NIST Definition of Cloud Computing, Version 15, 10-7-09. www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf.

[16] Mohiuddin Ahmed, Abu Sina Md. Raju Chowdhury, Mustaq Ahmed, Md. Mahmudul Hasan Rafee. An Advanced Survey on Cloud Computing and State-of-the-art Research Issues. International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012. View shared post

[17] Arian Stipic and Tomislav Bronzin. How cloud computing is (not) changing the way we do BI. May,2012

[18] Patrick Thibodeau. Amazon's data center outage reads like a thriller. December, 2009.
http://www.computerworld.com/s/article/9142154/Amazon_s_data_center_outage_reads_like_a_thriller.

[19] Gartner. Five Refining Attributes of Public and Private Cloud Computing. May, 2009.

[20] Eckerson, Wayne W. The Five Dimensions of Business Intelligence. 2005.

[21] Inmon, William H. Building the Data Warehouse. 2005.

[22] Eckerson, Wayne W. The Five Dimensions of Business Intelligence. 2005.

[23] Emerging Technologies Hype Cycle for 2013: Redefining the Relationship. http://public.brighttalk.com/resource/core/19507/august_21_hype_cycle_fenn_lehong_29685.pdf

[24] Potter, Ranoplh & Bezuidenhout, Brendon. Matching Business Intelligence with Cloud Computing. October, 2009. http://xqrx.com/writing/a_cloud.php.

[25] Golkar, Cyrus. Business and Technology Questions in Cloud Computing. July, 2009. http://www.b-eye-network.com/channels/1550/view/10905.

[26] Deshpande, Mukund & Joshi, Shreekanth. Incorporating Business Intelligence in the Cloud. August, 2009. http://www.b-eye-network.com/channels/1550/view/11143.

[27] Eckerson, Wayne. Implementing BI in the Cloud. June,2009. http://portals.tdwi.org/blogs/wayneeckerson/2009/06/implementing-bi-in-the-cloud.aspx.

[28] Recombinant Data Corp. Cloud Computing for Healthcare and Life Sciences Data Warehousing. 2009. http://www.b-eye-network.com/files/CloudComputing_WP.pdf.

[29] Dine, Stephen. B.I. in the Cloud. August, 2009.

[30] Wells, Dave. What's Up with Cloud Analytics. Decmber, 2009.

[31] Lounibos, Tom. SOASTA's 10,000 Hours in the Cloud. December, 2009. http://eclipsse.sys-con.com/node/1150203.

[32] http://public.brighttalk.com/resource/core/19507/august_21_hype_cycle_fenn_lehong_29685.pdf

[33] Donna Xu, Dongyao Wu, Xiwei Xu, Liming Zhu, Len Bass, "Making real time data analytics available as a service", ACM New York, NY, USA © 2015, May 2015

[34] Marcos D. Assuncao, Rodrigo N. Calherio, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya, "Big Data computing and clouds", Journal of Parallel and Distributed Computing, Vol. 79 Issue C, May 2015, Academic Press, Inc. Orlando, FL,USA

[35] Osman, A.; El-Refaey, M.; ElNaggar, A., "Towards real-time analytics in the cloud," Services (SERVICES), 2013 IEEE Ninth World Congress on , vol., no., pp.428,435, June 28 2013-July 3 2013

[36] Moore, P.; Xhafa, F.; Barolli, L.; Thomas, A., "Monitoring and detection of agitation in dementia: towards real-time and big-data solutions," P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference on , vol., no., pp.128,135, 28-30 Oct. 2013

[37] Kudva, V.D.; Nayak, P.; Rawat, A.; Anjana, G.R.; Sheetal Kumar, K.R.; Amrutur, B.; Mohan Kumar, M.S., "Towards a real-time campus-scale water balance monitoring system," VLSI Design (VLSID), 2015 28th International Conference on , vol., no., pp.87,92, 3-7 Jan. 2015

[38] Mohamed, N.; Al-Jaroodi, J., "Real-time big data analytics: applications and challenges," High Performance Computing & Simulation (HPCS), 2014 International Conference on , vol., no., pp.305,310, 21-25 July 2014

[39] Ediger, D.; Appling, S.; Briscoe, E.; McColl, R.; Poovey, J., "Real-time streaming intelligence: integrating graph and NLP analytics," High Performance Extreme Computing Conference (HPEC), 2014 IEEE , vol., no., pp.1,6, 9-11 Sept. 2014

[40] Cheng-Zhang Peng; Ze-Jun Jiang; Xiao-Bin Cai; Zhi-Ke Zhang, "Real-time analytics processing with MapReduce," Machine Learning and Cybernetics (ICMLC), 2012 International Conference on , vol.4, no., pp.1308,1311, 15-17 July 2012

[41] Kelly, C.; Craig, K.; Matthews, M., "Real-time predictive analytics to estimate air traffic flow rates," Integrated Communication, Navigation, and Surveillance Conference (ICNS), 2015 , vol., no., pp.N1-1,N1-12, 21-23 April 2015

[42] Osman, A.; El-Refaey, M.; ElNaggar, A., "Towards Real-Time Analytics in the Cloud," Services (SERVICES), 2013 IEEE Ninth World Congress on , vol., no., pp.428,435, June 28 2013-July 3 2013

[43] Berry, A.; Milosevic, Z., "Real-Time Analytics for Legacy Data Streams in Health: Monitoring Health Data Quality," Enterprise Distributed Object Computing Conference (EDOC), 2013 17th IEEE International , vol., no., pp.91,100, 9-13 Sept. 2013

[44] Han Hu; Yonggang Wen; Tat-Seng Chua; Xuelong Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," Access, IEEE , vol.2, no., pp.652,687, 2014

[45] Nirmala, M.B., "WAN Optimization Tools, Techniques and Research Issues for Cloud-Based Big Data Analytics," Computing and Communication Technologies (WCCCT), 2014 World Congress on , vol., no., pp.280,285, Feb. 27 2014-March 1 2014

[46] Chandarana, P.; Vijayalakshmi, M., "Big Data analytics frameworks," Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on , vol., no., pp.430,434, 4-5 April 2014

[47] Ben Ayed, A.; Ben Halima, M.; Alimi, A.M., "Big data analytics for logistics and transportation," Advanced Logistics and Transport (ICALT), 2015 4th International Conference on , vol., no., pp.311,316, 20-22 May 2015

[48] Yang Song; Alatorre, G.; Mandagere, N.; Singh, A., "Storage Mining: Where IT Management Meets Big Data Analytics," Big Data (BigData Congress), 2013 IEEE International Congress on , vol., no., pp.421,422, June 27 2013-July 2 2013

[49] Slavakis, K.; Giannakis, G.B.; Mateos, G., "Modeling and Optimization for Big Data Analytics: (Statistical) learning tools for our era of data deluge," Signal Processing Magazine, IEEE , vol.31, no.5, pp.18,31, Sept. 2014

[50] Hui Li; Xin Lu, "Challenges and Trends of Big Data Analytics," P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2014 Ninth International Conference on , vol., no., pp.566,567, 8-10 Nov. 2014

[51] Mukkamala, R.R.; Hussain, A.; Vatrapu, R., "Towards a Set Theoretical Approach to Big Data Analytics," Big Data (BigData Congress), 2014 IEEE International Confre on , vol., no., pp.629,636, June 27 2014-July 2 2014

List of Publications

1. Arvind Jindal and Dr. Inderveer Chana, “Real time Analytics for Non-web based applications over cloud”, International Journal of Cloud Computing and Services Science (IJ-ClOSER),IAES, Vol. 4 No.5 October 2015. [Communicated]