ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

Big Data - Solutions for RDBMS Problems - A Survey

S. Vikram Phaneendra[1], E. Madhusudhana Reddy[2]

Assistant Professor, Dept. of CSE, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh

Professor, Dept. of CSE, Madanapalle Institute of Technology & Science, Madanapalle-517325, Andhra Pradesh

Abstract: Now a day's increases shared data very fast due to social networking and mobile phone. In olden days the data is less and able to handle most popular RDBMS concepts, but recently it is difficult to handle this much of huge data through old RDBMS tools. To overcome this situation we told to prefer using of Big Data. In this paper, we will outline the origin and history of this new system to handle "Big Data". We look up to current popular big data systems, illustrated by Hadoop architecture and its current & future use-cases of this system, apache drill high level architecture, applications of Big Data and its challenges.

Keywords: Big Data, Hadoop, Apache Drill, RDBMS Solutions, Big Data Solutions.

I. INTRODUCTION

The data management firm has grown over the few decades, first and foremost Relational Data Base Management Systems (RDBMS) technology. Still today, the majority of backend systems are RDBMS for online digital data, financial system, medical, conveyance, insurance coverage, and telecommunication business. As of the sum of data collected, and analyzed in endeavors have increased many-folds in volume, variety and velocity of generation and consumption. Due to this the organizations have struggled with architectural limitations of traditional RDBMS architectures. Effect of this problem, the researchers are invented a new class of system, that has to be designed and implemented a new phenomenon of "Big Data" [3].

A. Sources of Big Data infrastructure

As of the popularity of Internet was one of the primary reason for rapid growth of communication and connectivity in the world, we saw emerge of the Big Data platforms in the Internet environment.

The online world around us is having becoming abnormal proliferation of data. In Facebook one million users in year 2004, it exceeds more than one billion users in year 2012, i.e. a thousand-fold increase in within 8 years span. Recently more than 60% of users access Facebook from their mobile phones. The data generated by social networks are proportional to number of contacts among users of the social network, rather than number of users. According to Metcalfe's Law [13], and its variants, N users of contacts are proportional to N*log N. thus, the growth of contacts, and conversations among users in a social network results data generation [3].

In 1998, Google was founded with the goal of organizing entire information in the world, became the dominant content discovery platform in the World Wide Web (WWW), managing man-power and semi-automated ways, such as web portals and directories. The Google faced challenges are, crawling the web, indexed, ranking of several billions of web pages difficult to solve with the existing data management systems economically. The amounts of data available on the web in Google's search index explode from 26 million pages to more than one trillion pages within a decade [14]. In addition to this data was "multi structured", i.e. having natural language text, images, audio, video, sensor information, animations, and structured data [3].

High Velocity: Due to the deluge of data from different data sources, enterprises are increasingly facing challenges. Because of increasing connectedness of people, applications, sensors, diversity and speed of data is very large. Analysis of this data with minimum delay is a challenging task [11].

What is Big Data

Big Data is which demands cost effective, innovative forms of information processing for enhanced insight and decision-making of high-volume, high-velocity, and high-variety [3].

Big Data is naturally huge amount of data able to process and which cannot be effectively, processed, captured and analyzed by traditional database and search tools in minimal amount of time. McKinsey estimates that, the "big" in Big Data would be anywhere between few dozens of terabytes to petabytes of the enterprises data. The

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

explosion of Big Data information is mainly due to the vast amount of data generated by social networking platforms, various mobile network devices, and sensors data and so on. Data analysts call it as "Digital Universe" [5].

Attributes of Big Data differs from other data in 5 dimensions such as Volume, Velocity, Variety, Value, and Complexity.

Value: The unstructured data is also having some valuable information, so mine such data from huge volumes of information is more considerable.

Complexity: Considering to social networks data, connections and associated data which describes more about relationship among the data.

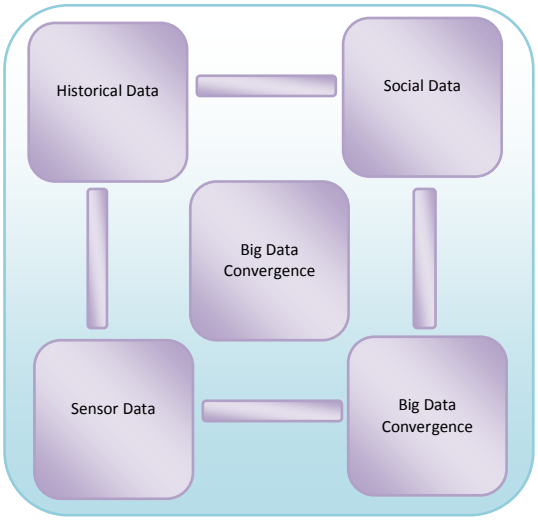

Figure 2: Convergence of data from different views

B. Versatile Technologies used in Big Data

From Big Data can derive main data by aggregating enormous amounts of data integrated from various sources. The following diagram shows versatile technologies used in Big Data.

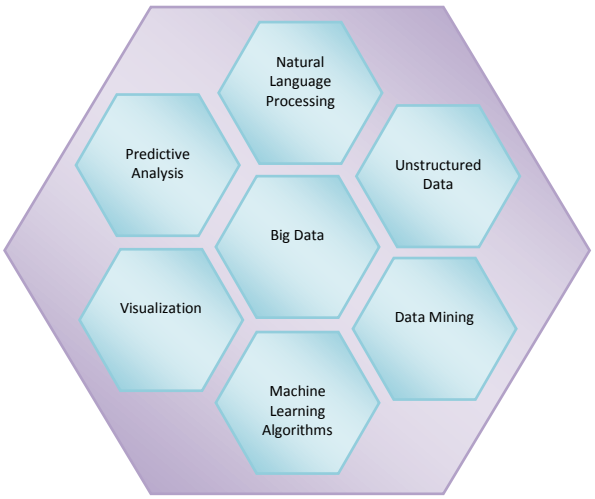

Figure 1: Various Technologies used in Big Data

Conventional sorting, searching and processing algorithms would not scale to handle the data in this rage, and that too most of it being unstructured. Most popular Big Data processing technologies include natural language processing algorithms, predictive modeling, and machine learning algorithms, unstructured programming and other artificial intelligence based techniques [5].
Bid Data is premeditated important for many enterprises, because any new product or new service will be eventually copied by competitors, but an organization can differentiate it by what it can do with the data it has. The following fig. shows the convergence of data from different views [5].

II. ARCHITECTURE

A. Hadoop

Hadoop is a group of applications projected to support parallel processing using the Google's MapReduce programming model [1] and is a quickly evolving ecosystem of components for implementing the Google Map Reduce Algorithms in a scalable model on commodity hardware. Users to process and store large volumes of data and analyzed it in ways not previously possible with less scalable solutions or standard SQL-based approached enabled by Hadoop.

I). Hadoop node types

Hadoop have different types of nodes within each Hadoop cluster; these include, NameNodes, DataNodes and EdgeNodes. Hadoop architecture is modular, allows individual components to be able to scale up and down as the needs of the industry requirement. The basic type nodes for a Hadoop cluster are as follows:

NameNode: The NameNode is the central location for data about the file system spreading in Hadoop environment. An environment can having one or two NameNodes, configuration allowed to provide minimal redundance between the NameNodes. The NameNode is link between clients of the Hadoop Distributed File System (HDFS) to locate intormation with the file system and provides updates for data that they have modified, added, moved and deleted.

DataNode: DataNodes build up the all most all of the servers contained in a Hadoop environment. General Hadoop environments would have more than one DataNodes and it may be increased up to thousands based on capacity and performance needed. The DataNode

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

servers have two functions: it having a portion of the data in the HDFS and it acts as a computing platform for running jobs running in local data within the HDFS.

EdgeNode: the EdgeNode is the access point for the users that need to utilize the Hadoop environment and external applications. The EdgeNode resides between the Hadoop cluster and the corporate network to provide access control, logging, policy enforcement and gateway services to the Hadoop environment. A distinctive Hadoop environment will have a minimum at least one EdgeNode and more are required on performance needs.

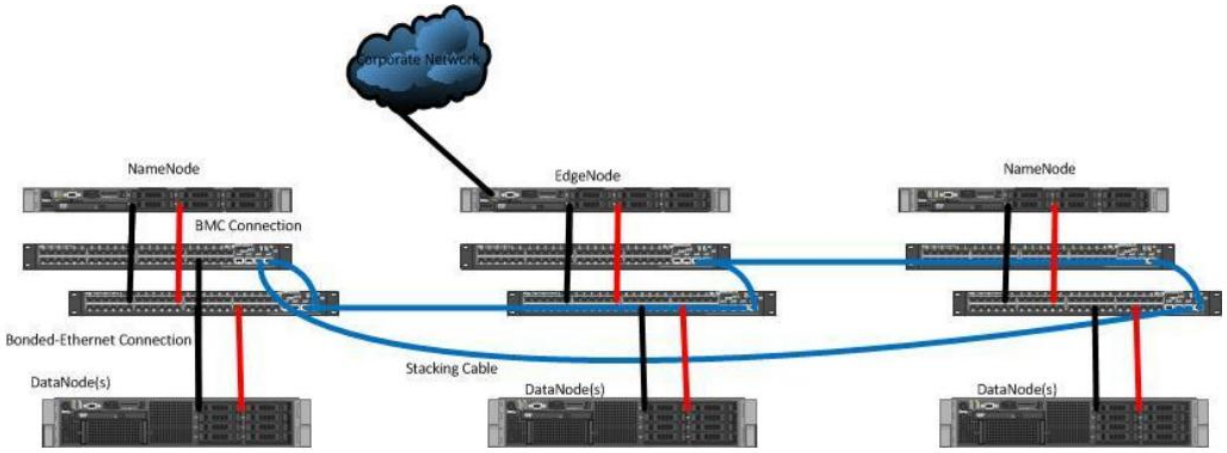

Figure 3: Node Types within Hadoop Clusters

II). Hadoop Uses

Actually the Hadoop was developed to be an open implementation of Google MapReduce and Google File System (GFS) [2]. As the ecosystem around Hadoop has fill-grown, a variety of tools have been developed to streamline data access, information access, security, specialized additions for industries, and data management. Inspite of this large ecosystem, there are several primary uses and workloads for Hadoop that can be outlined as:

Compute: A general use of Hadoop is as a distributed computing platform for processing of analyzing large amounts of information. The compute advantage is characterized by the need for huge number of CPUs and large amount of memory to store in process data. The Hadoop ecosystem provides the API required distributing and tracking workload as they are run on huge number of individual machines.

Storage: one basic component of the Hadoop ecosystem is HDFS. The HFS allowing users to have a distinct addressable namespace, stretch across many thousands of servers, creating in a single large file system. HDFS maintains replication of the data to avoid hardware failures do not lead to data loss. Most of the users it will use this scalable file system as a place to store large amounts of information that is accessed within processes run in Hadoop and or by external systems.

Database: Hadoop components that allows to be presented in SQL-Like interfaces. It allows standard SQL tools like SELECT, INSERT, DELETE, and UPDATE data within the Hadoop environment that too minimal code changes to existing applications. Many of the users will commonly use this approach for presenting data in SQL format for easy integration with existing systems and streamlined access by users.

III). What is Hadoop good for?

Hadoop is a collection of application projected to support parallel processing using the MapReduce programming model. When a actual MapReduce algorithm is released, Hadoop was subsequently developed around them, these tools are designed for specific uses. The actual use is for managing large data sets that needed to be easily searched. It having several other specific uses has emerged for Hadoop as an influential solution.

Large Data Sets: MapReduce coupled with HDFS is a victorious solution for storing huge volume of unstructured data.

Scalable Algorithm: For distributed processing capability of Hadoop, Any algorithm can scale too many cores of minimal inter-process communications.

Log Management: Hadoop is regularly used for analyzing and storing large sets of logs from different locations. In distributed environment log information is very necessary for tracing conversations among the nodes. It creates a very good area for manipulating, managing and analyzing different logs from diversity of sources within an organization, because the nature of distributed networks and it's scalability of Hadoop.

Extract-Transform-load (ETL) Platform: now a day almost all companies using many of the data warehouses and different relational database management system (RDBMS) platforms for their requirements within a

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

organization. Maintaining data up to data and synchronized these all different data can be a difficult process. Hadoop is a single solution for synchronize these all types of data and able to process and provide required information.

IV). Hadoop Adoption and use cases

From some years, Hadoop and other Big Data technologies have become most popular in non-internet based organization, as well as also struggled to handle the data deluge. Many of the organizations of infrastructure in various industries, such as finance, healthcare, insurance, retail, manufacturing, banking, advertizing and others have been almost fully digitized. Up to now, these organizations, data was stored in archival systems for recurrently retrieval purposes. The data is growing it is difficult to retrieve required information. However there was a growing realization among these organizations that this information can be utilized for gaining competitive advantage, improving customer experiences, and increasing process efficiencies [3].

The 3V's Volume, Velocity, and Variety of data, along with need to develop agile information driven applications, effects that the humans detecting patterns, analyzing and make indentify rich toolset data at hand. Conventional data exploration, visualization, e-commerce intelligence and reporting tolls are being adapted to co-exist with this new Big Data technologies; then implement advances in machine learning algorithms and methods, as well as abundant speed processing power and have able to deep and predictive analysis to used in information technology (IT) enterprises [3].

Industrial internet: the next leading edge

Most of the use cases of Big Data are analyzing customer behavior, their buying models, their likes and dislikes as posted in social networking media and their conversations, their visiting websites, geographic location from their mobile devices, the system generated data could be the next frontier for Big Data Systems. In addition to low cost sensors technology and minimum-range wireless connectivity has created possibility of real-time monitoring and chronological patterns of traditionally analog data sources [3].

The huge amount of data captured by the sensors; and prospect of storing and analyzing the data to make intelligent design, Operational decisions have created a new opportunity, now knowing by a new name as industrial internet [4].

B. Apache drill high level architecture

Earlier 2010, Google was published has seminal Dremel2 paper, it inaugurates two main innovations: generically handle nested data with column-striped representation and multilevel query execution trees, and those allowing for parallel processing of data broaden over thousands of computing nodes. These innovations are taken by the Apache Software Foundation in middle of the year 2012 and forming the core of a new incubator, Apache drill. At high level, 3 layers are in Apache Drill's Architecture.

User: It provides REST, Command Line Interface (CLI), JDBC/ODBC, etc., for human or application driven Interface.

Processing: Processing planar queries, execution, storage engines and allowable for pluggable query languages.

Data Sources: pluggable data origins either in a cluster setup or local, providing in place data processing.

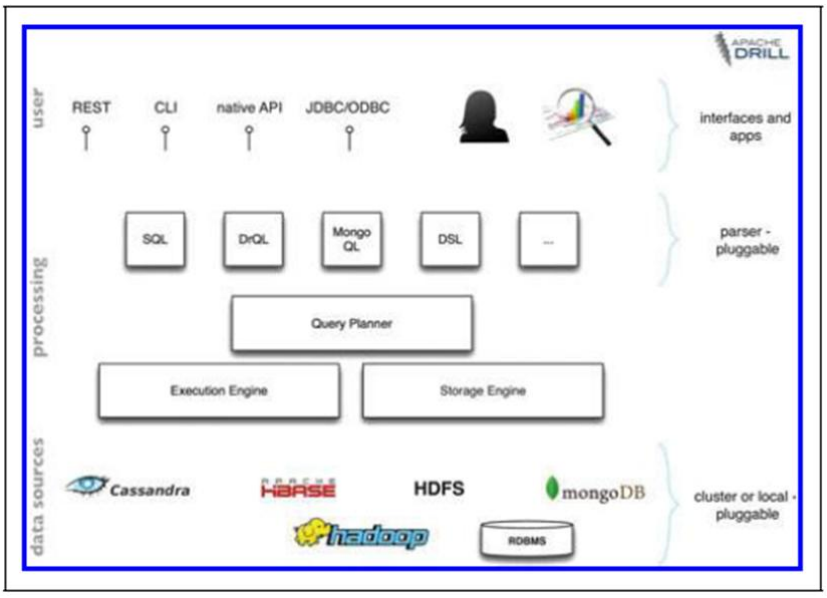

Figure 4: Apache Drill's Architecture

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

Main Features:

Extensible Rules: This Apache Drill Architecture is designed for capability of extensibility with interfaces and well-defined Application Programming Interface (API). It includes, starting from user layer, pluggable Query Language through the query API and also it supports User Defined Functions (UDF).

Complete Structured Query Language (SQL): Many settings are required in integrating Business Intelligence (BI) tools, Excel, SAP Crystal Reports, etc.

Nested data as a 1st class citizen: Apache Drill is flexible concerning to support different types data. As nested data is becoming widespread like JSON /BSON in document stores, XML, etc., and pulling down of nested data is error-prone. This architecture is supports nested data directly, effectively establishing an extension to most popular nested data formats.

Use but don't abuse schema: Different data sources increasingly, and this architecture do not have inflexible schemas; the schema may change quickly or differ in a per-record level. Apache Drill supports queries against unknown schemas and also the user able to define a schemas their requirement of let the Apache Drill discover it.

III. APPLICATIONS OF BIG DATA

We will analyze the impact and applications of Big Data of related technologies all over in various industries and technology domains.

A. Applications in industry domains [5]

Financial Industry:

✓ Better financial data management
✓ Investment banking using aggregated information from various sources likes financial forecasting, asset pricing and portfolio management.
✓ More accurate pricing adjustments based on vast amount of real-time data
✓ Stock advises based on huge amount of stock data analysis, unstructured data like social media content etc.
✓ Credit worthiness analysis by analyzing huge amount of customer transaction data from various sources
✓ Pro-active fraudulent transaction analysis
✓ Regulation conformance
✓ Risk analytics
✓ Trading analytics

Retail industry:

✓ Better analysis of supply chain data and touch points across Omnichannel operations
✓ Customer segmentation based on previous transactions and profi le information
✓ Analysis of purchase patterns and tailor made product off erings
✓ Unstructured data analysis from social media, multi-media to understand the tastes, preferences, and customer patterns and do sentiment analysis
✓ Targeted marketing based on user segmentation
✓ Competitor analysis

Mobility:

✓ Mining of customer location data, call patterns.
✓ Integrate with social media to provide location based services like sale off ers, friend alerts, points-of interest suggestions etc.
✓ Geo-location analysis:

Health care:

✓ Effective drug prescription by analyzing all structured and unstructured medical history and records of the patient
✓ Avoid un-necessary prescriptions

Insurance:

✓ Risk analysis of customer
✓ Analyzing cross-sell and up-sell opportunities based on customer spending patterns
✓ Insurance portfolio optimization and pricing optimization

B. Application across technology domains [5]

✓ Earthquake Analysis
✓ Search Engine improvements: New algorithms to analyze large unstructured data will be used. The algorithms will be artificial intelligence based working in parallel in multiple grids to process huge amount of data
✓ Business intelligence tools: Analytics tools will be able to provide new and creative visualizations to intuitively depict the meaning of the data
✓ Storage management tools: Private/cloud storage systems will undergo change to store huge amount of data
✓ Cloud computing: Cloud and social media play a vital role in handling Big Data. Cloud would be the platform of choice to store massive amount of data and to run the software as service to process the data
✓ ERP systems like CRM undergo great improvements. CRM system can help the on-call analysts to provide real-time customer offers, customer churn probability etc.
✓ Predictive analytics will be more effective by analyzing data from multiple dimensions

C. An interesting Case of Big Data is US 2012 Elections

Usually Big Data had a bid impact and re-define the way elections in US in the year 2012. This elections process elections team is used effectively Big Data for achieve victory. The democratic team was done data analysis and aggregated data from different data sources like voter list, social networking posts, fund raisers, etc., different tests were conducted to identify the voter's decision. Analyzing the Big Data was the main differentiator in alternation a good percentage of voters participating in elections.

IV. CHALLENGES

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

One of the computing areas that are attracting a lot of attention is 'Big Data'. What exactly is Big Data? As per Wikipedia, 'Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing apps [11].

Large enterprises across industries, from retail to financial services to manufacturing, are today actively exploring this new and exciting arena. At the same time, the veracity of inputs received from social media remains a matter of concern. There are also challenges in measuring the return on investment (i.e., ROI) from socio-business intelligence exercises: Statistically sound techniques for measuring ROI, even for simple matter such as advertising campaigns, are as yet not widely popular. Both these questions, i.e., efficiently establishing the veracity of social media inputs, as well as properly measuring ROI from socio-business intelligence, pose challenges for future research [6].

Challenges Storing and Maintaining the Big Data is a challenging task. The following challenges need to be faced by the enterprises or media when handling Big Data [7]:

- ✓ Data Privacy
- ✓ Capture
- ✓ Duration
- ✓ Storages
- ✓ Search
- ✓ Sharing
- ✓ Analysis
- ✓ Visualizations

REFERENCES

[1] The Google File System, http://research. google.com/archive/gfs.html, October 2003.
[2] MapReduce: Simplifi ed Data Processing on Large Clusters, http://research.google. com/archive/ mapreduce.html, December 2004
[3] Dr. Milind Bhandarkar, "*Big Data Systems: Past, Present & (possibly) Future, Cover Story*, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.
[4] Industrial Internet: Pushing the Boundaries of Minds and Machines, http://www.ge.com/docs/chapters/ Industrial_Internet.pdf, November 2012
[5] Shailesh Kumar Shivakumar,"*Big Data – A Big game changer*", Cover Story, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.
[6] Gautam Shroff,* Lipika Dey,** & Puneet Agarwal***, "*Socio-Business Intelligence Using Big Data*", Technical Trends, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.
[7] A Kavitha*, S Suseela**, and G Kapilya***, *Big Data*, Article, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.
[8] Michael Hausenblas and Jacques Nadeau, "*APACHE DRILL: Interactive Ad-Hoc Analysis at Scale*", DOI: 10.1089/big.2013.0011 _ MARY ANN LIEBERT, INC. _ VOL. 1 NO. 2 _ JUNE 2013 BIG DATA.
[9] Jyotiranjan Hota, *Adoption of In-Memory Analytics*, Article, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.
[10] Avinash Kadam, *Five Key Knowledge Areas for Risk Managers*, Article, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.
[11] Bipin Patwardhan* and Sanghamitra Mitra**, *Deriving Operational Insights from High Velocity Data*, CIO Perspective, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.
[12] Joey Jablonski, "*Introduction to Hadoop*, A Dell Technical White Paper", http://i.dell.com/sites/content/business/solutions/ whitepapers/en/Documents/hadoop-introduction.pdf.
[13] Metcalfe's Law Recurses Down the Long Tail of Social Networks, http://vcmike. wordpress.com/ 2006/08/18/metcalfesocial-networks/, April 2006
[14] We knew the web was big, http://googleblog.blogspot.com/2008/07/weknew-web-was-big.html, July 2008
[15] Pramod Taneja* and Prashant Wate**, "*Big Data Enabled Digital Oil Field*", Research Front, CSI Communication, ISSN 0970-647X | Volume No. 37 | Issue No. 1 | April 2013.

BIOGRAPHY

S. Vikram Phaneendra received B.Tech (Computer Science and Engineering) from JNTU, Hyderabad, M.Tech (Computer Science) from JNTUA, Anantapur. Currently he is working as Assistant Professor, Computer Science and Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. His research lie in the areas of Big Data, Cryptography & Network Security.

Dr. E. Madhusudhana Reddy received M.C.A from S.V. University, Tirupathi, M.Tech (IT) from Punjabi University, Patiala, M.Phil (Computer Science) from Madurai Kamaraj University, Madhurai, and Ph.D from S.V. University, Tirupathi. Currently he is working as Professor, Computer Science and Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. His research interest lie in the areas of Data Mining and Warehousing, Cryptography & Network Security and E-Commerce. He has published 39 research papers in National/ International Journals and Conferences. He is a Research Supervisor in JNTUH, JNTUA, JNTUAK, Andhra Pradesh, India. Also he is life member of Indian Society of Technical Education of India (ISTE), life member of Cryptography Research Society of India (CRSI) and fellow in ISET, India.