PhUSE US Connect 2018

Paper AB14

Data Science Applications and Scenarios

Giri Balasubramanian, PRA Health Sciences, Chennai, India
Edwin Ponraj Thangarajan, PRA Health Sciences, Chennai, India

ABSTRACT

Big Data developments have been centered mainly on the volume dimension of data, with frameworks such as Hadoop and Spark, capable of processing very large data sets in parallel. This paper focuses on the less researched dimensions of velocity and variety, which are characteristics of fast data applications. This paper proposes a general-purpose distributed platform to host and interconnect fast data applications, namely, those involving interacting resources in a heterogeneous environment such as the Internet of Things.

INTRODUCTION

Data is increasingly becoming cheap and important. We are now digitizing analog content that was created over centuries and collecting myriad new types of data from web logs, mobile devices, sensors, instruments, and transactions. A study estimates that 90 percent of the data in the world today has been created in the past two years and is increasing day by day in manifolds

At the same time, modern technologies are emerging to organize and make sense of this avalanche of data. We can now identify patterns and regularities in data of all sorts that allow us to advance scholarship, improve the human condition, and create commercial and social value. The rise of “big data” has the potential to deepen our understanding of phenomena ranging from physical and biological systems to human social and economic behavior.

Virtually every sector of the economy now has access to more data than would have been imaginable even a decade ago. Businesses today are accumulating new data at a rate that exceeds their capacity to extract value from it. The question facing every organization that wants to attract a community is how to use data effectively–not just their own data, but all the data that’s available and relevant.

Our ability to derive social and economic value from the newly available data is limited by the lack of expertise. Working with this data requires distinctive new skills and tools. The corpuses are often too voluminous to fit on a single computer, to manipulate with traditional databases or statistical tools, or to represent using standard graphics software. The data is also more heterogeneous than the highly curated data of the past. Digitized text, audio, and visual content, like sensor and weblog data, is typically messy, incomplete, and unstructured; it is often of uncertain provenance and quality; and frequently must be combined with other data to be useful. Working with user-generated data sets also raises challenging issues of privacy, security, and ethics.

By definition, Big Data, is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. In other words, big data is characterized by volume, variety (structured and unstructured data) velocity (high rate of changing) and veracity (uncertainty and incompleteness) and Value. By 2017, globally big data industry is expected to be USD 25 billion industry.

Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds, the speed at which credit card transactions are checked for fraudulent activities, or the milliseconds it takes trading systems to analyze social media networks to pick up signals that trigger decisions to buy or sell shares. Big data technology allows us now to analyze the data while it is being generated, without ever putting it into databases

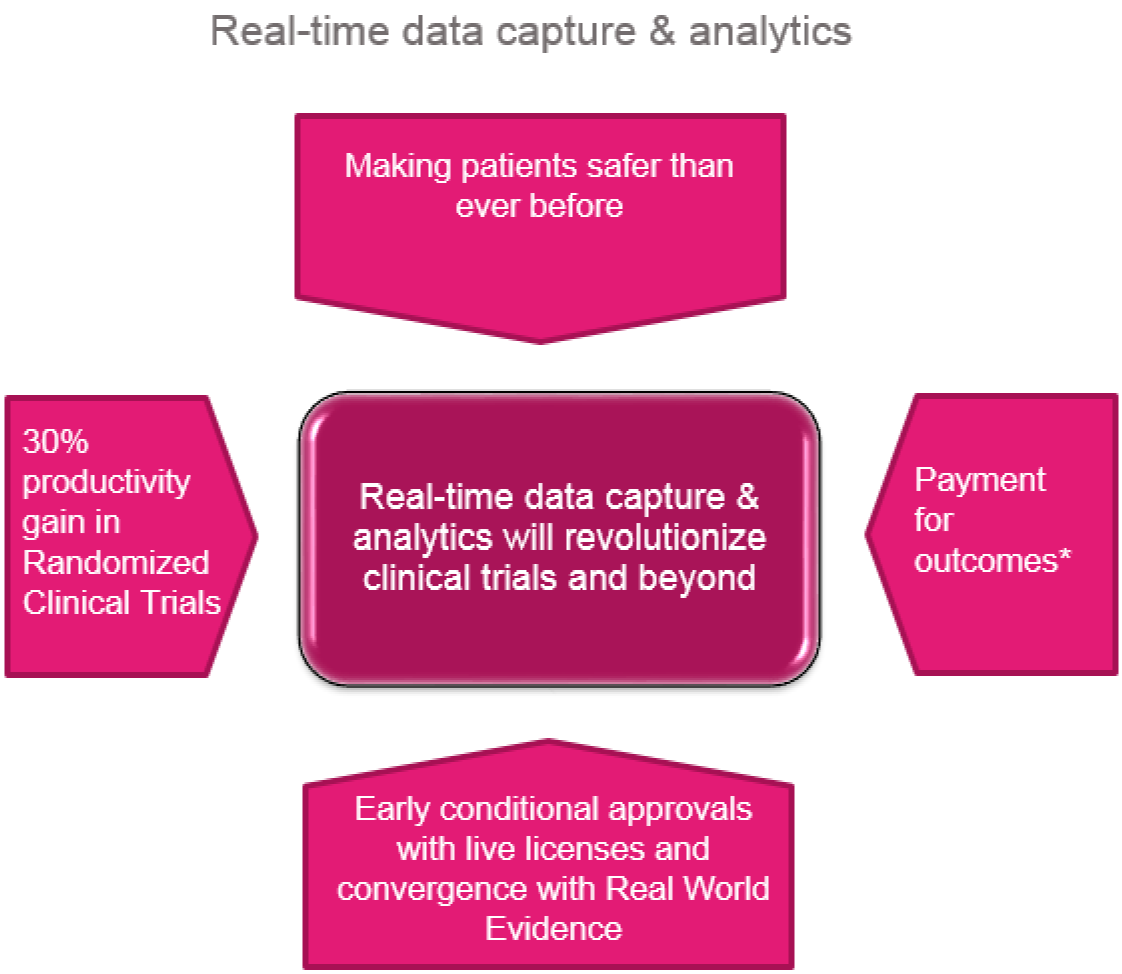
Variety refers to the several types of data we can now use. In the past we focused on structured data that neatly fits into tables or relational databases, such as financial data (e.g. sales by product or region). In fact, 80% of the world’s data is now unstructured, and therefore can’t easily be put into tables (think of photos, video sequences or social media updates). With big data technology we can now harness differed types of data (structured and unstructured) including messages, social media conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data. An important challenge, besides the large volumes and high-speed production rates (e.g., velocity), is raised by the enormous heterogeneity (e.g., variety) of such data.

Healthcare systems are also transformed by big data paradigm as data is generated from various sources such as electronic medical records systems, mobilized health records, personal health records, mobile healthcare monitors, genetic sequencing, and predictive analytics as well as a large array of biomedical sensors and smart devices.

To implement faster data science applications, any distributed data system is the ability to perform the required amount of data exchange and computation in the available timeframe, which translates into a required minimum data flow and processing rates. Big data scenarios turn this into a harder endeavor due to several reasons, including the following characteristics of data.

VOLUME	high volume of data (more data to process)
VELOCITY	high rate of incoming data (less time to process data)
VARIETY	data heterogeneity (more data formats or data sources to deal with)

The adoption of today's digital technologies offers a unique opportunity to revolutionize clinical trials with significant improvements in time, cost, and the quality of data collected through the introduction of real-time data capture.



*Cost pressures on healthcare systems will reach the point that pharmaceutical companies will increasingly be paid for value i.e. if the heath outcome of drug treatment is positive in the real-world.

With the advent of digital technologies, pharmaceutical companies are investing in several technologies such as eSource (e.g. electronic informed consent, and direct data entry into tablet computers), and remote patient monitoring (e.g. wearable or home-based medical devices transmitting patient data securely) that will enable real-time data capture and analytics. Real-time workflow data for study management will enable accurate trial status reporting, with benefits to drug supply reconciliation and wastage reduction, accurate and timely payments for investigational sites and CROs, and an overall improvement in study and program budget tracking.

The use of remote patient monitoring technologies has the potential for further cost reductions. For example, the typical Phase III protocol in 2012 had nearly 170 procedures performed on each patient across 11 visits. If multiple visits can be reduced, and technologies such as patient engagement apps introduced on the patient's own smartphone (Bring Your Own Device - BYOD) it seems likely that the new digital technologies will lead to enhanced patient engagement, compliance, retention, and even initial recruitment. This combined with richer, more frequent sensor data sets could lead to reduced patient sample sizes and eventually new digital biomarkers and endpoints. Perhaps the largest cost savings will be through earlier and better decision-making, such as prompt trial termination or re-design decisions. Consequently, 30% productivity gains are achievable in clinical trials by embracing the technologies that will enable real-time data capture and analytics.

Remote patient sensing technologies combined with patient engagement apps on patients' own smartphones during clinical trial phases will not only expedite the clinical trials process but will aggregate unique patient behavioral data sets and insights that will facilitate algorithm and services development to maximize effective drug use and resulting patient outcomes.

Information is being collected and generated from more sources than ever before, including sensors at the edge of IoT systems, social media, mobile devices, the web and traditional business data stores at a greater velocity and variety that needs to be processed using a different architecture rather than a traditional architecture to process the big data.

This paper would be going over few Data Science Application and Scenarios which are relevant to Clinical and Non-Clinical Industry segments to give you a sample of data which needs to be processed using a different architecture due to the arrival rate and variety of information that is captured as part of the deployment. Later section of the paper, dwells into the Edge Computing architecture that needs to be considered when the devices, technologies and networks are deployed to heterogeneous environment.

DATA SCIENCE APPLICATIONS

The data science applications are segregated into four major categories, namely statistical applications, business intelligence, data mining and data collaboration. Statistical applications focus on ad-hoc analysis and allow data scientists to do powerful things like run regressions and visualize data in a more easily digestible format. Business intelligence applications are essentially statistical applications focused on creating clear dashboards and distilling metrics. Data mining applications predict future trends and behaviors, allowing business to make proactive and present knowledge in the form which is easily understood to human. Data collaboration is another space that's likely to be more and more important in future as companies build out larger data science teams.

APPLICATIONS USES DATA SCIENCE

The following major areas of applications uses data science.

GENERAL	PHARMA
Internet Search	Patient Disease Identification/Diagnosis
Self-Driving Cars	Drug Discovery/Manufacturing
Robots	Personalized Treatment/Behavioral Modification
Delivery Logistics	Clinical Trial Research
Fraud and Risk Detection	Radiology and Radiotherapy
Airline Route Planning	Smart Electronic Health Records
Image Recognition	Epidemic Outbreak Prediction
Speech Recognition	Health Care Monitoring
Digital Advertisements	Genome Research

Data science is quickly becoming a game changer for the pharmaceutical and biotech industries. From drug discovery to getting the right treatments to the right patients at the right time, artificial intelligence is making medical research and treatment faster and more successful on a massive scale.

FAST DATA APPLICATIONS

Fast data has several inherent issues, in addition to those relating to big data. It depicts several scenarios in which massive quantities of data can be produced from a heterogeneous set of data sources, eventually with different formats and processing requirements. For simplicity, not all possible connections are depicted, but the inherent complexity of integrating all these systems and processing all the data they can produce is easy to grasp.

Most big data applications today use best-effort technologies such as Hadoop and Spark, in which immutable data is previously loaded into the processing nodes. This is suitable for applications in areas such as business analytics, which attempt to mine information that can be relevant in specific contexts and essentially just deal with the volume dimension of big data. However, this is not the case for applications where many data sets are produced, or many events occur frequently, in a heterogeneous ecosystem of producers and consumers. In these applications, processing needs to be performed as data are produced or events occur, therefore emphasizing the variety and velocity dimensions (fast data).

No matter which dimension we consider, “big” essentially means too complex, too much, too many and too fast to apply conventional techniques, technologies and systems, since their capabilities are not enough to handle such extraordinary requirements. This raises the problem of integrating heterogeneous interacting parties to a completely new level, in which conventional integration technologies (such as HTTP, XML, JSON, Web Services and RESTful applications) expose their limitations. These are based on technologies conceived initially for human interaction, with text as the main format and sub-second time scales and not for heavy duty, machine-level binary data exchange that characterizes computer-to-computer interactions, especially those involving big data. Innovative solutions are needed to deal with these integration problems, in what concerns fast data requirements. Unlike processing of large passive and immutable data sets, for which frameworks such as Hadoop are a good match, fast data scenarios consist of a set of active interacting peers, producing, processing and consuming data and event notifications.

DATA SCIENCE APPLICATION SCENARIO – I

[Non-Clinical Systems]

Integrated Vehicle Management System (IVHM) - is the process of assessing, preserving, and restoring system functionality across flight and ground systems.

Today, many of the world’s leading aerospace manufacturers make regular use of IVHM. This includes Boeing, Airbus, Rolls Royce, GE, Gulfstream, Embraer and Bombardier.

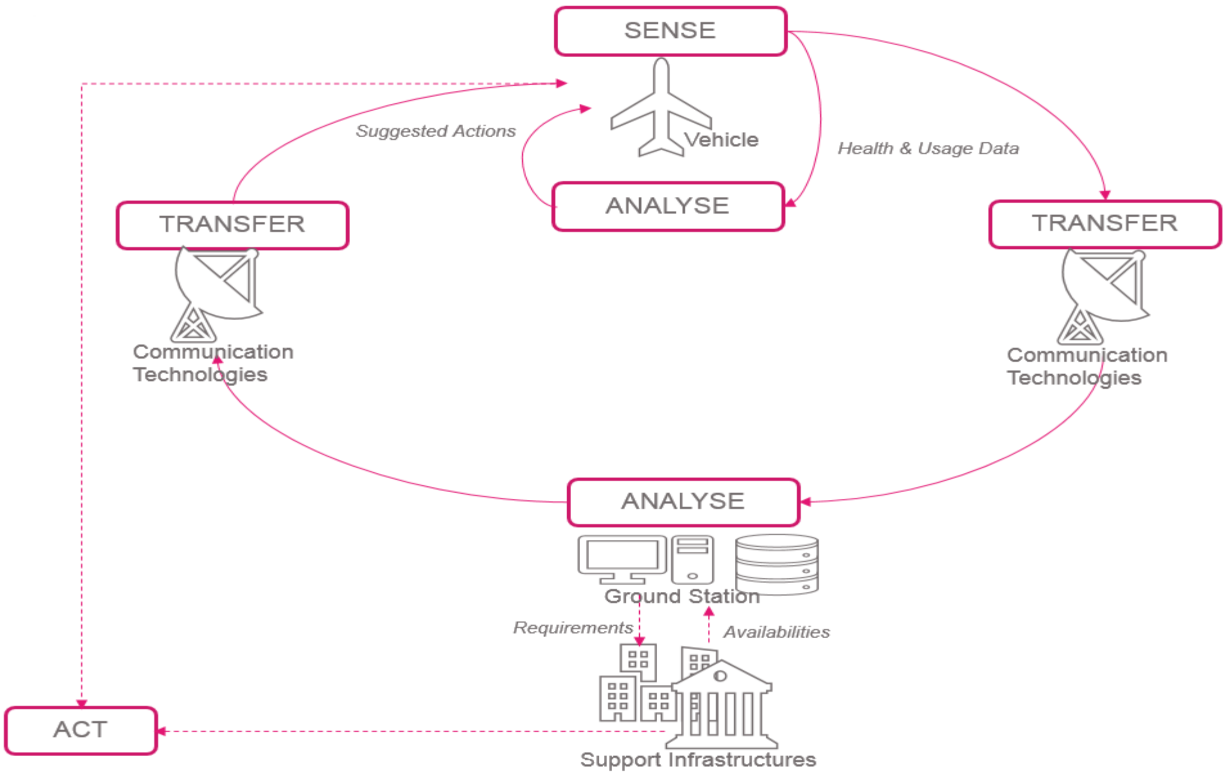
IVHM is defined as the transformation of system data into information to support operational decisions. For aerospace, this results in many business benefits such as minimized maintenance action/time, enhanced operational awareness, reduced inspections and troubleshooting, more efficient logistics operations, reduced schedule interruptions, and overall improved aircraft readiness, availability and safety.

On today’s smart aircraft platforms this is realized by collecting component/system health data via sensors from across an aircraft in near real-time during flight. Using diagnostic and prognostic algorithms, this data is converted into information that is used to make the best decisions on when to carry out maintenance and repair actions (such as replacement of a worn out or faulty part) that maximizes the aircraft’s availability and hence time operating safely in the air.

An IVHM system comprises a set of sensors and associated data processing hardware and software distributed between the vehicle and its support system. Here, for example, consider an aircraft. As illustrated in figure below, the IVHM system requires appropriate sensors to be positioned on critical components of the aircraft, monitoring the relevant subsystems (e.g. engine, propulsion, avionics, and structures) and state variables (e.g. temperature, pressure, speed, flowrate, and vibrations). The data collected by sensing devices are analyzed onboard the vehicle. At the same time, health and usage data are also transmitted to a ground support center where additional data analysis capabilities are deployed. In this case, wireless networks or more simple communication technologies are used to send the data from the aircraft to the remote support center so that analysis can still be performed in-flight.

Less critical data can be stored on the aircraft during the flight and accessed post flight at the ground station. Within the onboard and ground-based systems, vehicle health state is monitored continuously, and predictions are made regarding the remaining life. Then, if appropriate, actions are suggested to minimize the effect of faults or to repair/replace the failing components. In addition to this, planning and execution of actions can be arranged in conjunction with the support infrastructure which provides recovery and maintenance support. This example illustrates the typical configuration of an IVHM system.

At one end of the spectrum, all health management functions are incorporated onboard the vehicle, and at the other the data processing is entirely carried out with remote resources. Incorporating health management functions onboard, i.e. increasing vehicle autonomy, is motivated by a reduced dependence on data communication and therefore reduced operation costs and quicker response capabilities to unexpected events. On the other hand, diverting the analysis of health data to a remote support center provides enhanced fault forwarding, troubleshooting, and historical information support while reducing the amount of instrumentation and computer resources that need to be furnished onboard the vehicle.



At the core of IVHM is the realization that sensors and data buses onboard an aircraft can be storytellers of the aircraft's experience, condition and readiness for future missions. The challenge is to convert this data which is arriving at a higher velocity and variety into information that's understandable, specific, pertinent and accurate, and get it to mechanics quickly enough to maintain the system effectively. The onboard IVHM system continuously feeds data into a maintenance database, which enables pilots to perform corrective operations in flight and mechanics to schedule maintenance tasks before the plane lands.

The above is a classic example which has been extended to all kinds of transportation and other industries such as energy generation, manufacturing, dealing with volume, velocity and variety of data that must be processed at greater speeds to make it a more meaningful information and decision making. Digital technologies blended with IoT – Internet of Things has driven the industry to revisit the traditional system architecture to a more robust distributable architecture with splits the computing power across devices by adopting to Edge computing.

Edge computing entails processing and analyzing data closer to the source of where that data is collected. Instead of a device or sensor sending all its data over the internet to the cloud or an on-premise data center, it can:

- Process this data itself, essentially becoming its own mini data center.
- Deliver this data to a nearby computing device, such as a gateway networking device, a computer, or micro data center for analysis. This is sometimes called fog computing, though edge and fog computing are often used interchangeably.

With this new kind of architecture, a vast amount of processing power becomes decentralized from cloud service providers, which can help increase the speed of data analysis and decrease the load placed on internet networks to transmit enormous amounts of data.

DATA SCIENCE APPLICATION SCENARIO – II

[Clinical Systems – Health Care Monitoring]

Clinical trials, which seek to prove the safety and efficacy of new treatments through controlled testing on patients, are fraught with challenges that drive up R&D costs and slow the delivery of promising new treatments.

Clinical trials account for a staggering 40 percent of the pharma industry's research budget. Demonstrating a new treatment's efficacy can be particularly challenging for neurological and other diseases where symptoms vary widely, changes are subtle, and disease progression can be hard to assess. To determine efficacy, researchers gather evidence through clinic visits, where patients report on their symptoms, or paper diaries, in which patients record their medication regimen, symptoms, and observations. These approaches burden patients, contributing to the possibility of dropouts. In addition, the resulting evidence is subjective and often grows spotty during lengthy trials, providing a limited basis for analytics and decision making. And because patients may project a positive outcome during clinic visits, their reports may inadvertently produce biased results.

Remote monitoring with wearable devices offers new opportunities to advance clinical trials. By collecting data such as patients' movement activity, heart rate, and glucose levels, these devices can help produce consistent, objective evidence of the actual disease state and a treatment's impact. Kaiser Associates conducted a study for Intel projecting that up to 70 percent of clinical trials will incorporate wearable sensors by 2025. Leaders in the pharmaceutical, contract research, and medical device industries reported that they expect wearable devices to help reduce clinical trials costs, deliver higher-quality data, and speed time-to-results.

To achieve that value, pharmaceutical companies must capture, manage, and analyze vast amounts of data from wearables characterized by variety, velocity and volume. A typical phase 2 trial that runs for 6 months with 100 patients would generate over 200 billion data points. Developing objective criteria for assessing a treatment's impact is a crucial challenge and requires expertise as well as domain-specific experience.

Many pharmaceutical companies have started deploying the wearable device and mobile computing to capture the patient data at real time. Patients are monitored continuously in their home environments and can carry out predefined structured tests. Digital revolution with advent of IoT has enabled the clinical industry to explore new avenues of data capture for Rare Disease Patients who are mostly residing in their homes and it was a challenging task for the pharma companies to deploy necessary resource to conduct the clinical trials. Remote data collection has the potential to help reduce the frequency of clinic visits and simplify the tasks clinicians and patients must perform during each visit. These changes can help lower site costs while minimizing the burdens on patients, easing patient recruitment, and increasing retention.

Over recent years, it is observed with the development of mobile and wearable technologies to collect data from human vital signs and activities. Nowadays, wrist wearables including sensors (e.g., heart rate, accelerometer, and pedometer) that provide valuable data are common in market. During this development, it was identified interoperability challenges related to the collection and processing of data from wearable devices. Different vendors adopt specific approaches about the way data can be collected from wearables into third-party systems.

The presented Data Science Application scenario -II focus on the work which lies in using these novel health systems for sensing, and data processing to create some intelligence in the devices themselves and collaborate in providing advanced applications to the user before sending the data to next level for processing. Wearables are now built with sensors that has the necessary capability to collect and perform computing operation before the sending the patient data to smartphones

Smartphones can perform two different tasks: firstly, they send all the data to the server-side layer that processes, analyzes and stores the data, and shows the processing results; secondly, they are an alternative to processing and analyzing all the data, and presenting the processing results. The below table summarizes the contributions.

RESEARCH	BIOMEDICAL SIGNALS	DEVICES
Real-time streaming data in healthcare applications	Generic Biomedical signals	Generic Biomedical sensors
Recognition of activities and health monitoring	Heart biomedical signals	Smartphones & wearable devices
Long-term monitoring of respiration and pulse	Respiration and pulse	Non-contact sensors textile-integrated
Diabetes monitoring	Daily activity data	Smartphone & smartwatch
Active assistance	Activity and environment data	Wearable sensors and smartphone

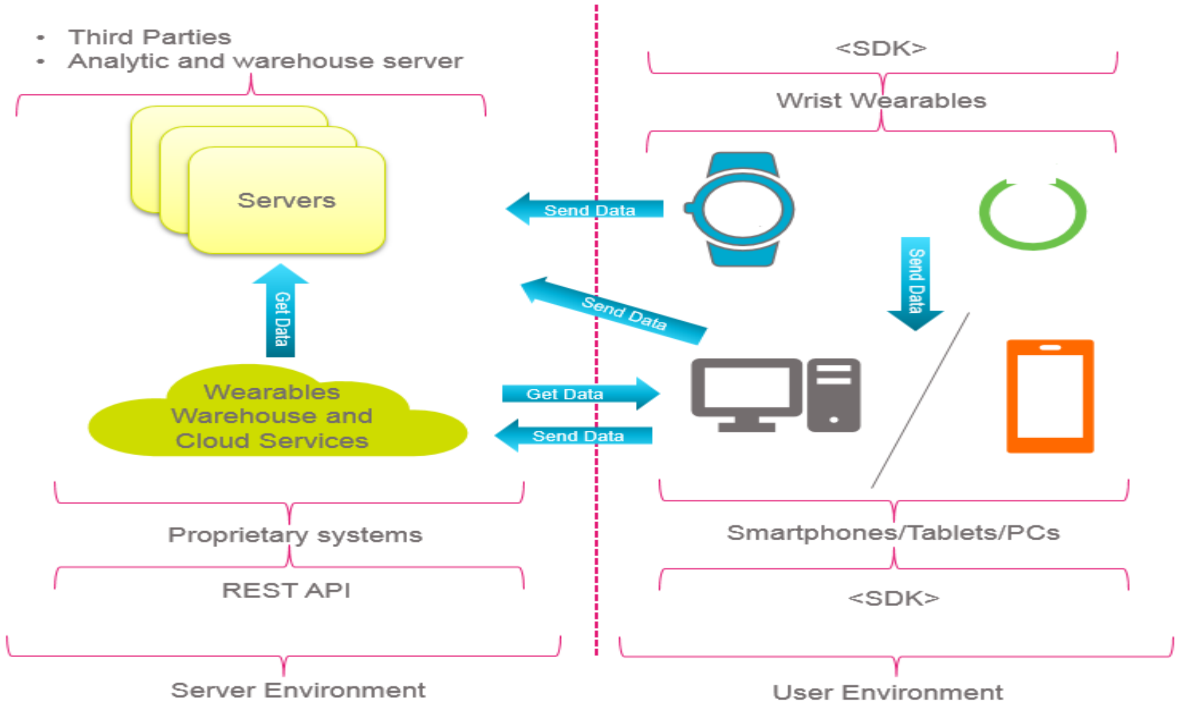
RESEARCH	BIOMEDICAL SIGNALS	DEVICES
Detect and prevent venous stasis	Pulse and blood flow data	Multi-sensor plethysmography device
Physiological data of elderly patients	Oxygen saturation level, Heart Rate	Biomedical sensors & smartphone
ECG Smart Healthcare monitoring	ECG signals	Wearable ECG sensors and Cloud for processing
Mobile medical computing systems	Medical signal and context information	Different sensors and actuators
Applications in the pervasive environment	Pulse rate, blood pressure, level of alcohol, etc.	Mobile healthcare

These biomedical sensors have a wide application in Clinical Trials and Health Care Monitoring. The solution integrates one or more wearable devices, a dedicated phone application with patient interaction, and a back-end cloud solution that stores big data and enables the development of novel algorithms.

One of the prime challenges of scientific healthcare applications is the streaming of data collected by many sensors deployed across the body. This current situation leads to novel scenarios where there is an intersection between the Internet of Things (with the quantity and quality of new available wearables), Big Data (the enormous amount of data collected by many heterogeneous sources) and real-time environments due to the specific features of the healthcare applications. This new situation is no longer manageable by traditional methods and new computing paradigms need to be defined.

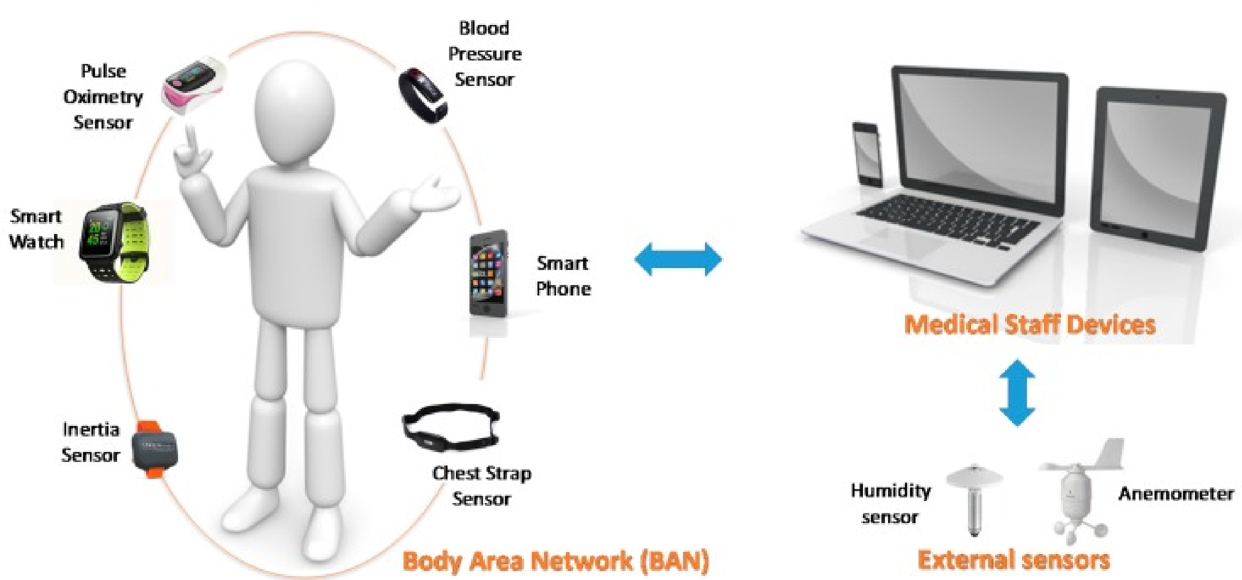
As part of Data Science Application, each one of these components is intended to perform some specific functions:

DATA COLLECTION & PROCESSING	Data is captured by the sensors available in wearable devices and smartphones. Data processing is done at sensor in order to create some intelligence in the devices themselves and collaborate in providing advanced applications to the user before transferring the full or processed data.
DATA TRANSFER	Data collected in wearables can be transferred to a computer or to a smartphone as an intermediate step towards its eventual transfer to a permanent data storage. This transfer can be produced through proprietary solutions or using third-party apps and programs.
PERMANENT DATA STORAGE	This function is related to the permanent storage of the data. Usually, this permanent storage is performed in proprietary servers, where third parties and final users can gain access to the data.
DATA ANALYSIS	This is related to the analytic processing of data to provide results of interest. Typically, this processing is performed in servers, both proprietary and third-party.



Modern activities of health monitoring have been revolutionized by development of IoT paradigm. Currently, the new smart sensor devices are changing the way users and professionals can work. In general terms, the combination of sensing technology and smart mobile devices can handle the computing capabilities of new healthcare applications which, not so long ago, had to be performed only in hospital environments in order to handle the velocity and variety of data that is being captured at any point of time in a heterogeneous fast data application.

The proposed framework enhances this health monitoring by means of leveraging the computing capabilities of modern wearables and other IoT devices for computing advanced medical applications. The main idea is to get some devices (sensors or wearables) of the BAN can take part in the application processing and then, to provide high level information to the medical staff's devices to perform further data analysis. The figure below shows the general scheme of the proposed computing elements considered by the framework. In contrast with other approaches, the criterion to be a part of this BAN is that the device is worn by the user, and not it's processing capability. This approach combines the 'things' of the BAN with sensing and computing capabilities in an integrated way. That is, there are thing with only sensing features (i.e., biosensors), there are things with only computing features (i.e., smartphones) and there are things with both sensing and computing features (i.e., smartwatches). All of them have communication capabilities to make a wireless network to share data.



As shown in the above figure, data acquisition is always made through devices of BAN. The preprocessing stages are serious candidate to be processed on the sensing devices and/or other devices of the BAN. In this way, the BAN supplies clean data instead of raw data. The next operations need some computing power to be performed. If it is possible to be computed inside the BAN, it provides health information rather than a data signal. In addition, it may reduce the amount of data to be transmitted and the bandwidth is most efficiently used.

In the above Data Science Application Scenario -II, a distributed framework that combines sensing and processing at different levels of the network to share the computing load among the available devices has been proposed to address this challenge. The IoT environments composed of wearables and other biosensors may benefit of it by allowing the processing of advanced applications with real-time constraints in a collaborative way. This distributed architecture enables to handle the velocity and variety of data that is acquired in a heterogeneous environment with some amount of pre-processing paving the way for edge computing in future.

The main advantages and novelties of the proposed system is the flexibility in the application execution by using resources from different available devices. In this way, the devices of BAN can provide shared computing resources that enables real-time monitoring and analysis of all acquired data.

FAST DATA APPLICATION DISTRIBUTED PLATFORM

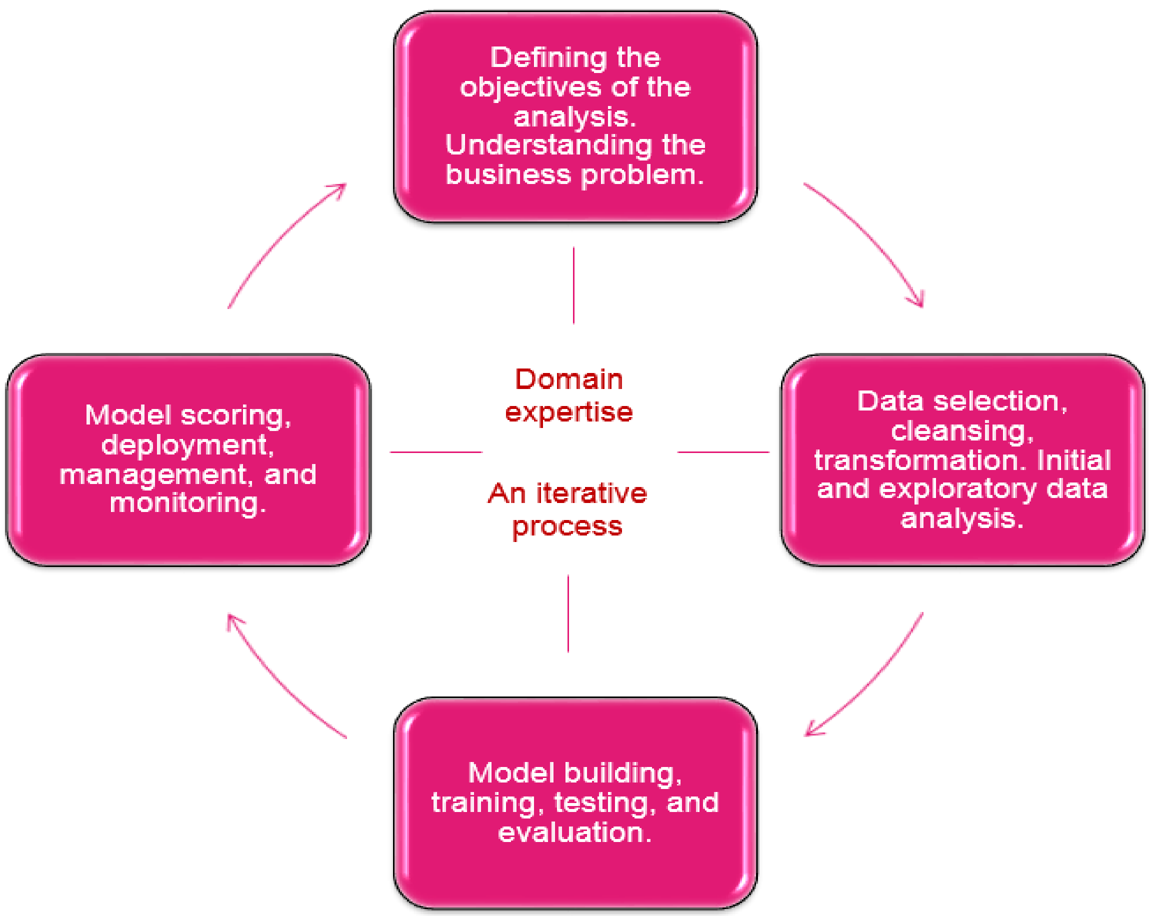
Scenarios I & II described above has given a good understanding on How IoT in a distributed heterogeneous environment has helped Data Science Applications which are deployed in Aeronautical, Oil & Gas and Health Care & Clinical Trial monitoring etc along with its inherent benefits offering exciting capabilities for businesses to continually unlock the potential of their data through techniques such as predictive analytics, machine learning, and data mining.

From a technical point of view, the work described in this paper is about data communication and homogenization, involving different types of devices: wearables, smartphones, computers and servers. This is situated in the current context of the Internet of Things (IoT) and Machine to Machine (M2M) standardization. There is a need for interoperability between different devices that can behave more or less as autonomous entities in an ecosystem. Communication is a main challenge, but other issues also are critical: security, privacy, trust, etc.

Data Science application architecture is guided to full fill the following requirements.

- Enable Apps to consume predictions and become smarter
- Bring predictive analytics to the IoT Edge
- Become easier, more accurate & faster to deploy and manage
- Fully support data science life cycle

Data Science Application Life Cycle is depicted in the figure below, which outlines the steps, from start to finish, that projects usually follow when they are executed.



Data Science Process

It is observed from the scenarios that the sensors are generating data constantly, and often analysis must be rapid. By the time the data makes its way to the cloud for analysis, the opportunity to act on it might be gone. This paper explains a new model for analyzing and acting on IoT data. It is either called edge computing or Fog computing.

- Analyzes the most time-sensitive data at the network edge, close to where it is generated instead of sending vast amounts of IoT data to the cloud.
- Acts on IoT data in milliseconds, based on policy.

- Sends selected data to the cloud for historical analysis and longer-term storage.

Capitalizing on the IoT requires a new kind of infrastructure. Today's cloud models are not designed for the volume, variety, and velocity of data that the IoT generates. Moving all data from these things to the cloud for analysis would require vast amounts of bandwidth.

Handling the volume, variety, and velocity of IoT data requires a new computing model. The main requirements are to:

Minimize latency	Milliseconds matter when you are trying to prevent manufacturing line shutdowns or restore electrical service. Analyzing data close to the device that collected the data can make the difference between averting disaster and a cascading system failure.
Conserve network bandwidth	Offshore oilrigs generate 500 GB of data weekly. Commercial jets generate 10 TB for every 30 minutes of flight. It is not practical to transport vast amounts of data from thousands or hundreds of thousands of edge devices to the cloud. Nor is it necessary, because many critical analyses do not require cloud-scale processing and storage.
Address security concerns	IoT data needs to be protected both in transit and at rest. This requires monitoring and automated response across the entire attack continuum: before, during, and after.
Operate reliably	IoT data is increasingly used for decisions affecting citizen safety and critical infrastructure. The integrity and availability of the infrastructure and data cannot be in question.
Collect and secure data across a wide geographic area with different environmental conditions	IoT devices can be distributed over hundreds or more square miles. Devices deployed in harsh environments such as roadways, railways, utility field substations, and vehicles might need to be ruggedized. That is not the case for devices in controlled, indoor environments.
Move data to the best place for processing	Which place is best depends partly on how quickly a decision is needed. Extremely time-sensitive decisions should be made closer to the things producing and acting on the data. In contrast, big data analytics on historical data needs the computing and storage resources of the cloud.

To quote Wikipedia, "Edge Computing is pushing the frontier of computing applications, data, and services away from centralized nodes to the logical extremes of a network. It enables analytics and knowledge generation to occur at the source of the data."

Fog Computing provides a unified solution at the edge for communications, device management, data harvesting, analysis and control. Fog Computing enables the deployment of a highly distributed but centrally managed infrastructure.

The IoT vision involves the connection of all kinds of physical devices to the virtual world, supporting their communication with existing Internet entities. No longer just people or software systems would exist in the virtual world, but also a myriad of devices that could be addressed, identified, located, sensed, actuated and, in general, interacted using information and communication technologies.

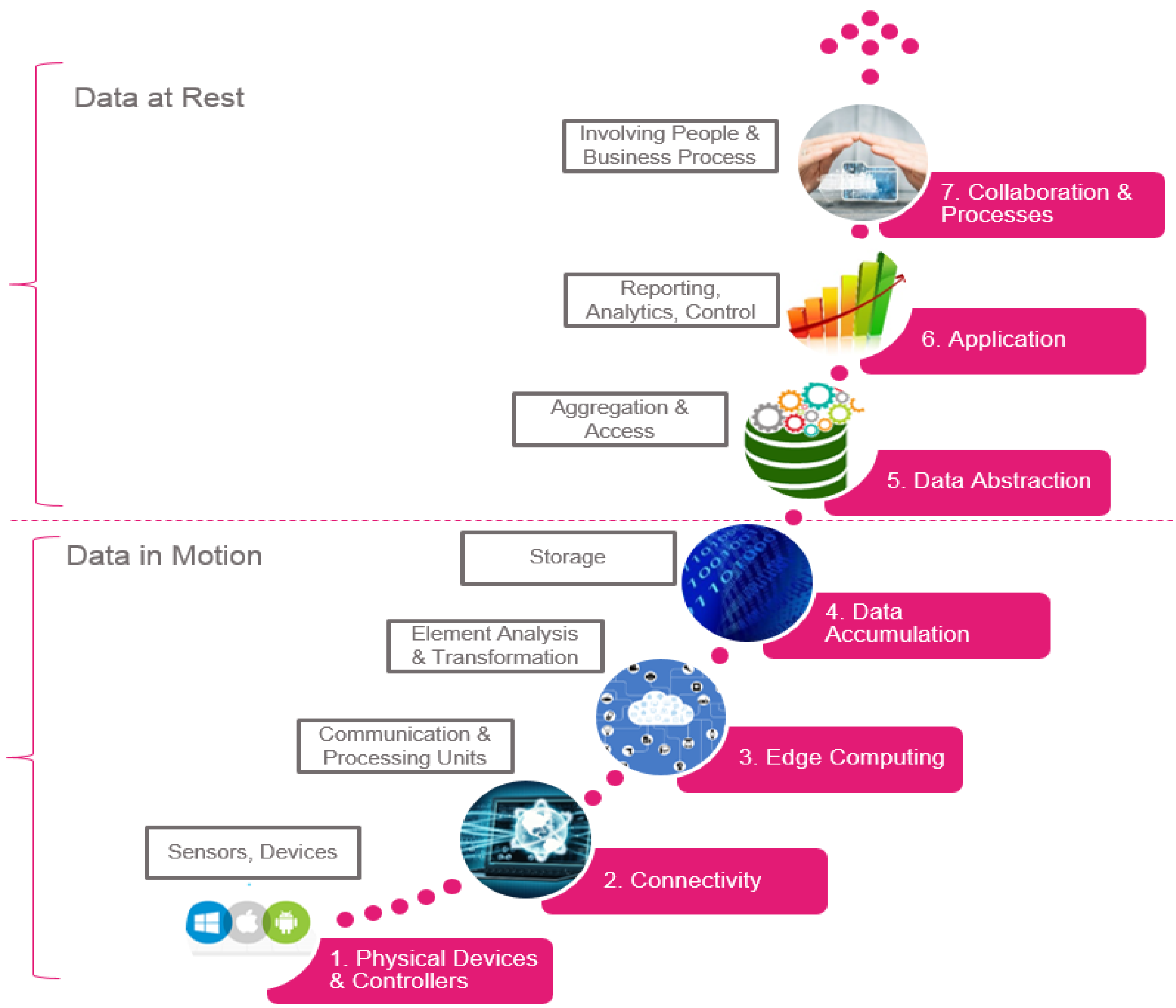
It is expected that the IoT has a strong impact on many aspects of human daily file in the near future. There is a common understanding that new technologies will soon enable the inclusion of communication modules, microprocessors and all kinds of electronic components as miniature pieces into everyday objects, making such objects smart entities. As a consequence, there will be a radical change from the 99% of things not connected to the Internet today to the 37 billion new things connected by 2020.

M2M is part of the vision of the IoT. Machines are considered in a broad sense as any type of device: computers, mobile devices, wearables, robots, cards, etc. The communication among machines can be performed in different ways using wireless or wired networks. Eventually, the M2M communications will allow end-users to get data about events of machines. M2M communications involve three phases: collection of data from sensors, transmission of the collected data to external systems and processing of data to provide some kind of result.

In recent years, main standardization bodies have launched several initiatives towards the standardization of the IoT and M2M domains. A main interest for the standardization bodies is the development of a framework or reference architecture that arranges the different issues involved. The three aforementioned initiatives are working towards such a result, but currently there is no clear proposal from any of them.

Cisco Systems, Inc. has published a white paper with a particular IoT Reference Model that deserves special attention.
It is made up of seven layers as shown in figure below:

1. Physical Devices & Controllers where sensors and actuators are situated;
2. Connectivity involving reliable and timely data transmission;
3. Edge (Fog) Computing involving initial data element analysis and transformation;
4. Data Accumulation to store data permanently;
5. Data Abstraction, providing services to aggregate and store data;
6. Application, related to the development of reporting, analytics and control; and
7. Collaboration & Processes, involving people and business processes.



This model groups layers under two different categories. Layers 1 to 4 are named as “Data in Motion” because data is processed and transformed while it is transferred from lower layers to upper layers. Layers 5 to 7 are named as “Data at Rest” because data is unchanging and stored in memory at each layer. Layer 3 is specifically tagged as “fog computing” because it refers to the idea that information processing can be initiated as early and as close to the edge of the network as possible. In this way the payload of the communications and processing in upper layers is reduced.

CONCLUSION

This paper analyses the less researched dimensions of velocity and variety in a Data Science Application that is involved in the collection and homogenization of data from devices say sensors, wearables of different providers in a heterogeneous fast data application environment. It highlights the issues about systems' interoperability and data integration while dealing with the different devices generating data constantly with variety and proposes a different computing architecture that is evolving at greater speed say Edge or Fog Computing that is gaining ground in IoT based industrial applications for large Data Science Applications.

The objective is to bring forward a new form of system for distributed systems, such as IoT, that meets the scalability and quality requirements, is flexible to meet the customers' exacting needs, comprehensive from the smallest device to the largest Information Technology system, and open so that everyone can extend the system.

REFERENCES

1. Collection and Processing of Data from Wrist Wearable Devices in Heterogeneous and Multiple-User Scenarios, Francisco de Arriba-Pérez, Manuel Caeiro-Rodríguez, Juan M. Santos-Gago Sensors (Basel) 2016 Sep; 16(9): 1538. Published online 2016 Sep 21
2. Cisco - The Internet of Things Reference Model. [(accessed on 24 May 2016)]. Available online: http://cdn.iotwf.com/resources/71/IoT_Reference_Model_White_Paper_June_4_2014.pdf.
3. Mattern F., Floerkemeier C. From Active Data Management to Event-Based Systems and More. Springer; Berlin, Germany: Heidelberg, Germany: 2010. From the Internet of Computers to the Internet of Things.
4. Hamida S., Hamida E., Ahmed B. A new mHealth communication framework for use in wearable WBANs and mobile technologies. Sensors.
5. Data Science and Big Data Computing Frameworks and Methodologies, Editor: Zaigham Mahmood,Department of Computing and Mathematics, University of Derby
6. Watson D., Piette M., Sezgen O. Machine to machine (M2M) technology in demand responsive commercial buildings; Proceedings of the 2004 ACEEE Summer Study on Energy Efficiency in Buildings; Pacific Grove, CA, USA. 23–27 August 2004.
7. Holler J., Tsiatsis V., Mulligan C. From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence. Academic Press; Cambridge, MA, USA: 2014.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Giri Balasubramanian, Edwin Ponraj Thangarajan
PRA Health Sciences
40, II Main Road, R.A. Puram
Chennai - 600 028, Tamilnadu, India
Email: BalasubramanianGiri@prahs.com, ThangarajanEdwin@prahs.com
Web: www.prahs.com