KSR INSTITUTE FOR ENGINEERING AND TECHNOLOGY

TIRUCHENGODE -637 215.

DEPARTMENT OF INFORMATION TECHNOLOGY

VOLUME 5
ISSUE 2

February
2018

DIGITIMES

Data Science for Engineers

KSR INSTITUTE FOR ENGINEERING AND TECHNOLOGY

Vision

To become a globally recognized Institution in Engineering Education, Research and Entrepreneurship.

Mission

M1	Accomplish quality education through improved teaching learning process
M2	Enrich technical skills with state of the art laboratories and facilities
M3	Enhance research and entrepreneurship activities to meet the industrial and societal needs

DEPARTMENT OF INFORMATION TECHNOLOGY

Vision

To produce competent Information Technology Professionals and Entrepreneurs with ethical values to meet the global challenges.

Mission

MD1	Impart quality education with ethical values in Information Technology through improved teaching learning process
MD2	Provide an ambient learning environment using state of the art laboratories and facilities
MD3	Encourage research and entrepreneurship activities to meet the dynamic needs of Information Technology industry and society

Program Educational Objectives (PEOs)

PEO	Key Words	Description
PEO 1	**Core Competency**	Graduates will be successful professionals in career by applying the knowledge of mathematics, science and engineering with appropriate techniques and modern tools.
PEO 2	**Professionalism**	Graduate will exhibit soft skills, professional and ethical values and thrust for continuous learning to maintain professionalism in the IT industries.
PEO 3	**Higher Studies and Entrepreneurship**	Graduates will engage in higher studies and outshine as entrepreneurs through life-long learning which leads to societal benefits.

DIGITIMES

Editorial

We would like to wholeheartedly thank our honorable Chairman, **Lion.Dr.K.S.Rangasamy** and vice chairman **Mr.R.Srinivasan,** and Principal **Dr.M.Venkatesan** for their continuous encouragement and constant support for bringing out the magazine. We profoundly thank our Head of the Department **Dr.P.MeenakshiDevi** for encouraging and motivating us to lead the magazine a successful one right from the beginning. DIGITIMES serves as a platform for updating and enhancing upcoming technologies in Information Technology. We are also grateful to all the contributors and faculty coordinator to bring this magazine.

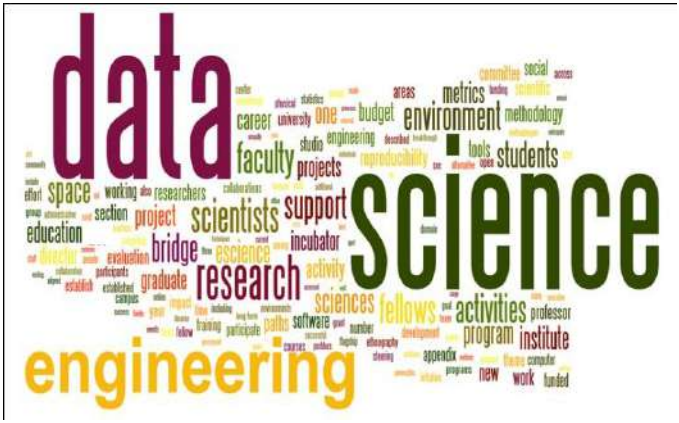
By,

Editorial Board

CONTENTS

DATA SCIENCE IN ENGINEERING

- **Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, and interact with data to create data products.
- Turn data into data products.



Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, structured or unstructured, similar to data mining.

The Data Science and Engineering (DSE) has been created to provide a comprehensive international forum for original results

in research, design, development, and assessment of technologies that timely address relevant challenges in data management and data-intensive applications.

Data Analytics vs. Statistical Analysis

Data Analytics	*Statistical Analysis*
❖ Utilizes data mining techniques	❖ Utilizes statistical and/or mathematical techniques
❖ Identifies inexplicable or novel relationships/trends.	❖ Used based on theoretical foundation.
❖ Seeks to visualize the data to allow the observation of relationships/trends	❖ Seeks to identify a significant level to address hypotheses or RQs

Data Science Applications

- Transaction Databases → Recommender systems (NetFlix), Fraud Detection (Security and Privacy)
- Wireless Sensor Data → Smart Home, Real-time Monitoring, Internet of Things
- Software Log Data → Automatic Trouble Shooting (Splunk)
- Genotype and Phenotype Data → Epic, 23andme, Patient-Centered Care, Personalized Medicine

Education for Data Science Roles

Eighty-eight percent of Data Scientists have a Master's Degree, and 46% have PhDs.

Other skills data scientists need include:

- In-depth knowledge of SAS and/or R. For Data Science, R is generally preferred. Science along with Java, Perl, C/C++.
- Hodoop platform: Although not always a requirement, knowing the Hodoop platform is still preferred for the field. Experience in Hive or Pig is a huge plus.
- SQL database/coding: Though No SQL and Hadoop are the major focus for data scientists; preferred candidates can write and execute complex queries in SQL.
- Working with unstructured data: It is extremely important that a Data Scientist is able to work with unstructured data—whether from social media, video feeds, audio, or other sources.

By

K.Vinitha, IV Year/IT

Without Your Involvement You Can't Succeed With Your Involvement You Can't Fail

-A.P.J.ABDUL KALAM

DATA SCIENCE VS BIG DATA

Data Science

Data scientists combine statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently to find patterns, along with the activities of cleansing, preparing, and aligning the data. Dealing with unstructured and structured data, Data Science is a field that encompasses anything related to data cleansing, preparation, and analysis. Data Science is an umbrella term for techniques used when trying to extract insights and information from data.

Big Data

"Big data is high-volume and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation." Big Data analytics find insights that help organizations make better business decisions.

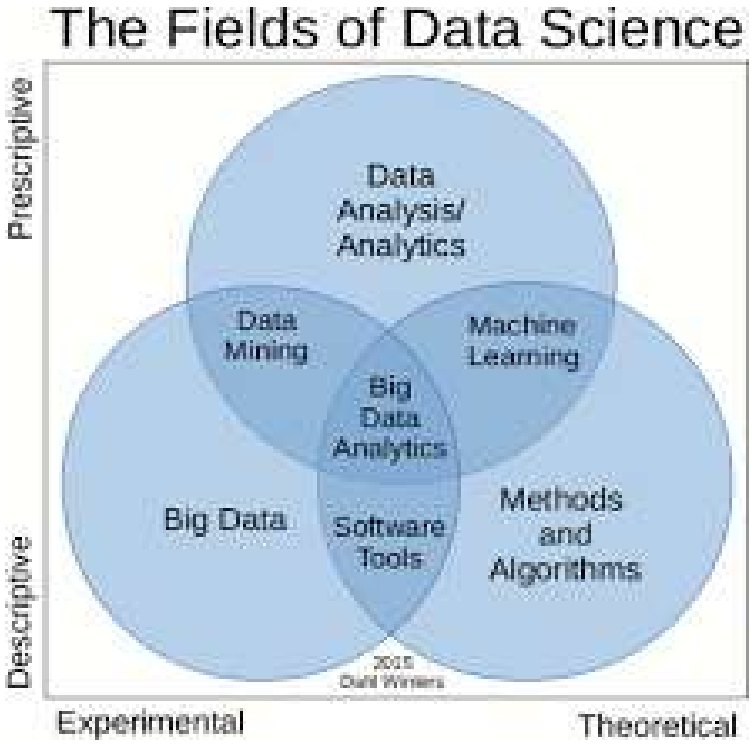
A buzzword that is used to describe immense volumes of data, both unstructured and structured, Big Data inundates organizations of all sizes on a day-to-day basis. Big Data refers to humongous volumes of data that cannot be effectively processed with traditional applications. The processing of Big Data begins with the raw data that isn't aggregated or organized—and is most often impossible to store in the memory of a single computer.

Data Analytics is the science of examining raw data with the purpose of finding patterns and drawing conclusions about that information by applying an algorithmic or mechanical process to derive insights. The work of a data analyst lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows; for example, running through a number of data sets to look for meaningful correlations between each other. Data Analytics is used in a number of

industries to enable organizations to make better decisions as well as verify and disprove existing theories or models.

By

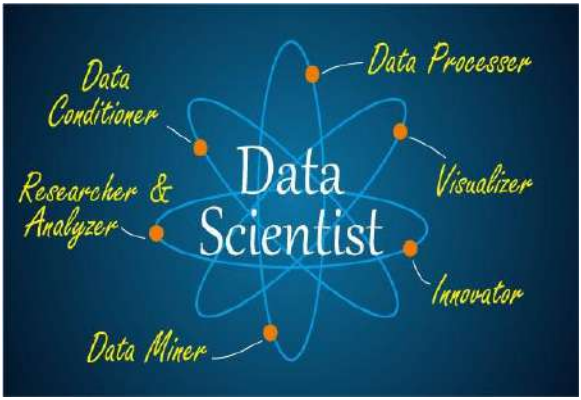
G.Dharani, III Year/IT



CATEGORIES OF DATA SCIENTISTS

❖ Statistician

They sometimes develop new statistical theories for big data that even traditional statisticians are not aware of. They are expert in statistical modeling, experimental design, sampling, clustering, data reduction, confidence intervals, testing, modeling, predictive modeling and other related techniques.



❖ Mathematician

NSA (national security agency) or defense/military people working on big data, astronomers, and operations research people doing analytic business optimization (inventory management and forecasting, pricing optimization, supply chain, quality control,

yield optimization) as they collect, analyze and extract value out of data.

❖ **Data processor**

Data engineering, Hardtop, database/memory/file systems optimization and architecture, API's, Analytics as a Service, optimization of data flows, data plumbing.

❖ **Innovators**

Those strong in business, ROI optimization, decision sciences, involved in some of the tasks traditionally performed by business analysts in bigger companies

❖ **Visualizer:** Those strong in visualization

❖ **Data miner:** Those strong in GIS, **spatial data**, data modeled by graphs, graph databases

By

P.Dharani, III Year/IT

Five things to remember about R

1. *Almost everything is a object*
2. *Almost everything is a vector*
3. *All commands are functions*
4. *Some commands produce different output*
5. *Know your default arguments*

WHY IS BIG DATA ANALYTICS IMPORTANT?

Big data analytics helps organizations harness their data and use it to identify new opportunities that lead to smarter business, more efficient operations, and higher profits.

Big data can be understood in the following ways:

1. **Cost reduction.** Big data technologies bring significant cost advantages when it comes to storing large amounts of data – plus identify more efficient ways of doing business.

2. **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, businesses are able to analyze information and make decisions.

3. **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want.

By
K.Gokul, IV Year/IT

BIG DATA AND THE INTERNET OF THINGS: TWO SIDES OF THE SAME COIN?

Read each statement and determine if it's referring to big data or the Internet of Things:

1. Every minute, we send 204 million emails, generate 1.8 million Facebook likes, send 278 thousand tweets, and upload 200 thousand photos to Facebook.

2. 12 million RFID tags (used to capture data and track movement of objects in the physical world) were sold in 2011. By 2021, it's estimated this number will increase to 209 billion as takes off.

3. The boom of [*big data or the Internet of Things?*] will mean that the amount of devices that connect to the internet will rise from about 13 billion today to 50 billion by 2020.

4. The [big data or the Internet of Things?] industry was expected to grow from US$10.2 billion in 2013 to about US$54.3 billion by 2017.

Here are the answers: *1 – big data; 2 – Internet of Things; 3 – Internet of Things; and 4 – big data.*

By
S.Vijay, IV Year/IT

Big Data Analytics in improving insurance claims data processing

Big Data analytics can make a big difference with insurance claims data:

Fraud –Most fraud solutions on the market today are rules-based. Unfortunately, it's too easy for fraudsters to manipulate and get around the rules. Predictive analysis uses a combination of rules, modeling, text mining, database searches and exception reporting to identify fraud sooner and more effectively at each stage of the claims cycle.

Subrogation – Opportunities for suborn often get lost in the sheer volume of data – most of it in the form of police records, adjuster notes and medical records. Text analytics searches through this unstructured data to find phrases that typically indicate a suborn case.

Settlement – To lower costs and ensure fairness, insurers often implement fast-track processes that settle claims instantly. But settling a claim on-the-fly can be costly if you overpay. By analyzing claims and claim histories, you can optimize the limits for instant payouts. Analytics can also shorten claims cycle times for higher customer satisfaction and reduced labor costs. It also

ensures significant savings on things such as rental cars for auto repair claims.

Loss reserve – When a claim is first reported, it is nearly impossible to predict its size and duration. But accurate loss reserving and claims forecasting is essential, especially in long-tail claims like liability and workers' compensation. Analytics can more accurately calculate loss reserve by comparing a loss with similar claims. Then, whenever the insurance claims data is updated, analytics can reassess the loss reserve, so we understand exactly how much money we need on hand to meet future claims.

Activity –Claims are usually assigned based on limited data – resulting in high reassignment rates that effect claim duration, settlement amounts and ultimately, the customer experience. Data mining techniques cluster and group loss characteristics to score, prioritize and assign claims to the most appropriate adjuster based on experience and loss type.

Litigation – Insurers can use analytics to calculate a litigation propensity score to determine which claims are more likely to result in litigation.

BY

B.Gokulkumar, IV Year/IT

If you want to shine like a sun

First burn like a sun

-A.P.J.ABDUL KALAM

ANALYTICS VIEW

Five views for approaches to analytics.

1. Open your mind

In today's economy, convention can no longer be a mainstay of business decisions. Companies that report a competitive advantage from using analytics are open to ideas that challenge their current practices. Because of that openness, the results of their analytics efforts have led to changes in how they do business.

2. Stop the insanity

Einstein defined insanity as doing the same thing over and over again and expecting different results. Many organizations have done little to nothing to improve their information management process but they expect big payoffs from their analytical efforts. Increased amounts of data have actually led to decreased insights. Companies must examine what they are doing at the front end of the analytics lifecycle where real change is needed to see desired results downstream.

3. Use more right brain

These days the talk is all about data scientists and quant's who live and breathe algorithms. And true, they can be the superheroes of big data, extracting new and meaningful insights out of

volumes of information. Open-mindedness and creativity play a key role in the effective use of analytics. Involve right-brained people in the analytical process to ask the outside-the-box questions.

4. Don't put all your eggs in…you know

A misconception about companies that are analytically driven is that they are completely analytically driven. That is, they make decisions solely by the numbers. Survey dispels that. True, companies more analytically mature are much more likely to use analytics to drive organizational strategy. But these organizations rely on a blend of intuition, experience and analytical results.

5. Stop dabbling

Even though analytics has become mainstream over the past 10 years and is often referred to as "table stakes," many organizations still run analytics initiatives on an adhoc or decentralized basis. If analytics is to be the golden goose that helps achieve and sustain competitive advantage, it's time to get serious about it. Companies struggling with their analytics initiatives have one major thing in common: they don't have a formal plan.

Learning gives creativity,
Creativity leads to thinking,
Thinking provides knowledge
Knowledge makes you great
-A.P.J.ABDUL KALAM

By

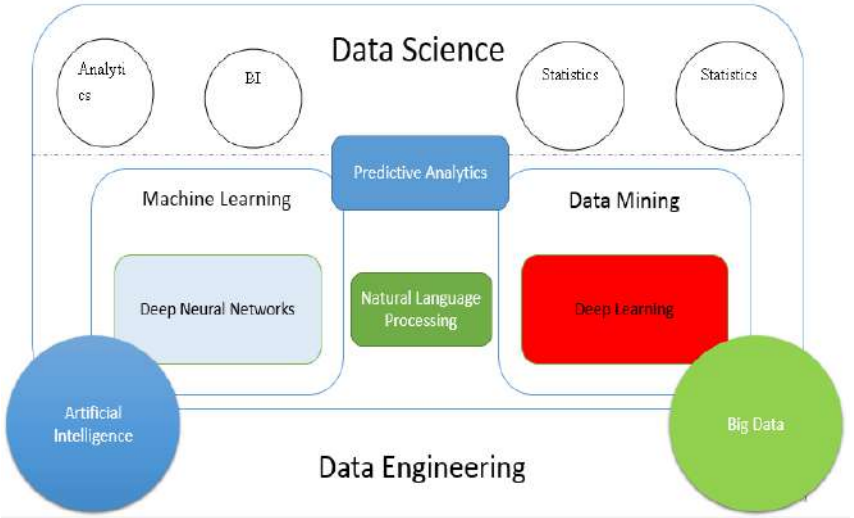
P.Hemalatha,III year/IT

CLOUD TOOLS

- Cloudability for Cloud Cost Analytics
- Cloudyn for Cloud Optimization
- Dell Boomi for Cloud Integration
- Enstratius For Cloud Infrastructure Management
- Informatica for Cloud Data Integration
- Mulesoft for Cloud Integration Service
- Opscode for Cloud Configuration Management
- Puppet Labs for Cloud Configuration Management
- Rightscale for Cloud Management
- Servicemesh for Enterprise Cloud Management

By

M.Karthika, II Year/IT



MAP REDUCE

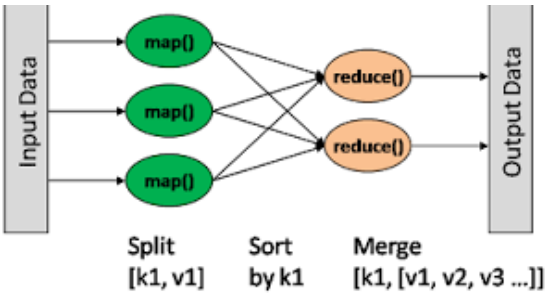
❖ **Map Reduce** is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

❖ Map Reduce is a framework for processing parallelizable problems across large datasets using a large number of computers (nodes) collectively referred to as a cluster in the components.

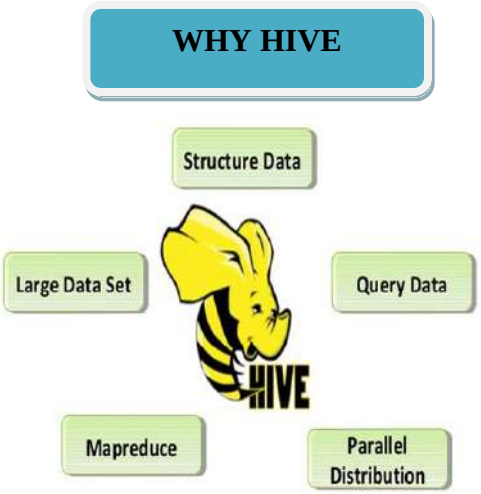
❖ "Map" step: Each worker node applies the "map()" function to the local data, and writes the output to a temporary storage. A master node ensures that only one copy of redundant input data is processed.

❖ "Shuffle" step: Worker nodes redistribute data based on the output keys (produced by the "map ()" function), such that all data belonging to one key is located on the same worker node.

❖ "Reduce" step: Worker nodes now process each group of output data, per key, in parallel.



By
G.Manibharathi,
II Year/IT



Apache Hive

It is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis. Apache Hive supports analysis of large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3 file system.

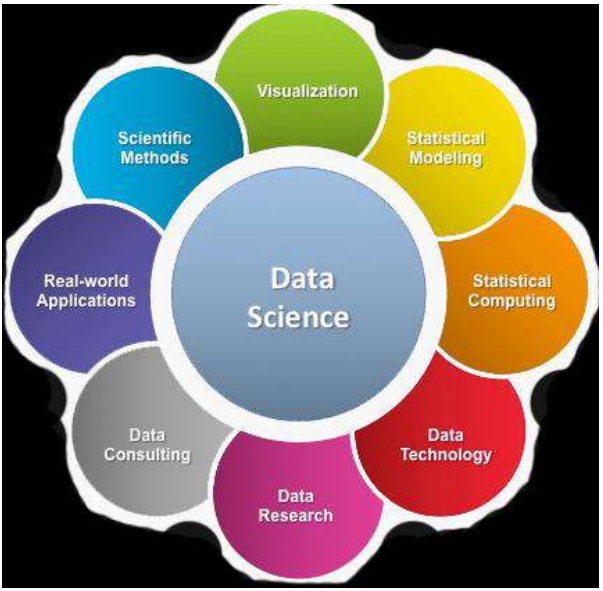
Other features of Hive includes

- Indexing to provide acceleration, index type including compaction and bitmap index as of 0.10, more index types are planned.
- Different storage types such as plain text, RCFile, HBase, ORC, and others.

- Metadata storage in a relational database management system, significantly reducing the time to perform semantic checks during query execution.
- Operating on compressed data stored into the Hardtop ecosystem using algorithms including DEFLATE,BWT, snappy, etc.Built-in user-defined functions (UDFs) to manipulate dates, strings, and other data-mining tools.
- Hive supports extending the UDF set to handle use-cases not supported by built-in functions.
- SQL-like queries (HiveQL), which are implicitly converted into Map Reduce or Tez, or Spark jobs.

By

R.Yuvapriya, II Year/IT



KAFKA - BIG DATA

WHAT IS KAFKA?

❖ KAFKA is an open-source stream processing platform developed by the Apache Software Foundation written in Scale and Java.

❖ The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds.

❖ Its storage layer is essentially a "massively scalable pub/sub message queue architected as a distributed transaction log," making it highly valuable for enterprise infrastructures to process streaming data.

❖ Additionally, Kafka connects to external systems (for data import/export) via Kafka Connect and provides Kafka Streams, a Java stream processing library.

❖ Due to its widespread integration into enterprise-level infrastructures, monitoring Kafka performance at scale has become an increasingly important issue.

❖ Monitoring end-to-end performance requires tracking metrics from brokers, consumer, and producers, in addition to monitoring Zoo Keeper which is used by Kafka for coordination

❖ There are currently several monitoring platforms to track Kafka performance, either open-source, like LinkedIn's Burrow, or paid, like Data dog.

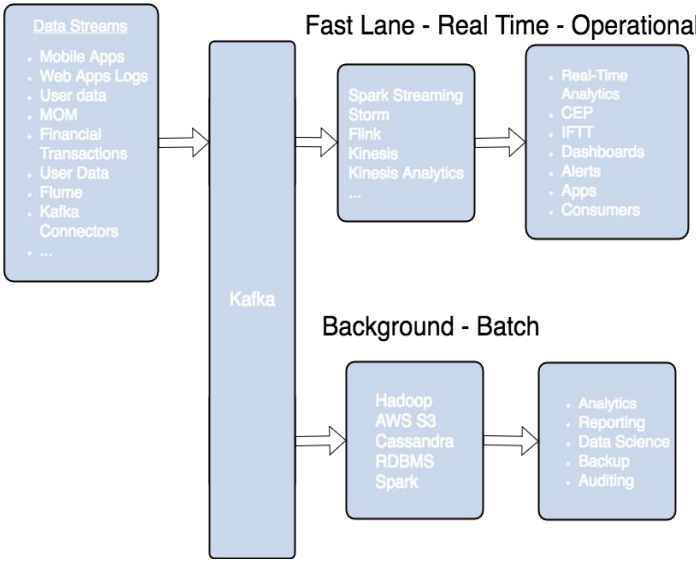
❖ In addition to these platforms, collecting Kafka data can also be performed using tools commonly bundled with Java, including Console.

❖ The major terms of Kafka's architecture are topics, records, and brokers. Topics consist of stream of records holding different information. On the other hand

KAFKA STREAMING ARCHITECTURE DIAGRAM

❖ Brokers are responsible for replicating the messages. There are four major APIs in Kafka: Streams API – This API converts the input streams to output and produces the result.

❖ Kafka has higher throughput, reliability, and replication characteristics, which makes it applicable for things like tracking service calls (tracks every call) or tracking IoT sensor data where a traditional MOM might not be considered. Connector API – Executes the reusable producer and consumer APIs that can link the topics to the existing applications.



WHY KAFKA?

❖Kafka is often used in real-time streaming data architectures to provide real-time analytics.

❖Since Kafka is a fast, scalable, durable, and fault-tolerant publish-subscribe messaging system, Kafka is used in use cases where JMS, Rabbit, and AMQP may not even be considered due to volume and responsiveness.

By

P Elangovan, III Year/IT

Be a hero. Always say,"I have no fear'

- SWAMI VIVEKANANDA

PIG-BIG DATA ANALYTICS

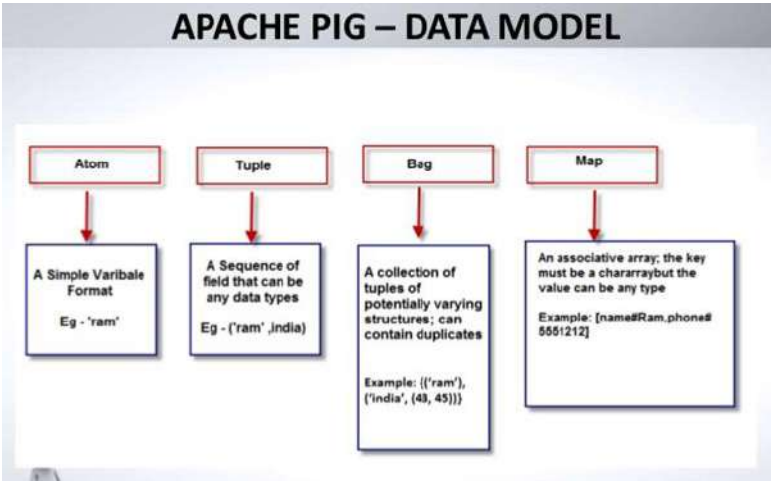
❖ Apache Pig was originally developed at Yahoo Research around 2006 for researches.

❖ **Pig** is a high level scripting language that is **used** with Apache **Hadoop**.

❖ It is also a procedural data flow language which is used by programmers and researchers.

❖ Pig is one of the components in hardtop ecosystem.

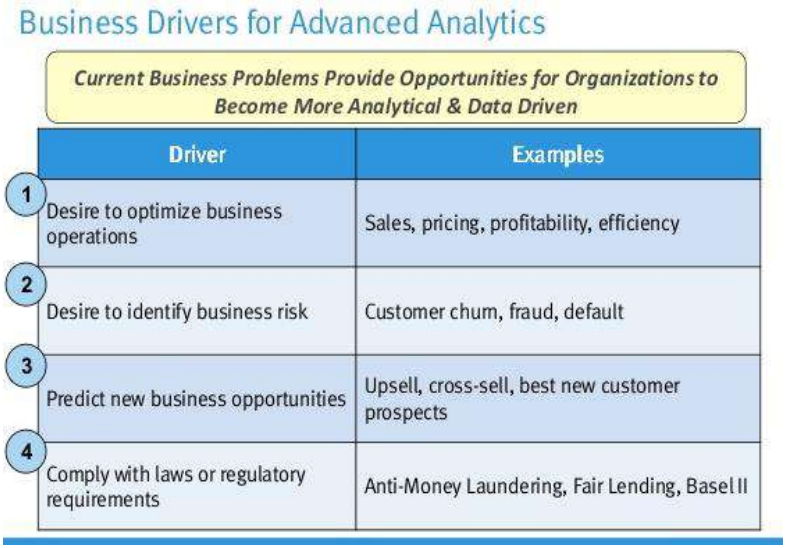
❖ Instead of writing Java code to implement MapReduce, one can opt between **Pig** Latin and Hive SQL languages to construct Map Reduce process.



- **Atom:** An atom is any single value, such as a string or a number.
- **Tuple:** A tuple is a record that consists of a sequence of fields. Each field can be of any type.
- **Bag:** A bag is a collection of non-unique tuples. The schema of the bag is flexible.
- **MAP**: It is an associative character array but the value can be of any type.

BY

S.MOHANRAJ, II Year/IT



R PROGRAMMING IN DATA SCIENCE

❖**R** is a free programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing.

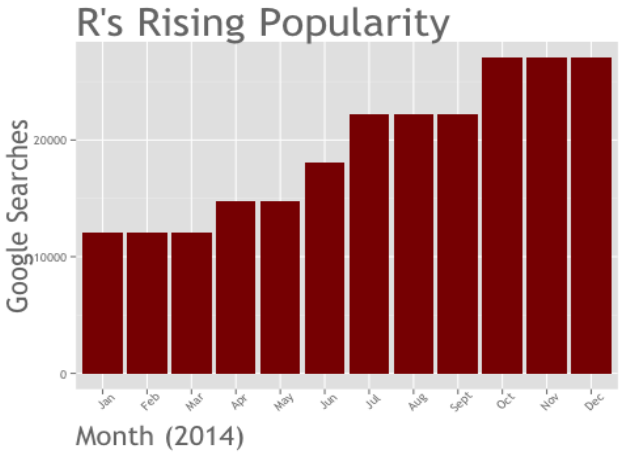
❖The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

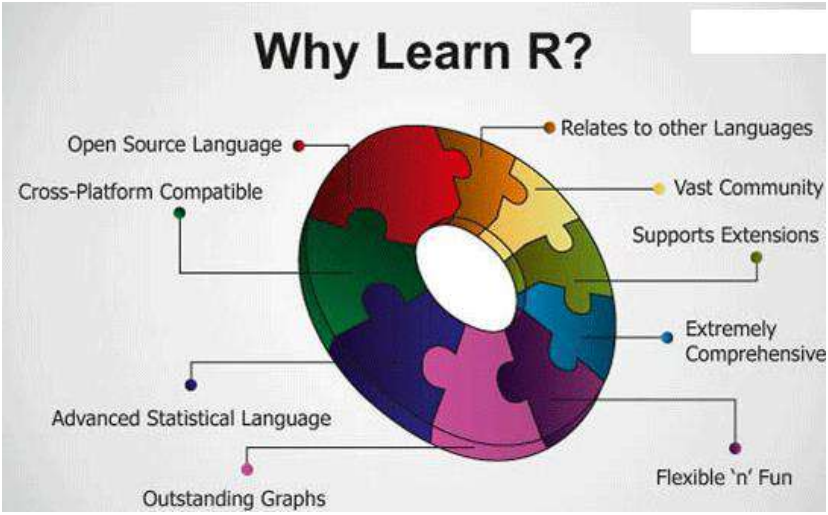
❖R is a GNU package.

❖The source code for the R software environment is written primarily in C, FORTRAN, and R.

❖It is an implementation of the S programming language combined with lexical scoping semantics inspired by scheme .

❖ S was created by John Chambers in 1976, while at Bell Labs.





Basic operations

- The most basic operation is a vector
- A vector can only contain objects of the same class.
- BUT: the one exception is a list, which is represented as a vector but can contain objects of different classes.

Objects:

R has five basic or "Atomic classes of objects:

1. Character
2. Numeric
3. Integer
4. Complex
5. Logical

By

K.Saraswathy,-II Year/IT

NASA: UNLOCKING THE SECRETS OF THE UNIVERSE WITH BIG DATA

❖ In NASA the actual rocket scientists use big data – close to an exabyte and counting – to keep hundreds of satellites in the air and fulfill its vision of "reach[ing] for new heights and reveal[ing] the unknown so that what we do and learn will benefit all humankind. NASA collects "big bang" data from across the solar system to unlock the secrets of the universe.

❖ Like the Square Kilometer Array project, this will aggregate data from tens of thousands of radio telescopes to figure out how the galaxy was formed at the "cosmic dawn?" NASA embraces extreme innovation in architecting its Storage and data management environment.

❖ It applies algorithms to data in tens of thousands of different formats from a range of spacecraft, including unmanned rovers and probes, earth-bound telescopes and observatories around the world – and archives it for future analysis as planetary science progresses.

By

A.Godson,IV Year/IT

DATA-DRIVEN "EARS" AND INSTINCT FOR THE MUSIC INDUSTRY

BIG DATA

The music industry has long been driven by instinct. Big-bet decisions on which acts to sign and promote were made by executives known for having great “ears” – i.e., an intuitive sense of who would top the charts and who was destined to a be one-hit (or no-hit) wonder. Big data is changing all that.

Record companies are pooling and correlating diverse data sets (downloads, social commentary, merchandise sales) and overlaying it with geographic and temporal data like concert locations and dates and TV appearances. Semantic and text analytics are necessary to understand that “bad” and “sick” are terms of endearment to music fans. But that’s what it takes to identify the next overnight sensation and define non-traditional metrics like engagement.

BY

N.Gnanasekar, IV Year/IT

HOW BIG DATA ANALYTICS IS USED TODAY

❖ As the technology that helps an organization to break down data silos and analyze data improves, business can be transformed in all sorts of ways. Today's advances in analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Face book.

❖ France's Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast. The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users. Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety. Proposed projects showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment.

By

S.Pavithra, IV Year/IT

BIG DATA IN VARIOUS DOMAINS

Travel and hospitality

Keeping customers happy is key to the travel and hotel industry, but customer satisfaction can be hard to gauge – especially in a timely manner. Resorts and casinos, for example, have only a short window of opportunity to turn around a customer experience that's going south fast. Big data analytics gives these businesses the ability to collect customer data, apply analytics and immediately identify potential problems before it's too late.

Health care

Big data is a given in the health care industry. Patient records, health plans, insurance information and other types of information can be difficult to manage – but are full of key insights once analytics are applied. That's why big data analytics technology is so important to heath care. By analyzing large amounts of information – both structured and unstructured – quickly, health care providers can provide lifesaving diagnoses or treatment options almost immediately.

Government

Certain government agencies face a big challenge: tighten the budget without compromising quality or productivity. This is particularly troublesome with law enforcement agencies, which

are struggling to keep crime rates down with relatively scarce resources. And that's why many agencies use big data analytics; the technology streamlines operations while giving the agency a more holistic view of criminal activity.

Retail

Customer service has evolved in the past several years, as savior shoppers expect retailers to understand exactly what and when they need it. Armed with endless amounts of data from customer loyalty programs, buying habits and other sources, retailers not only have an in-depth understanding of their customers, also predict trends, recommend new products and boost profitability

By

R.Rubini, IV Year/IT

Skill Set Needed for Data Science Engineers

- ❖Hadoop Certification & Experience(assumed)
- ❖EMC Data Scientist Certification
- ❖Visualization Expertise
- ❖Hands on with at least three Analytics Algorithms
- ❖Hands on with at least three big data sets in thr lab & with customer projects
- ❖Work with Data Scientists & Analytics groups

WHAT IS CLOUD SLANG?

Cloud Slang is an open source tool for orchestrating cutting edge technologies, such as Dockers and Core OS in an agent less manner. Use ready-made workflows or define your own custom ones. Cloud Slang workflows are reusable, shareable and easy to understand.

By

R.Swathi, IV YEAR/IT

KEY ROLES FOR A SUCCESSFUL ANALYTICS PROJECT

❖Business User – understands the domain area

❖Project Sponsor – Provides requirements

❖Project Manager – Ensure meeting objectives

❖Business Intelligence Analyst – provides business domain expertise based on deep understanding of the data

❖Database Administrative(DBA) – creates DB environment

❖Data Engineer – Provides technical skills, assists data management and extraction, supports analytic sandbox

❖Data Scientist – provides analytic techniques and modeling

IN –MEMORY ANALYTICS

❖ An approach to querying data when it resides in a computer's random access memory (RAM), as opposed to querying data that is stored on physical disks.

❖ BI and analytic applications have long supported caching data in RAM, but older 32-bit operating systems provided only 4 GB of addressable memory.

❖ In addition to providing incredibly fast query response times, in-memory analytics can reduce or eliminate the need for data indexing and storing pre-aggregated data in OLAP cubes or aggregate tables.

❖ This reduces IT costs and allows faster implementation of BI and analytic applications.

BY

V.Yaksheetha -II Year/IT

The capacity to learn is a gift; the ability to learn is a skill; the willingness to learn is a choice.

- BRAIN HERBERT

BIG DATA CHARACTERISTICS

1. Data volume

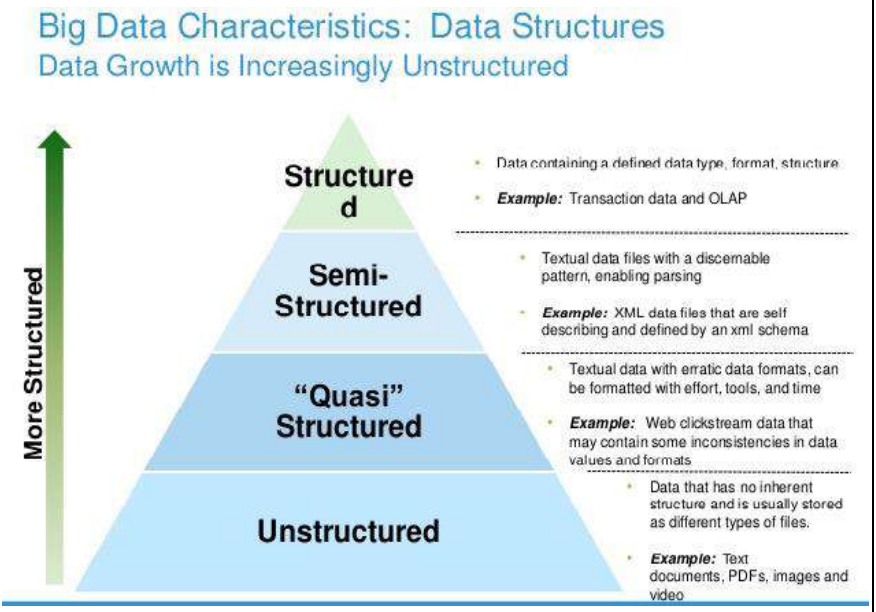
- ❖ High Volume
- ❖ 44Xincrease from 2009 to 2020(0.8 Zettabytes to 35.2zb)

2. Processing complexity

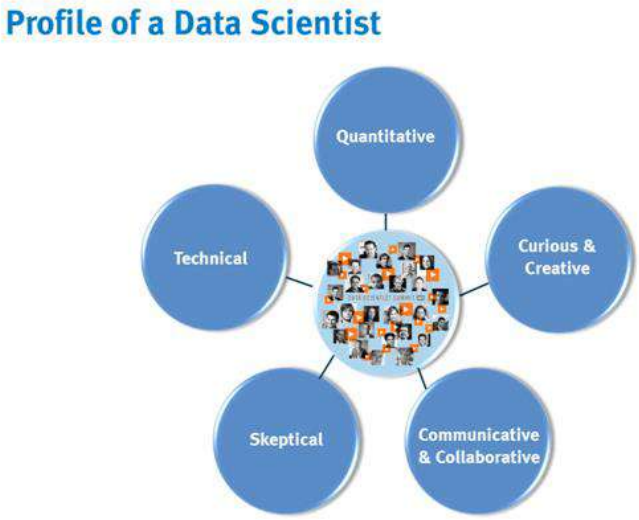
- ❖ Changing data structures
- ❖ Use cases warranting additional transformations and analytical techniques.

3. Data structure

- ❖ Greater variety of data structures to mine and analyze.



PROFILE OF A DATA SCIENTIST



Five Main Sets of Skills

- ❖Quantitative skill – e.g.,math,statistics
- ❖Technical aptitude – e.g., software engineering, programming
- ❖Skeptical mindset and critical thinking – ability to examine work critically
- ❖Curious and Creative – Passionate about data and finding creative solutions
- ❖Communicative and collaborative – can articulate ideas

By

M.Elankeni, III Year/IT

APACHE MADlib

Machine Learning at Scale

MAD stands for

- Magnetic,
- Agile ,
- Deep

Lib stands for library of

- Advanced
- Parallel& scalable
- In-database functions

Apache MADlib is an open source library for scalable in-database analytics. It provides data-parallel implementations of machine learning, mathematical and statistical methods on the Pivotal Greenplum PostgreSQL and the Apache HAWQ (incubating) Hadoop Native SQL platform.

MADlib uses the MPP architecture's full compute power to process very large data sets, whereas other products are limited by the amount of data that can be loaded into memory on a single node. MADLib algorithms are invoked from a familiar SQL interface so they are easy to use.

By

S.Mohana priya,III Year/IT

- Set of machine learning algorithms that leverage Hadoop to provide both data storage and the Mapreduce implementation.
- The mahout command itself a script that wraps the hadoop command and executes a requested algorithm from the Mahout job jar file.
- Scalable machine learning and data mining library fro Hadoop
- Support for four use cases

- Recommendation mining
- Classification
- Clustering
- Frequent item set

- Requires Hadoop infrastructure and java programming

By

R.Saravana priya,III Year/IT

BY

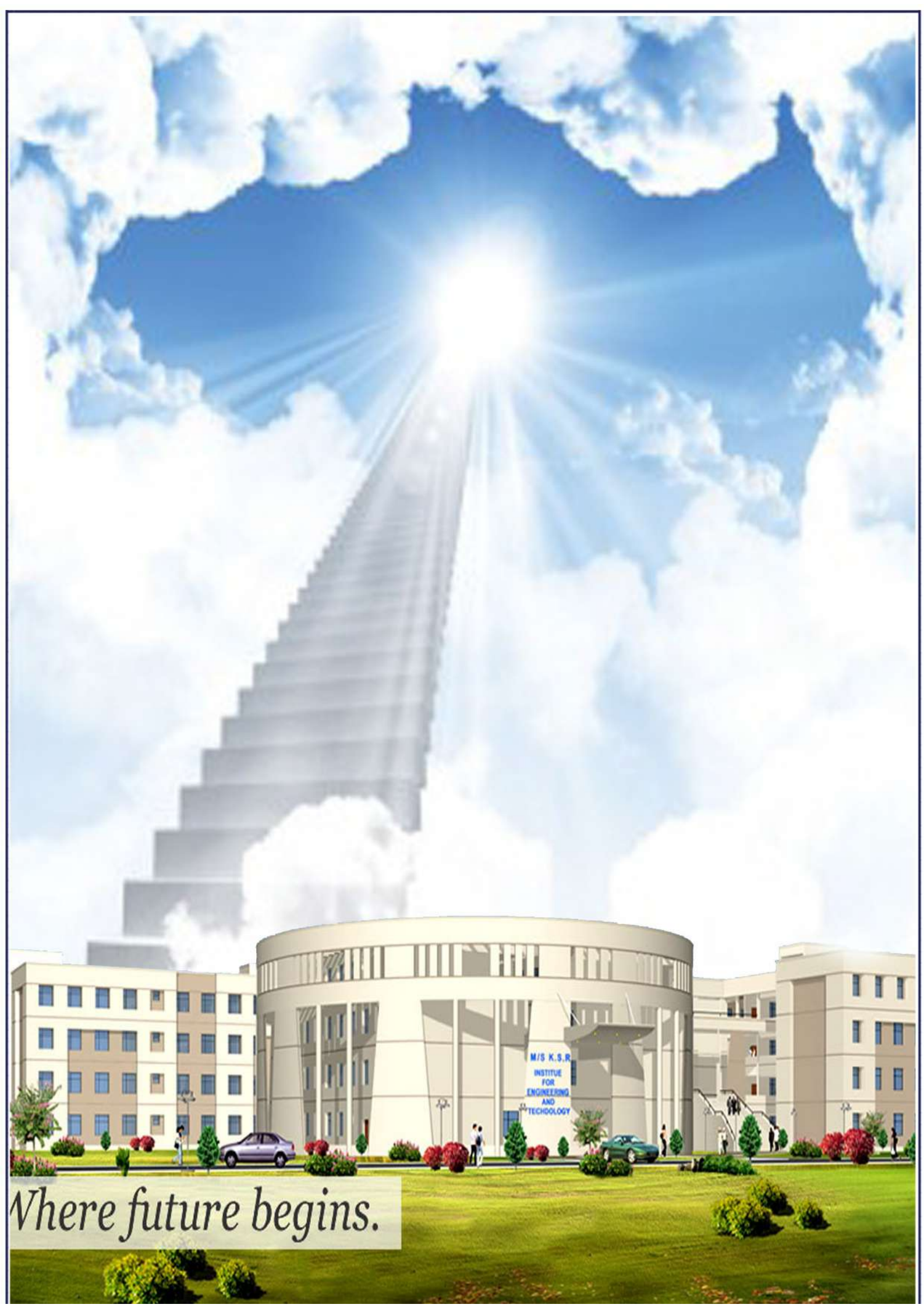
S.Karkulazhi,II Year/IT

Program Outcomes (POs)

PO1	**Engineering Knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the IT enabled solution of complex engineering problems.
PO2	**Problem Analysis:** Identify, analyze and provide solutions to the problems reaching substantiated IT enabled conclusions.
PO3	**Design/Development of Solutions:** Design solutions for complex engineering problems and design system components or processes that meet the desired needs within realistic constraints.
PO4	**Conduct Investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PO5	**Modern Tool Usage**: Create, select and apply appropriate techniques, resources and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
PO6	**The Engineer and Society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
PO7	**Environment and Sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
PO8	**Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of engineering practice.
PO9	**Individual and Team Work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO10	**Communication:** Communicate effectively on engineering activities with the engineering community and with society.
PO11	**Project Management and Finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO12	**Life Long Learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes(PSOs)

PSO1	**Programming Skill**	Work as Software Engineers for providing solutions to real world problems using programming languages and open source software.
PSO2	**Web Designing Skill**	Ability to use the web designing skill to establish new solutions for the societal needs.


M/S K.S.R.
INSTITUE
FOR
ENGINEERING
AND
TECHOLOGY
Where future begins.