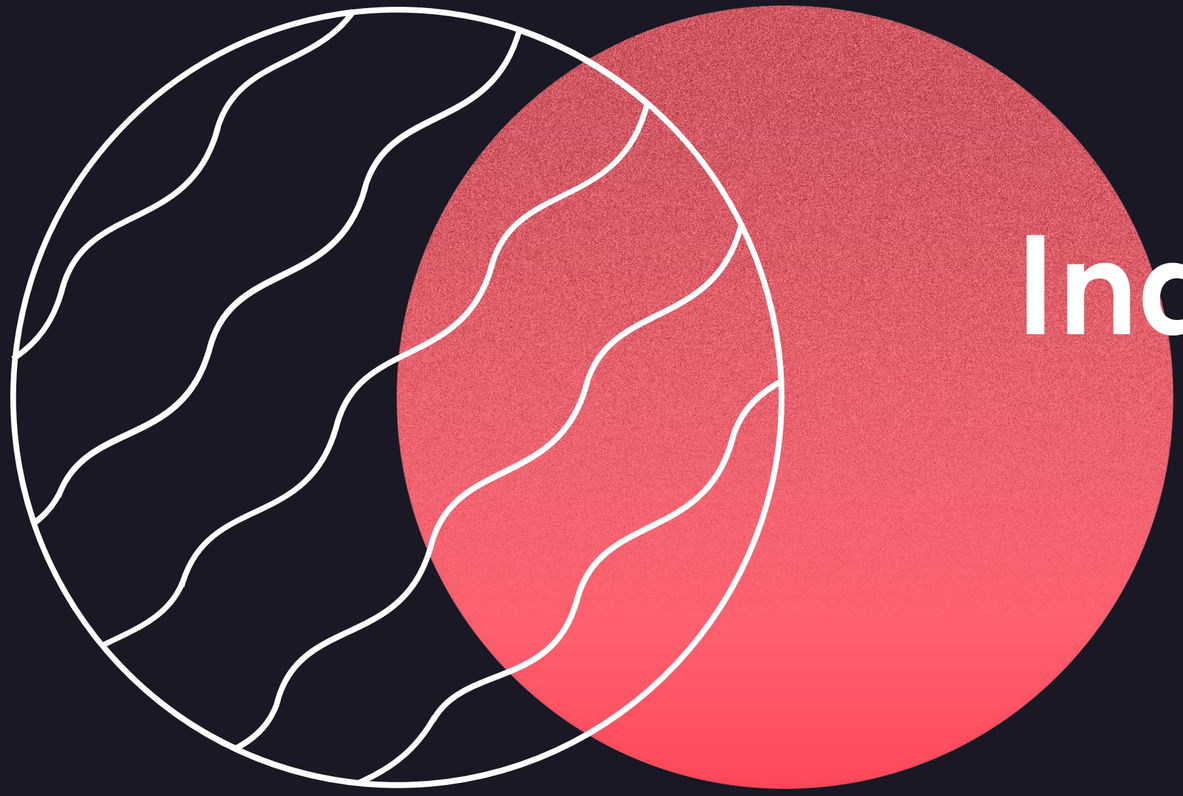


CSE1902
Industrial Internship
(Online Course)

Fall Semester 2021-22

by
Sharadindu Adhikari
19BCE2105

Course approved by CDC, VIT Vellore
VIVA-VOCE Presentation

Big Data (BDA) Foundation

Course by Digital Vidya & SSC NASSCOM

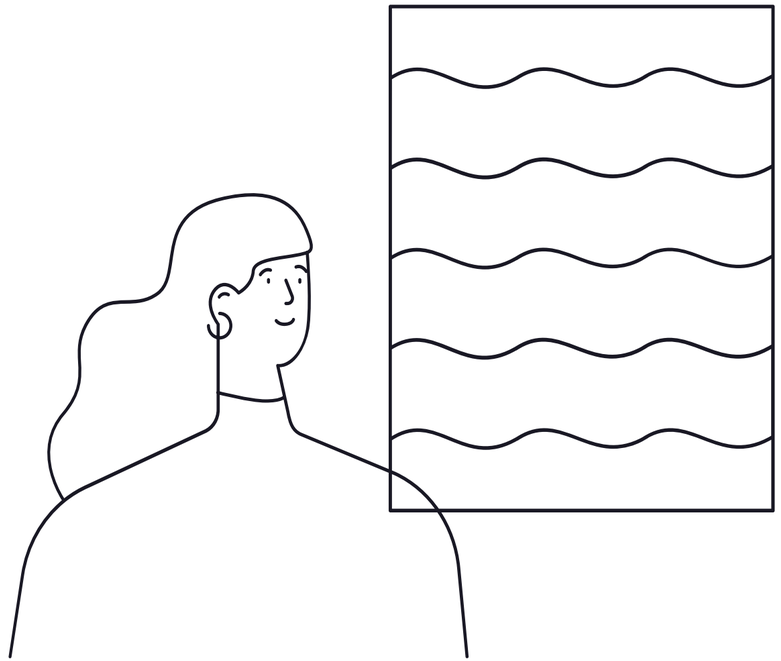
Course link: https://learn.futureskillsprime.in/pathways/ECL-1e02f39d-8c52-4cc3-8648-4243e1921506

Total duration of course lectures:	55 hours
Final exam duration:	60 minutes
Assignment hours:	12
Total number of assignments:	3
Total number of modules:	3
No. of attempts allowed:	1

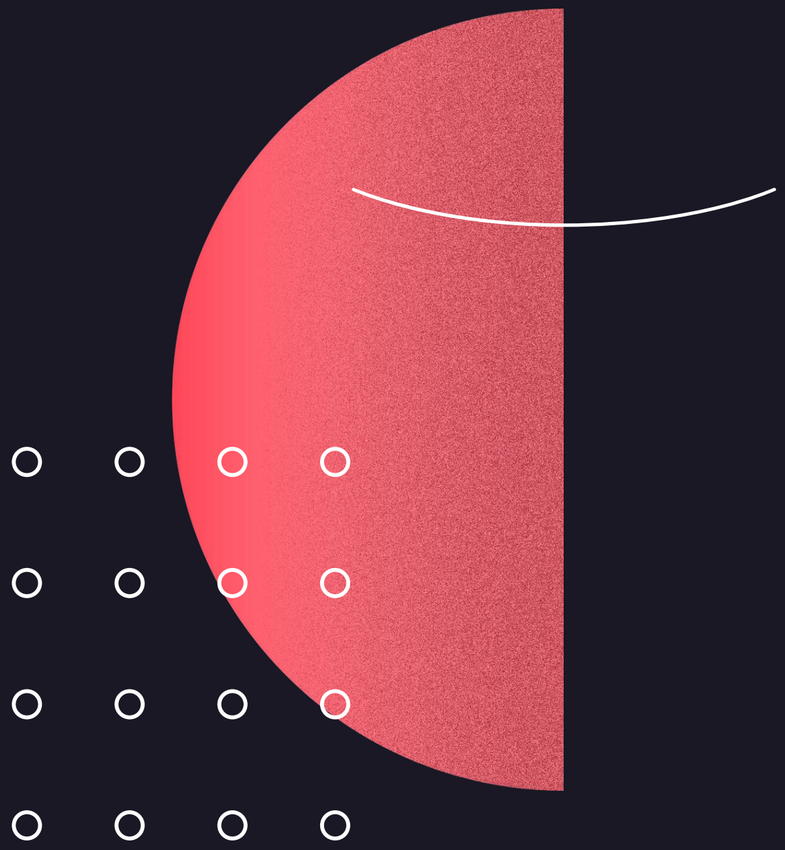
Introduction

The Big Data Foundation course was designed to establish an understanding of Big Data Analytics, Visualization, Data Processing and Management along with the knowledge of different Big Data platforms and their fundamentals. Foundational Curriculum for Big Data Analytics was aimed at upskilling those who have a basic understanding of programming and data sequences, to help them expand their knowledge and learn the fundamentals of Big Data Analytics technologies at a beginner level.

This Curriculum had been divided into three modules, of which the first is an introductory module, the second module focusses on big data fundamentals & tools and in the third module we had learned how data is processed and managed.



Abstract



The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of “Big Data,” including the recent announcement from the White House about new funding initiatives across different agencies, that target research for Big Data. While the promise of Big Data is real – for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 – there is no clear consensus on what is Big Data.

In fact, there have been many controversial statements about Big Data, such as “Size is the only thing that matters.” In this panel we will try to explore the controversies and debunk the myths surrounding Big Data.

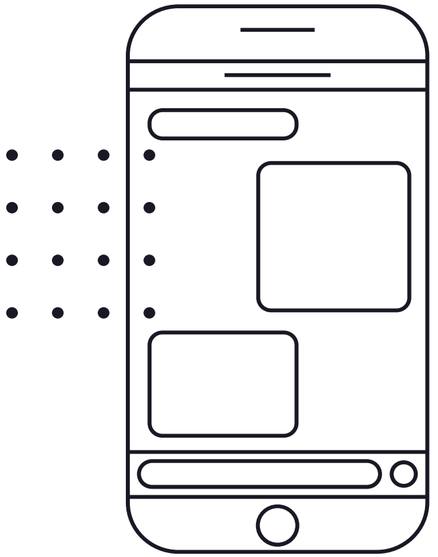
AIM

In this competitive business world, the benefits of Big Data cannot be underestimated. There are endless services offered by Big Data to the current market. If exploited properly, it can lead to substantial results. Almost every company is now moving towards Big Data Analytics due to numerous reasons. It is helping them in enhancing the overall growth of the organization.

I already had an idea about this from the couple of science forums I'm a part of. In order to acquire deeper understanding of the nuances, I'd set several aims in mind before approaching the course. I've discussed some of them below:

1. To develop better decision-making skills.

The main benefit of using Big Data Analytics is that it has boosted the decision-making process to a great extent. Rather than anonymously making decisions, companies are considering Big Data Analytics before concluding to any decision. A variety of customer-centric factors like what the customers want, the solution to their problems, analysing their needs according to the market trends, etc. are taken into account for a better decision-making process.



To understand how big data helps in better decision-making process, I've studied 2 different case studies: big data taking a gamble, and how it makes customers the king.

2. To understand the role of big data in greater innovations.

Innovations are crucial for the success of any organization. And to innovate, we need data, more and more data as it gives us the freedom to achieve the unthinkable. Big Data Analytics is used by various firms to create new products and services for their customers. Companies through Big Data, analyse different customers' opinions about their products and how their product is perceived. It gives them information about what they are lacking and what are the significant things to be kept in mind while developing any new product. This helps them in developing new products according to customer's requirement. Big Data Analytics gives the capabilities of thinking beyond the ordinary.

The case study I've read through regarding this discussed about IoT applications and Big Data's role in the invention of Driverless cars.

3. To learn about the influence of big data in education sector.

Big data benefits the education sector in managing the data related to students of an educational institute which is unmanageable. It is not used as it should be. Due to its huge size, it is hard for teachers to exploit it properly. Big Data Analytics has emerged as a boon to the education sector. It has started bringing the much-needed transformation in the education system and will surely take it to greater heights. Analysis of the capabilities of students based on the data can help teachers in nurturing their future in a better. Teachers are now aware of the student's strengths and weaknesses and can guide them accordingly.

4. To optimise product prices using big data.

The advantage of Big Data for companies is that they are using it to optimize the price they charge their customers. The goal is to set the prices in such a way that profit is maximized. Through Big Data they analyse the prices that have yielded the maximum profits to them under various historic market conditions. The aim is that the customer should get value for his money. As far as customers think that way the company will always keep growing. But to make a customer satisfied always, the company needs to make appropriate advancements in the product according to the trends in the market and Big Data facilitate them to do so.

5. Use-cases of big data in recommendation engines.

Imagine being able to have recommendations based on our previous as well as current choices made on various online platforms. Life is much easier when we have the option of choosing from the things we like. This is something that has changed the thinking of people towards various online platforms. They are now more comfortable being on these platforms.

The best example of a Big Data recommendation engine is that of various online shopping platforms. They analyse every customer's data and then recommend them accordingly. These recommendations are majorly based on the activities the customer did when he last visited the platform and his real-time activities. Also, suggestions are made to them based on a comparison between the customers who searched or bought familiar stuff. This is how online platforms have broken the physical barriers between them and their customers. Hasn't recommendation engines transformed the online shopping experience? It surely has.

6. To learn the roles of big data in life-saving application in healthcare industry.

The advent of Big Data Analytics has offered numerous benefits to the Healthcare Industry. It can be regarded as a Revolution in the Making. According to the Big Data Experts at QUANTZIG, "Big Data and Advanced Analytics may just be the answer to the hardest of Healthcare challenges". Big Data in healthcare would help practitioners to provide advanced and quality healthcare to their patients based on the electronic health records of the patient. It enhances the overall operational efficiency of the healthcare companies and has allowed them to make the required changes. Big Data Analytics would allow them to find a better cure for a disease by recognizing unknown connections and hidden patterns. Even a cure for a disease like cancer can be made possible by it.

Motivation

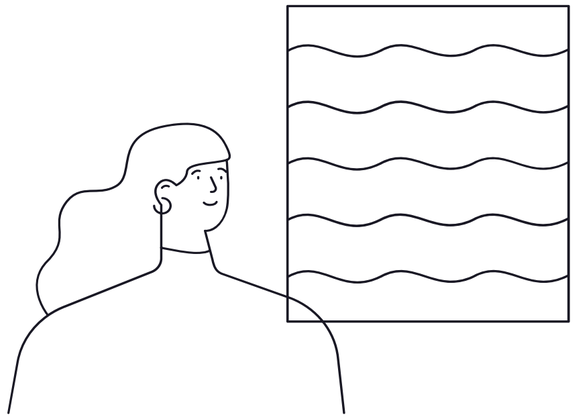
- As a person new to data science and interested in understanding why the Big Data Era has come to be, my primary motivation was to experience how one can perform predictive modelling and leverage graph analytics to model problems.

- And also to become conversant with the terminology and the core concepts behind big data problems, applications, and systems. To understand the dynamics and ask the right questions about data, communicate effectively with data scientists, and do basic exploration of large, complex datasets.

- To learn about how Big Data might be useful in my future business and career. And experience various data genres and management tools appropriate for each.

- To describe the reasons behind the evolving plethora of new big data platforms from the perspective of big data management systems and analytical tools. And become familiar with techniques using real-time and semi-structured data examples.

- And finally to learn techniques to extract value from existing untapped data sources and discovering new data sources.



Hardware Requirements

(A) Quad Core Processor (VT-x or AMD-V support recommended), 64-bit;

(B) 8 GB RAM;

(C) 20 GB disk free.

Software Requirements

This course relied on several open-source software tools, including Apache Hadoop. All required software had been downloaded and installed free of charge.

- Windows 7+
- Mac OS X 10.10+
- Ubuntu 14.04+
- CentOS 6+
- VirtualBox 5+

Background

The term Big Data was coined by Roger Mougalas back in 2005. However, the application of big data and the quest to understand the available data is something that has been in existence for a long time. As a matter of fact, some of the earliest records of the application of data to analyse and control business activities date as far back as 7,000 years.

In the same year (2005), Yahoo created the now open-source Hadoop with the intention of indexing the entire World Wide Web. Today, Hadoop is used by millions of businesses to go through colossal amounts of data.

1. Big Data phase 1.0

Data analysis, data analytics and Big Data originate from the longstanding domain of database management. It relies heavily on the storage, extraction, and optimization techniques that are common in data that is stored in Relational Database Management Systems (RDBMS). Database management and data warehousing are considered the core components of Big Data Phase 1.

2. Big Data phase 2.0

Since the early 2000s, the Internet and the Web began to offer unique data collections and data analysis opportunities. With the expansion of web traffic and online stores, companies such as Yahoo, Amazon and eBay started to analyse customer behaviour by analysing click-rates, IP-specific location data and search logs. This opened a whole new world of possibilities. From a data analysis, data analytics, and Big Data point of view, HTTP-based web traffic introduced a massive increase in semi-structured and unstructured data.

Background contd.

3. Big Data phase 3.0

Although web-based unstructured content is still the main focus for many organizations in data analysis, data analytics, and big data, the current possibilities to retrieve valuable information are emerging out of mobile devices.

BIG DATA PHASE 1	BIG DATA PHASE 2	BIG DATA PHASE 3
Period: 1970-2000	Period: 2000-2010	Period: 2010-present
DBMS-based, structured content: • RDBMS & data warehousing • Extract Transfer Load • Online Analytical Processing • Dashboards & scorecards • Data mining & statistical analysis	Web-based, unstructured content • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis	Mobile and sensor-based content • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization • Human-Computer-Interaction

Although it seems like big data has been around for a long time now and that we are getting closer to the pinnacle, big data may just be at its formidable stages. Big data in the near future may end up making big data now seem like a poultry amount.

Related work

The [1] study starts with an endeavour to comprehend contemporary choice emotionally supportive network which is a specific region of data frameworks with an emphasis on enhancing the choice making. Ackoff [2] contemplated that the objective of the administration data frameworks (MIS) was to make the data accessible to administrators for choice making purposes. Sadly, just few MIS were effective as the IT experts of the time did not comprehend the way of administrative work.

Alavi et al [3] thought that the MIS frameworks created were huge and unbendable and the reports produced were hard to fathom. Dearden [4] compressed the methodology identified with choice emotionally supportive networks. Gorry et al [5] authored the expression "choice emotionally supportive networks" in their paper and developed a structure for enhancing administration data frameworks utilizing Anthony's [6] classes of administrative movement and Simon's [7] scientific categorization of choice sorts. Sharp et al [8] contracted the definition to semiorganized administrative choices and proposed an extension that is pertinent today. There are the a lot of complex unstructured information accessible on web, consequently there is developing excitement for the idea of Big Data. Herbert A.

Simon [9] had proposed a behavioural model of level-headed decision which is utilized as a part of different commercial enterprises for choice making. The model was utilized with the contemporary choice making frameworks however with the blast of information in advanced time; huge information is the significant information supplier for the same.

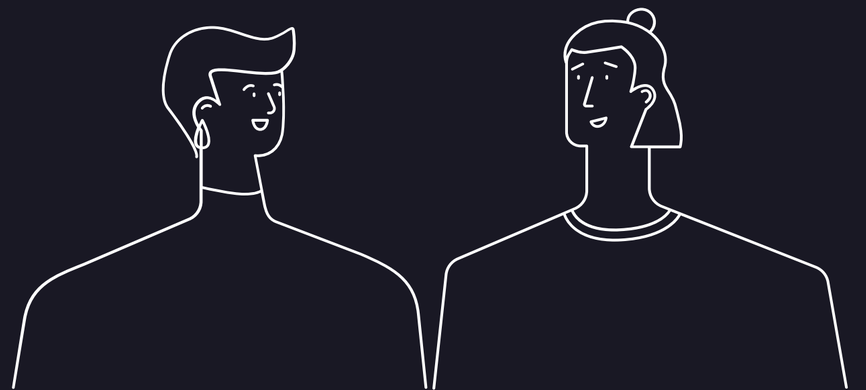
It has altered the exploratory examination, stargazing, training; human services and so forth [10-13]. McKinsey gauges [14] a funds of 300 billion dollars consistently in the only us by applying huge information idea. Such Big Data examination now drives almost every part of our advanced society, including portable administrations, retail, producing, money related administrations, life sciences, and physical sciences [15].

Description of Study and Gap Analysis

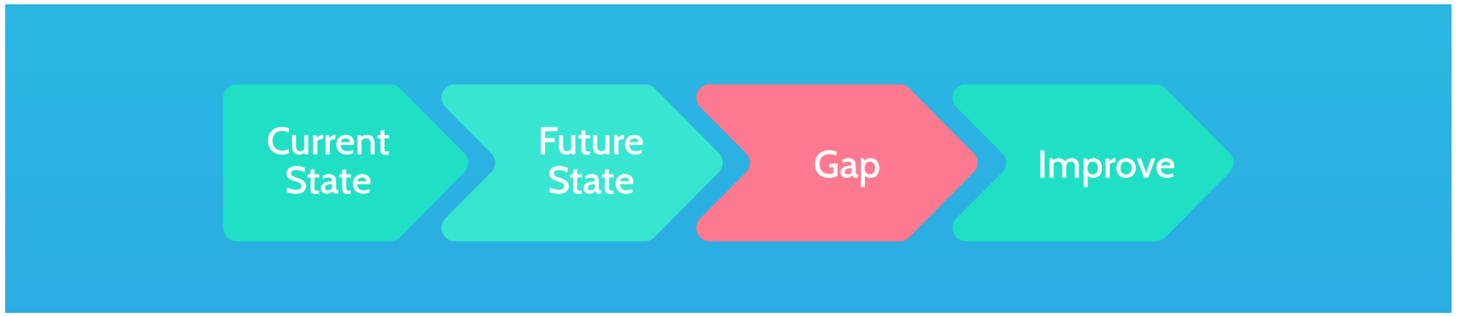
- As a data science enthusiast, the goal that I'd in mind while I picked up this course was to become conversant with the terminology and the core concepts behind big data problems, applications, and systems. And also, to learn one of the most common frameworks, Hadoop, that has made big data analysis easier and more accessible -- increasing the potential for data to transform our world.

- To learn about overcoming the problems with traditional system of grid computing. The need for multi-core distribution and how fallacies of Moore's law is limiting CPU speeds. And to learn more about distributed parallel processing: how a group of independent and geographically dispersed computer systems take part to solve a complex problem.

- And to be able to attain a level of understanding that would be helpful for my career going forward.



GAP ANALYSIS

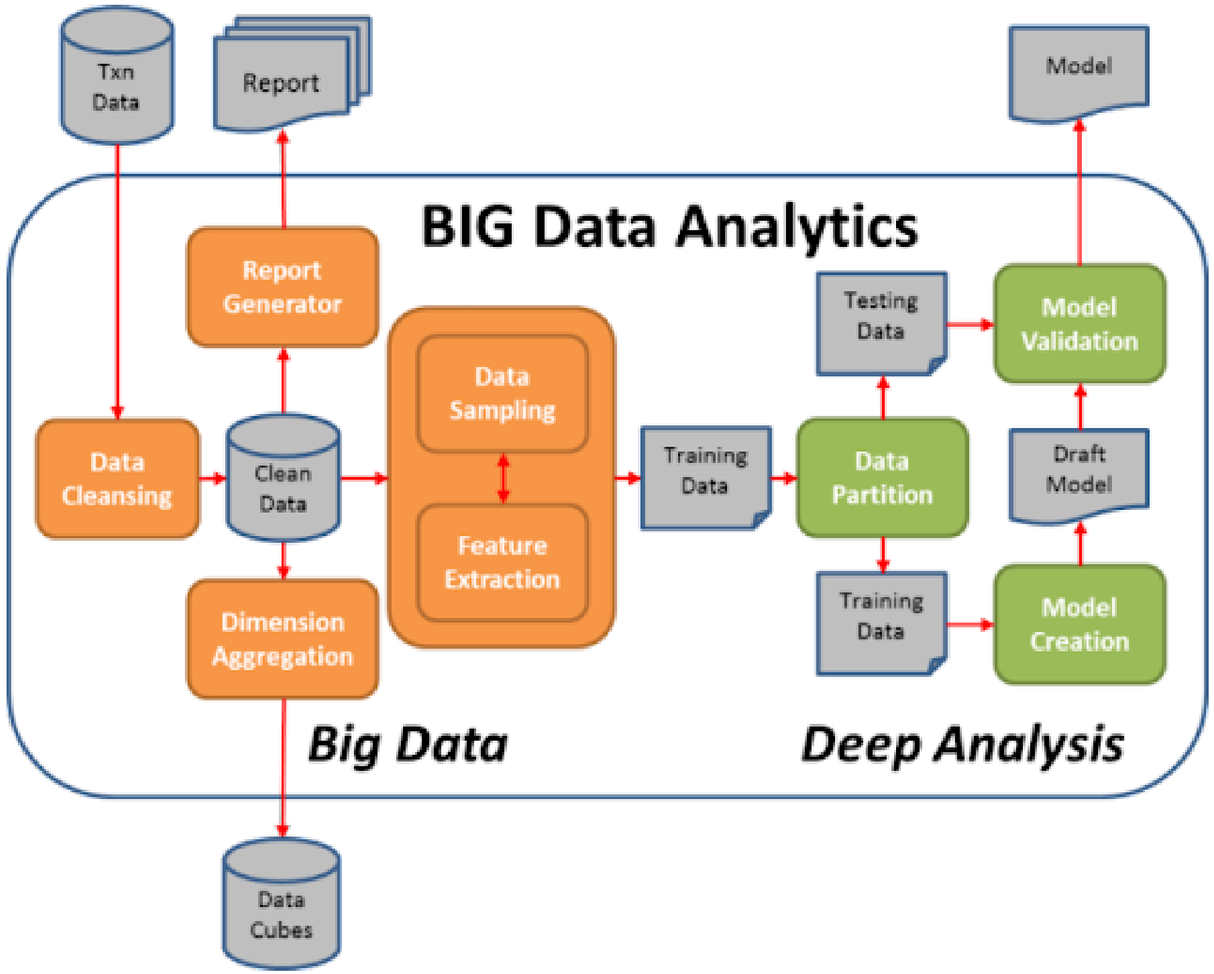


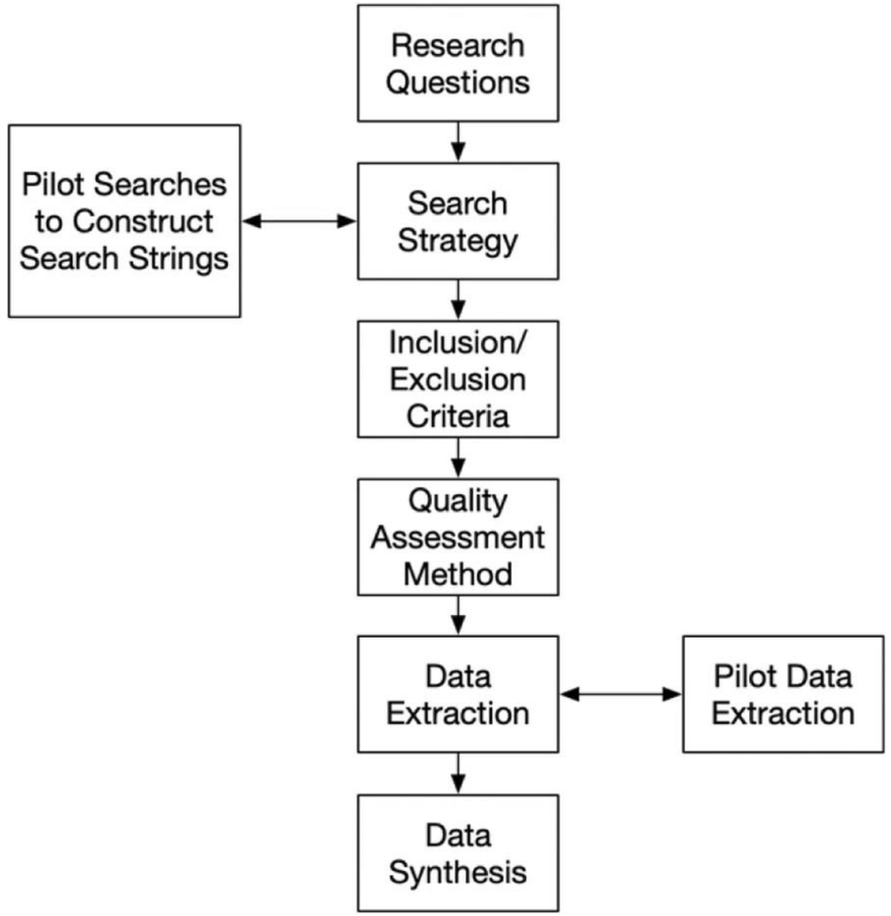
Week	Gaps
Week 1: Introduction to Big Data Analytics.	I had to learn to identify use cases from the Analytics industry and how BDA can be used to transform those businesses. The first week thus was fairly easy and the course material was very goof.
Week 2: BD Fundamentals and Platforms — leveraging customer behavioural analytics	Since I had no previous experience, I'd to learn Hadoop & HDFS from scratch. And perform a couple of tasks: like running Map and Reduce codes, performing data storage & retrieval operations, and processing batch operations. The assignment was out of scope and concepts took time to understand.
Week 3: BD Processing, Management and Analytics	This was the most challenging week. Questions were difficult and handling base cases and merging solutions was hard and time taking. On a sample WordCounter example, I had to perform Hadoop operations and a couple of processes on a provided COVID-19 dataset.

Gap Analysis contd.

Week 4: Big Data Business Model	None really. Had to learn about different business opportunities, maturity index, data transformations, and analysing different models of the big data industry. It was not that big of a deal, though an eye opener nonetheless.
Week 5: BDA Finale	Had to learn for the final quiz and concluding exam. Comparatively difficult problems that needed more intuition. Have to improve coding speed for these types of problems.

System Design Diag.



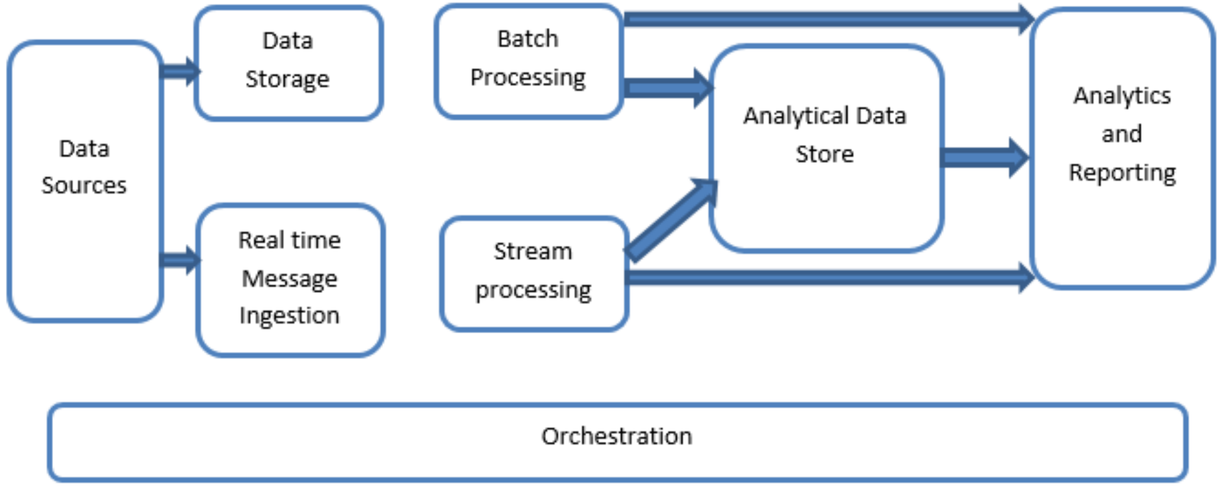


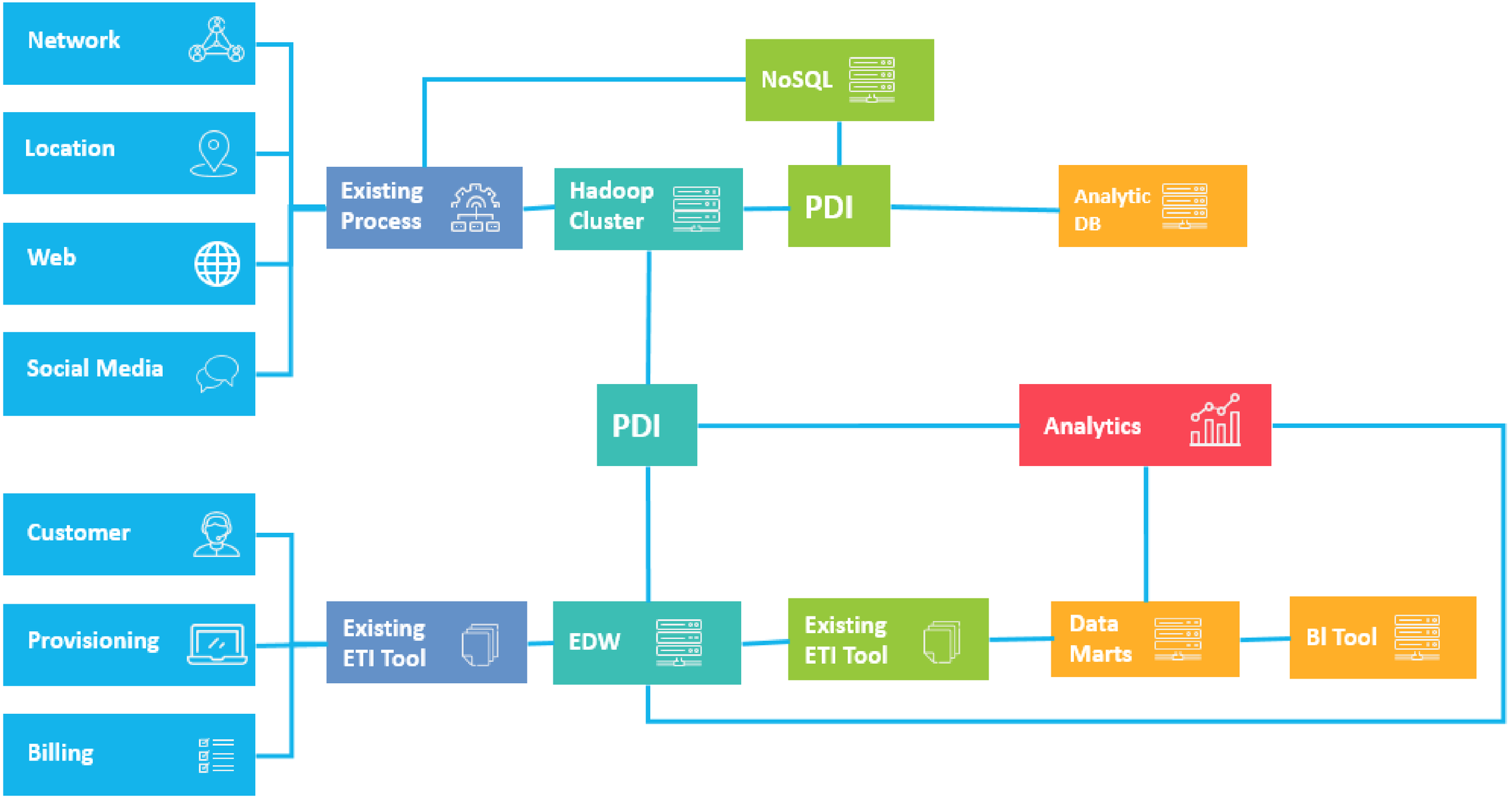
Architecture of work

Big data architecture can handle the processing, ingestion, and analysis of data that is too complex or large for traditional database systems. It is the overarching system used to manage large amounts of data to be analysed for business purposes, steer data analytics, and provide an environment in which big data analytics tools can extract vital business information, moreover big data architecture framework serves as a reference blueprint for big data infrastructures and solutions.

Big data solutions typically involve one or more of the following types of workload:

- Batch processing of big data sources at rest.
- Real-time processing of big data in motion.
- Interactive exploration of big data.
- Predictive analytics and machine learning.




Network
Location
Web
Social Media
Existing Process
NoSQL
Hadoop Cluster
PDI
Analytic DB
PDI
Analytics
Customer
Provisioning
Billing
Existing ETI Tool
EDW
Existing ETI Tool
Data Marts
BI Tool

Architecture contd.

Most big data architectures include some or all of the following components:

- Data sources: All big data solutions start with one or more data sources. Examples include:
  - Application data stores, such as relational databases.
  - Static files produced by applications, such as web server log files.
  - Real-time data sources, such as IoT devices.

- Data storage: Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a *data lake*. Options for implementing this storage include Azure Data Lake Store or blob containers in Azure Storage.

- Batch processing: Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to new files. Options include running U-SQL jobs in Azure Data Lake Analytics, using Hive, Pig, or custom Map/Reduce jobs in an HDInsight Hadoop cluster, or using Java, Scala, or Python programs in an HDInsight Spark cluster.

- Real-time message ingestion: If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing. This might be a simple data store, where incoming messages are dropped into a folder for processing. However, many solutions need a message ingestion store to act as a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics. Options include Azure Event Hubs, Azure IoT Hubs, and Kafka.

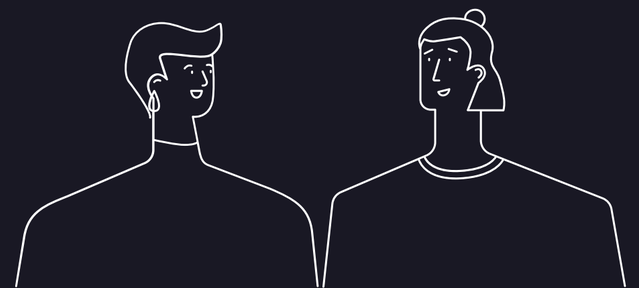
Architecture contd.

- Stream processing: After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. The processed stream data is then written to an output sink. Azure Stream Analytics provides a managed stream processing service based on perpetually running SQL queries that operate on unbounded streams. You can also use open source Apache streaming technologies like Storm and Spark Streaming in an HDInsight cluster.

- Analytical data store: Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools. The analytical data store used to serve these queries can be a Kimball-style relational data warehouse, as seen in most traditional business intelligence (BI) solutions.

- Analysis and reporting: The goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower users to analyse the data, the architecture may include a data modelling layer, such as a multidimensional OLAP cube or tabular data model in Azure Analysis Services. It might also support self-service BI, using the modelling and visualization technologies in Microsoft Power BI or Microsoft Excel.

- Orchestration: Most big data solutions consist of repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report or dashboard. To automate these workflows, you can use an orchestration technology such Azure Data Factory or Apache Oozie and Sqoop.



TECHNOLOGIES

Big Data Technologies are broadly classified into two categories.

1. Operational Big Data Technologies

Operational Big Data Technologies indicates the volume of data generated every day, such as online transactions, social media or any information from a particular company used for analysis by software based on big data technology. It acts as raw data to supply big data analysis technology. Few cases of Operational Big Data Technologies include information on MNC management, Amazon, Flipkart, Walmart, online ticketing for movies, flights, railways and more.

2. Analytical Big Data Technologies

Analytical Big Data Technologies concerns the advanced adjustment of Big Data Technologies, which is rather complicated than Operational Big Data. This category includes the real analysis of Big Data, which is essential to business decisions. Some examples in this area include stock marketing, weather forecasting, time series and medical records analysis.

Top 5 Current Big Data Technologies

1. Hadoop Ecosystem

Hadoop Framework was developed to store and process data with a simple programming model in a distributed data processing environment. The data present on different high-speed and low-expense machines can be stored and analyzed. Companies that have not explored Hadoop so far will most likely see its advantages and applications.

2. Artificial Intelligence

Artificial Intelligence is a broad bandwidth of computer technology that deals with the development of intelligent machines capable of carrying out different tasks typically requiring human intelligence. AI is revolutionizing the existing Big Data Technologies.

3. NoSQL Database

NoSQL includes a wide variety of different Big Data Technologies in the database, which are developed to design modern applications. It shows a non-SQL or non-relational database providing a method for data acquisition and recovery. They are used in Web and Big Data Analytics in real-time.

4. R Programming

R is one of the open-source Big Data Technologies and programming languages. The free software is widely used for statistical computing, visualization, unified development environments such as Eclipse and Visual Studio assistance communication. According to experts, it has been the world's leading language. The system is also widely used by data miners and statisticians to develop statistical software and mainly data analysis.

5. Data Lakes

Data Lakes means a consolidated repository for storage of all data formats at all levels in terms of structural and unstructured data. Data can be saved during Data accumulation as is without being transformed into structured data. It enables performing numerous types of Data analysis from dashboards and Data visualization to Big Data transformation in real-time for better business interference. Businesses that use Data Lakes stay ahead in the game from their competitors and carry out new analytics, such as Machine Learning, through new log file sources, data from social media and click-streaming.

EMERGING BIG DATA TECHNOLOGIES:

1. TensorFlow. It has a robust, scalable ecosystem of resources, tools, and libraries for researchers, allowing them to create and deploy powerful Machine Learning applications quickly.

2. Beam. Apache Beam offers a compact API layout to create sophisticated Parallel Data Processing pipelines through various Execution Engines or Runners. Apache Software Foundation developed these tools for Big Data in the year 2016.

3. Docker. It is one of the tools for Big Data that makes the development, deployment and running of container applications simpler. Containers help developers stack an application with all of the components they need, such as libraries and other dependencies.

4. Airflow. Apache Airflow is a Process Management and Scheduling System for the management of data pipelines. Airflow utilizes job workflows made up of DAGs (Directed Acyclic Graphs) tasks. The code description of workflows makes it easy to manage, validate and version a large amount of Data.

5. Kubernetes. It is one of the open-source tools for Big Data developed by Google for vendor-agnostic cluster and container management. It offers a platform for the automation, deployment, escalation and execution of container systems through host clusters.

6. Blockchain. It is the Big Data technology that carries a unique data safe feature in the digital Bitcoin currency so that it is not deleted or modified after the fact is written. It’s a highly secured environment and an outstanding option for numerous Big Data applications in various industries like baking, finance, insurance, medical and retail, to name a few.

To summarize, Big Data is still very much rising with more adoptions and more applications of existing Big Data technologies and the launch of newer solutions related to Big Data security, Cloud integrations, data mining etc.

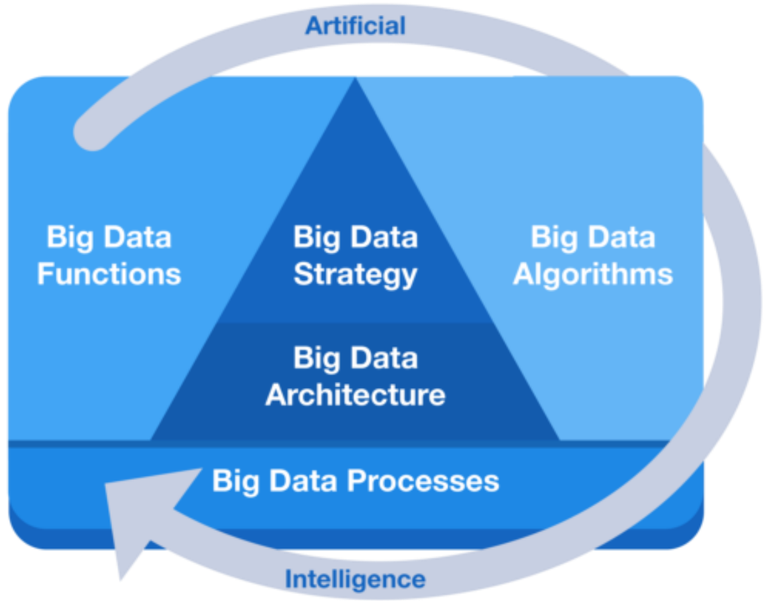
FRAMEWORK

The core objective of the Big Data Framework is to provide a structure for enterprise organisations that aim to benefit from the potential of Big Data. In order to achieve long-term success, Big Data is more than just the combination of skilled people and technology – it requires structure and capabilities.

The Big Data framework is a structured approach that consists of six core capabilities that organisations need to take into consideration when setting up their Big Data organization. The Big Data Framework consists of the following six main elements:

1. Big Data Strategy

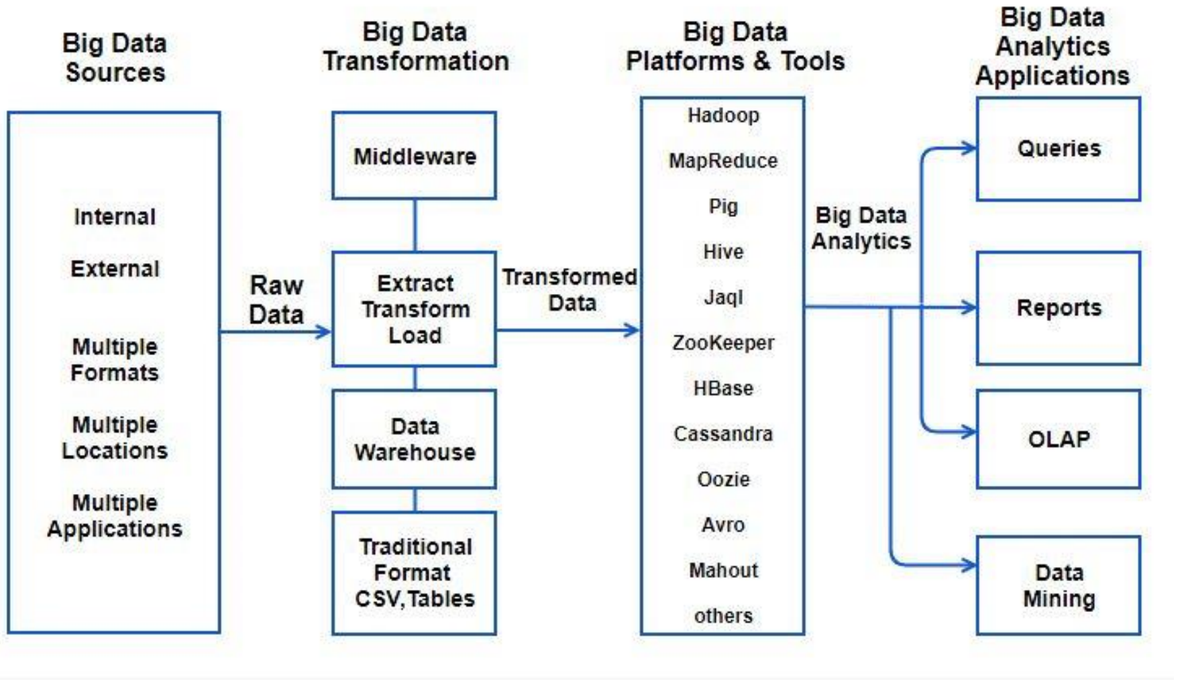
Data has become a strategic asset for most organisations. The capability to analyse large data sets and discern pattern in the data can provide organisations with a competitive advantage. Netflix, for example, looks at user behaviour in deciding what movies or series to produce. Alibaba, the Chinese sourcing platform, became one of the global giants by identifying which suppliers to loan money and recommend on their platform. Big Data has become Big Business.



2. Big Data Architecture

In order to work with massive data sets, organisations should have the capabilities to store and process large quantities of data. In order to achieve this, the enterprise should have the underlying IT infrastructure to facilitate Big Data. Enterprises should therefore have a comprehensive Big Data architecture to facilitate Big Data analysis

Framework contd.



3. Big Data Algorithms

A fundamental capability of working with data is to have a thorough understanding of statistics and algorithms. Big Data professionals therefore need to have a solid background in statistics and algorithms to deduct insights from data. Algorithms are unambiguous specifications of how to solve a class of problems. Algorithms can perform calculations, data processing and automated reasoning tasks. By applying algorithms to large volumes of data, valuable knowledge and insights can be obtained.

4. Big Data Processes

In order to make Big Data successful in enterprise organization, it is necessary to consider more than just the skills and technology. Processes can help enterprises to focus their direction. Processes bring structure, measurable steps and can be effectively managed on a day-to-day basis. Analysis becomes less dependent on individuals and thereby, greatly enhancing the chances of capturing value in the long term.

5. Big Data Functions

Big Data functions are concerned with the organisational aspects of managing Big Data in enterprises. This element of the Big Data framework addresses how organisations can structure themselves to set up Big Data roles and discusses roles and responsibilities in Big Data organisations. Organisational culture, organisational structures and job roles have a large impact on the success of Big Data initiatives.

CERTIFICATE

Certificate link: https://fsp-assessment-certificates.s3-ap-southeast-1.amazonaws.com/SharadinduAdhikari-64971411.pdf

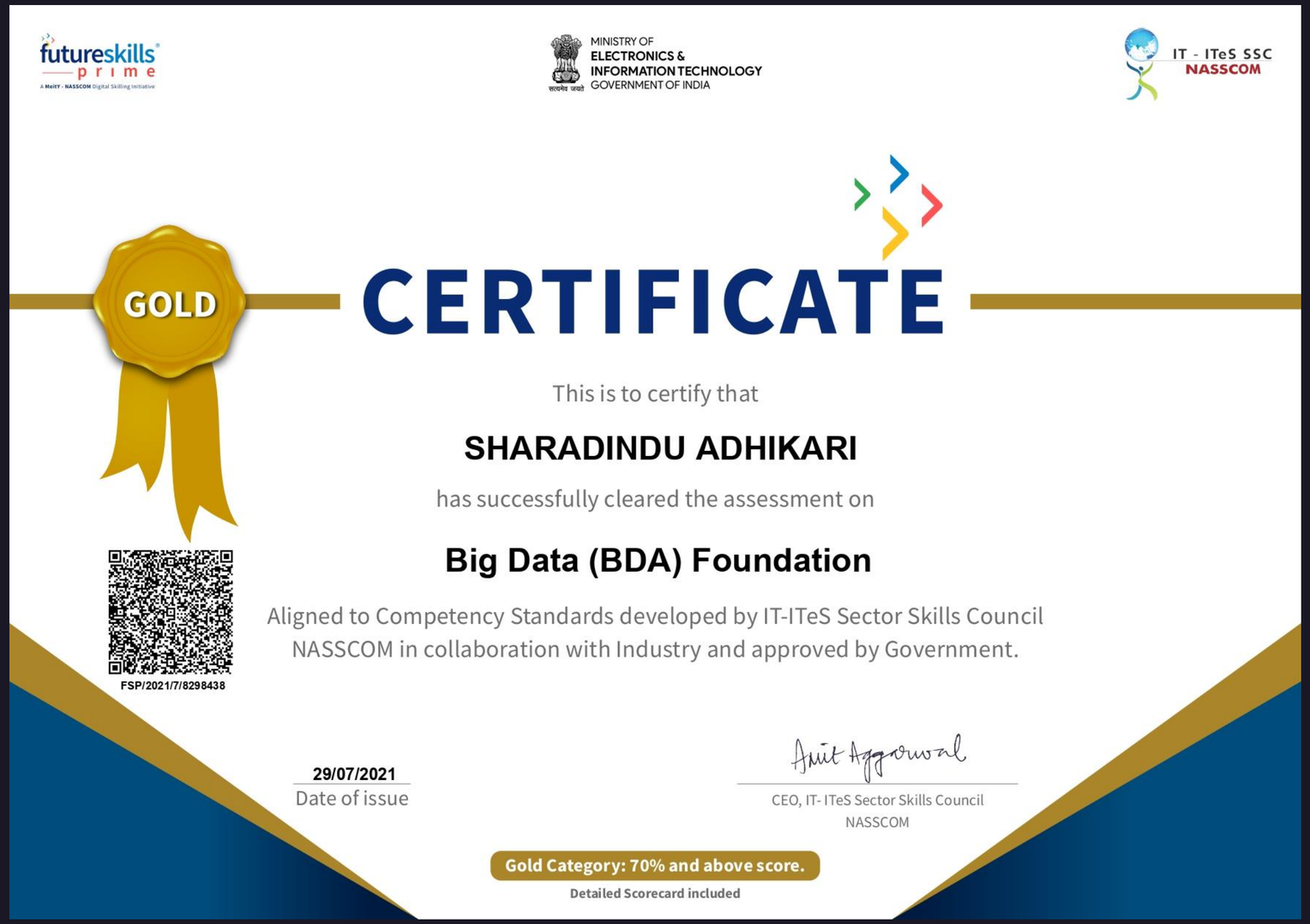
Course name: Big Data (BDA) Foundation

Course by: Digital Vidya & SSC Nasscom

Marks obtained: 80.00

Duration: 33 days

From 27th June 2021
To 29th July 2021



futureskills prime
A MeitY - NASSCOM Digital Skilling Initiative

MINISTRY OF
ELECTRONICS &
INFORMATION TECHNOLOGY
GOVERNMENT OF INDIA

IT - ITeS SSC
NASSCOM

Certificate Details

Candidate Name	Sharadindu Adhikari
Assessment/Course Name	Big Data (BDA) Foundation
Date of Issue	29/07/2021
Certification ID	FSP/2021/7/8298438
Category **Gold** >=70% / **Silver** 60%-69% / **Bronze** 50%-59%	Gold

FSP/2021/7/8298438

Assessment Score

Module Name/NOS ID	NSQF Level	Maximum Marks	Marks Obtained	Percentage
M001	NA	12.00	12.00	100.00
M002	NA	35.00	31.00	88.57
M003	NA	53.00	37.00	69.81
Total		100.00	80.00	80.00

Digital Vidya

Certificate of Completion

This is to certify that

Sharadindu Adhikari

has successfully completed

Big Data Foundation Course

Anuj Batra

Anuj Batra
CEO, Digital Vidya

Conclusion

The enormous blast of gadgets and information in the following couple of years will make the Big Data has one of the speediest territories of development for IT Industry (CAGR of 27-45% every year evaluated). As Services industry develops in the utilizing Big Data Analytics there will be parcel of Services and Products which are perfectly customized to utilization design comprehended utilizing Big Data investigation which as a part of turn will be utilized to drive and ad lib the conveyance cycle. This will leady to new Global Delivery models which will incorporate new advances and conveyance hubs that are disseminated universally driving development and giving bits of knowledge into marvel which were unexplored with conventional frameworks.

The Service Industry will have isolation of building information for investigation by outside administration suppliers yet they will have it translated by inward assets of big business. We feel this is a pattern yet can affirm after some examination. The Big Data Analytics Service Providers will have noteworthy business chances to assemble expansive datasets and determine surmising prompting associations ready to modify and offer administrations and items which will have the capacity to adapt to new requests of business for the Service business.

Scope for Future Work

Based on the analysis of the current literature, the following research directions should attract more attention in the future:

1. Distributed computation. Big data analytics need to find solutions to adapt to the real-time requirements of distribution applications. Currently, centralized computation has been widely used in big data analytics, which requires lots of additional resources to transfer and integrate data. At this point, distributed computation of big data may be a feasible way to meet the real-time requirements (Kambatla, Kollias, Kumar, & Grama, 2014);

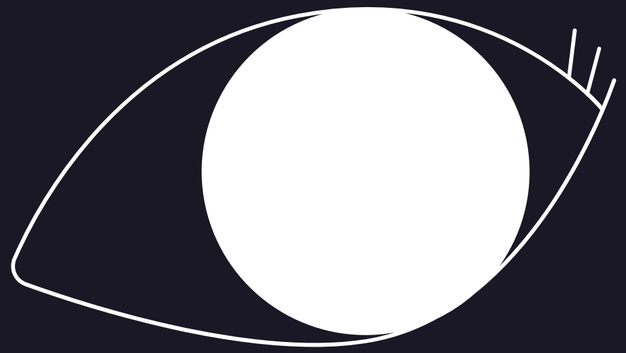
2. Integration of multiple programming models. The programming model is a powerful tool to implement the application logics. However, current programming models only focus on specific tasks. A single programming model may not be able to cope with the challenges of heterogeneous, distributed data sources in the future. A tighter integration of multiple programming models is required for future big data analytics. Apache YARN (Apache Hadoop NextGen MapReduce; Olston et al., 2003) is a pioneering example of this type of integration. As a cluster management technology, YARN enables the running of multiple programming models on the same cluster. This feature makes it easier for users to take advantage of cost-effective, linear-scale storage and processing;

3. Effective and standardized benchmarks. The research on big data benchmarks is still at inception. How to choose an efficient data set to represent the diversity and correlation of big data applications is still a big challenge for future benchmarks. More work needs to be done to facilitate a comparison of the performance of various frameworks;

4. Privacy and security. Traditional security mechanisms were designed to tackle the security problems of small-scale data. Some security issues of big data, such as secure computations in distributed programming frameworks and real-time security monitoring, have not been well studied in the existing literature. Future researchers need to identify security and privacy problems in the big data era and explore solutions to achieve a balance between data security and data processing efficiency.

REFERENCES

[1] http://strata.oreilly.com/2010/01/roger-magoulas-on-bigdata.html
[2] Ackoff, R.L. (1967) Management misinformation systems. Management Science, 14(4), 147-156.
[3] Alavi, M. and Carlson, P. (1992) A review of MIS research and disciplinary development. Journal of Management Information Systems, 8(4), 45-62.
[4] Dearden, J. (1972) MIS is a mirage. Harvard Business Review, 50(1), 90-99.
[5] Gorry, G.A. and Scott Morton, M.S. (1971) A framework for management information systems. Sloan Management Review, 13(1), 1-22.
[6] Anthony, R.N. (1965) Planning and Control Systems: A Framework for Analysis (Harvard University Graduate School of Business Administration, Cambridge, MA).
[7] Simon, H.A. (1977) The new science of management decision (rev. ed.). Englewood Cliffs, NJ: Prentice-Hall. (Original work published 1960).
[8] Keen and Scott Morton, Keen, P.G.W. and Scott Morton, M.S. (1978) Decision Support Systems: An Organizational Perspective. (Addison-Wesley, Reading, MA). [9] Herbert A. Simon, A Behavioral Model of Rational Choice, The Quarterly Journal of Economics, Vol. 69, No. 1. (Feb., 1955), pp. 99-118
[10] SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. Jan. 2008.
[11] Advancing Personalized Education. Computing Community Consortium. Spring 2011.
[12] Smart Health and Wellbeing. Computing Community Consortium. Spring 2011.
[13] A Sustainable Future. Computing Community Consortium. Summer 2011.
[14] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011
[15] Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the United States



Thank You!