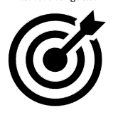
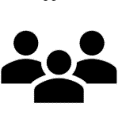
**IT PROF EL 6**

|  |  |  |
| --- | --- | --- |
| **Course** | DATA ANALYTICS | |
| **Sem/AY** | First Semester – | |
| **Module No.** | 1 and 2 | |
| **Lesson Title** | **Introduction to Big Data and Analytics** | **Big Data Analytics**  **Structured and Non-Structured**  **Business Intelligence, Data Warehouse and Hadoop** |
| **Week Duration** | **1-4** | |
| **Date** |  | |
| **Description of the Lesson** | **Big Data and Analytics**  This lesson will discuss the importance of Big Data and the different type of Analytics tools. The course will cover the types of Big Data. This also include the characteristics of Big Data.  **Business Intelligence, Data Warehouse and Hadoop**  This lesson will discuss difference of Business Intelligence, Data Warehouse and Hadoop. In this lesson it includes the technologies used to analyze data. this lesson will cover the components of Business Intelligence. This lesson discusses the Data Warehouse and designed to support the business decision and data consolidation using the ETL tools. | |

****

**Learning Outcomes**

|  |  |
| --- | --- |
| **Intended Learning Outcomes** | **Big Data and Analytics**  Students should be able to meet the following intended learning outcomes:   * Understand the uses Big Data in decision making * Characterize the types Big Data * Characterize types of Big Data   **Business Intelligence Data Warehouse and Hadoop**  Students should be able to meet the following intended learning outcomes:   * Understand the differences between Business Intelligence and Data Warehouse and Hadoop. * Understand the uses and benefits of Data Warehouse. * Understand the Hadoop Architecture. * Gain knowledge in File Transfer Protocol |
| **Targets/ Objectives** | **Big Data and Analytics**  At the end of the lesson, students should be able to:   * Identify what is Big Data and the purpose it Big Data in decision making. * Differentiate types Big Data * Distinguish the characteristics of Big Data * Identify the benefits of Big Data.   **Business Intelligence Data Warehouse and Hadoop**  At the end of the lesson, students should be able to:   * Describe the difference between Business Intelligence and Data Warehouse and Hadoop. * Differentiate the modern Data Warehouse * Describe the Hadoop architecture and framework. |



**Student Learning Strategies**

|  |  |
| --- | --- |
| **Online Activities (Synchronous/**  **Asynchronous)**  **Offline Activities**  **(e-Learning/Self-Paced)** | 1. Online Discussion via Google Meet   You will be directed to attend in an online class discussion on the Cognate/Course Track 3 (CPE 22). To have access on the Online Discussion, refer to the link that will be given by the instructor through Google Classroom  (For further instructions, refer to your Google Classroom and see the schedule of activities for this module).  These are the list of lessons provided on this SLM:  01 Big Data and Analytics  02 Business Intelligence, Data Warehouse and Hadoop   1. Learning Guide Questions:   **Big Data Analytics**   * What is Data and Big Data? * What are the types of Big Data? * What are the Characteristics of Big Data? * What are the importance of Big Data?   **Business Intelligence, Data Warehouse and Hadoop**   * What is Business Intelligence? * What are the differences between Business Intelligence and Data Warehouse and Hadoop? * What are the benefits of Data Warehouse? * What is Hadoop and its architecture? * What is ETL tools?   ***Note:*** *The insight that you will post on online discussion forum using Learning Management System (LMS) will receive additional scores in class participation.*  **Big Data Analytics: Lecture Guide**  **INTRODUCTION TO BIG DATA AND ANALYTICS: Lecture Guide**  **Data and Big Data**  is the data whose scale, diversity, and complexity require new architecture, re    **Figure 1: Big Data Landscape**  **History of Big Data**  Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.  Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.  The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it’s not just humans who are doing it.  With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.  While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.  **Benefits of Big Data and Data Analytics**   * Big data makes it possible for you to gain more complete answer because you more information. * More complete answer means more confidence in the data – which means a completely different approach to tacking problems.   **Types of Big Data**  **Structured**  Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.  **Unstructured**  Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.  **Structured vs Unstructured Data comparison**  **Figure 2: Structured Data vs. Unstructured Data**  **Human Generated Unstructured Data**   * Text files * Emails - e.g. including the visible “to”, “from”, “date / time”, “subject” entered to send an email) **Semi-Structured Data** * Social Media - (e.g. the number of “Likes”, “retweets”, “date”, “author” etc). * Websites * Mobile data – message, locations * Communication * Media * Business application   What is unstructured data?  **Figure 3: Types of Unstructured Data**  Machine-generated Unstructured Data   * **Satellite imagery** * **Scientific data** * **Digital surveillance**   **Semi-Structured**  Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data. Thus we come to the end of types of data.  Structured vs Unstructured vs Semi-structured data overview, differences and key characteristics  **Figure 4: Differences between Structured Data, Unstructured Data and Semi-Structured Data**  **V3_3_2Characteristics of Big Data 3V’s**   * **Variety** * **Velocity** * **Volume**   **Figure 5: Big Data 3V’s**   1. **Variety**   Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.   1. **Velocity**   Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.   1. **Volume**   Volume is one of the characteristics of big data. We already know that Big Data indicates huge ‘volumes’ of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data is stored in data warehouses. Thus comes to the end of characteristics of big data.  **Screen shot 2013-01-13 at 4.01.47 PM.png**  **Figure 6: Data Storage Growth (Volume)**  **The Importance of Big Data**   * Cost Saving * Time Reduction * Understand the market condition * Control online reputation * Using Big Data Analytics to Boost Customer Acquisition and Retention * Using Big Data Analytics to solve Advertisers problem and Offer Marketing Insights * Big Data Analytics as a Driver of Innovation and Product Development   **Business Intelligence**  Although Big Data and Business Intelligence are two technologies used to analyze data to help companies in the decision-making process, there are differences between both of them. They differ in the way they work as much as in the type of data they analyze.  Traditional BI methodology is based on the principle of grouping all business data into a central server. Typically, this data is analyzed in offline mode, after storing the information in an environment called **Data Warehouse.** The data is structured in a conventional relational database with an additional set of indexes and forms of access to the tables (multidimensional cubes).  Business Intelligence tools [20] | Download Scientific Diagram  **Figure 7: Business Intelligence**  These are the main differences between Big Data and Business Intelligence:   1. In a Big Data environment - information is stored on a distributed file system, rather than on a central server. 2. Big Data solution carry the processing function to the data, rather than the data to the functions. 3. Big Data can analyze data in different format, both structured and unstructured. 4. Data processed by Big Data solution can be historical or come from real-time sources. 5. Big Data technology used parallel mass processing (MPP).   What Is Business Intelligence? A Beginner&#39;s Guide in 2020  **Figure 8: Business Intelligence vs. Big Data**  **Data Warehouse**  A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. Data is populated into the Data Warehouse through the processes of extraction, transformation and loading (ETL tools). Data analysis tools, such as business intelligence software, access the data within the warehouse.  **Can Big Data replace an EDW?**  **Figure 9: Data Warehouse vs. Big Data**  Elements of Typical Data Warehouse   * Relational database * An extraction, loading, and transformation (ELT) solution for preparing the data for analysis * Statistical analysis, reporting, and data mining capabilities * Client analysis tools for visualizing and presenting data to business users * analytical applications that generate actionable information  (data science, AI , algorithm, or graph   Benefits of Data Warehouse   * Subject-oriented * Integrated * Nonvolatile   Benefits of Data Warehouse | Top 6 Useful Benefits of Data Warehouse  **Figure 10: Benefits of Data Warehouse**  **Modern Data Warehouse**   * Information Technology * Data Engineering * Business Analytic * Data Science Team   **Features of Modern Data Warehouse**   * A converged database that simplifies management of all data types and provides different ways to use data * self-service data ingestion and transformation services * support for SQL, machine learning, graph, and spatial processing * Automated management for simple provisioning, scaling, and administration   **HADOOP**  Hadoop is changing the perception of handling Big Data especially the unstructured data. Let’s know how Apache Hadoop software library, which is a framework, plays a vital role in handling Big Data. Apache Hadoop enables surplus data to be streamlined for any distributed processing system across clusters of computers using simple programming models. It truly is made to scale up from single servers to a large number of machines, each and every offering local computation, and storage space. Instead of depending on hardware to provide high-availability, the library itself is built to detect and handle breakdowns at the application layer, so providing an extremely available service along with a cluster of computers, as both versions might be vulnerable to failures.  **Hadoop Architecture**   1. Processing/Computation layer (MapReduce), and 2. Storage layer (Hadoop Distributed File System).   Hadoop Architecture  **Figure 10: Hadoop Architecture**  Apache Hadoop Architecture - HDFS, YARN &amp; MapReduce - TechVidvan  **Figure 11: Hadoop Framework**  **Hadoop Community Package Consist**   * File system and OS level abstraction * A MapReduce engine (either MapReduce or Yarn) * The Hadoop Distributed File System HDFS * Java Archive (JAR) files * Source code, documentation and a contribution section   **MapReduce**   * a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets) * large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.   **Hadoop Distributed Filed System (HDFS)**   * based on the Google File System (GFS) * provides a distributed file system that is designed to run on commodity hardware. * It provides high throughput access to application data and is suitable for applications having large datasets.   **Hadoop Framework**   1. **Hadoop Common-** Java libraries and utilities required by other Hadoop modules.   GitHub - apache/hadoop-common: Mirror of Apache Hadoop common  **Figure 12: Example of Hadoop Interface**   1. **Hadoop YARN-**  framework for job scheduling and cluster resource management.   What is Apache Hadoop YARN? - Intellipaat Blog  **Figure 13: Hadoop YARN Framework**  **How does Hadoop Work?**   1. Data is initially divided into directories and files. 2. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M). 3. files are then distributed across various cluster nodes for further processing 4. HDFS, being on top of the local file system, supervises the processing. 5. Blocks are replicated for handling hardware failure. 6. checking that the code was executed successfully. 7. Performing the sort that takes place between the map and reduce stages. 8. Sending the sorted data to a certain computer. 9. Writing the debugging logs for each job. |

 **Learning Resources**

|  |
| --- |
| **Book(s) Sources:**  Wiley, J. (2015). *Data Science and Big Data Analytics*.  **Electronic Sources:**  Retrieved from. https://libribook.com/ebook/4850/data-science-dummies-2nd-edition-pdf  Retrieved from. <https://www.snowflake.com/resource/cloud-data-analytics-for-dummies/>  Retrieved from. https://www.tutorialspoint.com/big\_data\_analytics/index.htm |