Lesson 1: Overview of Data Analysis

1.1. Introduction

The volume of data that one has to deal has exploded to unimaginable levels in the past decade, and at the same time, the price of data storage has systematically reduced. Private companies and research institutions capture terabytes of data about their users' interactions, business, social media, and also sensors from devices such as mobile phones and automobiles. The challenge of this era is to make sense of this sea of data. This is where big data analytics comes into picture.

Big Data Analytics largely involves collecting data from different sources, merge it in a way that it becomes available to be consumed by analysts and finally deliver data products useful to the organization business.

The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Big Data Analytics.

1.2. Definition of terms

There are a number of terms used in this unit which we need to familiarize ourselves with and some of which are described below:

1.2.1.Big Data

Big data is the collective name for the large amount of registered digital data and the equal growth thereof. The aim is to convert this stream of information into valuable information for the company.

1.2.2.Data Science

Data science is the process of deriving knowledge and insights from a huge and diverse set of data through organizing, processing and analyzing the data. It involves many different disciplines like mathematical and statistical modelling, extracting data from it source and applying data visualization techniques.

1.2.3.Data Analysis

Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.

1.2.4.Machine Learning

Machine Learning (ML) is that field of computer science with the help of which computer systems can provide sense to data in much the same way as human beings do.In simple words, ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method. The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.

1.2.5.Database Management Systems

A Database Management System (DBMS) is defined as the software system that allows users to define, create, maintain and control access to the database. A DBMS makes it possible for end users to create, read, update and delete data in database. It is a layer between programs and data.

1.2.6.Data warehouse

A data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization.

1.3. Characteristics of Big Data

Big data is characterized by the four V's. These V's stand for the four dimensions of Big Data: Volume, Velocity, Variety and Veracity.

1.3.1.Volume

Big Data is large in volume. It is estimated that we create 2.3 trillion gigabytes of data every day. And that will only increase. As the volume grows so rapidly, so does the need for new database management systems and IT employees. Millions of new IT jobs are expected to be created in the next few years to accommodate the Big Data flow.

1.3.2.Velocity

Velocity, or speed, refers to the enormous speed with which data is generated and processed. Until a few years ago, it took a while to process the right data and to surface the right information. Today, data is available in real time. This is not only a consequence of the speed of the internet, but also of the presence of Big Data itself. Because the more data we create, the more methods are needed to monitor all this data, and the more data is monitored. This creates a vicious circle.

1.3.3.Variety

The high speed and considerable volume are related to the variety of forms of data. Smart IT solutions are available today for all sectors, from the medical world to construction and business. Consider, for example, the electronic patient records in healthcare, which contribute to many trillions of gigabytes of data. When all parts of the world have the internet in the future, the volume and variety will only increase.

1.3.4.Veracity

How truthful Big Data is remains a difficult point. Data quickly becomes outdated and the information shared via the internet and social media does not necessarily have to be correct. Many managers and directors in the business community do not dare to make decisions based on Big Data. Data scientists and IT professionals have their hands full organizing and accessing the right data. It is very important that they find a good way to do this. Because if Big Data is organized and used in the right way, it can be of great value in our lives. From predicting business trends to preventing disease and crime

1.4. Types of Data analytics

There are four main types of big data analytics: diagnostic, descriptive, prescriptive, and predictive analytics. They use various tools for processes such as data mining, cleaning, integration, visualization, and many others, to improve the process of analyzing data and ensuring the company benefits from the data they gather.

1.4.1.Diagnostic analytics

Diagnostic analytics is one of the more advanced types of big data analytics that you can use to investigate data and content. Through this type of analytics, you use the insight gained to answer the question, "Why did it happen?" So, by analyzing data, you can comprehend the reasons for certain behaviors and events related to the company you work for, their customers, employees, products, and more.

Let's say there has been a drastic change in a product's sale even though you have not made any marketing changes to it. You would use diagnostic analytics to identify this anomaly and find the causal

relationship for such a change. Some tools and techniques used for such a task include: searching for patterns in the data sets, filtering the data, using probability theory, regression analysis, and more.

1.4.2.Descriptive analytics

Descriptive analytics is one of the most common forms of analytics that companies use to stay updated on current trends and the company's operational performances. It is one of the first steps of analyzing raw data by performing simple mathematical operations and producing statements about samples and measurements. After you identify trends and insight with descriptive analytics, you can use the other types of analytics to learn more about what causes those trends.

You will need to use descriptive analytics when dealing with finance, production, and sales. Some tasks that require this type of analytics include the production of financial reports and metrics, surveys, social media initiatives, and other business-related assignments.

1.4.3.Prescriptive analytics

Prescriptive analytics takes the results from descriptive and predictive analysis and finds solutions for optimizing business practices through various simulations and techniques. It uses the insight from data to suggest what the best step forward would be for the company.

Google is one of the many companies that use this type of analytics. They made use of it when designing their self-driving cars. These cars analyze data in real-time and make decisions based on prescriptive analytics.

1.4.4.Predictive analytics

As the name suggests, this type of data analytics is all about making predictions about future outcomes based on insight from data. In order to get the best results, it uses many sophisticated predictive tools and models such as machine learning and statistical modeling.

Predictive analytics is one of the most widely used types of analytics today.

1.5. Tools used in Data analysis

In this section, we discuss various tools used in big data analysis.

1.5.1.Apache Hadoop

Apache Hadoop is one of the main supportive element in Big Data technologies. It simplifies the processing of large amount of structured or unstructured data in a cheap manner. Hadoop is an open source project from apache that is continuously improving over the years. "Hadoop is basically a set of software libraries and frameworks to manage and process big amount of data from a single server to thousands of machines. It provides an efficient and powerful error detection mechanism based on application layer rather than relying upon hardware."

1.5.2.MapReduce

MapReduce was introduced by google to create large amount of web search indexes. It is basically a framework to write applications that processes a large amount of structured or unstructured data over the web. MapReduce takes the query and breaks it into parts to run it on multiple nodes. By distributed query processing it makes it easy to maintain large amount of data by dividing the data into several different machines. Hadoop MapReduce is a software framework for easily writing applications to

manage large amount of data sets with a highly fault tolerant manner. More tutorials and getting started guide can be found at Apache Documentation.

1.5.3.HDFS

HDFS (Hadoop distributed file system) is a java based file system that is used to store structured or unstructured data over large clusters of distributed servers. The data stored in HDFS has no restriction or rule to be applied, the data can be either fully unstructured of purely structured. In HDFS the work to make data senseful is done by developer's code only. Hadoop distributed file system provides a highly fault tolerant atmosphere with a deployment on low cost hardware machines. HDFS is now a part of Apache Hadoop project, more information and installation guide can be found at Apache HDFS documentation.

1.5.4.Hive

Hive was originally developed by Facebook, now it is made open source for some time. Hive works something like a bridge in between sql and Hadoop, it is basically used to make Sql queries on Hadoop clusters. Apache Hive is basically a data warehouse that provides ad-hoc queries, data summarization and analysis of huge data sets stored in Hadoop compatible file systems. Hive provides a SQL like called HiveQL query based implementation of huge amount of data stored in Hadoop clusters. In January 2013 apache releases Hive 0.10.0, more information and installation guide can be found at Apache Hive Documentation.

1.5.5.Pig

Pig was introduced by yahoo and later on it was made fully open source. It also provides a bridge to query data over Hadoop clusters but unlike hive, it implements a script implementation to make Hadoop data access able by developers and business persons. Apache pig provides a high level programming platform for developers to process and analyses Big Data using user defined functions and programming efforts. In January 2013 Apache released Pig 0.10.1 which is defined for use with Hadoop 0.10.1 or later releases. More information and installation guide can be found at Apache Pig Getting Started Documentation.

1.6. Traditional Analysis versus Data Science Approach

In a basic sense, measuring learning using a big data approach isn't too dissimilar from utilizing traditional approaches like the long-established Kirkpatrick, Phillips or Kaufman's models. When using these approaches, you start by generating a hypothesis that a change you are going to make to your workforce's learning will affect your organization's performance. You then measure a baseline, make the change and measure again to see how your baseline data has changed.

1.6.1.Relational Databases (SQL)

A relational schema is a set of relational tables and associated items that are related to one another. All of the base tables, views, indexes, domains, user roles, stored modules, and other items that a user creates to fulfill the data needs of a particular enterprise or set of applications belong to one schema. SQL provides a statement to define a schema.

1.6.2.Schema less and Column oriented Databases (No Sql)

We are using table and row based relational databases over the years, these databases are just fine with online transactions and quick updates. When unstructured and large amount of data comes into the

picture we needs some databases without having a hard code schema attachment. There are a number of databases to fit into this category, these databases can store unstructured, semi structured or even fully structured data.

Apart from other benefits the finest thing with schema less databases is that it makes data migration very easy. MongoDB is a very popular and widely used NoSQL database these days. NoSQL and schema less databases are used when the primary concern is to store a huge amount of data and not to maintain relationship between elements. "NoSQL (not only Sql) is a type of databases that does not primarily rely upon schema based structure and does not use Sql for data processing.

1.7. Opportunities in Data Science

In rapidly evolving industries, big data enables businesses to solve today's manufacturing challenges and to gain a competitive edge. With big data and analytics, companies have got a chance to make better real-time decisions about asset usage and operations scheduling. Below are the most in-demanding opportunities:

- Data Scientist
- Data Architect
- Business Intelligence developer
- Data Engineer
- Data Analyst
- Decision Scientist

Lesson 1: Review Questions

1. Discuss the Data Mining Life Cycle.
2. Discuss characteristics of big data.
3. Explain the benefits of various types of big data analytics.
4. Differentiate between traditional and big data business approach.