Research Paper 1

Karan Klair – c0735732

04th July 2019

Question 1:

WHY DO WE SAY THAT AZURE IS A "CLOUD OFFERING". DISCUSS THE FEATURES OF THIS:

Answer:

Cloud computing is the process of providing computational power, resources, infrastructure and various other services through the internet referred to as a **"Cloud"**. Microsoft Azure is an implementation of cloud computing. Microsoft has created the cloud computing service Microsoft Azure which provides all of the services of cloud computing such as Infrastructure as a Service (IaaS), Platform as a Service(PaaS),Platform as a Service(SaaS). Microsoft Azure is termed as a Cloud offering due to the following features: -

1)Azure users choose Virtual Hard Disk (VHD), which is equivalent to a Machine Instance, to create a VM. VHD can be pre-configured by Microsoft, the user or a third party.

2)Azure offers temporary storage through D drive, block storage through Page Blobs for VMs. Block Blobs and Files also serve as object storage. Supports relational databases; NoSQL and Big Data through Azure Table and HDInsight. Azure also offers site recovery, Import Export and Azure Backup for additional archiving and recovery options.

3)Microsoft offers Virtual Network (VNET) that offers users ability to create isolated networks as well as subnets, route tables, private IP address ranges and network gateways.

4)Microsoft's pricing is also pay-as-you-go, but they charge per minute, which provides a more exact pricing model. Azure also offers short term commitments with the option between pre-paid or monthly charges.

5)Users are billed a flat monthly rate rather than monthly usage.

Question 2:

WHAT IS DATA SCIENCE? WHAT ARE ITS TOOLS AND METHODOLOGIES.

Answer:

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.Data science uses the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems" The various tools are:-

(a) **Cloud AutoML** is part of Google's Machine Learning suite offerings that enables people with limited ML expertise to build high quality models. The first product, as part of the Cloud AutoML portfolio, is Cloud AutoML Vision. This service makes it simpler to train image recognition models.

(b) **Trifacta** is another startup with a heavy focus on data preparation. It has 3 product offerings:

Wrangler: A free stand-alone software. Allows up to 100MB of data Wrangler Pro: An upgraded version of the above. It allows both single and multi-user and the data volume limit is 40GB Wrangler Enterprise: The ultimate offering from Trifacta. It does not have any limit on the amount of data you process and allows unlimited users. Ideal for big organizations Trifacta offers a very intuitive GUI for performing data cleaning. It takes data as input and provides a summary with various statistics by column

(c.)**MLBase** is an open-source project developed by AMP (Algorithms Machines People) Lab at the University of California, Berkeley. The core idea behind this is to provide an easy solution for applying machine learning to large scale problems.

Methodologies:

1.) Linear regression is a technique that is appropriate to understand the association between one independent (or predictor) variable and one continuous dependent (or outcome) variable. For example, suppose we want to assess the association between total cholesterol (in milligrams per deciliter, mg/dL) and body mass index (BMI, measured as the ratio of weight in kilograms to height in meters2) where total cholesterol is the dependent variable, and BMI is the independent variable. In regression analysis, the dependent variable is denoted Y and the independent variable is denoted X. So, in this case, Y=total cholesterol and X=BMI.

When there is a single continuous dependent variable and a single independent variable, the analysis is called a simple linear regression analysis . This analysis assumes that there is a linear association between the two variables.

2.) Supervised Learning: Data scientist acts as a guide to teach the algorithm what conclusions it should come up with. It's similar to the way a child might learn arithmetic from a teacher. Supervised learning requires that the algorithm's possible outputs are already known and that the data used to train the algorithm is already labeled with correct answers. For example, a classification algorithm will learn to identify animals after being trained on a dataset of images that are properly labeled with the species of the animal and some identifying characteristics.

3.) UnSupervised Learning: is based on idea that a computer can learn to identify complex processes and patterns without a human to provide guidance along the way. Although unsupervised learning is prohibitively complex for some simpler enterprise use cases, it opens the doors to solving problems that humans normally would not tackle. While a supervised classification algorithm learns to ascribe inputted labels to images of animals, its unsupervised counterpart will look at inherent similarities between the images and separate them into groups accordingly, assigning its own new label to each group.

4.) Decision Trees:A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Question 3:

HOW DOES THE CLASSICAL/RELATIONAL DATA MODEL COMPARE TO THE "BIG DATA" MODEL. WHAT NEW PROBLEMS DOES BIG DATA SOLVE? WHY ARE WE DOING BIG DATA NOW (NOT 20 YEARS AGO)?

Answer:

The major difference between traditional data and big data are discussed below.

Data architecture Traditional data use centralized database architecture in which large and complex problems are solved by a single computer system. Centralised architecture is costly and ineffective to process large amount of data. Big data is based on the distributed database architecture where a large block of data is solved by dividing it into several smaller sizes.

Volume of data The traditional system database can store only small amount of data ranging from gigabytes to terabytes. However, big data helps to store and process large amount of data which consists of hundreds of terabytes of data or petabytes of data and beyond.

Types of data Traditional database systems are based on the structured data i.e. traditional data is stored in fixed format or fields in a file. Big data uses the semi-structured and unstructured data and improves the variety of the data gathered from different sources like customers, audience or subscribers. After the collection, Bid data transforms it into knowledge based information (Parmar Gupta 2015).

Data relationship In the traditional database system relationship between the data items can be explored easily as the number of informations stored is small. However, big data contains massive or voluminous data which increase

the level of difficulty in figuring out the relationship between the data items (Parmar Gupta 2015).

We are using Big data today because:

Big Data is an absolute technological requirement to the process the enormous data sets of today. There are enormous amounts of data, structured and unstructured.Data is being generated at an exponentially growing rate as the world becomes digitized in every facet of human activity. Traditional relational databases with normalized data models are elegant and self consistent but cannot scale to the size required for big data. All of this big data is being mined for trends, behavior, correlations, demographics and predictive models.

Question 4:

DISCUSS THE KINDS AND USES OF THE FOUR KINDS OF ANALYTICS:

Answer:

Descriptive analytics answers the question of what happened. For instance, a healthcare provider will learn how many patients were hospitalized last month. Descriptive analytics juggles raw data from multiple data sources to give valuable insights into the past. However, these findings simply signal that something is wrong or right, without explaining why. For this reason, highly data-driven companies do not content themselves with descriptive analytics only and prefer combining it with other types of data analytics.

Descriptive analytics Diagnostic analytics At this stage, historical data can be measured against other data to answer the question of why something happened. Companies go for diagnostic analytics as it gives in-depth insights into a particular problem

Predictive analytics tells what is likely to happen. It uses the findings of descriptive and diagnostic analytics to detect tendencies, clusters and exceptions, and to predict future trends, which makes it a valuable tool for forecasting.

Prescriptive analytics The purpose of prescriptive analytics is to literally prescribe what action to take to eliminate a future problem or take full advantage of a promising trend. forecasting is just an estimate, the accuracy of which highly depends on data quality and stability of the situation, so it requires careful treatment and continuous optimization.