Research Paper 1 - Azure Machine Learning Labs

Subhajit Chakraborthy

July 2019

1 Reflection and Learning Questions

1.1 QUESTION 1: WHY DO WE SAY THAT AZURE IS A "CLOUD OFFERING". DISCUSS THE FEATURES OF THIS:

Answer:

Cloud is the extension of local or on-premise computing. When we say we use cloud computing, we are using someone else's (generally a cloud service provider's) resources. These resources can be anything from just external storage space to remote infrastructure. The service provider charges users based on the usage of resources.

Microsoft Azure is a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through Microsoft-managed data centers. It provides software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS) and supports many different programming languages, tools and frameworks, including both Microsoft-specific and third-party software and systems.

Some Features of **AZURE** are:
i: Build websites with ASP.NET, PHP or Node.js
ii: Deploy and run Windows Server and Linux virtual machine
iii: Migrate applications and infrastructure
iv: SQL Database
v: Virtual Network
vi: Cloud Services
vii: Hadoop

1.2 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 is the same concept as data mining and big data: ”use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems”.

1. **Tools:**:

i) Algorithms.io
ii) Apache Hadoop
iii) Apache HBase
iv) BigML
v) Cascading
vi) Data Robot

2. **Methodologies**:

i) Linear regression : Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

ii) Logistic Regression : Logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be combined to model several classes of events such as determining whether an image contains a cat, dog, lion, etc... Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

iii) Time Series: A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

iv) Association Rules : Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

1.3 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.

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.

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.

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.

1.4 QUESTION 4: DISCUSS THE KINDS AND USES OF THE FOUR KINDS OF ANALYTICS:

Answer:

There are 4 types of analytics. Here, we start with the simplest one and go further to more the sophisticated. As it happens, the more complex an analysis is, the more value it brings.

1.Descriptive analytics: Descriptive analytics answers the question of what happened. For instance, a healthcare provider will learn how many patients were hospitalized last month; a retailer – the average weekly sales volume; a manufacturer – a rate of the products returned for a past month, etc. Let us also bring an example from our practice: a manufacturer was able to decide on focus product categories based on the analysis of revenue, monthly revenue per product group, income by product group, total quality of metal parts produced per month.

2.Diagnostic analytics: At this stage, historical data can be measured against other data to answer the question of why something happened. Thanks to diagnostic analytics, there is a possibility to drill down, find out dependencies and identify patterns. Companies go for diagnostic analytics as it gives in-depth insights into a particular problem. At the same time, a company should have detailed information at their disposal otherwise data collection may turn out to be individual for every issue and time-consuming.

3.Predictive analytics: 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. Despite numerous advantages that predictive analytics brings, it is essential to understand that 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.

4.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. An example of prescriptive analytics from our project portfolio: a multinational company was able to identify opportunities for repeat purchases based on customer analytics and sales history.