BWT-Data Science Task 15

Bisma Fajar

July 2024

1 Introduction to Data Science Life Cycle

This process can be broken down into 5 stages:

- Capturing
- Processing
- Analysis
- Communication
- Maintenance

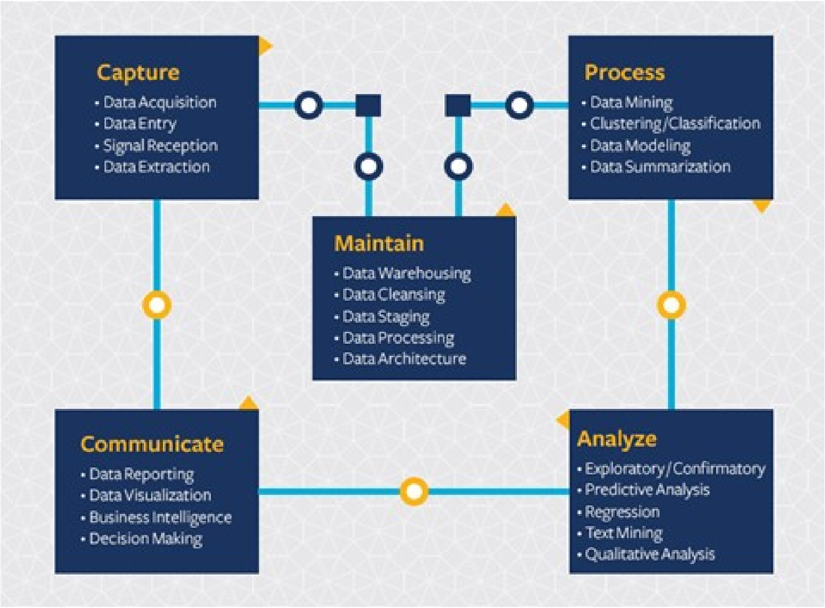

Figure 1: Data Science Life Cycle

1.1 Capturing

It's practically two stages combined into one: acquiring the data and defining the purpose and problems that need to be addressed. Defining the goals of the project will require deeper context into the problem or question.
Questions a data scientist may ask:

- Has this problem been approached before? What was discovered?
- Is the purpose and goal understood by all involved?
- Is there ambiguity and how to reduce it?
- What are the constraints?
- What will the end result potentially look like?
- How much resources (time, people, computational) are available?

Next is identifying, collecting, then finally exploring the data needed to achieve these defined goals.
Questions a data scientist may ask about the data:

- What data is already available to me?
- Who owns this data?
- What are the privacy concerns?
- Do I have enough to solve this problem?
- Is the data of acceptable quality for this problem?
- If I discover additional information through this data, should we consider changing or redefining the goals?

1.2 Processing

The processing stage of the lifecycle focuses on discovering patterns in the data as well as modeling. Some techniques used in the processing stage require statistical methods to uncover the patterns. Typically, this would be a tedious task for a human to do with a large data set and will rely on computers to do the heavy lifting to speed up the process. This stage is also where data science and machine learning will intersect.
Common techniques used in this stage are covered in the ML for Beginners curriculum.

- **Classification:** Organizing data into categories for more efficient use.
- **Clustering:** Grouping data into similar groups.
- **Regression:** Determine the relationships between variables to predict or forecast values.

1.3 Maintaining

Maintenance is an ongoing process of managing, storing and securing the data throughout the process of a project and should be taken into consideration throughout the entirety of the project.

1.4 Storing Data

Considerations of how and where the data is stored can influence the cost of its storage as well as performance of how fast the data can be accessed. Decisions like these are not likely to made by a data scientist alone but they may find themselves making choices on how to work with the data based on how it's stored.

1.4.1 On premise vs off premise vs public or private cloud

On premise refers to hosting managing the data on your own equipment, like owning a server with hard drives that store the data, while off premise relies on equipment that you don't own, such as a data center. The public cloud is a popular choice for storing data that requires no knowledge of how or where exactly the data is stored, where public refers to a unified underlying infrastructure that is shared by all who use the cloud. Some organizations have strict security policies that require that they have complete access to the equipment where the data is hosted and will rely on a private cloud that provides its own cloud services.

1.4.2 Cold vs hot data

When training your models, you may require more training data. If you're content with your model, more data will arrive for a model to serve its purpose. In any case the cost of storing and accessing data will increase as you accumulate more of it. Separating rarely used data, known as cold data from frequently accessed hot data can be a cheaper data storage option through hardware or software services. If cold data needs to be accessed, it may take a little longer to retrieve in comparison to hot data.

1.5 Managing Data

As we work with data we discover that some of the data needs to be cleaned using some of the techniques covered in the lesson focused on data preparation to build accurate models. When new data arrives, it will need some of the same applications to maintain consistency in quality. Some projects will involve use of an automated tool for cleansing, aggregation, and compression before the data is moved to its final location. Azure Data Factory is an example of one of these tools.

1.6 Securing Data

One of the main goals of securing data is ensuring that those working it are in control of what is collected and in what context it is being used. Keeping data secure involves limiting access to only those who need it, adhering to local laws and regulations, as well as maintaining ethical standards, as covered in the ethics lesson.

Here's some things that a team may do with security in mind:

- Confirm that all data is encrypted
- Provide customers information on how their data is used
- Remove data access from those who have left the project
- Let only certain project members alter the data