Task 1: The Data Analytics Life Cycle
By Eric Williams

## A: PHASES OF THE DATA ANALYTIC LIFE CYCLE

The data analytics life cycle is made up of seven parts. I will list them here and give a summary of them, as well as describe my personal experience and expertise with each part:

1. **Business Understanding**. In this phase, stakeholders and upper-level management meet to define what analysis needs to be done to benefit the company. They evaluate what depth the analysis will need and what the goals are.

   **My expertise**: I personally have no expertise in this area, as I have no business background and have never been a stakeholder or upper-level manager. However, I sometimes do data analysis on my own personal projects, such as for my word count when writing books or measuring my progress toward my fitness goals. In this situation, I have to decide what I want to know from the data I have collected. When analyzing my writing, I like to look at my word count over time and see how my everyday life impacted my productivity.

2. **Data Acquisition**. In this phase, the analyst gathers all the data they need from various sources available to them. Usually it is supplied by the company. The data could be taken from databases, an Application Programming Interface, a data stream, or another source.

   **My expertise**: I personally have no expertise in professionally acquiring data because I am relatively new to data analytics. However, I have created and gathered data on my writing productivity when I wrote my first and second books. Acquiring the data was relatively simple because I kept it in a word document and updated it every day after I finished writing.

3. **Data Cleaning**. In this phase, the data is cleaned and changed to be analyzed. A lot of times, raw data is not formatted in a way that it can be analyzed easily. Sometimes, as an analyst, you are given too much data and must pare it down to just the useful data that you need to meet your goals.

   **My expertise**: I haven not worked as a data analyst, so I have not done any professional data cleaning. However, when analyzing my own data for my own projects, I often have to format the data properly so I can look for trends. When analyzing trends in my writing while writing my first book, I copied the data over from a document into a spreadsheet. However, I had to clean the word count data to just the raw numbers. Spreadsheets have a hard time analyzing numbers when there are strings attached in the same cell, so I had to manually delete the strings to just leave the numbers and dates, which

comprised the useful data in this case. So I have a small amount of expertise on the concept of data cleaning, but I have not performed data cleaning in Python before.

4. **Data Exploration**. In this phase, the analyst explores the data to look for patterns, trends, and relationships. This can be done by creating graphs and other visualizations, which can make relationships easier to identify.

   **My expertise**: Again, I have not done this professionally, but when analyzing my book writing data, this is the phase when I create scatterplots, histograms, and bar graphs for my word count over time in Excel. I identify the trends and try to picture what the most accurate representation of my data would be. I have a little bit of expertise on this as a general concept, but have not yet developed the programming skills to do this in a professional setting.

5. **Data Modeling**. In this phase, the analyst develops the models (mathematical or computational) in order to predict future data, or to determine what conclusions can be drawn from the data.

   **My expertise**: I have no professional experience with this, but I also, unfortunately, have no amateur experience with this either. Advanced models and future predictions are well beyond the scope of the data modeling I have done in the past. I have no expertise in this, outside of some basic modeling in statistics classes (no programming).

6. **Data Mining/Machine Learning**. In this phase, the analyst “mines” the data, meaning they use the data to try and find the useful “nuggets” that can be useful for interpreting the data and trends. In this step, the trends and data are interpreted and reports are given to the stakeholders and upper management who requested the data analysis. Machine Learning can also be part of the data mining process.

   **My expertise**: I have no professional or amateur expertise with this, though I suppose when I analyze my own data as a hobbyist, this is when I determine my findings and draw conclusions from the data I have analyzed.

7. **Reporting and Visualization**. In this phase, the last step is to interpret and communicate the findings of the analyst. This is when you get to tell the story of what you found by analyzing the data. This can be done in a report, a presentation, with a Tableau dashboard, or in a number of ways.

   **My expertise**: Personally, I have never done this professionally, but this is when I take the visualizations I have made and write an article about it for my writing blog. I present the data to the world by publishing the numbers, data, and findings online. I would say I have some expertise with this, but I have never presented data I analyzed professionally. So it is accurate to say my experience is quite limited.

## A1: PROPOSAL TO GAIN EXPERTISE

1. **Business Understanding**

   **How I might gain expertise**: I could gain expertise by working at a company that allows me to sit in on their meetings while they make business decisions. Alternatively, if I ever rise to the level of upper management or stakeholder, I might be able to be part of the process myself.

2. **Data Acquisition**

   **How I might gain expertise:** I could gain expertise by working at a company that has data that needs to be analyzed. It would then be part of my job to acquire the data the company has out of a spreadsheet or an API.

3. **Data Cleaning**

   **How I might gain expertise**: I could gain expertise by downloading a dataset from the internet and properly preparing it for analysis. Raw data is rarely ready to be analyzed, so I could find datasets online and determine how they need to be changed so that they are compatible with Python code.

4. **Data Exploration**

   **How I might gain expertise**: I could gain expertise by using my programming skills in Python to create scatter plots, histograms, or box and whisker plots using downloaded datasets.

5. **Data Modeling**

   **How I might gain expertise**: I could gain expertise by downloading a dataset and performing linear regressions and clustering analysis in Python.

6. **Data Mining/Machine Learning**

   **How I might gain expertise**: I could gain expertise by creating a machine learning model in Python. I could take a dataset and train the model using Python code to sort the data into categories.

7. **Reporting and Visualization**

   **How I might gain expertise**: I could gain expertise by working in Tableau with a dataset to create visualizations. I could also practice creating a report or an interactive dashboard that would be useful for stakeholders.

## A2: HOW THE GOAL AND MISSION HELP THE ANALYST

Any organization that requires data analytics would have both a goal and a mission.

1. **How an organization's goal helps the analyst identify the business requirement**

   A goal may be something simple, such as "increase revenue by 20% this quarter." An organization may have multiple goals, or constantly evolving ones. The analyst should always know what the goal of an organization or company is to ensure the project helps them reach the goal. The analyst should use the goal of the organization as a guide for his data analysis, to ensure the analysis is helpful and contributes toward achieving the goal. The business requirement for a data analyzing project should help the organization reach the goals they have set.

2. **How an organization's mission helps the analyst identify the business requirement**

   The mission of an organization is usually much more broad than a goal. A mission is the overall, long-term vision for the future. For example, Microsoft's mission is to "to empower every person and every organization on the planet to achieve more" (Satya, 2017). Thus, if a data analyst is working for Microsoft, they should ensure their analysis contributes to the overall mission to empower every person on the planet to achieve more. If the analysis is unrelated to the mission or does not support the mission, then it is likely not going to be a valuable contribution to the company. Understanding the specifics of the mission should provide valuable insight into the business requirement of your analysis.

## B: DATA ANALYTICS LIFE CYCLE

For my data analytics tool, I choose Python. Python is an object-oriented programming language that can be used for a wide range of projects. It is especially powerful for data analysis using packages such as Pandas, NumPy, Matplotlib, and Scikit-learn.

**How Python might be used in Data Exploration for an organization like Microsoft**
If I was working as a data analyst for a company like Microsoft, I might be assigned to analyze survey responses describing which operating system consumers prefer and why. I could use Matplotlib in Python to create bar graphs and histograms to explore the relationship between consumers of different ages and their preferences for different operating systems. I could explore the data by creating visualizations to look for relationships between what kinds of people prefer the newer operating systems and which do not. I could then use this data to suggest updates to the operating systems to appeal to specific demographics, or to target marketing to desired demographics.

**B1: RISKS**

Three risks I might encounter while analyzing consumer data with Python for Microsoft would be:

1. **Data Privacy**
   When collecting information about consumers, it is important to not mishandle data that consumers would want to make public. Although it could be valuable to know a consumer's birthdate and address, the specifics are probably not essential for data analysis in Python. Simple information such as age and the state in which they reside would be enough information to identify their demographic.

2. **Security Breaches**
   Part of the danger mentioned above when storing sensitive information is that if you experience a data breach, someone could steal private consumer data. If someone is able to access your consumer information, it is best to not have personal data you do not need so that it cannot be stolen and abused. Although the data is necessary to run the Python code, it should be gathered and stored carefully so it is not misused.

3. **Bias/Discrimination**
   Sometimes, analyzing data on people can lead to biased results. This can especially be a problem if the analysis is not representative of the population. This could occur because of improper sampling in the surveys sent out, or perhaps certain demographics would be less likely to respond to a survey. This could lead to skewed, biased results that would show in the data analysis, creating visualizations in Python that are not actually representative of the population.

**B2: DESCRIPTION OF PROBLEM**

If I were doing data analysis for Microsoft, it is entirely plausible that they would want me to create an interactive dashboard for the data analysis that includes a state map that displays data when selected. This would be very difficult to create in Python and would be much better suited for a program like Tableau. Although it is possible to program interactive elements in Python, this technical problem would be much easier and less time-consuming to perform in software that is built for easily creating interactive material, not just analyzing data and creating simple visualizations.

**C: TOOL SELECTION PROCESS**

It is important to select the right tool for the right data analyzing job. Some tools are better for dealing with structured data, others are better at cleaning data, and others are better at displaying data. Thus, it is important to understand what your end goal for data visualization is and work backward to determine which tool will help you create it. For example, if you are working with a relational database, SQL can be a great tool. If you are doing statistical analysis,

R can be a great programming language. If you have specific data science goals that require analyzing large datasets and producing simple regressions or representations, Python is the ideal tool. Tableau is idea if you want to create an interactive visual dashboard.

In the process I chose in Part B, I chose to use Python for Data Exploration based on consumer survey responses about their preference regarding operating systems. If there are a lot of surveys given that result in a massive dataset, Python would be an excellent tool for analyzing that data set. With Matplotlib in Python, it would be easy to explore the data visually and create simple graphs and charts.

**C2: FINDINGS**

Because the problem I chose was creating an interactive state map visual dashboard, the solution to the problem is to use Tableau instead of Python. Tableau is great at working with nearly any database and its simple drag-and-drop features make it very interactive and easy to create visualizations.

As mentioned above, Python would be an ideal tool for analyzing a large dataset and for creating basic visualizations, but it is not necessarily the right tool for the job when creating an interactive dashboard.

**C3: POTENTIAL ETHICAL RISKS**

When using Python to handle consumer data, there are three potential ethical risks:

1. **Ethical Problem: Not being candid about known or suspected limitations, defects, or biases in the data.**

   This can affect the reliability of the statistical analysis. Being aware of and up-front about the limitations or problems with the analysis can help your analysis maintain integrity. In my specific example of using Python to analyze consumer data, if there is missing data, if the data is skewed, or if there were any problems with accurate survey representation when collecting the data, those issues should be disclosed as a disclaimer on why the data might not accurately represent reality.

2. **Ethical Problem: Not supporting valid inferences, transparency, and sound science.**

   For example, assume the consumers in the survey were not aware that they were participating in a survey to be analyzed. This lack of transparency can be considered unethical. Similarly, if you are collecting and analyzing data to exploit your consumers or clients or deceive the scientific community by offering inaccurate data or conclusions, that would be unethical. In this example, it would be important to uphold scientific

standards while also keeping the interests of the consumer in mind rather than exploiting them for profit.

3. **Ethical Problem: Not protecting and respecting the rights and interests of the consumers.**

   In this example, it would be unethical to collect and store personal information on the consumers that is not necessary for the data analysis. Gathering data like birth dates, social security numbers, or home addresses would not be helpful in understanding the demographics that prefer certain operating systems over others. Similarly, if sensitive data is not stored carefully and securely, it could lead to a data breach that could negatively affect the consumers. It is important to only gather the data needed and secure it properly.

Sources

Satya, N. (2017). *Hit Refresh: The Quest to Rediscover Microsoft's Soul and Imagine a Better Future for Everyone*

All other sources used were WGU course materials.