In Task 1, I will draw heavily from my professional experience as a data analyst. I have over eight years in the tech industry. I began my career at Amazon, working with large datasets about the North American warehouse network. My responsibilities included identifying missing, damaged, and illegal inventory. Subsequently, I transitioned to a role as a project manager and analyst in Brand Protection, where I assisted customers in safeguarding their intellectual property rights. This involved detecting risks, fraudulent activities, and counterfeit products that could adversely affect Amazon and its customers. I work as a data analyst at Starbucks' research and development lab. In this role, I contribute to test preparation, analyze test results, and present key findings and insights utilizing Microsoft Excel and Tableau.

I will discuss each phase of the data life cycle, drawing on my expertise in data analytics and the tools I employ. I will explain why these tools are integral to my professional career and the ethical considerations I uphold. Throughout this task, I will incorporate my experiences as a data analyst with over eight years of experience.

The **Business Understanding** phase is the foundation of the data analytics lifecycle. In this phase, we define key business questions that guide our analysis, ensuring the entire process aligns with the organization's goals and objectives. This step involves collaborating with stakeholders to identify the problem, determine the data requirements, and outline the data collection and analysis methods. Analysts can navigate datasets more effectively by establishing a clear focus and ensuring that their work directly addresses the organization's needs. (*Course* |, n.d.)

In my current role, I analyze data to help our organization gain insights into retail store operations. For example, I examine metrics such as order fulfillment times, errors in the fulfillment process, and anomalies that indicate underlying issues or opportunities for improvement. I contribute to refining processes and driving operational efficiency by focusing on these areas. To maintain alignment with stakeholders, we hold weekly meetings every Monday to review assignments, discuss ongoing projects, and address potential roadblocks. These discussions help us refine our focus, assess data from the previous week, and make any necessary adjustments to our approach.

By gaining expertise in this area, you can Improve your communication skills, learn to effectively communicate with stakeholders to understand their needs and translate business goals and analytical tasks. Study business fundamentals to better understand the organizations' industry, operations, and strategic objectives. Learn project management techniques and develop skills to organize and prioritize tasks, ensuring the analysis stays focused and relevant.

The **Data Acquisition** phase involves gathering data required for analysis. This step is crucial, ensuring analysts have the necessary information to address the business problem or research question. Data comes from various databases, APIs, external files, or real-time streaming systems. Analysts must ensure the data is accurate, complete, and relevant. (*Course* |, n.d.)

In my current role, I retrieve data using SQL and Python on Microsoft Azure via Databricks notebooks. My experience includes writing ETL pipelines to transform and load data into a usable format and creating automation tasks to retrieve data at regular intervals (e.g., every hour or every 24 hours). I also write targeted queries to extract specific subsets of data for analysis,

such as insights into particular tests. These skills ensure efficient and reliable access to data, a foundational step for meaningful analytics.

To develop expertise in this phase, one could learn SQL and Python and write queries and scripts to extract data from relational and non-relational databases. One could also understand data storage systems and gain familiarity with cloud platforms like Microsoft Azure, AWS, or Google Cloud. One could also automate data retrieval and explore tools for data acquisition, such as cron jobs, Apace Airflow, or custom Python scripts. (*Get to Know Azure | Microsoft Azure*, n.d.) (*Google Cloud Documentation*, n.d.)

The goal and mission of an organization significantly influence data acquisition. For Example, if the mission is to improve customer satisfaction, the analyst would prioritize collecting data from customer feedback forms, surveys, and social media interactions. This ensures the data gathered aligns with the strategic objectives and drives actionable insights.

The **Data Cleaning** phase involves preparing the raw data to ensure it is accurate, complete, and consistent. This phase is vital because high-quality analysis depends on reliable and clean data. Typical tasks in this phase include identifying and handling missing values, correcting inconsistencies, standardizing formats, and removing outliers or irrelevant entries. A thorough data cleaning process ensures the data is suitable for analysis and decision-making. (*Course |*, n.d.)

In my current role, I frequently encounter data inconsistencies when retrieving information from our databases. This is because the data is inputted in real-time by a team of over 45 individuals, leading to variability in formatting and occasional errors. For instance, data tied to a specific Project ID may contain duplicate entries, missing values, or test attempts flagged as errors. My responsibility is to identify and resolve these issues. I rely on Microsoft Excel to spot inconsistencies, standardize data formats, and remove erroneous entries, ensuring the cleaned data is ready for analysis.

To develop expertise in this phase, Learn data cleaning tools and techniques and familiarize yourself with tools like Python, R, or Excel. Understand data quality, study common data issues, and how they affect analysis outcomes. Automate cleaning processes, build scripts or use tools to streamline repetitive cleaning tasks.

The **Data Exploration** phase focuses on summarizing and understanding the data that has been gathered and cleaned. This critical step involves examining the data to uncover initial patterns, relationships, and anomalies. The goal is to better understand the dataset's structure, trends, and potential issues. Typical tasks include calculating descriptive statistics (e.g., mean, median, and standard deviation), visualizing data distributions, and identifying correlations or outliers. (*Course |*, n.d.)

In my professional experience, I have utilized Microsoft Excel pivot tables extensively during the Data Exploration phase. Pivot tables allow me to quickly summarize large datasets and identify trends that merit further investigation. For instance, when analyzing test results, I often create pivot tables to aggregate data by categories, such as test type or period, and calculate averages or counts. This work usually reveals key patterns, such as an unexpected spike in test

failures, which prompts deeper analysis. Additionally, I have used visual tools like scatterplots and histograms to identify correlations and spot outliers that may affect the validity of subsequent studies.

To build expertise in this phase, you could Learn data visualization tools, such as Tableau Power BI, or Python libraries, such as Matplotlib and Seaborn. You could also develop statistical skills, gain proficiency in applying descriptive statistics, and understand their implications. Finally, you could work with real-world data, practicing exploring messy, real-world datasets to get accustomed to handling practical challenges.

The Data Exploration phase should align with the organization's goals to ensure relevant and actionable insights. For example, the exploration phase might focus on uncovering trends highlighting process inefficiencies if the organization's mission is to improve operational efficiency. In my role, exploring test result data to identify trends or anomalies directly supports the organization's goal of maintaining high-quality standards and improving overall performance.

**Predictive Modeling** uses statistical algorithms and machine learning techniques to forecast future outcomes or behaviors based on historical data patterns. This phase is pivotal for making data-driven decisions, allowing organizations to anticipate challenges and opportunities. Predictive modeling often involves regression, classification, and time-series analysis to identify trends and predict outcomes with measurable accuracy. (*Course* |, n.d.)

While my current role does not heavily involve predictive modeling, I plan to build expertise in this area by learning Python and applying regression models. For example, I intend to undertake a project predicting inventory shortages based on historical trends. This project will allow me to explore techniques such as linear regression to identify patterns in inventory data and machine learning algorithms to enhance predictive accuracy. By integrating historical data with predictive analytics, I aim to develop actionable forecasts supporting better planning and decision-making.

To gain expertise in this phase, you can learn key tools and techniques and develop proficiency in Python or R, focusing on libraries like Scikit-learn, TensorFlow, or PyTorch for Machine Learning. Practice real-world problems and take on projects or Kaggle competitions to apply predictive modeling to practical scenarios. Understand the theoretical foundations, study statistical concepts like linear and logistic regression, and advanced machine learning algorithms like random forests and neural networks. (*PyTorch Documentation — PyTorch 2.5 Documentation*, n.d.)

The organization's mission and goals provide direction for predictive modeling efforts. For instance, a retail organization aiming to optimize inventory management might prioritize predictive models to forecast demand fluctuations. In my planned project, predicting inventory shortages would align with an organization’s goal of minimizing disruptions and improving operational efficiency. Tailoring predictive models to these objectives ensures that the analysis delivers maximum value.

The **Data Mining and Machine Learning** phase focuses on discovering meaningful patterns, correlations, and insights within large datasets using statistical methods and advanced machine-learning techniques. This phase enables analysts to extract actionable knowledge from

complex datasets, uncover hidden trends, and develop predictive or prescriptive models. Standard methods include clustering, classification, association rule mining, and anomaly detection. (*Course* |, n.d.)

Building expertise in this area, you could learn core machine learning techniques, studying foundational methods like clustering, classification, and anomaly detection using Python libraries such as Scikit-Learn or TensorFlow. You could also work on practical projects, engage in hands-on projects, or participate in competitions to apply data mining and machine learning techniques to real-world problems. Understand theoretical concepts, study books, and online courses to grasp statistical concepts and machine-learning principles. (*User Guide*, n.d.)

The **reporting and visualization** phase of the data analytics lifecycle focuses on presenting the findings clearly and in a digestible manner. This phase tells the story of the data, allowing stakeholders to make informed decisions based on the insights uncovered during analysis. Microsoft Excel, Power BI, Tableau, Google Looker, and Elastic Kibana are commonly used to create impactful visualizations and real-time dashboards. These tools enable analysts to connect to databases, write SQL, Python, or R code, and transform complex data into accessible formats through charts, graphs, and interactive elements. (*Course* |, n.d.)

In my role, I regularly utilize Microsoft Excel and Power BI to create visual reports for stakeholders and team members. For example, I use bar graphs to illustrate trends, tables to present detailed numerical data, and timeline sliders or search bars to enable interactive exploration of key metrics. These visualizations help translate intricate datasets into meaningful insights, ensuring that everyone can understand and act upon the findings regardless of technical expertise.

To gain expertise in this phase, one could Learn visualization tools and develop proficiency in tools like Power BI, Tableau, or Looker to create professional dashboards. Enhance design skills by studying data visualization principles to create clear, aesthetically pleasing, and effective visuals. Practice coding for visualization and learn Python for custom visualizations and advanced analytics.

A potential problem with reporting and visualization is that my team has to be mindful of the audience that will read our reports. We have procedures in place to ensure that there are multiple versions of our Microsoft Excel pivot tables and dashboards. They have hidden or removed data so that no one outside the organization and team sees it. This is due to an issue where another team read our data and made a costly decision based on misunderstanding that data.

There are many different risks with how I use my tools and techniques in my current data analyst role. Our use of Data Bricks does not include a test / non-live database. Instead, I create a copy of the ETL notebook I want to change, make the edits in my copy, run the query, and ensure the results are what I want to see. After trusting and verifying the results, I go back and make the live changes to the ETL pipeline, which then updates each time on the hour.

Other risks are outside of my control, too. Since I work at such a large company, multiple organizations and teams are changing our databases, which can dramatically impact how we retrieve our data. They can also break the ETL pipelines we have written. An example of this is

when an update made by an engineering team outside my organization prevented one of the ETLs in our pipeline from working correctly.

A third risk of relying too much on SQL and Python is automation. A goal for my data team is to automate as much data cleaning as possible. However, careful consideration needs to be made. While we retrieve much of the same data each time for our stakeholders, how the data must be presented and manipulated differs. This means that automation might adjust a .csv file in a way that is irrelevant to what we need and can even make the data inaccurate and ineffective. Writing automation scripts for each use case scenario is necessary.

Knowing SQL and Python is necessary for my role because careful gathering and retrieval of our data is necessary for success. Thankfully, we have notebooks with Data Bricks and can save and share the most common and useful bits of code with the team. Something about our use of Data Bricks is how you can write a notebook that runs code entries in sequence. An example is an SQL query to retrieve the data, followed by Python code that will modify it, such as cleaning it. Then, we download it as a .csv file and work on the data visualization in Microsoft Excel Pivot tables.

SQL is a compelling programming language because its syntax is simple but powerful. It is used in many different ways to retrieve the data we need, manipulate it, and mold it to fit the needs of our teams and stakeholders. I use SQL to recover data from multiple databases and tables and combine them into one report. I then use it to make dashboards and Excel pivot tables.

SQL enables efficient retrieval of structured data, which is critical for creating timely reports that drive business decisions. An example of this is extracting sales trends to help optimize inventory management.

Python is used because it efficiently performs optimizations, runs automation, and even creates data visualization. It has a large, well-known, and maintained library set, such as Numpy, pandas, scipy, sci-kit, and TensorFlow. Python is also significant in that it works with many different IDEs. I use Data Bricks, for example. It can also run locally and as a web app. (*Tutorials*, n.d.)

The selection of tools involves considerations such as scalability, team familiarity, and integration with existing systems. The course material highlights that SQL supports efficient relational data handling and integrates seamlessly into established analytics pipelines.

An ethical problem with any tool is data privacy and protection: Knowing when not to share that data with those outside of your organization or even company. Knowing the best practices to keep your data secure and not accidentally leaking that data is essential. Examples include someone asking you for data or seeing exciting customer information or upcoming new products. Still, it is crucial not to share that with friends and family. My role as a data analyst allows me a lot of privilege in knowing sensitive information that is not yet available to the public.

It is essential to use appropriate data for what your stakeholders are requesting. Do not find easy or quick shortcuts for gathering the data. Trust and verify your data, and present it in the most honest way possible. Stakeholders want to see the truth in the data to make the best-informed decision.

Be candid and open about data limitations. Speak up quickly and often, and ask questions when you see data that is error-prone, does not look right, or returns results that are not expected. I see this all the time in my data analyst roles: data that is wildly different than what was expected, and then I need to go back to my stakeholders and discuss my findings and how to proceed. We devise an action plan to figure out the best way to work with this data, and sometimes, that data needs to be thrown out altogether and started over again.

*Course |. (n.d.).* *https://apps.cgp-oex.wgu.edu/wgulearning/course/course-v1:WGUx+OEX0343+v01/block-v1:WGUx+OEX0343+v01+type@sequential+block@543516264ee84ad4b89a1449a5d2db90/block-v1:WGUx+OEX0343+v01+type@vertical+block@63e2765421dc476a8611a55973cafd50*

*Get to know Azure | Microsoft Azure*. (n.d.). https://azure.microsoft.com/en-us/explore/

*What is Databricks?* (n.d.). Databricks on AWS.

https://docs.databricks.com/en/introduction/index.html

*PyTorch documentation — PyTorch 2.5 documentation*. (n.d.).

https://pytorch.org/docs/stable/index.html

*User Guide*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/user_guide.html

*Tutorials*. (n.d.). TensorFlow. https://www.tensorflow.org/tutorials

*What is Databricks?* (n.d.). Databricks on AWS.

https://docs.databricks.com/en/introduction/index.html

*Google Cloud Documentation*. (n.d.). Google Cloud. https://cloud.google.com/docs

*NumPy documentation — NumPy v2.2 Manual*. (n.d.). https://numpy.org/doc/stable/

*pandas documentation — pandas 2.2.3 documentation*. (n.d.). https://pandas.pydata.org/docs/