Robert Pavlik

D596 – The Data Analytics Journey

LFN1 Task 1: The Data Analytics Life Cycle

A.

The seven phases of the data analytics life cycle are Business Understanding, Data Acquisition, Data Cleaning, Data Exploration, Predictive Modeling, Data Mining, and Reporting and Visualization. My current position is as a Software Engineer at a Tax Company. I work daily on designing, maintaining, and improving forms the company uses to process and store people's and place tax data. At work, I not only write code but also work a lot in MS SQL Server to retrieve and manipulate data.

- Business Understanding – Also known as the discovery phase. This is where the stakeholders are identified, and the project's goal and scope are clearly defined.
    - Reflection – In my current role I engage in the business objectives by aligning my design and improvement task with the company’s tax processing needs.
    - Gain Expertise - By working with different departments at my company and attending cross-functional meetings I would be able to gain a better understanding of the departments that my company deploys.
- Data Acquisition – Also known as the collecting phase. This is where all the relevant data from various sources is collected that will be used later in the life cycle process.
    - Reflection – My daily work with MS SQL Server gives me a strong foundation in the data acquisition process. Working with complex SQL queries to pull data from multiple sources that can be used to show a larger picture when it comes to processing tax information.
    - Gain Expertise - The ability to learn different data integration tools and techniques such as ETL tools, and data APIs.
- Data Cleaning – This phase ensures all data that is collected is relevant to the project objectives. Handling any missing values, outliers, and inconsistencies is critical for ensuring accurate information.
    - Reflection – Dealing with a lot of data daily, I must keep an eye out for inconsistent and inaccurate data to ensure that the forms are processing the proper and correct data.
    - Gain Expertise – Gaining more practice working with complex datasets can help in getting better at ensuring that only accurate data is being used for the project.
- Data Exploration – During this phase all the relevant and accurate data is analyzed to discover patterns, and relationships and gain insight.
    - Reflection – Working on tax forms I can see daily the relationships in data like those between local tax data and tax rates and how it can affect customers and counties.
    - Gain Expertise - Taking courses like this degree program on statistics and EDA to strengthen my understanding of data distribution and relationships.
- Predictive Modeling – This phase allows the analyst a chance to build and validate models using the data to predict future outcomes.

    - Reflection – At work, I do not personally use predictive modeling tools. I do understand SQL and data flow which I think will help me when learning more about predictive techniques in this course.

    - Gain Expertise - Studying and learning algorithms and concepts of machine learning and practicing building predictive models using analyst tools would be a great way to get a better understanding of Predictive Modeling

- Data Mining – Analysts use computers and machine learning to look through large amounts of data to identify patterns and gain knowledge.

    - Reflection – Using SQL to process large datasets gives me a good starting point for understanding the phase of the process.

    - Gain Expertise – This phase doesn't happen at my current job, so I think the best way for me to gain experience and a better understanding of this process would be to take these classes and learn the process that data analysts go through during the data mining phase.

- Reporting and Visualization – During this phase, the analyst presents the data with graphs and interactive dashboards to inform the other stakeholders of the results. This will help them also see the trends and patterns to identify actionable insight.

    - Reflection – Designing the interface on in-house forms pulls multiple data points together to give a picture of the information needed to perform various tax-related tasks. I then present the final product to the different teams that will be using this form to process their daily task. My technical communication skills are critical to showing the other stakeholders the changes made and the new way the data is being presented.

    - Gain Expertise - During this degree program learning tools like Tableau will assist me in learning how to create dashboards that can be used to summarize complex data and allow the stakeholders the chance to explore the data results.

The goals and missions of an organization assist in helping the analyst identify business requirements by ensuring that the efforts are aligned with the company's strategic objectives. Being an analyst or a software engineer, understanding this mission helps in identifying which data elements are most critical, how they should be processed, and what insights would be most valuable to stakeholders. This alignment ensures that the analytical solutions developed are not only technically sound but also contribute to the company's success.

B.

During the data cleaning phase of the data analytics life cycle using SQL to query, update, and manipulate data can be useful in identifying and correcting errors, handling missed values, removing duplicates, and ensuring data consistency. Using things such as JOINs, UPDATEs, and DELETEs makes SQL ideal for efficiently cleaning and preparing the data for the next step in the life cycle.

In my current company, we use queries to ensure that the data being processed is the correct data and in the correct format. Things such as Social Security numbers, mailing addresses, and customers' names; can be frequently entered into the database incorrectly, and cleaning the data to ensure that they are displayed and displayed correctly in the product is crucial to ensuring that the results and processing are compliant with regulatory requirements.

- Human Error – Writing SQL is just like any other programming language proper syntax and formatting are critical to ensuring data integrity. A poorly written query can inadvertently corrupt data, leading to inaccurate analysis and loss of critical information.

- Performance Issues – Processing and Data Cleaning on a large dataset can be very resource intensive, which in turn can lead to slow performance and even crash the servers if not handled properly. Complex queries are known to consume large amounts of significant memory and processing power, potentially affecting the performance of other operations being run on the database. Things like running a full table scan can affect the performance of other tasks being run especially during peak company hours.

- Scalability Limitations – While SQL can be a powerful tool it also might not be the most efficient way to process large datasets. As volumes of data grow, queries can become harder to manage and optimize. Which in turn would lead to longer processing times and increase the difficulty in maintaining data integrity.

- An organization's data can come from many sources analyzing that data is crucial to an organization. SQL can be a tool to assist in doing that task, but it doesn't come without risk or potential problems. This can especially be true when an organization is dealing with large complex datasets. As an organization grows so does the amount of data being stored. SQL queries that were once quick and efficient can start to become slow and resource-intensive as the dataset gets larger. There is also the risk of inadvertently altering or losing critical data when using SQL to process and clean datasets. Human error and improper training can lead to incorrect conditions when using an UPDATE or DELETE statement. This corrupts data and makes it unreliable for analysis. Proper training, careful planning, and testing of the SQL queries can help to mitigate these risks.

  One way my current organization helps to elevate these types of risks is by using different environments of data to test and run queries and operations on datasets before releasing them to the production environment. Everything is tested and reviewed before it is merged into the production environment to ensure complete data integrity.

C.

Selecting SQL for data cleaning involves evaluating various factors, such as the nature of the data, the existing infrastructure, the technical expertise of the team, and the specific requirements of the data that needs to be cleaned for the task.

- If the data is stored in a relational database, SQL is the ideal choice due to it being able to interact with such databases efficiently.
- SQL is a well-established language, and many data professionals are familiar with the syntax and formatting of the language. Having people on the team who already know this information can reduce any learning curve that might come with using a tool that the team is not familiar with.
- SQL is compatible with most database management systems or DBMS; this ensures that it can integrated into the existing data workflow with minimal disruption.
- When using efficient and optimized queries SQL can handle volumes of data effectively. Which can make it well suited to deal with medium to large datasets.

SQL provides some powerful capabilities for data manipulation. Filtering, joining, and aggregating data can be critical when it comes to the data-cleaning process. SQL's ability to handle complex queries makes it suitable for identifying and rectifying data quality issues such as duplicates, inconsistencies, and missing values. There are also built-in functions such as sting manipulation and pattern matching which can be particularly useful for standardizing data formats.

Using SQL for data cleaning can provide several key results such as improved data quality, efficiency, and consistency. By using SQL to clean data organizations can ensure that their datasets have improved

reliability and accuracy for analysis, which can lead to better decision making. Tasks that would be time-consuming and probably error-prone if done manually can be automated using SQL which would save time and reduce the likelihood of human error.

Standardizing data formats and cleaning operations using SQL helps maintain consistency across different datasets. This can be critical when an organization needs to combine data from multiple sources, ensuring that all the data is adhering to the same standards.

Using SQL during the data cleaning process can open data to potential ethical problems. Data privacy and security are crucial to any organization to protect its data. If proper access controls are not in place this can lead to unauthorized personnel gaining access to sensitive information, and privacy violations. Organizations must ensure that only authorized and knowledgeable personnel can execute data-cleaning operations. The removal of certain outliers during data cleaning without proper justification can lead to the data being skewed. This would cause an ethical issue since the data is being manipulated in a way that misrepresents the original information. Another ethical implication can happen when a SQL query is written poorly or inefficiently. This can lead to data loss or corruption of the data which can be critical to the decision-making process or compliance.

Being aware of these privacy and ethical concerns, organizations can implement safeguards and use best practices to ensure that SQL is used responsibly and ethically in the data cleaning process. This can include also using strong access control, documenting the data cleaning procedure, and regularly auditing data handling practices.