Team challenges

Title	Overview
1 - Configure the environment and raw import	Review the files that you need to import into Synapse Analytics, configure your solution accordingly, and complete the full import.
2 - Optimize data load	Create a data loading pipeline that provides a repeatable import process and meets the RTO requirements of a 60-minute full rebuild of the warehouse.
3 - Optimize performance of existing queries and create new queries	Uncover query performance issues and craft queries that help WWI unlock new insights into both historical and new data.
4 - Manage and monitor the solution	Protect WWI's data with an end-to-end security configuration for the data warehouse. Address the CIO's concerns about WWI's ability to monitor the data pipeline by providing visibility into each process and configuring alerts as needed.

Introduction

Wide World Importers is in the process of building a new, modern analytics solution. The old analytics solution was built using an older, on-premises version of SQL Server and was based on the relational engine only. WWI's top management expects the new solution to support the strategic move towards near-real-time data analysis.

The core objective of this PoC challenge is to prove that Azure Synapse Analytics is the right platform to be used by the new solution.

Completing the challenges

Work as a team to complete the challenges listed below. Pay attention to the background story for each challenge, as they contain insights into the customer's pain points and what they want to solve. Successful teams collaborate on understanding each challenge, then divide and conquer to work in parallel as much as possible.

You have the freedom to choose the solution your team believes will best fit WWI's needs. However, you must be able to explain the thought process behind the decisions to your coach.

> **IMPORTANT TECHNICAL NOTE**
>
> Your workplace for the challenges is the `wwi_poc` schema in your SQL pool.
>
> In case you need to create additional tables, you must use the same schema (`wwi_poc`) for them.

1 - Configure the environment and raw import

Background story

WWI began an internal initiative to modernize their outdated, on-premises data and analytics platform by migrating a year of historical data to Azure Synapse Analytics. They were sold on the promise of a highly-scalable, thoroughly modern, unified data warehouse and analytics system that can store their old data and capture and process new data as it arrives. Not only that, but they would be able to unlock new insights by integrating non-relational data and be prepared to delve into machine learning when they are ready to tackle that challenge.

However, that promise never came. After importing each month's-worth of sales data, the data engineering team would execute queries over the data set in the SQL pool. After a few iterations of this process, the queries began to be painfully slow to execute. Each iteration of the data import took longer because they truncated the tables and re-imported all of the data up to the next month. Finally, the team gave up on the process and decided to seek outside help.

Wide World Importers wants your help proving that Synapse Analytics is the right platform for their needs. They have invested a fair amount of resources to this project already, and have gotten the go-ahead from leadership and the board to fully migrate to a new solution over the next eight to twelve months.

Technical details

Sales data is currently being inserted into the SQL pool. About 57% of the data is already in the internal tables of the SQL Pool. This roughly covers Jan 2014 to April 2017. The remainder of the data is in external CSV and Parquet files.

One of WWI's large LOB systems had a major outage during the month of May 2017. Data was exported using an alternative approach and is available in CSV files. Starting June 2017, data is available as Parquet files.

Also, customer data is only partially imported. Issues with the processing of customer information prevented a complete import of customer data.

WWI resources

WWI loaded their data to the primary ADLS Gen2 account for the Synapse Analytics workspace.

You can find the CSV files for May 2017 in the following path: `wwi-02 / sale-poc`.

You can find the Parquet files in the following paths:

- `wwi-02 / sale-small / Year=2017 / Quarter=Q2 / Month=6`
- `wwi-02 / sale-small / Year=2017 / Quarter=Q3`
- `wwi-02 / sale-small / Year=2017 / Quarter=Q4`
- `wwi-02 / sale-small / Year=2018`
- `wwi-02 / sale-small / Year=2019`

You can find the complete customer data in the following path: `wwi-02 / data-generators / generator-customer.csv`. The file should be approximately 140 MB in size.

> **IMPORTANT TECHNICAL NOTE**
>
> Do not use other files from the data lake to import sales data as they will invalidate the results of the PoC.

Success criteria

- **All data** is migrated to the SQL pool. This is a raw import, which means that your focus is not on repeatability of the data load process.
- There are no time constraints on the data loading operation, but be mindful of leaving time for the remaining challenges.
    - Consider working with a subset of both the CSV and Parquet files as you iterate through your data loading process. Test your assumptions with sample sizes before loading the entire data set.

2 - Optimize data load

Background story

Importing all of the existing data is only part of the data load story. Wide World Importers has a recovery-time objective (RTO) of 60 minutes for a full rebuild of the warehouse. If, during testing and experimentation, for example, data gets corrupted in the SQL pool, they want to be able to completely rebuild the warehouse for a new iteration of testing. The data engineers lack confidence in the warehouse and have grown accustomed to iterating over snapshots of their data throughout the process. This RTO gives them the confidence to proceed with further testing and configuration of the system.

In addition to the RTO requirements, top management is demanding more and more a departure from the traditional "analyze today, yesterday's data". The goal is to significantly reduce the gap between the moment data is generated and the moment it ends up in dashboards.

Data post May 2017 is now coming in as a continuous stream of Parquet files. Propose and implement a data lake architecture where top management can get data with various compromises between speed of delivery and accuracy/completeness. Provide a bronze level where freshly collected sales data is analyzed using Synapse SQL Serverless and exposed into dashboards. Provide a silver level where data quality has been increased via data engineering. Finally, provide the gold level where top-quality data has been persisted in a Synapse SQL Pool.

Success criteria

- You have created a data loading pipeline that provides a repeatable import process and meets the RTO requirements of a 60-minute full rebuild of the warehouse.
- You have proven this process by wiping out the database (excluding pre-existing data) and conducting a full import with predictable results and processing time.
- You store all raw data in a bronze folder, cleaned up data in a silver folder, and all fully-transformed data stores in the SQL pool.
- The data loading resource takes priority over all other resources connected to the SQL pool.

> **IMPORTANT TECHNICAL NOTE**
>
> Observe the important detail in the second success criteria: you should NOT take into account the pre-existing data when estimating and demonstrating the required RTO.

3 - Optimize performance of existing queries and create new queries

Background story

Wide World Importers currently uses an on-premises SQL Server for their sales data mart. They come from a strictly traditional relational database background and are unfamiliar with MPP systems like Synapse Analytics. As such, they are confused as to why their T-SQL queries are not performing as expected.

These business-critical queries take a long time to complete, which makes them nervous about executing the same queries against a full load of their data. As such, the current query speeds do not meet their business needs, nor do they give WWI confidence that they will be able to realistically execute new types of queries to meet emerging requirements.

Technical details

WWI has supplied the following business-critical queries that they currently use and suffer significant performance issues:

Query wish list

Leadership wants to see some early, tangible benefit from the data modernization effort. They've been sold on the "art of the possible" and how Synapse helps unlock new insights on their data. These queries should have a visual component that gets decision makers excited about where the company is headed and have instant transference of complicated sales data into easy-to-understand market insights. They have described the following queries they'd like to see in the new system:

- What is the dynamic of the year-on-year sales profit across individual countries?
- Which are the most profitable countries?
- What is the evolution rate (increase in frequency of purchases and the overall value of those purchases) month-over-month for customers overall?
    - What is the evolution rate for individual customers?
- We tend to see more sales during the week vs. the weekend. Can you identify the customers who make more than the average number of weekend purchases, where they shop, and what their top products are?
- What percentage of our customers prefer to buy from the same set of stores?
- We have seasonal products that sell very well. We want to know what percentage of our customers have a strong preference for purchasing seasonal products vs. non-seasonal products during the seasonal days? Is there a stronger preference overall for seasonal products sold during one season over the other?
- We would like to see the top 10 products that have a combination of the highest profit and highest number of sales each month.
    - Can you identify products that have the highest profit, but do not meet the top 10 criteria in sales so we can build a "hit list" of products we want to focus on promoting to our customers?

Success criteria

- Queries based off the "query wish list". These queries must execute at human-interactive speeds (ideally in under 1 minute per query).
    - Create compelling visualizations for the new queries.
    - Implement RBAC on the new reports, showing information pertinent to the logged in user, based on the criteria outlined above.

> **INSIDER TIP**
>
> We've learned that WWI is running several PoC projects in parallel. Our sources tell us that one of the most important criteria to select the winner team will be the execution time of these queries. You should do your best to get the shortest execution times possible.

4 - Manage and monitor the solution

Background story

Wide World Importers has a commitment to their customers that they will protect their information from data leakage. They also have a commitment to higher authorities that they will maintain industry-standard compliance in handling and storing data, and to their shareholders that they will be responsible stewards of sensitive competitive data. All of these commitments have understandably given the CISO a lot of angst around ensuring absolute protection of the data and systems that access it, throughout the architecture's components. This new analytics solution is one part of a company-wide security initiative that the CISO is driving. There is an additional level of scrutiny on this project because it is a departure from their trusted on-premises sphere of influence. There is still an inherent mistrust of the cloud by some in WWI's leadership. The CISO wants to be able to assure them that the solution meets several key security requirements, including securing end-to-end data process, from the external files all the way to the serving layer.

In addition, one of the current major pain points for the CIO is the limited capabilities of insights into various data processes. The CIO expects the new solution to significantly increase the visibility into all the data processes developed as part of the new analytical solution.

Technical details

As you and your team plan the security and monitoring aspects of the solution, keep in mind that best practices must be followed for all components of your data pipeline. Data must be encrypted in-transit and at rest. You must follow least privilege access guidelines and make sure that auditing and monitoring are key aspects of your delivery.

Success criteria

- Implement and demonstrate end-to-end security measures for the data warehouse rebuild process.
    - The problem of customer PII (Personally Identifiable Information) is addressed.
    - Least-privilege access is incorporated.
    - Secrets are encrypted and not available in clear text, anywhere in the configuration.
        - Want to maintain exclusive control over the keys to used to encrypt the data warehouse data at rest. They do not want Microsoft or any other entity to provide or have access to these keys.
    - RBAC is implemented at both the data source (only system-level accounts access the source data), and at the serving layer. If you do not fully implement RBAC, you must be able to explain to WWI how you plan to implement it.
    - Will need to prove the flexibility to assign users to groups who should have access to the workspace, as well those that might have elevated permissions, such as those who can administer the entire Synapse Workspace, or manage just the Spark or SQL Pools or use Pipelines.
- Monitor all data processes and react to potential problems that might occur.
    - Monitor for suspicious traffic against the storage account and receive alerts for any flagged incidents.