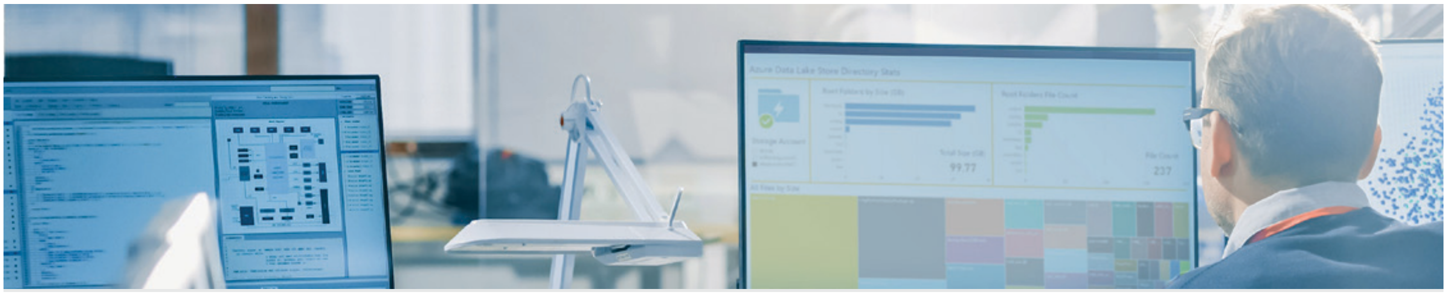


TRANSFORMING DATA ORCHESTRATION TO STREAMLINE DATA AVAILABILITY

Migrating from SQL on premise to SQL Azure

The IT services division a large technology manufacturer supports the maintenance and repair of technology devices sold around the world. They were using a legacy solution to track hardware telemetry data such as ttechnical support and repair data, customer history, and device information. This legacy solution was unable to meet real-time data requirements due to on-premise limitations for data scale and load. There were terabytes worth of data that needed to be optimized for performance and scaled to account for growth. The company needed to ingest data from multiple internal and external vendors, each with a unique data set with distinct schemas, at different intervals. For example, some vendors provided a traditional daily batch schedule while others utilized a micro-batch approach where data sets could-be ingested within minutes or seconds.

The company had stored these data sets with several terabytes worth of data residing on multiple SQL servers. Data issues increased due to the lack of speed for accessing data, making it difficult to support the large workload. The on-premise servers were showing their limitations for future scalability and sustainability day after day.

THE SOLUTION:

The company wanted to begin migrating data providers and consumers from the on-premise SQL based platform to SQL on Azure within 4 months of beginning the data engineering project. Catapult leveraged multiple Microsoft analytics tools such as Azure Data Lake Storage Gen 2, Data Factory, Databricks, and Synapse DW to create a more scalable solution for the massive data project.

Catapult created three parallel tracks: ingestion, user onboarding, and downstream orchestration. The approach enabled the organization to begin receiving data from vendors into the new data lake while still supporting users that were in transition to the environment. In addition, data pipelines were built to allow downstream subscribers to continue receiving curated data sets.

Data is ingested into the Azure Data Lake Gen 2 (bronze zone) using event-based orchestration. The data is transformed, cleansed, and aggregated using Data Factory and Databricks (silver zone). Processed data is delivered to the curated directories in the lake (gold zone) and is available for output to downstream consumers such as Synapse DW, Power BI and Azure Machine Learning.

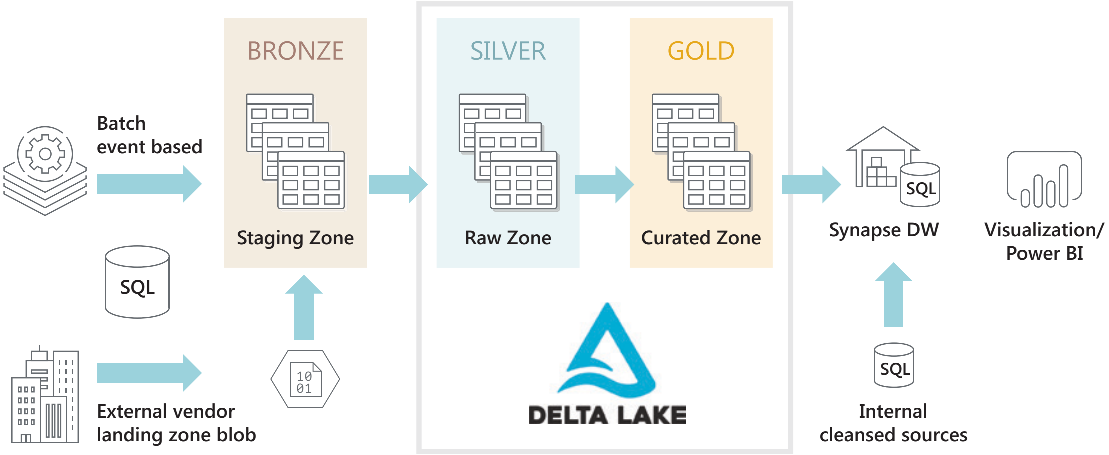
A delta lake was delivered as a layer above the lake to allow data processing to be streamed or batched. A delta lake prevents corruption by supporting atomic transactions that assure changes succeed completely or fail completely.

RESULTS:

Catapult migrated the on-premise data SQL Server to a data warehouse hosted in Azure. This new environment allows the company to scale as needed for business continuity and supports large future workloads. Moving the servers into the cloud also improved access speed for users. The new data lake is also capable of ingesting data regardless of the frequency of the data provided by vendors.

Within 8 months since the launch date, 3.5 trillion records were ingested. In the same period, approximately 1 petabyte of diagnostic telemetry data has been ingested into the new lake. On any given day, data providers may trigger several hundred or several thousand events. Each data set ingested triggers an event and each event may be associated with millions of records to process.

- Unified data and AI with one consolidated repository on Azure for users
- Disaster Recovery support maintained by Azure storage
- Data snapshots enabling consumers to access and revert to earlier versions
- Accurate and consistent data availability even in the case of a cluster or hardware failure
- Azure Event Grid and Data Factory to process data in real time as it is ingested
- A real time event-based ingestion solution delivers a real time view of data sources
- Bring your own data (BYOD) capability enables ad hoc analysis and machine learning
- A flexible design allows the organization to prioritize which data sets need to be available


Batch
event based
SQL
External vendor
landing zone blob
10
01
BRONZE
Staging Zone
SILVER
Raw Zone
GOLD
Curated Zone
DELTA LAKE
SQL
Synapse DW
SQL
Internal
cleansed sources
Visualization/
Power BI