Data warehousing and analytics for sales and marketing
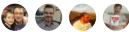
09/15/2018 • 5 minutes to read • Contributors 🧑 🧑 🧑 🧑

In this article

This example scenario demonstrates a data pipeline that integrates large amounts of data from multiple sources into a unified analytics platform in Azure. This specific scenario is based on a sales and marketing solution, but the design patterns are relevant for many industries requiring advanced analytics of large datasets such as e-commerce, retail, and healthcare.

This example demonstrates a sales and marketing company that creates incentive programs. These programs reward customers, suppliers, salespeople, and employees. Data is fundamental to these programs, and the company wants to improve the insights gained through data analytics using Azure.

The company needs a modern approach to analysis data, so that decisions are made using the right data at the right time. The company's goals include:

- Combining different kinds of data sources into a cloud-scale platform.
- Transforming source data into a common taxonomy and structure, to make the data consistent and easily compared.
- Loading data using a highly parallelized approach that can support thousands of incentive programs, without the high costs of deploying and maintaining on-premises infrastructure.
- Greatly reducing the time needed to gather and transform data, so you can focus on analyzing the data.

Relevant use cases

This approach can also be used to:

- Establish a data warehouse to be a single source of truth for your data.
- Integrate relational data sources with other unstructured datasets.
- Use semantic modeling and powerful visualization tools for simpler data analysis.

Architecture
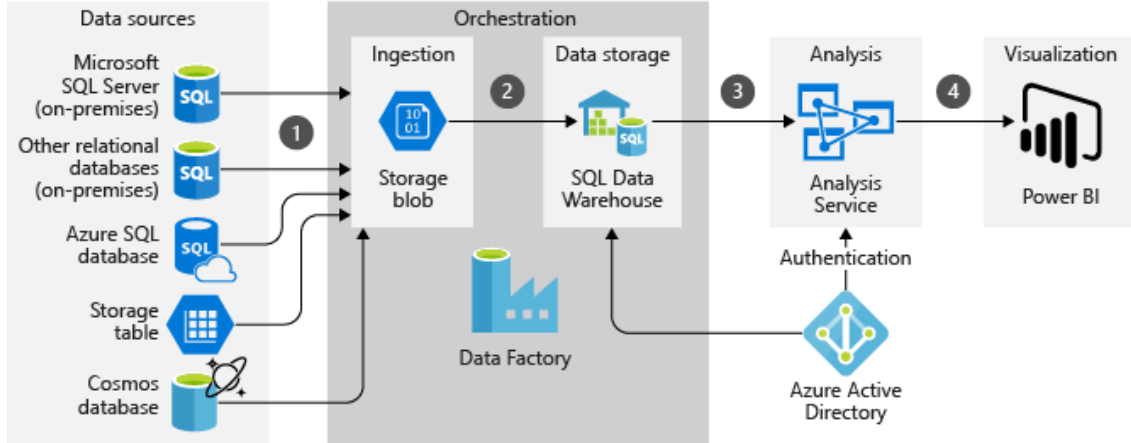
The data flows through the solution as follows:

1. For each data source, any updates are exported periodically into a staging area in Azure Blob storage.
2. Data Factory incrementally loads the data from Blob storage into staging tables in SQL Data Warehouse. The data is cleansed and transformed during this process. Polybase can parallelize the process for large datasets.
3. After loading a new batch of data into the warehouse, a previously created Analysis Services tabular model is refreshed. This semantic model simplifies the analysis of business data and relationships.
4. Business analysts use Microsoft Power BI to analyze warehoused data via the Analysis Services semantic model.

Components

The company has data sources on many different platforms:

- SQL Server on-premises
- Oracle on-premises
- Azure SQL Database
- Azure table storage
- Cosmos DB

Data is loaded from these different data sources using several Azure components:

- Blob storage is used to stage source data before it's loaded into SQL Data Warehouse.
- Data Factory orchestrates the transformation of staged data into a common structure in SQL Data Warehouse. Data Factory uses Polybase when loading data into SQL Data Warehouse to maximize throughput.
- SQL Data Warehouse is a distributed system for storing and analyzing large datasets. Its use of massive parallel processing (MPP) makes it suitable for running high-performance analytics. SQL Data Warehouse can use PolyBase to rapidly load data from Blob storage.
- Analysis Services provides a semantic model for your data. It can also increase system performance when analyzing your data.
- Power BI is a suite of business analytics tools to analyze data and share insights. Power BI can query a semantic model stored in Analysis Services, or it can query SQL Data Warehouse directly.
- Azure Active Directory (Azure AD) authenticates users who connect to the Analysis Services server through Power BI. Data Factory can also use Azure AD to authenticate to SQL Data Warehouse via a service principal or Managed identity for Azure resources.

Alternatives

- The example pipeline includes several different kinds of data sources. This architecture can handle a wide variety of relational and non-relational data sources.

- Data Factory orchestrates the workflows for your data pipeline. If you want to load data only one time or on demand, you could use tools like SQL Server bulk copy (bcp) and AzCopy to copy data into Blob storage. You can then load the data directly into SQL Data Warehouse using Polybase.

- If you have very large datasets, consider using Data Lake Storage, which provides limitless storage for analytics data.

- An on-premises SQL Server Parallel Data Warehouse appliance can also be used for big data processing. However, operating costs are often much lower with a managed cloud-based solution like SQL Data Warehouse.

- SQL Data Warehouse is not a good fit for OLTP workloads or data sets smaller than 250GB. For those cases you should use Azure SQL Database or SQL Server.

- For comparisons of other alternatives, see:
  - Choosing a data pipeline orchestration technology in Azure
  - Choosing a batch processing technology in Azure
  - Choosing an analytical data store in Azure
  - Choosing a data analytics technology in Azure

Considerations

The technologies in this architecture were chosen because they met the company's requirements for scalability and availability, while helping them control costs.

- The massively parallel processing architecture of SQL Data Warehouse provides scalability and high performance.
- SQL Data Warehouse has guaranteed SLAs and recommended practices for achieving high availability.
- When analysis activity is low, the company can scale SQL Data Warehouse on demand, reducing or even pausing compute to lower costs.
- Azure Analysis Services can be scaled out to reduce response times during high query workloads. You can also separate processing from the query pool, so that client queries aren't slowed down by processing operations.
- Azure Analysis Services also has guaranteed SLAs and recommended practices for achieving high availability.
- The SQL Data Warehouse security model provides connection security, authentication and authorization via Azure AD or SQL Server authentication, and encryption. Azure Analysis Services uses Azure AD for identity management and user authentication.

Pricing

Review a pricing sample for a data warehousing scenario via the Azure pricing calculator. Adjust the values to see how your requirements affect your costs.

- SQL Data Warehouse allows you to scale your compute and storage levels independently. Compute resources are charged per hour, and you can scale or pause these resources on demand. Storage resources are billed per terabyte, so your costs will increase as you ingest more data.
- Data Factory costs are based on the number of read/write operations, monitoring operations, and orchestration activities performed in a workload. Your Data Factory costs will increase with each additional data stream and the amount of data processed by each one.
- Analysis Services is available in developer, basic, and standard tiers. Instances are priced based on query processing units (QPUs) and available memory. To keep your costs lower, minimize the number of queries you run, how much data they process, and how often they run.
- Power BI has different product options for different requirements. Power BI Embedded provides an Azure-based option for embedding Power BI functionality inside your applications. A Power BI Embedded instance is included in the pricing sample above.

Next Steps

- Review the Azure reference architecture for automated enterprise BI, which includes instructions for deploying an instance of this architecture in Azure.

- Read the Maritz Motivation Solutions customer story. That story describes a similar approach to managing customer data.
- Find comprehensive architectural guidance on data pipelines, data warehousing, online analytical processing (OLAP), and big data in the Azure Data Architecture Guide.