Integrating Data in a Multi-cloud Environment

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Azure SQL Data Warehouse for business analytics

Azure Data Factory for integrating data sources

Eliminating data silos

Integrating data across multiple cloud platforms

Data Silo

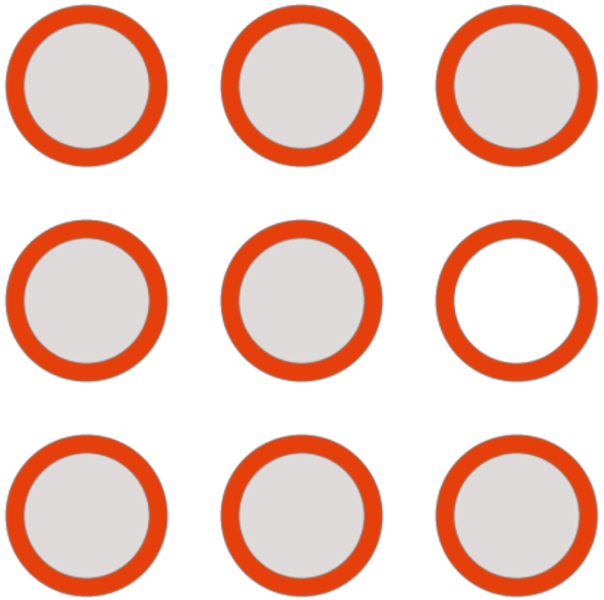
An isolated repository of enterprise data, unconnected to other data repositories, and unavailable for use by most users in the organization

Common Forms of Data Silos

Standalone relational databases

Block storage such as persistent disks

Data warehouses holding semi-structured data



Data silos pose many serious problems

- No single source of truth
- Stale, out-of-date, modified data
- Hard to connect the dots
- Political turf battles over ownership
- Storage costs can be significant
- Audit and other gray areas

Data silos are a big problem - solve either by integrating the silos, or by using data lakes

Data Lake

A single repository for all enterprise data - structured or unstructured, batch or streaming, raw or transformed, on-cloud or on-premises

Data Warehouse

Structured data store used for analytical processing and reporting; usually hold transformed data fed in from disparate sources via ETL pipelines

ETL Pipelines

Programs or scripts with business logic to automatedly *e*xtract data from disparate sources, *t*ransform it to satisfy a schema, then *l*oad it into a data warehouse

Two Approaches to Solving Silos

ETL Pipelines + Data Warehouses

Connect up silos

Extract-Transform-Load (ETL) pipelines

Output of ETL pipelines into data warehouse

Fine for hybrid on-cloud and on-premises

Integration and phased migration of legacy data

Data Lakes

Eliminate silos entirely

Single repository for all enterprise data

Data lake holds raw data in unprocessed form

Best for cloud-first and cloud-only

One-off migration of legacy data

Azure Data Products

Azure SQL Database

Azure Data Warehouse

Azure Data Lake

Azure Data Factory

Power BI

Azure SQL Database

Managed relational database service on the Microsoft Azure cloud platform ideal for transaction processing applications.

Azure Data Warehouse

Azure's flagship data warehouse offering that competes directly with Google BigQuery and Amazon Redshift.

Azure Data Lake

Platform-as-a-Service offering that ties Azure Data Lake Storage and Data Lake Analytics; supports language called U-SQL.

Azure Data Factory

Managed service meant for building complex, hybrid ETL pipelines that integrate data silos and can include Hadoop and Machine Learning transformations.

Power BI

Business analytics app with powerful visualization and data exploration capabilities; closely integrated with Microsoft and Azure data services.

Two Approaches to Solving Silos

ETL Pipelines + Data Warehouses

Azure Data Factory and Azure Data Warehouse

Data Lakes

Azure Data Lake

Data lakes eliminate data silos

Single source of truth

Hold raw data in all formats

Cost-effective

Access and audit controls built-in

Visualization and insight generation

Data Lakes vs. Data Warehouses

Data Lakes	Data Warehouses
Raw data in native form	Ingested data in transformed form
Schemaless, or schema-on-read	Predefined schema-on-write
Can hold entirely unstructured data	Structured or semi-structured data
All data in enterprise	Data required for analytics (OLAP)
Optimized for fast ingestion, cheap storage	Optimized for structured retrieval and analytics

Demo

Provisioning an Azure SQL Data Warehouse

Demo

Moving data from blob storage to Azure SQL Data Warehouse

Demo

Moving data from AWS S3 storage to Azure SQL Data Warehouse

Summary

Azure SQL Data Warehouse for business analytics

Azure Data Factory for integrating data sources

Eliminating data silos

Integrating data across multiple cloud platforms