Azure Data Lake Zone and Tool Selection

Version 1.0

Proposed Options

Full Diagram | "The Kitchen Sink"

Option 1 | Current Zones

Option 2 |Pause Curated Zone

Option 3 | Future Curated and Exploratory Zones

Options | Decision Making Questions

Why is there a curated zone?

What business needs does the curated zone resolve, or provide?

What value (today) does the curated zone provide?

Is there an immediate, or within 1 to 2 year need to hire data scientists?

Can the loading of SQL Data Warehouse bypass the curated zone?

Can validation be moved to the staging zone?

Can the Trade Compliance FTP process be moved to the staging zone?

Do you have an Archive Data Zone?

Zone Explanation and Purpose

Raw Data Zone

- Exact copy of source data in native format (aka master dataset in the batch layer)
- Immutable to change
- History retained indefinitely
- Data access is highly limited to a few users
- Everything downstream can be regenerated from the Raw Zone

Transient / Temp Zone

- Selectively utilized
- Separation of "new data" from "raw data" to ensure data consistency
- Transient low-latency data (aka speed layer)
- Data duality validations

Master Data Zone

- Reference Data

User Drop Zone

- Manually generated data

Staged Data Zone

- Data staged for a specific purpose or application

Standardized Raw Data Zone

- Raw data which varies in format, or schema, such as JSON which is standardized into columns and rows (aka "semantic normalization")
- File consolidations of data (i.e. to overcome performance issues with many small files

Archive Data Zone

- Active archive of aged data, available for querying when needed

Analytics Sandbox Zone

- Workspace for exploratory data science and analytics
- Valuable efforts are productionized into the Curated Data Zone

Curated Data Zone

- Cleansed and transformed data, organized for optimal data delivery (aka serving layer)
- Supports "Big Data Self Service"
    - Design elements for creating a common data platform for all data.
      This is a 5 to7 year journey with the goal for business user to self-deploy
      working with IT teams as needed to run analytics and reporting
      on top of the common data platform – Ralph Kimball
- Standard security, change management, and governance

Azure Tool Explanation and Purpose

Data Lake Storage
Massively scalable data lake storage

Data Lake Analytics
Distributed analytics service that makes big data easy

SQL Data Warehouse
Elastic data warehouse as a service with enterprise-class features (MMP)

SQL Database
Managed relational SQL database as a service

Data Factory
Hybrid data integration at enterprise scale, made easy

Databricks
Fast, easy, and collaborative Apache Spark-based analytics platform

IoT Hub, IoT Edge, and IoT Central
Connect, monitor, and manage billions of IoT assets including the cloud and SaaS

Machine Learning
Open and elastic AI development spanning the cloud and the edge

Event Hub
Receive telemetry from millions of devices

Stream Analytics
Real-time data stream processing from millions of IoT Technologies

Analysis Services
Enterprise-grade analytics engine as a service

Power BI
Suite of business analytics tools that deliver insights throughout your organization

Data Catalog
Get more value from your enterprise data assets by cataloging your meta-data

Key Vault
Safeguard and maintain control of keys and other secrets

Scheduler
Run your jobs on simple, or complex recurring schedules

** Below | Not on the Diagrams **

Storage
Durable, highly available, and massively scalable cloud storage (blob, file, tables, and queues)

SQL Server Stretch Database
Dynamically stretch on-premises SQL Server databases to Azure

SQL Elastic Pools
Simple cost-effective solution for managing and scaling multiple databases that have varying and unpredictable demand

ExpressRoute
Dedicated private network fiber connections to Azure

Azure Functions
Process events with serverless code

Data Factory Integration Runtimes
- SSIS Integration Runtime

On-Premises Data Gateways
- Power BI
- Data Factory