Design and implemen

Azure Storage Accounts



This is a service on the Azure platform that allows you to store different data objects.



Blob storage – This is Microsoft's object storage solution on the cloud.



You can use this service for storing large amounts of unstructured data.

Azure Data Lake Storage Gen2



This service provides features that are pertinent to big data analytics.



This service is built on top of Azure Blob Storage.



You can use this as a central repository for storing your structured and unstructured data.



Azure Synapse



This is an enterprise analytics service.



Here you can store your data in data warehouses and query the data accordingly.



It also has Spark capabilities and Pipelines for data integration.



Azure Data Factory



This is a cloud-based ETL tool. It also acts as a data integration service.



You can create workflows when it comes to orchestrating data movement and transforming data.



Has support for a variety of data stores.



Azure Stream Analytics



This is a fully managed stream processing engine.



Here you can analyze and process large volumes of streaming data.



You can ingest data from different data sources.



Azure Databricks



This is a set of tools that can be used for building, deploying, sharing and maintaining enterprise-grade data solutions at scale.



You can use this service to process and analyze your data..



This service makes use of Apache Spark that can be used to process data.



Design and implement data storage

- Azure Synapse Analytics

Best Practices

Design

Distributed Tables

Hash-distribution

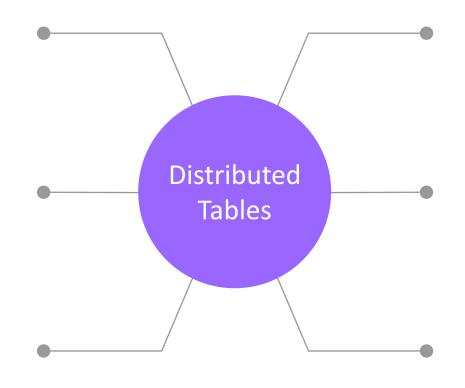
This improves query performance on large tables.

Round-Robin distribution

This helps in improving loading speed.

Data Movement

With hash-distribution,
SQL Analytics has
knowledge on the rows in
the distribution.



Choose Round-Robin

No joining key, no good candidate for hash distribution.

Fact Tables

Best to use hashdistribution.

Dimension

Best to use replicated tables.



Design

Hash-distributed Column

1 Choosing the right column

Ensure that data gets distributed across the distributions

2 Data Skew

If the data is not distributed properly it can lead to data skew.

3 Data Movement

Choose a column that minimizes data movement – Is used in JOIN, GROUP BY clauses.



Design and Develop Data Processing

- Azure Data Factory

Mapping Data Flows

Mapping Data Flows



This feature helps to visualize the data transformations in Azure Data Factory.



You can write the required transformation logic without actually writing any code.



The data flows run on Apache Spark clusters. Azure Data Factory will manage the transformations in the data flow.



Mapping Data Flows



Debug Mode – You can actually see the results of the data flow while designing the flow.



In the debug mode session, the data flow will run interactively on the Spark cluster.



In the debug mode, you will be charged on an hourly basis for the active cluster.



Design and Develop Data Processing - Scala, Notebooks and Spark

Lake databases

Lake Databases

1 Spark pool

These are the databases and tables created in the Spark pool.

2 Shared

These tables can also be accessed from the Serverless SQL pool.

3 Data

Here the data for the databases are stored in data lake storage. Supported file formats are Parquet, Delta and CSV.

Design and Develop Data Processing

- Azure Databricks

More on clusters

Databricks

Clusters

Compute

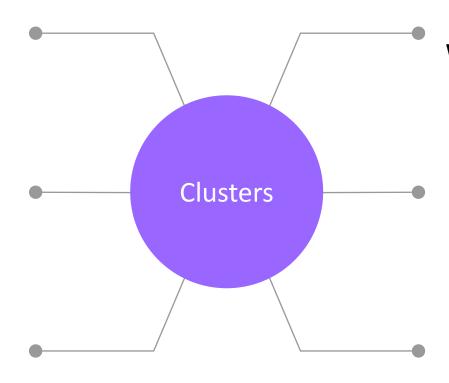
Your workloads run on the cluster compute resources.

Notebooks

You define the workloads as a set of commands in Notebooks.

Cluster types

You have your all-purpose clusters and the job clusters.



Terminate

When you terminate the cluster, the details of the cluster are retained for 30 days.

Pin

Pin the cluster if you want the details of the terminated cluster for a longer duration.

Job Cluster

Here the cluster terminates once the job is complete.

Cluster Access Mode

1 Single User

Supported Languages – Python, SQL, Scala, R

2 Shared

Needs a Premium plan. You can have data isolation amongst users. Python and SQL supported.

3 No Isolation Shared

Supported Languages – Python, SQL, Scala, R.

Cluster Access Mode

1 Driver node

This maintains the state information for the notebooks attached to the cluster. It also maintains the SparkContext.

Worker node

This runs the executors for running the distributed workloads.

3 Spot Instances

To save on costs, you can also make use of Spot Instances.

Design and Implement Data Security

Dynamic Data Masking

Dynamic Data Masking

Exposure

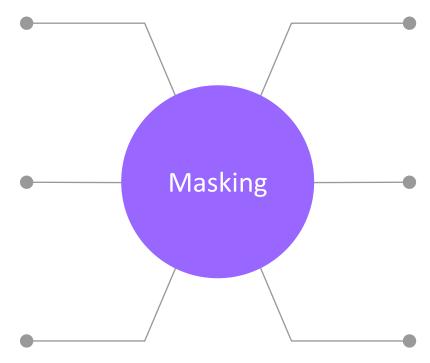
Here you can limit the exposure of data.

Rule

You can create rules to mask the data.

Credit Card masking rule

This is used to mask the column that contain credit card details. Here only the last four digits of the field are exposed.



Email

Here first letter of the email address is exposed. And the domain name of the email address is replaced with XXX.com.

Custom text

Here you decide which characters to expose for a field.

Random number

Here you can generate a random number for the field.



Data classification

Data

Classification

1 Classify

This feature provides capabilities for discovering, classifying, labelling, and reporting the sensitive data in your databases.

2 Discovery

The data discovery feature can scan the database and identify columns that contains sensitive data. You can then view and apply the recommendations accordingly.

3 Sensitivity labels

You can then apply sensitivity labels to the column. This helps to define the sensitivity level of the data stored in the column.

Monitor and optimize data storage and data processing

Microsoft Purview

Microsoft Purview



This is a unified data governance service that helps to track your data assets across multi-cloud, on-premises and software-as-a-service data stores.



Data Map – This helps to provide metadata about your enterprise data.



Data Estate Insights – This can give an overview of your data.



Azure Stream Analytics Metrics

Monitoring

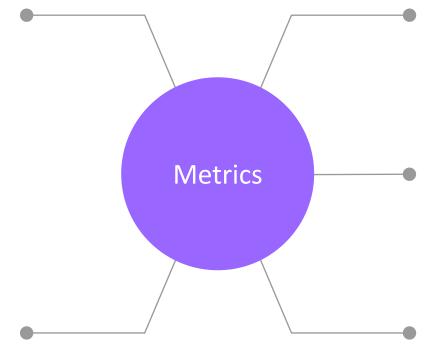
Azure Stream Analytics

Backlogged Input Events

Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. For this you can decide to scale up the number of streaming units.

Data Conversion Errors

Number of output events that could not be converted to the expected output schema.



Early Input Events

Events whose application timestamp is earlier than their arrival time by more than 5 minutes.

Late Input Events

Events that arrived later than the configured late arrival tolerance window.

Out-of-Order Events

Number of events received out of order that were either dropped or given an adjusted timestamp, based on the Event Ordering Policy.



Azure Stream Analytics Monitoring

1 Watermark delay

The maximum watermark delay across all partitions of all outputs in the job.

If the value of the Watermark Delay is greater than 0, it could be due to many reasons.

2 Delays

Inherent processing delay of the streaming pipeline. Clock skew of the processing node generating the metric

3 Resources

There are not enough processing resources in your Stream Analytics job.