

Create Deployment Architecture

Capstone Step 7

Decision Making Process

This process was fairly straightforward, as I now have experience using each of the tool required.

1. I did have to decide between using **Apache Airflow** vs **Azure Data Factory** for orchestration of ETL pipeline tasks

I chose **Azure Data Factory** because I knew it would ease any pain encountered when authenticating connections between services

2. I successfully imported and queried the data using both:

Databricks using a Pyspark kernel

Azure Data Studio using a SQL kernel

The speeds were roughly the same, but VM core quota issues with **Databricks** frustrated me, and I chose to serve the data with **Azure Data Studio**

3. I was unable to create a **Microsoft Power BI** resource without using an organization email account

Provisioning business intelligence software is overkill for my requirements, as **Azure Data Studio** extension **SandDance** accommodates plotting.

4. I then realized that if and when I want to move to a non-Azure infrastructure, I would want to be cloud-agnostic,

so I converted **Azure Data Studio** + **SandDance** to **MySQL** + **Jupyter**

Comparing speeds with **Databricks** again shows **MySQL** to be superior (at least for the resources I have configured):

Benchmarking MySQL vs Databricks (pypark.sql query)

[illegible]

MySQL Workbench

AzureMySQL AzureMySQL AzureMySQL

File Edit View Query Database Server Tools Scripting Help

count_yearmonth_dets SQL File 1*

Limit to 1000 rows

```

1 • SELECT
2   CONCAT(BEGIN_YEARMONTH,BEGIN_NEWID) AS BEGIN_FULLDATE,
3   CONCAT(END_YEARMONTH,END_NEWID) AS END_FULLDATE
4 FROM (SELECT
5   BEGIN_YEARMONTH,
6   CASE WHEN CHAR_LENGTH(BEGIN_DAY) = 1 THEN CONCAT('0',BEGIN_DAY) ELSE BEGIN_DAY END AS BEGIN_NEWID,
7   END_YEARMONTH,
8   CASE WHEN CHAR_LENGTH(END_DAY) = 1 THEN CONCAT('0',END_DAY) ELSE END_DAY END AS END_NEWID
9 FROM defaultdb.details) d
  
```

Query Statistics

Timing (as measured at client side):
Execution time: 0:00:0.15163422

Timing (as measured by the server):
Execution time: 0:00:0.00151632
Table lock wait time: 0:00:0.00000000

Errors:
Had Errors: NO
Warnings: 0

Joins per Type:
Full table scans (Select_scan): 1
Joins using table scans (Select_full_join): 0
Joins using range search (Select_full_range_join): 0
Joins with range checks (Select_range_check): 0
Joins using range (Select_range): 0

Sorting:
Sorted rows (Sort_rows): 0
Sort merge passes (Sort_merge_passes): 0
Sorts with ranges (Sort_range): 0
Sorts with table scans (Sort_scan): 0

Index Usage:
No Index used

Other Info:
Event Id: 47
Thread Id: 272

Result 1

Query Completed

Read Only

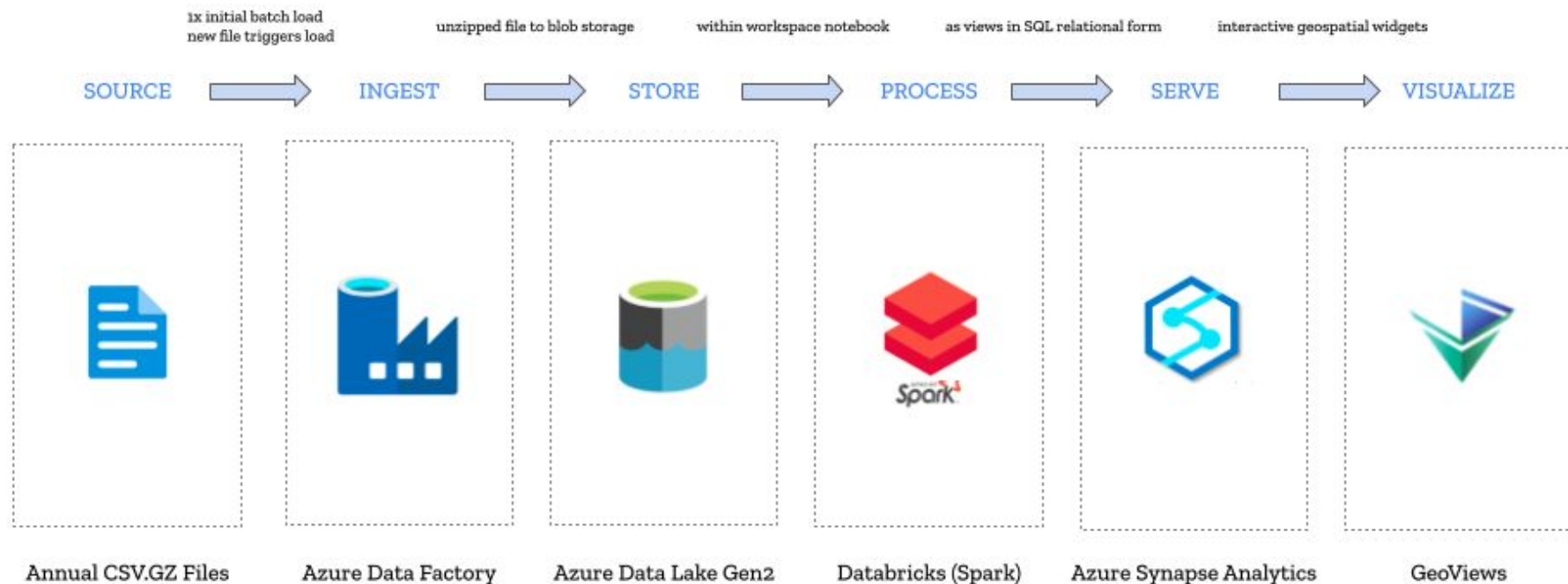
Same Query
Same Results
MySQL 75% Faster

.15 secs
s -> .65 secs

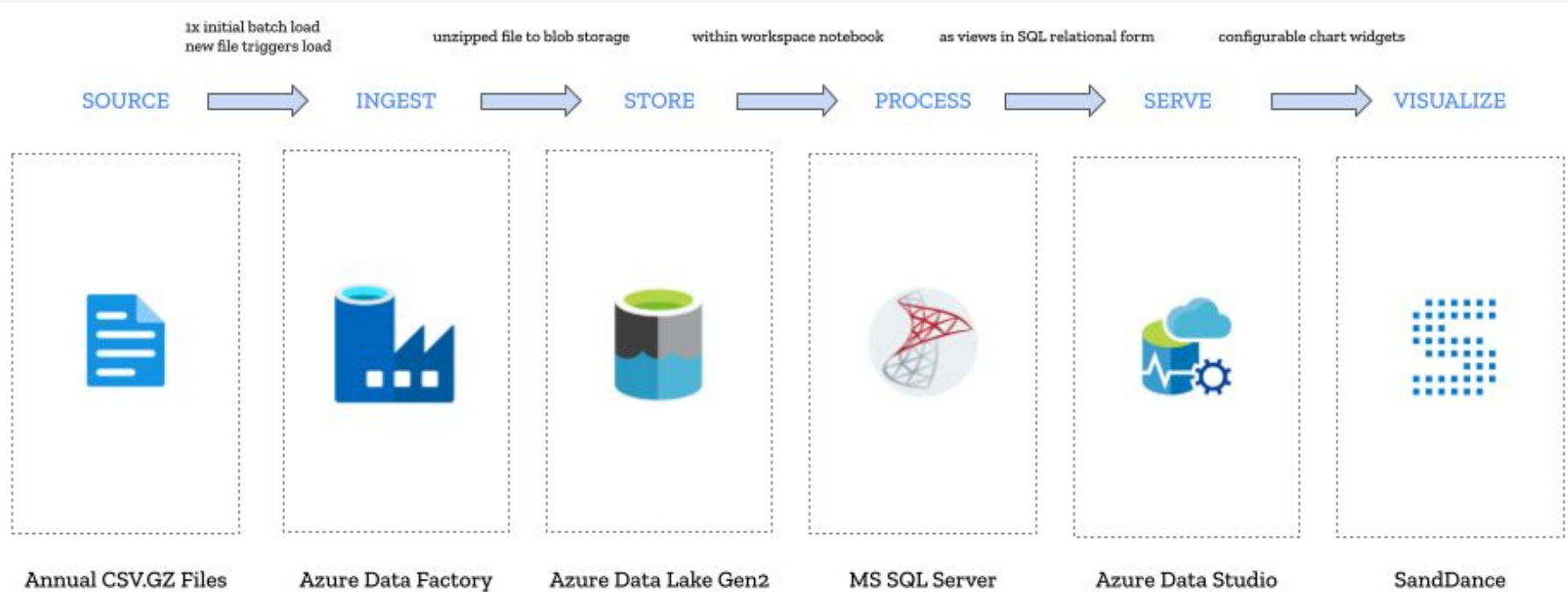
Benchmarking MySQL vs Databricks (dataframe query)

[illegible]

Initial Architecture



Intermediate Architecture



Final Decision Architecture

1x initial batch load
new file triggers load

unzipped file to blob storage

within workspace notebook

as views in SQL relational form

plot using python libraries via notebook

SOURCE



INGEST



STORE



PROCESS



SERVE



VISUALIZE



Annual CSV.GZ Files



Azure Data Factory



Azure Data Lake Gen2



MySQL



MySQL Workbench



Jupyter Notebooks

Resource Deployment Exports

The ability to deploy and export the resources was possible for all resources except **Azure Data Factory**, which lacks the feature.

```
databricks_template.json
{
  "$schema": "https://schema.management.azure.com/schemas/2019-04-01/deploymentTemplate.json#",
  "contentVersion": "1.0.0.0",
  "parameters": {
    "workspaces.databricksname": {
      "defaultValue": "databricksname",
      "type": "String"
    }
  },
  "resources": [
    {
      "apiVersion": "2021-04-01-preview",
      "location": "australiaeast",
      "name": "[parameters('workspaces.databricksname')]",
      "properties": {
        "authorizations": [
          {
            "principalId": "9a7da9f6-d333-4348-988a-e267929b0eb0",
            "roleDefinitionId": "b6a6657-a8ff-443c-a75c-2f8dcdb35",
            "createdBy": {},
            "createdDateTime": "2022-06-13T12:09:33.6569284Z",
            "managedResourceGroup": "[concat('/subscriptions/55f807fa-c5eb-44dd-8357-76a176efc946/resourceGroups/databricks-rg-', parameters('workspaces.databricksname'), '-yadcdkhubois')]",
            "parameters": {
              "enablePublicIp": {
                "type": "Bool",
                "value": false
              },
              "natGatewayName": {
                "type": "String",
                "value": "nat-gateway"
              },
              "requireEncryption": {
                "type": "Bool",
                "value": false
              },
              "publicIpName": {
                "type": "String",
                "value": "nat-gw-public-ip"
              },
              "requireInfrastructureEncryption": {
                "type": "Bool",
                "value": false
              },
              "storageAccountName": {
                "type": "String",
                "value": "dstorage2cd76pvhsk"
              },
              "storageAccountSkuname": {
                "type": "String",
                "value": "Standard_GS"
              },
              "webAddressPrefix": {
                "type": "String",
                "value": "10.139"
              },
              "updatedBy": {}
            },
            "sku": {
              "name": "standard"
            },
            "type": "Microsoft.Databricks/workspaces"
          }
        ],
        "variables": {}
      }
    }
  ]
}
```

```
storage_template.json
{
  "$schema": "https://schema.management.azure.com/schemas/2019-04-01/deploymentTemplate.json#",
  "contentVersion": "1.0.0.0",
  "parameters": {
    "variables": {}
  },
  "resources": [
    {
      "type": "Microsoft.Storage/storageAccounts",
      "apiVersion": "2022-09-01",
      "name": "pipelinestorageaccount",
      "location": "australiaeast",
      "sku": {
        "name": "Standard_IR5",
        "tier": "Standard"
      },
      "kind": "StorageV2",
      "properties": {
        "endpointsType": "Standard",
        "defaultObjectEncryption": false,
        "publicNetworkAccess": "Enabled",
        "allowCrossTenantReplication": true,
        "minimumTlsVersion": "TLS1.2",
        "allowBlobPublicAccess": true,
        "allowSharedKeyAccess": true,
        "isInferencing": true,
        "networkDefault": {
          "type": "AzureServices",
          "virtualNetworkRules": [
            {
              "ipAddress": [
                "defaultAction": "Allow"
              ],
              "supportHttpTrafficOnly": true,
              "encryption": {
                "requireInfrastructureEncryption": false,
                "services": [
                  {
                    "keyType": "Account",
                    "enabled": true
                  },
                  {
                    "keyType": "Account",
                    "enabled": true
                  }
                ],
                "blob": {
                  "keyType": "Account",
                  "enabled": true
                },
                "keySource": "Microsoft.Storage"
              },
              "accessTier": "Hot"
            }
          ]
        },
        "type": "Microsoft.Storage/storageAccounts/blobServices",
        "apiVersion": "2022-09-01",
        "name": "pipelinestorageaccount/default",
        "dependsOn": [
          "[resourceId('Microsoft.Storage/storageAccounts', 'pipelinestorageaccount')]"
        ],
        "sku": {
          "name": "Standard_IR5",
          "tier": "Standard"
        },
        "properties": {
          "changeFeed": {
            "enabled": false
          },
          "restorePolicy": {
            "enabled": false
          },
          "containerDeleteRetentionPolicy": {
            "enabled": false
          },
          "cors": {
            "corsRules": [
              {
                "deleteRetentionPolicy": {
                  "allowDeleteRetentionPolicy": false,
                  "enabled": false
                }
              }
            ]
          }
        },
        "variables": {}
      }
    }
  ]
}
```

```
synapse_template.json
{
  "$schema": "https://schema.management.azure.com/schemas/2019-04-01/deploymentTemplate.json#",
  "contentVersion": "1.0.0.0",
  "parameters": {
    "workspaces.synapseaccountname": {
      "defaultValue": "synapseaccount",
      "type": "String"
    },
    "storageaccount.pipelinestorageaccount.externalId": {
      "defaultValue": "/subscriptions/55f807fa-c5eb-44dd-8357-76a176efc946/resourceGroups/storagexpaas-rg/",
      "type": "String"
    },
    "variables": {}
  },
  "resources": [
    {
      "type": "Microsoft.Synapse/workspaces",
      "apiVersion": "2021-06-01",
      "name": "[concat(parameters('workspaces.synapseaccountname'))]",
      "location": "australiaeast",
      "identity": {
        "type": "SystemAssigned"
      },
      "properties": {
        "defaultDataLakeStorage": {
          "resourceId": "[parameters('storageaccounts.pipelinestorageaccount.externalId')]",
          "createOnDemand": true,
          "accountUrl": "https://pipelinestorageaccount.dfs.core.windows.net",
          "filesystem": "pipelinestorageaccount",
          "encryption": {
            "type": "Microsoft.Storage/storageAccounts/blobServices",
            "apiVersion": "2022-09-01",
            "name": "[concat(parameters('workspaces.synapseaccountname'), '-dev.azure.synapse')]",
            "initialWorkspacesObjectID": "4f82496c-d58b-48f0-a02f-5226f72dc6d4",
            "azureADOnlyAuthentication": false,
            "trustedServiceBypassEnabled": false
          },
          "type": "Microsoft.Synapse/workspaces/auditingSettings",
          "apiVersion": "2021-06-01",
          "name": "[concat(parameters('workspaces.synapseaccountname'), '/default')]",
          "dependsOn": [
            "[resourceId('Microsoft.Synapse/workspaces', parameters('workspaces.synapseaccountname'))]"
          ],
          "properties": {
            "retentionDays": 9,
            "auditActionGroups": [
              {
                "isStorageSecondaryFile": false,
                "isAzureMonitorExportEnabled": false,
                "state": "Disabled"
              },
              {
                "storageAccountSubscriptionId": "00000000-0000-0000-0000-000000000000"
              }
            ]
          }
        },
        "type": "Microsoft.Synapse/workspaces/azureADOnlyAuthentications",
        "apiVersion": "2021-06-01",
        "name": "[concat(parameters('workspaces.synapseaccountname'), '/default')]",
        "dependsOn": [
          "[resourceId('Microsoft.Synapse/workspaces', parameters('workspaces.synapseaccountname'))]"
        ]
      }
    }
  ]
}
```