Niloofar Moosavi

Big Data and Cloud - DE23

Architecture

The key components include:

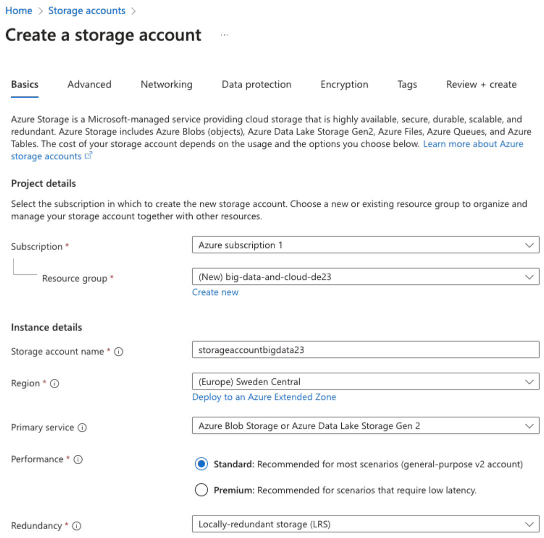
1. **Azure Subscription**: A subscription was created to manage all the resources.

2. **Azure Resource Group**: A resource group was established to organize all related services.

3. **Azure Storage Account**: A storage account was created with three containers:

   - **StructuredData**: Contains a CSV file with sales data from Amazon, sourced from Kaggle.
   - **UnstructuredData**: Contains an XML log file that was generated in Python.
   - **Curated**: Transformed data is stored here in Parquet format.

   Creating Azure Data Lake Storage:

   Create a storage account

   Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. Learn more about Azure storage accounts

   Project details

   Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

   Subscription *: Azure subscription 1

   Resource group *: (New) big-data-and-cloud-de23

   Instance details

   Storage account name *: storageaccountbigdata23

   Region *: (Europe) Sweden Central

   Primary service: Azure Blob Storage or Azure Data Lake Storage Gen 2

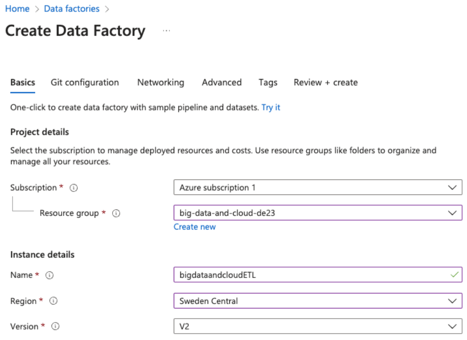
   Performance *:
   ● **Standard:** Recommended for most scenarios (general-purpose v2 account)
   ○ **Premium:** Recommended for scenarios that require low latency.

   Redundancy *: Locally-redundant storage (LRS)

4. **Azure Data Factory (ADF)**: ADF is utilized to orchestrate data movement and transformation. Data from the storage account is extracted, transformed, and loaded into the curated folder in Parquet format.

   Creating Azure Data Factory:

   - The screenshot below demonstrates the initial setup process for ADF, including selecting the appropriate resource group and naming the data factory.



One-click to create data factory with sample pipeline and datasets. Try it

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * : Azure subscription 1

Resource group * : big-data-and-cloud-de23

Create new

Instance details

Name * : bigdataandcloudETL

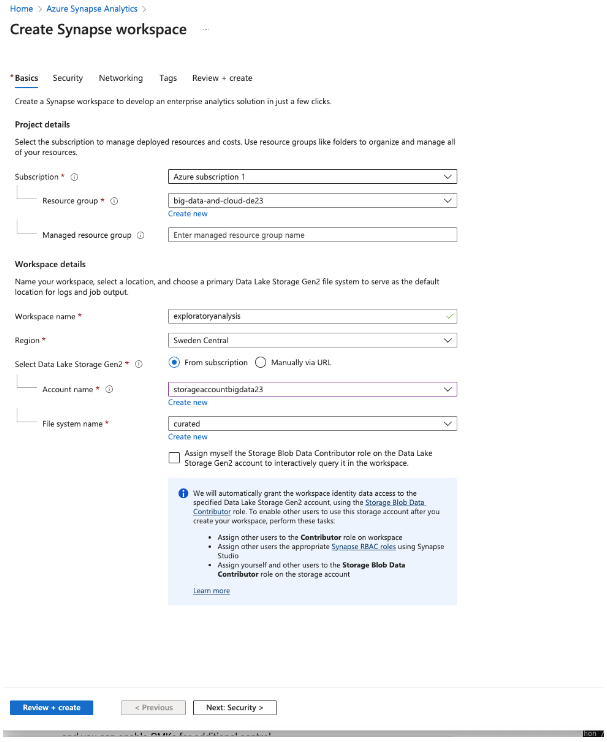
Region * : Sweden Central

Version * : V2

5. **Azure Synapse Analytics**: Synapse is used for data exploration and querying. It allows running Spark SQL to create tables on top of the Parquet files and analyze the transformed data.

   Creating Synapse:

   - The screenshot illustrates the setup of Synapse, including the selection of performance tiers and configurations for serverless resources.



Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all of your resources.

Subscription * : Azure subscription 1

Resource group * : big-data-and-cloud-de23

Create new

Managed resource group : Enter managed resource group name

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * : exploratoryanalysis

Region * : Sweden Central

Select Data Lake Storage Gen2 * : ● From subscription ○ Manually via URL

Account name * : storageaccountbigdata23

Create new

File system name * : curated

Create new

□ Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.

We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the Storage Blob Data Contributor role. To enable other users to use this storage account after you create your workspace, perform these tasks:

- Assign other users to the **Contributor** role on workspace
- Assign other users the appropriate Synapse RBAC roles using Synapse Studio
- Assign yourself and other users to the **Storage Blob Data Contributor** role on the storage account

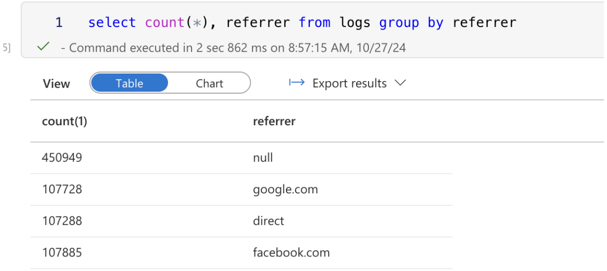
Learn more

   Data Exploration:

   I used this sql query to create a table on top of the data

```
%%sql
CREATE TABLE logs USING PARQUET
LOCATION
'abfss://curated@storageaccountbigdata23.dfs.core.windows.net/logs'
```

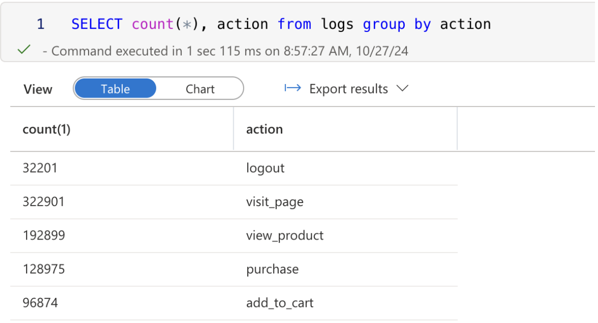
   Here are some example results from the log data:



```
select count(*), referrer from logs group by referrer
```

✓ - Command executed in 2 sec 862 ms on 8:57:15 AM, 10/27/24

count(1)	referrer
450949	null
107728	google.com
107288	direct
107885	facebook.com

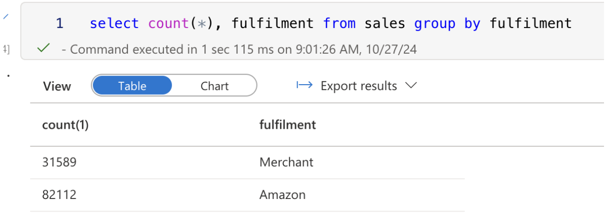


```
SELECT count(*), action from logs group by action
```

✓ - Command executed in 1 sec 115 ms on 8:57:27 AM, 10/27/24

count(1)	action
32201	logout
322901	visit_page
192899	view_product
128975	purchase
96874	add_to_cart

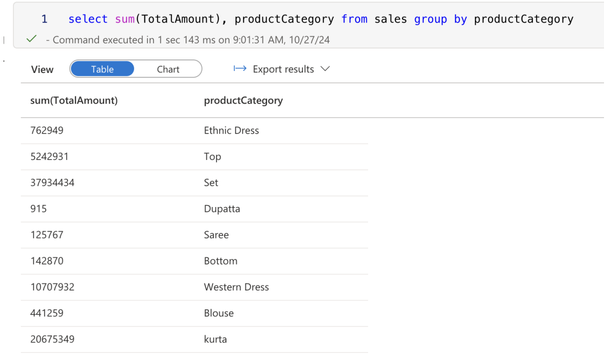
And some example results from the sales data:



```
select count(*), fulfilment from sales group by fulfilment
```

✓ - Command executed in 1 sec 115 ms on 9:01:26 AM, 10/27/24

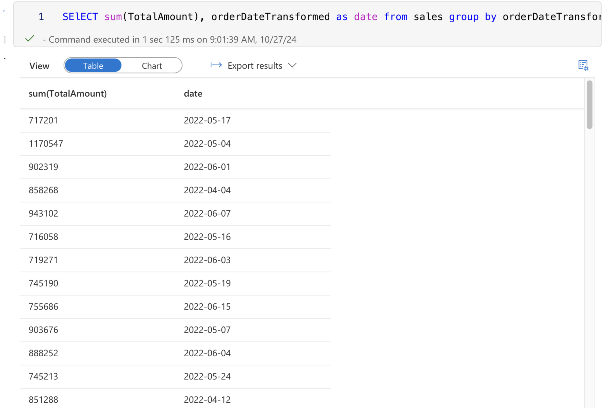
count(1)	fulfilment
31589	Merchant
82112	Amazon



```
select sum(TotalAmount), productCategory from sales group by productCategory
```

✓ - Command executed in 1 sec 143 ms on 9:01:31 AM, 10/27/24

sum(TotalAmount)	productCategory
762949	Ethnic Dress
5242931	Top
37934434	Set
915	Dupatta
125767	Saree
142870	Bottom
10707932	Western Dress
441259	Blouse
20675349	kurta



```
SElECT sum(TotalAmount), orderDateTransformed as date from sales group by orderDateTransfor
```

✓ - Command executed in 1 sec 125 ms on 9:01:39 AM, 10/27/24

sum(TotalAmount)	date
717201	2022-05-17
1170547	2022-05-04
902319	2022-06-01
858268	2022-04-04
943102	2022-06-07
716058	2022-05-16
719271	2022-06-03
745190	2022-05-19
755686	2022-06-15
903676	2022-05-07
888252	2022-06-04
745213	2022-05-24
851288	2022-04-12

6. **Power BI**: Power BI is employed to create dashboards and visualizations based on the transformed data stored in Azure Synapse.

Data Management

Data management for this project was achieved through a comprehensive ETL (Extract, Transform, Load) pipeline built in Azure Data Factory (ADF). This pipeline was designed to ensure data quality, facilitate efficient transformations, and prepare datasets for advanced analysis. Below are the detailed steps undertaken:

1. Extraction:

    - The sales data, sourced from a structured CSV file, was ingested into Azure Data Factory from the StructuredData container in the Azure Storage Account.

    - The XML log data, from the UnstructuredData container, was also ingested for further processing.

2. Data Transformation

    Sales data:

    - Type Casting: The quantity column was cast to integer and the amount column to double to standardize the data types for accurate calculations.

    - Data Filtering: A set of filters was applied to ensure data consistency and reliability: Only rows where Amount > 0 and Qty > 0 were retained.

    - Null or empty values in the currency field were excluded (!isNull(currency)).

    - Column Selection and Renaming: Only necessary columns were selected and renamed to align with naming conventions. This step reduced redundancy and improved clarity for downstream processes.

    - Derived Columns: Additional calculated fields were introduced: totalAmount was computed as the product of price and quantity.

    - Date Transformation: The date column was reformatted from dd-mm-yy to the standardized yyyy-mm-dd format to ensure compatibility with analysis tools.

    log data:

    - Data Flattening: The nested structure of the XML log data was flattened to extract relevant details into a tabular format.

    - Column Selection: A subset of critical columns was chosen for further analysis to minimize processing overhead.

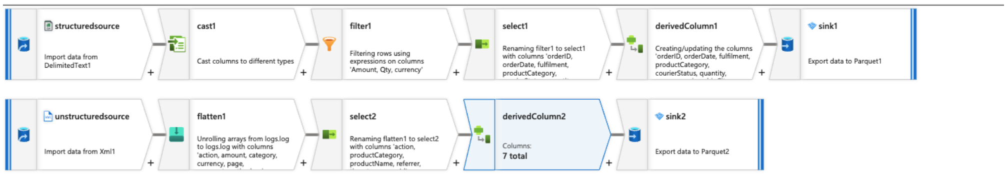
    - Derived Field: A unique identifier, logId, was created by concatenating userId and timestamp. This step ensured a unique key for tracking individual log events across the dataset.

3. Loading and Storage

    The transformed datasets were stored in the Curated container of the Azure Storage Account in Parquet format. This format was chosen for its efficient storage and querying capabilities.

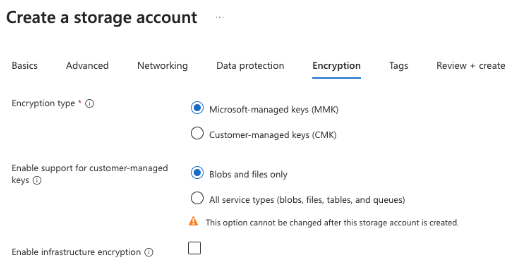
ETL Process Visual Representation

To illustrate the data management workflow, the following diagram summarizes the key steps involved in the ETL pipeline:


structuredsource
Import data from DelimitedText1
cast1
Cast columns to different types
filter1
Filtering rows using expressions on columns 'Amount, Qty, currency'
select1
Renaming filter1 to select1 with columns 'orderID, orderDate, fulfilment, productCategory,
derivedColumn1
Creating/updating the columns 'orderID, orderDate, fulfilment, productCategory, courierStatus, quantity,
sink1
Export data to Parquet1
unstructuredsource
Import data from Xml1
flatten1
Unrolling arrays from logs.log to logs.log with columns 'action, amount, category, currency, page,
select2
Renaming flatten1 to select2 with columns 'action, productCategory, productName, referrer,
derivedColumn2
Columns:
7 total
sink2
Export data to Parquet2

Security

- Data in the storage account is encrypted at rest.



Create a storage account

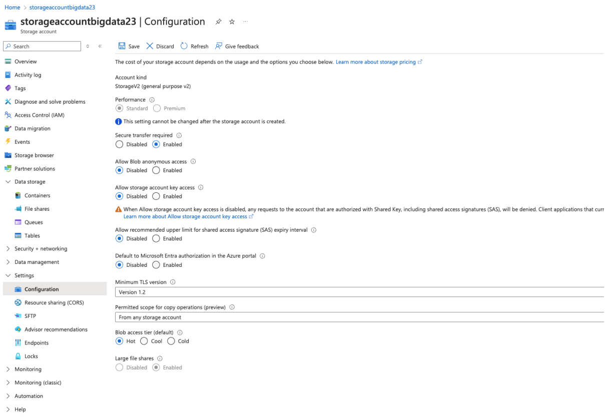
Encryption type *
● Microsoft-managed keys (MMK)
○ Customer-managed keys (CMK)

Enable support for customer-managed keys
● Blobs and files only
○ All service types (blobs, files, tables, and queues)
This option cannot be changed after this storage account is created.

Enable infrastructure encryption □

Morevoer, Azure uses Secure Socket Layer (SSL)/Transport Layer Security (TLS) protocols to encrypt data as it travels between applications and Azure services. This safeguards data from interception during transmission.

- To enhance the security posture of the data architecture, key access to the storage account was disabled. The screenshot highlights this configuration, ensuring that access is managed solely through role-based permissions.



storageaccountbigdata23 | Configuration
Storage account

The cost of your storage account depends on the usage and the options you choose below. Learn more about storage pricing

Account kind
StorageV2 (general purpose v2)

Performance
● Standard ○ Premium
This setting cannot be changed after the storage account is created.

Secure transfer required
○ Disabled ● Enabled

Allow Blob anonymous access
● Disabled ○ Enabled

Allow storage account key access
● Disabled ○ Enabled
When Allow storage account key access is disabled, any requests to the account that are authorized with Shared Key, including shared access signatures (SAS), will be denied. Client applications that curr
Learn more about Allow storage account key access

Allow recommended upper limit for shared access signature (SAS) expiry interval
● Disabled ○ Enabled

Default to Microsoft Entra authorization in the Azure portal
● Disabled ○ Enabled

Minimum TLS version
Version 1.2

Permitted scope for copy operations (preview)
From any storage account

Blob access tier (default)
● Hot ○ Cool ○ Cold

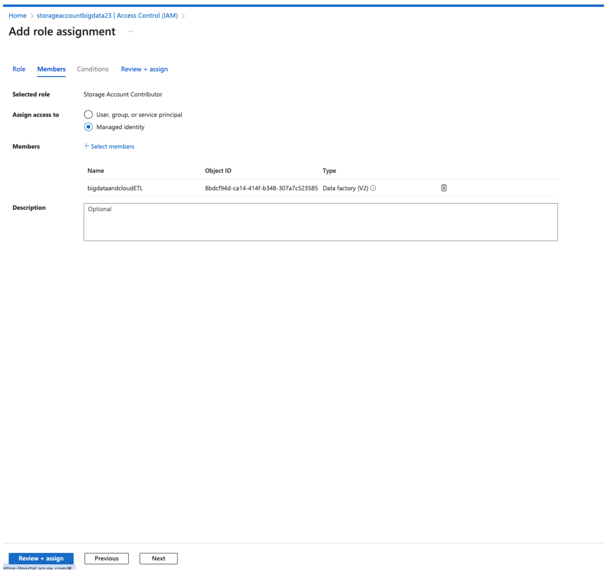
Large file shares
○ Disabled ● Enabled

- RBAC was used to access data in the blob storage:

  **Data factory's access to data:

  - To ensure proper security measures, the following settings were configured:
    - **Blob Contributor Role**: This role allows ADF to access the storage account without exposing key access. The screenshot shows the assignment of the Blob Contributor role

to the managed identity of ADF.



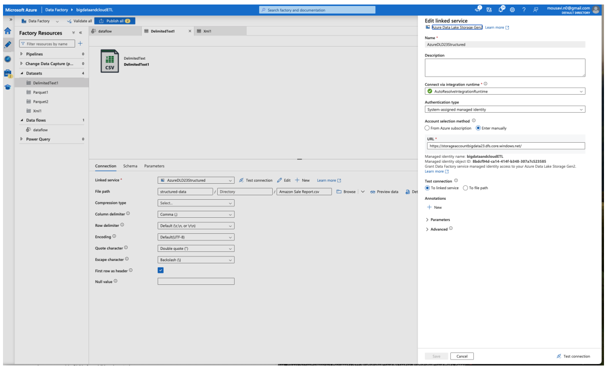
Selected role: Storage Account Contributor

Assign access to: ○ User, group, or service principal ● Managed identity

Name	Object ID	Type
bigdataandcloudETL	8bdcf94d-ca14-414f-b348-307a7c523585	Data factory (V2)

Description: Optional

- **Dataset Authentication**: The authentication type is selected to be 'system-assigned-managed-identity'.



Linked service *: AzureDLD23Structured

File path: structured-data / Directory / Amazon Sale Report.csv

Compression type: Select...

Column delimiter: Comma (,)

Row delimiter: Default (\r,\n, or \r\n)

Encoding: Default(UTF-8)

Quote character: Double quote (")

Escape character: Backslash (\)

First row as header: ☑

Null value:

Name *: AzureDLD23Structured

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: System-assigned managed identity

Account selection method: ○ From Azure subscription ● Enter manually

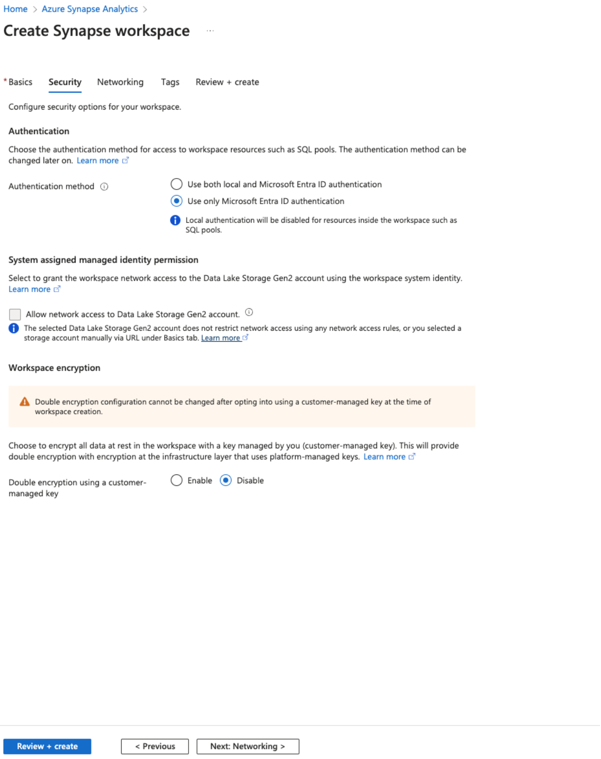
URL *: https://storageaccountbigdata23.dfs.core.windows.net/

Managed identity name: **bigdataandcloudETL**
Managed identity object ID: **8bdcf94d-ca14-414f-b348-307a7c523585**
Grant Data Factory service managed identity access to your Azure Data Lake Storage Gen2.
Learn more

Test connection: ● To linked service ○ To file path

Annotations: + New

RBAC for Synapse: The screenshot shows how access is managed, ensuring only authorized users can interact with Synapse resources.



Configure security options for your workspace.

Authentication

Choose the authentication method for access to workspace resources such as SQL pools. The authentication method can be changed later on. Learn more

Authentication method: ○ Use both local and Microsoft Entra ID authentication ● Use only Microsoft Entra ID authentication

Local authentication will be disabled for resources inside the workspace such as SQL pools.

System assigned managed identity permission

Select to grant the workspace network access to the Data Lake Storage Gen2 account using the workspace system identity. Learn more

□ Allow network access to Data Lake Storage Gen2 account.

The selected Data Lake Storage Gen2 account does not restrict network access using any network access rules, or you selected a storage account manually via URL under Basics tab. Learn more

Workspace encryption

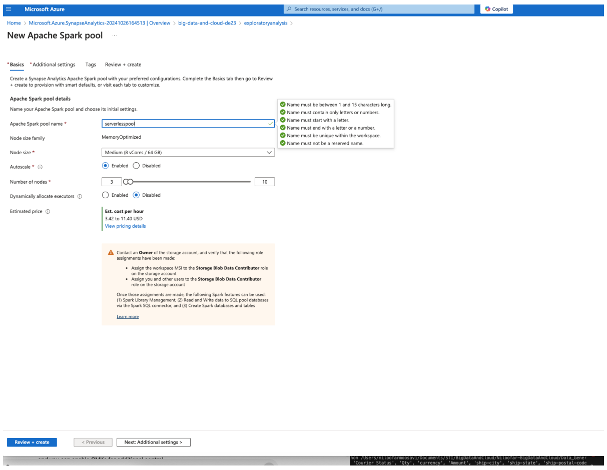
Double encryption configuration cannot be changed after opting into using a customer-managed key at the time of workspace creation.

Choose to encrypt all data at rest in the workspace with a key managed by you (customer-managed key). This will provide double encryption with encryption at the infrastructure layer that uses platform-managed keys. Learn more

Double encryption using a customer-managed key: ○ Enable ● Disable

Scalability

Serverless SQL Pool: Utilizing the serverless option in Synapse allows for on-demand querying of data stored in the Parquet files without the need for provisioning dedicated resources. The screenshot displays the setup options available for creating serverless SQL pools, emphasizing scalability, cost-effectiveness, and flexibility in querying large datasets.



Home > Microsoft.Azure.SynapseAnalytics-20241026164513 | Overview > big-data-and-cloud-de23 > exploratoryanalysis >

New Apache Spark pool ...

Create a Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + create to provision with smart defaults, or visit each tab to customize.

Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name *: serverlesspool

- ✓ Name must be between 1 and 15 characters long.
- ✓ Name must contain only letters or numbers.
- ✓ Name must start with a letter.
- ✓ Name must end with a letter or a number.
- ✓ Name must be unique within the workspace.
- ✓ Name must not be a reserved name.

Node size family: MemoryOptimized

Node size *: Medium (8 vCores / 64 GB)

Autoscale *: ● Enabled ○ Disabled

Number of nodes *: 3 – 10

Dynamically allocate executors: ○ Enabled ● Disabled

Estimated price: **Est. cost per hour** 3.42 to 11.40 USD View pricing details

Contact an **Owner** of the storage account, and verify that the following role assignments have been made:

- Assign the workspace MSI to the **Storage Blob Data Contributor** role on the storage account
- Assign you and other users to the **Storage Blob Data Contributor** role on the storage account

Once those assignments are made, the following Spark features can be used: (1) Spark Library Management, (2) Read and Write data to SQL pool databases via the Spark SQL connector, and (3) Create Spark databases and tables

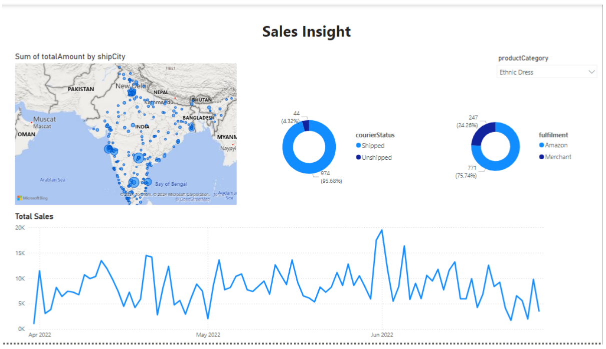
Learn more

```
hon /Users/niloofarmoosavi/Documents/STI/BigDataAndCloud/Niloofar-BigDataAndCloud/Data_Gener
 'Courier Status', 'Qty', 'currency', 'Amount', 'ship-city', 'ship-state', 'ship-postal-code
```

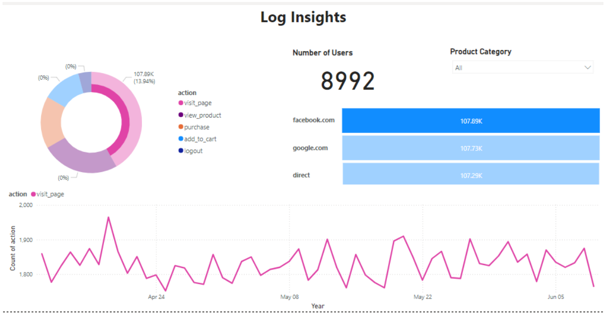
Power BI Report

Sales Insights:

- Finally, the Power BI dashboard provides visual insights derived from the sales and logs data.



Sales Insight
Sum of totalAmount by shipCity
PAKISTAN
New Delhi
NEPAL
Kathmandu
BHUTAN
BANGLADESH
INDIA
Muscat
OMAN
MYANM
Arabian Sea
Bay of Bengal
productCategory
Ethnic Dress
44 (4.32%)
974 (95.68%)
courierStatus
Shipped
Unshipped
247 (24.26%)
771 (75.74%)
fulfilment
Amazon
Merchant
Total Sales
20K
15K
10K
5K
0K
Apr 2022
May 2022
Jun 2022



Log Insights
(0%)
(0%)
107.89K (13.94%)
(0%)
action
visit_page
view_product
purchase
add_to_cart
logout
Number of Users
Product Category
All
facebook.com 107.89K
google.com 107.73K
direct 107.29K
action visit_page
Count of action
Apr 24
May 08
May 22
Jun 05
Year