

---

## Large Scale Data Processing

---

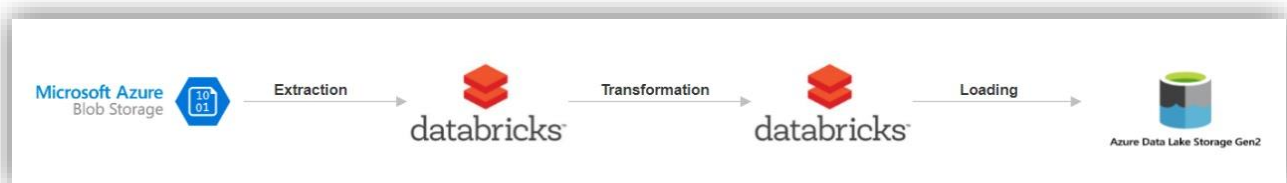
### Project Overview

In today's data-driven world, the ability to efficiently process and analyze large volumes of data is crucial for businesses to gain insights and make informed decisions. This project aims to leverage the power of Azure Databricks and PySpark to perform large-scale data processing tasks, including Extract, Transform, and Load (ETL) operations, on massive datasets. By utilizing Databricks clusters, we ensure scalability and parallel processing capabilities.

### About Project

This project leverages Azure Databricks and PySpark for large-scale data processing on a sample CSV file containing employee data. The dataset consists of approximately 100,000 records with 11 fields including employee ID, full name, job title, department, gender, ethnicity, age, hire date, annual salary, bonus percentage, country, and exit date.

### Architectural Diagram



## Key-Components/Requirements of the projects

### 1. Azure Databricks:

- Azure Databricks provides a cloud-based platform for big data analytics and machine learning. It offers a collaborative environment for data engineers, data scientists, and analysts to work together seamlessly.
- Databricks provides managed Spark clusters, eliminating the need for infrastructure management and allowing teams to focus on data processing tasks.

### 2. PySpark:

- PySpark is the Python API for Apache Spark, a powerful open-source framework for distributed data processing. PySpark simplifies development tasks by providing a Python interface to Spark's capabilities.
- With PySpark, developers can write concise and expressive code to perform complex data transformations, aggregations, and analytics on large datasets.

### 3. ETL Operations:

- **Extract:** Data ingestion from various sources such as databases, data lakes, streaming platforms, or external APIs.
- **Transform:** Data transformation tasks including cleansing, filtering, aggregating, joining, and enriching datasets to prepare them for analysis.
- **Load:** Storing processed data into target systems such as data warehouses, data lakes, or serving layers for downstream consumption.

#### 4. Scalability with Databricks Clusters:

- Databricks clusters dynamically allocate computational resources based on workload requirements, ensuring optimal performance and resource utilization.
- Autoscaling capabilities automatically adjust cluster size to accommodate changes in workload demand, allowing for seamless scalability without manual intervention.
- Databricks Runtime optimizes performance with built-in optimizations, caching, and tuning for various workloads, resulting in faster processing times.

## Azure Resources Used for this Project

### 1. Azure Data Lake Storage Gen2:

- This is where the Transformed data is Loaded. Azure Data Lake Storage Gen2 provides a scalable and secure platform for storing large volumes of data. It enables us to manage, access, and analyse data effectively

### 2. Azure Blob Storage:

- This is where the raw data is stored. Azure Blob Storage integral to Microsoft Azure's storage service, is a cloud-based solution tailored for managing vast amounts of unstructured data, encompassing both text and binary data. Termed "Blob" for "Binary Large Object," it signifies a compilation of binary data treated as a singular entity within a database.

### 3. Databricks Cluster

- An Azure Databricks cluster process the data depending on the user instructions in the Azure Notebook. It serves as a computational resource facilitating the processing of extensive data and execution of analytics workloads through the Apache Spark platform within the Microsoft Azure cloud.

# How It works

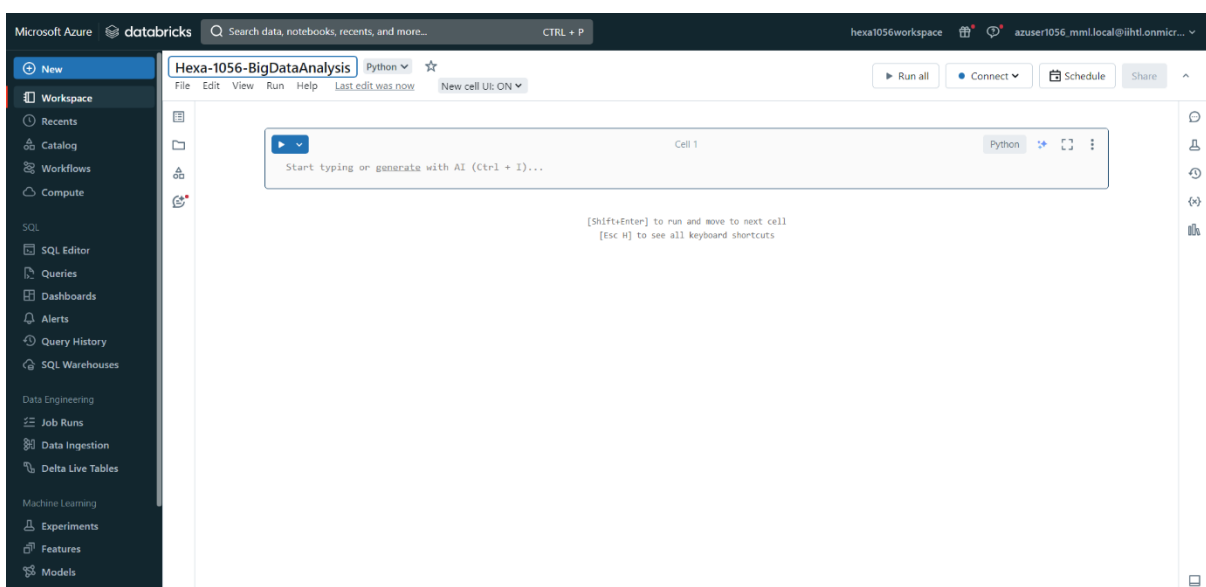
## 1. Setting Up Azure Databricks Environment:

- Sign in to the Azure portal and create an Azure Databricks workspace. Configure workspace settings, including pricing tier, region, and workspace name

The screenshot shows the 'Create an Azure Databricks workspace' page in the Microsoft Azure portal. The page has a blue header with the Microsoft Azure logo and a search bar. Below the header, there's a breadcrumb trail: 'Home > Azure Databricks >'. The main title is 'Create an Azure Databricks workspace'. There are tabs for 'Basics', 'Networking', 'Encryption', 'Tags', and 'Review + create'. The 'Basics' tab is selected. Under 'Project Details', there's a section for 'Subscription' and 'Resource group'. The 'Subscription' dropdown is set to 'Azure subscription 1' and the 'Resource group' dropdown is set to 'rg-azuser1056\_mml.local-SHf8B'. Below this, there's a 'Create new' link. Under 'Instance Details', there's a section for 'Workspace name', 'Region', 'Pricing Tier', and 'Managed Resource Group name'. The 'Workspace name' dropdown is set to 'hexa1056workspace'. The 'Region' dropdown is set to 'Central India'. The 'Pricing Tier' dropdown is set to 'Premium (+ Role-based access controls)'. There's a blue information box below the pricing tier dropdown that says: 'We selected the recommended pricing tier for your workspace. You can change the tier based on your needs.' The 'Managed Resource Group name' field is empty. At the bottom, there are three buttons: 'Review + create', '< Previous', and 'Next: Networking >'.

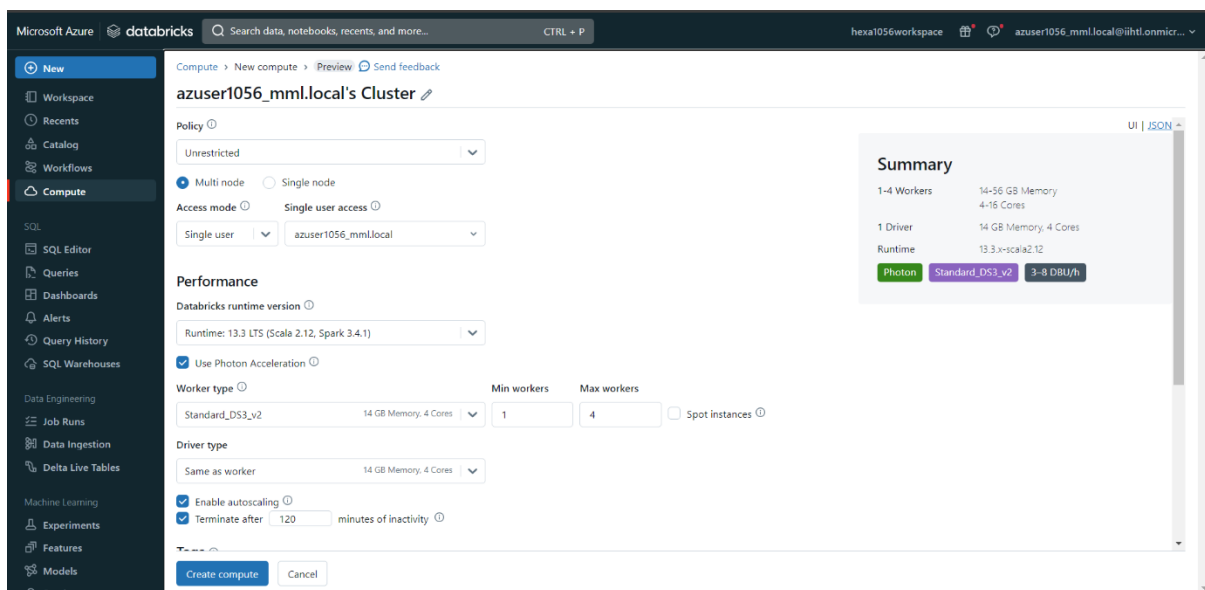
## 2. Developing PySpark Notebooks:

- Create a new PySpark notebook within the Databricks workspace. Begin writing PySpark code to perform ETL operations, data transformations, and other data processing tasks.



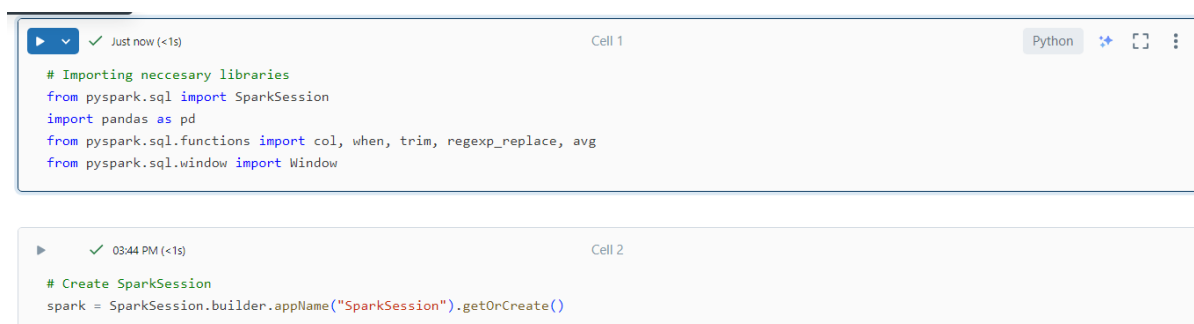
### 3. Create Cluster and Connecting to notebook

- The cluster is created with 4 working nodes and autoscaling is enabled which automatically adjust cluster size to accommodate changes in workload demand, allowing for seamless scalability without manual intervention.



### 4. Importing Necessary libraries and Creating Spark Session:

- Use `SparkSession.builder` to configure and create a `SparkSession`. specify the application name using `.appName()` and configure any additional Spark options using `.config()`. Finally, call `.getOrCreate()` to either create a new `SparkSession`



## 5. Extracting Data from Source storage

- Connecting data source (Azure Blob Storage) by mounting it to the Databricks File System (DBFS) to simplify data access
- It helps to retrieve raw data for processing and analysis within the PySpark environment.

```
03:47 PM (13s) Cell 3

# 1) Extracting the data from blob storage
# Mounting the blob storage with Azure databricks

dbutils.fs.mount(
    source = "wasbs://hexa1056sourcecontainer@hexa1056sourcestorage.blob.core.windows.net",
    mount_point="/mnt/blobStorage",
    extra_configs={"fs.azure.account.key.hexa1056sourcestorage.blob.core.windows.net":
        "ZfVh3TIuEgvwdltdhS1/3N1JVDwbu9+jWwn746GixRU1Tghc2ru1Tk7wlrj/a3BN+gCiytmo9hQh+AStTnd5Ag=="})

True
```

```
03:48 PM (<1s) Cell 4

# Listing the File information to get file path

dbutils.fs.ls('/mnt/blobStorage')

[FileInfo(path='dbfs:/mnt/blobStorage/Employee_data.csv', name='Employee_data.csv', size=10037905, modificationTime=1708941245000)]
```

```
3 minutes ago (10s) Cell 5

# Reading the data of blob storage and converting it to spark RDD

path = "dbfs:/mnt/blobStorage/Employee_data.csv"
RDD = spark.read.csv(path, header=True,inferSchema=True)

(2) Spark Jobs

RDD: pyspark.sql.dataframe.DataFrame = [EEID: integer, Full Name: string ... 10 more fields]
```

3 minutes ago (1s) Cell 6 Python

```
RDD.display()
```

(1) Spark Jobs

	EEID	Full Name	Job Title	Department	Gender	Ethnicity	Age
1	98278	Samaira Raj	Human Resources Manager	Human Resources	Male	White	46
2	19840	Lavanya Hayer	Data Analyst	Research and Development	Female	null	55
3	22487	Shayak Raval	Financial Analyst	Finance	Female	White	29
4	17160	Inaaya Bala	Software Engineer	Research and Development	Male	null	65
5	10385	Aarav Garde	Data Analyst	Research and Development	Male	White	56
6	28297	Romil Keer	Customer Service Representative	Customer Service	null	null	19
7	21123	Pranav Chadha	Human Resources Manager	Human Resources	Male	Hispanic	46

10,000 rows | Truncated data | 1.07 seconds runtime Refreshed 3 minutes ago

## 6. Transforming the raw data

- Utilize PySpark DataFrame transformations and functions to cleanse, transform, and prepare the data for analysis.
- Implement business logic and data processing steps to transform raw dataset up to mark for data analysis purpose.

### Transformations done:-

#### ▪ Removing the Duplicate records

```
▶ Just now (4s) Cell 7

# Removing the Duplicate data

print(RDD.count())
RDD=RDD.distinct()
print(RDD.count())

▶ (5) Spark Jobs

▶ RDD: pyspark.sql.dataframe.DataFrame = [EEID: integer, Full Name: string ... 10 more fields]
100000
99996
```

#### ▪ Handling anonymous data

```
▶ Just now (3s) Cell 8

# Removing the anonymous data

print(RDD.count())
RDD = RDD.na.drop(["any", subset=["EEID"]])
print(RDD.count())

▶ (6) Spark Jobs

▶ RDD: pyspark.sql.dataframe.DataFrame = [EEID: integer, Full Name: string ... 10 more fields]
99996
99984
```



- Removing Extra spaces form the data

▶ Just now (1s)

Cell 10

Python

+

⌵

⋮

# Removing Leading and Trailing spaces from the data

RDD = RDD.withColumn("Full Name", trim("Full Name"))  
RDD = RDD.withColumn("Job Title", trim("Job Title"))  
RDD = RDD.withColumn("Department", trim("Department"))  
RDD = RDD.withColumn("Gender", trim("Gender"))  
RDD = RDD.withColumn("Ethnicity", trim("Ethnicity"))  
RDD = RDD.withColumn("Country", trim("Country"))  
  
RDD.display()

▶ (2) Spark Jobs

▶ RDD: pyspark.sql.dataframe.DataFrame = [EEID: integer, Full Name: string ... 10 more fields]

Table ⌵ +

New result table: OFF ⌵

	EEID	Full Name	Job Title	Department	Gender	Ethnicity	Age
1	46400	Riya Grover	Data Analyst	Marketing	Female	Asian	25
2	61085	Ira Wable	Project Manager	Engineering	Female	Asian	18
3	92676	Parinaaz Karpe	Software Engineer	IT	Female	Asian	44
4	38396	Taran Butala	Software Engineer	Research and Development	Female	White	68
5	45876	Jayant Devan	Sales Representative	Sales	Female	Black	24
6	69030	Abram Mani	Software Engineer	IT	Male	Hispanic	44
7	22677	Samiha Vasa	Sales Reoresentative	Sales	Female	null	70

⬇ ⌵ 10,000 rows | Truncated data | 1.46 seconds runtime

Refreshed now

- Filling null values with proper messages and data

▶ Just now (<1s)

Cell 11

Python

+

⌵

⋮

# Filling the null values with proper message

RDD = RDD.na.fill(value="Not Known",subset=["Full Name"])  
RDD = RDD.na.fill(value="Not Known",subset=["Job Title"])  
RDD = RDD.na.fill(value="Not Known",subset=["Department"])  
RDD = RDD.na.fill(value="Prefer Not to say",subset=["Gender"])  
RDD = RDD.na.fill(value="Not Known",subset=["Ethnicity"])  
RDD = RDD.na.fill(value="Not Known",subset=["Country"])  
RDD = RDD.na.fill(value=0,subset=["Bonus %"])  
RDD = RDD.withColumn('Hire Date',when(col('Hire Date').isNull(),('No data provided')).otherwise(col('Hire Date')))  
RDD = RDD.withColumn('Exit Date',when(col('Exit Date').isNull(),('Currently Working')).otherwise(col('Exit Date')))

▶ RDD: pyspark.sql.dataframe.DataFrame = [EEID: integer, Full Name: string ... 10 more fields]

▶

✓ Just now (1s)

Cell 12

RDD.display()

▶ (2) Spark Jobs

Table ▾ +

New result table: OFF ▾

	EEID	Full Name	Job Title	Department	Gender	Ethnicity	Age
1	46400	Riya Grover	Data Analyst	Marketing	Female	Asian	25
2	61085	Ira Wable	Project Manager	Engineering	Female	Asian	18
3	92676	Parinaaz Karpe	Software Engineer	IT	Female	Asian	44
4	38396	Taran Butala	Software Engineer	Research and Development	Female	White	68
5	45876	Jayant Devan	Sales Representative	Sales	Female	Black	24
6	69030	Abram Mani	Software Engineer	IT	Male	Hispanic	44
7	22677	Samiha Vasa	Sales Representative	Sales	Female	Not Known	70

⬇

▾

10,000 rows | Truncated data | 1.15 seconds runtime

Refreshed now

▶ ▾

✓ Just now (6s)

Cell 13

Python

# Filling the numerical column's null values with proper average values

window\_spec = Window.partitionBy()

RDD = RDD.withColumn('Age', when(col('Age').isNull(), avg(col('Age')).over(window\_spec)).otherwise(col('Age')))

average\_salaries = RDD.groupBy("Country", "Department").avg("Annual Salary")

RDD = RDD.join(average\_salaries, ["Country", "Department"], "left").withColumnRenamed("avg(Annual Salary)", "Average Salary")

RDD = RDD.withColumn('Annual Salary', when(col('Annual Salary').isNull(), col('Average Salary')).otherwise(col('Annual Salary'))).drop("Average Salary")

RDD.display()

▶ (8) Spark Jobs

average\_salaries: pyspark.sql.dataframe.DataFrame = [Country: string, Department: string ... 1 more field]

RDD: pyspark.sql.dataframe.DataFrame = [Country: string, Department: string ... 10 more fields]

Table ▾ +

New result table: ON ▾

Search

🔍

📄

	Country	Department	EEID	Full Name	Job Title	Gender	Ethnicity
1	Canada	Marketing	46400	Riya Grover	Data Analyst	Female	Asian
2	UK	Engineering	61085	Ira Wable	Project Manager	Female	Asian
3	USA	IT	92676	Parinaaz Karpe	Software Engineer	Female	Asian
4	Canada	Research and Development	38396	Taran Butala	Software Engineer	Female	White
5	Australia	Sales	45876	Jayant Devan	Sales Representative	Female	Black
6	Canada	IT	69030	Abram Mani	Software Engineer	Male	Hispanic
7	Canada	Sales	22677	Samiha Vasa	Sales Representative	Female	Not Known
8	USA	Customer Service	94951	Kabir Dey	Customer Service Representative	Prefer Not to say	Not Known

- Renaming USA to US to make dataset consistent

```
Cell 14
Just now (2s)

# Renaming USA as US

RDD=RDD.withColumn('Country',when(RDD.Country=='USA',regexp_replace(RDD.Country,'USA','US')).otherwise(RDD.Country))
print(RDD.select("Country").distinct().collect())
```

(3) Spark Jobs

RDD: pyspark.sql.dataframe.DataFrame = [Country: string, Department: string ... 10 more fields]

[Row(Country='US'), Row(Country='UK'), Row(Country='Canada'), Row(Country='Australia')]

- Statistical data

Cell 15  
Python  
Just now (4s)

```
# Statistical Data for analysis

RDD.describe("Age", "Annual Salary", "Bonus %").display()
```

(8) Spark Jobs

	summary	Age	Annual Salary	Bonus %
1	count	99984	99984	99984
2	mean	43.937610017602815	79966.66046657386	5.0034777564409945
3	stddev	15.330310539728506	23091.820104429615	2.8844491871334133
4	min	18.0	40000.48	0.0
5	max	70.0	119999.56	10.0

5 rows | 4.24 seconds runtime

Refreshed now

## 7. Loading Data into Sink Storage

- To store the transformed data we need to create sink storage (Azure Data Lake) and Container
- Connecting data source (ADLS) by mounting it to the Databricks notebook to load the data

[Home](#) > [Storage accounts](#) >

Create a storage account

Basics | Advanced | Networking | Data protection | Encryption | Tags | Review

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 \*

rg-azuser1056\_mmllocal-SHFs8

Create new

**Instance details**

Storage account name ⓘ \*

hexa1056sinkstorage

Review

< Previous

Next : Advanced >

[Give feedback](#)

[https://portal.azure.com/#k](#)

[Home](#) > [Storage accounts](#) >

Create a storage account

Basics | Advanced | Networking | Data protection | Encryption | Tags | Review

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 \*

rg-azuser1056\_mmllocal-SHFs8

Create new

**Instance details**

Storage account name ⓘ \*

hexa1056sinkstorage

Review

< Previous

Next : Advanced >

[Give feedback](#)

[https://portal.azure.com/#k](#)

Microsoft Azure Search resources, services, and docs (G+)

Home > hexa1056sinkstorage\_1708941446843 | Overview > hexa1056sinkstorage | Containers >

### hexa1056sinkcontainer

Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

**Overview**

Diagnose and solve problems

Access Control (IAM)

**Settings**

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

**Authentication method:** Access key (Switch to Microsoft Entra user account)

**Location:** hexa1056sinkcontainer

Search blobs by prefix (case-sensitive)  ☐ Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> Transformed_data						...

Just now (11s) Cell 16

```
# 3) Loading Data in Azure Data Lake

# Mounting the sink storage(Azure Data Lake) with Azure databricks

dbutils.fs.mount(
    source = "wasbs://hexa1056sinkcontainer@hexa1056sinkstorage.blob.core.windows.net",
    mount_point="/mnt/blobStorage1",
    extra_configs={"fs.azure.account.key.hexa1056sinkstorage.blob.core.windows.net":
        "EWHBv0tr0q7pGTHFI1i0tqk9iU+JTba373jdDdBQG0ylx6Nj9wVmtoAjiHtjd5tvcUvfZpXDTm3u+AStZniWHw=="})

True
```

04:37 PM (<1s) Cell 17

```
dbutils.fs.ls("/mnt/blobStorage1")
```

[FileInfo(path='dbfs:/mnt/blobStorage1/Transformed\_data/', name='Transformed\_data/', size=0, modificationTime=0)]

04:37 PM (1s) Cell 18

```
# Converting RDD to pandas dataframe

pandas_df=RDD.toPandas()
```

(1) Spark Jobs

5 minutes ago (1s) Cell 19 Python

```
# Loading the transformed data in Azure Data Lake

pandas_df.to_csv('/dbfs/mnt/blobStorage1/Transformed_data/Transformed_Data.csv',index=False)
```

- Data Successfully Loaded

Microsoft Azure

Search resources, services, and docs (G+J)

azuser1056 mm1.local@...

BHT (BHTLONMICROSOFT.COM)

Home > hexa1056sinkstorage | Containers >

hexa1056sinkcontainer

Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: hexa1056sinkcontainer / Transformed\_data

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[.]						...
Transformed_Data.csv	26/02/2024, 16:45:15	Hot (Inferred)		Block blob	11.16 MiB	Available

## 8. Unmounting the Source and Sink storage

Just now (21s)

Cell 20

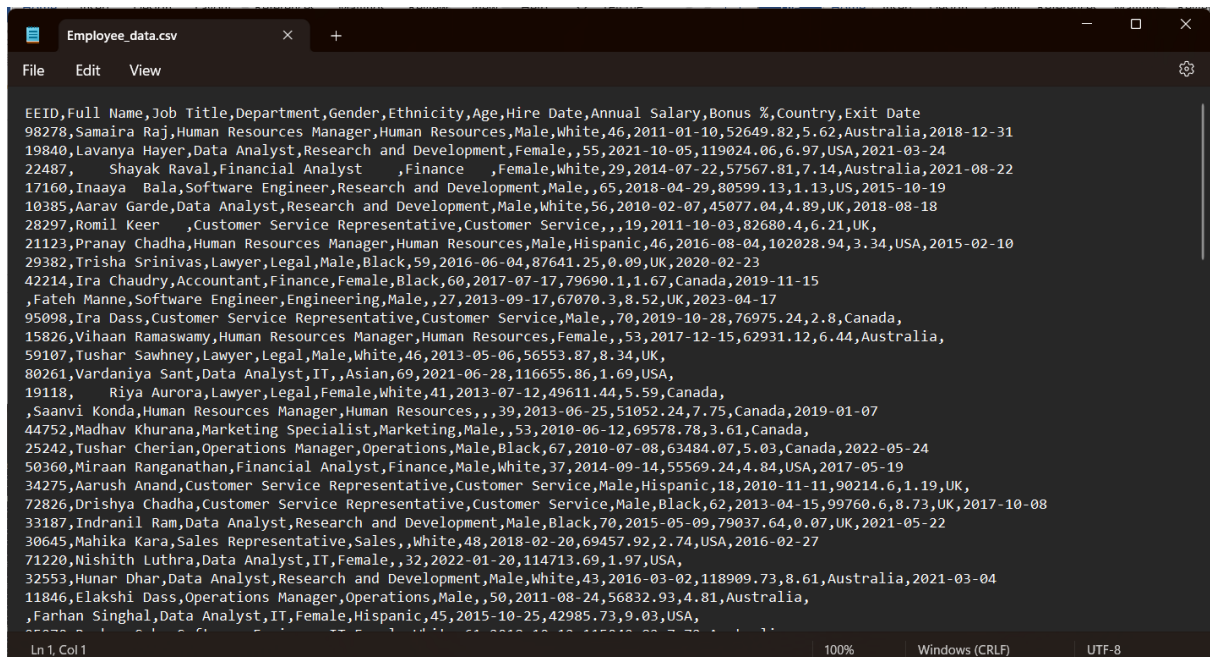
```
# Unmounting the source and sink storage

dbutils.fs.unmount("/mnt/blobStorage")
dbutils.fs.unmount("/mnt/blobStorage1")

/mnt/blobStorage has been unmounted.
/mnt/blobStorage1 has been unmounted.

True
```

# Raw Data



EEID	Full Name	Job Title	Department	Gender	Ethnicity	Age	Hire Date	Annual Salary	Bonus %	Country	Exit Date
98278	Samaira Raj	Human Resources Manager	Human Resources	Male	White	46	2011-01-10	52649.82	5.62	Australia	2018-12-31
19840	Lavanya Hayer	Data Analyst	Research and Development	Female		55	2021-10-05	119024.06	6.97	USA	2021-03-24
22487	Shayak Raval	Financial Analyst	Finance	Female	White	29	2014-07-22	57567.81	7.14	Australia	2021-08-22
17160	Inaaya Bala	Software Engineer	Research and Development	Male		65	2018-04-29	80599.13	1.13	US	2015-10-19
10385	Aarav Garde	Data Analyst	Research and Development	Male	White	56	2010-02-07	45077.04	4.89	UK	2018-08-18
28297	Romil Keer	Customer Service Representative	Customer Service			19	2011-10-03	82680.4	6.21	UK	
21123	Pranay Chadha	Human Resources Manager	Human Resources	Male	Hispanic	46	2016-08-04	102028.94	3.34	USA	2015-02-10
29382	Trisha Srinivas	Lawyer	Legal	Male	Black	59	2016-06-04	87641.25	0.09	UK	2020-02-23
42214	Ira Chaudry	Accountant	Finance	Female	Black	60	2017-07-17	79690.1	1.67	Canada	2019-11-15
	Fateh Manne	Software Engineer	Engineering	Male		27	2013-09-17	67070.3	8.52	UK	2023-04-17
95098	Ira Dass	Customer Service Representative	Customer Service	Male		70	2019-10-28	76975.24	2.8	Canada	
15826	Vihaan Ramaswamy	Human Resources Manager	Human Resources	Female		53	2017-12-15	62931.12	6.44	Australia	
59107	Tushar Sawhney	Lawyer	Legal	Male	White	46	2013-05-06	56553.87	8.34	UK	
80261	Vardaniya Sant	Data Analyst	IT		Asian	69	2021-06-28	116655.86	1.69	USA	
19118	Riya Aurora	Lawyer	Legal	Female	White	41	2013-07-12	49611.44	5.59	Canada	
	Saanvi Konda	Human Resources Manager	Human Resources			39	2013-06-25	51052.24	7.75	Canada	2019-01-07
44752	Madhav Khurana	Marketing Specialist	Marketing	Male		53	2010-06-12	69578.78	3.61	Canada	
25242	Tushar Cherian	Operations Manager	Operations	Male	Black	67	2010-07-08	63484.07	5.03	Canada	2022-05-24
50360	Miraan Ranganathan	Financial Analyst	Finance	Male	White	37	2014-09-14	55569.24	4.84	USA	2017-05-19
34275	Aarush Anand	Customer Service Representative	Customer Service	Male	Hispanic	18	2010-11-11	90214.6	1.19	UK	
72826	Drishya Chadha	Customer Service Representative	Customer Service	Male	Black	62	2013-04-15	99760.6	8.73	UK	2017-10-08
33187	Indranil Ram	Data Analyst	Research and Development	Male	Black	70	2015-05-09	79037.64	0.07	UK	2021-05-22
30645	Mahika Kara	Sales Representative	Sales		White	48	2018-02-20	69457.92	2.74	USA	2016-02-27
71220	Nishith Luthra	Data Analyst	IT	Female		32	2022-01-20	114713.69	1.97	USA	
32553	Hunar Dhar	Data Analyst	Research and Development	Male	White	43	2016-03-02	118909.73	8.61	Australia	2021-03-04
11846	Elakshi Dass	Operations Manager	Operations	Male		50	2011-08-24	56832.93	4.81	Australia	
	Farhan Singhal	Data Analyst	IT	Female	Hispanic	45	2015-10-25	42985.73	9.03	USA	

# Transformed Data

Country	Department	EEID	Full Name	Job Title	Gender	Ethnicity	Age	Hire Date	Annual Salary	Bonus %	Exit Date
Canada	Marketing	46400	Riya Grover	Data Analyst	Female	Asian	25.0	2014-09-26	75499.29	3.83	2020-08-21
UK	Engineering	61085	Ira Wable	Project Manager	Female	Asian	18.0	2017-08-29	110198.75	2.09	2017-04-02
US	IT	92676	Parinaaz Karpe	Software Engineer	Female	Asian	44.0	2016-10-31	41026.82	5.31	2015-11-24
Canada	Research and Development	38396	Taran Butala	Software Engineer	Female	White	68.0	2011-07-04	105485.69	8.69	2023-01-10
Australia	Sales	45876	Jayant Devan	Sales Representative	Female	Black	24.0	2014-05-23	57866.08	4.18	Currently Working
Canada	IT	69030	Abram Mani	Software Engineer	Male	Hispanic	44.0	2011-03-16	48409.17	8.63	Currently Working
Canada	Sales	22677	Samiha Vasa	Sales Representative	Female	Not Known	70.0	2021-11-09	104495.8	5.83	2018-11-24
US	Customer Service	94951	Kabir Dey	Customer Service Representative	Prefer Not to say	Not Known	40.0	2015-10-30	49184.54	8.37	2023-10-04
UK	Sales	30669	Tarini Brahmhatt	Sales Representative	Male	White	64.0	2021-02-18	64526.56	9.09	2022-04-18
UK	Customer Service	39341	Kismat Keer	Customer Service Representative	Female	Asian	40.0	2011-06-02	110918.42	4.94	Currently Working
UK	Sales	44291	Aradhya Chacko	Sales Representative	Prefer Not to say	Not Known	29.0	2019-12-18	71443.58	1.7	2017-07-05
Canada	Engineering	86184	Stuvan Dutt	Software Engineer	Female	Black	69.0	2015-07-14	41516.36	5.26	Currently Working
Canada	Research and Development	20783	Rasha Aurora	Data Analyst	Female	White	40.0	2014-09-26	118531.4	2.32	2018-10-23
US	Operations	78333	Ritvik Lanka	Customer Service Representative	Prefer Not to say	Not Known	43.0	2012-06-01	90268.73	8.09	Currently Working
Canada	Legal	38418	Ryan Sura	Lawyer	Prefer Not to say	Not Known	18.0	2013-05-31	46707.41	4.8	2016-12-22
US	Operations	45764	Adira Sahni	Customer Service Representative	Male	Not Known	26.0	2011-05-21	78001.94	2.36	2019-07-17
Australia	IT	86860	Navya Shere	Data Analyst	Male	Hispanic	49.0	2017-01-16	71266.08	4.21	2019-05-05
US	Legal	45864	Farhan Balay	Lawyer	Male	White	22.0	2020-05-13	116167.98	5.35	2019-10-25
UK	Finance	31359	Anvi Ahluwalia	Accountant	Prefer Not to say	Asian	25.0	2010-08-25	101658.98	9.25	2017-06-24
UK	Operations	27549	Emir Tailor	Operations Manager	Prefer Not to say	Black	64.0	2019-04-08	68657.94	5.17	2019-09-30
UK	Engineering	75249	Arnav Mand	Project Manager	Male	Not Known	67.0	2021-04-03	111875.16	8.71	Currently Working
Australia	Marketing	47025	Samar Chowdhury	Marketing Specialist	Male	Not Known	60.0	2010-08-19	78734.94	9.19	Currently Working
Australia	Sales	12513	Ranbir Kibe	Sales Representative	Prefer Not to say	Hispanic	30.0	2022-11-24	53666.9	2.53	Currently Working
Australia	Human Resources	80691	Indrans Lad	Human Resources Manager	Male	White	66.0	2012-03-17	103161.84	0.66	2023-09-28
Canada	Operations	88868	Biju Sani	Operations Manager	Male	Hispanic	70.0	2020-01-06	59208.66	7.47	2017-07-28
UK	Finance	46756	Dhruv Chakraborty	Financial Analyst	Female	Not Known	32.0	2017-12-17	82979.87	1.98	2022-11-08

Ln 8, Col 110

100%Unix (LF)UTF-8

## Conclusion

In conclusion, this project successfully demonstrated the utilization of Azure Databricks and PySpark for large-scale data processing on a sample CSV file containing employee data. Through the implementation of Extract, Transform, Load (ETL) operations, the dataset was cleaned, transformed, and prepared for further analysis or reporting. Overall, this project serves as a practical demonstration of how Azure Databricks and PySpark can be effectively utilized for large-scale data processing.