Analyzing Salary data

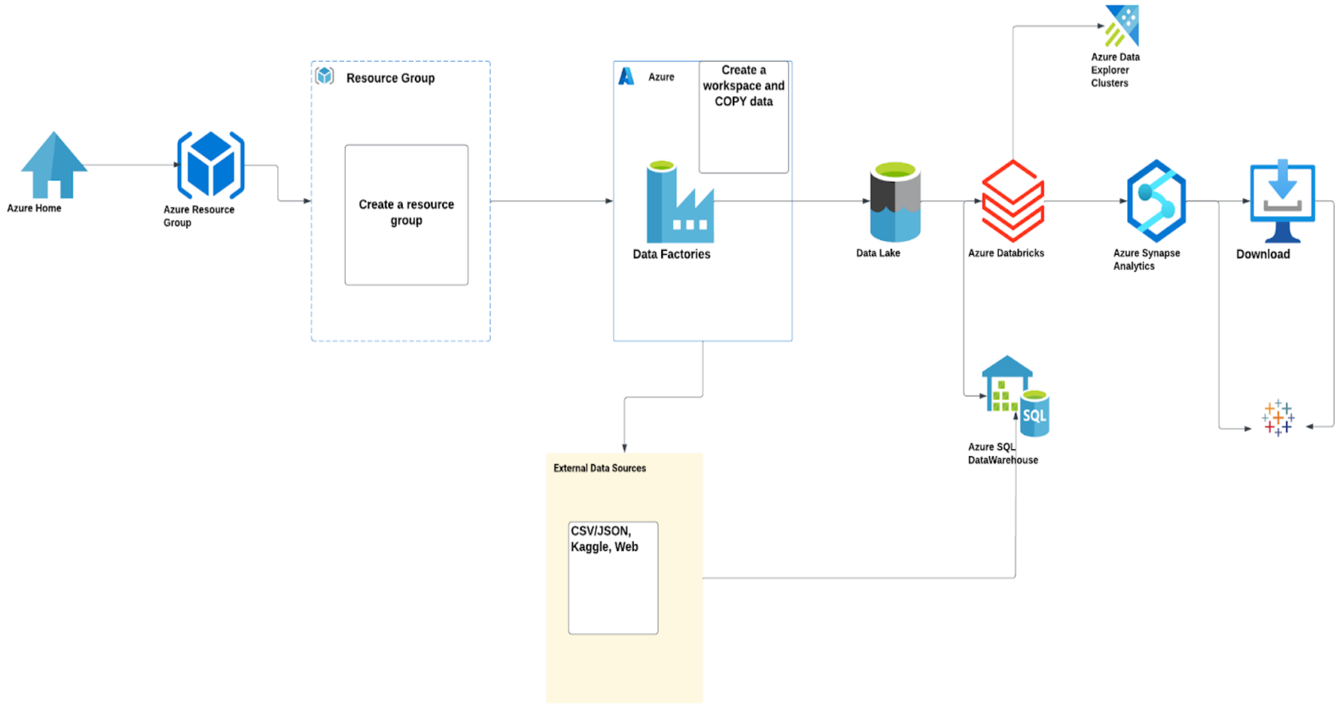
Data set: [Click to checkout Data]

About the Data: This dataset offers an extensive compilation of salary data spanning diverse industries and global regions. Gathered from reputable employment platforms and surveys, it encompasses information on job titles, compensation, industry sectors, geographical locations, and more. We can utilize this dataset for a thorough examination of trends in the job market, a comparative analysis of salaries across various professions, and to make well-informed decisions regarding career paths or hiring strategies. The dataset has been meticulously cleaned and preprocessed to facilitate analysis, and it is accessible under an open license for research and data analysis endeavors.

Tools used for Project:

- Azure Services - Data Factory, Databricks, Synapse Analytics
- Tableau
- Lucid Chart

Process Flow:



Implementing Azure Analytics with Tableau involves a series of well-defined steps for seamless integration and effective data analysis. Here is an elaboration of the key steps:

Azure Resource Group Creation:
Start by establishing an Azure Resource Group. This serves as the foundational container for organizing and managing resources related to your analytics project on the Azure platform.

Azure Data Factory Workspace Setup:
Create an Azure Data Factory workspace, which acts as the orchestration hub for managing and orchestrating data workflows.
Within this workspace, set up a container with two directories, "raw data" and "transformed data," providing structured storage for data at different stages of processing.

Data Ingestion with Azure Data Factory:
Utilize the "COPY data" feature in Azure Data Factory to connect to the raw data source, typically hosted on a Git repository.
Direct the data to the desired destination, often referred to as the "sink," and store it in a CSV format after undergoing necessary transformations.

Azure Databricks Workspace Creation:
Establish an Azure Databricks workspace, a collaborative environment for big data analytics, and create a new App registration within it.
Extract essential credentials such as Client ID, Tenant IP, and secret key during the App registration process. These credentials are vital for seamless data access between Azure Data Factory and Databricks.

Direct Data Ingestion into Azure Databricks:
As an alternative to using Azure Data Factory, the data file can be ingested directly into Azure Databricks, eliminating the need for an intermediate step. This involves reading the data into a tabular format suitable for analysis.

Analyzing Salary data

Data Exploration and Analysis in Azure Databricks:

Leverage SQL and/or Spark within the Azure Databricks workspace to access and manipulate the tabular data.

Create a new notebook in the workspace to load necessary Python libraries, inspect the data for inconsistencies, and perform statistical analysis.

Data Visualization and Analysis with Tableau and Azure Synapse Analytics:

Utilize Azure Synapse Analytics for efficient data warehousing and analysis.

Develop visualizations and dashboards using Tableau, connecting to the processed data in Azure Synapse Analytics to gain meaningful insights.

Leverage the combined capabilities of Tableau and Azure Synapse Analytics for advanced data visualization, analysis, and reporting.

Tableau Integration:

Download Tableau Desktop and the required drivers to establish connectivity with Azure Databricks.

Alternatively, download the data from Azure Databricks and use it directly integration to Tableau for creating visualizations and dashboards

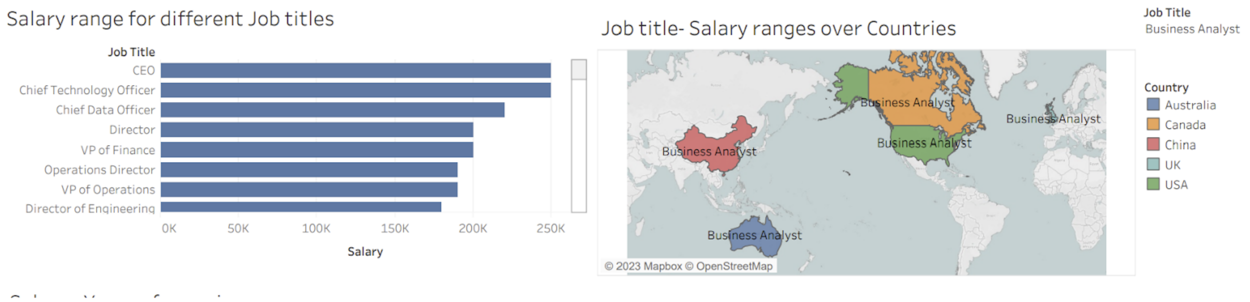
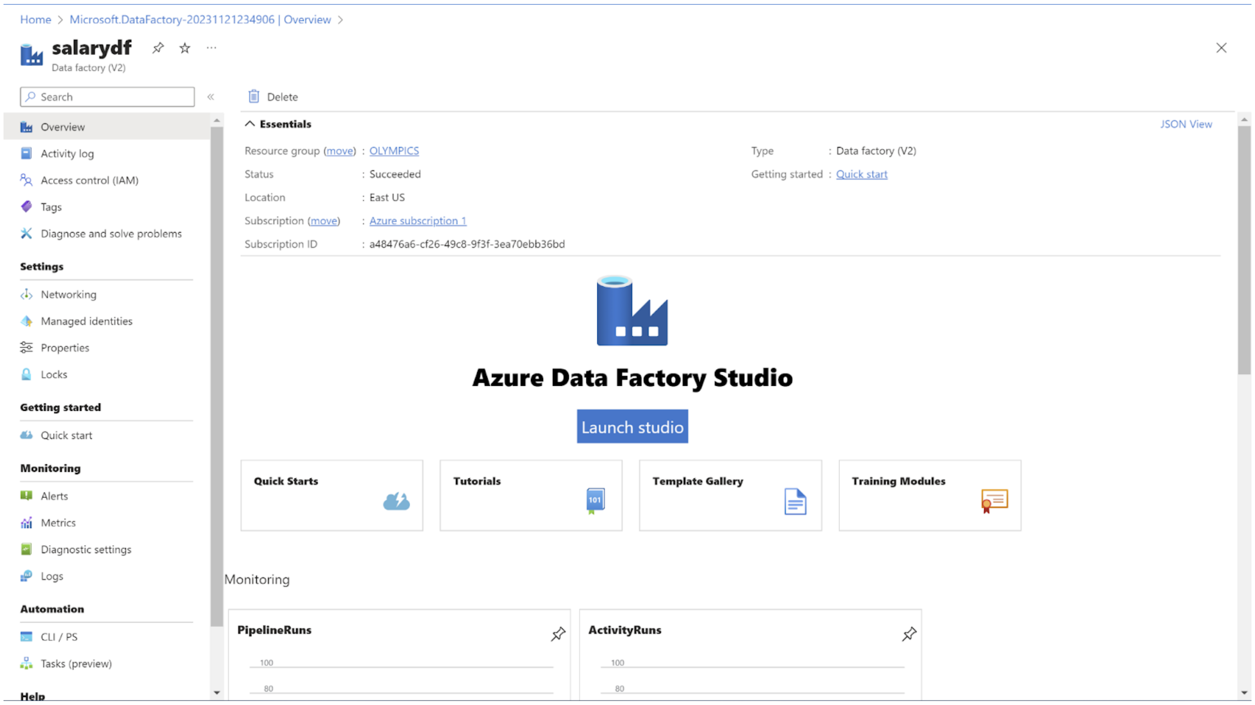


Tableau visualizations

<u>Screenshots Below show the Process followed after creating DataFactory Workspace</u>

Analyzing Salary data

- Create a workspace in Data Factory Studio(Note: Resource group should be created before this)



Home > Microsoft.DataFactory-20231121234906 | Overview >

salarydf
Data factory (V2)

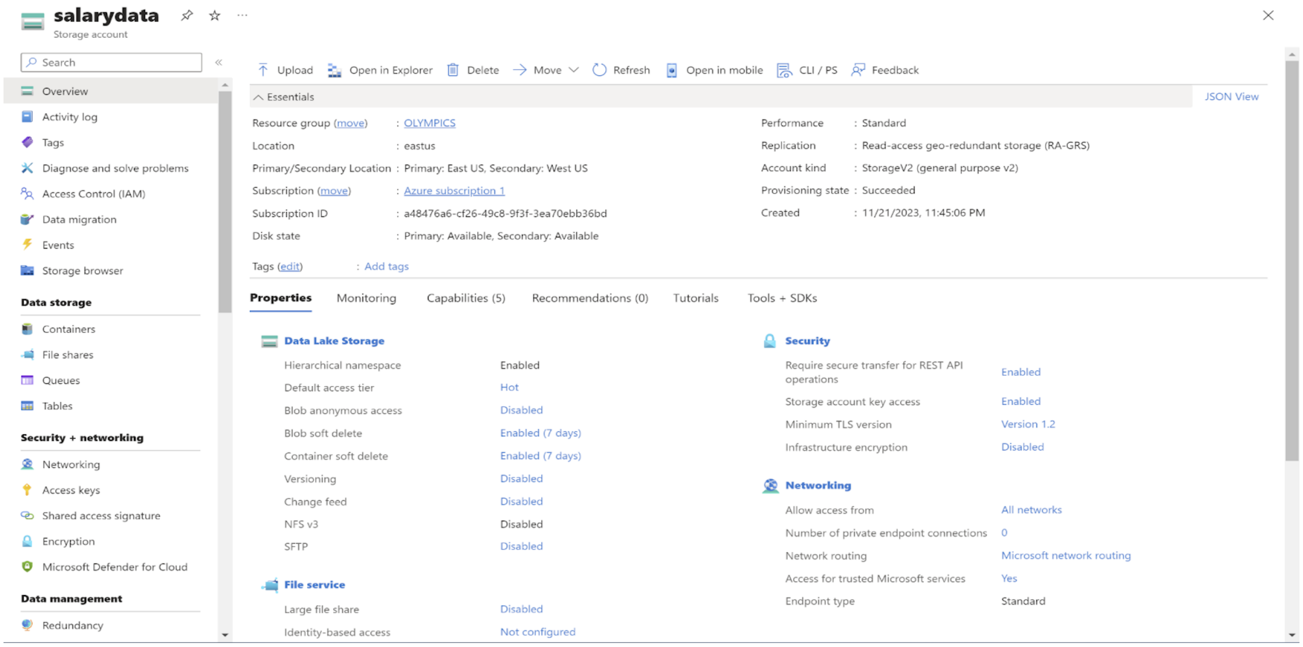
Delete

Essentials | JSON View

Resource group (move) : OLYMPICS
Status : Succeeded
Location : East US
Subscription (move) : Azure subscription 1
Subscription ID : a48476a6-cf26-49c8-9f3f-3ea70ebb36bd

Type : Data factory (V2)
Getting started : Quick start

- **We can see the details here data factory is attached to resource group and at left we can see containers -select to create**



salarydata
Storage account

Upload | Open in Explorer | Delete | Move | Refresh | Open in mobile | CLI / PS | Feedback

Essentials | JSON View

Resource group (move) : OLYMPICS
Location : eastus
Primary/Secondary Location : Primary: East US, Secondary: West US
Subscription (move) : Azure subscription 1
Subscription ID : a48476a6-cf26-49c8-9f3f-3ea70ebb36bd
Disk state : Primary: Available, Secondary: Available

Performance : Standard
Replication : Read-access geo-redundant storage (RA-GRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 11/21/2023, 11:45:06 PM

Tags (edit) : Add tags

Properties | Monitoring | Capabilities (5) | Recommendations (0) | Tutorials | Tools + SDKs

Data Lake Storage

Hierarchical namespace	Enabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Enabled (7 days)
Container soft delete	Enabled (7 days)
Versioning	Disabled
Change feed	Disabled
NFS v3	Disabled
SFTP	Disabled

File service

Large file share	Disabled
Identity-based access	Not configured

Security

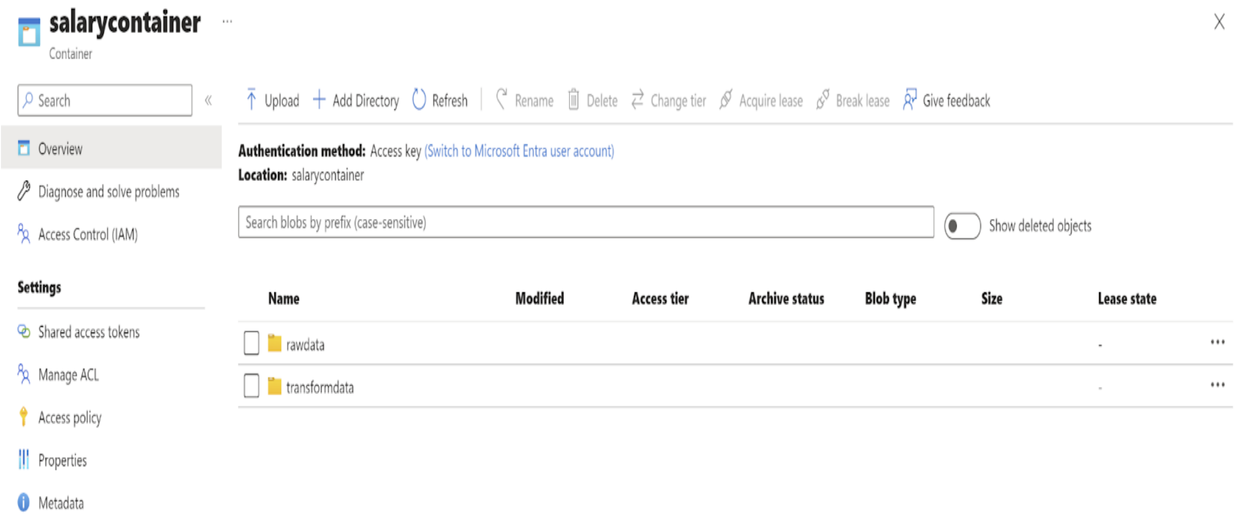
Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2
Infrastructure encryption	Disabled

Networking

Allow access from	All networks
Number of private endpoint connections	0
Network routing	Microsoft network routing
Access for trusted Microsoft services	Yes
Endpoint type	Standard

- **Create two directories in containers**

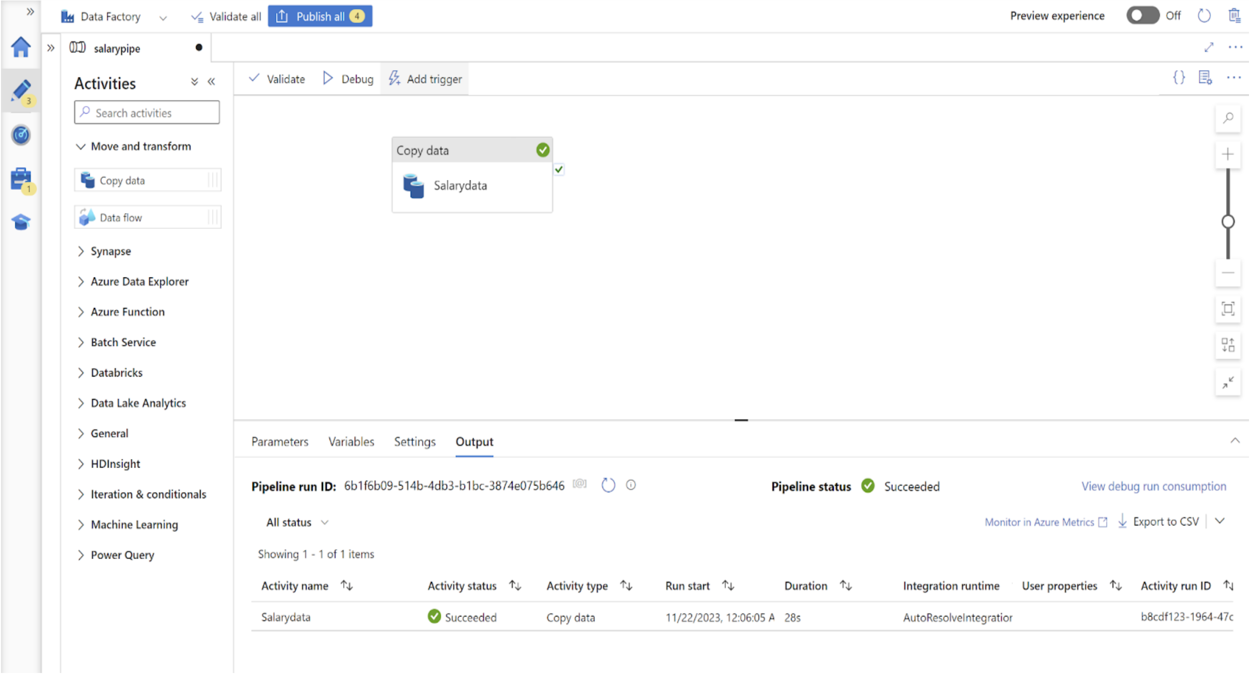
Analyzing Salary data



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

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
rawdata						-
transformdata						-

- **Create a pipeline and utilize COPY Data to load data then validate, Debug and published**



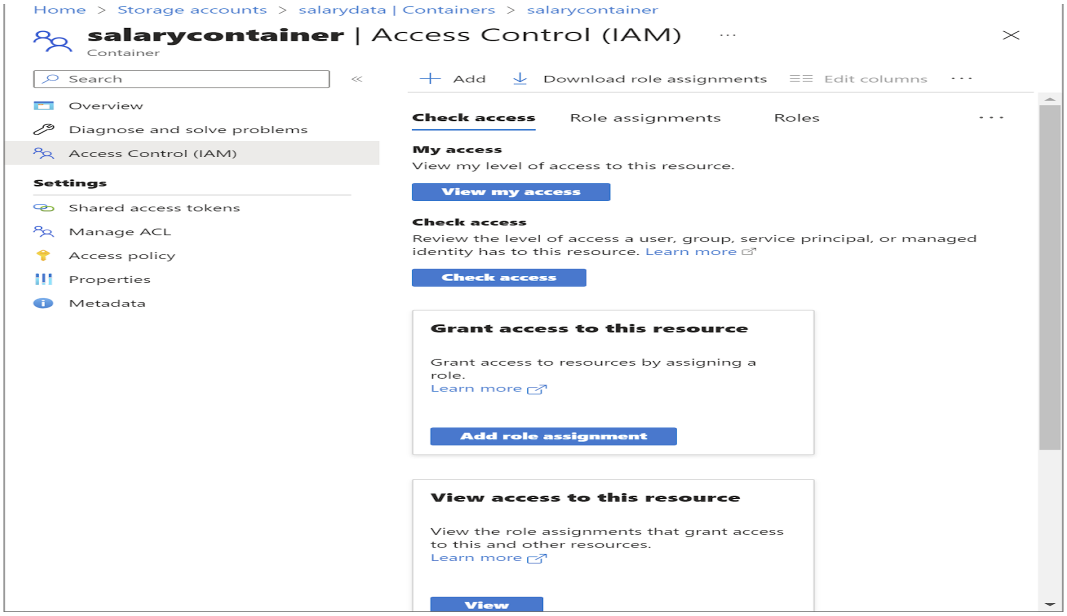
Pipeline run ID: 6b1f6b09-514b-4db3-b1bc-3874e075b646

Pipeline status ✓ Succeeded

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Salarydata	✓ Succeeded	Copy data	11/22/2023, 12:06:05 A	28s	AutoResolveIntegratior		b8cdf123-1964-47c

- **Add access to Azure Blob container**



Home > Storage accounts > salarydata | Containers > salarycontainer

salarycontainer | Access Control (IAM)

Container

My access

View my level of access to this resource.

Check access

Review the level of access a user, group, service principal, or managed identity has to this resource. Learn more

Grant access to this resource

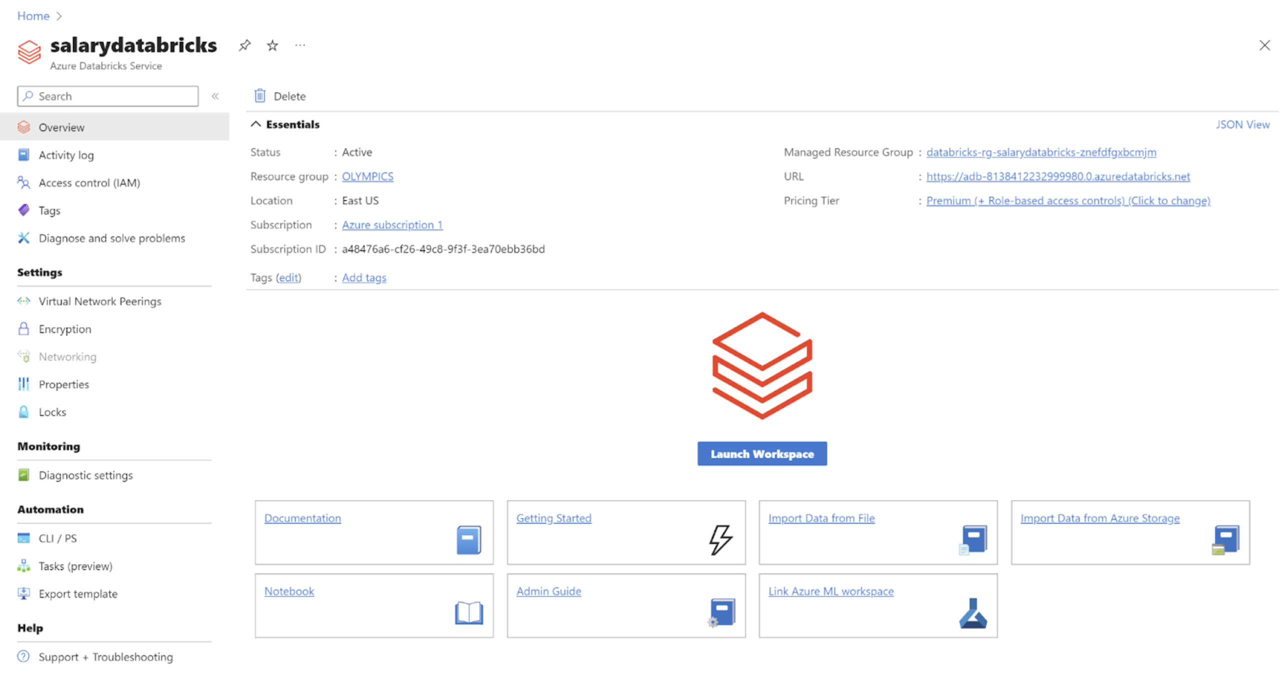
Grant access to resources by assigning a role.
Learn more

View access to this resource

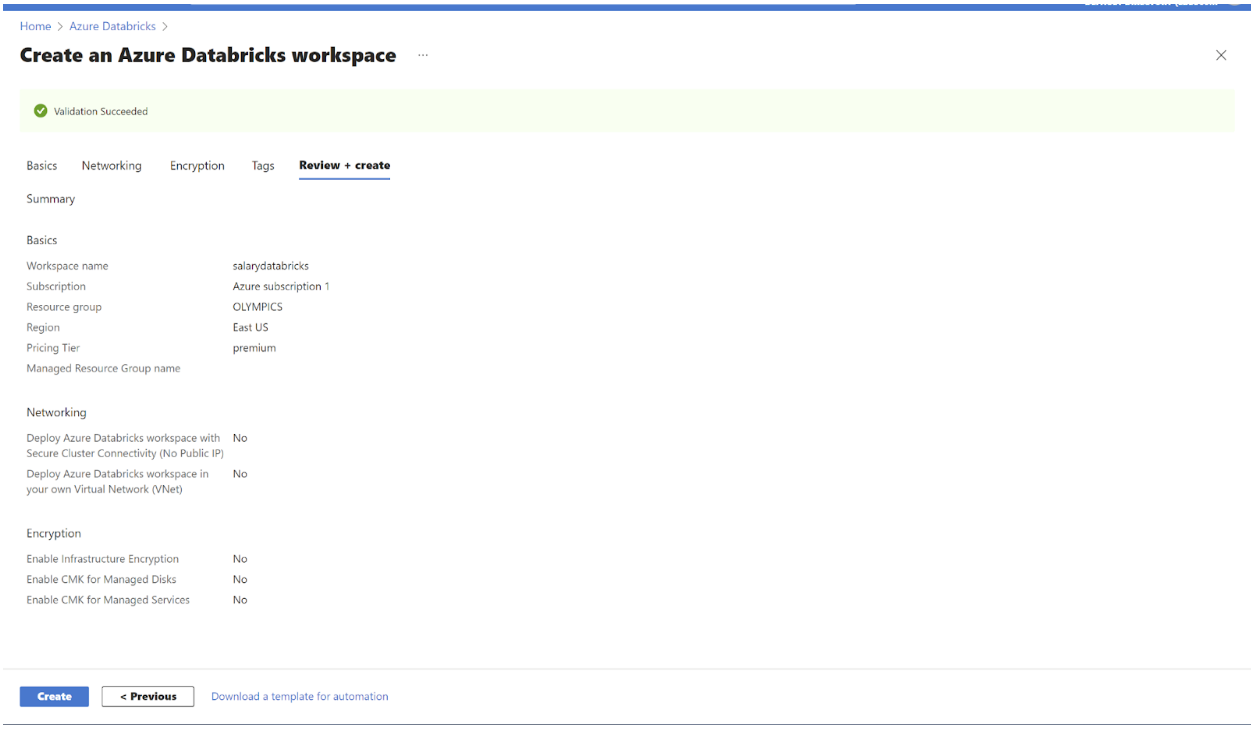
View the role assignments that grant access to this and other resources.
Learn more

- **Create DataBricks workspace**

Analyzing Salary data



Status	: Active	Managed Resource Group	: databricks-rg-salarydatabricks-znefdfgxbcmjm
Resource group	: OLYMPICS	URL	: https://adb-8138412232999980.0.azuredatabricks.net
Location	: East US	Pricing Tier	: Premium (+ Role-based access controls) (Click to change)
Subscription	: Azure subscription 1		
Subscription ID	: a48476a6-cf26-49c8-9f3f-3ea70ebb36bd		
Tags (edit)	: Add tags		


Home > Azure Databricks >
Create an Azure Databricks workspace
Basics
Networking
Encryption
Tags
Review + create
Create
< Previous
Download a template for automation

Validation Succeeded

Summary

Basics

Workspace name	salarydatabricks
Subscription	Azure subscription 1
Resource group	OLYMPICS
Region	East US
Pricing Tier	premium
Managed Resource Group name	

Networking

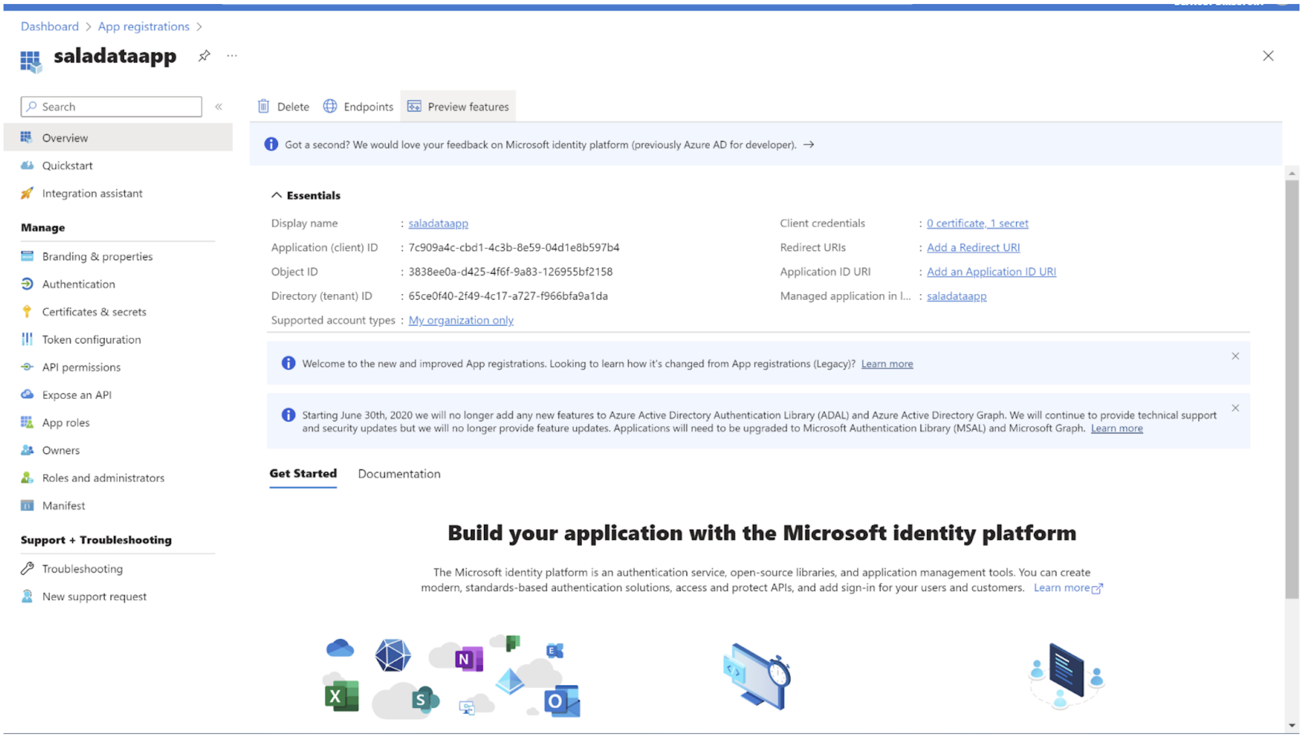
Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)	No
Deploy Azure Databricks workspace in your own Virtual Network (VNet)	No

Encryption

Enable Infrastructure Encryption	No
Enable CMK for Managed Disks	No
Enable CMK for Managed Services	No

- **Create a new App registration to mount data factory in databricks**

Analyzing Salary data



Got a second? We would love your feedback on Microsoft identity platform (previously Azure AD for developer). →

Display name : saladataapp
Application (client) ID : 7c909a4c-cbd1-4c3b-8e59-04d1e8b597b4
Object ID : 3838ee0a-d425-4f6f-9a83-126955bf2158
Directory (tenant) ID : 65ce0f40-2f49-4c17-a727-f966bfa9a1da
Supported account types : My organization only

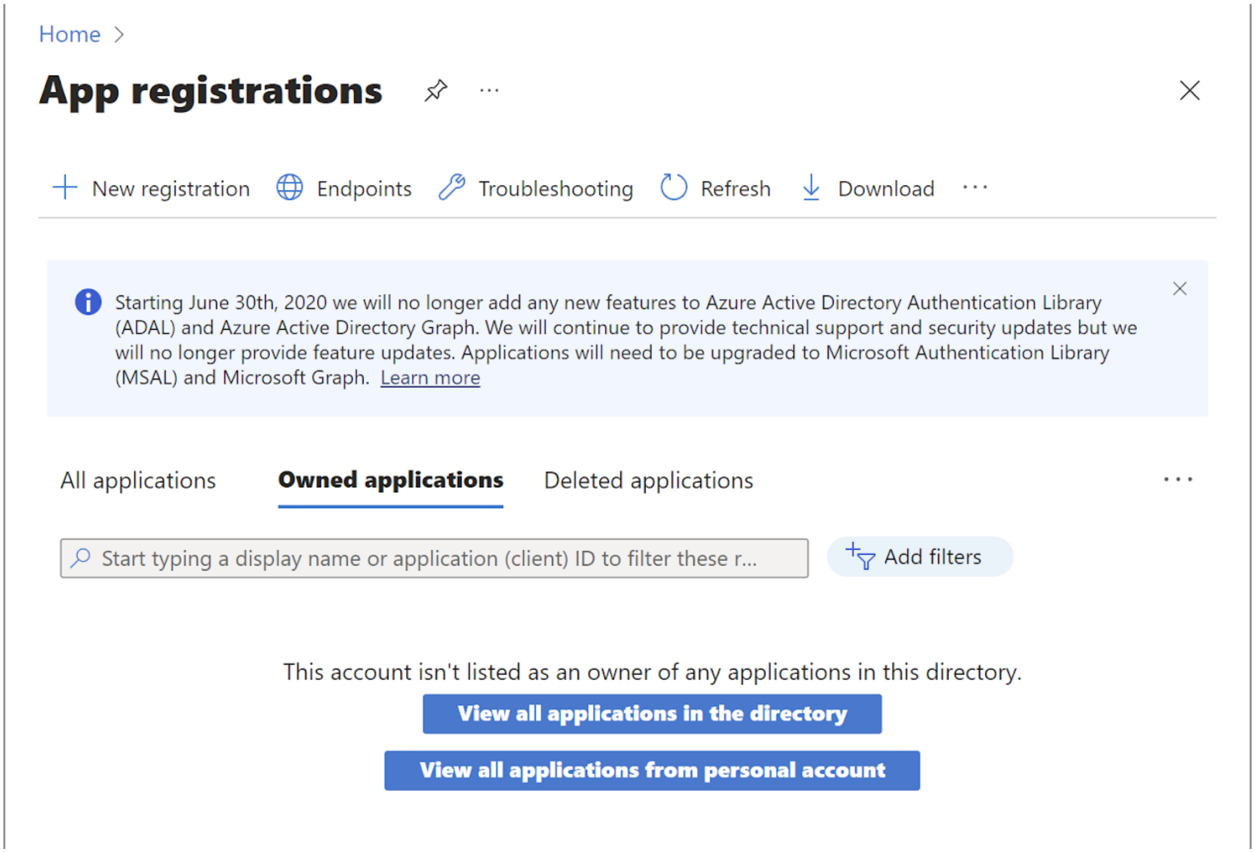
Client credentials : 0 certificate, 1 secret
Redirect URIs : Add a Redirect URI
Application ID URI : Add an Application ID URI
Managed application in l... : saladataapp

Welcome to the new and improved App registrations. Looking to learn how it's changed from App registrations (Legacy)? Learn more

Starting June 30th, 2020 we will no longer add any new features to Azure Active Directory Authentication Library (ADAL) and Azure Active Directory Graph. We will continue to provide technical support and security updates but we will no longer provide feature updates. Applications will need to be upgraded to Microsoft Authentication Library (MSAL) and Microsoft Graph. Learn more

Build your application with the Microsoft identity platform

The Microsoft identity platform is an authentication service, open-source libraries, and application management tools. You can create modern, standards-based authentication solutions, access and protect APIs, and add sign-in for your users and customers. Learn more



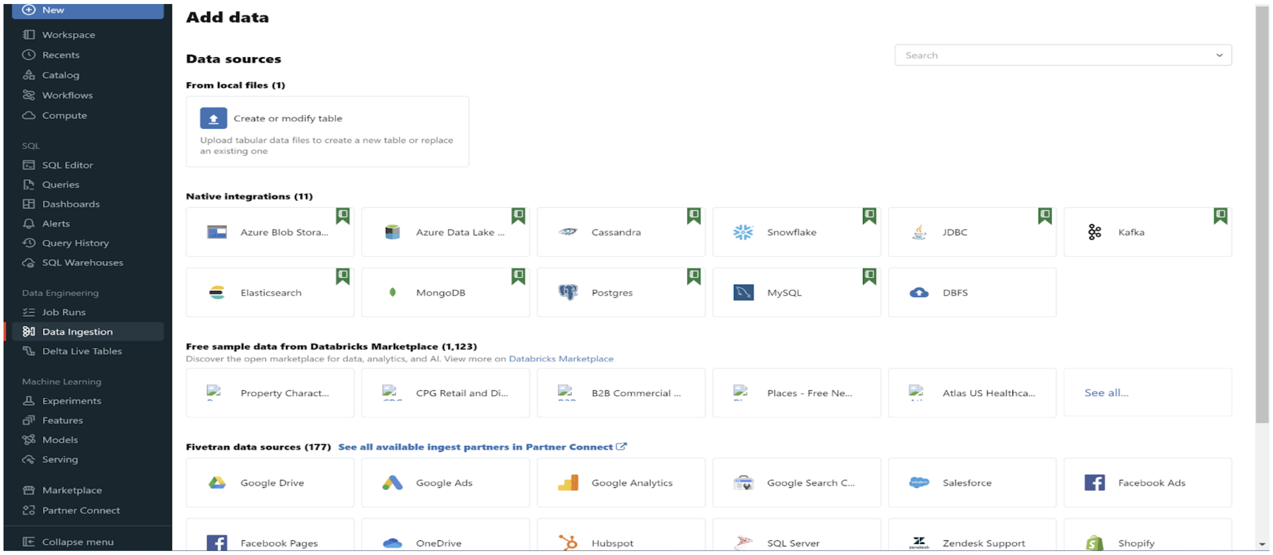
Starting June 30th, 2020 we will no longer add any new features to Azure Active Directory Authentication Library (ADAL) and Azure Active Directory Graph. We will continue to provide technical support and security updates but we will no longer provide feature updates. Applications will need to be upgraded to Microsoft Authentication Library (MSAL) and Microsoft Graph. Learn more

Start typing a display name or application (client) ID to filter these r...

This account isn't listed as an owner of any applications in this directory.

Analyzing Salary data

- **Else add data through files/ various integrations**



Add data

Data sources

From local files (1)

Create or modify table

Upload tabular data files to create a new table or replace an existing one

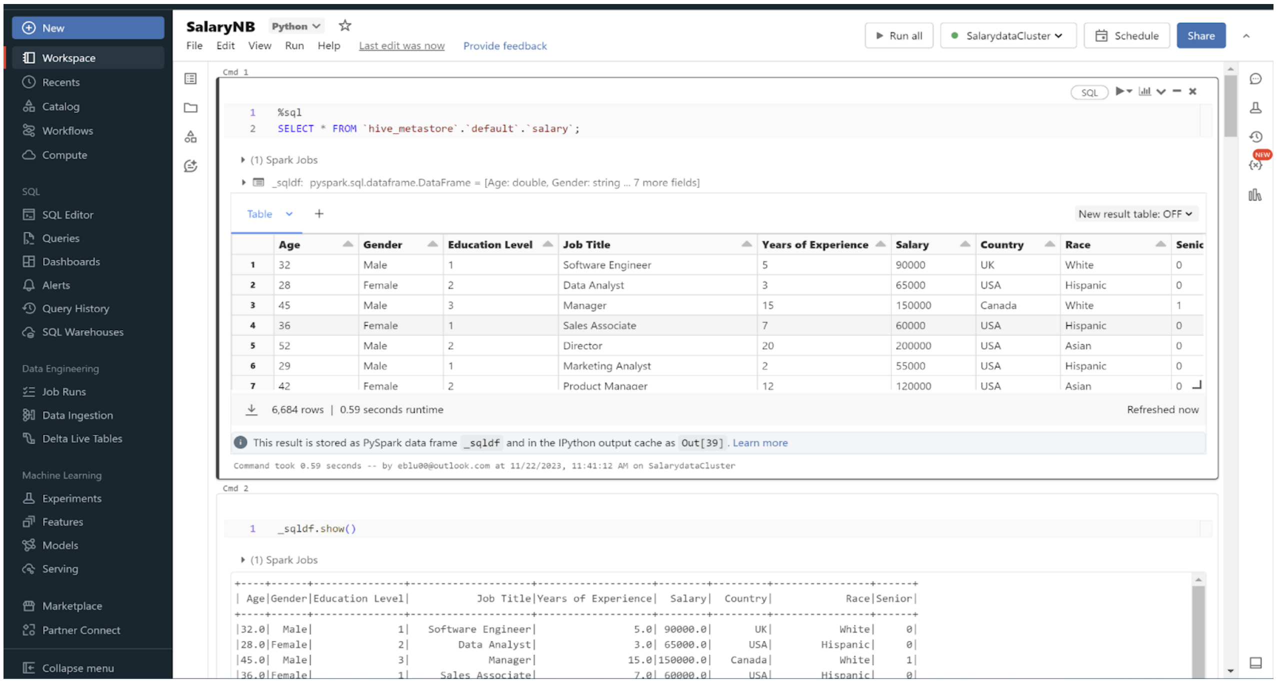
Native integrations (11)

Free sample data from Databricks Marketplace (1,123)

Discover the open marketplace for data, analytics, and AI. View more on Databricks Marketplace

Fivetran data sources (177) See all available ingest partners in Partner Connect

- **Create a new notebook >Python and ensure your cluster is selected. Then load data used PySpark**



SalaryNB

Cmd 1

```
%sql
SELECT * FROM `hive_metastore`.`default`.`salary`;
```

(1) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [Age: double, Gender: string ... 7 more fields]

	Age	Gender	Education Level	Job Title	Years of Experience	Salary	Country	Race	Senic
1	32	Male	1	Software Engineer	5	90000	UK	White	0
2	28	Female	2	Data Analyst	3	65000	USA	Hispanic	0
3	45	Male	3	Manager	15	150000	Canada	White	1
4	36	Female	1	Sales Associate	7	60000	USA	Hispanic	0
5	52	Male	2	Director	20	200000	USA	Asian	0
6	29	Male	1	Marketing Analyst	2	55000	USA	Hispanic	0
7	42	Female	2	Product Manager	12	120000	USA	Asian	0

6,684 rows | 0.59 seconds runtime

Refreshed now

This result is stored as PySpark data frame `_sqldf` and in the IPython output cache as `Out[39]`. Learn more

```
Command took 0.59 seconds -- by eblu00@outlook.com at 11/22/2023, 11:41:12 AM on SalarydataCluster
```

Cmd 2

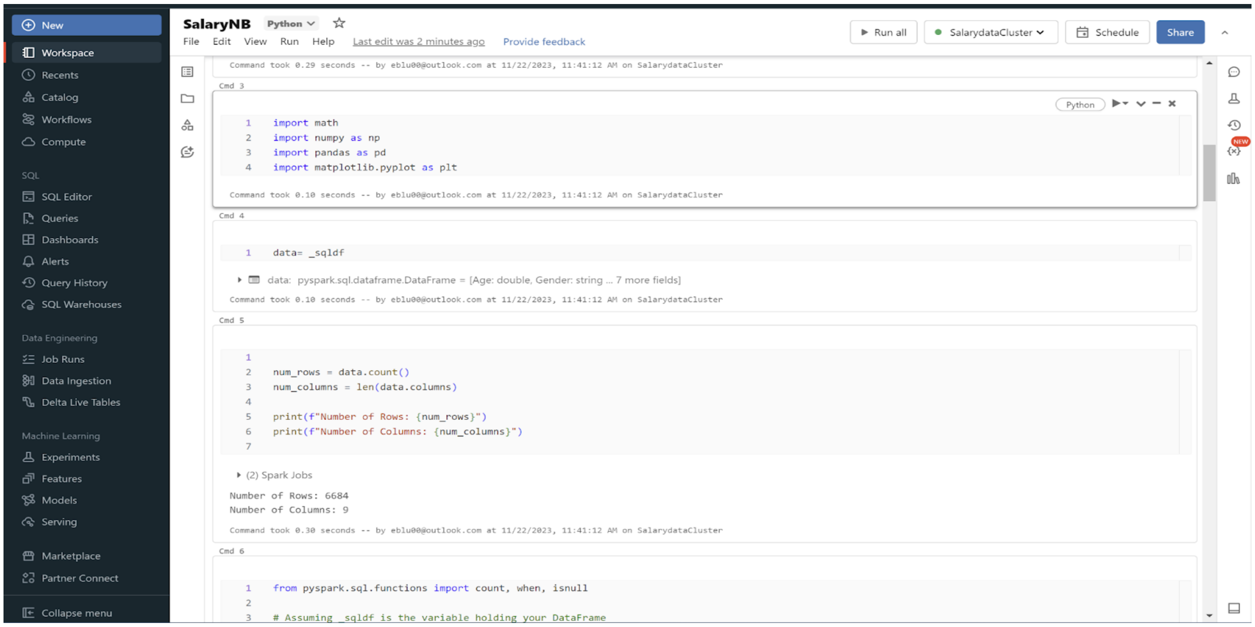
```
_sqldf.show()
```

(1) Spark Jobs

```
+----+------+---------------+-----------------+-------------------+--------+-------+---------------+------+
| Age|Gender|Education Level|        Job Title|Years of Experience|  Salary|Country|           Race|Senior|
+----+------+---------------+-----------------+-------------------+--------+-------+---------------+------+
|32.0|  Male|              1|Software Engineer|                5.0| 90000.0|     UK|          White|     0|
|28.0|Female|              2|     Data Analyst|                3.0| 65000.0|    USA|       Hispanic|     0|
|45.0|  Male|              3|          Manager|               15.0|150000.0| Canada|          White|     1|
|36.0|Female|              1|  Sales Associate|                7.0| 60000.0|    USA|       Hispanic|     0|
```

Analyzing Salary data

- **Using Spark Dataframe we performed Analysis**



```
Command took 0.29 seconds -- by eblu00@outlook.com at 11/22/2023, 11:41:12 AM on SalarydataCluster
Cmd 3
import math
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

Command took 0.10 seconds -- by eblu00@outlook.com at 11/22/2023, 11:41:12 AM on SalarydataCluster
Cmd 4
data= _sqldf

data:  pyspark.sql.dataframe.DataFrame = [Age: double, Gender: string ... 7 more fields]
Command took 0.10 seconds -- by eblu00@outlook.com at 11/22/2023, 11:41:12 AM on SalarydataCluster
Cmd 5

num_rows = data.count()
num_columns = len(data.columns)

print(f"Number of Rows: {num_rows}")
print(f"Number of Columns: {num_columns}")


(2) Spark Jobs
Number of Rows: 6684
Number of Columns: 9
Command took 0.30 seconds -- by eblu00@outlook.com at 11/22/2023, 11:41:12 AM on SalarydataCluster
Cmd 6
from pyspark.sql.functions import count, when, isnull

# Assuming _sqldf is the variable holding your DataFrame
```

Tableau:

The Tableau process undertaken involved the utilization of a dataset in CSV format. To initiate the analysis, the dataset was uploaded onto Tableau using the Spatial file upload option. This method allows for the seamless integration of spatial data, providing a rich foundation for exploration and visualization.

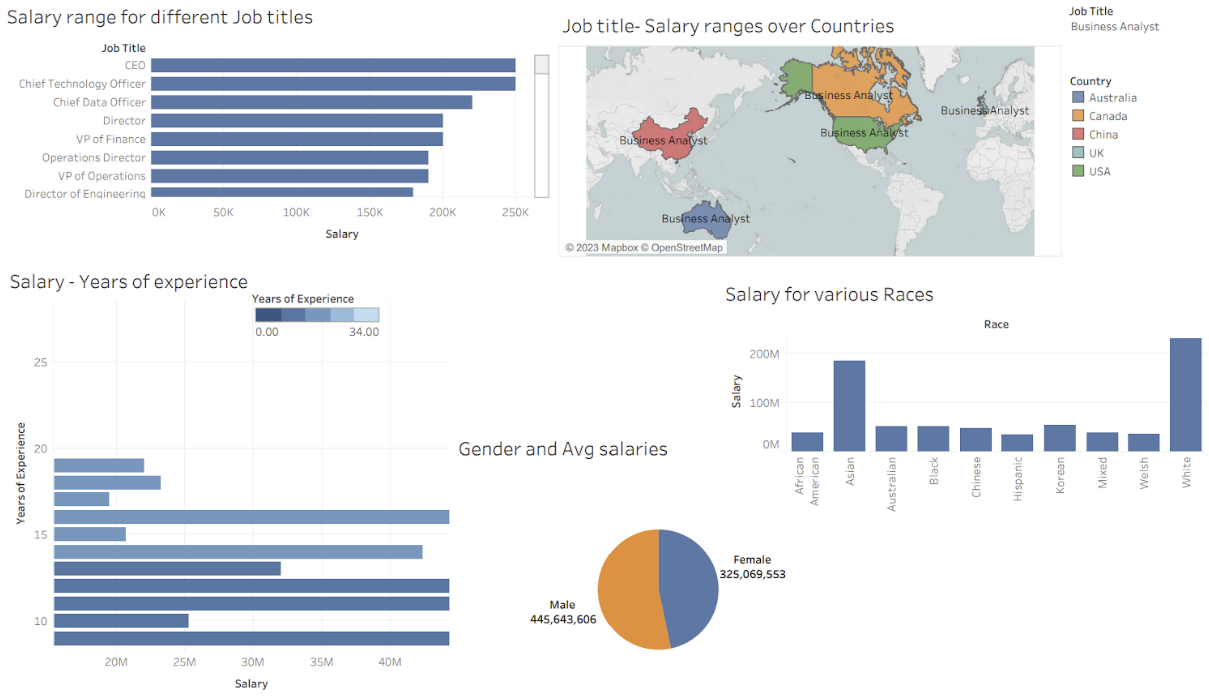
Once the dataset was successfully uploaded, the next step involved extracting the data into the Tableau workspace. Extraction allows for faster data processing within Tableau, enhancing the overall efficiency of the analysis.

Subsequently, the analytical exploration of the dataset took place. This was accomplished by creating separate sheets within Tableau, each focusing on specific features or aspects of the data. Breaking down the analysis into individual sheets enables a more granular examination of each variable or category, facilitating a deeper understanding of the dataset.

Concurrently, a key aspect of the process involved the creation of a dashboard. The dashboard serves as a consolidated and visually intuitive representation of the analyzed data. It acts as a user-friendly interface that brings together insights from various sheets, making it easy to communicate and comprehend the findings.

The emphasis on creating an easily explainable dashboard suggests a focus on clarity and accessibility. By designing the dashboard with a user-friendly approach, the intent is to make the insights readily understandable to a diverse audience, ensuring that the analytical results are effectively communicated.

Tableau Viz:



Let's break down each visualization

Bar Graph - Salary Range by Job Title:

- The bar graph illustrates the salary ranges associated with different job titles.
- Notably, the 'CEO' position stands out with the highest remuneration, followed by positions such as CTO, CDO, Director, and VP.
- A trend emerges where executive-level titles consistently command higher salaries on average.

Analyzing Salary data

Map - Average Salary of Selected Job Titles Worldwide:

- The map provides a global perspective, displaying the average salary for selected job titles across different countries, with USD as the base currency.
- The color-coded map facilitates easy differentiation between countries.
- The observation indicates that the USA generally offers more competitive salaries compared to other countries.

Bar Graph - Salary vs. Experience:

- The bar graph explores the relationship between work experience and salary.
- Contrary to expectations, it reveals that salaries vary irrespective of experience levels. Some individuals with less experience earn higher salaries, while those with more experience may earn lower salaries.
- The range of 10-15 years of experience emerges as a period with more competitive salaries.

Pie Charts - Gender Disparity in Salary:

- The pie charts aim to understand salary disparities between male and female employees.
- The observation indicates that, on average, females earn a comparatively lower salary than males.
- However, the difference, while present, is not overwhelmingly significant, suggesting a potential for equalization in the future.

Bar Chart - Salary vs. Race:

- The bar chart examines salary discrepancies based on race.
- Clear trends emerge, indicating that individuals belonging to the white and Asian races generally earn higher salaries compared to other racial groups.

In conclusion, the data visualizations present a comprehensive snapshot of salary dynamics across job titles, global locations, experience levels, gender, and race. Executives consistently earn higher salaries, the USA stands out for competitive wages, experience alone does not linearly determine earnings, gender pay gaps exist but are not overwhelming, and disparities based on race are evident, with white and Asian individuals earning higher salaries on average. These findings provide valuable insights for understanding and addressing various aspects of salary inequality.