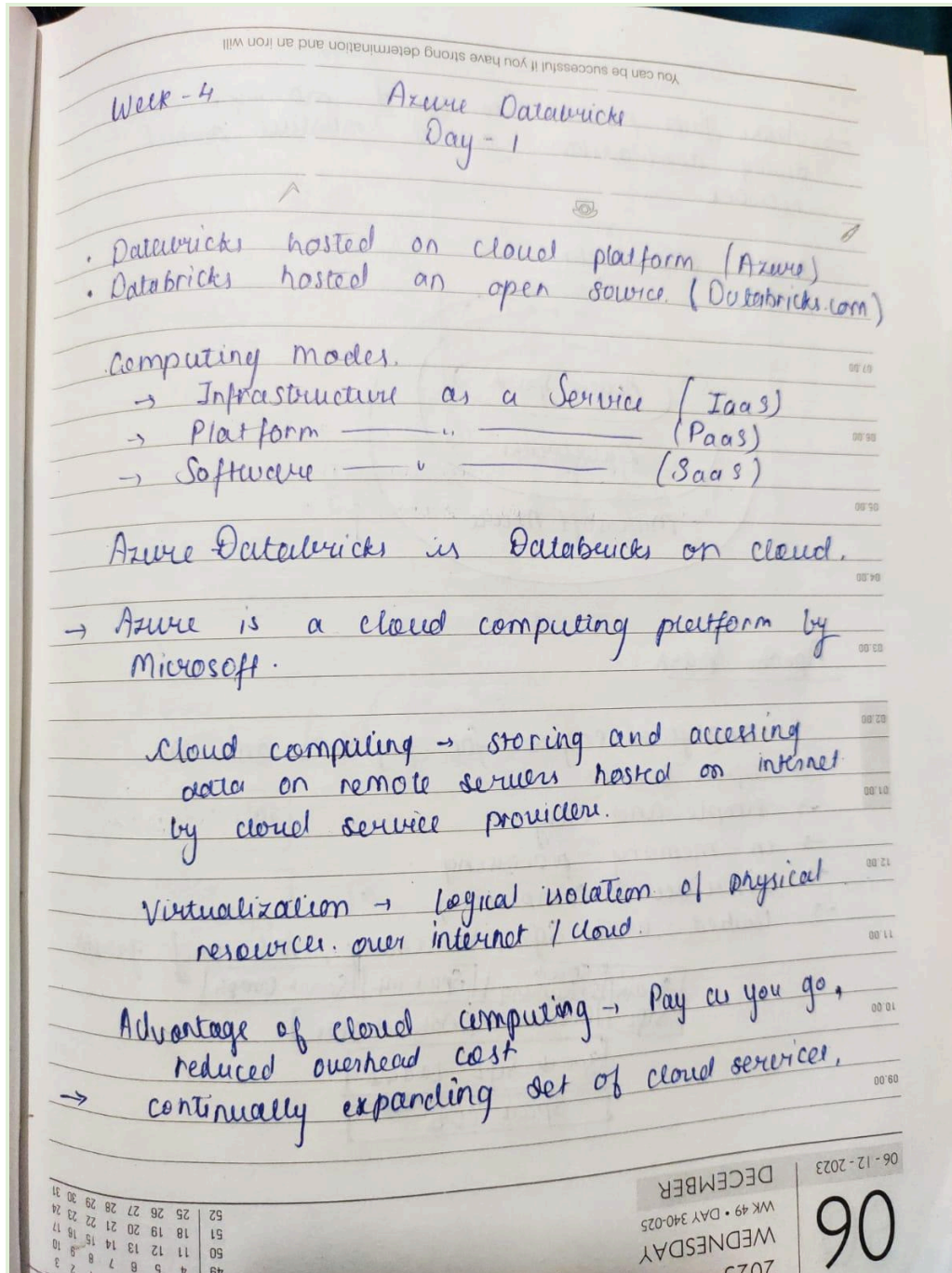


Azure Databricks Assignment - 1

Mitushi Vishwakarma

- Introduction to databricks
- Notes :



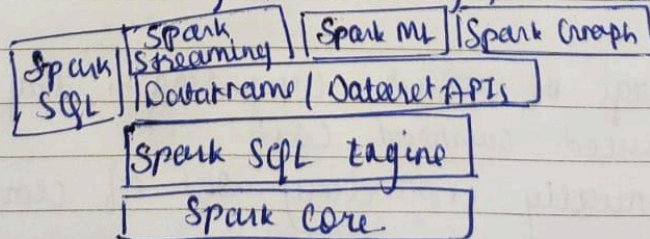
- Azure gives freedom to build, manage and deploy applications on a massive global network

Azure Databricks

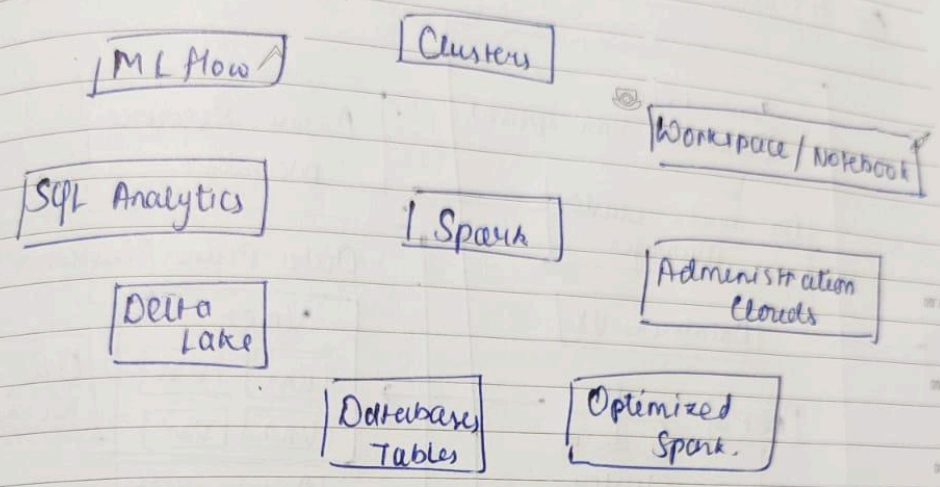


Apache Spark :-

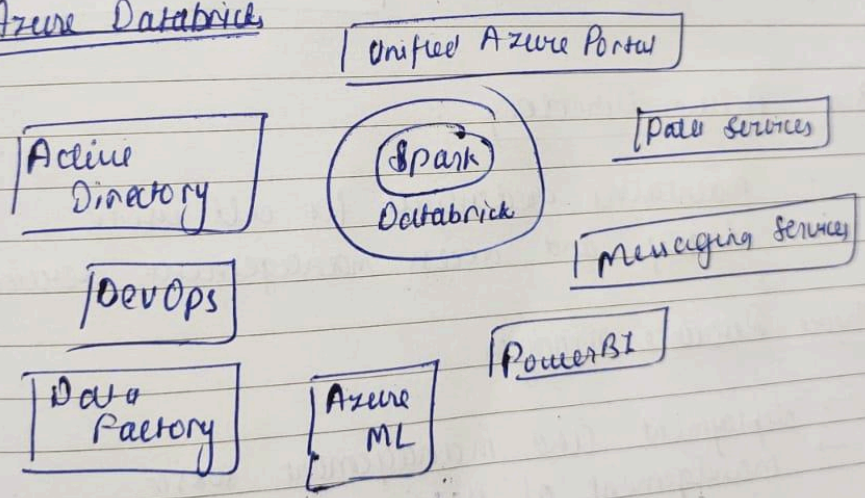
- analytical engine for big data and ML
- Open source
- simple and easy
- in-memory processing
- distributed computing
- unified with SQL, Streaming, ML and graphs



Databricks



Azure Databricks

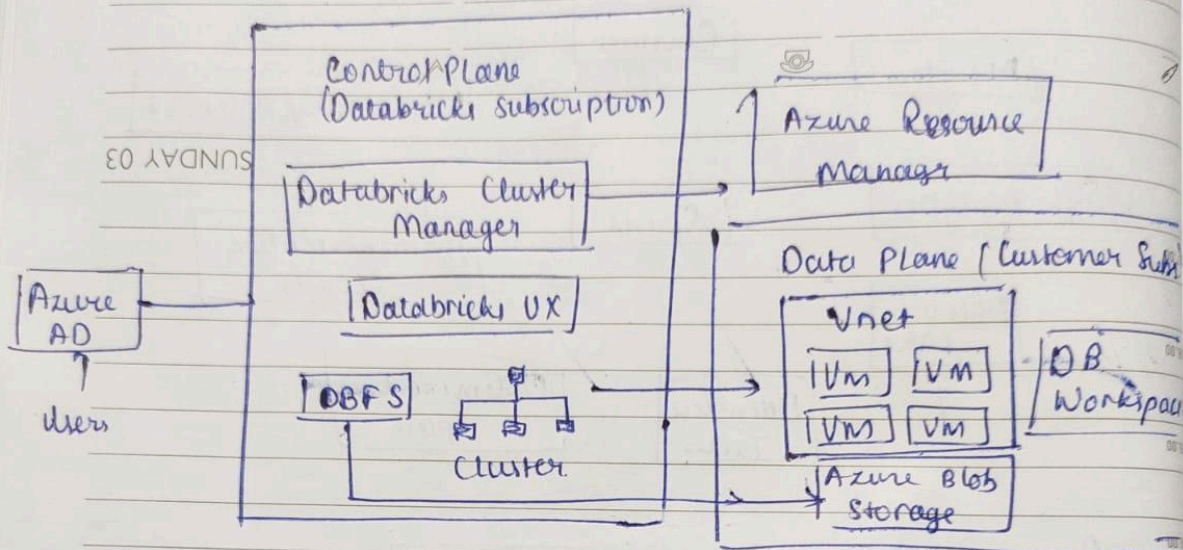


50	11	12	13	14	15	16	17
51	18	19	20	21	22	23	24
52	25	26	27	28	29	30	31

AD → Active Directory

Individual commitment to a group effort is what makes a company work

AD Architecture



Azure Active Directory :

- maintains credentials for all users
- identity and access management service

Azure Resource Manager :

- deployment and management service
- management of resources in Azure account

Azure Blob Storage :-

- Scalable storage in cloud.

02 - 12 - 2023

02

DECEMBER

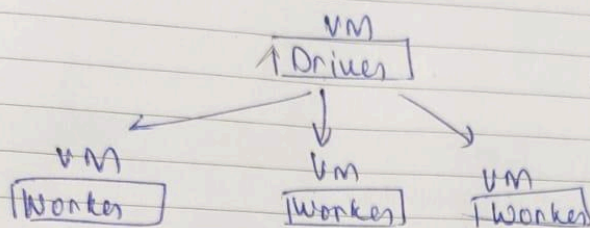
WK 48 • DAY 336-029

SATURDAY

2023

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Databricks Clusters



Driver sends data to Workers.

We run query on databricks notebook it connects to driver then to workers and gives o/p.

Cluster Types

All purpose

Created manually
expensive
shared
persistent
for interactive workloads

Job Cluster

Created by job
cheaper
isolated for job
terminate at end of job
for automation

Configuration of cluster decides cost.

Single / Multi Node. → multiple virtual machines

↓
less cost
single VMs

DECEMBER
WK 48 • DAY 335-030
FRIDAY
2023

01-12-2023

01

52	51	50	49	48
25	18	19	12	4
26	19	20	13	5
27	20	21	14	6
28	21	22	15	7
29	22	23	16	8
30	23	24	17	9
31	24	25	18	10

● Setting up Azure Databricks Workspace

The image shows two screenshots from the Azure portal. The top screenshot displays a list of existing Azure Databricks workspaces. The bottom screenshot shows the 'Create an Azure Databricks workspace' wizard, specifically the 'Project Details' and 'Instance Details' sections.

Existing Azure Databricks Workspaces

Name	Type	Resource group	Location	Subscription
AzureDataBricks_1098	Azure Databricks Service	rg-azuser1098_mml.local-1c29r	Central India	Azure subscription 1
azure_databricks_stan	Azure Databricks Service	azure-rg	East US	Azure subscription 1
AzureWSTest-AUS	Azure Databricks Service	rg-azuser1092_mml.local-gv7uw	Australia Central	Azure subscription 1
databricks_hexa	Azure Databricks Service	rg-azuser1062_mml.local-xrIDN	Central India	Azure subscription 1
DatabricksTraining-1083	Azure Databricks Service	rg-azuser1083_mml.local-5QutK	East US	Azure subscription 1
DatabricksTraining-1083-EA	Azure Databricks Service	rg-azuser1083_mml.local-5QutK	East Asia	Azure subscription 1
DB_learning	Azure Databricks Service	rg-azuser1089_mml.local-22dNT	Central India	Azure subscription 1
db_ws_siva	Azure Databricks Service	rg_siva	East US	Azure subscription 1
DBOne1099	Azure Databricks Service	rg-azuser1099_mml.local-zlwGt	West Europe	Azure subscription 1
Dbrickslearn1086_AustraliaE	Azure Databricks Service	rg-azuser1086_mml.local-qvNvp	Australia East	Azure subscription 1
hexa-deb-1054	Azure Databricks Service	rg-azuser1062_mml.local-vrIDN	South India	Azure subscription 1

Create an Azure Databricks workspace - Project Details

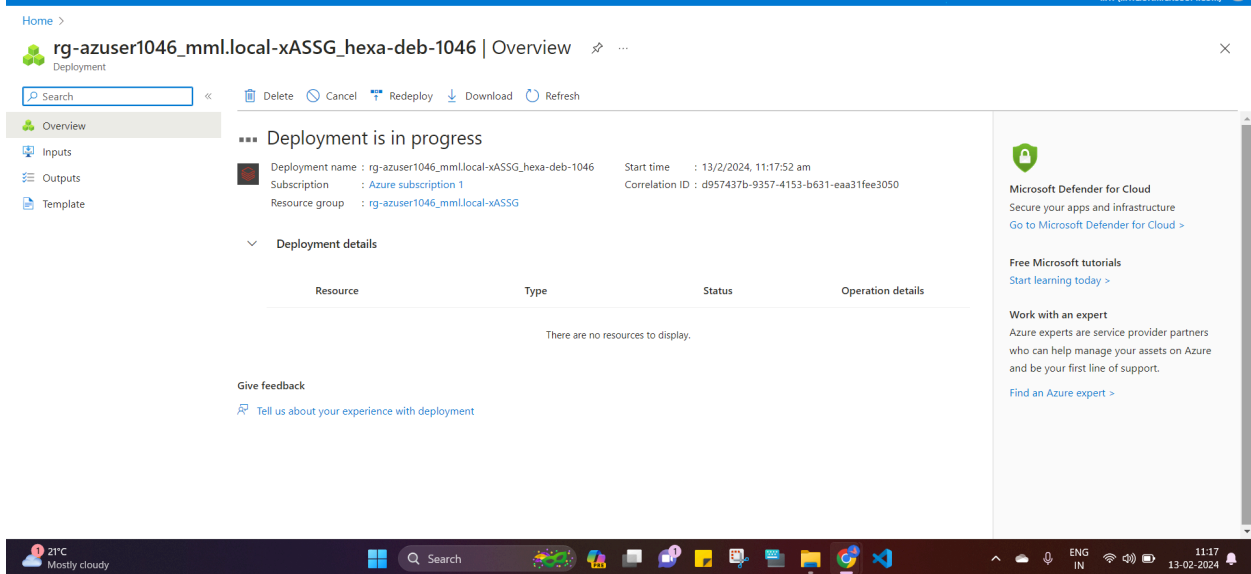
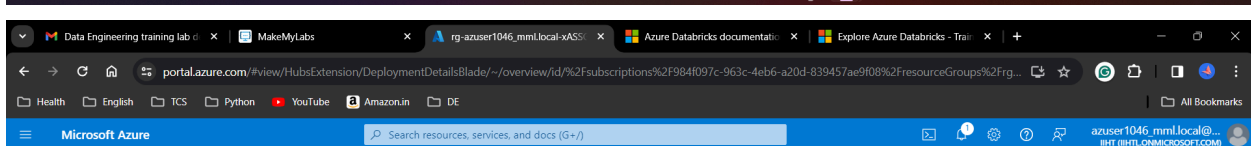
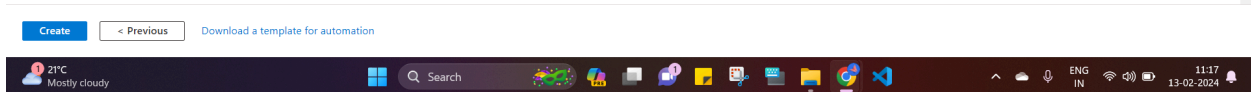
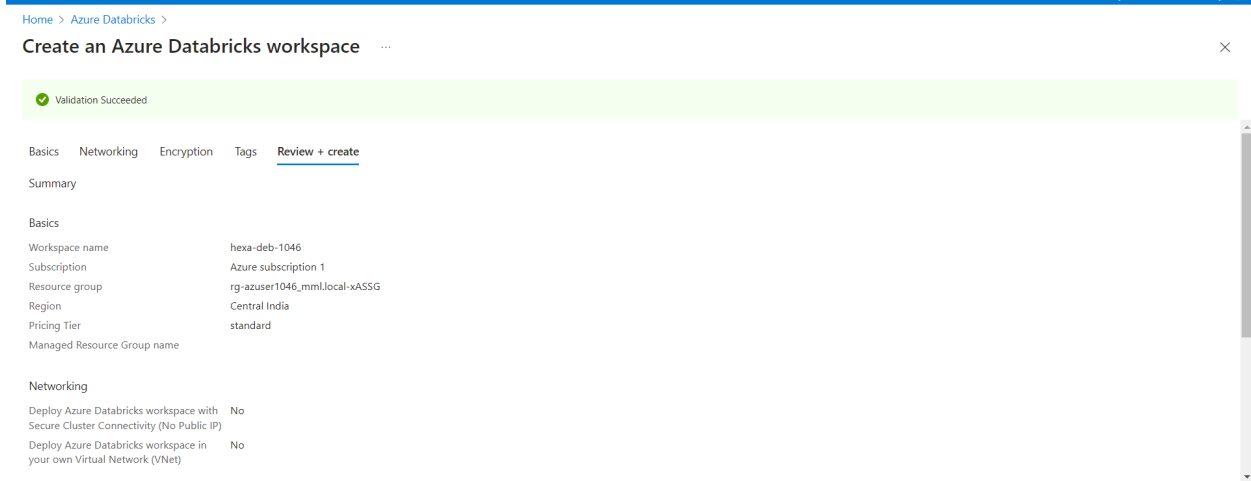
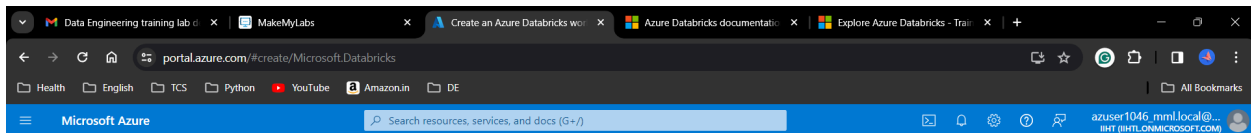
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription: Azure subscription 1
Resource group: rg-azuser1046_mml.local-xASSG [Create new](#)

Create an Azure Databricks workspace - Instance Details

Workspace name: hexa-deb-1046
Region: Central India
Pricing Tier: Standard (Apache Spark, Secure with Microsoft Entra ID)
Managed Resource Group name: Enter name for managed resource group

Buttons: Review + create, < Previous, Next: Networking >



portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions%2F984f097c-963c-4eb6-a20d-839457ae9f08%2FresourceGroups%2Frg-azuser1046_mml.local-xASSG_hexa-deb-1046 | Overview

Microsoft Azure

Home > rg-azuser1046_mml.local-xASSG_hexa-deb-1046 | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Your deployment is complete

Deployment name : rg-azuser1046_mml.local-xASSG_hexa-deb-1046 Start time : 13/2/2024, 11:17:52 am
Subscription : Azure subscription 1 Correlation ID : d957437b-9357-4153-b631-eea31fee3050
Resource group : rg-azuser1046_mml.local-xASSG

Deployment details

Next steps

Go to resource

Give feedback

Tell us about your experience with deployment

Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
[Set up cost alerts >](#)

Microsoft Defender for Cloud
Secure your apps and infrastructure
[Go to Microsoft Defender for Cloud >](#)

Free Microsoft tutorials
[Start learning today >](#)

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
[Find an Azure expert >](#)

21°C Mostly cloudy 11:20 13-02-2024

portal.azure.com/#@ihl.onmicrosoft.com/resource/subscriptions/984f097c-963c-4eb6-a20d-839457ae9f08/resourceGroups/rg-azuser1046_mml.local-xASSG/providers/... | Overview

Microsoft Azure

Home > rg-azuser1046_mml.local-xASSG_hexa-deb-1046 | Overview >

hexa-deb-1046
Azure Databricks Service

Search

Delete

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Virtual Network Peerings

Encryption

Networking

Properties

Locks

Automation

CLI / PS

Tasks (preview)

Export template

Essentials

Status : Active

Managed Resource Group : [databricks-rg-hexa-deb-1046-geybk32eodeus](#)

Resource group : [rg-azuser1046_mml.local-xASSG](#)

URL : [https://adb-8517674426995631.11.azuredatabricks.net](#)

Location : Central India

Pricing Tier : [Standard \(Apache Spark, Secure with Microsoft Entra ID\) \(Click to see details\)](#)

Subscription : [Azure subscription 1](#)

Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08

Tags (edit) : [Add tags](#)

[Launch Workspace](#)

[Upgrade to Premium](#)

[Documentation](#)

[Getting Started](#)

[Import Data from File](#)

[Import Data from Azure Storage](#)

21°C Mostly cloudy 11:21 13-02-2024

- **Setting up Azure Databricks workspace and configuring clusters**

This screenshot shows the Azure Databricks cluster configuration page for a new cluster. The cluster is named 'azuser1046_mml.local@iihtl.onmicrosoft.com's Cluster'. The configuration is set to 'Multi node' with 'Single user access'. The 'Performance' section shows 'Databricks Runtime Version' as '13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12)' and 'Use Photon Acceleration' is checked. The 'Worker type' is 'Standard_D4ads_v5' with '16 GB Memory, 4 Cores'. The 'Driver type' is also 'Standard_D4ads_v5' with '16 GB Memory, 4 Cores'. The 'Summary' box on the right indicates 1 Worker, 16 GB Memory, 4 Cores, and 1 Driver, 16 GB Memory, 4 Cores. The 'Runtime' is '13.3.x-scala2.12'. The 'Photon' and 'Standard_D4ads_v5' buttons are highlighted. The 'Terminate' button is visible.

This screenshot shows the same Azure Databricks cluster configuration page, but the cluster is now in a 'Running' state, indicated by a green checkmark next to the cluster name. The configuration details remain the same as in the previous screenshot. The 'Workers' column now shows '1' worker, and the 'Current' column shows '1' current worker. The 'Spot instances' checkbox is still unchecked. The 'Summary' box and 'Photon'/'Standard_D4ads_v5' buttons are still present. The 'Terminate' button is still visible.

- **Creating a Databricks notebook and Implementing Databricks for real-time data processing**

The screenshot displays the Databricks web interface. The browser address bar shows the URL: `adb-8517674426995631.11.azuredatabricks.net/?o=8517674426995631#notebook/3569440616992487`. The notebook is titled "Untitled Notebook 2024-02-13 11:32:23" and is in Python mode. The left sidebar contains navigation options like Workspace, Recents, Catalog, Workflows, Compute, Data Engineering, Job Runs, Delta Live Tables, Machine Learning, Experiments, Features, Models, and Serving. The main area shows a code cell with the following Python code:

```
1 import pyspark
2 from pyspark.sql import SparkSession
3
4 # Initializing Spark Session
5 spark = SparkSession.builder.appName("Manipulation in dataframes").getOrCreate()
6 # Creating dataframe
7 data = [
8     ('Mitushi', 22, 'F', 1000),
9     ('Vishesh', 24, 'M', 2000),
10    ('Tanisha', 22, 'F', 3000),
11    ('Zamran', 38, 'M', 5000)
12 ]
13 columns = ["Name", "Age", "Gender", "Salary"]
14 df = spark.createDataFrame(data=data, schema=columns)
15 df.show()
```

Below the code, the output of the `df.show()` command is displayed as a table:

Name	Age	Gender	Salary
Mitushi	22	F	1000
Vishesh	24	M	2000
Tanisha	22	F	3000
Zamran	38	M	5000

The bottom of the screen shows a Windows taskbar with the date and time: 11:33, 13-02-2024.