



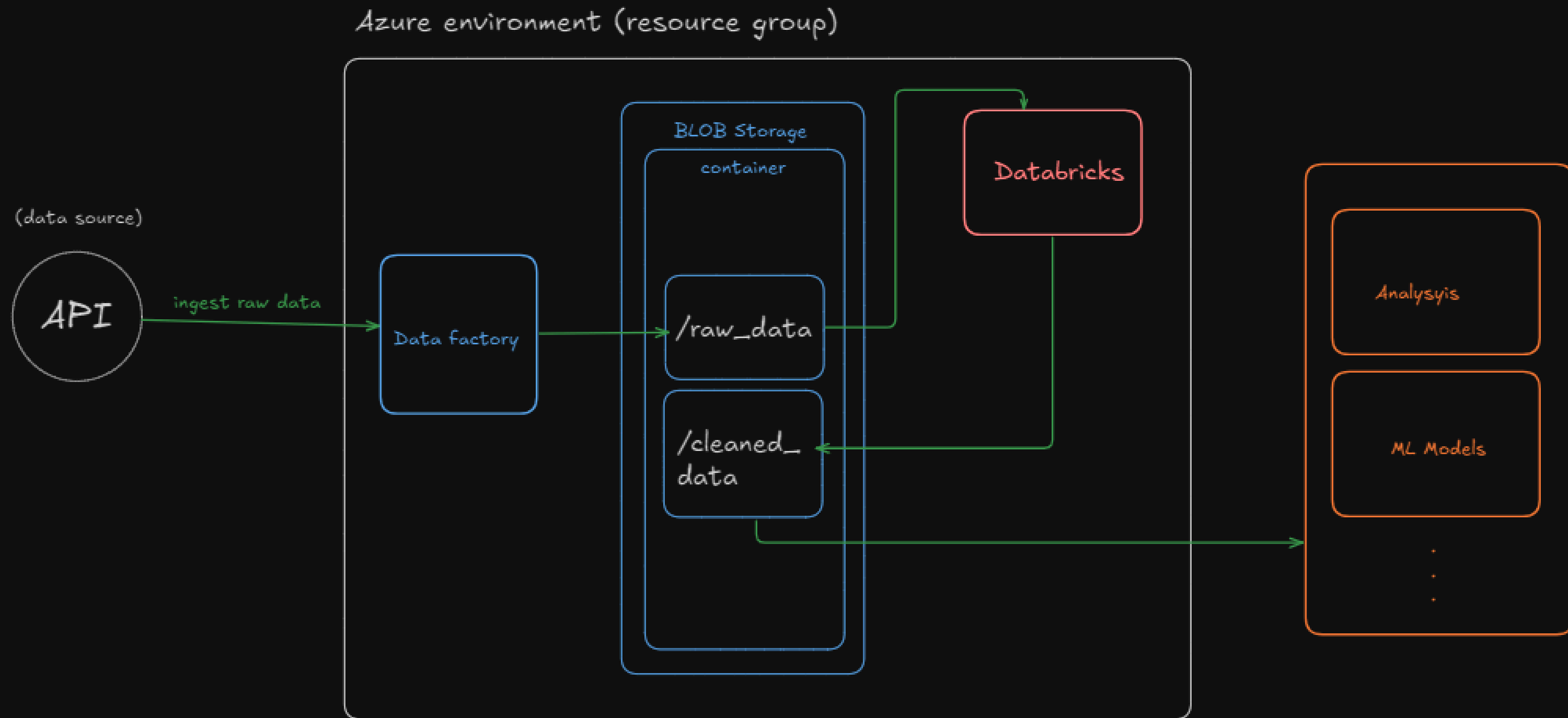
# ***DATA ENGINEERING***

———— **FUTURENSE** ————

AJAY CHELLIAH, JAI ADITHYA, MEENU, KRUSHNA KOUSHIK, DINESH KUMAR, VISHAL, GAUTHAM  
KRISHNA



# DATA ENGINEERING ARCHITECTURE (AZURE)



# DATA BRICKS PIPELINE (INGESTION)

Microsoft Azure Data Factory | futureuse-usp-df

Search factory and documentation

meenusree@lpu.in  
LOVELY PROFESSIONAL UNIVERSITY

Data Factory | Validate all | Publish all

pipeline1

Validate | Debug | Trigger (1)

Expand resources pane

Copy data | candidate\_application\_tracker | Notebook | transform\_candidate\_application\_tracker

Copy data | leads\_generated | Notebook | transform\_leads\_generated

Copy data | phone\_metrics | Notebook | transform\_phone\_metrics

Copy data | tokens\_paid | Notebook | transform\_token\_paid

Copy data | campaign\_performance | Notebook | transform\_campaign\_performance

Copy data | webinar\_leads

Parameters | Variables | Settings | Output

# AZURE BLOB STORAGE

Microsoft Azure

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

Copilot

meenusree@lpu.in  
LOVELY PROFESSIONAL UNIVERS...

Home > Storage accounts > futureseuspmain | Containers >

futureseuspmaincontainer

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: futureseuspmaincontainer

Search blobs by prefix (case-sensitive)

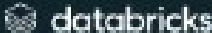
Show deleted objects

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

# DATABRICKS (DATA TRANSFORMATION)

## CLUSTER


Microsoft Azure



Search data, notebooks, recents, and more...

CTRL + P

futureense\_usp\_databricks



New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Machine Learning

Playground

Experiments

Features

Models

Serving

Compute

meenu avuthu's Cluster

Configuration

Notebooks (0)

Libraries

Event log

Spark UI

Driver logs

Metrics

Apps

Spark compute UI - Master

Multi node

Single node

Access mode

Single user access

Single user

meenu avuthu

Performance

Databricks Runtime Version

15.3 (includes Apache Spark 3.5.0, Scala 2.12)

Use Photon Acceleration

Node type

Standard\_DS3\_v2

14 GB Memory, 4 Cores

Terminate after

120

minutes of inactivity

Tags

No custom tags

Automatically added tags

Advanced options

More

Terminate

Edit

Summary

1 Driver

14 GB Memory, 4 Cores

Runtime

15.3.x-scala2.12

Photon

Standard\_DS3\_v2

1.5 DBU/h

# DATABRICKS WORKSPACE

Microsoft Azure databricks

Search data, notebooks, recents, and more...CTRL + P

futureuse\_osp\_databricks

New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Machine Learning

Playground

Experiments

Features

Models

Serving

Recents

Name	Last viewed	Type
phone_metrics	9 hours ago	Notebook
campaign_performance	12 hours ago	Notebook
leads_generated	13 hours ago	Notebook
tokens_paid	13 hours ago	Notebook
playground	16 hours ago	Notebook
candidate_application_tracker	16 hours ago	Notebook

# DATA BRICKS NOTEBOOK

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

futureense\_usp\_databricks

New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Machine Learning

Playground

Experiments

Features

Models

Serving

campaign\_performance

Python

File Edit View Run Help

Last edit was 2 days ago

Provide feedback

Run all

meenu awuthu's Cluster

Schedule

Share

13 hours ago (4s)

3

```
import pandas as pd
from io import BytesIO
from azure.storage.blob import BlobServiceClient

connection_string = "DefaultEndpointsProtocol=https;AccountName=futureenseuspmain;AccountKey=VsBCM36R1E0RjT4R/eHIUbe0w2S3kep8TnIHonQnqaHeJdcCdp2fphkR58T+Rkt6xfDv9bcHgtvn+AStkoRmbA==;
EndpointSuffix=core.windows.net"
container_name = "futureenseuspmaincontainer"
input_blob_name = "raw_data/campaign_performance.csv"
blob_service_client = BlobServiceClient.from_connection_string(connection_string)
blob_client = blob_service_client.get_blob_client(container=container_name, blob=input_blob_name)
blob_data = blob_client.download_blob().readall()

df = pd.read_csv(BytesIO(blob_data))
print("Raw Data:")
print(df.head())
```

Raw Data:

	dates	...	adset_name
0	2024-04-29	...	GHAT/GRE
1	2024-05-02	...	USP-2_KA_TN_050424
2	2024-05-02	...	USP-2_KA_TN_050424
3	2024-05-02	...	USP-2_KA_TN_050424
4	2024-05-02	...	USP-2_AP_TS_050424

[5 rows x 11 columns]

13 hours ago (<1s)

4

```
df.info()
```

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 28534 entries, 0 to 28533

Data columns (total 11 columns):

#	Column	Non-Null Count	Dtype
0	dates	28534 non-null	object
1	campaign name	28534 non-null	object

# CONNECTING DATA TO POWERBI FROM BLOB STORAGE

