Here's a **10-step hands-on assignment guide for Objective 1** (**Test Data Management**) tailored for execution on the **Microsoft Fabric** platform. These steps will walk you through generating synthetic data, managing production samples, anonymizing data, and injecting edge cases — all using Microsoft Fabric tools like **Lakehouse**, **Notebooks**, and **Dataflows Gen2**.

Hands-On Assignment Guide (Objective 1: Test Data Management)

Platform: Microsoft Fabric
Target Skills: Synthetic data creation, anonymization, edge-case data handling, using Spark & Lakehouse in Fabric

---

✅ Step 1: Set Up Your Workspace in Microsoft Fabric

- Open Microsoft Fabric.
- Create a new **Workspace** (e.g., `Test Data Management Lab`).
- Enable **Lakehouse**, **Notebooks**, and **Data Pipelines** options for the workspace.

✅ Step 2: Create a Lakehouse

- In the workspace, click on **New** > **Lakehouse**.
- Name it (e.g., `TestLakehouse`).
- This will serve as the data storage layer (Delta format) for your test datasets.

✅ Step 3: Launch a Spark Notebook

- In your Lakehouse, click **New Notebook**.
- Attach it to the `TestLakehouse`.
- Set language to **PySpark**.
- Rename the notebook to `SyntheticDataGenerator`.

✅ Step 4: Generate Synthetic Data using Faker

Paste and run the following code in your notebook:

Python

```python
from faker import Faker
import pandas as pd
from pyspark.sql import SparkSession

fake = Faker()
records = [{'id': i, 'name': fake.name(), 'email': fake.email(), 'dob': fake.date_of_birth()} for i in range(1000)]
pdf = pd.DataFrame(records)
df = spark.createDataFrame(pdf)

df.write.format("delta").mode("overwrite").saveAsTable("synthetic_users")
```

✅ This writes a table named `synthetic_users` to the Lakehouse.

✅ Step 5: Create a Dataflow Gen2 to Pull Sample Production Data

- Go to **Dataflows Gen2** > **New dataflow**.
- Choose a source (e.g., Azure SQL DB, SharePoint, or upload a CSV).
- Select a small table to simulate production data.
- Apply basic transformations (e.g., rename columns).

- **Load data into the Lakehouse** created earlier (e.g., table name: `prod_sample_users`).

✅ Step 6: Anonymize Production Sample Data in a Notebook

Create a new notebook named `AnonymizationNotebook`. Use the following Spark code:

Python

```python
import hashlib
from pyspark.sql.functions import udf, col

def hash_value(val):
    return hashlib.sha256(val.encode()).hexdigest() if val else None

hash_udf = udf(hash_value)

df = spark.read.table("prod_sample_users")
df_anon = df.withColumn("email", hash_udf(col("email"))) \
            .withColumn("name", hash_udf(col("name")))

df_anon.write.format("delta").mode("overwrite").saveAsTable("anon_users")
```

✅ This creates an anonymized version of your production data.

✅ Step 7: Inject Edge Cases into Synthetic Data

In your `SyntheticDataGenerator` notebook, add this:

Python

```python
from pyspark.sql import Row

# Create edge cases
edge_cases = [
    Row(id=None, name="", email="invalid_email", dob=None),  # Nulls and invalids
    Row(id=9999, name="Test User", email="test@example.com", dob="3000-01-01"),  # Future DOB
    Row(id=-1, name="Negative ID", email="neg@example.com", dob="2000-01-01")  # Invalid ID
]

df_edge = spark.createDataFrame(edge_cases)
df_combined = df.union(df_edge)

df_combined.write.format("delta").mode("overwrite").saveAsTable("synthetic_users_with_edge_cases")
```

✅ Step 8: Visualize Tables in Lakehouse Explorer

- Go to the **Lakehouse Explorer**.
- Validate your tables:
  - `synthetic_users`
  - `prod_sample_users`
  - `anon_users`
  - `synthetic_users_with_edge_cases`
- Click on each table to preview the data and schema.

✅ Step 9: Schedule Data Generation with Data Pipelines

- Go to **Data Pipelines** > **New pipeline**.
- Add a **Notebook activity** and choose `SyntheticDataGenerator`.

- Add a trigger (e.g., daily or manual run).
- Save and publish the pipeline.

✅ Step 10: Document Your Work in OneLake Notes / OneNote

- Create a `README` style document or **OneNote page** linked to your workspace.
- Summarize:
  - What each table represents
  - How anonymization and edge cases are handled
  - Screenshots

🏁 Outcome

By completing these 10 steps, you will have:

- A synthetic dataset for testing
- Sample production data pulled securely
- Anonymized data compliant with privacy rules
- Edge cases injected for robust test scenarios
- Automation and scheduling using Data Pipelines
