**End-to-End Project for Data Pipeline and Reporting**

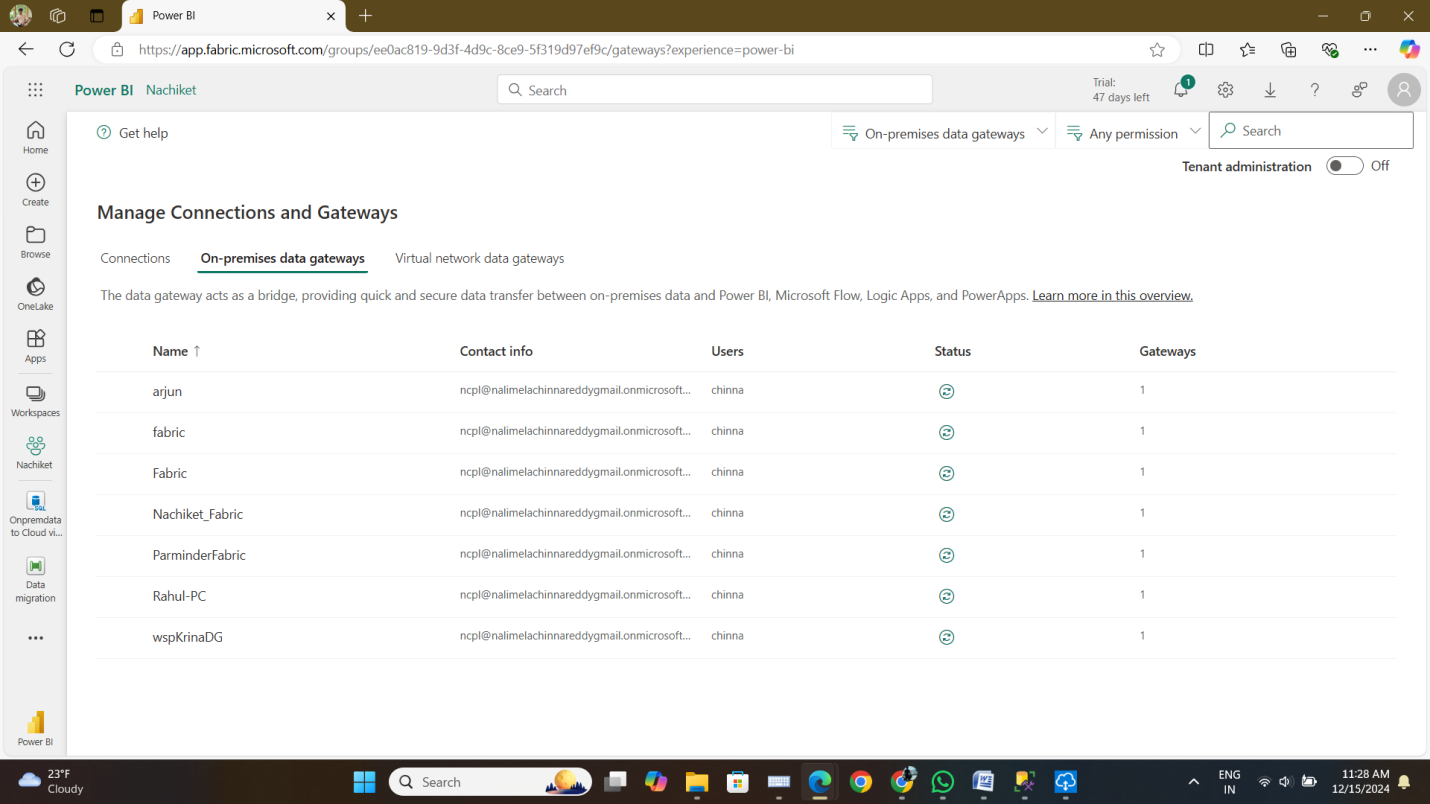
**Nachiket Prajapati**

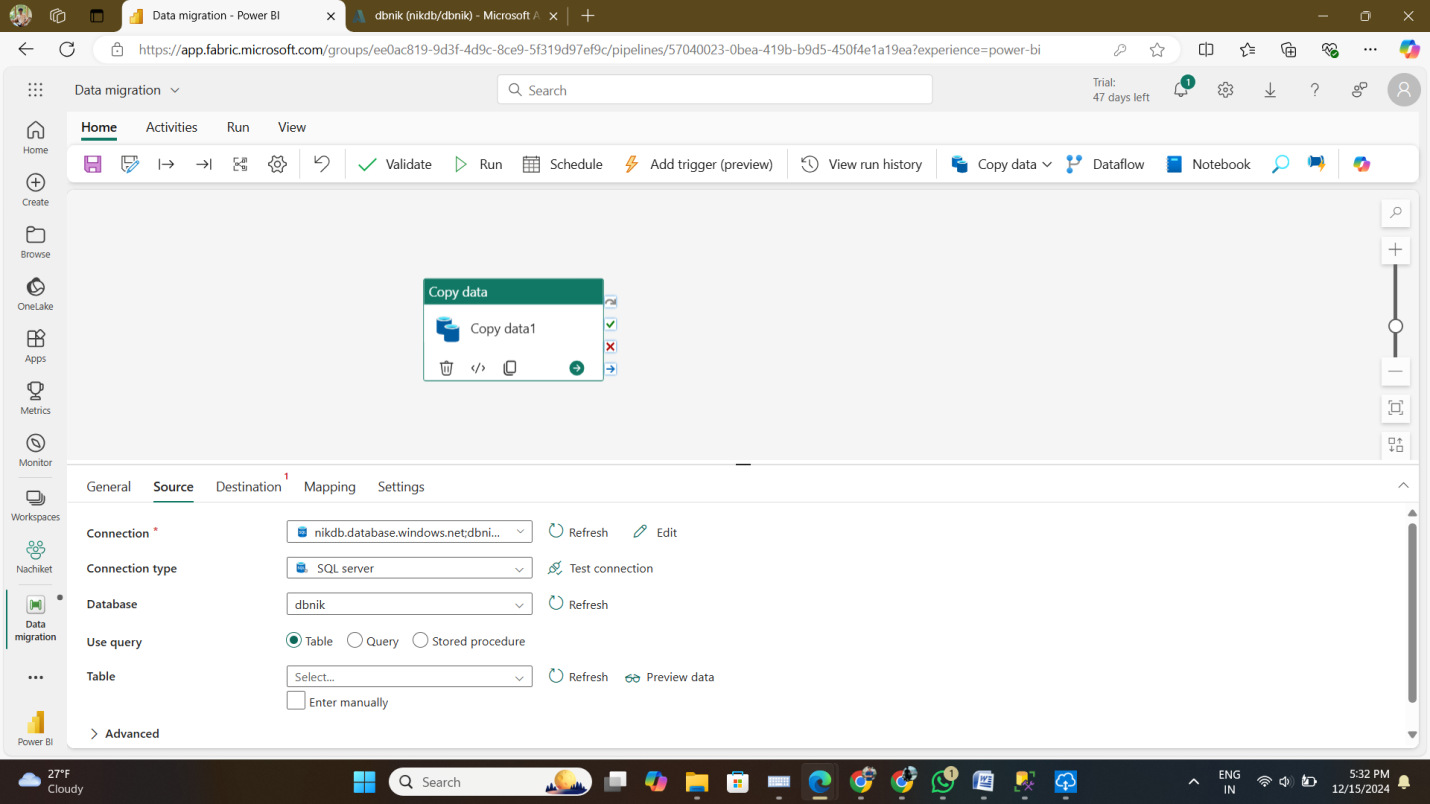
**Batch:- NCPL (C-19)**

**Objective:** Build a complete data pipeline from on-premises data to cloud storage, process and clean data, and create reports using the Azure Data Lakehouse and Fabric.

### **Project Outline and Expected Outputs**

1. **Data Ingestion**
   * **Objective:** Ingest data from an on-premises environment to an Azure Data Lake Gen2 storage account.





* + **Expected Output:**
    - A folder named **Bronze** in the Data Lake Gen2 storage account containing raw data in its original format.
    - Verification logs or metadata confirming successful data ingestion.

1. **Data Cleaning and Transformation**
   * **Objective:** Apply data cleanup techniques and convert the raw data into Delta format.
   * **Expected Output:**
     + A folder named **Silver** in the Data Lake Gen2 storage account containing cleaned and transformed data in Delta format.
     + A summary of cleanup operations performed (e.g., removal of duplicates, handling null values).
2. **Data Processing for SCD Type 1**
   * **Objective:** Use data from the **Silver** folder to implement Slowly Changing Dimension (SCD) Type 1 logic and store the results in a new Delta format.
   * **Expected Output:**
     + A folder named **Gold** in the Data Lake Gen2 storage account containing processed data with SCD Type 1 implementation in Delta format.
     + A log or report detailing the changes made during the SCD Type 1 transformation.
3. **Lakehouse Shortcut**
   * **Objective:** Create a shortcut in the Lakehouse to access the data stored in the Data Lake Gen2 storage account.
   * **Expected Output:**
     + A functional shortcut in the Lakehouse pointing to the **Gold** folder in the Data Lake Gen2 storage account.
     + Verification that the shortcut provides seamless access to the data.
4. **Reporting in Fabric**
   * **Objective:** Use the Lakehouse data to generate a report in Fabric.
   * **Expected Output:**
     + A comprehensive report based on the data in the Lakehouse.
     + Visualizations, insights, and metrics relevant to the data processed.

**Deliverables:**

* Screenshots or proof of each step (e.g., folders in Data Lake, data transformations, and reports in Fabric).
* A brief document summarizing the process, tools, and techniques used.

**Note to Trainees:**

* Focus on creating an automated and efficient pipeline.
* Ensure proper documentation of all steps for future reference.
* Use Azure Data Factory, Delta Lake, and other Azure tools as needed for implementation.