A PROJECT REPORT
ON
NEXUS DATALENS – AN END-TO-END DATA ENGINEERING PROJECT

Submitted in partial fulfilment of the
Requirements for the award of the Degree of

MASTER OF SCIENCE (INFORMATION TECHNOLOGY) 2024-25

PREPARED BY
RUKSANA SHAIKH (6716)

UNDER THE ESTEEMED GUIDANCE OF
PROF. JEBA ROSELET

DEPARTMENT OF INFORMATION TECHNOLOGY
RAMNIRANJAN JHUNJHUNWALA COLLEGE OF ARTS, SCIENCE & COMMERCE (EMPOWERED AUTONOMOUS), GHATKOPAR (W) 400 086
MUMBAI, MAHARASHTRA

(Affiliated to University of Mumbai)
Hindi Vidya Prachar Samiti's
Ramniranjan Jhunjhunwala College of Arts, Science and Commerce
(Empowered Autonomous)
Ghatkopar (West), Mumbai – 86

Hindi Vidya Prachar Samiti's
Ramniranjan Jhunjhunwala College of Arts, Science and Commerce
(Empowered Autonomous)
Ghatkopar (West), Mumbai 400 086

CERTIFICATE

*This is to certify that the project entitled "**Nexus DataLens**" is bonafide work of **Ruksana Shaikh (6716)** submitted in partial fulfillment of the requirements for the award of the degree of Master of Science in Information Technology from University of Mumbai during the academic year 2024-2025*

_____ _____

Internal Guide External Examiner

_____ _____

Coordinator CollegeSeal

R. J. COLLEGE of Arts, Science & Commerce (Empowered Autonomous)
(Hindi Vidya Prachar Samiti's **RAMNIRANJAN JHUNJHUNWALA COLLEGE of Arts, Science & Commerce)**
Opposite Ghatkopar Railway Station, Ghatkopar (West), Mumbai - 400086, Maharashtra, INDIA.
Website: www.rjcollege.edu.in • Email : rjcollege@rjcollege.edu.in • Tel No. + 91 22 25151763
College is recognized under Section 2(f) & 12(B) of the UGC Act, 1956
Affiliated to UNIVERSITY OF MUMBAI || NAAC Re-Accredited 'A' Grade (CGPA: 3.10)

Ref: 1322

04th December 2024

To,
Mr. Gaurav Bhoi and Ms. Ruksana Shaikh
Department of Information Technology
Ramniranjan Jhunjhunwala College
Ghatkopar (W),
Mumbai 400 086

Re: Approval of funding for your Master's Research Project

Dear Mr. Bhoi & Ms. Shaikh,

We are pleased to inform you that the Research Advisory Committee has approved your proposal for a Master's Research Project titled **"Nexus DataLens (An end-to-end data analytics project using AWS)"** under the guidance of Ms. Jeba Roselet. Consequently, a seed money of not more than Rs. 24,850/- has been granted which you have to spend before 15th March 2025 as much as possible.

Under this project, you are entitled to travel in public transport (second class) within the city. Also, there is an upper limit of Rs. 500/- per day to be spent on lodging and boarding. The grant money cannot be used to meet the publication cost of the research papers.

It is necessary to strictly adhere to the guidelines for the release/ reimbursements of grants attached along with this letter.

Wishing you the best for your research work.

Sincerely Yours,

Principal
PRINCIPAL
RAMNIRANJAN JHUNJHUNWALA COLLEGE
OF ARTS, SCIENCE & COMMERCE (AUTONOMOUS)
Ghatkopar (W), Mumbai - 400 086, Maharashtra, INDIA

ABSTRACT

In the modern era of data-driven decision-making, organizations require efficient data pipelines to ingest, transform, and analyze data at scale. Traditional data processing methods often involve manual intervention, inefficiencies, and scalability challenges, making automated solutions essential for ensuring accuracy, speed, and reliability.

This project, Nexus DataLens, is an end-to-end data analysis and transformation platform that leverages Microsoft Azure services to automate data workflows. The system allows users to upload structured datasets (CSV files) through a web-based frontend, which are then processed via an automated Azure-based data pipeline. Key services include Azure Data Factory (ADF) for data ingestion, Azure Data Lake for storage, and Azure Databricks for data transformation using Apache Spark. The transformed data is stored back in Azure Data Lake, allowing users to securely download the processed dataset through the frontend.

The primary objective of this project is to enhance data processing efficiency, scalability, and security while minimizing manual effort. The system enables seamless automation of ETL (Extract, Transform, Load) processes, ensuring secure, scalable, and optimized data management. The cloud-based architecture supports large datasets while maintaining data security, compliance, and accessibility.

Nexus Data Lens has applications in business intelligence, reporting, data warehousing, and enterprise-level data integration. It can be integrated with SQL databases, NoSQL stores, and AI/ML pipelines to facilitate advanced analytics and decision-making. The solution provides a robust, scalable, and automated approach to handling business-critical data, making it a powerful tool for organizations seeking efficiency in data processing and management.

ACKNOWLEDGEMENT

The success and outcome of this project were possible by the guidance and support from many people. I am incredibly privileged to have got this all along with the achievement of my project. It required a lot of effort from everyone involved in this project with me and I would like to thank them.

I express my sincere gratitude to **Mrs. Bharthi Bhole**, Head of the Department, for providing me with the opportunity to undertake this project. Her unwavering support, leadership, and encouragement have been instrumental in the successful completion of this work. The valuable resources and guidance provided by her played a crucial role in shaping the direction of this project.

I extend my heartfelt appreciation and deep respect to my esteemed professor, **Prof. Jeba Roselet**, whose keen interest and meticulous attention to detail ensured the precision and quality of this project. Her invaluable insights, constant guidance, and constructive feedback throughout every phase of the project have been pivotal in refining the outcome. Her dedication and mentorship have been a source of immense motivation and learning.

I am truly grateful for the enriching experience and knowledge gained through this project, which would not have been possible without the support and encouragement of my mentors.

Then I would like to thank my parents and friends who have helped me with their valuable suggestions and guidance has been helpful in various phases of the completion of the project.

Gaurav Ramesh Bhoi
Ruksana Shaikh

Date: _____

Index

INTRODUCTION

Data is one of the most valuable assets for businesses and organizations, serving as the foundation for informed decision-making, strategic planning, forecasting, and trend analysis. However, raw data is often unstructured, inconsistent, and requires extensive transformation before it can deliver meaningful insights. Manual data processing methods introduce challenges such as errors, inefficiencies, delays, and scalability limitations, making them unsuitable for modern, data-driven enterprises.

To address these challenges, this project presents Nexus DataLens—a fully automated, scalable, and cloud-based data pipeline that enables users to seamlessly upload, transform, and retrieve structured datasets. The system provides a web-based frontend where users can upload CSV files, which are then processed through an Azure-based data workflow. Key components include:

- Azure Data Factory (ADF) for automated data ingestion
- Azure Data Lake for secure and scalable storage
- Azure Databricks for powerful data transformation and ETL (Extract, Transform, Load) operations

Once the transformation is complete, the processed data is stored back in Azure Data Lake, allowing users to download the refined dataset from the frontend. The solution ensures data security, accuracy, and efficiency, while eliminating the need for manual intervention.

By leveraging cloud computing and big data technologies, this project aims to streamline data workflows, enhance scalability and performance, and provide organizations with a robust tool for managing large datasets. With applications in business intelligence, analytics, and data warehousing, Nexus Data Lens serves as a cutting-edge solution for enterprises looking to optimize their data transformation processes.

LITERATURE SURVEY

1. Introduction

In today's data-driven landscape, organizations are increasingly adopting cloud-based solutions to address the challenges of traditional Extract, Transform, Load (ETL) processes. Manual intervention, scalability limitations, and inefficiencies in handling large datasets have led to the development of modern, automated data pipelines. This literature survey focuses on existing methodologies, tools, and best practices for incremental ETL pipelines and cloud-based data integration, drawing insights from official Microsoft documentation and resources. It also highlights gaps in current approaches and justifies the adoption of Microsoft Azure services for the proposed solution.

2. Existing Solutions in Data Processing

a. Incremental Data Processing with Azure Data Factory

Incremental data processing has become a cornerstone of efficient data pipelines, enabling organizations to transfer only new or updated records instead of reloading entire datasets. According to Microsoft's official documentation on Azure Data Factory (ADF) , incremental loads can be implemented using techniques like Change Data Capture (CDC) , watermark columns, or timestamp-based filtering. ADF's ability to orchestrate complex workflows ensures seamless data movement between on-premises systems and cloud storage. Additionally, the integration of Self-Hosted Integration Runtime allows secure connectivity for hybrid environments, making it ideal for scenarios where data resides both on-premises and in the cloud.

b. Scalable Storage with Azure Data Lake Gen2

Azure Data Lake Gen2 is designed to handle massive volumes of structured and unstructured data while maintaining cost efficiency and performance. As outlined in Microsoft's Azure Data Lake Gen2 documentation , its hierarchical file system supports a layered architecture (Bronze, Silver, Gold), ensuring clarity and governance throughout the data pipeline. This approach enables organizations to store raw data, perform transformations, and generate business-ready datasets in a scalable and secure manner. Furthermore, the integration of role-based access control (RBAC) and encryption ensures compliance with industry standards.

c. Advanced Transformations with Azure Databricks

Azure Databricks, powered by Apache Spark , provides a collaborative environment for performing high-performance data transformations. According to Microsoft's Azure Databricks documentation , it supports distributed processing for large datasets, enabling operations like filtering, aggregations, joins, and machine learning model training. The platform's interactive notebooks allow teams to write code in Python, Scala, or SQL, fostering collaboration and flexibility. Additionally, its seamless integration with Azure Data Lake Gen2 ensures efficient data ingestion and output storage.

d. Unified Analytics with Azure Synapse Analytics

Azure Synapse Analytics combines big data processing with enterprise data warehousing, offering a unified platform for analytics. As described in Microsoft's Synapse Analytics documentation , it supports both batch and real-time queries, enabling users to analyze data at scale. Its integration with Power BI facilitates the creation of interactive dashboards and reports, empowering stakeholders to make data-driven decisions. Synapse Analytics also supports advanced analytics use cases, such as predictive modeling and trend analysis, making it a versatile tool for modern data pipelines.

e. Visualization with Power BI

Power BI serves as a powerful visualization tool for transforming processed data into actionable insights. According to Microsoft's Power BI documentation , it enables users to create interactive dashboards and reports that can be shared across teams. Its real-time connectivity to Azure services like Synapse Analytics and Data Lake Gen2 ensures timely and accurate visualizations. Additionally, Power BI's self-service capabilities empower non-technical users to explore data independently, fostering a data-driven culture within organizations.

3. Identification of Gaps

While Microsoft's ecosystem offers robust tools for building data pipelines, certain challenges remain:

- Complexity in Implementation : Configuring and integrating services like Azure Data Factory, Data Lake, and Databricks requires technical expertise, which may pose a barrier for smaller teams.
- Cost Management : While Azure services are scalable, managing costs for large-scale data pipelines can be challenging without proper optimization.
- Security Concerns : Although Azure provides tools like Azure Active Directory and Azure Key Vault for securing sensitive data, ensuring compliance with regulations like GDPR and HIPAA requires careful planning and implementation.
- Real-Time Processing Limitations : While incremental loads reduce latency, achieving true real-time analytics often requires additional configurations and optimizations.

These gaps highlight the need for a well-designed, end-to-end solution that addresses these challenges while leveraging the strengths of Microsoft Azure.

4. Justification of the Proposed Approach

The proposed solution leverages Microsoft Azure's ecosystem to build a comprehensive, scalable, and secure data pipeline. By utilizing the following components, the project ensures efficient data processing and actionable insights:

- Azure Data Factory : Automates incremental data ingestion from an on-premises SQL Server database to Azure Data Lake Gen2, reducing manual effort and ensuring data consistency.
- Azure Data Lake Gen2 : Provides a scalable storage solution with a layered architecture (Bronze, Silver, Gold), ensuring data governance and accessibility.
- Azure Databricks : Performs advanced transformations using Apache Spark, enabling high-performance processing of large datasets.
- Azure Synapse Analytics : Supports both batch and real-time analytics, bridging the gap between data engineering and business intelligence.
- Power BI : Visualizes insights through interactive dashboards, facilitating data-driven decision-making.
- Azure Active Directory and Key Vault : Ensure secure access control and protection of sensitive data, addressing compliance requirements.

This approach aligns with Microsoft's best practices and addresses the limitations of traditional ETL processes by integrating automation, scalability, and security into a unified framework.

5. Conclusion

The literature reviewed in this section highlights the capabilities of Microsoft Azure services in addressing modern data engineering challenges. By leveraging Azure Data Factory, Data Lake Gen2, Databricks, Synapse Analytics, and Power BI, organizations can build efficient, scalable, and secure data pipelines. While certain challenges remain, such as cost management and real-time processing, the proposed solution aims to bridge these gaps by implementing a robust, end-to-end architecture. This ensures that the pipeline meets the demands of today's data-driven organizations while adhering to industry standards and best practices.

USE CASE SUMMARY

Industry-level use cases include real-time inventory management in retail, fraud detection in finance, predictive maintenance in manufacturing, smart grid analytics in energy, and network performance monitoring in telecommunications. These use cases leverage the scalability, automation, and advanced analytics capabilities of Azure to drive business value.

The incremental load process from SSMS to Azure is highly versatile and can be applied across various industries to address specific business challenges. Whether it's real-time inventory management in retail, fraud detection in finance, or predictive maintenance in manufacturing, this architecture supports scalable, secure, and automated data pipelines.

Use Case: Real-Time Movie Inventory Management

- Scenario: A company needs to synchronize its on-premises inventory database with a cloud-based analytics platform to monitor levels in near real-time.
- How It Works:
    - Incremental loads transfer only updated inventory records (e.g., new stock arrivals, sold items) to Azure Data Lake.
    - Azure Databricks processes the data to calculate metrics like low stock alerts, reorder points, and sales trends.
    - Power BI visualizes the data for store managers and supply chain teams.
- Benefits:
    - Reduces overstock and stockouts by providing timely insights.
    - Enables predictive analytics for demand forecasting using AI/ML models.

PROJECT OVERVIEW

Nexus DataLens is an end-to-end data pipeline that uses Azure Data Factory for orchestration, Azure Data Lake Gen2 for storage, Azure Databricks for transformation, and Azure Synapse Analytics and Power BI for analysis and reporting. It supports incremental loads from an on-premises SQL Server database, ensuring scalability, automation, security, and real-time insights.By implementing an incremental load process, you ensure efficient data transfer while maintaining data integrity. The layered architecture (Bronze, Silver, Gold) ensures data quality and governance, while Azure Databricks and Azure Synapse Analytics provide robust transformation and analytics capabilities. Finally, Power BI enables interactive visualization for actionable insights, making this solution highly valuable for data-driven decision-making across industries.

1. Source: On-Premises SQL Server Database

- Purpose: The starting point of the data pipeline is an on-premises SQL Server database, which serves as the source of raw data.
- Data Characteristics: This database likely contains structured data that needs to be processed and analysed.
- Incremental Load: Instead of performing a full data transfer, the solution focuses on incremental loading, where only new or updated records are transferred to the cloud. This approach reduces bandwidth usage and processing time.

2. Orchestration: Azure Data Factory (ADF)

- Role: Azure Data Factory acts as the orchestrator for the entire data pipeline.
- Tasks:
    - Triggering Incremental Loads: ADF schedules or triggers the incremental load process based on predefined intervals (e.g., hourly, daily) or events.
    - Data Movement: It transfers data from the on-premises SQL Server to Azure Data Lake Gen2.
    - Pipeline Monitoring: ADF provides visibility into the status of the pipeline, including success or failure notifications.

3. Storage: Azure Data Lake Gen2

- Purpose: Azure Data Lake Gen2 serves as the primary storage layer for the raw and transformed data.
- Architecture Layers:
    - Bronze Layer: Raw data ingested from the SQL Server database is stored here without any transformations. This ensures that the original data is preserved for future use.
    - Silver Layer: Data is cleaned and standardized in this layer. For example, missing values may be handled, duplicates removed, and data formatted consistently.

- Gold Layer: Business-ready data is stored here after further transformations. This layer contains curated datasets that are optimized for analytics and reporting.

4. Transformation: Azure Databricks

- Role: Azure Databricks is used for data transformation and enrichment.
- Tasks:
  - Processing Bronze Layer Data: Raw data from the Bronze Layer is read and transformed into the Silver Layer.
  - Advanced Transformations: Complex operations like filtering, aggregating, joining datasets, or applying machine learning models can be performed using Apache Spark.
  - Scalability: Azure Databricks offers scalable compute resources to handle large datasets efficiently.

5. Analysis and Reporting: Azure Synapse Analytics and Power BI

- Azure Synapse Analytics:
  - Purpose: Serves as a unified analytics platform for querying and analyzing the transformed data.
  - Tasks:
    - Supports both batch and real-time analytics.
    - Enables SQL queries on the Gold Layer data.
    - Integrates with other Azure services for advanced analytics.
- Power BI:
  - Purpose: Provides interactive dashboards and reports for business users.
  - Tasks:
    - Visualizes insights derived from the Gold Layer data.
    - Allows users to explore data through interactive charts, graphs, and tables.
    - Supports self-service analytics for non-technical stakeholders.

6. Security and Governance

- Azure Active Directory (AAD):
  - Purpose: Manages user identities and access control.
  - Tasks:

- Provides single sign-on (SSO) for secure access to Azure resources.
- Enforces role-based access control (RBAC) to restrict access to sensitive data.

- Azure Key Vault:
  - Purpose: Securely stores encryption keys, secrets, and certificates.
  - Tasks:
    - Encrypts data at rest and in transit.
    - Manages authentication credentials for various services in the pipeline.

7. Incremental Load Process

The core functionality of your project is the incremental load , which ensures that only new or updated data is transferred from the SQL Server database to Azure Data Lake Gen2. This process involves:

1. Identifying Changes: Using techniques such as change data capture (CDC), timestamp columns, or log files to detect new or modified records.
2. Extracting Data: Querying the SQL Server database to retrieve only the changed data.
3. Loading Data: Transferring the extracted data to Azure Data Lake Gen2 using Azure Data Factory.
4. Updating Layers: Updating the Bronze, Silver, and Gold layers incrementally to reflect the latest changes.

8. Benefits of the Architecture

- Scalability: Handles large volumes of data with ease due to Azure's cloud infrastructure.
- Automation: Reduces manual intervention through automated ETL processes.
- Security: Ensures data privacy and compliance with tools like Azure Active Directory and Azure Key Vault.
- Real-Time Insights: Supports near-real-time data processing and analytics.
- Flexibility: Easily integrates with other Azure services for advanced analytics and reporting.

Cloud Services

1. SQL Server Management Studio (SSMS)

SSMS is a comprehensive tool for managing and querying on-premises SQL Server databases. It enables users to execute complex SQL queries, manage database schemas, and track changes using features like Change Data Capture (CDC). SSMS ensures the source database is optimized for incremental data extraction, making it a critical starting point for your ETL pipeline.

Overview:

- Purpose: SSMS is a tool used to manage and interact with SQL Server databases. It serves as the interface for querying, managing, and maintaining the source data stored in the on-premises SQL Server database.

Key Features:

- Database Management:
    - Execute SQL queries to extract, update, or delete data.
    - Manage tables, views, stored procedures, and indexes.
- Data Exploration:
    - Explore and analyze data using graphical query builders or custom SQL scripts.
- Change Data Capture (CDC):
    - Enables tracking of changes (inserts, updates, deletes) in the database, which is critical for implementing incremental loads.
- Backup and Recovery:

- Perform regular backups to prevent data loss.
- Restore databases from backups in case of failure.

Role in Project :

- SSMS is used to manage the on-premises SQL Server database, ensuring that it is properly configured for incremental data extraction.
- It provides the mechanism to identify new or updated records using techniques like CDC or timestamp-based filtering.

2. Azure Data Factory (ADF)

Azure Data Factory automates and orchestrates the movement and transformation of data across various environments. It supports incremental loads by identifying new or updated records using watermark columns or CDC mechanisms. ADF's scheduling and monitoring capabilities ensure seamless data flow from on-premises systems to Azure services.

Overview :

- Purpose : ADF is a cloud-based data integration service that orchestrates and automates data movement and transformation workflows.

Key Features :

- Pipeline Orchestration :
  - Automates the ETL process by scheduling and triggering pipelines.
  - Supports event-based triggers (e.g., when new data arrives) or time-based triggers (e.g., daily or hourly).
- Data Movement :
  - Transfers data between on-premises systems (via the Self-Hosted Integration Runtime) and Azure services.
  - Handles large-scale data transfers efficiently.
- Monitoring :
  - Provides a visual interface to monitor pipeline runs, track errors, and view performance metrics.
- Incremental Load Support :
  - Facilitates incremental data extraction by leveraging watermark columns (e.g., timestamps) or CDC mechanisms.

Role in Project :

- ADF extracts incremental data from the SQL Server database and loads it into Azure Data Lake Gen2.

- It orchestrates the entire ETL workflow, ensuring seamless data flow between services.

3. Azure Data Lake Gen2

Azure Data Lake Gen2 provides scalable, cost-efficient storage for structured and unstructured data. Its hierarchical file system organizes data into Bronze, Silver, and Gold layers, ensuring clarity and governance. With built-in security and integration with analytics tools, it serves as the backbone of your data pipeline.

Overview :

- Purpose : Azure Data Lake Gen2 is a scalable storage solution designed for big data analytics. It combines the capabilities of Azure Blob Storage with a hierarchical file system.

Key Features :

- Scalability :
    - Stores petabytes of structured, semi-structured, and unstructured data.
    - Automatically scales to accommodate growing data volumes.
- Layered Architecture:
    - Bronze Layer: Raw data ingested from the source (SQL Server).
    - Silver Layer: Cleaned and standardized data after initial transformations.
    - Gold Layer: Curated, business-ready data optimized for analytics.
- Cost Efficiency:
    - Offers tiered storage options (hot, cool, archive) to optimize costs based on data access patterns.
- Security:
    - Supports encryption at rest and in transit.
    - Integrates with Azure Active Directory for fine-grained access control.

Role in Project:

- Acts as the primary storage layer for raw, intermediate, and transformed data.
- Ensures data is securely stored and readily accessible for downstream processing.

4. Azure Databricks

Azure Databricks leverages Apache Spark to perform high-performance data transformations and advanced analytics. It processes raw data from the Bronze Layer into enriched datasets in the Silver Layer, supporting operations like filtering, aggregations, and joins. Its collaborative notebooks enable teams to build and refine workflows efficiently.

Overview :

- Purpose : Azure Databricks is a collaborative analytics platform powered by Apache Spark. It is used for high-performance data transformation and advanced analytics.

Key Features :

- Apache Spark :
    - Performs distributed data processing for large datasets.
    - Supports batch and real-time transformations.
- Notebooks :
    - Interactive notebooks (Python, Scala, SQL) allow users to write and execute code collaboratively.
- Machine Learning :
    - Integrates with MLflow for building, training, and deploying machine learning models.
- Scalability :
    - Automatically scales compute resources based on workload demands.
- Integration :
    - Seamlessly integrates with Azure Data Lake Gen2 for reading and writing data.

Role in Project :

- Transforms raw data from the Bronze Layer into cleaned and enriched data in the Silver Layer.
- Performs advanced transformations, such as aggregations, joins, and feature engineering, to prepare data for analysis.
- Can be used to build AI/ML models if needed.

5. Azure Synapse Analytics

Azure Synapse Analytics combines big data processing with enterprise data warehousing for unified analytics. It allows users to run SQL queries on transformed data and supports real-time analytics. By integrating with Power BI, it bridges the gap between data analysis and visualization for actionable insights.

Overview :

- Purpose : Azure Synapse Analytics is a unified analytics service that combines big data processing with enterprise data warehousing.

Key Features :

- SQL Analytics :

    - Allows users to run SQL queries on large datasets stored in Azure Data Lake.
    - Supports both serverless and provisioned SQL pools.
- Big Data Processing :
    - Integrates with Apache Spark for distributed data processing.
- Integration :
    - Connects seamlessly with Power BI for visualization.
    - Works with Azure Data Lake Gen2 for data storage.
- Scalability :
    - Scales automatically to handle massive workloads.
- Real-Time Analytics :
    - Supports streaming data ingestion and analysis.

Role in Project :

- Analyzes the Gold Layer data to generate insights.
- Supports ad-hoc queries and reporting for business stakeholders.
- Serves as a bridge between the data pipeline and visualization tools like Power BI.

6. Power BI

Power BI transforms processed data into interactive dashboards and reports for business stakeholders. Its self-service capabilities empower non-technical users to explore trends and patterns visually. With real-time connectivity to Azure services, Power BI ensures timely and accurate decision-making through intuitive visualizations.

Overview :

- Purpose : Power BI is a business analytics tool used to create interactive dashboards and reports.

Key Features :

- Visualization :
    - Build charts, graphs, and tables to visualize data trends and patterns.
    - Supports drill-down and cross-filtering for deeper insights.
- Self-Service Analytics :
    - Empowers non-technical users to explore data without requiring extensive coding skills.

- Integration :
    - Connects directly to Azure Synapse Analytics or Azure Data Lake Gen2 for real-time data access.
- Collaboration :
    - Share dashboards and reports with teams via the Power BI Service.
- Customization :
    - Create custom visuals or use pre-built templates.

Role in Project :

- Visualizes insights derived from the Gold Layer data.
- Enables business users to make data-driven decisions through interactive dashboards.

7. Azure Active Directory (AAD)

Azure Active Directory provides secure identity management and access control for Azure resources. It enforces role-based permissions and supports multi-factor authentication to protect sensitive data. AAD ensures compliance with industry standards while enabling seamless single sign-on across services.

Overview:

- Purpose: AAD is Microsoft's identity and access management service, providing secure access to Azure resources.

Key Features:

- Single Sign-On (SSO):
    - Allows users to log in once and access multiple Azure services.
- Role-Based Access Control (RBAC)
    - Assigns permissions based on user roles (e.g., admin, contributor, reader).
- Multi-Factor Authentication (MFA):
    - Adds an extra layer of security by requiring additional verification steps.
- Audit Logs:
    - Tracks user activities and access attempts for compliance purposes.

Role in Your Project:

- Ensures secure access to Azure services like ADF, Data Lake, and Synapse Analytics.
- Prevents unauthorized access to sensitive data.

8. Azure Key Vault

Azure Key Vault securely stores encryption keys, secrets, and certificates to safeguard sensitive information. It integrates with other Azure services to manage credentials and API keys without exposing them. By centralizing secret management, Key Vault enhances the security and compliance of your data pipeline.

8. Azure Key Vault

Overview:

- Purpose: Azure Key Vault is a service for securely storing and managing secrets, encryption keys, and certificates.

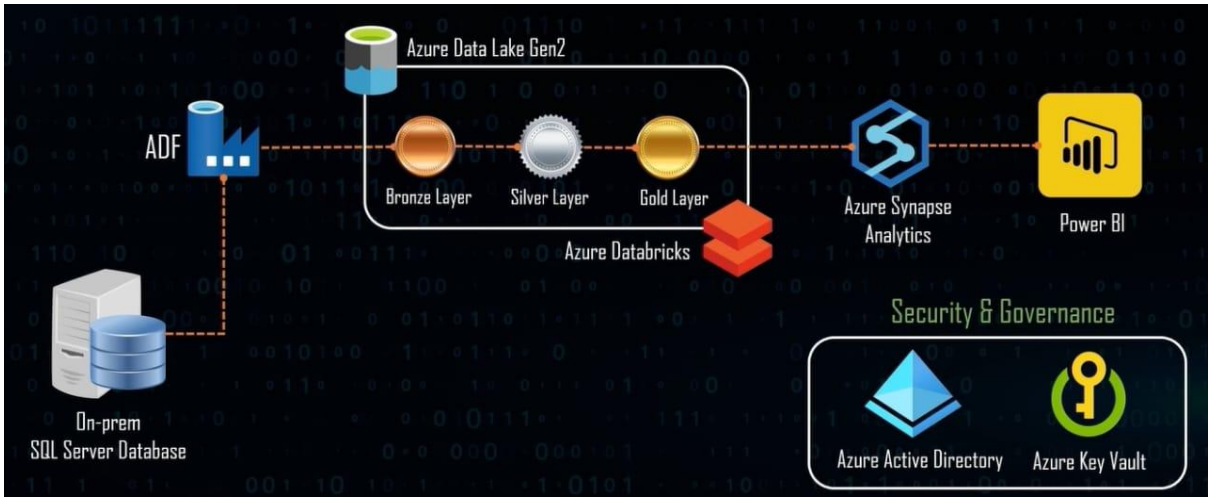
Key Features:

- Secret Management :
    - Stores credentials, API keys, and connection strings securely.
- Encryption Keys :
    - Manages encryption keys for data at rest and in transit.
- Access Control :
    - Restricts access to secrets and keys based on user roles.
- Integration :
    - Works seamlessly with other Azure services like ADF and Data Lake.

Role in Project :

- Protects sensitive information such as database credentials and API keys.
- Ensures compliance with security standards like GDPR or HIPAA.

ARCHITECTURE



1. Data Source: On-Premises SQL Server Database

- The journey begins with an on-premises SQL Server database , which serves as the source of raw data. This database contains structured data that is critical for business operations and analytics.
- To ensure efficiency, the system leverages incremental loading techniques such as Change Data Capture (CDC) or timestamp-based filtering to identify only new or updated records. This minimizes the volume of data transferred and reduces processing overhead.

2. Data Ingestion via Azure Data Factory (ADF)

- Azure Data Factory (ADF) acts as the orchestrator of the entire data pipeline. It automates the extraction of incremental data from the SQL Server database and ensures secure transfer to Azure Data Lake Gen2 .
- ADF supports both scheduled and event-driven triggers, enabling real-time or near-real-time data ingestion. Additionally, it uses the Self-Hosted Integration Runtime to facilitate seamless connectivity between on-premises systems and the cloud.
- During this phase, ADF ensures data integrity by logging pipeline runs, monitoring errors, and providing visibility into the status of data movement.

3. Storage and Layered Architecture in Azure Data Lake Gen2

- Once the data reaches Azure Data Lake Gen2 , it is stored in a layered architecture to maintain clarity, governance, and scalability:
    - Bronze Layer : This layer stores raw, unprocessed data ingested directly from the SQL Server database. It serves as a "single source of truth" for all incoming data.
    - Silver Layer : Raw data from the Bronze Layer is cleaned, standardized, and enriched using tools like Azure Databricks . Operations such as removing

duplicates, handling missing values, and formatting inconsistencies are performed here.

- Gold Layer : The Silver Layer data is further transformed into curated, business-ready datasets optimized for analytics and reporting. This layer contains high-quality data tailored for specific use cases.

- Azure Data Lake Gen2's hierarchical file system and scalability ensure efficient storage and retrieval of large datasets.

4. Transformation with Azure Databricks

- Azure Databricks , powered by Apache Spark , performs advanced data transformations on the Bronze and Silver Layer data. It processes large volumes of data in parallel, enabling high-performance operations such as filtering, aggregations, joins, and feature engineering.
- Azure Databricks' collaborative notebooks allow data engineers and analysts to write code in languages like Python, Scala, or SQL, making it flexible for various transformation needs.
- The transformed data is then written back to Azure Data Lake Gen2, specifically into the Gold Layer, ready for downstream analytics.

5. Analysis with Azure Synapse Analytics

- The Gold Layer data is analyzed using Azure Synapse Analytics , a unified analytics service that combines big data processing with enterprise data warehousing.
- Synapse Analytics supports both batch and real-time analytics , enabling users to run complex SQL queries or perform exploratory data analysis on large datasets.
- It integrates seamlessly with Power BI , allowing for smooth transitions between data analysis and visualization.

6. Visualization with Power BI

- Insights derived from Azure Synapse Analytics are visualized using Power BI , a powerful business intelligence tool.
- Power BI creates interactive dashboards and reports, enabling business users to explore trends, patterns, and key metrics through charts, graphs, and tables.
- Its self-service capabilities empower non-technical stakeholders to make data-driven decisions without requiring extensive technical expertise.

7. Security and Governance

- Throughout the pipeline, security and governance are ensured using two key services:
    - Azure Active Directory (AAD) : Manages user identities and enforces role-based access control (RBAC). It ensures that only authorized users can access sensitive data and resources.

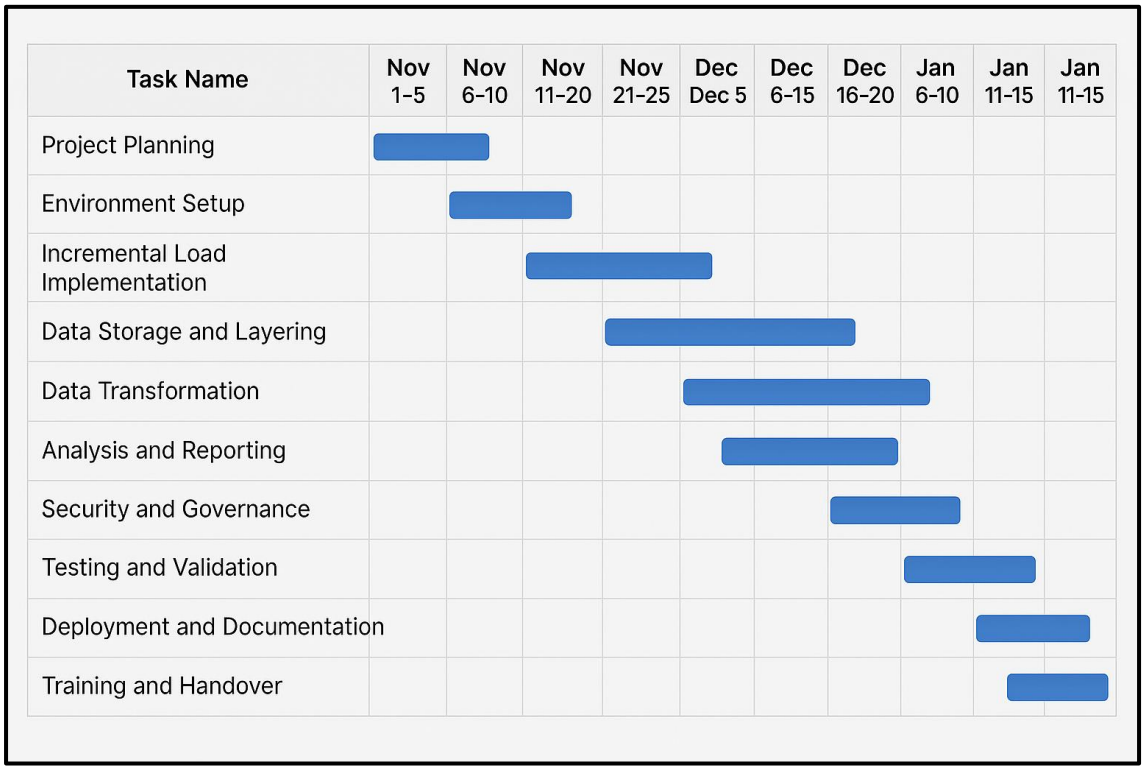
- Azure Key Vault : Securely stores encryption keys, secrets, and certificates, protecting sensitive information such as database credentials and API keys. This ensures compliance with industry standards like GDPR, HIPAA, or SOC 2.

8. End-to-End Workflow Summary

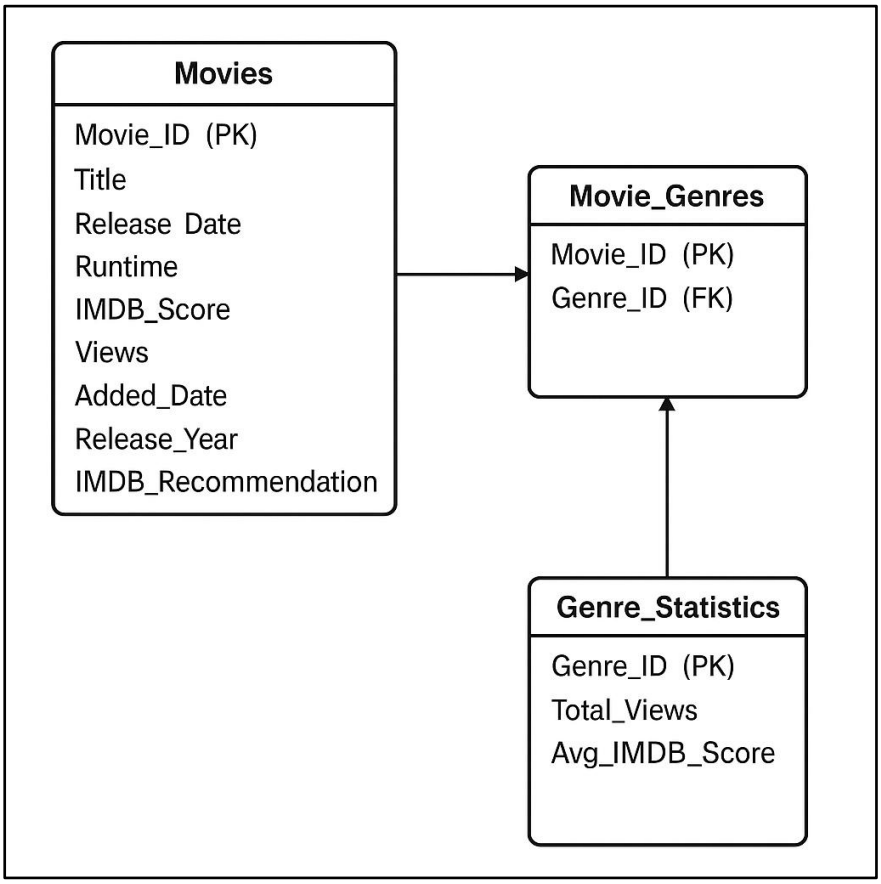
- The entire workflow starts with data extraction from the on-premises SQL Server database using Azure Data Factory.
- Incrementally loaded data is stored in Azure Data Lake Gen2, where it progresses through the Bronze, Silver, and Gold layers.
- Azure Databricks transforms the data, while Azure Synapse Analytics analyzes it for deeper insights.
- Finally, Power BI visualizes the results, enabling stakeholders to make informed decisions.
- Security and governance are maintained throughout the pipeline using Azure Active Directory and Azure Key Vault.

SYSTEM DESIGN

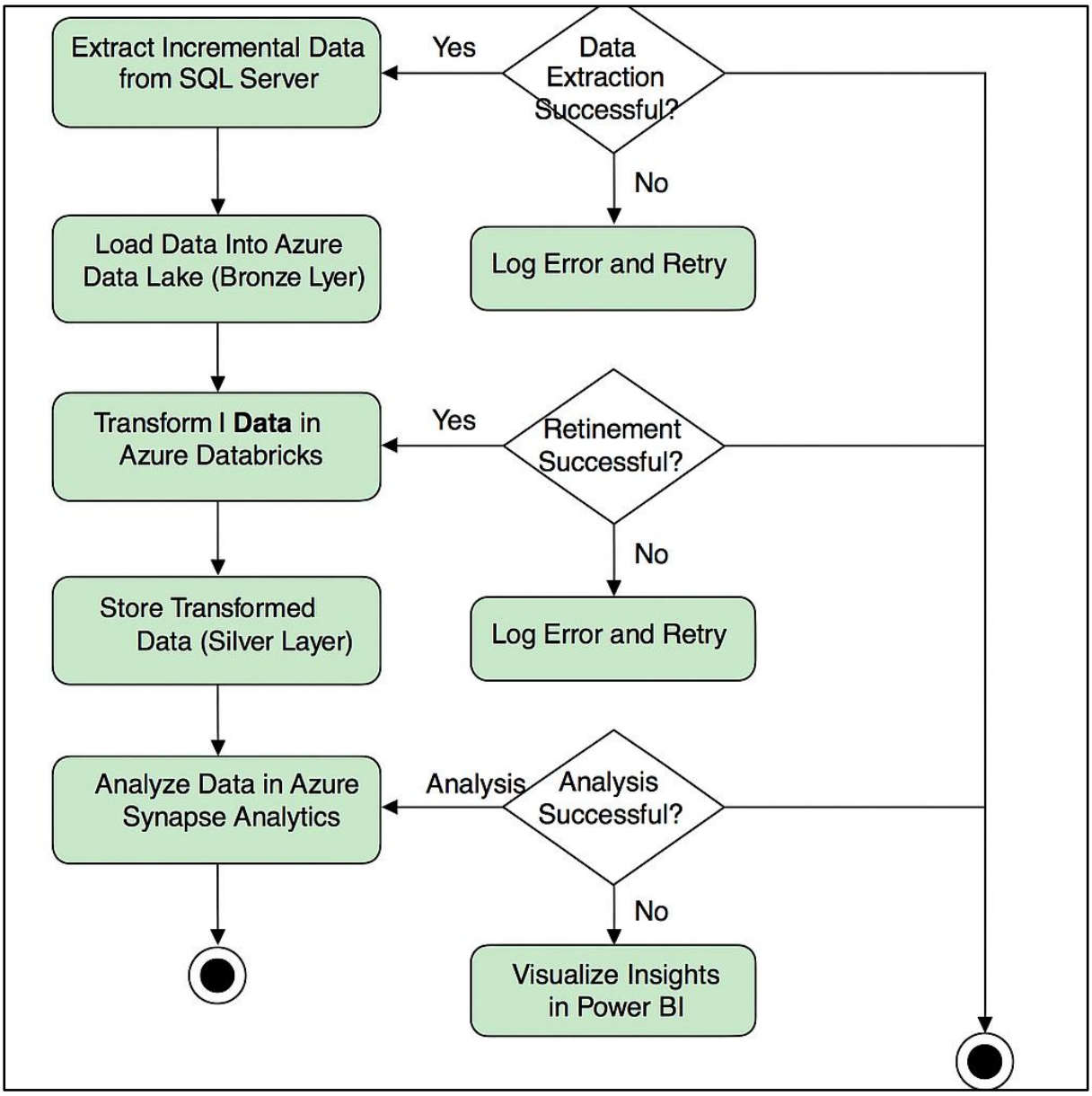
Gannt Chart



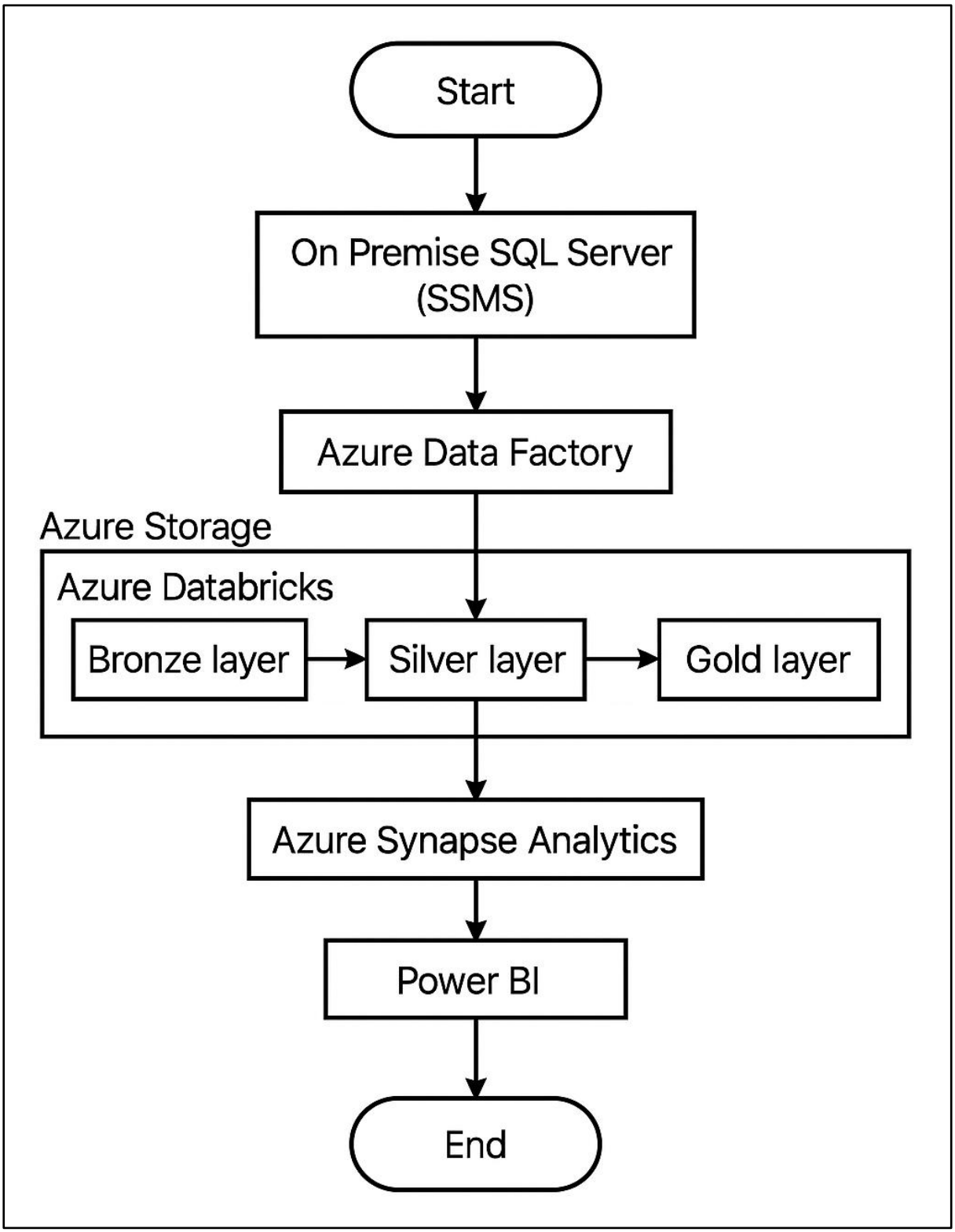
ER Diagram



UML Diagram


Extract Incremental Data from SQL Server
Yes
Data Extraction Successful?
No
Load Data Into Azure Data Lake (Bronze Lyer)
Log Error and Retry
Transform I Data in Azure Databricks
Yes
Retinement Successful?
No
Store Transformed Data (Silver Layer)
Log Error and Retry
Analyze Data in Azure Synapse Analytics
Analysis
Analysis Successful?
No
Visualize Insights in Power BI

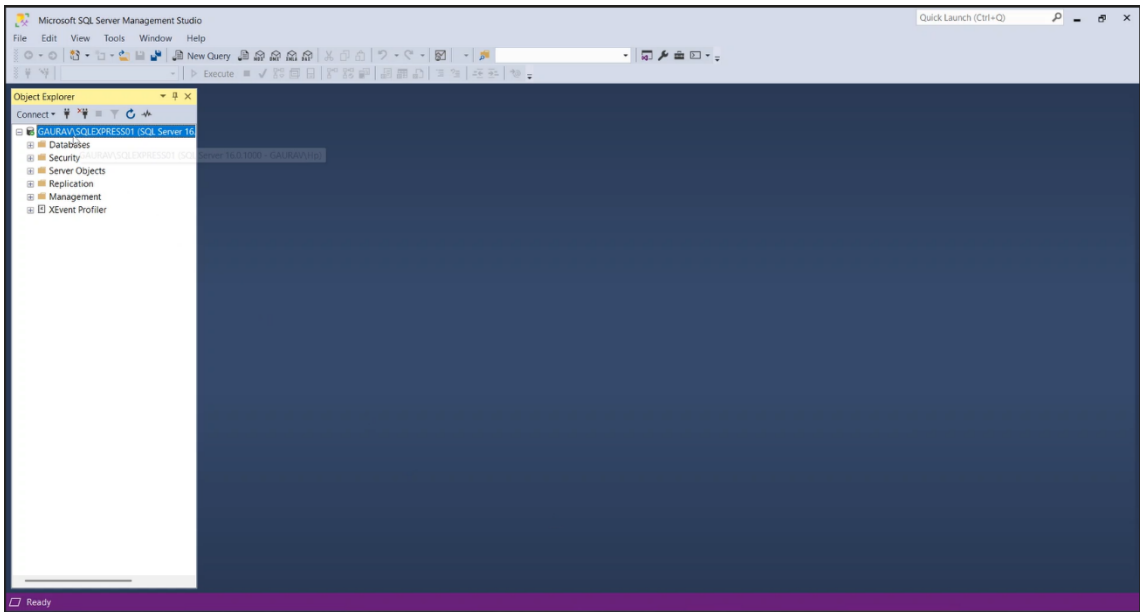
Activity Diagram



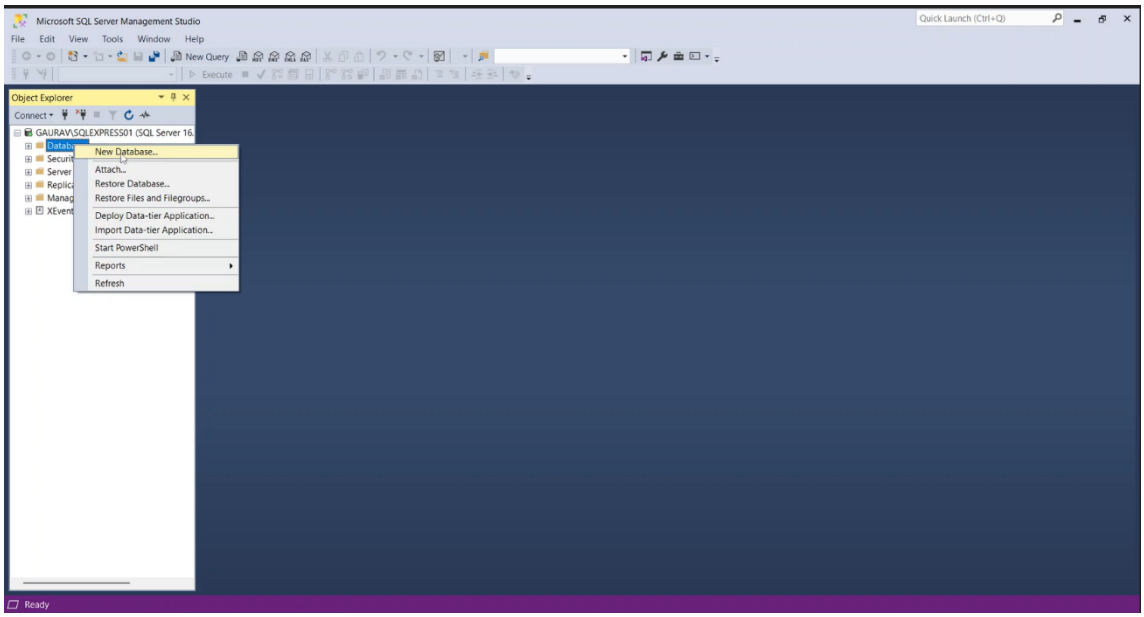
IMPLEMENTATION

Setting Up SQL server Management Server

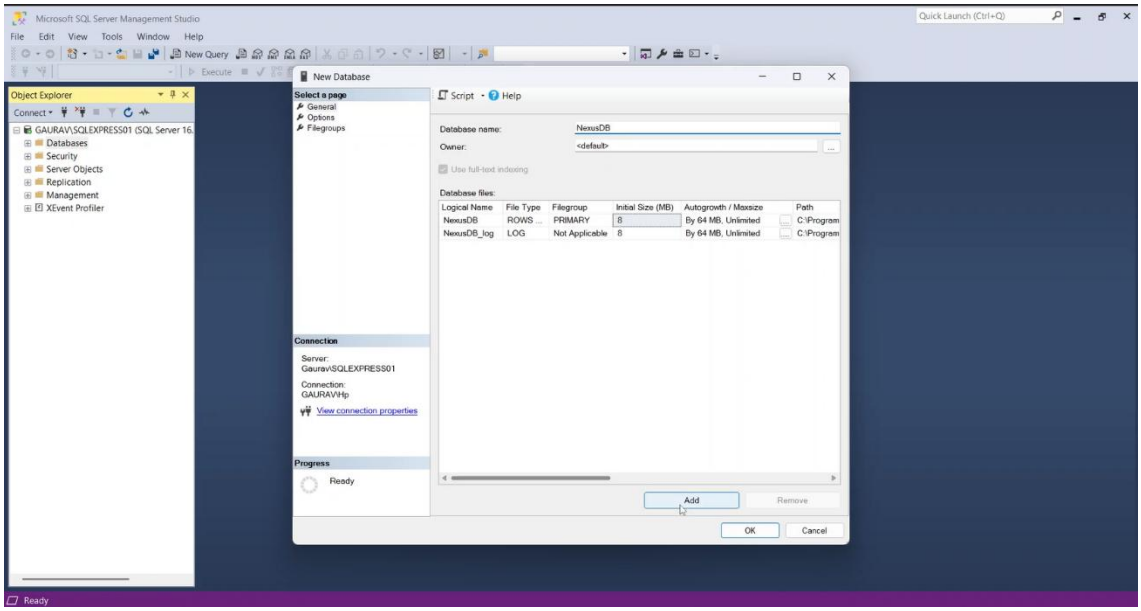
Microsoft SQL Server Management Studio (SSMS) connected to the local SQL Server instance "GAURAV\SQLEXPRESS01". The Object Explorer displays key components like Databases, Security, Server Objects, and Management.



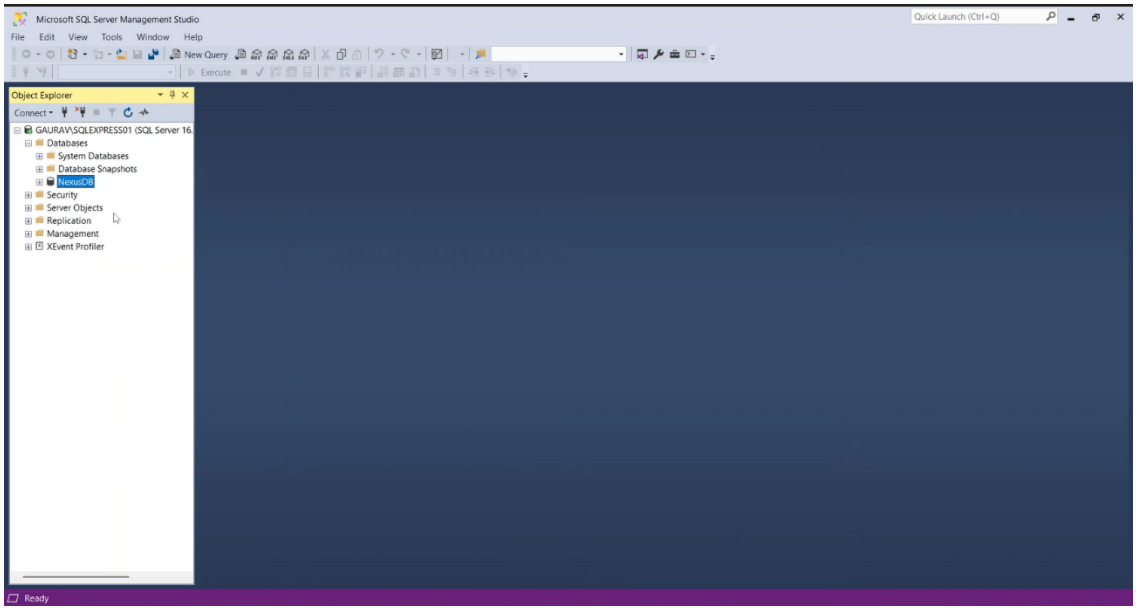
In SQL Server Management Studio, the context menu shows the option to create a new database on the connected server "GUARAV\SQLEXPRESS01".



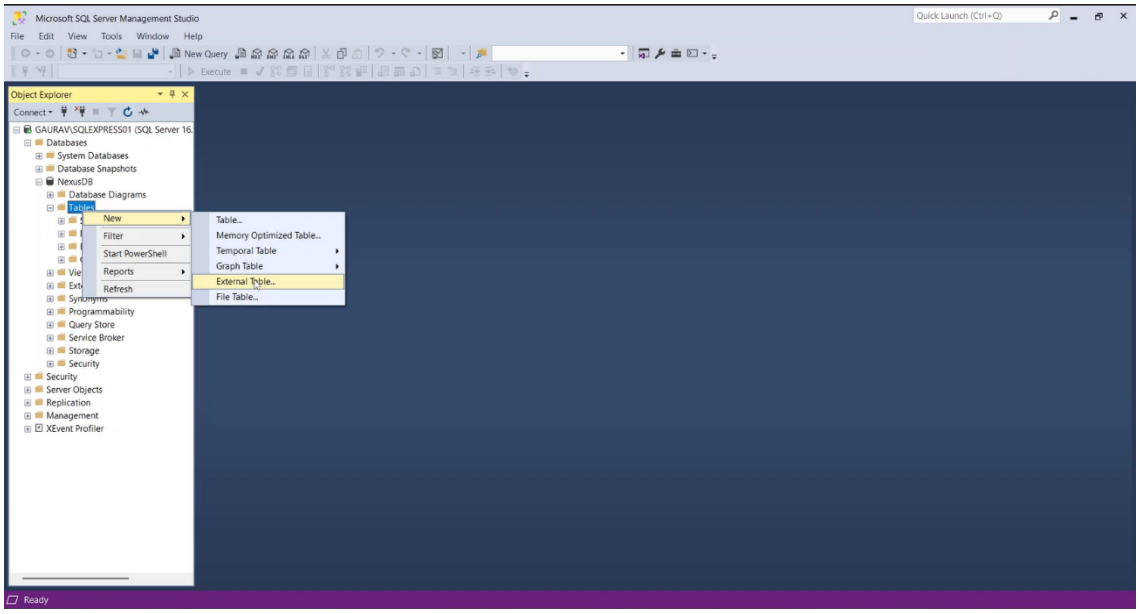
A new database named "NexusDB" is being created in SQL Server Management Studio, with default data and log file settings configured.



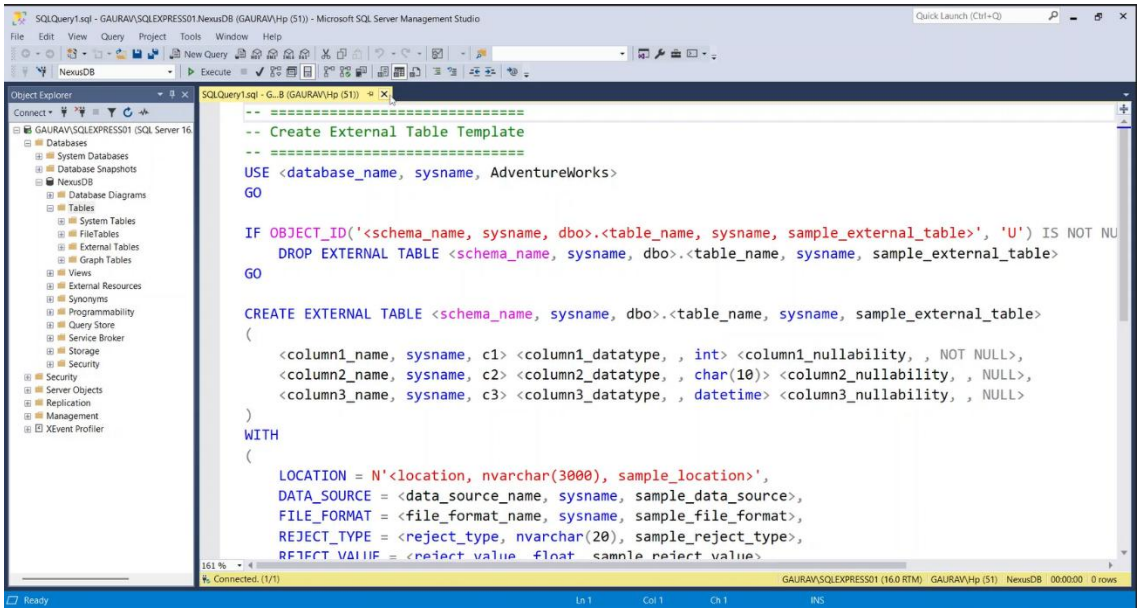
The newly created database "NexusDB" is now visible under the Databases section in SQL Server Management Studio.



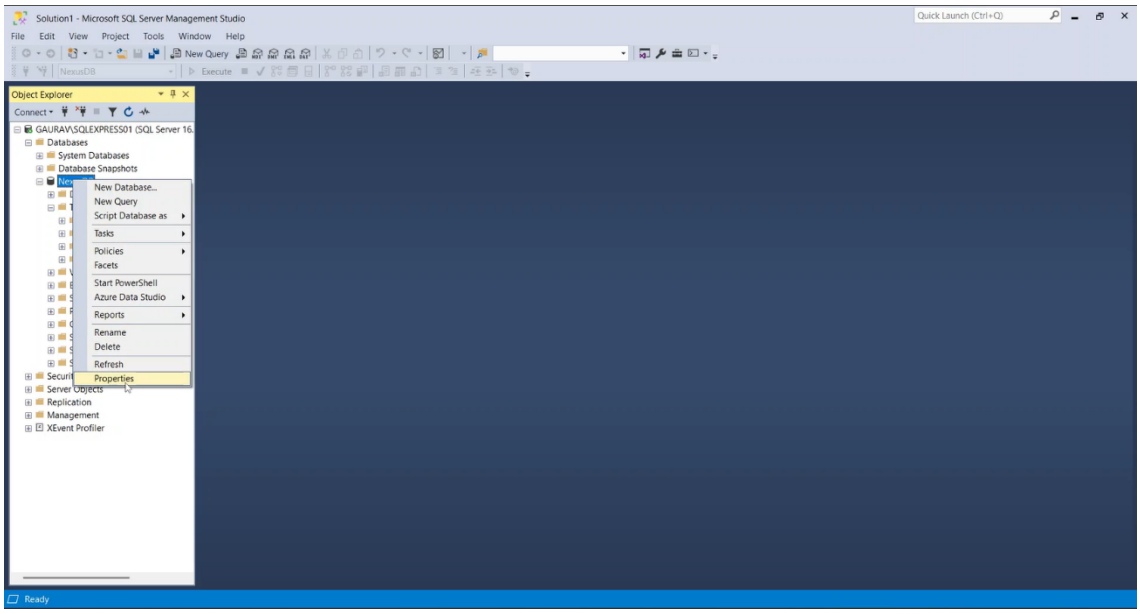
A right-click on the "Tables" folder in the NexusDB database shows the context menu to create a new external table.



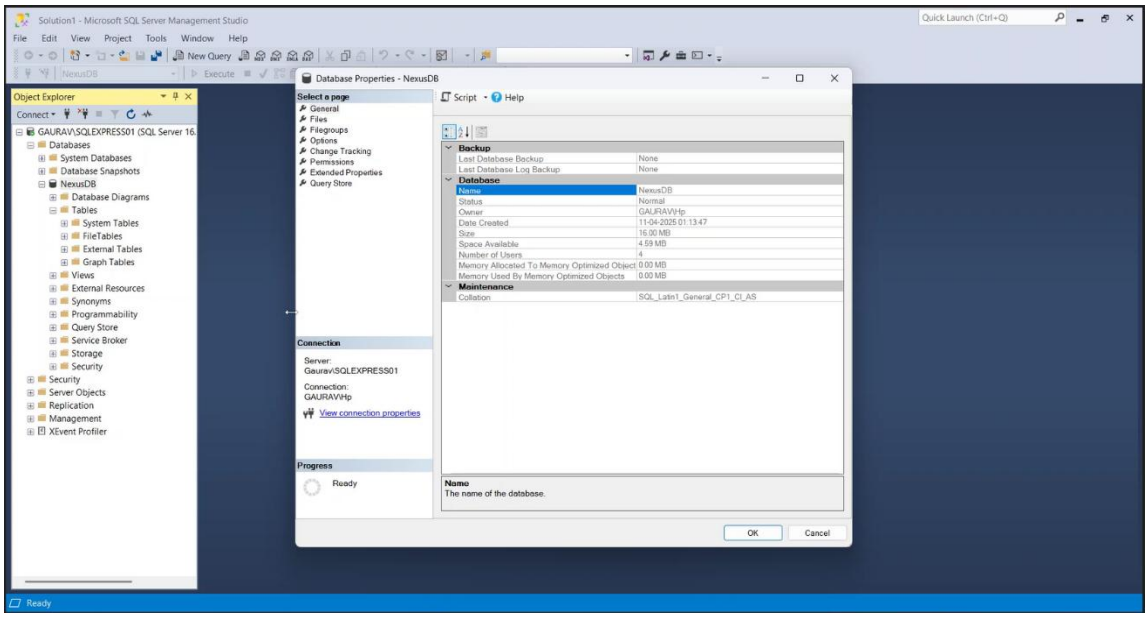
The SQL script shown is a template for creating an external table in SQL Server, including optional steps to drop the table if it exists.



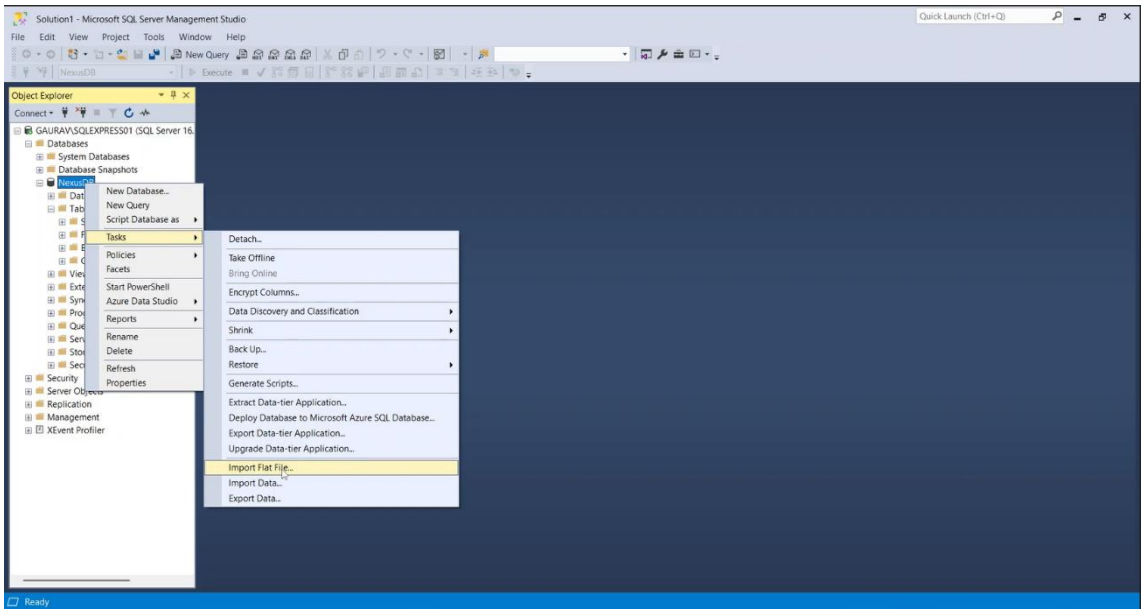
The user is right-clicking on the server node in SQL Server Management Studio to access the Properties option from the context menu.



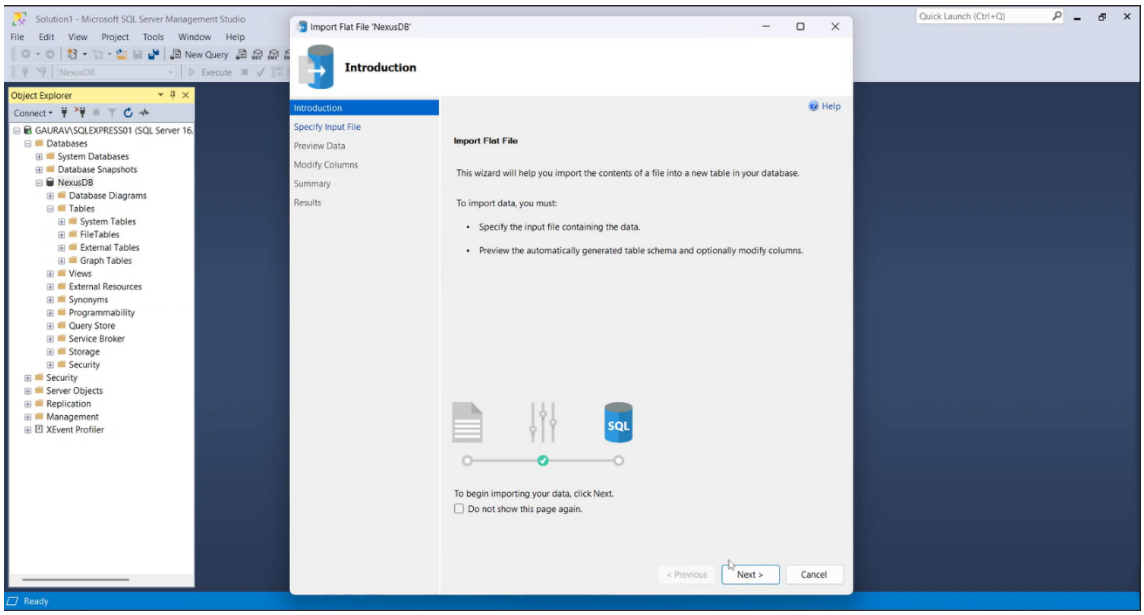
The Database Properties window for *NexusDB* is open in SQL Server Management Studio, showing detailed information like size, owner, creation date, and collation settings.



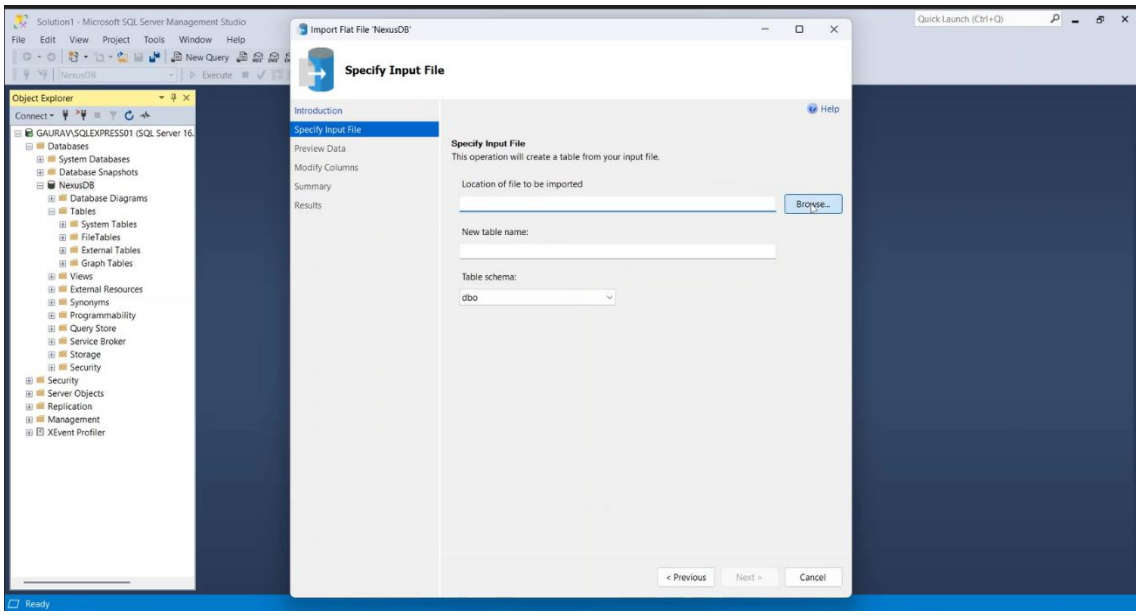
The Tasks menu for the *NexusDB* database is open in SQL Server Management Studio, with the Import Flat File option highlighted to import data from a flat file.



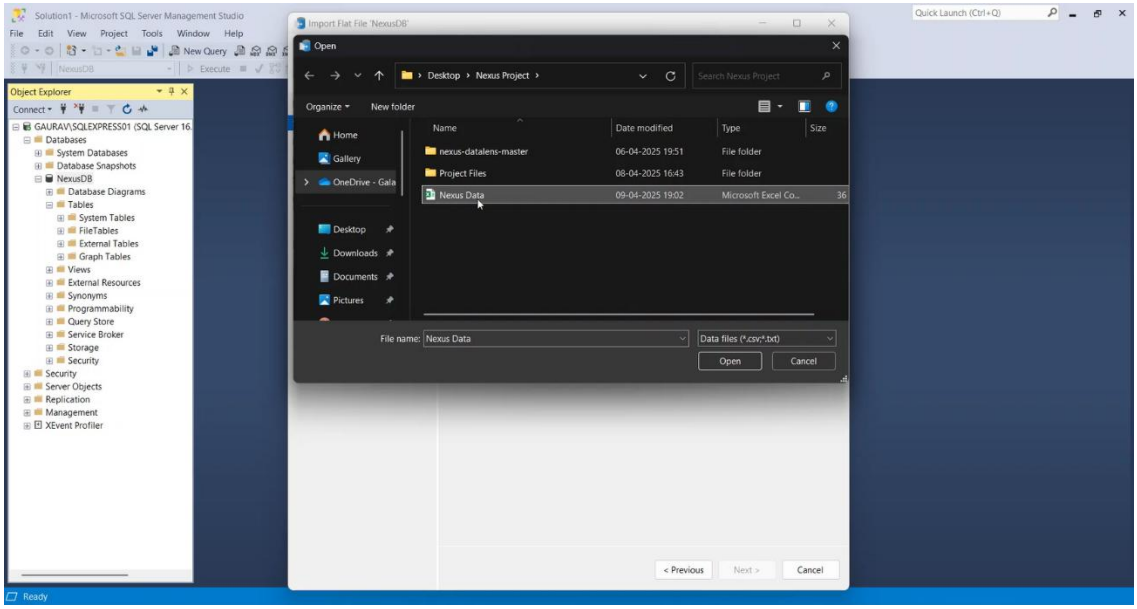
The Import Flat File wizard has opened in SQL Server Management Studio to guide the user through importing a file into the NexusDB database.



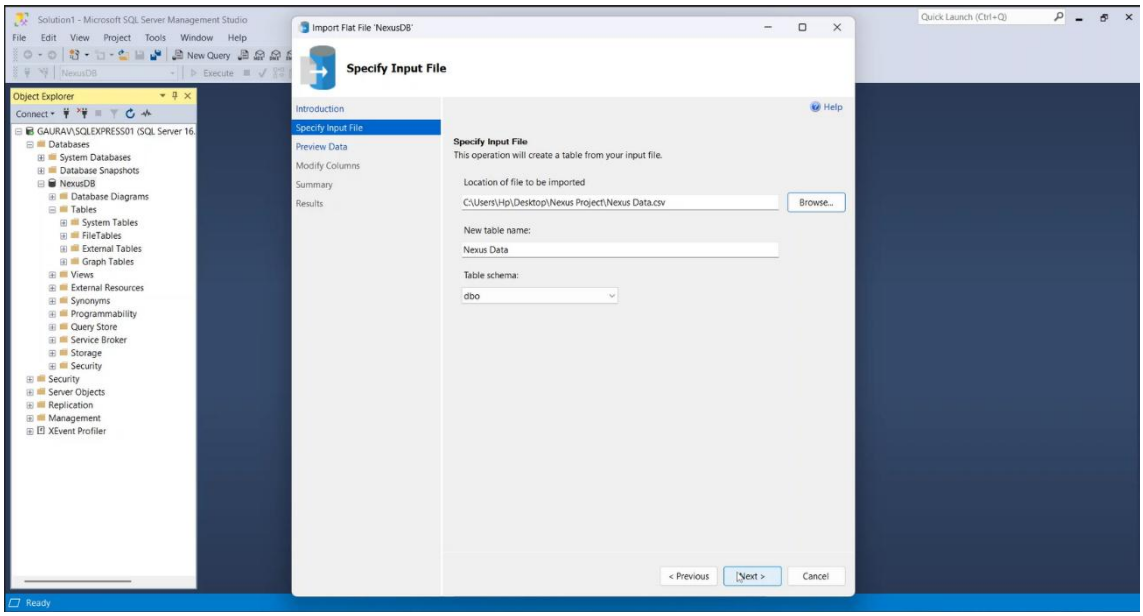
The user to browse and select a flat file to import, specify a new table name, and choose the table schema for the NexusDB database.



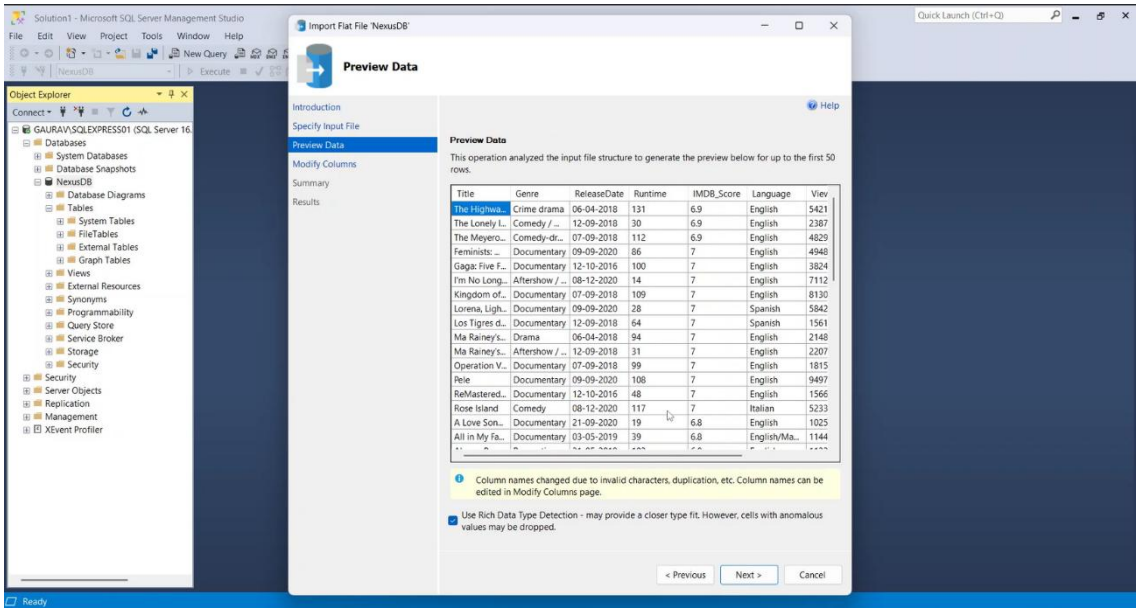
This window shows the user selecting an Excel file named "Nexus Data" to import into the SQL Server database.



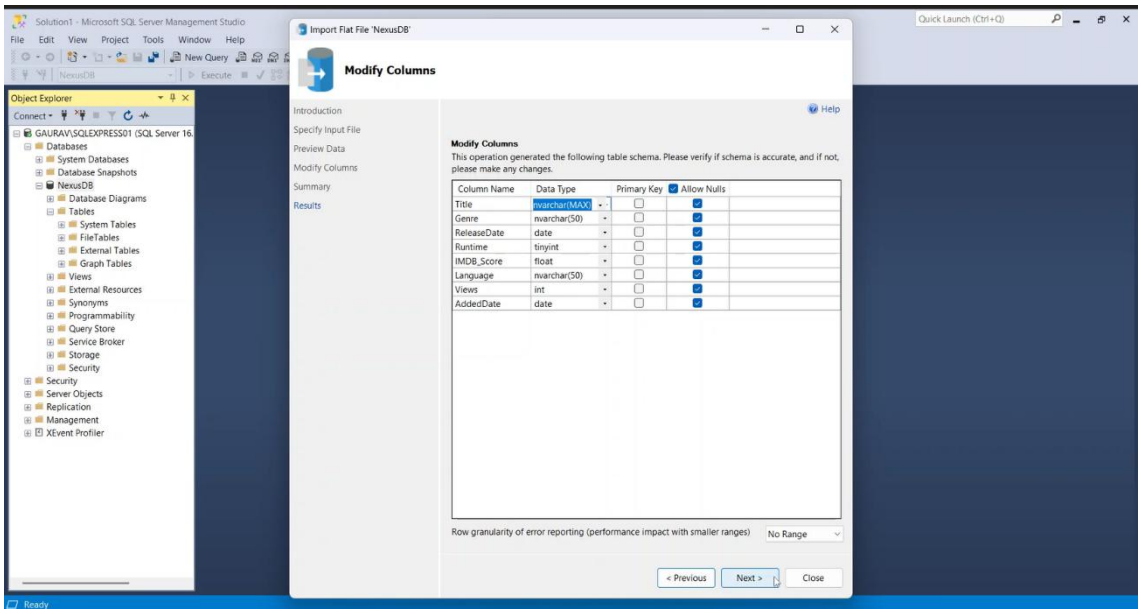
The user has selected the CSV file "Nexus Data.csv" for import and is ready to proceed to the next step to create a new table named "Nexus Data" in the dbo schema.



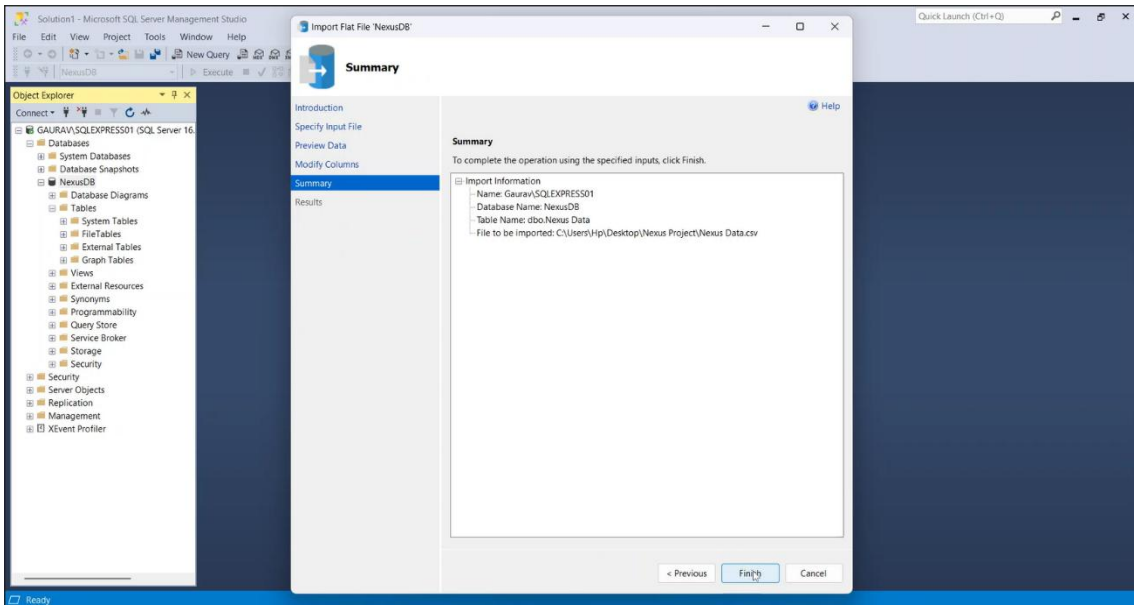
The import wizard displays a preview of the CSV data, showing column names and sample rows while notifying that some column names were adjusted for compatibility.



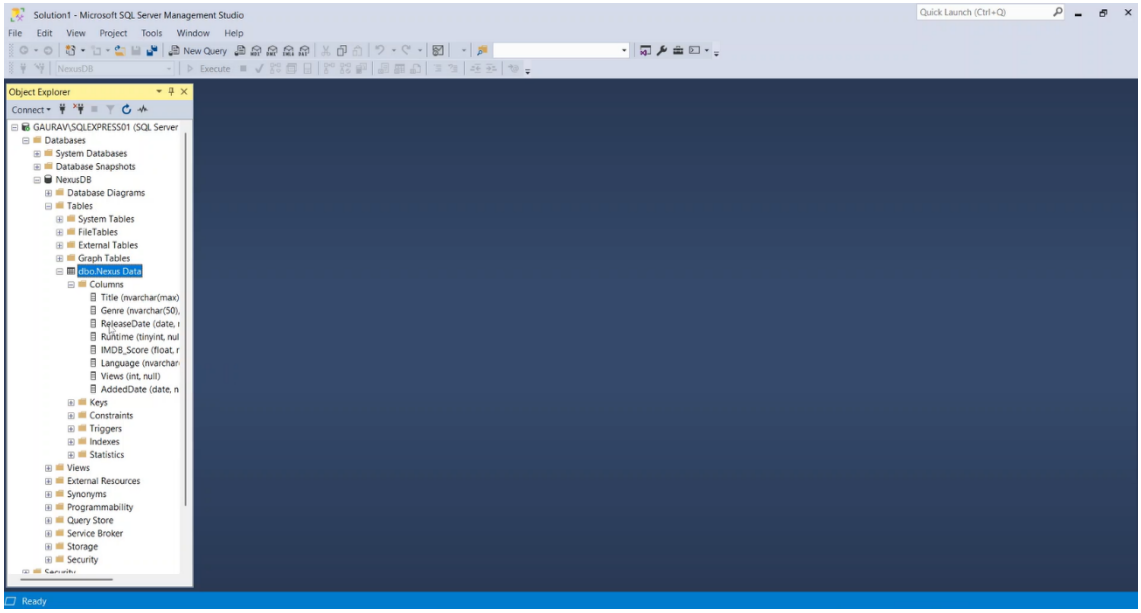
This step lets you review and modify column names, data types, nullability, and primary key settings before importing the data into the SQL table.



This summary screen confirms all import settings before finalizing the process to load the CSV file into the specified SQL Server table.



The table Nexus Data with its columns has been successfully created and is now listed under Tables in the NexusDB database.



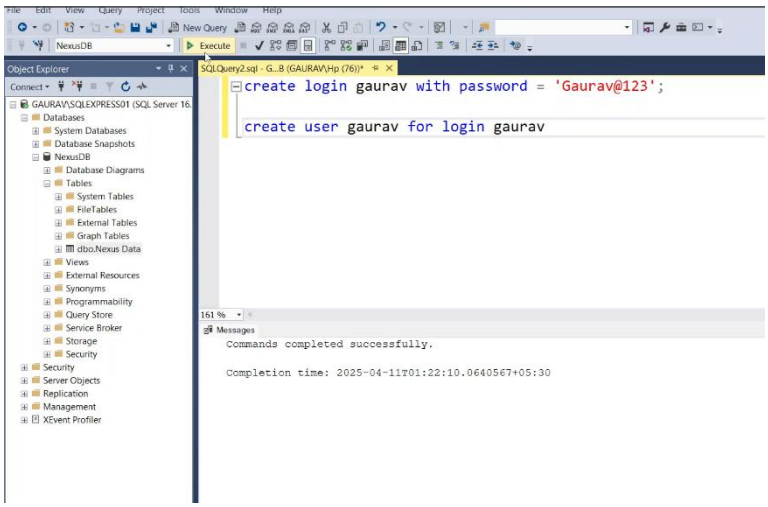
Create User in SSMS

Open SQL Server Management Studio (SSMS) connected to a server named GAURAV\SQLEXPRESS01.

```
create login gaurav with password = 'Gaurav@123';
```

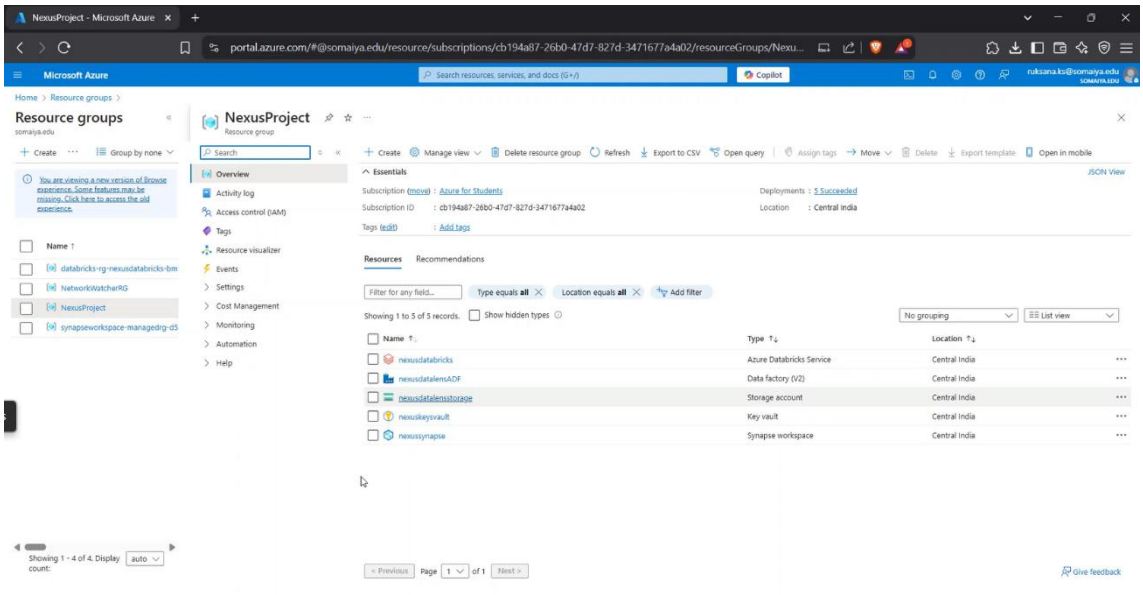
Create a Database User for the Login

```
create user Gaurav for login gaurav
```

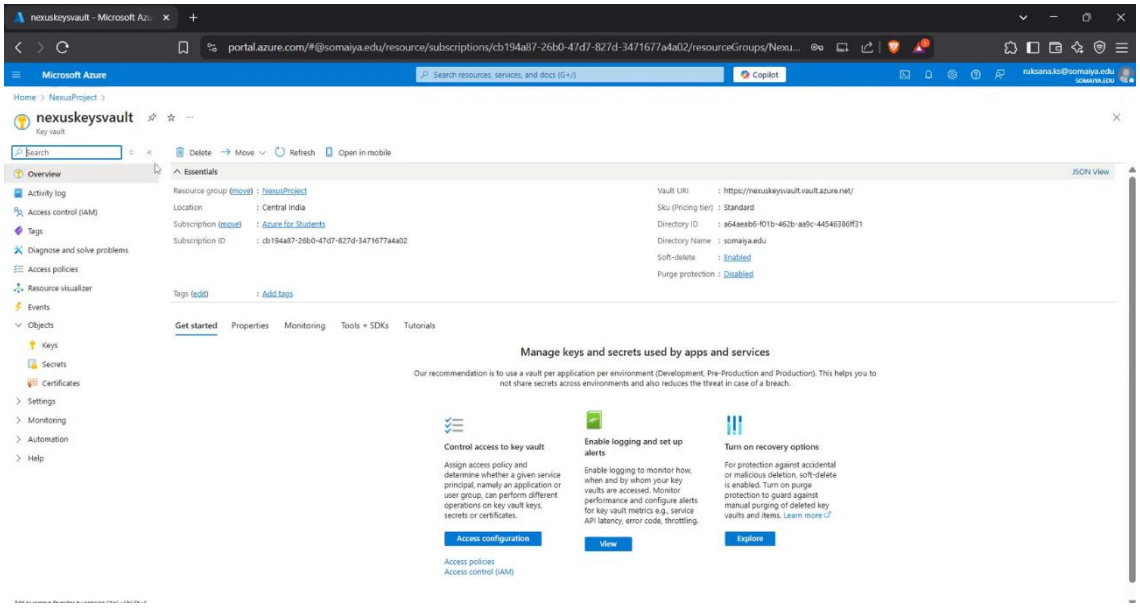


SETUP KEYVAULT

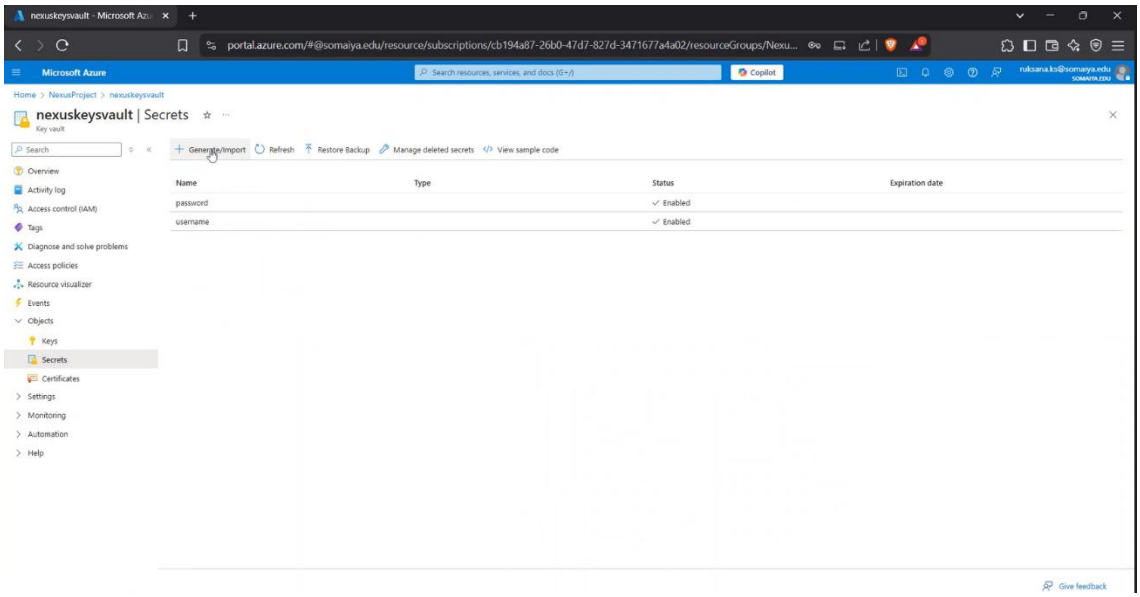
“NexusProject" resource group in Microsoft Azure, listing its associated resources like Databricks, Data Factory, Storage Account, Key Vault, and Synapse Workspace. All resources are deployed in the Central India region under the Azure for Students subscription.



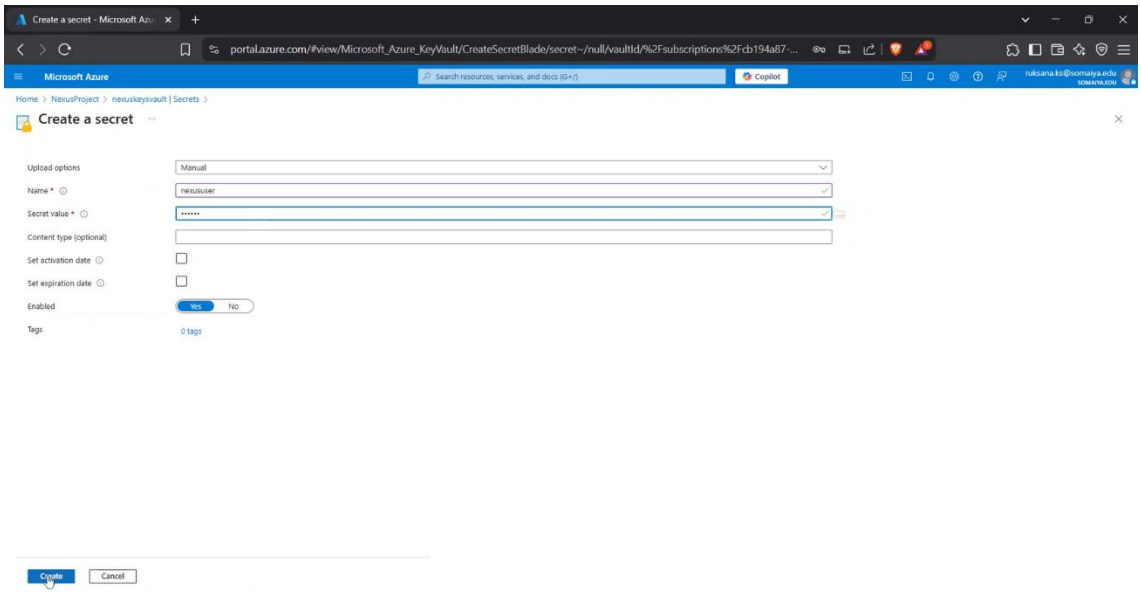
The image shows the Azure Key Vault "**nexuskeyvault**" overview page under the "**NexusProject**" resource group, displaying its configuration and security management options.



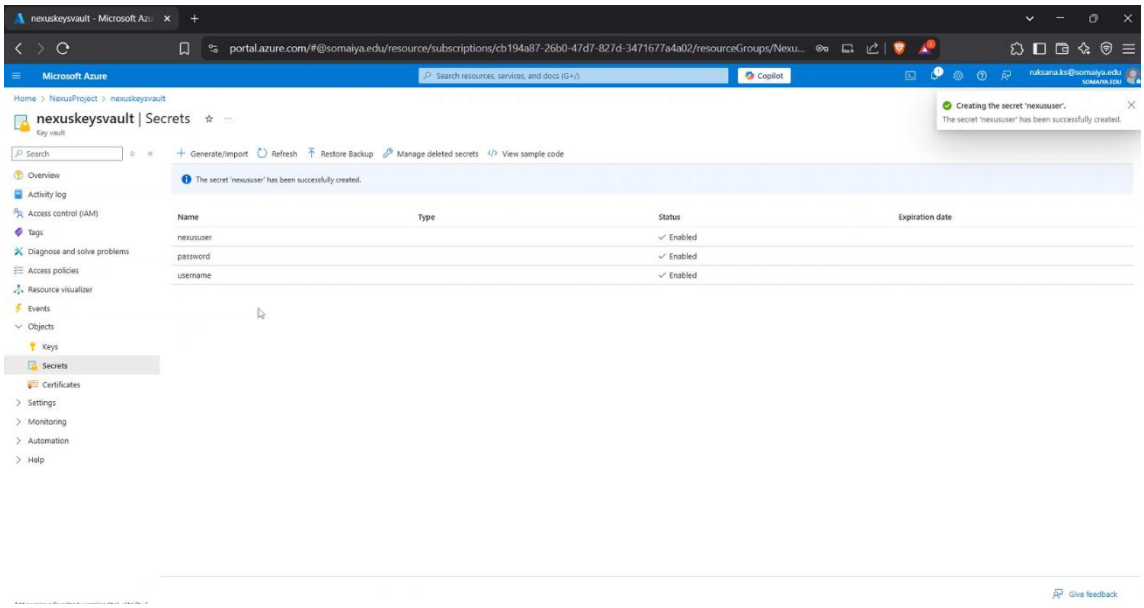
The image shows the "**Secrets**" section of the Azure Key Vault "**nexuskeyvault**," listing two enabled secrets named "**password**" and "**username**."



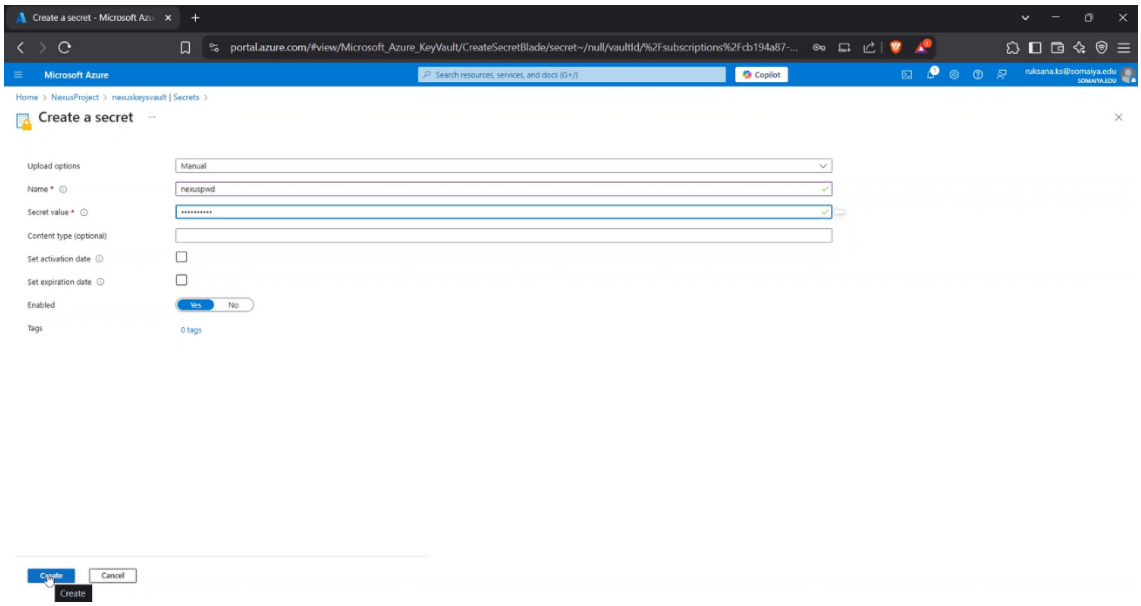
The image shows the creation of a new secret named "**newuser**" in the Azure Key Vault, with manual upload and enabled status.



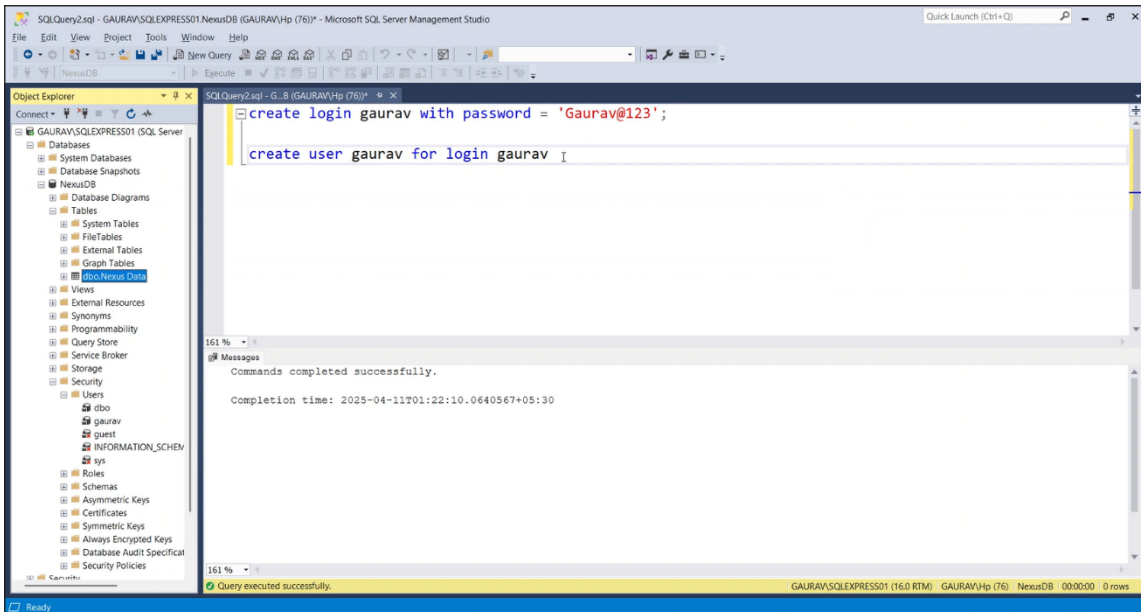
The secret "**newuser**" has been successfully created and added to the list of enabled secrets in the Azure Key Vault.



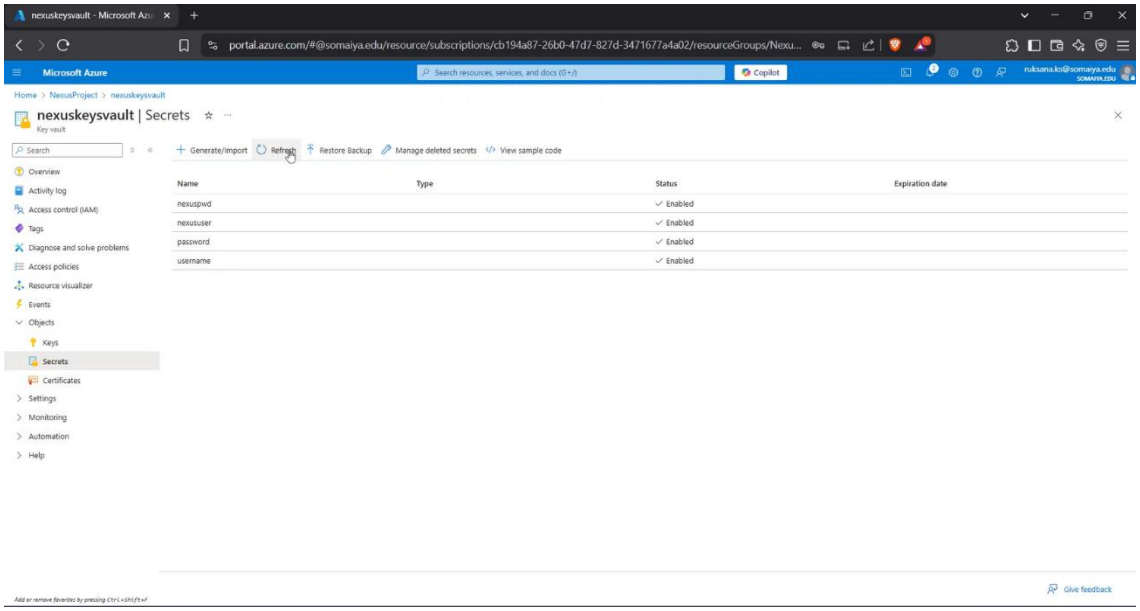
The creation of a new secret named "**newuserpwd**" with a specified value in Azure Key Vault, ready to be saved by clicking "**Create".**



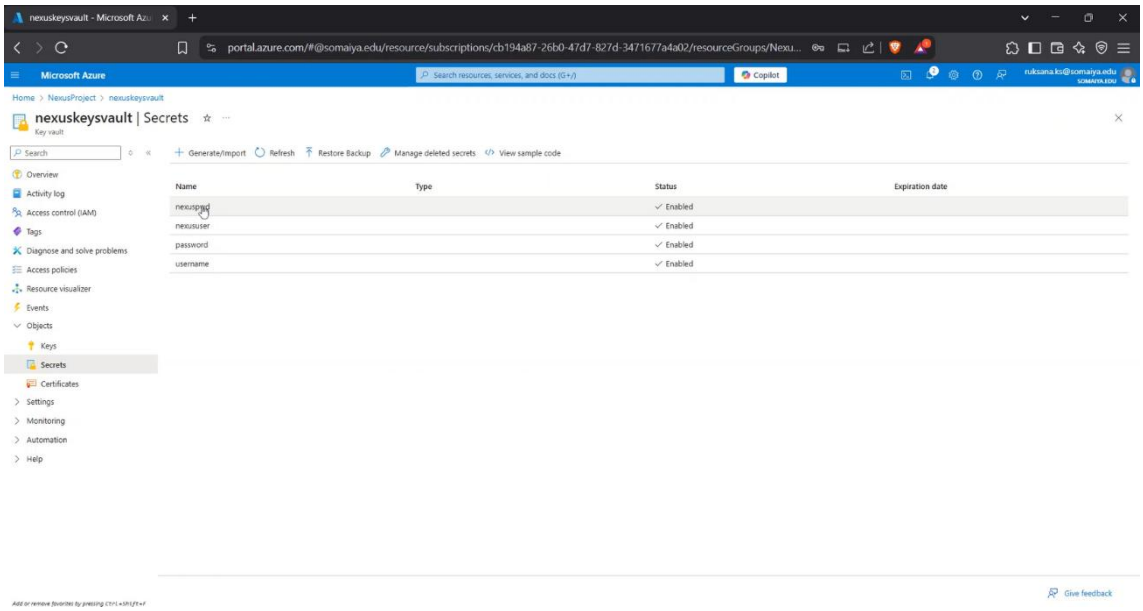
The SQL script creates a login named "**gaurav**" with a password and then creates a corresponding user for that login in the **NexusDB** database.



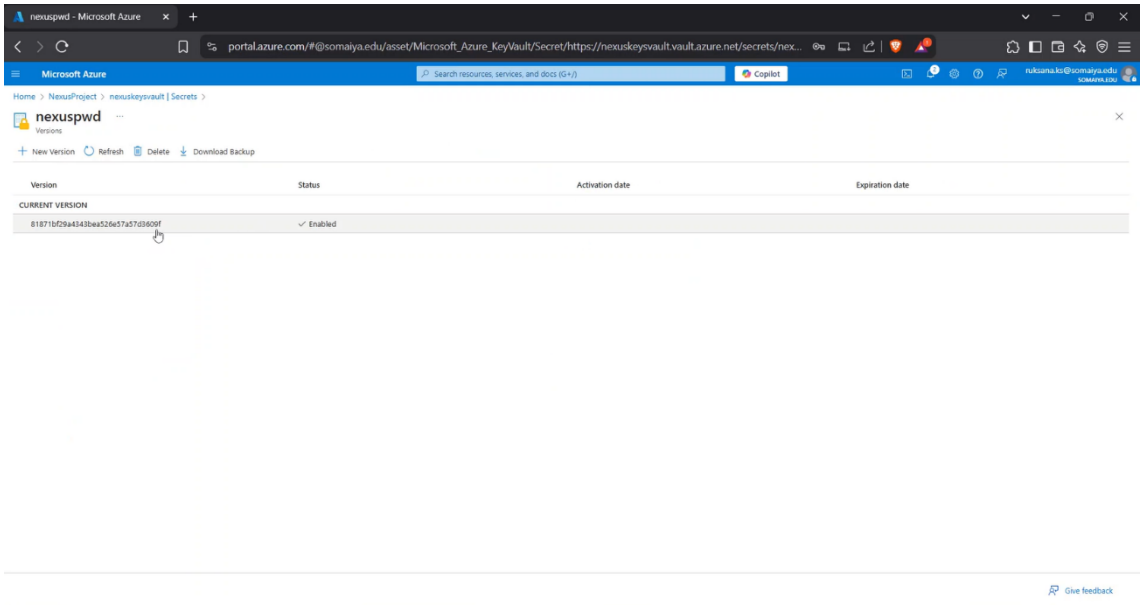
The Azure Key Vault now contains four enabled secrets**: nexuspwd, nexususer, password,** and **username.**



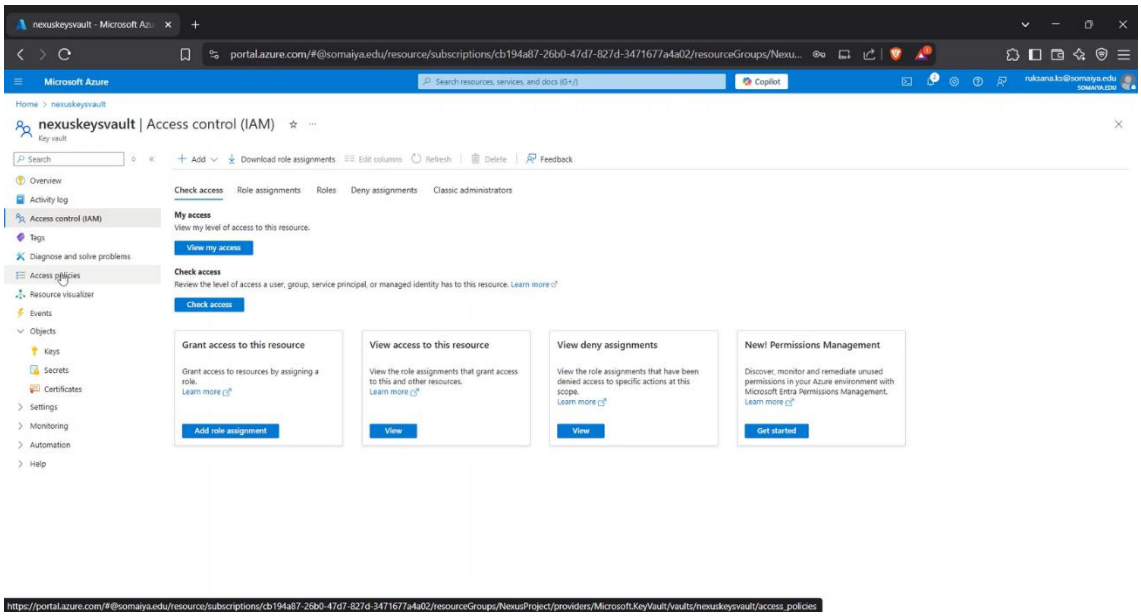
The secret named **nexuspwd** in the Azure Key Vault is selected, with all four listed secrets currently **enabled**.



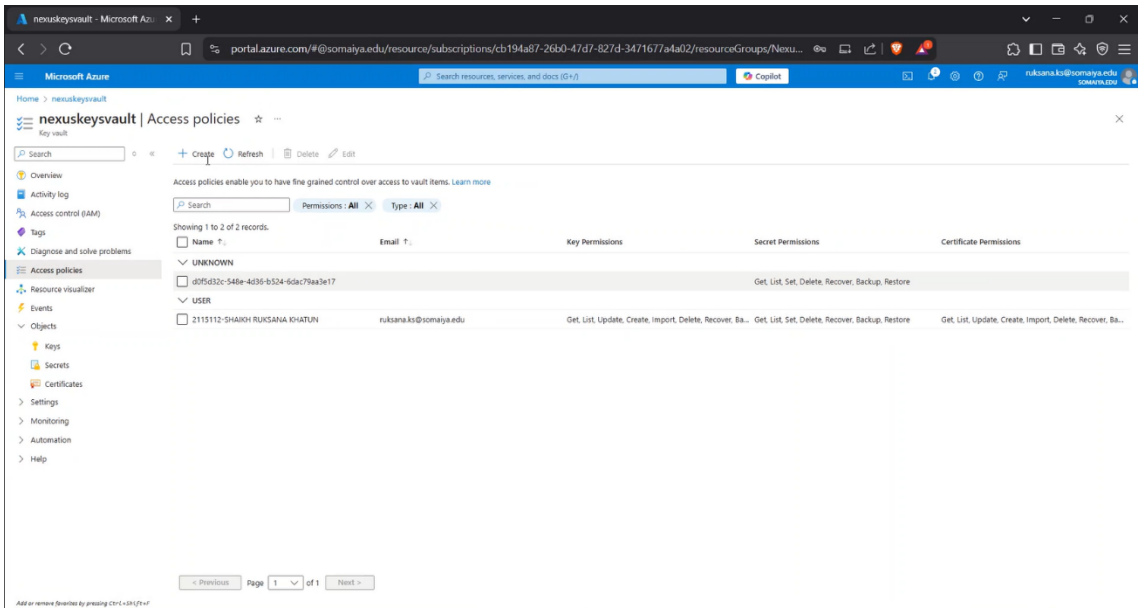
The secret **nexuspwd** in Azure Key Vault has one enabled version with no set activation or expiration date.



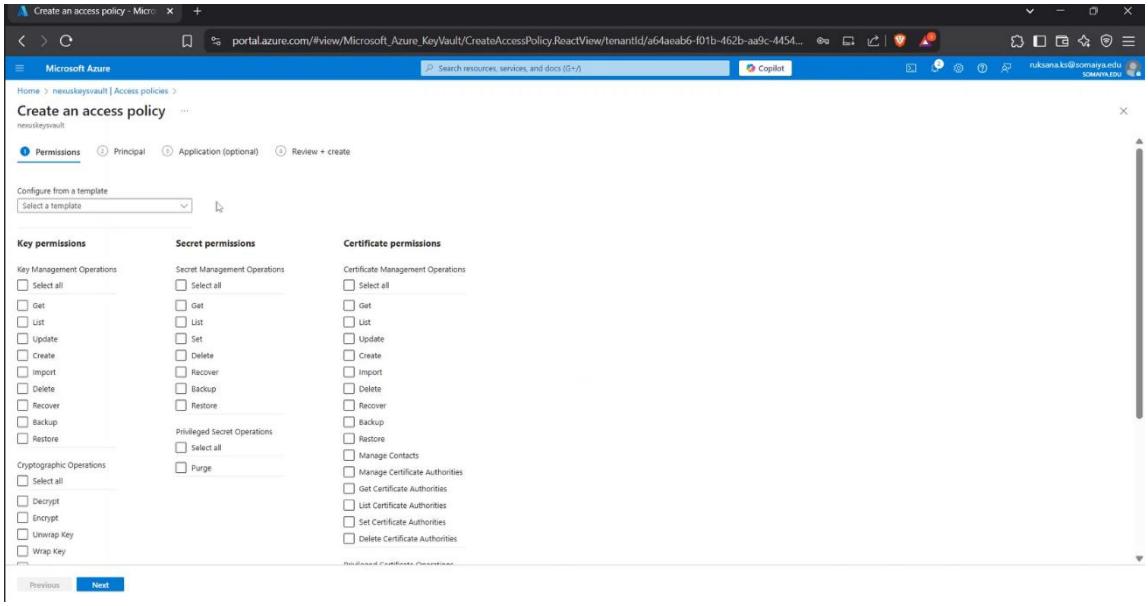
The **Access control (IAM)** page for the **Azure Key Vault nexuskeyvault** allows managing role-based access, including checking access, assigning roles, and viewing permissions.



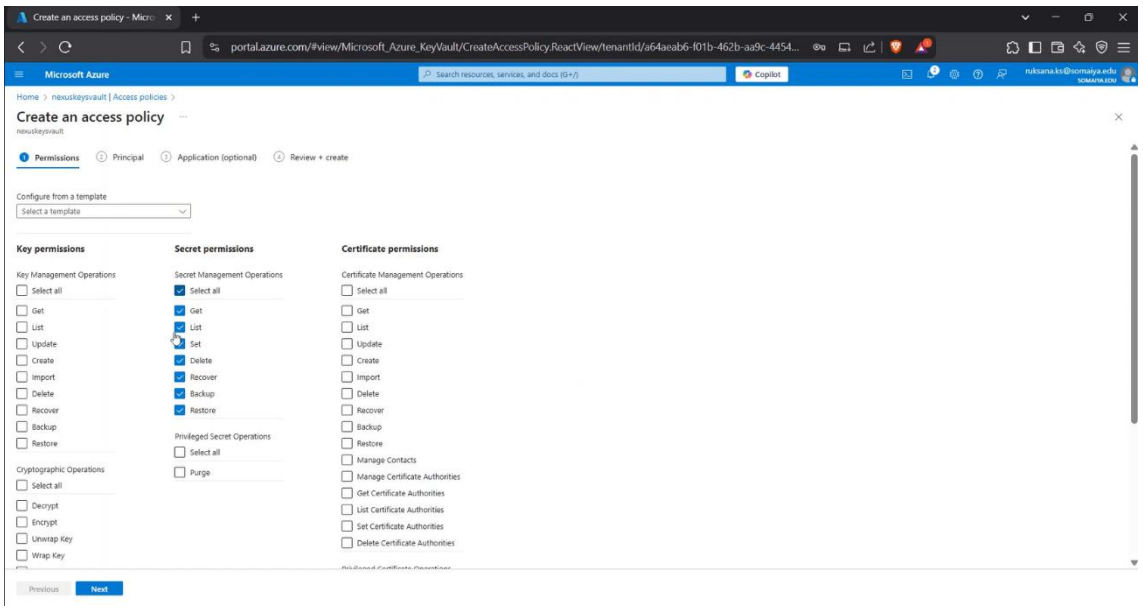
The Access policies page of nexuskeyvault displays users with specific permissions to manage keys, secrets, and certificates, with an option to create new policies.



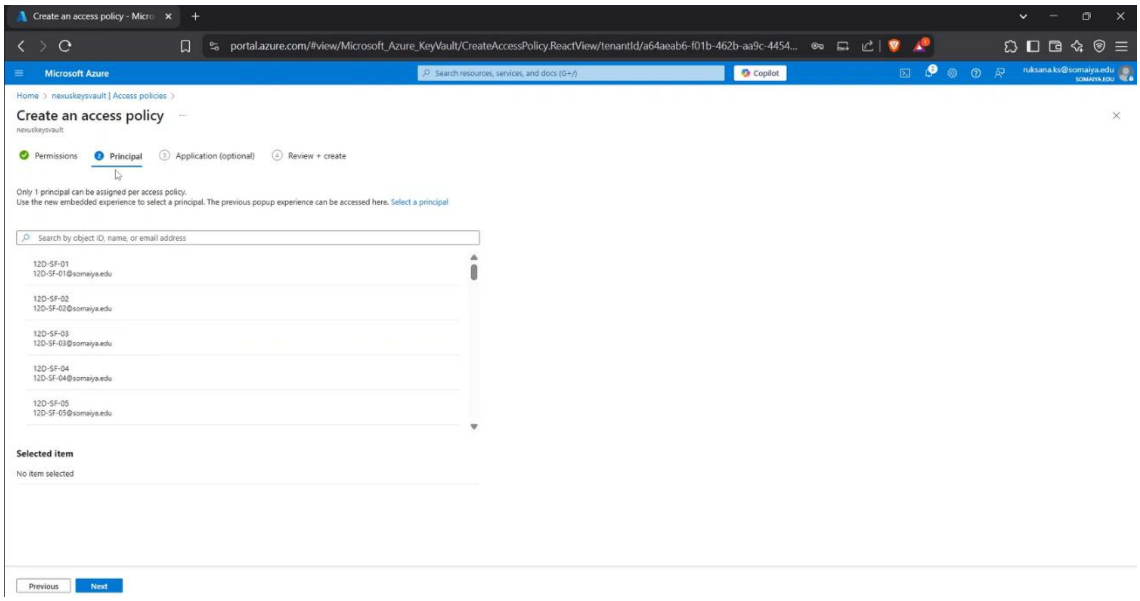
Azure allows configuring permissions for a new Key Vault access policy, including key, secret, and certificate operations.



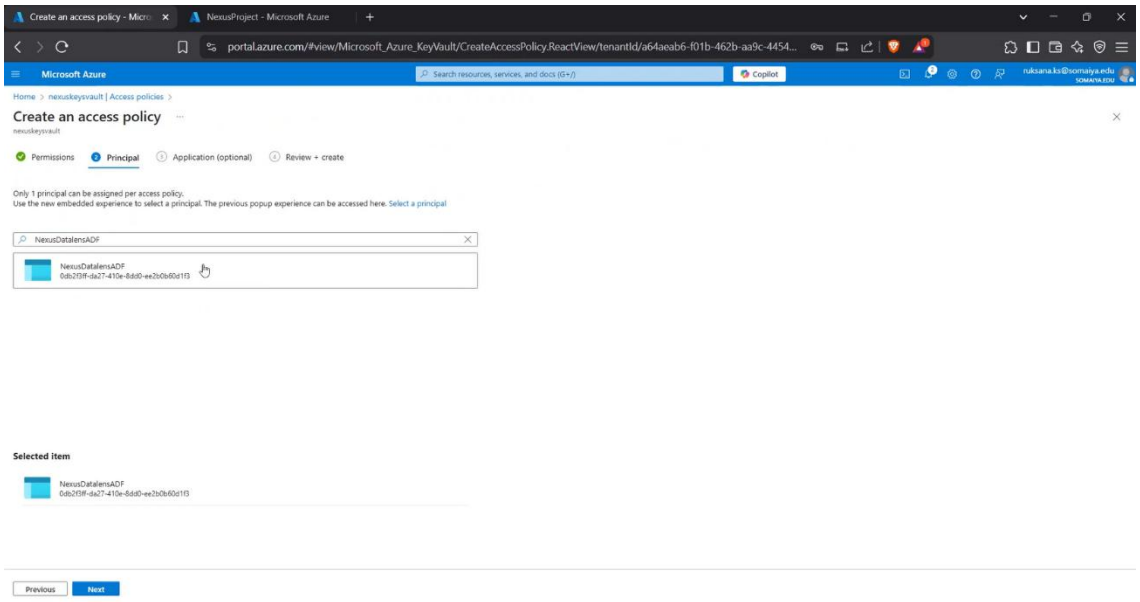
Secret permissions **"Get," "List,"** and **"Delete"** are selected while creating a Key Vault access policy in Azure.



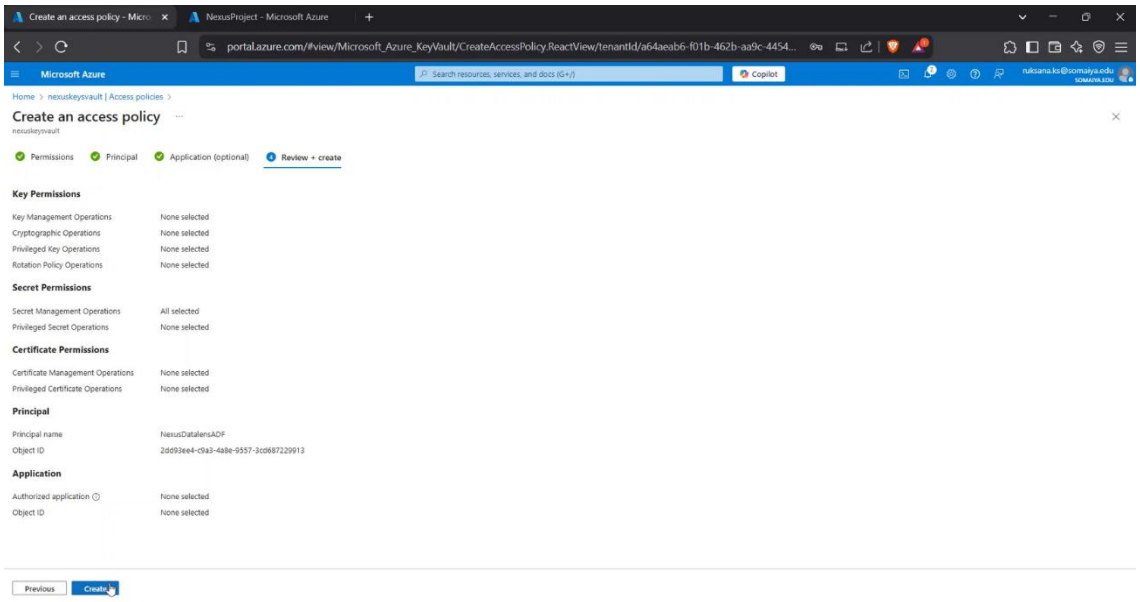
It allows selecting a principal (user, app, or group) to assign the access policy in Azure Key Vault.



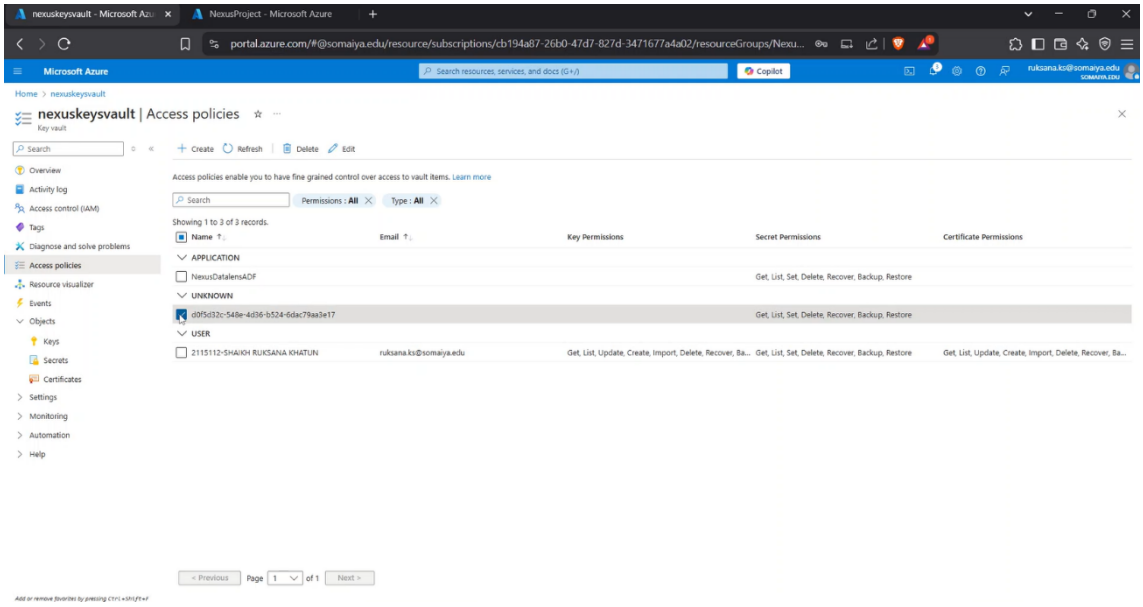
The principal **NexusDataLensADF** is selected to assign permissions in the Azure Key Vault access policy.



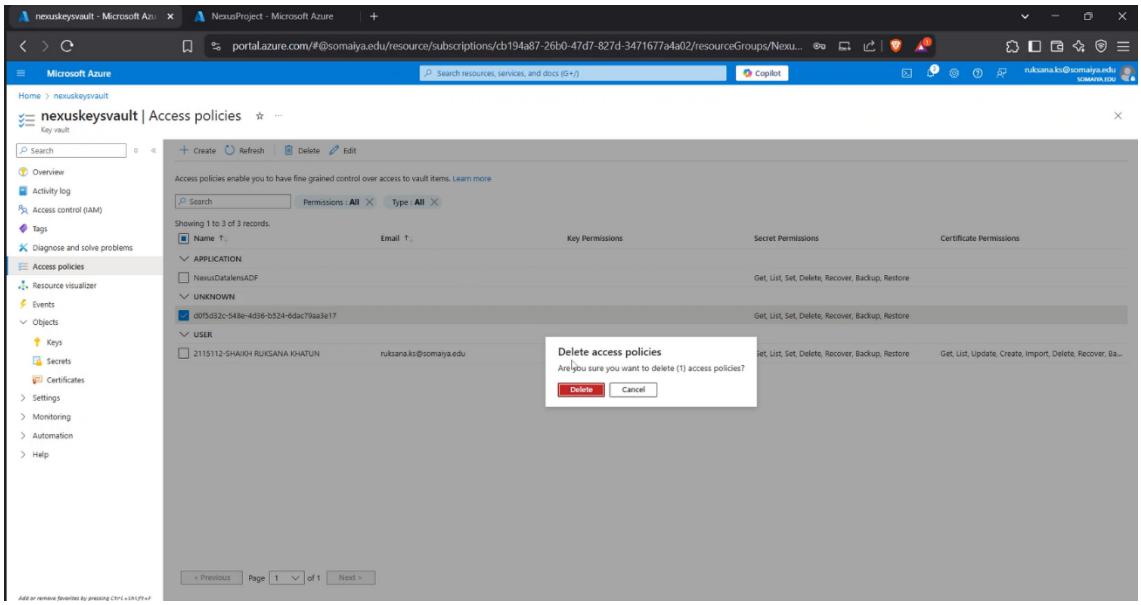
The final review screen confirms that all secret permissions are granted to the principal **NexusDataLensADF**, and the access policy is ready to be created.



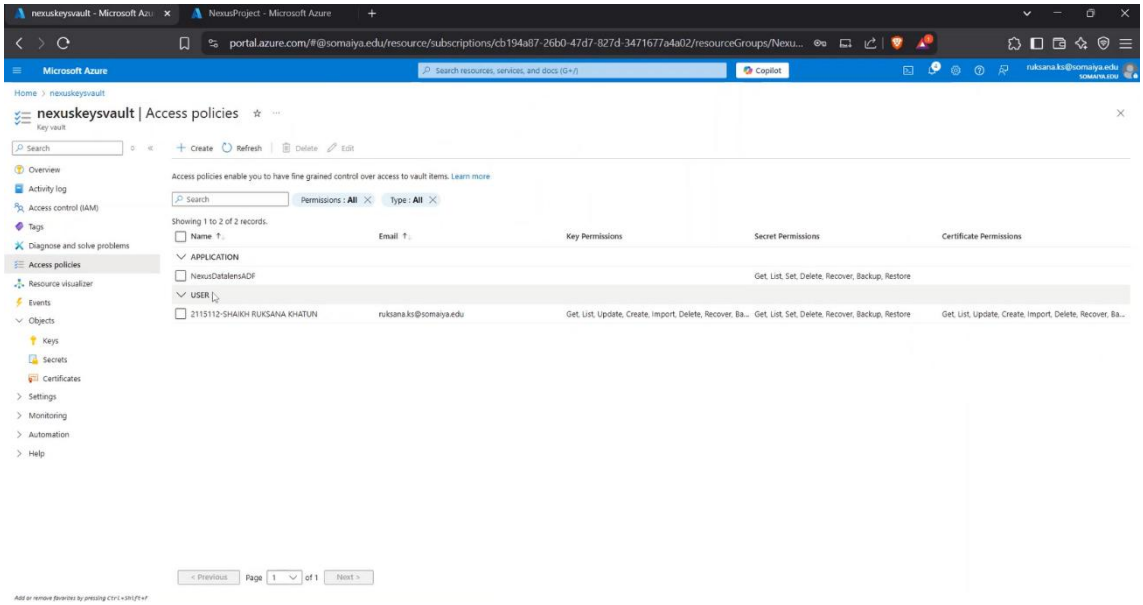
A new access policy for the **NexusDataLensADF** application has been successfully added with full secret permissions (Get, List, Set, Delete, Recover, Backup, Restore).



A confirmation dialog is shown to delete one selected access policy from the Azure Key Vault.

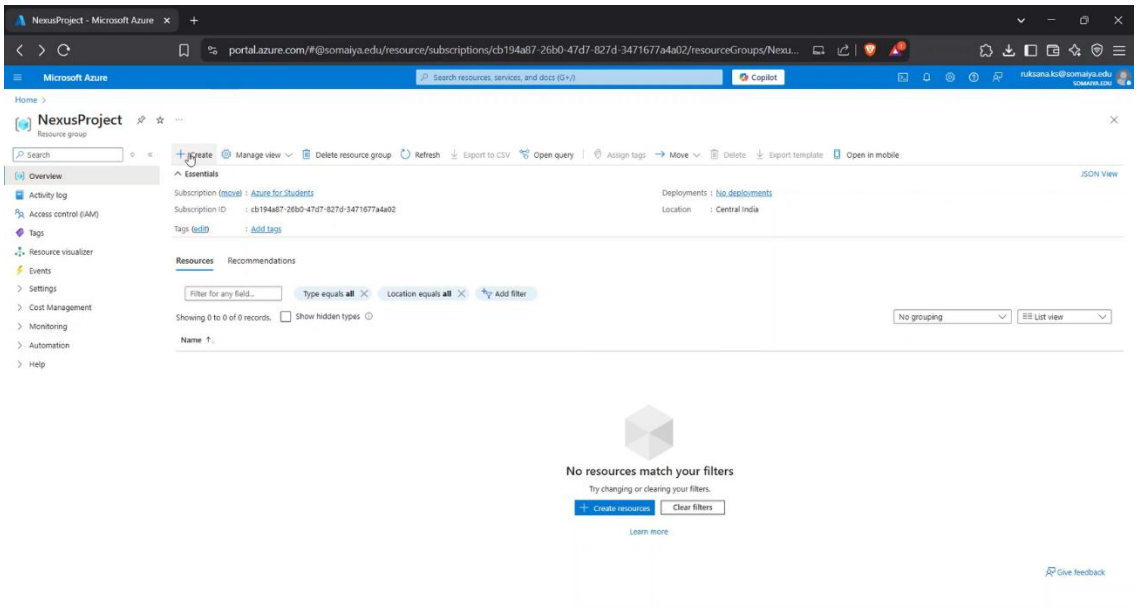


The access policy was successfully deleted, leaving two remaining entries for the application "**NexusDataInesADF**" and the user "**SHAIKH RUKSANA KHATUN**."

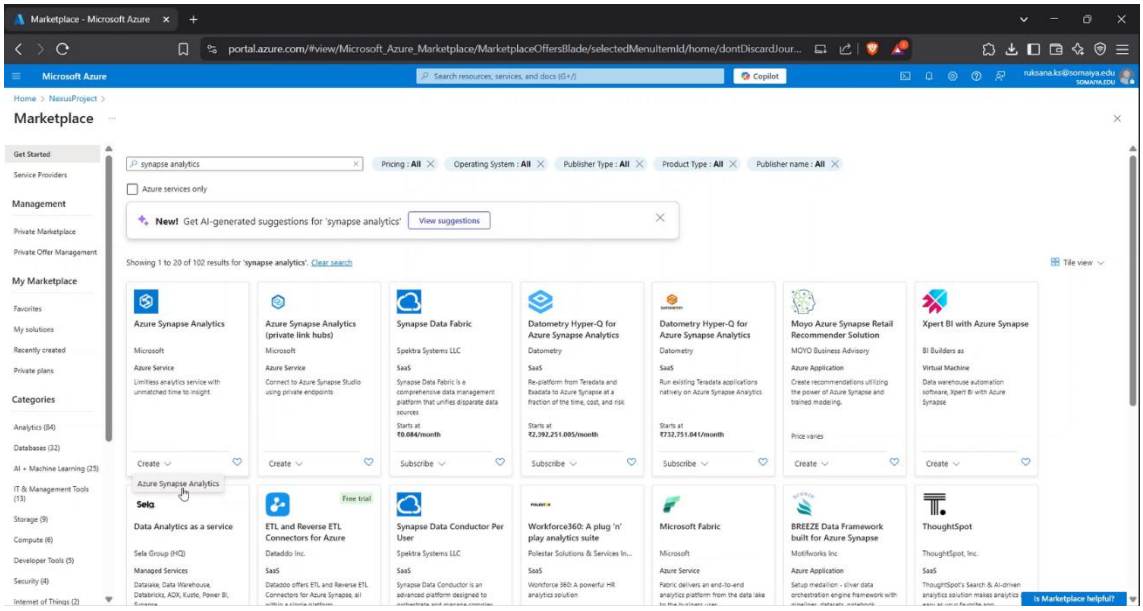


SETTING RESOUCE GROUP

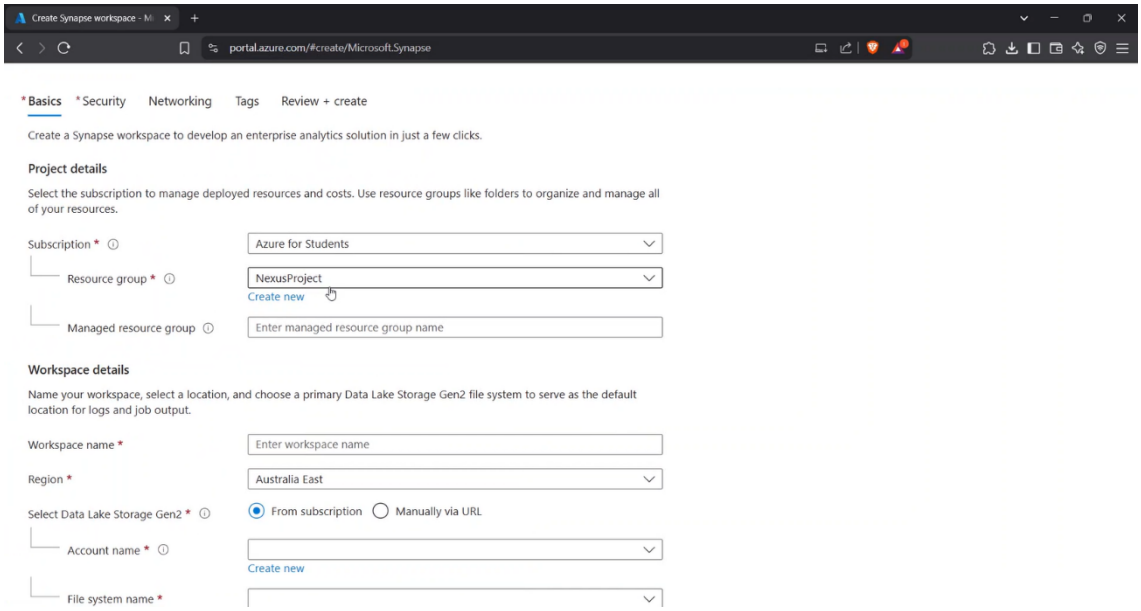
The "**NexusProject**" resource group currently has no visible resources due to active filters, with an option to create new resources or clear filters.



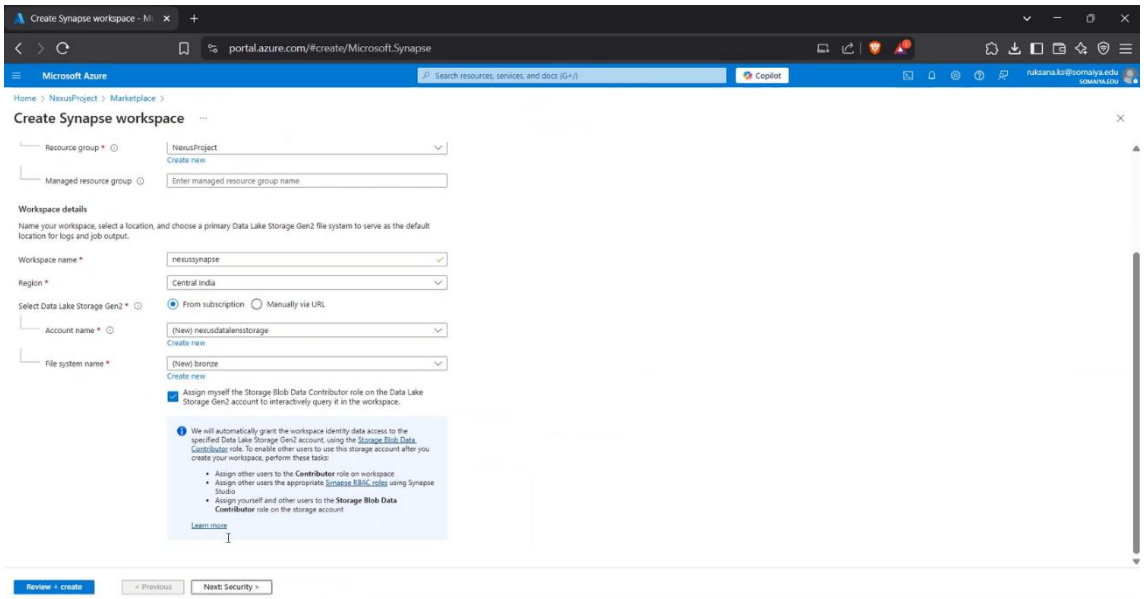
The Azure Marketplace displays various Synapse Analytics-related services and solutions available for creation or subscription within the "**NexusProject**" environment.



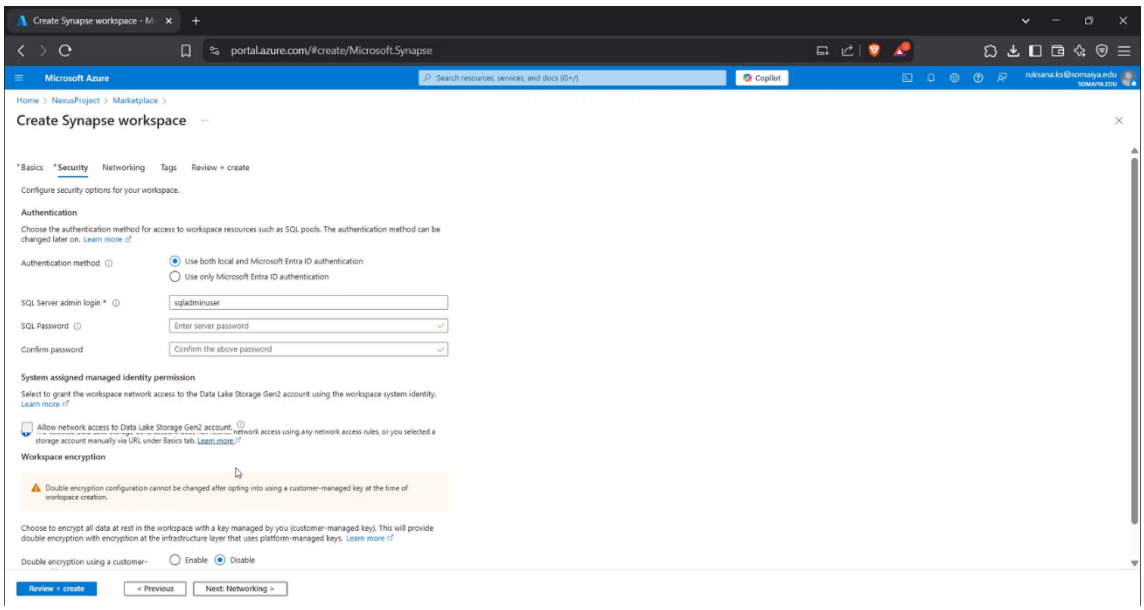
The form is being filled to create a new Synapse workspace under the **"NexusProject"** resource group with specified subscription, region, and linked Data Lake Storage Gen2 account.



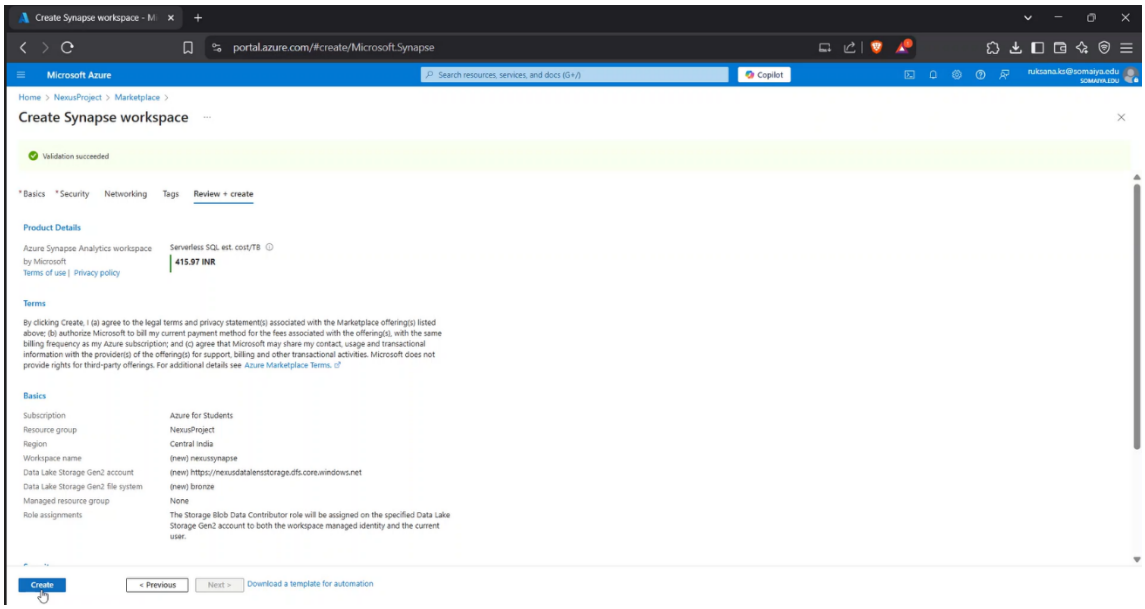
A Synapse workspace named "**nexussynapse**" is being created in Central India, linked to the "**nexusdatalensstorage**" Data Lake Storage with "**bronze**" as the file system.



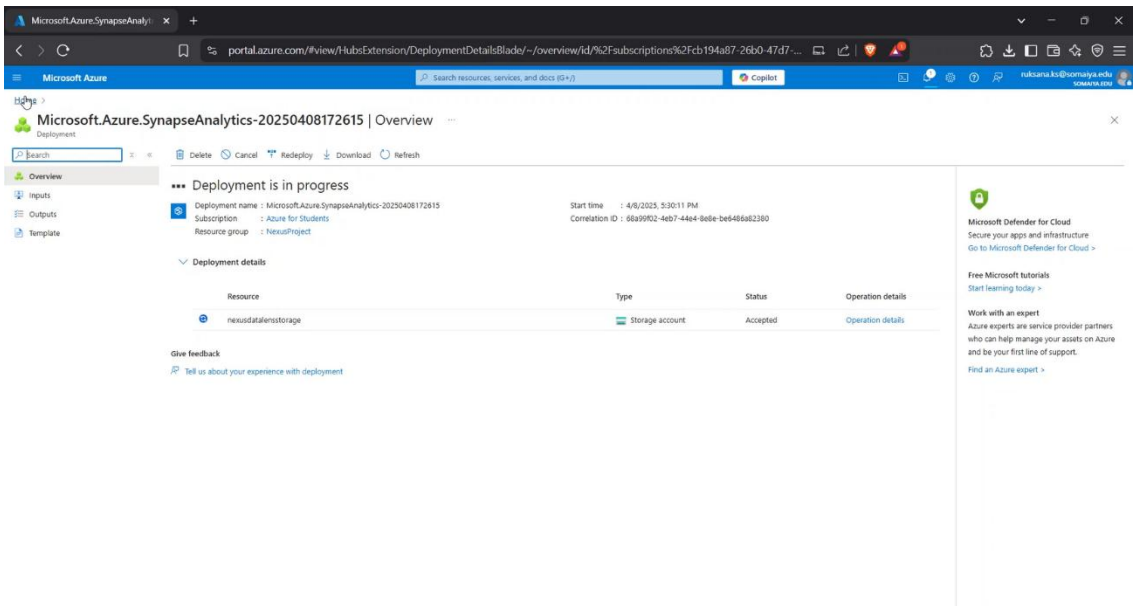
The Synapse workspace is being configured with SQL authentication, network access to Data Lake enabled, and default platform-managed encryption selected.



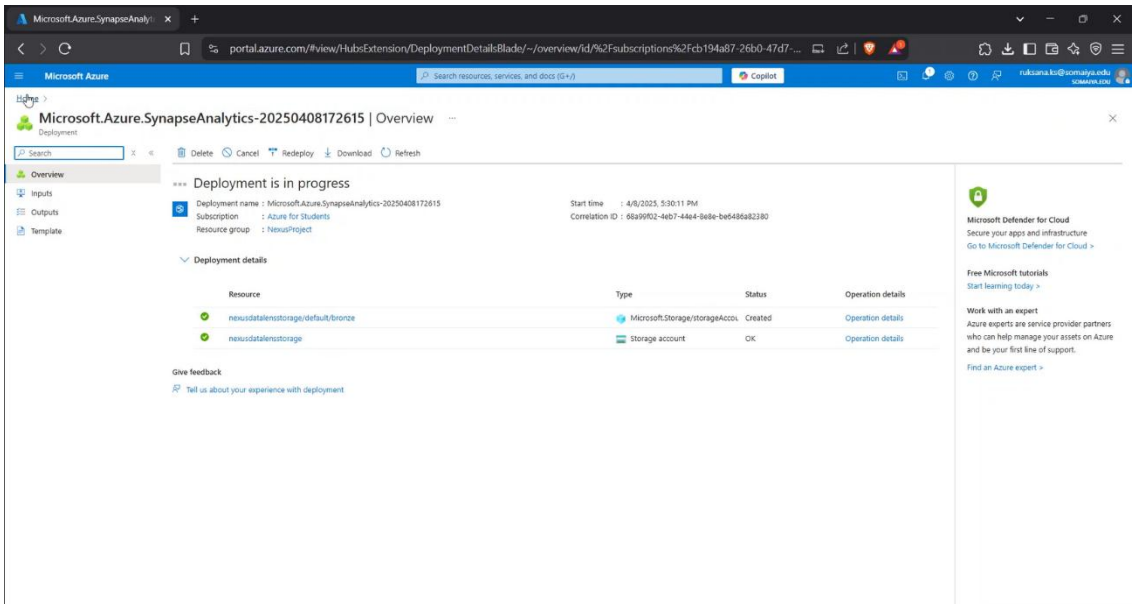
The **Synapse workspace setup review** has **succeeded**, and the user is ready to create it with specified configuration and pricing.



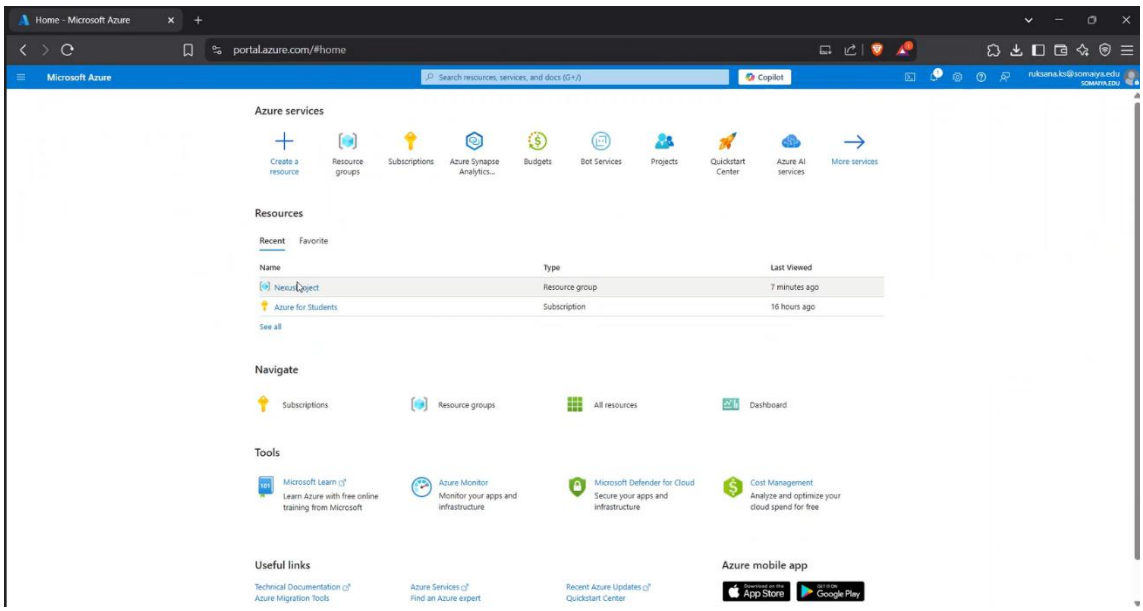
The Synapse workspace deployment is currently in progress with the storage account resource accepted under the **NexusProject resource group**.



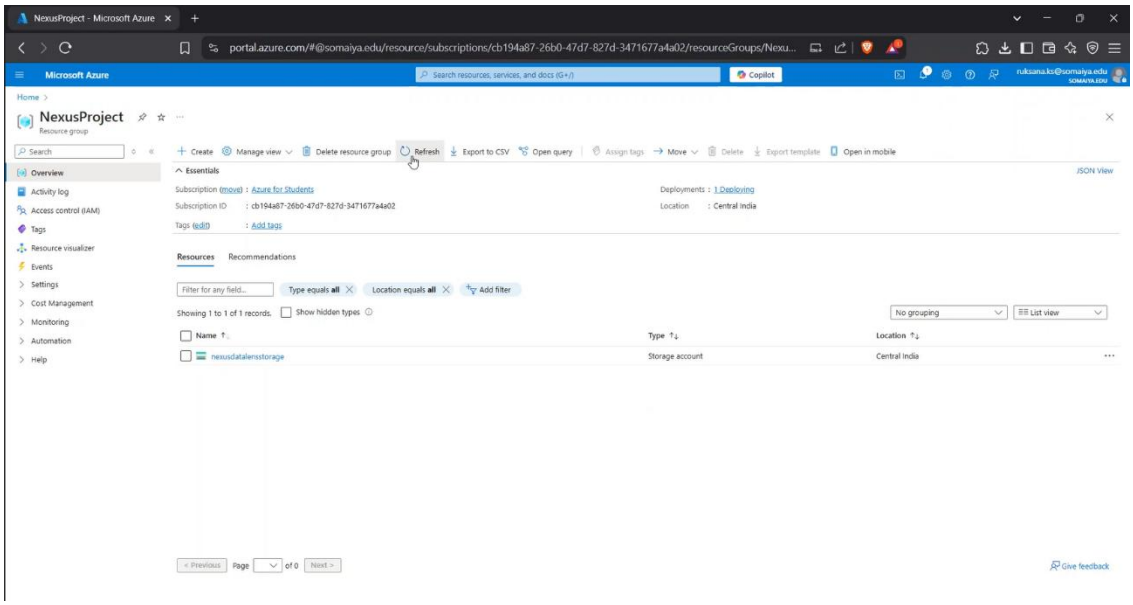
The Synapse workspace deployment is in **progress**, with the storage container bronze created and the storage account status marked as OK.



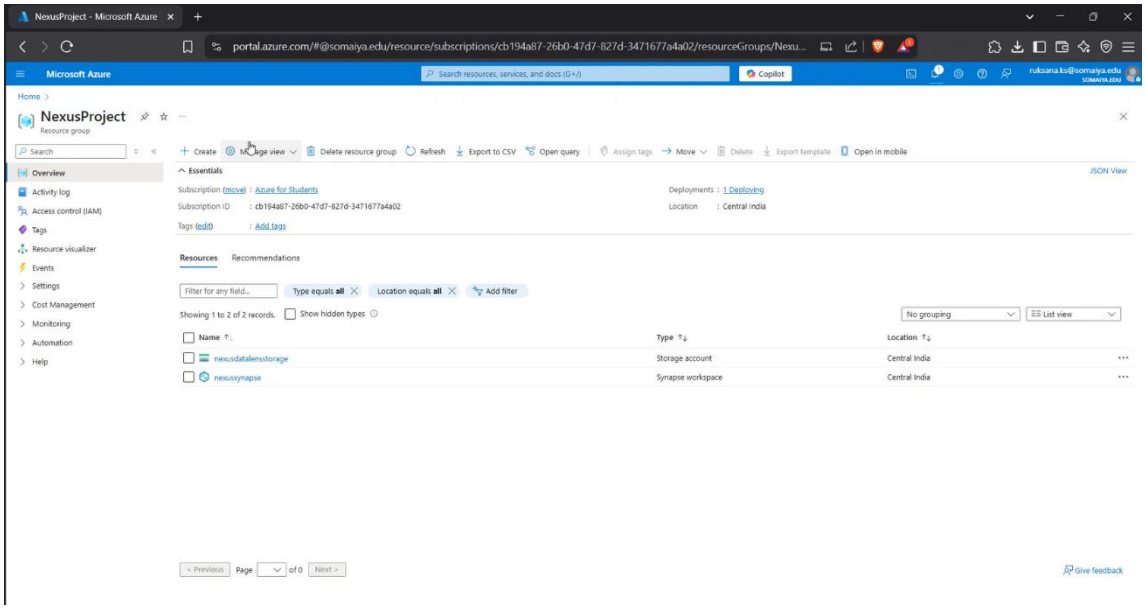
The Azure portal home page displays recent resources, with "**NexusProject" (a resource group)** being the most recently accessed.



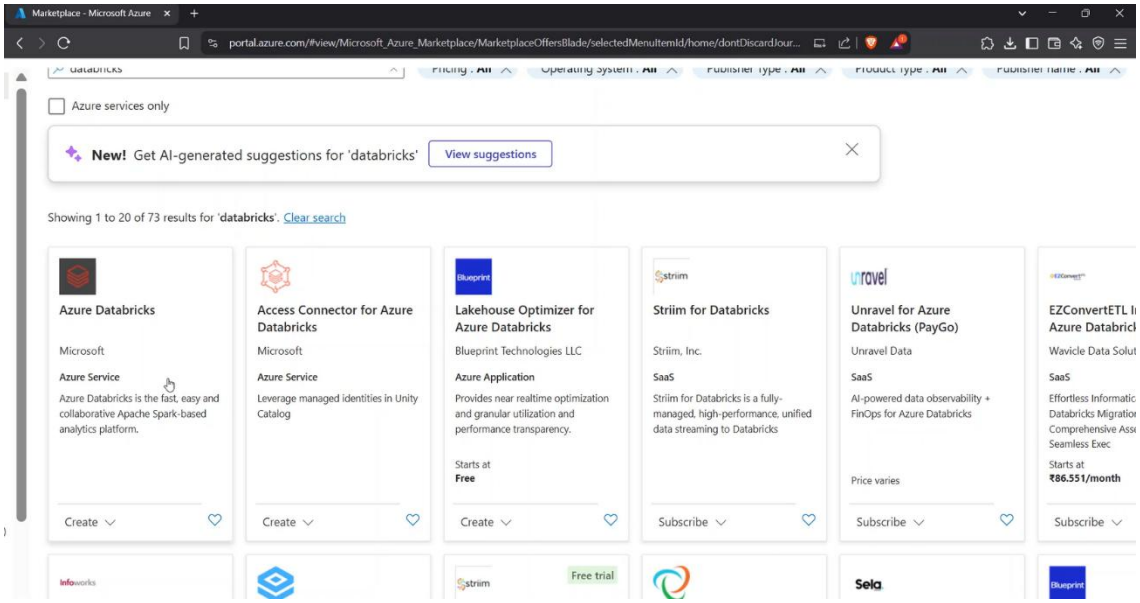
The image shows an Azure portal page displaying the "**NexusProject**" resource group with one storage account resource named "**nexusdatastenstorage**" located in Central India.



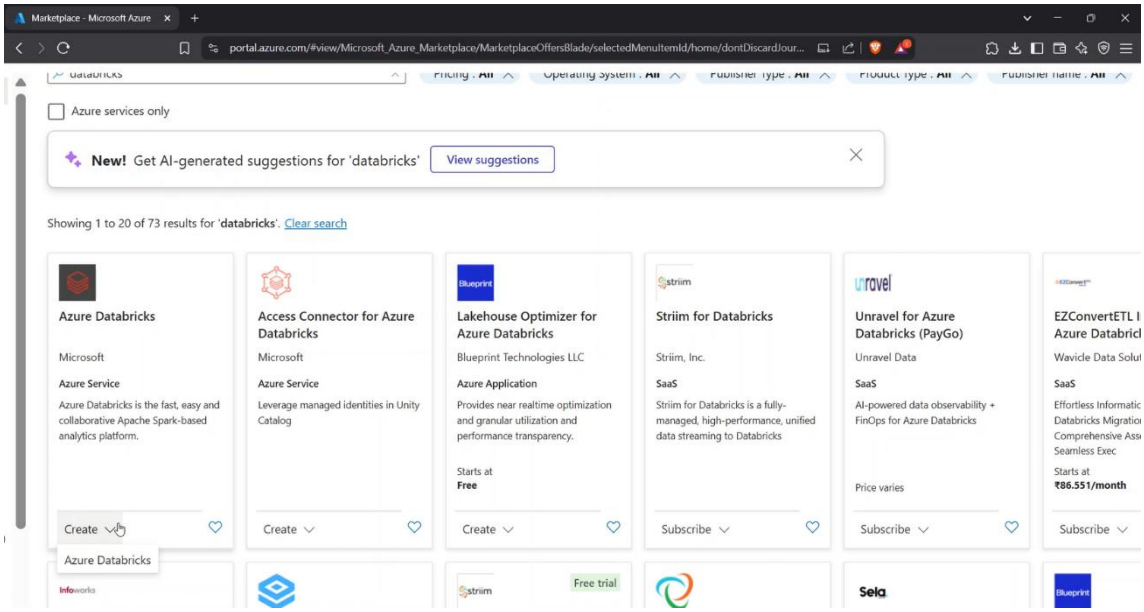
The updated Azure portal view shows the "**NexusProject**" resource group now containing two resources: a storage account ("**nexusdatastenstorage**") and a Synapse workspace ("**nexussynapse**"), both located in Central India.



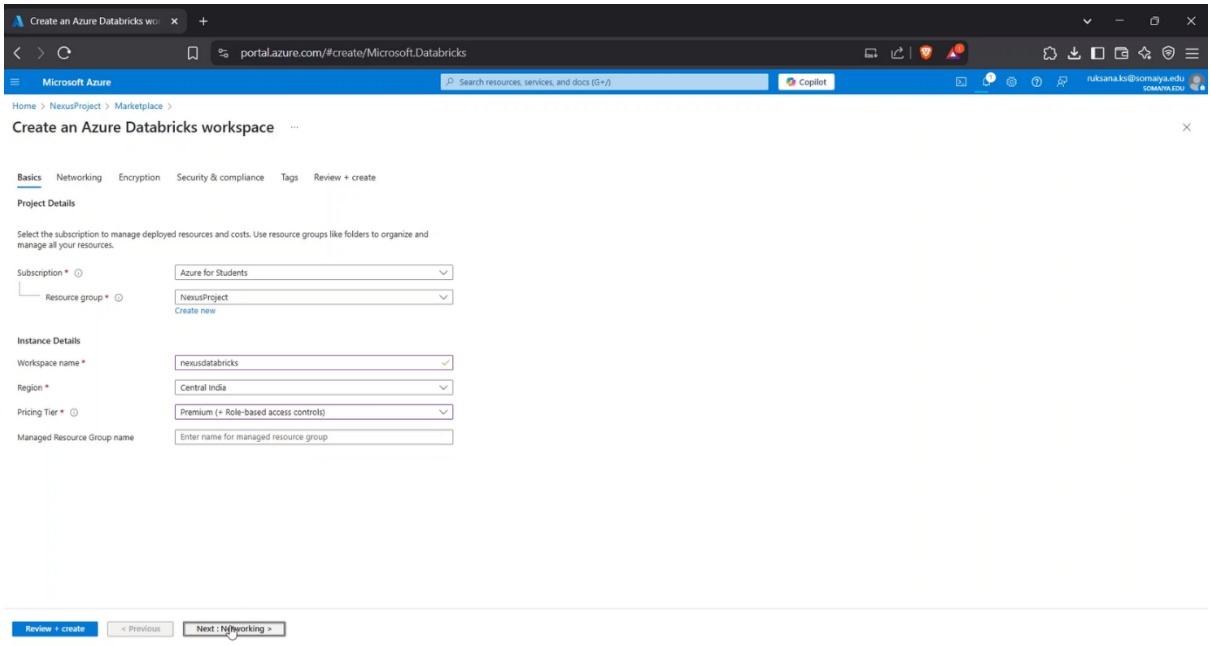
The Azure Marketplace displays various Databricks-related services, including "**Azure Databricks**" by Microsoft for Spark-based analytics and other integration and optimization tools.



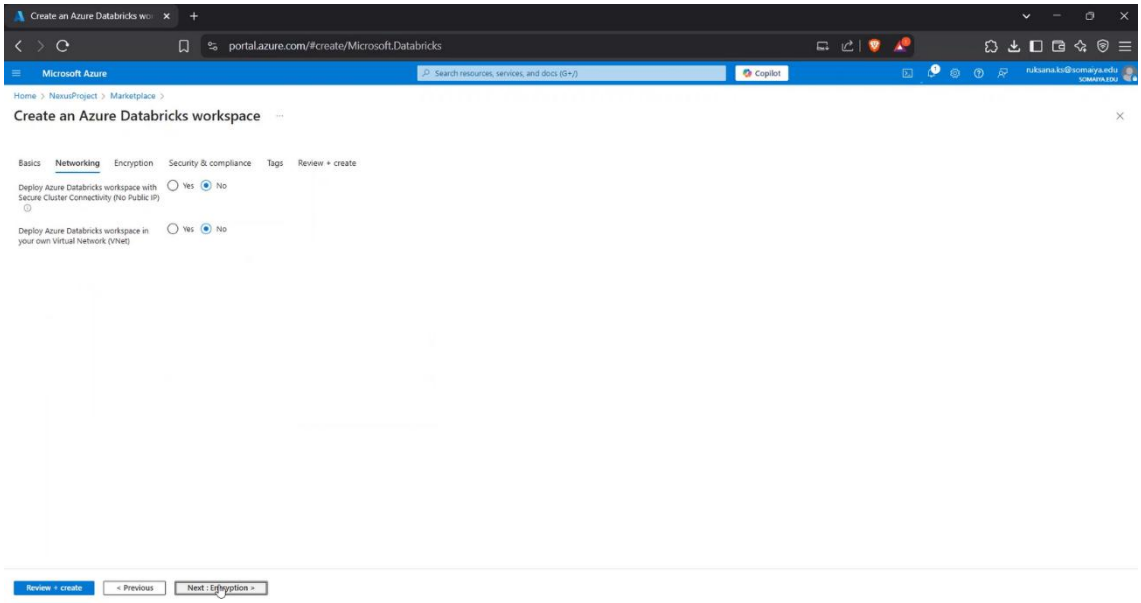
The user is about to create an "**Azure Databricks**" workspace from the Azure Marketplace by clicking the "**Create**" button under the Azure Databricks service.



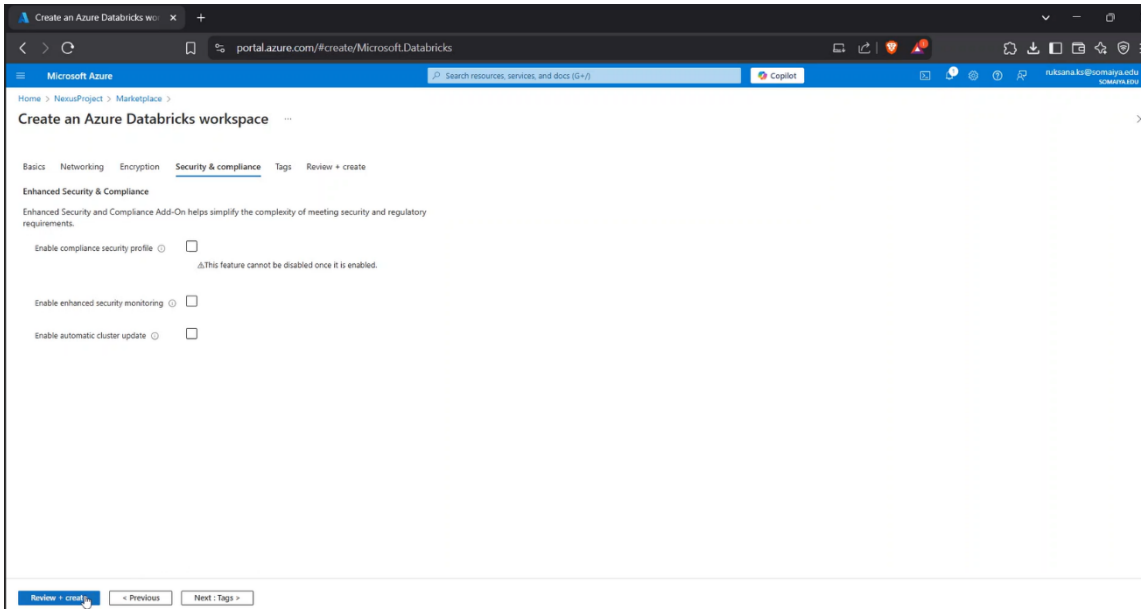
The user is configuring the basic settings to create an Azure Databricks workspace named "**nexusdatabricks**" in the "**NexusProject**" resource group within the Central India region.



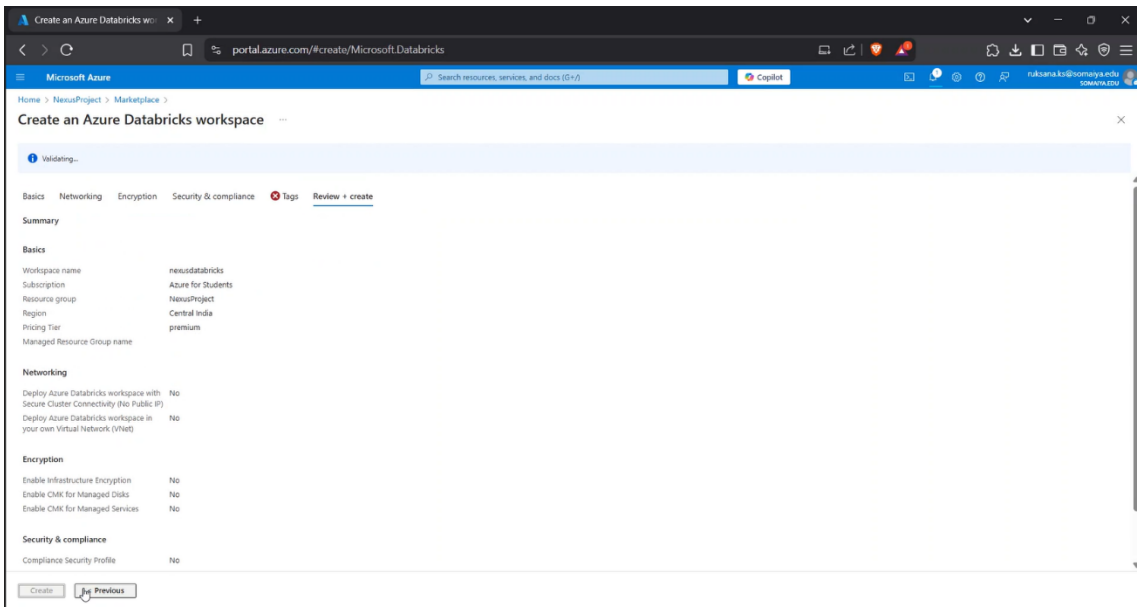
The user is configuring networking options for the Azure Databricks workspace, choosing whether to enable secure cluster connectivity and virtual network deployment.



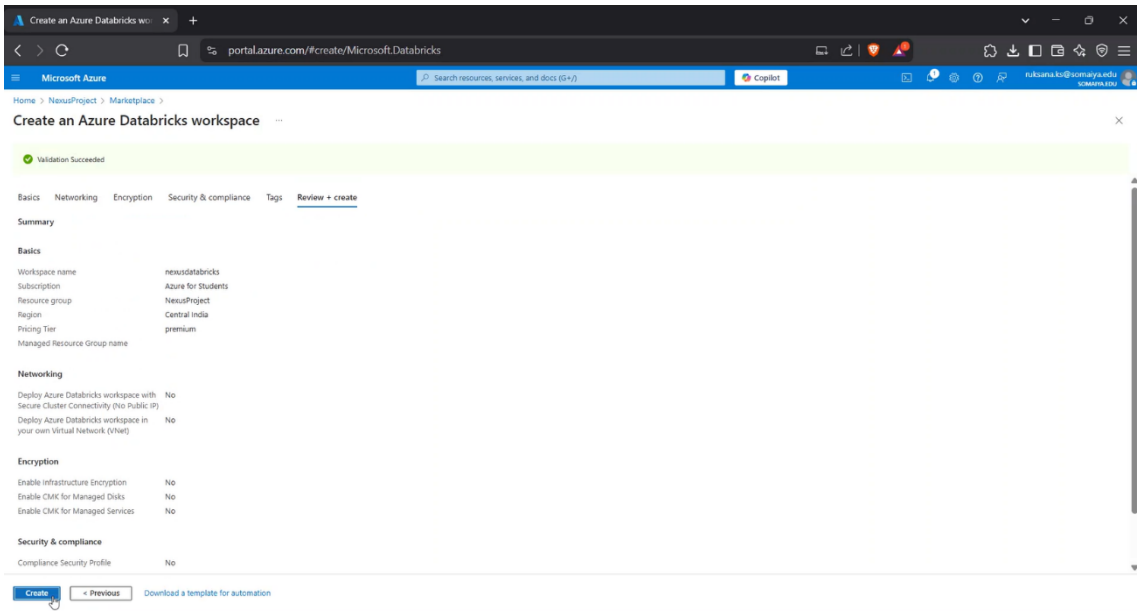
The user is on the Security & Compliance tab while creating an Azure Databricks workspace, with optional settings for compliance, monitoring, and automatic cluster updates before proceeding to review and create.



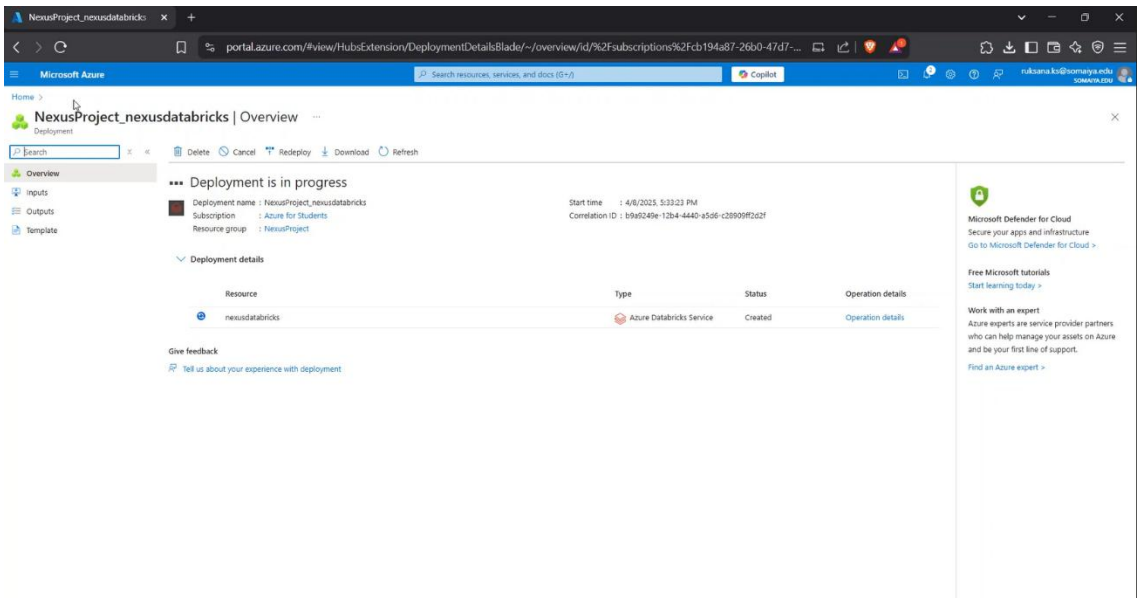
The user is on the final "**Review + create**" step for setting up an Azure Databricks workspace, reviewing all configuration details before deployment.



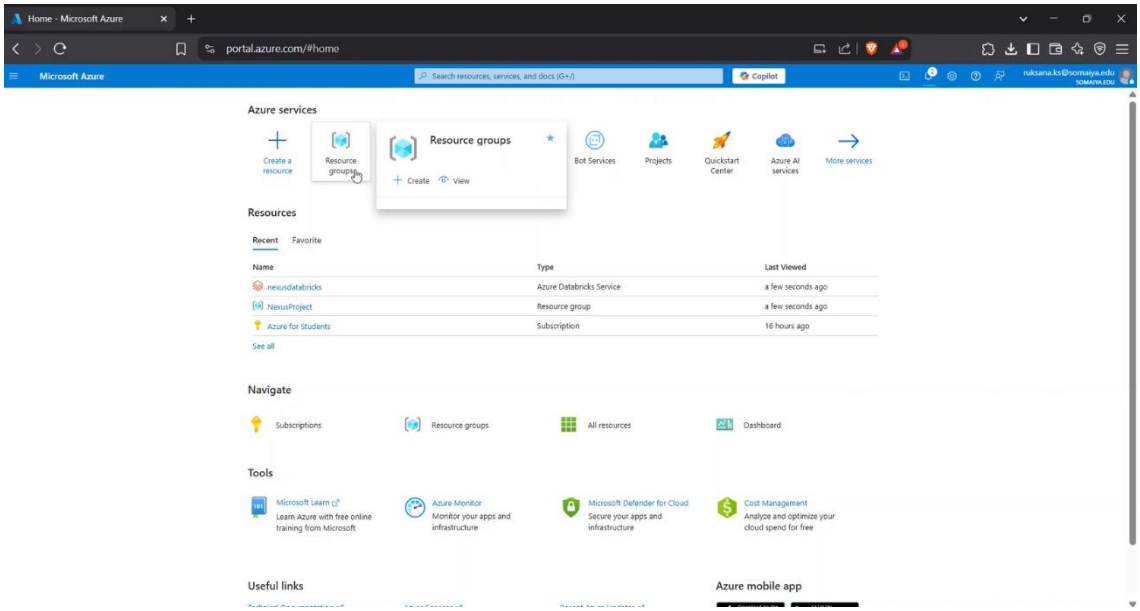
The user has successfully validated the configuration and is ready to create the **Azure Databricks workspace** by clicking the "**Create**" button.



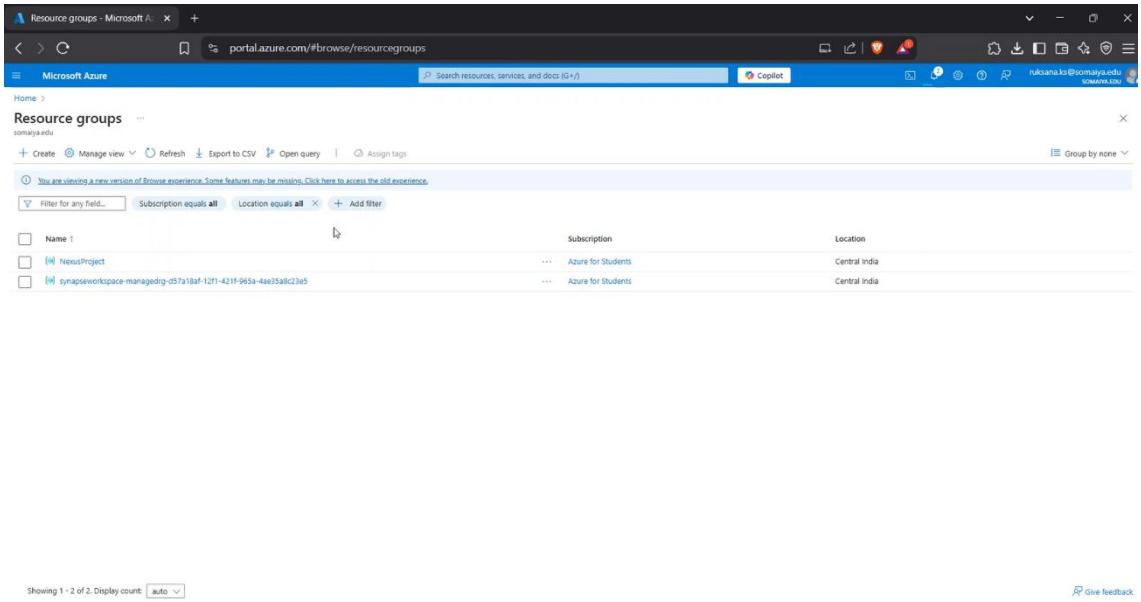
The deployment of the Azure Databricks workspace named "**nexusdatabricks**" is currently in progress within the "**NexusProject**" resource group.



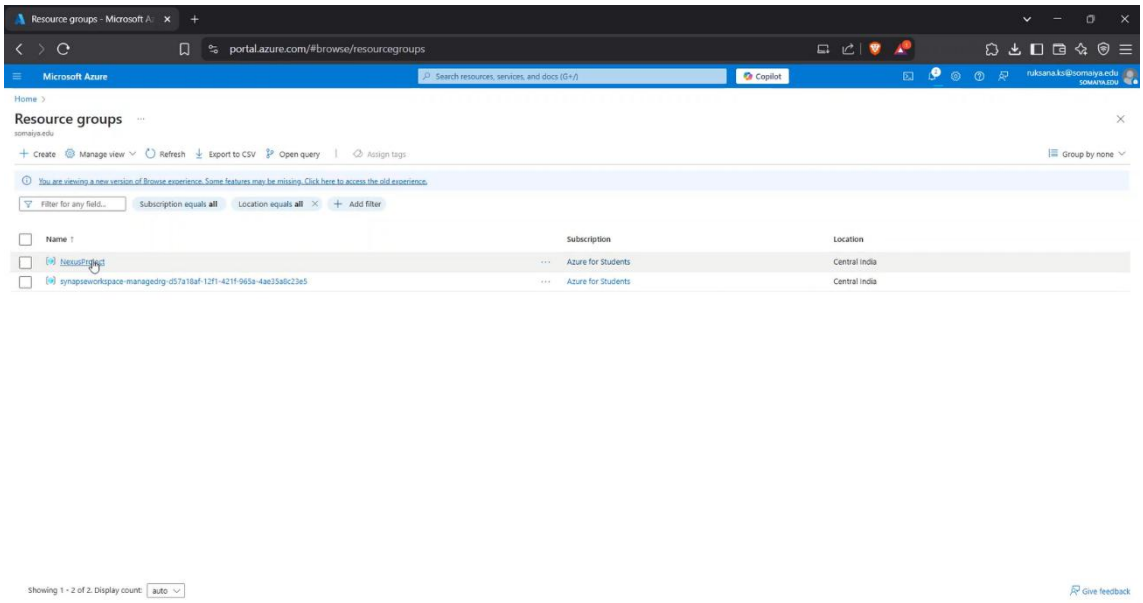
On the Azure portal home screen click on the "**Resource groups**" option to view or **create a new resource group**.



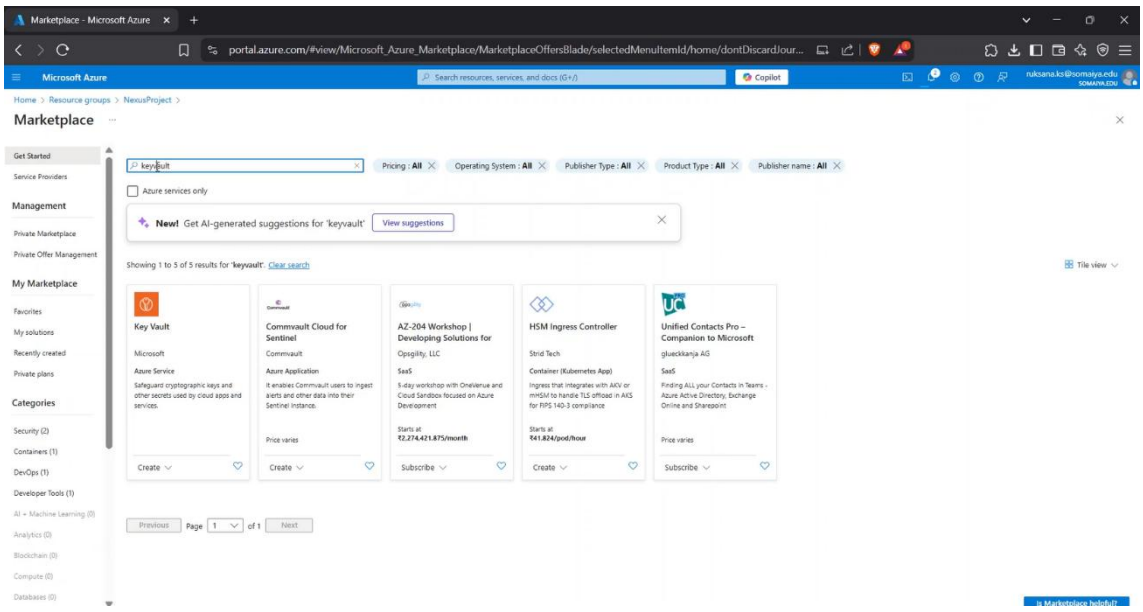
The "**Resource groups**" page in the Azure portal displays existing resource groups under the "**Azure for Students**" subscription, with both located in Central India.



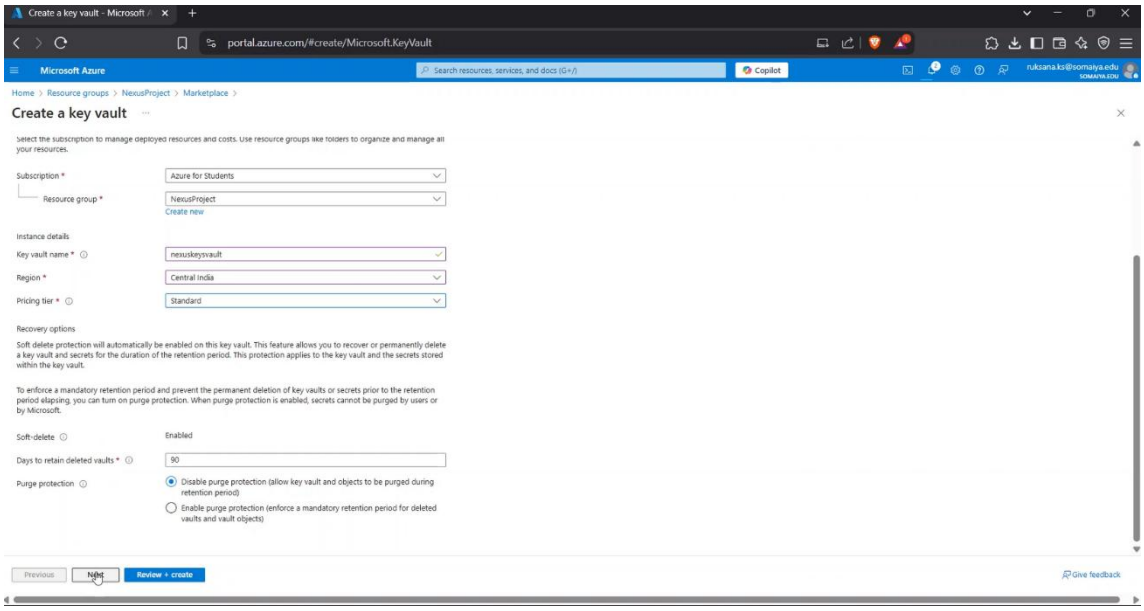
The user selects the "**NexusProject**" resource group from the list of available groups under the "**Azure for Students**" subscription in the Central India region.



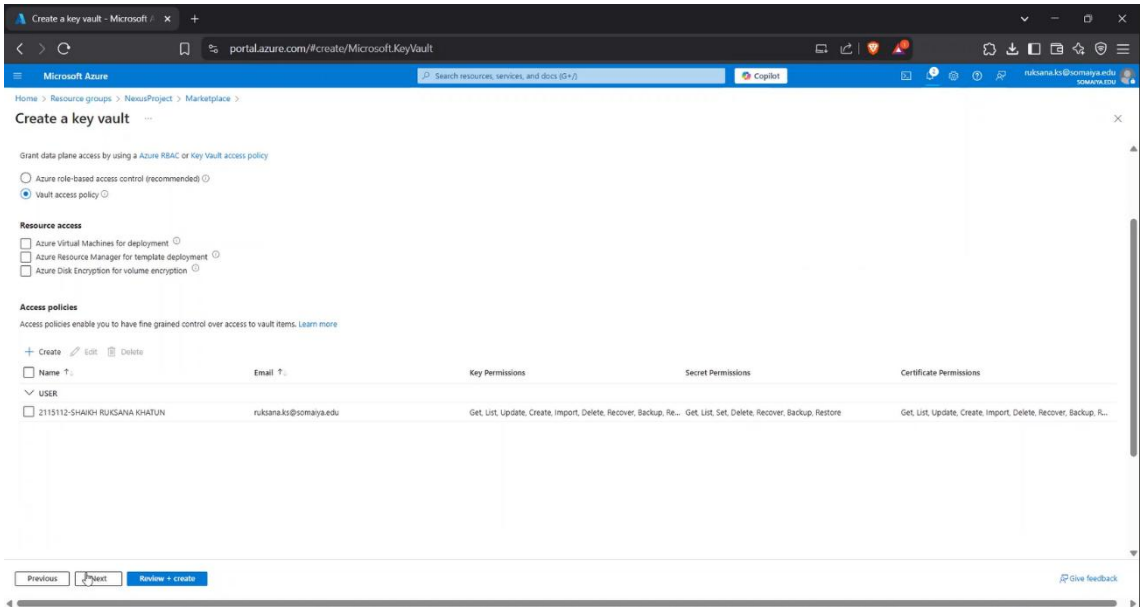
The user searches for "**Key Vault**" in the Azure Marketplace under the NexusProject resource group to **create a new key vault service**.



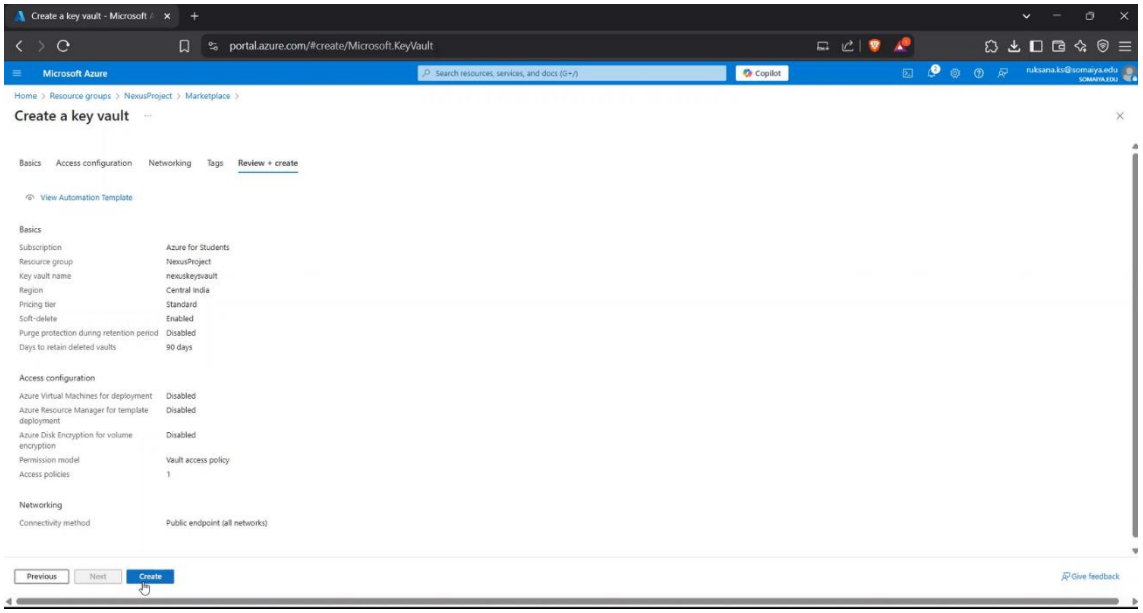
The user is configuring and creating a new Azure Key Vault named "**nexuskeyvault**" under the "**NexusProject**" resource group in the Central India region.



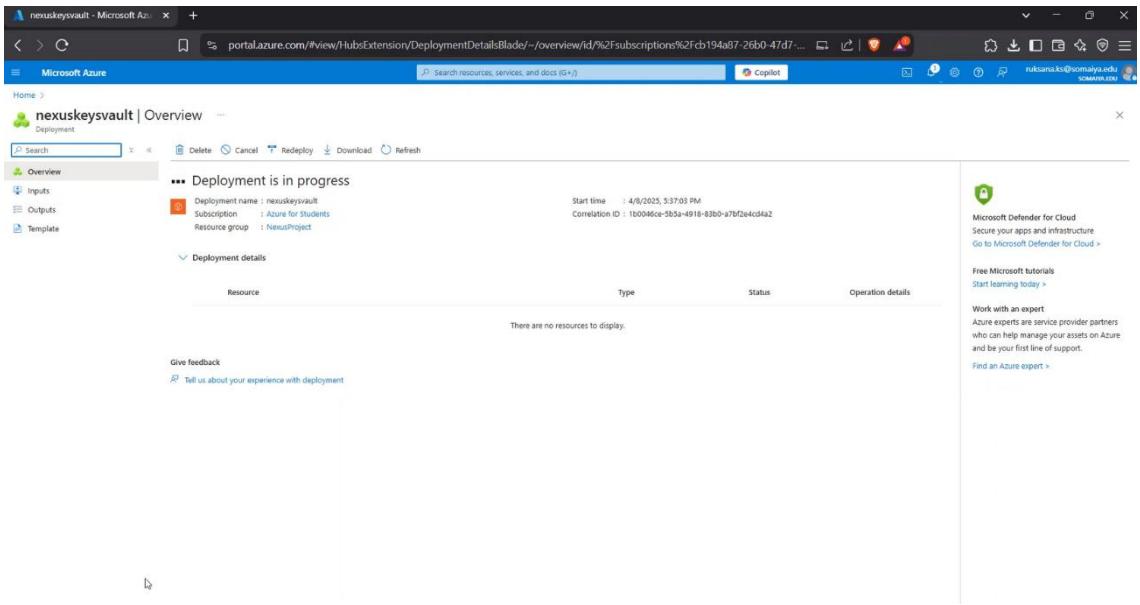
The user is assigning full key, secret, and certificate permissions to a specified user while setting access policies for the new Azure Key Vault before final review and creation.



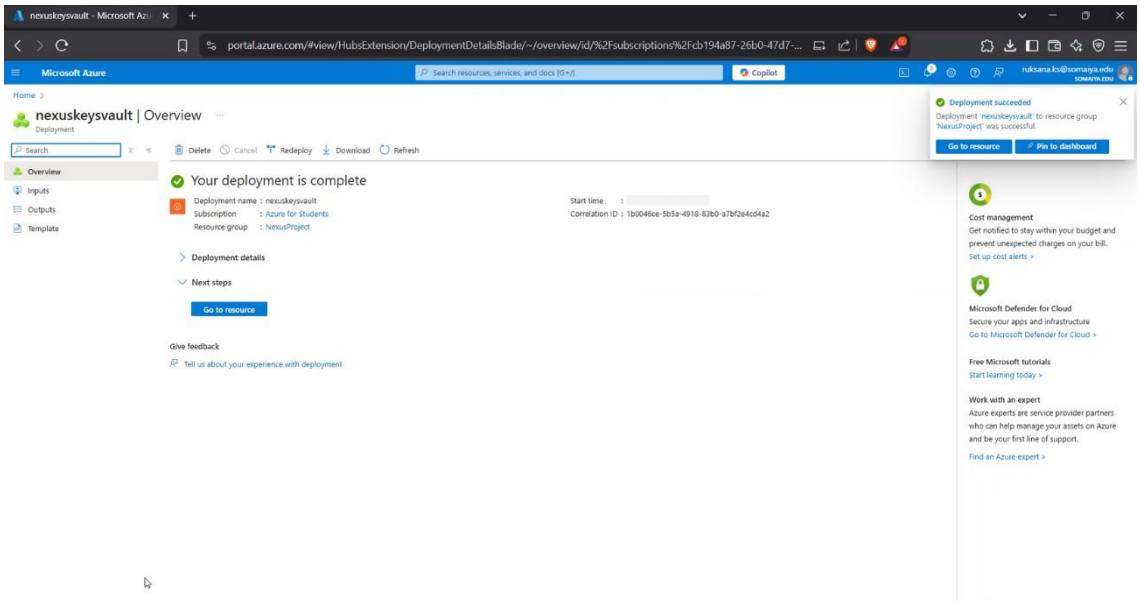
The user is finalizing the configuration and clicking "**Create**" to deploy a new Azure Key Vault with specified settings and access policies.



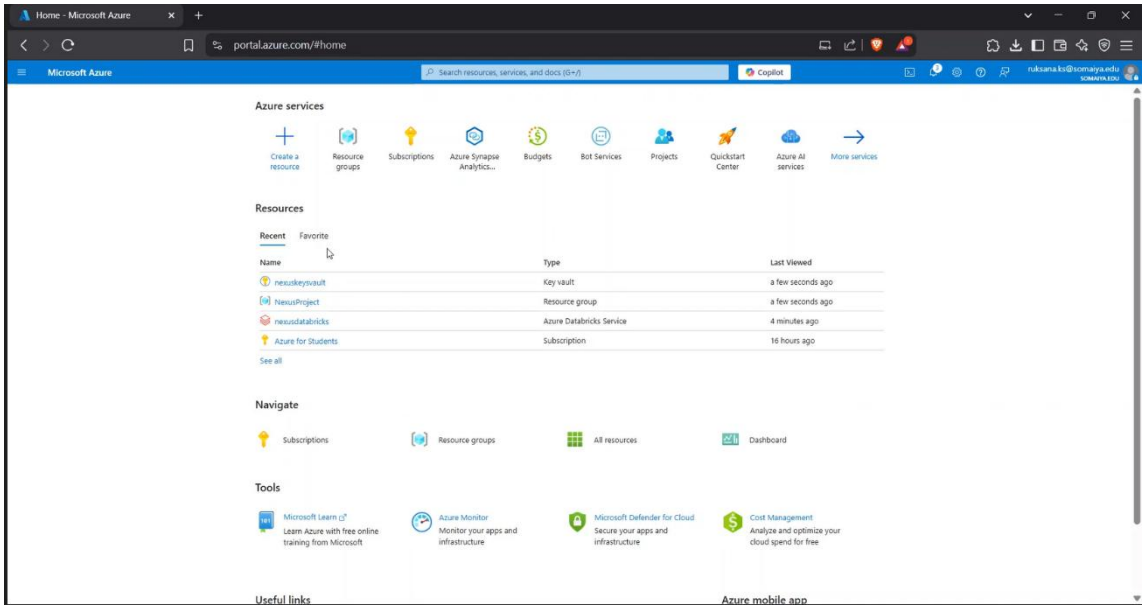
The deployment of the Azure Key Vault named "**nexuskeyvault**" is currently in progress within the "**NexusProject**" resource group.



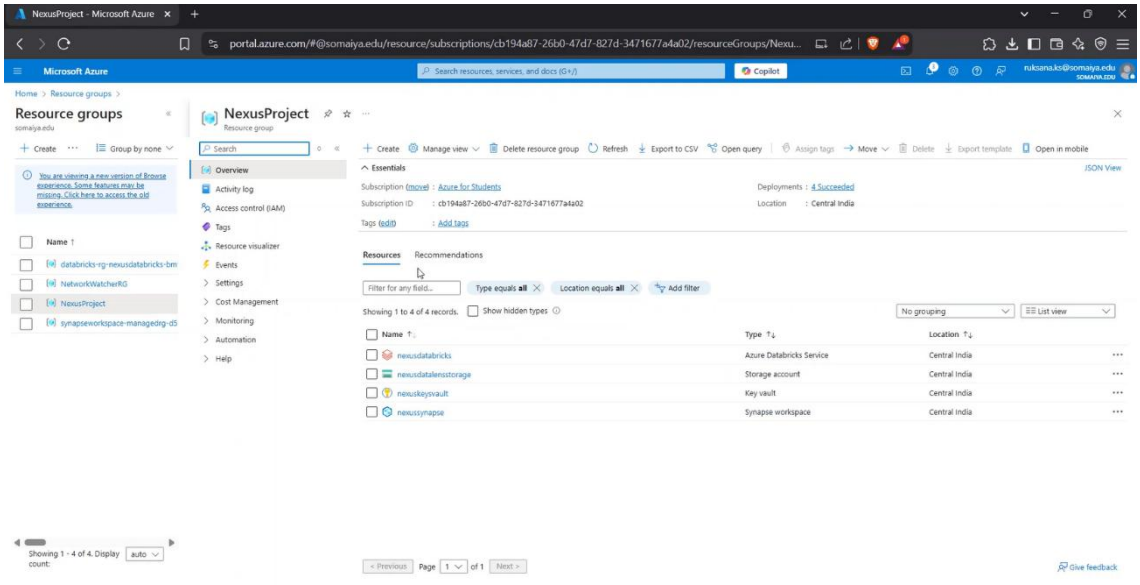
The deployment of the "**nexuskeyvault**" in the "**NexusProject**" resource group has completed successfully



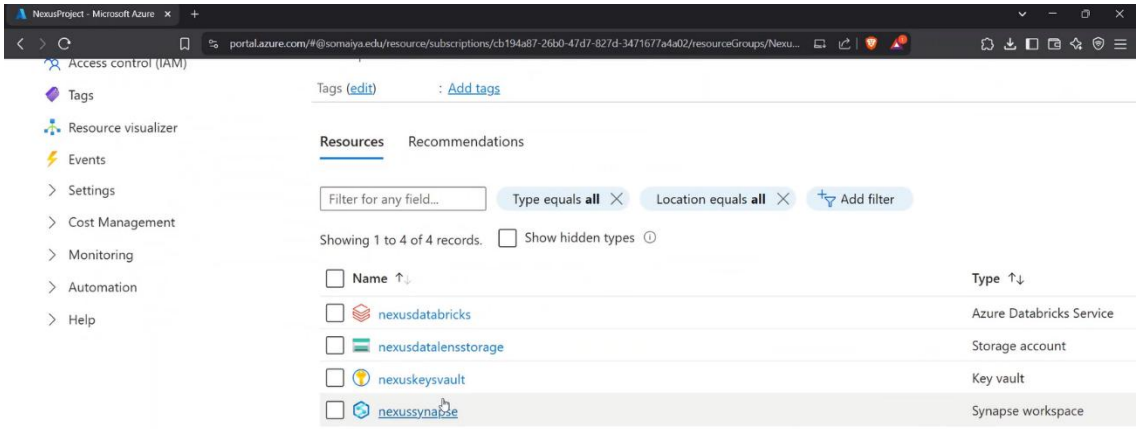
The Azure portal home screen shows the recently accessed resources, including the newly created Key Vault "**nexuskeyvault**" under the "**NexusProject**" resource group.



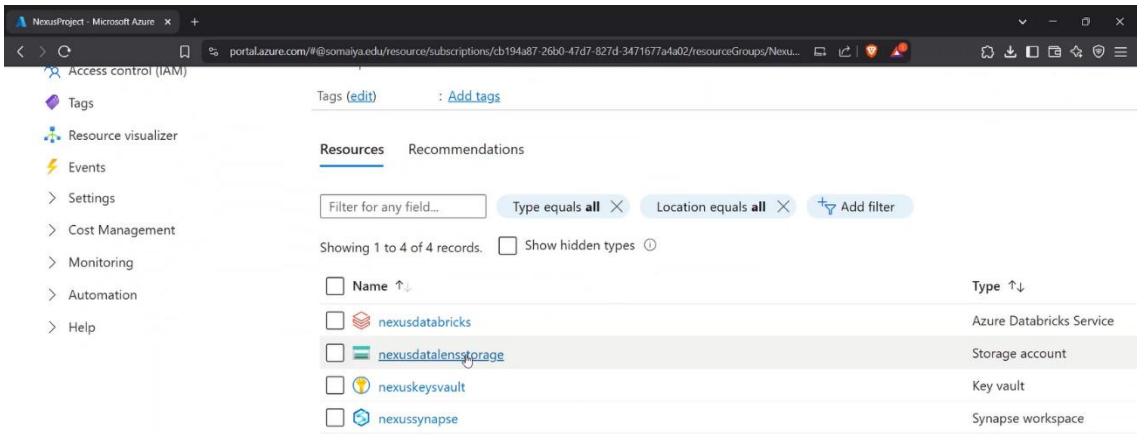
The "**NexusProject**" resource group in Azure contains four resources: a Databricks service, a storage account, a key vault, and a Synapse workspace, all located in Central India.



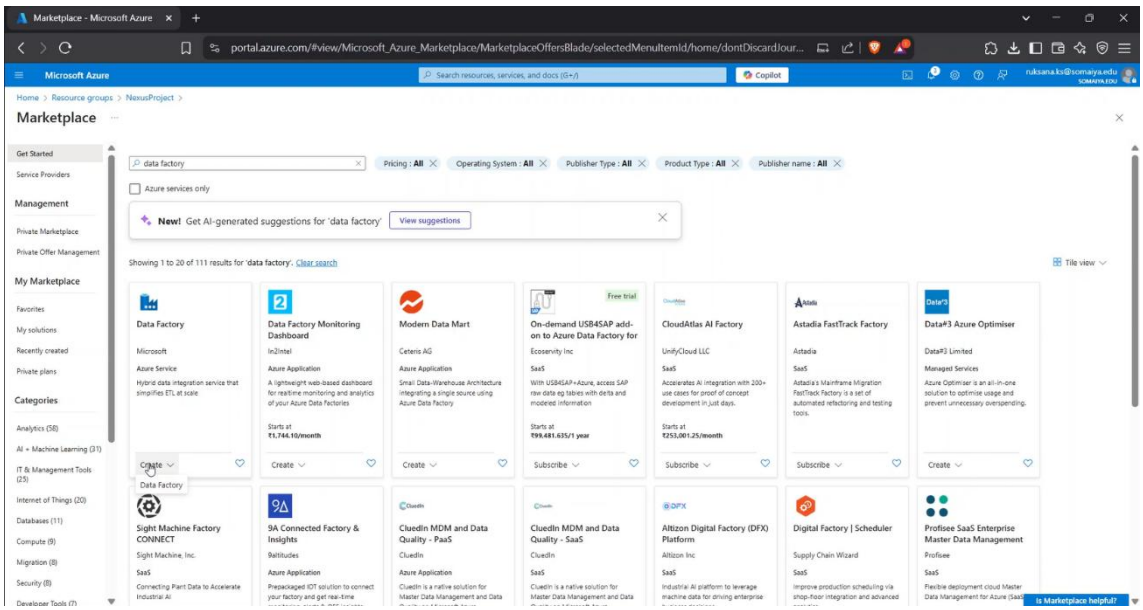
The Azure portal displaying four resources under the "**NexusProject**" resource group, including Databricks, a storage account, a key vault, and a Synapse workspace.



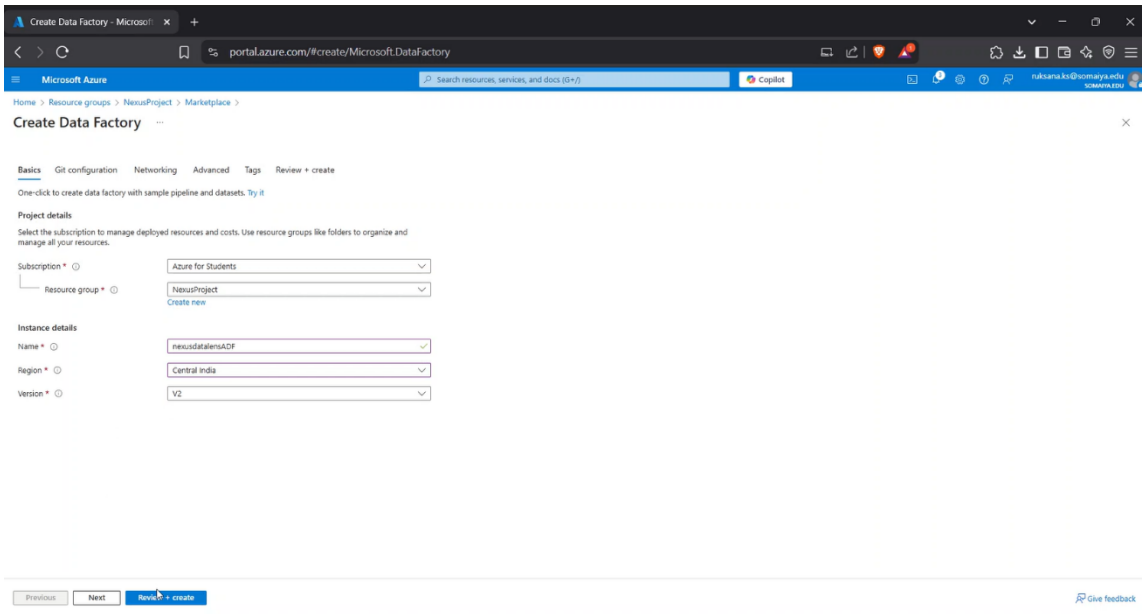
The Azure portal displays four resources in the "**NexusProject"** resource group: Databricks, a storage account, a key vault, and a Synapse workspace.



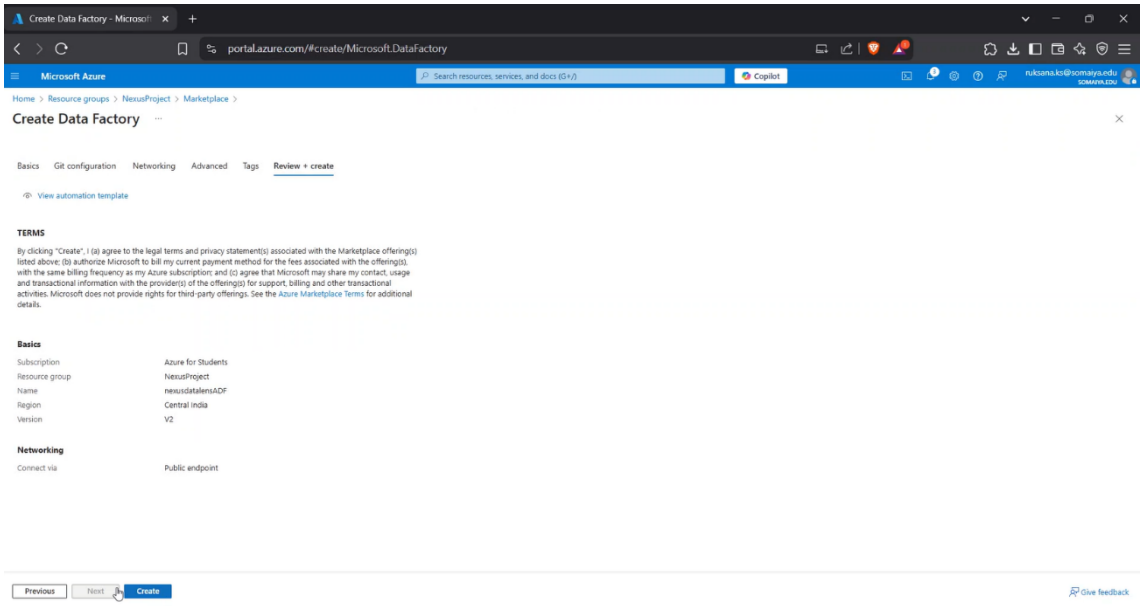
The Azure Marketplace with search results for "**Data Factory**" ,displaying various related services and solutions available for deployment.



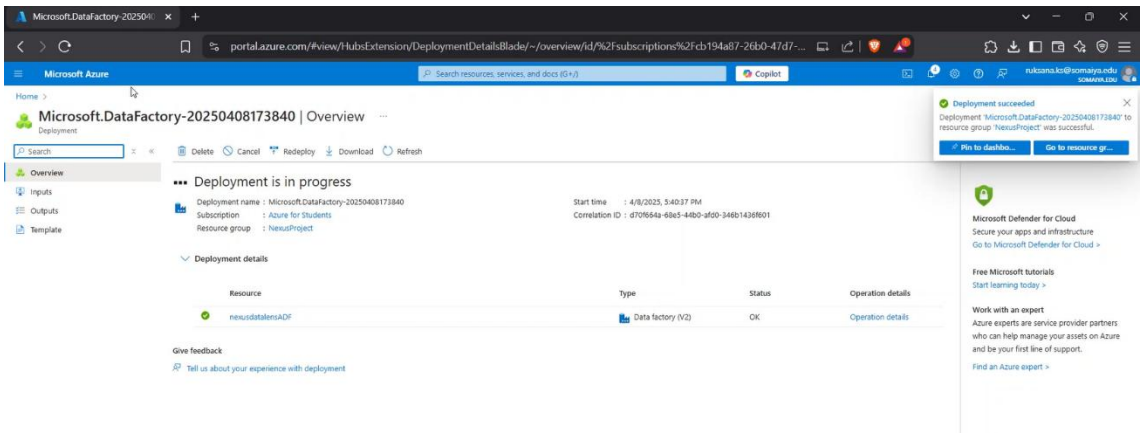
The "**Create Data Factory**" setup page in Azure, where the user is configuring a new Data Factory named "**nexusdatainsADF**" under the "**NexusProject**" resource group in the Central India region.



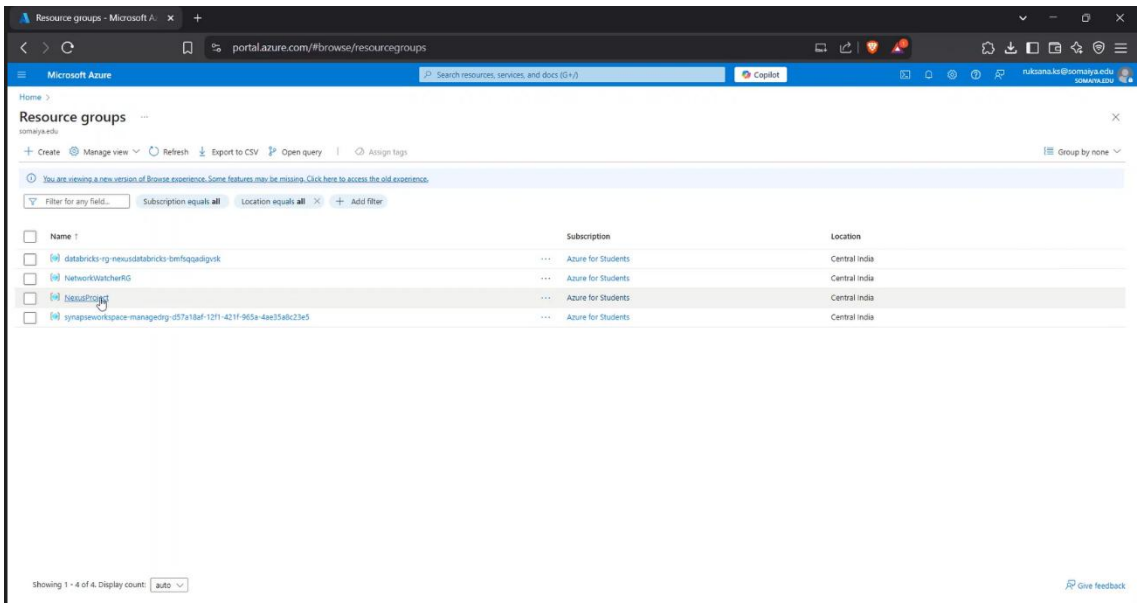
The image shows the final "**Review + create**" step in Azure for deploying a Data Factory instance named "**nexusdatainsADF**" under the "**NexusProject**" resource group using a public endpoint in the Central India region.



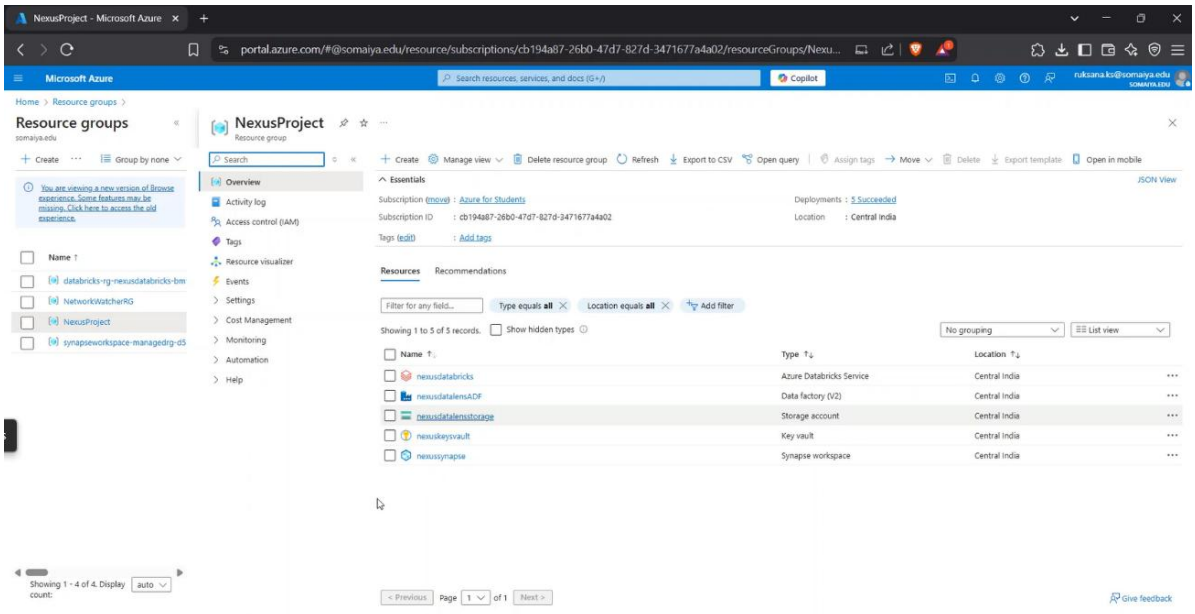
The image shows a successful deployment notification for the Data Factory resource "**nexusdatainsADF**" under the "**NexusProject**" resource group in Azure.



The image shows the Azure portal listing four resource groups under the "**Azure for Students**" subscription, all located in the Central India region.

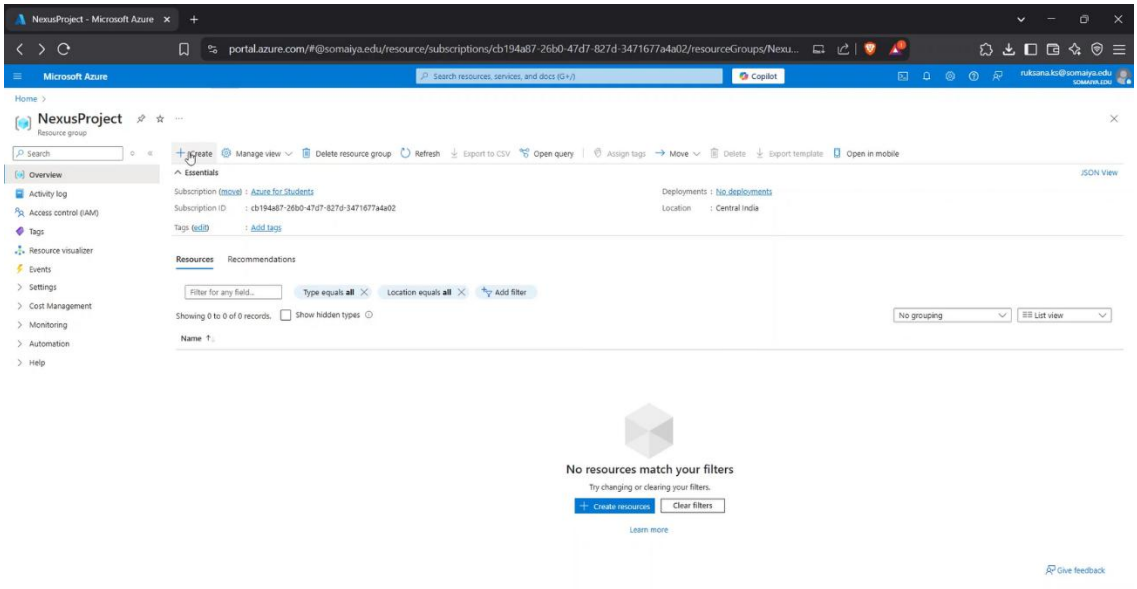


Complete setup of resource group

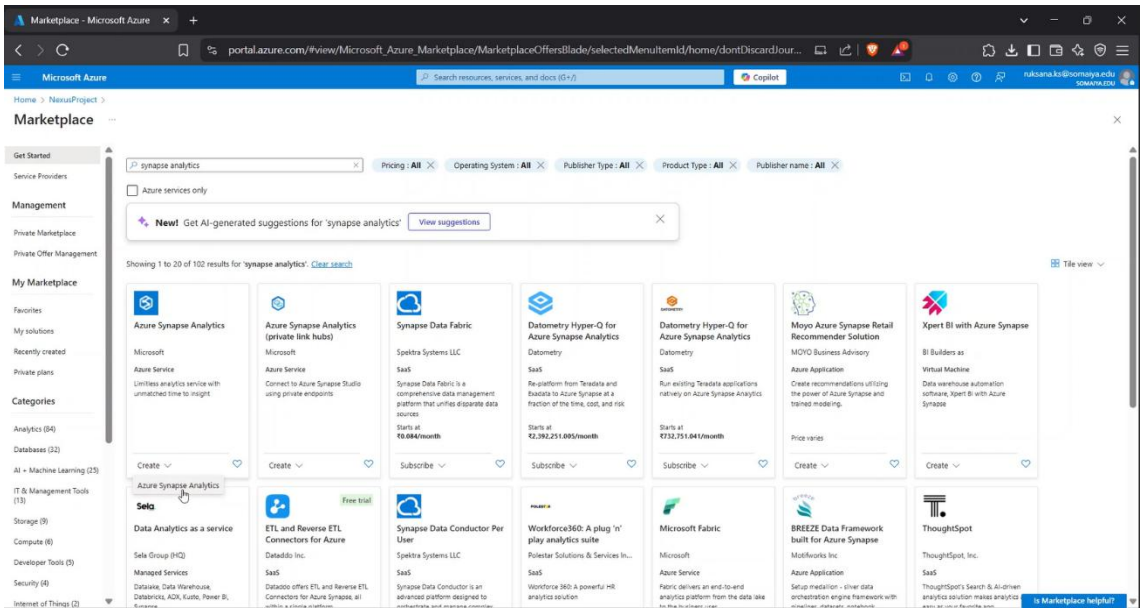


SETTING RESOUCE GROUP

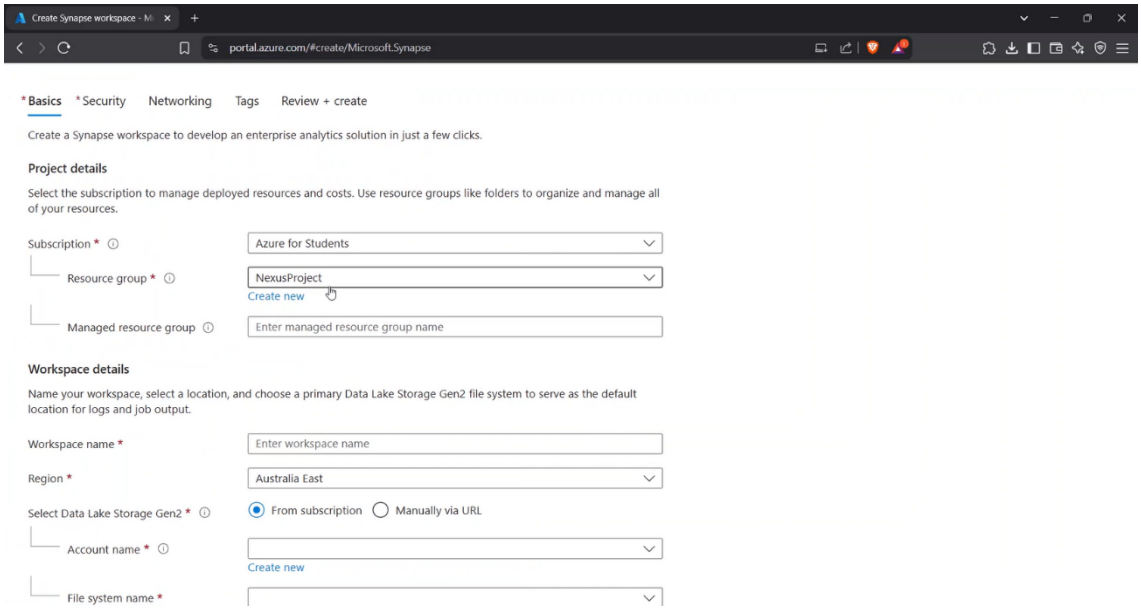
The "**NexusProject**" resource group currently has no visible resources due to active filters, with an option to create new resources or clear filters.



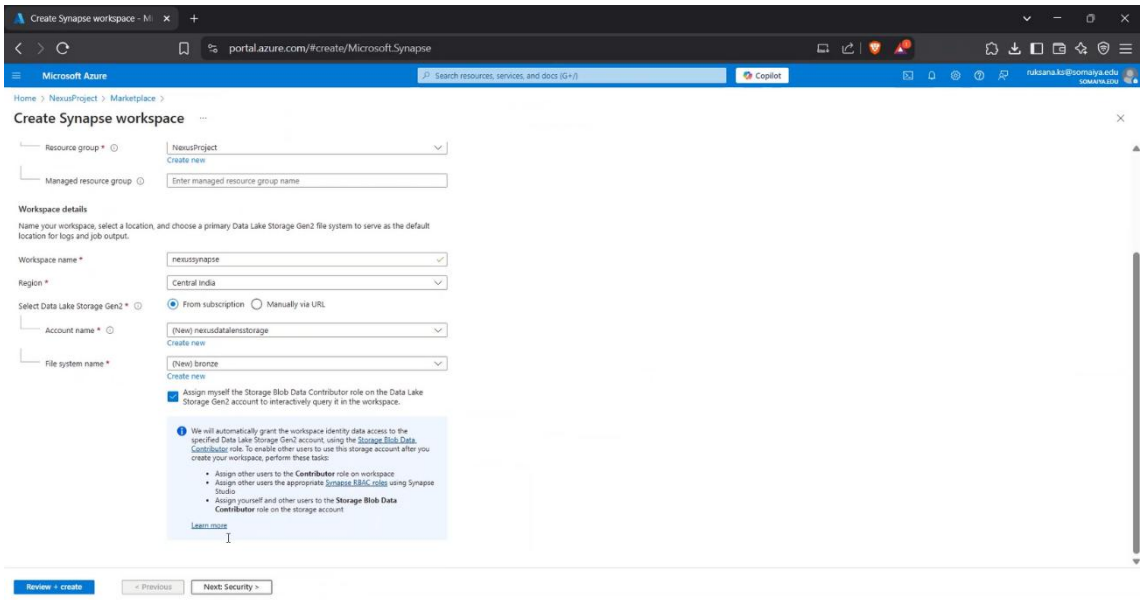
The Azure Marketplace displays various Synapse Analytics-related services and solutions available for creation or subscription within the "**NexusProject**" environment.



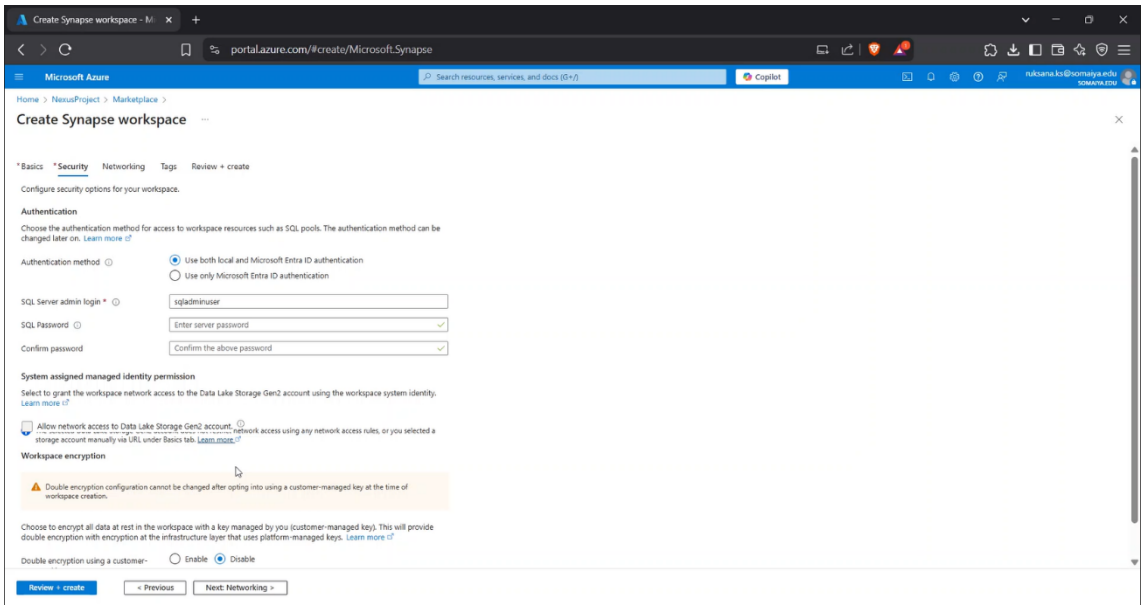
The form is being filled to create a new Synapse workspace under the **"NexusProject"** resource group with specified subscription, region, and linked Data Lake Storage Gen2 account.



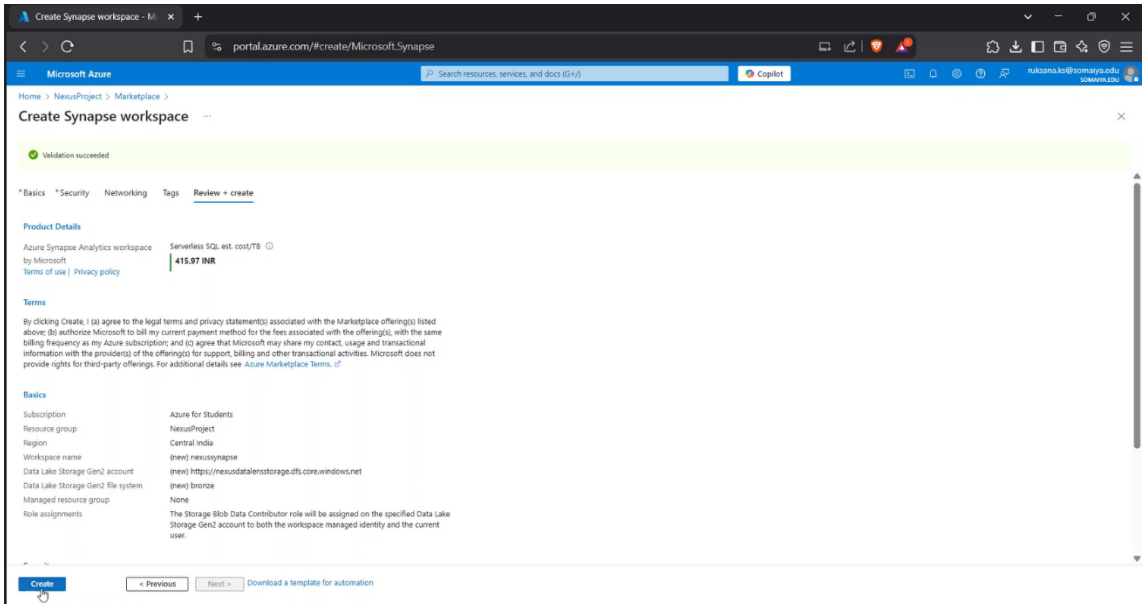
A Synapse workspace named "**nexussynapse**" is being created in Central India, linked to the "**nexusdatalensstorage**" Data Lake Storage with "**bronze**" as the file system.



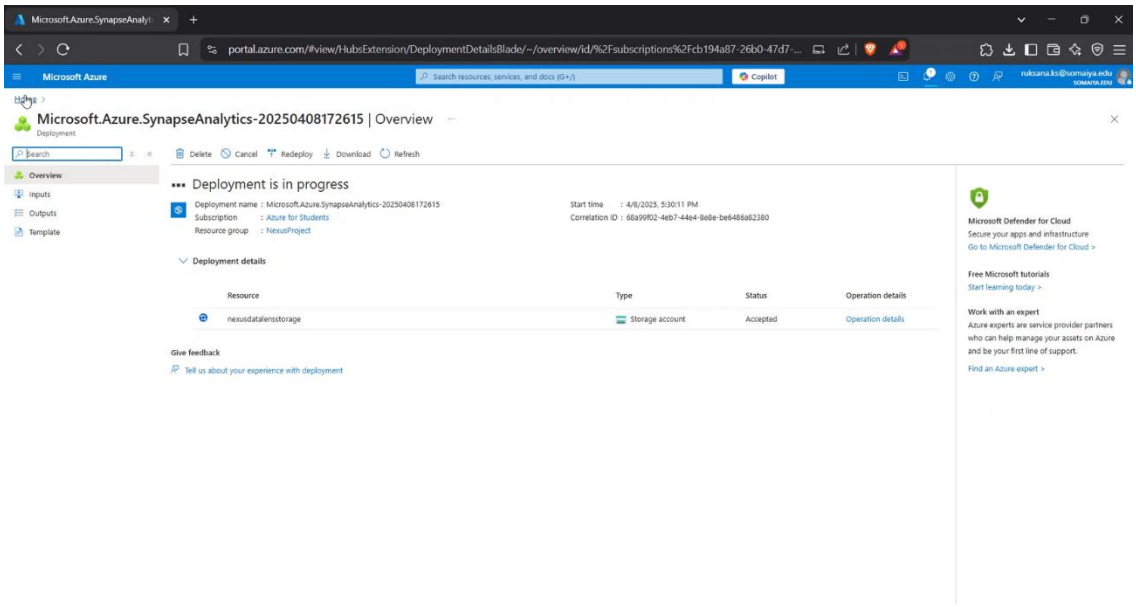
The Synapse workspace is being configured with SQL authentication, network access to Data Lake enabled, and default platform-managed encryption selected.



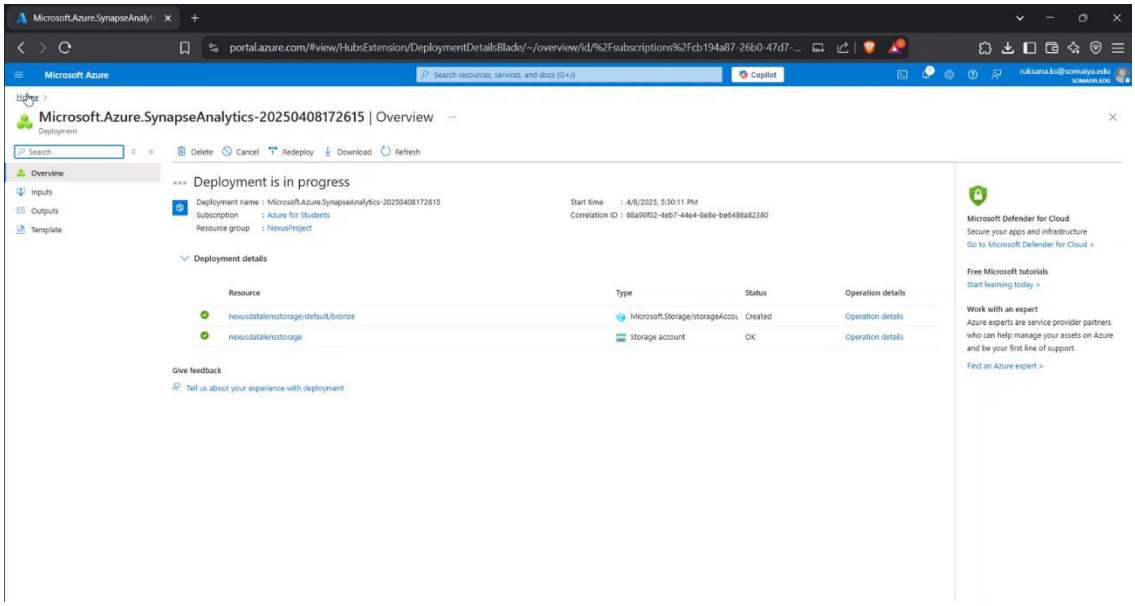
The **Synapse workspace setup review** has **succeeded**, and the user is ready to create it with specified configuration and pricing.



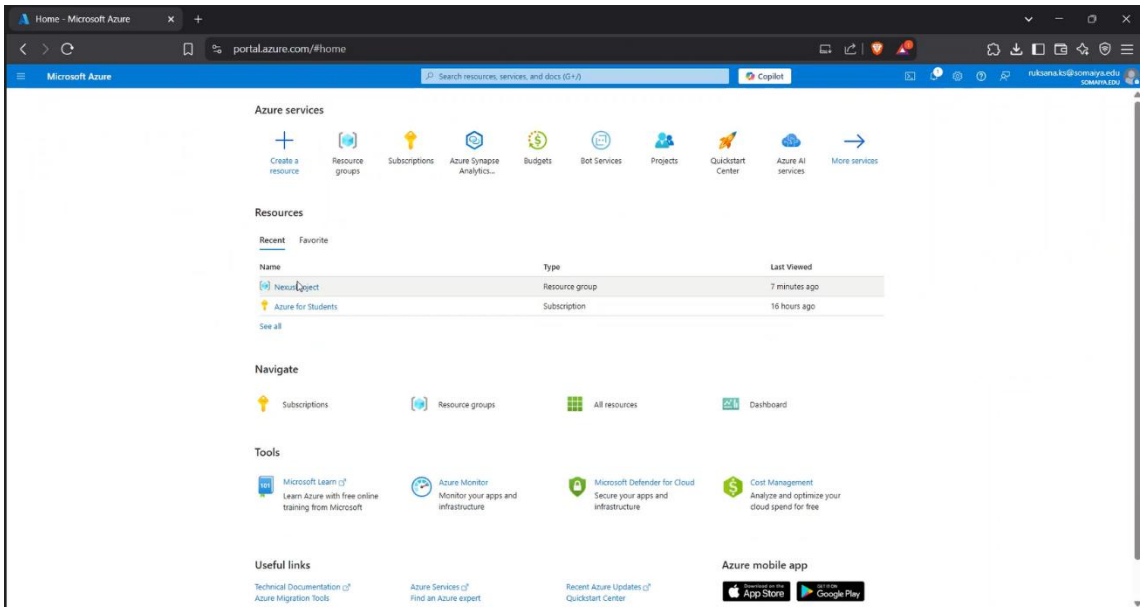
The Synapse workspace deployment is currently in progress with the storage account resource accepted under the **NexusProject resource group**.



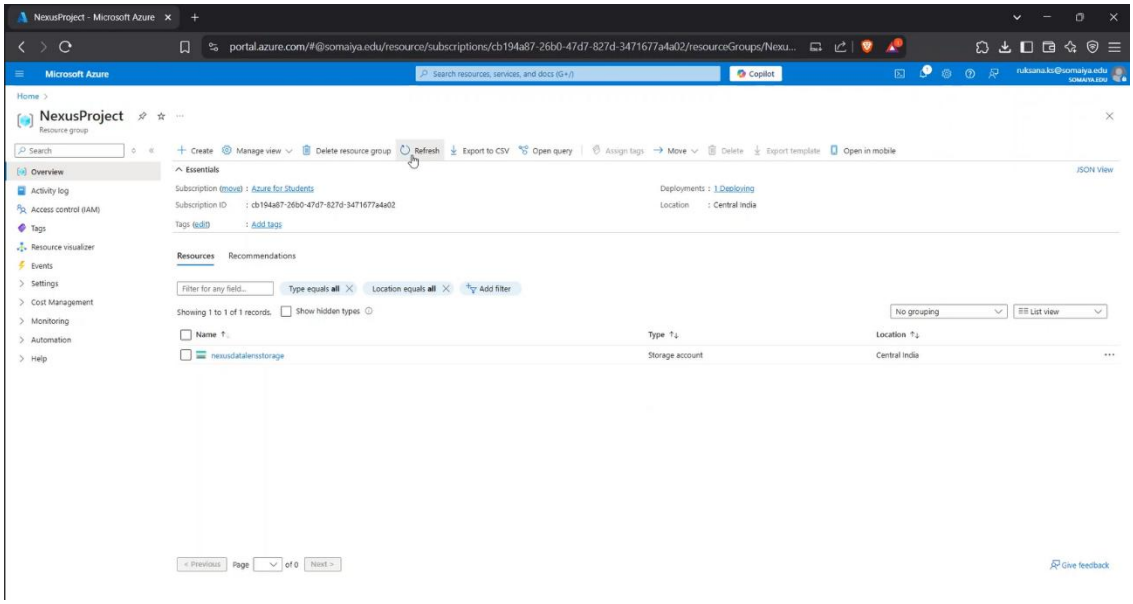
The Synapse workspace deployment is in **progress**, with the storage container bronze created and the storage account status marked as OK.



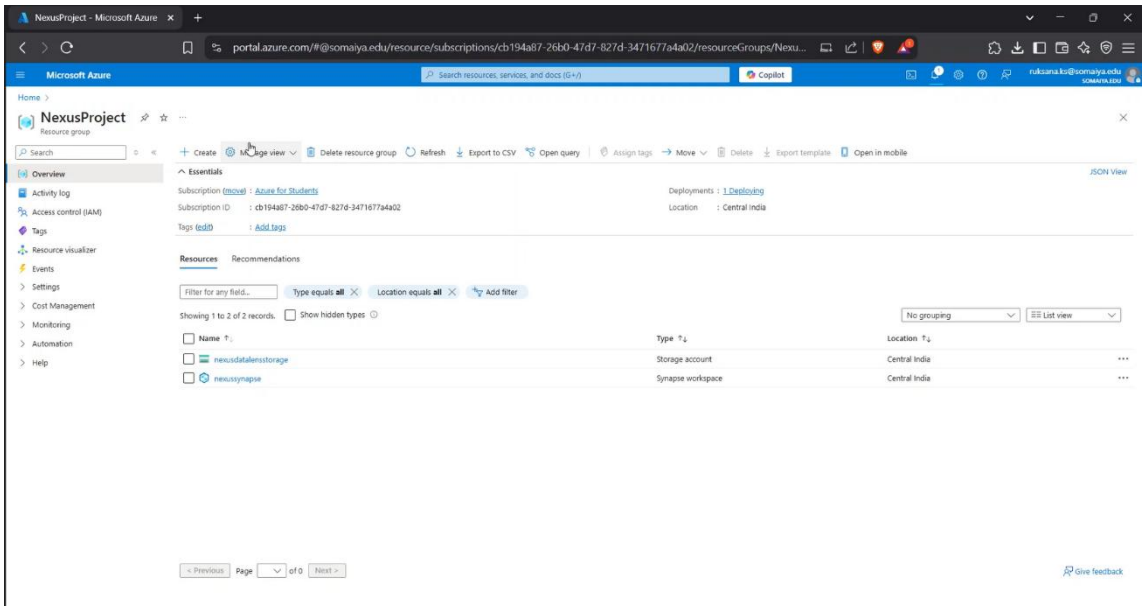
The Azure portal home page displays recent resources, with "**NexusProject" (a resource group)** being the most recently accessed.



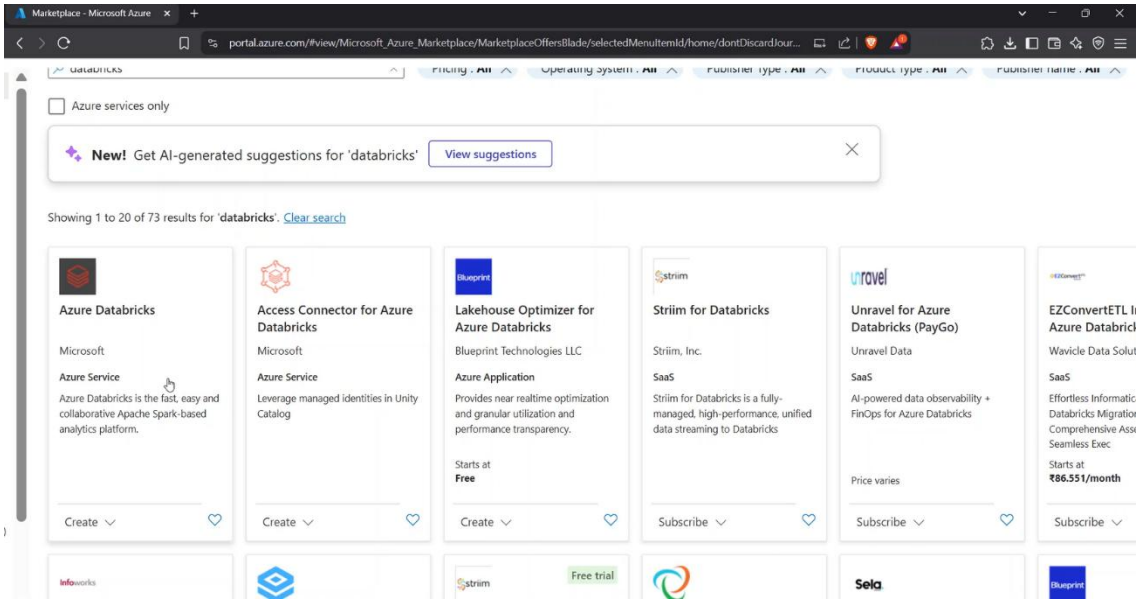
The image shows an Azure portal page displaying the "**NexusProject**" resource group with one storage account resource named "**nexusdatastenstorage**" located in Central India.



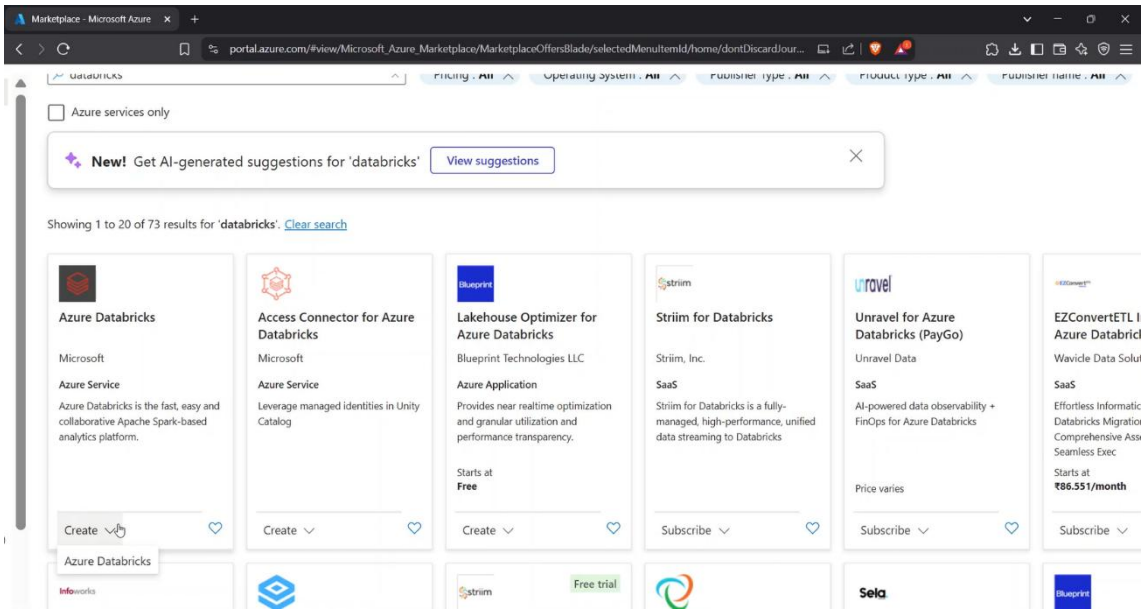
The updated Azure portal view shows the "**NexusProject**" resource group now containing two resources: a storage account ("**nexusdatastenstorage**") and a Synapse workspace ("**nexussynapse**"), both located in Central India.



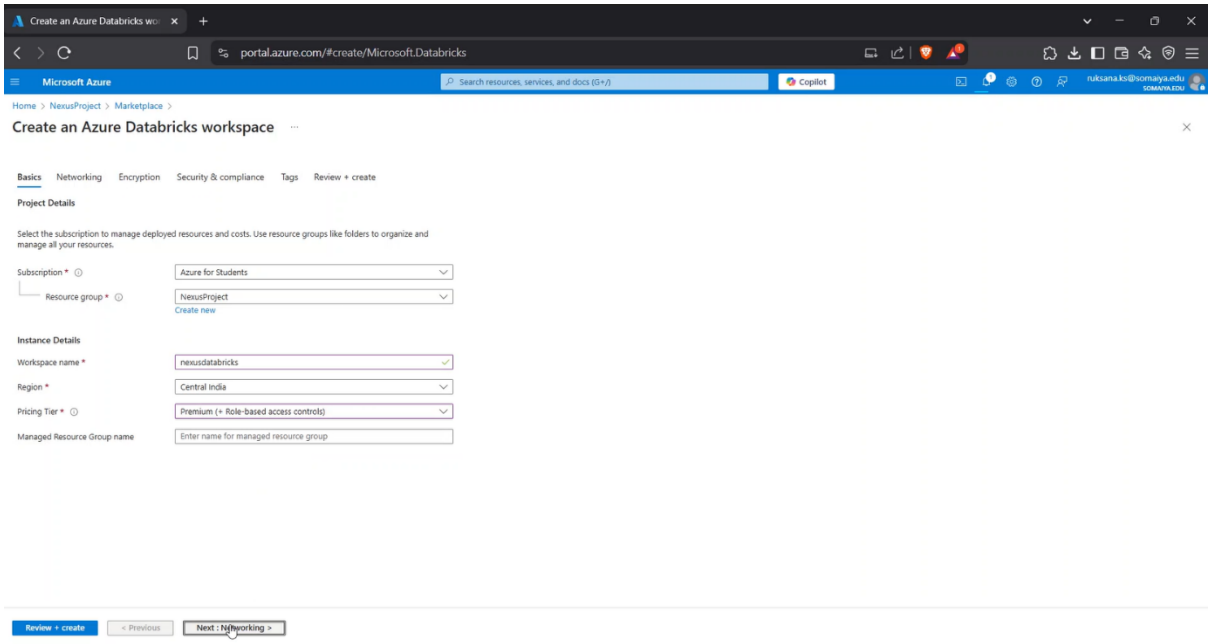
The Azure Marketplace displays various Databricks-related services, including "**Azure Databricks**" by Microsoft for Spark-based analytics and other integration and optimization tools.



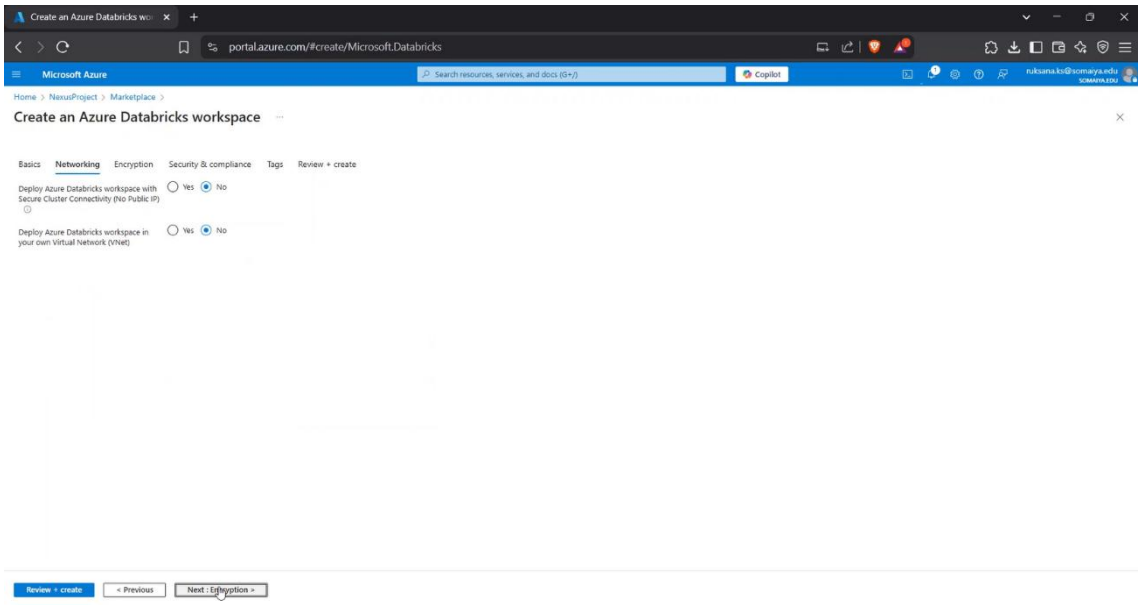
The user is about to create an "**Azure Databricks**" workspace from the Azure Marketplace by clicking the "**Create**" button under the Azure Databricks service.



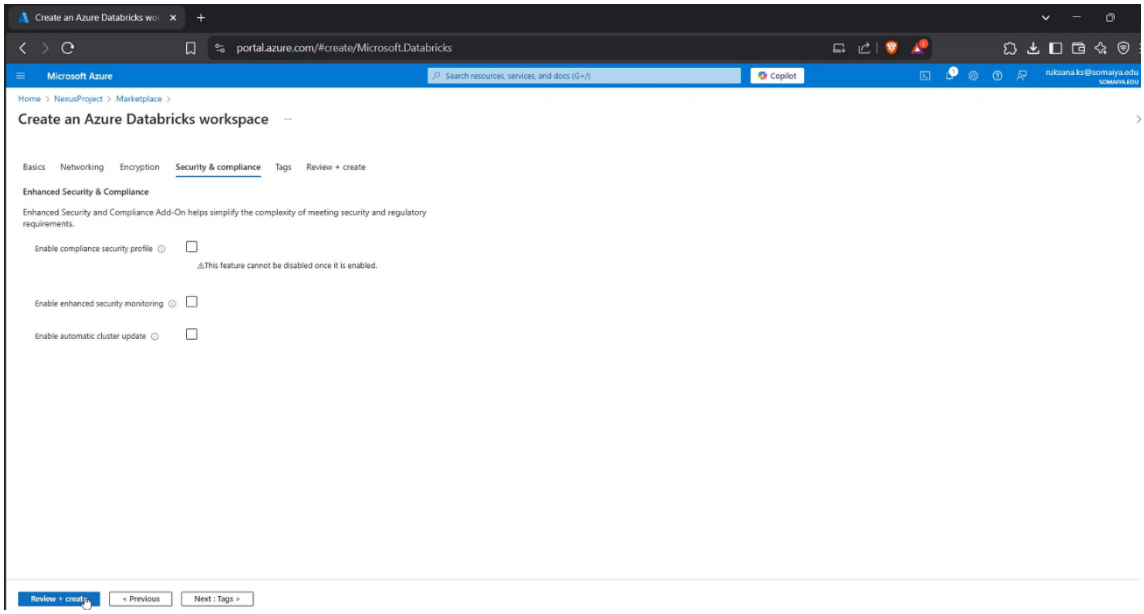
The user is configuring the basic settings to create an Azure Databricks workspace named "**nexusdatabricks**" in the "**NexusProject**" resource group within the Central India region.



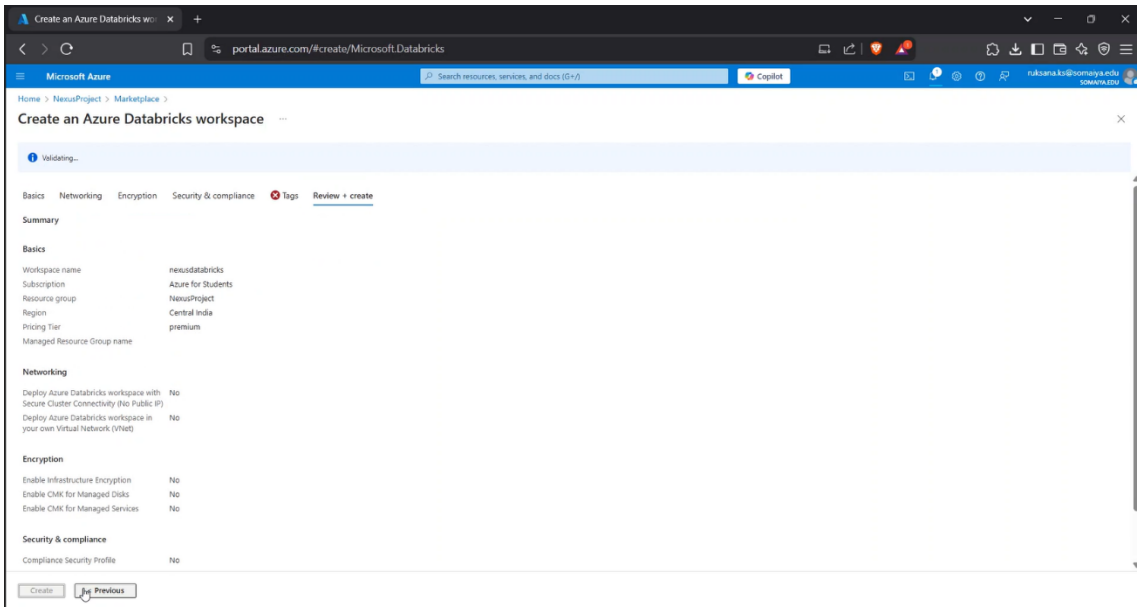
The user is configuring networking options for the Azure Databricks workspace, choosing whether to enable secure cluster connectivity and virtual network deployment.



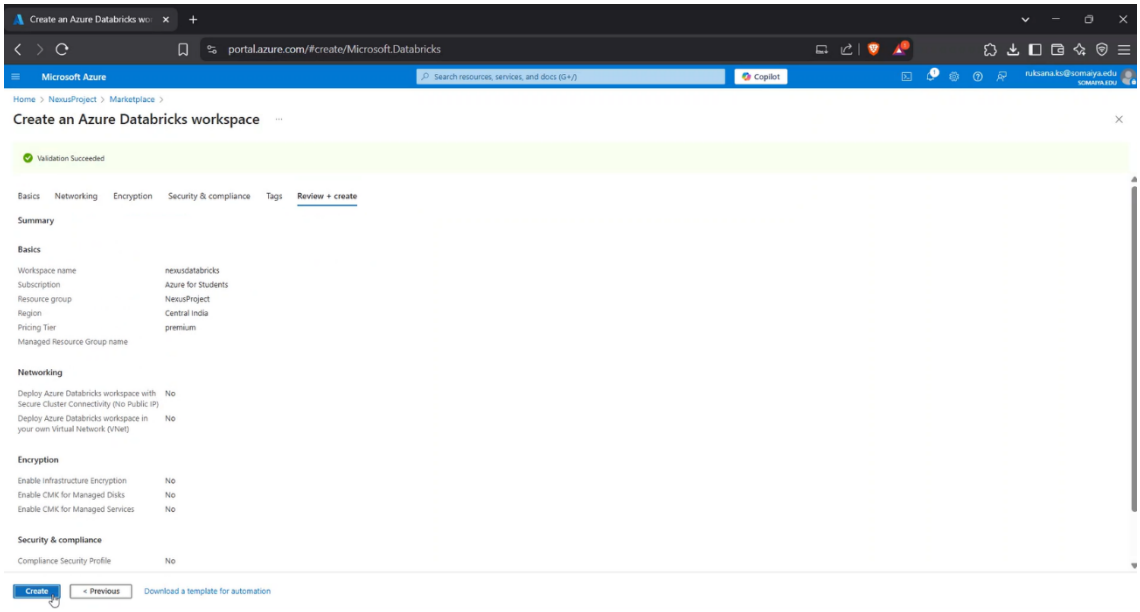
The user is on the Security & Compliance tab while creating an Azure Databricks workspace, with optional settings for compliance, monitoring, and automatic cluster updates before proceeding to review and create.



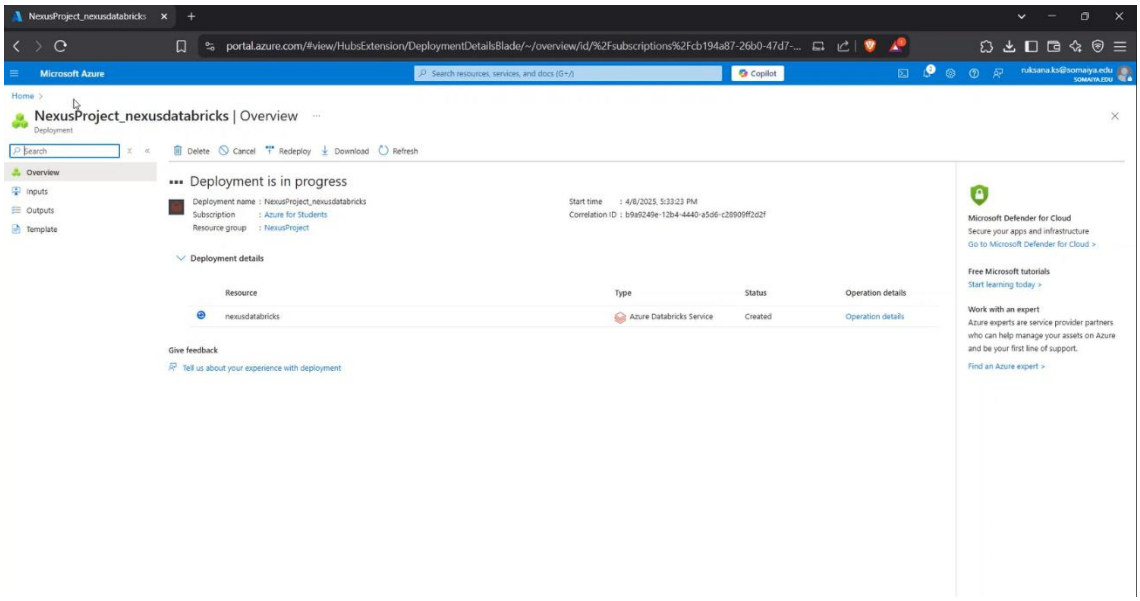
The user is on the final "**Review + create**" step for setting up an Azure Databricks workspace, reviewing all configuration details before deployment.



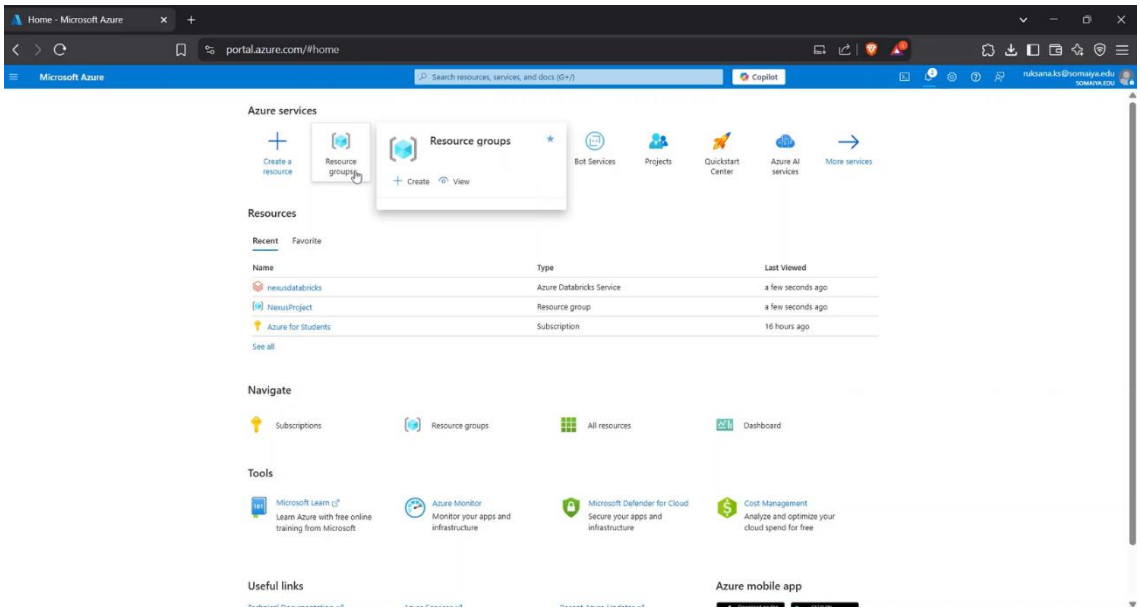
The user has successfully validated the configuration and is ready to create the **Azure Databricks workspace** by clicking the "**Create**" button.



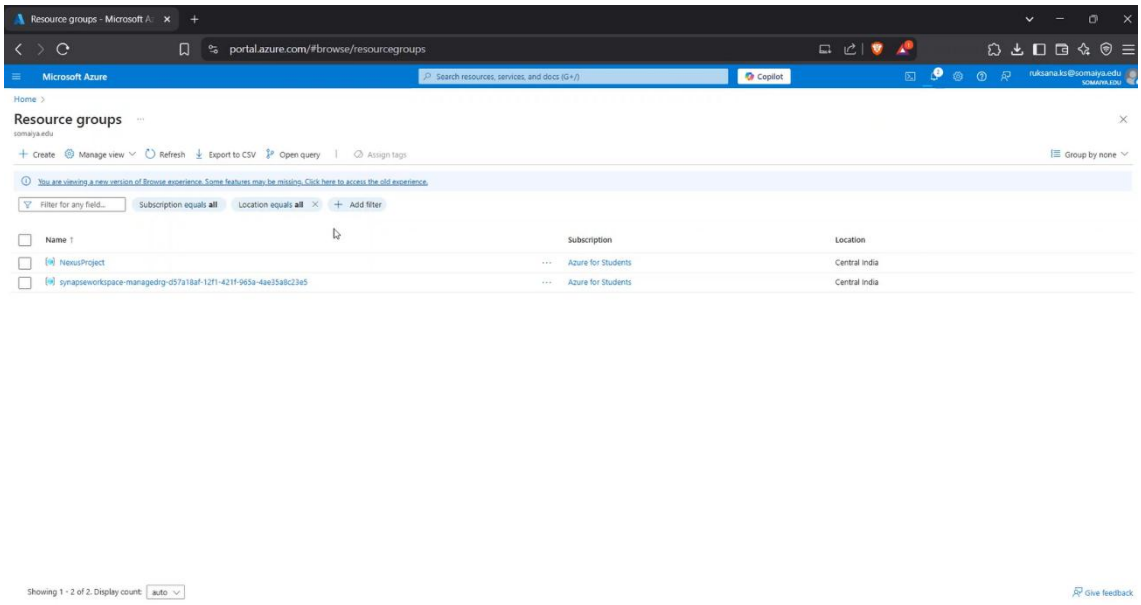
The deployment of the Azure Databricks workspace named "**nexusdatabricks**" is currently in progress within the "**NexusProject**" resource group.



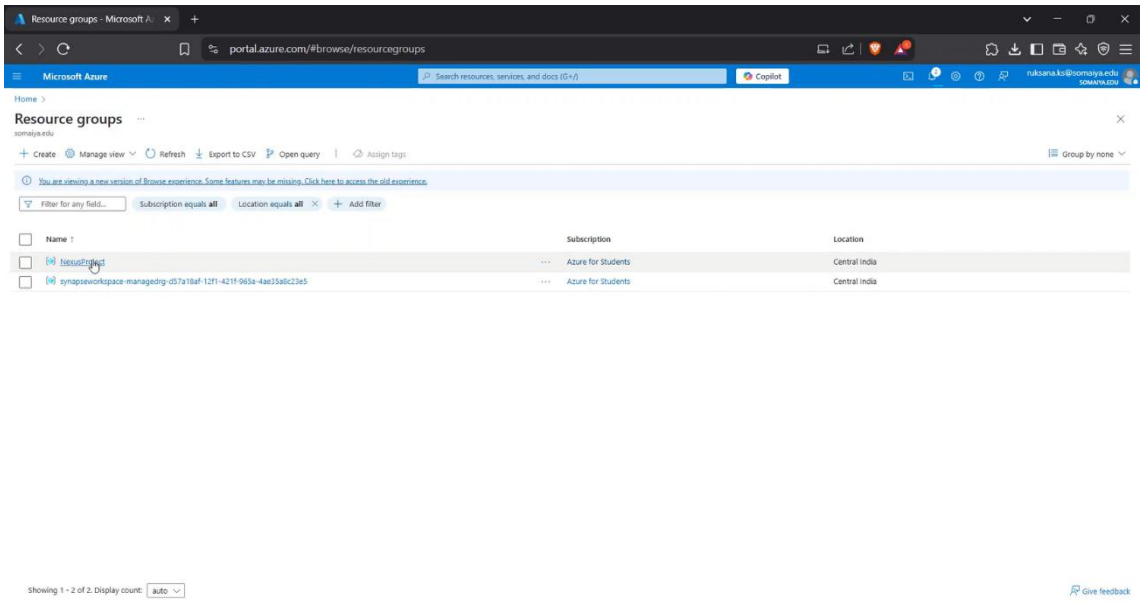
On the Azure portal home screen click on the "**Resource groups**" option to view or **create a new resource group**.



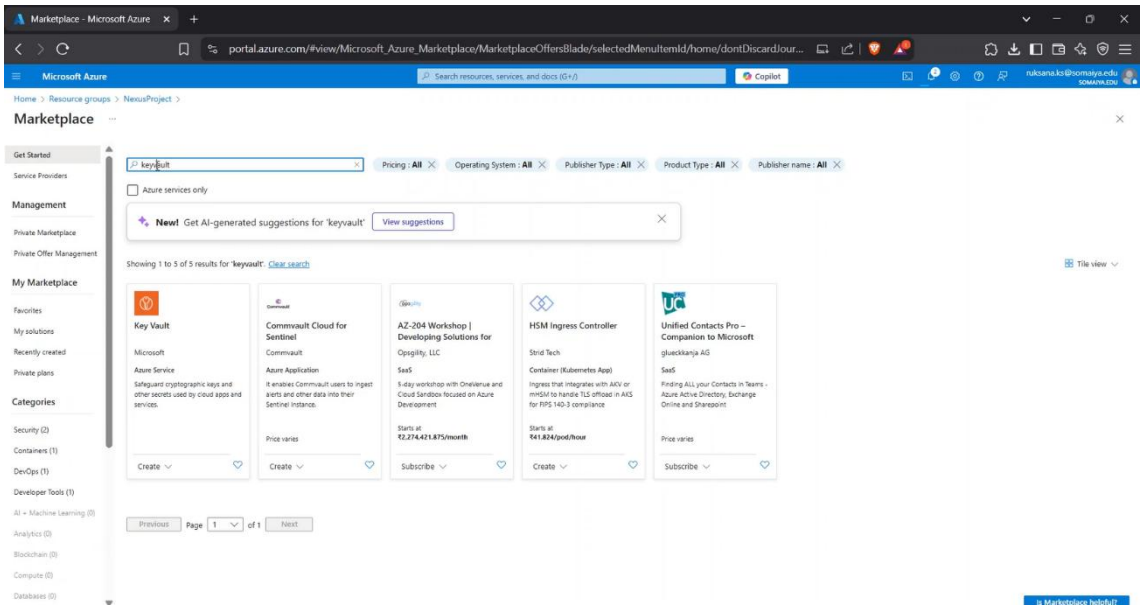
The "**Resource groups**" page in the Azure portal displays existing resource groups under the "**Azure for Students**" subscription, with both located in Central India.



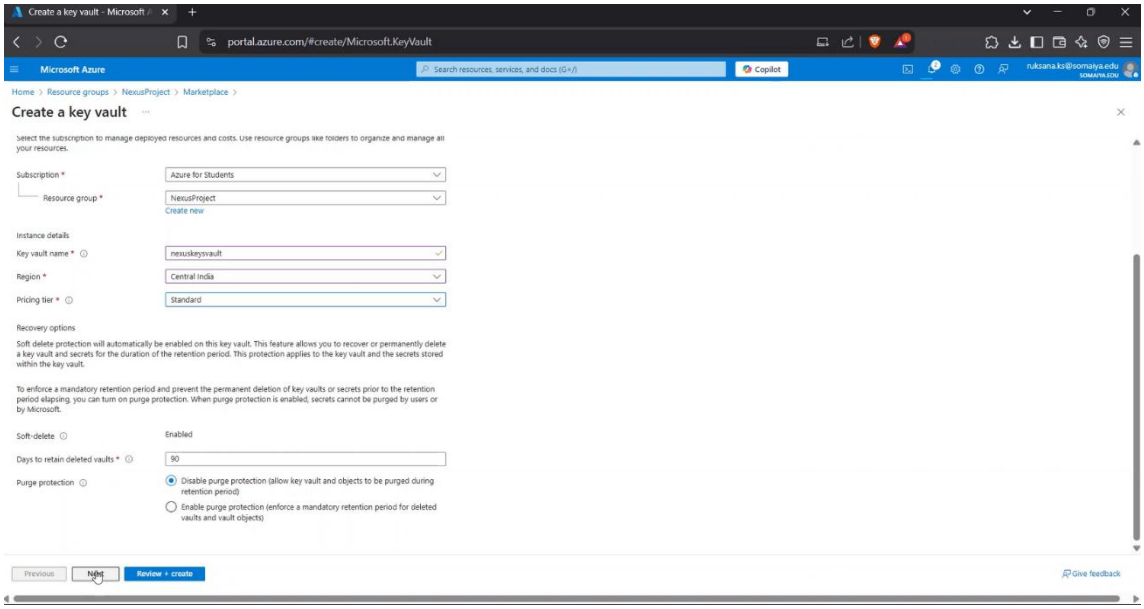
The user selects the "**NexusProject**" resource group from the list of available groups under the "**Azure for Students**" subscription in the Central India region.



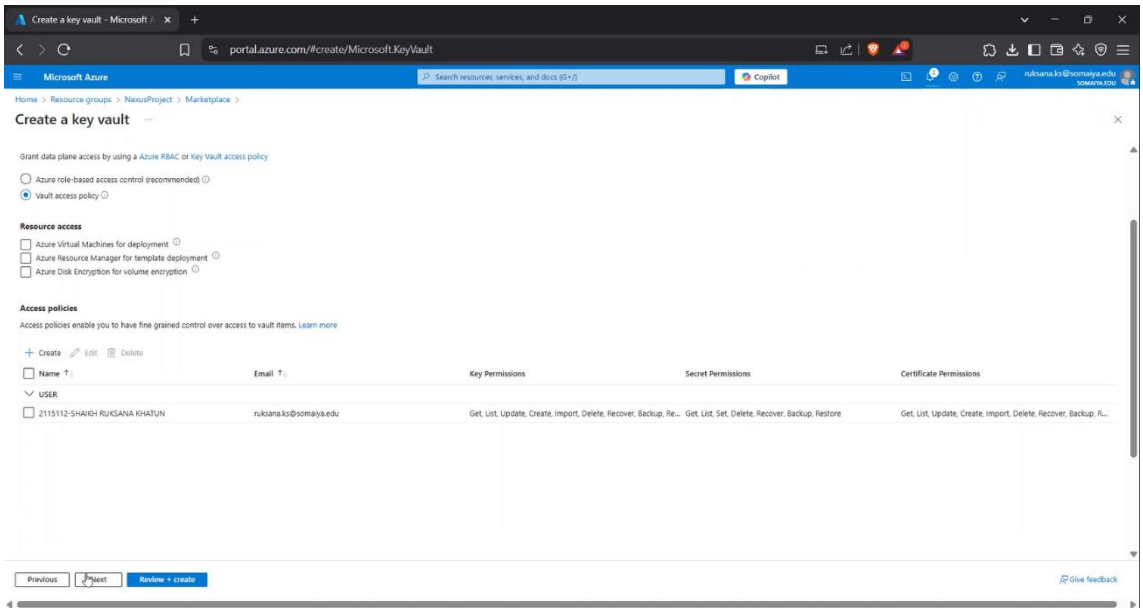
The user searches for "**Key Vault**" in the Azure Marketplace under the NexusProject resource group to **create a new key vault service**.



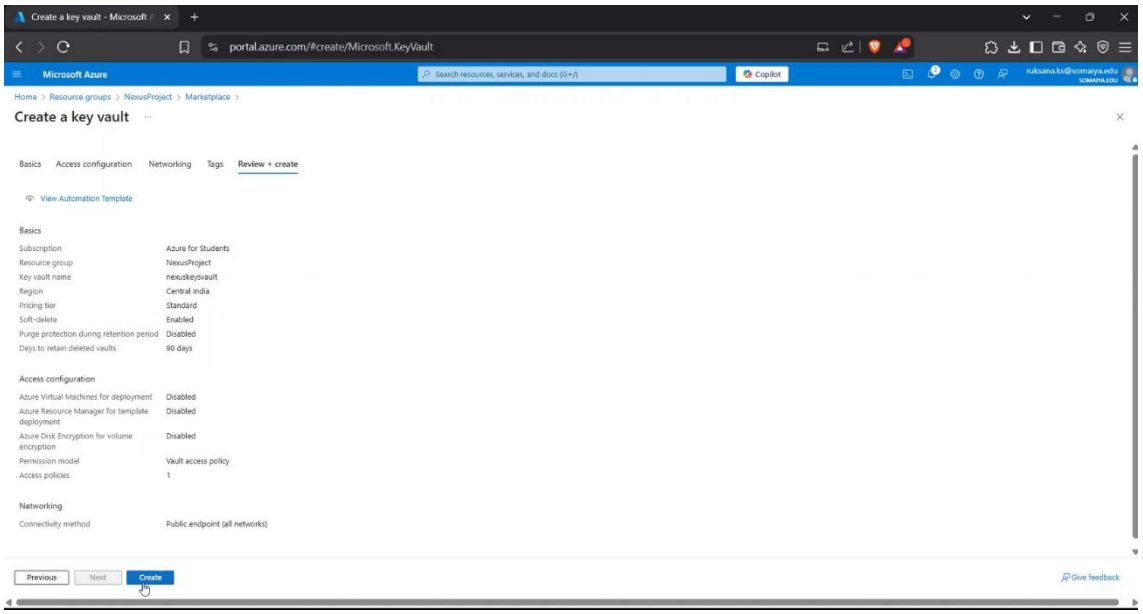
The user is configuring and creating a new Azure Key Vault named "**nexuskeyvault**" under the "**NexusProject**" resource group in the Central India region.



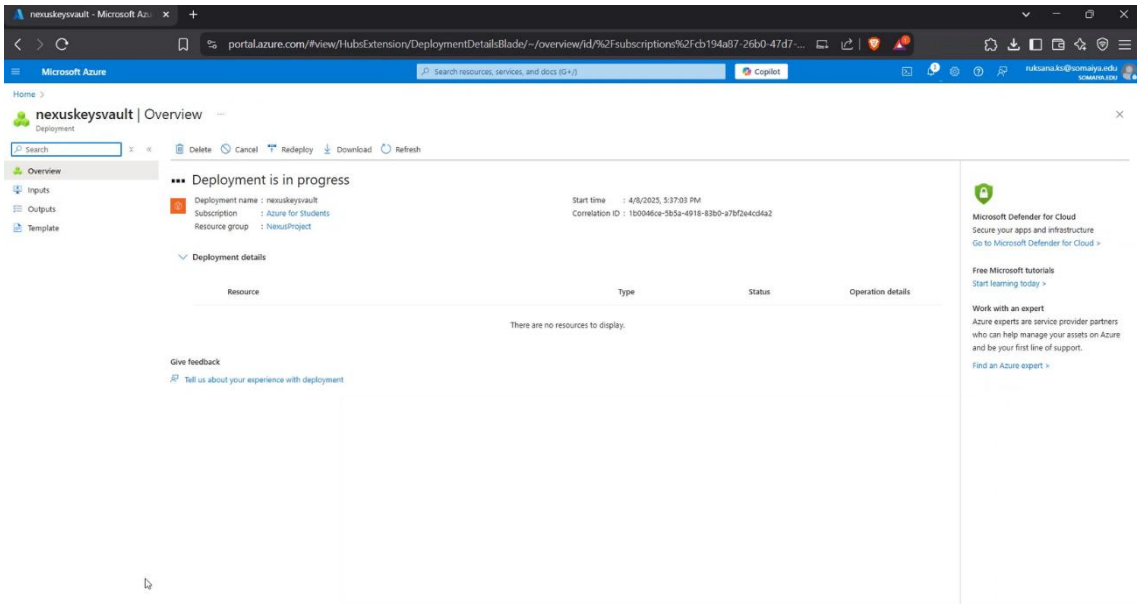
The user is assigning full key, secret, and certificate permissions to a specified user while setting access policies for the new Azure Key Vault before final review and creation.



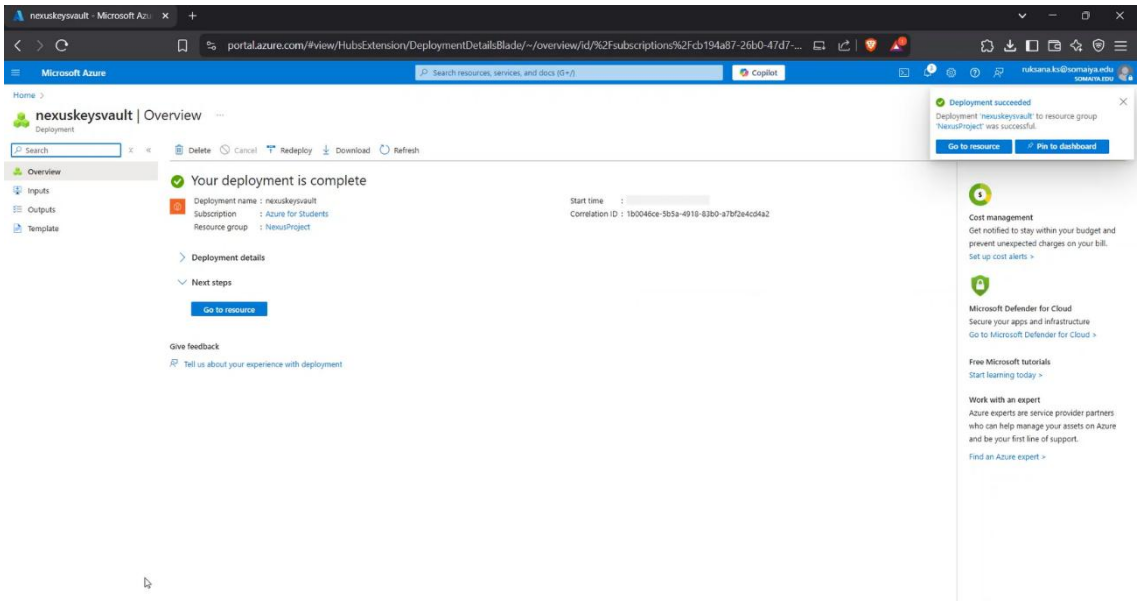
The user is finalizing the configuration and clicking "**Create**" to deploy a new Azure Key Vault with specified settings and access policies.



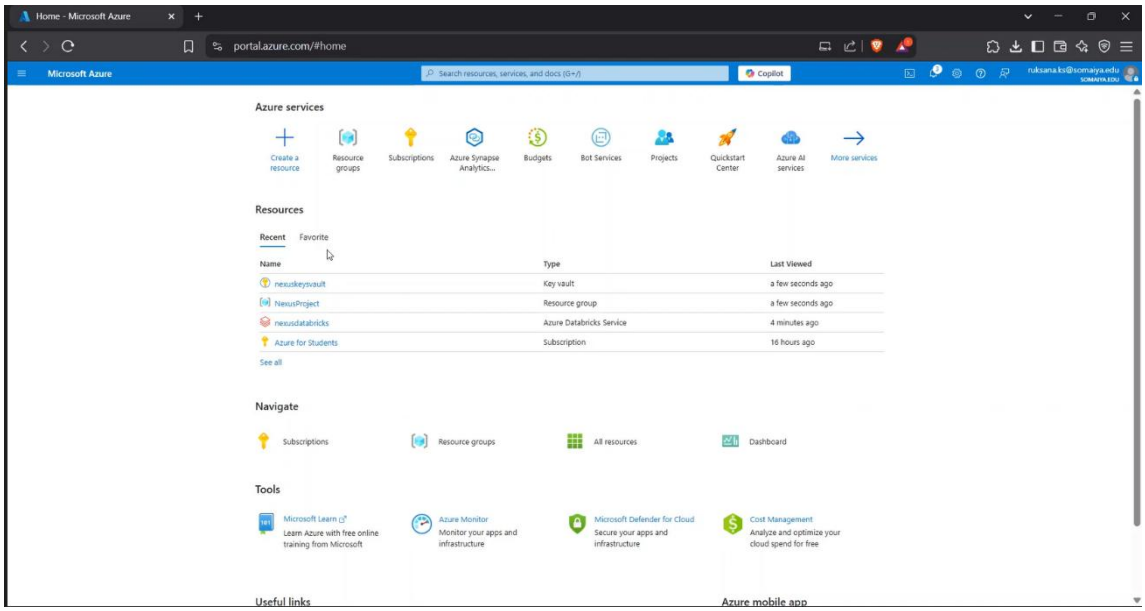
The deployment of the Azure Key Vault named "**nexuskeyvault**" is currently in progress within the "**NexusProject**" resource group.



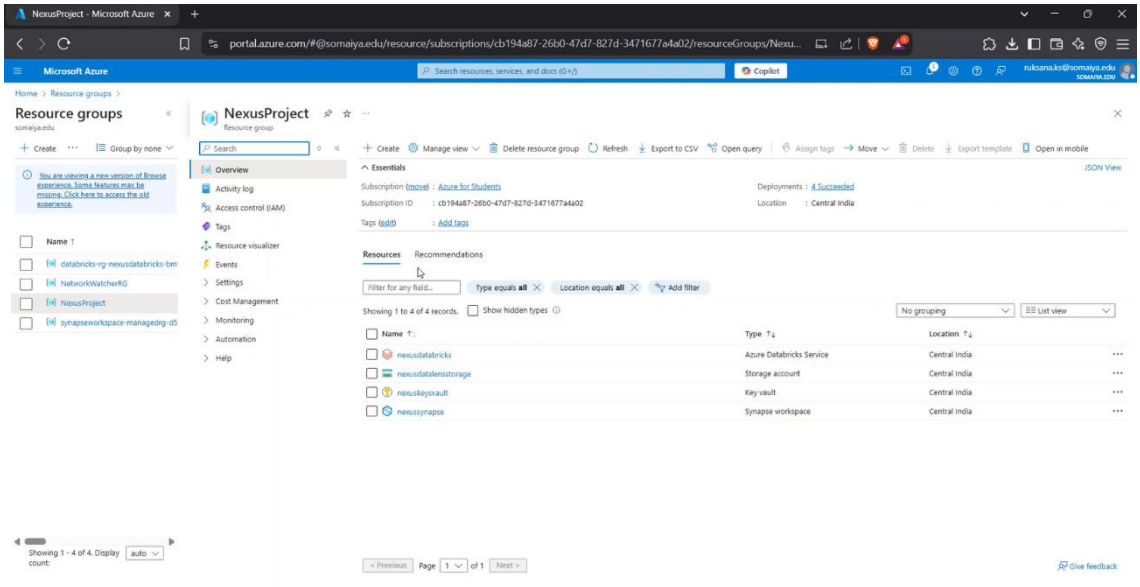
The deployment of the "**nexuskeyvault**" in the "**NexusProject**" resource group has completed successfully



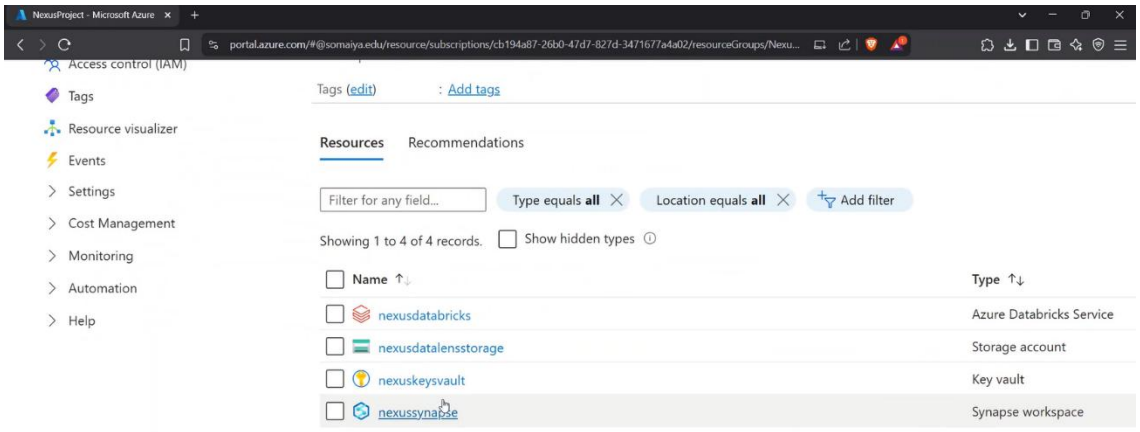
The Azure portal home screen shows the recently accessed resources, including the newly created Key Vault "**nexuskeyvault**" under the "**NexusProject**" resource group.



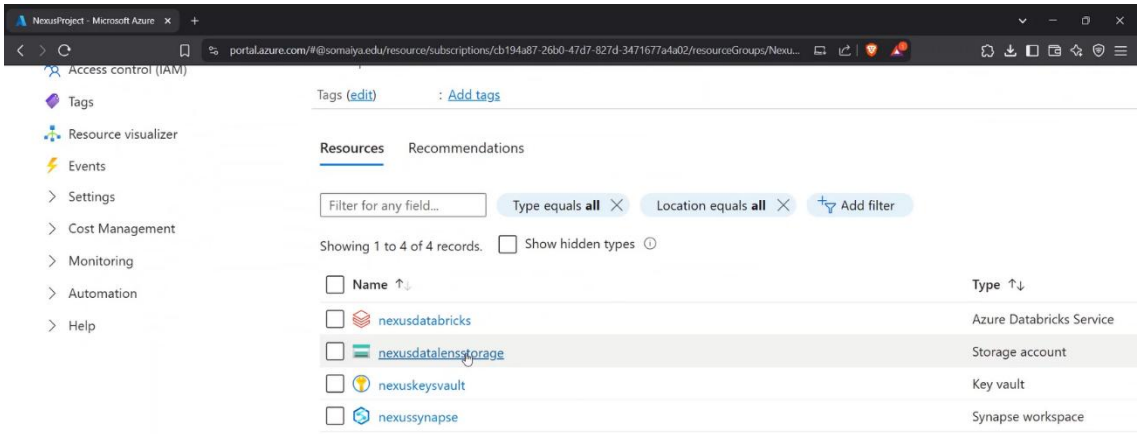
The "**NexusProject**" resource group in Azure contains four resources: a Databricks service, a storage account, a key vault, and a Synapse workspace, all located in Central India.



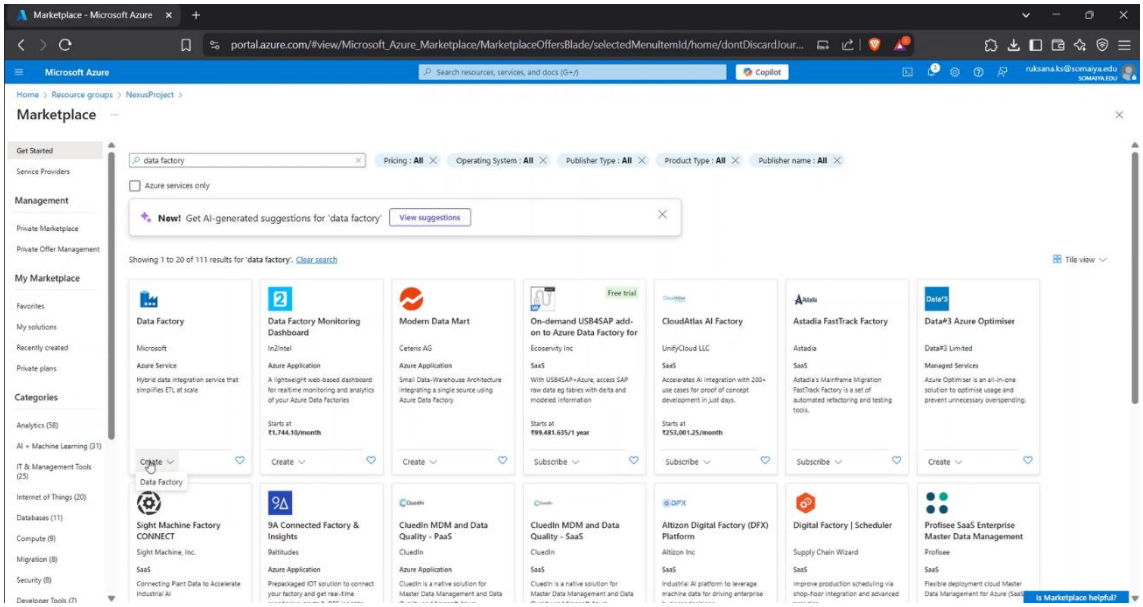
The Azure portal displaying four resources under the "**NexusProject**" resource group, including Databricks, a storage account, a key vault, and a Synapse workspace.



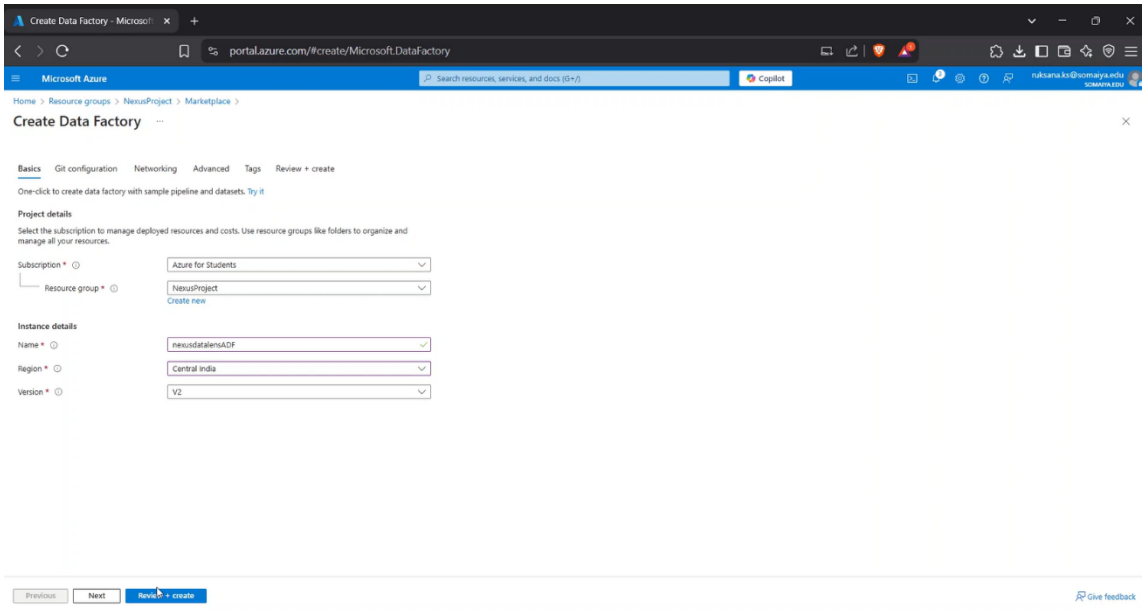
The Azure portal displays four resources in the "**NexusProject"** resource group: Databricks, a storage account, a key vault, and a Synapse workspace.



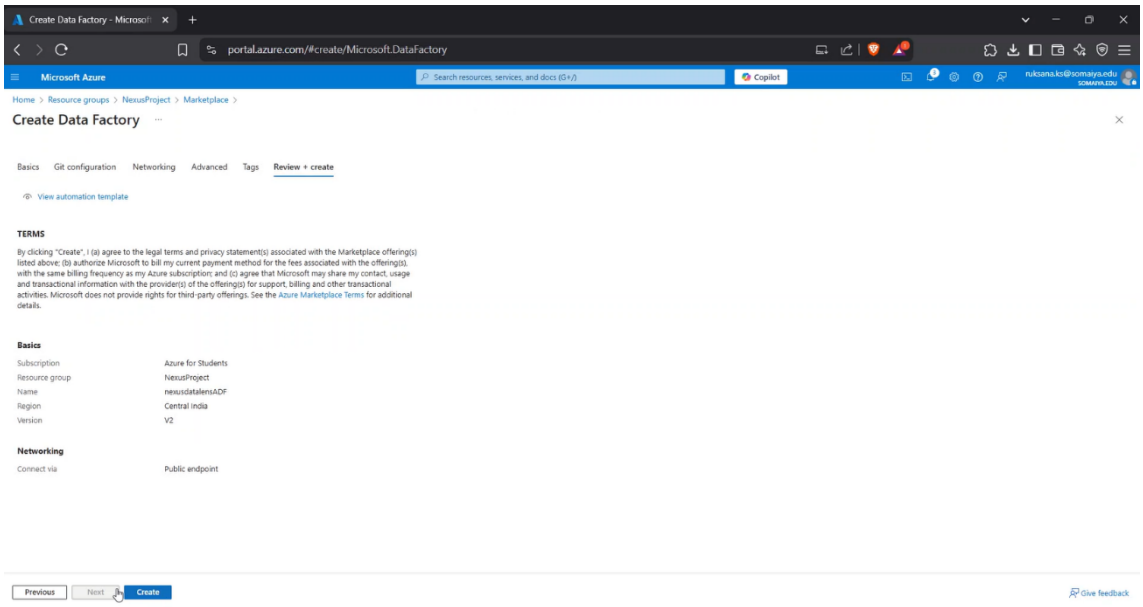
The Azure Marketplace with search results for "**Data Factory**" ,displaying various related services and solutions available for deployment.



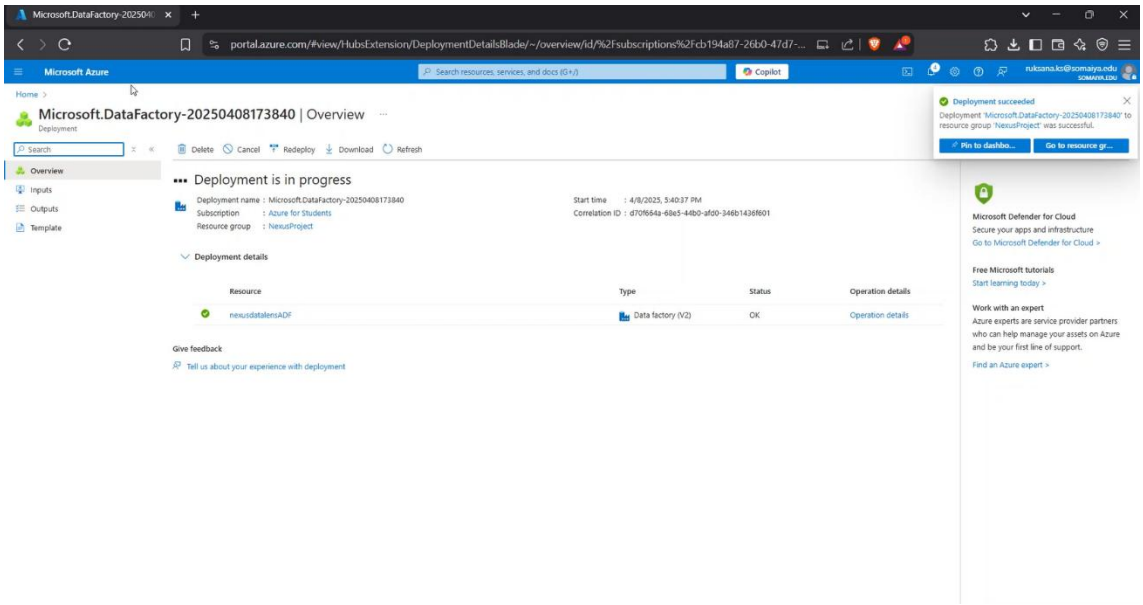
The "**Create Data Factory**" setup page in Azure, where the user is configuring a new Data Factory named "**nexusdatainsADF**" under the "**NexusProject**" resource group in the Central India region.



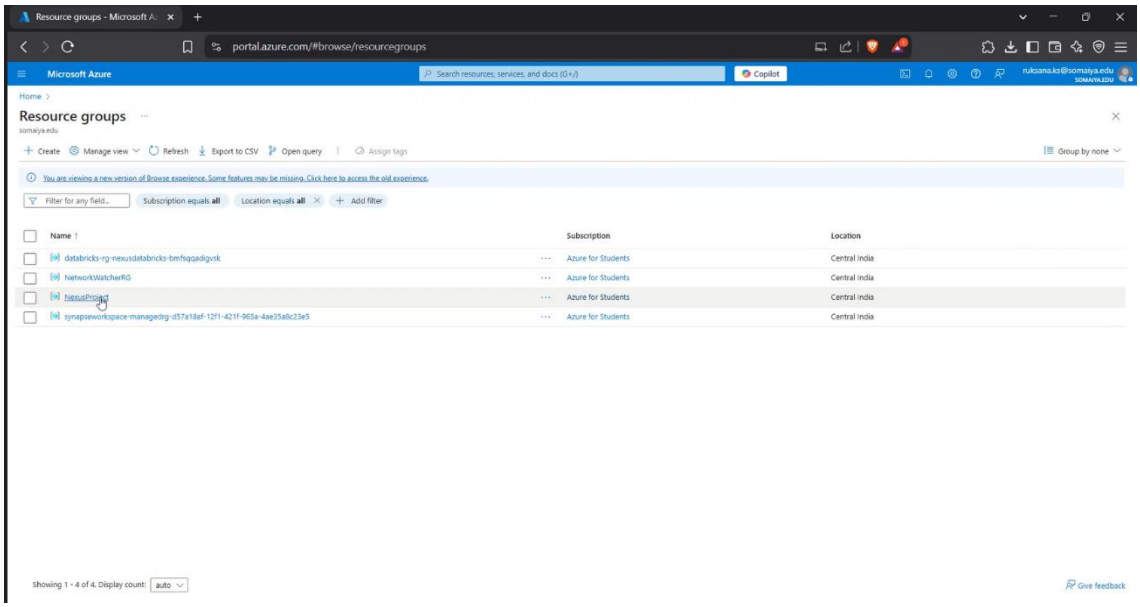
The image shows the final "**Review + create**" step in Azure for deploying a Data Factory instance named "**nexusdatainsADF**" under the "**NexusProject**" resource group using a public endpoint in the Central India region.



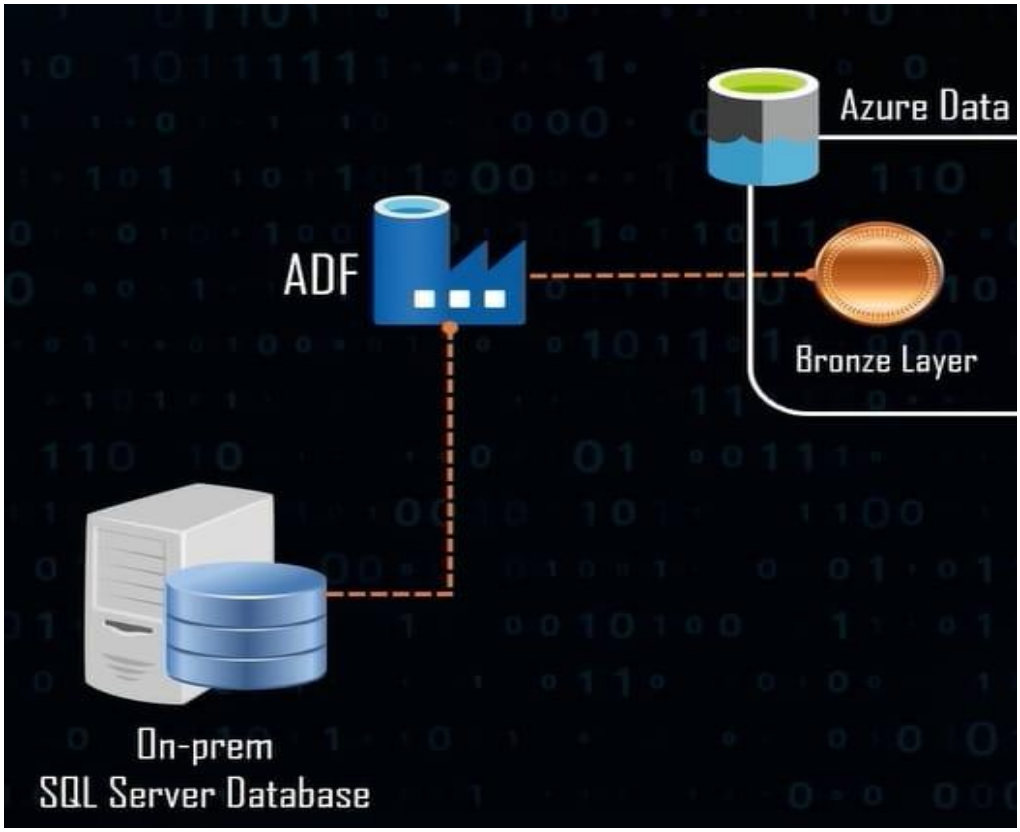
The image shows a successful deployment notification for the Data Factory resource "**nexusdatainsADF**" under the "**NexusProject**" resource group in Azure.



The image shows the Azure portal listing four resource groups under the "**Azure for Students**" subscription, all located in the Central India region.



DATA INJECTION - Using SQL Server Management Studio (SSMS)



Open SQL Server Management Studio (SSMS)

Open Azure Portal

- Go to https://portal.azure.com.
- Sign in with your Azure credentials.

Navigate to Azure Data Factory

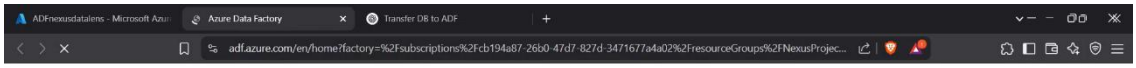
- In the Azure portal search bar at the top, type Data Factories.
- Click on Data Factories from the search results.

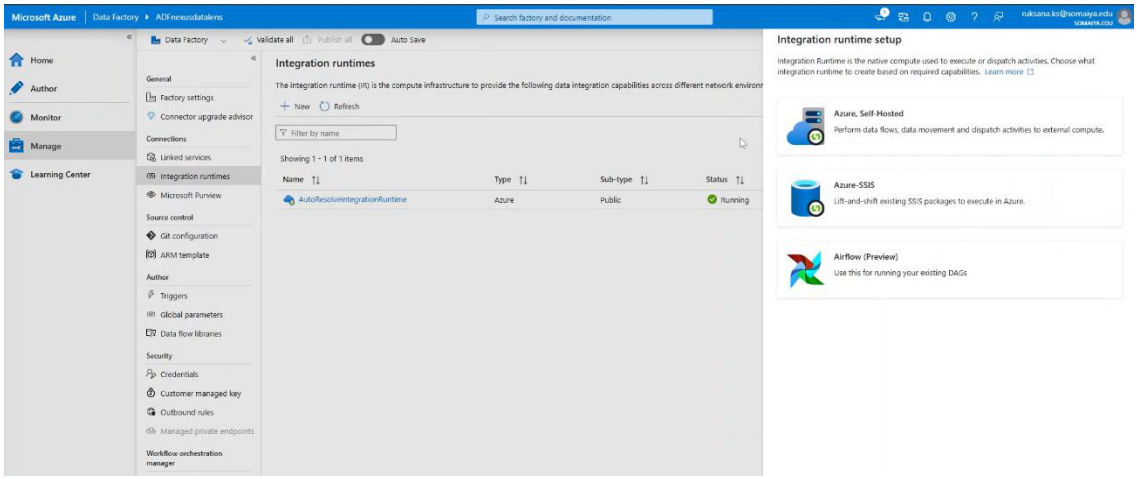
Select Your Data Factory Instance

- In the list of existing Data Factories, select your Data Factory instance (in this case, probably one associated with NexusProject as seen in the URL).

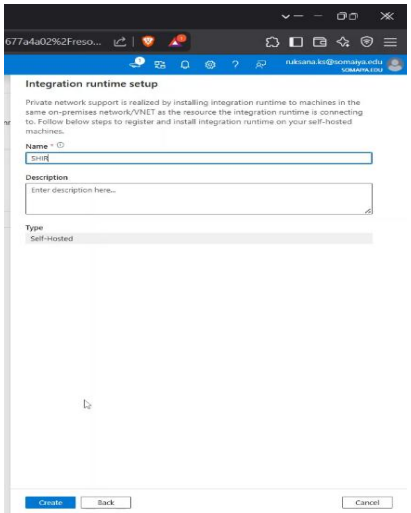
Launch Azure Data Factory Studio

- After selecting the instance, click the 'Open Azure Data Factory Studio' button from the resource overview page.

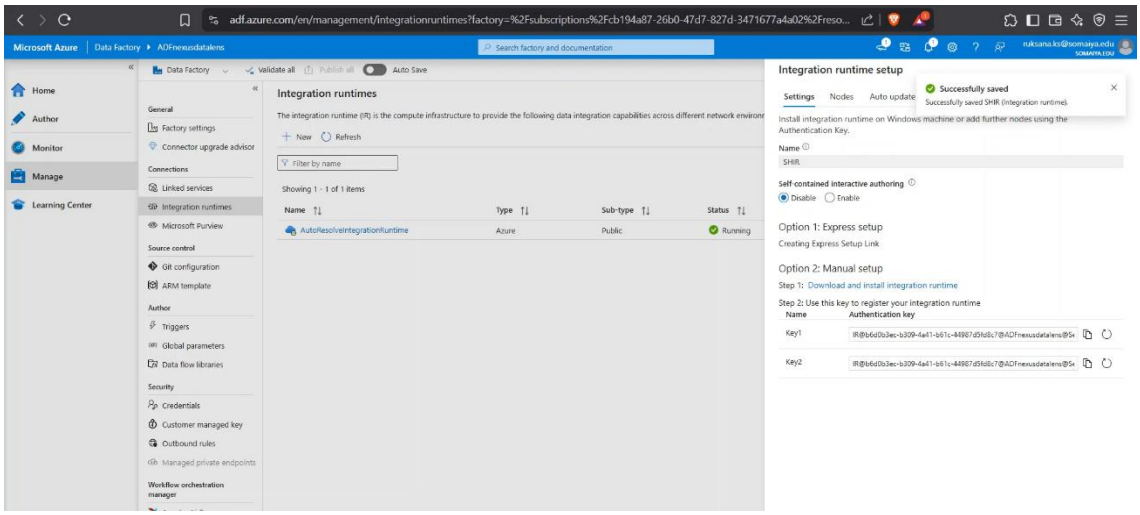




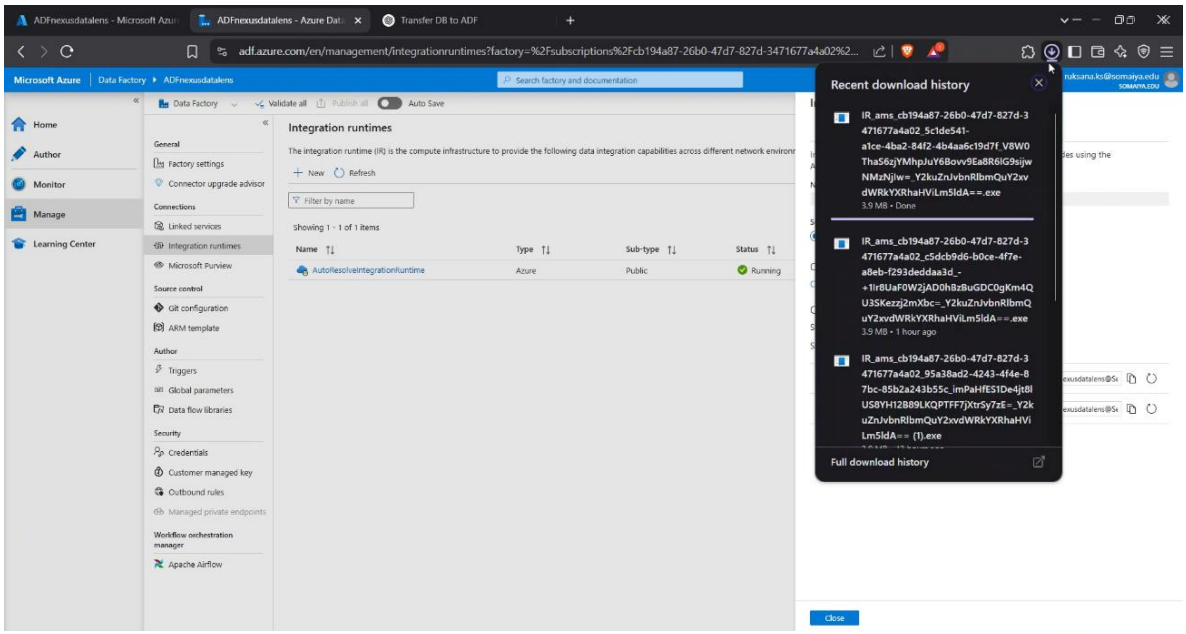
Create a new Self-hosted Integration Runtime in Azure Data Factory by entering name **SHIR** and type Self-Hosted clicking Create.



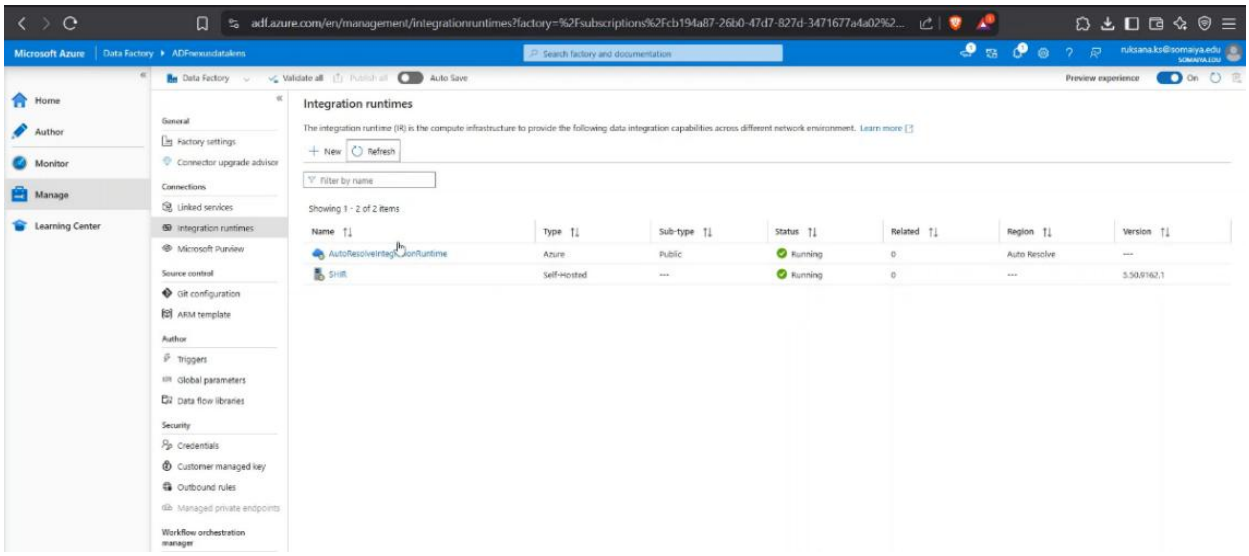
Self-hosted Integration Runtime (SHIR) is created and provides authentication keys for registering it on a Windows machine via express or manual setup.



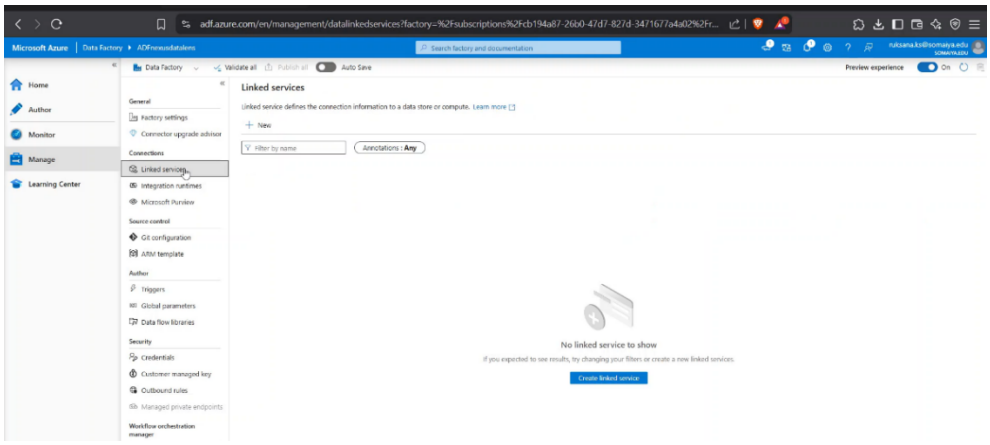
Azure Data Factory's Integration runtimes tab with a running Azure runtime. On the right, the recent downloads panel shows multiple .exe files, likely installers for a Self-hosted Integration Runtime (SHIR).



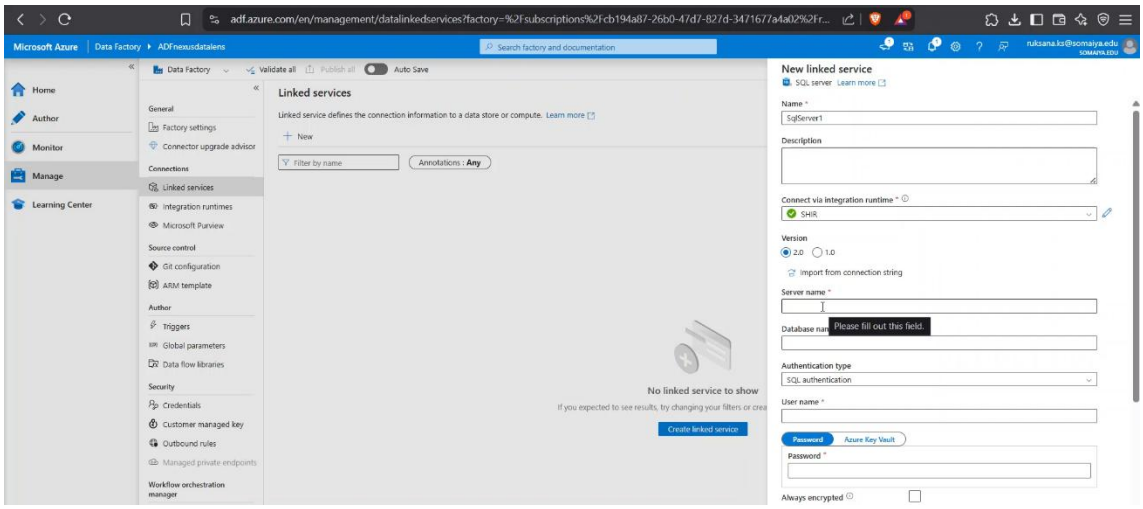
The Azure Data Factory now shows two integration runtimes: **AutoResolveIntegrationRuntime (Azure-hosted)** and **SHIR (Self-hosted)**, both in a running state. This setup enables data integration both in the cloud and on-premises.



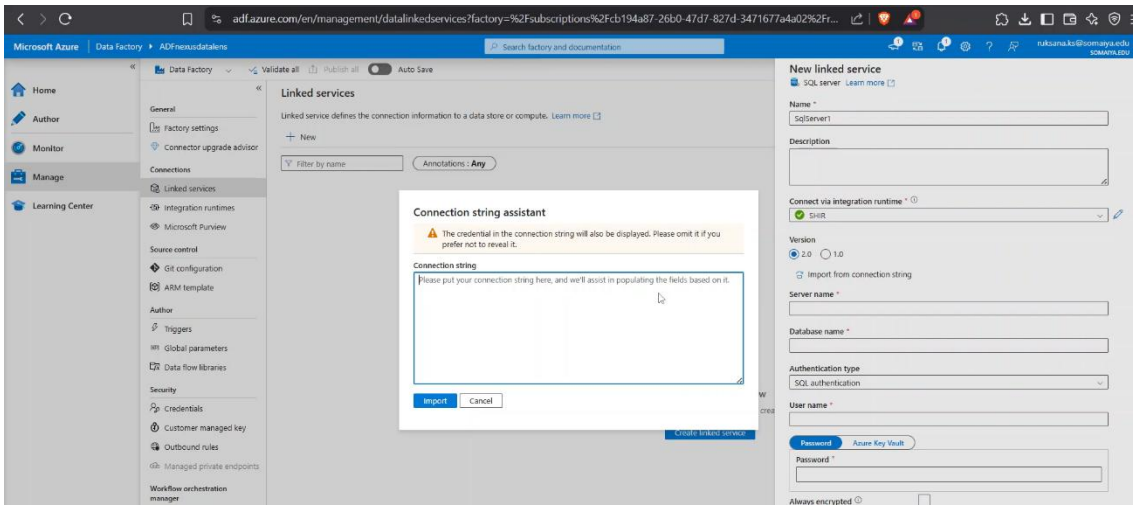
The Linked services tab in Azure Data Factory, which currently has no connections configured. Linked services are used to define connection info for data sources or compute environments.



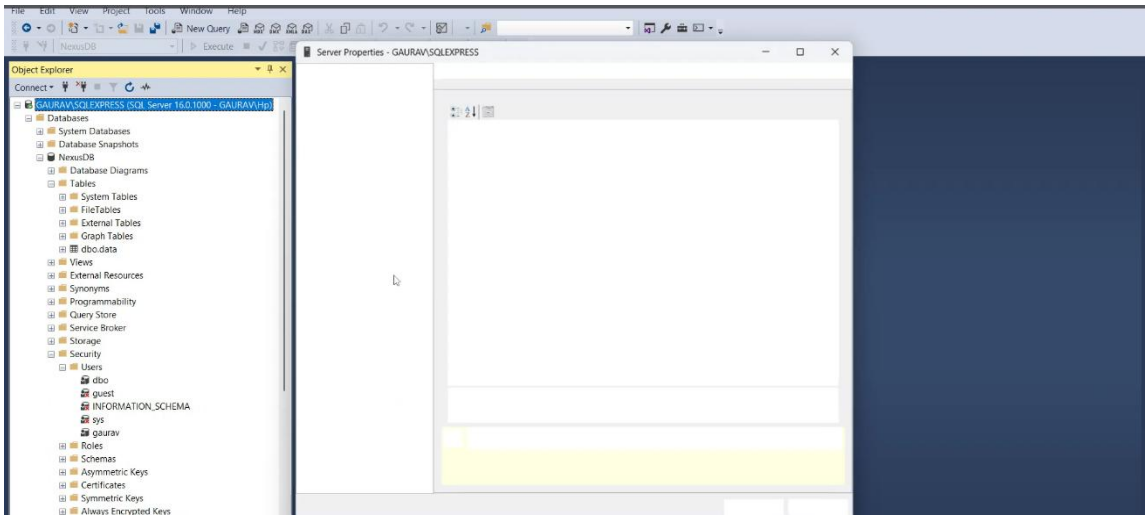
The process of creating a new linked service for a SQL Server in Azure Data Factory. The user is required to input connection details such as server name, database name, and authentication type.



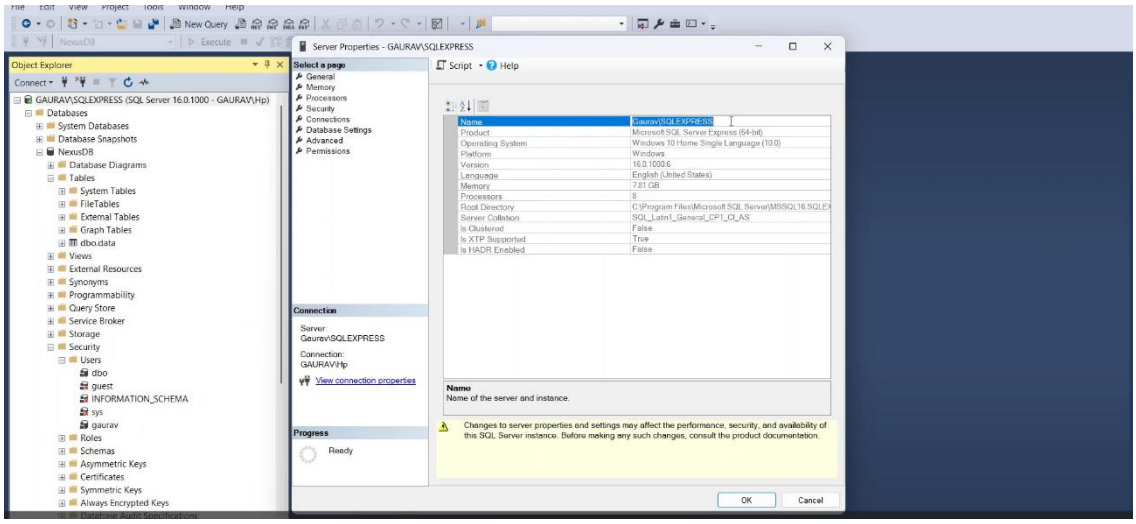
The Connection String Assistant in Azure Data Factory while creating a new linked service. It prompts the user to input a connection string to auto-populate SQL Server connection details securely.



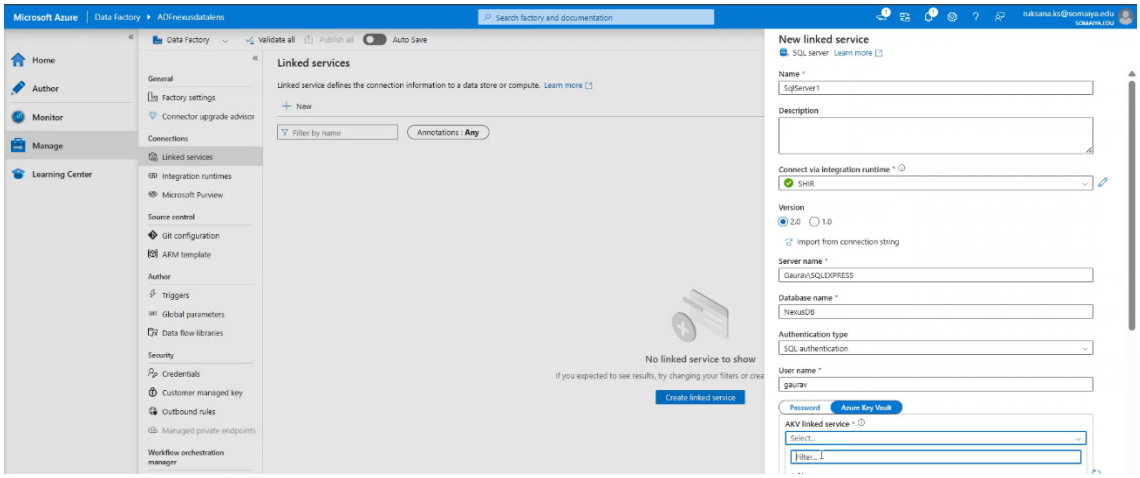
SQL Server Management Studio (SSMS) with the NexusDB database expanded, displaying its tables, views, and other objects.



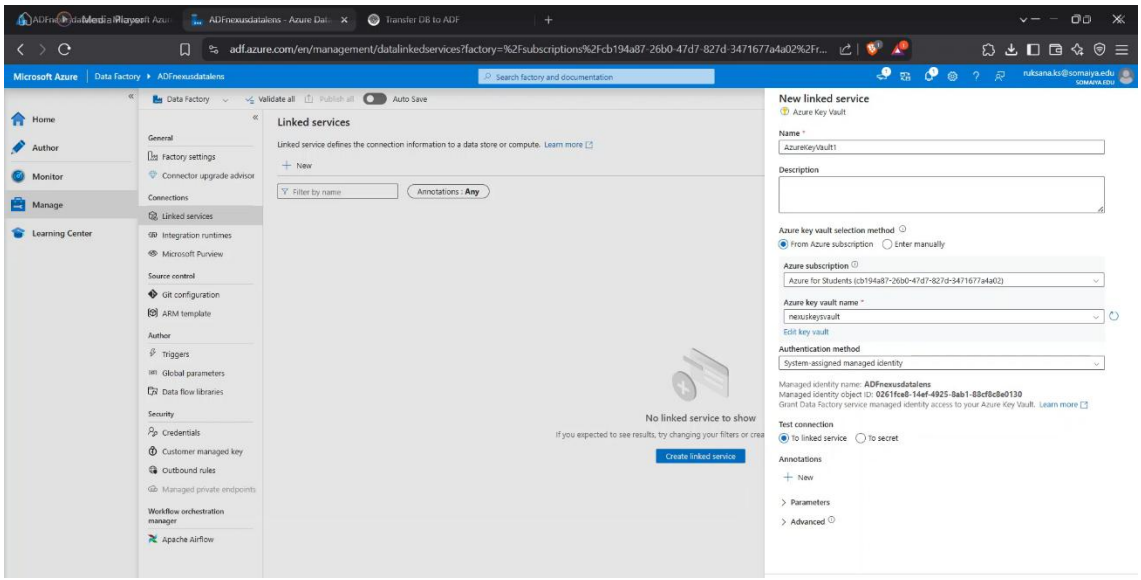
The Server Properties window in SSMS for the instance "**GAURAV\SQLEXPRESS**"(Server Name), showing system details like SQL Server version, operating system, memory, and collation. It helps users verify configuration and environment settings.



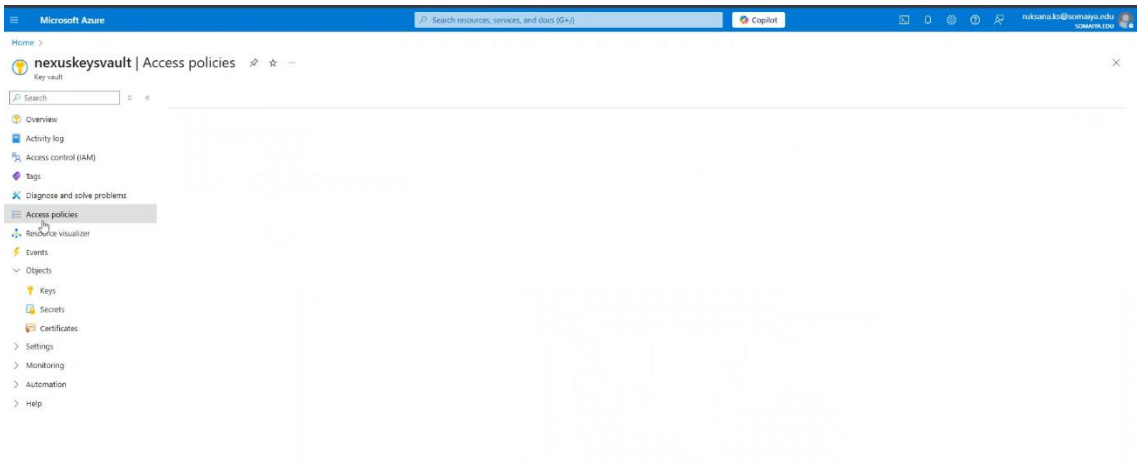
The configuration of a new linked service in Azure Data Factory, connecting to a SQL Server instance named "**Gaurav\SQLEXPRESS**" with the database "**NexusDB**". SQL authentication is used with the username "**gaurav**", and an option to select an Azure Key Vault for the password is being displayed.



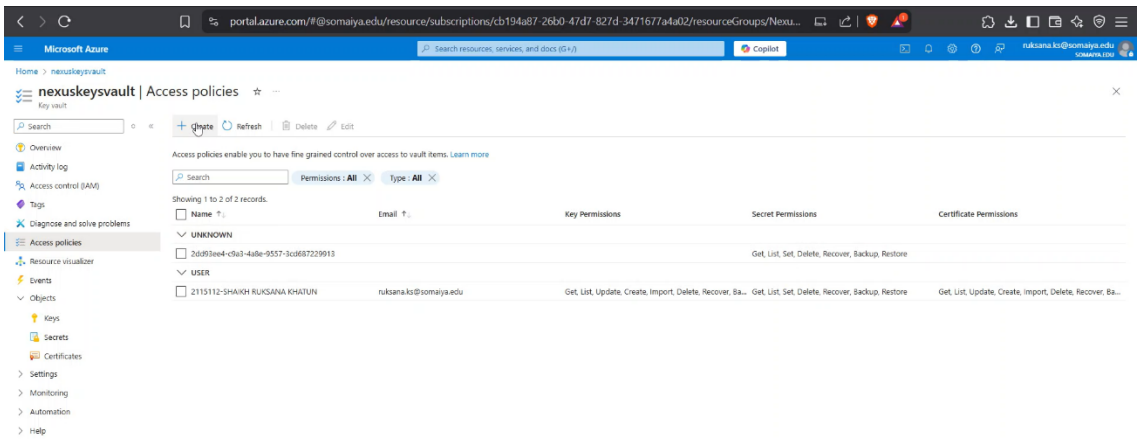
The process of creating a new linked service in Azure Data Factory to connect with an Azure Key Vault. The right pane captures the configuration details, including selecting the subscription, key vault name, and using system-assigned managed identity for authentication.



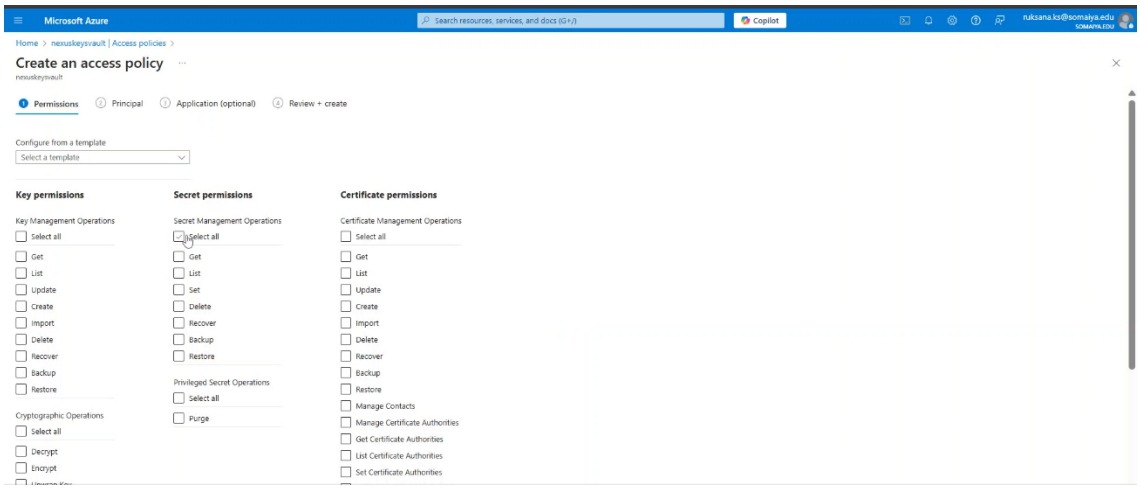
The Access policies section of the Azure Key Vault named "**nexuskeyvault**". Here, you can configure permissions for users and services to access secrets, keys, and certificates within the vault.



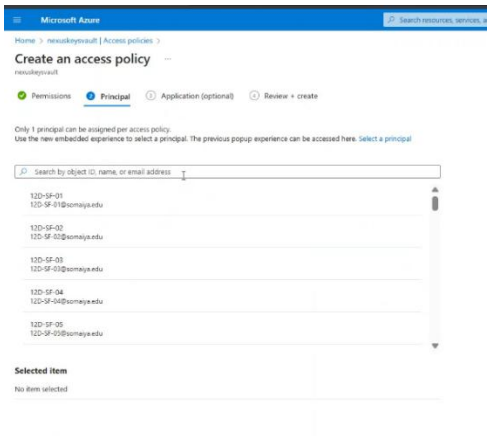
The Access policies list for the nexuskeyvault, showing users and their assigned permissions. It includes detailed rights such as Get, List, Set, Delete, Recover, Backup, and Restore for secrets, keys, and certificates.



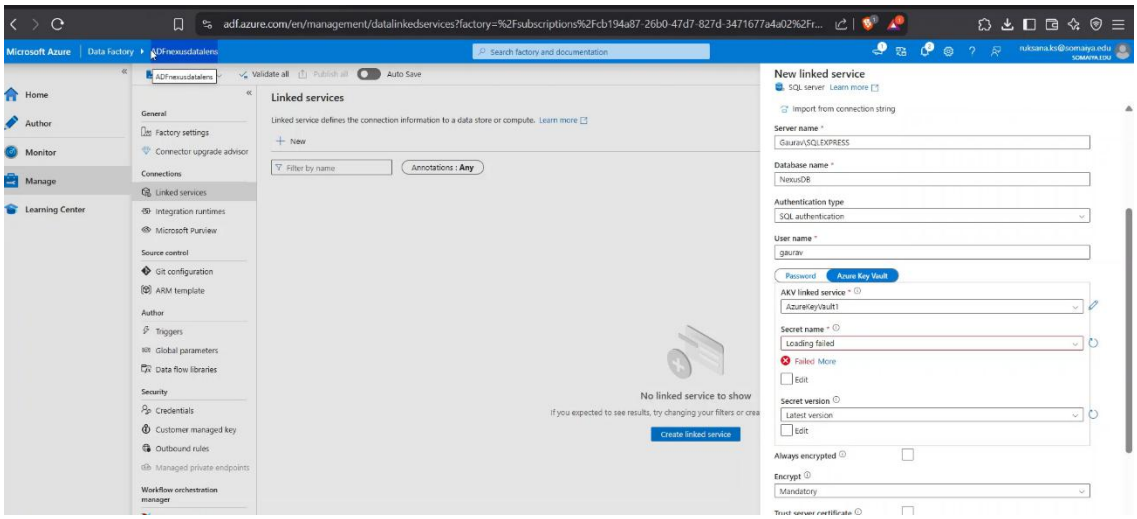
Create an access policy page for Azure Key Vault, where you can assign specific permissions for keys, secrets, and certificates. It allows fine-grained control, such as selecting operations like Get, Set, Delete, or Purge under each category.



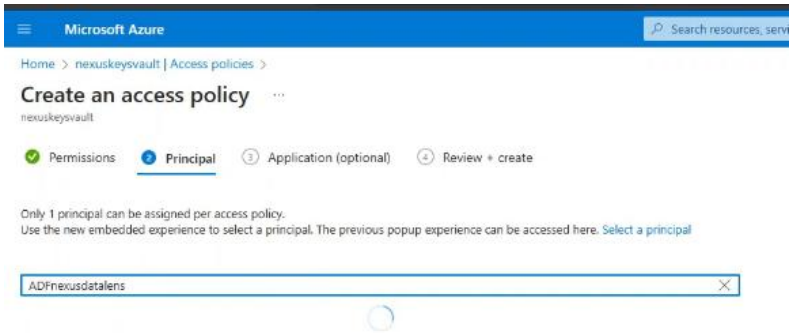
The Principal selection step while creating an access policy in Azure Key Vault. Here, you choose a user or service principal by searching with name, object ID, or email to assign the defined permissions.



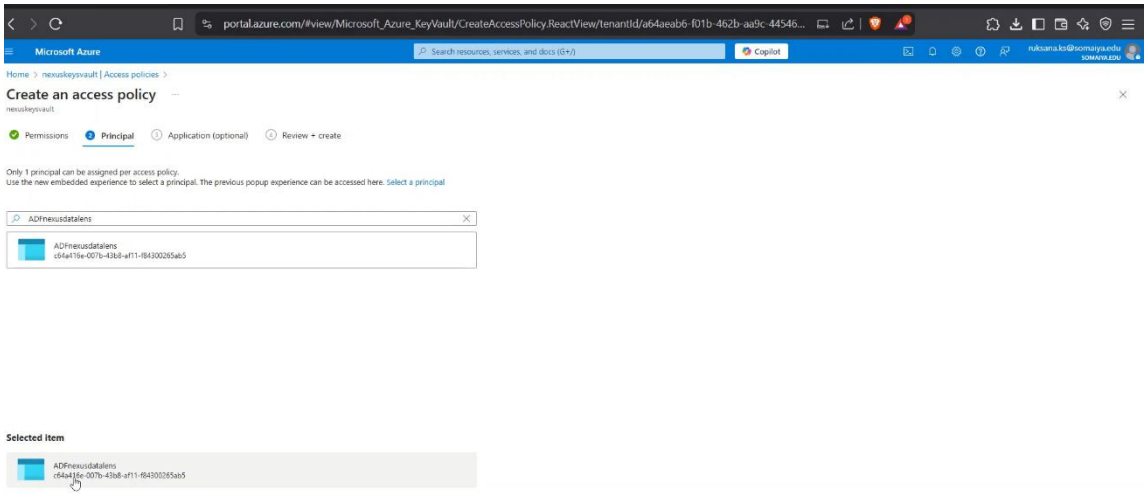
The process of linking a SQL Server in Azure Data Factory using SQL authentication with the password fetched from Azure Key Vault.



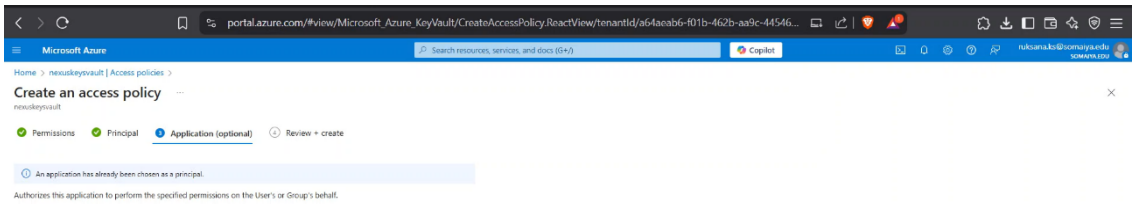
The selection of the "**ADFnexusdatalens"** Data Factory as a principal while creating an access policy in Azure Key Vault.



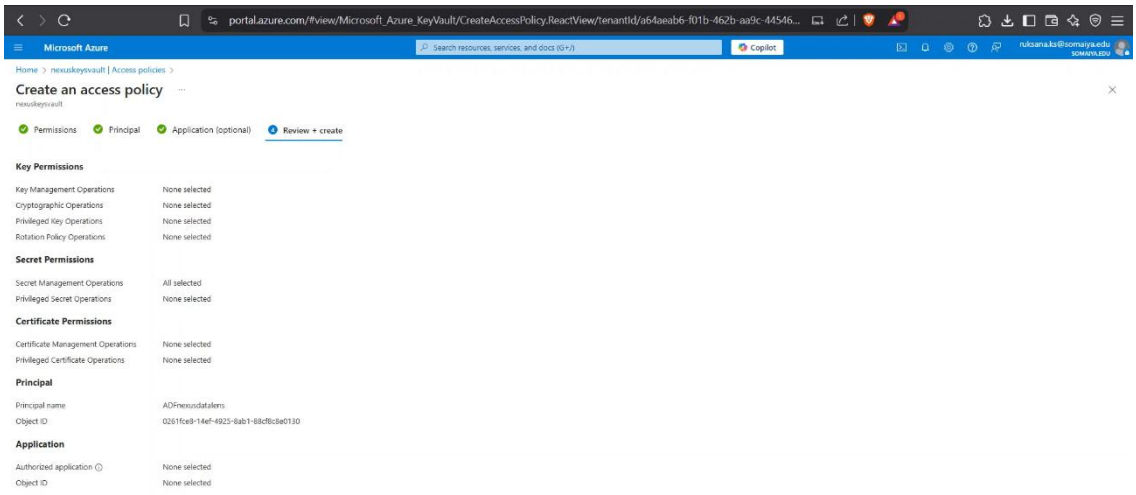
The Data Factory instance **ADFnexusdatalens** has been **successfully selected** as the principal for the Key Vault access policy.



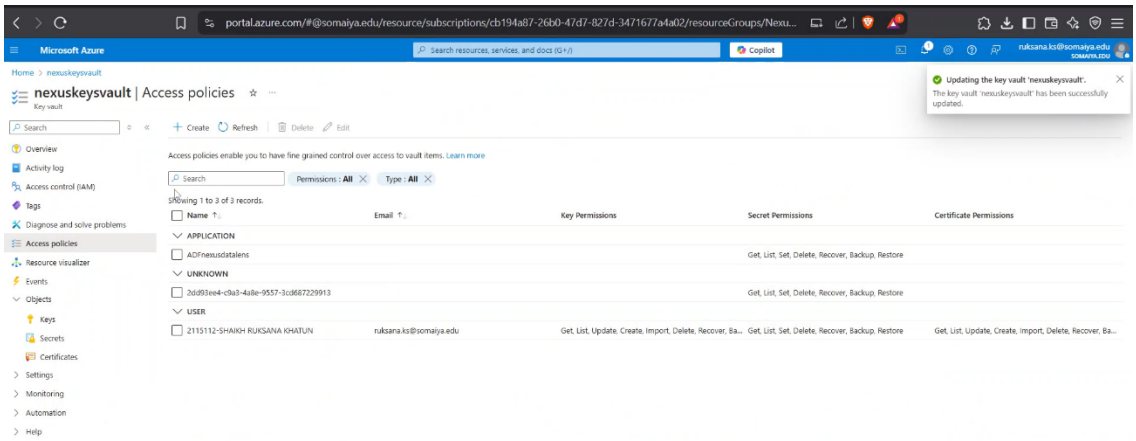
This step confirms that an application (**ADFnexusdatalens**) has already been selected as the principal to authorize Key Vault access.



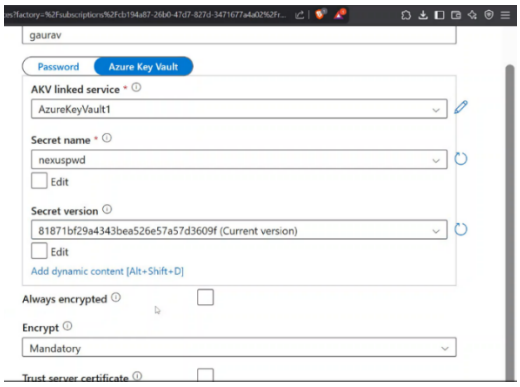
The final review of the access policy, showing that only **secret permissions** are granted to the **ADFnexusdatalens principal.**

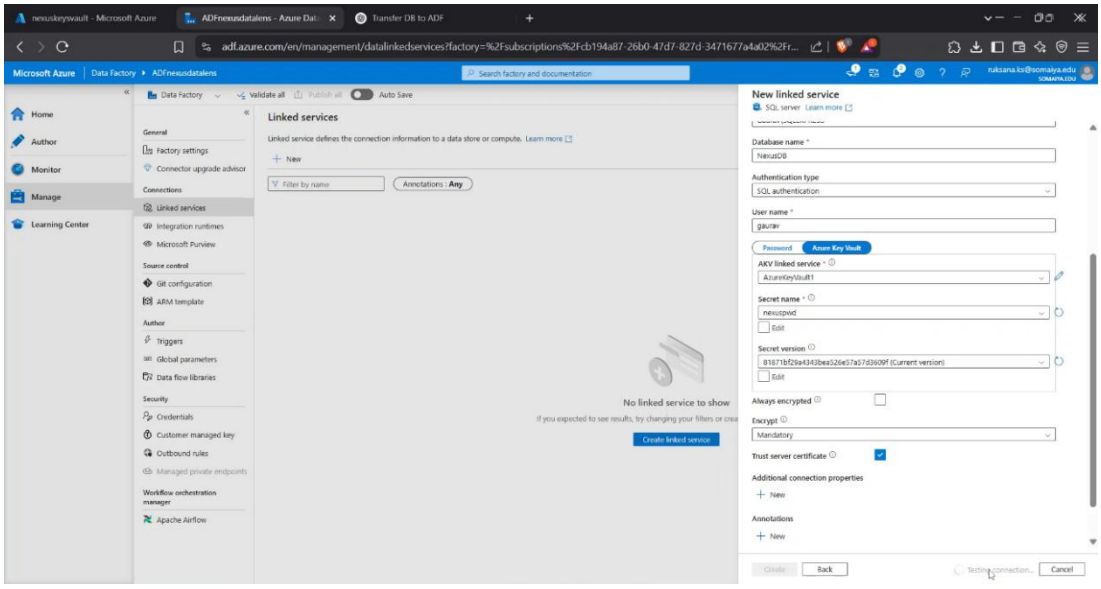


The access policy for the Key Vault "**nexuskeyvault**" has been successfully updated, granting secret permissions to **ADFnexusdatalens.**

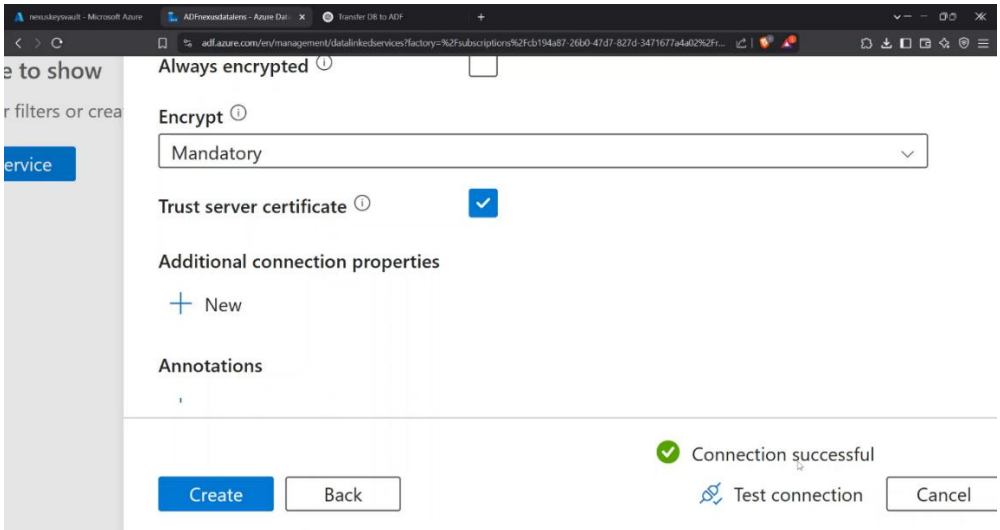


The Azure Key Vault secret named "**nexuspwd**" with a specified version is selected from **AzureKeyVault1** for secure integration in the linked service.

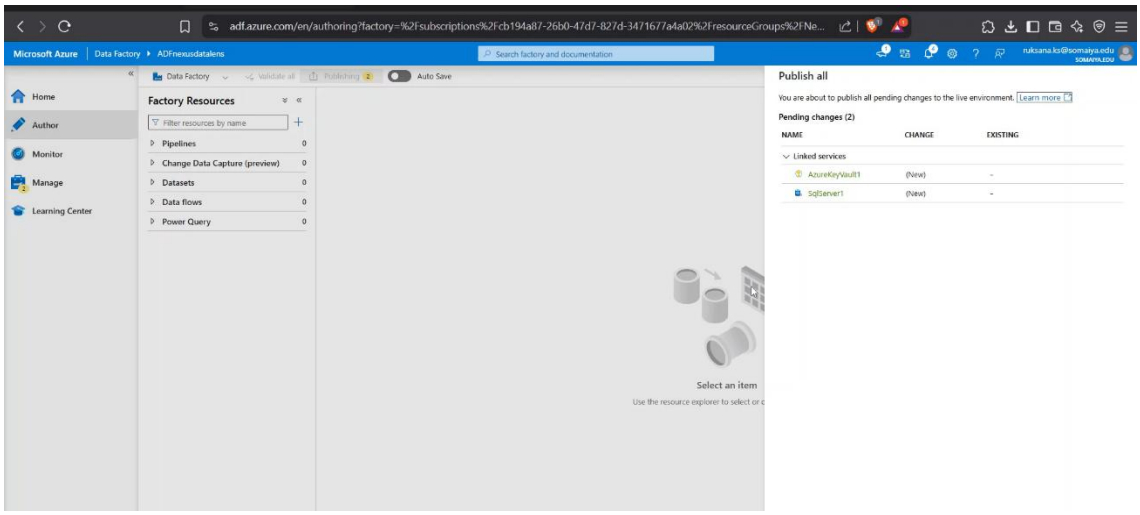


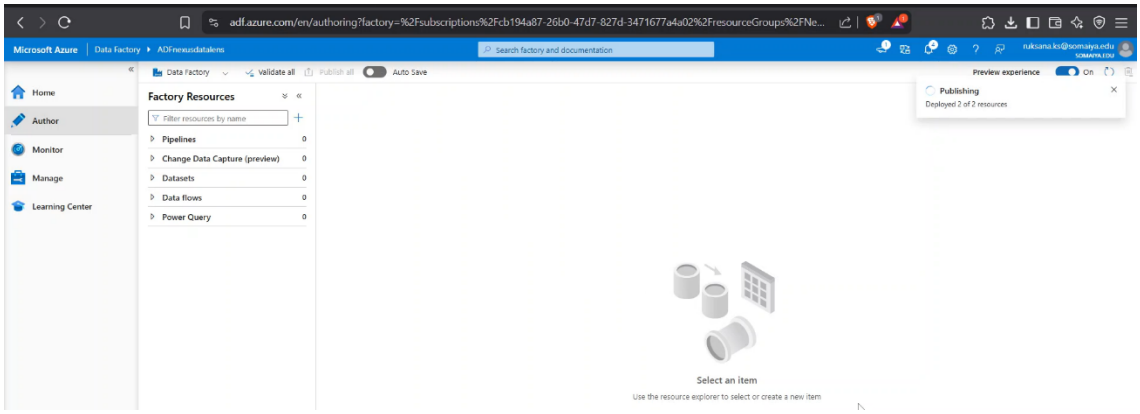


The database connection configuration is complete with encryption and certificate trust, and the **connection test** was **successful**.

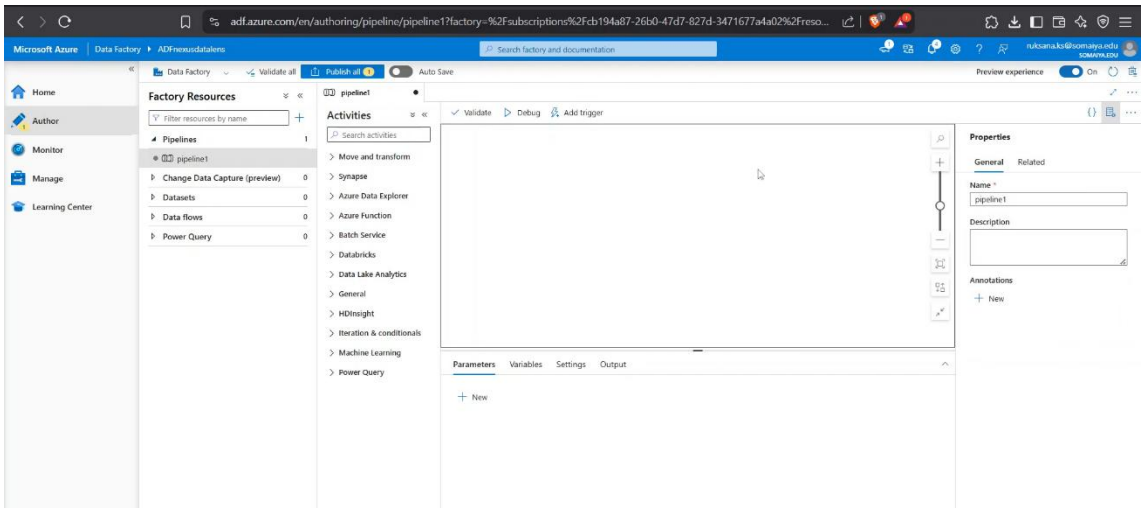


Two new linked services, **AzureKeyVault1** and **SqlServer1**, are ready to be published to the live Azure Data Factory environment.

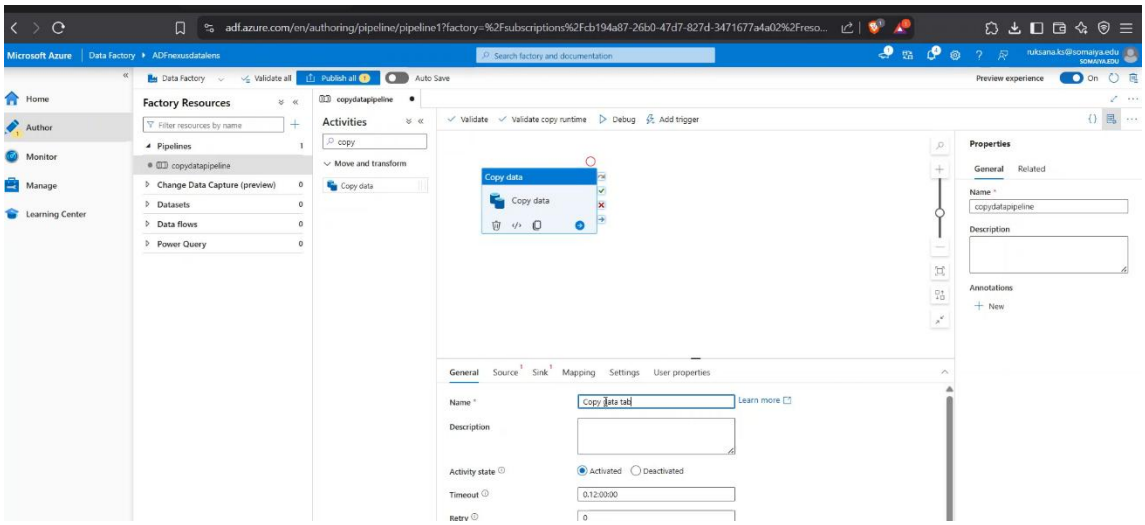




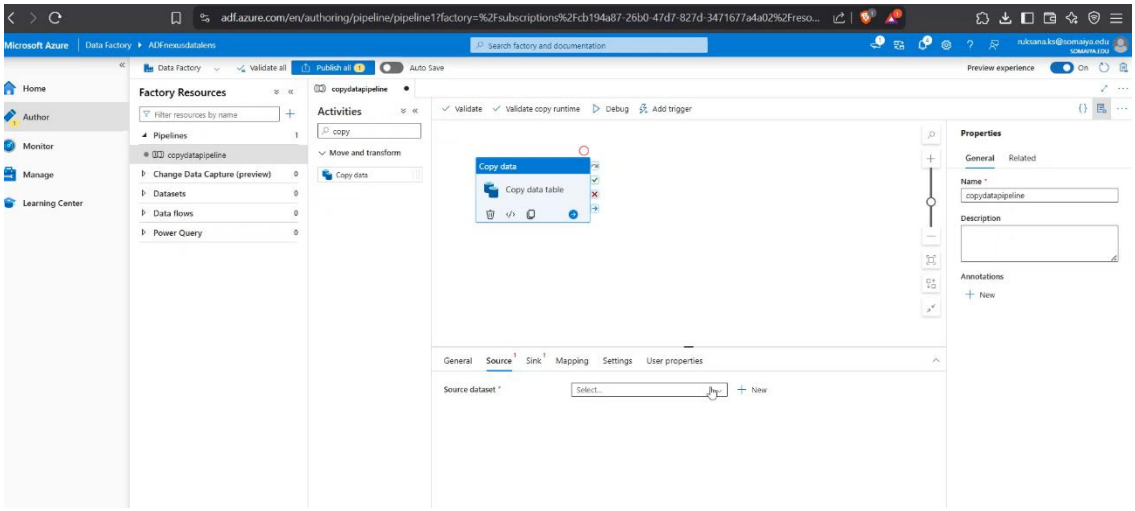
A new pipeline named "**pipeline1**" is being created in Azure Data Factory, ready for activities and configurations.



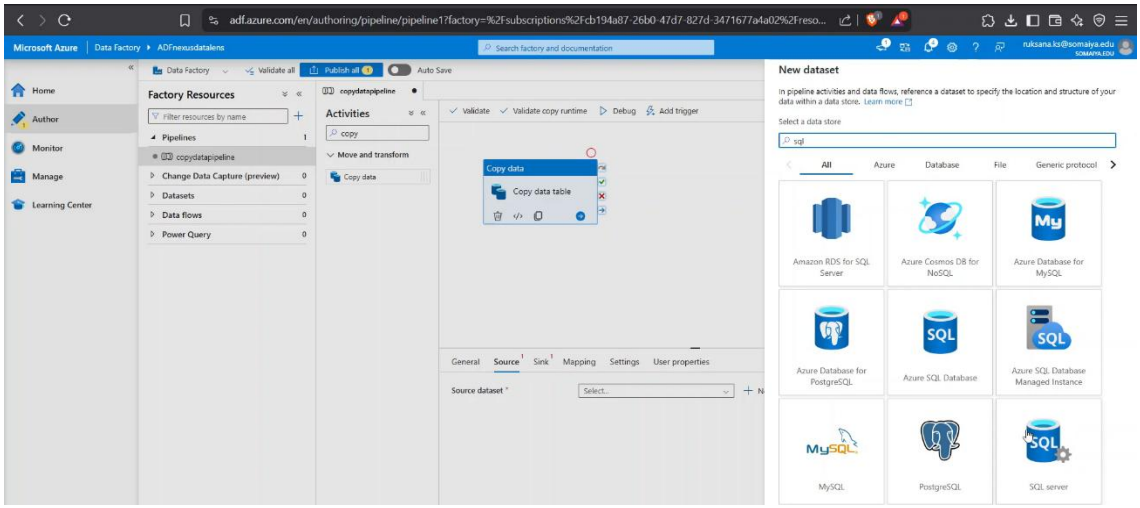
A pipeline named "**copydatapipeline**" with a "**Copy data**" activity titled "**Copy data task**" is being configured in Azure Data Factory.



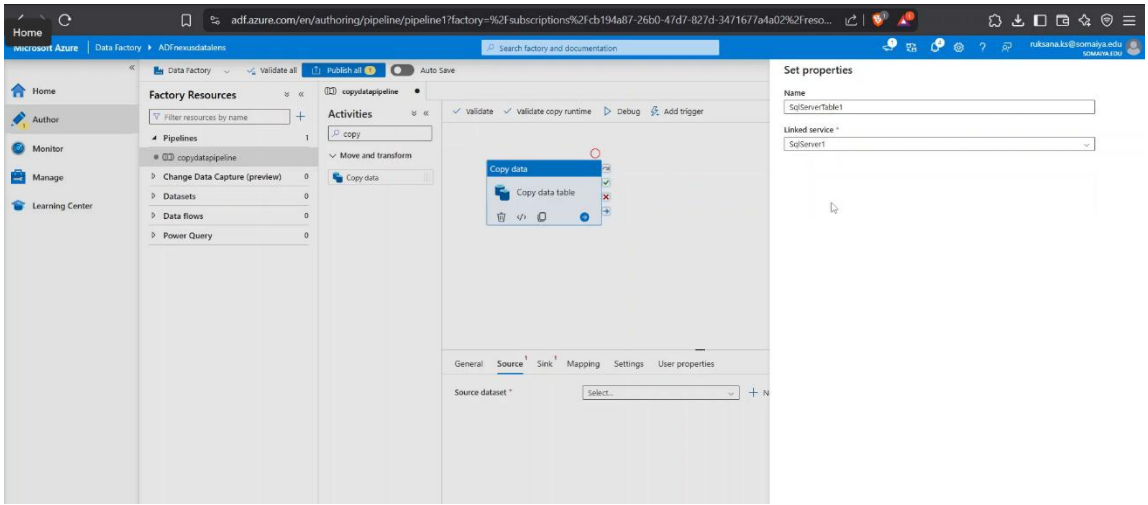
In Azure Data Factory, the source dataset for the "**Copy data table**" activity in the "**copydatapipeline**" is being selected.

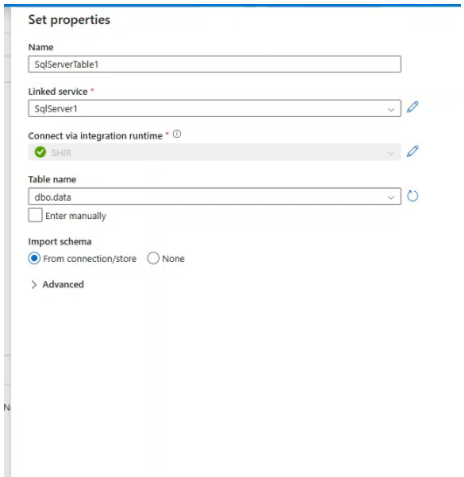


In Azure Data Factory, a new source dataset is being created by selecting a data store type, with options like SQL Server, MySQL, and Azure SQL Database shown.

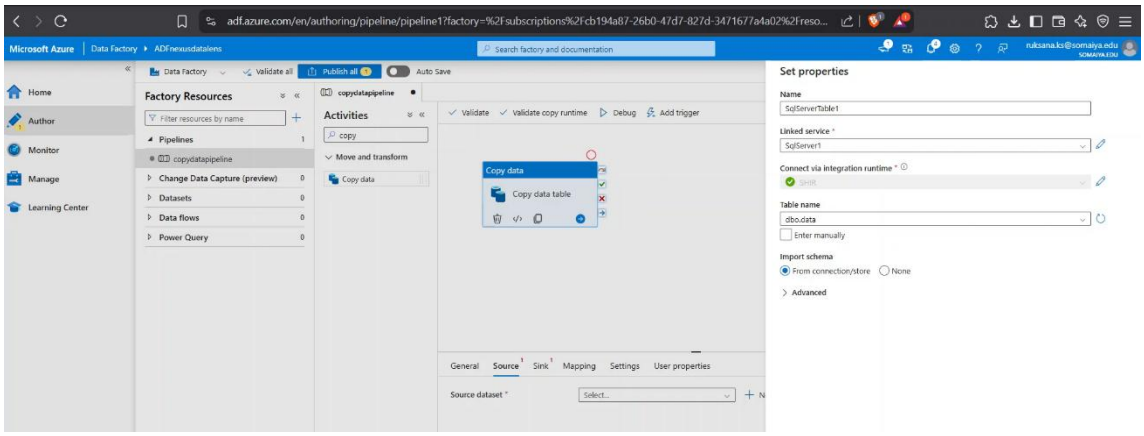


In Azure Data Factory, a source dataset named "**SqlServerTable1**" is being created using the linked service "**SqlServer1**".

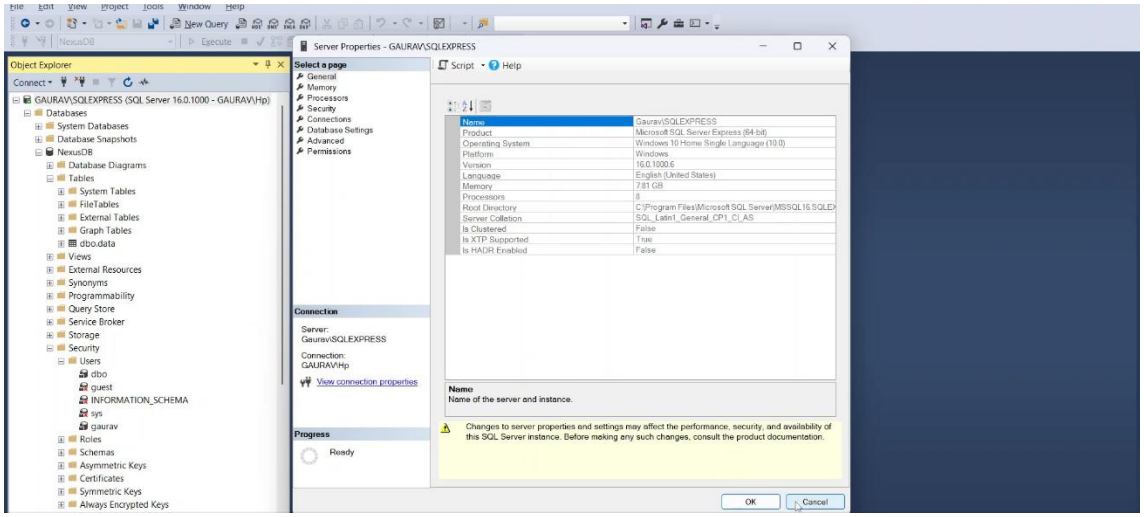




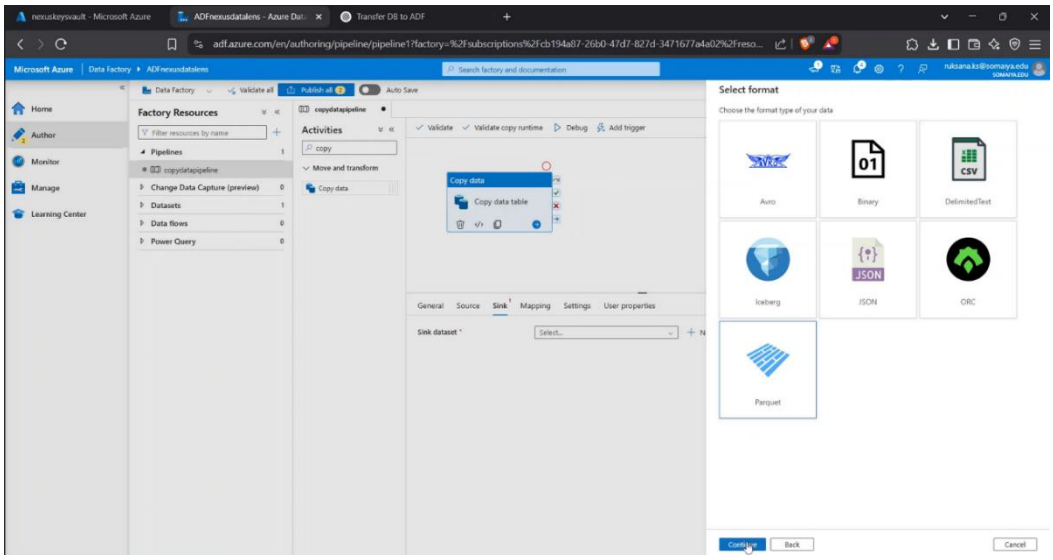
The source dataset "**SqlServerTable1**" is configured to connect to the "**SqlServer1**" linked service and pull data from the "**db.data**" table using integration runtime**.**



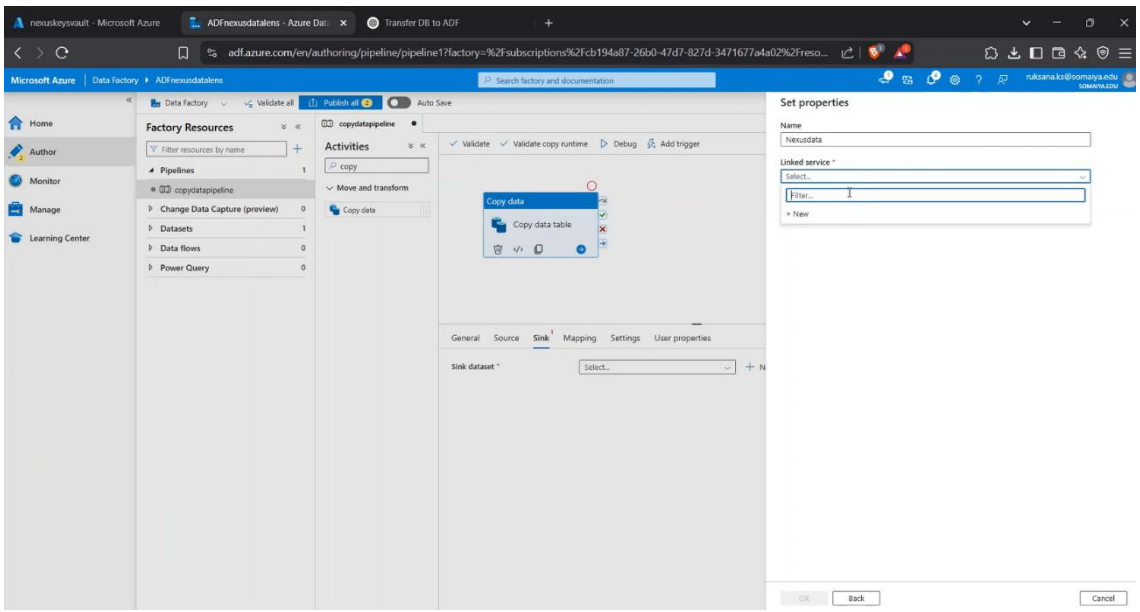
The SQL Server Management Studio (SSMS) is displaying the server properties for the local instance "**GAURAV\SQLEXPRESS**" running SQL Server Express on a Windows machine.



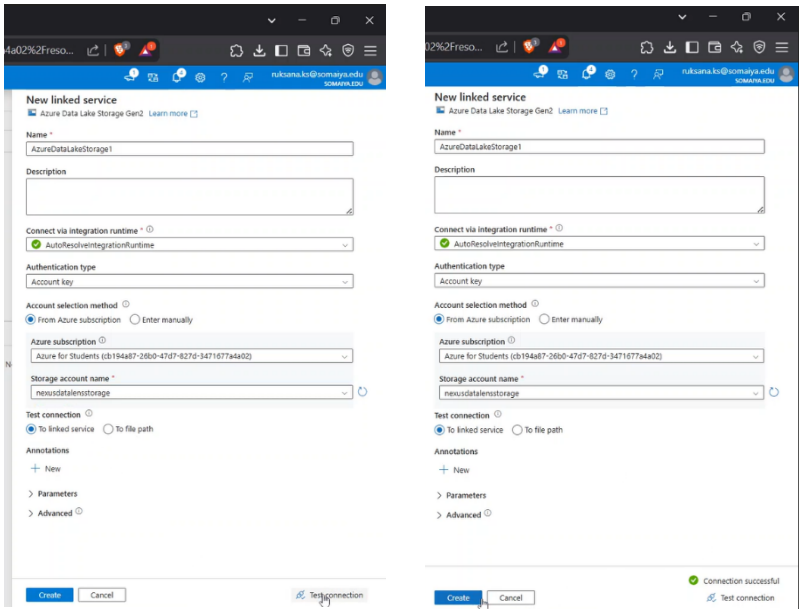
In Azure Data Factory, the user is selecting the **Parquet** format for the sink dataset while configuring a Copy Data activity in the pipeline.



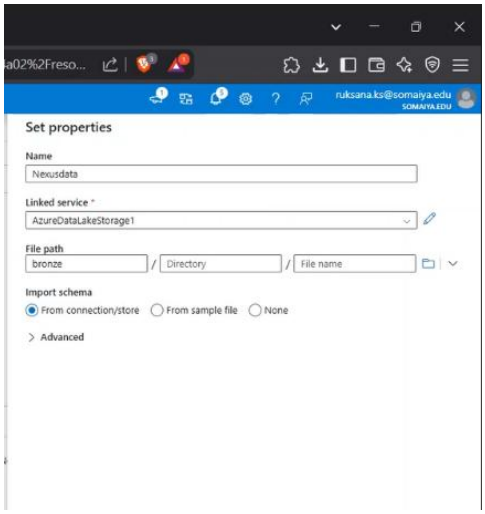
The user is setting properties for the sink dataset named ***Nexusdata*** by selecting or creating a linked service in Azure Data Factory.



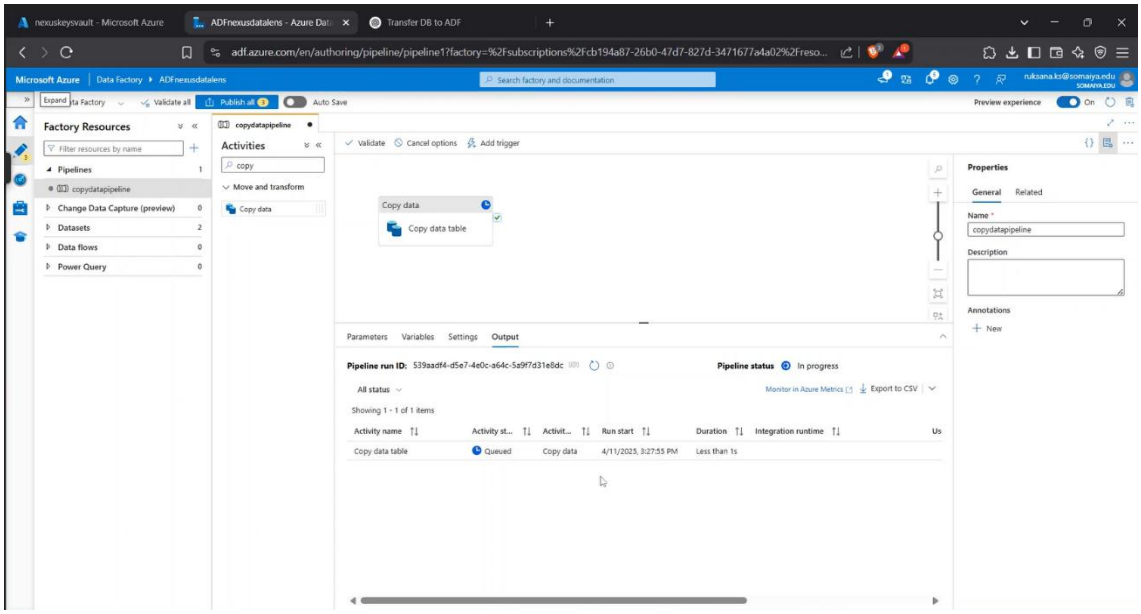
The user is configuring a new linked service to connect **Azure Data Factory** with **Azure Data Lake Storage Gen2** using an account key and is testing the connection.



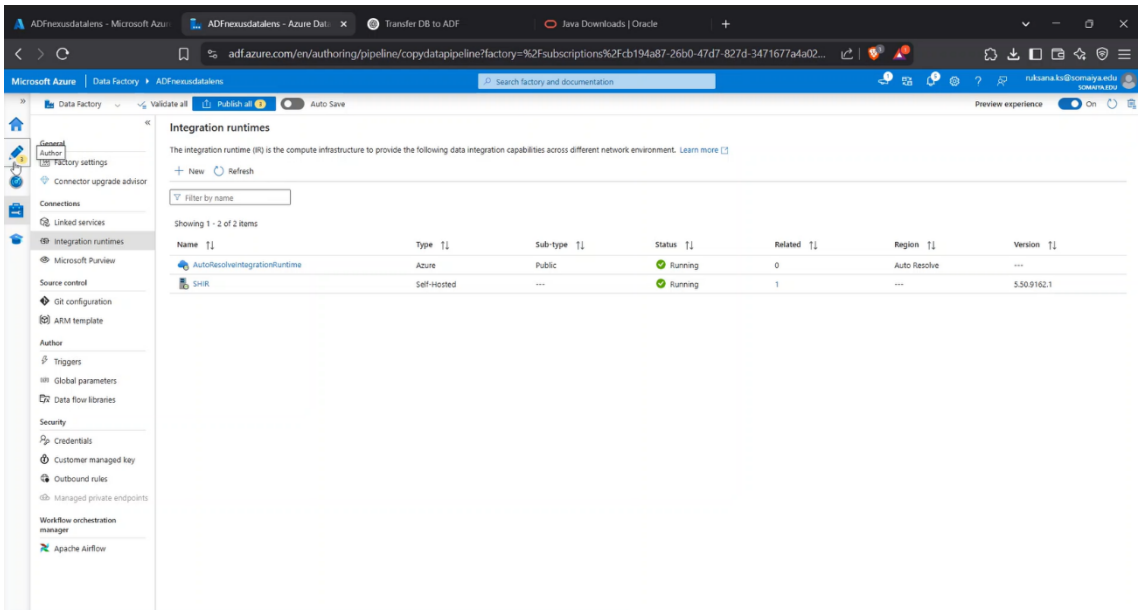
The user is setting the sink dataset properties to write data into the "**bronze**" directory of **Azure Data Lake Storage Gen2** using the linked service.



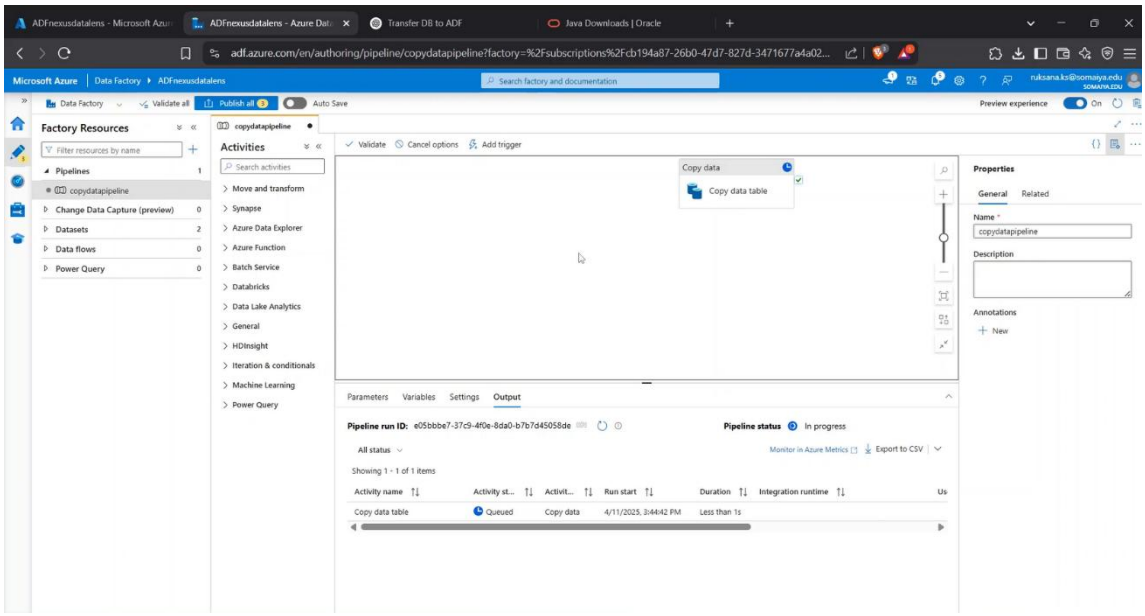
The pipeline run for copying data has started and the "**Copy data table**" activity is currently in the **Queued state**.



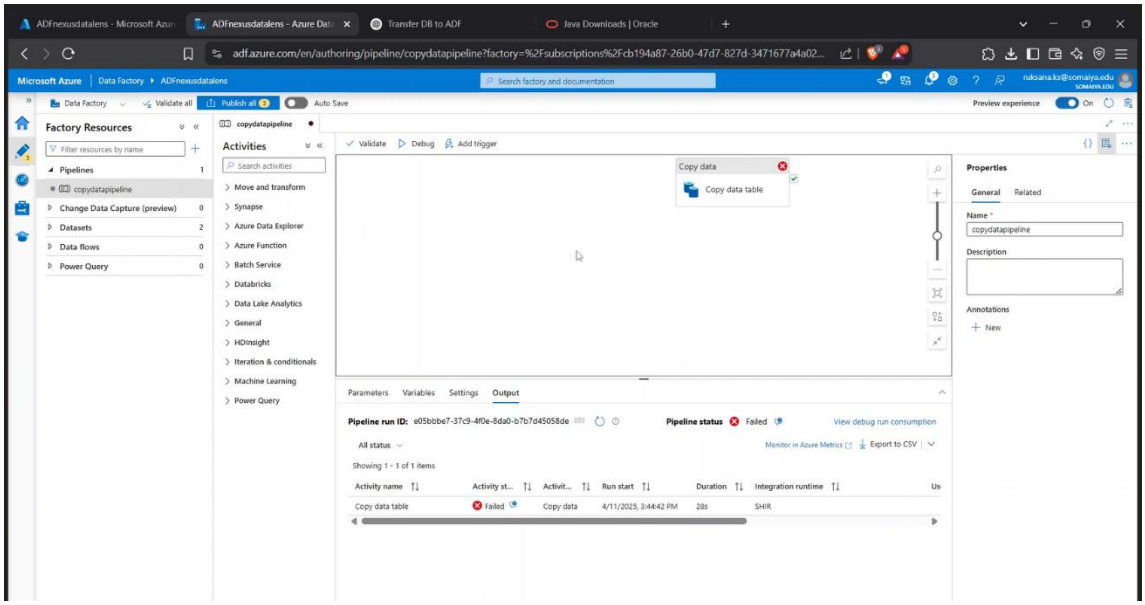
The Integration Runtimes panel shows both **Azure** and **Self-hosted runtimes** are **running**, enabling data movement across cloud and on-premises environments.



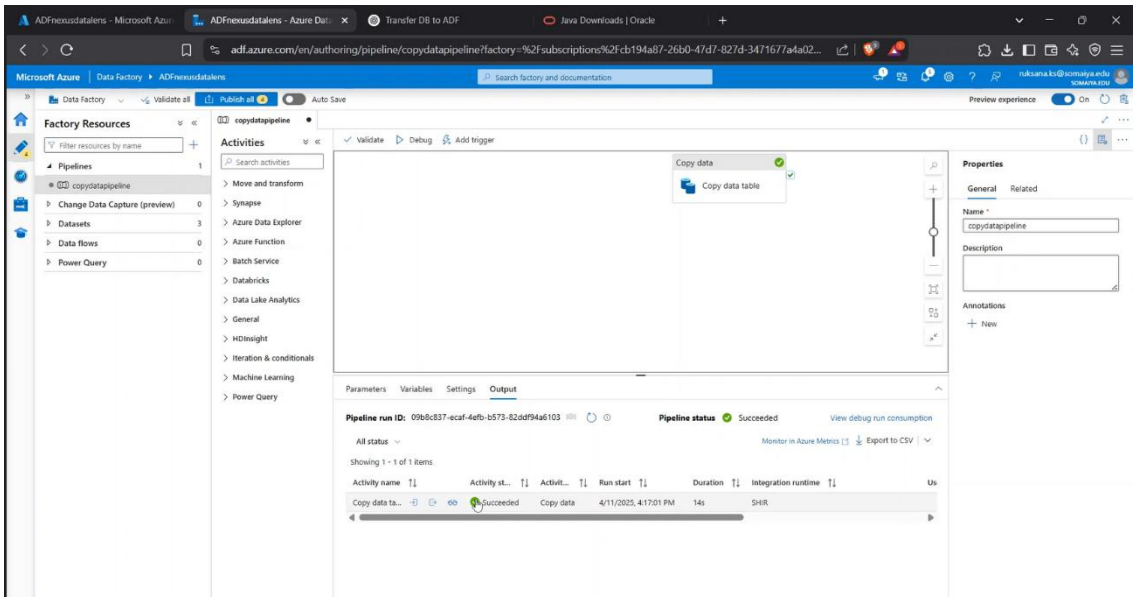
The pipeline run in Azure Data Factory is currently in progress, with the "**Copy data table**" activity queued for execution.



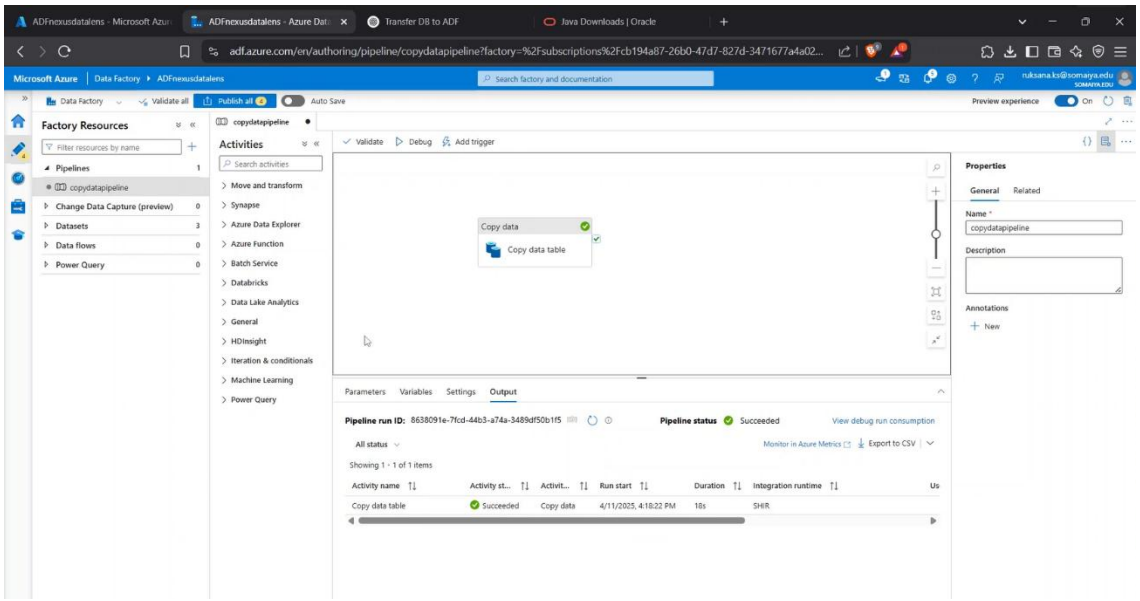
The Azure Data Factory pipeline run has failed during the "**Copy data table**" activity using the self-hosted integration runtime (SHIR).



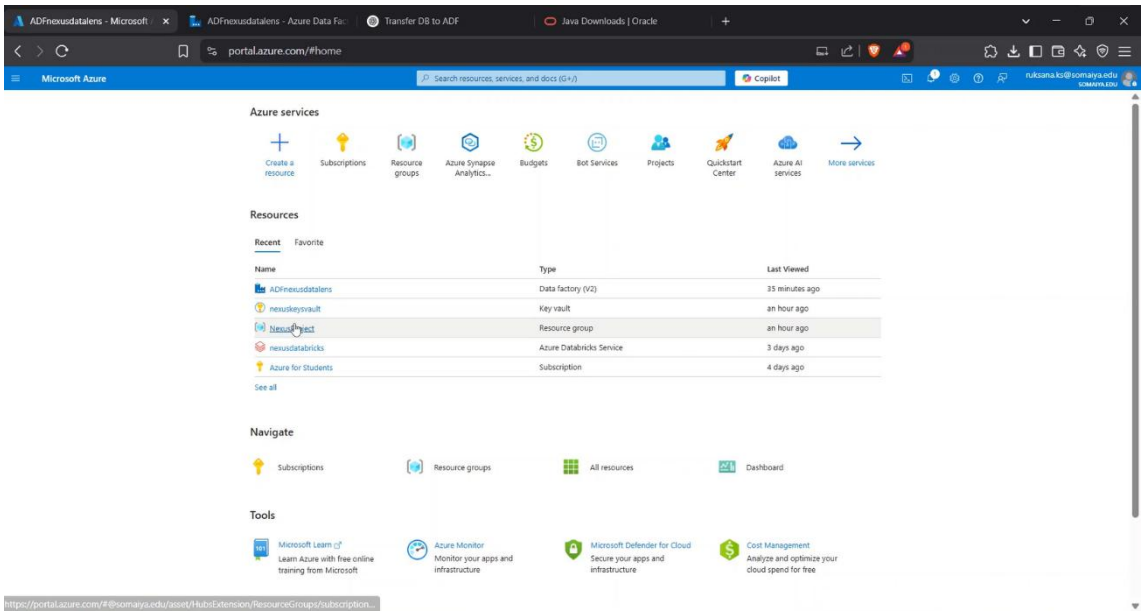
The Azure Data Factory pipeline **successfully completed** the "**Copy data table**" activity using the self-hosted integration runtime (SHIR).



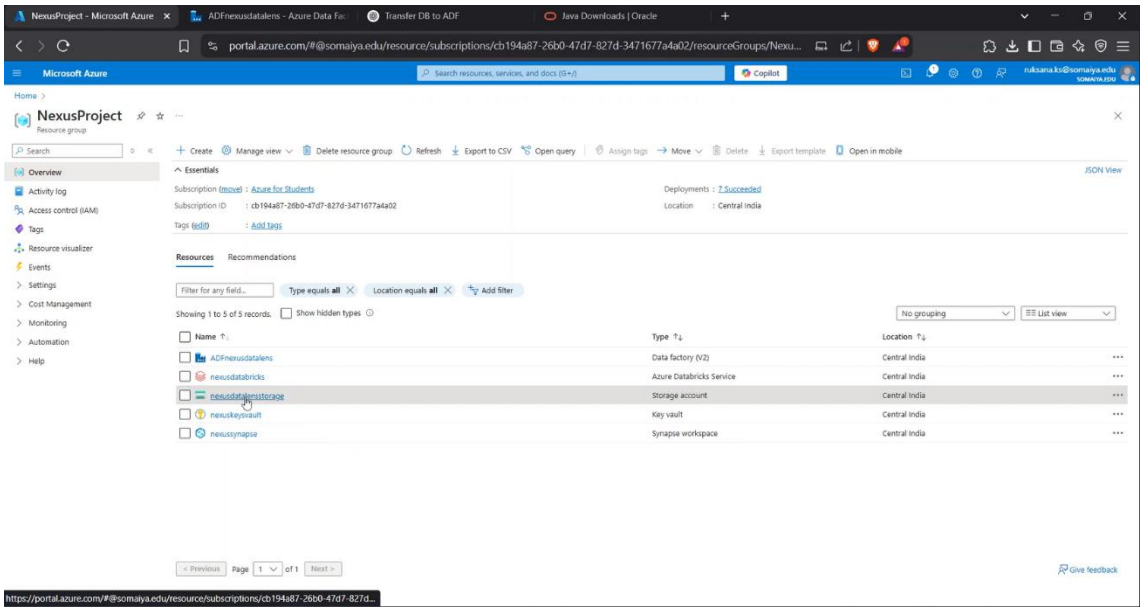
The "**copydatapipeline**" in Azure Data Factory successfully completed the data transfer using SHIR in 18 seconds.



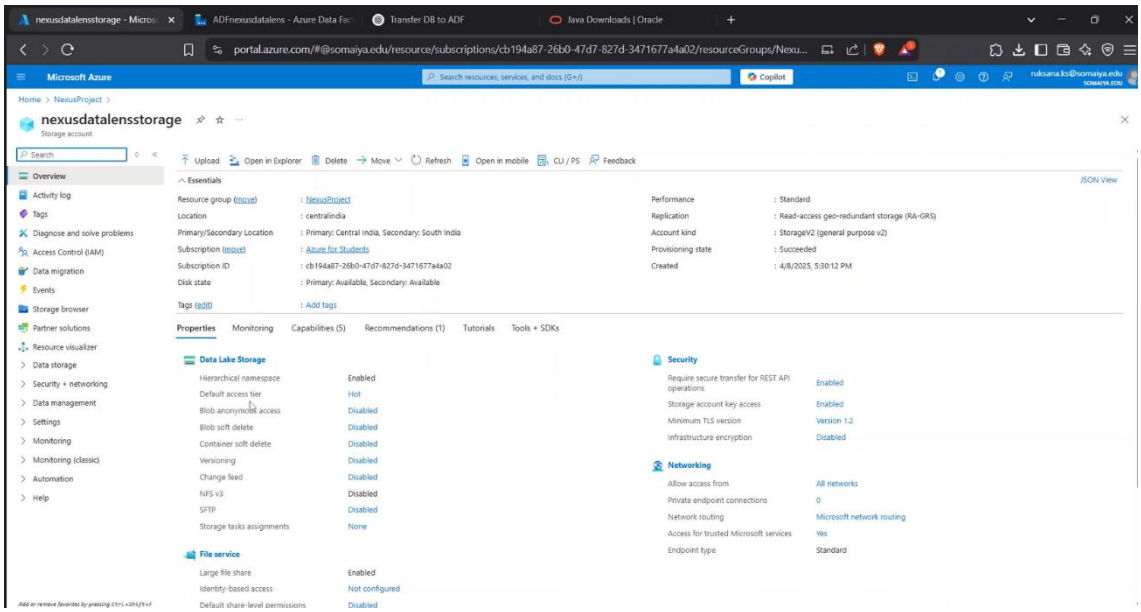
The Azure portal home page displays active resources including a Data Factory, Key Vault, Resource Group, and Databricks service under the user's subscription.



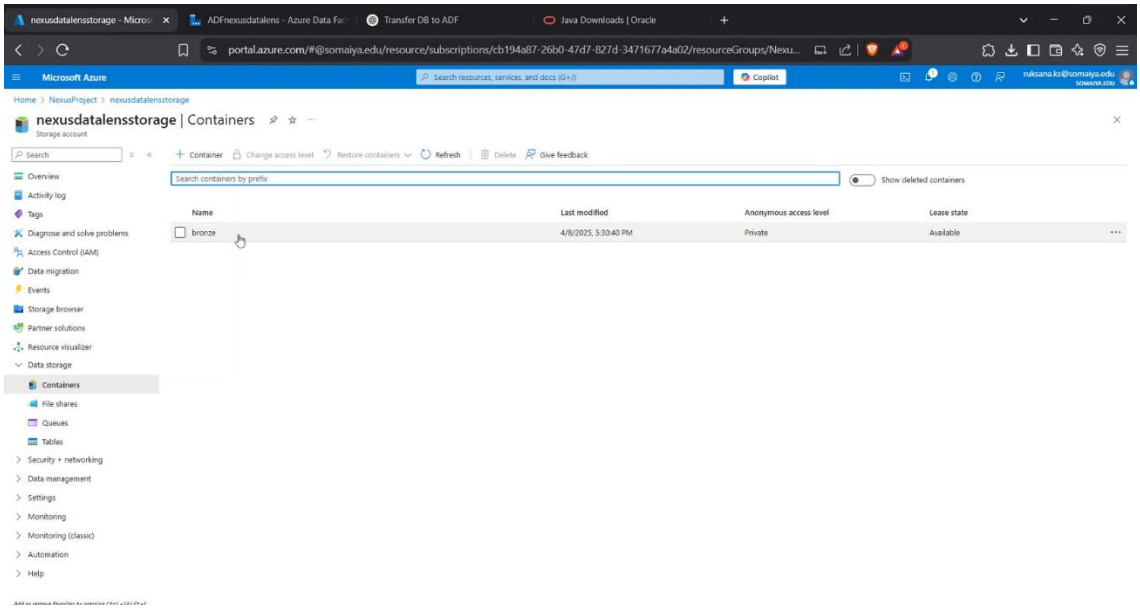
The NexusProject resource group in Azure contains various deployed resources including Data Factory, Databricks, a Storage Account, a Key Vault, and a Synapse workspace in the Central India region.



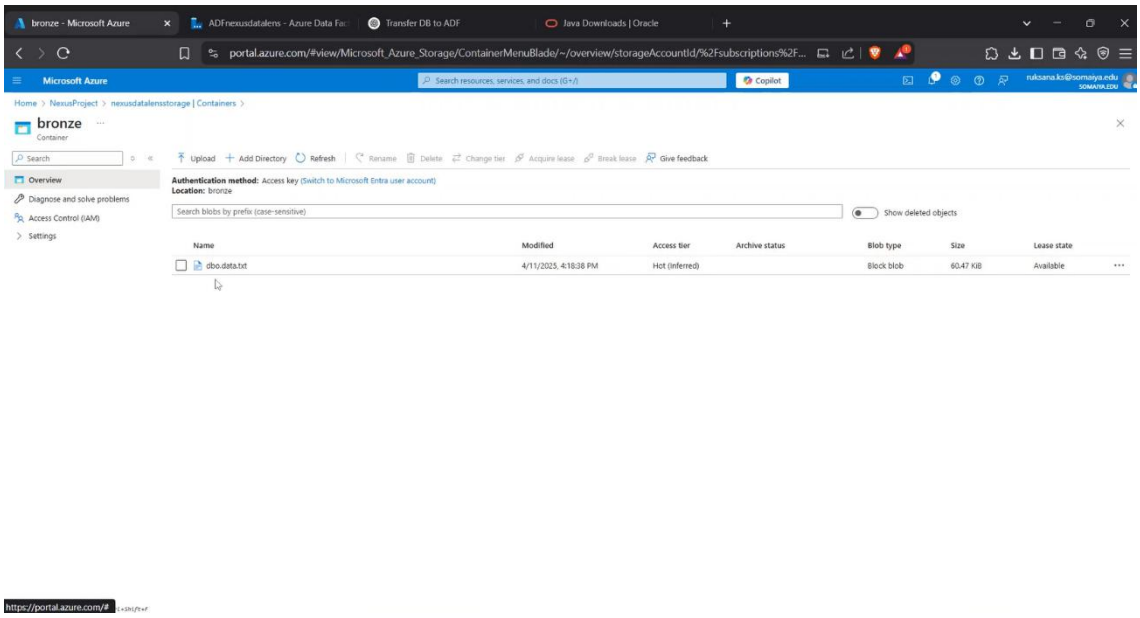
The **nexusdatalakestorage** is a successfully provisioned Azure Data Lake Storage Gen2 account with hierarchical namespace enabled and various features like blob soft delete and versioning disabled.



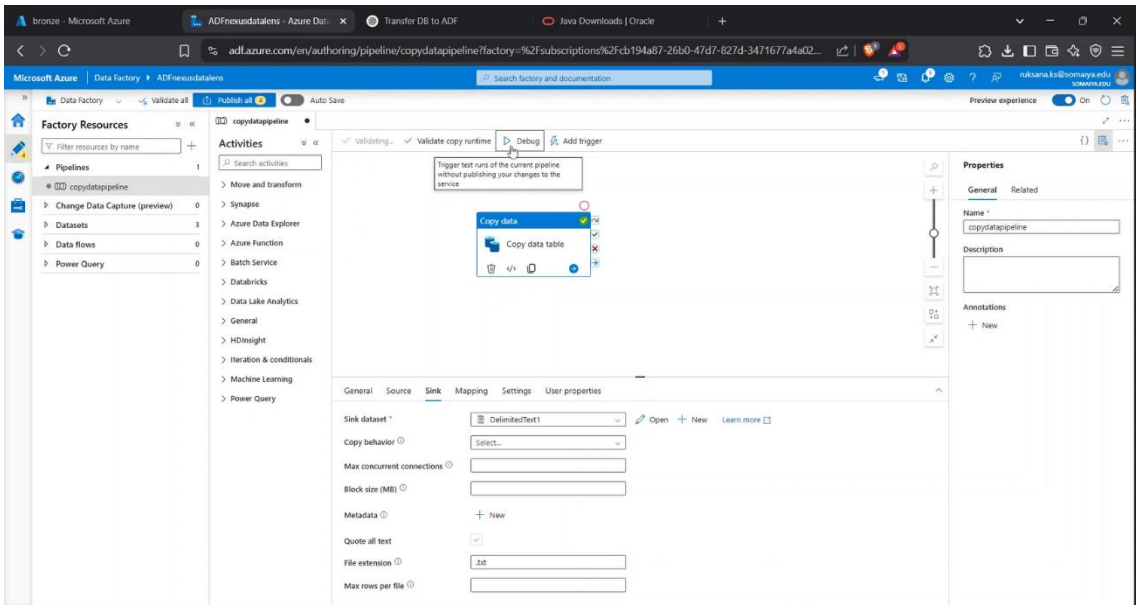
The **nexusdatalakestorage** account contains a single container named bronze with private access and is currently in an available lease state.



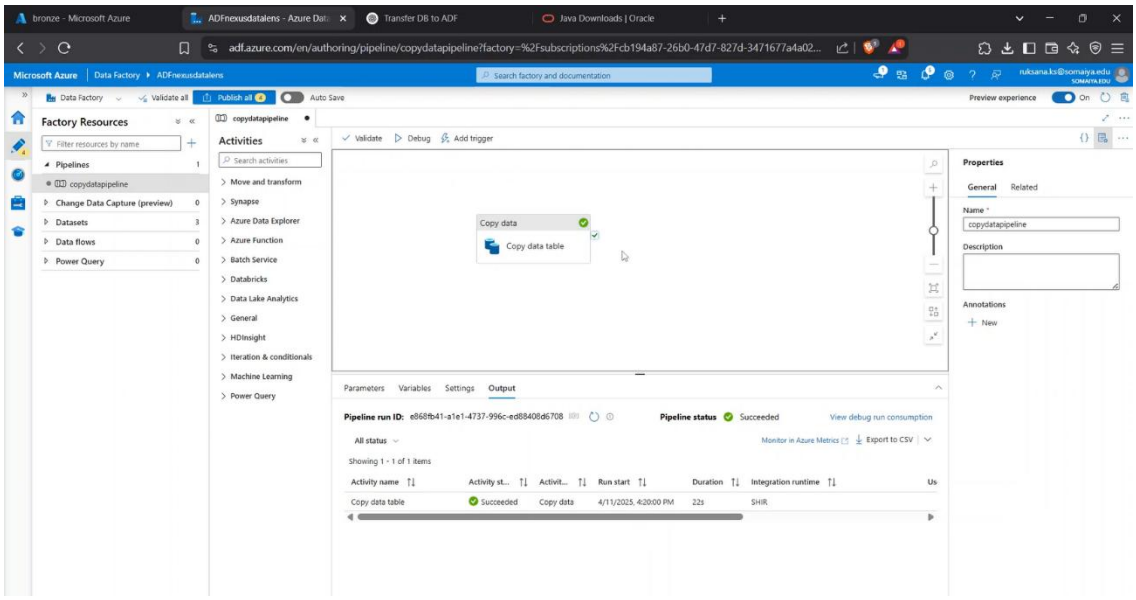
The bronze container in the nexusdatalakestorage account contains a single block blob file named **dbo.data.txt**, stored in the hot access tier and currently available.



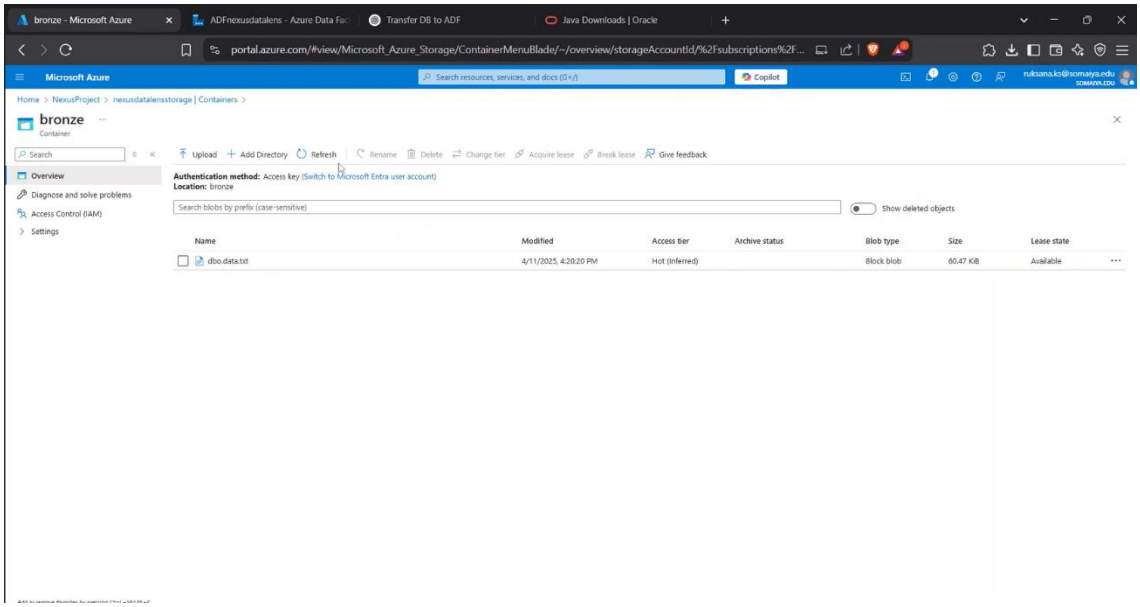
The sink settings of the "**Copy data**" activity in the Azure Data Factory pipeline are configured to write data to a dataset named **DelimitedText1**.



The Azure Data Factory pipeline "**copydatapipeline**" successfully completed the "**Copy data table**" activity in 22 seconds.

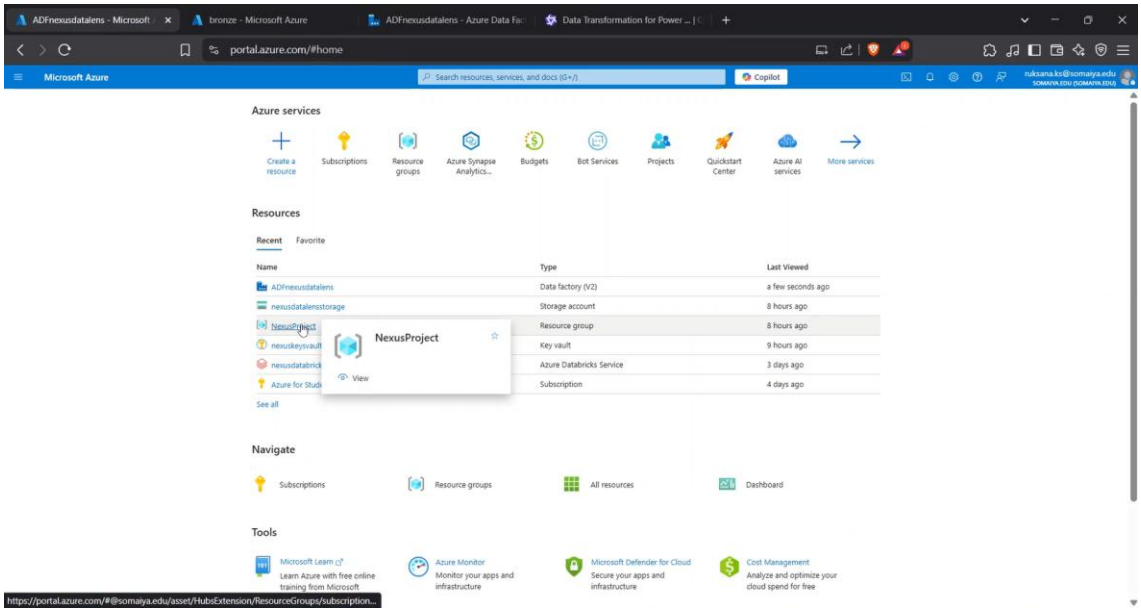


The Azure Storage container "**bronze**" contains a file named **dbo.data.txt**, uploaded on 4/11/2025, with a size of 60.47 KB and access tier set to hot.

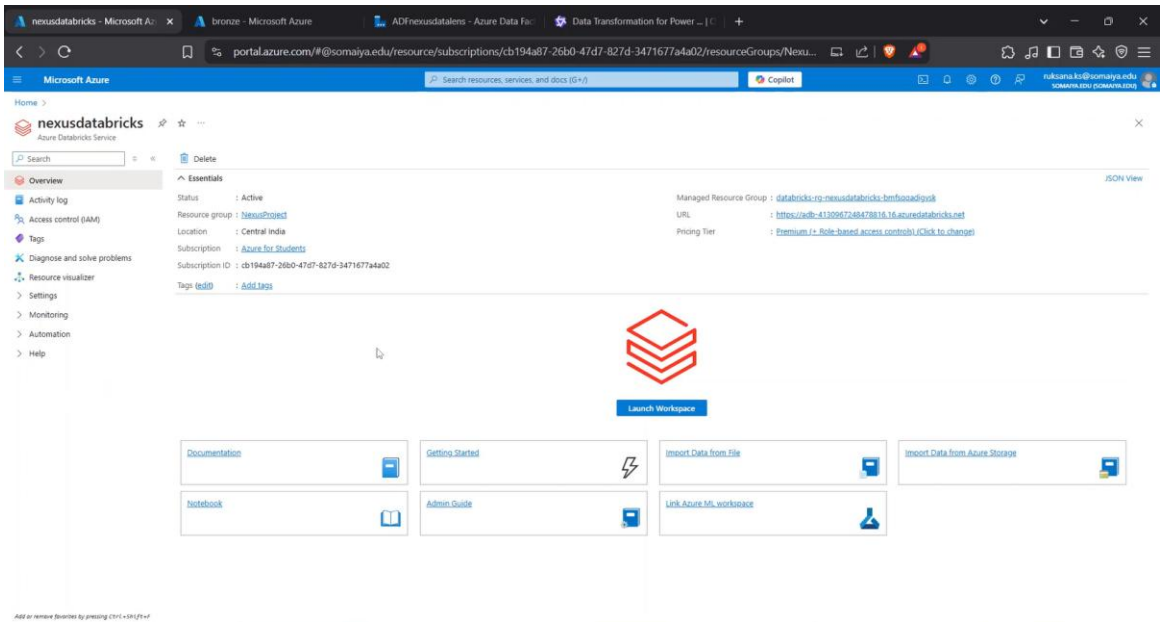


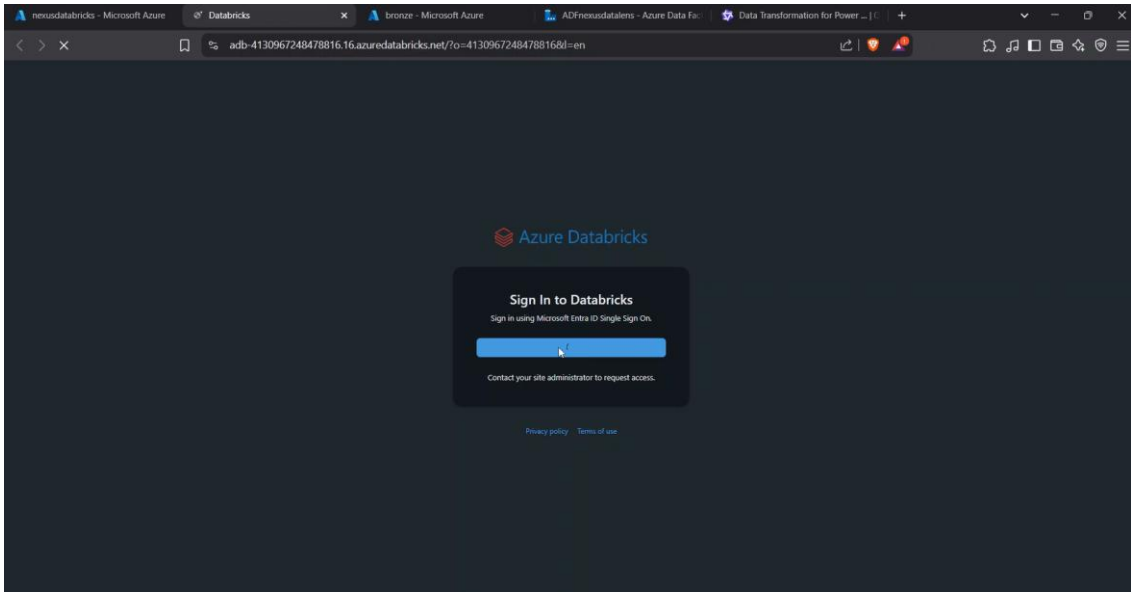
Transformation

The Azure portal screen shows a list of recently accessed resources, including a Data Factory, Storage Account, and Databricks service. Click on the "**NexusDatabricks**" resource group to view or manage related Azure resources.

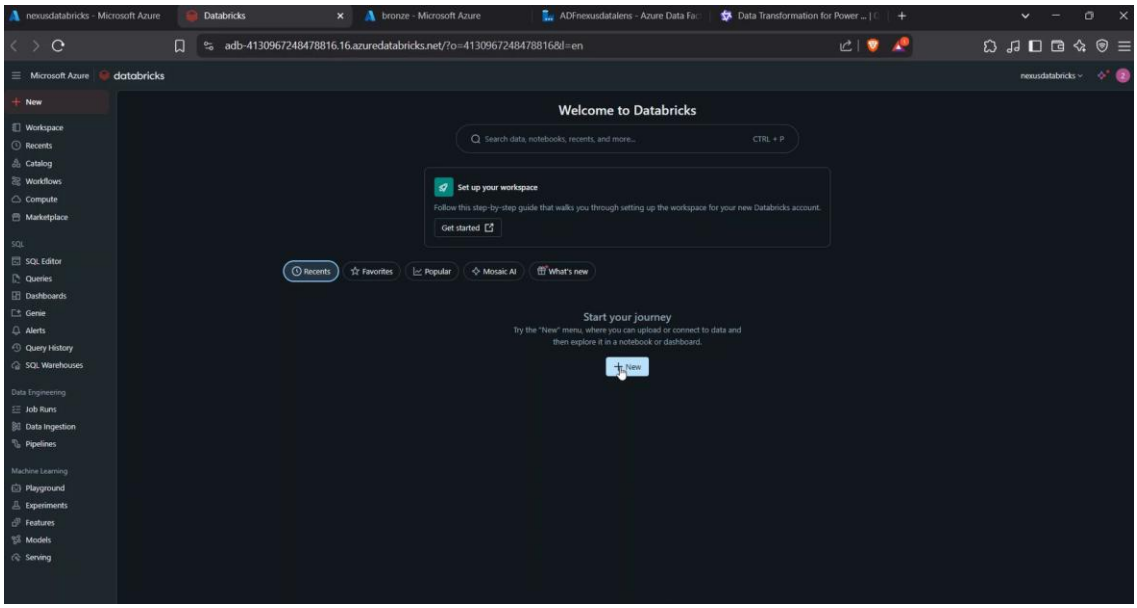


The Azure Databricks service overview for the "**nexusdatabricks**" resource, with an active status, **launch the workspace**.

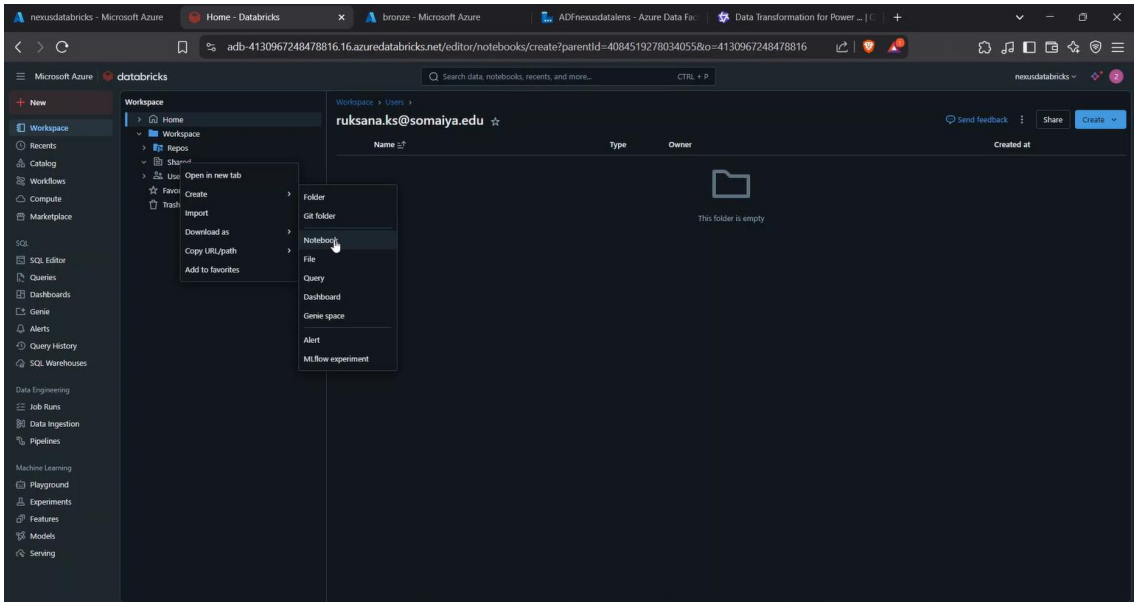




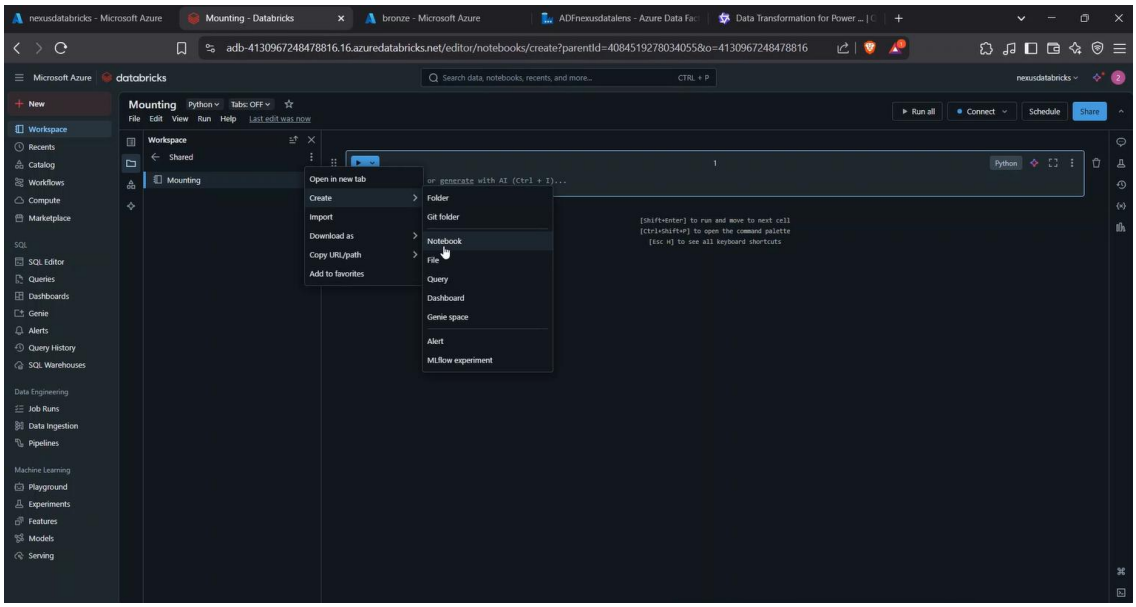
The Databricks workspace home screen where users can start their data journey by creating notebooks, dashboards, or other resources via the "**New**" button.



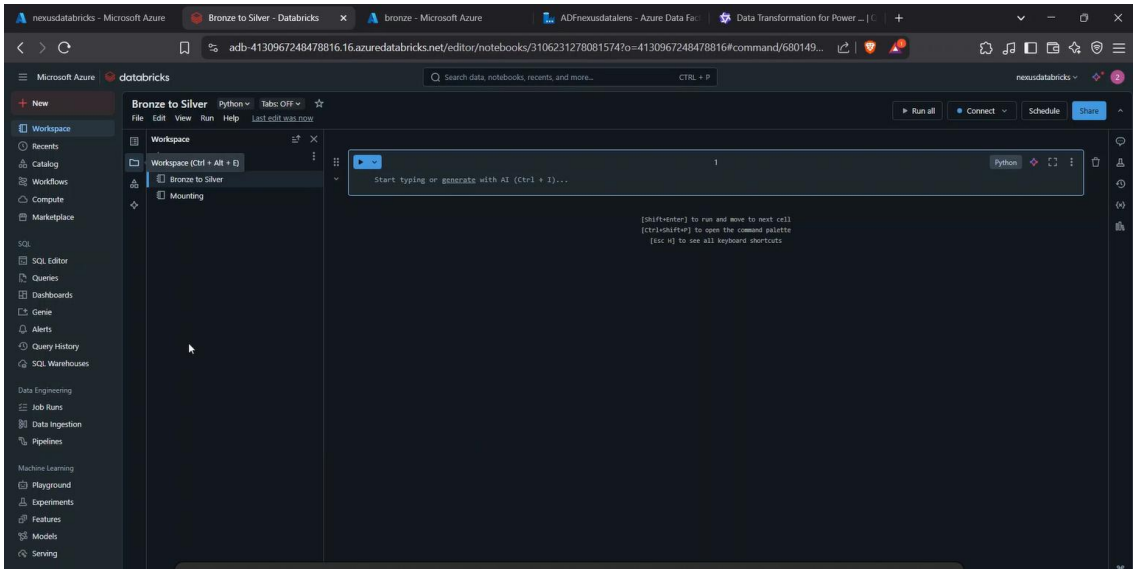
Creata a new Databricks notebook from the Workspace section by selecting "**Create > Notebook.**"



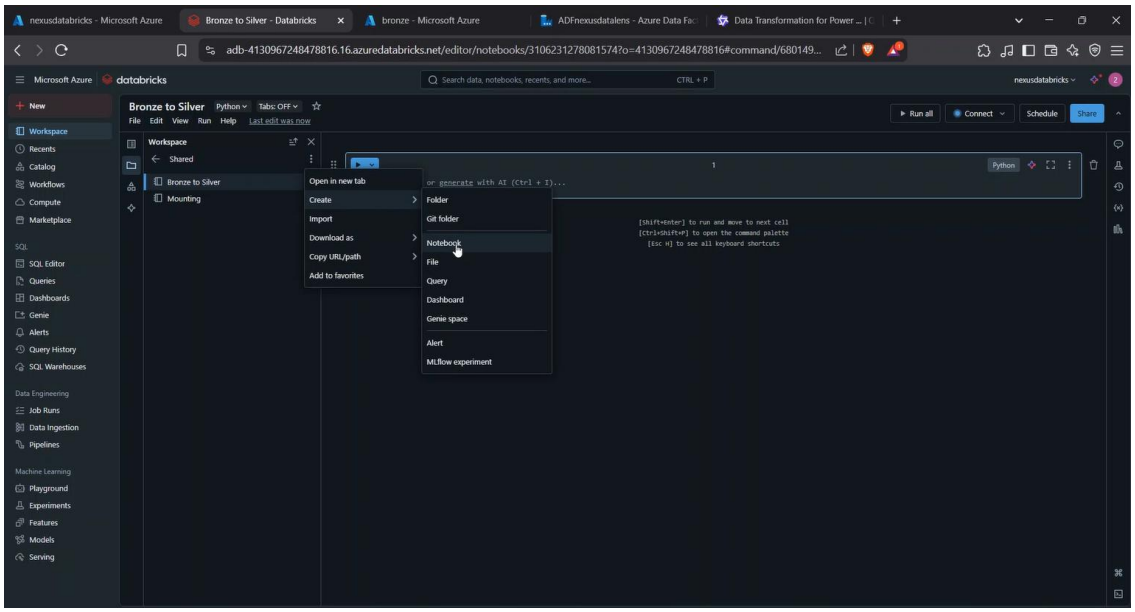
Open Databricks notebook then naming the notebook as “**Mounting**” and using the workspace sidebar menu to create another new notebook.



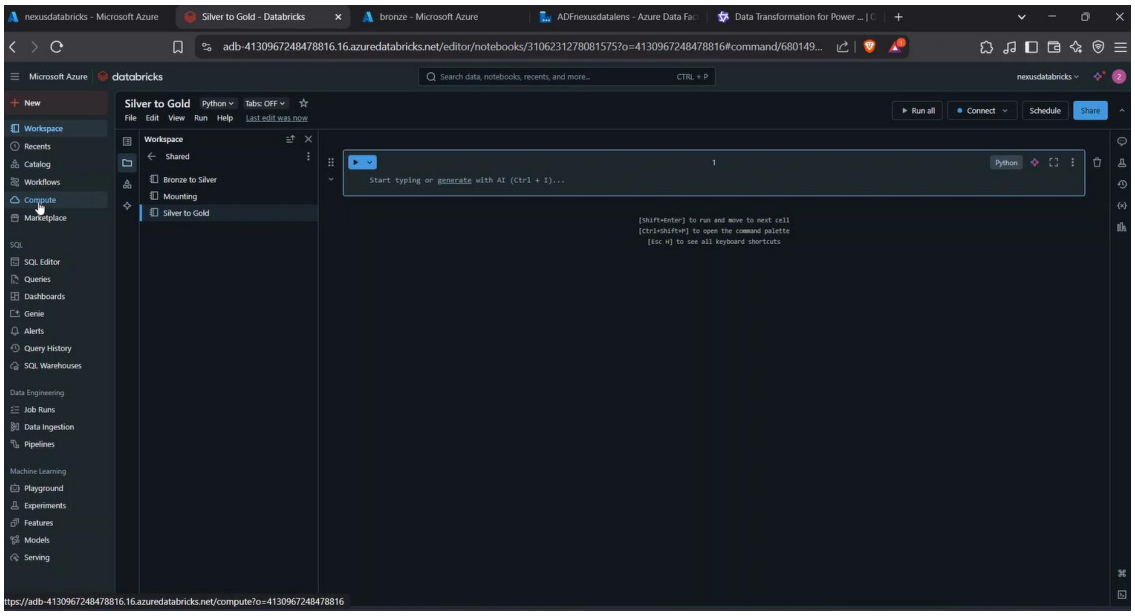
A Databricks notebook named "**Bronze to Silver**" open and ready for code input in the Python language.



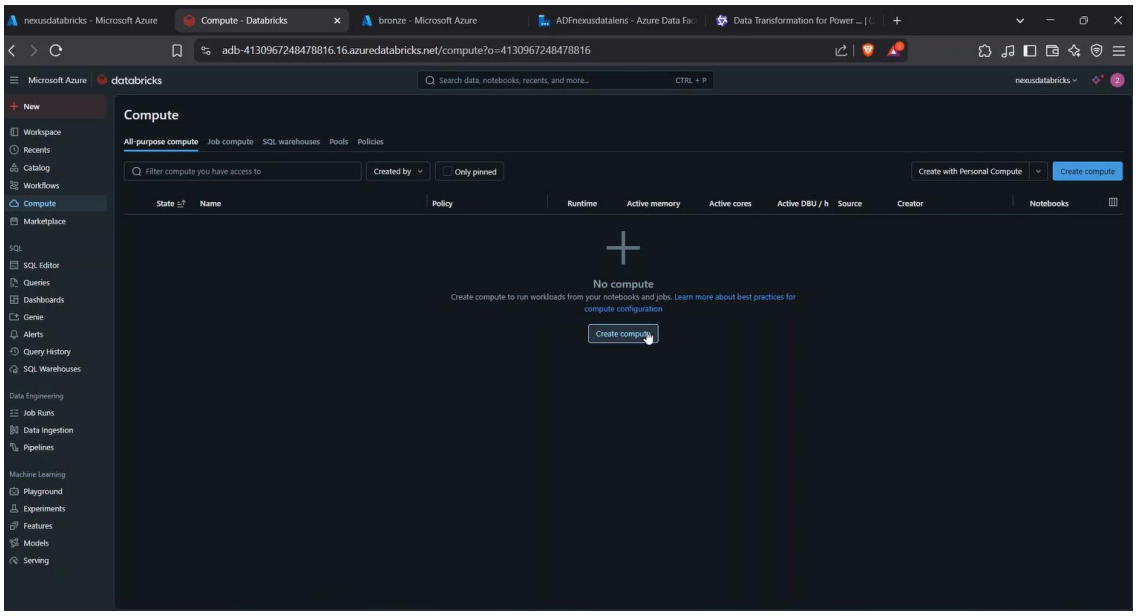
The user in Databricks selecting the option to create a new notebook from the "**Create**" menu under the Workspace.



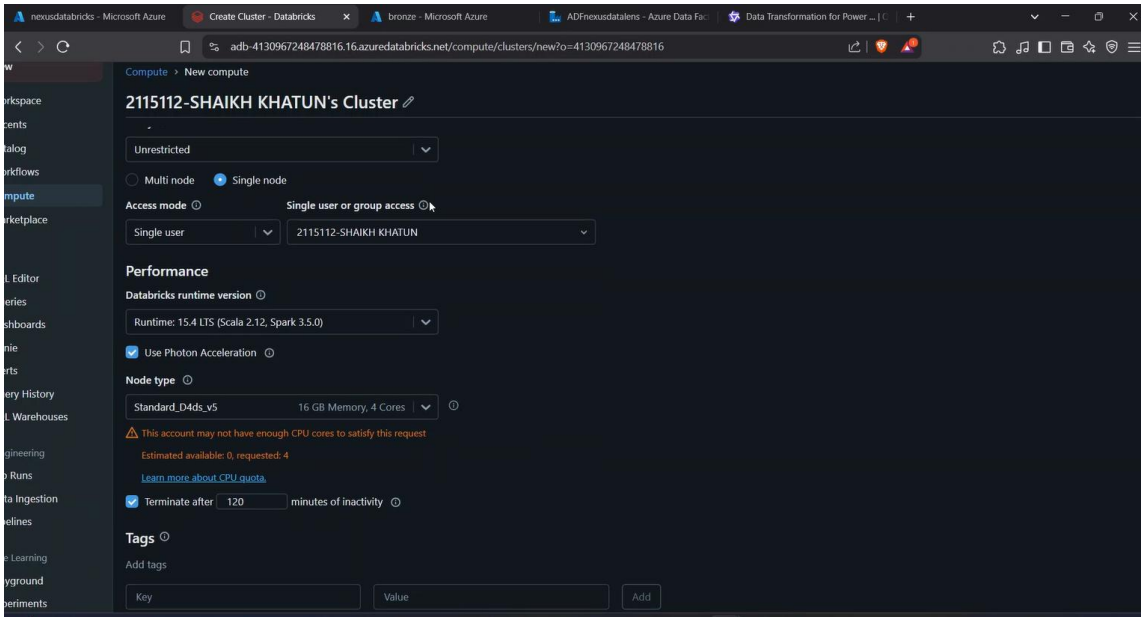
A Databricks notebook named " **Silver to Gold**" open and ready for code input in the Python language.



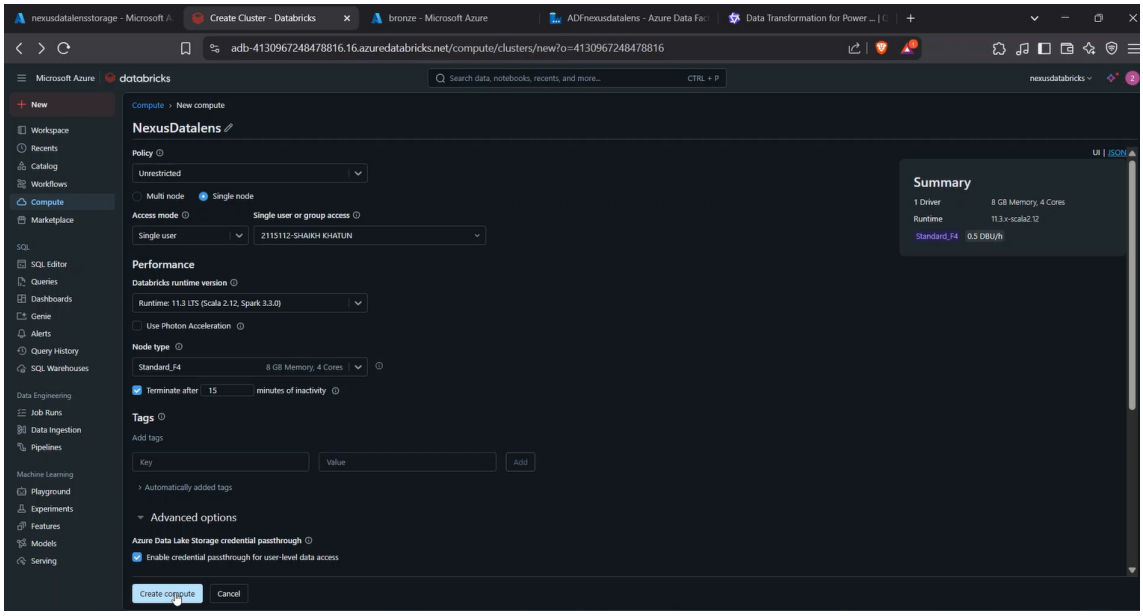
The Databricks **Compute** page in Azure, where a user is about to create a new compute cluster by clicking **Create compute**.



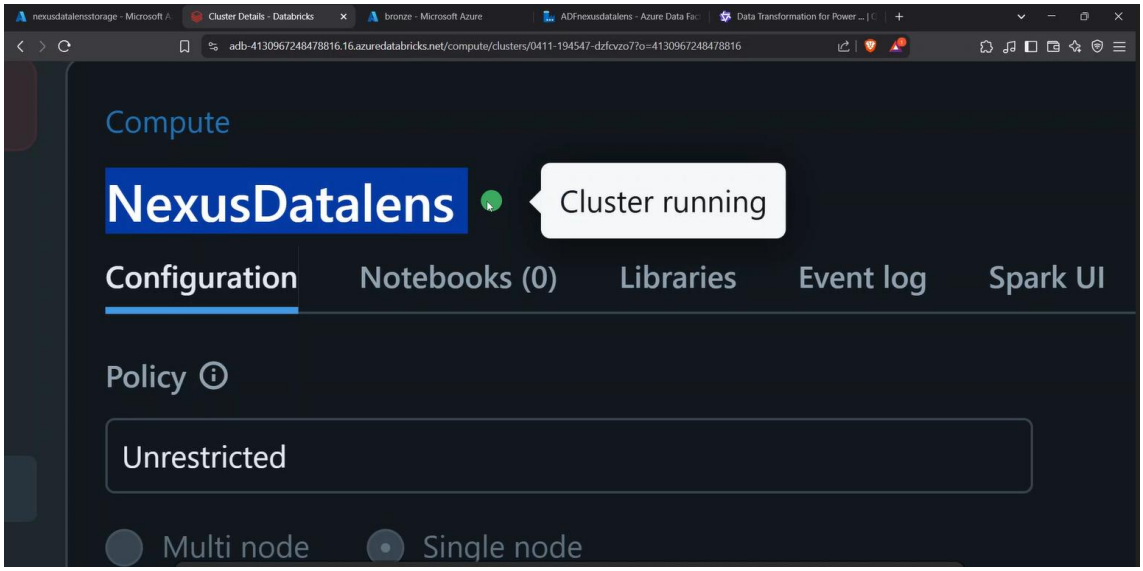
The configuration page for creating a new Databricks cluster in Azure, where the user is setting options like cluster name, access mode, runtime version, and node type.Select "**Single node**"



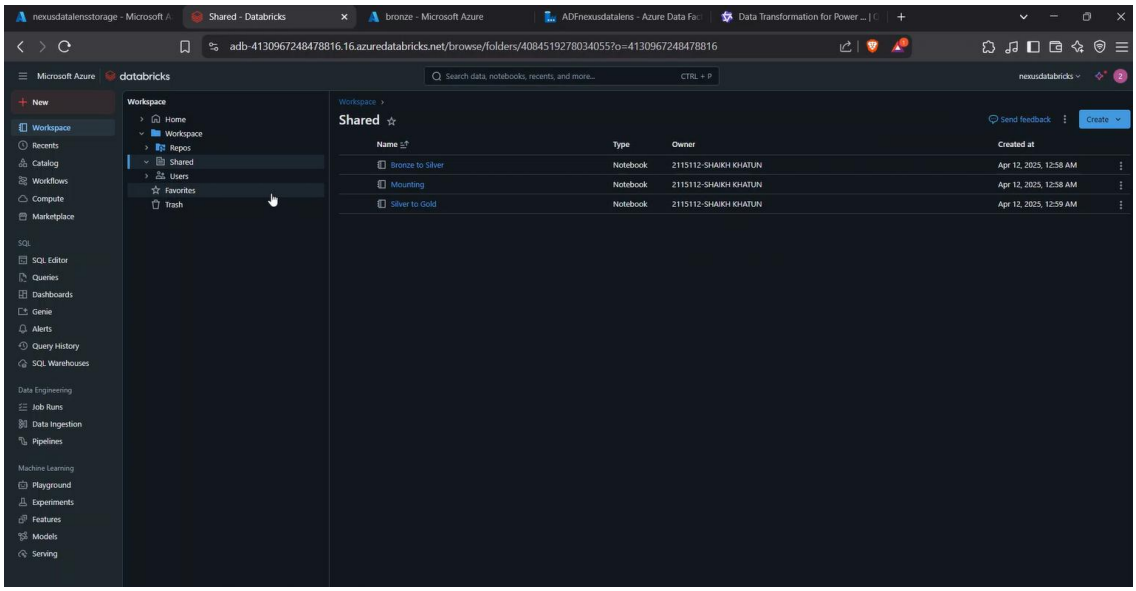
The final step of creating a single-node Azure Databricks cluster named **NexusDatalens** with **Runtime 11.3 LTS (Scala 2.12, Spark 3.3.0), Standard_F4** node type (8 GB memory, 4 cores), 15-minute auto-termination, and Azure Data Lake credential passthrough enabled before clicking **Create Compute**.



This confirms that **the NexusDatalens** Databricks cluster is successfully running with an **Unrestricted** policy and accessible configuration options.



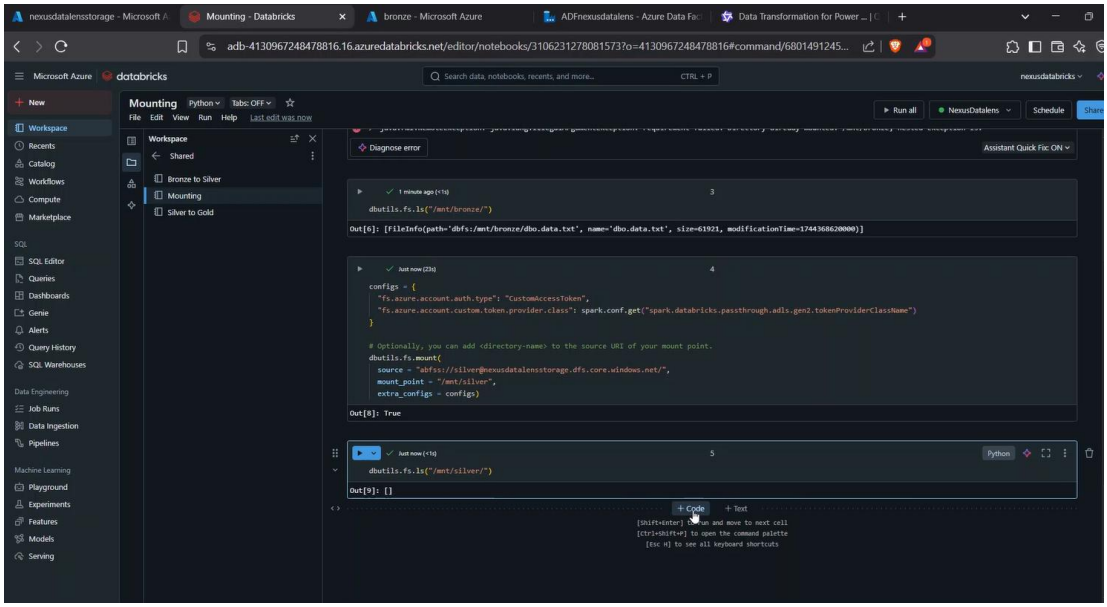
The **Shared** workspace folder in Azure Databricks containing three notebooks: **Bronze to Silver**, **Mounting**, and **Silver to Gold**, all owned by the same user.

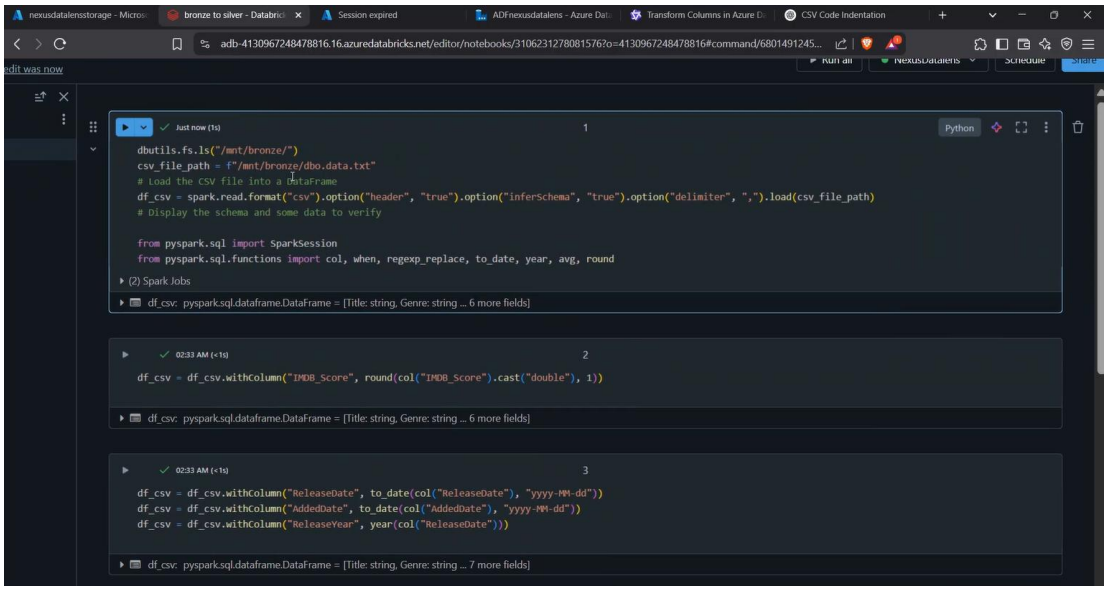


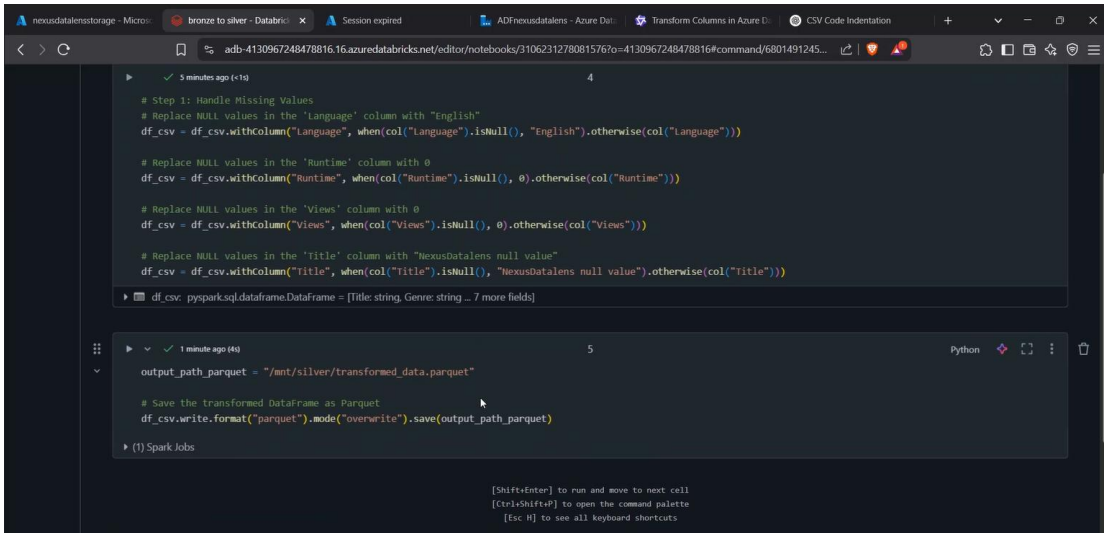
Mounting Code

```
# Databricks notebook source
# MAGIC %md
# MAGIC Mounting Bronze continer

# COMMAND ----------

# Databricks notebook source
configs = {
  "fs.azure.account.auth.type": "CustomAccessToken",
  "fs.azure.account.custom.token.provider.class": spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}


# COMMAND ----------

# Optionally, you can add <directory-name> to the source URI of your mount point.
dbutils.fs.mount(
  source = "abfss://bronze@nexusdatalensstorage.dfs.core.windows.net/",
  mount_point = "/mnt/bronze",
  extra_configs = configs)

# COMMAND ----------

dbutils.fs.ls("/mnt/bronze/")

# COMMAND ----------

# MAGIC %md
# MAGIC Mounting Silver continer

# COMMAND ----------

configs = {
  "fs.azure.account.auth.type": "CustomAccessToken",

  "fs.azure.account.custom.token.provider.class": spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

# Optionally, you can add <directory-name> to the source URI of your mount point.
dbutils.fs.mount(
  source = "abfss://silver@nexusdatalensstorage.dfs.core.windows.net/",
  mount_point = "/mnt/silver",
  extra_configs = configs)

# COMMAND ----------

dbutils.fs.ls("/mnt/silver/")

# COMMAND ----------

# MAGIC %md
# MAGIC Mounting Gold Continer

# COMMAND ----------

configs = {
  "fs.azure.account.auth.type": "CustomAccessToken",
  "fs.azure.account.custom.token.provider.class": spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

# Optionally, you can add <directory-name> to the source URI of your mount point.
dbutils.fs.mount(
  source = "abfss://gold@nexusdatalensstorage.dfs.core.windows.net/",
  mount_point = "/mnt/gold",
  extra_configs = configs)
```

Code Screenshot:



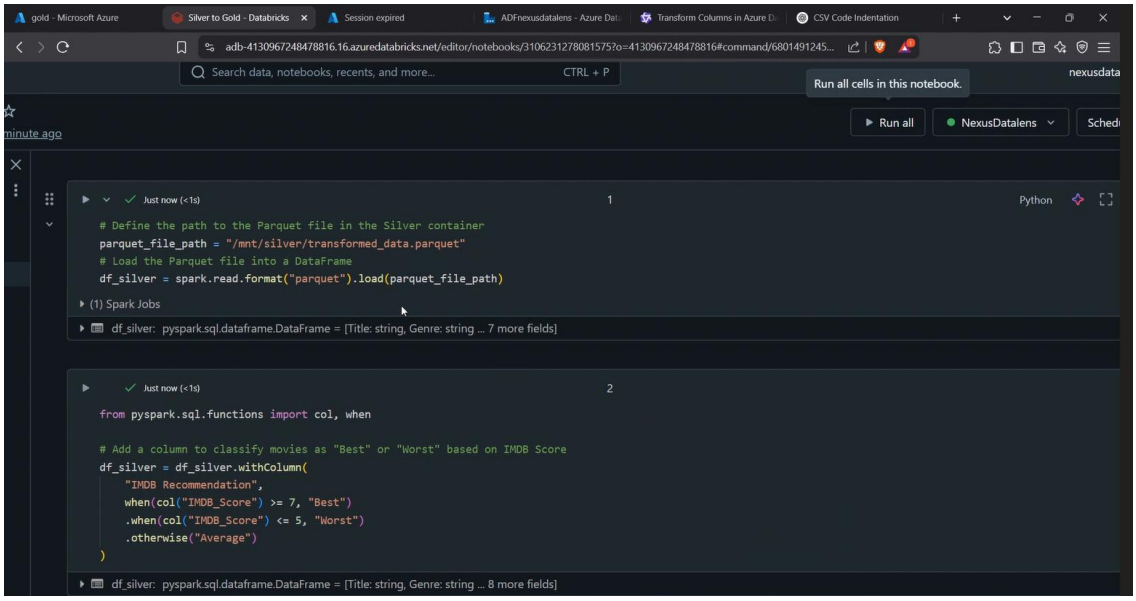


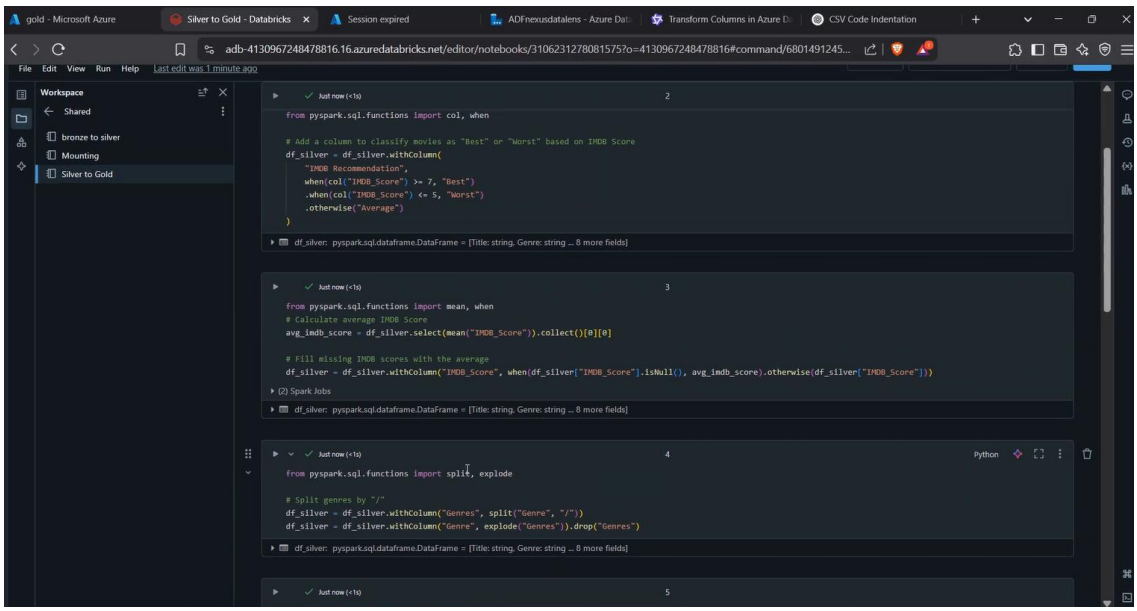


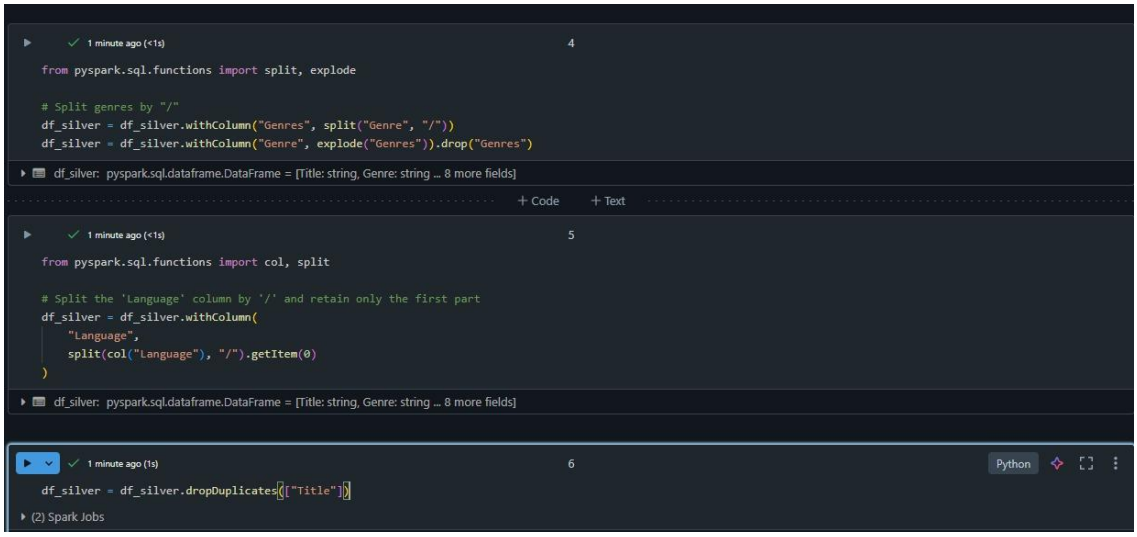
Silver to Gold Code

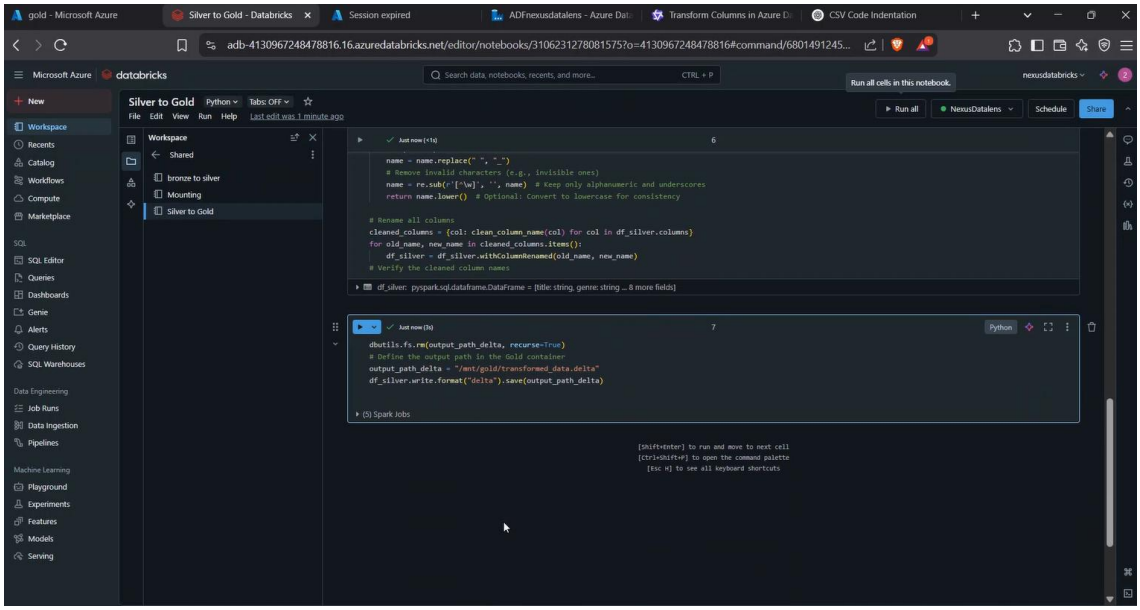
```
# Databricks notebook source
# Define the path to the Parquet file in the Silver container
parquet_file_path = "/mnt/silver/transformed_data.parquet"
# Load the Parquet file into a DataFrame
df_silver = spark.read.format("parquet").load(parquet_file_path)

# COMMAND ----------

from pyspark.sql.functions import col, when

# Add a column to classify movies as "Best" or "Worst" based on IMDB Score
df_silver = df_silver.withColumn(
    "IMDB Recommendation",
    when(col("IMDB_Score") >= 7, "Best")
    .when(col("IMDB_Score") <= 5, "Worst")
    .otherwise("Average")
)

# COMMAND ----------

from pyspark.sql.functions import mean, when
# Calculate average IMDB Score
avg_imdb_score = df_silver.select(mean("IMDB_Score")).collect()[0][0]

# Fill missing IMDB scores with the average
df_silver = df_silver.withColumn("IMDB_Score", when(df_silver["IMDB_Score"].isNull(), avg_imdb_score).otherwise(df_silver["IMDB_Score"]))

# COMMAND ----------

from pyspark.sql.functions import split, explode

# Split genres by "/"
df_silver = df_silver.withColumn("Genres", split("Genre", "/"))
df_silver = df_silver.withColumn("Genre", explode("Genres")).drop("Genres")


# COMMAND ----------

from pyspark.sql.functions import col, split

# Split the 'Language' column by '/' and retain only the first part
df_silver = df_silver.withColumn(
    "Language",
    split(col("Language"), "/").getItem(0)
)

# COMMAND ----------

import re

# Function to clean column names
def clean_column_name(name):
    # Remove leading/trailing whitespace
    name = name.strip()
    # Replace spaces with underscores
    name = name.replace(" ", "_")
    # Remove invalid characters (e.g., invisible ones)
    name = re.sub(r'[^\w]', '', name)  # Keep only alphanumeric and underscores
    return name.lower()  # Optional: Convert to lowercase for consistency

# Rename all columns
cleaned_columns = {col: clean_column_name(col) for col in df_silver.columns}
for old_name, new_name in cleaned_columns.items():
    df_silver = df_silver.withColumnRenamed(old_name, new_name)
# Verify the cleaned column names

# COMMAND ----------

dbutils.fs.rm(output_path_delta, recurse=True)
# Define the output path in the Gold container
output_path_delta = "/mnt/gold/transformed_data.delta"
df_silver.write.format("delta").save(output_path_delta)
```

Code Screenshot:





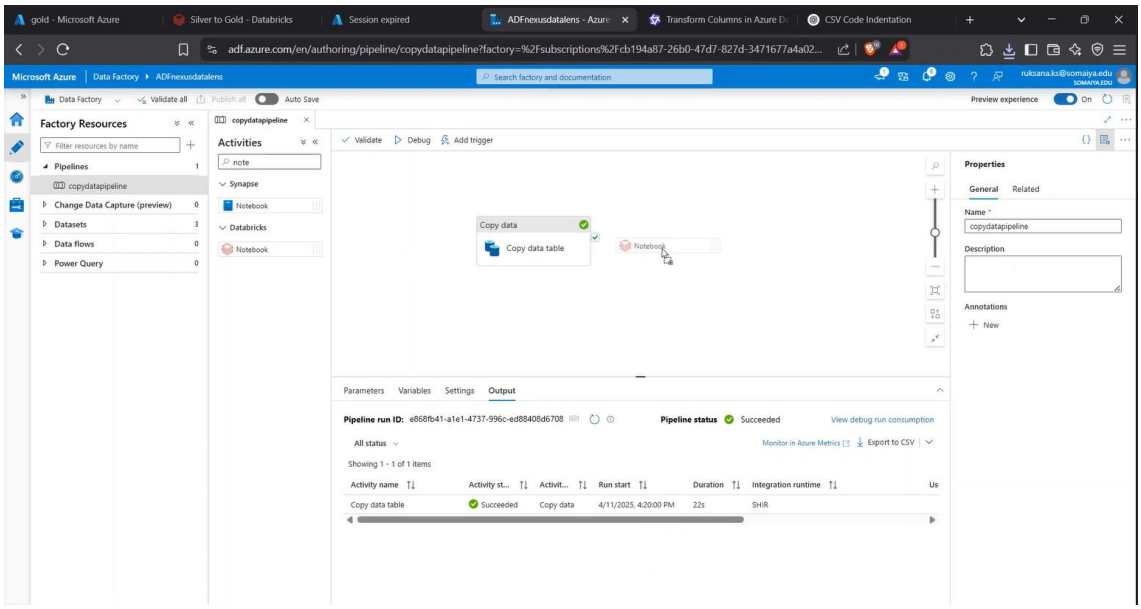




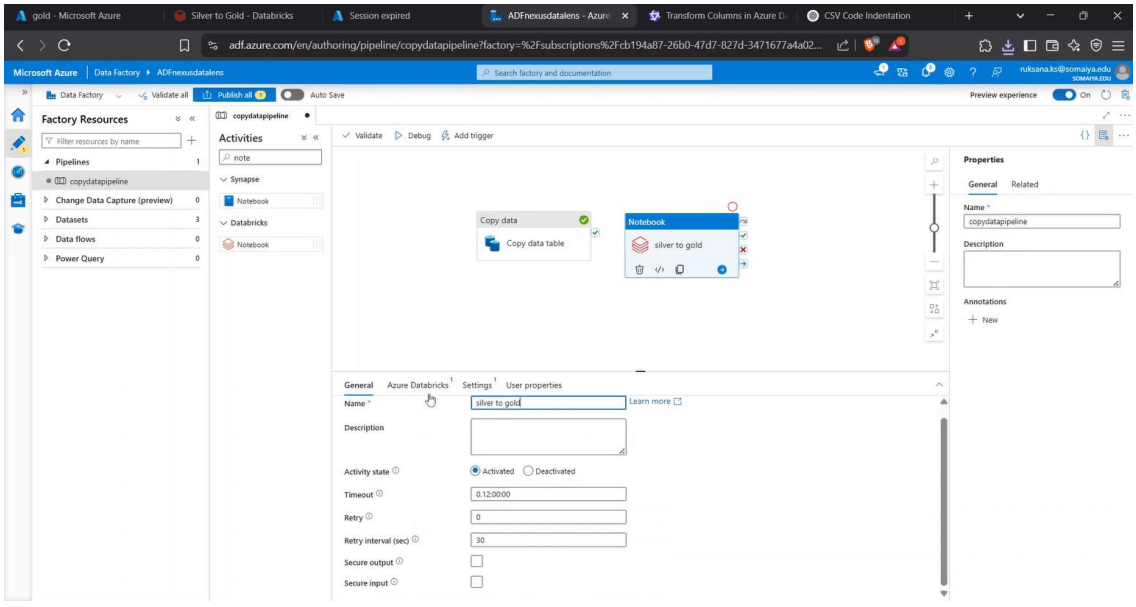
Bronze to Silver Code:

```
# Databricks notebook source
dbutils.fs.ls("/mnt/bronze/")
csv_file_path = f"/mnt/bronze/dbo.data.txt"
# Load the CSV file into a DataFrame
df_csv = spark.read.format("csv").option("header", "true").option("inferSchema", "true").option("delimiter", ",").load(csv_file_path)
# Display the schema and some data to verify

from pyspark.sql import SparkSession
from pyspark.sql.functions import col, when, regexp_replace, to_date, year, avg, round

# COMMAND ----------

df_csv = df_csv.withColumn("IMDB_Score", round(col("IMDB_Score").cast("double"), 1))


# COMMAND ----------

df_csv = df_csv.withColumn("ReleaseDate", to_date(col("ReleaseDate"), "yyyy-MM-dd"))
df_csv = df_csv.withColumn("AddedDate", to_date(col("AddedDate"), "yyyy-MM-dd"))
df_csv = df_csv.withColumn("ReleaseYear", year(col("ReleaseDate")))


# COMMAND ----------

# Step 1: Handle Missing Values
# Replace NULL values in the 'Language' column with "English"
df_csv = df_csv.withColumn("Language", when(col("Language").isNull(), "English").otherwise(col("Language")))

# Replace NULL values in the 'Runtime' column with 0
df_csv = df_csv.withColumn("Runtime", when(col("Runtime").isNull(), 0).otherwise(col("Runtime")))

# Replace NULL values in the 'Views' column with 0
df_csv = df_csv.withColumn("Views", when(col("Views").isNull(), 0).otherwise(col("Views")))

# Replace NULL values in the 'Title' column with "NexusDatalens null value"
df_csv = df_csv.withColumn("Title", when(col("Title").isNull(), "No Title").otherwise(col("Title")))

# COMMAND ----------

output_path_parquet = "/mnt/silver/transformed_data.parquet"

# Save the transformed DataFrame as Parquet
df_csv.write.format("parquet").mode("overwrite").save(output_path_parquet)
```

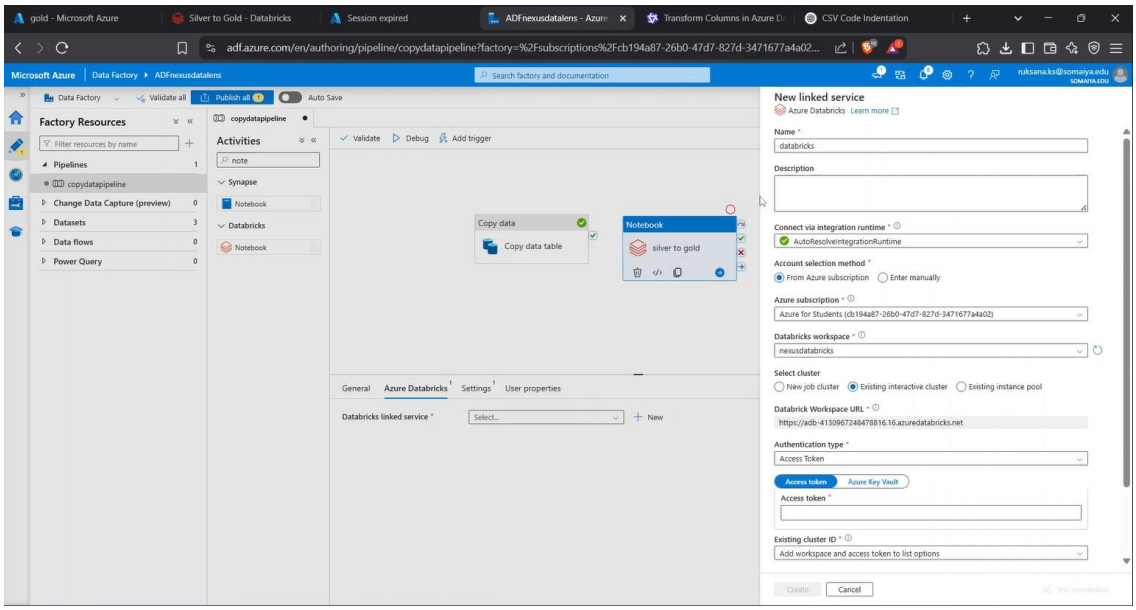
An Azure Data Factory pipeline named **copydatapipeline**, where a **Copy data** activity has successfully run to copy a data table in 22 seconds.



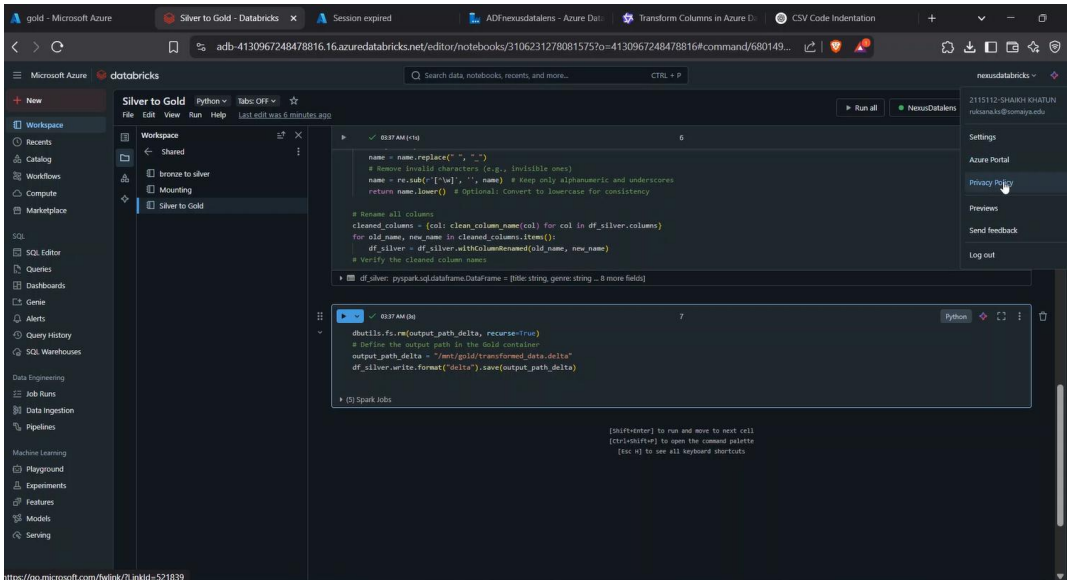
An Azure Data Factory pipeline where a **Databricks Notebook activity** named **Silver to Gold** is being configured. It follows a successfully executed **Copy data** activity, enabling a notebook-driven data transformation step.



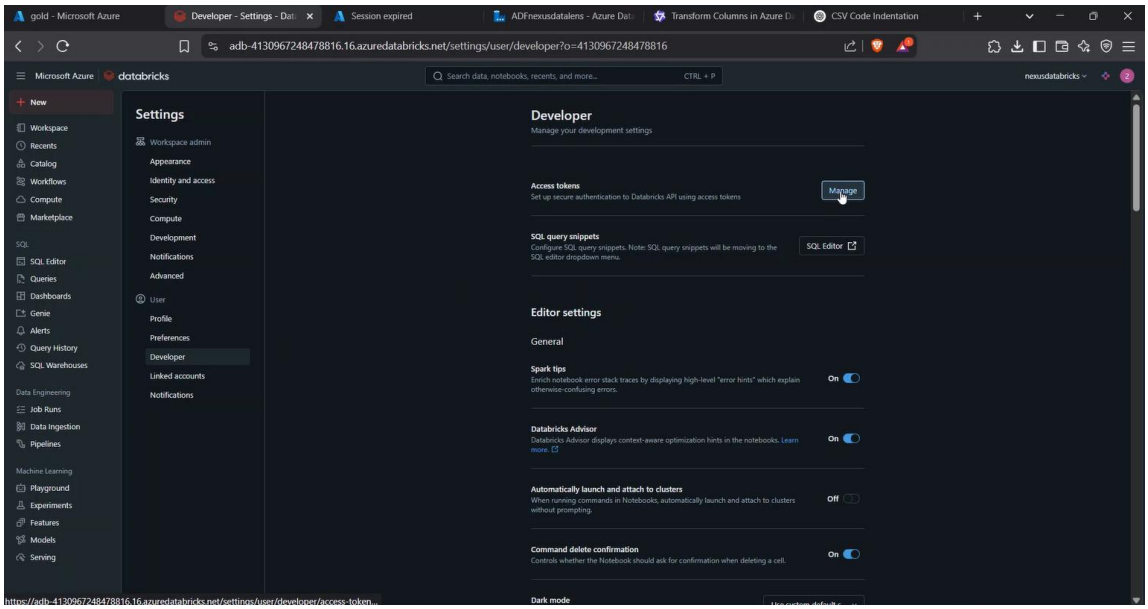
The setup of a new Azure Databricks linked service in Azure Data Factory, including configuration of workspace, authentication, and cluster options.



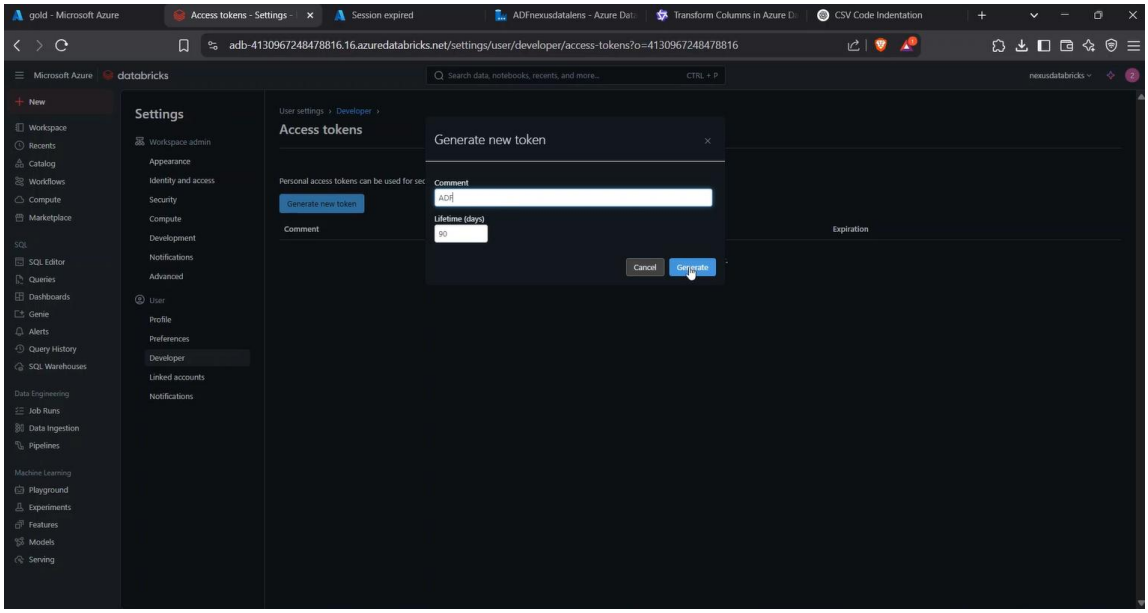
The image shows a Databricks notebook transforming and writing a cleaned DataFrame from Silver to Gold storage using PySpark.



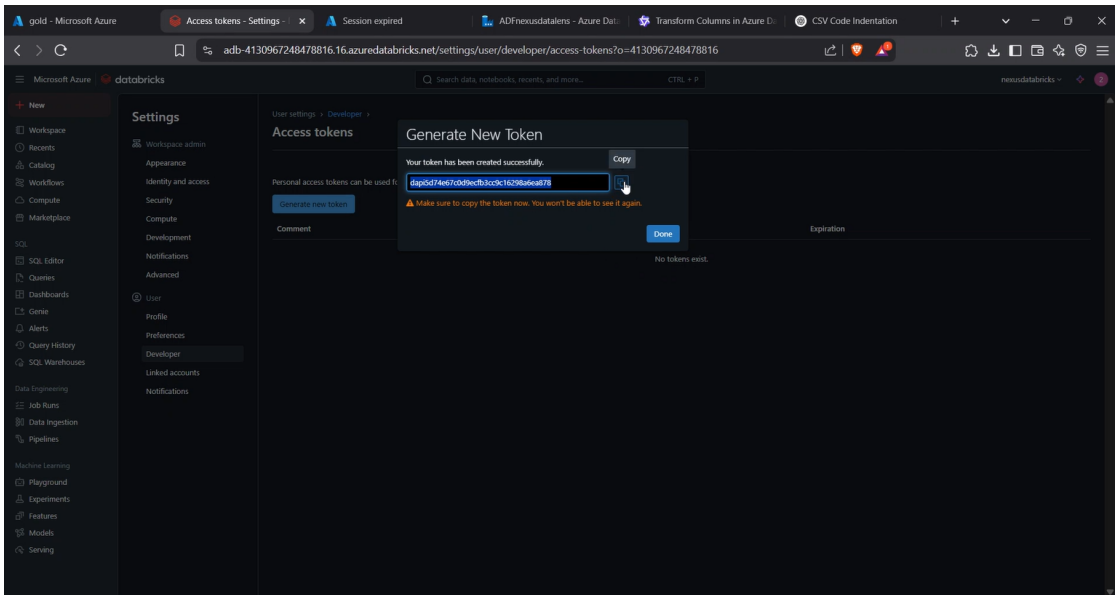
The Developer settings page in Databricks where users can manage access tokens and configure editor settings like Spark tips and dark mode.



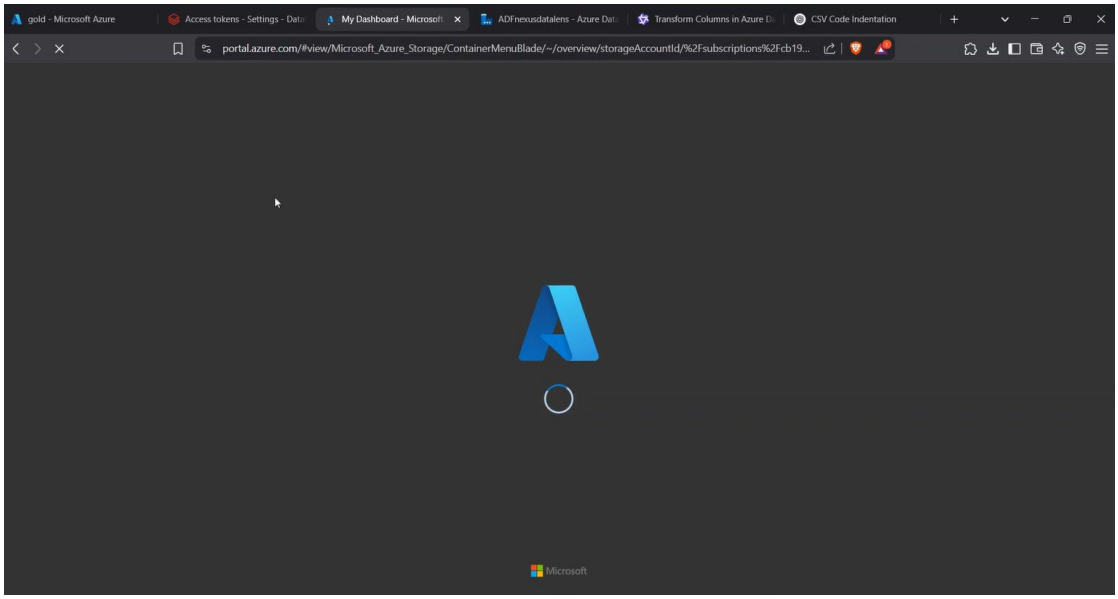
The image shows the process of generating a new personal access token in Databricks, which is used for secure authentication in integrations like Azure Data Factory.



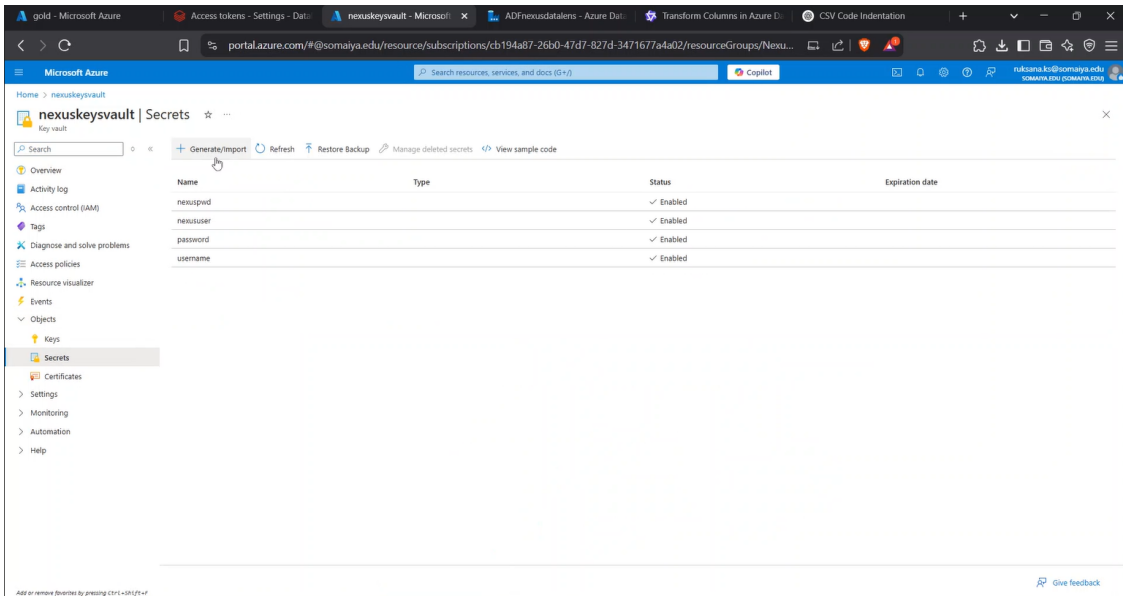
The image shows a newly generated Databricks access token, which must be copied immediately as it won't be visible again for future use.



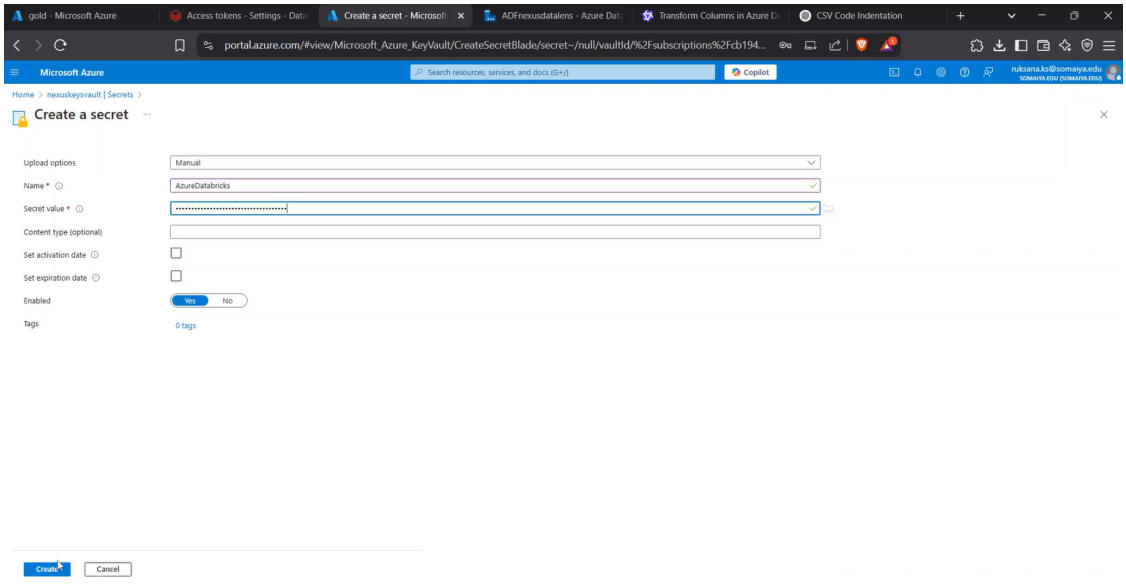
The Azure portal loading screen, likely indicating access to a storage container or resource for integration with Databricks or Data Factory.



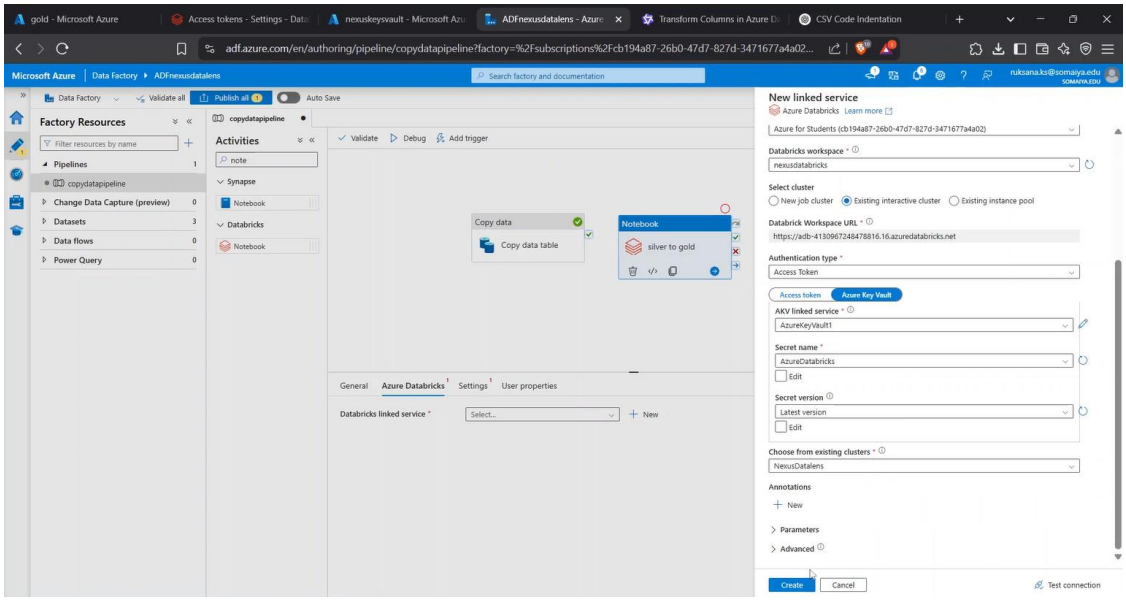
The image shows Azure Key Vault with secrets (like usernames and passwords) stored and enabled for secure access by integrated services like Azure Data Factory and Databricks.



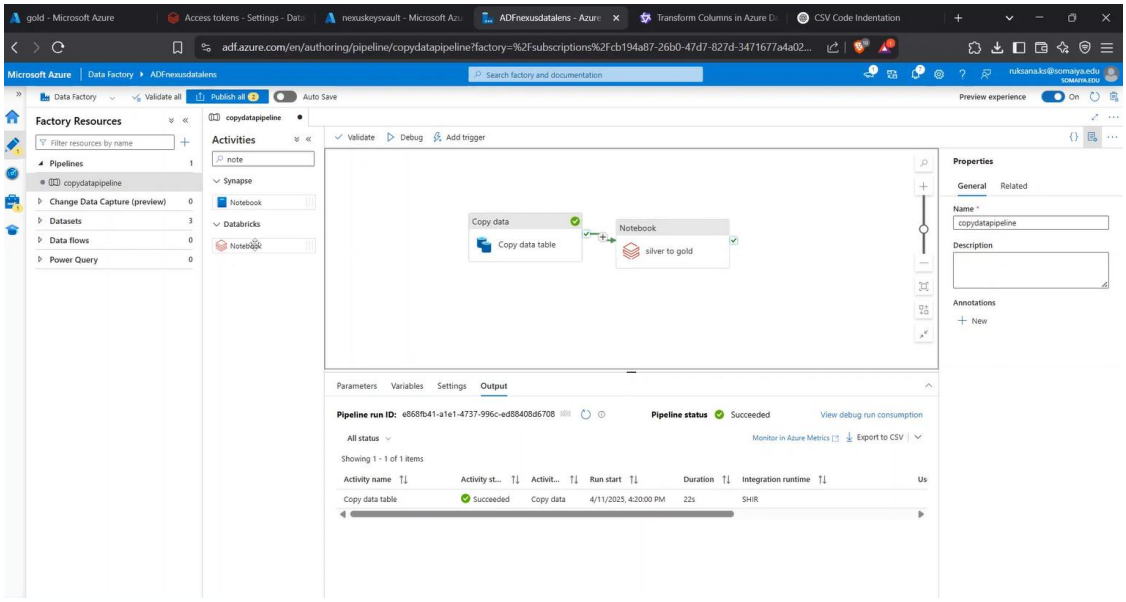
The image shows the creation of a new secret named "**AzureDatabricks**" in Azure Key Vault with manual upload and the secret value entered.

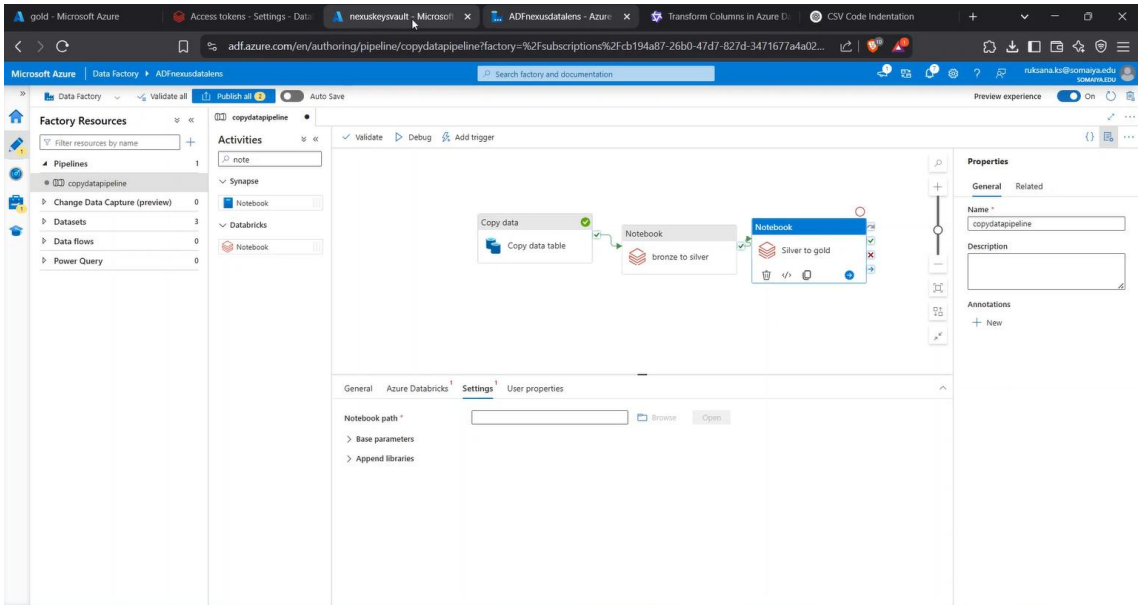


The creation of a new linked service in Azure Data Factory to connect with an Azure Databricks workspace (**nexusdatabricks)** using Azure Key Vault for token authentication. The Key Vault (**Azurekeyvault1**) is configured to retrieve the token stored under the secret name **AzureDatabricks**.

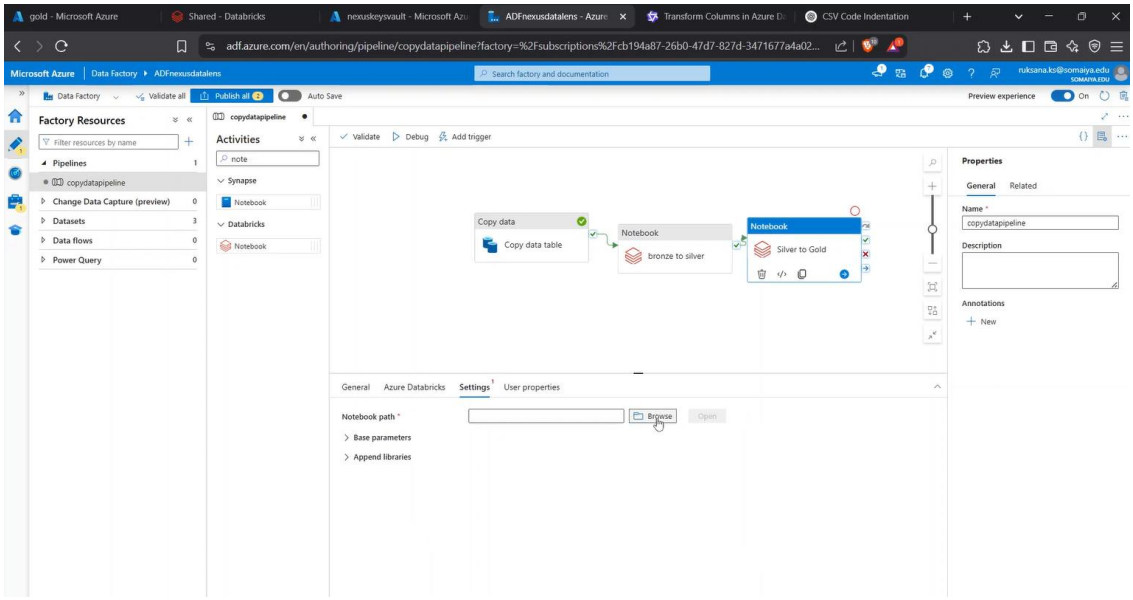


The screenshot shows a successful connection test to the Databricks linked service **(databricks)** in the Azure Data Factory pipeline named **copydatapipeline.**

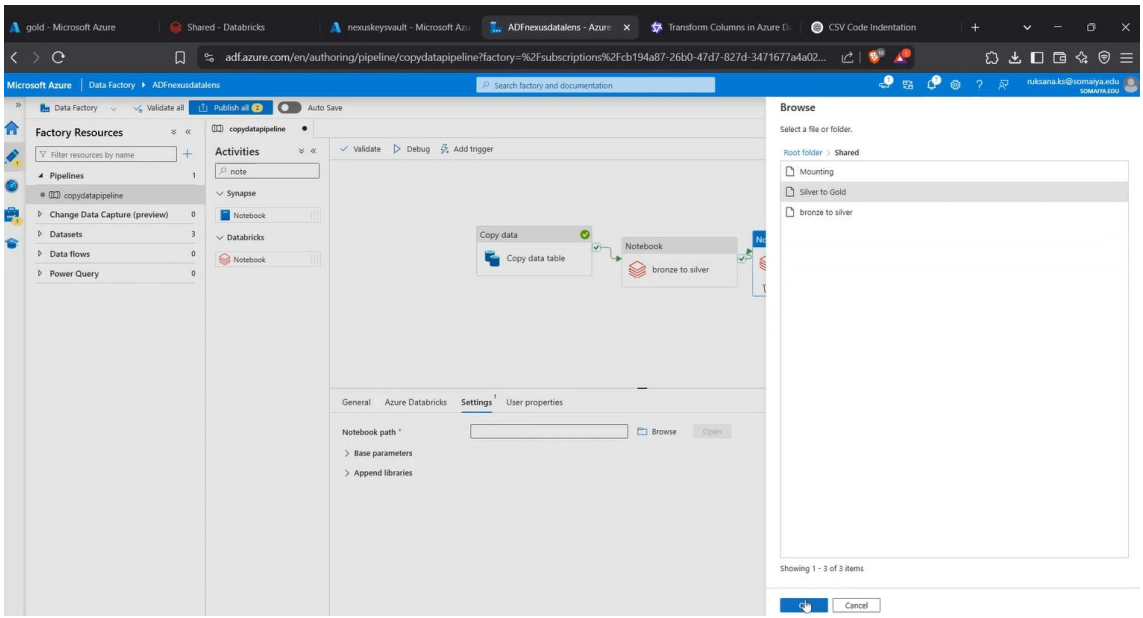




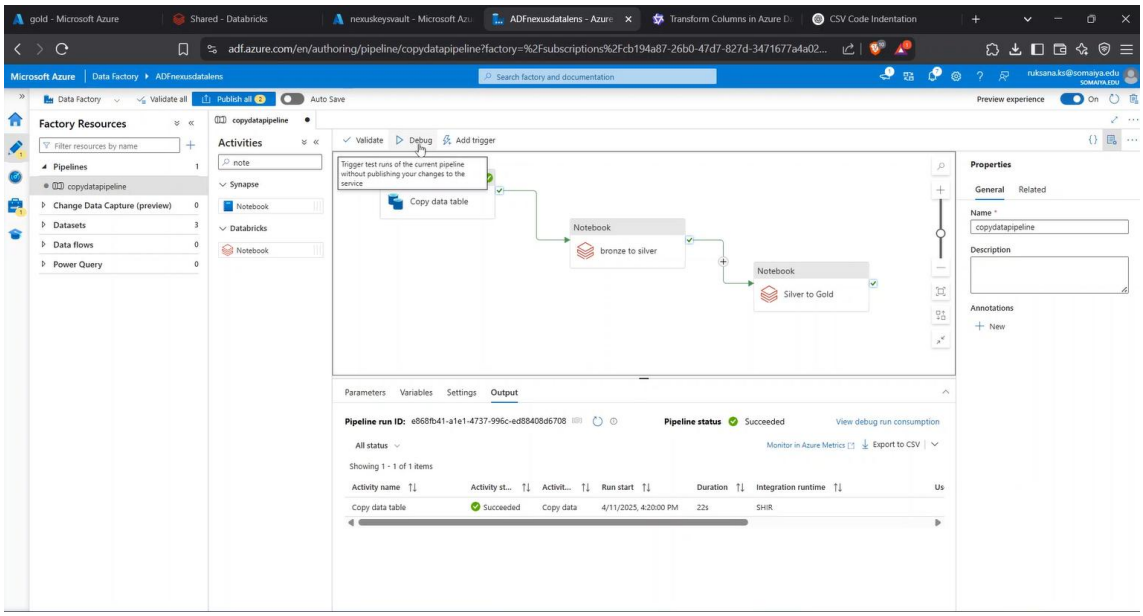
The Azure Data Factory pipeline copydatapipeline successfully connects to the databricks linked service to run two sequential notebooks: "**bronze to silver**" and "**silver to gold**"



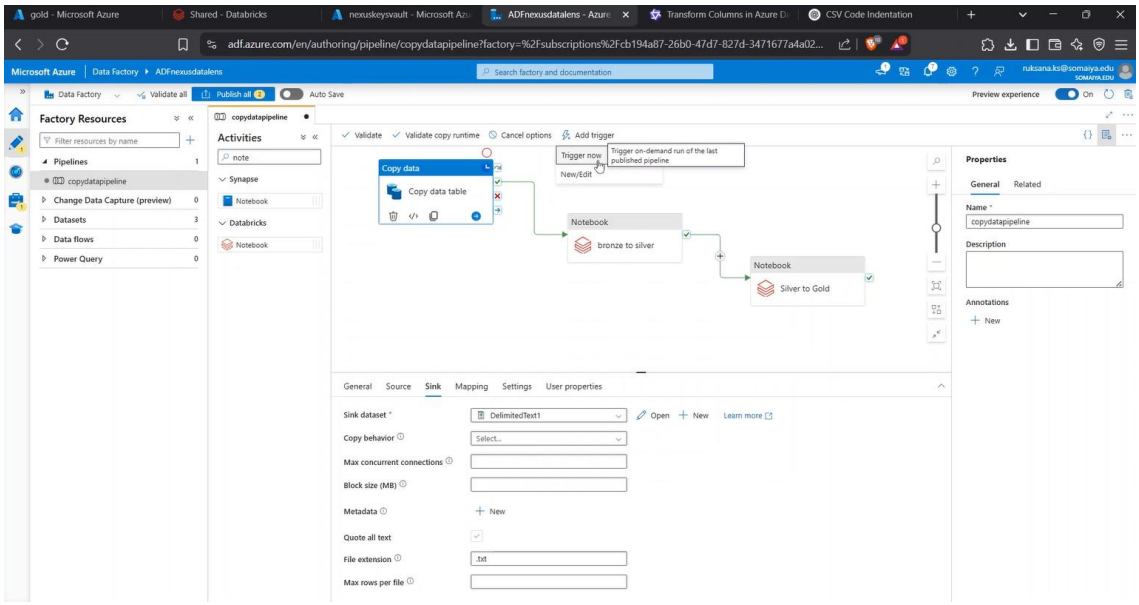
The selection of the **bronze to silver** Databricks notebook from the Shared folder in Azure Data Factory for use in the pipeline.



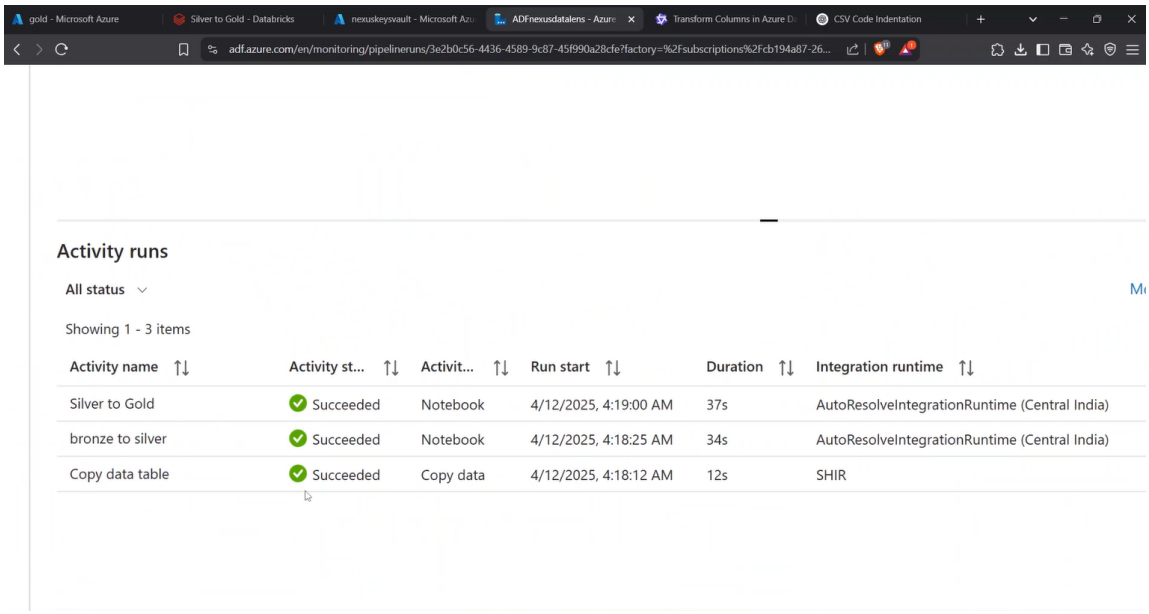
The Azure Data Factory pipeline **copydatapipeline** has successfully run all activities—copying data, and executing the **bronze to silver** and **silver to gold** Databricks notebooks.



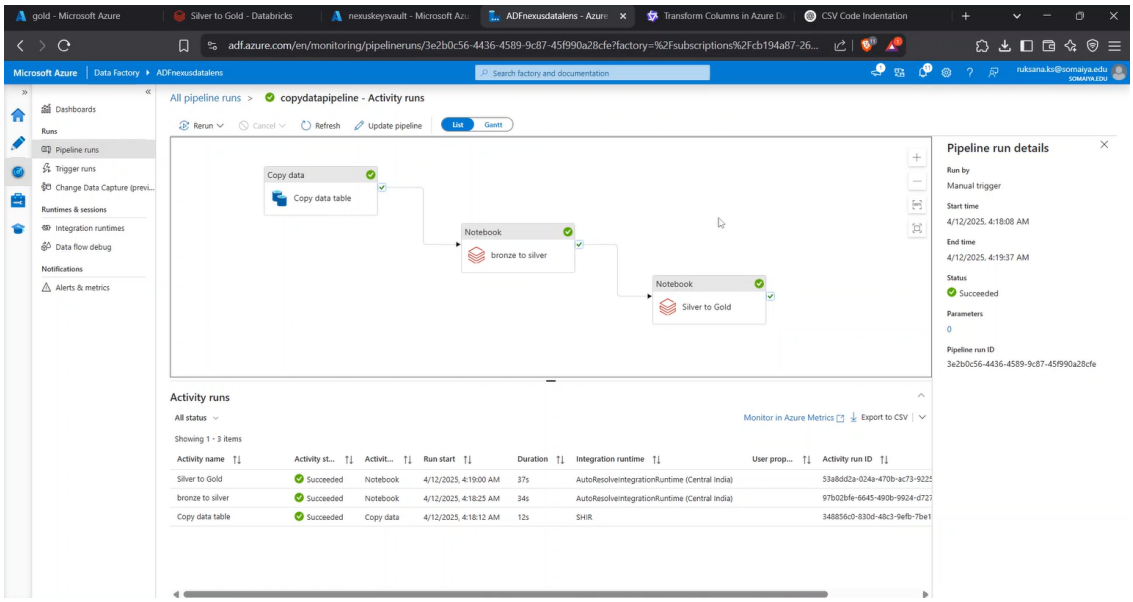
This Azure Data Factory pipeline copies data to a table, then processes it through two Databricks notebooks to transform it from bronze to silver, and finally to gold.



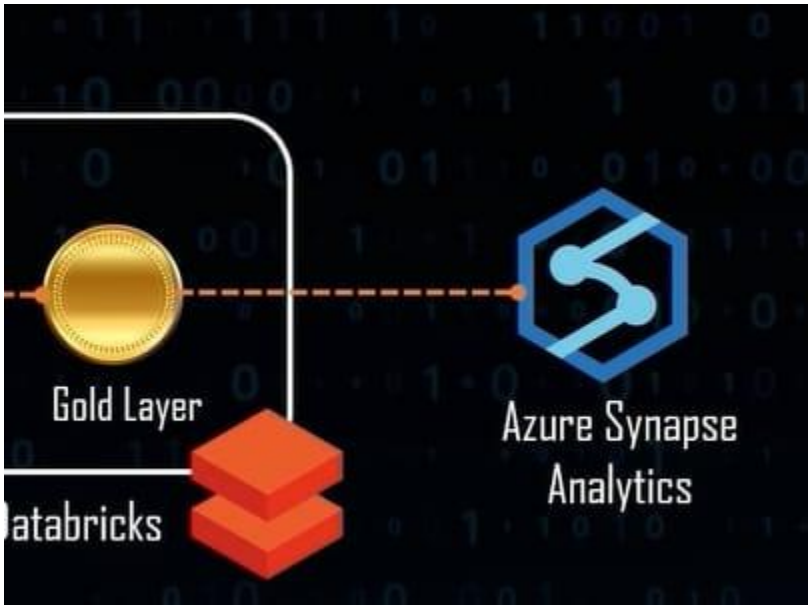
The pipeline successfully executed all three activities—data copy, bronze to silver, and silver to gold transformations—within a span of one minute.



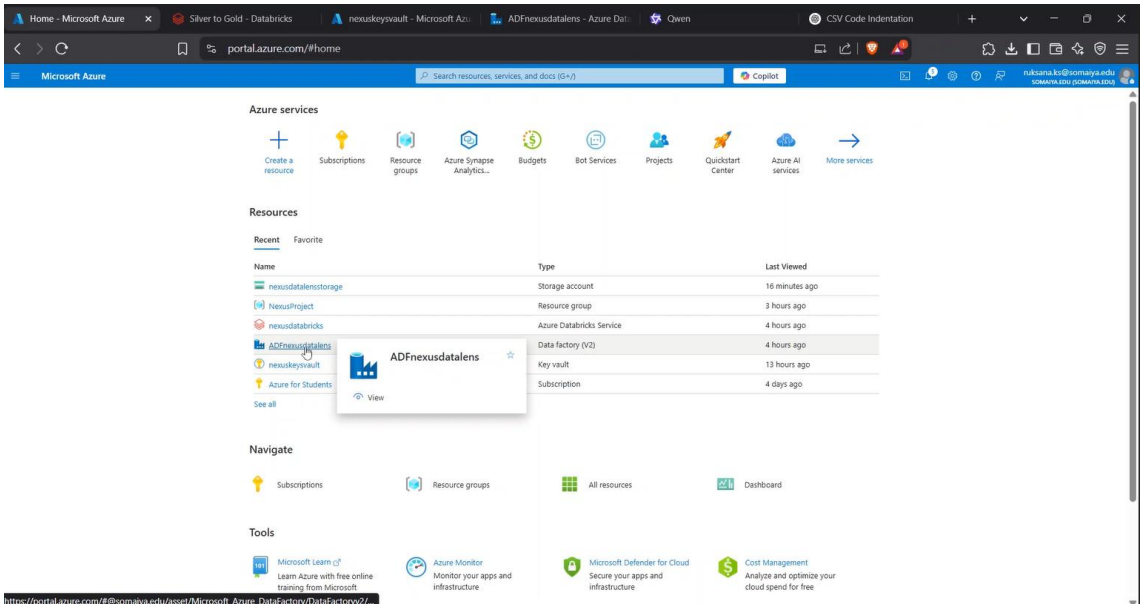
The manually triggered ADF pipeline completed successfully, executing the copy, bronze-to-silver, and silver-to-gold notebook transformations in sequence.



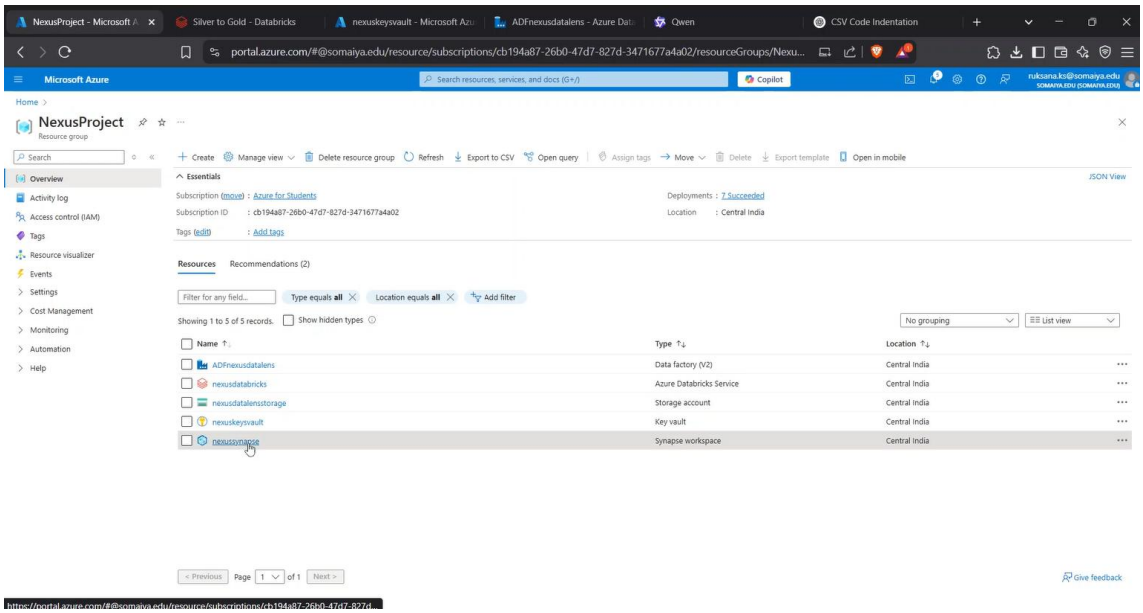
SYNAPSE



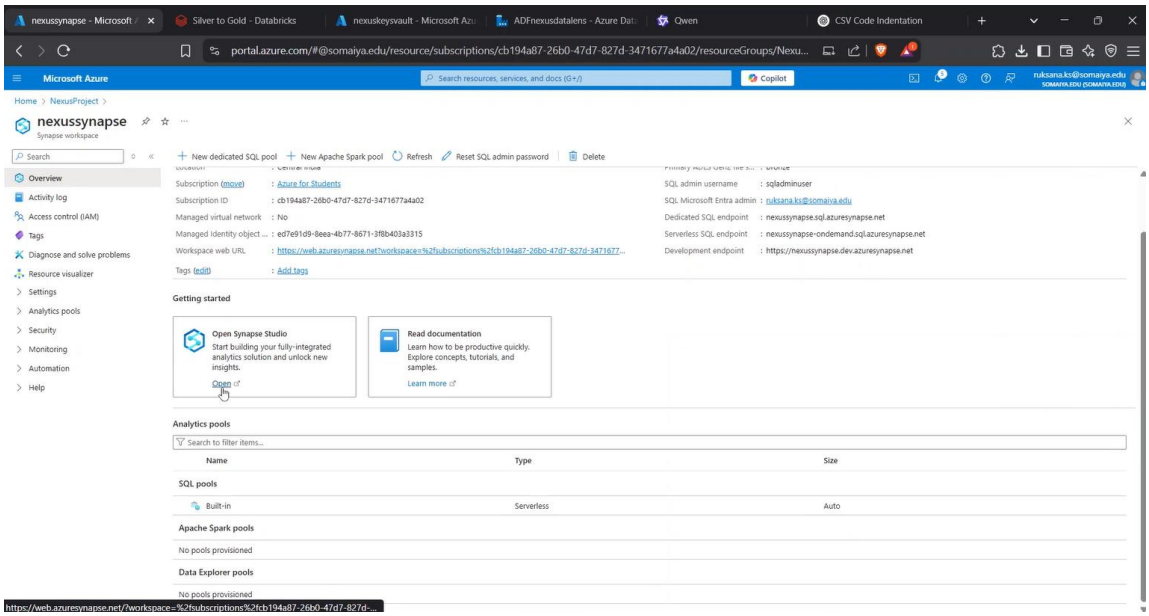
The Azure portal homepage displays recent resources, highlighting the selected Data Factory instance "**NexusProject**" for further access or management.



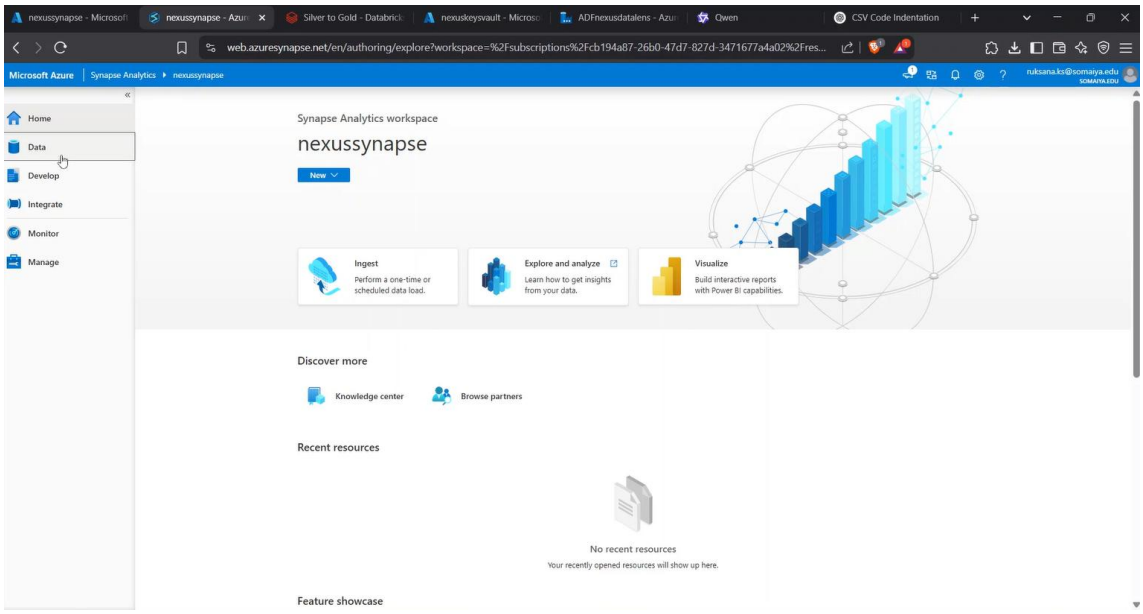
The NexusProject resource group in Azure contains five resources, including Data Factory, Databricks, Storage Account, Key Vault, and Synapse Workspace, all located in Central India.



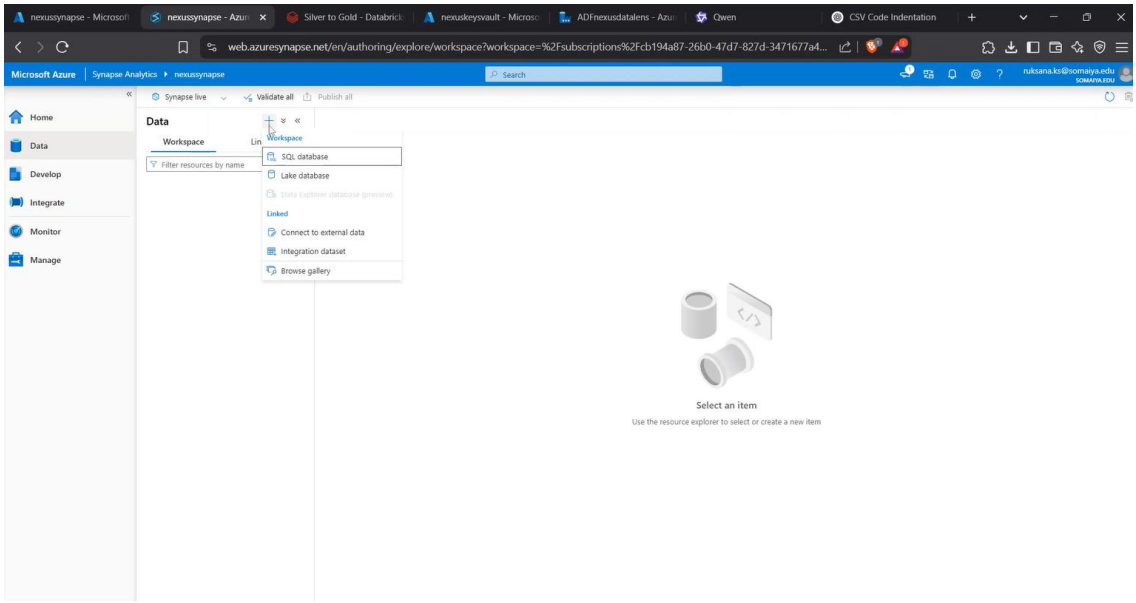
The Synapse workspace "**nexussynapse**" is ready for use with serverless SQL enabled, offering quick access to Synapse Studio for data analytics and development.



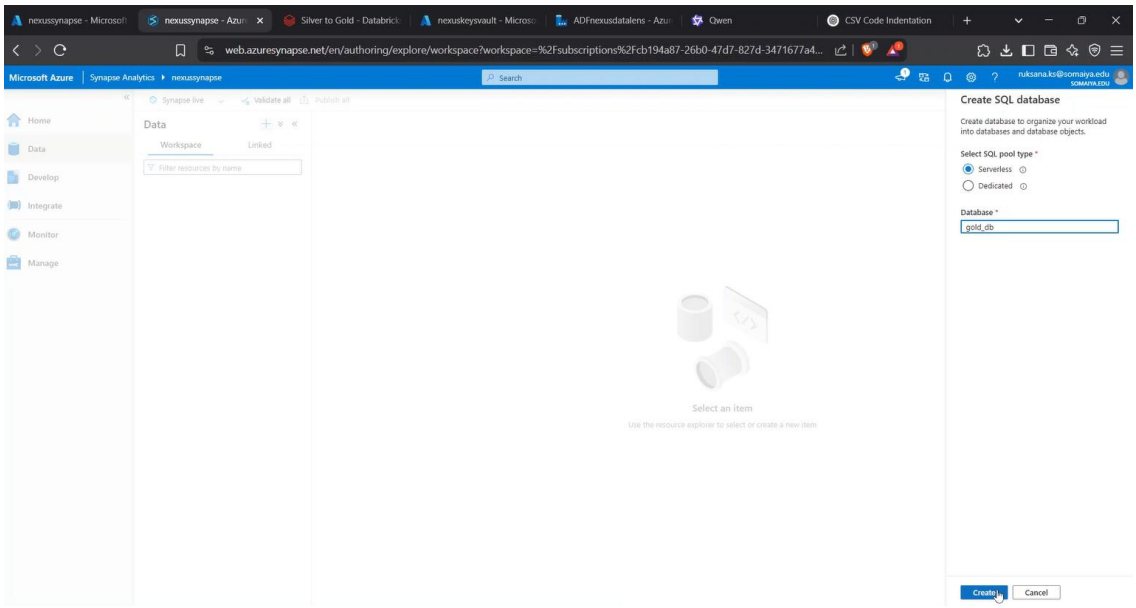
The Synapse Analytics workspace "**nexussynapse**" dashboard offers options to ingest, explore, and visualize data, with navigation to data, development, and integration features.



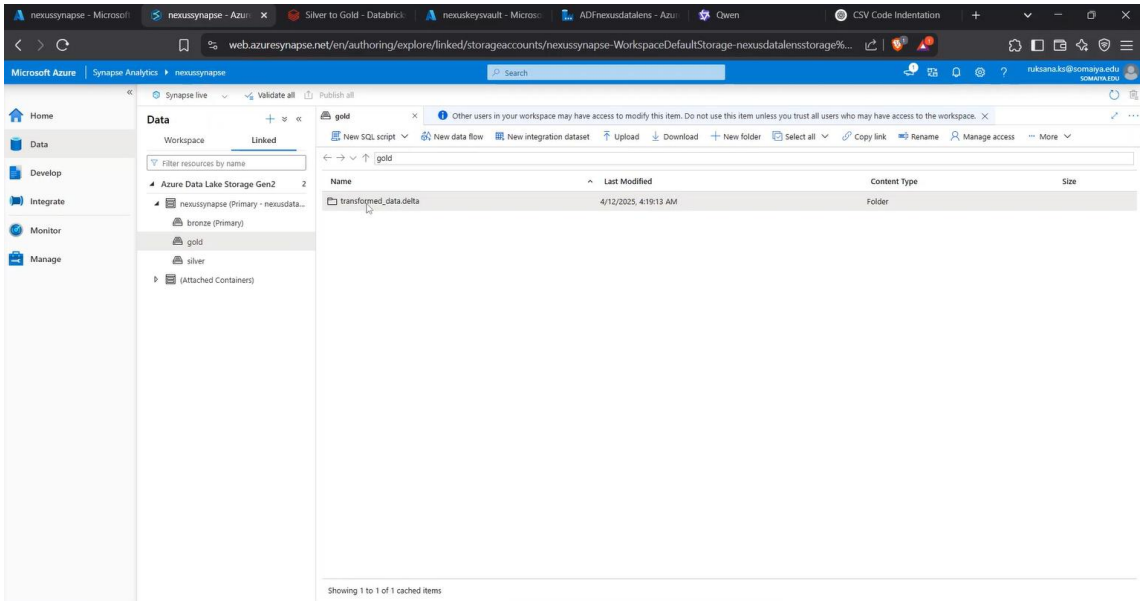
The image shows the "**Data**" pane in Azure Synapse Studio, where the user is selecting to add a new SQL or Lake database under the Workspace tab.



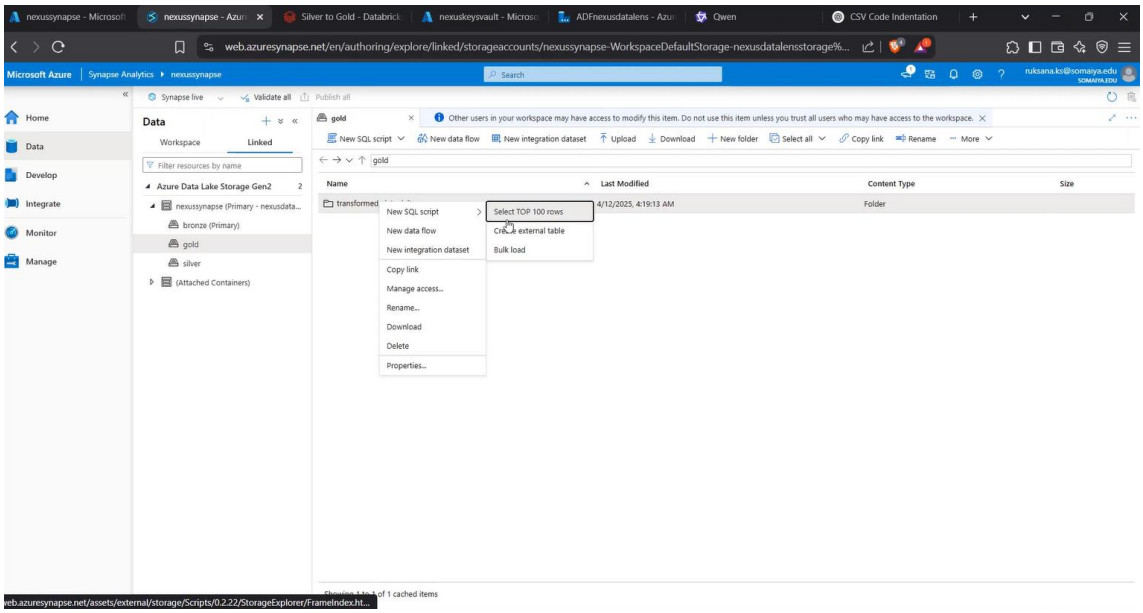
The user creating a new serverless SQL database named **gold_db** in Azure Synapse Studio by clicking the **Create button.**



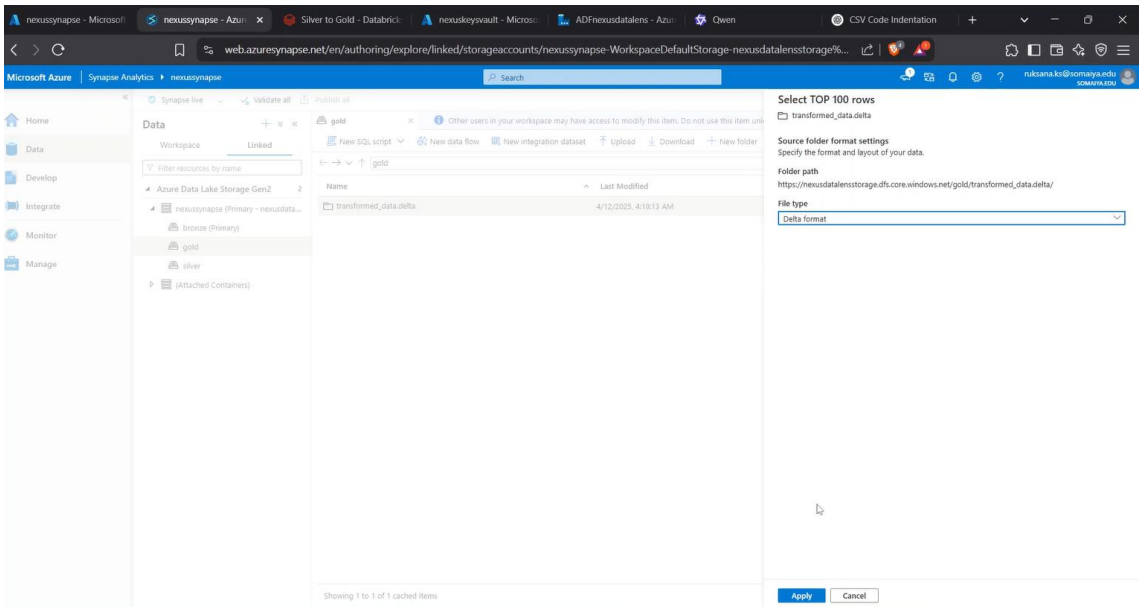
The image shows the gold container in Azure Data Lake Storage Gen2 within Synapse Studio, containing a folder named **transformed_data.delta**.



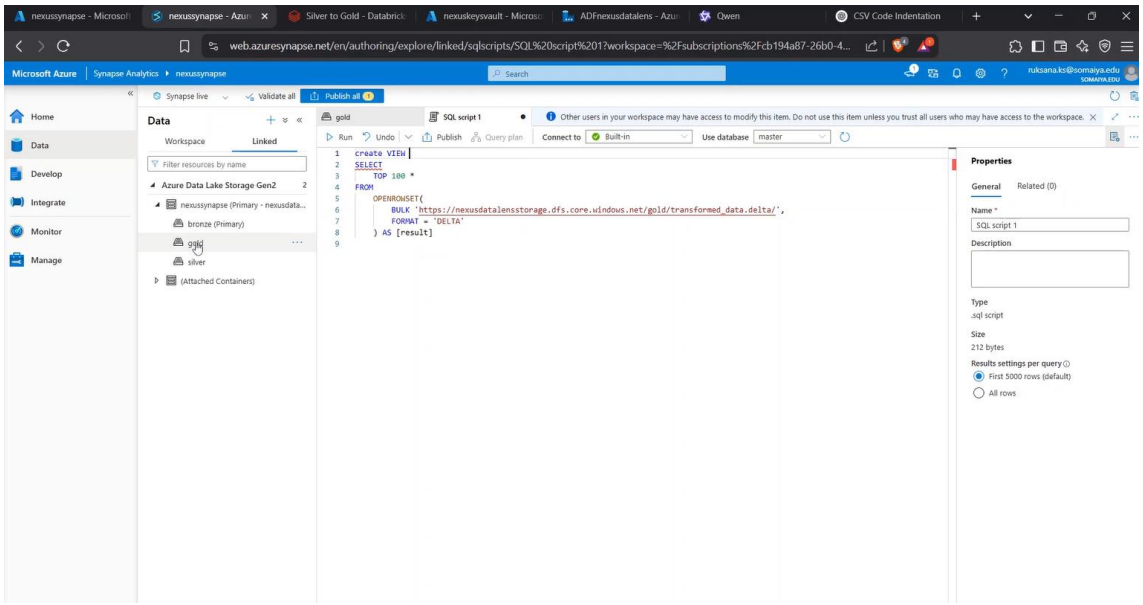
The user right-clicking on the transformed_data.delta folder in Azure Synapse Studio and selecting the "**Select TOP 100 rows**" SQL script option.



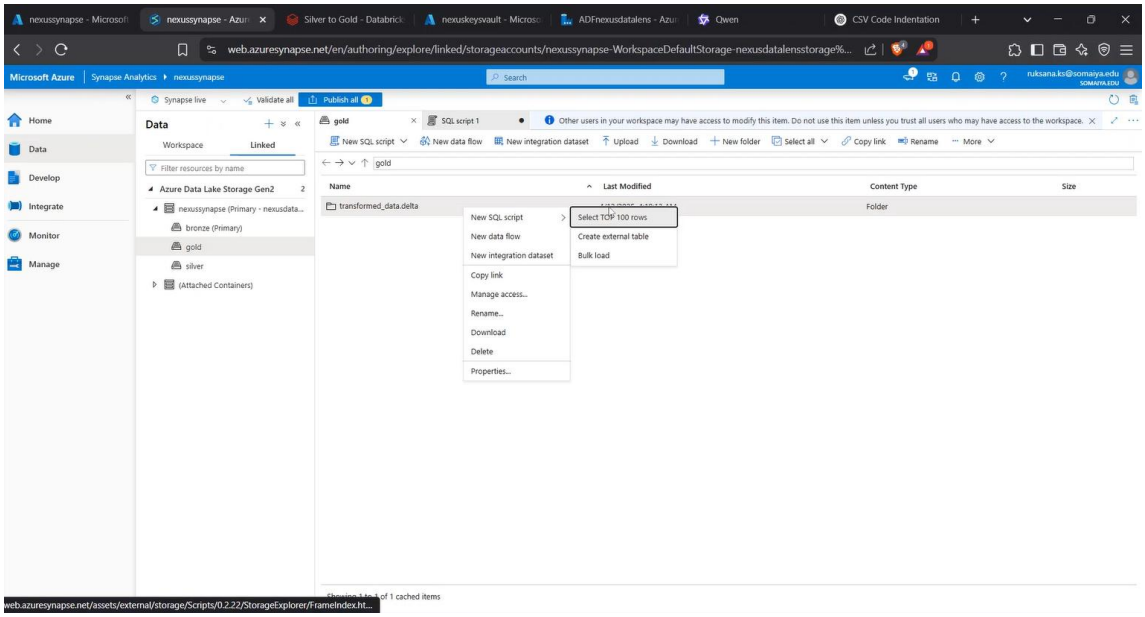
The image shows the user configuring the file type as Delta format for the folder **transformed_data.delta** before querying the top 100 rows in Azure Synapse Studio.



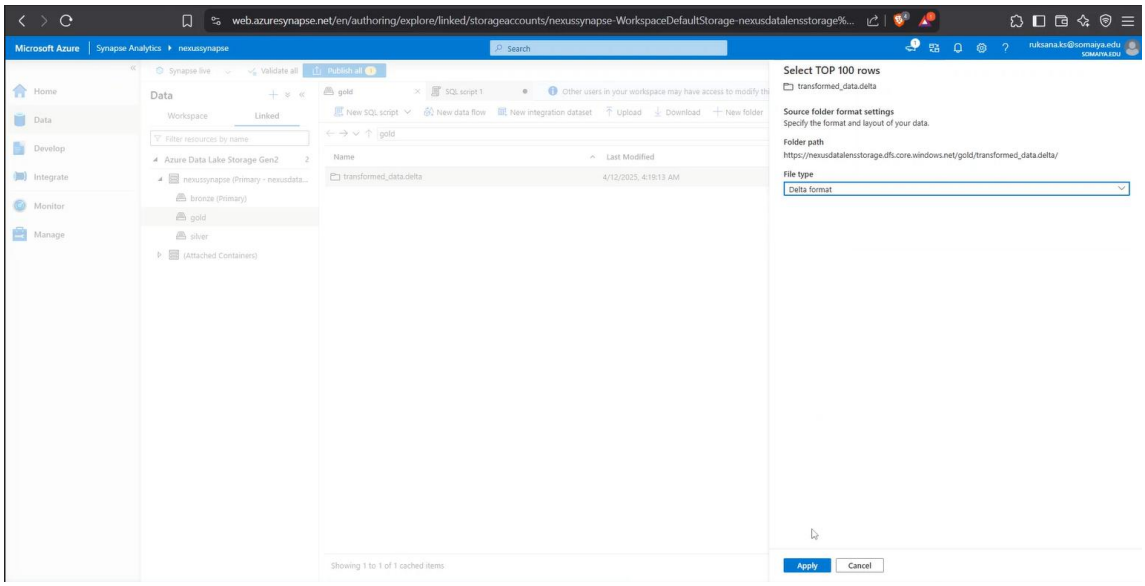
The image shows a SQL script in Azure Synapse Studio that creates a view to select the top 100 rows from a Delta format file located in the **gold/transformed_data.delta** folder.



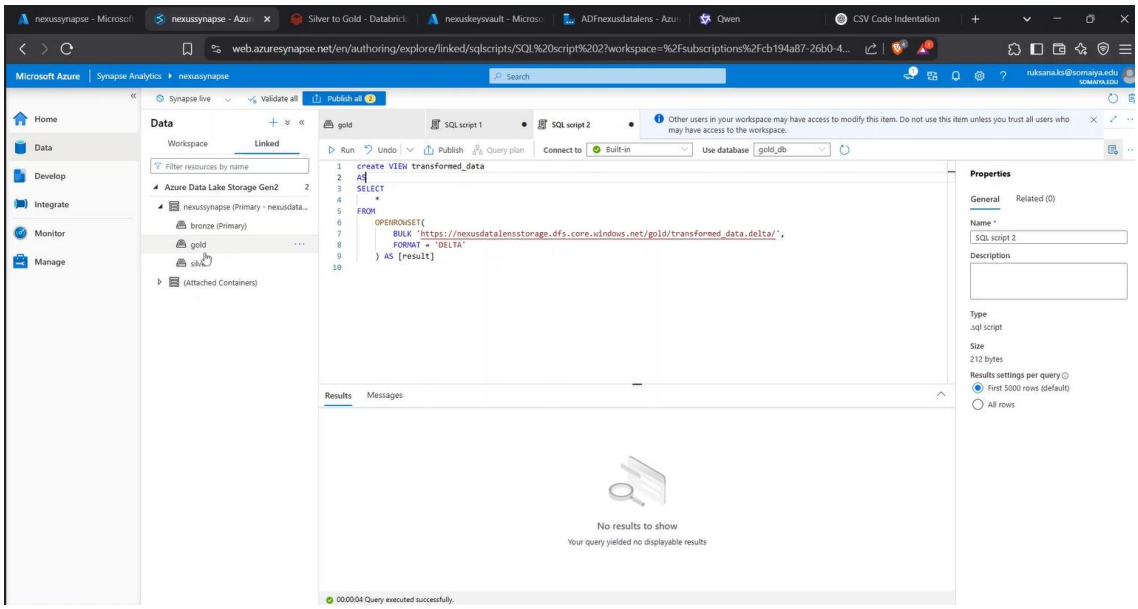
The user again right-clicking on the transformed_data.delta folder in Azure Synapse Studio and selecting "**Select TOP 100 rows**" to preview data from the Delta table.



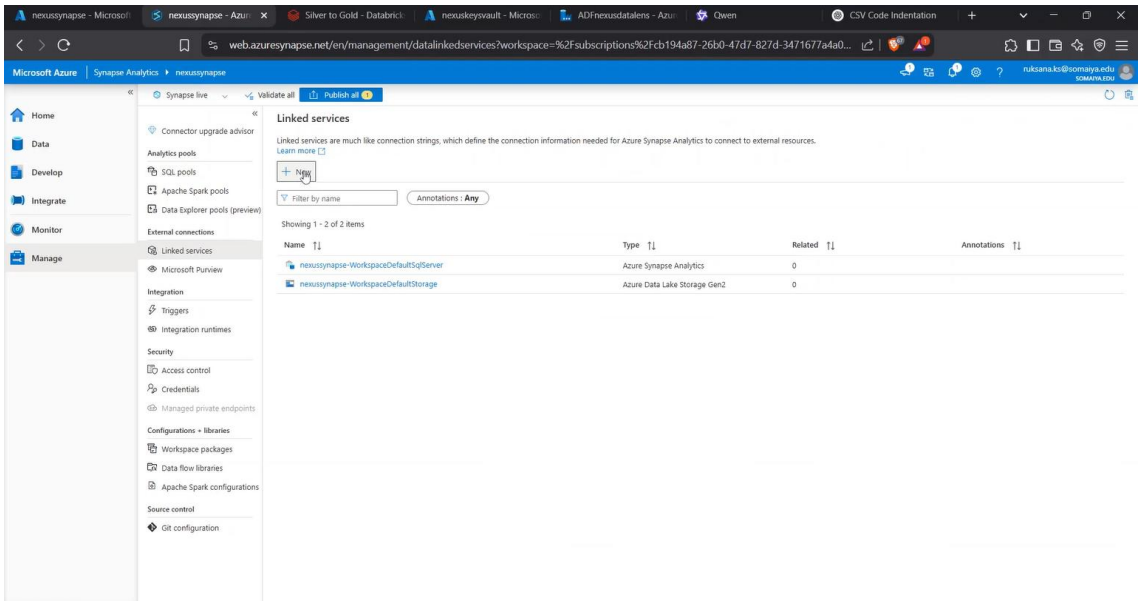
The image shows the process of selecting the top 100 rows from a Delta format file **(transformed_data.delta)** in Azure Synapse Analytics.



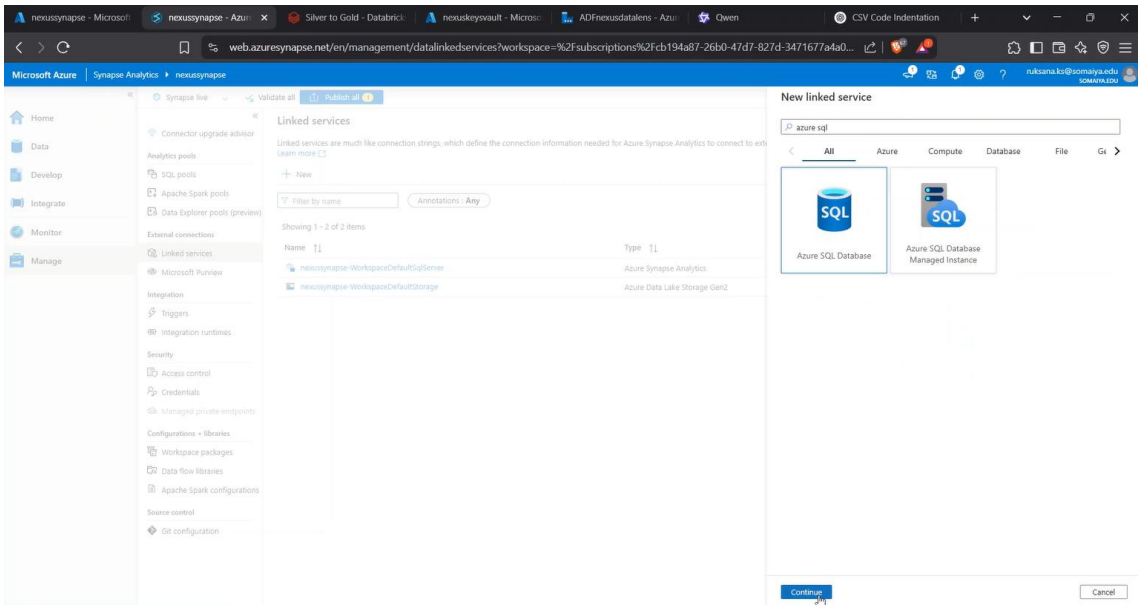
The image shows a SQL script in Azure Synapse that creates a view (**transformed_data**) by querying Delta format data from a specified path in the Data Lake using **OPENROWSET.**



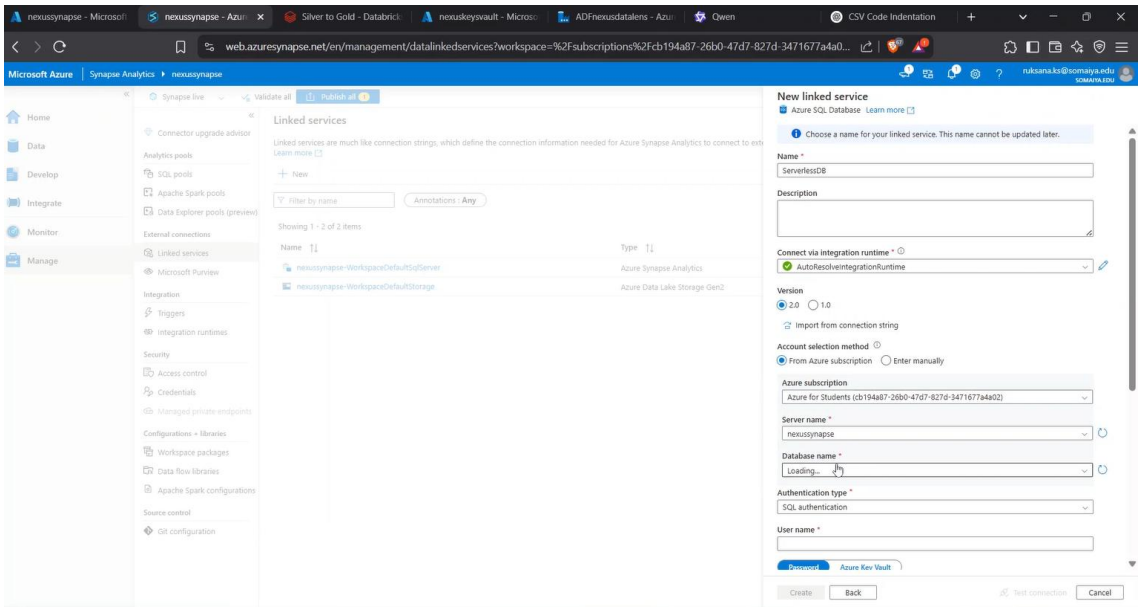
The image shows the "**Linked services**" section in Azure Synapse Analytics where users can add new connections to external resources by clicking the **"+ New"** button.



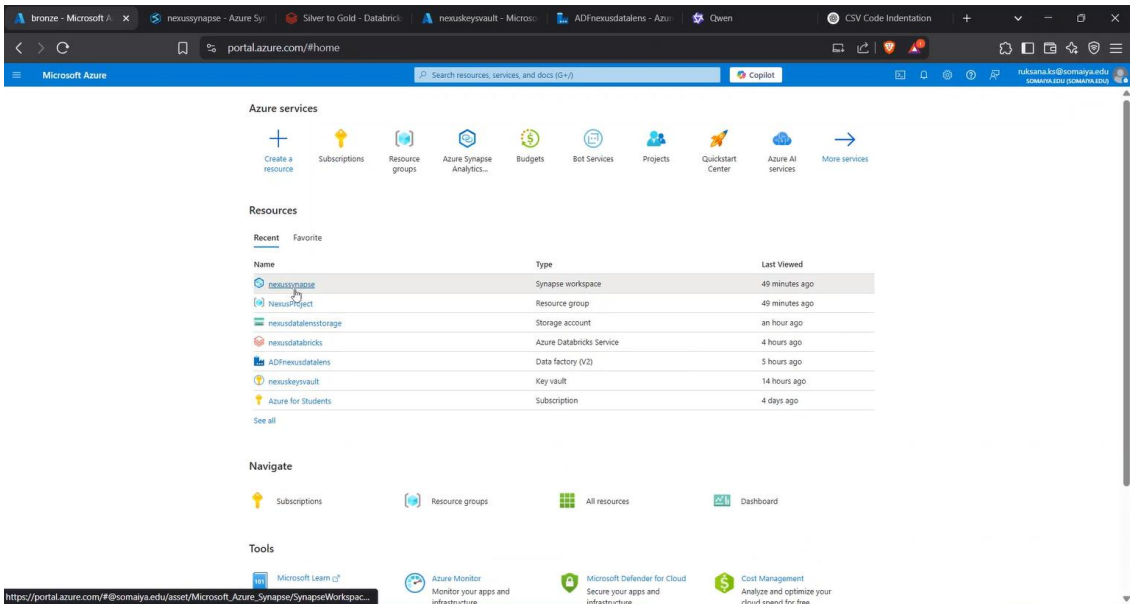
The image shows the selection of "**Azure SQL Database**" as a new linked service in Azure Synapse Analytics, with the user about to click "**Continue**" to proceed with the setup.



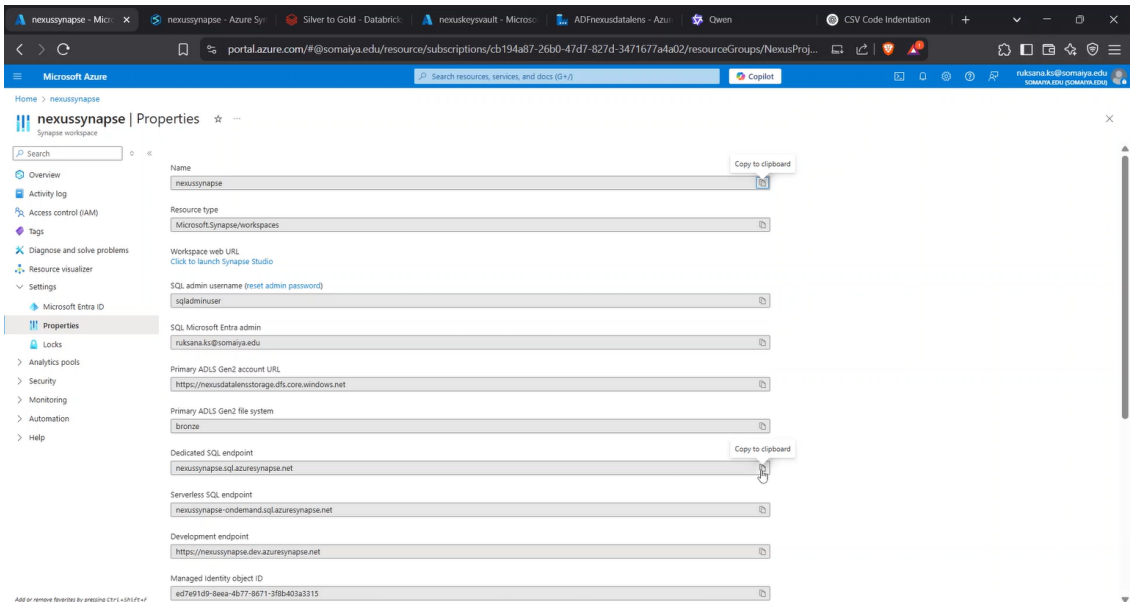
The configuration of a new linked service in Azure Synapse for connecting to an Azure SQL Database using SQL authentication.



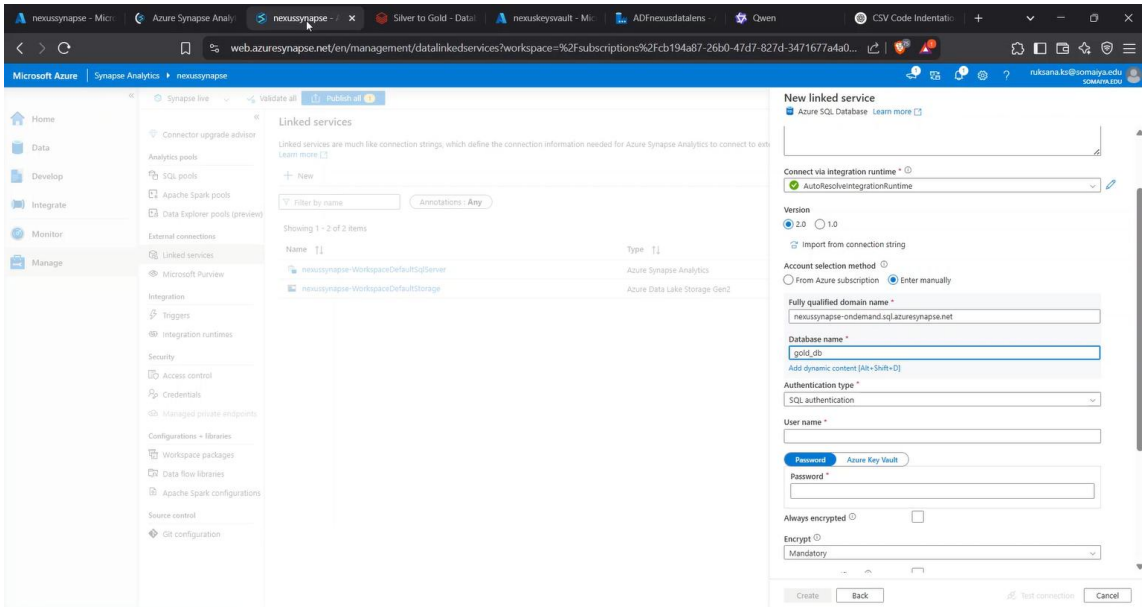
The Azure portal home displays recently accessed resources, highlighting a Synapse workspace named "**nexusynapse**" as the most recently viewed.



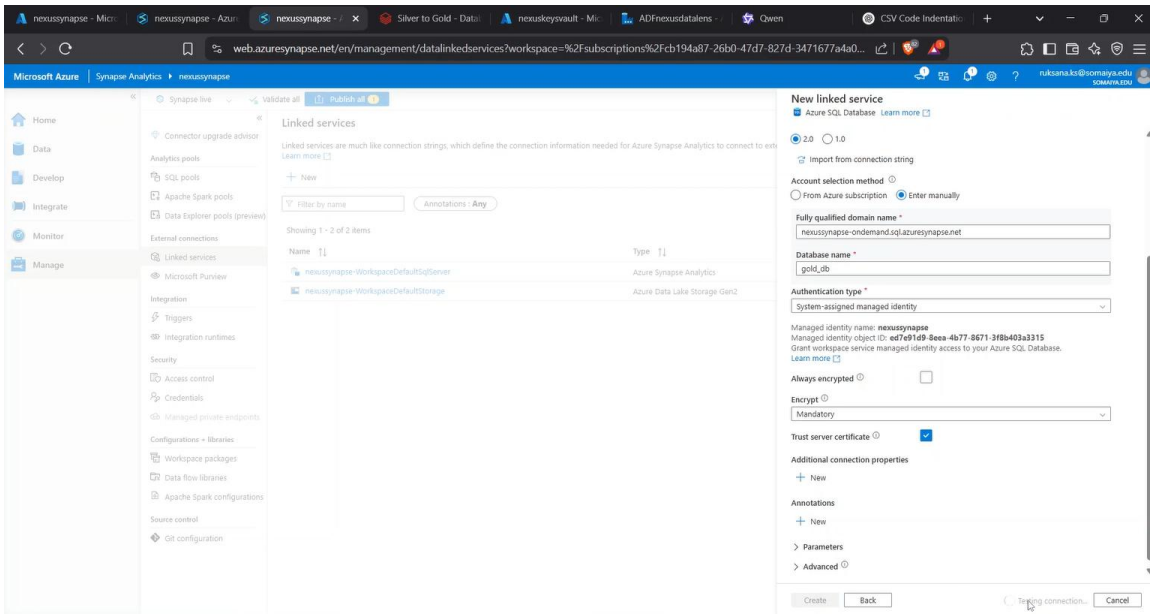
The properties page of the Synapse workspace "**nexusynapse**" displays key configuration details like workspace URL, SQL admin, endpoints, and managed identity.



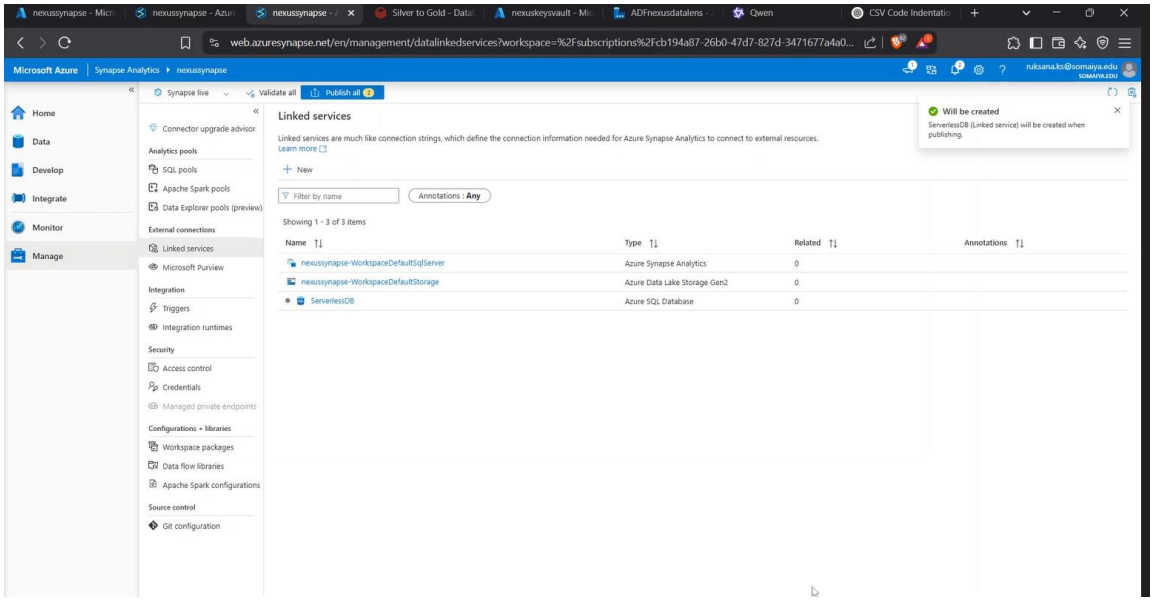
Creating a new linked service using SQL authentication by specifying the domain, database name **(gold_db),** and credentials for Azure SQL Database connection.



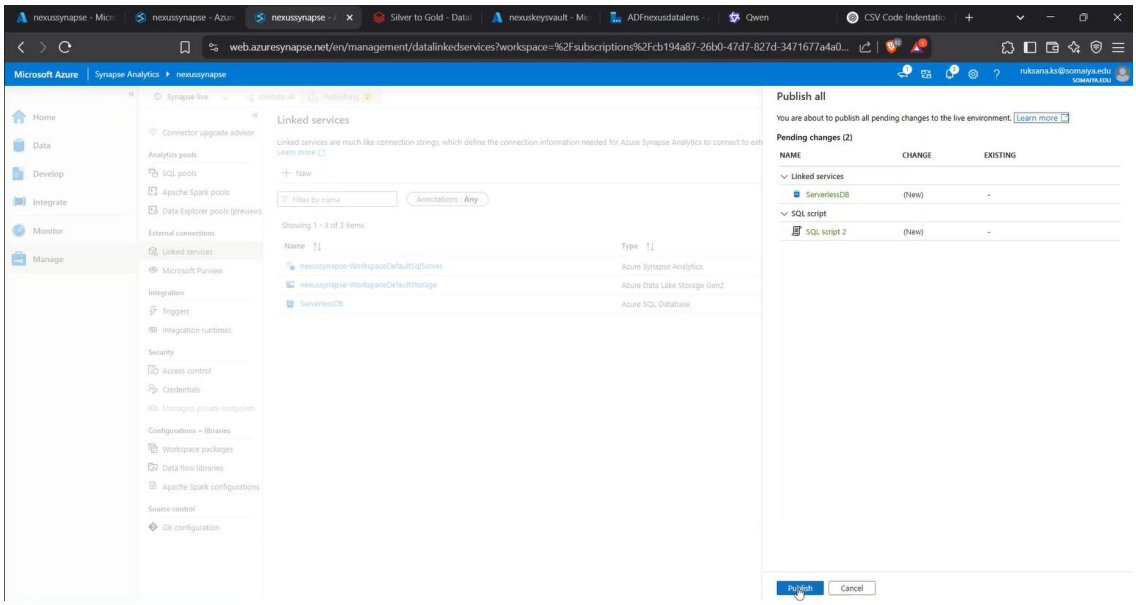
A new linked service for an Azure SQL Database named **gold_db** is being configured using a managed identity for authentication.



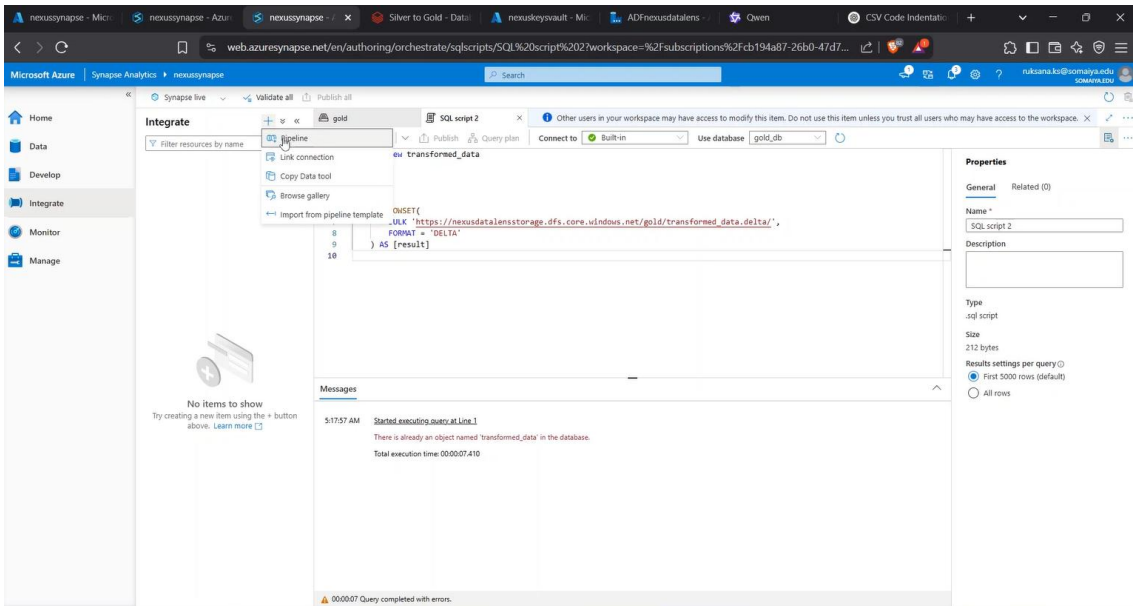
The "**Linked services**" view shows that the new linked service ServerlessDB will be created upon publishing.



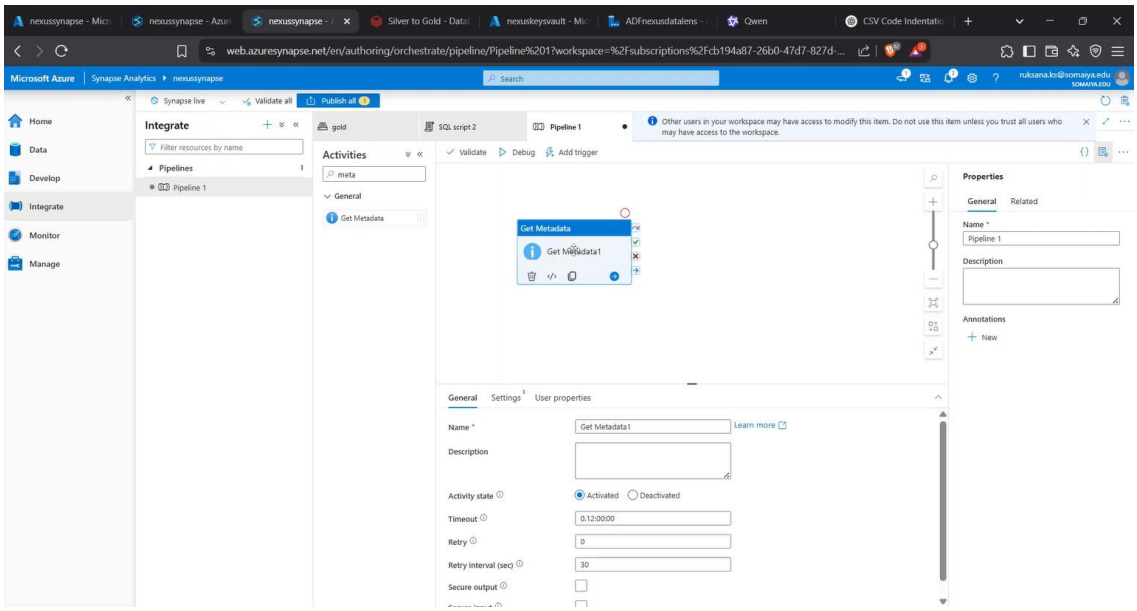
The user is about to publish changes including a new linked service **ServerlessDB** and a new SQL script **SQL script 2** to the Synapse workspace.



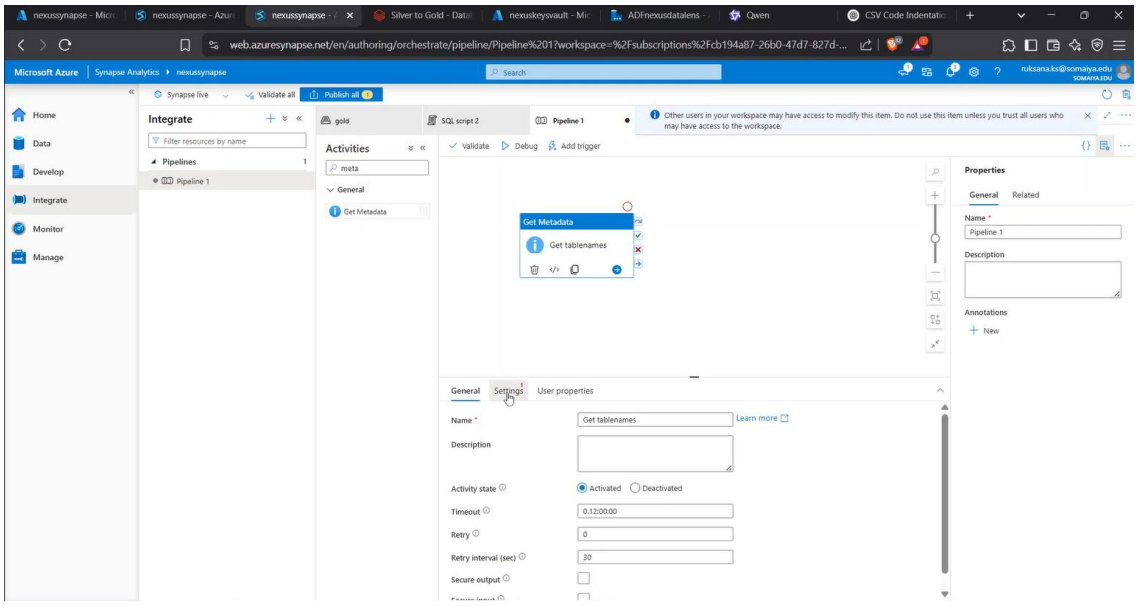
The SQL script failed with an error indicating that the table transformed_data already exists in the gold_db database.



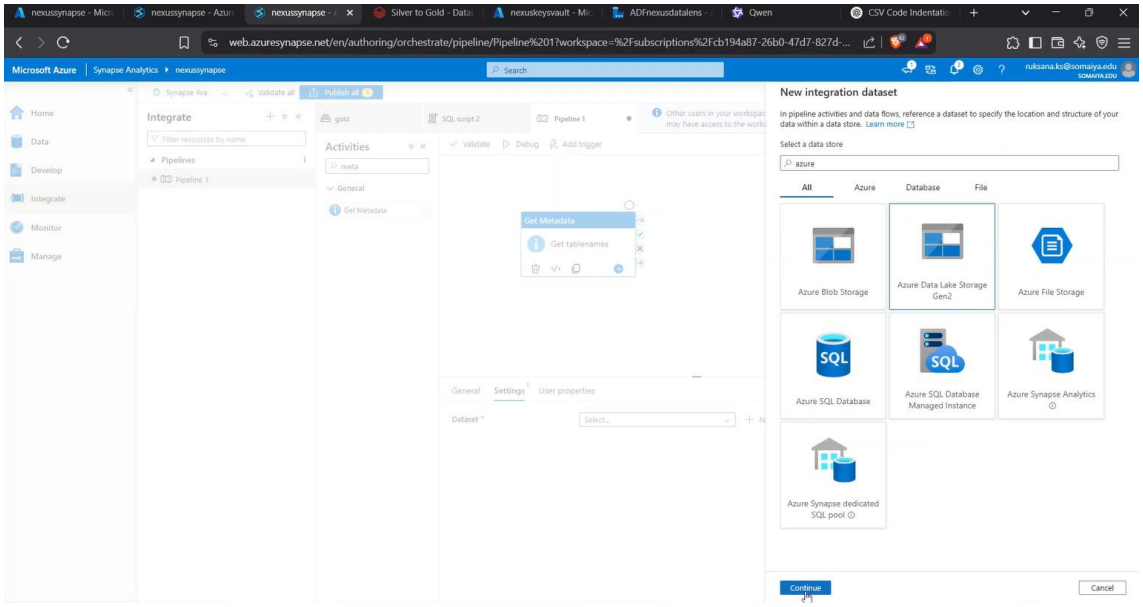
The "**Get Metadata1**" activity is activated in the pipeline with default execution settings like timeout and retry configured.



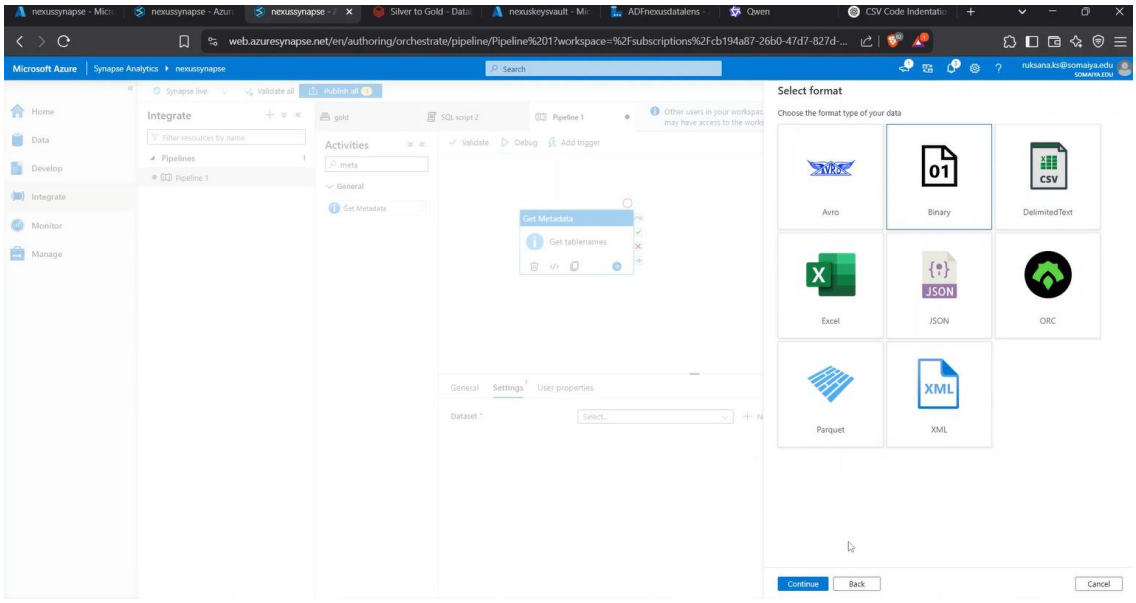
The "**Get Metadata**" activity is named "**Get tablenames**" and is configured with default timeout, retry, and output settings.



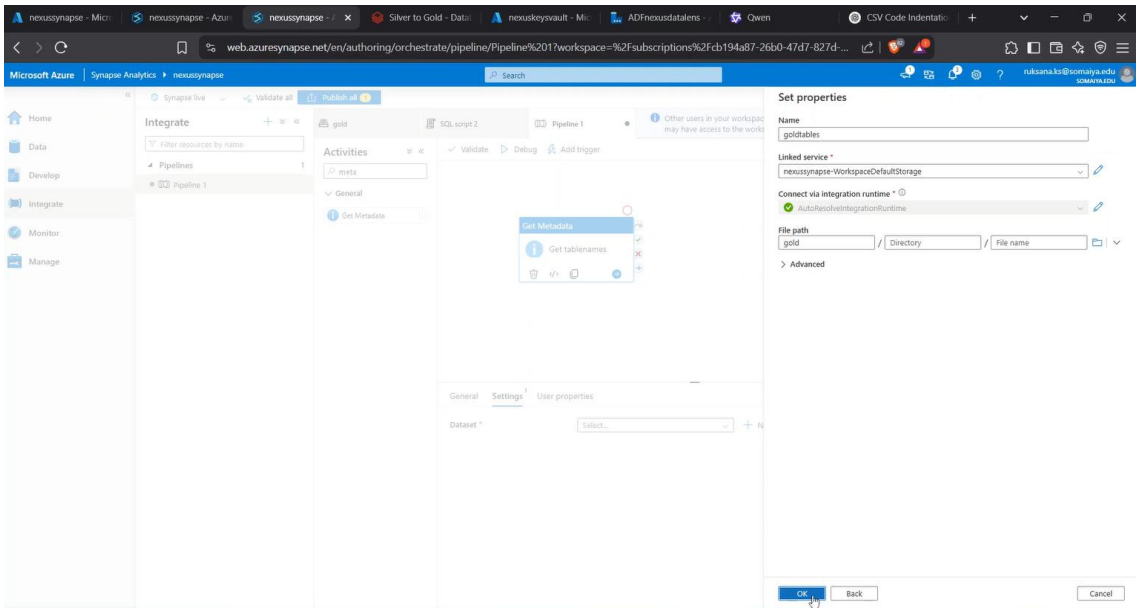
The user is creating a new integration dataset by selecting "**Azure Data Lake Storage Gen2**" as the **data store type**.



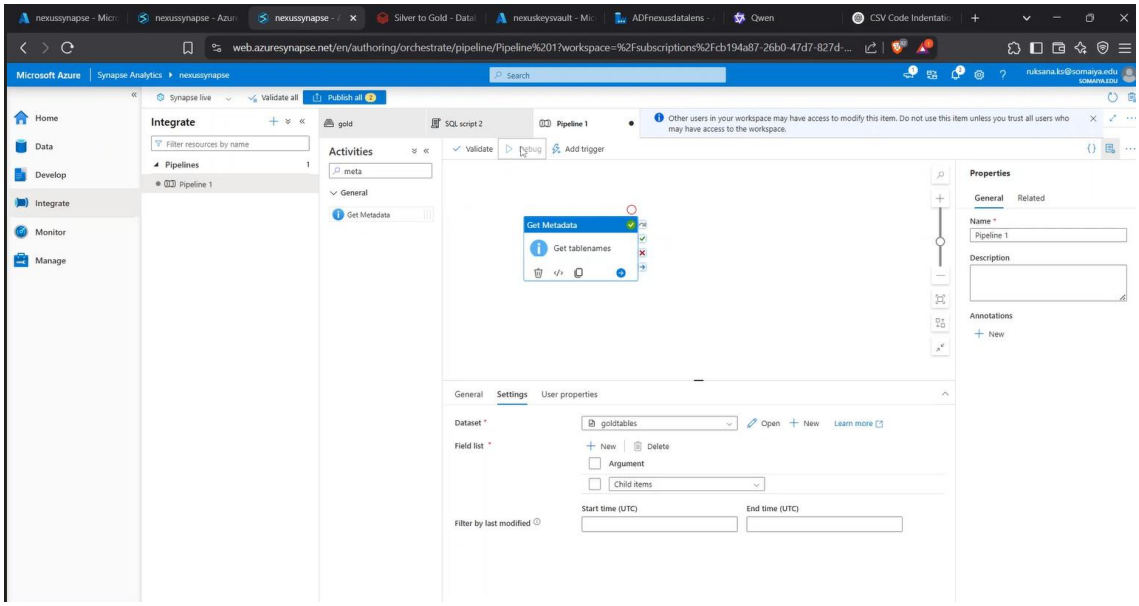
The data format selection screen shows various options, with "**Binary**" currently selected for the dataset configuration.



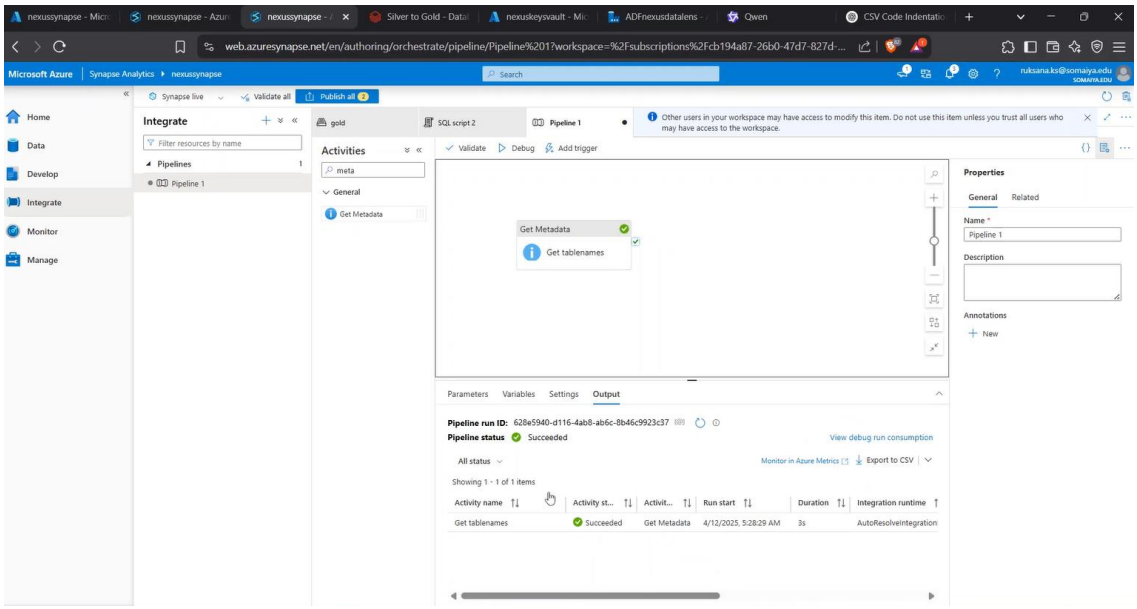
The "**goldtables**" dataset is configured to access the "**gold**" directory in Azure Data Lake Storage via the specified linked service.



The pipeline's "**Get Metadata**" activity is configured to retrieve child items (table names) from the "**goldtables**" dataset.

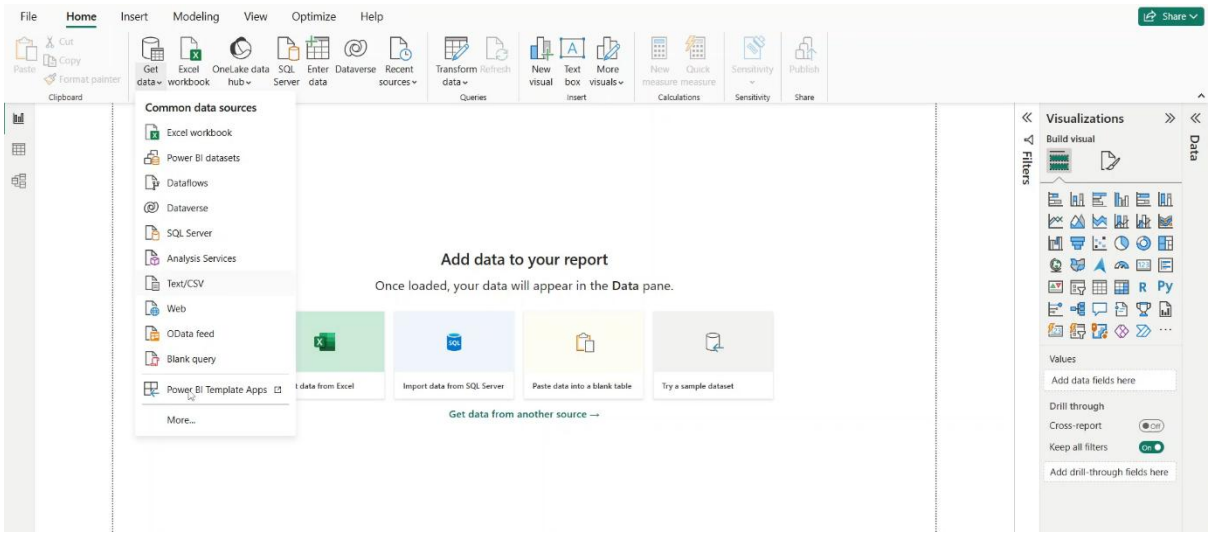


The Synapse pipeline "**Pipeline 1**" ran successfully, retrieving table names using a "**Get Metadata**" activity.

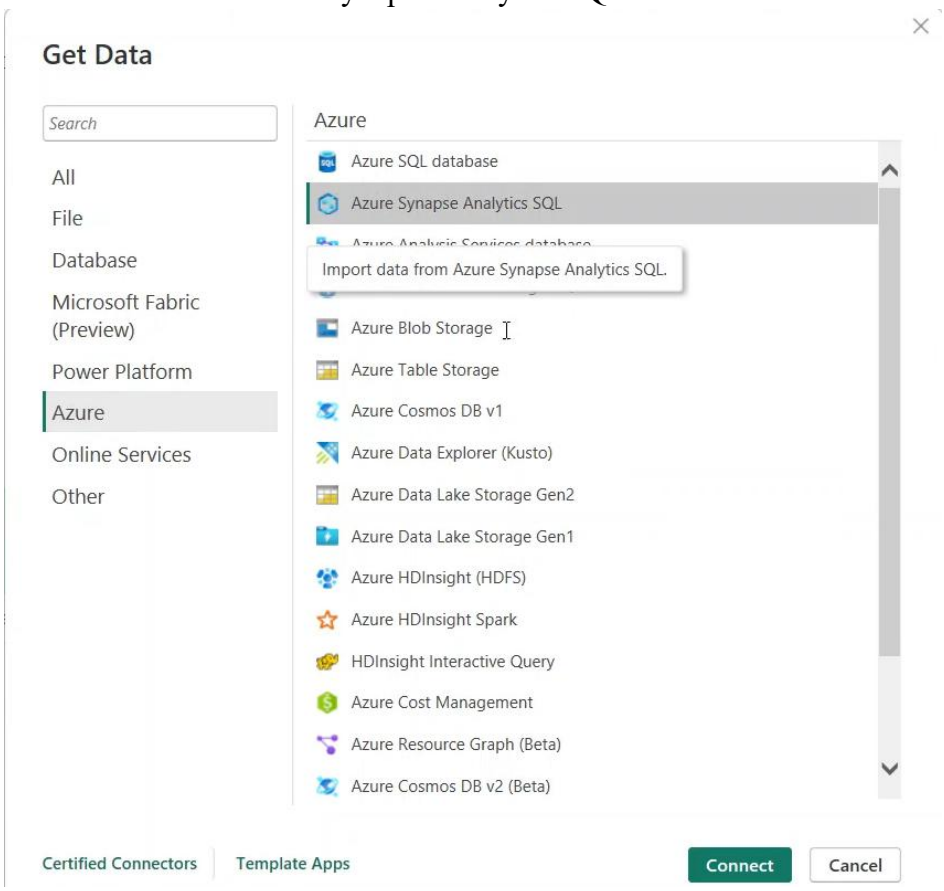


Power Bi Setup

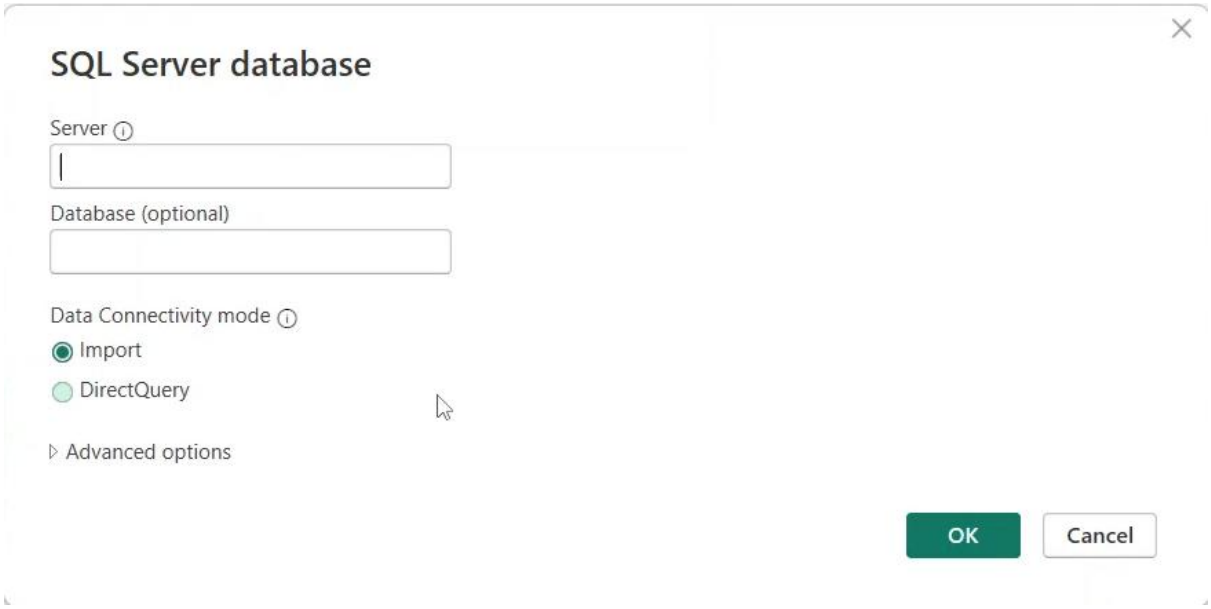
Install Power Bi and open it
in Navbar You will see click on Get Data > More



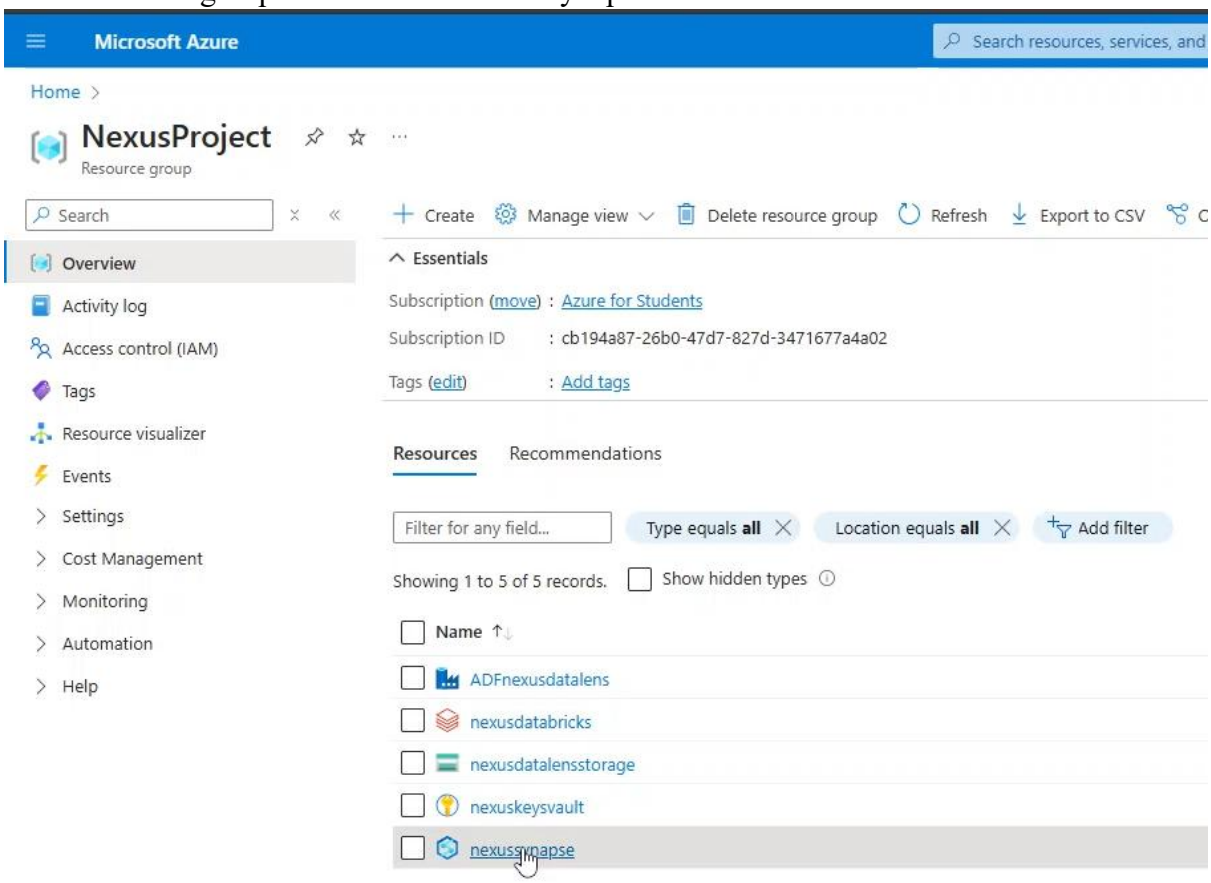
Click On Azure > Azure Synapse Analytics SQL > Connect



This Window will open You need to paste serverless endpoint of Azure synapse analytics for that

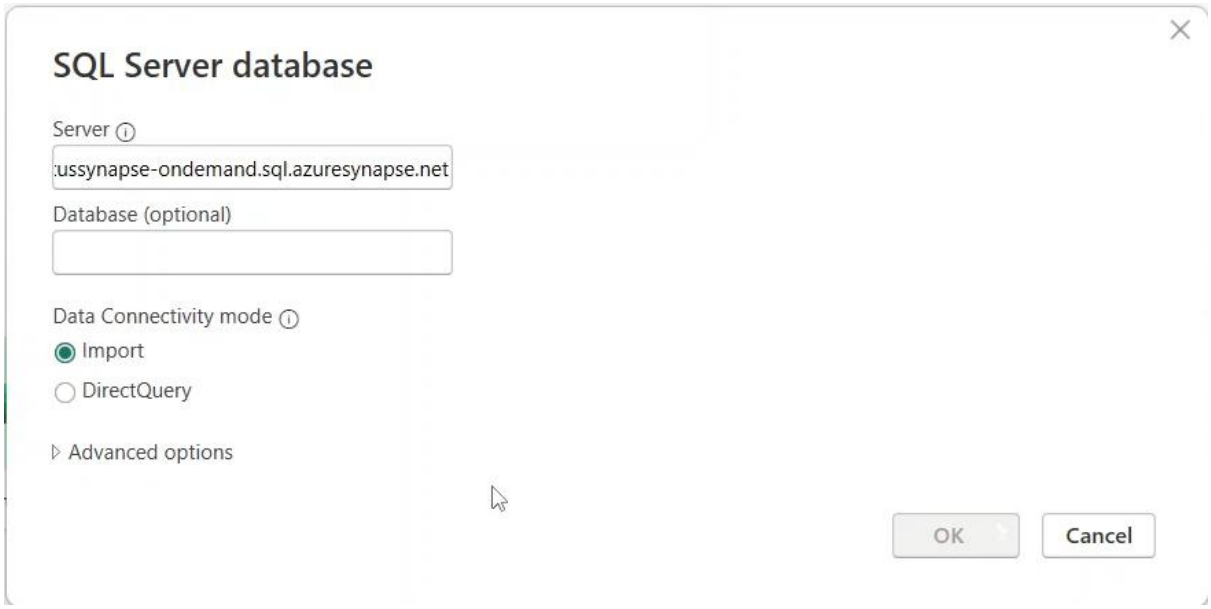


Go to resource group and click on > nexussynapse

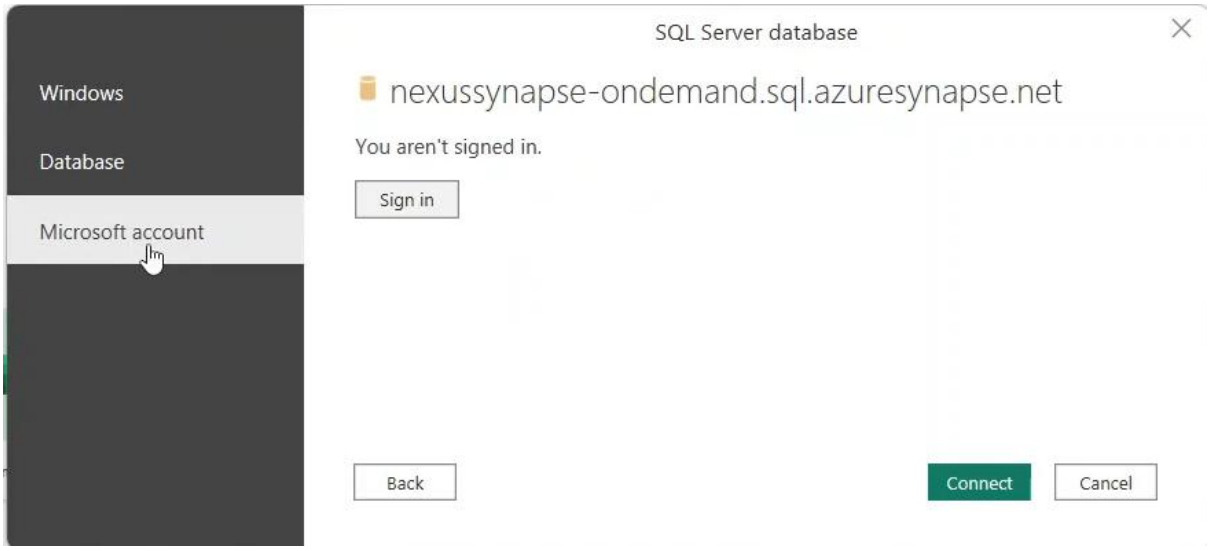


Got to properties you will get link > copy that link

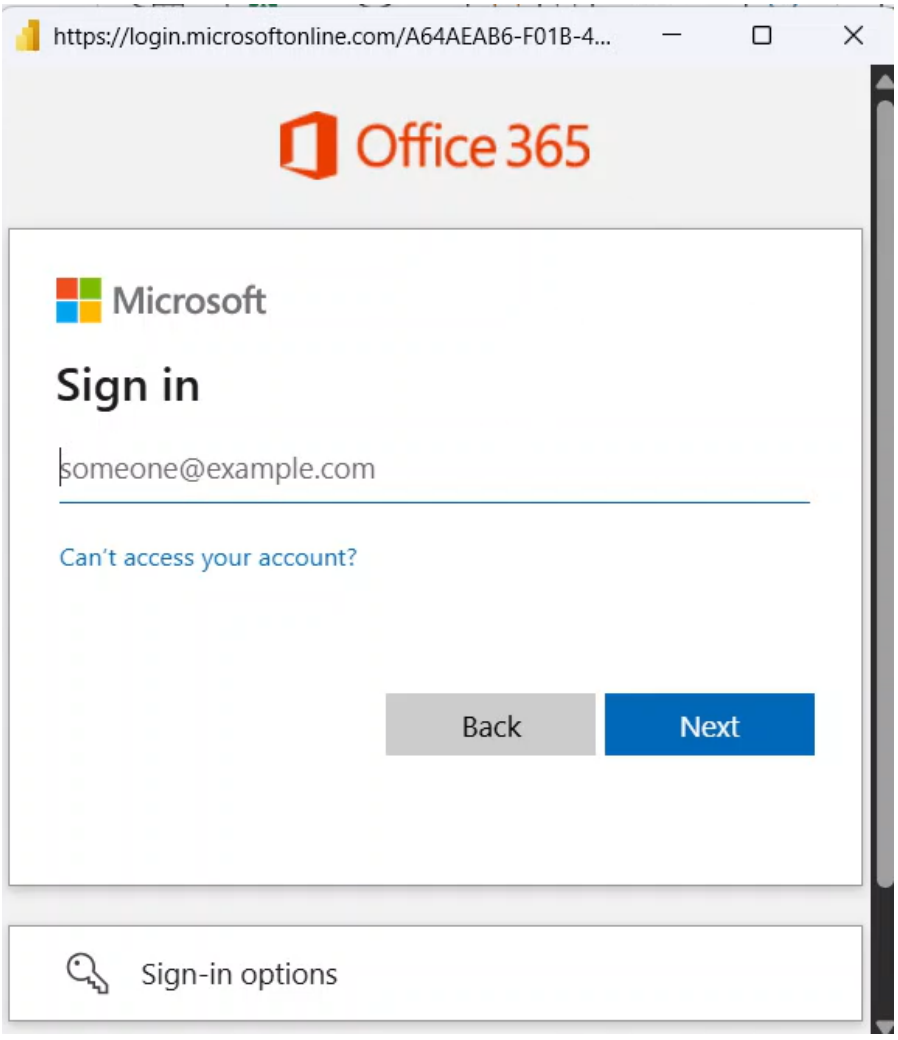
Paste it over here > OK



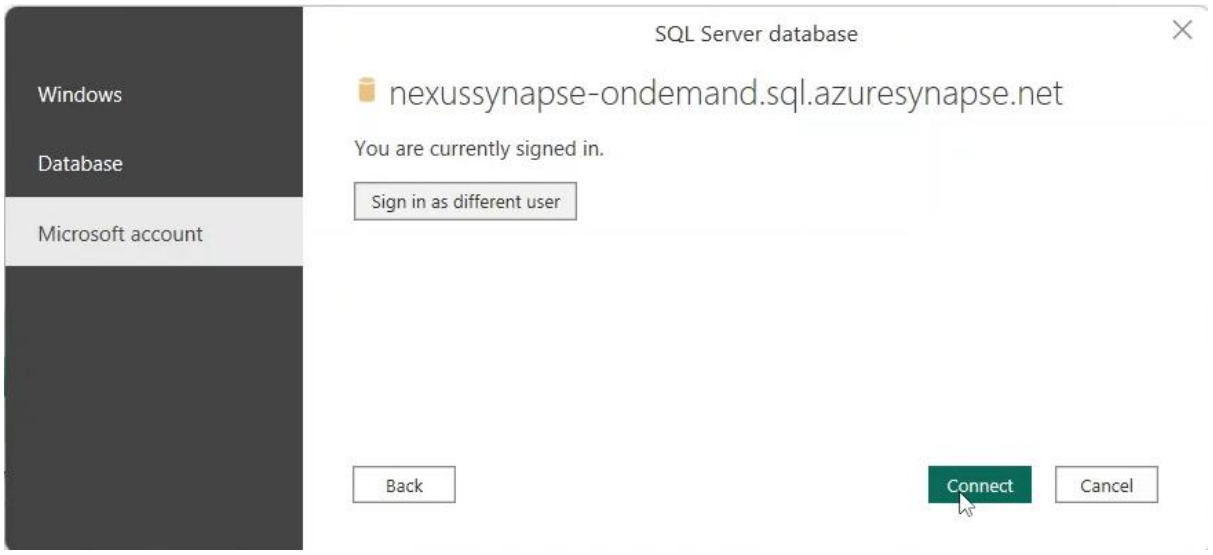
Click on Microsoft Account > Sign in



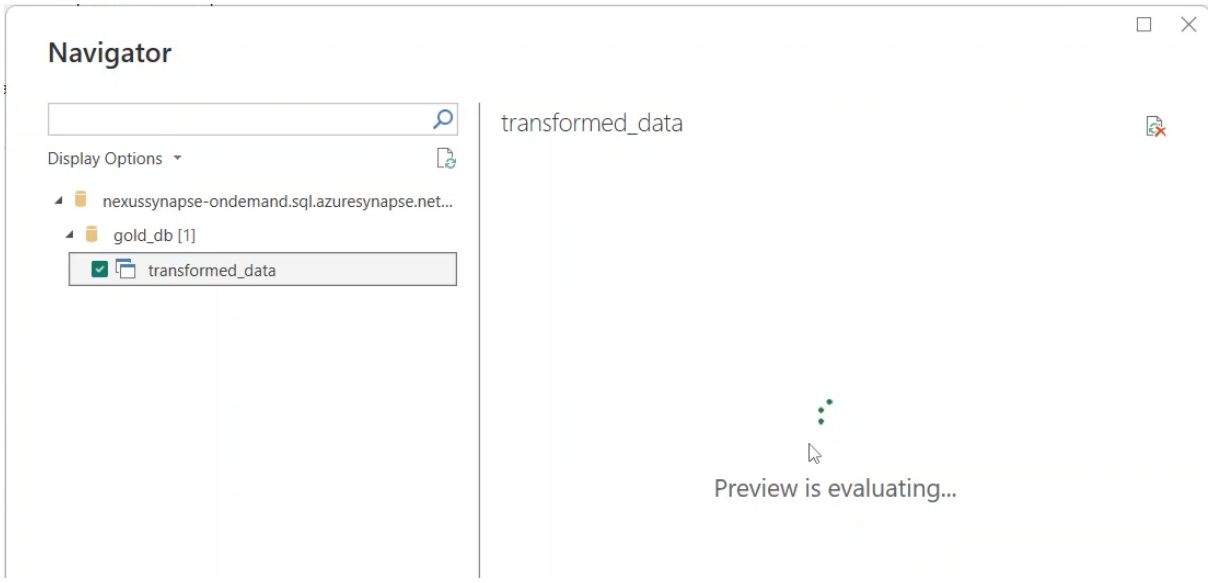
Sign in with same Azure account in which you are making Pipeline



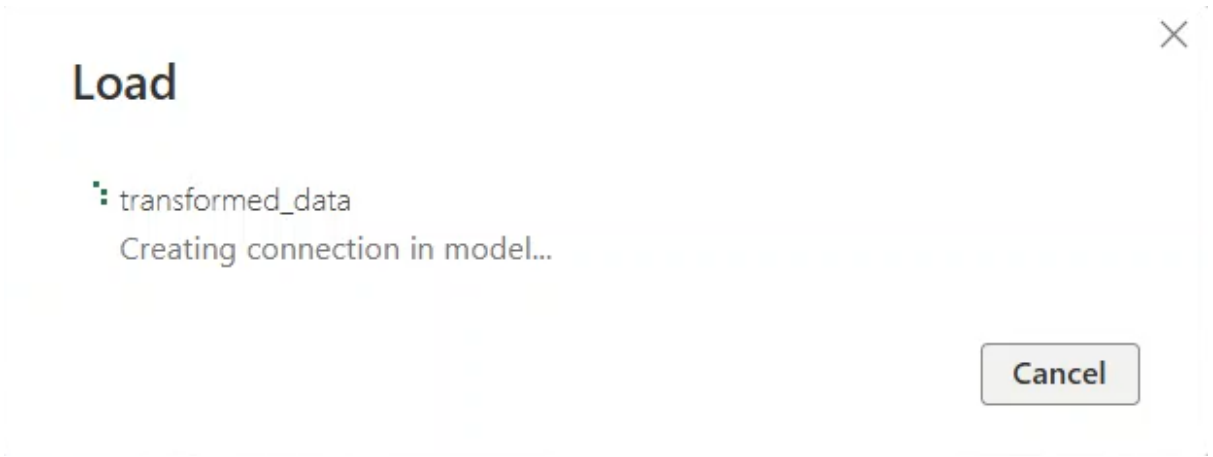
Click on connect



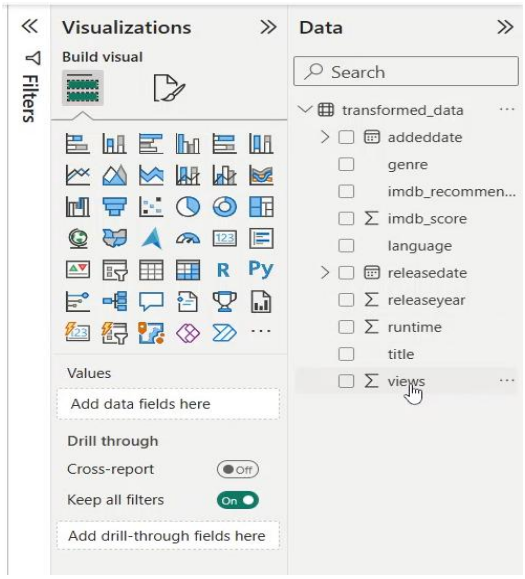
Navigator Tab will get open in which you will able to see Transform Data table Check the checkbox



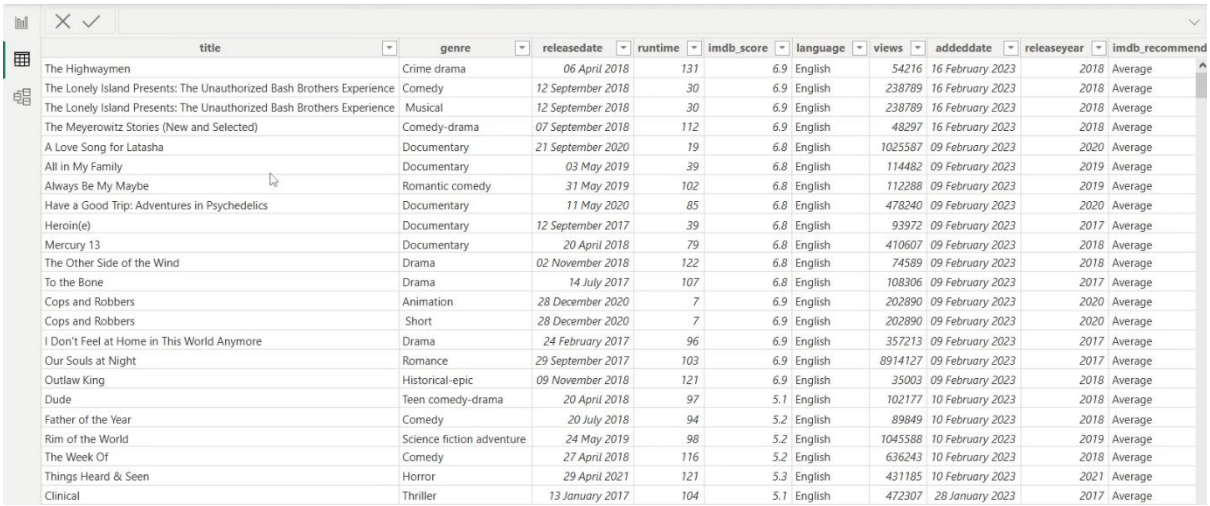
Wait for some time to load the data



Once it get loaded it will show us Data Coloums in data Tab

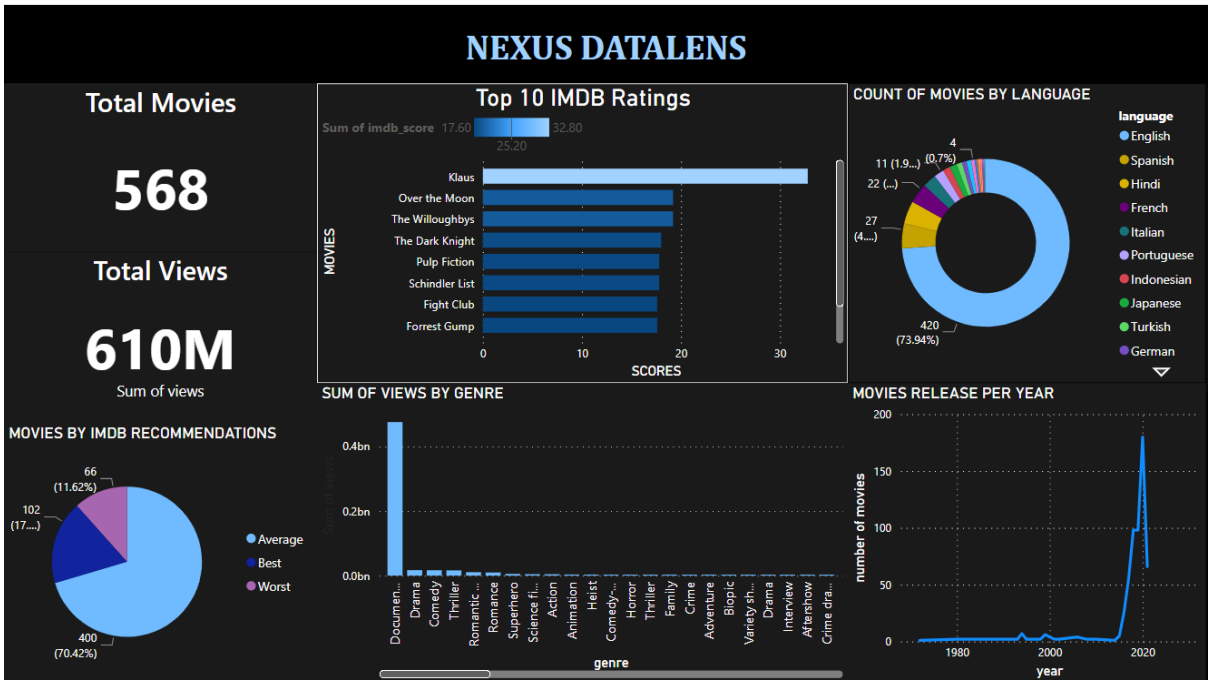


You can see the transform data table in table view



title	genre	releasedate	runtime	imdb_score	language	views	addeddate	releaseyear	imdb_recommend
The Highwaymen	Crime drama	06 April 2018	131	6.9	English	54216	16 February 2023	2018	Average
The Lonely Island Presents: The Unauthorized Bash Brothers Experience	Comedy	12 September 2018	30	6.9	English	238789	16 February 2023	2018	Average
The Lonely Island Presents: The Unauthorized Bash Brothers Experience	Musical	12 September 2018	30	6.9	English	238789	16 February 2023	2018	Average
The Meyerowitz Stories (New and Selected)	Comedy-drama	07 September 2018	112	6.9	English	48297	16 February 2023	2018	Average
A Love Song for Latasha	Documentary	21 September 2020	19	6.8	English	1025587	09 February 2023	2020	Average
All in My Family	Documentary	03 May 2019	39	6.8	English	114482	09 February 2023	2019	Average
Always Be My Maybe	Romantic comedy	31 May 2019	102	6.8	English	112288	09 February 2023	2019	Average
Have a Good Trip: Adventures in Psychedelics	Documentary	11 May 2020	85	6.8	English	478240	09 February 2023	2020	Average
Heroin(e)	Documentary	12 September 2017	39	6.8	English	93972	09 February 2023	2017	Average
Mercury 13	Documentary	20 April 2018	79	6.8	English	410607	09 February 2023	2018	Average
The Other Side of the Wind	Drama	02 November 2018	122	6.8	English	74589	09 February 2023	2018	Average
To the Bone	Drama	14 July 2017	107	6.8	English	108306	09 February 2023	2017	Average
Cops and Robbers	Animation	28 December 2020	7	6.9	English	202890	09 February 2023	2020	Average
Cops and Robbers	Short	28 December 2020	7	6.9	English	202890	09 February 2023	2020	Average
I Don't Feel at Home in This World Anymore	Drama	24 February 2017	96	6.9	English	357213	09 February 2023	2017	Average
Our Souls at Night	Romance	29 September 2017	103	6.9	English	8914127	09 February 2023	2017	Average
Outlaw King	Historical-epic	09 November 2018	121	6.9	English	35003	09 February 2023	2018	Average
Dude	Teen comedy-drama	20 April 2018	97	5.1	English	102177	10 February 2023	2018	Average
Father of the Year	Comedy	20 July 2018	94	5.2	English	89849	10 February 2023	2018	Average
Rim of the World	Science fiction adventure	24 May 2019	98	5.2	English	1045588	10 February 2023	2019	Average
The Week Of	Comedy	27 April 2018	116	5.2	English	636243	10 February 2023	2018	Average
Things Heard & Seen	Horror	29 April 2021	121	5.3	English	431185	10 February 2023	2021	Average
Clinical	Thriller	13 January 2017	104	5.1	English	472307	28 January 2023	2017	Average

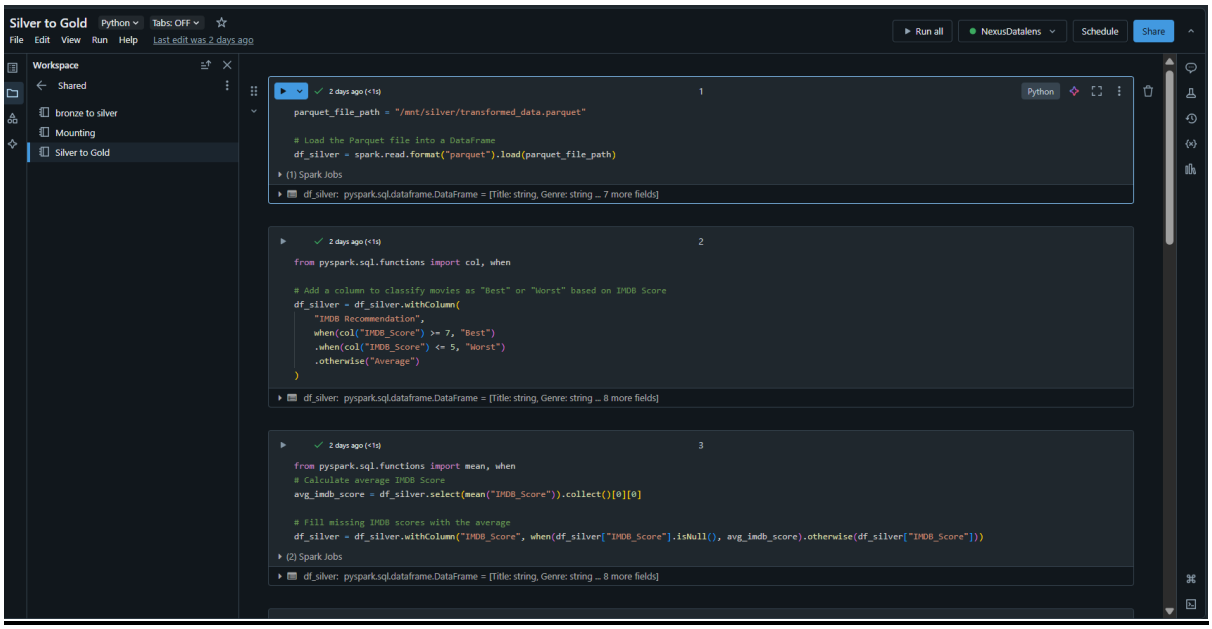
Create the Visualized Dashboard



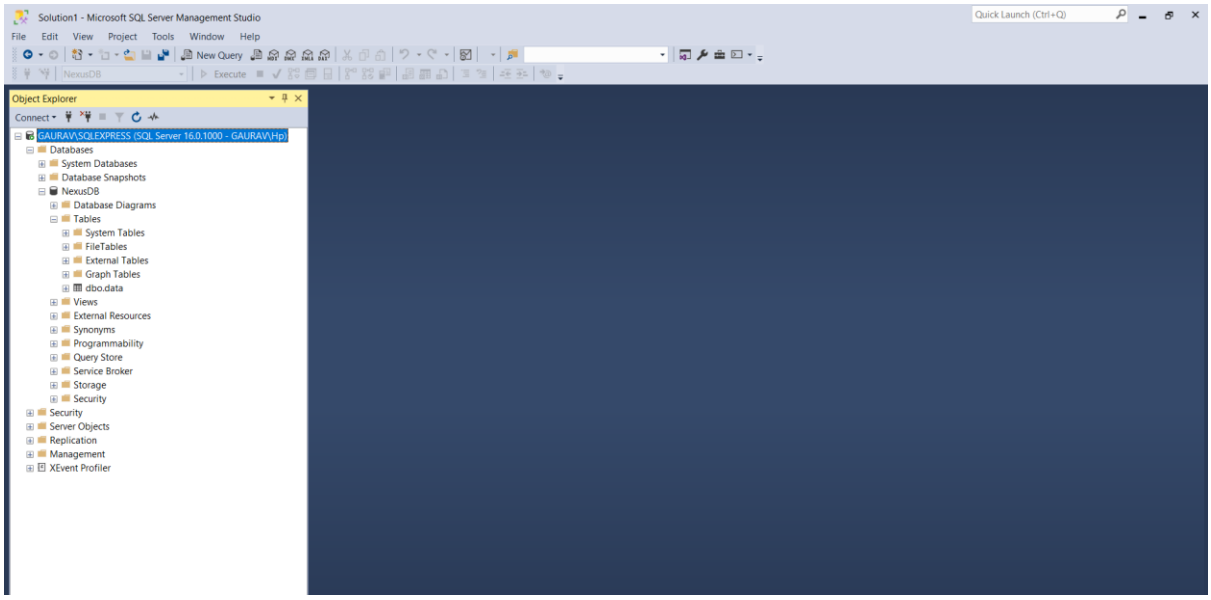
PIPELINE TESTING

Perquisites

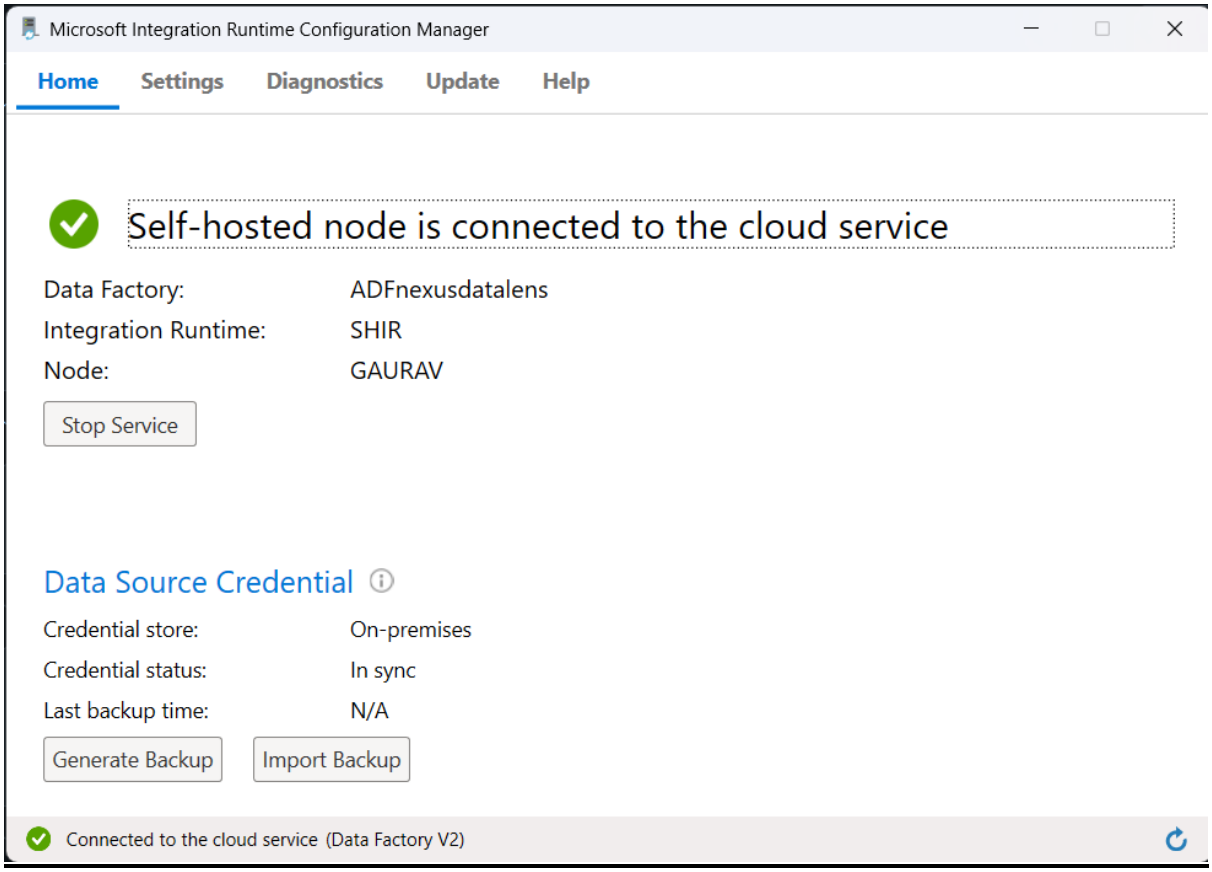
- Make sure your cluster is connected and attach to the Notebookbook.



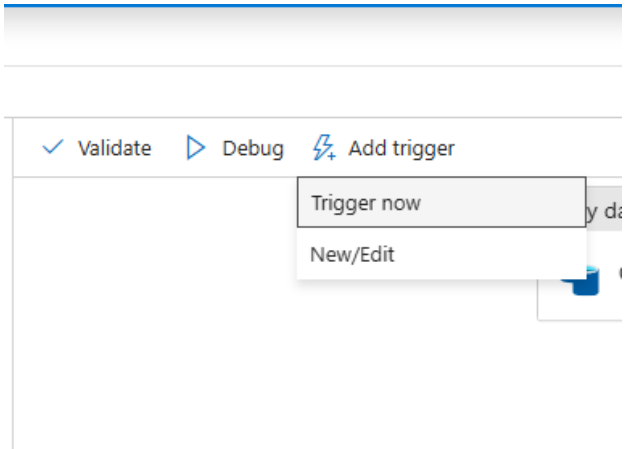
- SSMS Server should be online.

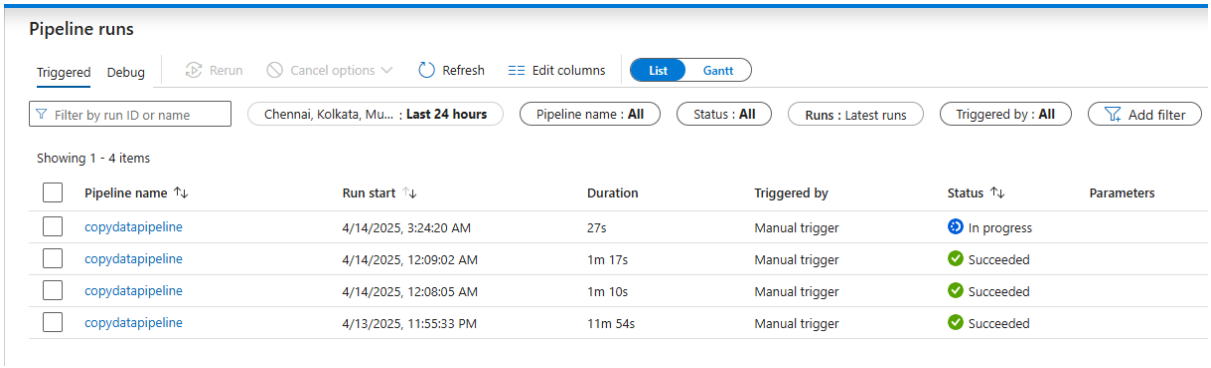


- On-premises Integration runtime is Connected

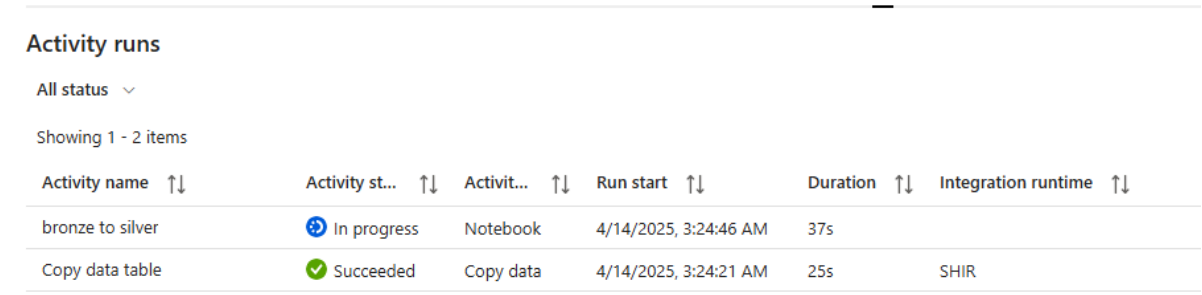


To run the Pipeline you need to Auther > copydatapipeline > Trigger > Trigger Now

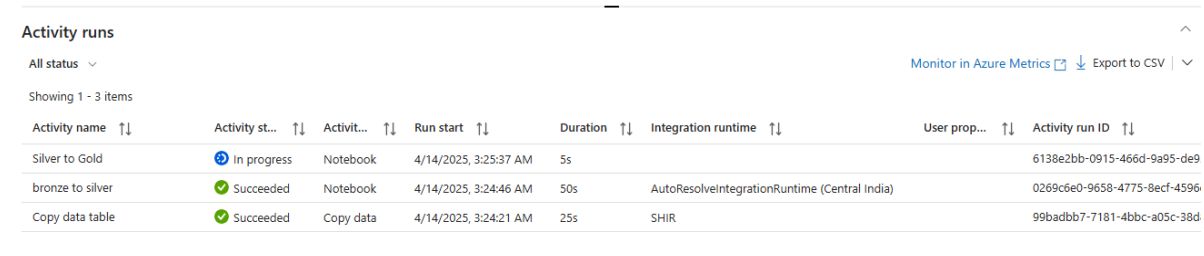




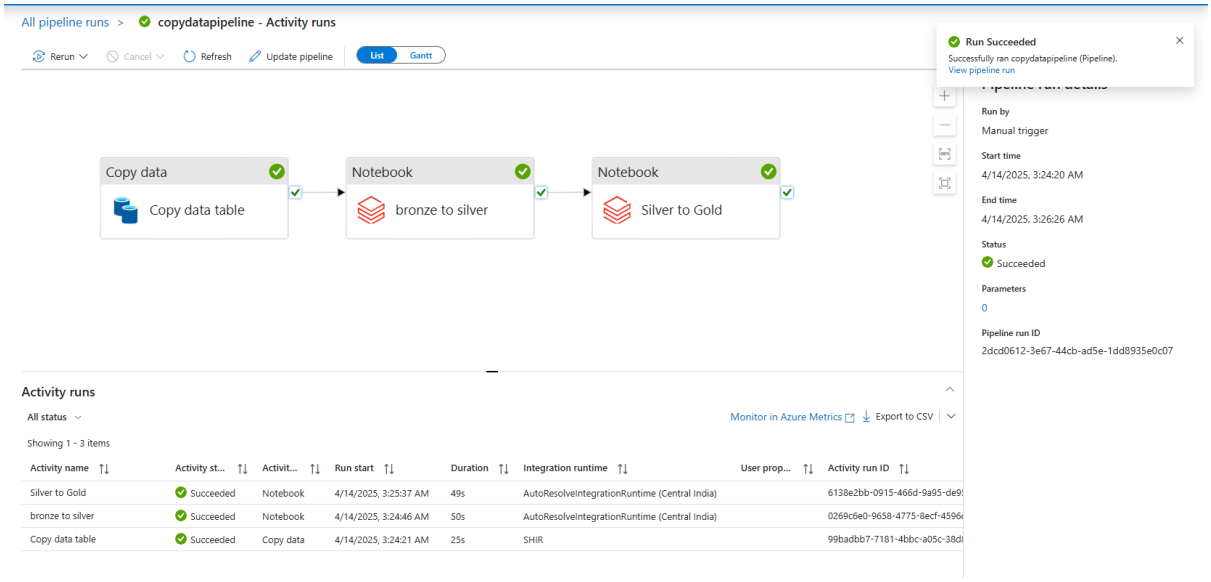
You can see the progress in activity runs
copy data table is Succeeded



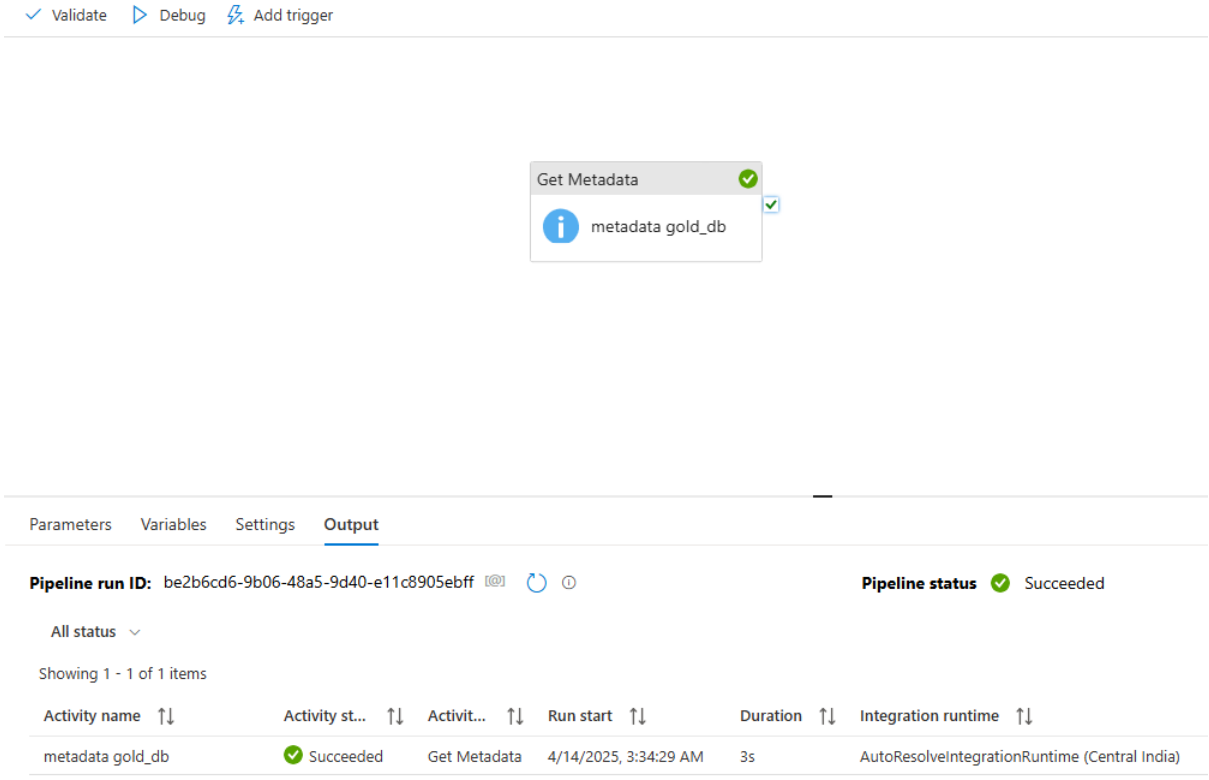
Bronze to silver notebook is succeeded



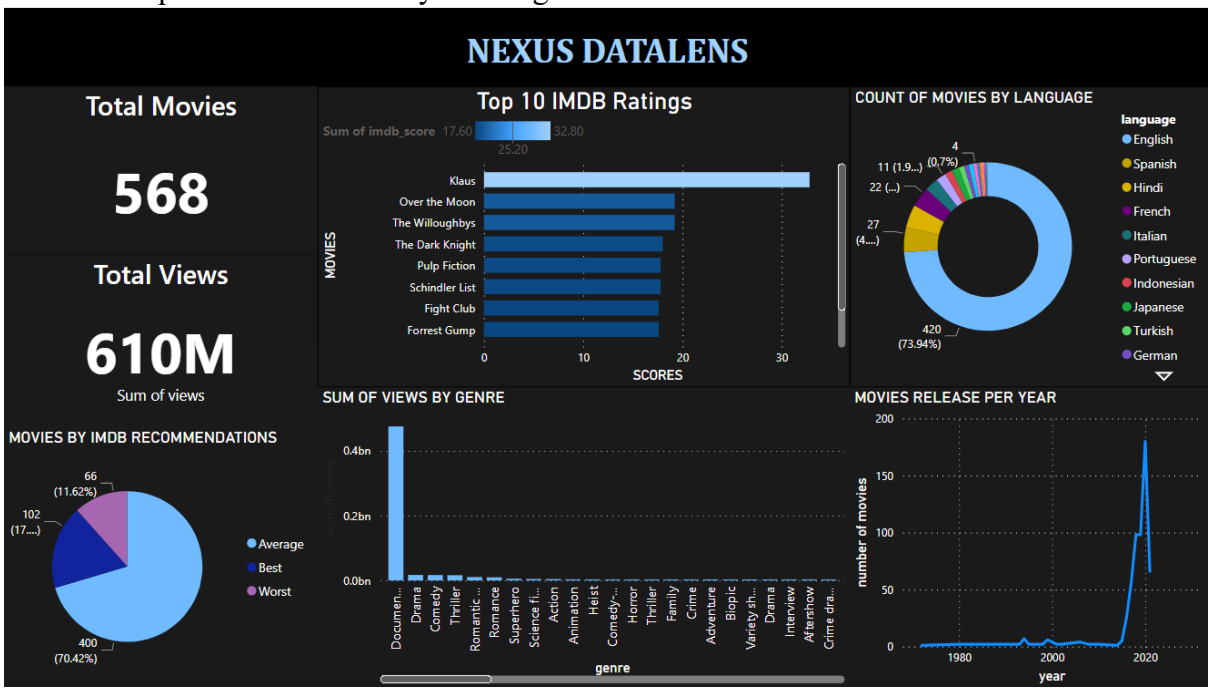
Silver to Gold notebook succeeded



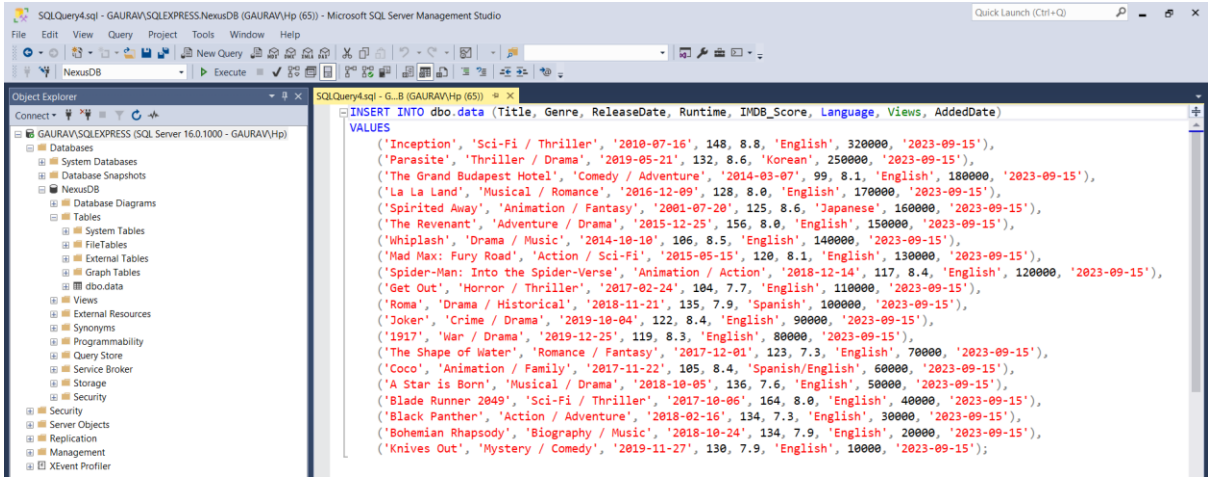
No need to rerun the synapse analytics pipeline it will get the views automatically with the help of scripts



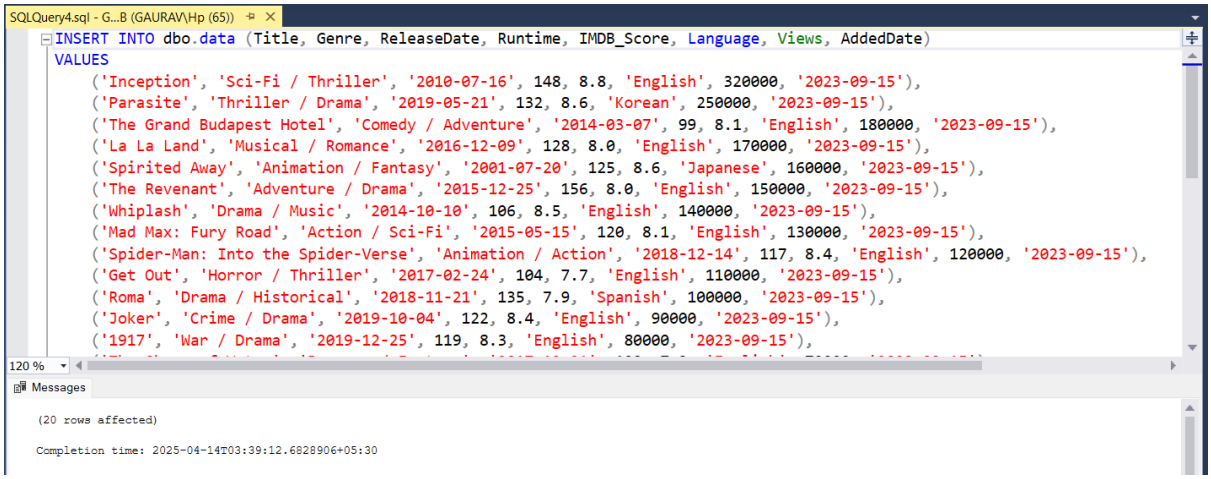
Refresh the power bi dashboard you will get the Visualized data



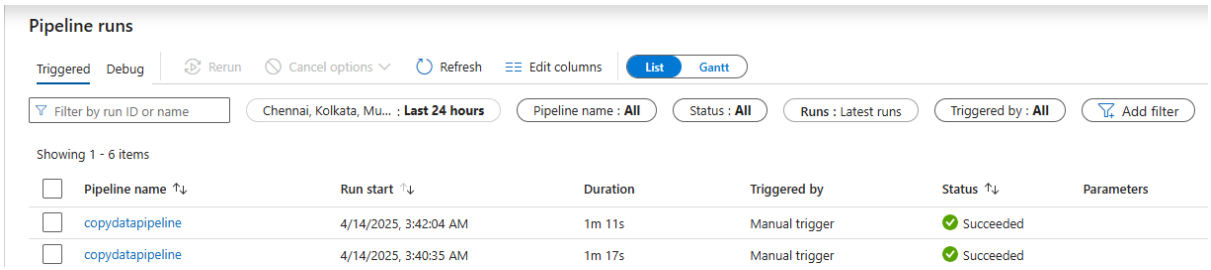
Got to SQL Server Management Studio (SSMS) for Incremental Data



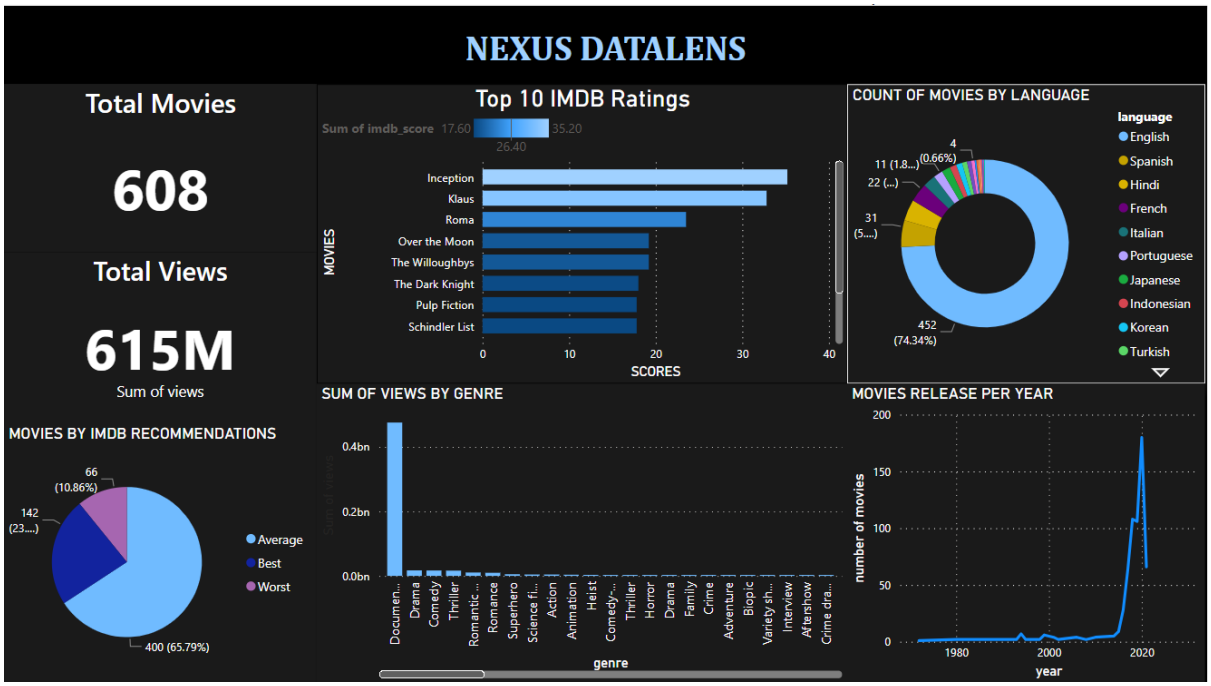
Run the Insert Query



Go to Azure Data factory and re run the Pipeline



Updated Data In the visualized format



FUTURE SCOPE

The current implementation of an incremental ETL pipeline using Azure services lays a strong foundation for data integration, transformation, and analytics. However, there are numerous opportunities to extend and enhance the system to meet evolving business requirements and leverage cutting-edge technologies. The following areas represent the future scope of the project:

1. Integration with Advanced Analytics and Artificial Intelligence

- Machine Learning Models : Extend the pipeline to incorporate predictive analytics and anomaly detection using Azure Machine Learning or AutoML . This will enable proactive decision-making based on historical trends and real-time insights.
- Natural Language Processing (NLP) : Integrate NLP capabilities to analyze unstructured data such as customer feedback, logs, or social media feeds stored in Azure Data Lake.
- Recommendation Systems : Build recommendation engines tailored to specific industries, such as personalized product suggestions in retail or treatment recommendations in healthcare.

2. Real-Time Data Processing

- Streaming Pipelines : Transition to real-time data processing by integrating Azure Event Hubs or Apache Kafka . This will support use cases like IoT data ingestion, fraud detection, and live monitoring of operational metrics.
- Delta Lake : Implement Delta Lake in Azure Data Lake Gen2 to enable ACID transactions and incremental updates for large-scale datasets, ensuring consistency and reliability in real-time workflows.

3. Scalability and Performance Optimization

- Serverless Architecture : Adopt serverless components such as Azure Functions or Databricks Serverless to reduce infrastructure management overhead and optimize costs.
- Auto-Scaling : Enable auto-scaling for compute resources in Azure Databricks and Azure Synapse Analytics to handle peak workloads efficiently.
- Data Partitioning : Optimize storage and query performance by partitioning data in Azure Data Lake based on attributes like date, region, or category.

4. Multi-Cloud and Hybrid Cloud Support

- Multi-Cloud Integration : Extend the pipeline to integrate with other cloud platforms like AWS or Google Cloud, providing redundancy and flexibility for global operations.
- Hybrid Cloud Enhancements : Strengthen hybrid cloud capabilities by supporting additional on-premises systems or edge devices for data ingestion and processing.

5. Enhanced Security and Compliance

- Advanced Encryption : Implement advanced encryption techniques, such as Homomorphic Encryption , to secure sensitive data during processing.
- Data Masking : Use Azure SQL Database Dynamic Data Masking to protect sensitive information during visualization and reporting.
- Compliance Automation : Automate compliance checks for regulations like GDPR, HIPAA, and CCPA using tools like Azure Policy and Compliance Manager .

6. User-Friendly Interfaces and Self-Service Analytics

- Custom Dashboards : Develop role-specific dashboards in Power BI tailored to executives, analysts, and engineers, ensuring relevant insights for each user group.
- Self-Service Tools : Enable non-technical users to create their own reports and analyses using Power Query or Power BI Templates .
- Chatbots for Insights : Integrate chatbots powered by Azure Bot Service to allow users to query data and receive insights via natural language.

7. Blockchain for Data Integrity

- Immutable Logs : Use blockchain technology to maintain immutable logs of data transformations, ensuring transparency and traceability throughout the pipeline.
- Smart Contracts : Implement smart contracts to automate workflows such as data validation, approval processes, and compliance checks.

8. Edge Computing and IoT Integration

- Edge Analytics : Process data at the edge using Azure IoT Edge to reduce latency and bandwidth usage, enabling real-time insights for IoT devices.
- IoT Data Pipelines : Extend the pipeline to ingest and process data from IoT devices, supporting use cases like predictive maintenance, smart city analytics, and environmental monitoring.

9. Cross-Industry Applications

The project can be adapted to serve diverse industries, each with unique requirements:

- Retail : Inventory optimization, demand forecasting, and personalized marketing.
- Healthcare : Patient data analytics, disease prediction, and telemedicine insights.
- Finance : Fraud detection, risk analysis, and regulatory reporting.
- Manufacturing : Predictive maintenance, supply chain optimization, and quality control.

10. Cost Optimization

- Cost Monitoring : Use Azure Cost Management to monitor and optimize resource usage, ensuring cost efficiency.
- Reserved Instances : Leverage reserved instances for predictable workloads to reduce costs.
- Spot Instances : Use Azure Spot VMs for non-critical tasks like batch processing or testing.

11. Collaboration and DevOps Enhancements

- CI/CD Pipelines : Implement continuous integration and deployment (CI/CD) for ETL pipelines using Azure DevOps or GitHub Actions , ensuring seamless updates and version control.
- Collaborative Notebooks : Use collaborative notebooks in Azure Databricks to enable teams to work together on data transformations and analyses.

12. Sustainability and Green Computing

- Energy-Efficient Workflows : Optimize workflows to minimize energy consumption, aligning with sustainability goals.
- Carbon Tracking : Use Azure's Sustainability Calculator to track and reduce the carbon footprint of the data pipeline.

13. Data Governance and Metadata Management

- Data Catalog : Implement Azure Purview to create a centralized data catalog for metadata management and data discovery.
- Data Lineage : Track data lineage to understand how data flows through the pipeline and identify dependencies.

14. Global Expansion

- Multi-Region Deployment : Deploy the pipeline across multiple Azure regions to ensure high availability and low latency for global users.
- Localization : Support localization for dashboards and reports to cater to international audiences.

15. Open Source Contributions

- Open Source Tools : Contribute to open-source projects related to data engineering, such as Apache Spark or Delta Lake, to enhance community-driven innovation.
- Reusable Components : Package reusable components of the pipeline (e.g., transformation scripts) as open-source libraries for broader adoption.

CONCLUSION

Nexus Datalens offers a comprehensive, end-to-end data analysis solution designed to simplify the way users interact with and extract value from their data. With an intuitive, user-friendly interface, it allows seamless data uploads, efficient processing, and insightful visualizations, ensuring accessibility for both technical and non-technical users. Advanced security measures safeguard sensitive information, while robust analytical capabilities enable businesses to uncover trends, identify patterns, and make data-driven decisions with confidence. By leveraging cutting-edge technologies such as AI-driven insights, automation, and scalable cloud infrastructure, Nexus Datalens optimizes complex data workflows, enhances operational efficiency, and unlocks the full potential of data, driving innovation and business growth.

REFERENCES

1. SQL Server Management Studio (SSMS)
   https://learn.microsoft.com/en-us/ssms/sql-server-management-studio-ssms
2. Azure Data Factory
   https://learn.microsoft.com/en-us/azure/data-factory/
3. Azure Data Lake Storage Gen 2
   https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction
4. Azure Data Bricks
   https://learn.microsoft.com/en-us/azure/databricks/
5. Azure Synapse Analytics
   https://learn.microsoft.com/en-us/azure/synapse-analytics/
6. Power Bi
   https://learn.microsoft.com/en-us/power-bi/
7. Key Vault
   https://learn.microsoft.com/en-us/azure/key-vault/
8. Active Directory
   https://learn.microsoft.com/en-us/troubleshoot/windows-server/active-directory/active-directory-overview
9. Udemy Courses
   - End to End Azure Data Engineering Real Time Project
     https://www.udemy.com/course/end-to-end-azure-data-engineering-real-time-project/?couponCode=NVDIN35
   - Azure Data Factory +Synapse Analytics End to End ETL project
     https://www.udemy.com/course/azure-data-factory-synapse-analytics-end-to-end-etl-project/?couponCode=NVDIN35