

“AZURE DATA FACTORY”

-

**“MICROSOFT CLOUD SERVICE
FOR DATA INTEGRATION – ETL
and ELT.”**

PREPARED BY
Ankur Soni (2018BTechCSE125)

FACULTY GUIDE
Dr. Suman Saha



**Department of Compute Science Engineering
Institute of Engineering and Technology (IET)
JK Lakshmipat University Jaipur**

May 2022

“AZURE DATA FACTORY”

-

**“MICROSOFT CLOUD SERVICE
FOR DATA INTEGRATION – ETL
and ELT.”**

Practice School-2

Submitted in partial fulfilment of the requirements

for the degree of

Bachelor of Technology in Computer Science Engineering

By:

Ankur Soni (2018BTechCSE125)

Faculty Guide:

Dr. Suman Saha



**Department of Computer Science Engineering
Institute of Engineering and Technology (IET)
JK Lakshmipat University Jaipur**

May 2022

CERTIFICATE

This is to certify that the Practice School-2 project work entitled “**Azure Data Factory – Microsoft cloud service for data integration – ETL & ELT**” submitted by **Ankur Soni (2018BTechCSE125)**, towards the partial fulfillment of the requirements for the degree of **Bachelor of Technology in Computer Science Engineering** of JK Lakshmipat University Jaipur is the record of work carried out by them under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for Practice School-II examination.



Dr. Suman Saha

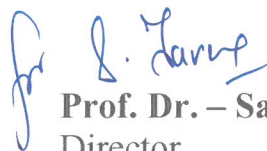
Associate Professor,
Department of Computer Science
Engineering
Institute of Engineering & Technology
(IET)
JK Lakshmipat University, Jaipur



Dr. Gireesh Kaushik

Co-ordinator PS2 and Asst. Professor,
Department of Computer Science
Engineering
Institute of Engineering & Technology
(IET)
JK Lakshmipat University, Jaipur

Date of Submission: 27-05-2022



Prof. Dr. – Sanjay Goel

Director
Institute of Engineering & Technology
JK Lakshmipat University Jaipur

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to Dr. Suman Saha, Associate Professor, Department of Computer Science & Engineering for his valuable time and guidance that made the project work a success.

My heartfelt thanks to Mr. Mayank Jhamb and Mr. Masthan Reddy Konda for his support during the project.. Thanking all my friends and all those who have helped me carrying out this work directly or indirectly without whom working on this project work was not possible.

I would also like to thank sincerely to Dr. Sonal Jain, Head of CSE Department and Dr. Sanjay Goel, Director, Institute of Engineering and Technology, JK Lakshmipat University for giving us a platform to carry out the project.



Sincerely yours,

Ankur Soni (2018BTechCSE125)

TABLE OF CONTENTS

Topic	Page No.
CERTIFICATE	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	1
LIST OF ACRONYMS	3
CHAPTER 1: INTRODUCTION	
1.1 About Organization	4
1.2 Purpose	5
1.3 Objectives	5
1.4 Solutions	6
CHAPTER 2: TECHNOLOGY STACK	
3.1 Microsoft Azure	8
3.2 Azure Data Factory (ADF)	8
3.3 Azure Synapse Analytics (ASA)	9
3.4 Databricks	10
3.5 Azure Logic Apps (ALA)	10
CHAPTER 3: WORKFLOW / RESULTS	11
CHAPTER 4: CONCLUSION	22
CHAPTER 5: FUTURE WORK	24
REFERENCES	27

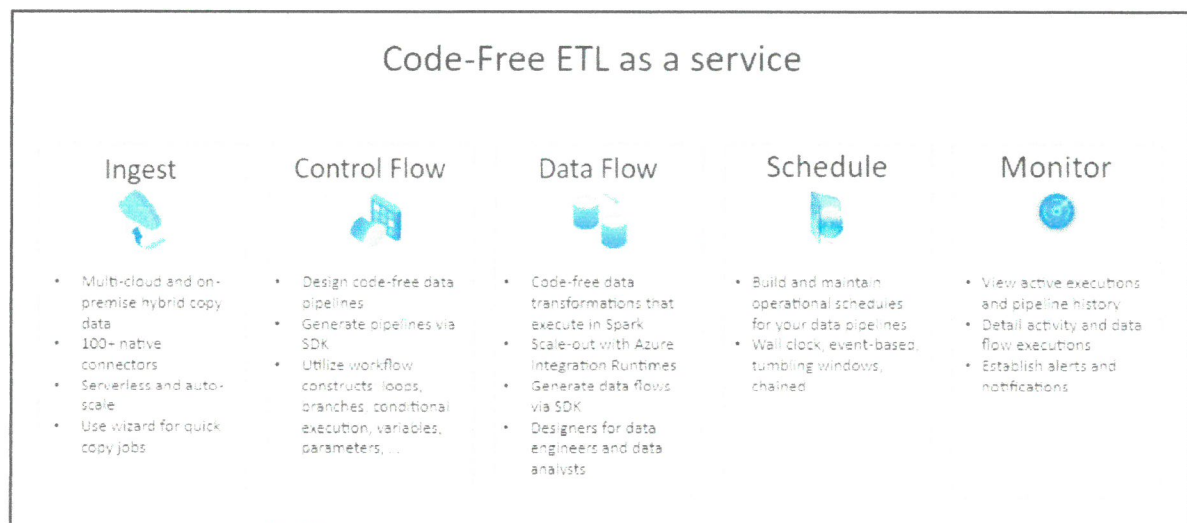
ABSTRACT

In the world of big data, raw, unorganized data is often stored in relational, non-relational, and other storage systems. However, on its own, raw data doesn't have the proper context or meaning to provide meaningful insights to analysts, data scientists, or business decision makers. Big data requires a service that can orchestrate and operationalize processes to refine these enormous stores of raw data into actionable business insights.

Azure Data Factory is a managed cloud service that's built for these complex hybrid extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects.

How Does It Work?

Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers.



Top-level concepts

An Azure subscription might have one or more Azure Data Factory instances (or data factories). Azure Data Factory is composed of below key components.

- **Pipelines** : The graphical user interface where you place widgets and draw data paths.
- **Activities** : A graphical widget that does something to your data.

- **Datasets** : An explicitly defined set of data that ADF can operate on.
- **Linked services** : The connection information that allows ADF to access a specific outside data resource.
- **Integration Runtimes** : A glue/gateway layer that lets ADF talk to software outside of itself.
- **Triggers** : Triggers represent the unit of processing that determines when a pipeline execution needs to be kicked off. There are different types of triggers for different types of events.

These components work together to provide the platform on which you can compose data-driven workflows with steps to move and transform data.

Usage Scenarios

For example, imagine a gaming company that collects petabytes of game logs that are produced by games in the cloud. The company wants to analyze these logs to gain insights into customer preferences, identify up-sell and cross-sell opportunities and provide a better experience to its customers.

To analyze these logs, the company needs to use reference data such as customer information, game information, and marketing campaign information that is in an on-premises data store. The company wants to utilize this data from the on-premises data store, combining it with additional log data that it has in a cloud data store.

Azure Data Factory is the platform that solves such data scenarios. It is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale. Using Azure Data Factory, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores. You can build complex ETL processes that transform data visually with data flows or by using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database.

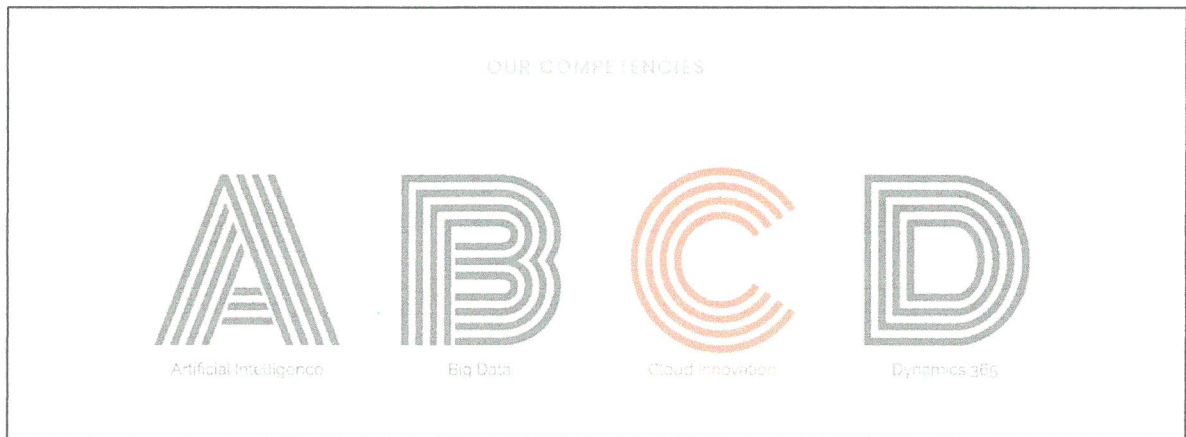
Additionally, you can publish your transformed data to data stores such as Azure Synapse Analytics for business intelligence (BI) applications to consume. Ultimately, through Azure Data Factory, raw data can be organized into meaningful data stores and data lakes for better business decisions.

List of Acronyms

Abbreviation	Full form
Wip	Wipro
Ap	Apollo Pharmacy
Az	Azure
Adf	Azure Data Factory
Rg	Resource group
La	Logic Apps
SHIR	Self-hosted Integration Runtime
ADLS	Azure Datalake Storage
ASA	Azure Synapse Analytics
AKv	Azure Key Vault
AM	Azure Monitor
AAD	Azure Active Directory
Vm	Virtual machine

CHAPTER 1 : INTRODUCTION

1.1 About Organization



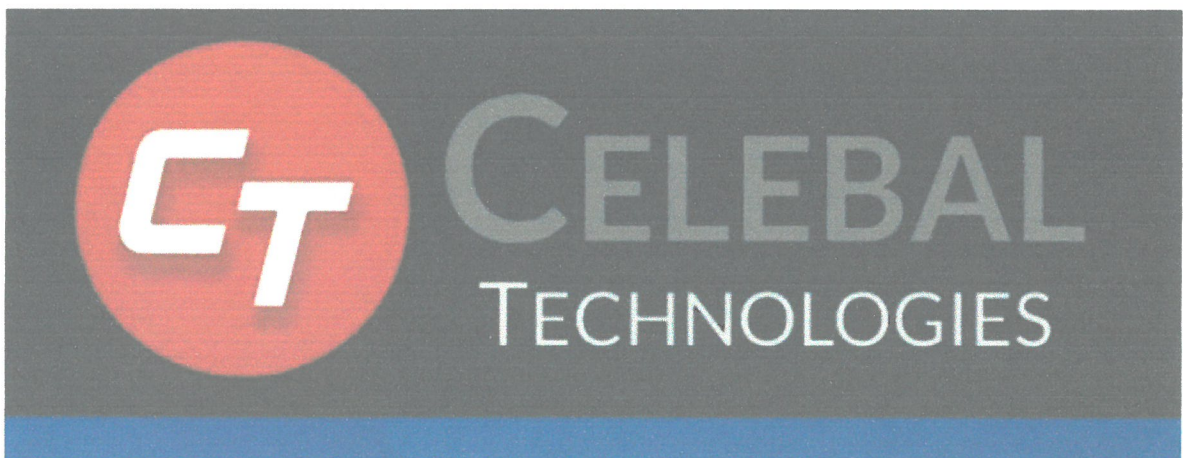
A- ARTIFICIAL INTELLIGENCE

B- BIG DATA

C- CLOUD INNOVATION

D- DYNAMIC 365

Celebal Technologies is a premier software services company in the field of Data Science, Big Data, and Enterprise Cloud. We help you achieve a competitive advantage with intelligent data solutions, built using cutting-edge technology. Our core offerings are around Cloud Innovation, Supply Chain Analytics, Chat Bots, Power Platforms, and Data Analytics. Our solutions can help you accelerate decision-making and take giants leaps in your digital transformation journeys.



1.2 Purpose



Wipro Limited is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship.

The objective is to build a modern data platform that can help Wipro to move their data to the cloud and get better insights from all the data being generated to enable them to make informed decisions in charting out their growth story.

1.3 Objectives

OBJECTIVE 1: Setting up the Azure environment for Wipro to connect with and move their data from multiple data sources like Json files, API's, SQL databases etc to the azure cloud for multiple tenants.

OBJECTIVE 2: Setting up the Data Ingestion layer for moving data from the respective data sources to the the staging area.

OBJECTIVE 3: Setting up the Data Analytics layer for doing all the analytics work and pass that data to the Data Visualization layer to find the meaningful insights.

OBJECTIVE 4: Setting up the Azure logic app so that the tech team of Wipro can check the pipeline status (For ex: Data moved per day, pipeline got successful or failed, amount of data moved in Gb's or Tb's etc) via mail so that they won't need to open their azure portal to check pipeline status every single day.

1.4 Solution

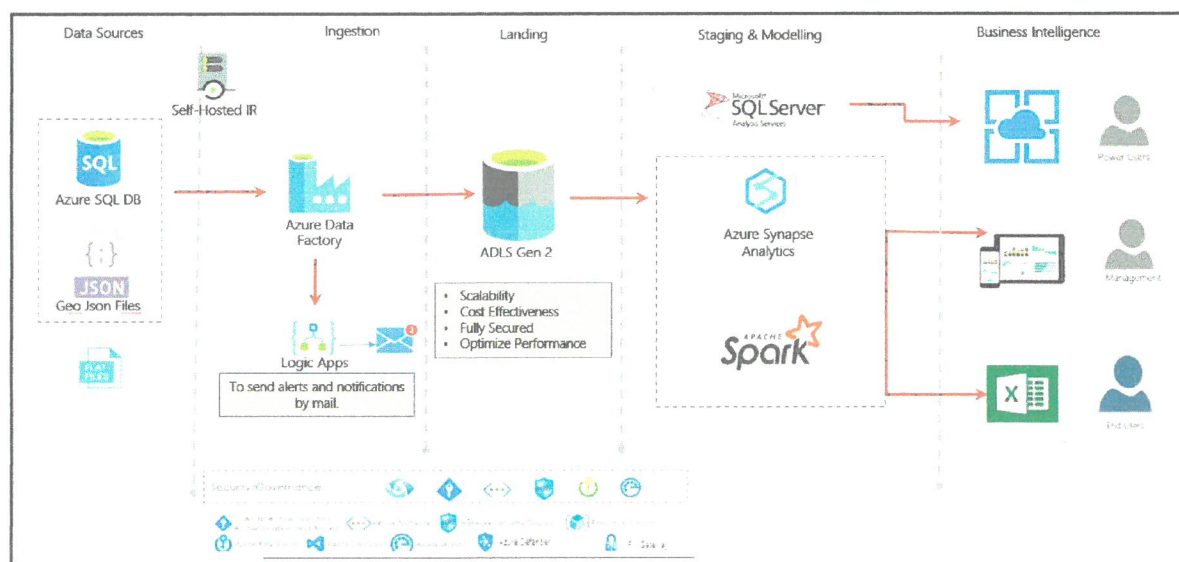


Figure 1.3: Technical Architecture

- **Resource group** - We have to create a single resource group with given naming convention. That holds all the related resources to the solution, in this case it will be holding ADF and Azure SQL database.
- **Azure Data Factory** - Azure Data Factory is Azure's cloud ETL service for scale-out serverless data integration and data transformation. Both the linked services be it SHIR or Azure SQL DB services will be created in ADF.
- **Self-Hosted Integration Runtime (SHIR) Linked Service** - SHIR serves as a secure gateway between internal and external networks, it passes encrypted credentials to access internal resources and encrypts outbound data. In this case on successful registration the connection between on-prem and ADF will be established.
- **ADLS Gen 2** - Azure Data Lake Storage Gen2 (also known as ADLS Gen2) is a next-generation data lake solution for big data analytics. Azure Data Lake Storage Gen2 builds Azure Data Lake Storage Gen1 capabilities—file system semantics, file-level security, and scale—into Azure Blob storage, with its low-cost tiered storage, high availability, and disaster recovery features.
- **Azure Synapse Analytics** - Azure Synapse Analytics is a limitless analytics service that brings together data integration, enterprise data warehousing and big data analytics. It gives you the freedom to query data on your terms, using either serverless or dedicated resources—at scale.

- **Azure SQL DB Linked Service**– This serves as a gateway to connect to our SQL DB that we created on the Azure portal.
- **Key vault** - Azure Key Vault is a cloud service for securely storing and accessing secrets. A secret is anything that you want to tightly control access to, such as API keys, passwords, certificates, or cryptographic keys. Key Vault service supports two types of containers: vaults and managed hardware security module(HSM) pools.
- **Azure SQL DB**– We have to create a SQL database ,Azure SQL Database is a fully managed database service, which means that Microsoft operates SQL Server for you, and ensures its availability and performance. SQL Database also includes innovative features to enhance your business continuity, such as built-in high availability.
- **Apache Spark** - Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its **in-memory cluster computing** that increases the processing speed of an application.

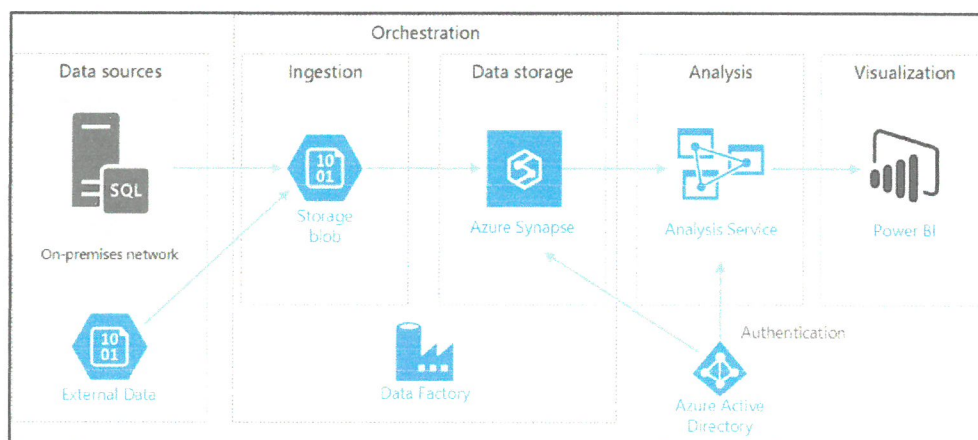
CHAPTER 2 : TECHNOLOGY STACK

2.1 Microsoft Azure

At its core, Azure is a public cloud computing platform—with solutions including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) that can be used for services such as **analytics, virtual computing, storage, networking, and much more.**

2.2 Azure Data Factory (ADF)

Azure Data Factory is a Microsoft cloud service offered by the Azure platform that allows data integration from many different sources. Azure Data Factory is a perfect solution when in need of building hybrid extract-transform-load (ETL), extract-load-transform (ELT) and data integration pipelines.

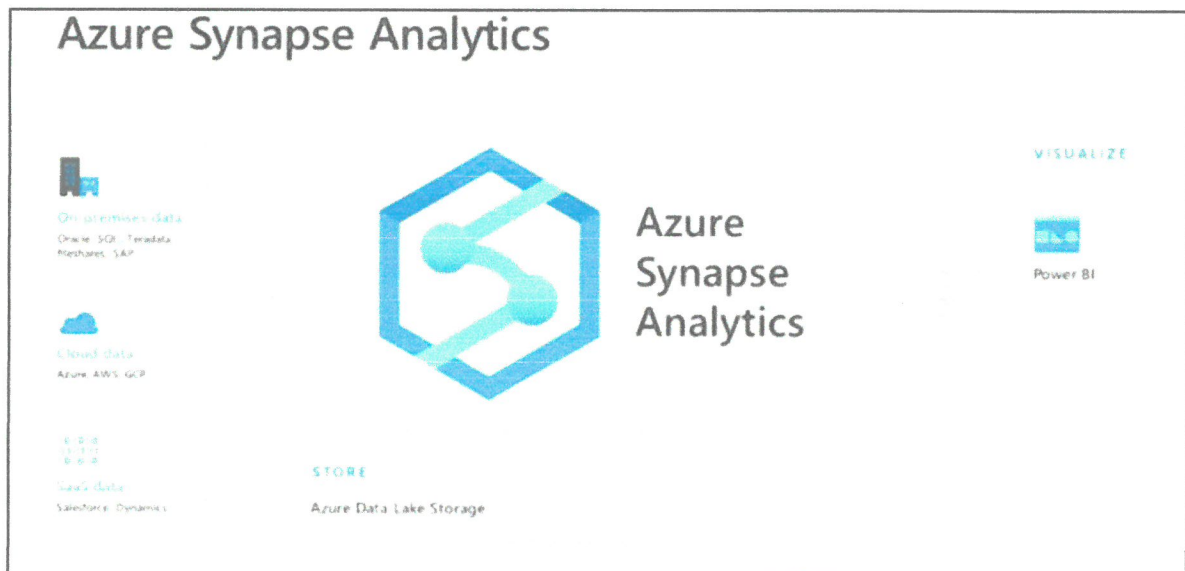


2.3 Azure Synapse Analytics

Azure Synapse Analytics is a limitless analytics service that brings together data integration, enterprise data warehousing and big data analytics. It gives you the freedom to query data on your terms, using either serverless or dedicated resources—at scale. Azure Synapse brings these worlds together with a unified experience to ingest, explore, prepare, transform, manage and serve data for immediate BI and machine learning needs.

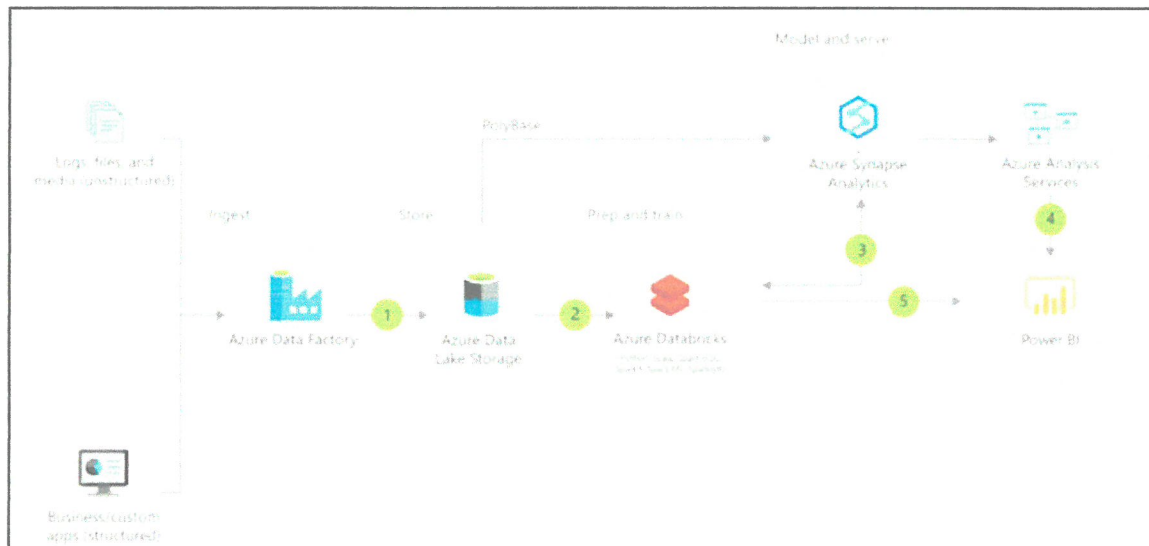


Azure Synapse Analytics



2.4 Databricks (Azure)

Azure Databricks, is a fully managed service which provides powerful ETL, analytics, and machine learning capabilities. Databricks provides a unified, open platform for all your data. It empowers data scientists, data engineers and data analysts with a simple collaborative environment to run interactive and scheduled data analysis workloads.



2.5 Azure Logic Apps (ALA)

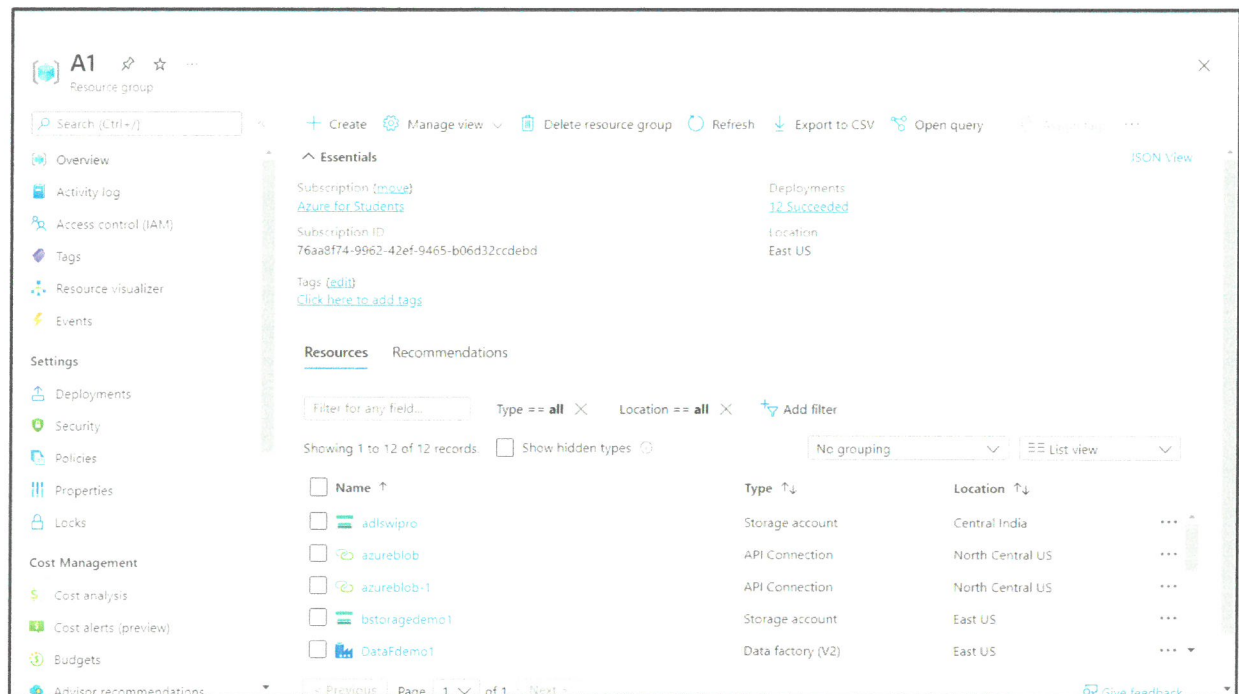
Azure Logic Apps is a cloud-based platform for creating and running automated workflows that integrate your apps, data, services, and systems. With this platform, you can quickly develop highly scalable integration solutions for your enterprise and business-to-business (B2B) scenarios.

The following list describes just a few example tasks, business processes, and workloads that you can automate using the Azure Logic Apps service:

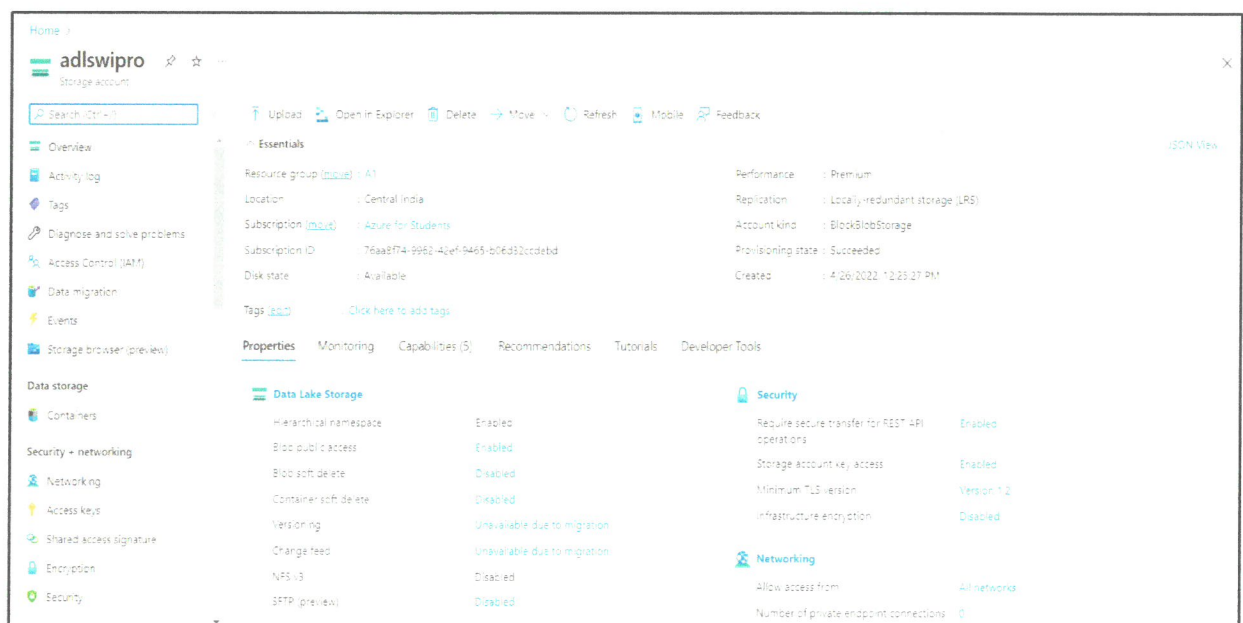
- Schedule and send email notifications using Office 365 when a specific event happens, for example, a new file is uploaded.
- Route and process customer orders across on-premises systems and cloud services.
- Move uploaded files from an SFTP or FTP server to Azure Storage.
- Monitor tweets, analyze the sentiment, and create alerts or tasks for items that need review.

CHAPTER 3 : WORKFLOW & RESULT

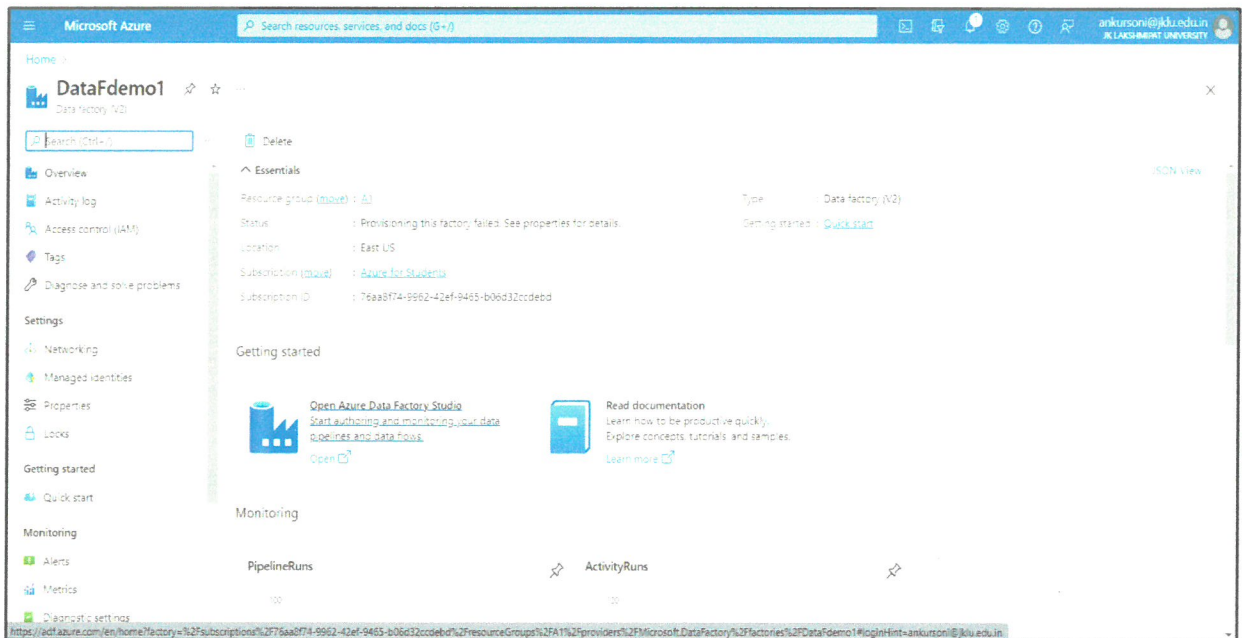
1. Create a resource group on Azure to deploy the resources on that.



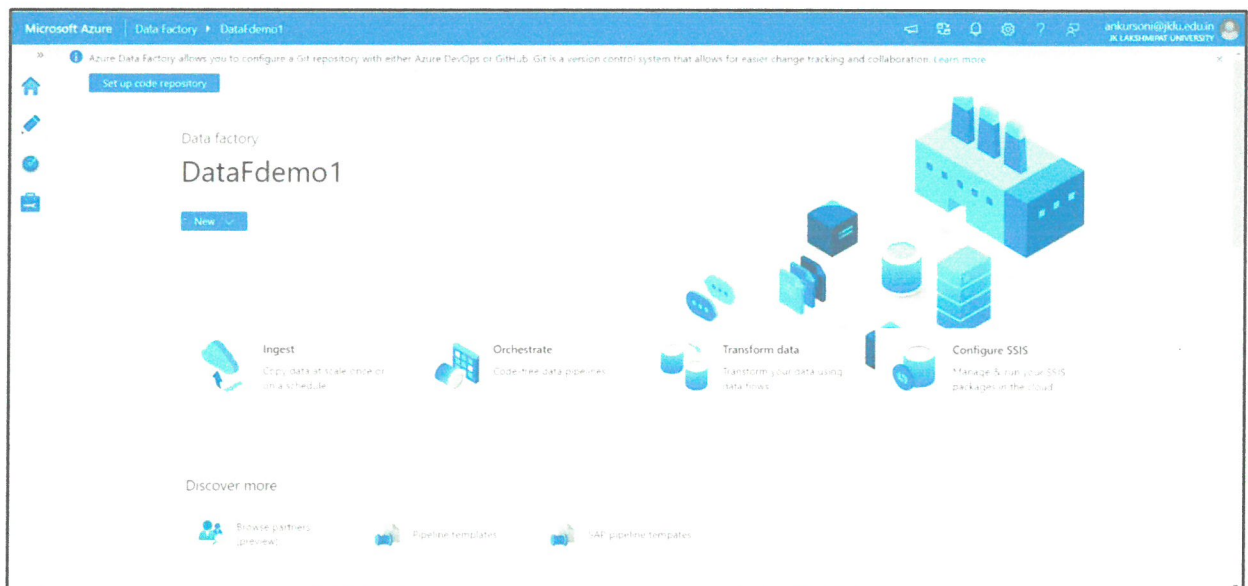
2. Creating storage account in the Resource group for the staging purpose.

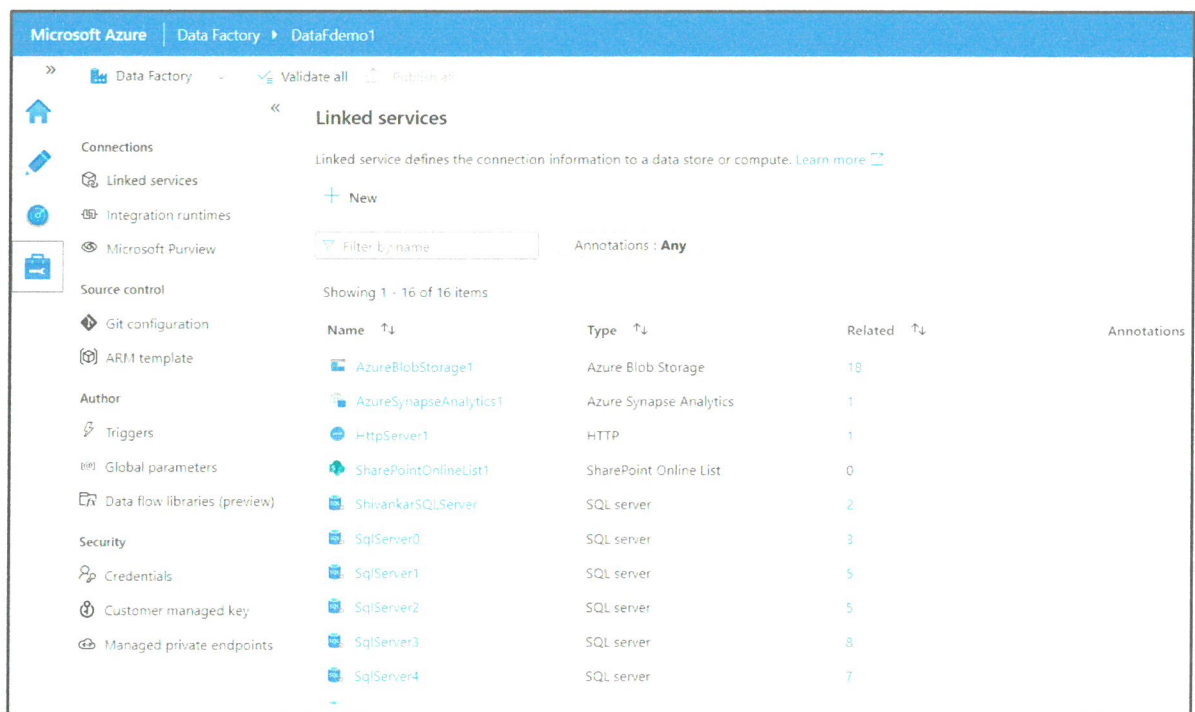
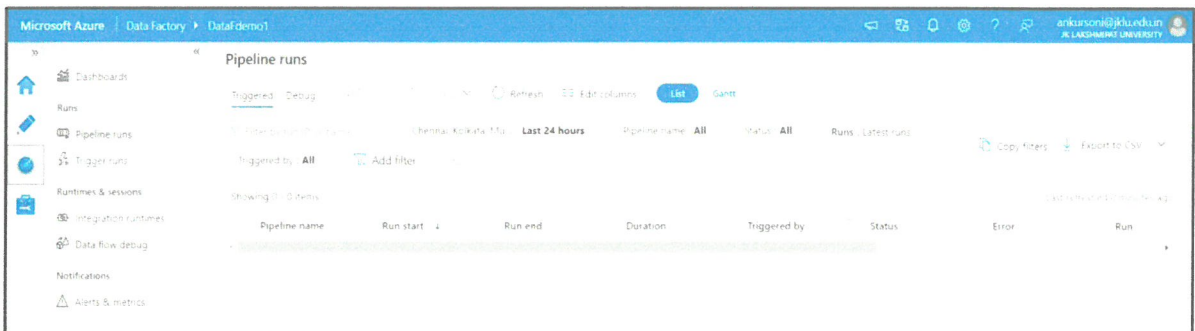
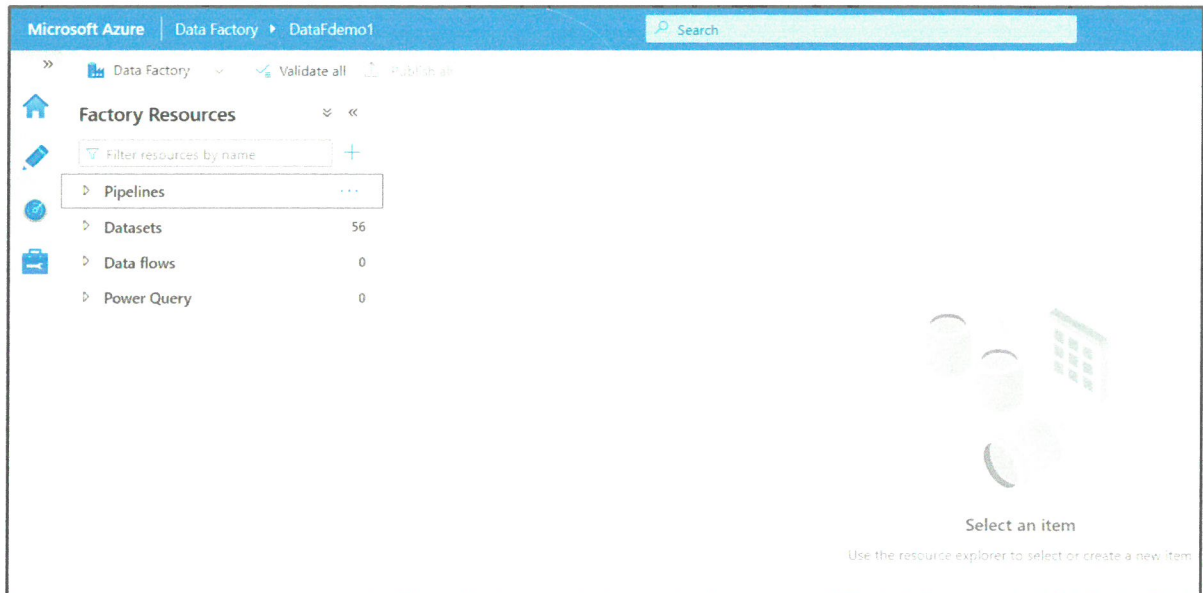


3. Creating data factory for ETL purpose and to perform all the data movement related activity from the source to the destination.

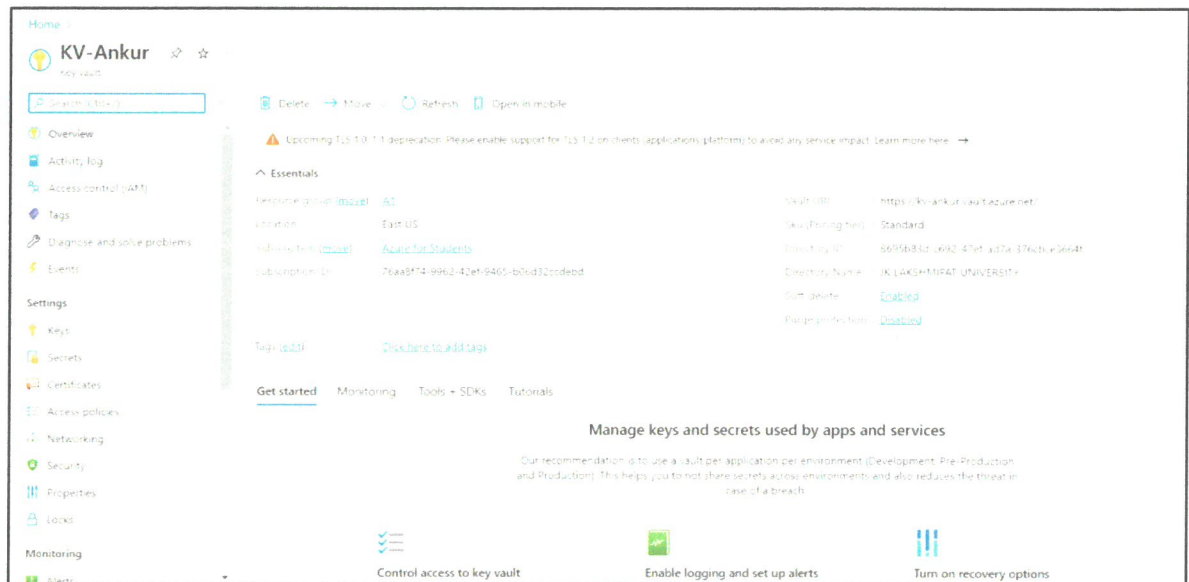


How the Data Factory looks like.

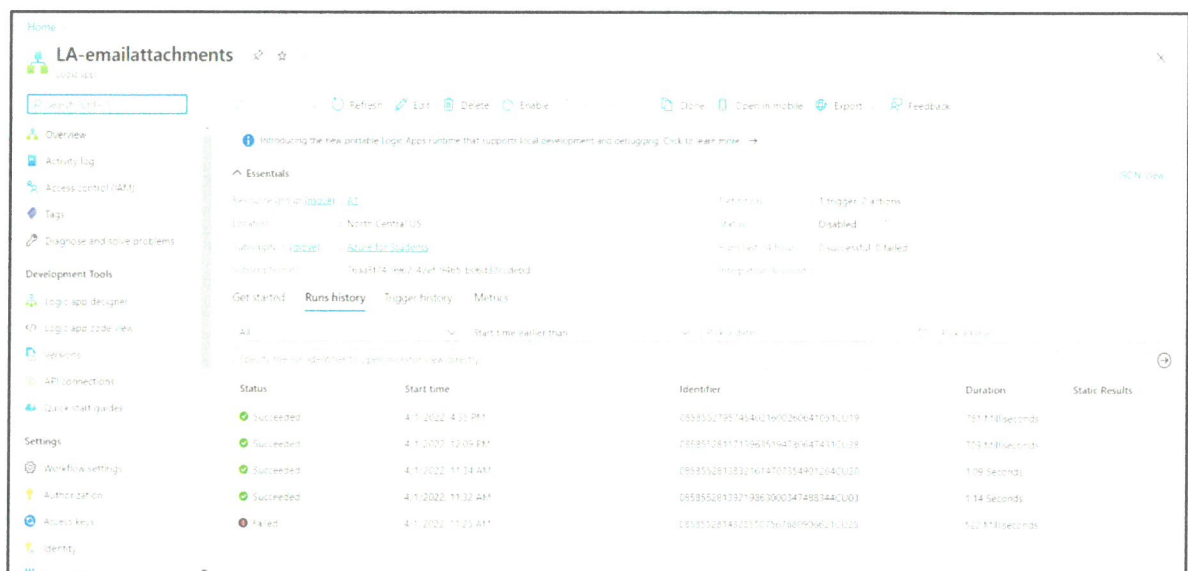




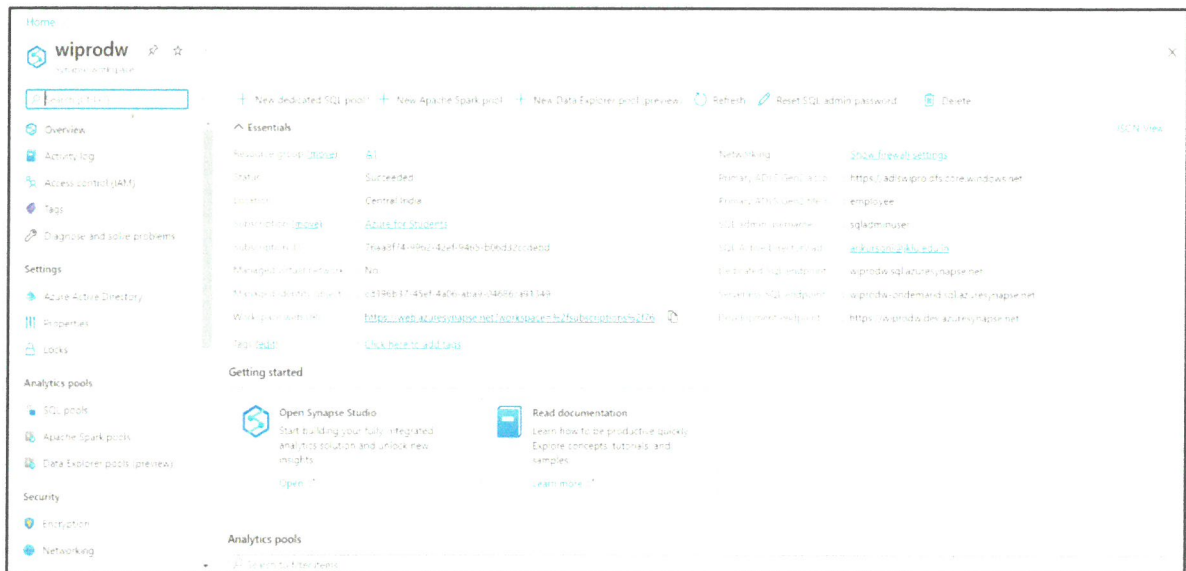
4. Create Azure Key Vault to store all the secrets (credentials) that we'll require in making linked services for the data sources.



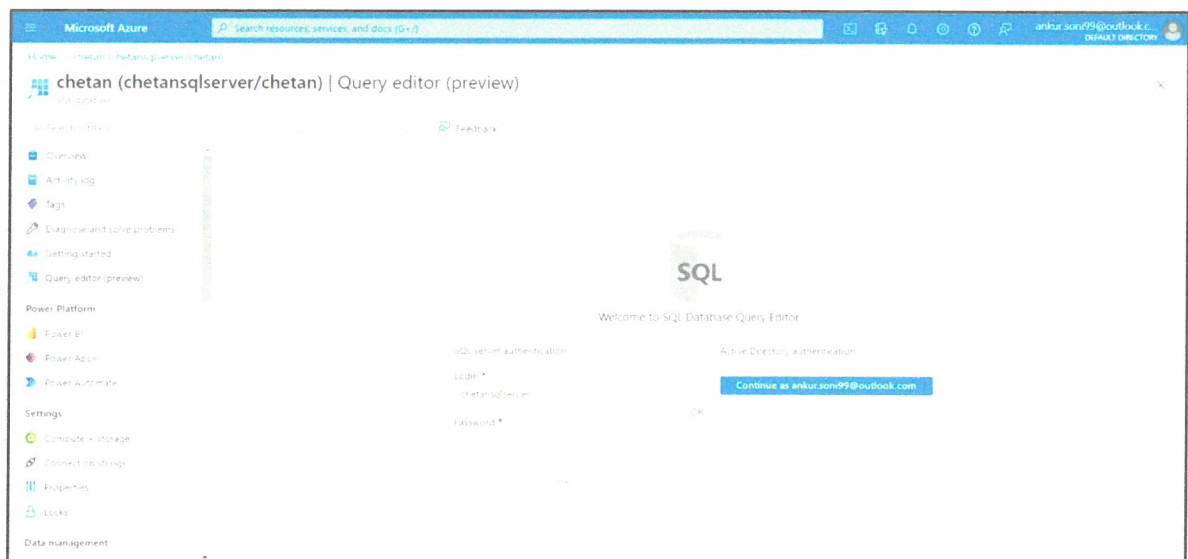
5. Create logic app so that the user can monitor the pipeline status and other information about the pipeline via mail.



6. Create synapse workspace so that data can be moved here for all the analytical operations that need to be performed.



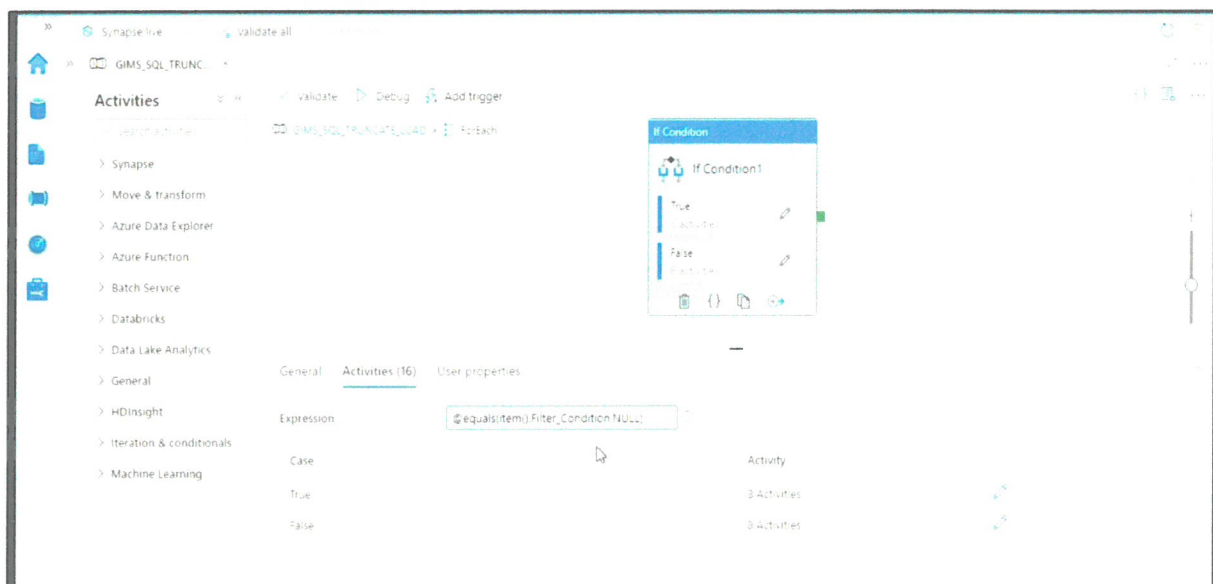
7. Create Sql DB so that we can connect to client's data and do all the desired operations like data movement (ETL or ELT).



8. Now we'll create the ETL pipelines for moving the data from the data source to it's destination.



Here we are doing truncate load which means truncate all the data if present into the destination area and then load the whole data from the lookup table.



Here we are putting a condition in the IF-Condition activity and if it satisfies then the activities in the True section will get executed or else the activities of the False section.



Here we are executing the True Section Activities which includes moving data from SQL database to ADLS (staging area) and then to synapse.

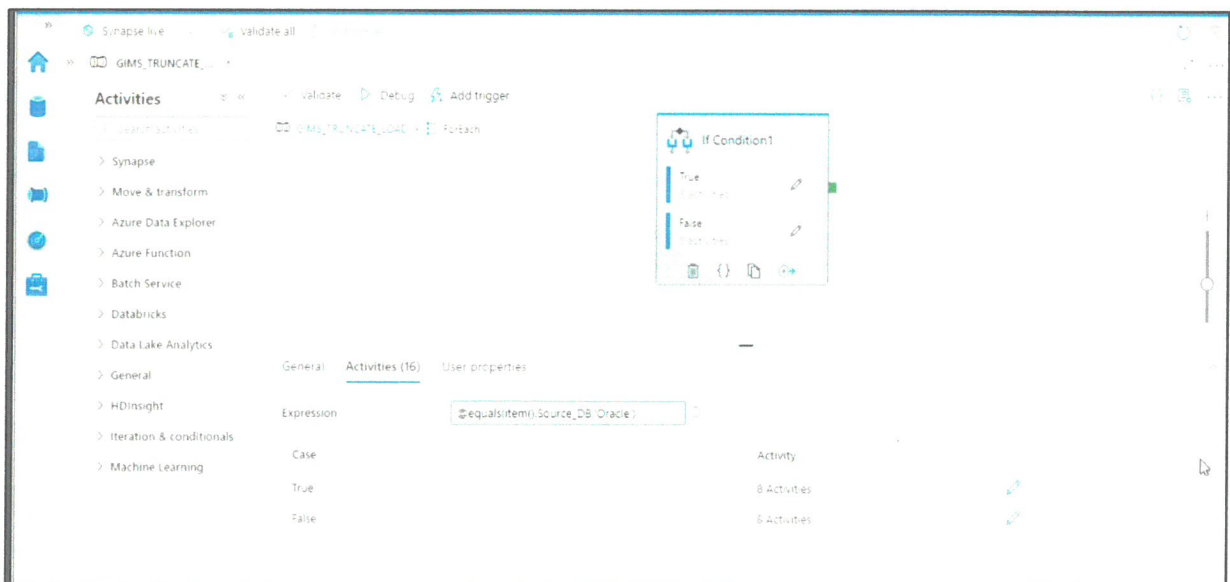


Here If the condition returns false then these activities will get executed which is the same scenario as mentioned above i.e. moving data from SQL to ADLS and then from ADLS to Synapse Analytics.

9. Now, we'll do the truncate load for Oracle Database as the data source.



Here we are fetching all the information stored in lookup table so that we can use those as a parameter to filter our data in the data movement process.



Here we are checking that if the Source_DB is equals to 'Oracle' then the activities in the True section will get executed else if Source_DB is something else then all the activities under the False section will get executed.



Here as the condition satisfied we are moving the data from Oracle to Azure Datalake storage (ADLS) then from ADLS to synapse for further process of analytics.

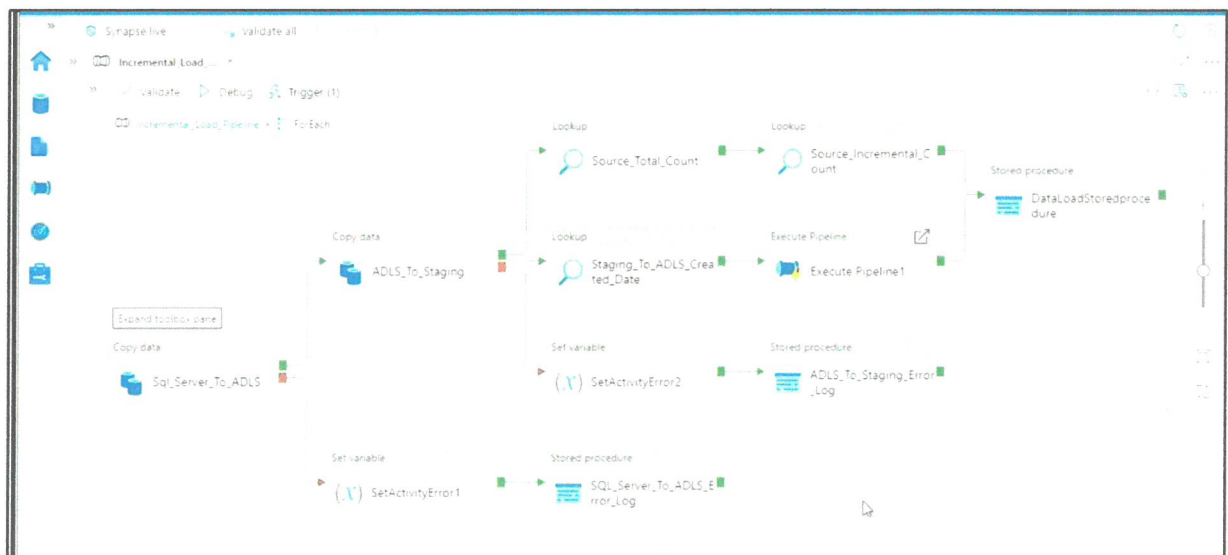


Here when the condition got failed then we are copying the data from SAP (data source) to ADLS as staging and then from ADLS to Synapse for the rest of the analytics.

10. Now after the truncate load from various data sources and we'll do Incremental load which is to copy the new/changed data in the database to the destination so that data gets updated everytime it changes in the source.



Here we are picking up the data from the lookup table such as 'lastModifiedDate' which is the date when the data got last modified and 'lastUpdatedDate' which is the date mentioned in the lookup table which is about when we last did the incremental load. So the gap between these two dates provides us the new data which needs to be copied into the destination table.



Here for each time we are checking and then running the pipeline so that we can get the new data of every table and copy it to ADLS first and then to synapse. And then updating the 'lastUpdatedDate' in the lookup table through stored procedure.

RESULTS:

When moving workloads (data and code) to Microsoft Azure, the ideal solution to migrate these existing ETLs is to Azure Data Factory (ADF). ADF provides for the automatic movement and transformation of data at scale in the cloud.

Data Factory, as a complex ETL process, moves, transforms, and controls data. The tool is highly automated and helps to orchestrate your data efficiently. So, Azure Data Factory performance allows spending minimum time setting the tool up, thus having more time to get insights. And You can easily migrate ETL workloads from on-premises EDWs and Data Lakes, to the Azure cloud. ADF use can be used to deploy, run, and manage ETL packages.

And as Azure Data Factory offers a global cloud presence, with data movement available in over **25 countries and protected by Azure security infrastructure** which makes it highly accessible.

CHAPTER 4 : CONCLUSION

Data Factory is a great tool for the rapid transition of data onto the cloud. Data Factory provides easy data integration on-premises and on the cloud. Data Factory offers you the possibility to easily integrate cloud data with on-premises data. It's unique in its ease of use yet its ability to transform and enrich complex data. It delivers data integration which is scalable, available and at low costs. Today this service is a crucial building block in any data platform & machine learning project.

4.1 What doesn't Azure Data Factory do?

Data Factory isn't SSIS (SQL Server Integration Services) in the cloud. It has less database specific features and focuses on supporting broader data transformation & movements (incl. big datasets, incl. data lake operations).

Data Factory can, however, run your SSIS packages in the Cloud (once build in SSIS). This allows to leverage Data Factory's scalability with SSIS's advanced ETL features.

		Small data	Close to real-time data (every minute)	Big Data
Azure Data Factory (ADF V2)		Yes	Yes	Yes
Integration (SSIS)	Services	Yes	No	No

4.2 How does it differ from other ETL Tools?

Data Factory is one option to use as cloud ETL (or ELT) tool. There are some features that distinguish Azure Data Factory from other tools.

- It also has the ability to run SSIS packages
- It auto-scales (fully managed PaaS product) based on the given workload.
- It allows to run up to once per minute
- It bridges on-premise & Azure Cloud seamlessly through a gateway
- It can handle big data volumes
- It can connect & work together with other compute services (Azure Batch, HDInsights) to even run truly big data computations during ETL.

Synapse Analytics is an integrated platform service from Microsoft Azure that combines the capabilities of data warehousing, data integrations, ETL pipelines, analytics tools & services, the scale for big-data capabilities, visualization & dashboards.

4.3 Where it fits?

The most common business use-cases for Azure Synapse Analytics are:

- Data Warehouse: Ability to integrate with various data platforms and services.
- Descriptive/Diagnostic Analytics: Use T-SQL queries against the Synapse database to perform data exploration and discovery.
- Realtime Analytics: Azure Synapse Link enables integration with disparate operational data sources to implement real-time analytics solutions.
- Advanced Analytics: Uses Azure Databricks to support decision-making by leveraging Azure Databricks.
- Reporting & Visualization: Integrate with PowerBI to empower and enhance business decision-making.

CHAPTER 5 : FUTURE WORK

5.1 About Organization



Apollo Pharmacy is a part of Apollo Hospitals - India's trusted pharmacy. It is India's first and largest branded pharmacy network, with over 4500 outlets in key locations. Accredited with - International Quality Certification, Apollo Pharmacy offers genuine medicines round-the-clock, through our 24-hour Pharmacies. Apollo Pharmacy also provides customer care any time of the day. Quality is the cornerstone of our existence. Apollo Pharmacy is well stocked with medicines, OTC and FMCG products, manned by a team of qualified and experienced staff. Apollopharmacy.in has more than 5000 products in various categories like Vitamins and Supplements, Baby Care, Personal Care, Health & Nutrition Foods and OTC.

5.2 Objectives

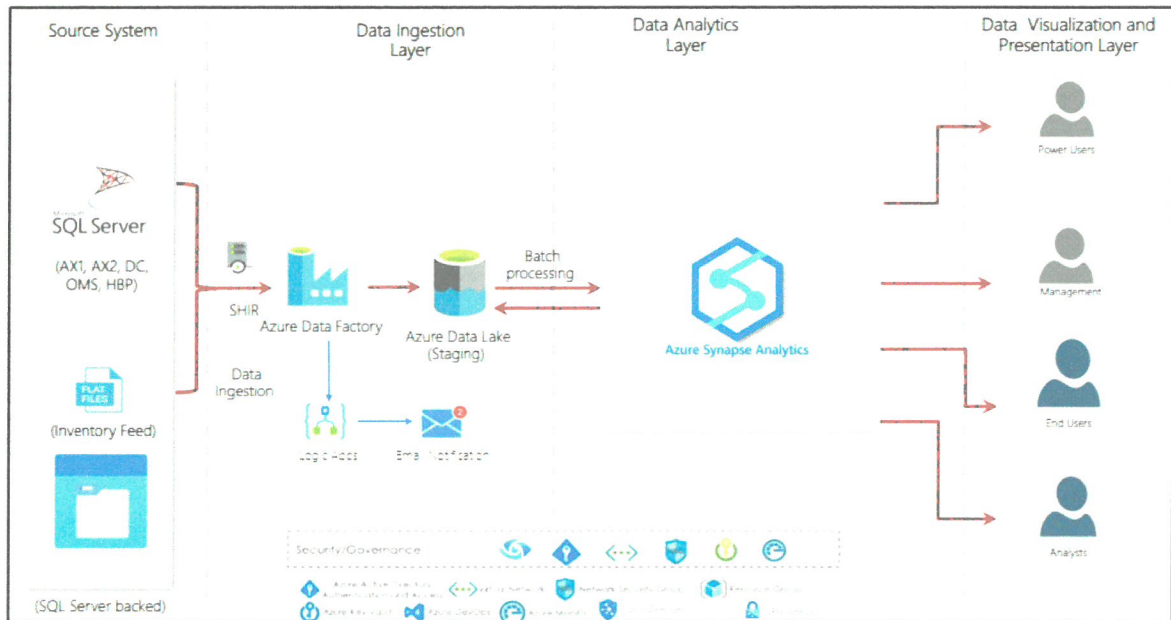
OBJECTIVE 1: Setting up the Azure environment to connect with and move data from multiple data sources like API's, SQL databases etc for multiple tenants.

OBJECTIVE 2: Setting up the Data Ingestion layer for moving data from the respective data sources to the the staging area.

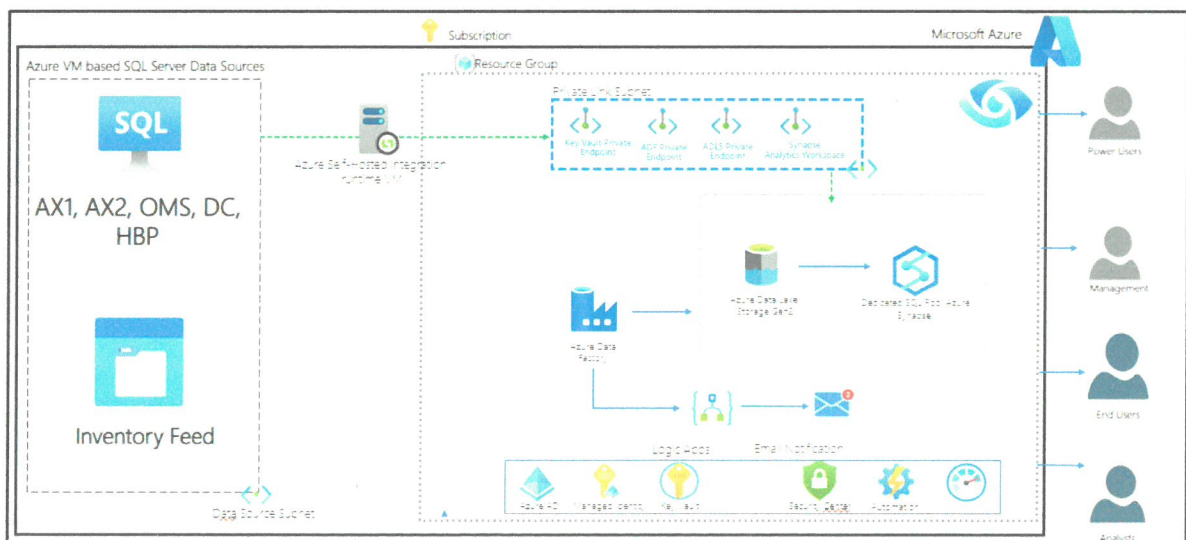
OBJECTIVE 3: Setting up the Data Analytics layer for doing all the analytics work and pass that data to the Data Visualization layer to find the meaningful insights.

5.3 Solution

1. SOLUTION ARCHITECTURE



2. NETWORK ARCHITECTURE



3. TECHNICAL SOLUTION

- Connectivity of source systems will be done, and respective data will be pulled into ADLS Gen 2 & Azure Synapse leveraging Azure Data Factory
- For on-prem Data sources connectivity, Self-Hosted Integration Runtime will be setup in an Azure VM. This SHIR VM will act as a connectivity bridge between the on-prem SQL servers and Azure Data Factory.
- The structured & unstructured data will be pulled into ADLS Gen 2 via Azure Data Factory & will further be loaded into Azure Synapse SQL DWH (SQL Dedicated Pool).
- Full Load and incremental data pipelines will be built in Azure Data Factory for both transactional and master data tables for each data source.
- ROW level Security will be implemented in Azure Synapse.
- For Inventory Feed Data Source, CT Team would require walkthrough of the Source, to ascertain the connectivity setup and establishing mechanism for bulk load and incremental load data pipelines.
- Setting up Azure Purview for data governance and management over Azure Storage, Azure Data Factory and Azure Synapse.

REFERENCES

1. <https://celebaltech.com>
2. https://en.wikipedia.org/wiki/Microsoft_Azure
3. <https://docs.microsoft.com/en-us/azure/data-factory/>
4. <https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities?tabs=data-factory>
5. <https://docs.microsoft.com/en-us/azure/synapse-analytics/>
6. <https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/>
7. <https://docs.microsoft.com/en-us/azure/logic-apps/logic-apps-overview>
8. <https://docs.microsoft.com/en-us/azure/key-vault/general/>
9. <https://dev.mysql.com/doc/>
10. <https://docs.microsoft.com/en-us/azure/azure-resource-manager/>



✉ hr@celebaltech.com
enterprisesales@celebaltech.com

🌐 www.celebaltech.com

☎ 0141-4700659

📍 India: A-2 7th floor, UDB Corporate Tower, J.L.N. Marg, Jaipur

STRICTLY CONFIDENTIAL

Dated: 26th May, 2022

Ongoing Internship Certificate

To Whomsoever It May Concern

This is to certify that **Ankur Soni** had undertaken Internship with Celebal Technologies, Jaipur, starting from 17th January 2022 in **Data Integration** Department.

During his 4 months of Internship, he was working under the mentorship of **Mr. Mayank Kumar Jhamb**.

Best Regards

Tejal

Tejal Mangal
Sr. Executive
Human Resources
Celebal Technologies