

PS II - REPORT

ETL using Azure Data Factory, Databricks and SSMS

**PREPARED BY:
Naveen Pratap Singh (2018BTECHCSE009)**

**Faculty Guide:
Dr Suman Shah**



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

May 2022

PS II - REPORT

ETL using Azure Data Factory, Databricks and SSMS

Practice School-2

Submitted in partial fulfilment of the requirements

for the degree of

Bachelor of Technology in Computer Science Engineering

By:

Naveen Pratap Singh (2018BTECHCSE009)

Faculty Guide:

Dr Suman Shah



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

May 20

CERTIFICATE

This is to certify that the Practice School-2 project work entitled "ETL using Azure data factory, Databricks and SSMS" submitted by Naveen Pratap Singh (2018BTECHCSE009) towards the partial fulfilment 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 Shah

Associate Professor

Department of Computer Science
Engineering

Institute of Engineering & Technology
(IET)

JK Lakshmipat University Jaipur



Dr. Gireesh Kaushik

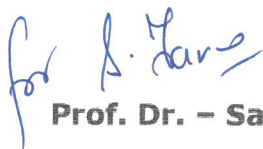
Co-ordinator PS2 and Associate
Professor,

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 start by thanking my honourable faculty Dr Suman Shah who has provided me with the necessary guidance and information needed to complete this internship report. I would like to thank PS II coordinator Dr Gireesh Kaushik for all his support, necessary tips and guidelines during the activation period.

Also, I would like to thank my reporting manager Mr. Anil Sharma Associate Big Data Engineer (Celebal Technologies) for teaching me all the basics of my tasks and patiently showing me every elaborate detail and guiding me throughout my internship period. I am very much thankful of Celebal Technologies, its management, especially the team of Big Data Department for their cordial support in my learning and completing this internship.

Lastly, I would like to present my gratitude to my friends who joined the Celebal technologies together with me and helped me through any difficulty I faced with all their support.

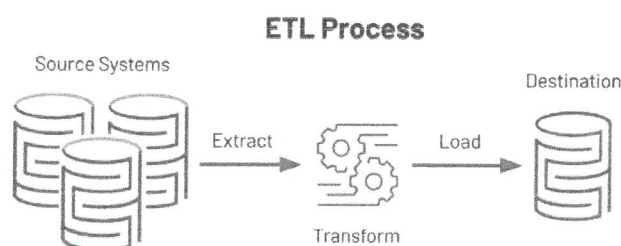
Naveen

Sincerely yours,

Naveen Pratap Singh (2018BTECHCSE009)

ABSTRACT

As the amount of data, data sources, and data types at organizations grow, the importance of making use of that data in analytics, data science and machine learning initiatives to derive business insights grows as well. The need to prioritize these initiatives puts increasing pressure on the data engineering teams because processing the raw, messy data into clean, fresh, reliable data is a critical step before these initiatives can be pursued. ETL, which stands for extract, transform, and load, is the process data engineers use to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems.



How Does ETL Work?

Extract

The first step of this process is extracting data from the target sources that are usually heterogeneous such as business systems, APIs, sensor data, marketing tools, and transaction databases, and others. As you can see, some of these data types are likely to be the structured outputs of widely used systems, while others are semi-structured JSON server logs.

Transform

The second step consists of transforming the raw data that has been extracted from the sources into a format that can be used by different applications. In this stage, data gets cleansed, mapped and transformed, often to a specific schema, so it meets operational needs. This process entails several types of transformation that ensure the quality and integrity of data. Data is not usually loaded directly into the target data source, but instead it is common to have it uploaded into a staging database. This step ensures a quick roll back in case something does not go as planned. During this stage, you have the possibility to generate audit reports for regulatory compliance or diagnose and repair any data issues.

Load

Finally, the load function is the process of writing converted data from a staging area to a target database, which may or may not have previously existed. Depending on the requirements of the application, this process may be either quite simple or intricate. Each of these steps can be done with ETL tools or custom code.

What is an ETL pipeline?

An ETL pipeline (or data pipeline) is the mechanism by which ETL processes occur. Data pipelines are a set of tools and activities for moving data from one system with its method of data storage and processing to another system in which it can be stored and managed differently. Moreover, pipelines allow for automatically getting information from many disparate sources, then transforming and consolidating it in one high-performing data storage.

LIST OF FIGURES

Figures	Page Numbers
Fig 1 SQL task sheet	14
Fig 2 ADF task sheet	14
Fig 3 Azure data factory	15
Fig 4 ADF 101 sheet list	15
Fig 5 ADF 101 task lists	16
Fig 6 Databricks learning platform	16
Fig 7 Databricks	17
Fig 8 Databricks	17
Fig 9 Azure services	20
Fig 10 Azure storage account	20
Fig 11 Azure data factory creation	21
Fig 12 Azure Data factory overview	21
Fig 13 Linked service creation: azure SQL database to ADF	22
Fig 14 Linked service creation: azure blob storage to ADF	22
Fig 15 Copy activity demonstration source	23
Fig 16 Copy activity sink operation	23
Fig 17 Databricks	24
Fig18 Databricks	24
Fig 19 azure storage container	25
Fig 20 SSMS	26
Fig 21 On premises Self Hosted linked service	27
Fig 22 Self hosted linked service creation	27

Fig 23 Azure storage container	28
Fig 24 SSMS	29
Fig 25 Databricks platform	30
Fig 26 Microsoft azure	30

LIST OF FLOWCHARTS

Names	Page Number
• FLOW CHART 1	13
• FLOW CHART 2	18
• FLOW CHART 3	26

LIST OF ACRONYMS

- ETL: EXTRACT, TRANSFORM, LOAD
- ADF: AZURE DATA FACTORY
- SQL: STRUCTURED QUERY LANGUAGE
- SSMS: SQL SERVER MANAGEMENT STUDIO
- BLOB: AZURE BLOB STORAGE
- DB: DATABRICKS

CONTENT

Title	Page No
CERTIFICATE	2
ACKNOWLEDEMENT	3
ABSTRACT	4
LIST OF FIGURES	5
LIST OF FLOWCHARTS	7
LIST OF ACRONYMS	8
1. INTRODUCTION	11
1.1 PIDILITE 11	
1.1.1 METHODOLOGY	11
1.2 TRANSORG SVPOC 11	
1.2.1 METHODOLOGY	11
2. JOURNEY IN CELEBAL	13
2.1 WEEK 1	13
2.2 WEEK 2	13
2.3 WEEK 3	14
2.4 WEEK 4	14
2.5 WEEK 5	14
2.6 WEEK 6	15
2.7 WEEK 7	15
2.8 WEEK 8	16
2.9 WEEK 9	16
2.10 WEEK 10	17
2.11 WEEK 11	17
3. PROJECT INVOIVED	18
3.1 PIDILITE	18

3.1.1 WORKFLOW	19
3.1.2 PROCEDURE USED	20
3.2 TRANSORG SVPOC	26
3.2.1 WORKFLOW	26
3.2.2 PROCEDURE USED	26
4. TECHNOLOGY USED	29
4.1 SQL SERVER MANAGEMENT STUDIO	29
4.2 DATBRICKS	29
4.3 AZURE	30
5. FUTURE WORK	31
6. REFERENCES	31

CHAPTER 1 : INTRODUCTION

Celebal Technologies is a premier software services company in the field of Data Science, Big Data, and Enterprise Cloud. It helps achieve a competitive advantage with intelligent data solutions, built using cutting-edge technology. The solutions provided are powered by Robotics, Artificial Intelligence, and Machine Learning algorithms which offer improved business efficiency in this interconnected world. Core offerings are around Cloud Innovation, Supply Chain Analytics, Chat Bots, Power Platforms, and Data Analytics. Solutions provided can help accelerate decision making and take giants leaps in the digital transformation journeys.

The mission is to make data simple and easy to understand for all organizations. Celebal is committed to providing solutions powered by modern cloud and artificial intelligence that integrate with traditional enterprise software. The tailor-made solutions help enterprises maximize productivity, improve speed and accuracy.

Help modernize legacy and on-premises applications by leveraging modern cloud paradigms such as Micro services, Advanced Analytics, API Management, and other Cloud-native services.



1.1 PIDILITE PROJECT:

A brief introduction:

Pidilite project has many aspects. The client data comes from REST API's and applications like Samsung Knox. The goal of the project is to Extract the given data and perform some transformations and make the presentable enough for business intelligence.

1.1.1 Methodology Adopted:

1. Usage of Azure Data Factory to Fetch the data from on Azure SQL Server Database.
2. Fetching Data from Azure data factory to Databricks.
3. Performing the necessary transformations on the data
4. Uploading the data back to the Azure data factory blob storage for the usage of the power BI team.

1.2 TRANSORG SVPOC PROJECT:

A brief introduction:

This project deals with performing the meaningful operations on the transaction data given by the client to Celebal technologies.

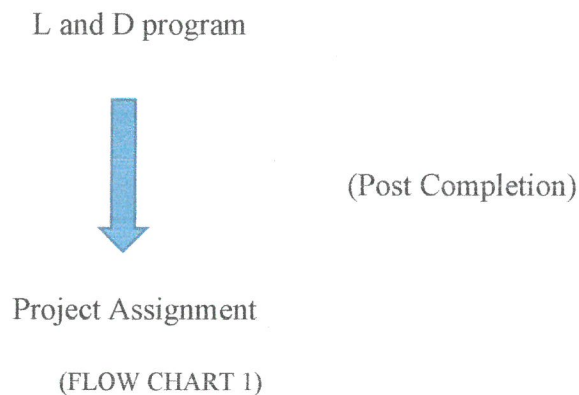
1.1.1 Methodology Adopted:

1. Fetching the data from SQL server using the SQL credentials
2. Performing necessary operations.
3. Passing the data to other teams for business intelligence operations.

CHAPTER 2 : JOURNEY IN CELEBAL

Celebal Technologies not only aims to achieve the highest efficiency in the different sections on which it's working but also the growth of employees as an individual. All the newly hired interns are to undergo a training process that involves them solving various sheets and questions related to their field. In the case of me the sheets were related to the big data field.

Workflow Of the interns:



Journey in the Celebal technologies can be summarized below:

2.1 • WEEK 1 (17th Jan – 22nd January)

Week 1 had the orientation program which focused on the introduction of the interns to the company policies and officials.

Introduction and sessions on the usage of Zoho took place. Zoho is an employee portal which records the attendance and work progress of the employee every day if they are working from remote or if even in the company.

After everything was setup, I was given our reporting managers/industry experts based on the domain that was assigned to me.

2.2 • WEEK 2 (24th Jan – 29th January)

Week 2 was all about completing the SQL task sheet that was given to me. It included two sections one of which was the reading material for the concepts and other was the practice questions related to it that were either from the company's portal or hacker rank.

Every day the reporting manager/industry expert would take a note of what I via email and initiated helping me out in terms of any queries.

The sheet was called SQL 100

B		C	D	E
1	Sub Topics	Study Links		Duration
23	Advanced joins in sql server	https://www.youtube.com/watch?v=Ga9C4BtD9E8	7th Day	
24	Union and Union All in SQL Server	https://www.youtube.com/watch?v=3uF8P7Q0T6U&list=PL9695876C417268A8&index=7		
25	Union and Union All in SQL Server	https://www.youtube.com/watch?v=3uF8P7Q0T6U&list=PL9695876C417268A8&index=7		
26	SQL Practice Questions			
27	Select All Task Hackerrank	https://www.hackerrank.com/challenges/select-all-task		
28	Select by ID Task Hackerrank	https://www.hackerrank.com/challenges/select-by-id		
29	X	https://www.hackerrank.com/challenges/select-by-id		
30	Employee Names	https://www.hackerrank.com/challenges/employee-names		
31	Japanese Cities	https://www.hackerrank.com/challenges/japanese-cities		
32	Employee Salary	https://www.hackerrank.com/challenges/employee-salary		
33	Japanese Cities	https://www.hackerrank.com/challenges/japanese-cities		
34	Weather Observation Station 1	https://www.hackerrank.com/challenges/weather-observation-station-1		
35	Weather Observation Station 3	https://www.hackerrank.com/challenges/weather-observation-station-3		
36	Weather Observation Station 4	https://www.hackerrank.com/challenges/weather-observation-station-4		
37	Weather Observation Station 5	https://www.hackerrank.com/challenges/weather-observation-station-5		
38	Weather Observation Station 5	https://www.hackerrank.com/challenges/weather-observation-station-5		
CT SQL100				
Level 2 Questions		Sheet 1		

Fig 1 SQL task sheet

2.3 • WEEK 3 (31st Jan – 5th Feb)

Week 3 basically revolved around me completing the advance SQL sheet which involved many good topics like sub queries, stored procedures, CTE, dynamic SQL correlated sub queries etc.

And I was assigned a Bata bricks boot camp course related to Data bricks simultaneously to pursue to build up my data bricks skills.

2.4 • WEEK 4 (7th Feb – 12th Feb)

Week 4 mainly consisted ADF 100 sheet. I setup my Microsoft azure account and was given a brief intro about them via my reporting manager. Further I was given more sheets like ADF100, ADF101

Difference between Normal container and Datalake container	https://www.linkedin.com/pulse/azure-data-engineering-series-part-1-blob-storage-vs-lake-grandy/?trk=public_profile_article
Different types of Performance and access Tiers	https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tier#text=Azure%20storage%20offers%20differ
Replication models (LRS, ZRS, RA-GRS)	https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy
ACL in Storage account	https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control
Azure Active Directory (AD, AAD) Tutorial Identity and Access Management Service	https://www.youtube.com/watch?v=Ma7VAQE7ga4
Introduction To Azure Data Factory	https://www.youtube.com/watch?v=Mc9Ira8WZU&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=1
Top Level Concepts in Azure Data Factory	https://www.youtube.com/watch?v=Pd1QZ0Y9eg&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=2
Deploy Azure Data factory	https://www.youtube.com/watch?v=x0yeMHT2ak&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=3
Pipelines and Activities in Azure Data Factory	https://www.youtube.com/watch?v=V_7bN4MqzA&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=5
Linked Services and Datasets in Azure Data Factory	https://www.youtube.com/watch?v=Ud_7wL24M&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=6
Triggers in azure data factory	https://www.youtube.com/watch?v=UfdR0d_ghaU&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=7
Schedule Trigger in Azure Data Factory	https://www.youtube.com/watch?v=xqpmDfG8Vp&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=8
Tumbling Window Trigger	https://www.youtube.com/watch?v=vrug-C_NDUI&list=PLMwWz7eqEaTJffbb8zVOv9C0tal1FO&index=9

Fig 2 ADF task sheet

2.5 • WEEK 5 (14th Feb – 19th Feb)

In this week I did the tasks related to the ADF 100 Sheet using Microsoft Azure.

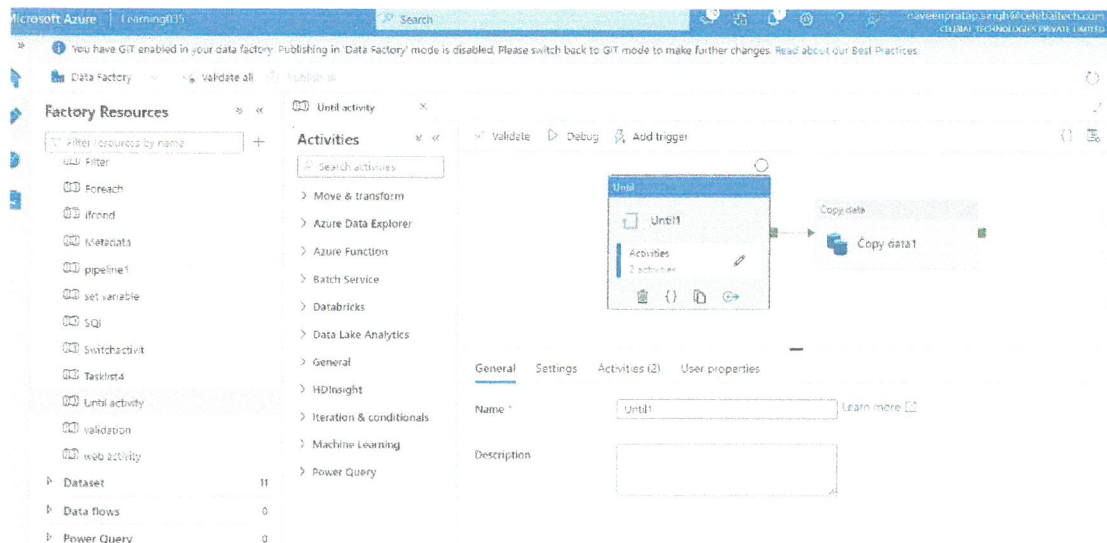


Fig 3 Azure data factory

2.6 • WEEK 6 (21th Feb – 26th Feb)

Week 6 included me doing the ADF 101 sheet list.

Sl. No.	Task Topic	Video Link
1	Integration Runtime	https://www.youtube.com/watch?v=8rJ0mVAswfc&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=12
2	Azure Integration Runtime	https://www.youtube.com/watch?v=MQ84n4Al_p4&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=13
3	Self Hosted Integration Runtime	https://www.youtube.com/watch?v=ilGex7ZHABl&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=14
4	Setting up Self hosted Integration Runtime	https://www.youtube.com/watch?v=RI1b-pRO4LE&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=15
5	Shared Self hosted Integration Runtime	https://www.youtube.com/watch?v=9BvU_NpntSg&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=16
6	Parameterize Linked Service	https://www.youtube.com/watch?v=M22Mj0rcBcs&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=17
7	Transformation Activity	https://www.youtube.com/watch?v=wsOYvT4Dpw&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=43
8	Stored Procedure	https://www.youtube.com/watch?v=4Npu4F6dgMo&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=44
9	Different Author Modes	https://www.youtube.com/watch?v=4E98C4Pdlp8&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=72
10	Setup Azure DevOps Git Repository	https://www.youtube.com/watch?v=vPUnURzXJSQ&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=74
11	Use Azure key Vault to store the secret	https://www.youtube.com/watch?v=yKTgc33G1KU&list=PLMWaZteqtEaLTjffbbBzVOv9C0otal1FO&index=84

Fig 4 ADF 101 sheet list

2.7 • WEEK 7 (28th Feb – 5th March)

Week 7 included the completion of the ADF 101 tasks and related queries given by us to our reporting managers.

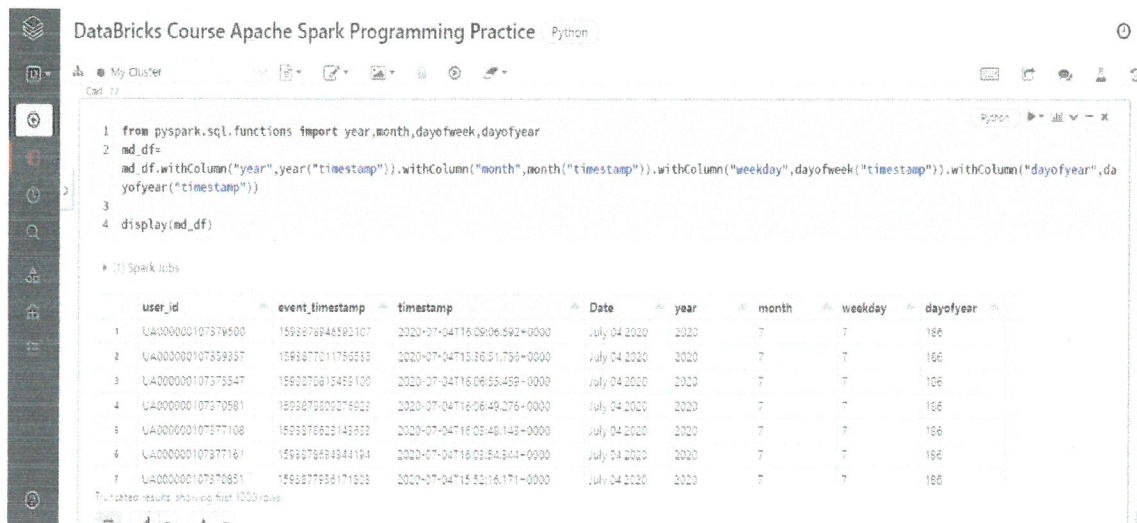


Fig 7 Databricks

2.10 • WEEK 10 (21th March – 26th March)

Data Bricks Capstone project on the partner's academy.

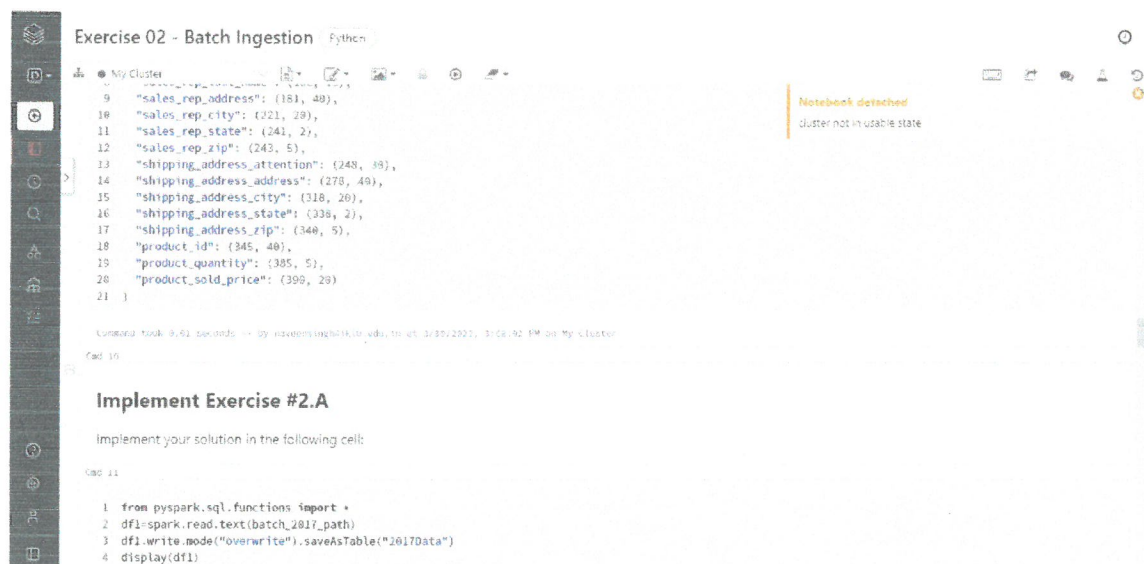


Fig 8 Databricks

2.11 • WEEK 11 (28th March – 2nd April)

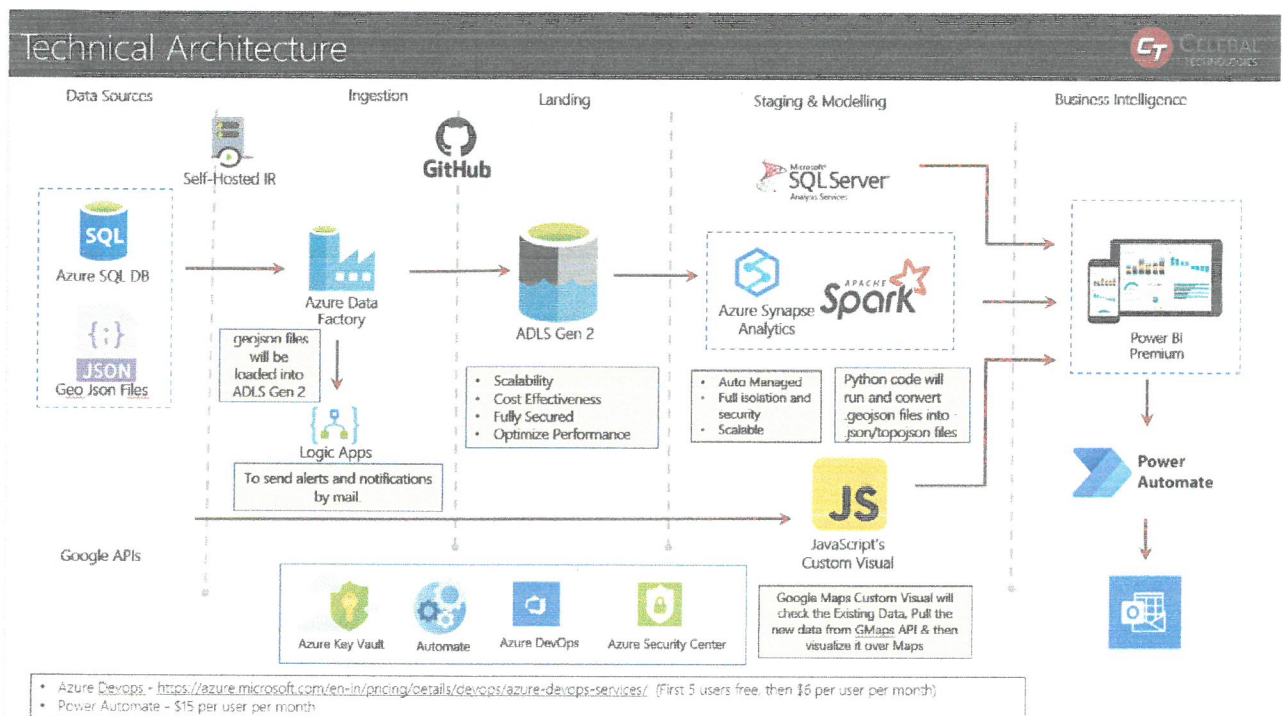
Revision of the skills acquired so far by completing tasks given by reporting manager.

CHAPTER 3 : PROJECTS INVOLVED

3.1 PIDILITE:

It is a project which involves fetching the data from rest API or the Samsung Knox application or the azure SQL database of the client. The workflow was divided into many teams concerning the ETL and analysis of data. My work has been listed below.

Architecture of the project PIDILITE



FLOW CHART 1: PIDILITE PROJECT 2022

3.1.1 FLOW OF THE PROJECT:

1. The data comes from the Azure SQL database



Azure SQL DB

2. Data is being fetched into the Azure data factory. And stored in the ADLS gen 2.



Data Factory

3. The data is further sent to Databricks for transformations and modelling.

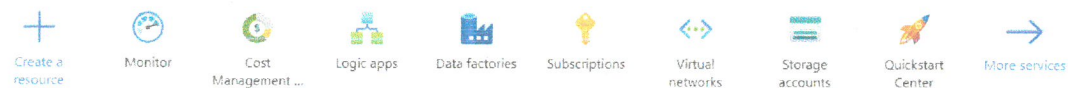


4. Data is sent to the other teams for further business analysis.

3.1.2 PROCEDURE USED:

- i) Creation of the azure blob storage which will be our database for accessing the client data over cloud.

Azure services



Resources

Recent Favorite

Name	Type	Last Viewed
Azure for Students	Subscription	3 days ago
learning3102	Storage account	3 days ago
Learningdf31	Data factory (V2)	3 days ago
email310000	Logic App (Standard)	6 days ago
Naveen30rg	Resource group	a week ago

[See all](#)

Navigate

Fig 9 Azure services

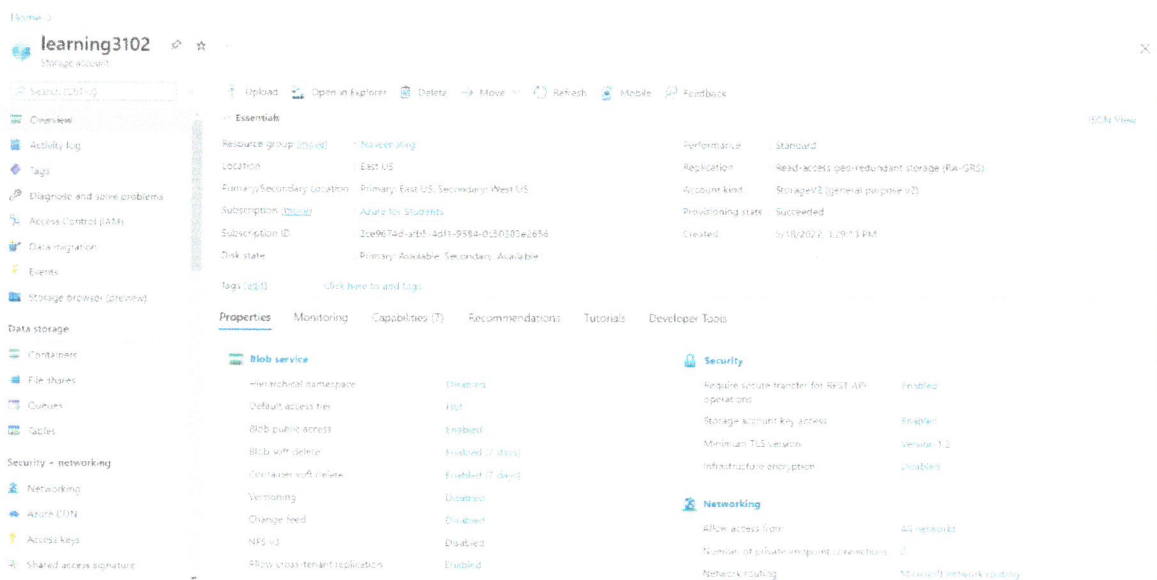


Fig 10 Azure storage account

- ii) Creation of the azure data factory for the ETL of data. Basically, we are using ADF as a movement tool for now.

The screenshot shows the 'Create Data Factory' wizard in the Azure portal. The 'Basics' tab is selected, showing fields for 'Subscription' (Azure for Students), 'Resource group' (Create new), 'Name' (empty), 'Region' (East US), and 'Version' (V1 (Recommended)). At the bottom, there are buttons for 'Review + create', '< Previous', and 'Next: Get configuration >'. The taskbar at the bottom shows the time as 10:07 PM.

Fig 11 Azure data factory creation

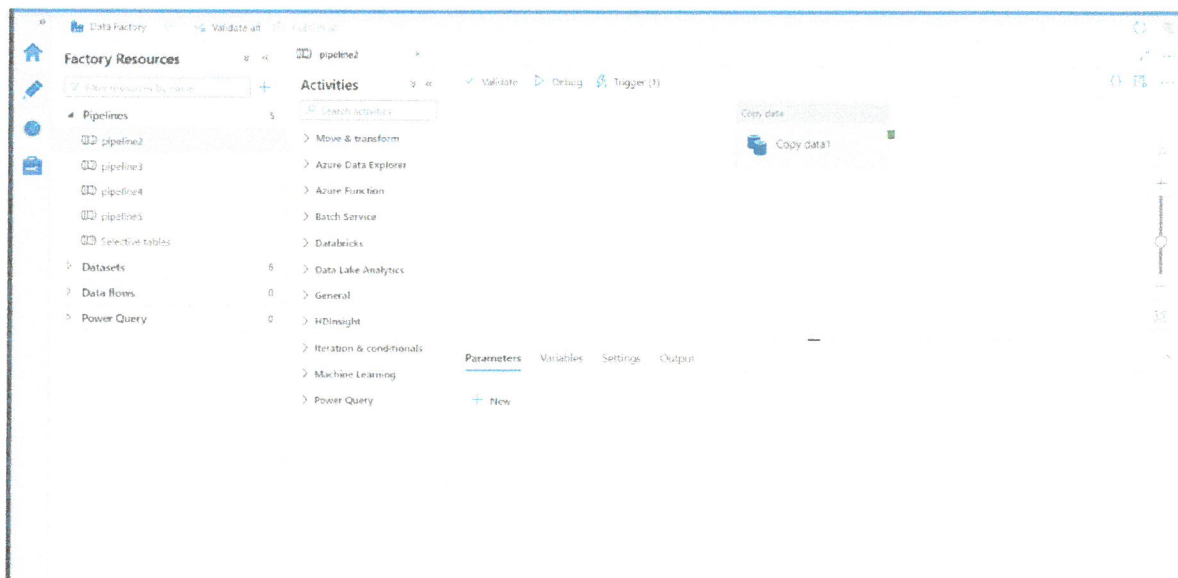


Fig 12 Azure Data factory overview

- iii) The formation of the linked services which connect the ADF to azure SQL database. Linked services basically point to the databases where data is stored when we create a dataset in the ADF.

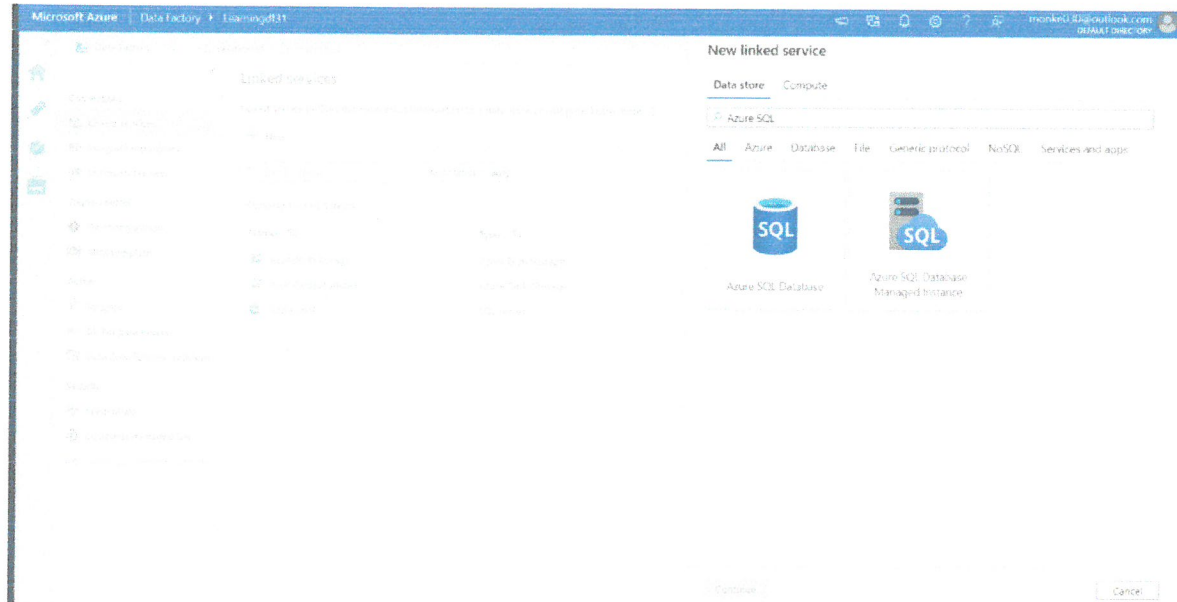


Fig 13 Linked service creation: azure SQL database to ADF

- iv) The formation of the linked services which connect the ADF to Azure blob storage. Azure blob storage is the default cloud storage of the ADF. It gives a default integration runtime namely Auto Resolve Integration Runtime.

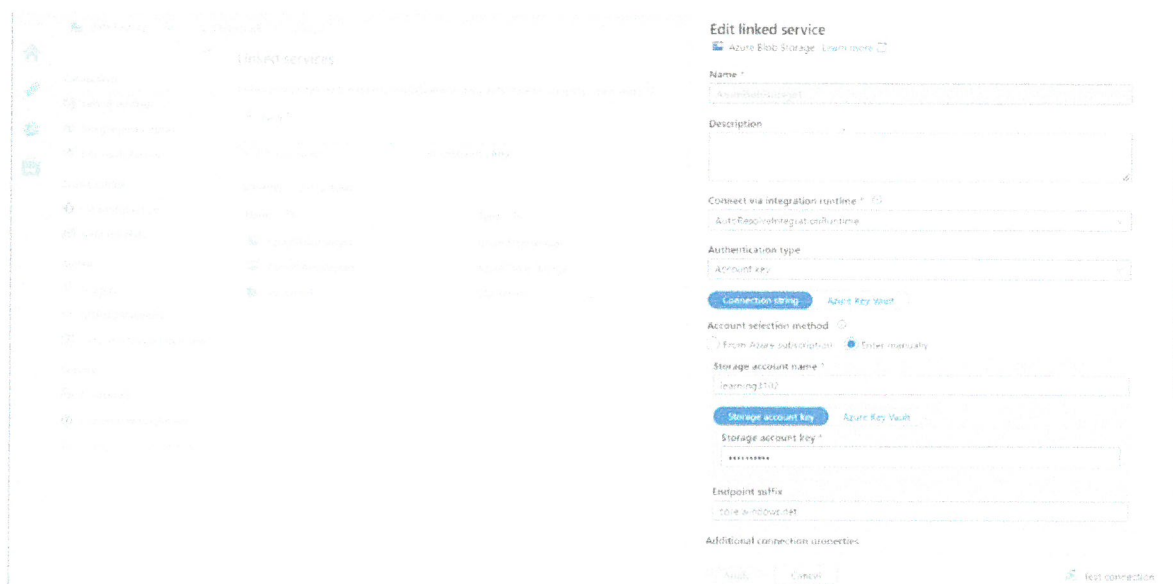


Fig 14 Linked service creation: azure blob storage to ADF

- v) Getting the data into blob storage via copy activity. Copy activity is a strong activity in ADF. It lets us move a huge chunk of data efficiently and super-fast.

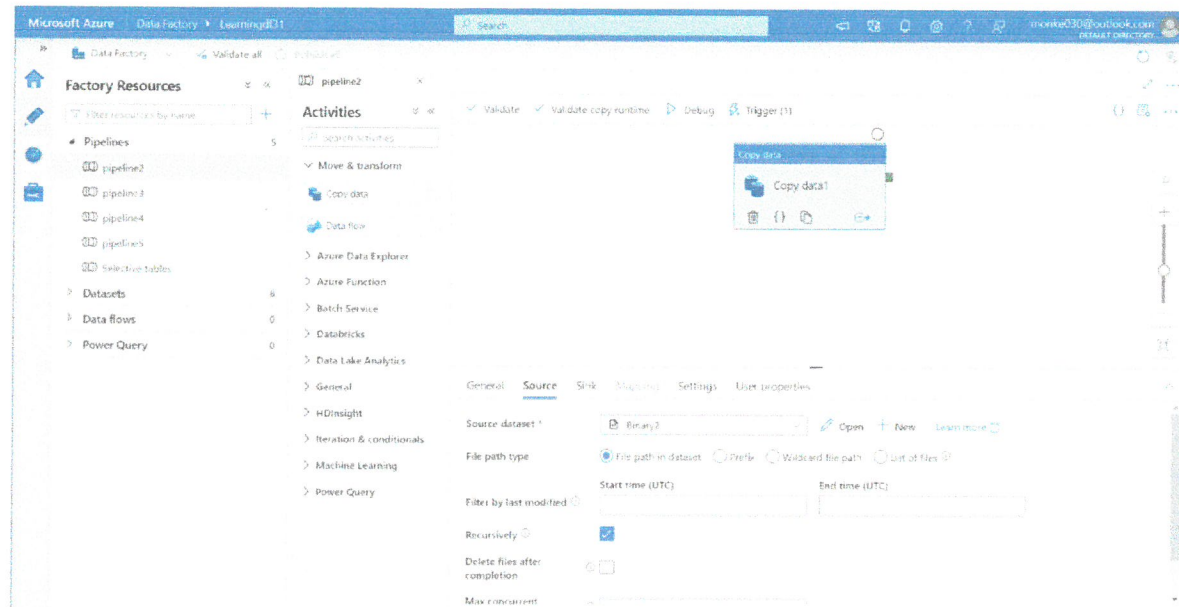


Fig 15 Copy activity demonstration source

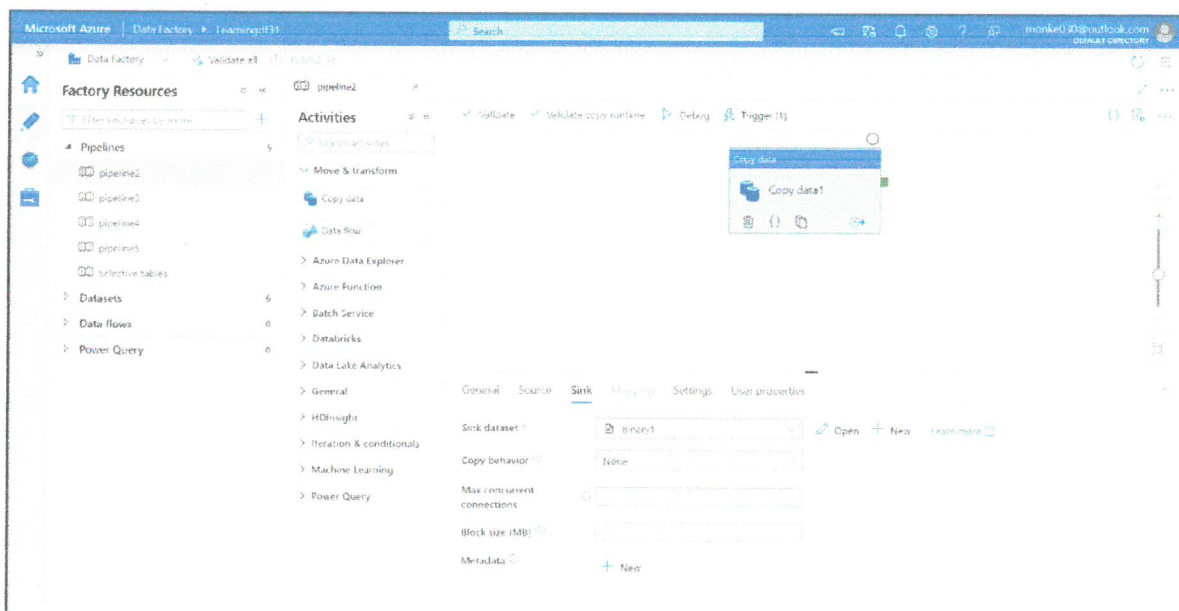


Fig 16 Copy activity sink operation

vi) Performing the necessary modelling and transformation in data bricks.

Create Fact and Dim Tables Python

```

1 new_df=spark.sql("""select
  sales_rep_id,sales_rep_first_name,sales_rep_last_name,sales_rep_address,sales_rep_city,sales_rep_state,sales_rep_zip,ingest_file_name,sales_rep_vsn,ingested_at from
  (batch_rep_view)""")
2 new_df=new_df.withColumn("error_ssn_format",col("sales_rep_vsn").like("%s-%s"))
3 new_df=new_df.withColumn("sales_rep_vsn",new_df["sales_rep_vsn"].cast("long").withColumn("sales_rep_zip",new_df["sales_rep_zip"].cast("int"))
4 new_df=new_df.dropDuplicates(["sales_rep_first_name"])
5 display(new_df)

```

sales rep id	sales rep first name	sales rep last name	sales rep address	sales rep city	sales rep state	sales rep zip	ingest file no
1d7493e-1a72-4d42-bd0f-40e85d3d326	Aaron	Bull	PO Box 927	Danier	CO	80127	dchf0bascade
d1d2701a-8005-4000-9250-620305542000	Armed	West	2775 Curve Court	Greenwater	FL	32609	dchf0bascade
45910d45-1d59-4d06-8210-4d8539d95e0	Alex	Cypeland	566 Motor Court S	Escondido	CA	94022	dchf0bascade
9e975d4d-425e-4730-9936-55020704200	Alexa	Rios	778 Balconia Place S	Seale	FL	32713	dchf0bascade
70c41025-4f21-4570-0050-5008002d8465	Alyson	Kennedy	PO Box 591	Durham	NC	27712	dchf0bascade
915d8b51-0f44-478a-525e-6d8b6f43847	Andrea	Kenny	PO Box 241	Denton	TX	77034	dchf0bascade
40717f14-67a0-429d-4504-07561b0d119f	Andrew	Rosen	3701 North 4 Avenue	Starkport	LA	70154	dchf0bascade

```

1 new_df.write.format("delta").mode("overwrite").option("overwriteSchema",True).saveAsTable(f"{user_db}.(sales_reps_table)")
2

```

Fig 17 Databricks

Fact and Dim Tables Python

```

1 order_df=order_df.withColumn("submitted_at",col("submitted_at").cast("timestamp")).withColumn("shipping_address_zip",col("shipping_address_zip").cast("integer"))
2 order_df=order_df.dropDuplicates(["sales_rep_id"])
3 order_df=order_df.withColumn("submitted_vyyy_mm",)
4 display(order_df)
5 order_df.printSchema()

```

```

root
 |-- submitted_at: timestamp (nullable = true)
 |-- order_id: string (nullable = true)
 |-- customer_id: string (nullable = true)
 |-- sales_rep_id: string (nullable = true)
 |-- shipping_address_attention: string (nullable = true)
 |-- shipping_address_address: string (nullable = true)
 |-- shipping_address_city: string (nullable = true)
 |-- shipping_address_state: string (nullable = true)
 |-- shipping_address_zip: integer (nullable = true)
 |-- ingest_file_name: string (nullable = true)
 |-- ingested_at: timestamp (nullable = true)

```

submitted at	order id	customer id	sales rep id	shipping address attention	shipping
1 null	c7691051-5731-40be-990f-913a779e79ed	144e154f-910a-4063-10a9-01850a0e0005	05227a10-445e-423e-d19d-0504043e0800	Ruby Weber	717 Colum
2 null	c7602131-5731-4796-b10d-41a9d40b4401	0a70e30e-111e-4593-9640-76da20890a0a	0714327-0000-9471-9e0b-071514701032	Stanley Key	577 Roman
3 null	c7601a01-6655-40d3-a610-97010c1a1001	91706812-b070-4612-b070-921445744004	05770e02-0a77-4004-9108-1920c1787832	Elizabeth Myers	804 E 21st
4 null	c7601a01-6655-40d3-a610-97010c1a1001	0122e80e-0d30-420e-b07a-91116020a709	002e1500-3e12-4010-b071-505207021807	Kenneth Anglin	422 W River
5 null	c7601a01-6655-40d3-a610-97010c1a1001	91a1774a-6074-101a-0700-00d12127e702	0a20e80a-1d11-4050-90a1-5e023a0e0770	Ernie Savage	663 S Main
6 null	c7601a01-6655-40d3-a610-97010c1a1001	0a0e705a-0b70-df11-0090-100a17100000	1b413100-0171-400a-4000-400401010401	Rene Sanders	7101 Mary
7 null	c7601a01-6655-40d3-a610-97010c1a1001	01114070-0014-4710-0000-000000000000	11210e70-0a10-4070-0070-000000000000	Robert Danner	555 E Main

Fig 18 Databricks

- vii) Saving the data and storing it back to the blob storage for further analysis by business intelligent teams.

Home > learning3102 > learning

learning Container

Search (Ctrl+F)

Upload Change access level Refresh Icons Share Blob Acquire Blob Set Blob Versioning Blob Access Analytics

Authentication method: Access key (Switch to Azure AD User Account)

Location: learning

Search blobs by prefix (non-sensitive) Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
friends						...
friends1M						...
learning21M						...
learning1M						...
out1						...
out2						...
out3						...
salary						...
salary1M						...
salary100						...
students1M						...
xq100.xlsx	5/23/2022 1:06:55 PM	Hot (Inferred)		Block Blob	49.09 KB	Available

Fig 19 azure storage container

3.2 PROJECT TRANSORG SVPOC:

3.2.1 Workflow:

SQL Server  SSMS (Modelling and transformation of data)

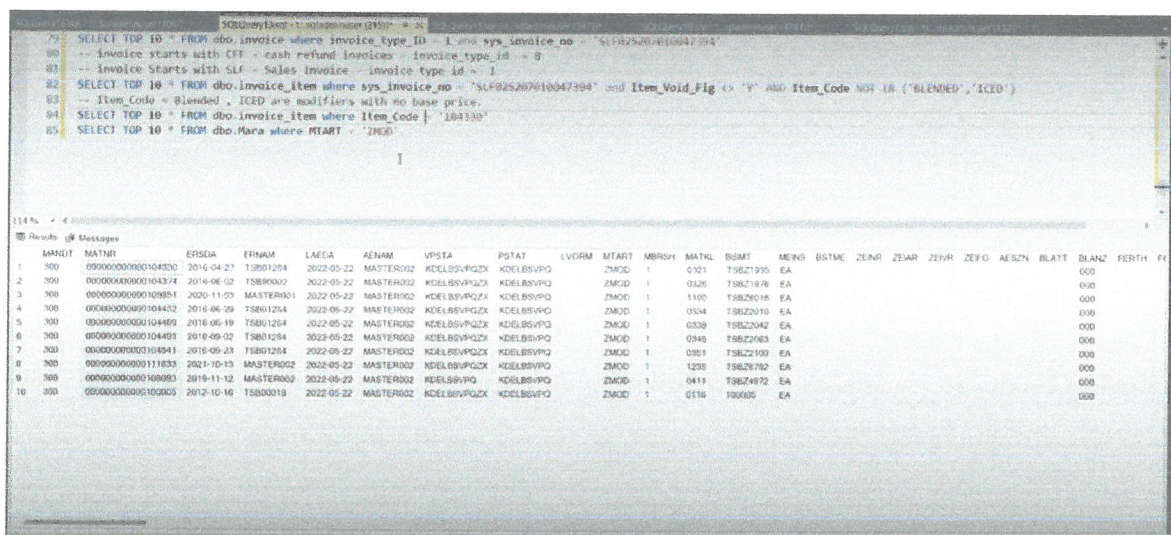


Azure Blob storage

FLOW CHART 3

3.2.2 Methodology:

1. Running Queries over the given dataset. And creating another unique table which will contain a detailed summary and transformed data of the client.



```
79 SELECT TOP 10 * FROM dbo.invoice where invoice_type_id = 1 and sys_invoice_no = '51182520/010047394'
80 -- invoice starts with CFI - cash refund invoices - invoice type id = 9
81 -- invoice starts with SLI - Sales Invoice - invoice type id = 1
82 SELECT TOP 10 * FROM dbo.invoice_item where sys_invoice_no = '51182520/010047394' and Item_Void_flg <> 'Y' AND Item_Code NOT IN ('BLENDED','ICED')
83 -- Item_Code = Blended , ICED are modifiers with no base price.
84 SELECT TOP 10 * FROM dbo.invoice_item where Item_Code = '104330'
85 SELECT TOP 10 * FROM dbo.Mara where MTART = 'ZMCO'
```

	AMNDT	MATNR	ERSDA	FINAM	LAEDA	AENAM	VPSTA	PSTAT	LVORM	MTART	MBRSH	MATKL	BSMT	MENA	BSTME	ZEARN	ZEAR	ZEAR	ZEPF	AESZ	BLATT	BLANT	PERTH	F0
1	300	0000000000000104330	2016-04-27	T5801204	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0121	T5827195	EA											
2	300	000000000000104374	2016-06-02	T5800002	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0326	T58271876	EA											
3	300	000000000000104851	2020-11-03	MASTER001	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	1100	T5826018	EA											
4	300	00000000000104432	2016-06-29	T5801204	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0304	T5822010	EA											
5	300	00000000000104489	2016-06-19	T5801204	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0338	T5822042	EA											
6	300	00000000000104493	2016-06-02	T5801204	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0301	T5822190	EA											
7	300	00000000000104541	2016-06-23	T5801204	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0340	T5822163	EA											
8	300	00000000000111833	2021-10-13	MASTER002	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	1203	T5826750	EA											
9	300	00000000000108093	2018-11-12	MASTER002	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0411	T5824972	EA											
10	300	00000000000109005	2012-10-16	T5800019	2022-05-22	MASTER002	KDELBSVQZK	KDELBSVQ	ZMCO	1	0116	T00005	EA											

Fig 20 SSMS

2. Moving data to azure Blob storage:

2.1 Creating a runtime integration service in the machine where the data is hosted. Commonly called on premise machine. Runtime integration basically helps move out your data to the cloud from a private or public workspace.

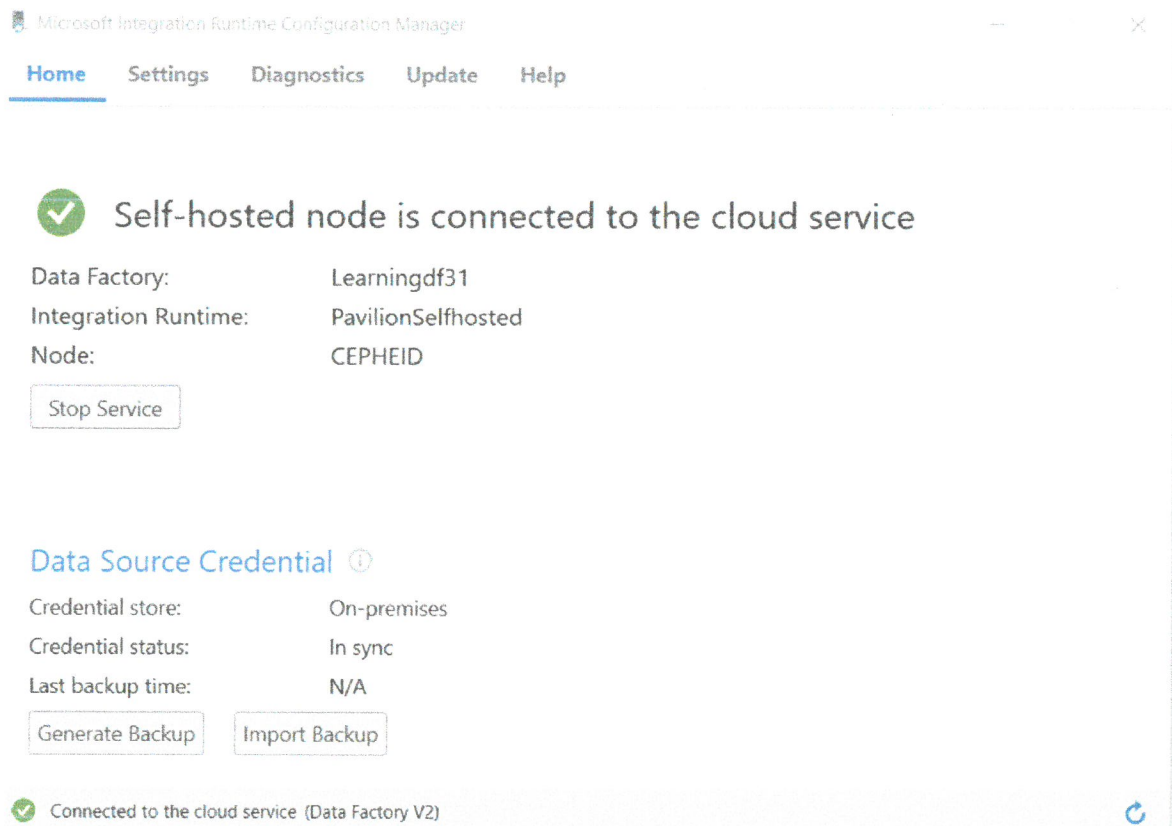


Fig 21 On premises Self Hosted linked service

2.2 Creating a linked service in azure so that we can access the integration runtime on pc.

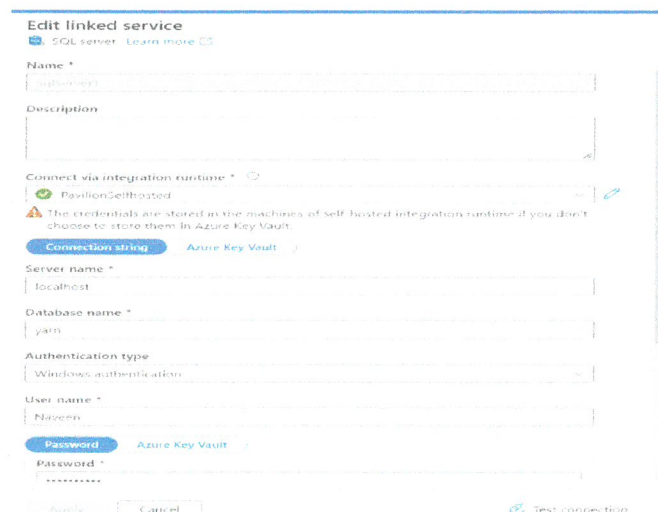


Fig 22 Self hosted linked service creation

2.2.2 Finally using copy data to move data into azure blob storage.

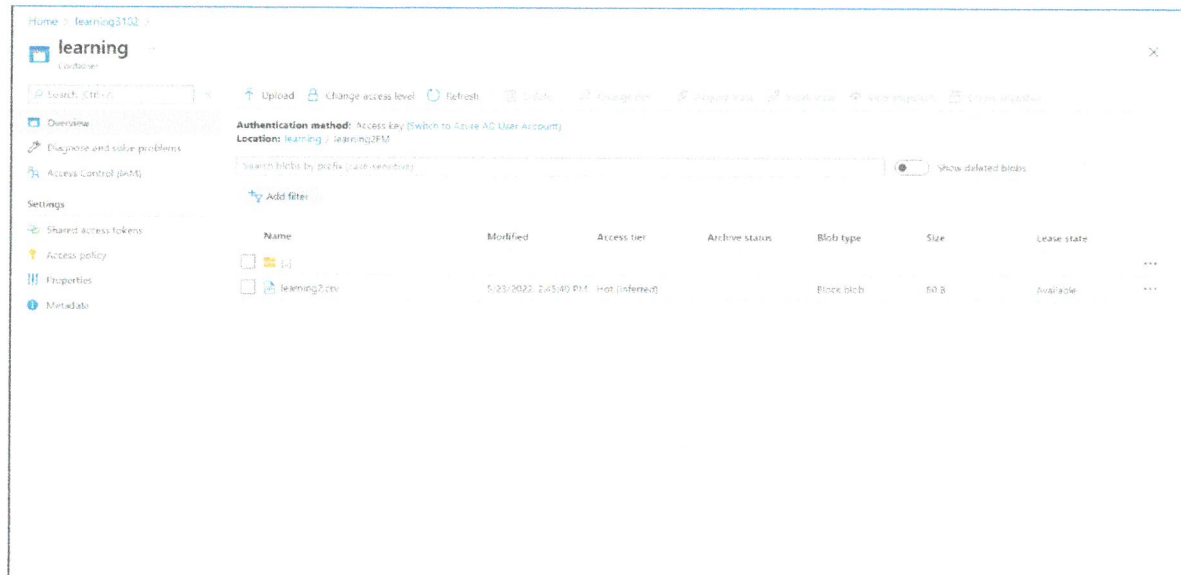


Fig 23 Azure storage container

CHAPTER 4 : TECHNOLOGIES USED

4.1 • SQL Server Management Studio

SQL Server Management Studio (SSMS) is an integrated environment for managing any SQL infrastructure, from SQL Server to Azure SQL Database. SSMS provides tools to configure, monitor, and administer instances of SQL Server and databases. Use SSMS to deploy, monitor, and upgrade the data-tier components used by your applications, and build queries and scripts.

Use SSMS to query, design, and manage your databases and data warehouses, wherever they are - on your local computer, or in the cloud.

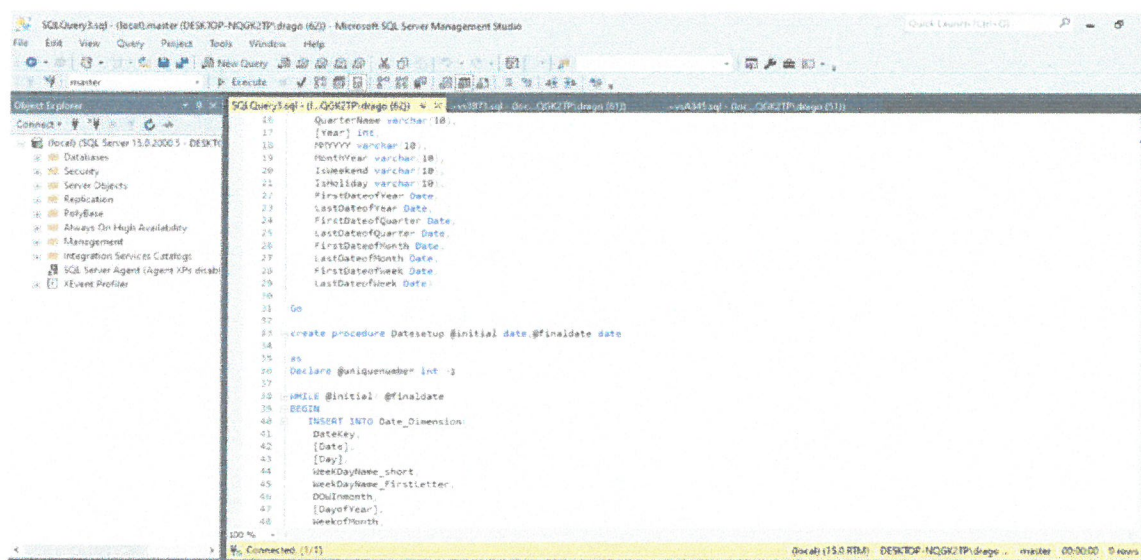


Fig 24 SSMS

4.2 • Data bricks

A data lake house unifies the best of data warehouses and data lakes in one simple platform to handle all your data, analytics, and AI uses cases. It's built on an open and reliable data foundation that efficiently handles all data types and applies one common security and governance approach across all of your data and cloud platforms.

• Data bricks Community Edition

The Data bricks Community Edition is the free version of **our cloud-based big data platform**. Its users can access a micro-cluster as well as a cluster manager and notebook environment. All users can share their notebooks and host them free of charge with Databricks. We hope this will enable everyone to create new and exciting content that will benefit the entire Apache Spark™ community.

The Databricks Community Edition also comes with a rich portfolio of award-winning training resources that will be expanded over time, making it ideal for developers, data scientists, data engineers and other IT professionals to learn Apache Spark.

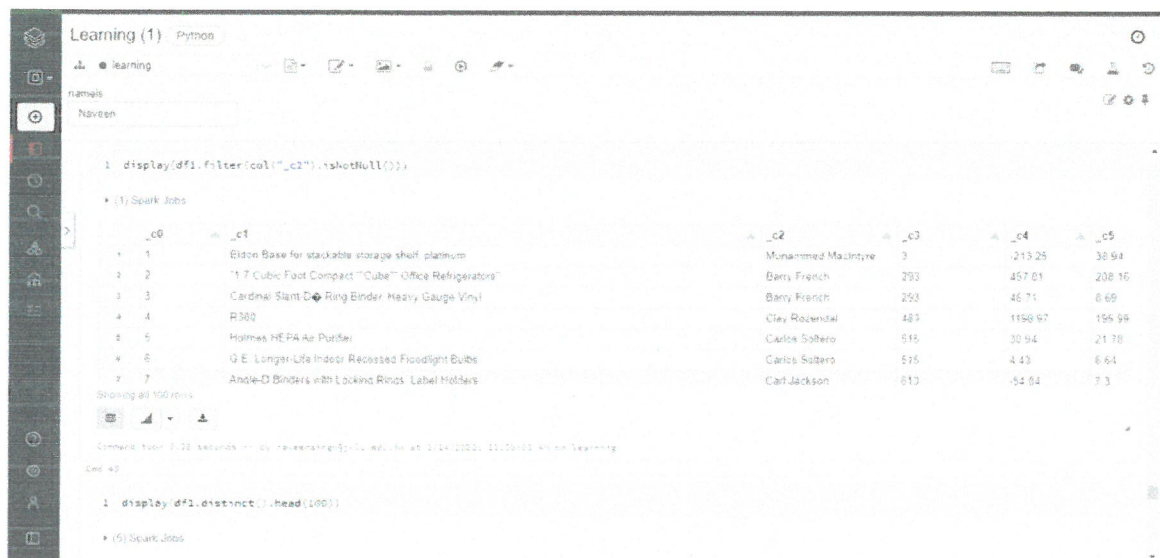


Fig 25 Databricks platform

4.3 • Microsoft Azure

The Azure cloud platform is more than 200 products and cloud services designed to help you bring new solutions to life—to solve today's challenges and create the future. Build, run and manage applications across multiple clouds, on-premises and at the edge, with the tools and frameworks of your choice.

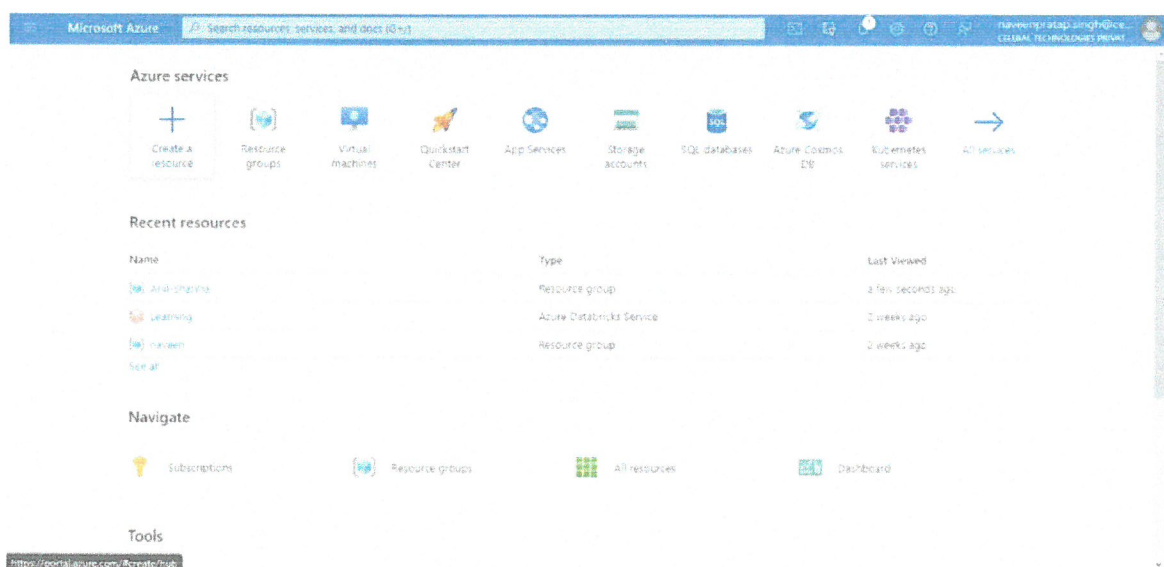


Fig 26 Microsoft azure

CHAPTER 5 : FUTURE WORK

Would be attaining the skills related to azure data factory and data bricks. Will focus on obtaining projects which challenge my current knowledge and contributes to my learning and development.

PIDILITE:

It is a project that requires a lot of updates from the client side and often many things are added in the The project every month. So will have the opportunity to enhance my skills on that.

REFERENCES

<https://celebaltech.com>

[Download SQL Server Management Studio \(SSMS\) - SQL Server Management Studio \(SSMS\) | Microsoft Docs](#)

[Databricks - The Data and AI Company](#)

[Databricks Community Edition FAQs](#)

Databricks Apache Spark Programming course

[What is ETL? - Databricks](#)

<https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/>

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities?tabs=data-factory>



✉ 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

CHAPTER 6 : To Whomsoever It May Concern

This is to certify that **Naveen Pratap Singh** had undertaken Internship with Celebal Technologies, Jaipur, starting from 17th January 2022 in **Big Data & EDW** Department.

During his 4 months of Internship, he was working under the mentorship of **Mr. Anil Rajkumar Sharma**.

Best Regards

Tejal

Tejal Mangal
Sr. Executive
Human Resources
Celebal Technologies