PROJECT REPORT BY:

COHORT: INTCDB22DW145

MEMBERS:

Rajesh Adam (2158777)

Pratibha Pratibha (2165699)

Vaishnavi Desale (2149310)

Amogh Raina (2160927)

Viraj kadlag (2165174)

MENTOR: Saha Mithun

TRAINER: Ravikanth Varigonda

COACH: Rao Kottakota Tejeswara

ACKNOWLEDGEMENT

We take this opportunity to express our sincere thanks and deep gratitude to all those people who extended their wholehearted cooperation and have helped in completing this project successfully. We would like to express my gratitude to my mentor, **Mr. Mithun Saha** for constantly motivating and sharing his experience in this field.

Special thanks to the entire IT team of **COGNIZANT** and our trainer for providing us with the knowledge and resources that have proved to be of immense help to complete this project.

We feel pride and privilege in expressing our deep sense of gratitude to all those who have helped me in presenting this assignment.

ABSTRACT

The mission of the Sales Data Warehouse project is to provide strategic and tactical support to all departments and divisions of a media company through the acquisition and analysis of data pertaining to their customers and markets. This project helps to identify areas of readership and marketing through creation of a Data Warehouse that will provide a company with a better understanding of its customers and markets.

Overall, the project has identified three basic needs:

- Acquiring and maintaining core data about the households, individuals, and businesses within the market area.
- Acquiring and maintaining transactional data on the households, individuals, and businesses within the market area.
- Acquiring and implementing the tools needed to effectively manipulate and access the core and transactional data.

This project provides a wide variety of benefits to a number of business units within a media company. These benefits are expected to help drive marketing and readership, as well as improve productivity and increase revenue.

TABLE OF CONTENTS

CHAPTER 1

ABOUT THE PROJECT

Chapter 1: ABOUT THE PROJECT

2.1. PURPOSE AND SCOPE OF THE PROJECT

Daily sales report is the core-tracking element for controlling cash, sales performance and forecasting the future sales. The purpose of the daily sales report is to be informed about the level of activity taking place within the sales organization and to make sure that each salesperson is meeting or exceeding the sales target.

The overriding business objective is to increase net paid circulation. Some of the secondary objectives are reducing churn and increasing retention, profiling best customers so we can market to their non-subscribing look-alikes, improving efficiency in sales channels, monitoring sales and marketing activities, tracking subscriber behavior across multiple subscriptions and households, targeting sales and retention efforts, and focusing sales on high impact channels, prospects and offers.

CHAPTER 2

LITERATURE REVIEW

CHAPTER 3: LITERATURE REVIEW

3.1 SQL SERVERS

SQL stands for Structured Query Language. It is used for storing and managing data in relational database management systems (RDMS). It is a standard language for Relational Database Systems. It enables a user to create, read, update and delete relational databases and tables.

SQL Server is a relational database management system, or RDBMS, developed and marketed by Microsoft.

Similar to other RDBMS software, SQL Server is built on top of SQL, a standard programming language for interacting with relational databases. SQL Server is tied to Transact-SQL, or T-SQL, Microsoft's implementation of SQL that adds a set of proprietary programming constructs.

Microsoft provides both data management and business intelligence (BI) tools and services together with SQL Server.

For data management, SQL Server includes SQL Server Integration Services (SSIS), SQL Server Data Quality Services, and SQL Server Master Data Services. To develop databases, SQL Server provides SQL Server Data tools; and to manage, deploy, and monitor databases SQL Server has SQL Server Management Studio (SSMS).

For data analysis, SQL Server offers SQL Server Analysis Services (SSAS). SQL Server Reporting Services (SSRS) provides reports and visualization of data. The Machine Learning Services technology appeared first in SQL Server 2016 which was renamed from the R Services.

3.2 Data Warehouse and ETL Concepts

3.2.1 Data Warehousing

A **data warehouse** is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence. Business analysts, data engineers, data scientists, and decision makers access the data through business intelligence (BI) tools, SQL clients, and other analytics applications

The Data Warehouse is a collection of data in support management decision processes, which is:

- Subject-Oriented Integrated
- Integrated
- Time-Variant and Non-Volatile

A **Data Warehouse** is separate from DBMS, it stores a huge amount of data, which is typically collected from multiple heterogeneous sources like files, DBMS, etc. The goal is to produce statistical results that may help in decision makings.It usually contains historical data derived from transaction data and other sources.

Need for Data Warehouse:

- An ordinary Database can store MBs to GBs of data and that too for a specific purpose. For storing data of TB size, the storage shifted to Data Warehouse.
- Besides this, a transactional database doesn't offer itself to analytics. To effectively perform analytics, an organization keeps a central Data Warehouse to closely study its business by organizing, understanding, and using its historic data for taking strategic decisions and analyzing trends.

Characteristics of Data Warehouse:

- Separate DSS database
- Only for Storage of data, no new data is created
- Integrated and Structured data Historical data and Read-only
- Various levels of summarization
- Subject-Oriented and Easily accessible

Features of Data Warehouse:

- Strategic enterprise level decision support
- Multi-Dimensional view on the enterprise data
- Caters to the entire spectrum of management
- Descriptive, Standard business terms
- High degree of Scalability and high analytical capability
- Historical data only.

Fact and Dimension Tables in a Data Warehouse:

- Fact Table or Reality Tables:
    - Fact table contains the measuring of the attributes of a dimension table.
    - In fact table, There is less attributes than dimension tables.
    - Fact table forms a vertical table.
    - It comes after the dimension table.
    - The attribute format of fact table is in numerical format and text format.
- Dimension Tables:
    - Dimension table contains the attributes on that truth table and calculates the metric.
    - While in the dimension table, There is more attributes than the fact table.
    - While in the dimension table, There is less records than in the fact table.
    - The dimension table forms a horizontal table.
    - While the attribute format of the dimension table is in text format.

3.2.2 ETL Concepts

ETL stands for Extraction, Transformation and Loading. These processes are required for transferring the operational data to data warehouse or presentation data.

Extraction: It is a process which involves reading and understanding the source data. It also involves copying the source data which can be put into a staging area for further manipulation.

Transformation:

- Cleansing data- Correcting spelling, checking missing data , checking and resolving the domain conflicts.
- Combination of data from multiple sources.
- Deduplication of data
- Assignment of warehouse keys.

Loading: It is basically the process of loading the data in the data warehouse to each of the data marts. Indexing should be there in the data mart before the arrival of data for better query performance. Thus, the loaded data is indexed and supplied for publishing.

Need for ETL in a Data Warehouse:

- Businesses have data in multiple databases with different codification and formats.
- Mergers and acquisitions have also created disparities in data representation.
- Transformation is required to convert and summarize operational data into a consistent, business-oriented form.
- Compute any derived information and aggregates.
- To maintain historic data in retrievable form

Reporting concepts:

Reporting is a board BI category and there are plenty of options and modes for its generation, definition, design, formatting and propagation.

A successful reporting platform implementation in a business intelligence environment requires great attention to be paid from both the business end users and IT professionals. The fact is that the reporting layer is what business users might consider a data warehouse system and if they do not like it, they will not use it.

The report types can be divided into several groups:

- Standard , static reports
- Ad Hoc reports
- Interactive . multidimensional OLAP reports
- Dashboards
- Technology Reports
- Write-back reports Reporting platforms:
    - IBM Cognos
    - SAP
    - SAS
    - Tableau
    - Microsoft Business intelligence
    - QlikView

3.3 Microsoft - AZURE

Microsoft Azure, formerly known as Windows Azure, is Microsoft's public cloud computing platform. It provides a range of cloud services, including compute, analytics, storage and networking. Users can pick and choose from these services to develop and scale new applications, or run existing applications in the public cloud.

The Azure platform aims to help businesses manage challenges and meet their organizational goals. It offers tools that support all industries -- including e-commerce, finance and a variety of Fortune 500 companies -- and is compatible with open source technologies. This provides users with the flexibility to use their preferred tools and technologies. In addition, Azure offers 4 different forms of cloud computing: infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS) and serverless.

Azure Services Used in this project:

- **Azure Blob Storage -** Is a service for storing large amounts of unstructured object data, such as text or binary data. You can use Blob storage to expose data publicly to the world, or to store application data privately. Common uses of Blob storage include:

- Serving images or documents directly to a browser
- Storing files for distributed access
- Streaming video and audio
- Storing data for backup and restore, disaster recovery, and archiving
- Storing data for analysis by an on-premises or Azure-hosted service

Azure Blob Storage helps you create data lakes for your analytics needs and provides storage to build powerful cloud-native and mobile apps. Optimize costs with tiered storage for your long-term data and flexibly scale up for high-performance computing and machine learning workloads.

- **Azure SQL Database -** Azure SQL Database is a fully managed platform as a service (PaaS) database engine that handles most of the database management functions such as upgrading, patching, backups, and monitoring without user involvement. Azure SQL Database is always running on the latest stable version of the SQL Server database engine and patched OS with 99.99% availability. With Azure SQL Database, you can create a highly available and high-performance data storage layer for the applications and solutions in Azure. SQL Database can be the right choice for a variety of modern cloud applications because it enables you to process both relational data and non-relational structures, such as graphs, JSON, spatial, and XML.
- **Azure Data Factory -** 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.

  ADF can pull data from the outside world (FTP, Amazon S3, Oracle, and many more), transform it, filter it, enhance it, and move it along to another destination. In my work for a health-data project we are using ADF to drive our data flow from raw ingestion to polished analysis that is ready to display.

Getting ADF to do real work for you involves the following layers of technology, listed from the highest level of abstraction that you interact with down to the software closest to the data.

  - Pipeline, the graphical user interface where you place widgets and draw data paths
  - Activity, a graphical widget that does something to your data
  - Source and Sink, the parts of an activity that specify where data is coming from and going to
  - Data Set, an explicitly defined set of data that ADF can operate on
  - Linked Service, the connection information that allows ADF to access a specific outside data resource
  - Integration Runtime, a glue/gateway layer that lets ADF talk to software outside of itself

- **ADF Pipeline Creation -** A data factory might have one or more pipelines. A pipeline is a logical grouping of activities that performs a unit of work. Together, the activities in a pipeline perform a task. For example, a pipeline can contain a group of activities that ingests data from an Azure blob, and then runs a Hive query on an HDInsight cluster to partition the data. The benefit of this is that the pipeline allows you to manage the activities as a set instead of managing each one individually. The activities in a pipeline can be chained together to operate sequentially, or they can operate independently in parallel.
- **ADF Mapping Data Flow** - Create and manage graphs of data transformation logic that you can use to transform any-sized data. You can build-up a reusable library of data transformation routines and execute those processes in a scaled-out manner from your ADF pipelines. Data Factory will execute your logic on a Spark cluster that spins-up and spins-down when you need it. You won't ever have to manage or maintain clusters.
- **Azure Resource Group -** A resource group is a container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group.
- **Azure Linked Service -** Linked services are much like connection strings, which define the connection information needed for the service to connect to external resources.

CHAPTER 3

HARDWARE AND SOFTWARE REQUIREMENTS

CHAPTER 4: HARDWARE AND SOFTWARE REQUIREMENTS

Hardware and software requirements of any project must be satisfied, so that the virtual environment can be set up on any machine to run the project. So, in this section the software and the hardware requirements are discussed completely.

4.1 HARDWARE REQUIREMENTS

The program can run on a machine with following configurations:

- Windows 10 Operating system/ Linux/ Ubuntu
- Strong Internet connection
- Ram>4gb

4.2 SOFTWARE REQUIREMENTS

The following section will cover aspects related to Enterprise Data Warehouse for Daily Sales Report application.

The following are the modules in this proposed system

- **Store** - This Module stores the details on store ID, name, branch available stock and product ID
- **Salesperson** - This Module stores the details on sales person, name, supervisor ID, data of joining and goal ID
- **product** - This Module maintains the details on product ID, Product Name, price, manufacturing date and expiry date
- **Sales** - This module stores the details on sales ID, product ID, store ID, salesperson ID, sold quantity, Amount and date of sales
- **Goal** - This module stores the details on goal ID, goal name, goal expiry date, goal amount and goal quantity per year

SQL Management Studio - creating on premises Databases and Star Schema creation and uploading those databases on Azure SQL Database server.

4.3 SYSTEM ENVIRONMENT

- The primary focus should be the technical aspect of EIM, i.e. data warehousing concepts. Business functions and calculations are secondary.
- Use lexicons for table/column names for consistency.

- Common dimensions like customer should be the same for all business areas. One of the teams should take care of developing/refreshing common dimensions.
- The DW Data model should be a physical one, i.e. should identify table/column names along with size, primary keys (surrogate keys), business keys and foreign keys.
- Data model is not required for staging tables. However these tables don't need to have surrogate/foreign keys.
- Date dimension should be a common one across all business areas.
- All dimensions and facts should be prefixed with DIM and FCT respectively.
- Suffix surrogate keys, business keys and foreign keys with SK, BK and FK respectively.
- Insert and Update Timestamps apart from the start and end dates for dimensions
- Lexicon for ETL
- Microsoft Azure

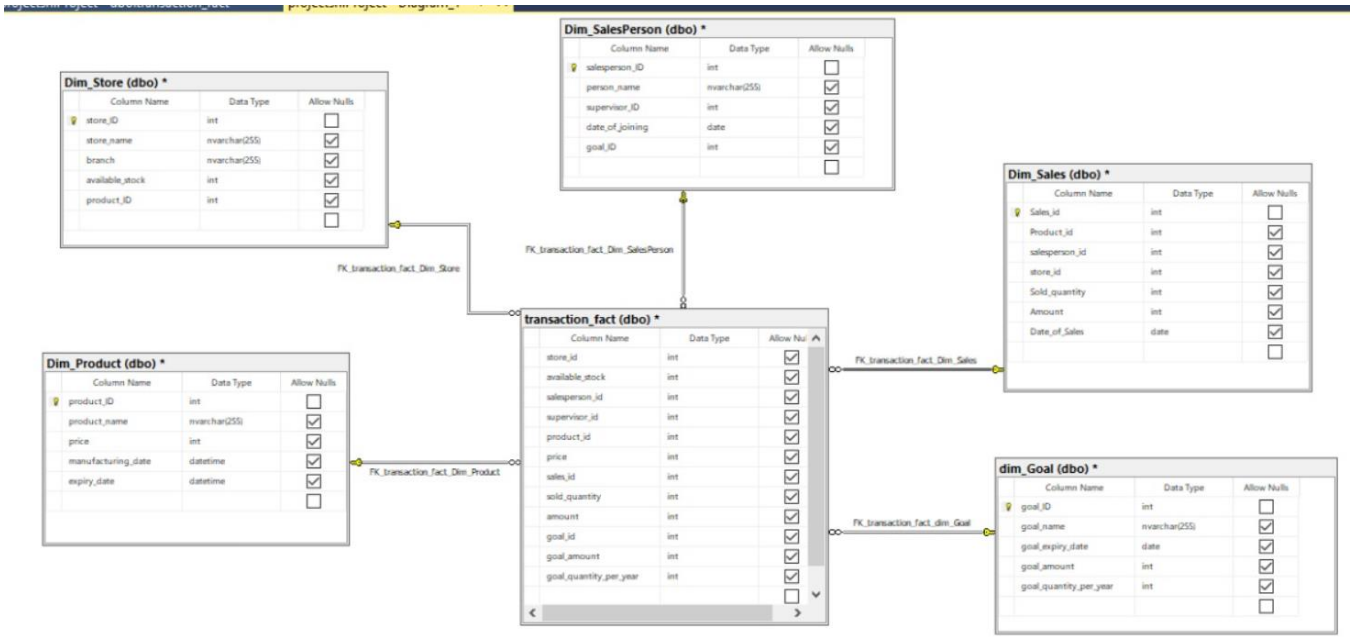
CHAPTER 4

ENVIRONMENTAL SETUP

CHAPTER 5: ENVIRONMENTAL SETUP

5.1 ARCHITECTURE DIAGRAM

5.1.1 STAR SCHEMA



5.1.2 PHYSICAL ARCHITECTURE

Physical Architecture:

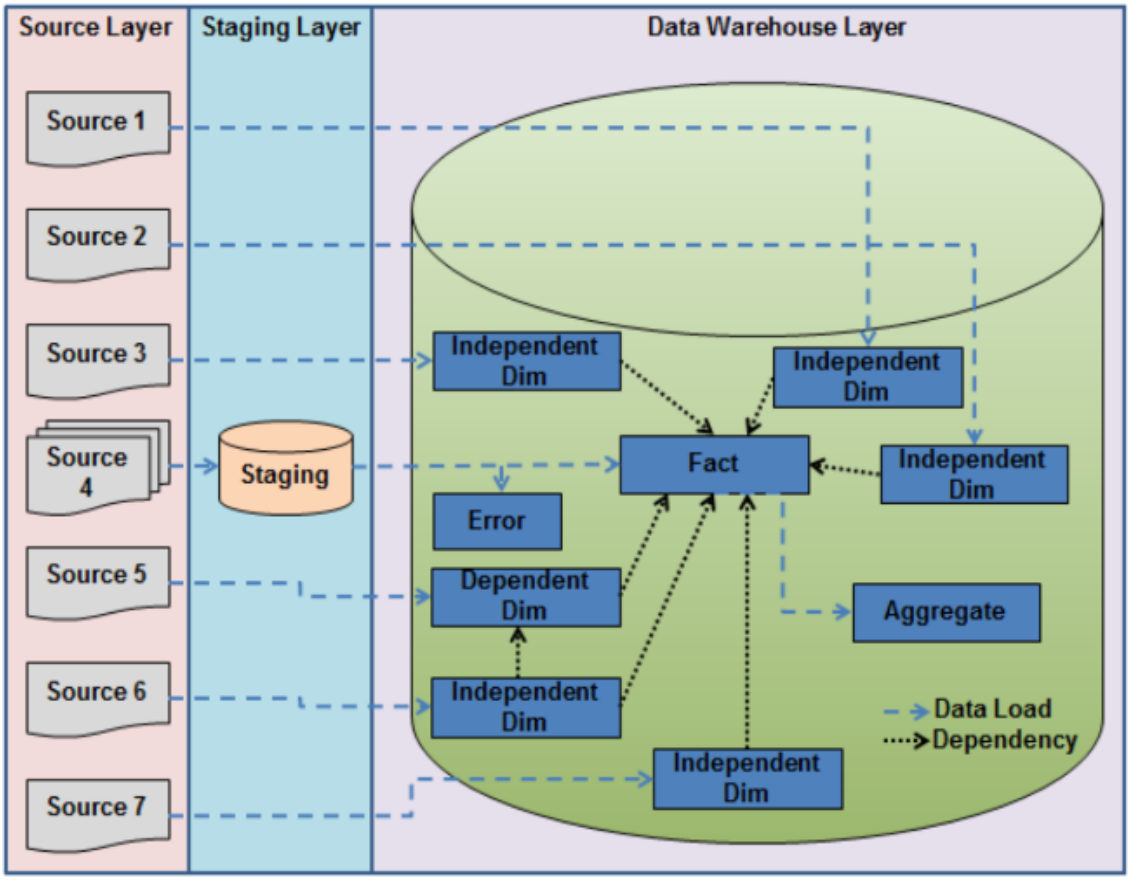


Fig. 2

5.1.3 LOGICAL ARCHITECTURE

The Logical Architecture defines the Processes (the activities and functions) that are required to provide the required User Services. Many different Processes must work together and share information to provide a User Service. The Processes can be implemented via software, hardware, or firmware. Logical Architecture is

independent of technologies and implementations.

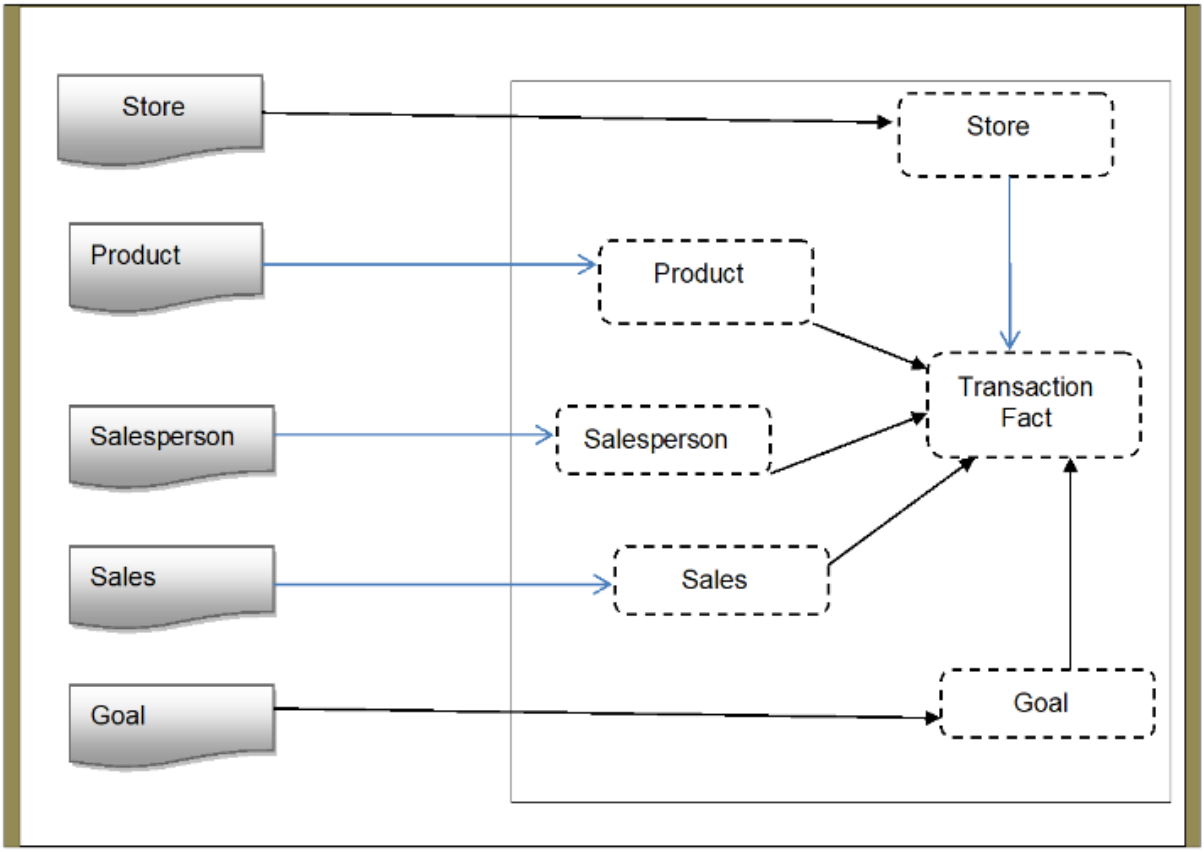


Fig. 3

5.1.4 DATA WAREHOUSE LIFE CYCLE

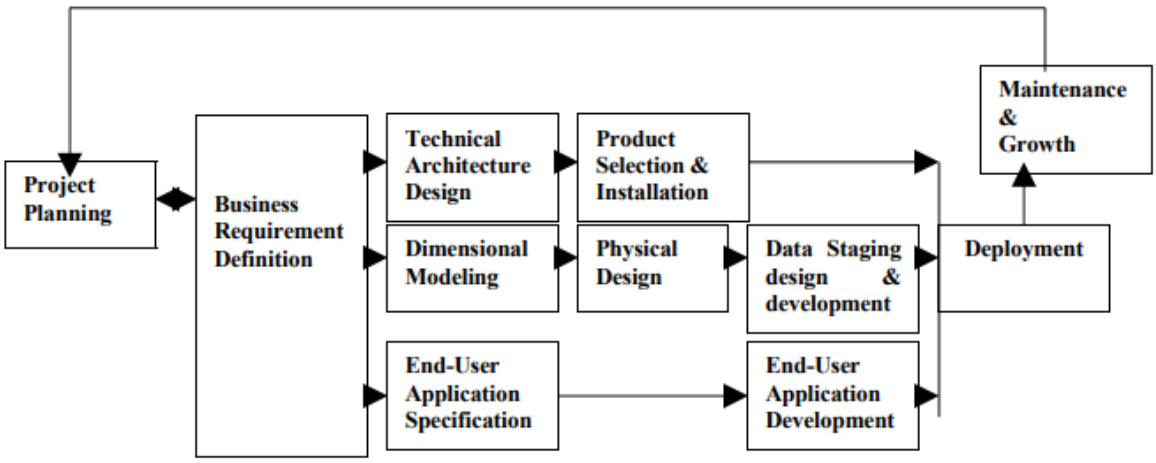


Fig. 4

CHAPTER 5

PROJECT MODELING AND FUNCTIONAL REQUIREMENTS

CHAPTER 6: PROJECT MODELING AND FUNCTIONAL REQUIREMENTS

6.1 Project Planning:

Daily sales report is the core-tracking element for controlling cash, sales performance and forecasting the future sales. The purpose of the daily sales report is to be informed about the level of activity taking place within the sales organization and to make sure that each salesperson is meeting or exceeding the sales target.

6.2 Business Requirements

The data warehousing system is being developed to enhance the decision support and generating customized reports for Daily Sales, and help in core-tracking elements for controlling cash, sales performance and forecasting the future sales.

6.3 Dimensional Modeling

- **SQL Server Management Studio (SSMS)**

  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.

6.1.1 Tables formed in SSMS

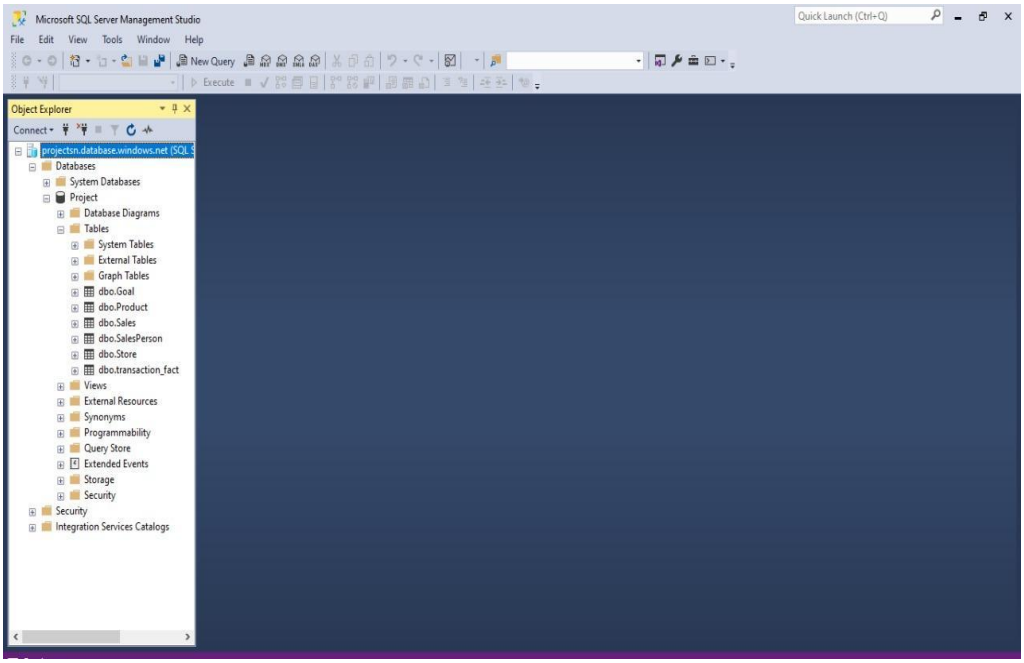

Microsoft SQL Server Management Studio
File Edit View Tools Window Help
New Query
Execute
Quick Launch (Ctrl+Q)
Object Explorer
Connect
projectsn.database.windows.net (SQL S
Databases
System Databases
Project
Database Diagrams
Tables
System Tables
External Tables
Graph Tables
dbo.Goal
dbo.Product
dbo.Sales
dbo.SalesPerson
dbo.Store
dbo.transaction_fact
Views
External Resources
Synonyms
Programmability
Query Store
Extended Events
Storage
Security
Security
Integration Services Catalogs

Fig. 5

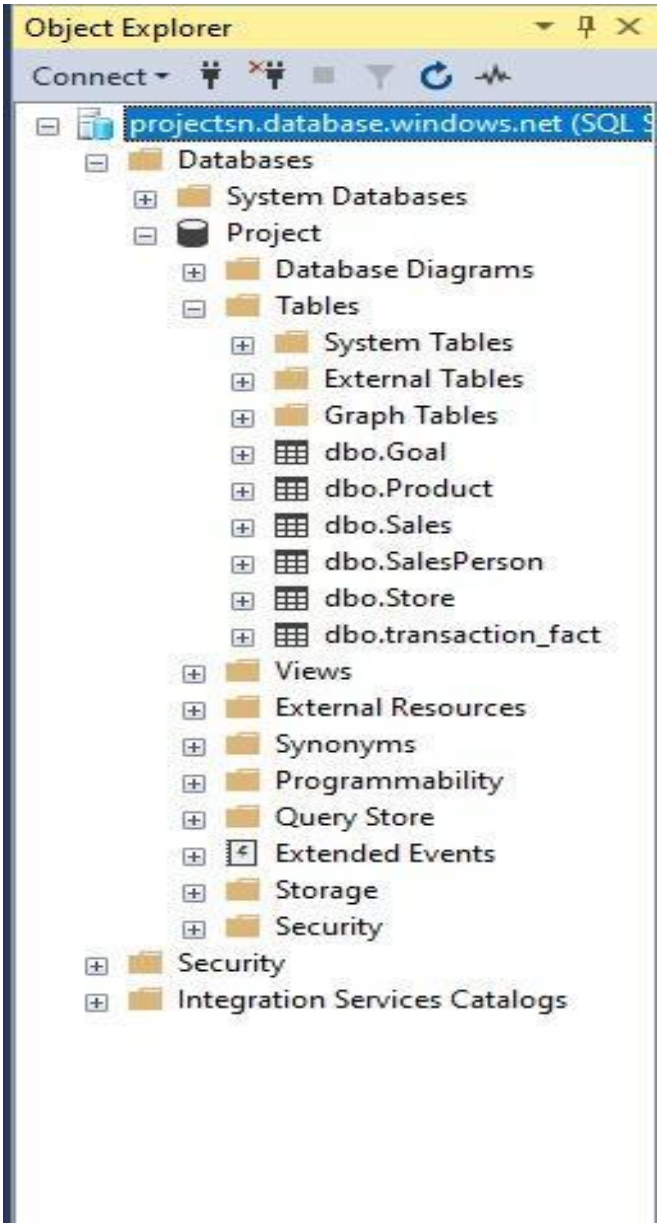

Object Explorer
Connect
projectsn.database.windows.net (SQL S
Databases
System Databases
Project
Database Diagrams
Tables
System Tables
External Tables
Graph Tables
dbo.Goal
dbo.Product
dbo.Sales
dbo.SalesPerson
dbo.Store
dbo.transaction_fact
Views
External Resources
Synonyms
Programmability
Query Store
Extended Events
Storage
Security
Security
Integration Services Catalogs

Fig. 6

6.1.2 Project Source Data:

store_ID	store_name	branch	available_stock	product_ID	branch_id
1245	nike	delhi	100	3	5
2176	puma	noida	65	5	6
4567	crocs	chennai	250	6	3
8274	bata	ranchi	300	8	7
1267	haldirams	haridwar	500	9	8
3790	burgerkings	noida	400	10	9
9836	h&m	gurgaon	70	12	10
6032	casio	delhi	50	13	5
1176	samsung	cp	20	14	12
4455	appple	mumbai	10	15	23
8547	sony	jaipur	80	16	34
1002	jbl	haryana	90	17	54
1996	boat	punjab	450	18	22
5082	lays	banglore	1000	19	11
6722	addidas	delhi	5	20	5

Table 1 - Store Table

Sales id	Product id	sales pers	store id	Sold quan	Amount	Date of Sale	Goal id
456	3	101	1245	90	135000	8/30/201	11
123	5	102	2176	60	360000	1/7/2022	13
789	6	103	4567	230	460000	7/26/2022	15
528	8	104	8274	280	280000	12/15/2021	33
252	9	105	1267	500	2500	4/25/2022	55
741	10	106	3790	390	78000	5/13/2021	66
564	12	107	9836	60	54000	1/13/2022	77
456	13	108	6032	40	160000	10/9/2022	88
262	14	109	1176	15	117000	11/19/2021	99
900	15	110	4455	8	800000	6/25/2022	22
290	16	111	8547	70	1400000	12/13/2022	44
299	17	112	1002	80	1280000	6/10/2022	23
230	18	113	1996	85	127500	7/20/2022	34
951	19	114	5082	900	45000	8/21/2021	54
159	20	115	6722	3	36000	6/30/2022	74

Table 2 - Sales Table

salesperson_ID	person_name	supervisor_ID	date_of_joining	goal_ID
101	amogh	4	5/29/2000	11
102	viraj	5	2/20/2020	13
103	rajesh	7	5/25/2010	15
104	vaishnavi	6	11/12/2007	33
105	yogesh	2	5/23/2010	55
106	ramu	8	2/26/1999	66
107	pratibha	9	12/9/2018	77
108	tanmay	20	6/24/2017	88
109	tushar	13	10/5/2016	99
110	aakash	45	6/14/2015	22
111	ramdev	11	7/2/2012	44
112	modi	74	7/8/2013	23
113	rahul	1	4/11/2020	34
114	indira	3	12/6/2021	54
115	srk	10	5/9/2019	74

Table 3 - Salesperson Table

goal_ID	goal_name	goal_expiry_date	goal_amount	goal_quantity_per_year
11	A	7/4/2021	80	78
13	B	12/14/2021	55	50
15	C	1/30/2022	200	170
33	D	12/26/2021	250	200
55	E	2/1/2022	500	500
66	F	12/14/2021	489	450
77	G	4/24/2022	400	390
88	H	12/31/2021	44	40
99	I	3/25/2022	40	35
22	J	4/15/2022	15	10
44	K	9/10/2021	4	3
23	L	3/5/2022	70	50
34	M	2/15/2022	50	30
54	N	6/8/2022	65	60
74		10/6/2021	5	3

Table 4 - Goal Table

product_ID	product_name	price	manufacturing_date	expiry_date
3	jordans	15000	8/26/2021	10/27/2028
5	evos	6000	6/30/2022	4/23/2030
6	flip flops	2000	5/11/2022	12/28/2025
8	sandal	1000	7/7/2021	7/7/2025
9	ladoo	50	4/25/2022	6/28/2022
10	whooper	200	2/12/2021	2/13/2021
12	jeans	900	7/12/2021	4/23/2024
13	watch	4000	4/4/2022	12/10/2023
14	note s20	78000	7/12/2021	2/18/2031
15	iphone 14	100000	4/7/2022	4/28/2032
16	sony TV	200000	6/18/2022	12/7/2037
17	jbl pulse speaker	16000	5/26/2022	2/23/2024
18	boat nirvana	1500	2/13/2022	8/20/2024
19	magic masala	50	7/21/2021	1/31/2022
20	yeezy's	12000	1/30/2022	1/12/2025

Table 5 - Product Table

6.4 FUNCTIONAL REQUIREMENTS - AZURE SQL DATABASE

This part of the project deals with Azure SQL Database, Azure Blob Storage and creating the pipelines in Azure Data Factory and automate the whole process

6.2.1 Uploading the source data in Azure Blob Storage

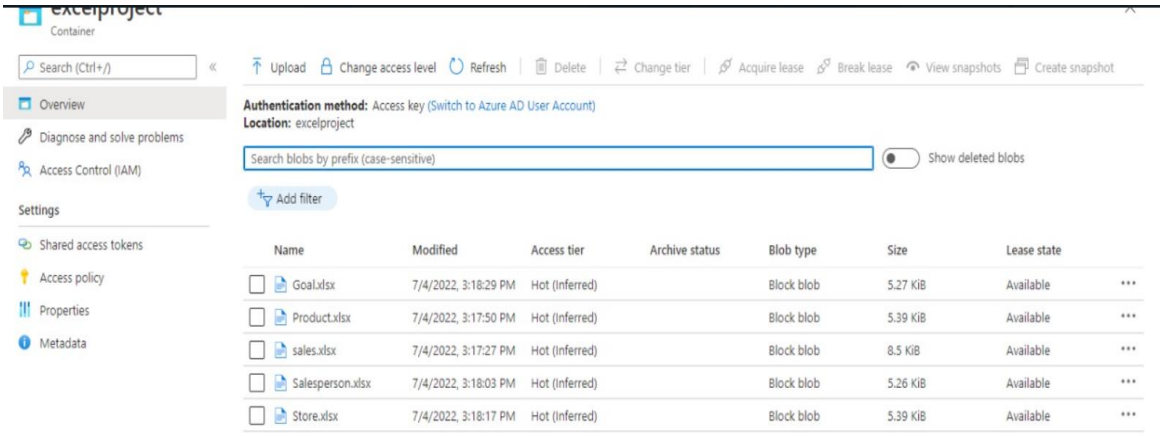
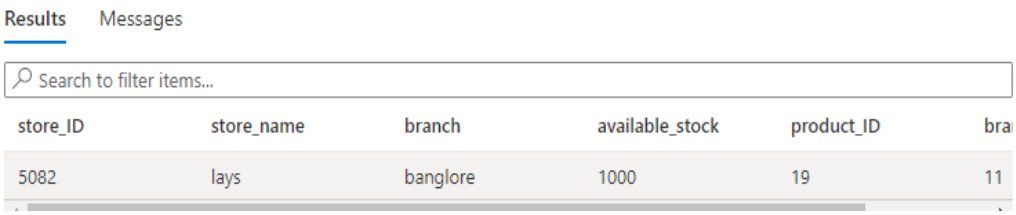


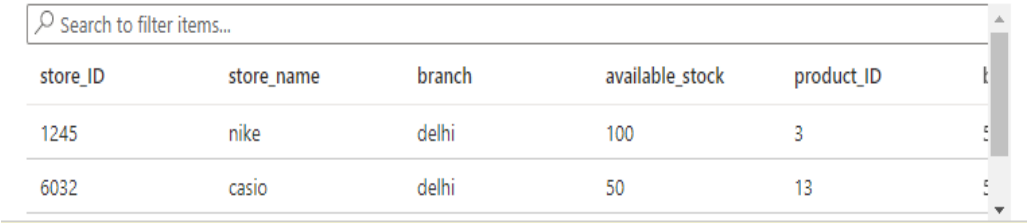
Fig. 7 - Blob Storage Account

6.2.2 Functional Requirements of database modules

- **Store Module**:
  - **Target1** - Produce report on store details for each branch having maximum stock
  - Transformation : Aggregrator , Sub Queries
  - Query:
    - select * from store where available_stock=(select max(available_stock) from store);



  - **Target 2:** To fetch store details having more than 3 products in each branch
  - Transformation : Aggregator , Filter , Subqueries
  - Query:
    - select * from store where branch in (select branch from store group by branch having count(product_id)>3);



  - **Target 3**: Dimension Store is used to get the Store Details of the product for daily sales report
  - Transformation : join
  - Query:
    - Select * from Store s join Product p on s.product_id = p.product_id ;

store_ID	store_name	branch	available_stock
1245	nike	delhi	100
2176	puma	noida	65

- **Salesperson module** :
    - **Target 1**: Report salesperson who is having experience more than 15 yrs
    - Transformation : Filter
    - Query:
        - SELECT * FROM salesperson WHERE year(getdate())-year(date_of_joining) > 15;

Search to filter items...

salesperson_ID	person_name	supervisor_ID	date_of_joining	goal_ID
101	amogh	4	2000-05-29T00:00:00.00000...	11
106	ramu	8	1999-02-26T00:00:00.00000...	66

    - **Target2**: To display Salesperson details based on i) Name ii) Date of Joining
    - Transformation : Filter
    - Query:
        - select person_name,date_of_joining from salesperson;

person_name	date_of_joining
amogh	2000-05-29T00:00:00.0000000
viraj	2020-02-20T00:00:00.0000000

    - **Target 3**: To fetch list of salesperson who is having more than 1 goal
    - Transformation : Filter
    - Query:
        - SELECT person_name FROM salesperson where salesperson_id in (select salesperson_id from salesperson group by salesperson_id having count(goal_id)>1);

Results Messages

Search to filter items...

person_name
amogh
arjun

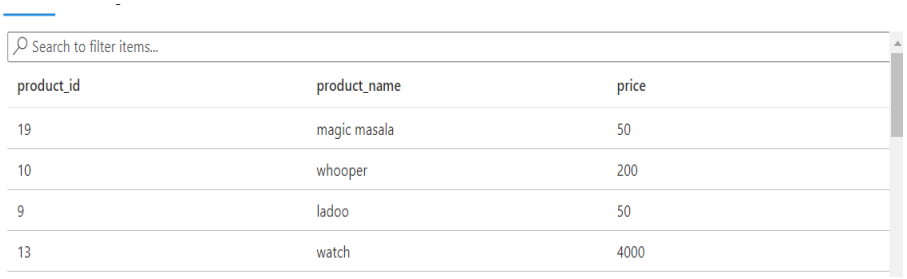
- **Target 4**: Dimension Salesperson is used to get the Salesperson information
- Transformation :
- Query:
  - Select * from Salesperson;

salesperson_ID	person_name	supervisor_ID	date_of_joining	goal_ID
101	amogh	4	2000-05-29T00:00:00....	11
102	viraj	5	2020-02-20T00:00:00....	13
103	rajesh	7	2010-05-25T00:00:00....	15

- **Product Module** :
  - **Target1** : Report product information of each store
  - Transformation : join
  - Query :
    - Select p.product_id , p.product_name , p.price , p.manufacturing_date , p.expiry_date , s.store_name from product as p join store as s on p.product_id = s.product_id;

3	jordans	15000	2021-08-26T00:00:0...	2028-10-27T00:00:0..
5	evos	6000	2022-06-30T00:00:0...	2030-04-23T00:00:0..
6	flip flops	2000	2022-05-11T00:00:0...	2025-12-28T00:00:0..
8	sandal	1000	2021-07-07T00:00:0	2025-07-07T00:00:0

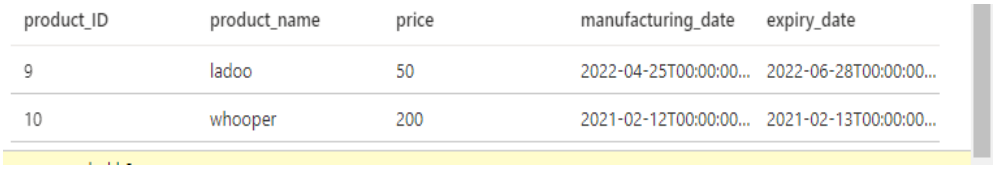
  - **Target 2**: To fetch product details based on expiry Date and manufacturing date.
  - Transformation : Filter
  - Query:
    - select product_id,product_name,price from product order by expiry_date,manufacturing_date;



Search to filter items...

product_id	product_name	price
19	magic masala	50
10	whooper	200
9	ladoo	50
13	watch	4000

- **Target3**: To display product which is yet to expire and display based on price
- Transformation : Filter
- Query:
  - select * from product where expiry_date < convert(date, getdate(), 101) order by price;



product_ID	product_name	price	manufacturing_date	expiry_date
9	ladoo	50	2022-04-25T00:00:00...	2022-06-28T00:00:00...
10	whooper	200	2021-02-12T00:00:00...	2021-02-13T00:00:00...

- **Target 4**: Dimension Product is used to get the product details.
- Transformation : Filter
- Query:
  - Select * from product;

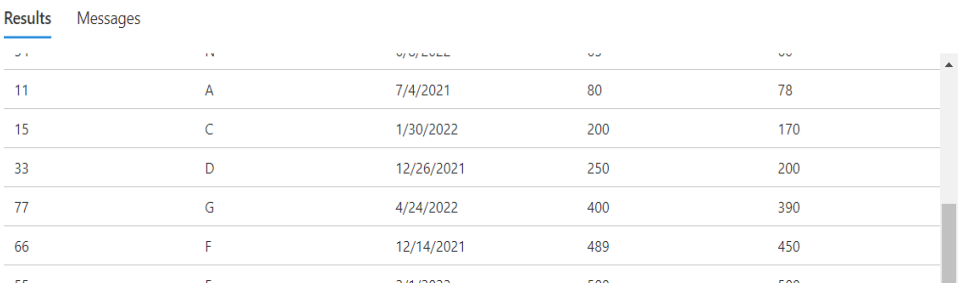
product_ID	product_name	price	manufacturing_date	expiry_date
3	jordans	15000	2021-08-26T00:00:00...	2028-10-27T00:00:0(
5	evos	6000	2022-06-30T00:00:00...	2030-04-23T00:00:0(

- **Sales Module**:
  - **Target 1**: To store salesperson who made the maximum sales each day
  - Transformation : Joins , Filter , Aggregate
  - Query:
    - Select sp.person_name from salesperson as sp join sales as s on sp.salespeson_id=s.salesperson_id where s.amount=(select max(amount) from sales);

person_name
ramdev

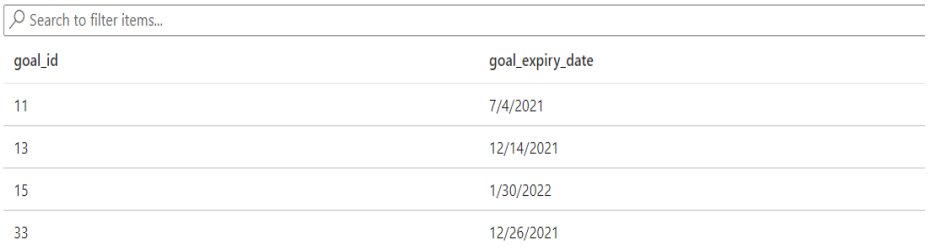
- **Target 2**: To fetch second largest product price sold for current month
- Transformation : Filter
- Query :
  - select top(1) amount from (select top(2) amount from sales where substring(date_of_sales,1,2) = substring(convert(varchar, getdate(), 101),1,2) order by amount desc) as sales order by amount;
- **Target 3**: To generate a store which sold MANY products.
- Transformation: Filter
- Query :
  - select * from sales where sales_id in (select sales_id from sales group by sales_id having count(*)>1);

- **Goal Module**:
  - **Target1**: To fetch the goal details based on quantity
  - Transformation: Filter
  - Query:
    - select * from goal order by goal_quantity_per_year;



  - **Target3**: To check the goal expiry date for a goal number
  - Transformation: Filter
  - Query:

- select goal_id,goal_expiry_date from goal;



Search to filter items...

goal_id	goal_expiry_date
11	7/4/2021
13	12/14/2021
15	1/30/2022
33	12/26/2021

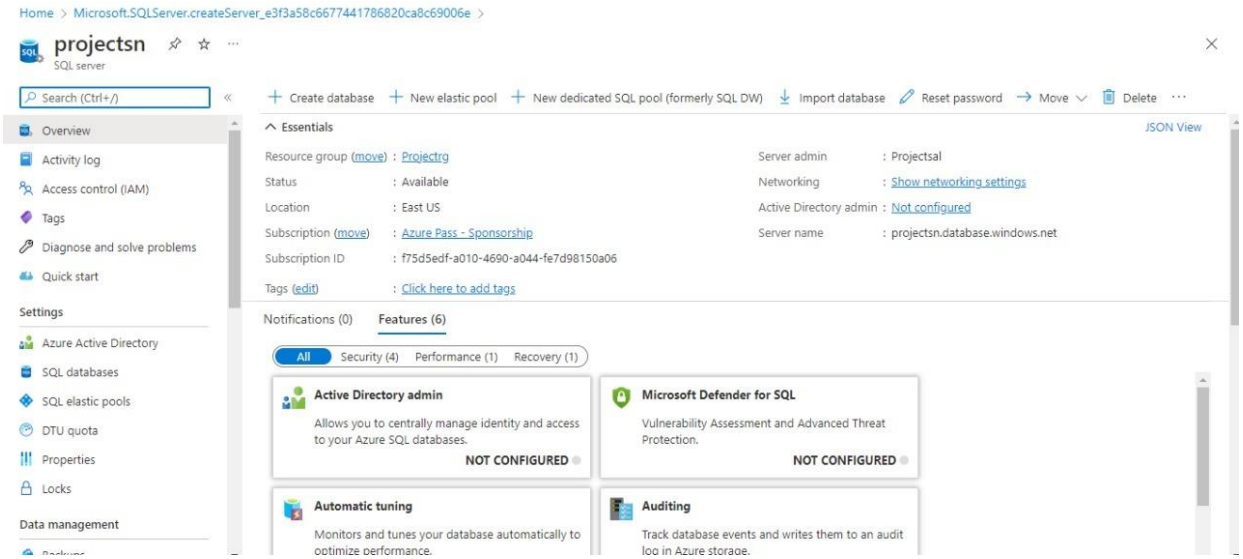
6.5 AZURE DATA FACTORY - PIPELINE CREATION AND DATA FLOW

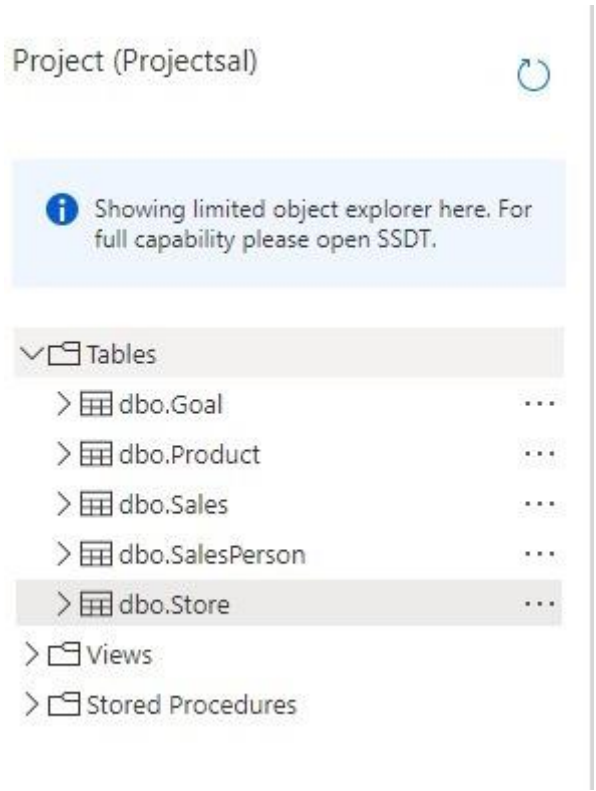
- A data factory can have one or more pipelines. A pipeline is a logical grouping of activities that together perform a task. The activities in a pipeline define actions to perform on your data. For example, you may use a copy activity to copy data from a SQL Server database to an Azure Blob Storage. Then, use a Hive activity that runs a Hive script on an Azure HDInsight cluster to process/transform data from the blob storage to produce output data. Finally, use a second copy activity to copy the output data to an Azure Synapse Analytics on top of which business intelligence (BI) reporting solutions are built.
- An activity can take zero or more input datasets and produce one or more output datasets. The following diagram shows the relationship between pipeline, activity, and dataset in Data Factory:
- A pipeline allows you to manage activities as a set instead of each one individually. For example, you can deploy, schedule, suspend, and resume a pipeline, instead of dealing with activities in the pipeline independently.
- Data Factory supports two types of activities: data movement activities and data transformation activities. Each activity can have zero or more input datasets and produce one or more output datasets.

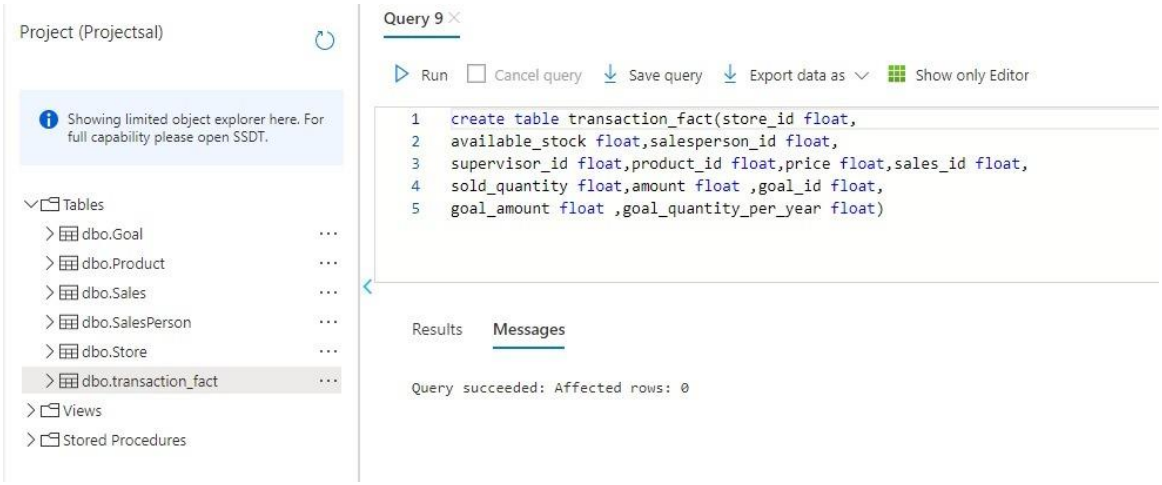
CHAPTER 6

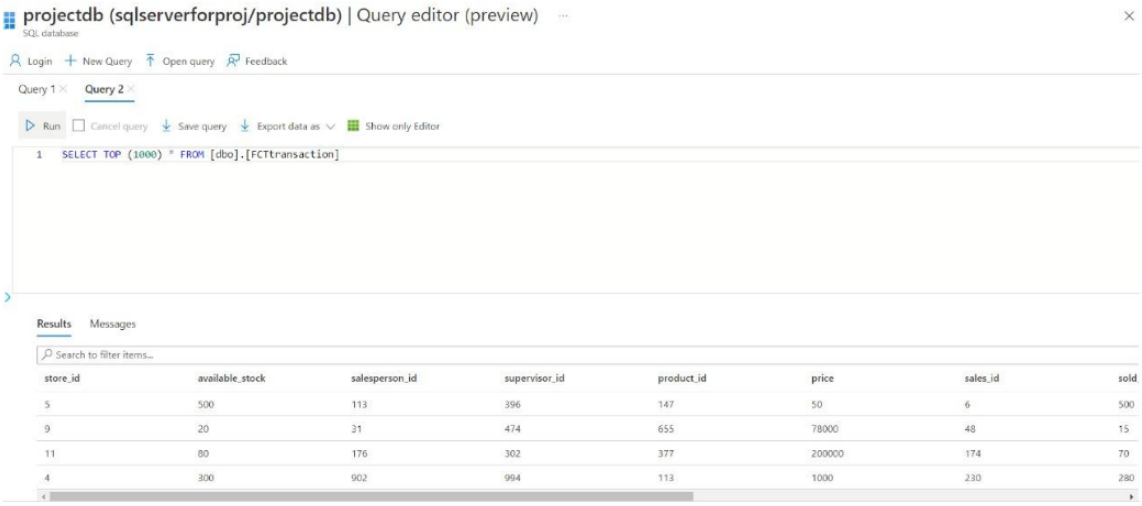
GUI SNAPSHOTS

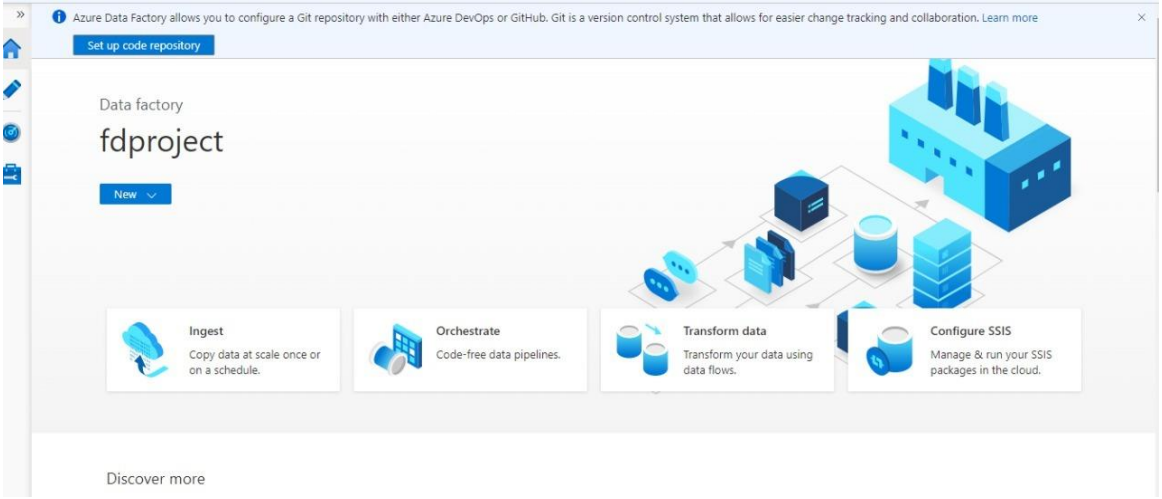
CHAPTER 7 : GUI SNAPSHOTS

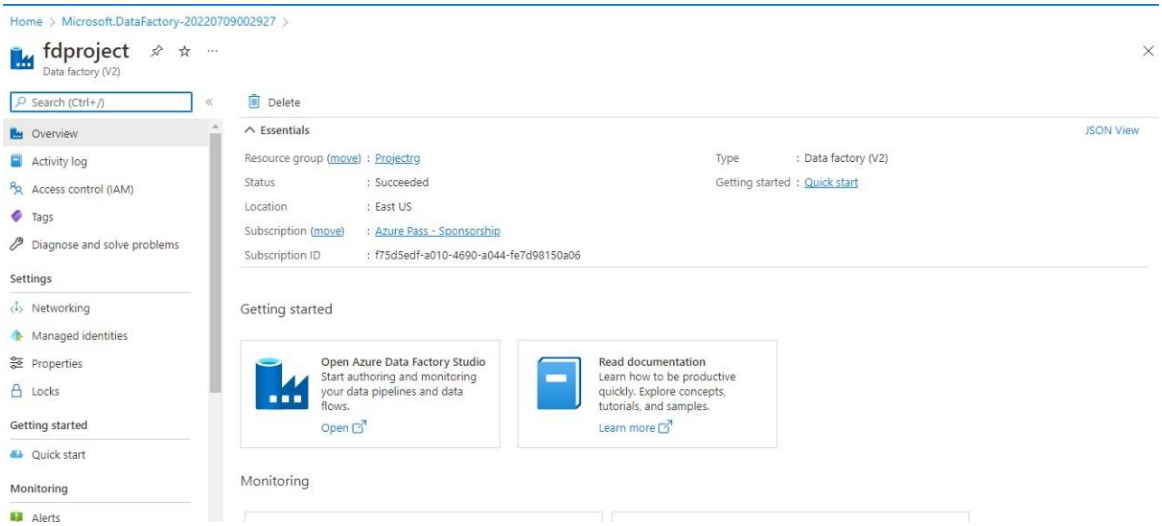


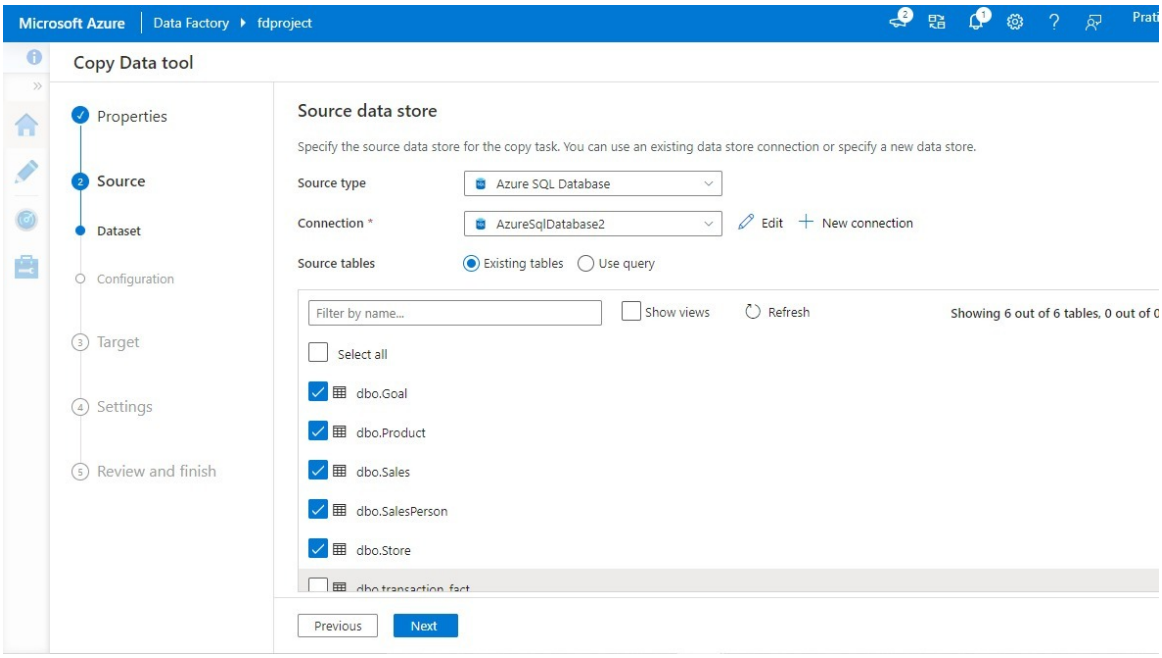


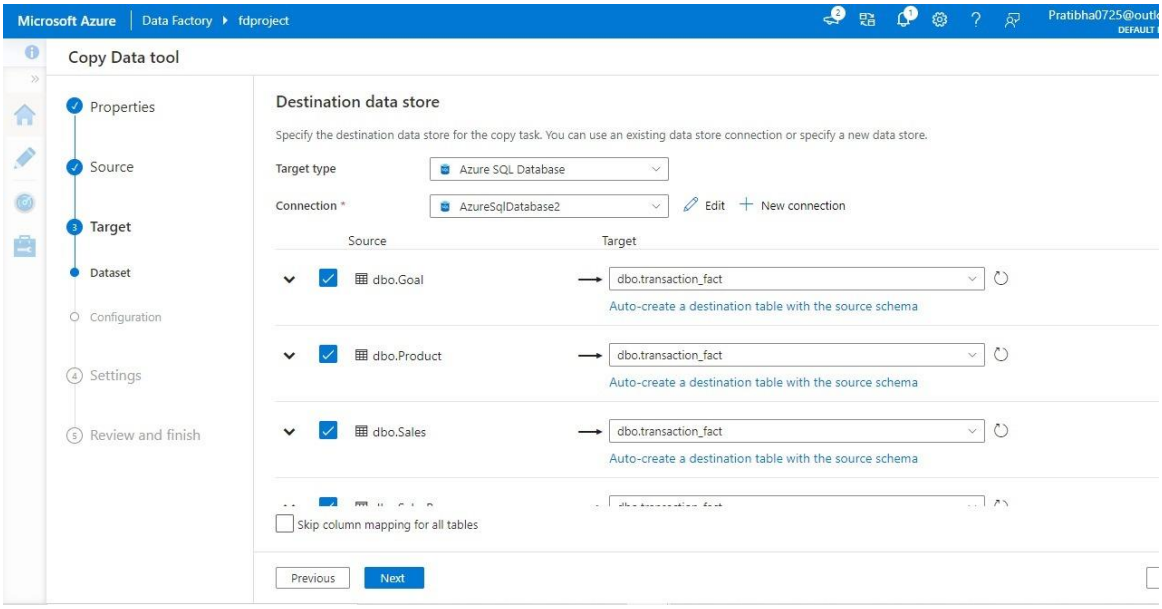

Project (Projectsal)
Showing limited object explorer here. For full capability please open SSDT.
Tables
dbo.Goal
dbo.Product
dbo.Sales
dbo.SalesPerson
dbo.Store
dbo.transaction_fact
Views
Stored Procedures
Query 9
Run
Cancel query
Save query
Export data as
Show only Editor
1 create table transaction_fact(store_id float,
2 available_stock float,salesperson_id float,
3 supervisor_id float,product_id float,price float,sales_id float,
4 sold_quantity float,amount float ,goal_id float,
5 goal_amount float ,goal_quantity_per_year float)
Results
Messages
Query succeeded: Affected rows: 0

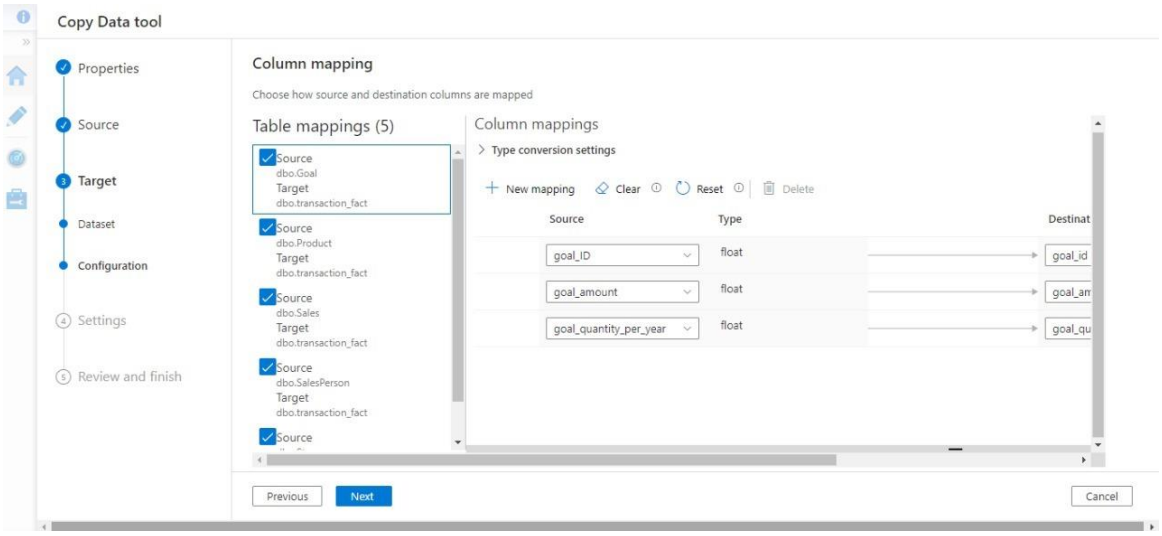

projectdb (sqlserverforproj/projectdb) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
Query 1
Query 2
Run
Cancel query
Save query
Export data as
Show only Editor
1 SELECT TOP (1000) * FROM [dbo].[FCTtransaction]
Results
Messages
Search to filter items...
store_id available_stock salesperson_id supervisor_id product_id price sales_id sold
5 500 113 396 147 50 6 500
9 20 31 474 655 78000 48 15
11 80 176 302 377 200000 174 70
4 300 902 994 113 1000 230 280

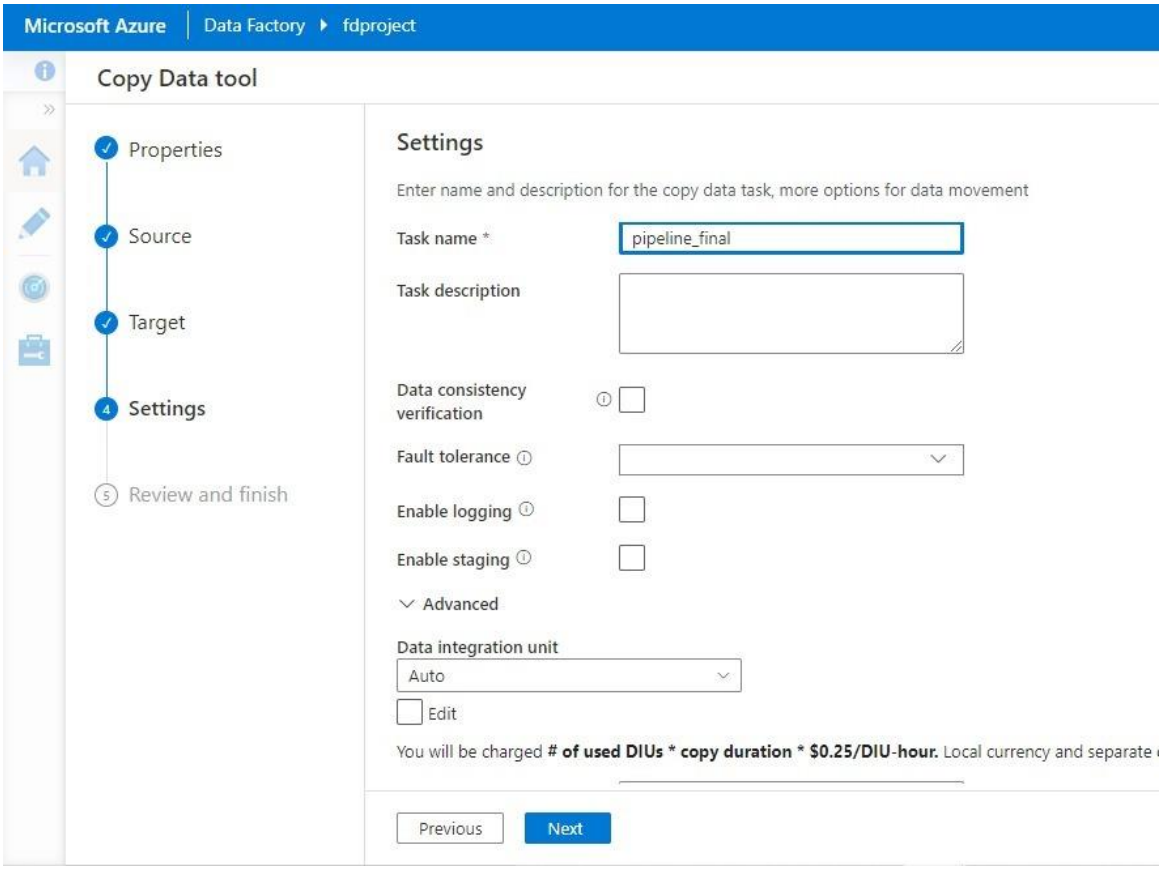

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration. Learn more
Set up code repository
Data factory
fdproject
New
Ingest
Copy data at scale once or on a schedule.
Orchestrate
Code-free data pipelines.
Transform data
Transform your data using data flows.
Configure SSIS
Manage & run your SSIS packages in the cloud.
Discover more

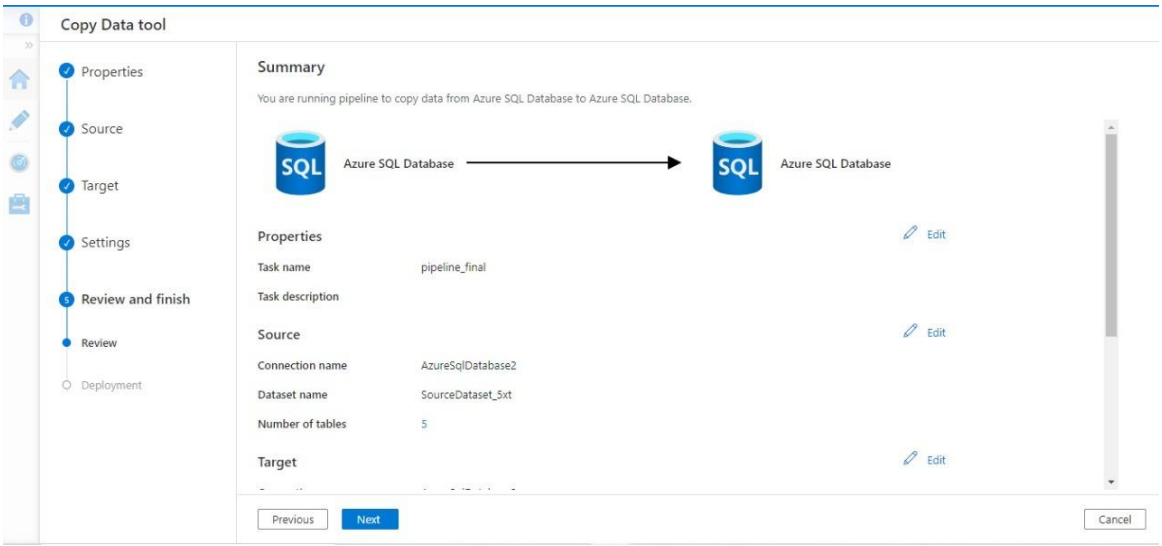

Home > Microsoft.DataFactory-20220709002927 >
fdproject
Data factory (V2)
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Networking
Managed identities
Properties
Locks
Getting started
Quick start
Monitoring
Alerts
Delete
Essentials
JSON View
Resource group (move) : Projectrg
Status : Succeeded
Location : East US
Subscription (move) : Azure Pass - Sponsorship
Subscription ID : f75d5edf-a010-4690-a044-fe7d98150a06
Type : Data factory (V2)
Getting started : Quick start
Getting started
Open Azure Data Factory Studio
Start authoring and monitoring your data pipelines and data flows.
Open
Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more
Monitoring

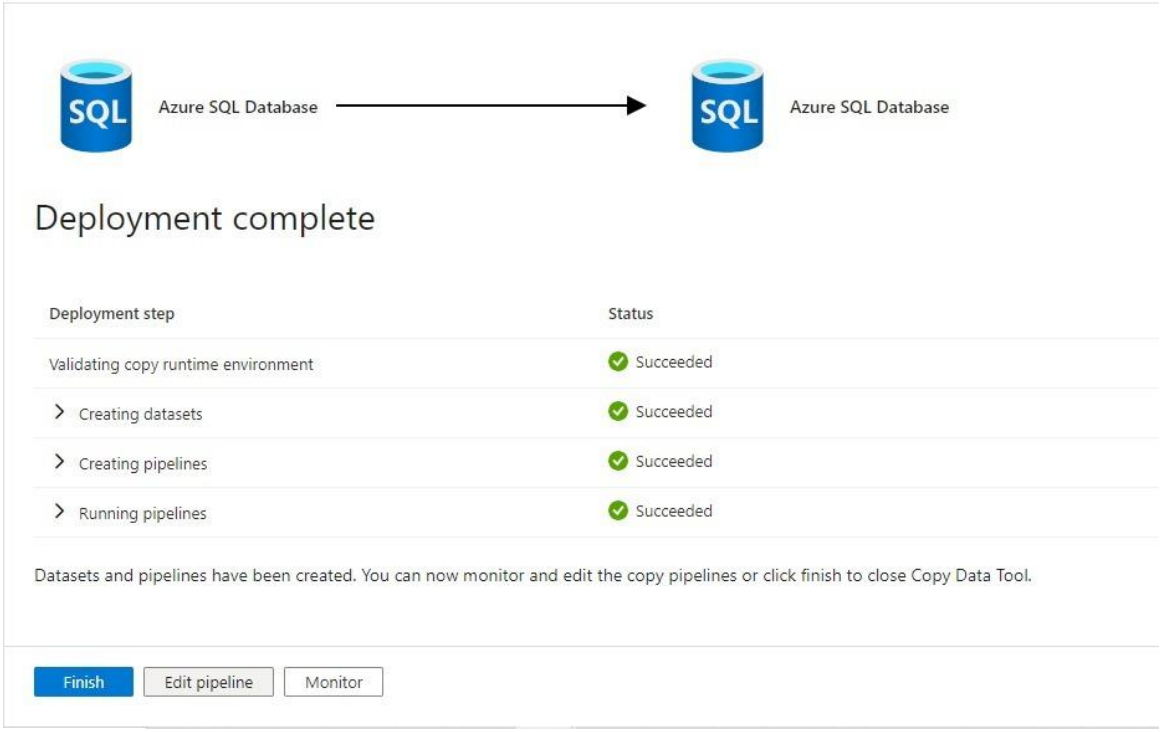

Microsoft Azure
Data Factory
fdproject
Copy Data tool
Properties
Source
Dataset
Configuration
Target
Settings
Review and finish
Source data store
Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.
Source type
Azure SQL Database
Connection *
AzureSqlDatabase2
Edit
New connection
Source tables
Existing tables
Use query
Filter by name...
Show views
Refresh
Showing 6 out of 6 tables, 0 out of
Select all
dbo.Goal
dbo.Product
dbo.Sales
dbo.SalesPerson
dbo.Store
Previous
Next

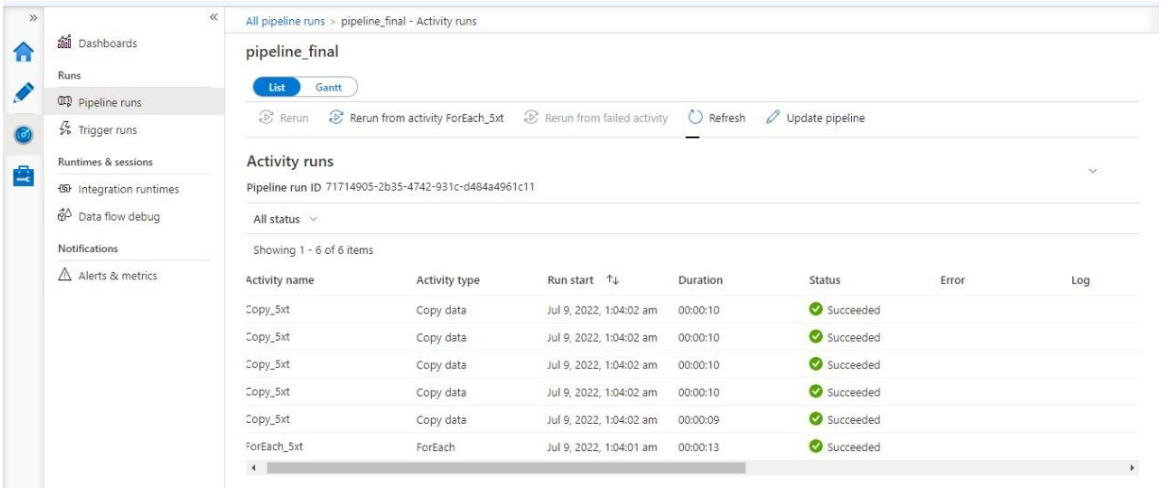

Microsoft Azure
Data Factory
fdproject
Copy Data tool
Properties
Source
Target
Dataset
Configuration
Settings
Review and finish
Destination data store
Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.
Target type
Azure SQL Database
Connection *
AzureSqlDatabase2
Edit
New connection
Source
Target
dbo.Goal
dbo.transaction_fact
Auto-create a destination table with the source schema
dbo.Product
dbo.transaction_fact
Auto-create a destination table with the source schema
dbo.Sales
dbo.transaction_fact
Auto-create a destination table with the source schema
Skip column mapping for all tables
Previous
Next

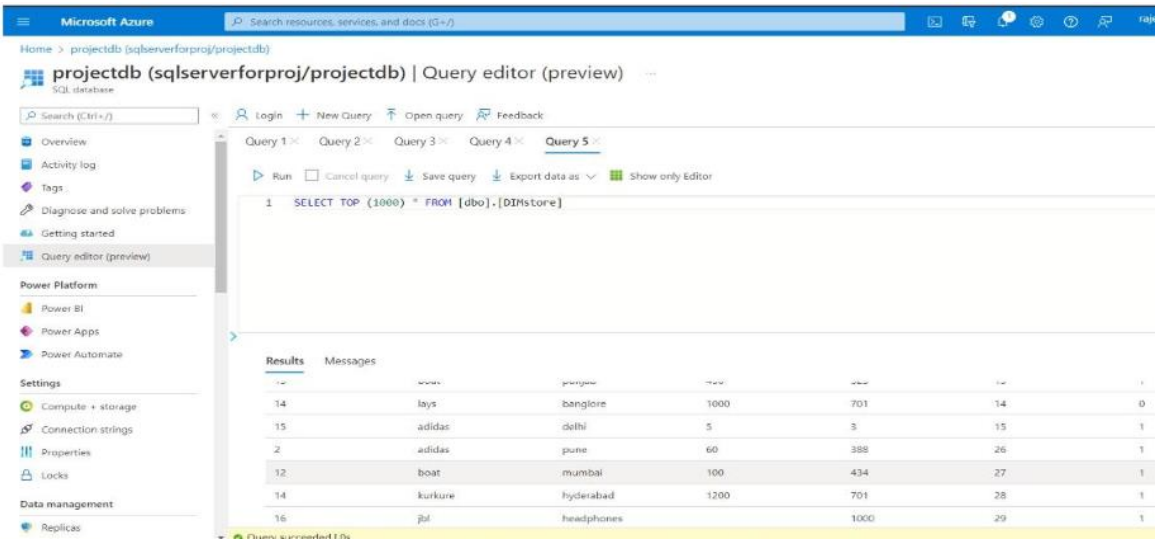

Copy Data tool
Properties
Source
Target
Dataset
Configuration
Settings
Review and finish
Column mapping
Choose how source and destination columns are mapped
Table mappings (5)
Source
dbo.Goal
Target
dbo.transaction_fact
Source
dbo.Product
Target
dbo.transaction_fact
Source
dbo.Sales
Target
dbo.transaction_fact
Source
dbo.SalesPerson
Target
dbo.transaction_fact
Source
Column mappings
Type conversion settings
New mapping
Clear
Reset
Delete
Source
Type
Destinat
goal_ID
float
goal_id
goal_amount
float
goal_amt
goal_quantity_per_year
float
goal_qu
Previous
Next
Cancel

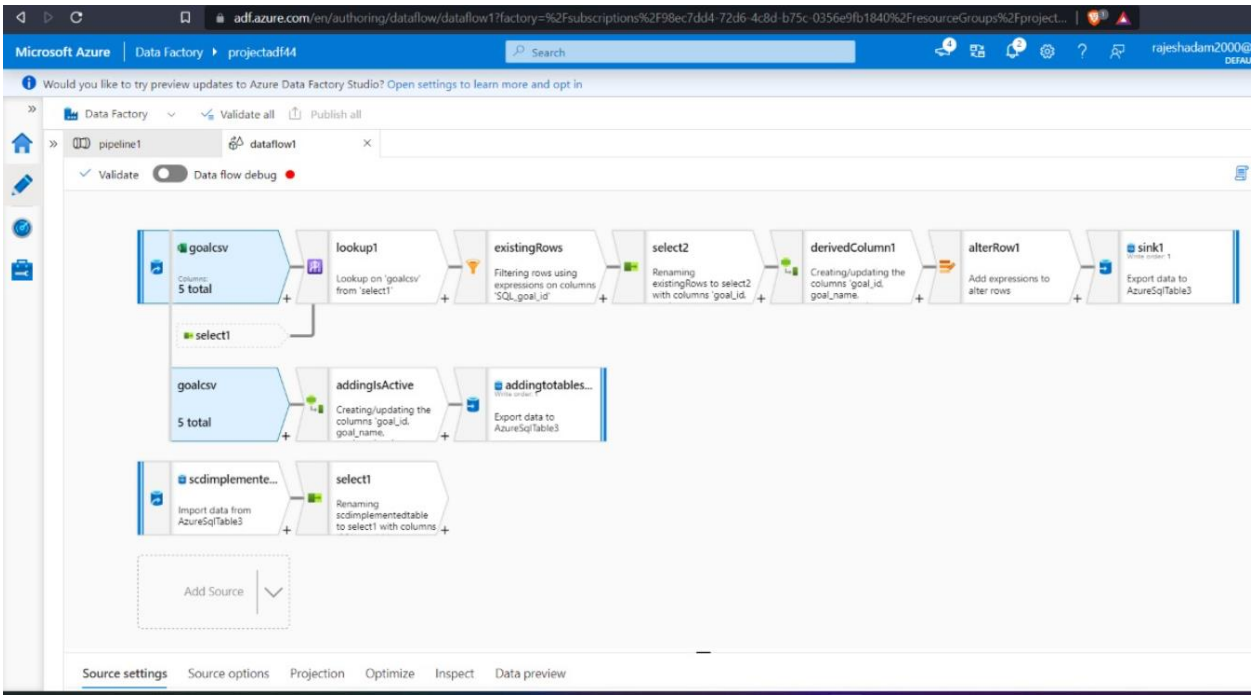

Microsoft Azure
Data Factory
fdproject
Copy Data tool
Properties
Source
Target
Settings
Review and finish
Settings
Enter name and description for the copy data task, more options for data movement
Task name *
pipeline_final
Task description
Data consistency verification
Fault tolerance
Enable logging
Enable staging
Advanced
Data integration unit
Auto
Edit
You will be charged # of used DIUs * copy duration * $0.25/DIU-hour. Local currency and separate
Previous
Next


Copy Data tool
Properties
Source
Target
Settings
Review and finish
Review
Deployment
Summary
You are running pipeline to copy data from Azure SQL Database to Azure SQL Database.
Azure SQL Database
Azure SQL Database
Properties
Edit
Task name
pipeline_final
Task description
Source
Edit
Connection name
AzureSqlDatabase2
Dataset name
SourceDataset_5xt
Number of tables
5
Target
Edit
Previous
Next
Cancel


Azure SQL Database
Azure SQL Database
Deployment complete
Deployment step
Status
Validating copy runtime environment
Succeeded
Creating datasets
Succeeded
Creating pipelines
Succeeded
Running pipelines
Succeeded
Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.
Finish
Edit pipeline
Monitor


Dashboards
Runs
Pipeline runs
Trigger runs
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
All pipeline runs > pipeline_final - Activity runs
pipeline_final
List
Gantt
Rerun
Rerun from activity ForEach_5xt
Rerun from failed activity
Refresh
Update pipeline
Activity runs
Pipeline run ID 71714905-2b35-4742-931c-d484a4961c11
All status
Showing 1 - 6 of 6 items
Activity name
Activity type
Run start
Duration
Status
Error
Log
Copy_5xt
Copy data
Jul 9, 2022, 1:04:02 am
00:00:10
Succeeded
Copy_5xt
Copy data
Jul 9, 2022, 1:04:02 am
00:00:10
Succeeded
Copy_5xt
Copy data
Jul 9, 2022, 1:04:02 am
00:00:10
Succeeded
Copy_5xt
Copy data
Jul 9, 2022, 1:04:02 am
00:00:10
Succeeded
Copy_5xt
Copy data
Jul 9, 2022, 1:04:02 am
00:00:09
Succeeded
ForEach_5xt
ForEach
Jul 9, 2022, 1:04:01 am
00:00:13
Succeeded


Microsoft Azure
Search resources, services, and docs (G+/)
Home > projectdb (sqlserverforproj/projectdb)
projectdb (sqlserverforproj/projectdb) | Query editor (preview)
SQL database
Search (Ctrl+/)
Login
New Query
Open query
Feedback
Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)
Power Platform
Power BI
Power Apps
Power Automate
Settings
Compute + storage
Connection strings
Properties
Locks
Data management
Replicas
Query 1
Query 2
Query 3
Query 4
Query 5
Run
Cancel query
Save query
Export data as
Show only Editor
1 SELECT TOP (1000) * FROM [dbo].[DIMstore]
Results
Messages
14 lays banglore 1000 701 14 0
15 adidas delhi 5 3 15 1
2 adidas pune 60 388 26 1
12 boat mumbai 100 434 27 1
14 kurkure hyderabad 1200 701 28 1
16 jbl headphones 1000 29 1


adf.azure.com/en/authoring/dataflow/dataflow1?factory=%2Fsubscriptions%2F98ec7dd4-72d6-4c8d-b75c-0356e9fb1840%2FresourceGroups%2Fproject...
Microsoft Azure
Data Factory
projectadf44
Search
Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in
Data Factory
Validate all
Publish all
pipeline1
dataflow1
Validate
Data flow debug
goalcsv
5 total
lookup1
Lookup on 'goalcsv' from 'select1'
existingRows
Filtering rows using expressions on columns 'SQL_goal_id'
select2
Renaming existingRows to select2 with columns 'goal_id.
derivedColumn1
Creating/updating the columns 'goal_id. goal_name.
alterRow1
Add expressions to alter rows
sink1
Export data to AzureSqlTable3
select1
goalcsv
5 total
addingIsActive
Creating/updating the columns 'goal_id. goal_name.
addingtotables...
Export data to AzureSqlTable3
scdimplemente...
Import data from AzureSqlTable3
select1
Renaming scdimplementedtable to select1 with columns
Add Source
Source settings
Source options
Projection
Optimize
Inspect
Data preview

CHAPTER 7

REFERENCES

CHAPTER 8: REFERENCES

8.1 WEBSITES

- https://en.wikipedia.org/wiki/Cognizant
- https://www.w3schools.com/sql
- https://www.guru99.com/data-warehousing.html
- https://www.udemy.com/

8.2 COURSES

Courses were provided by the company so as to learn about all the mentioned