DATA WEARHOUSEING

BANK DATABASE

Moolchandani, Vanshita
vanshi@bu.edu

INTRODUCTION

In the dynamic and data-driven landscape of the banking industry, effective management and utilization of vast amounts of information are crucial for making informed business decisions. Data warehousing plays a pivotal role in addressing the complex data needs of banks, providing a centralized and organized repository for comprehensive analysis and reporting.

DATA SOURCES

The data encompasses various facets of a financial system and customer relationship management, featuring information from 4,500 distinct accounts. These accounts are characterized by account_id and district_id. Additionally, the dataset encompasses details on over 5,300 clients, each identified by a client_id and accompanied by information such as gender, age, social details, and contact information.

The transaction table within the dataset is extensive, containing a substantial 1,056,320 rows. This table provides transaction-specific data, including trans_id, account_id, transaction amount, transaction type, and date-related information.

Furthermore, the dataset includes a loan table comprising 682 rows. This table offers valuable insights into loans, with details such as loan_id, account_id, loan amount, loan duration, installment payments, and loan status. The comprehensive nature of this data enables a thorough analysis of the financial system and customer interactions within the studied context.

Here is a short overview of datasets:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	client_id	sex	date of bir	age	social	first	middle	last	account_id	phone	email	address_1	address_2	zipcode	district_id
2	C00000001	Female	#########	29	926-93-215	Emma	Avaya	Smith	A00000576	367-171-68	emma.smi	387 Wellin	Unit 1	47246	18
3	C00000002	Male	2/4/1965	54	806-94-572	Noah	Everest	Thompson	A00003818	212-423-77	noah.thom	75 W. Berkshire St.		10040	1
4	C00000003	Female	10/9/1960	59	614-70-910	Olivia	Brooklynn	Johnson	A00000704	212-425-69	olivia.john	36 Second St.		10162	1
5	C00000004	Male	12/1/1976	43	580-20-341	Liam	Irvin	White	A00002378	951-567-89	liam.white	7607 Sunnyslope Stre		49047	5
6	C00000005	Female	7/3/1980	39	536-14-580	Sophia	Danae	Williams	A00002632	428-265-15	sophia.wil	755 Galvin Street		40852	5
7	C00000006	Male	9/22/1939	80	430-17-582	Mason	Javen	Lopez	A00001972	813-629-50	mason.lop	8266 Arnol	Unit 6	44987	12
8	C00000007	Male	1/25/1949	71	305-80-425	Jacob	Khai	Lee	A00001539	836-845-81	jacob.lee7	827 N. Green Lake Cou		44359	15
9	C00000008	Female	2/21/1958	61	425-96-635	Ava	Eliora	Brown	A00000793	413-444-92	ava.brown	817 Joy Ridge Dr.		1101	51
10	C00000009	Male	#########	64	832-31-725	William	Marek	Gonzalez	A00002484	781-995-61	william.ma	68 Carriage Road		1901	60
11	C00000010	Male	5/1/1963	56	295-22-612	Ethan	Veer	Harris	A00001695	508-902-55	ethan.harr	754 Grandrose St.		2740	57
12	C00000011	Female	8/22/1970	49	666-36-299	Isabella	Haidyn	Jones	A00001726	508-688-64	isabella.jo	873 Randall Mill Dr.		2741	57
13	C00000012	Male	2/20/2001	18	771-44-399	James	Amauri	Clark	A00002881	431-869-38	james.clar	275 Fairground Drive		42009	40
14	C00000013	Female	5/29/1994	25	618-33-880	Mia	Hawa	Miller	A00002357	840-168-69	mia.miller	7429 Lindon Drive		47052	54

	loan_id	account_id	amount	duration	payments	status	fulldate	loan_distri	purpose
1	loan_id	account_id	amount	duration	payments	status	fulldate	loan_distri	purpose
2	L00005657	A00003354	4980	12	415	A	7/5/2014	59	car
3	L00006234	A00006061	5148	12	429	C	5/28/2018	1	car
4	L00006699	A00008330	7656	24	319	A	3/30/2014	1	car
5	L00006688	A00008268	8616	24	359	C	9/26/2017	1	car
6	L00006312	A00006453	10944	36	304	C	5/17/2018	1	car
7	L00006019	A00005033	11400	12	950	A	7/25/2016	1	car
8	L00006229	A00006040	11736	24	489	A	10/2/2014	1	car
9	L00005103	A00000666	12540	12	1045	A	#########	64	debt_conso
10	L00006539	A00007559	12792	12	1066	C	7/11/2018	1	car
11	L00005718	A00003637	14028	12	1169	A	9/1/2017	1	car
12	L00006064	A00005215	14628	12	1219	A	8/4/2014	1	car

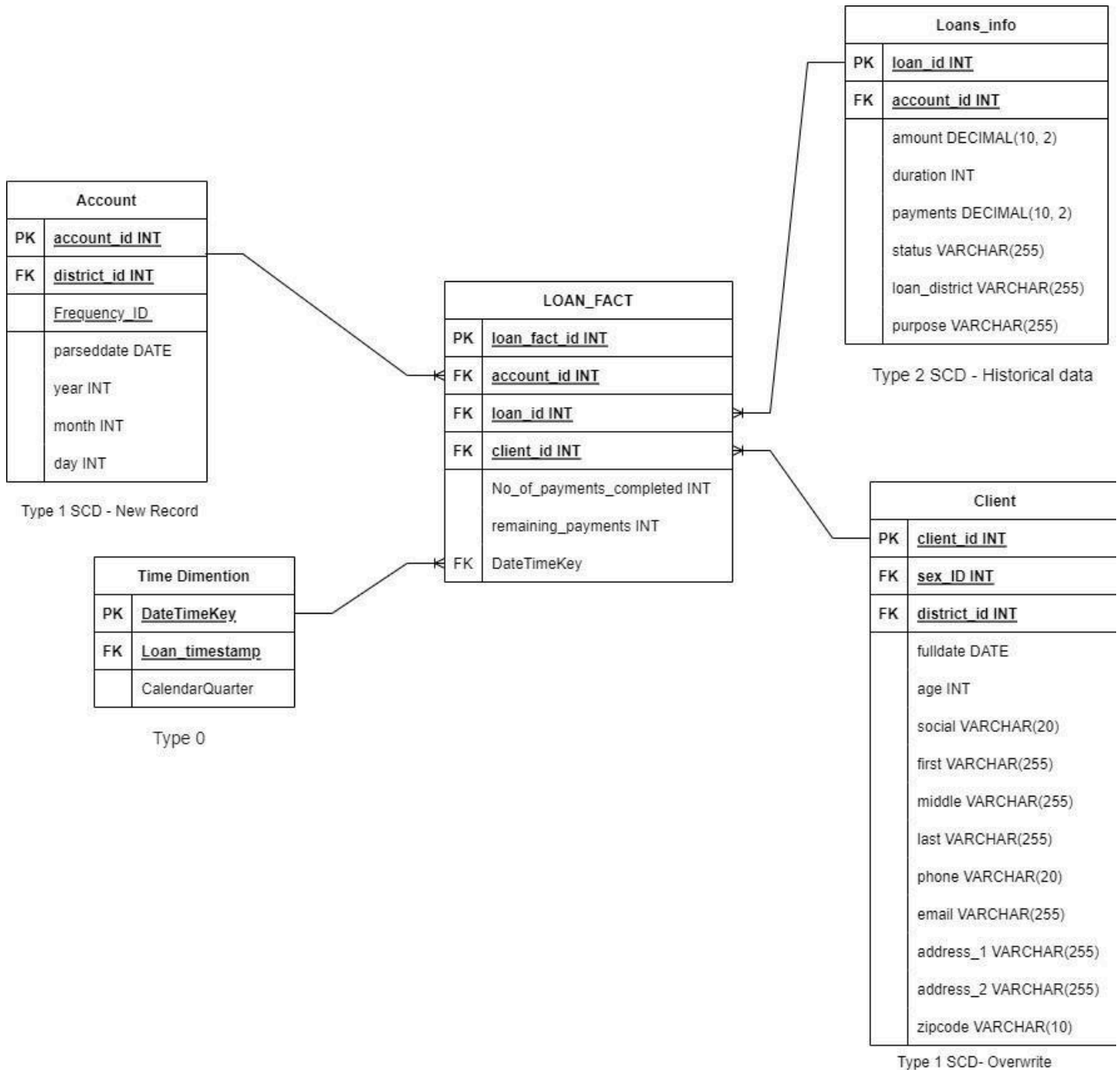|	trans_id	source_acc	trans_type	trans_oper	amount	balance	bank	trans_amo	year	month	day	fulldate	fulltime	fulldatewithtime
1	trans_id	source_acc	trans_type	trans_oper	amount	balance	bank	trans_amo	year	month	day	fulldate	fulltime	fulldatewithtime
2	T00695247	A00002378	TT01	TO01	700	700			2013	1	1	1/1/2013	11:02:40	2013-01-01T11:02:40
3	T00171812	A00000576	TT01	TO01	900	900			2013	1	1	1/1/2013	8:23:33	2013-01-01T08:23:33
4	T00207264	A00000704	TT01	TO01	1000	1000			2013	1	1	1/1/2013	15:19:21	2013-01-01T15:19:21
5	T01117247	A00003818	TT01	TO01	600	600			2013	1	1	1/1/2013	10:42:35	2013-01-01T10:42:35
6	T00579373	A00001972	TT01	TO01	400	400			2013	1	2	1/2/2013	11:39:25	2013-01-02T11:39:25
7	T00771035	A00002632	TT01	TO01	1100	1100			2013	1	2	1/2/2013	15:15:12	2013-01-02T15:15:12
8	T00452728	A00001539	TT01	TO01	600	600			2013	1	3	1/3/2013	9:34:29	2013-01-03T09:34:29
9	T00725751	A00002484	TT01	TO01	1100	1100			2013	1	3	1/3/2013	9:36:35	2013-01-03T09:36:35
10	T00497211	A00001695	TT01	TO01	200	200			2013	1	3	1/3/2013	15:22:23	2013-01-03T15:22:23
11	T00232960	A00000793	TT01	TO01	800	800			2013	1	3	1/3/2013	15:08:16	2013-01-03T15:08:16
12	T00505240	A00001726	TT01	TO01	1000	1000			2013	1	3	1/3/2013	13:54:49	2013-01-03T13:54:49
13	T00144541	A00000485	TT01	TO01	300	300			2013	1	4	1/4/2013	16:40:53	2013-01-04T16:40:53
14	T00637741	A00002177	TT01	TO01	800	800			2013	1	4	1/4/2013	10:30:59	2013-01-04T10:30:59
15	T00689827	A00002357	TT01	TO01	800	800			2013	1	4	1/4/2013	16:29:16	2013-01-04T16:29:16

Architectural Design
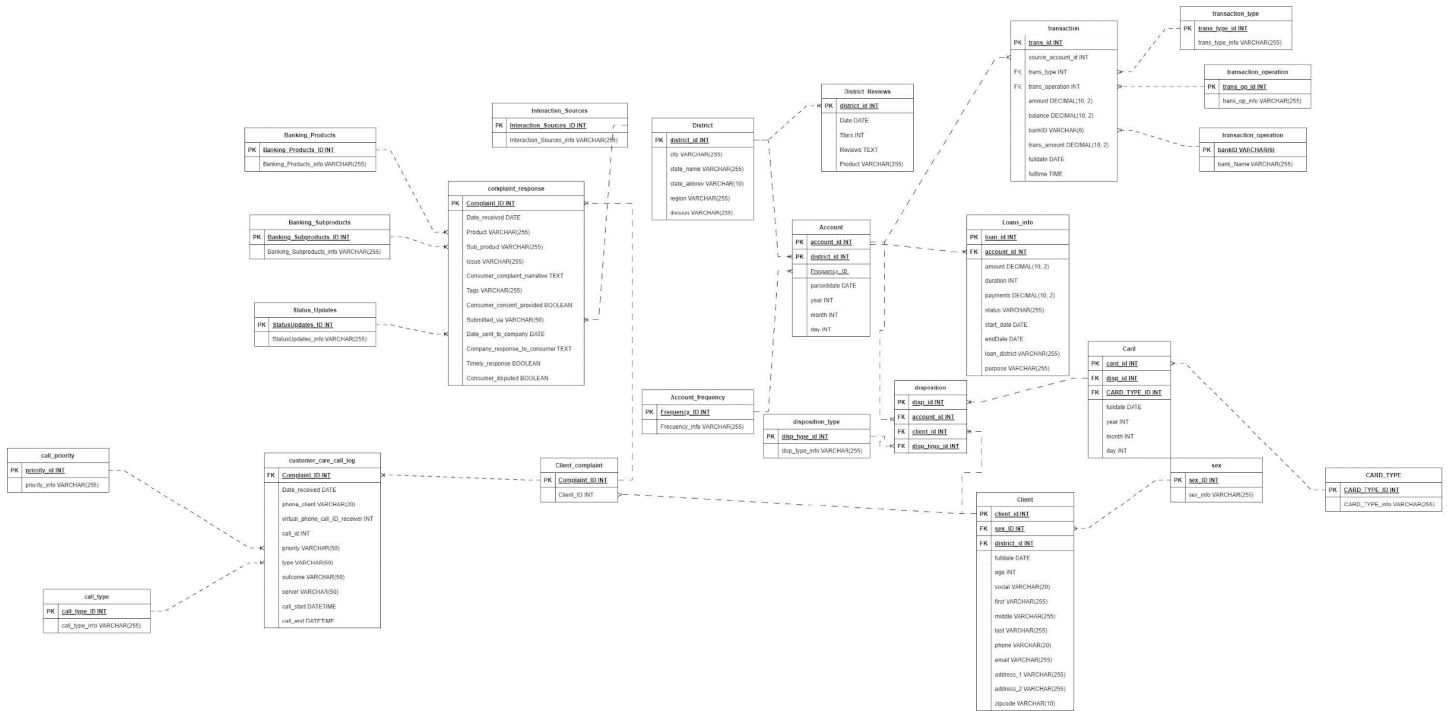
Updated ERD with transaction table normalized, based on Bank details
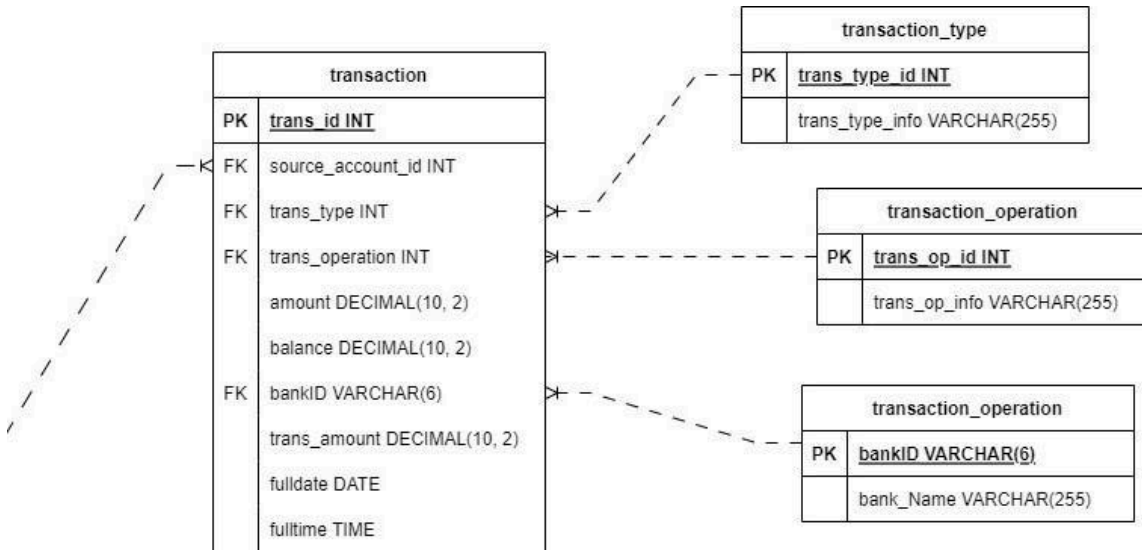
Slow Changing Diagram:

For the delta report in this project, I have specifically chosen one out of the three fact tables to concentrate on.

Loans_info

PK	loan_id INT
FK	account_id INT
	amount DECIMAL(10, 2)
	duration INT
	payments DECIMAL(10, 2)
	status VARCHAR(255)
	loan_district VARCHAR(255)
	purpose VARCHAR(255)

Type 2 SCD - Historical data

Account

PK	account_id INT
FK	district_id INT
	Frequency_ID
	parseddate DATE
	year INT
	month INT
	day INT

Type 1 SCD - New Record

LOAN_FACT

PK	loan_fact_id INT
FK	account_id INT
FK	loan_id INT
FK	client_id INT
	No_of_payments_completed INT
	remaining_payments INT
FK	DateTimeKey

Client

PK	client_id INT
FK	sex_ID INT
FK	district_id INT
	fulldate DATE
	age INT
	social VARCHAR(20)
	first VARCHAR(255)
	middle VARCHAR(255)
	last VARCHAR(255)
	phone VARCHAR(20)
	email VARCHAR(255)
	address_1 VARCHAR(255)
	address_2 VARCHAR(255)
	zipcode VARCHAR(10)

Type 1 SCD- Overwrite

Time Dimention

PK	DateTimeKey
FK	Loan_timestamp
	CalendarQuarter

Type 0

Here we have considered new attributes according to changes made after loading SCD Type 0 Type 1 and Type2.

SQL QUERIES

1. To normalize the data
   EXAMPLE:

```sql
SELECT * FROM Client;

SELECT DISTINCT sex FROM Client;

CREATE TABLE sex
(
        sex_ID NNNVARCHAR(3) PRIMARY KEY,
        sex_info NNNVARCHAR(20)
);

INSERT INTO sex VALUES
('S01','Male'),
('S02','Female');


UPDATE Client
SET sex_ID = 'S01'
WHERE sex_ID = 'Male';

UPDATE Client
SET sex_ID = 'S02'
WHERE sex_ID = 'Female';


ALTER TABLE Client
ALTER COLUMN sex_ID NNNVARCHAR(3);

ALTER TABLE Client
ADD CONSTRAINT sex_ID_FK
FOREIGN KEY (sex_ID)
REFERENCES Sex(sex_ID);

SELECT DISTINCT sex_ID FROM Client;
```

Normalized Transaction table

```
Select * from bank

Select * from [transaction]
```

68 %

Results   Messages

	ID	bank_name
4	4	JPMorgan Chase
5	5	U.S. Bancorp
6	6	Charles Schwab Corp.
7	7	Goldman Sachs
8	8	Bank of America
9	9	PNC Financial Services
10	10	TD Bank
11	11	State Street Corp.

	trans_id	source_account_id	trans_type	trans_operation	amount	balance	bank	trans_amount	fulldate	fulltime	BANK_ID
13	T00637741	A00002177	TT01	TO01	800	800	NULL	NULL	2019-01-04	10:30:59	NULL
14	T00689827	A00002357	TT01	TO01	800	800	NULL	NULL	2019-01-04	16:29:16	NULL
15	T00846006	A00002881	TT01	TO01	700	700	NULL	NULL	2019-01-04	9:29:01	NULL
16	T00637742	A00002177	TT01	TO02	5123	5923	Stat...	62457513	2019-01-05	16:44:42	11
17	T02908688	A00009635	TT01	TO01	400	400	NULL	NULL	2019-01-05	13:32:11	NULL
18	T00232961	A00000793	TT01	TO02	3401	4201	Gol...	6149286	2019-01-05	11:56:11	7
19	T00192096	A00000652	TT01	TO01	700	700	NULL	NULL	2019-01-05	9:28:49	NULL
20	T00542215	A00001844	TT01	TO01	500	500	NULL	NULL	2019-01-06	14:38:54	NULL

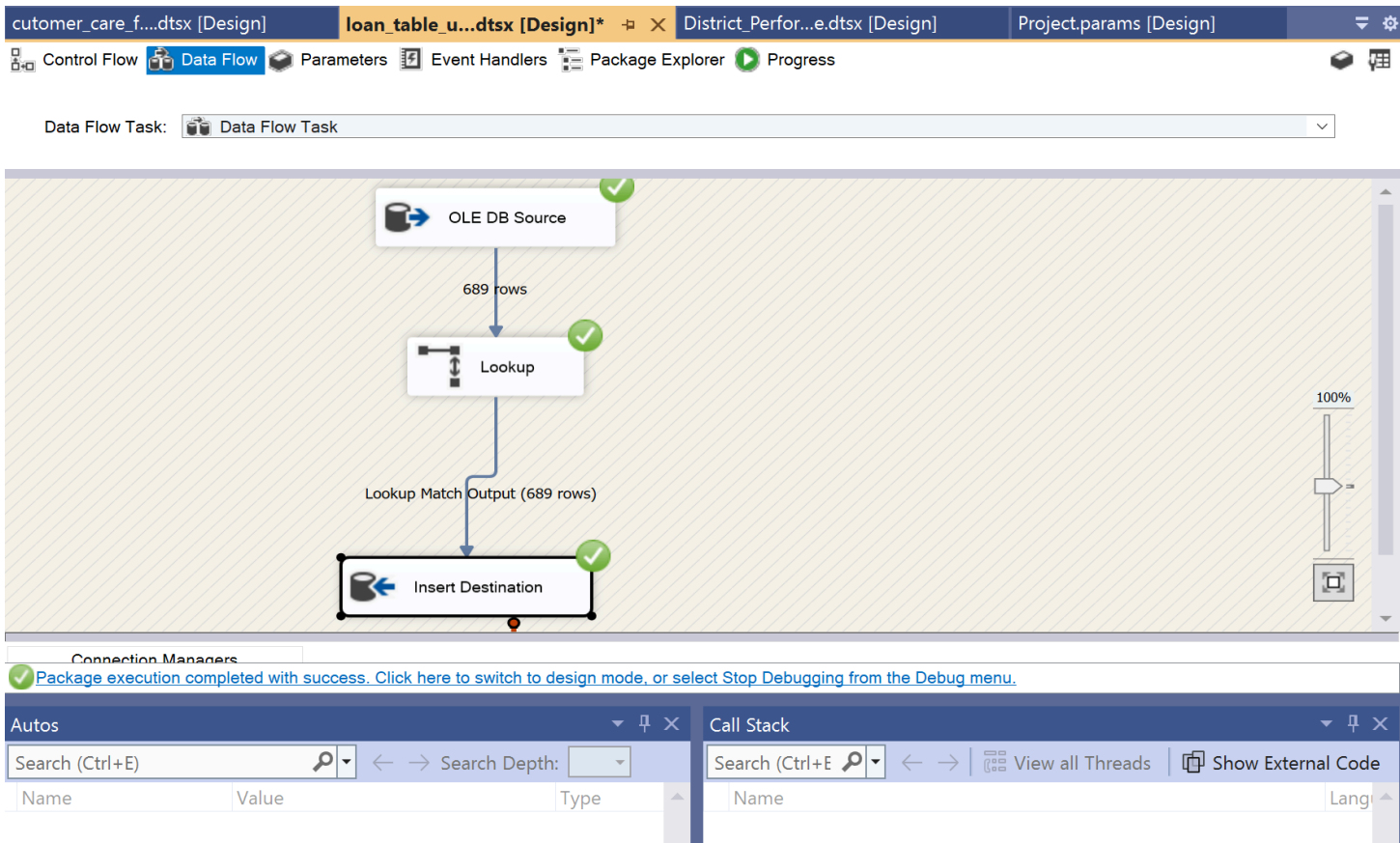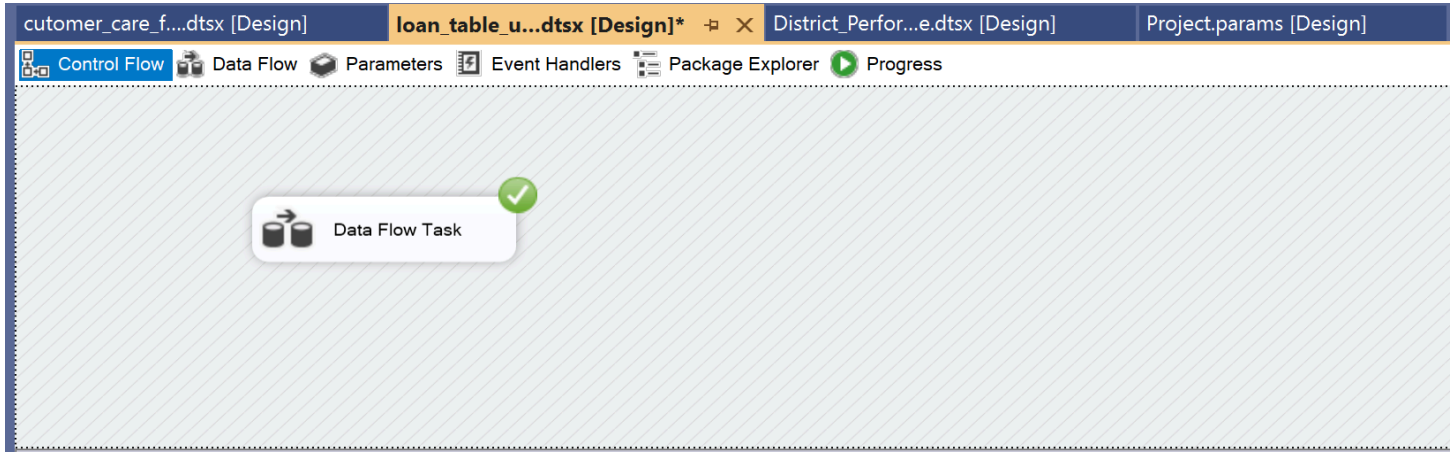Query executed successfully.   localhost (16.0 RTM)   LAPTOP-RBLOJAJG\vansh ...   project_CS7779_bank_da...   00:00:06   734,057 rows

Normalization of Loans_Info using SSIS

```sql
SELECT * FROM Loan_Account
```

90 %

Results | Messages

	Loan_Acc_key	Loan_id	account_id
1	1	L00004959	A00000002
2	2	L00004961	A00000019
3	3	L00004962	A00000025
4	4	L00004967	A00000037
5	5	L00004968	A00000038
6	6	L00004973	A00000067
7	7	L00004986	A00000097
8	8	L00004988	A00000103
9	9	L00004989	A00000105

Fact Tables:

Measures in Loan_Fact

loan_fact_id (Primary Key): A unique identifier for each record in the table.

total_payments: Represents the total payments made towards the loan. This is likely an aggregate function, summarizing the overall payments received.

remaining_payments: Represents the remaining payments yet to be made on the loan. Similar to "total_payments," this is likely an aggregate function indicating the outstanding balance.
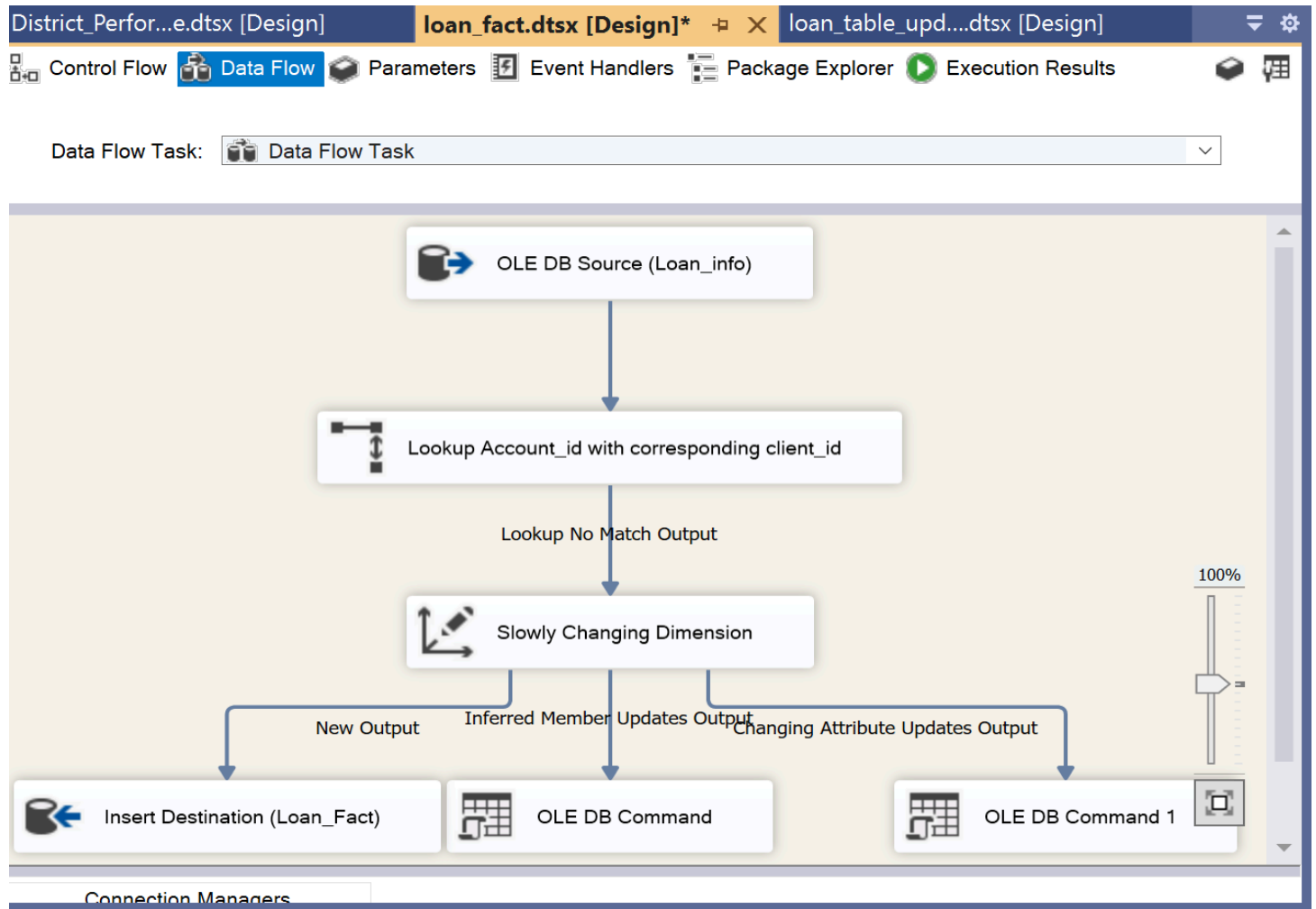
```sql
SELECT TOP 5 * FROM LOAN_FACT
```

90 %

Results | Messages

	loan_fact_id	loan_id	account_id	client_id	No_of_payments_completed	remaining_payments	start_date	DATETIMEKEY
1	1	L00004959	A00000002	C00000002	23	0	2020-01-05	2020-01-05 11:02:40.000
2	2	L00004961	A00000019	C00000025	8	0	2022-04-29	2022-04-29 08:23:33.000
3	3	L00004962	A00000025	C00000031	0	30276	2023-12-08	2023-12-08 15:19:21.000
4	4	L00004967	A00000037	C00000045	2	0	2018-10-14	2018-10-14 10:42:35.000
5	5	L00004968	A00000038	C00000046	20	0	2018-04-19	2018-04-19 11:39:25.000

SSIS (Integration Package)

Control Flow   **Data Flow**   Parameters   Event Handlers   Package Explorer   Execution Results

Data Flow Task: Data Flow Task

OLE DB Source (Loan_info)

Lookup Account_id with corresponding client_id

Lookup No Match Output

Slowly Changing Dimension

New Output   Inferred Member Updates Output   Changing Attribute Updates Output

100%

Insert Destination (Loan_Fact)   OLE DB Command   OLE DB Command 1

Connection Managers

Measures in District_Performance_Fact Table:

District_Performance_id (Primary Key): A unique identifier for each record in the table.

District_ID: Identifies the district for which the performance metrics are recorded.

Total_no_of_clients: Represents the total number of clients associated with the respective district.

Total_no_of_Complaints: Indicates the total number of complaints received by the district.

No_of_Solved_Complaints: Represents the count of complaints that have been successfully resolved.

No_of_Unsolved_Complaints: Represents the count of complaints that remain unresolved.

Avg_Response_Time_For_Solved_Complaints: Captures the average response time for resolving complaints that have been marked as solved. The response time is measured in minutes

```
SELECT TOP 5 * FROM District_Performance_Fact
```

10 %

Results | Messages

	District_Performance_id	District_ID	Total_no_of_clients	Total_no_of_Complaints	No_of_Solved_Complaints	No_of_Unsolved_Complaints	Avg_Response_Time_For_Solved_Complaints
1	309	23	53	242	41	187	1
2	310	46	74	304	66	224	2
3	311	69	61	264	66	187	1
4	312	29	51	231	49	172	2
5	313	75	55	262	50	199	1

SSIS (Integration Package)

Execute SQL Task Editor

Configure the properties required to run SQL statements and stored procedures using the selected connection.

General
Parameter Mapping
Result Set
Expressions

∨ General	
Name	**Execute SQL Task 1**
Description	**Execute SQL Task**
∨ **Options**	
TimeOut	**0**
CodePage	**1252**
TypeConversionMode	**Allowed**
∨ **Result Set**	
ResultSet	**None**
∨ **SQL Statement**	
ConnectionType	**OLE DB**
Connection	**LocalHost.project_CS7779_bank_database_D**
SQLSourceType	**Direct input**
SQLStatement	**exec PopulateDistrictPerformance**
IsQueryStoredProcedure	False
BypassPrepare	**True**

Name
Specifies the name of the task.

Browse...   Build Query...   Parse Query

OK   Cancel   Help

Measures in Customer_care_Fact Table:

Customer_care_Fact_id (Primary Key): A unique identifier for each record in the table.

Complaint_id: Identifies the complaint associated with a particular record.

call_time: Captures the time when the call related to the complaint was made.

```sql
SELECT TOP 5 * FROM Customer_care_Fact
```

90 %

Results | Messages

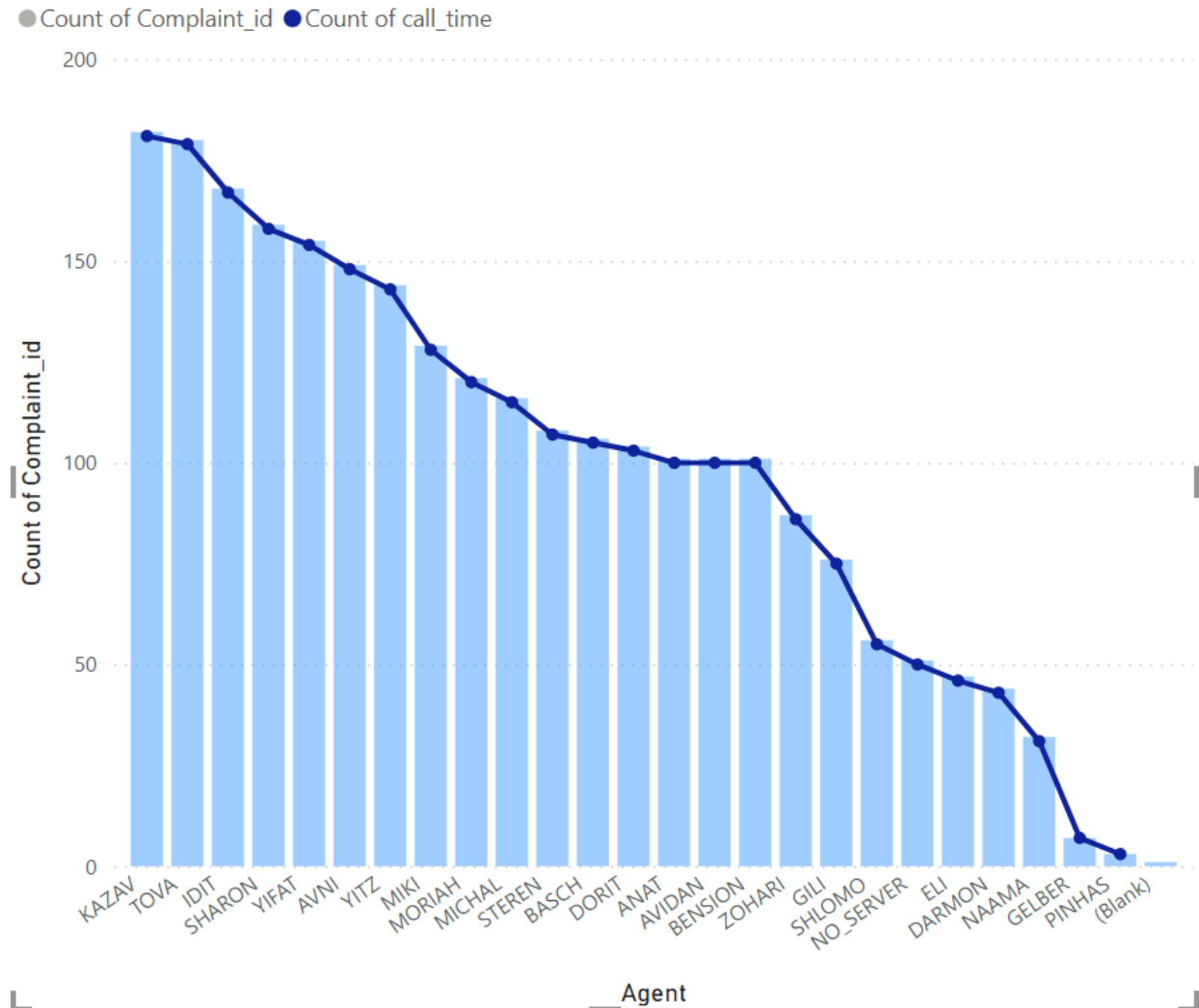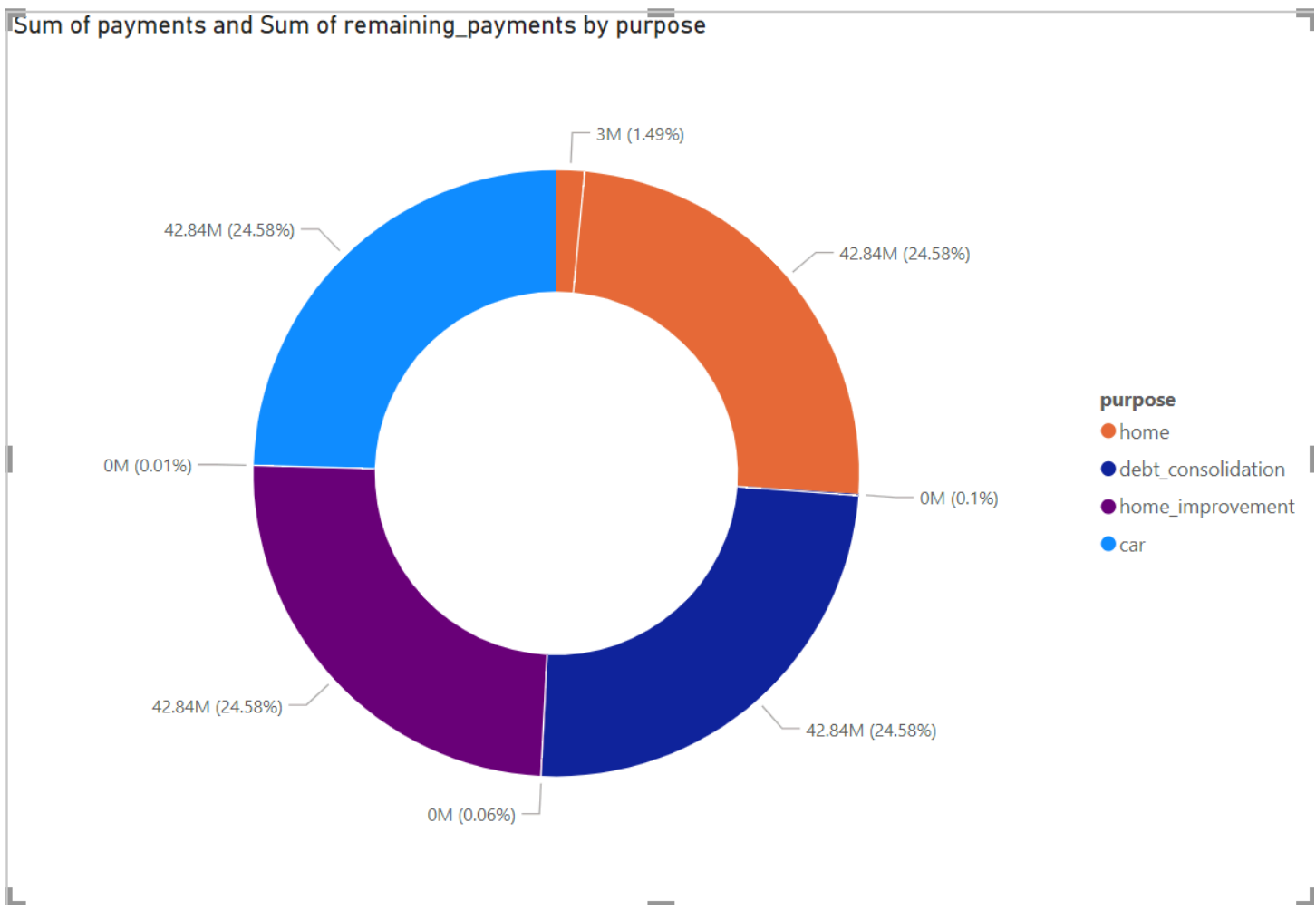|	Customer_care_Fact_id	Complaint_id	date_recieved	Agent	call_time
1	1	CR2406263	2023-05-27	MICHAL	00:06:12.0000000
2	2	CR2405641	2023-03-27	TOVA	00:17:48.0000000
3	3	CR2405629	2023-03-27	YIFAT	00:12:37.0000000
4	4	CR2400594	2023-03-23	AVNI	00:01:19.0000000
5	5	CR2399607	2023-03-22	STEREN	00:06:57.0000000

SSIS (Integration Package):
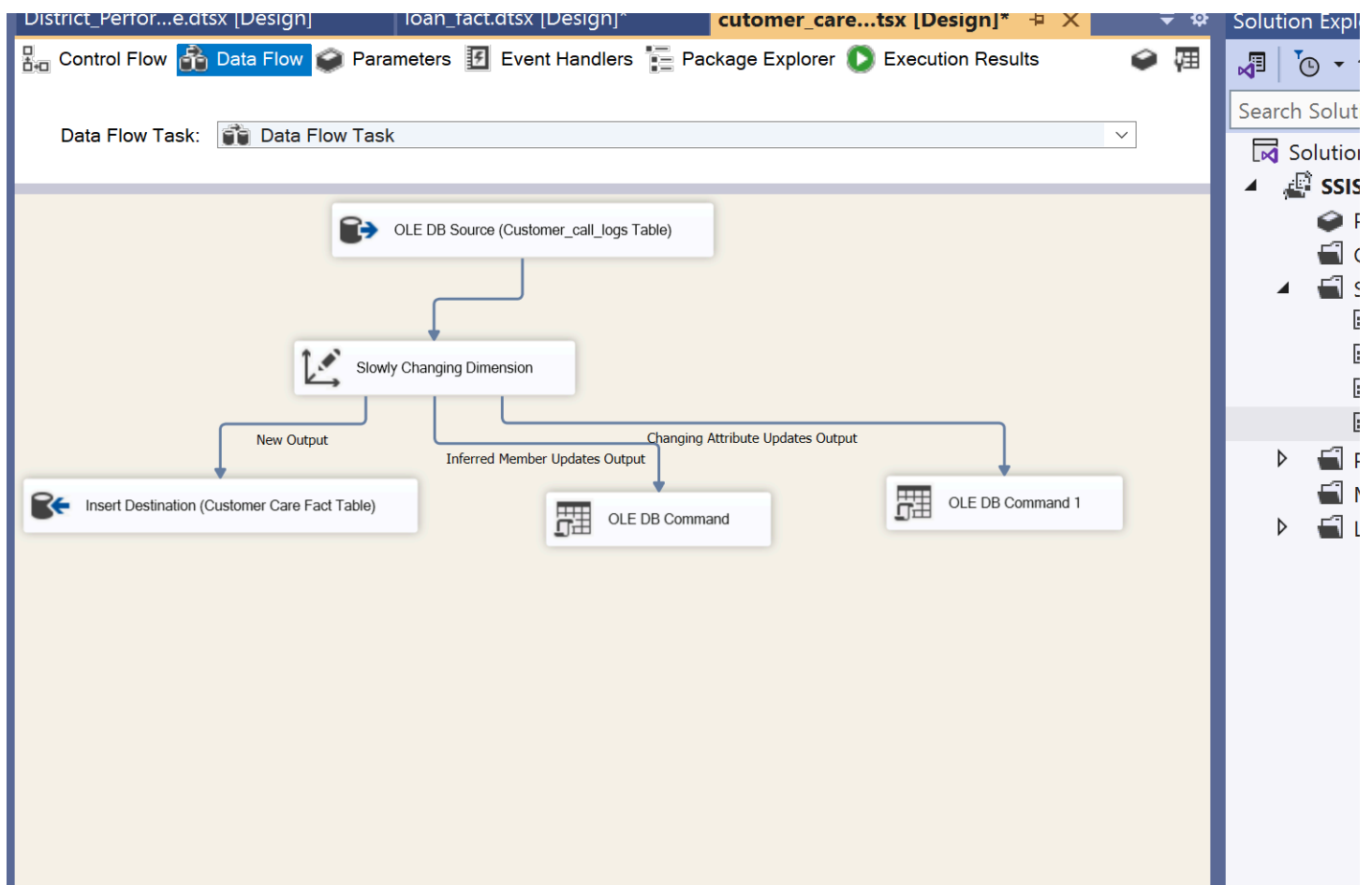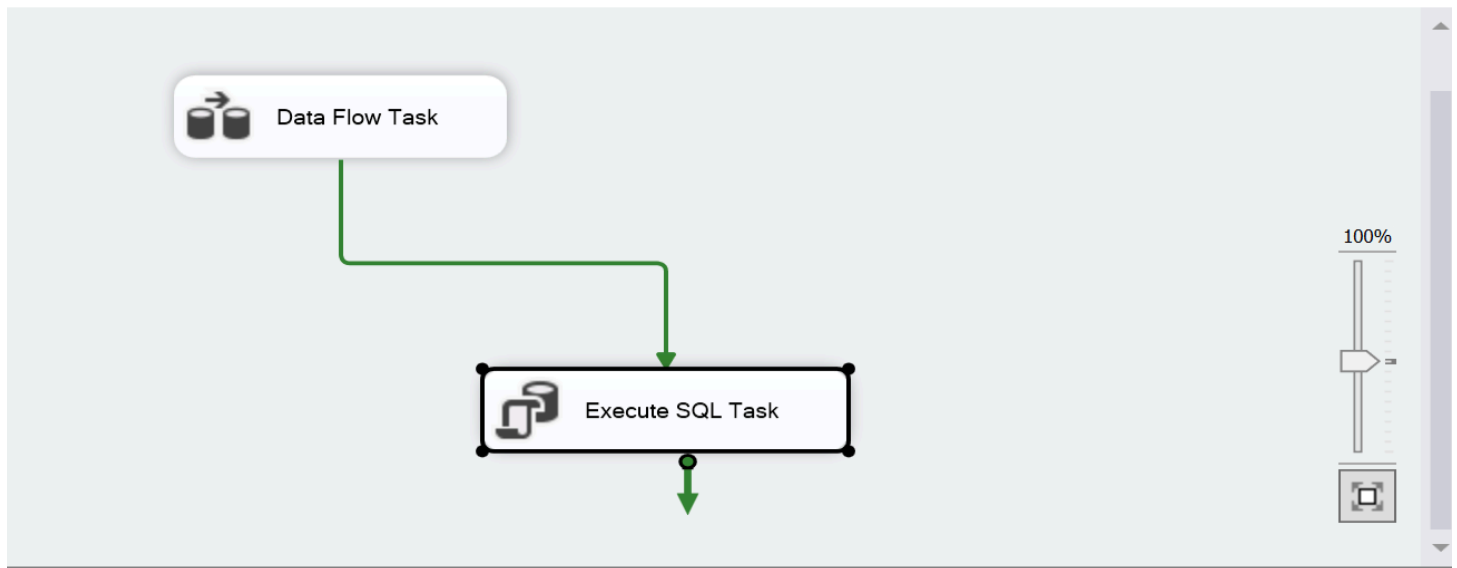
ANALYSIS

BUSINESS CASE 1: Check agent performance based on the number of call they answered and time to resolve the issue.

Count of Complaint_id and Count of call_time by Agent

BUSINESS CASE 2: LOANS ISSUED BY CLIENTS FOR DIFFERENT PURPOSE WITH PAYMENTS DUE.

Sum of payments and Sum of remaining_payments by purpose

3M (1.49%)

42.84M (24.58%)

42.84M (24.58%)

0M (0.01%)

0M (0.1%)

42.84M (24.58%)

42.84M (24.58%)

0M (0.06%)

purpose
- home
- debt_consolidation
- home_improvement
- car

BUSINESS CASE 3 : Number of transaction based upon the type of tranasction

Count of trans_operation by trans_op_info