

Part 1- Final Project Report

+ Data Staging for New York (Talend ETL Job):

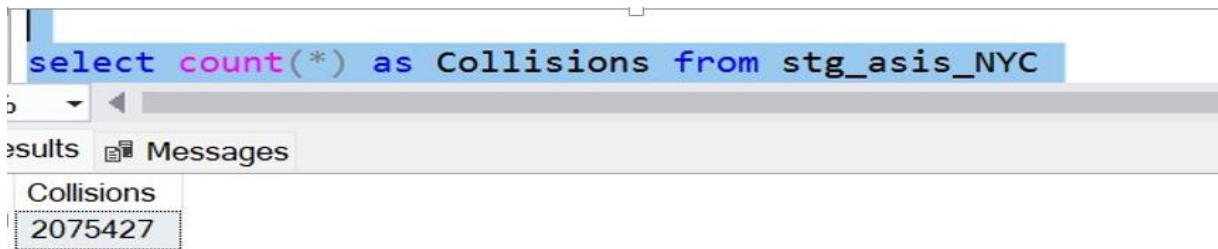
1. Job NYC_MVC_Staging-



The screenshot shows the Talend ETL Job Designer for the job 'Job NYC_MVC_Staging'. The job flow starts with a 'NYC_Collision' connector, followed by a 'tMap_1' component, and then an 'MSSQL_server' connector. The 'tMap_1' component is configured with two output streams: 'row1 (Main)' and 'row2 (Reject order:2)'. The 'row1 (Main)' stream is mapped to the 'MSSQL_server' connector, and the 'row2 (Reject order:2)' stream is mapped to a 'tLogRow_1' component. The job execution log shows the following details:

```
Starting job NYC_MVC_Staging at 16:38 07/04/2024.
[statistics] connecting to socket on port 3883
[statistics] connected
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| CRASH_DATE | CRASH_TIME | BOROUGH | ZIP_CODE | LATITUDE | LONGITUDE | LOCATION | ON_STREET_NAME | CROSS_STREET_NAME | OFF_STREET_NAME | NU |
|-----+-----+-----+-----+-----+-----+-----+-----+-----+
[statistics] disconnected
Job NYC_MVC_Staging ended at 16:40 07/04/2024. [Exit code = 0]
```

COUNT:-



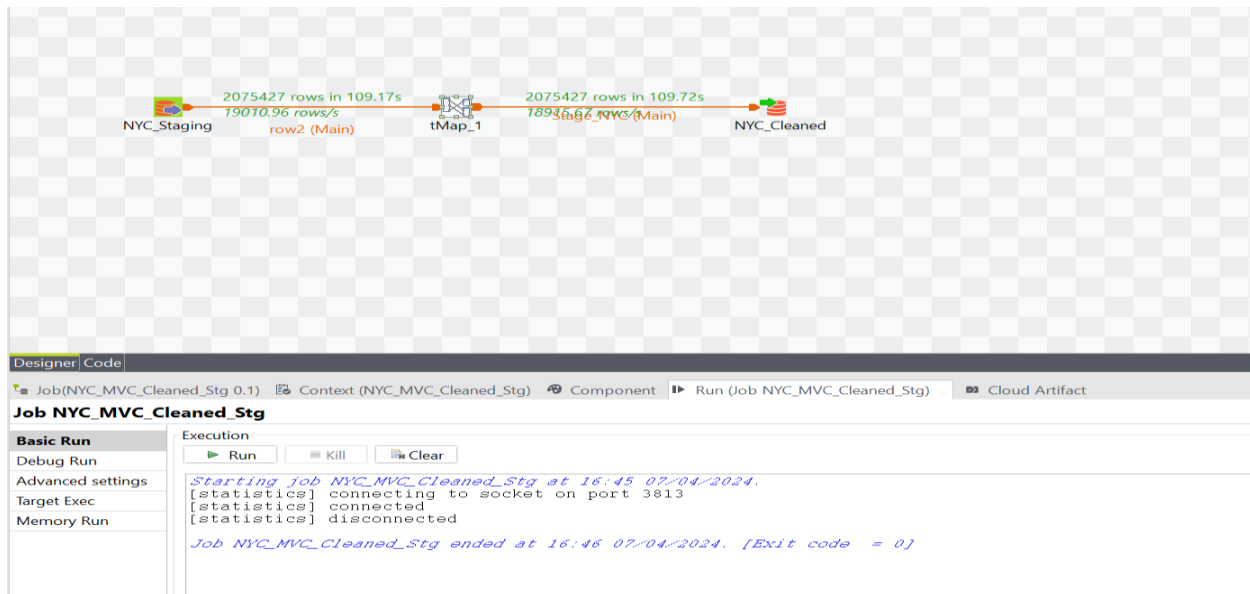
The screenshot shows a SQL query and its results. The query is:

```
select count(*) as Collisions from stg_asis_NYC
```

The results show a single row with the value 2075427.

Collisions
2075427

2. Job NYC_MVC_Cleaned_Stg-



The screenshot shows the Talend ETL Job Designer for the job 'Job NYC_MVC_Cleaned_Stg'. The job flow starts with a 'NYC_Staging' connector, followed by a 'tMap_1' component, and then an 'NYC_Cleaned' connector. The 'tMap_1' component is configured with two output streams: 'row1 (Main)' and 'row2 (Main)'. The 'row1 (Main)' stream is mapped to the 'NYC_Cleaned' connector, and the 'row2 (Main)' stream is mapped to the 'NYC_Staging' connector. The job execution log shows the following details:

```
Starting job NYC_MVC_Cleaned_Stg at 16:45 07/04/2024.
[statistics] connecting to socket on port 3813
[statistics] connected
[statistics] disconnected
Job NYC_MVC_Cleaned_Stg ended at 16:46 07/04/2024. [Exit code = 0]
```

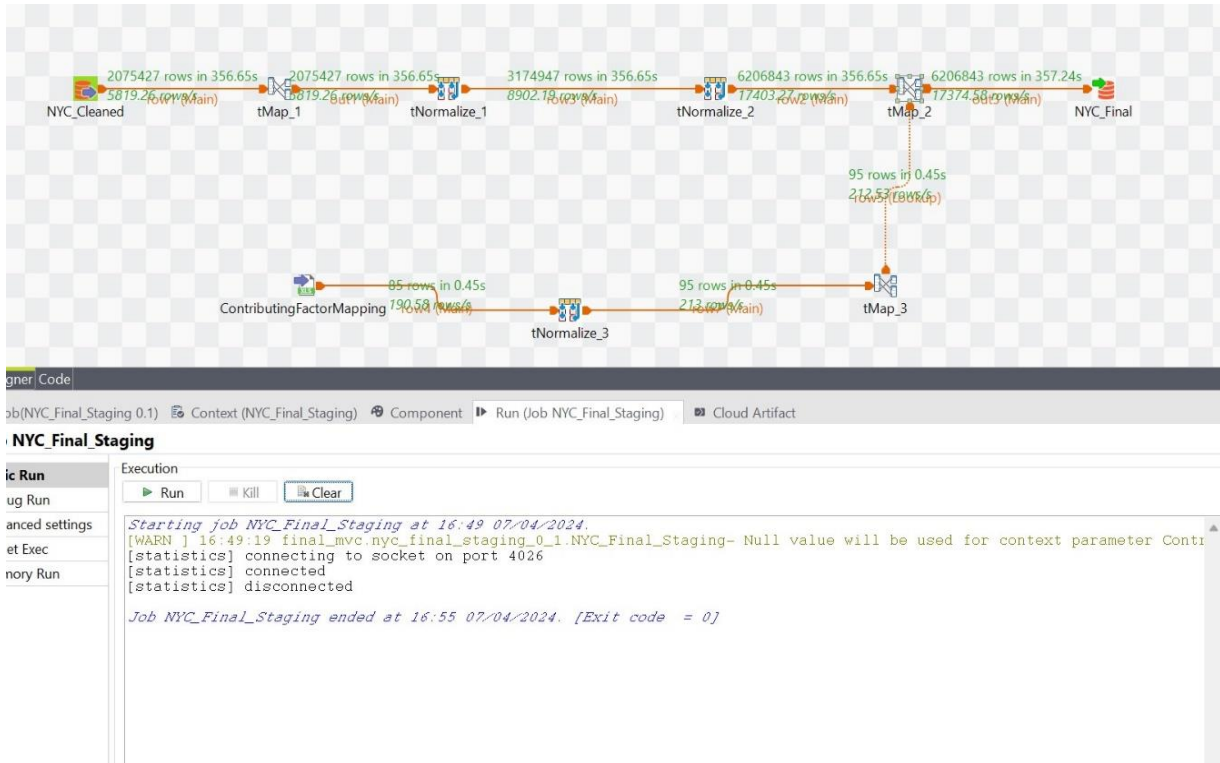
COUNT:-

```
select count(*) as Collisions from stg_Cleaned_NYC
```

Results Messages

Collisions
2075427

3. Job NYC_Final_Staging-



COUNT:-

```
select count(*) as Collisions from stg_Final_NYC
```

Results Messages

Collisions
6206843

❖ Stage Table Loading for NYC:-

Select TOP 1000 * from stg_asis_NYC;
 Select TOP 1000 * from stg_Cleaned_NYC;
 Select TOP 1000 * from stg_Final_NYC;

```
select TOP 1000 * from stg_asis_NYC
select TOP 1000 * from stg_Cleaned_NYC
select TOP 1000 * from stg_Final_NYC
```

100 %

Results Messages

	COLLISION_ID	CRASH_DATE	CRASH_TIME	BOROUGH	ZIP_CODE	LATITUDE	LONGITUDE	LOCATION	ON_STREET_NAME	CROSS_STREET_NAME	OFF_STREET_NAME	NUMBER_OF_
1	3978006	2018-09-08 00:00:00.000	14:21	BROOKLYN	11218	40.63578	-73.97597	(40.63578, -73.97597)	DITMAS AVENUE	EAST 3 STREET		0
2	3972258	2018-08-30 00:00:00.000	16:05	BROOKLYN	11205	40.69593	-73.95269	(40.695927, -73.95269)			128 NOSTRAND AVENUE	0
3	3977841	2018-09-07 00:00:00.000	16:50	MANHATTAN	10026	40.80392	-73.95013	(40.803925, -73.950134)			149 WEST 118 STREET	0
4	3974532	2018-09-05 00:00:00.000	9:05	MANHATTAN	10022	40.75837	-73.971	(40.758373, -73.971)	LEXINGTON AVENUE	EAST 53 STREET		0
5	3973611	2018-09-04 00:00:00.000	9:20			40.71906	-73.8332	(40.719055, -73.8332)	GRAND CENTRAL PKWY			0
6	3980755	2018-09-14 00:00:00.000	13:25			40.85208	-73.82685	(40.85208, -73.82685)	BRUCKNER EXPRESSWAY			1
7	3977877	2018-09-07 00:00:00.000	9:30	QUEENS	11421	40.69366	-73.85217	(40.69366, -73.85217)	WOODHAVEN BOULEVARD	JAMAICA AVENUE		5
8	3981396	2018-09-13 00:00:00.000	11:00	QUEENS	11372	40.75063	-73.89207	(40.750633, -73.89207)			35-46 74 STREET	0

	COLLISION_ID	CRASH_DATE	CRASH_TIME	BOROUGH	ZIP_CODE	LATITUDE	LONGITUDE	LOCATION	ON_STREET_NAME	CROSS_STREET_NAME	OFF_STREET_NAME	NUMBER_OF_PER
1	3978006	2018-09-08 00:00:00.000	14:21:00	BROOKLYN	11218	40.63578	-73.97597	(40.63578, -73.97597)	DITMAS AVENUE	EAST 3 STREET	NA	0
2	3972258	2018-08-30 00:00:00.000	16:05:00	BROOKLYN	11205	40.69593	-73.95269	(40.695927, -73.95269)	NA	NA	128 NOSTRAND AVENUE	0
3	3977841	2018-09-07 00:00:00.000	16:50:00	MANHATTAN	10026	40.80392	-73.95013	(40.803925, -73.950134)	NA	NA	149 WEST 118 STREET	0
4	3974532	2018-09-05 00:00:00.000	9:05:00	MANHATTAN	10022	40.75837	-73.971	(40.758373, -73.971)	LEXINGTON AVENUE	EAST 53 STREET	NA	0
5	3973611	2018-09-04 00:00:00.000	9:20:00	NA	-1	40.71906	-73.8332	(40.719055, -73.8332)	GRAND CENTRAL PKWY	NA	NA	0
6	3980755	2018-09-14 00:00:00.000	13:25:00	NA	-1	40.85208	-73.82685	(40.85208, -73.82685)	BRUCKNER EXPRESSWAY	NA	NA	1
7	3977877	2018-09-07 00:00:00.000	9:30:00	QUEENS	11421	40.69366	-73.85217	(40.69366, -73.85217)	WOODHAVEN BOULEVA...	JAMAICA AVENUE	NA	5
8	3981396	2018-09-13 00:00:00.000	11:00:00	QUEENS	11372	40.75063	-73.89207	(40.750633, -73.89207)	NA	NA	35-46 74 STREET	0

	COLLISION_ID	CRASH_DATE	CRASH_TIME	BOROUGH	ZIP_CODE	LATITUDE	LONGITUDE	LOCATION	ON_STREET_NAME	CROSS_STREET_NAME	OFF_STREET_NAME	IS PEDESTRIAN	NUMBER_OF_P
1	3632807	2017-03-15 00:00:00.000	18:04:00	MANHATTAN	10065	40.76802	-73.97029	(40.768024, -73.97029)	5 AVENUE	EAST 65 STREET	NA	N	0
2	3632807	2017-03-15 00:00:00.000	18:04:00	MANHATTAN	10065	40.76802	-73.97029	(40.768024, -73.97029)	5 AVENUE	EAST 65 STREET	NA	N	0
3	3632807	2017-03-15 00:00:00.000	18:04:00	MANHATTAN	10065	40.76802	-73.97029	(40.768024, -73.97029)	5 AVENUE	EAST 65 STREET	NA	N	0
4	3632807	2017-03-15 00:00:00.000	18:04:00	MANHATTAN	10065	40.76802	-73.97029	(40.768024, -73.97029)	5 AVENUE	EAST 65 STREET	NA	N	0
5	3630739	2017-03-12 00:00:00.000	18:00:00	NA	-1	40.63023	-74.14209	(40.63023, -74.14209)	NA	NA	42 WALKER ST...	Y	1
6	3636707	2017-03-21 00:00:00.000	10:30:00	MANHATTAN	10018	40.75209	-73.98352	(40.75209, -73.98352)	NA	NA	34 WEST 39 ST...	N	0
7	3630417	2017-03-12 00:00:00.000	11:00:00	NA	-1	40.5854	-74.1685	(40.5854, -74.1685)	NA	NA	2505 RICHMOND...	N	0

Query executed successfully.

localhost (15.0 RTM) APEXyugal (62) Final_MVC 00:00:00 3,000 rows

+ Data Staging for Chicago (Talend ETL Job):-

1. Job Chicago_MVC_Staging-

The diagram shows an ETL job flow: Chicago (Source) → tMap_1 (Transformer) → MSSQL_Server_Chicago (Target). Statistics for Chicago: 817723 rows in 98.77s, 817723 rows in 99.42s, 0 rows in 99.42s. Statistics for tMap_1: 817723 rows in 99.8s, 81793 rows in 99.8s. A red arrow labeled 'row2 (Reject order:2)' points to tLogRow_1.

Designer | Code

Job(Chicago_MVC_staging 0.1) Context (Chicago_MVC_staging) Component Run (Job Chicago_MVC_staging) Cloud Artifact

Job Chicago_MVC_staging

Basic Run | Execution | Debug Run | Advanced settings | Target Exec | Memory Run

Run Kill Clear

Starting job Chicago_MVC_staging at 17:08 07/04/2024.
[statistics] connecting to socket on port 3643
[statistics] connected

CRASH_RECORD_ID	CRASH_DATE_EST_I	CRASH_DATE	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	WEATHER_CONDIT
[statistics] disconnected
Job Chicago_MVC_staging ended at 17:10 07/04/2024. [Exit code = 0]

Count:-

select count(*) as Collisions from stg_asis_Chicago

Results Messages

Collisions
817723

2. Job Chicago_Cleaned_Staging-

The diagram shows an ETL job flow: ASIS_Chicago (Source) → tMap_2 (Transformer) → Cleaned_Chicago (Target). Statistics for ASIS_Chicago: 817723 rows in 53.23s, 1536265 rows in 53.23s. Statistics for tMap_2: 817723 rows in 54.08s, 1512034 rows in 54.08s.

Designer | Code

Job(Chicago_Cleaned_Staging 0.1) Context (Chicago_Cleaned_Staging) Component Run (Job Chicago_Cleaned_Staging) Cloud Artifact

Job Chicago_Cleaned_Staging

Basic Run | Execution | Debug Run | Advanced settings | Target Exec | Memory Run

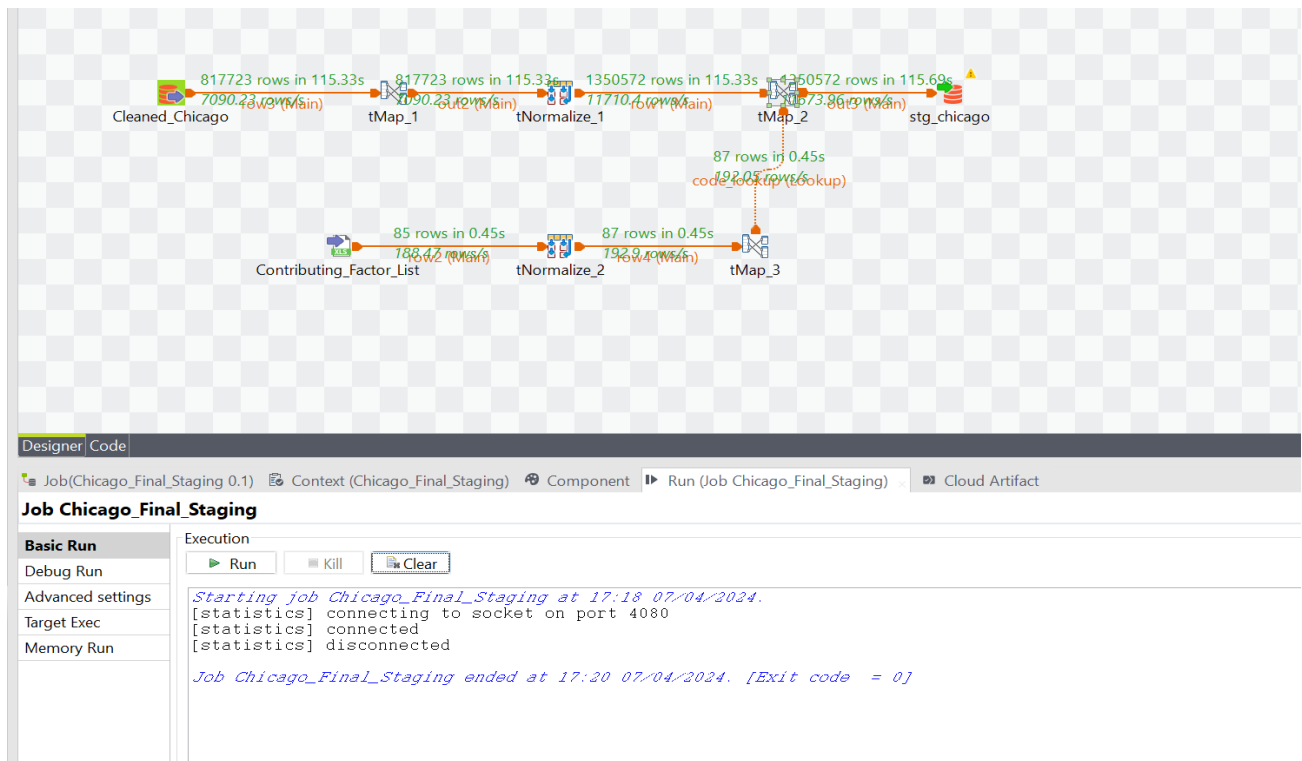
Run Kill Clear

Starting job Chicago_Cleaned_Staging at 17:13 07/04/2024.
[statistics] connecting to socket on port 3388
[statistics] connected
[statistics] disconnected
Job Chicago_Cleaned_Staging ended at 17:14 07/04/2024. [Exit code = 0]

Count:-

<pre>select count(*) as Collisions from stg_Cleaned_Chicago</pre>	
121 %	
Results	Messages
Collisions	
1	817723

3. Job Chicago_Final_Staging-



Count:-

<pre>select count(*) as Collisions from stg_Final_Chicago</pre>	
21 %	
Results	Messages
Collisions	
1	1350572

❖ Stage Table Loading for Chicago:-

```
Select TOP 1000 * from stg_asis_Chicago;
```

```
Select TOP 1000 * from stg_Cleaned_Chicago;
```

```
Select TOP 1000 * from stg_Final_Chicago;
```

<

+ Data Staging for Austin (Talend ETL Job):

1. Job Austin_MVC_Staging-

Designer | Code

Job(Austin_MVC_Staging 0.1) Context (Austin_MVC_Staging) Component Run (Job Austin_MVC_Staging) Cloud Artifact

Job Austin_MVC_Staging

Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run

Execution

Run Kill Clear

Starting job Austin_MVC_Staging at 17:26 07/04/2024.
[statistics] connecting to socket on port 3939
[statistics] connected

crash_id	crash_fatal_flg	crash_date	crash_time	case_id	rpt_latitude	rpt_longitude	rpt_block_num	rpt_street_pfx	rpt_street

[statistics] disconnected
Job Austin_MVC_Staging ended at 17:26 07/04/2024. [Exit code = 0]

2. Job Austin_Final_Staging-

Designer | Code

Job(Austin_Final_Staging 0.1) Context (Austin_Final_Staging) Component Run (Job Austin_Final_Staging) Cloud Artifact

Job Austin_Final_Staging


Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run

Execution

Run Kill Clear

Starting job Austin_Final_Staging at 17:30 07/04/2024.
[statistics] connecting to socket on port 3803
[statistics] connected
[statistics] disconnected
Job Austin_Final_Staging ended at 17:30 07/04/2024. [Exit code = 0]

3. Job Austin_MVC_Cleaned_Staging-

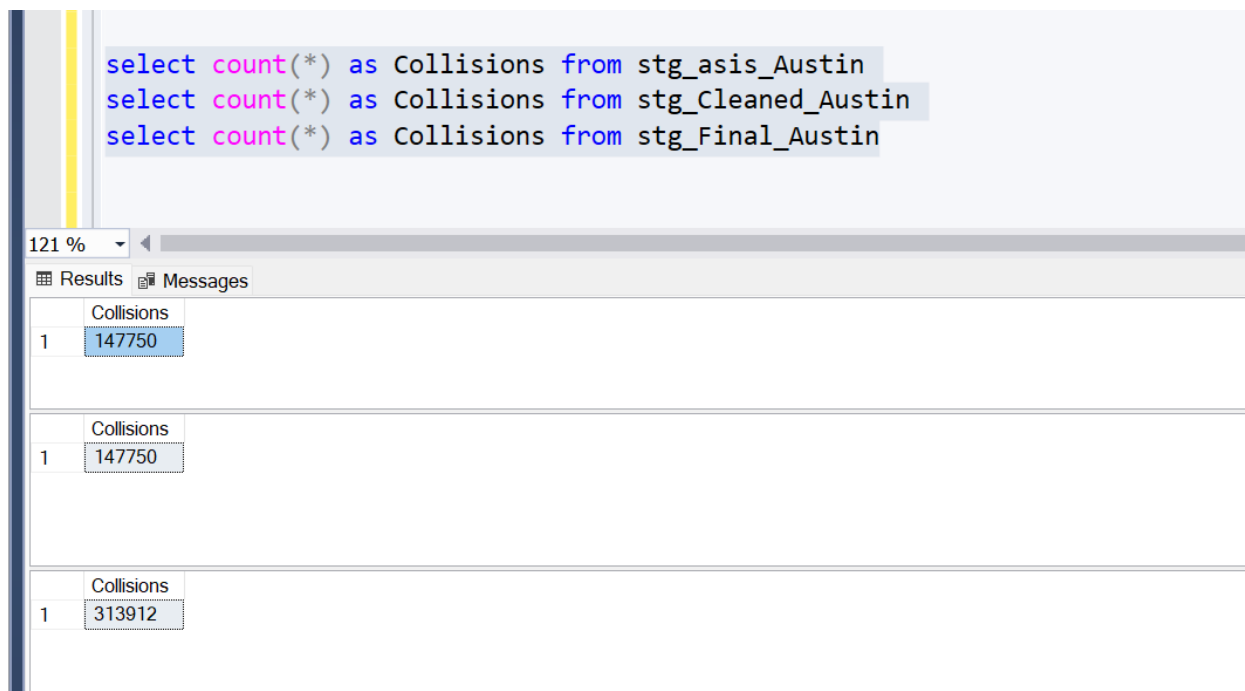


The diagram shows a data flow from MSSQL_Server_Austin to tMap_3, and then to Austin_Cleaned. The flow is labeled with row counts and durations: 147750 rows in 15.34s, 9632.94 rows in 15.34s, 147750 rows in 16.44s, and 8987.76 rows in 16.44s. Below the diagram is the execution log for the job Austin_MVC_Cleaned_Stg.

Execution log for job Austin_MVC_Cleaned_Stg:

```
Starting job Austin_MVC_Cleaned_Stg at 17:28 07/04/2024.
[statistics] connecting to socket on port 3741
[statistics] connected
[statistics] disconnected
Job Austin_MVC_Cleaned_Stg ended at 17:28 07/04/2024. [Exit code = 0]
```

Counts:-



The image shows a SQL query and its results. The query is:

```
select count(*) as Collisions from stg_asis_Austin
select count(*) as Collisions from stg_Cleaned_Austin
select count(*) as Collisions from stg_Final_Austin
```

The results are displayed in a table with the following data:

Collisions
147750

Collisions
147750

Collisions
313912

❖ Stage Table Loading for Austin:-

Select TOP 1000 * from stg_asis_Austin;

Select TOP 1000 * from stg_Cleaned_Austin;

Select TOP 1000 * from stg_Final_Austin;

121 %

Results Messages

	crash_id	crash_fatal_flg	crash_date	crash_time	case_id	rpt_latitude	rpt_longitude	rpt_block_num	rpt_street_ptx	rpt_street_name	rpt_street_sfx	crash_speed_limit	road_constr_zone_flg	latitude	long	
1	13762420	N	2014-03-30 10:58:00.000	10:58:00	140890874	NULL	NULL			3707 MANCHACA	RD	10	N	NULL	NU	
2	13777334	N	2014-03-27 13:07:00.000	13:07:00	140860852	NULL	NULL	3400		PALM WAY TO MOPAC NB RAMP		50	N	30.404	-97	
3	13777441	N	2014-03-28 15:42:00.000	15:42:00	140871196	NULL	NULL	8704		BALCONES CLUB DR	DR	-1	N	30.43798	-97	
4	13797332	N	2014-04-09 14:09:00.000	14:09:00	140991015	NULL	NULL	8000		E US 290 HWY SVRD EB		60	N	30.32762	-97	
5	13795604	N	2014-04-07 18:00:00.000	18:00:00	140971248	NULL	NULL	200	W	BEN WHITE	BLVD	-1	N	30.22516	-97	
6	13785070	N	2014-03-31 03:26:00.000	03:26:00	140900191	NULL	NULL	8700	S	IH 35 SVRD		50	N	30.16984	-97	
7	13790426	N	2014-04-04 14:34:00.000	14:34:00	140941160	NULL	NULL	4000		S FM 973 RD	RD	-1	N	30.1932	-97	
8	13795213	N	2014-04-18 02:06:00.000	02:06:00	141080164	NULL	NULL	1500	E	ANDERSON	LN	65	N	30.33279	-97	
	crash_id	crash_fatal_flg	crash_date	crash_time	case_id	rpt_latitude	rpt_longitude	rpt_block_num	rpt_street_ptx	rpt_street_name	rpt_street_sfx	crash_speed_limit	road_constr_zone_flg	latitude	longi	
1	18314841	N	2021-06-16 17:03:00.000	17:03:00	211671067	-1	-1	4600	N	IH 35	FWY	40	N	30.307	-97.7	
2	18359631	N	2021-07-05 17:33:00.000	17:33:00	211861561	-1	-1	5700	NA	E MARTIN LUTHER KING JR BLVD	NA	55	N	30.28595	-97.6	
3	18322172	N	2021-06-13 14:55:00.000	14:55:00	211641024	-1	-1	7100	N	CAPITAL OF TEXAS	NA	55	N	30.37146	-97.7	
4	18339872	N	2021-06-17 14:30:00.000	14:30:00	-1	30.33935	-97.70079	-1	NA	NOT REPORTED	NA	55	Y	30.33937	-97.7	
5	18344800	N	2021-06-29 20:56:00.000	20:56:00	211801393	-1	-1	10600	NA	NOT REPORTED	NA	65	N	30.39522	-97.7	
6	18326110	N	2021-06-22 16:57:00.000	16:57:00	211731048	-1	-1	5600	N	MOPAC NB	EXPY	60	N	30.33574	-97.7	
7	18288588	N	2021-06-02 00:06:00.000	00:06:00	211521653	-1	-1	6000	NA	NOT REPORTED	NA	60	Y	30.31807	-97.7	
8	18321630	N	2021-06-20 16:29:00.000	16:29:00	211711131	-1	-1	10200	NA	S IH 35 SVRD SB	NA	50	N	30.15356	-97.7	
	crash_id	crash_fatal_flg	crash_date	crash_time	case_id	rpt_latitude	rpt_longitude	rpt_block_num	rpt_street_ptx	rpt_street_name	rpt_street_sfx	crash_speed_limit	road_constr_zone_flg	latitude	longitude	street_name
1	18314841	N	2021-06-16 17:03:00.000	17:03:00	211671067	-1	-1	4600	N	IH 35	FWY	40	N	30.307	-97.71196	IH0035
2	18314841	N	2021-06-16 17:03:00.000	17:03:00	211671067	-1	-1	4600	N	IH 35	FWY	40	N	30.307	-97.71196	IH0035
3	18359631	N	2021-07-05 17:33:00.000	17:33:00	211861561	-1	-1	5700	NA	E MARTIN LU...	NA	55	N	30.28...	-97.66697	E MARTIN...
4	18359631	N	2021-07-05 17:33:00.000	17:33:00	211861561	-1	-1	5700	NA	E MARTIN LU...	NA	55	N	30.28...	-97.66697	E MARTIN...
5	18359631	N	2021-07-05 17:33:00.000	17:33:00	211861561	-1	-1	5700	NA	E MARTIN LU...	NA	55	N	30.28...	-97.66697	E MARTIN...
6	18322172	N	2021-06-13 14:55:00.000	14:55:00	211641024	-1	-1	7100	N	CAPITAL OF T...	NA	55	N	30.37...	-97.78593	SL0360
7	18322172	N	2021-06-13 14:55:00.000	14:55:00	211641024	-1	-1	7100	N	CAPITAL OF T...	NA	55	N	30.37...	-97.78593	SL0360

Dimensional model-Part 1

