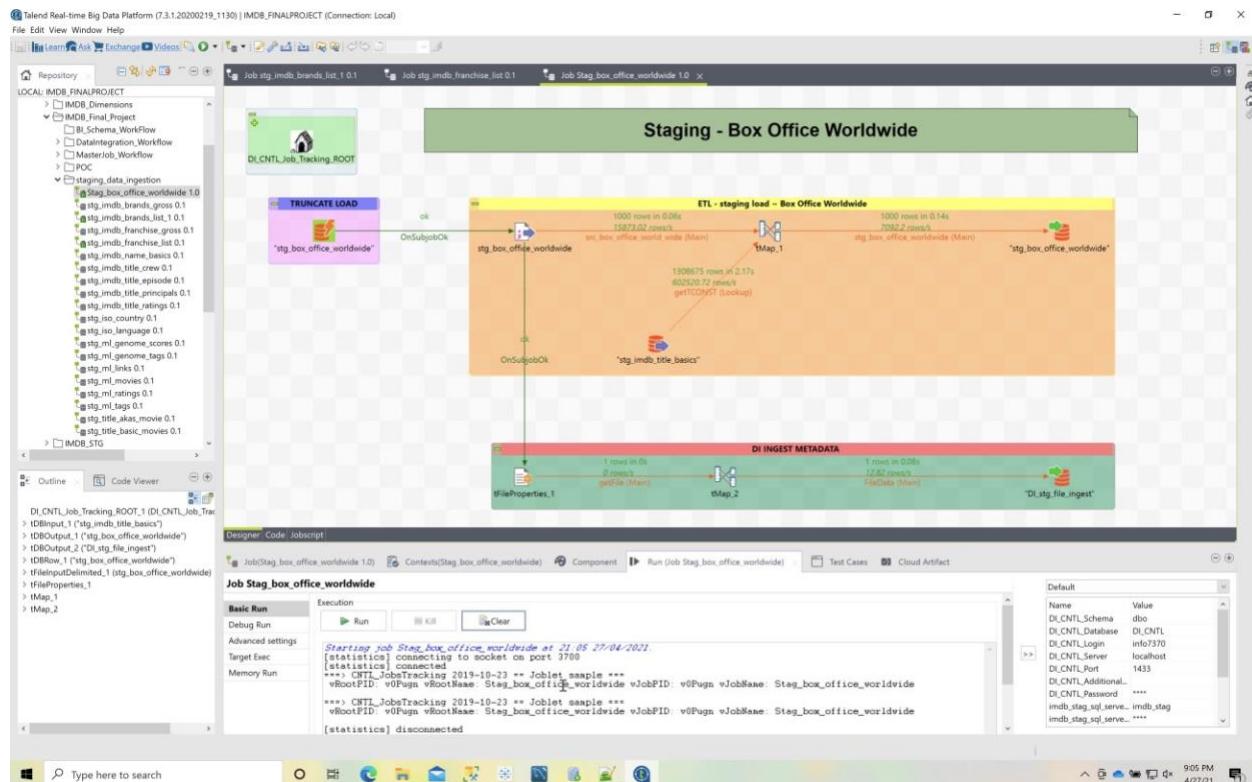


# Staging Screen Shots- Talend

## 1. Stag\_box\_office\_worldwide



Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

SQLQuery1.sql - localhost.imdb\_stag [info7370 (54)] - Microsoft SQL Server Management Studio (Administrator)

```
select * from [dbo].[stg_box_office_worldwide]
```

Results Messages

	Office_ Worldwide	BoxOffice_Rank	Title	WorldWide_LifetimeGross	Domestic_LifetimeGross	Domestic_Pct	Foreign_LifetimeGross	Foreign_Pct	Release_Year	BDR_input	BDR_Sk	Di_Creat	DT
1	1	1	Avengers: Endgame	279700564	68573000	0.307	193047594	0.693	2019	www09	9	NKXHu	2021-04-27 20:42:28.710
2	2	2	Avatar	279043000	70507625	0.272	202951375	0.728	2009	www09	9	NKXHu	2021-04-27 20:42:28.710
3	3	3	Titanic	1885190	68853944	0.3	183037884	0.547	1997	www09	9	NKXHu	2021-04-27 20:42:28.710
4	4	4	Star Wars: Episode VI - The Force Awakens	18262025	59882025	0.453	170300959	0.547	2015	www09	9	NKXHu	2021-04-27 20:42:28.710
5	5	5	Avengers: Infinity War	151547954	678015482	0.331	1389544272	0.669	2018	www09	9	NKXHu	2021-04-27 20:42:28.710
6	6	6	Jurassic World	167504037	65270625	0.38	1618130512	0.61	2016	www09	9	NKXHu	2021-04-27 20:42:28.710
7	7	7	Spider-Man: Homecoming	1603610	54582024	0.323	162203722	0.572	2017	www09	9	NKXHu	2021-04-27 20:42:28.710
8	8	8	The Avengers	151817988	62357910	0.41	89545078	0.59	2012	www09	9	NKXHu	2021-04-27 20:42:28.710
9	9	9	Furious 7	151934751	36307020	0.233	1162404965	0.767	2015	www09	9	NKXHu	2021-04-27 20:42:28.710
10	10	10	Deadpool	140230033	4779518	0.29	103711871	0.571	2019	www09	9	NKXHu	2021-04-27 20:42:28.710
11	11	11	Avengers: Age of Ultron	140280888	4900588	0.327	94800000	0.673	2015	www09	9	NKXHu	2021-04-27 20:42:28.710
12	12	12	Black Panther	140389583	7005996	0.52	64885398	0.48	2018	www09	9	NKXHu	2021-04-27 20:42:28.710
13	13	13	Star Wars: Episode VII - The Force Awakens	140400398	5944200	0.34	9844200	0.76	2017	www09	9	NKXHu	2021-04-27 20:42:28.710
14	14	14	Star Wars: Episode VIII - The Last Jedi	140527356	13253000	0.445	72368607	0.538	2017	www09	9	NKXHu	2021-04-27 20:42:28.710
15	15	15	Jurassic World: Fallen Kingdom	140847944	417719760	0.319	891748184	0.681	2018	www09	9	NKXHu	2021-04-27 20:42:28.710
16	16	16	Frozen	140820282	4073000	0.313	88064275	0.687	2013	www09	9	NKXHu	2021-04-27 20:42:28.710
17	17	17	Beauty and the Beast	140820798	56452128	0.302	89094505	0.643	2017	www09	9	NKXHu	2021-04-27 20:42:28.710
18	18	18	Incredibles 2	140807998	808081744	0.49	63423615	0.51	2018	www09	9	NKXHu	2021-04-27 20:42:28.710
19	19	19	The Fate of the Furious	140805118	22608385	0.183	105994703	0.817	2017	www09	9	NKXHu	2021-04-27 20:42:28.710
20	20	20	Minions	140805123	4800000	0.207	62335267	0.681	2013	www09	9	NKXHu	2021-04-27 20:42:28.710
21	21	21	Captain America: Civil War	140806400	115930897	0.305	83864570	0.71	2015	www09	9	NKXHu	2021-04-27 20:42:28.710
22	22	22	Spider-Man: Far From Home	140829629	40804349	0.354	745211944	0.646	2018	www09	9	NKXHu	2021-04-27 20:42:28.710
23	23	23	Avengers: Endgame	140830007	3880000	0.242	745211944	0.529	2018	www09	9	NKXHu	2021-04-27 20:42:28.710
24	24	24	The Lord of the Rings: The Return of the King	140219401	37784905	0.331	794373495	0.669	2003	www09	9	NKXHu	2021-04-27 20:42:28.727
25	25	25	Spider-Man: Far From Home	140326303	113187998	0.345	741389511	0.658	2019	www09	9	NKXHu	2021-04-27 20:42:28.727
26	26	26	Captain Marvel	140348844	4260000	0.378	781403685	0.623	2019	www09	9	NKXHu	2021-04-27 20:42:28.727
27	27	27	Game of Thrones: Dark of the Moon	140374079	3623000	0.318	781403685	0.568	2011	www09	9	NKXHu	2021-04-27 20:42:28.727
28	28	28	Skyfall	140861013	30460277	0.275	80420708	0.725	2012	www09	9	NKXHu	2021-04-27 20:42:28.727

localhost (15.0 RTM) | info7370 (54) | imdb\_stag | 00:00:00 | 1.000 rows

Time:

SQLQuery3.sql - loc\_cntl [info7370 (53)] - SQLQuery1.sql - loc\_tag [info7370 (54)] - Microsoft SQL Server Management Studio (Administrator)

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where job_name = 'Stag_box_office_worldwide'
```

Results Messages

	job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
1	Stag_box_office_worldwide	Y	Stag_box_office_worldwide	success	0.05	2021-04-27 21:05:08.400	end	IMDB_FINALPROJECT	1.0

# Staging Screen Shots- Talend

## DI INJECT:

The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left shows the database structure, including the 'imdb\_stag' database and its tables. The central pane displays the results of a query:

```
select * from DI_stg_file_ingest
```

The results grid shows one row of data:

File Group	Filename	FileSize	RowVersioned	Di_JobID	Di_CreatedOn
BOX OFFICE World Wide	C:\INFO\7379\dvdr-spring2021\odbc\FinalSource..	1000	1000	NHruAR	2021-04-27 21:11:41.677

At the bottom, a message indicates "Query executed successfully."

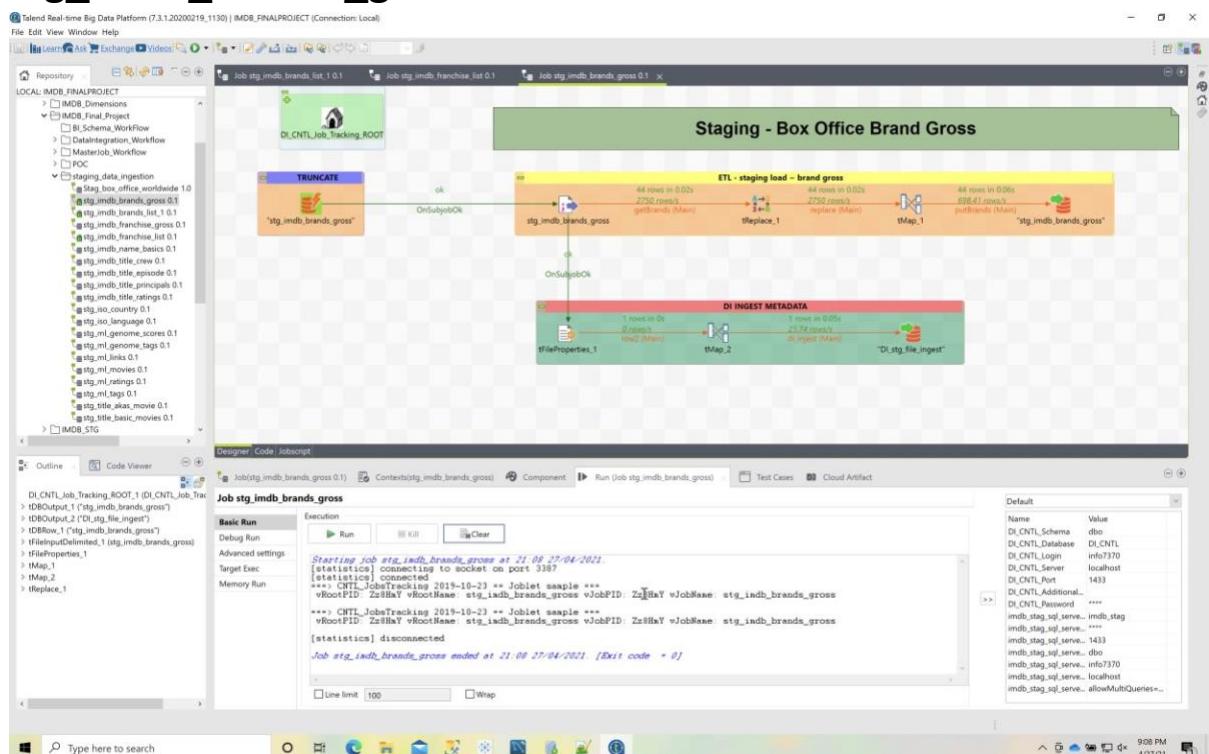
**data consistency:** data is loaded from file to stag table, data remains consistent with sk values and referential integrity

**reject:** for this table we are not supposed to do any rejects.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component and adding cols using tmap for create\_date and DI\_ETL\_ID which is process id, other than that we have also completed bonus part where in we are populating DI\_stg\_inject table.

# Staging Screen Shots- Talend

## 2. stg\_imdb\_brands\_gross



The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. A query window titled 'SQLQuery1.sql - localhost.imdb\_stag (info7370 (54))' is open, displaying the results of the query 'select \* from [dbo].[stg\_imdb\_brands\_gross]'. The results table has 44 rows and includes columns such as 'Brand Name', 'Total\_Revenue', 'Number\_of\_Releases', 'Top\_Release', 'Lifetime\_Gross', 'SQR\_SK', 'Dt\_JobID', and 'Dt\_Create\_DT'. The data includes various movie studios and their financial performance.

	Brand Name	Total_Revenue	Number_of_Releases	Top_Release	Lifetime_Gross	SQR_SK	Dt_JobID	Dt_Create_DT
1	Marvel Comics	14222386206	80	Avengers Endgame	698173000	13	9tB6lC	2021-04-27 20:57:47.943
2	Legendary Pictures	891814262	55	Jurassic World	852279525	13	9tB6lC	2021-04-27 20:57:47.943
3	Luciferfilm	620288628	39	Star Wars Episode VII - The Force Awakens	791254750	13	9tB6lC	2021-04-27 20:57:47.943
4	Pixar	607657062	25	Incredibles 2	908381144	13	9tB6lC	2021-04-27 20:57:47.943
5	DC Comics	571315101	44	The Dark Knight	533343058	13	9tB6lC	2021-04-27 20:57:47.943
6	DreamWorks Animation	504882000	14	Shrek	908381144	13	9tB6lC	2021-04-27 20:57:47.943
7	Bad Robot	307707801	15	Star Wars Episode VI - The Force Awakens	936462255	13	9tB6lC	2021-04-27 20:57:47.943
8	Vertigo Entertainment	305748002	40	Frozen	327481748	13	9tB6lC	2021-04-27 20:57:47.943
9	Warner Bros. Animation Studio	290164000	12	The Secret Life of Pets	368384330	13	9tB6lC	2021-04-27 20:57:47.943
10	Illumination Entertainment	281064046	9	Get Out	178040968	13	9tB6lC	2021-04-27 20:57:47.943
11	Blumhouse Productions	208893819	47	Teenage Mutant Ninja Turtles	19124754	13	9tB6lC	2021-04-27 20:57:47.943
12	Hawkins	193071489	16	Spider-Man: Homecoming	190243170	13	9tB6lC	2021-04-27 20:57:47.943
13	Nickelodeon	183071489	13	Spider-Man: Into the Spider-Verse	190243170	13	9tB6lC	2021-04-27 20:57:47.943
14	Sony Pictures Animation	191873863	21	Spider-Man: Into the Spider-Verse	190243170	13	9tB6lC	2021-04-27 20:57:47.943
15	Walt Disney Media	184283216	38	The Chronicles of Narnia: The Lion, the Witch a...	261719867	13	9tB6lC	2021-04-27 20:57:47.943
16	Blaauw Media	174042403	13	Ice Age: Dawn of the Dinosaurs	261719867	13	9tB6lC	2021-04-27 20:57:47.943
17	Stephen King	172944893	49	1	327481748	13	9tB6lC	2021-04-27 20:57:47.943
18	MTV	152046830	36	The Longest Yard	181198460	13	9tB6lC	2021-04-27 20:57:47.943
19	Paramount Pictures	152046830	17	Teenage Mutant Ninja Turtles	181198460	13	9tB6lC	2021-04-27 20:57:47.943
20	Saturday Night Live - Alumni Debate	1023546031	30	Bravehearts	189198725	13	9tB6lC	2021-04-27 20:57:47.943
21	Dark Horse Comics	847917338	16	300	210944905	13	9tB6lC	2021-04-27 20:57:47.943
22	The Big Picture Deep	8089389	8	Ala in Wonderland	257708862	13	9tB6lC	2021-04-27 20:57:47.943
23	Warner Animation Group	764039796	8	The Lego Movie	90580338	13	9tB6lC	2021-04-27 20:57:47.943
24	Tyler Perry	768833862	16	Madagascar	684847075	13	9tB6lC	2021-04-27 20:57:47.943
25	CBS Films	603201187	29	Scary Stories to Tell in the Dark	684847075	13	9tB6lC	2021-04-27 20:57:47.943
26	John Goodman	646264825	13	The Croods	227471707	13	9tB6lC	2021-04-27 20:57:47.943
27	Robert Liodum	64549188	6	The Bounce Ultimatum	227471707	13	9tB6lC	2021-04-27 20:57:47.943
28	Nicholas Sparks	574702829	11	The Notebook	8102787	13	9tB6lC	2021-04-27 20:57:47.943

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

TIME

The screenshot shows a SQL Server Management Studio (SSMS) window titled 'TIME'. The Object Explorer on the left shows a database named 'di\_cntl' containing several tables and stored procedures related to IMDB data. The central pane displays a T-SQL query:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_imdb_brands_gross'
```

The results pane at the bottom shows one row of data:

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_imdb_brands_gross	Y	stg_imdb_brands_gross	success	0.01	2021-04-27 21:13:20.017	end	IMDB_FINALPROJECT	0.1

# Staging Screen Shots- Talend

## DI\_INJECT:

The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left shows the database structure, including the 'imdb\_stag' database and its tables. The central pane displays a query results grid for the 'DI\_stg\_file\_ingest' table. The results show two rows of data, both of which are rejects. The first row is for 'BOX OFFICE World Wide' with File Group 'BOX OFFICE' and File Name 'C:\INFO\7379\dwh-spring2021-odbc\FinalSource...'. The second row is for 'BOX OFFICE Map' with File Group 'BOX OFFICE Map' and File Name 'C:\INFO\7379\dwh-spring2021-odbc\FinalSource...'. Both rows have RowStatus '44', DI\_JobID 'NNHuAR', and DI\_CreatedOn '2021-04-27 21:11:41.677'. The status column for both rows is 'REJECTED'. The bottom status bar indicates the query was executed successfully.

File Group	File Name	FileRows	RowStatus	DI_JobID	DI_CreatedOn
BOX OFFICE	C:\INFO\7379\dwh-spring2021-odbc\FinalSource...	1000	44	NNHuAR	2021-04-27 21:11:41.677
BOX OFFICE Map	C:\INFO\7379\dwh-spring2021-odbc\FinalSource...	44	REJECTED	FFPHO	2021-04-27 21:13:20.017

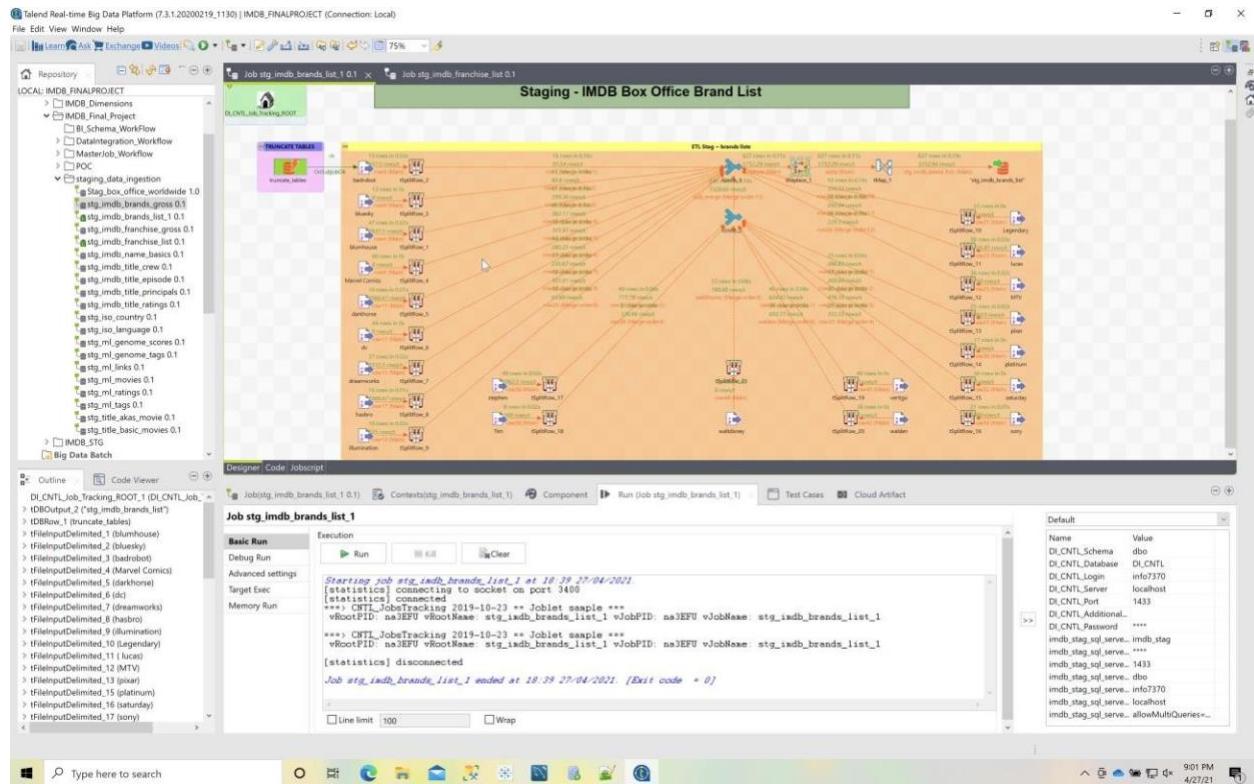
**data consistency:** data is loaded from file to stag table, data remains consistent with sk values and referential integrity.

**reject:** for this table we are not supposed to do any rejects.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component and adding cols using tmap for create\_date and DI\_ETL\_ID which is process id, other than that we have also completed bonus part where in we are populating DI\_stg\_inject table.

## 3. stg\_imdb\_brands\_list

# Staging Screen Shots- Talend



# Staging Screen Shots- Talend

SQlQuery1.sql - localhost.imdb\_stag (Info7370 [54]) - Microsoft SQL Server Management Studio (Administrator)

File Edit View Query Project Tools Window Help

Object Explorer

localhost (SQL Server 15.0.2080.9 - Info7370)

- Databases
  - System Databases
  - AdventureworksLT
  - AdventureworksDW
  - AdventureworksDW2019
  - Chinook
  - di\_cmt
  - GraduateDataModel
  - imbd\_stag
  - Database Diagrams
  - Tables
    - FileTables
    - Extended Tables
    - Graph Tables
    - dbo.DL\_stg\_file\_ingest
    - dbo.stg\_box\_office\_worldwide
    - dbo.stg\_box\_office\_worldwide\_all
    - dbo.stg\_imdb\_brands\_gross
    - dbo.stg\_imdb\_brands\_list
    - dbo.stg\_imdb\_box\_office\_gross
    - dbo.stg\_imdb\_franchise\_list
    - dbo.stg\_imdb\_name\_basics
    - dbo.stg\_imdb\_name\_basics\_rejects
    - dbo.stg\_imdb\_title\_akas
    - dbo.stg\_imdb\_title\_akas\_rejects
    - dbo.stg\_imdb\_title\_principals
    - dbo.stg\_imdb\_title\_principals\_rejects
    - dbo.stg\_imdb\_title\_ratings
    - dbo.stg\_imdb\_title\_ratings\_rejects
    - dbo.stg\_no\_country
    - dbo.stg\_no\_language
    - dbo.stg\_ml\_genome\_scores
    - dbo.stg\_ml\_genome\_tags
    - dbo.stg\_ml\_links
    - dbo.stg\_ml\_movies
    - dbo.stg\_ml\_ratings
    - dbo.stg\_ml\_tags
  - Views
  - Materialized Views
  - Synonyms
  - Programmability
  - Service Broker

SQLQuery1.sql - loc.tag (Info7370 [54])

```
select * from [dbo].[stg_imdb_brands_list]
```

I

Results (109 rows affected)

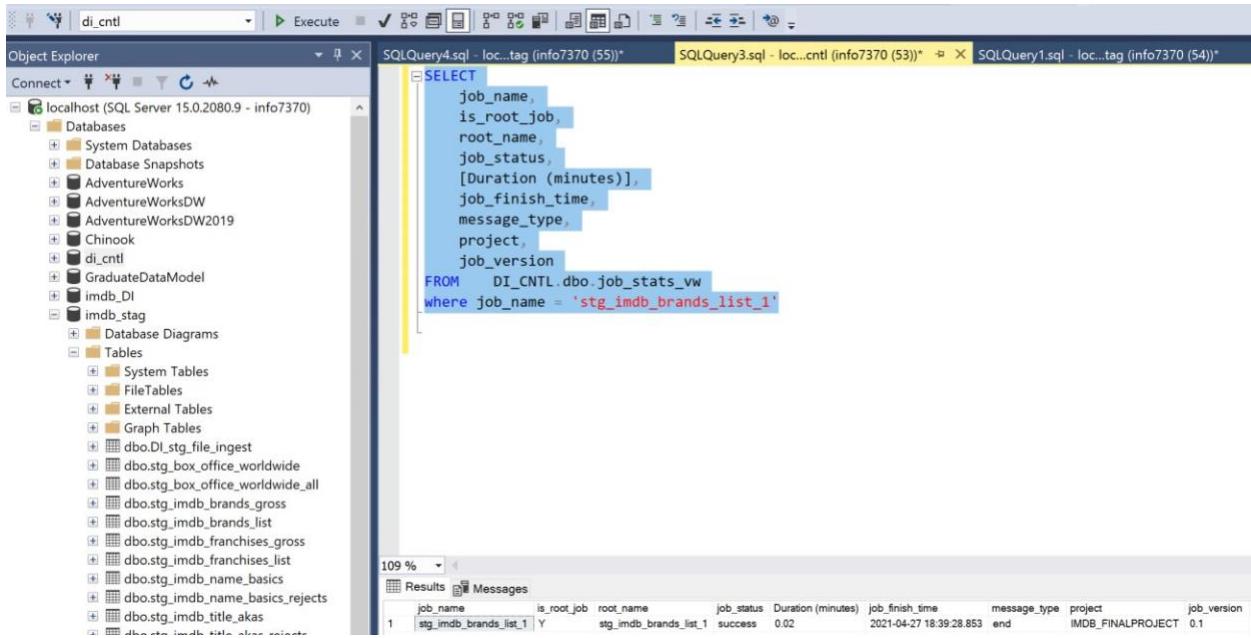
Brands_Len_SK	Brand_Name	Release_Fmt	Release_Name	Lifetime_Gross	Wk_Theaters	Opening_Gross	Open_Theaters	Release_Date	Deadline	DOI_Sr	Di_JobID	Di_Creat	DT
1	Bad Robot	1	Star Wars: Episode VII - The Force Awakens	93666225	4134	24796675	4134	2015-12-18 00:00:00.000	Wat Disney Studios Motion Pictures	1	na3FVU	2021-04-27 18:39:28.713	
2	Bad Robot	2	Star Wars: Episode IX - The Rise of Skywalker	615202542	4405	177383854	4053	2019-12-20 00:00:00.000	Wat Disney Studios Motion Pictures	1	na3FVU	2021-04-27 18:39:28.713	
3	Bad Robot	3	Star Trek	287730219	4053	76243289	3845	2009-08-09 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
4	Bad Robot	4	Star Trek Into Darkness	238770593	3907	76243289	3308	2013-05-16 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
5	Bad Robot	5	Mission: Impossible - Fallout	220169104	4395	61236834	4386	2018-07-27 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
6	Bad Robot	6	Mission: Impossible - Ghost Protocol	208937603	3585	12785204	425	2011-12-16 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
7	Bad Robot	7	Mission: Impossible - Rogue Nation	188200763	5549	55200000	5549	2015-07-29 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
8	Bad Robot	8	Star Trek Beyond	158848340	3623	8033211	3623	2016-07-22 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
9	Bad Robot	9	Super 8	127004179	3424	354511910	3379	2011-06-10 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
10	Bad Robot	10	Cloverfield	80044521	4241	40393701	4241	2008-09-12 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
11	Bad Robot	11	10 Cloverfield Lane	70828293	3427	24727437	3391	2016-03-11 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
12	Bad Robot	12	Morning Glory	31011732	2344	925288	2616	2010-11-10 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
13	Bad Robot	13	Jay Ride	214949	2052	7347298	2497	2001-09-28 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
14	Bad Robot	14	Clover	21754844	2893	10225238	2899	2014-11-09 00:00:00.000	Paramount Pictures	1	na3FVU	2021-04-27 18:39:28.713	
15	Bad Robot	15	Inferno	1420655	110	47000	5	2015-09-19 00:00:00.000	Sony Pictures Classics	1	na3FVU	2021-04-27 18:39:28.713	
16	Blue Sky	1	Ice Age: Dawn of the Dinosaurs	188677305	4102	41990862	4098	2009-07-17 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
17	Blue Sky	2	Ice Age: The Meltdown	188332000	4822	48220004	4822	2006-07-21 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
18	Blue Sky	3	Ice Age: Collision Course	178381405	3345	48312454	3316	2013-05-15 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
19	Blue Sky	4	Ice Age: Continental Drift	161321943	3884	48629259	3881	2012-07-13 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
20	Blue Sky	5	Horton Hears a Who!	148420301	3581	48220004	3581	2008-05-16 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
21	Blue Sky	6	Rio	140161809	3842	38225962	3828	2011-04-19 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
22	Blue Sky	7	Rio 2	121138408	3975	39278989	3944	2014-04-11 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
23	Blue Sky	8	The Peanuts Movie	123719520	4422	44220004	4422	2015-02-13 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
24	Blue Sky	9	Robots	128205912	3776	36045301	3776	2005-03-11 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
25	Blue Sky	10	Epic	16711682	3884	33010631	3882	2013-06-24 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
26	Blue Sky	11	Ferdinand	84419295	5830	1486888	3621	2017-09-22 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
27	Blue Sky	12	Spies in Disguise	64575703	2002	13554768	2019	2019-12-26 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	
28	Blue Sky	13	Ice Age: Collision Course	640300305	3997	21370906	3992	2018-04-22 00:00:00.000	Twentieth Century Fox	1	na3FVU	2021-04-27 18:39:28.713	

Query executed successfully.

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

TIME:



The screenshot shows a SQL Server Management Studio (SSMS) interface. The Object Explorer on the left shows a connection to 'localhost (SQL Server 15.0.2080.9 - info7370)' with several databases listed, including 'AdventureWorks', 'AdventureWorksDW', 'AdventureWorksDW2019', 'Chinook', 'di\_ctrl', 'GraduateDataModel', 'imdb\_DL', and 'imdb\_stag'. The 'Tables' node under 'imdb\_stag' contains many tables related to IMDB data. The central pane displays a T-SQL query:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_imdb_brands_list_1'
```

The 'Results' tab at the bottom shows the output of the query:

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_imdb_brands_list_1	Y	stg_imdb_brands_list_1	success	0.02	2021-04-27 18:39:28.853	end	IMDB_FINALPROJECT	0.1

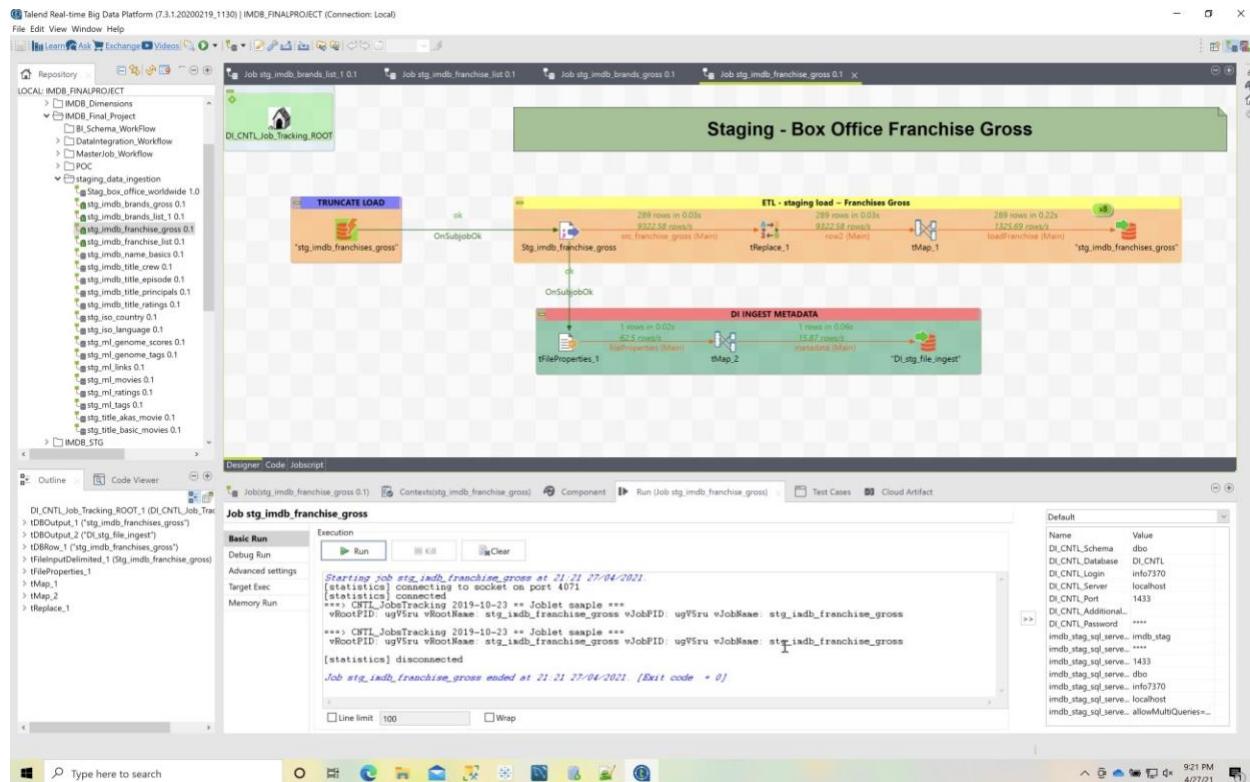
**data consistency:** data is loaded from multiple files, to make consistent data we used tReplace

**reject:** for this table we are not supposed to do any rejects.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component, tUnite to merge all the files into a single result set table also, added cols using tmap for create\_date and DI\_ETL\_ID which is process id, other than that we have also completed bonus part where in we are populating DI\_stg\_injest table.

## 4. stg\_imdb\_franchise\_gross

# Staging Screen Shots- Talend

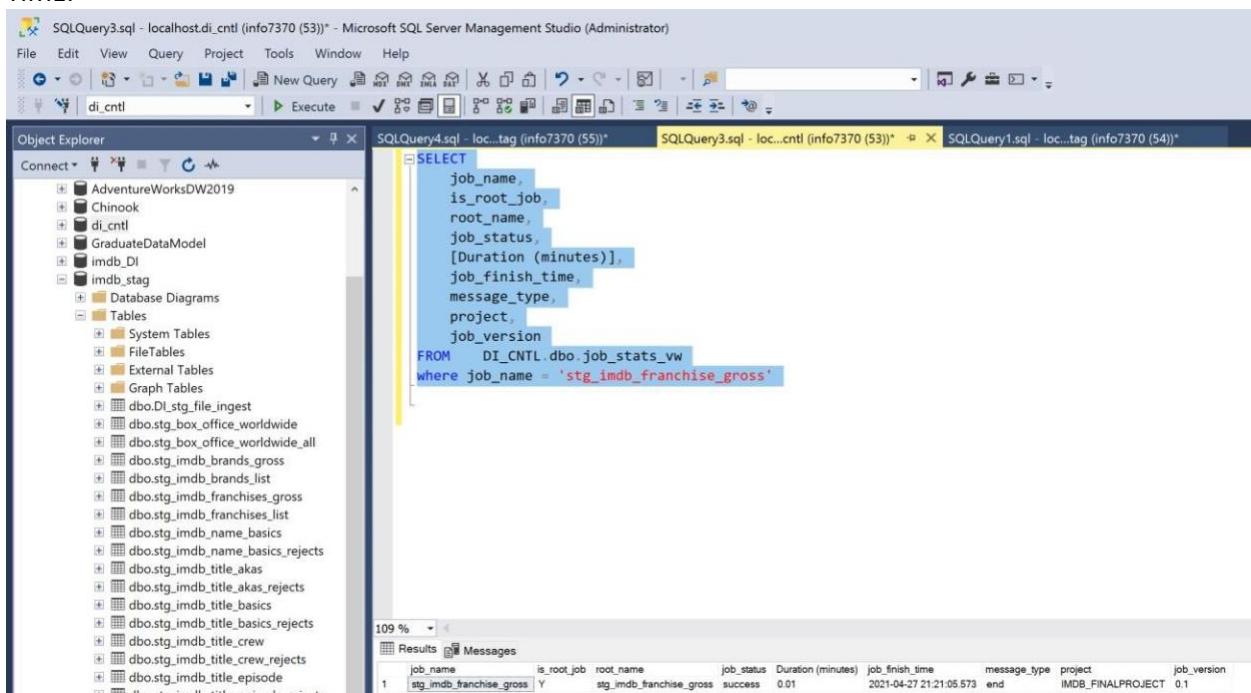


Franchise_Gross	Franchise	Title	Revenue	Number_of_Releases	Type
856373000		Avengers: Endgame	13	ug7fu	2021-04-27 21:21:05.393
93662225		Star Wars: Episode VI - The Force Awakens	13	ug7fu	2021-04-27 21:21:05.397
54308043		The Lion King	13	ug7fu	2021-04-27 21:21:05.397
100000000		Harry Potter and the Deathly Hallows: Part 2	13	ug7fu	2021-04-27 21:21:05.397
856373000		Avengers: Endgame	13	ug7fu	2021-04-27 21:21:05.397
403708375		Spider-Man: Homecoming	13	ug7fu	2021-04-27 21:21:05.397
200000000		Despicable Me	13	ug7fu	2021-04-27 21:21:05.397
533040588		The Dark Knight	13	ug7fu	2021-04-27 21:21:05.397
387017219		Harry Potter and the Deathly Hallows: Part 2	13	ug7fu	2021-04-27 21:21:05.397
412863408		Wonder Woman	13	ug7fu	2021-04-27 21:21:05.397
100000000		Transformers: Revenge of the Fallen	13	ug7fu	2021-04-27 21:21:05.397
652270625		Middle Earth	13	ug7fu	2021-04-27 21:21:05.397
377027525		The Lord of the Rings: The Return of the King	13	ug7fu	2021-04-27 21:21:05.397
100000000		Transformers: Dark of the Moon	13	ug7fu	2021-04-27 21:21:05.397
402111870		Transformers: Revenge of the Fallen	13	ug7fu	2021-04-27 21:21:05.397
423319182		Pirates of the Caribbean: Dead Man's Chest	13	ug7fu	2021-04-27 21:21:05.397
100000000		Transformers: Dark of the Moon	13	ug7fu	2021-04-27 21:21:05.397
447232047		The Hunger Games: Catching Fire	13	ug7fu	2021-04-27 21:21:05.397
257730019		Shrek	13	ug7fu	2021-04-27 21:21:05.397
100000000		Star Trek	13	ug7fu	2021-04-27 21:21:05.397
140138863		The Hunger Games: Catching Fire	13	ug7fu	2021-04-27 21:21:05.397
100000000		The Dark Knight	13	ug7fu	2021-04-27 21:21:05.397
137199346		Troy	13	ug7fu	2021-04-27 21:21:05.397
122026440		Toy Story 4	13	ug7fu	2021-04-27 21:21:05.397
343026088		Despicable Me 2	13	ug7fu	2021-04-27 21:21:05.397
368060585		Transformers: Revenge of the Fallen	13	ug7fu	2021-04-27 21:21:05.397
533040588		Transformers: Dark of the Moon	13	ug7fu	2021-04-27 21:21:05.397
100000000		Transformers: The Last Knight	13	ug7fu	2021-04-27 21:21:05.397
119379588		Transformers: The Last Knight	13	ug7fu	2021-04-27 21:21:05.397
100000000		Transformers: The Last Knight	13	ug7fu	2021-04-27 21:21:05.397
330360194		Transformers: The Last Knight	13	ug7fu	2021-04-27 21:21:05.397
377027525		Batman v Superman: Dawn of Justice	13	ug7fu	2021-04-27 21:21:05.397
100000000		Batman v Superman: Dawn of Justice	13	ug7fu	2021-04-27 21:21:05.397
100000000		Indiana Jones and the Kingdom of the Crystal Skull	13	ug7fu	2021-04-27 21:21:05.397
317701119		Indiana Jones and the Kingdom of the Crystal Skull	13	ug7fu	2021-04-27 21:21:05.397

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

TIME:



The screenshot shows a Microsoft SQL Server Management Studio (Administrator) window. The Object Explorer on the left lists several databases and objects under 'di\_cntl'. The central pane displays a T-SQL query:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL dbo.job_stats_vw
where job_name = 'stg_imdb_franchise_gross'
```

The results pane at the bottom shows one row of data:

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
1 stg_imdb_franchise_gross	Y	stg_imdb_franchise_gross	success	0.01	2021-04-27 21:21:05.573	end	IMDB_FINALPROJECT	0.1

# Staging Screen Shots- Talend

## DI INJECT:

The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left lists databases like AdventureworksDW2019, Chinook, di\_cnf1, GraduateDataModel, imbd\_DI, and imbd\_stag. The current connection is to imbd\_stag. The central pane displays the results of a query:

```
select * from DI_stg_file_inject
```

The results grid shows three rows of data:

DI_stg_file_inject	File_Group	Filename	FileRows	RowInserted	DI_ArtID	DI_CreatedDate
1	BOX OFFICE World Wide	C:\INFO\7370\dwh-spring2021\odbc\FinalSource..	44	44	FPhHO	2021-04-27 21:13:20.017
2	2	BOX OFFICE Mojo	C:\INFO\7370\dwh-spring2021\odbc\FinalSource..	44	FPhHO	2021-04-27 21:13:20.017
3	3	BOX OFFICE Mojo	C:\INFO\7370\dwh-spring2021\odbc\FinalSource..	289	ugvSu	2021-04-27 21:21:08.573

At the bottom, it says "Query executed successfully."

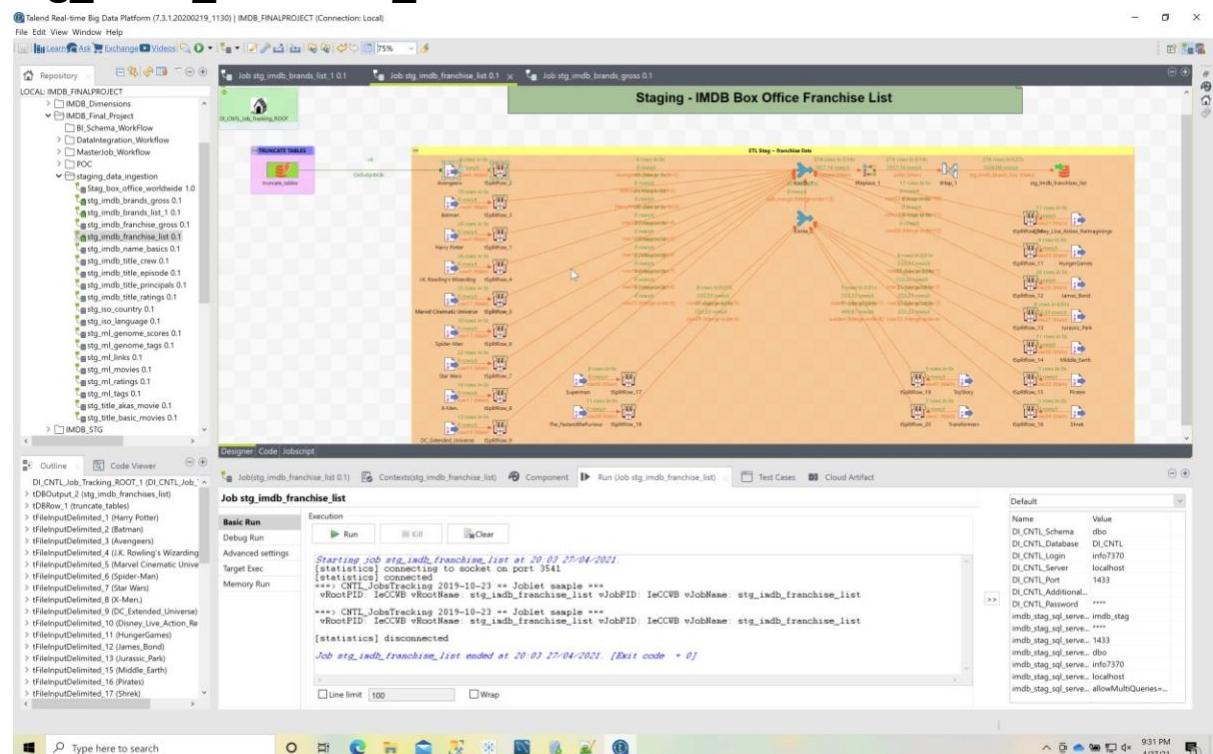
**data consistency:** data is loaded from the file, we are using tReplace to make data clean and consistent, identity column making the stg consistent.

**reject:** for this table we are not supposed to do any rejects.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component, we have added cols using tmap for create\_date and DI\_ETL\_ID which is process id, other than that we have also completed bonus part where in we are populating DI\_stg\_inject table.

# Staging Screen Shots- Talend

## 5. stg\_imdb\_franchise\_list



The screenshot shows the SSMS interface with the query results for the 'stg\_imdb\_franchise\_list' table. The results table contains 274 rows of franchise data, including columns such as Franchise\_Lm\_SK, Franchise, Release\_Rank, Release\_Name, Lifetime\_Gross, etc. The message bar at the bottom indicates the query was executed successfully.

Franchise_Lm_SK	Franchise	Release_Rank	Release_Name	Lifetime_Gross	Mer_Theater	Opening_Gross	Open_Theaters	Release_Date	Distributor	SOR_SK	Di_WebID	Di_Creation_DT
1	Avengers	1	Avengers: Endgame	4662	357118000	4662	4662	2019-04-26 00:00:00.000	Walt Disney Studios Motion Pictures	1	WCCWB	2021-04-27 20:00:00.383
2	Avengers	2	Avengers: Infinity War	270464000	270464000	4714	270464000	2019-04-26 00:00:00.000	Walt Disney Studios Motion Pictures	1	WCCWB	2021-04-27 20:00:00.383
3	Avengers	3	The Avengers	4349	207438700	4349	207438700	2010-04-04 00:00:00.000	Walt Disney Studios Motion Pictures	1	WCCWB	2021-04-27 20:00:00.383
4	Avenger	4	Avengers: Age of Ultron	4276	192771100	4276	192771100	2015-05-01 00:00:00.000	Walt Disney Studios Motion Pictures	1	WCCWB	2021-04-27 20:00:00.383
5	Batman	5	The Dark Knight	16349	163490000	16349	16349	2008-07-19 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
6	Batman	6	The Dark Knight Rises	4404	160872900	4404	160872900	2012-07-20 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
7	Batman	7	Batman v Superman: Dawn of Justice	330070194	4256	160677374	4242	2016-03-25 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
8	Batman	8	Batman v Superman: Justice League	22011	40400000	22011	40400000	2017-03-17 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
9	Batman	9	Batman Begins	200247714	2008	4745640	2008	2005-06-15 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
10	Batman	10	Batman Returns	184011112	2893	52784433	2842	1996-06-16 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
11	Batman	11	Batman Returns	175792634	4088	53035488	4088	2017-02-10 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
12	Batman	12	Batman Returns	162320000	2500	52644101	2500	1995-07-28 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
13	Batman	13	Batman & Robin	107201956	2942	42672805	2934	1997-06-20 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
14	Batman	14	Batman Mask of the Phantasm	5617391	1508	1189075	1508	1983-12-25 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
15	Batman	15	Batman Returns	373000000	1125	0	0	2000-07-28 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
16	Batman	16	The Dark Knight2017 Re-release	1510986	0	0	0	2012-07-19 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
17	Batman	17	Batman Begins2012 Re-release	1500858	0	0	0	2012-07-19 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
18	Batman	18	The Dark Knight2012 Re-release	2700000	0	0	0	2012-07-19 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
19	Batman	19	Batman2019 Re-release	159419	488	159419	488	2019-04-04 00:00:00.000	Fathom Events	1	WCCWB	2021-04-27 20:00:00.383
20	Batman	20	Batman Returns2019 Re-release	70642	507	0	0	2019-06-08 00:00:00.000	Fathom Events	1	WCCWB	2021-04-27 20:00:00.383
21	Batman	21	Batman Returns2019 Re-release	36017	424	38014	424	2019-06-14 00:00:00.000	Fathom Events	1	WCCWB	2021-04-27 20:00:00.383
22	Batman	22	Batman Returns2019 Re-release	28987	0	0	0	2019-06-14 00:00:00.000	Fathom Events	1	WCCWB	2021-04-27 20:00:00.383
23	Batman	23	Batman Mask of the Phantasm2019 Re-release	17813	761	0	0	2018-11-12 00:00:00.000	Fathom Events	1	WCCWB	2021-04-27 20:00:00.383
24	Harry Potter	24	Harry Potter and the Deathly Hallows: Part 2	38107819	4376	160186427	4376	2011-07-20 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
25	Harry Potter	25	Harry Potter and the Half-Blood Prince	31978769	3202	9028000	3202	2009-07-18 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
26	Harry Potter	26	Harry Potter and the Prisoner of Azkaban	30159917	4455	73857172	4455	2004-07-01 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
27	Harry Potter	27	Harry Potter and the Chamber of Secrets	29583305	4126	12697372	4126	2001-07-19 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383
28	Harry Potter	28	Harry Potter and the Order of the Phoenix	25204720	4269	7718841	4269	2007-07-12 00:00:00.000	Warner Bros.	1	WCCWB	2021-04-27 20:00:00.383

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

TIME:

The screenshot shows the SQL Server Management Studio interface. In the Object Explorer on the left, there are several databases listed: AdventureWorksDW2019, Chinook, di\_cntl, GraduateDataModel, imdb\_DL, and imdb\_stag. Under the imdb\_stag database, there are various tables and views, including stg\_imdb\_franchise\_list. In the center, three tabs are open: SQLQuery4.sql, SQLQuery3.sql, and SQLQuery1.sql. The SQLQuery3.sql tab contains a SELECT statement:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_imdb_franchise_list'
```

In the bottom right pane, the Results tab displays the output of the query:

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_imdb_franchise_list	Y	stg_imdb_franchise_list	success	0.01	2021-04-27 20:03:50.587	end	IMDB_FINALPROJECT	0.1

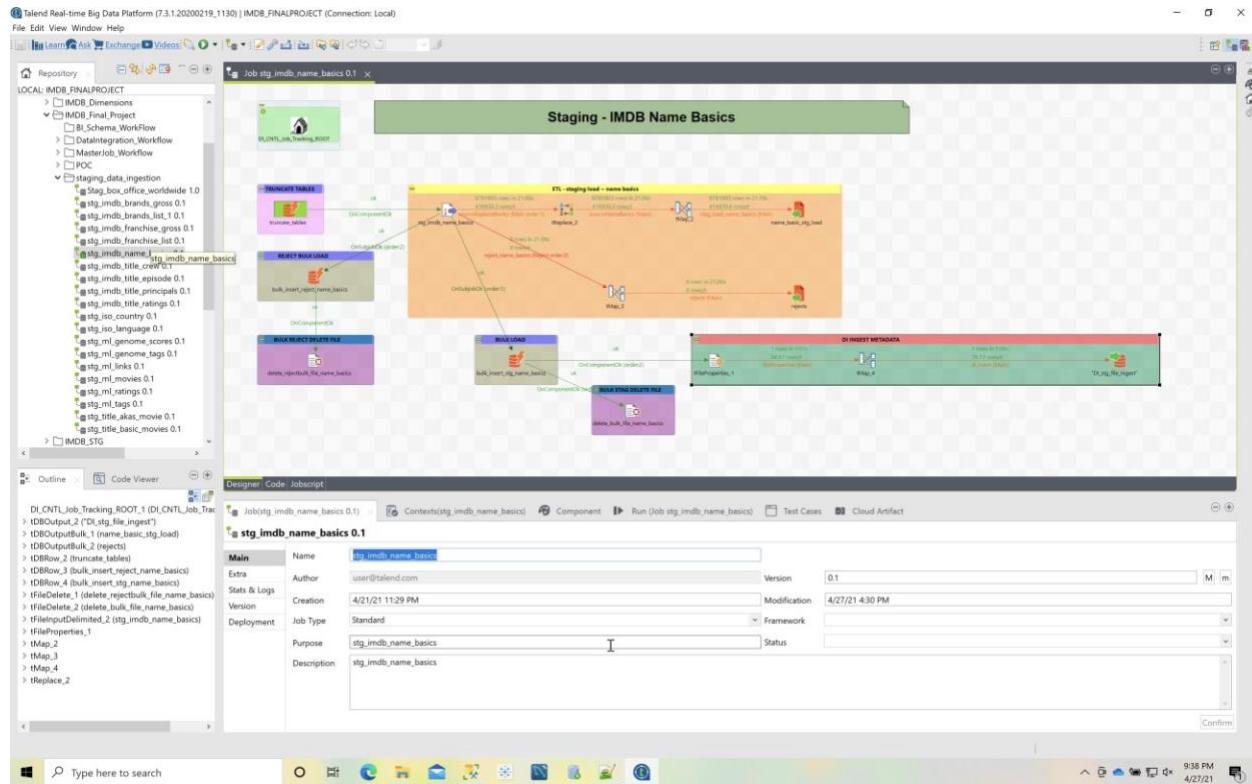
**data consistency:** data is loaded from multiple files, to make consistent data we have used tReplace component to clean the data.

**reject:** for this table we are not supposed to do any rejects.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component, tUnite to merge all the files into a single result set table also, added cols using tmap for create\_date and DI\_ETL\_ID which is process id, other than that we have also completed bonus part where in we are populating DI\_stg\_injest table.

# Staging Screen Shots- Talend

## 6. stg\_imdb\_name\_basics & Rejects



# Staging Screen Shots- Talend

SQLQuery1.sql - localhost.imdb\_stag (Info7370 (54)) - Microsoft SQL Server Management Studio (Administrator)

```
select * from [dbo].[stg_imdb_name_basics]
```

Object Explorer

- AdventureWorksDW2019
- Chinook
- di\_cntl
- GraduateDataModel
- imdb\_DL
- imdb\_stag
- Database Diagrams
- Tables
  - System Tables
  - FileTables
  - External Tables
  - Graph Tables
  - dbo.DL\_stg\_file\_ingest
  - dbo.stg\_box\_office\_worldwide
  - dbo.stg\_box\_office\_worldwide\_all
  - dbo.stg\_imdb\_brands\_gross
  - dbo.stg\_imdb\_brands\_list
  - dbo.stg\_imdb\_franchises\_gross
  - dbo.stg\_imdb\_franchises\_list
  - dbo.stg\_imdb\_name\_basics
  - dbo.stg\_imdb\_name\_rejects
  - dbo.stg\_imdb\_title\_akas
  - dbo.stg\_imdb\_title\_akas\_rejects
  - dbo.stg\_imdb\_title\_basics
  - dbo.stg\_imdb\_title\_rejects
  - dbo.stg\_imdb\_title\_crew
  - dbo.stg\_imdb\_title\_episode
  - dbo.stg\_imdb\_title\_episode\_rejects
  - dbo.stg\_imdb\_title\_principals
  - dbo.stg\_imdb\_title\_rejects
  - dbo.stg\_imdb\_title\_ratings
  - dbo.stg\_imdb\_title\_settings\_rejects
  - dbo.stg\_imdb\_no\_country
  - dbo.stg\_imdb\_no\_language
  - dbo.stg\_ml\_genome\_scores
  - dbo.stg\_ml\_genome\_tags
  - dbo.stg\_ml\_movies
  - dbo.stg\_ml\_ratings
  - dbo.stg\_ml\_tags
  - Views
  - External Resources
  - Synonyms
  - Assemblies
  - Seniors
  - Services
  - Broker
  - Storage
  - Security
- inf07370
- inf07370\_output
- NYPD
- NYPD\_TST

Results Messages

table_5k_recnum	primaryName	birthYear	deathYear	birtYear_cstr	deathYear_cstr	primaryProfession	knownForTitles	sort_sh	di_job	di_create_dt
1	Lauren Bacall	1914	1984	1914	1984	actress,soundtrack,make-up,department	2007177,2002395,0111767,2002782	1	987Inv	2021-04-27 00:00:00
2	Brigitte Bardot	1934	2014	1934	2014	actress,soundtrack,music,department	2004442,2005745,0093048,0040189	1	987Inv	2021-04-27 00:00:00
3	John Barrymore	1882	1942	1882	1942	actor,soundtrack,writer	2007723,2008079,0097975,2002782	1	987Inv	2021-04-27 00:00:00
4	John Barrymore	1882	1942	1882	1942	actor,soundtrack,writer	2007723,2008079,0097975,2002782	1	987Inv	2021-04-27 00:00:00
5	Ingrid Bergman	1915	1982	1915	1982	actress,soundtrack,producer	2003483,2003855,009109,00303787	1	987Inv	2021-04-27 00:00:00
6	Humphrey Bogart	1899	1957	1899	1957	actor,soundtrack,producer	2003370,2003453,0040997,0042165	1	987Inv	2021-04-27 00:00:00
7	Richard Burton	1925	1984	1925	1984	actor,soundtrack,producer	2005777,2005749,0001184,2002893	1	987Inv	2021-04-27 00:00:00
8	James Cagney	1899	1986	1899	1986	actor,soundtrack,director	2005376,2003187,0040241,0020370	1	987Inv	2021-04-27 00:00:00
9	Bette Davis	1908	1989	1908	1989	actress,soundtrack,make-up,department	2005487,2003120,0040280,2003140	1	987Inv	2021-04-27 00:00:00
10	Doris Day	1922	2019	1922	2019	soundtrack,actress,producer	2004817,2005943,009109,0030172	1	987Inv	2021-04-27 00:00:00
11	Marlene Dietrich	1897	1962	1897	1962	actress,soundtrack,make-up,department	2005451,2005745,009109,0030172	1	987Inv	2021-04-27 00:00:00
12	Bette Davis	1908	1989	1908	1989	actress,soundtrack,make-up,department	2005487,2003120,0040280,2003140	1	987Inv	2021-04-27 00:00:00
13	James Dean	1931	1955	1931	1955	actor,musician,writer	2004541,2004581,0040228,005023	1	987Inv	2021-04-27 00:00:00
14	James Dean	1931	1955	1931	1955	actor,musician,writer	2004541,2004581,0040228,005023	1	987Inv	2021-04-27 00:00:00
15	Georges Delerue	1925	1982	1925	1982	cognac,soundtrack,music,department	2008946,2009783,0098320,0037345	1	987Inv	2021-04-27 00:00:00
16	Markie Dietrich	1901	1982	1901	1982	actress,soundtrack,make-up,department	2005121,2005121,000111,2002116	1	987Inv	2021-04-27 00:00:00
17	Kathleen Freeman	1912	1984	1912	1984	actress,soundtrack,make-up,department	2005121,2005121,000111,2002116	1	987Inv	2021-04-27 00:00:00
18	Federico Fellini	1920	1993	1920	1993	writer,director,assistant,director	2004738,20050783,0007120,0030779	1	987Inv	2021-04-27 00:00:00
19	Henry Fonda	1905	1982	1905	1982	actor,producer,soundtrack	2003231,2004116,0005083,0003044	1	987Inv	2021-04-27 00:00:00
20	Clark Gable	1901	1960	1901	1960	actor,soundtrack,producer	2002752,2002382,0003116,0031181	1	987Inv	2021-04-27 00:00:00
21	Judy Garland	1922	1969	1922	1969	soundtrack,actress	2004732,2005031,0002138,2003199	1	987Inv	2021-04-27 00:00:00
22	Jerry Goldsmith	1929	2004	1929	2004	music,department,soundtrack,composer	2010982,2011771,0011488,0112178	1	987Inv	2021-04-27 00:00:00
23	Cary Grant	1904	1986	1904	1986	actor,soundtrack,producer	2004728,2003259,0006023,0003125	1	987Inv	2021-04-27 00:00:00
24	Pita Hayworth	1918	1987	1918	1987	actress,soundtrack,producer	2003723,2004025,0005698,0004952	1	987Inv	2021-04-27 00:00:00

Activate Windows  
Go to Settings to activate Windows

localhost (15.0 RTM) info7370 (54) imdb\_stag\_000123 8,781,803 rows

Ready Type here to search

9:53 PM 4/27/21

TIME:

SQLQuery4.sql - loc\_tag (Info7370 (55))

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_imdb_name_basics'
```

Object Explorer

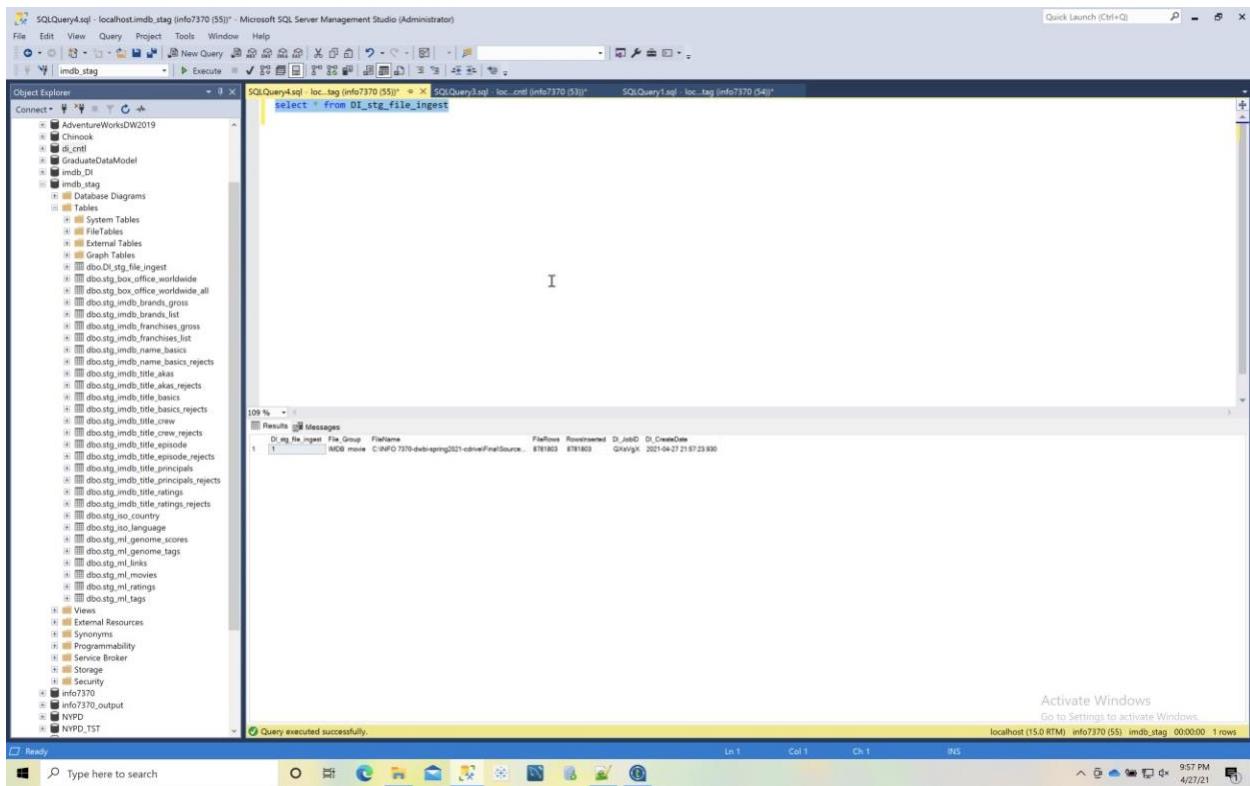
- AdventureWorksDW2019
- Chinook
- di\_cntl
- GraduateDataModel
- imdb\_DL
- imdb\_stag
- Database Diagrams
- Tables
  - System Tables
  - FileTables
  - External Tables
  - Graph Tables
  - dbo.DL\_stg\_file\_ingest
  - dbo.stg\_box\_office\_worldwide
  - dbo.stg\_box\_office\_worldwide\_all
  - dbo.stg\_imdb\_brands\_gross
  - dbo.stg\_imdb\_brands\_list
  - dbo.stg\_imdb\_franchises\_gross
  - dbo.stg\_imdb\_franchises\_list
  - dbo.stg\_imdb\_name\_basics
  - dbo.stg\_imdb\_name\_rejects
  - dbo.stg\_imdb\_title\_akas
  - dbo.stg\_imdb\_title\_akas\_rejects
  - dbo.stg\_imdb\_title\_basics
  - dbo.stg\_imdb\_title\_rejects
  - dbo.stg\_imdb\_title\_crew
  - dbo.stg\_imdb\_title\_episode
- inf07370
- inf07370\_output
- NYPD
- NYPD\_TST

Results Messages

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_imdb_name_basics	Y	stg_imdb_name_basics	success	1.00	2021-04-27 21:37:41.613	end	IMDB_FINALPROJECT	0.1

# Staging Screen Shots- Talend

## DI INJECT:



**data consistency:** data is loaded from tsv files, to make consistent data we have used tReplace component to clean the data.

**reject:** for rejects we are checking trim columns and check each row against schema structure, we have not received any rejects for this table.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component.

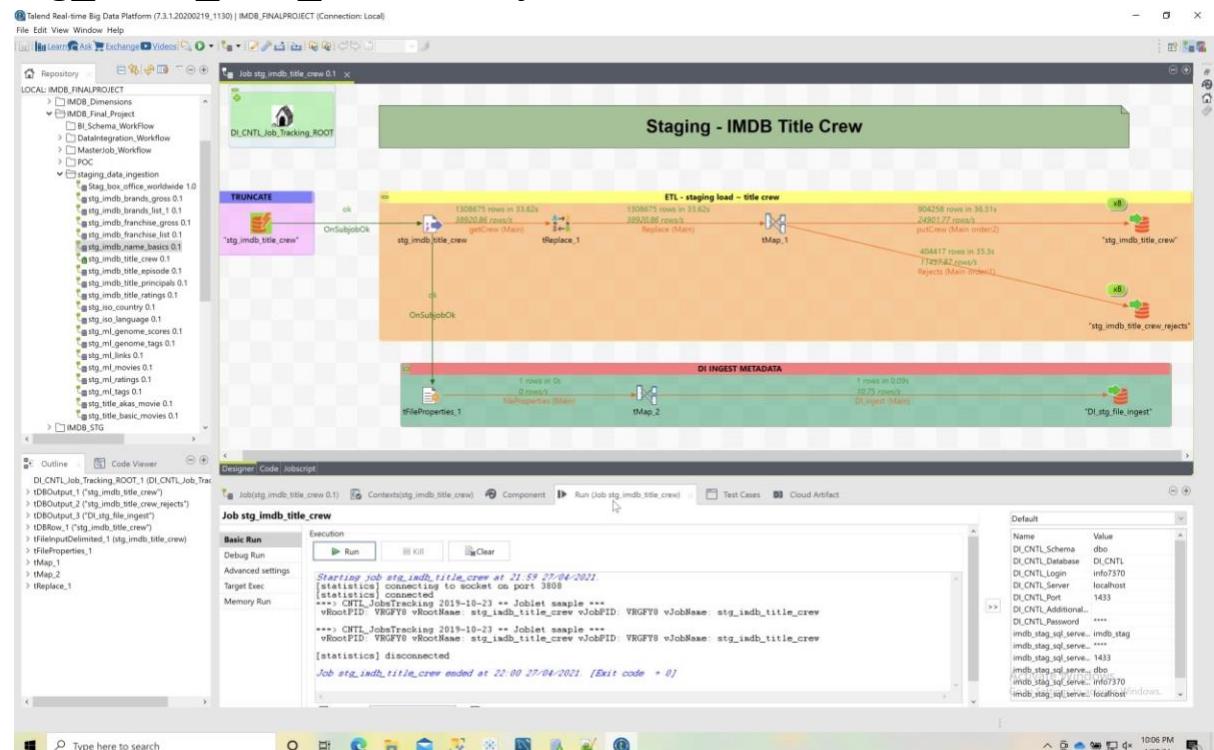
For performance we are using bulk components,

For DI standard we are deleting bulk file which is created in the process

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 7. stg\_imdb\_title\_crew & Rejects



title_sk	name	director	writers	sdr_sk	di_jobid	di_create_dt	di_update_dt
1	1618	nm000009	nm008156	4	VRFQYB	2021-04-21 09:29:587	
2	1692	nm000036	nm005960	4	VRFQYB	2021-04-21 09:29:587	
3	1693	nm000036	nm005960	4	VRFQYB	2021-04-21 09:29:587	
4	1694	nm000091	nm017588	4	VRFQYB	2021-04-21 09:29:587	
5	1695	nm000108	nm005960	4	VRFQYB	2021-04-21 09:29:587	
6	1696	nm000109	nm005960	4	VRFQYB	2021-04-21 09:29:587	
7	1697	nm000110	nm005960	4	VRFQYB	2021-04-21 09:29:587	
8	1698	nm000111	nm005960	4	VRFQYB	2021-04-21 09:29:587	
9	1699	nm000112	nm005960	4	VRFQYB	2021-04-21 09:29:587	
10	1700	nm000132	nm005960	4	VRFQYB	2021-04-21 09:29:587	
11	1701	nm000132	nm017588	4	VRFQYB	2021-04-21 09:29:583	
12	1702	nm000138	nm017588	4	VRFQYB	2021-04-21 09:29:583	
13	1703	nm000140	nm017588	4	VRFQYB	2021-04-21 09:29:583	
14	1704	nm000174	nm047118	4	VRFQYB	2021-04-21 09:29:583	
15	1705	nm000181	nm008310	4	VRFQYB	2021-04-21 09:29:583	
16	1706	nm000182	nm008310	4	VRFQYB	2021-04-21 09:29:583	
17	1707	nm000189	nm005717	4	VRFQYB	2021-04-21 09:29:583	
18	1708	nm000190	nm047118	4	VRFQYB	2021-04-21 09:29:583	
19	1709	nm000192	nm024040	4	VRFQYB	2021-04-21 09:29:583	
20	1710	nm000193	nm024040	4	VRFQYB	2021-04-21 09:29:583	
21	1711	nm000211	nm017588	4	VRFQYB	2021-04-21 09:29:583	
22	1712	nm000215	nm047118	4	VRFQYB	2021-04-21 09:29:583	
23	1713	nm000216	nm047118	4	VRFQYB	2021-04-21 09:29:583	
24	1714	nm000230	nm017588	4	VRFQYB	2021-04-21 09:29:583	
25	1715	nm000233	nm005717	4	VRFQYB	2021-04-21 09:29:583	
26	1716	nm000242	nm005717	4	VRFQYB	2021-04-21 09:29:583	
27	1717	nm000242	nm017588	4	VRFQYB	2021-04-21 09:29:583	
28	1718	nm000247	nm156608, nm005680, nm002304	4	VRFQYB	2021-04-21 09:29:583	

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

The screenshot shows a Microsoft SQL Server Management Studio window with three tabs open:

- SQLQuery1.sql - localhost.imdb\_stag (info7370 (54)) - Microsoft SQL Server Management Studio (Administrator)
- SQLQuery4.sql - loc\_tag (info7370 (55))\*
- SQLQuery3.sql - loc\_cntl (info7370 (53))\*

The Object Explorer on the left shows the database structure for the imdb\_stag database, including tables like stg\_imdb\_title\_crew\_rejects.

The Results pane displays the output of the following query:

```
select * from [dbo].[stg_imdb_title_crew_rejects]
```

The results show approximately 404,417 rows of data, each containing columns such as title\_id, genre, directors, writers, Dr\_Report\_Code, SOR\_SK, Dr\_JobID, D\_Create\_DT, and VRSF19.

At the bottom of the Results pane, it says "Query executed successfully."

TIME:

The screenshot shows a Microsoft SQL Server Management Studio window with three tabs open:

- SQLQuery4.sql - loc\_tag (info7370 (55))\*
- SQLQuery3.sql - loc\_cntl (info7370 (53))\*
- SQLQuery1.sql - loc.tag (info7370 (54))\*

The Object Explorer on the left shows the database structure for the loc\_cntl database, including tables like DI\_CNTL.dbo.job\_stats\_vw.

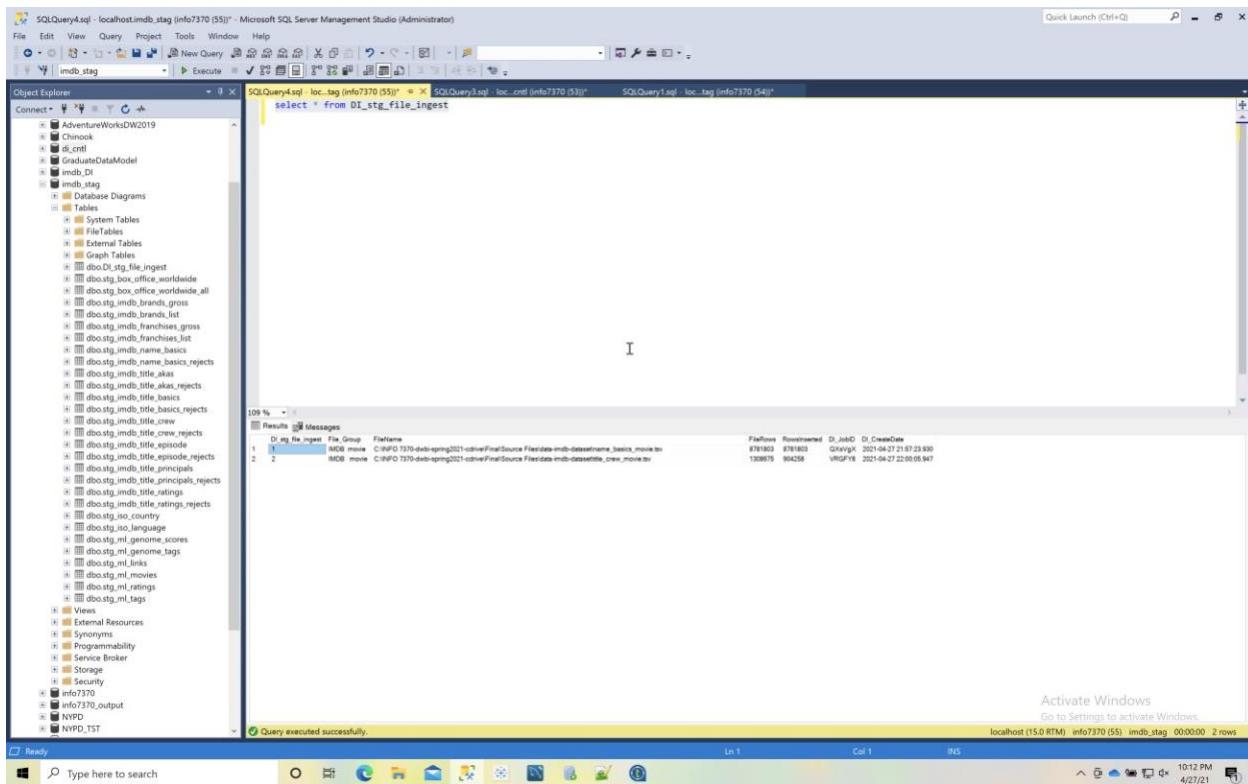
The Results pane displays the output of the following query:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM
    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_imdb_title_crew'
```

The results show one row of data with columns: job\_name, is\_root\_job, root\_name, job\_status, Duration (minutes), job\_finish\_time, message\_type, project, and job\_version.

# Staging Screen Shots- Talend

## DI INJECT:



**data consistency:** data is loaded from tsv files, to make consistent data we have used tReplace component to clean the data.

**reject:** for rejects we are checking trim columns and check each row against schema structure, we are also using referential integrity reject for col di\_reject\_code

```
(Relational.ISNULL(Replace.directors) && !(Relational.ISNULL(Replace.writers))) ? -997 :  
(Relational.ISNULL(Replace.writers) && !(Relational.ISNULL(Replace.directors))) ? -996 :
```

```
((Relational.ISNULL(Replace.directors)) && (Relational.ISNULL(Replace.writers))) ? -995 : 0
```

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component.

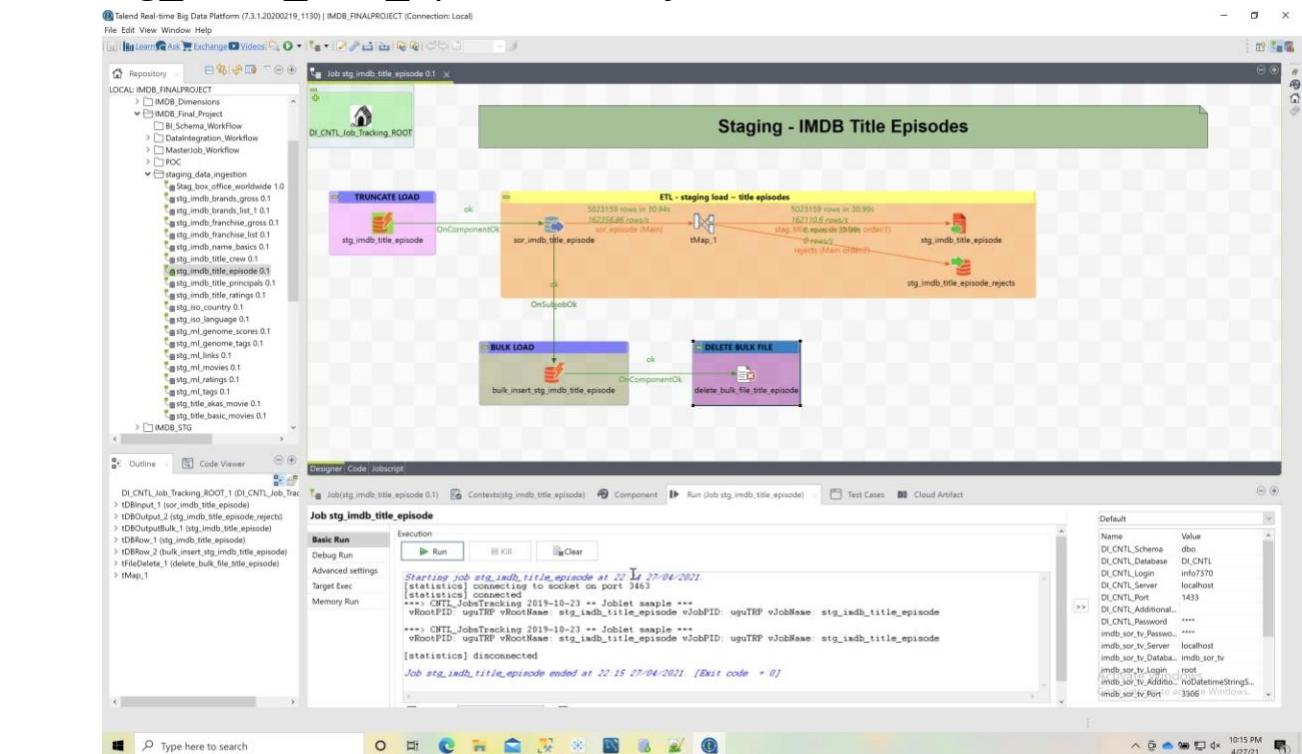
For performance we are using bulk components,

For DI standard we are deleting bulk file which is created in the process

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 8. stg\_imdb\_title\_episode & Rejects



The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface with a query window running the following SQL statement:

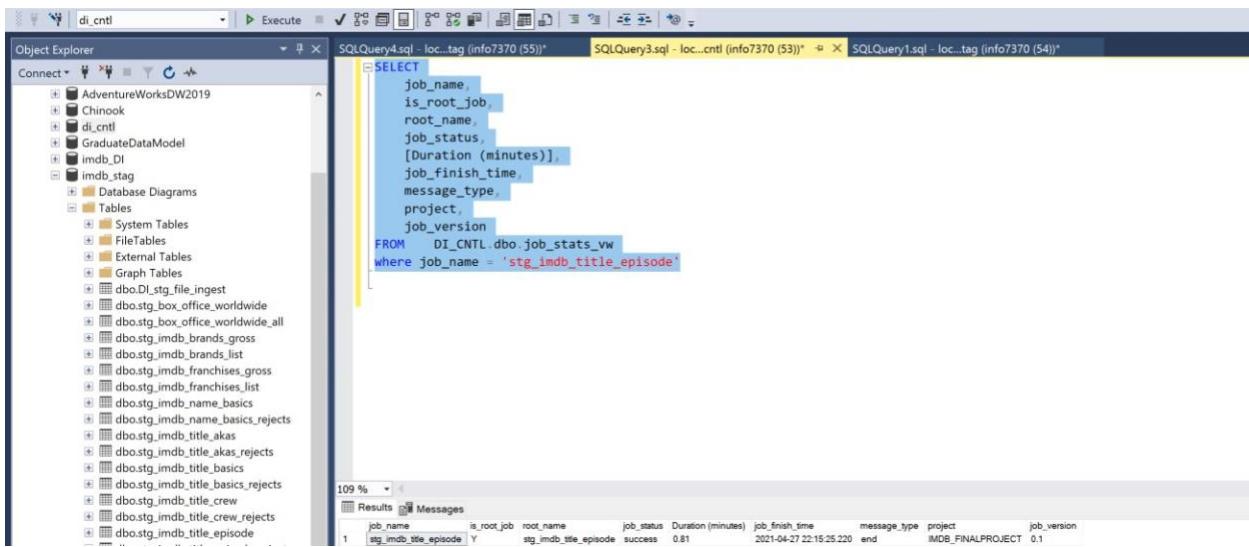
```
select * from [dbo].[stg_imdb_title_episode]
```

The results grid displays the following columns:

row_id	sk	name	parentName	seasonNumber	episodeNumber	seasonNumber_char	episodeNumber_char	SOR_sk	D_JOBID	DI_Created_DT
1	1	80041951	80041938	1	9	1		1	ugvTRP	2021-04-27 00:00:00.000
2	2	80042416	80041935	1	17	1	17	1	ugvTRP	2021-04-27 00:00:00.000
3	3	80042416	80041935	1	18	1	18	1	ugvTRP	2021-04-27 00:00:00.000
4	4	80043426	80040501	2	42	3	42	1	ugvTRP	2021-04-27 00:00:00.000
5	5	80043431	80040515	2	16	2	16	1	ugvTRP	2021-04-27 00:00:00.000
6	6	80043710	80040515	2	4	2	4	1	ugvTRP	2021-04-27 00:00:00.000
7	7	80043710	80040515	2	3	3	3	1	ugvTRP	2021-04-27 00:00:00.000
8	8	80044083	80040862	1	6	1	6	1	ugvTRP	2021-04-27 00:00:00.000
9	9	80044688	80042423	2	16	2	16	1	ugvTRP	2021-04-27 00:00:00.000
10	10	80044688	80042423	2	45	3	45	1	ugvTRP	2021-04-27 00:00:00.000
11	11	80045019	80040132	4	11	4	11	1	ugvTRP	2021-04-27 00:00:00.000
12	12	80045980	80042424	2	3	2	3	1	ugvTRP	2021-04-27 00:00:00.000
13	13	80046000	80042425	4	5	4	5	1	ugvTRP	2021-04-27 00:00:00.000
14	14	80046150	80347198	NULL	NULL	NULL	NULL	1	ugvTRP	2021-04-27 00:00:00.000
15	15	80046855	80040484	1	4	1	4	1	ugvTRP	2021-04-27 00:00:00.000
16	16	80047180	80040484	2	20	5	20	1	ugvTRP	2021-04-27 00:00:00.000
17	17	80047810	80747402	3	35	3	35	1	ugvTRP	2021-04-27 00:00:00.000
18	18	80047882	80547746	1	15	1	15	1	ugvTRP	2021-04-27 00:00:00.000
19	19	80047901	80547746	2	9	2	9	1	ugvTRP	2021-04-27 00:00:00.000
20	20	80047981	80548913	3	6	1	6	1	ugvTRP	2021-04-27 00:00:00.000
21	21	80048067	80546887	2	20	2	20	1	ugvTRP	2021-04-27 00:00:00.000
22	22	80048302	80547795	1	6	1	6	1	ugvTRP	2021-04-27 00:00:00.000
23	23	80048302	80547795	6	11	6	11	1	ugvTRP	2021-04-27 00:00:00.000
24	24	80048378	80547701	1	6	1	6	1	ugvTRP	2021-04-27 00:00:00.000
25	25	80048442	80547702	1	3	1	3	1	ugvTRP	2021-04-27 00:00:00.000
26	26	80048442	80547702	1	10	1	10	1	ugvTRP	2021-04-27 00:00:00.000
27	27	80048883	80989126	8	42	8	42	1	ugvTRP	2021-04-27 00:00:00.000
28	28	80048883	80989126	1	20	1	20	1	ugvTRP	2021-04-27 00:00:00.000

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend



The screenshot shows a SQL Server Management Studio (SSMS) interface. In the Object Explorer, there are several databases listed: AdventureWorksDW2019, Chinook, di\_cntl, GraduateDataModel, imdb\_DL, and imdb\_stag. Under the di\_cntl database, there are tables like System Tables, FileTables, External Tables, Graph Tables, and various specific tables related to the IMDB dataset. A query window is open with the following T-SQL code:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_imdb_title_episode'
```

The Results tab shows one row of data from the query:

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_imdb_title_episode	Y		success	0.81	2021-04-27 22:15:25.220	end	IMDB_FINALPROJECT	0.1

**data consistency:** data is loaded from sor tv mysql db data is consistent using sk column

**reject:** for rejects we are checking trim columns and check each row against schema structure, we have not received any rejects for this table.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component.

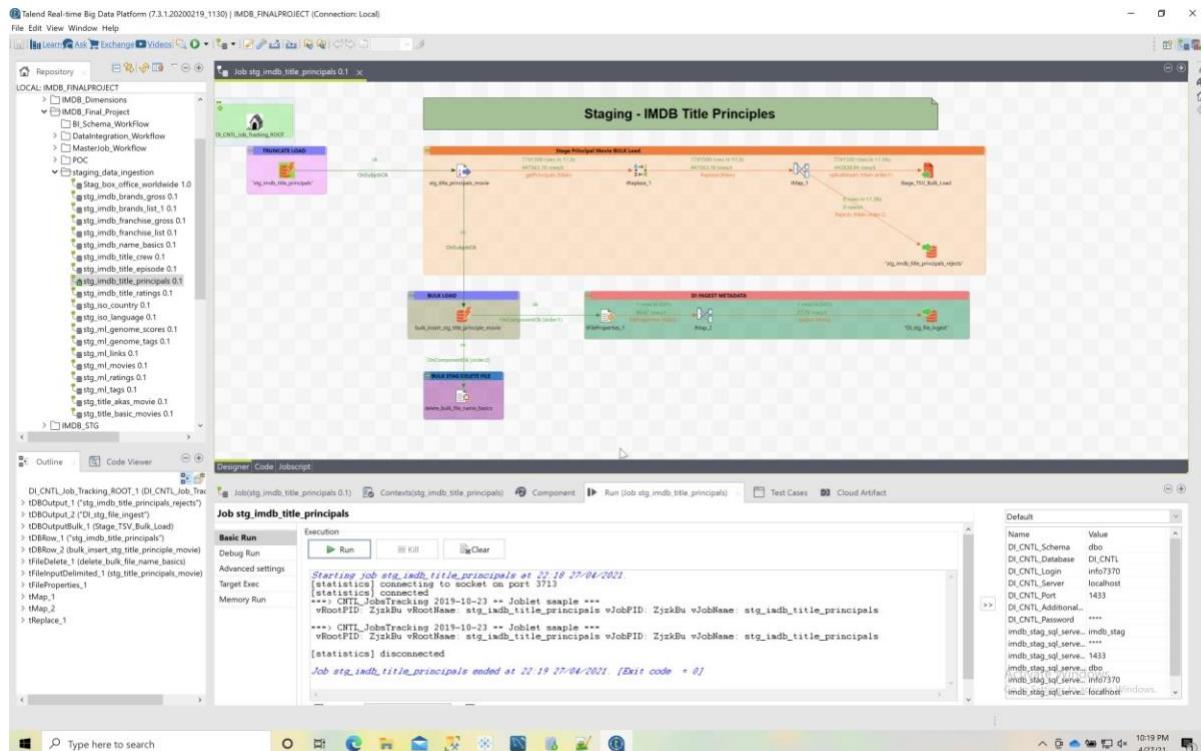
For performance we are using bulk components,

For DI standard we are deleting bulk file which is created in the process

We are also capturing DI ingest metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 9. stg\_imdb\_title\_principals & Rejects



SQLQuery1.sql - localhost.imdb\_stag (info7370 (54)) - Microsoft SQL Server Management Studio (Administrator)

```
select * from [dbo].[stg_imdb_title_principals]
```

Object Explorer

	id	name	ordering	parent	category	job	characters	SCR_SK	Dir_JobID	Dir_Create_DT
1	1	00000001	1	nm108870	self	NULL	["Self"]	5	ZjzBu	2021-04-27 00:00:00.000
2	2	10000001	2	nm0000890	director	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
3	3	00000001	3	nm0000890	cinematographer	director of photography	NULL	5	ZjzBu	2021-04-27 00:00:00.000
4	4	10000001	4	nm0000890	editor	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
5	5	00000002	2	nm133271	composer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
6	6	10000003	1	nm0071526	director	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
7	7	10000003	2	nm0071526	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
8	8	10000003	3	nm133271	composer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
9	9	10000003	4	nm044200	editor	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
10	10	10000003	5	nm0071526	editor	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
11	11	10000004	2	nm133271	composer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
12	12	10000005	1	nm044482	actor	NULL	["Blackmail"]	5	ZjzBu	2021-04-27 00:00:00.000
13	13	10000005	2	nm0005690	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
14	14	10000005	3	nm0005690	director	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
15	15	10000005	4	nm024379	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
16	16	10000006	1	nm0071526	director	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
17	17	10000006	2	nm0071526	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
18	18	10000007	2	nm018947	actor	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
19	19	10000007	3	nm0005690	director	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
20	20	10000007	4	nm0005690	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
21	21	10000007	5	nm024379	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
22	22	10000008	1	nm005328	actor	NULL	["Screwing Man"]	5	ZjzBu	2021-04-27 00:00:00.000
23	23	10000008	2	nm005328	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
24	24	10000008	3	nm037485	cinematographer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
25	25	10000009	1	nm0005690	actress	NULL	["Miss Sunshine-Holbrook (Miss Jerry)"]	5	ZjzBu	2021-04-27 00:00:00.000
26	26	10000009	2	nm0005690	producer	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000
27	27	10000009	3	nm130758	actor	NULL	["Cheaney Deppe - The Director of the New York..."]	5	ZjzBu	2021-04-27 00:00:00.000
28	28	10000009	4	nm0085156	director	NULL	NULL	5	ZjzBu	2021-04-27 00:00:00.000

Activate Windows

Go to Settings to activate Windows.

# Staging Screen Shots- Talend

TIME:

```

SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_imdb_title_principals'
  
```

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_imdb_title_principals	Y		success	0.83	2021-04-27 22:19:44.603	end	IMDB_FINALPROJECT	0.1

DI INJECT:

```

select * from DI_stg_file_ingest
  
```

Di.stg_file_ingest	File Group	Filename	FileRows	RowVersioned	Di.JobID	Di.CreateDate
1	IMDB movie	C:\INFO\T370\delsr\spring2021\cdlne\FinalSource\Feed\data\imdb-datasetname_basics_movie.tsv	6791803	8191803	GKvVgk	2021-04-27 21:57:23.930
2	IMDB movie	C:\INFO\T370\delsr\spring2021\cdlne\FinalSource\Feed\data\imdb-datasetfile_crew_movie.tsv	1309875	904258	vRfGfYb	2021-04-27 22:00:05.947
3	IMDB movie	C:\INFO\T370\delsr\spring2021\cdlne\FinalSource\Feed\data\imdb-datasetfile_principals_movie.tsv	714500	714500	ZMqBx	2021-04-27 22:19:44.603
4	IMDB movie	C:\INFO\T370\delsr\spring2021\cdlne\FinalSource\Feed\data\imdb-datasetfile_ratings_movie.tsv	273403	373403	KsPjY	2021-04-27 22:27:24.033

**data consistency:** data is loaded from tsv files, to make consistent data we have used tReplace component to clean the data.

**reject:** for rejects we are checking trim columns and check each row against schema structure, we have not received any rejects for this table.

**structural integration processes:** initially we are truncating the table, we are doing

# Staging Screen Shots- Talend

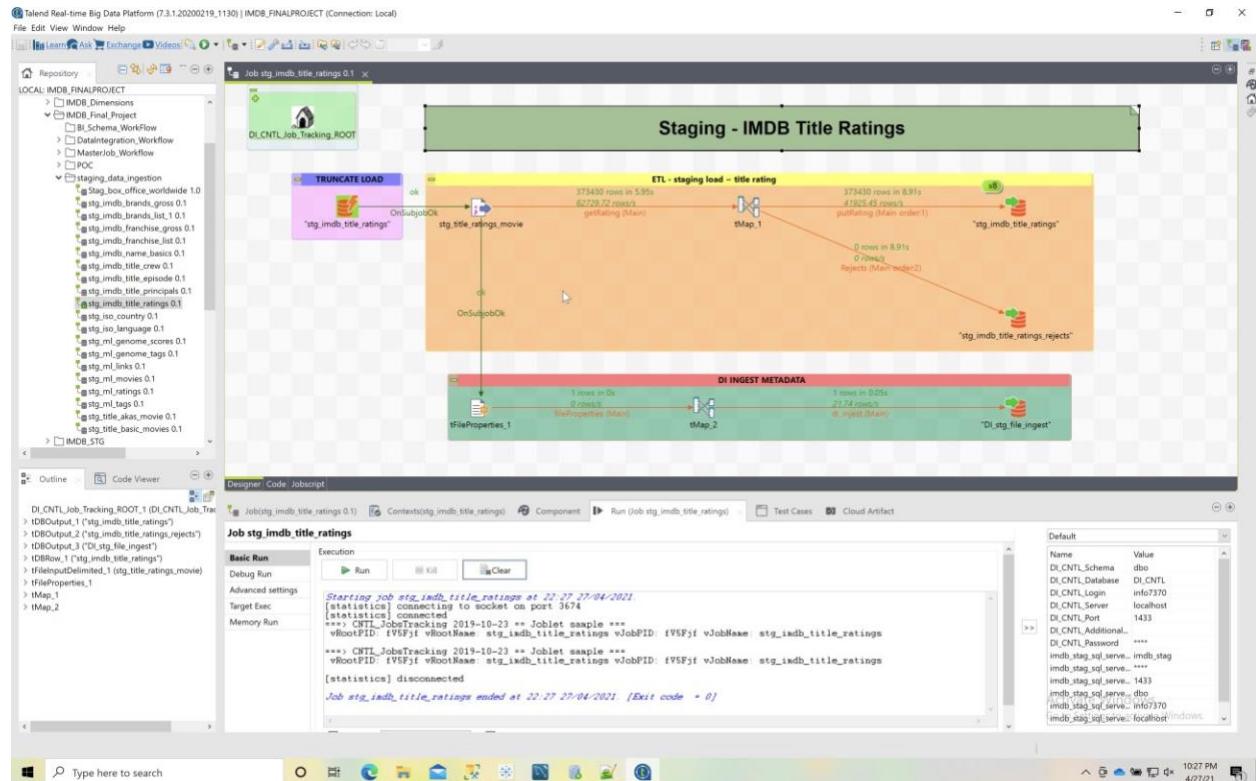
data cleansing using tReplace component.

For performance we are using bulk components,

For DI standard we are deleting bulk file which is created in the process

We are also capturing DI ingest metadata and storing in the table which is bonus point.

## 10. stg\_imdb\_title\_ratings & Rejects



# Staging Screen Shots- Talend

SQLQuery1.sql - localhost.imdb\_stag (Info7370 (54)) - Microsoft SQL Server Management Studio (Administrator)

```
select * from [dbo].[stg_imdb_title_ratings]
```

Results Messages

	id	start_time	averageRating	numVotes	GOR_SK	Di_JobID	Di_CreatedDT
1	6566	2000001	5.7	1035	6	N/SP	2021-04-27 22:27:17.817
2	6569	2000002	6.1	198	6	N/SP	2021-04-27 22:27:17.817
3	6538	2000003	6.5	1328	6	N/SP	2021-04-27 22:27:17.827
4	6540	2000004	5.2	20	6	N/SP	2021-04-27 22:27:17.827
5	6538	2000005	6.1	2108	6	N/SP	2021-04-27 22:27:17.827
6	6539	2000006	5.3	116	6	N/SP	2021-04-27 22:27:17.827
7	6541	2000007	5.9	548	6	N/SP	2021-04-27 22:27:17.827
8	6541	2000009	5.4	1792	6	N/SP	2021-04-27 22:27:17.827
9	6542	2000009	5.9	158	6	N/SP	2021-04-27 22:27:17.827
10	6543	2000010	5.9	489	6	N/SP	2021-04-27 22:27:17.827
11	6544	2000011	5.2	259	6	N/SP	2021-04-27 22:27:17.827
12	6545	2000012	7.4	10198	6	N/SP	2021-04-27 22:27:17.827
13	6546	2000013	5.7	1060	6	N/SP	2021-04-27 22:27:17.827
14	6547	2000014	7.2	4408	6	N/SP	2021-04-27 22:27:17.827
15	6548	2000015	6.1	808	6	N/SP	2021-04-27 22:27:17.827
16	6549	2000016	5.9	1188	6	N/SP	2021-04-27 22:27:17.827
17	6550	2000017	4.7	208	6	N/SP	2021-04-27 22:27:17.827
18	6552	2000018	5.3	474	6	N/SP	2021-04-27 22:27:17.827
19	6554	2000019	5.3	18	6	N/SP	2021-04-27 22:27:17.827
20	6555	2000020	5.5	459	6	N/SP	2021-04-27 22:27:17.827
21	6558	2000022	5.1	879	6	N/SP	2021-04-27 22:27:17.827
22	6560	2000023	5.7	1125	6	N/SP	2021-04-27 22:27:17.827
23	6561	2000024	4.8	28	6	N/SP	2021-04-27 22:27:17.827
24	6563	2000025	4.2	29	6	N/SP	2021-04-27 22:27:17.827
25	6564	2000026	5.7	1300	6	N/SP	2021-04-27 22:27:17.827
26	6565	2000027	5.6	917	6	N/SP	2021-04-27 22:27:17.827
27	6566	2000028	5.1	148	6	N/SP	2021-04-27 22:27:17.827
28	6568	2000029	5.9	2829	6	N/SP	2021-04-27 22:27:17.827

Activate Windows  
Go to Settings to activate Windows  
localhost (15.0 RTM) - info7370 (54) - imdb\_stag - 00:00:02 - 373,430 rows

## TIME

SQLQuery4.sql - loc...tag (Info7370 (55))\*

SQLQuery3.sql - loc...cntl (Info7370 (53))\*

SQLQuery1.sql - loc...tag (Info7370 (54))\*

SELECT  
job\_name,  
is\_root\_job,  
root\_name,  
job\_status,  
[Duration (minutes)],  
job\_finish\_time,  
message\_type,  
project,  
job\_version  
FROM \_DI\_CNTL dbo.job\_stats\_vw  
where job\_name = 'stg\_imdb\_title\_ratings'

Results Messages

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_imdb_title_ratings	Y	stg_imdb_title_ratings	success	0.16	2021-04-27 22:24.033	end	IMDB_FINALPROJECT	0.1

# Staging Screen Shots- Talend

## DI INJECT:

The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left lists databases, including AdventureWorksDW2019, Chinook, dlt\_crm, dlt\_crmDataModel, dlt\_DI, imdb\_stag, and several system tables like Database Diagrams, Tables, External Tables, and Graph Tables. The central pane displays a query result grid for a table named DI\_stg\_file\_ingest. The grid has columns: Di\_ag\_file\_inject, File\_Group, FileName, FileOffset, RowInserted, Di\_JobID, and Di\_CreateDate. There are four rows of data, each corresponding to a file named C:\INFO\7370\dlt-spring2021\cdlive\FinalSource\Files\data-imdb-datasetname\_basics\_movie.tsv. The bottom status bar indicates the query was executed successfully at 10:31 PM on 4/27/21.

Di_ag_file_inject	File_Group	FileName	FileOffset	RowInserted	Di_JobID	Di_CreateDate
1	1	IMDB movie	C:\INFO\7370\dlt-spring2021\cdlive\FinalSource\Files\data-imdb-datasetname_basics_movie.tsv	8781803	8781803	QKvVgX 2021-04-27 21:57:23.930
2	2	IMDB movie	C:\INFO\7370\dlt-spring2021\cdlive\FinalSource\Files\data-imdb-datasetname_crew_movie.tsv	1389875	90428	VHqG7Y 2021-04-27 22:00:05.847
3	3	IMDB movie	C:\INFO\7370\dlt-spring2021\cdlive\FinalSource\Files\data-imdb-datasetname_principals_movie.tsv	7741600	7741600	ZBqyA 2021-04-27 22:18:44.683
4	4	IMDB movie	C:\INFO\7370\dlt-spring2021\cdlive\FinalSource\Files\data-imdb-datasetname_retag_movie.tsv	273620	273620	KSPY 2021-04-27 22:27:24.033

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

**reject:** for rejects we are checking trim columns and check each row against schema structure, we have not received any rejects for this table.

**structural integration processes:** initially we are truncating the table.

For Performance we have enabled parallelization to speed up the load

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 11. stg\_iso\_country

**Talend Real-time Big Data Platform (7.3.1.20200219\_1130) IMDB\_FINALPROJECT (Connection: Local)**

**Job stg\_iso\_country 0.1**

```

graph LR
    Start((Start)) --> Truncate[Truncate Load]
    Truncate --> Load[stg_iso_country]
    Load --> OnSubjobOk1[OnSubjobOk]
    OnSubjobOk1 --> ETL[ETL - staging load - iso country]
    ETL --> OnSubjobOk2[OnSubjobOk]
    OnSubjobOk2 --> DI[DI INGEST METADATA]
    DI --> OnSubjobOk3[OnSubjobOk]
    OnSubjobOk3 --> End((End))
    
```

**Designer Code: Jobscript**

```

Job stg_iso_country
Basic Run
Execution
Run | Clear
Starting job stg_iso_country at 22:33:27/04/2021
[statistics] connecting to socket on port 3927
[statistics] connected
--> JobTracking 2019-10-23 ** Joblet saplio ***
<--> JobTracking 2019-10-23 ** Joblet saplio ***
<--> CNTL_JobTracking 2019-10-23 ** Joblet saplio ***
<--> CNTL_JobTracking 2019-10-23 ** Joblet saplio ***
[statistics] disconnected
Job stg_iso_country ended at 22:33:27/04/2021. [Exit code = 0]

```

**Default**

Name	Value
DI_CNTL_Schema	dbo
DI_CNTL_Database	DI_CNTL
DI_CNTL_Login	info7370
DI_CNTL_Server	localhost
DI_CNTL_Port	1433
DI_CNTL_Password	****
imdb_stg_file_ingest	steg
imdb_stg_server	***
imdb_stg_sql_server	1433
imdb_stg_sql_server_dbo	imdb_stg
imdb_stg_sql_server_info7370	info7370
imdb_stg_sql_server_localhost	localhost

**SQL Server Management Studio (Administrator)**

Object Explorer

Results

```

select * from [dbo].[stg_iso_country]

```

Table: stg\_iso\_country

Country_SK	country_name	alpha_2	alpha_3	country_code	iso_3166_2	region	sub_region	intermediate_region	region_code	sub_region_code	intermediate_region_code	SOR_SK	Di_JobID	Di_Create_DT
1	Afghanistan	AF	AFG	4	ISO 3166-2:AF	Asia	Southern Asia	142	34	NULL	7	70519	2021-04-27 22:33:41 800	
2	Albania	AL	ALB	8	ISO 3166-2:AL	Europe	Eastern Europe	190	154	NULL	7	70519	2021-04-27 22:33:41 800	
3	Algeria	DZ	DZA	12	ISO 3166-2:DZ	Africa	Southern Africa	193	19	NULL	7	70519	2021-04-27 22:33:41 800	
4	Angola	AO	AGO	24	ISO 3166-2:AO	Africa	Central Africa	2	19	NULL	7	70519	2021-04-27 22:33:41 800	
5	Anguilla	AI	ASM	16	ISO 3166-2:AI	Oceania	Polynesia	9	61	NULL	7	70519	2021-04-27 22:33:41 800	
6	Anguilla	AD	AND	20	ISO 3166-2:AD	Europe	Western Europe	158	19	NULL	7	70519	2021-04-27 22:33:41 800	
7	Anguilla	AO	AGO	24	ISO 3166-2:AO	Africa	Sub-Saharan Africa	2	202	17	7	70519	2021-04-27 22:33:41 800	
8	Anguilla	AI	ASM	16	ISO 3166-2:AI	Oceania	Polynesia	19	419	29	7	70519	2021-04-27 22:33:41 800	
9	Anguilla	AO	AGO	24	ISO 3166-2:AO	Africa	Sub-Saharan Africa	142	142	142	7	70519	2021-04-27 22:33:41 800	
10	Anguilla	AO	AGO	28	ISO 3166-2:AO	Americas	Latin America and the Caribbean	19	419	29	7	70519	2021-04-27 22:33:41 800	
11	Anguilla	AI	ASM	32	ISO 3166-2:AI	Oceania	Polynesia	19	5	5	7	70519	2021-04-27 22:33:41 800	
12	Anguilla	AM	ABW	92	ISO 3166-2:AM	Europe	Western Europe	142	142	142	7	70519	2021-04-27 22:33:41 800	
13	Anguilla	AW	ABW	93	ISO 3166-2:AW	Americas	Latin America and the Caribbean	19	419	29	7	70519	2021-04-27 22:33:41 800	
14	Anguilla	AU	AUS	36	ISO 3166-2:AU	Oceania	Australia and New Zealand	9	53	NULL	7	70519	2021-04-27 22:33:41 800	
15	Anguilla	AT	AUT	40	ISO 3166-2:AT	Europe	Western Europe	190	193	NULL	7	70519	2021-04-27 22:33:41 800	
16	Anguilla	AZ	AZE	31	ISO 3166-2:AZ	Africa	Eastern Africa	142	145	NULL	7	70519	2021-04-27 22:33:41 800	
17	Anguilla	BS	BHS	44	ISO 3166-2:BS	Americas	Latin America and the Caribbean	19	419	29	7	70519	2021-04-27 22:33:41 800	
18	Anguilla	BH	BHR	48	ISO 3166-2:BH	Americas	Western Asia	142	145	NULL	7	70519	2021-04-27 22:33:41 800	
19	Anguilla	BD	BDE	50	ISO 3166-2:BD	Americas	South Asia	142	34	NULL	7	70519	2021-04-27 22:33:41 800	
20	Anguilla	BB	BRB	52	ISO 3166-2:BB	Americas	Latin America and the Caribbean	19	419	29	7	70519	2021-04-27 22:33:41 800	
21	Anguilla	BY	BLR	112	ISO 3166-2:BY	Europe	Eastern Europe	150	151	NULL	7	70519	2021-04-27 22:33:41 800	
22	Anguilla	BR	BRA	64	ISO 3166-2:BR	Americas	South America	150	151	NULL	7	70519	2021-04-27 22:33:41 800	
23	Anguilla	BZ	BZL	64	ISO 3166-2:BZ	Americas	Latin America and the Caribbean	19	419	13	7	70519	2021-04-27 22:33:41 800	
24	Anguilla	BJ	BEN	204	ISO 3166-2:BJ	Africa	Sub-Saharan Africa	2	202	11	7	70519	2021-04-27 22:33:41 800	
25	Anguilla	BM	BMR	64	ISO 3166-2:BM	Americas	Caribbean	19	21	NULL	7	70519	2021-04-27 22:33:41 800	
26	Anguilla	BT	BTN	64	ISO 3166-2:BT	Asia	Southern Asia	142	34	NULL	7	70519	2021-04-27 22:33:41 800	
27	Anguilla	BO	BOL	68	ISO 3166-2:BO	Americas	Latin America and the Caribbean	19	419	5	7	70519	2021-04-27 22:33:41 800	
28	Anguilla	BW	BES	535	ISO 3166-2:BW	Americas	Caribbean	19	419	29	7	70519	2021-04-27 22:33:41 800	

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

TIME:

```

SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where   job_name = 'stg_iso_country'
  
```

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_iso_country	Y	stg_iso_country	success	0.01	2021-04-27 22:33:41.940	end	IMDB_FINALPROJECT	0.1

DI INJECT:

```

select * from DI_stg_file_ingest
  
```

File_Group	Filname	Status	Reasoned	D_JobID	D_CreatedDate
1	C:\INFO\7370\dbs\spring2021\ctvne\finalSource\Franchise\midb-datasetname_basics_movies.m	1	0	8741803	2021-04-27 21:53:23.830
2	C:\INFO\7370\dbs\spring2021\ctvne\finalSource\Franchise\midb-datasetname_crew_movies.m	1	0	1308175	2021-04-27 22:05:547
3	C:\INFO\7370\dbs\spring2021\ctvne\finalSource\Franchise\midb-datasetname_principals_movies.m	1	0	7741900	2021-04-27 22:19:44.603
4	C:\INFO\7370\dbs\spring2021\ctvne\finalSource\Franchise\midb-datasetname_settings_movies.m	1	0	373420	2021-04-27 22:27:24.033
5	ISO REFERENCE courses	1	0	249	2021-04-27 22:38:41.927

Activate Windows  
Go to Settings to activate Windows  
localhost (15.0 RTM) info7370 (55) imbd\_stag 00:00:00 5 rows

# Staging Screen Shots- Talend

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

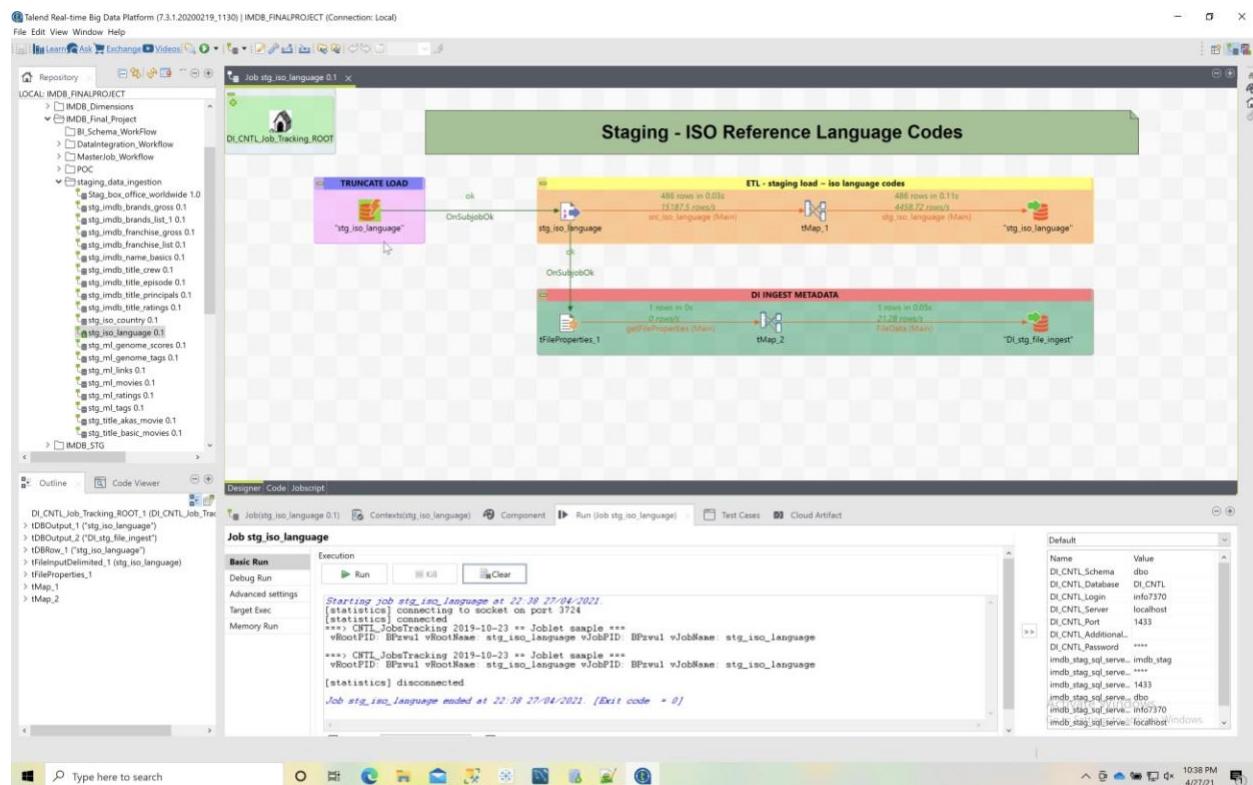
**reject:** No rejects required for this table.

**structural integration processes:** initially we are truncating the table.

For Performance we have enabled parallelization to speed up the load

We are also capturing DI ingest metadata and storing in the table which is bonus point.

## 12. stg\_iso\_language



# Staging Screen Shots- Talend

SQLQuery1.sql - localhost.imdb\_stag (info7370 (54)) - Microsoft SQL Server Management Studio (Administrator)

```
select * from [dbo].[stg_iso_language]
```

Language_SK	enpld_1	enpld_2	enpld_3	enpld_4	Language_Name	SOP_SK	DI_JobID	DI_Created
1	azb	ab	ab		Azeri	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
2	azk	ak	ak		Azerbaijan	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
3	ace	ac	ac		Acadmic	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
4	ada	ad	ad		Adangme	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
5	ady	adghe	adghe		Adyghe, Adygei	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
6	af	afghan	afghan		Afghan language	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
7	afz	afzani	afzani		Afzani	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
8	an	an-	an-		Afhaanian	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
9	an-	an-	an-		An-	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
10	ara	ar-	ar-		Arabic	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
11	aka	ak-	ak-		Akan	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
12	ak-	ak-	ak-		Akkadian	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
13	ala	al-	al-		Albanian	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
14	ale	al-	al-		Ala-	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
15	alg	al-	al-		Algnpran languages	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
16	am-	at-	at-		Albanian	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
17	am-	am-	am-		Am-	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
18	arg	arg	arg		English, Old (ca 450-1100)	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
19	arp	arp-	arp-		Angika	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
20	arp-	arp-	arp-		Angika languages	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
21	ara	ar-	ar-		Arabic	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
22	arc	ar-	ar-		Official Aramaic (700-300 BCE) Imperial Aramean	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
23	arg	ar-	ar-		Armenian	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
24	am-	am-	am-		Armenian	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
25	am-	am-	am-		Mapudungun, Maputun	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
26	arp	arp-	arp-		Arpaki	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
27	ar-	ar-	ar-		Aruban languages	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713
28	amr	am-	am-		Arwae	BPM001	2021-04-27 22:38:24.713	2021-04-27 22:38:24.713

Activate Windows  
Go to Settings to activate Windows  
localhost (15.0 RTM) info7370 (54) imdb\_stag 00:00:00 - 486 rows

TIME:

SQLQuery3.sql - localhost.di\_cntl (info7370 (53)) - Microsoft SQL Server Management Studio (Administrator)

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM
    DI_CNTL.dbo.job_stats_vw
where
    job_name = 'stg_iso_language'
```

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_iso_language	Y	stg_iso_language	success	0.02	2021-04-27 22:38:24.820	end	IMDB_FINALPROJECT	0.1

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

## DI INJECT:

The screenshot shows the Microsoft SQL Server Management Studio interface. In the Object Explorer, the database 'imdb\_stag' is selected. In the center pane, a query window displays the following SQL code:

```
select * from DI_stg_file_Ingest
```

The results pane shows a table with 6 rows, representing the metadata for the 'IMDb movie' dataset. The columns are:

FileRow	RowNumber	Di_JobID	Di_CreateDate
1	1	IMDb movie	C:\INFO\7370-ds-spring2021-cdm\FinalSource\Files\ds-imdb-datasetname_basic_movie.tsv
2	2	IMDb movie	C:\INFO\7370-ds-spring2021-cdm\FinalSource\Files\ds-imdb-datasettitle_crew_movie.tsv
3	3	IMDb movie	C:\INFO\7370-ds-spring2021-cdm\FinalSource\Files\ds-imdb-datasettitle_genome_movie.tsv
4	4	IMDb movie	C:\INFO\7370-ds-spring2021-cdm\FinalSource\Files\ds-imdb-datasettitle_rating_movie.tsv
5	5	ISO REFERENCE sources	C:\INFO\7370-ds-spring2021-cdm\FinalSource\Files\ds_iso_lrcources_is_all.tsv
6	6	ISO REFERENCE language-codes	C:\INFO\7370-ds-spring2021-cdm\FinalSource\Files\ds_is_language-codes_is.tsv

At the bottom of the results pane, it says "Query executed successfully."

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

**reject:** No rejects required for this table.

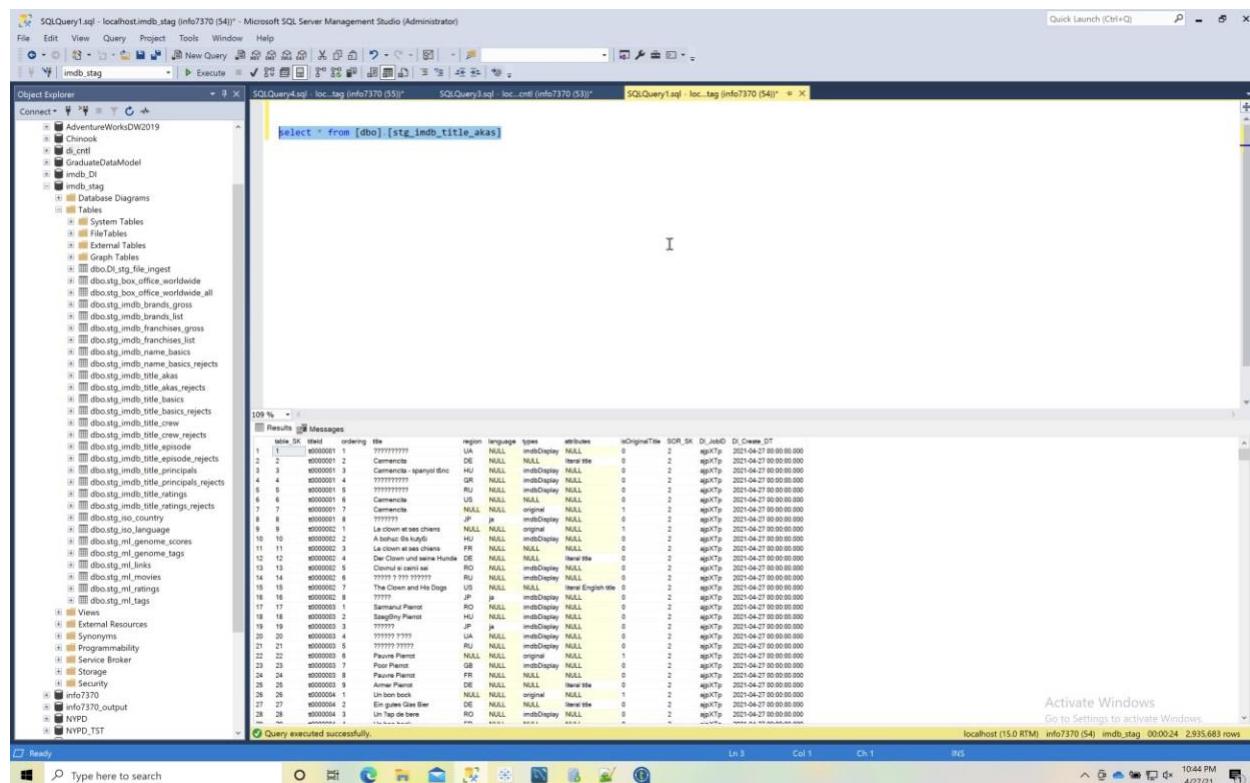
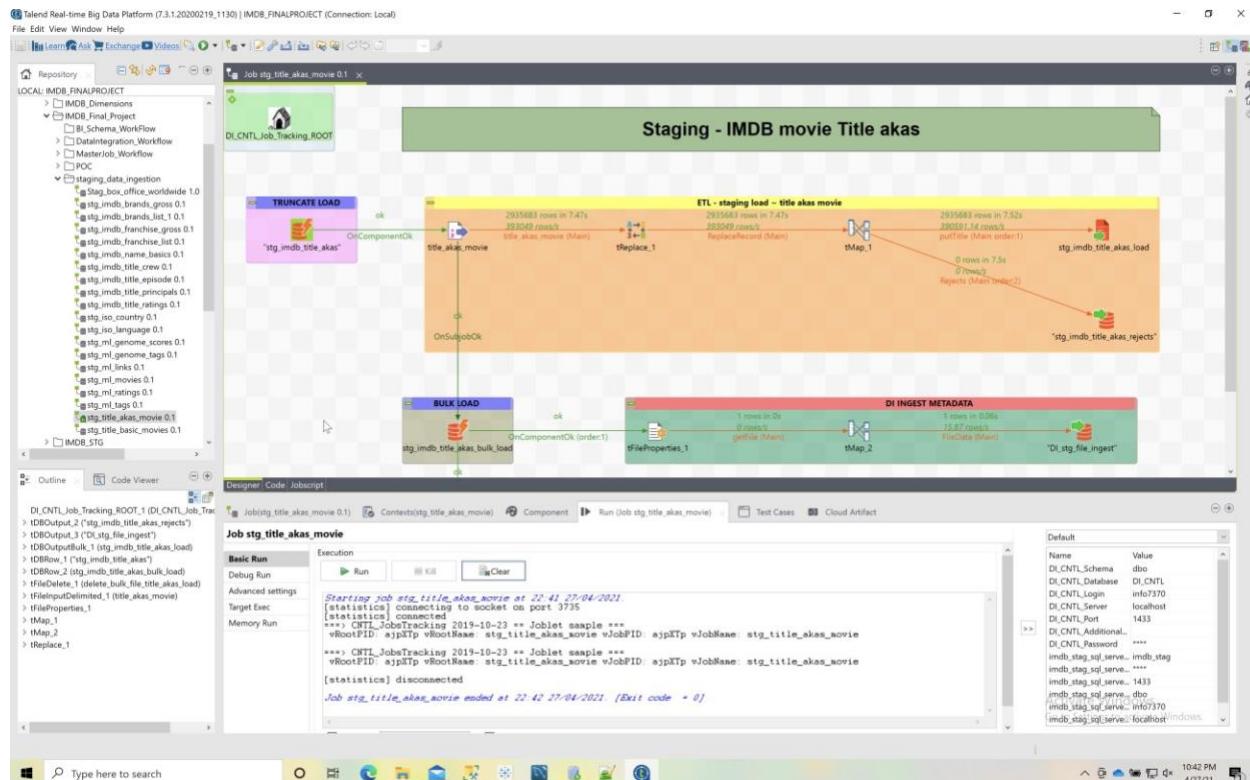
**structural integration processes:** initially we are truncating the table.

For Performance we have enabled parallelization to speed up the load

We are also capturing DI inject metadata and storing in the table which is bonus point.

## 13. stg\_title\_akas\_movie & Rejects

# Staging Screen Shots- Talend



Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

TIME:

The screenshot shows a SQL Server Management Studio (SSMS) interface. The Object Explorer on the left lists databases like AdventureWorksDW2019, Chinook, di\_cntl, GraduateDataModel, imdb\_DL, and imdb\_stag. The di\_cntl database is selected. The center pane displays a T-SQL query:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_title_akas_movie'
```

The results pane at the bottom shows one row of data:

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
1 stg_title_akas_movie	y	stg_title_akas_movie	success	0.35	2021-04-27 22:42:08.587	end	IMDB_FINALPROJECT	0.1

# Staging Screen Shots- Talend

## DI INJECT:

The screenshot shows a Microsoft SQL Server Management Studio window. The Object Explorer on the left shows a database named 'imdb\_stag' containing various tables and objects. The main pane displays the results of a query:

```
select * from DI_stg_file_ingest
```

The results grid shows the following data:

FileRowID	FileProcessed	Di_JobID	Di_CreatedOn	
1	C:\INFO\7370-debs-spring2021\cdm\FinalSource\Files\files\imdb-datasette\titles.movie.tsv	1308176	VRSPYR	2021-04-27 22:00:59,847
2	IMDb movie	774100	ZJBLB	2021-04-27 22:19:44,603
3	IMDb movie	379430	YVSPYR	2021-04-27 22:24:03,233
4	IMDb movie	249	TYSPYR	2021-04-27 22:30:47,827
5	ISO REFERENCE countries	488	BPewc1	2021-04-27 22:34:24,820
6	ISO REFERENCE language-codes	2035683	apXtp	2021-04-27 22:42:08,570
7	IMDb movie			

Below the results grid, a message says "Query executed successfully."

**data consistency:** data is loaded from tsv files, to make consistent data we have used tReplace component to clean the data and sk col

**reject:** for rejects we are checking trim columns and check each row against schema structure, we have not received any rejects for this table.

**structural integration processes:** initially we are truncating the table, we are doing data cleansing using tReplace component.

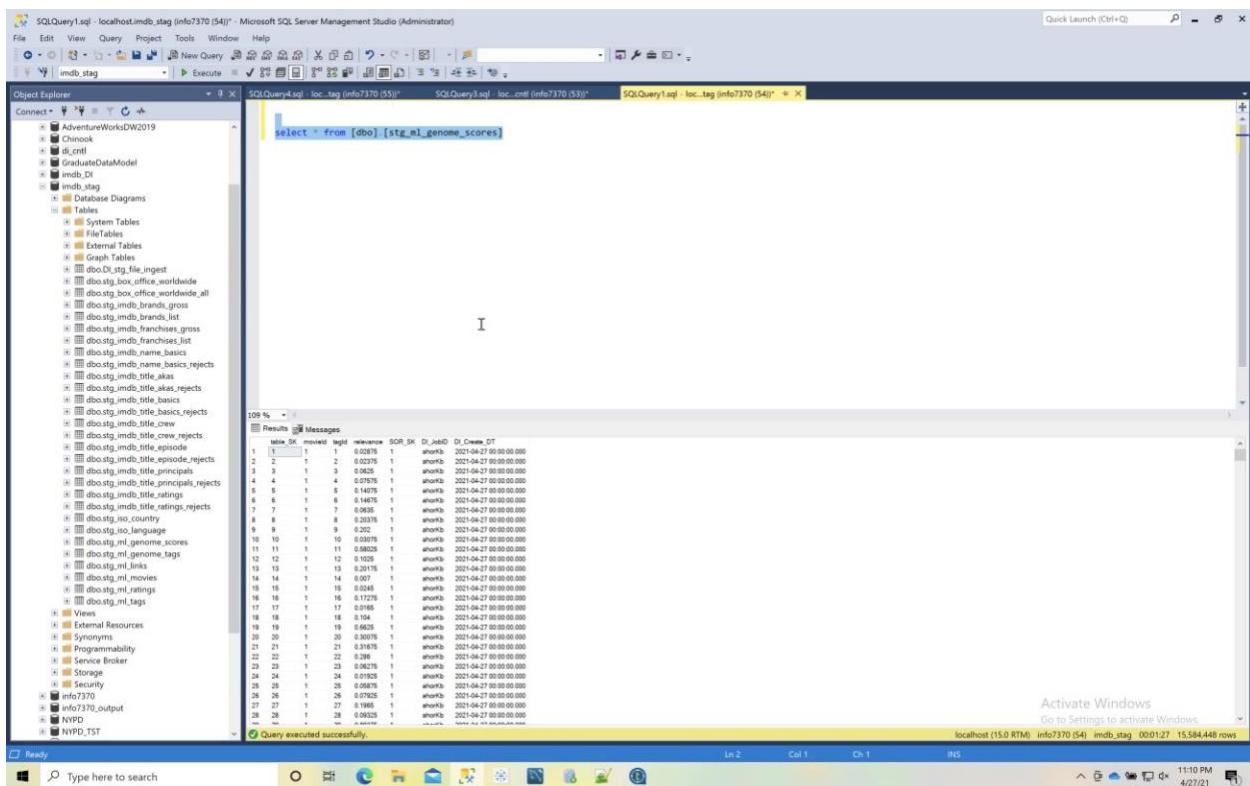
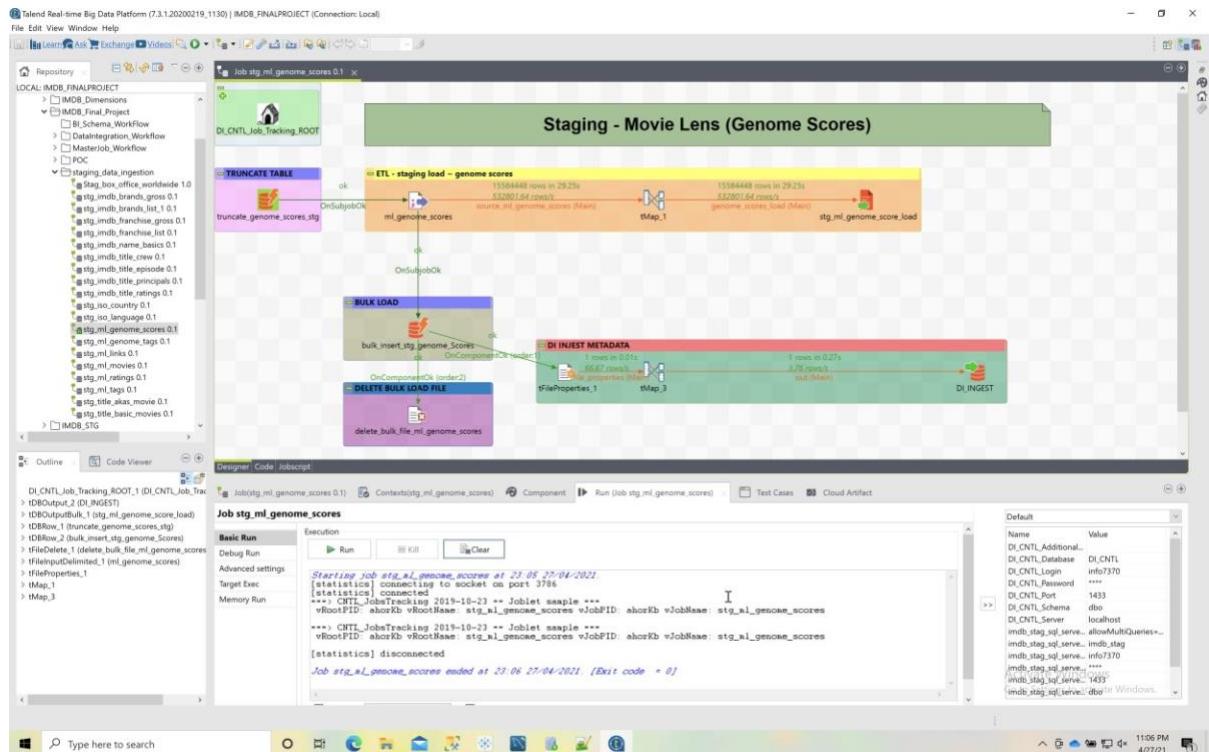
For performance we are using bulk components,

For DI standard we are deleting bulk file which is created in the process

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 14. stg\_ml\_genome\_scores



# Staging Screen Shots- Talend

TIME:

```

SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_ml_genome_scores'
  
```

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_ml_genome_scores	Y	stg_ml_genome_scores	success	1.31	2021-04-27 23:06:41.037	end	IMDB_FINALPROJECT	0.1

DI INJECT:

```

select * from DI_stg_file_ingest
  
```

Di_stg_file_ingest	File_Group	File_name
1	IMDB	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\cdm\imdb\datasette\_basic_movie.tsv
2	IMDB	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\cdm\imdb\datasette\_crew_movie.tsv
3	IMDB	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\cdm\imdb\datasette\_principal_movie.tsv
4	IMDB	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\cdm\imdb\datasette\_titles_movie.tsv
5	ISO REFERENCE	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\cdm\iso\iso_countries.iso
6	ISO REFERENCE	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\cdm\iso\iso_language-codes.iso
7	IMDB	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\cdm\imdb\datasette\_akas_movie.tsv
8	Movie Lens	C:\INFO\7370\dbs\spring2021\cdm\final\source\file\ml\genome-scores.csv

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

# Staging Screen Shots- Talend

**reject:** No reject required for this table.

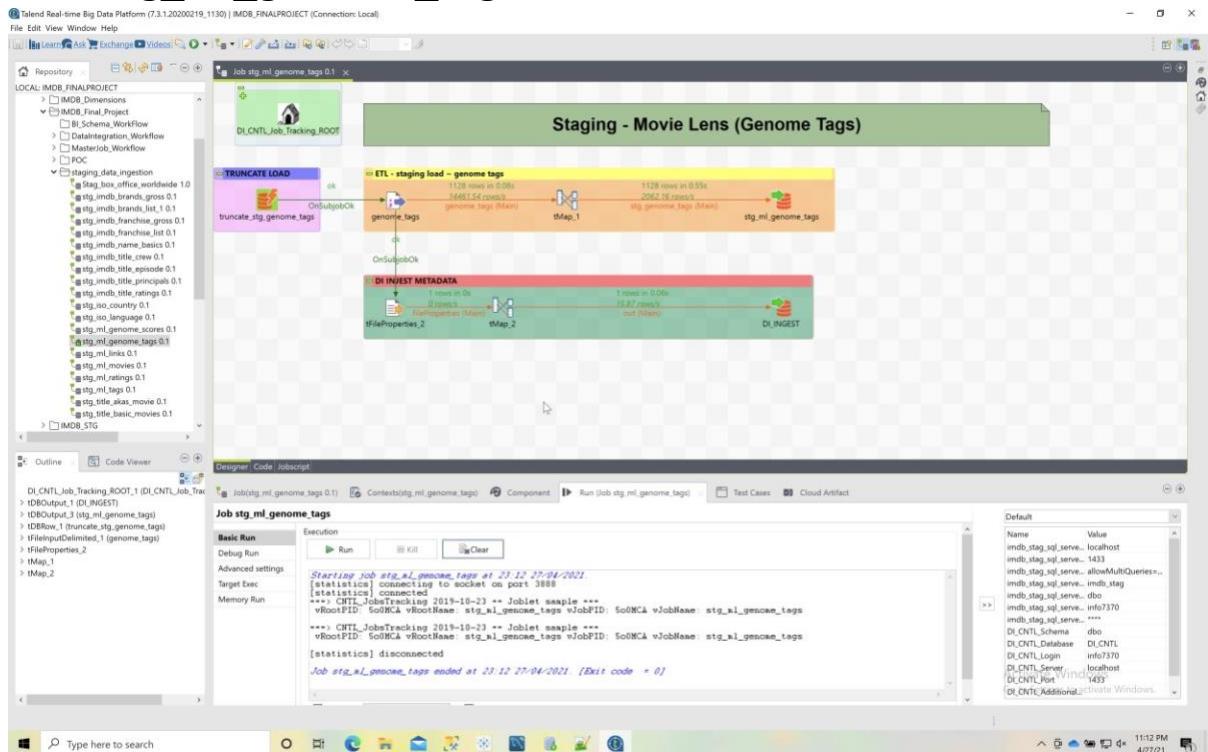
**structural integration processes:** initially we are truncating the table

For performance we are using bulk components,

For DI standard we are deleting bulk file which is created in the process

We are also capturing DI ingest metadata and storing in the table which is bonus point.

## 15. stg\_ml\_genome\_tags



# Staging Screen Shots- Talend

```
select * from [dbo].[stg_mi_genome_tags]
```

row_sk	tag_id	tag	SDR_SK	DI_JobID	DI_Created_DT
1	1	007 (series)	1	S00MCA	2021-04-27 23:12:40.143
2	2	007 (series)	1	S00MCA	2021-04-27 23:12:40.143
3	3	18th century	1	S00MCA	2021-04-27 23:12:40.143
4	4	1950s	1	S00MCA	2021-04-27 23:12:40.143
5	5	1930s	1	S00MCA	2021-04-27 23:12:40.143
6	6	1950s	1	S00MCA	2021-04-27 23:12:40.143
7	7	1960s	1	S00MCA	2021-04-27 23:12:40.143
8	8	1970s	1	S00MCA	2021-04-27 23:12:40.143
9	9	1980s	1	S00MCA	2021-04-27 23:12:40.143
10	10	19th century	1	S00MCA	2021-04-27 23:12:40.143
11	11	3d	1	S00MCA	2021-04-27 23:12:40.143
12	12	70mm	1	S00MCA	2021-04-27 23:12:40.143
13	13	8mm	1	S00MCA	2021-04-27 23:12:40.143
14	14	911	1	S00MCA	2021-04-27 23:12:40.143
15	15	aereman	1	S00MCA	2021-04-27 23:12:40.143
16	16	aereman studios	1	S00MCA	2021-04-27 23:12:40.143
17	17	adult	1	S00MCA	2021-04-27 23:12:40.143
18	18	abund	1	S00MCA	2021-04-27 23:12:40.143
19	19	actor	1	S00MCA	2021-04-27 23:12:40.143
20	20	action packed	1	S00MCA	2021-04-27 23:12:40.143
21	21	adaption	1	S00MCA	2021-04-27 23:12:40.143
22	22	adapted from book	1	S00MCA	2021-04-27 23:12:40.143
23	23	adapted from comic	1	S00MCA	2021-04-27 23:12:40.143
24	24	adapted from game	1	S00MCA	2021-04-27 23:12:40.143
25	25	addiction	1	S00MCA	2021-04-27 23:12:40.143
26	26	adolescence	1	S00MCA	2021-04-27 23:12:40.143
27	27	adult	1	S00MCA	2021-04-27 23:12:40.143
28	28	adultery	1	S00MCA	2021-04-27 23:12:40.143

Query executed successfully.

TIME:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM
    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_mi_genome_tags'
```

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
stg_mi_genome_tags	Y	stg_mi_genome_tags	success	0.02	2021-04-27 23:12:40.487	end	IMDB_FINALPROJECT	0.1

# Staging Screen Shots- Talend

## DI INJECT:

The screenshot shows the Microsoft SQL Server Management Studio interface. In the Object Explorer on the left, a database named 'imdb\_stag' is selected. In the center, a query window displays the following T-SQL code:

```
select * from DI_stg_file_ingest
```

The results grid shows a table with 9 rows, titled 'Results'. The columns are 'FileIndex', 'RowNumbered', 'Di.JobID', 'Di.CreateDate', 'Filename', and 'FileGroup'. The data is as follows:

FileIndex	RowNumbered	Di.JobID	Di.CreateDate	Filename	FileGroup
1	1	8791803	QInqK	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\cds-imdb-datasetname_basic_movie.tsv	IMDB movie
2	2	9542936	VtRgYb	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\cds-imdb-datasette_crew_movie.tsv	IMDB movie
3	3	7193005	ZdhuJ	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\cds-imdb-datasette_credits_movie.tsv	IMDB movie
4	4	373430	PvPFy	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\cds-imdb-datasette_genre_movie.tsv	IMDB movie
5	5	249	TsHv	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\cds-imdb-datasette_genre_movie.tsv	IMDB movie
6	6	480	SpPwU	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\cds-imdb-datasette_genre_movie.tsv	IMDB movie
7	7	2935683	apXtp	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\cds-imdb-datasette_genre_movie.tsv	IMDB movie
8	8	15594448	shnkb	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\genome-scores.csv	Movie Lens genome-scores
9	9	1128	SeMCA	C:\INFO\7370-debt-spring2021-cdm\FinalSource\Files\genome-tags.csv	Movie Lens genome-tags

At the bottom of the results grid, it says 'Query executed successfully.'

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

**reject:** No rejects required for this table.

**structural integration processes:** initially we are truncating the table.

For Performance we have enabled parallelization to speed up the load

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 16. stg\_ml\_links

Talend Real-time Big Data Platform (7.3.1.20200219\_1130) | IMDB\_FINALPROJECT (Connection: Local)

**Staging - Movie Lens (Links)**

Job stg\_ml\_links

Basic Run

```

Starting job stg_ml_links at 23-16-27-04-2021.
[statistics] connecting to socket on port 3770
*** CNTL_JobTracking 2019-10-23 -- Joblet sample ***
vRootID: 63Ra0R vRootName: stg_ml_links vJobID: 63Ra0R vJobName: stg_ml_links
*** CNTL_JobTracking 2019-10-23 -- Joblet sample ***
vRootID: 63Ra0R vRootName: stg_ml_links vJobID: 63Ra0R vJobName: stg_ml_links
[statistics] disconnected
Job stg_ml_links ended at 23-16-27-04-2021. [Exit code = 0]

```

Default

Name	Value
DL_CNTL_Schema	dbo
DL_CNTL_Database	DL_CNTL
DL_CNTL_Login	info7370
DL_CNTL_Server	localhost
DL_CNTL_Port	1433
DL_CNTL_Authentication	Windows Authentication
DL_CNTL_Password	****
imdb_stag_sql_seve...	imdb_stag_sql_seve...
imdb_stag_sql_seve...	****
imdb_stag_sql_seve...	1433
imdb_stag_sql_seve...	dbo
imdb_stag_sql_seve...	info7370
imdb_stag_sql_seve...	localhost

SQL Server Management Studio (Administrator)

Object Explorer

Results

```

select * from [dbo].[stg_ml_links]

```

109 %

id	SK	merged	imdb	imdb_name	SOR_SK	D_JobID	D_Created_DT
1	1	0114709	862	00114709	1	63Ra0R	2021-04-27 23:16:48.373
2	2	0113497	884	00113497	1	63Ra0R	2021-04-27 23:16:48.373
3	3	0113228	1560	00113228	1	63Ra0R	2021-04-27 23:16:48.373
4	4	0113229	1560	00113229	1	63Ra0R	2021-04-27 23:16:48.373
5	5	0113041	11865	00113041	1	63Ra0R	2021-04-27 23:16:48.373
6	6	0112277	949	00112277	1	63Ra0R	2021-04-27 23:16:48.373
7	7	0112278	1000	00112278	1	63Ra0R	2021-04-27 23:16:48.373
8	8	0112302	4532	00112302	1	63Ra0R	2021-04-27 23:16:48.373
9	9	011476	3091	0011476	1	63Ra0R	2021-04-27 23:16:48.373
10	10	0112348	9087	00112348	1	63Ra0R	2021-04-27 23:16:48.373
11	11	0112348	9087	00112348	1	63Ra0R	2021-04-27 23:16:48.373
12	12	0113066	12116	00113066	1	63Ra0R	2021-04-27 23:16:48.373
13	13	0113067	12789	00113067	1	63Ra0R	2021-04-27 23:16:48.373
14	14	0113987	1085	00113987	1	63Ra0R	2021-04-27 23:16:48.373
15	15	0121760	1408	001121760	1	63Ra0R	2021-04-27 23:16:48.373
16	16	0121761	1408	001121761	1	63Ra0R	2021-04-27 23:16:48.373
17	17	0114388	4584	00114388	1	63Ra0R	2021-04-27 23:16:48.373
18	18	0113101	1	00113101	1	63Ra0R	2021-04-27 23:16:48.373
19	19	0113102	878	00113102	1	63Ra0R	2021-04-27 23:16:48.373
20	20	0113845	11817	00113845	1	63Ra0R	2021-04-27 23:16:48.373
21	21	0113161	812	00113161	1	63Ra0R	2021-04-27 23:16:48.373
22	22	0113162	812	00113162	1	63Ra0R	2021-04-27 23:16:48.373
23	23	0113401	9691	00113401	1	63Ra0R	2021-04-27 23:16:48.373
24	24	0114168	12968	001114168	1	63Ra0R	2021-04-27 23:16:48.373
25	25	0114169	12968	001114169	1	63Ra0R	2021-04-27 23:16:48.373
26	26	0140557	1620	001140557	1	63Ra0R	2021-04-27 23:16:48.373
27	27	0114011	828	00114011	1	63Ra0R	2021-04-27 23:16:48.373
28	28	0114012	828	00114012	1	63Ra0R	2021-04-27 23:16:48.373
29	29	0114013	828	00114013	1	63Ra0R	2021-04-27 23:16:48.373

Query executed successfully.

TIME:

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

```

SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM   DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_mi_links'
  
```

	job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
1	stg_mi_links	Y	stg_mi_links	success	0.04	2021-04-27 23:16:50.297	end	IMDB_FINALPROJECT	0.1

## DI INJECT:

```

select * from DI_STG_file_ingest
  
```

	Di_Aggr_Ingest	File_Group	File_Name	FileRows	RowInserted	Di.JobID	Di.CreateDate
1	1	IMDb movie	C:\INFO\7370\delspring2021\cine\final\source\Filereads\imdb-datasource_imdb_movies.br	879163	879163	01091X	2021-04-27 23:16:50.00
2	2	IMDb movie	C:\INFO\7370\delspring2021\cine\final\source\Filereads\imdb-datasource_imdb_movies.br	138000	138000	1V9P9X	2021-04-27 23:16:50.447
3	3	IMDb movie	C:\INFO\7370\delspring2021\cine\final\source\Filereads\imdb-datasource_principals_movies.br	774500	774500	ZpDvE	2021-04-27 23:16:50.603
4	4	IMDb movie	C:\INFO\7370\delspring2021\cine\final\source\Filereads\imdb-datasource_rating_movies.br	373450	373450	4t5PF	2021-04-27 23:16:50.833
5	5	IMDb REVERSE courses	C:\INFO\7370\delspring2021\cine\final\source\Filereads\ac_immlanguage-codes_no.br	249	249	h0mH	2021-04-27 23:16:50.837
6	6	IMDb REVERSE courses	C:\INFO\7370\delspring2021\cine\final\source\Filereads\ac_immlanguage-codes_no.br	486	486	BPrnC	2021-04-27 23:16:50.820
7	7	MovieLens genre-score	C:\INFO\7370\delspring2021\cine\final\source\Filereads\imdb-datasource_akas_movies.br	2939863	2939863	sgtP7r	2021-04-27 23:16:50.870
8	8	MovieLens genre-score	C:\INFO\7370\delspring2021\cine\final\source\Filereads\imdb-datasource_akas_movies.br	15524446	15524446	qfJL4P	2021-04-27 23:16:50.827
9	9	MovieLens genre-score	C:\INFO\7370\delspring2021\cine\final\source\Filereads\imdb-datasource_akas_movies.br	1128	1128	SuMNC	2021-04-27 23:16:49.487
10	10	MovieLens links	C:\INFO\7370\delspring2021\cine\final\source\filelinks.csv	82423	82423	8JRaP	2021-04-27 23:16:50.280

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

**reject:** No rejects required for this table.

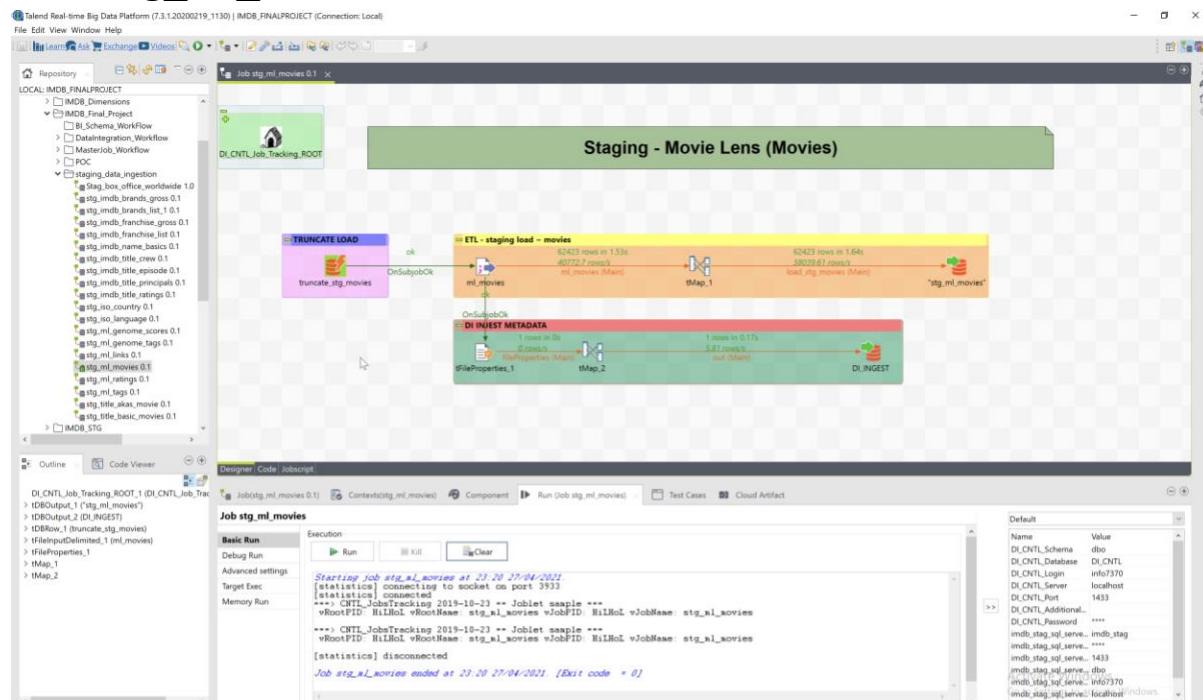
**structural integration processes:** initially we are truncating the table.

For Performance we have enabled parallelization to speed up the load

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 17. stg\_ml\_movies



Object Explorer

SQLQuery1.sql - localhost\imdb\_stag [info7370 (54)]\* - Microsoft SQL Server Management Studio (Administrator)

SQLQuery2.sql - loc..tag [info7370 (55)]\* - Microsoft SQL Server Management Studio (Administrator)

SQLQuery3.sql - loc..cntf [info7370 (53)]\* - Microsoft SQL Server Management Studio (Administrator)

SQLQuery1.sql - loc..tag [info7370 (54)]\* - Microsoft SQL Server Management Studio (Administrator)

select \* from [dbo].[stg\_ml\_movies]

select \* from [dbo].[titles]

select \* from [dbo].[titles] where title like '%(1995)'

title_id	title	year	genres	score	di_jobid	di_create_dt
1	Toy Story (1995)	1995	Adventure/Animation/Children/Comedy/Fantasy	8.2	HUH01	2021-04-27 23:20:880
2	Jumanji (1995)	1995	Adventure/Children/Fantasy	7.8	HUH01	2021-04-27 23:20:880
3	Citizen Ruth (1995)	1995	Comedy/Drama	7.7	HUH01	2021-04-27 23:20:880
4	Father of the Bride Part I (1995)	1995	Comedy	7.6	HUH01	2021-04-27 23:20:880
5	Sabrina (1995)	1995	Adventures/Thriller	7.5	HUH01	2021-04-27 23:20:880
6	Tam and Huck (1995)	1995	Adventure/Children	7.4	HUH01	2021-04-27 23:20:880
7	9 to 5 (1995)	1995	Adventures/Thriller	7.3	HUH01	2021-04-27 23:20:880
8	Godzilla (1995)	1995	Action/Adventure/Thriller	7.2	HUH01	2021-04-27 23:20:880
9	American President, The (1995)	1995	Comedy/Drama/Romance	7.1	HUH01	2021-04-27 23:20:880
10	10 (1995)	1995	Comedy/Drama/Romance	7.0	HUH01	2021-04-27 23:20:880
11	11 (1995)	1995	Comedy/Drama/Romance	6.9	HUH01	2021-04-27 23:20:880
12	Home Alone 2: Lost in New York (1995)	1995	Comedy/Family	6.8	HUH01	2021-04-27 23:20:880
13	Babe (1995)	1995	Action/Adventure/Children	6.7	HUH01	2021-04-27 23:20:880
14	Hour (1995)	1995	Drama	6.6	HUH01	2021-04-27 23:20:880
15	Die Hard with a Vengeance (1995)	1995	Action/Adventures/Romance	6.5	HUH01	2021-04-27 23:20:880
16	Cocino (1995)	1995	Crime/Drama	6.4	HUH01	2021-04-27 23:20:880
17	Sense and Sensibility (1995)	1995	Drama/Romance	6.3	HUH01	2021-04-27 23:20:880
18	Age of Innocence (1995)	1995	Comedy	6.2	HUH01	2021-04-27 23:20:880
19	Ace Ventura: When Nature Calls (1995)	1995	Comedy	6.1	HUH01	2021-04-27 23:20:880
20	Money Train (1995)	1995	Action/Crime/Gangsters/Thriller	6.0	HUH01	2021-04-27 23:20:880
21	Die Hard (1995)	1995	Comedy/Thriller	5.9	HUH01	2021-04-27 23:20:880
22	Cop Out (1995)	1995	Crime/Drama/Horror/Mystery/Thriller	5.8	HUH01	2021-04-27 23:20:880
23	Assassins (1995)	1995	Action/Crime/Thriller	5.7	HUH01	2021-04-27 23:20:880
24	White Palace (1995)	1995	Drama	5.6	HUH01	2021-04-27 23:20:880
25	Leaving Las Vegas (1995)	1995	Drama/Romance	5.5	HUH01	2021-04-27 23:20:880
26	Die Hard (1995)	1995	Drama	5.4	HUH01	2021-04-27 23:20:880
27	Die Hard With a Vengeance (1995)	1995	Action/Thriller	5.3	HUH01	2021-04-27 23:20:880
28	Persuasion (1995)	1995	Drama/Romance	5.2	HUH01	2021-04-27 23:20:880

TIME:

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

```

SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM    DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_ml_movies'
  
```

	job_name	is_root_job	root_name	job_status	Duration (minutes)	job_finish_time	message_type	project	job_version
1	stg_ml_movies	Y	stg_ml_movies	success	0.04	2021-04-27 23:20:27.600	end	IMDB_FINALPROJECT	0.1

## DI INJECT:

File Group	FileName	FileSize	LastModified	JobID	CreateDate
1	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\imdb-datasasename_basic_movies.csv	871833	2021-04-27 23:57:23.890		
2	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\imdb-datasename_movie.csv	934258	2021-04-27 22:00:05.947		
3	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\imdb-datasename_principals_movies.csv	7741502	2021-04-27 22:19:44.863		
4	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\imdb-datasename_releases_movies.csv	1046805	2021-04-27 22:20:05.532		
5	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\iso_all.csv	249	2021-04-27 22:33:41.877		
6	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\iso_languages_codes_iso.csv	481	2021-04-27 22:38:24.827		
7	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\iso_movies.csv	2966873	2021-04-27 22:38:41.570		
8	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\genome_scores.csv	1558448	2021-04-27 23:08:41.907		
9	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\genome_legs.csv	1128	2021-04-27 23:12:40.487		
10	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\links.csv	82423	2021-04-27 23:20:27.587		
11	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\movies.csv	82423	2021-04-27 23:20:27.587		
12	C:\INFO\7370\delspring2021\odbc\FinalSource\Flatfile\links.csv	82423	2021-04-27 23:22:17.287		

Query executed successfully.

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

**reject:** No rejects required for this table.

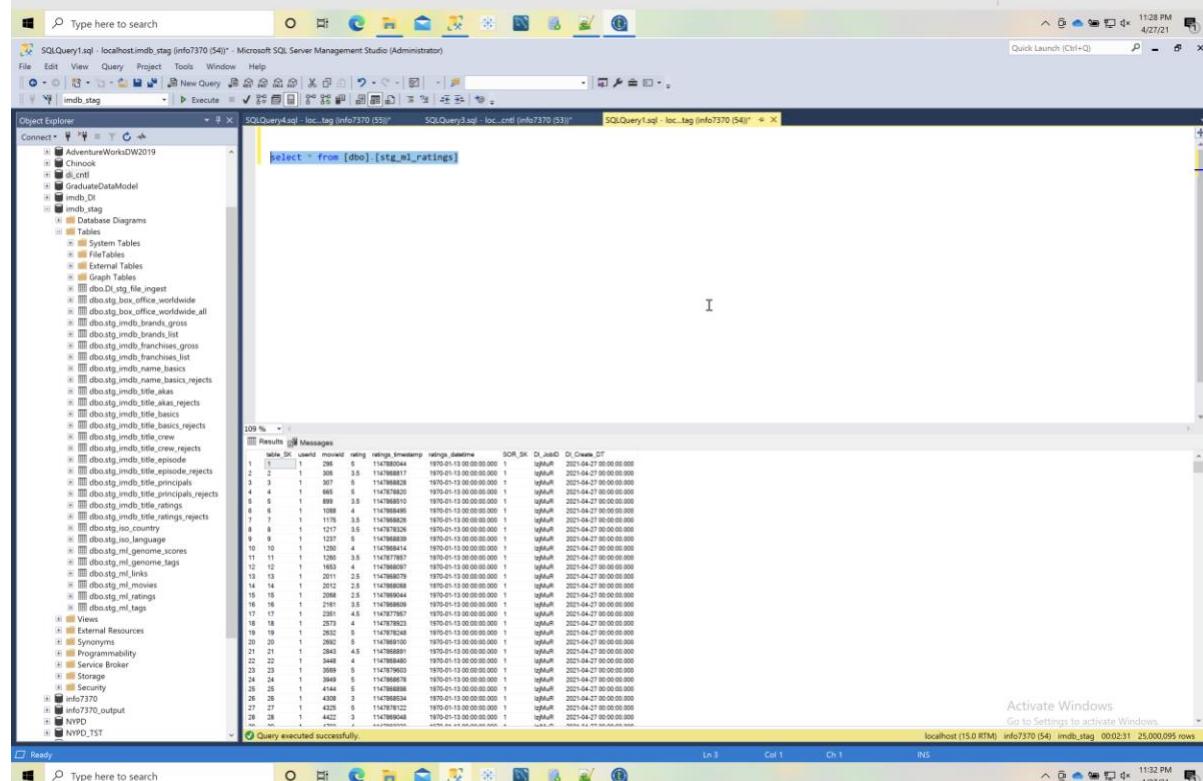
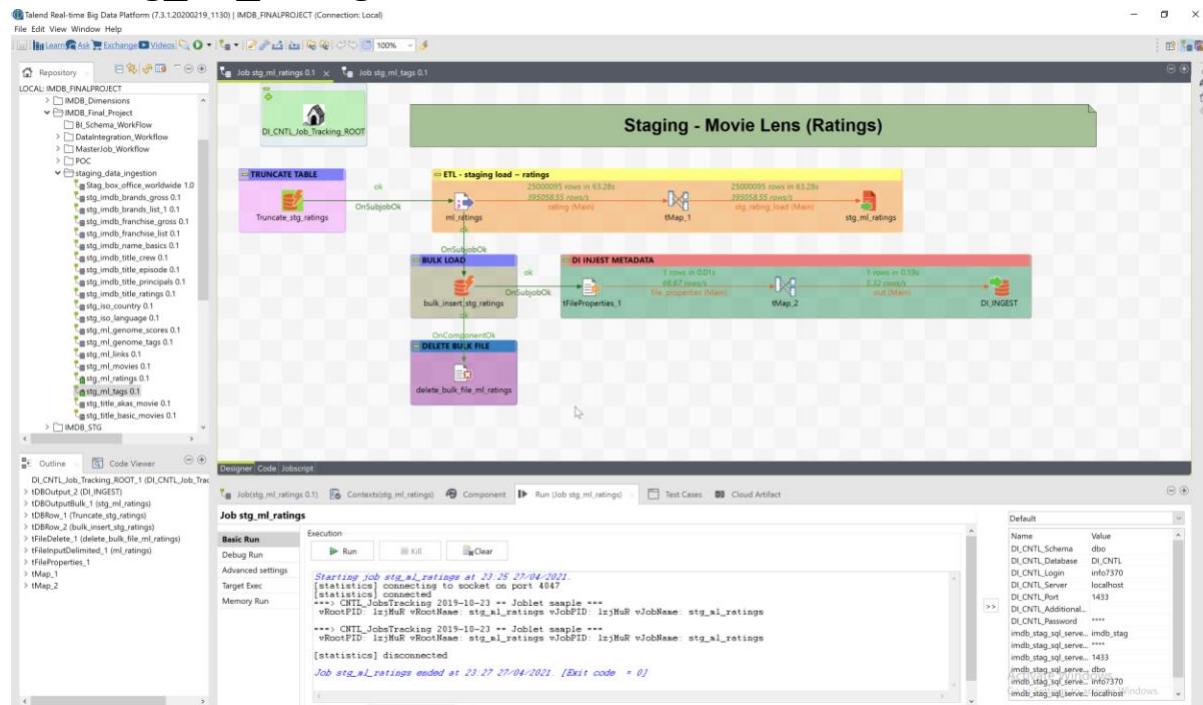
**structural integration processes:** initially we are truncating the table.

For Performance we have enabled parallelization to speed up the load

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 18. stg\_ml\_ratings



TIME:

Submitted By: Shweta Gupta

# Staging Screen Shots- Talend

```

SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_start_time,
    message_type,
    project,
    job_version
FROM DI_CNTL.dbo.job_stats_vw
WHERE job_name = 'stg_ml_ratings'
  
```

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_start_time	message_type	project	job_version
stg_ml_ratings	Y		success	2.62	2021-04-27 23:27:54.820	end	IMDB_FINALPROJECT	0.1

```

select * from DI_stg_file_ingest
  
```

File Group	Filename	FileSize	Rowcount	D_Job	D_CreatedOn
IMDB movie	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\imdb\deservertime_0012_movie.btr	130000	130000	8781803	8781803
IMDB movie	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\imdb\deservertime_0012_movie.btr	7741500	7741500	7741500	7741500
IMDB movie	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\imdb\deservertime_principals_movie.btr	379430	379430	8781803	8781803
IMDB movie	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\imdb\deservertime_rating_movie.btr	246	246	8781803	8781803
ISO REFERENCE courses	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\iso_3166language-codes_no_btr	486	486	8781803	8781803
IMDB movie	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\ac_001language-codes_no_btr	2939883	2939883	8781803	8781803
Movie Lens genome-actors	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\akas_movie.btr	15500444	15500444	8781803	8781803
Movie Lens genome-actors	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\lego.btr	1128	1128	8781803	8781803
Movie Lens links	C:\INFO\7370\dbs\spring02\1\di\file\final\source\Filereads\links.csv	62423	62423	8781803	8781803
Movie Lens links	C:\INFO\7370\dbs\spring02\1\di\file\final\source\linkedin.csv	62423	62423	8781803	8781803
Movie Lens tags	C:\INFO\7370\dbs\spring02\1\di\file\final\source\taginfo.csv	1093860	1093860	8781803	8781803
Movie Lens ratings	C:\INFO\7370\dbs\spring02\1\di\file\final\source\filereads.csv	2600098	2600098	8781803	8781803

Activate Windows  
Go to Settings to activate Windows.  
localhost (15.0 RTM) info7370 (55) imdb\_stag 00:00:00 14 rows

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

**reject:** No reject required for this table.

**structural integration processes:** initially we are truncating the table

For performance we are using bulk components,

For DI standard we are deleting bulk file which is created in the process

We are also capturing DI inject metadata and storing in the table which is bonus point.

# Staging Screen Shots- Talend

## 19. stg\_ml\_tags

TIME:

The screenshot shows the Object Explorer on the left with the database 'imdb\_stag' selected. The 'Tables' node is expanded, showing numerous tables related to the IMDB dataset. The 'Results' pane on the right displays the output of a SQL query:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_finish_time,
    message_type,
    project,
    job_version
FROM   DI_CNTL.dbo.job_stats_vw
where job_name = 'stg_ml_tags'
```

The results show one row for the 'stg\_ml\_tags' job, which is a root job ('Y') named 'stg\_ml\_tags'. It completed successfully with a duration of 0.64 minutes at 2021-04-27 23:26:46.977. The project is 'IMDB\_FINALPROJECT' and the job version is 0.1.

DI INJECT:

The screenshot shows the Object Explorer on the left with the database 'imdb\_stag' selected. The 'Tables' node is expanded, showing the 'DI\_STG\_FILE\_INGEST' table. The 'Results' pane on the right displays the output of a SQL query:

```
select * from DI_STG_FILE_INGEST
```

The results show 14 rows of file ingestion details. The table has columns: Di\_Ag\_Rule, File Group, Filename, FileIndex, RowNumbered, Di\_JobID, Di\_CreatedOn, and Di\_UpdatedOn. The data includes various file paths for IMDB movie, title, and genre files, along with their respective file indices and timestamp details.

**data consistency:** data is loaded from tsv files, to make consistent data we have used sk column.

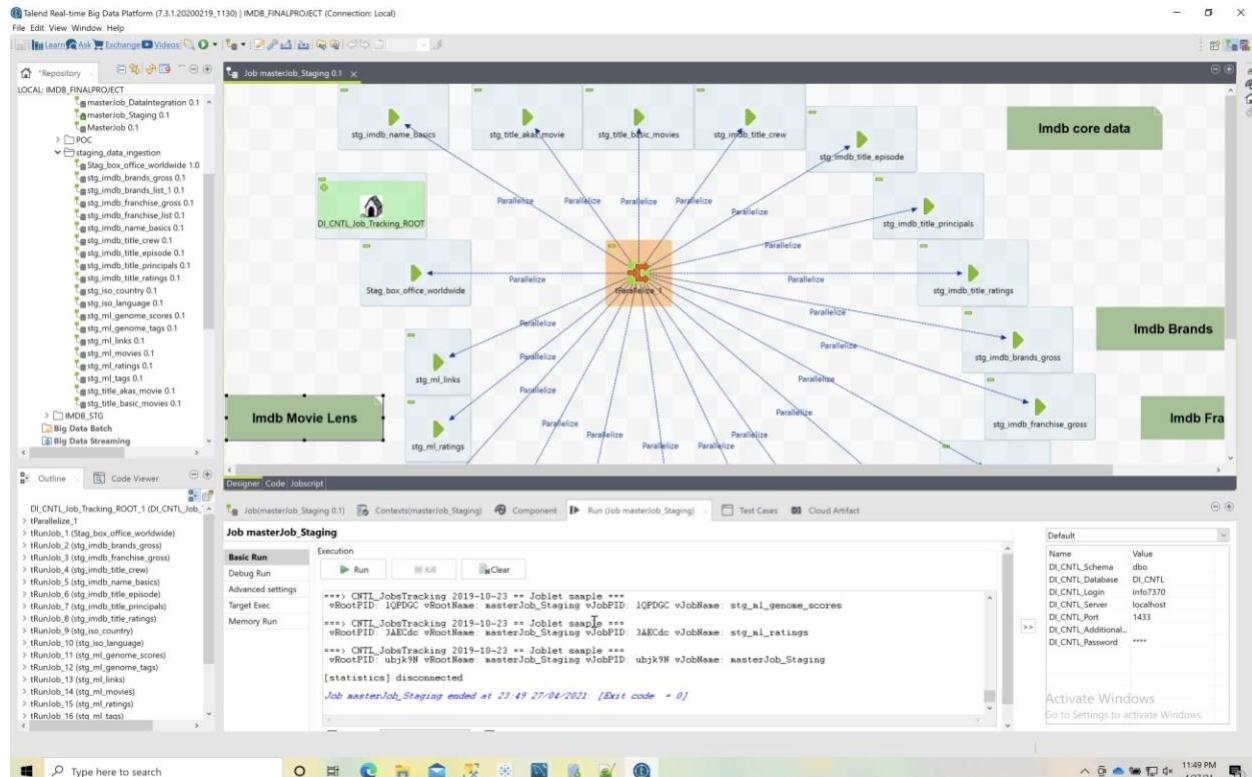
**reject:** No rejects required for this table.

**structural integration processes:** initially we are truncating the table.

# Staging Screen Shots- Talend

For Performance we have enabled parallelization to speed up the load  
We are also capturing DI inject metadata and storing in the table which is bonus point.

## STAGING MASTER:



TIME: 7.54min

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The left pane is the Object Explorer, displaying the database structure with objects like 'di\_cntl', 'AdventureWorksDW2019', and 'Chinook'. The right pane is the SQL Query Editor, showing a query to select job statistics from the 'DI\_CNTL.dbo.job\_stats\_vw' view. The query is as follows:

```
SELECT
    job_name,
    is_root_job,
    root_name,
    job_status,
    [Duration (minutes)],
    job_start_time,
    message_type,
    project,
    job_version
FROM
    DI_CNTL.dbo.job_stats_vw
```

The results pane shows the following table:

job_name	is_root_job	root_name	job_status	Duration (minutes)	job_start_time	message_type	project	job_version
masterJob_Staging	Y	masterJob_Staging	success	7.54	2021-04-27 23:49:09.570	end	IMDB_FINALPROJECT	0.1