Project:

How to Turn Data Into Insights From Ingestion

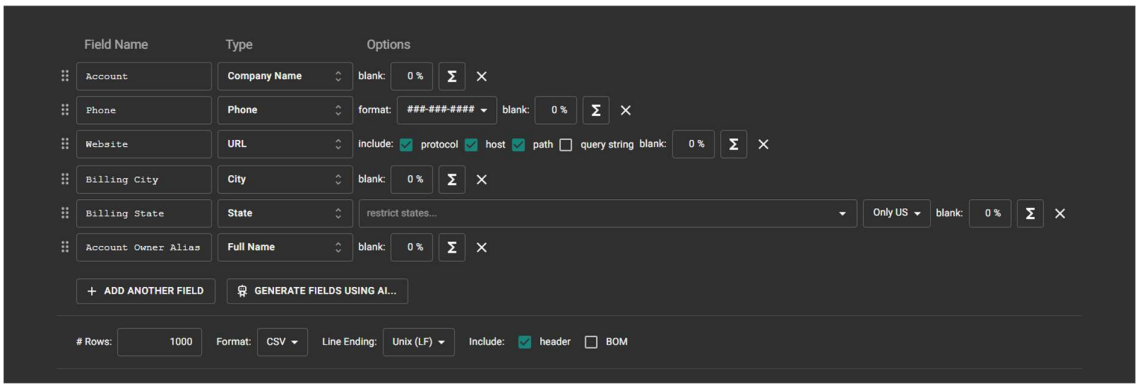
Deliverables:

Part 1: Generating Data and Configuring Your Tools

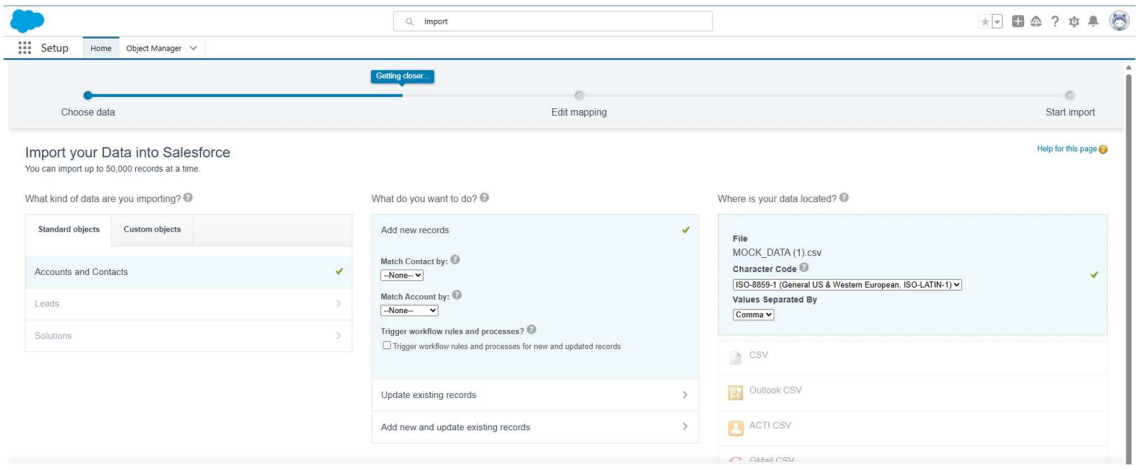
Data Generation :

Created a salesforce enterprise account using the provided link

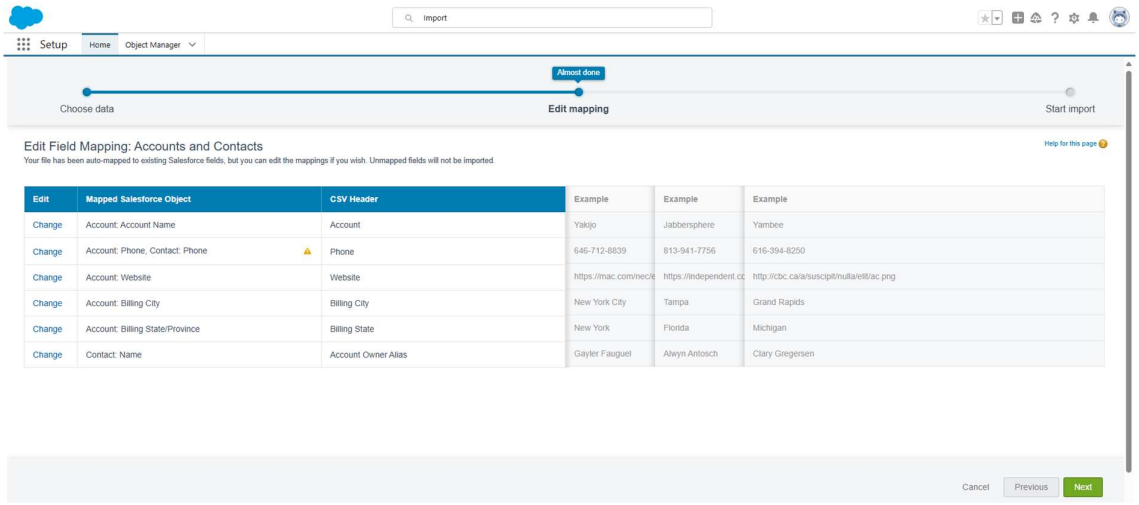
Generated a data set in Mockeroo using the below format with 1000 rows

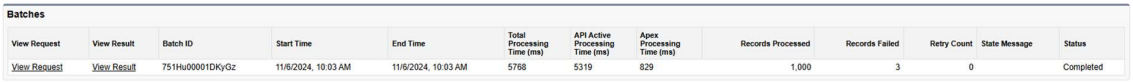


Successfully imported generated data in Salesforce.



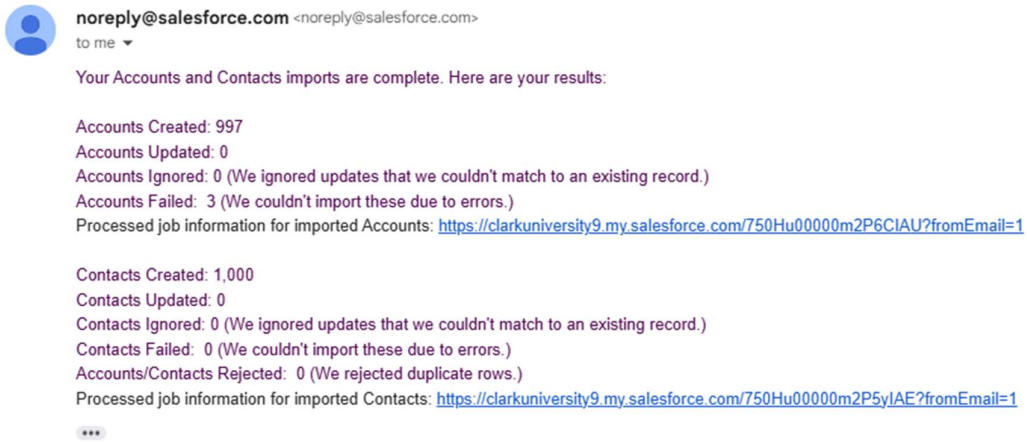
Mapped Salesforce Object to respective CSV header in order to import successfully without any errors.





Batches

View Request	View Result	Batch ID	Start Time	End Time	Total Processing Time (ms)	API Active Processing Time (ms)	Apex Processing Time (ms)	Records Processed	Records Failed	Retry Count	State Message	Status
View Request	View Result	751Hu00001DKyGz	11/6/2024, 10:03 AM	11/6/2024, 10:03 AM	5768	5319	829	1,000	3	0		Completed



noreply@salesforce.com <noreply@salesforce.com>
to me

Your Accounts and Contacts imports are complete. Here are your results:

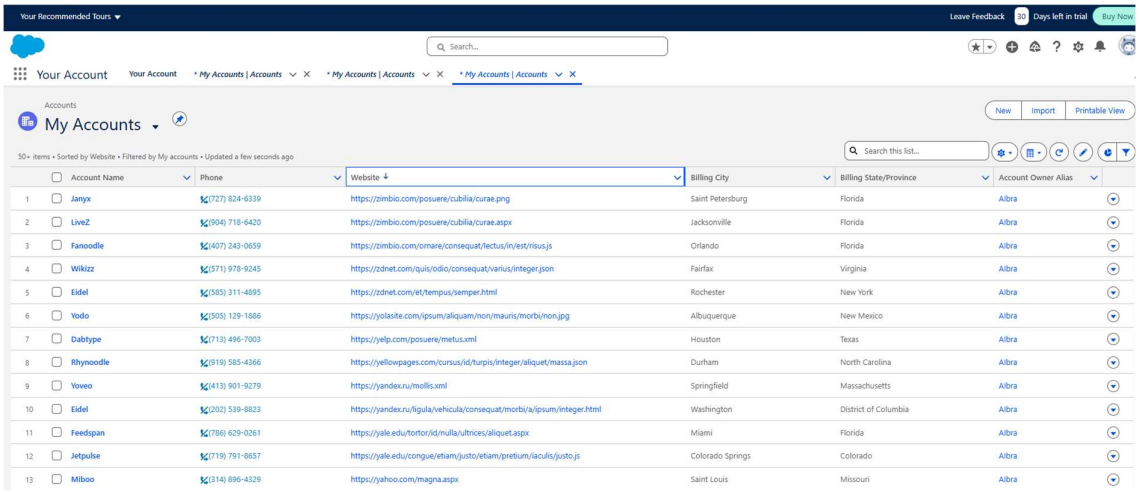
Accounts Created: 997
Accounts Updated: 0
Accounts Ignored: 0 (We ignored updates that we couldn't match to an existing record.)
Accounts Failed: 3 (We couldn't import these due to errors.)
Processed job information for imported Accounts: https://clarkuniversity9.my.salesforce.com/750Hu00000m2P6CIAU?fromEmail=1

Contacts Created: 1,000
Contacts Updated: 0
Contacts Ignored: 0 (We ignored updates that we couldn't match to an existing record.)
Contacts Failed: 0 (We couldn't import these due to errors.)
Accounts/Contacts Rejected: 0 (We rejected duplicate rows.)
Processed job information for imported Contacts: https://clarkuniversity9.my.salesforce.com/750Hu00000m2P5yIAE?fromEmail=1

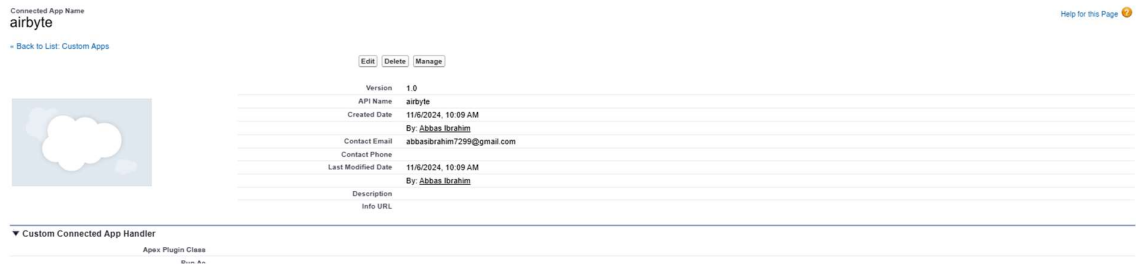
As soon as I tried importing the dataset in Salesforce based on mapping the objects, it got processed, and received a response mail from Salesforce confirming that all the contacts data got imported successfully but as for accounts 3 records didn't get imported successfully. While scrutinizing those three records, they were found to be duplicates and that's the reason they didn't get imported into Salesforce.

	A	B	C	D	E	F	G
	BillingCity	BillingState	Name	Phone	Website	Salesforce.com ID	Errors
	Montgomery	Alabama	Yabox	(334) 555-2578	https://narod.ru/elit/ac.jpg		DUPLICATES_DETECTED:You are creating a duplicate record. We recommend you use an existing record instead.:--
	Philadelphia	Pennsylvania	Babblestorm	(215) 378-7988	http://salon.com/ullamcorper.png		DUPLICATES_DETECTED:You are creating a duplicate record. We recommend you use an existing record instead.:--
	Houston	Texas	Reallinks	(713) 807-7999	http://google.com.br/vulputate/justo/in/blandit/ultrices/enim/lorem.json		DUPLICATES_DETECTED:You are creating a duplicate record. We recommend you use an existing record instead.:--

After importing the generated dataset into my accounts, we could find our imported data below in the accounts tab



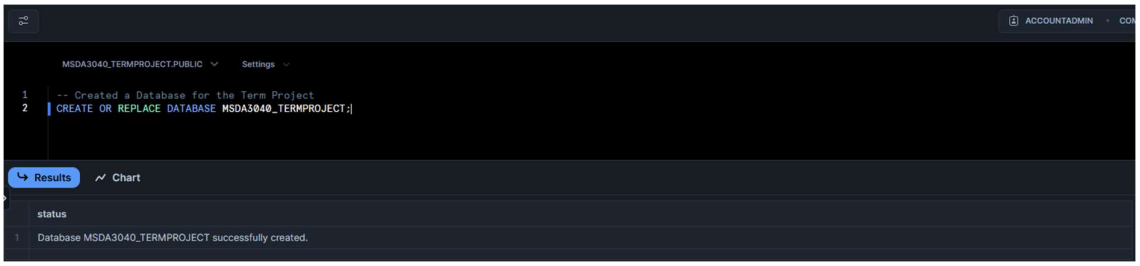
After this stage, I have set up an Airbyte connection in Salesforce



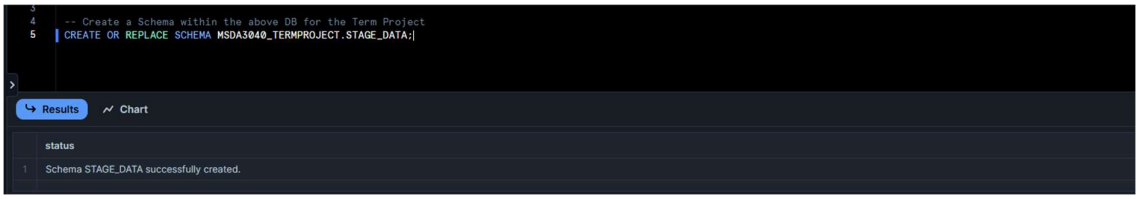
Data Preparation:

In this stage, I've created and configured database objects in Snowflake Datawarehouse so that it can be set up as our destination stage. Raw data will be stored inside the staging table and later transformed to store inside the production table for downstream users.

Below SQL script shows that I've created database in the Snowflake



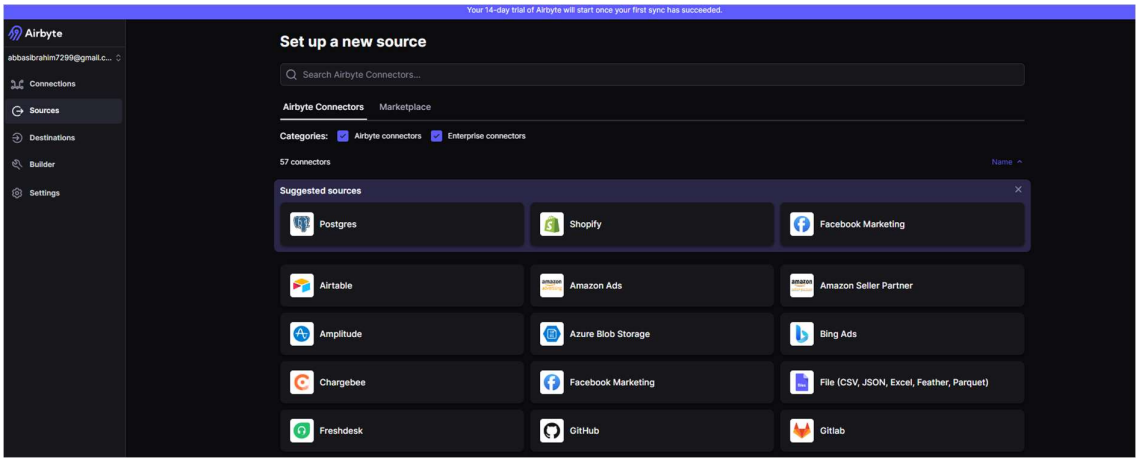
Then, created a schema named “MSDA3040_TERMPROJECT.STAGE_DATA” within the above Database in order to store the raw data. Below SQL script creates a SQL schema.

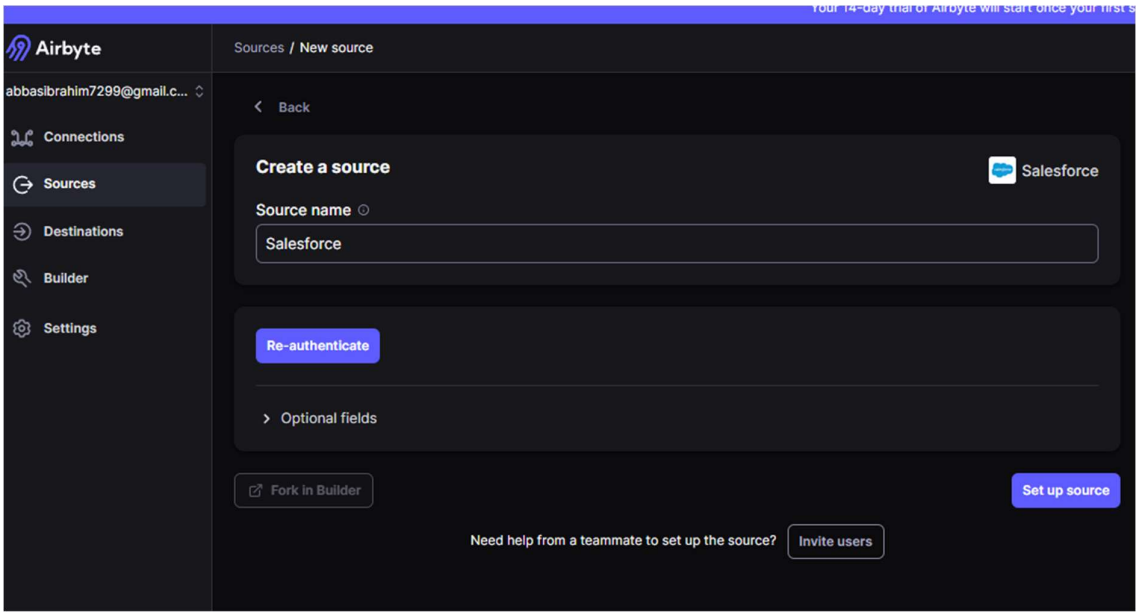


Data Loading:

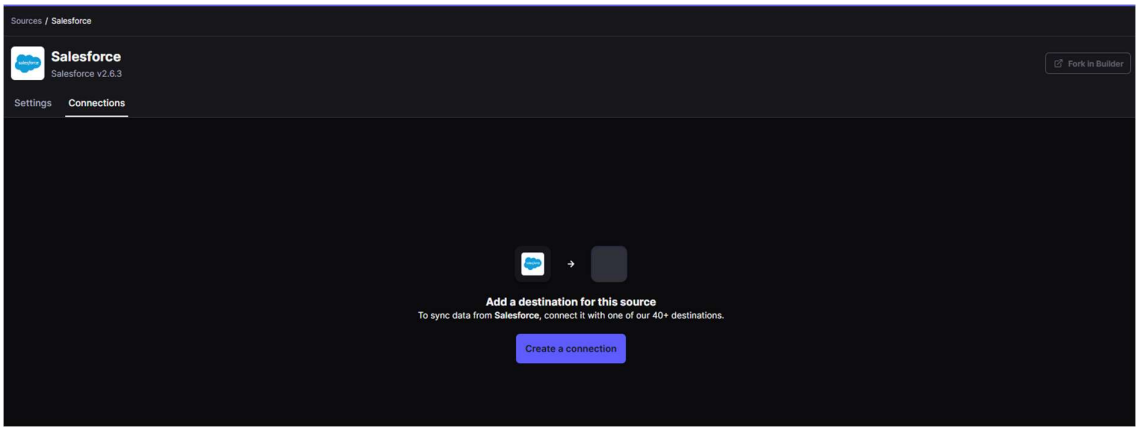
Created an Airbyte account to set up the data source and destination to extract and load the data from its source to the destination.

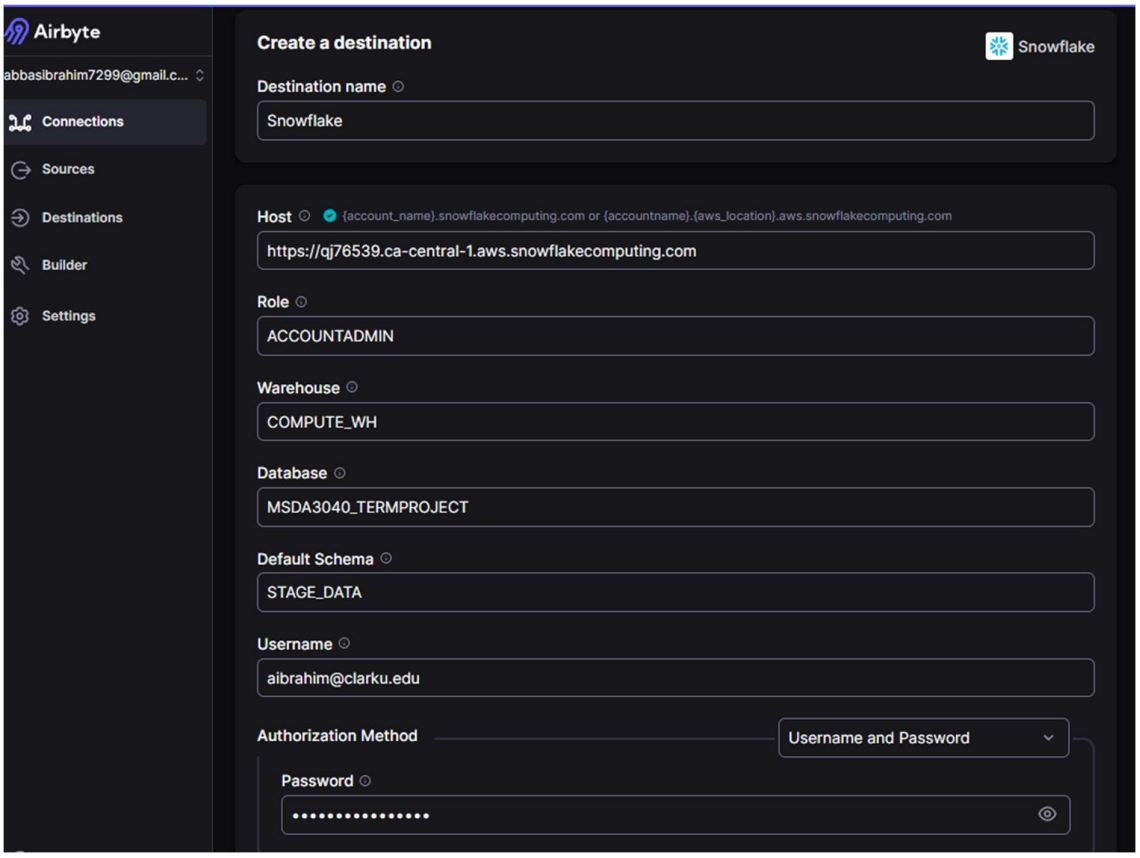
Selected “Salesforce” as ETL tool’s source and have authenticated and configured “Salesforce” has the source for our project.

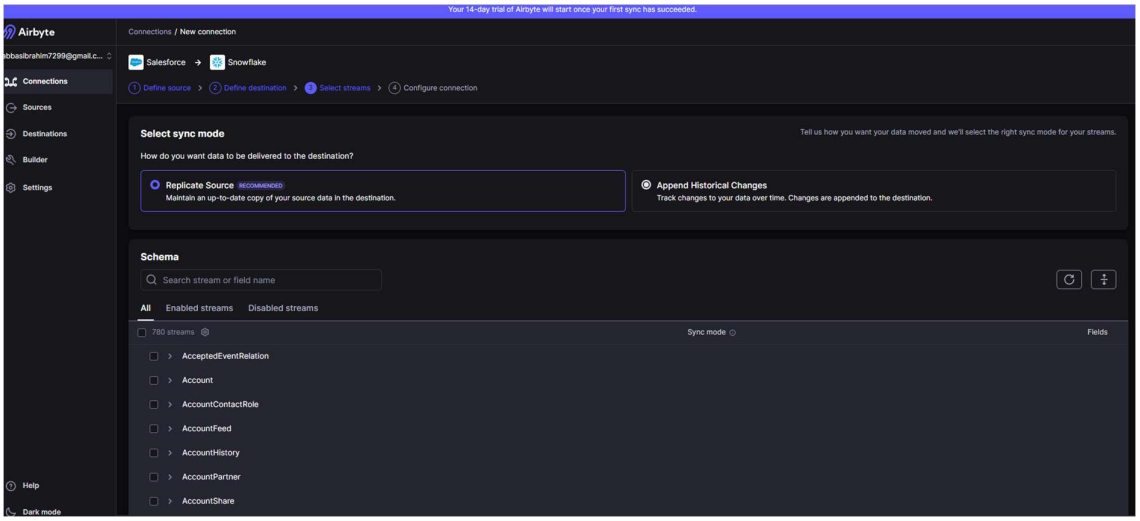


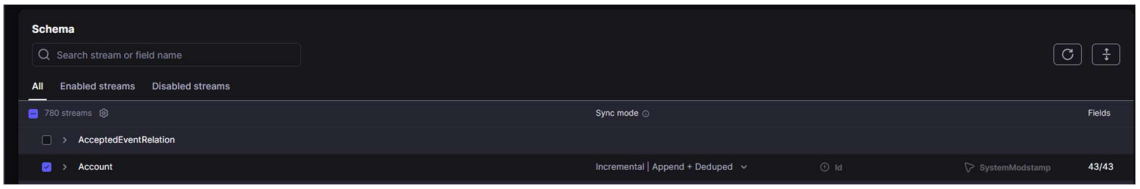


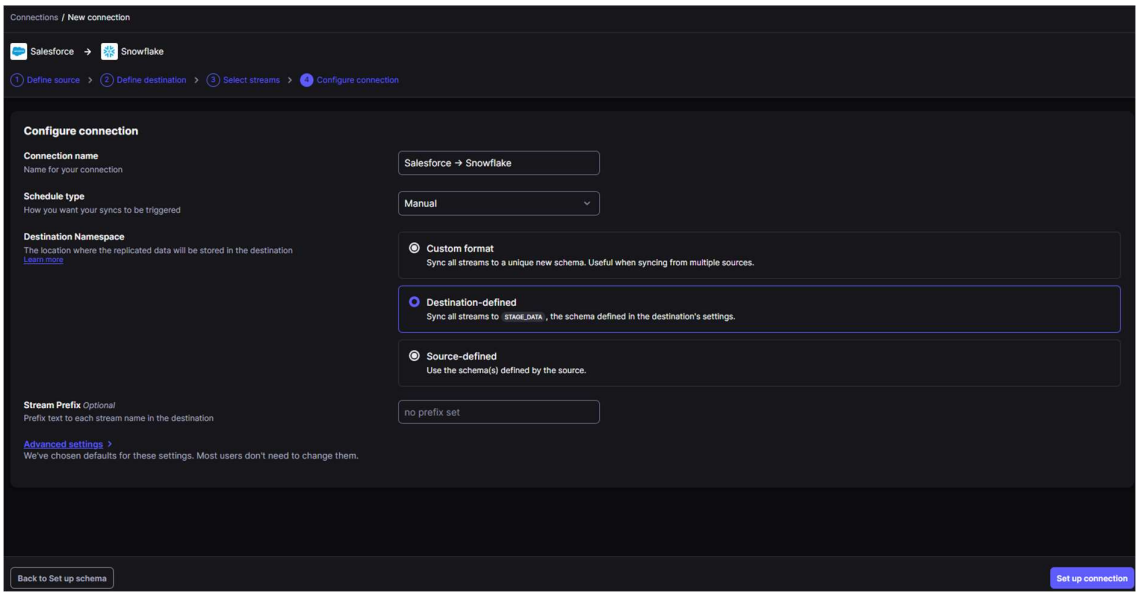
After successfully setting up the source, added the ETL's destination as Snowflake and filled in the below credentials to accurately load the data to its destination. After filling in all those details, I configured the destination and changed replication frequency to manual and manually selected the fields like ID, Name, Billing City, Billing State, Billing Phone, and Website to load these fields only into the destination data warehouse, also timestamp has been included automatically by the ETL tool itself.



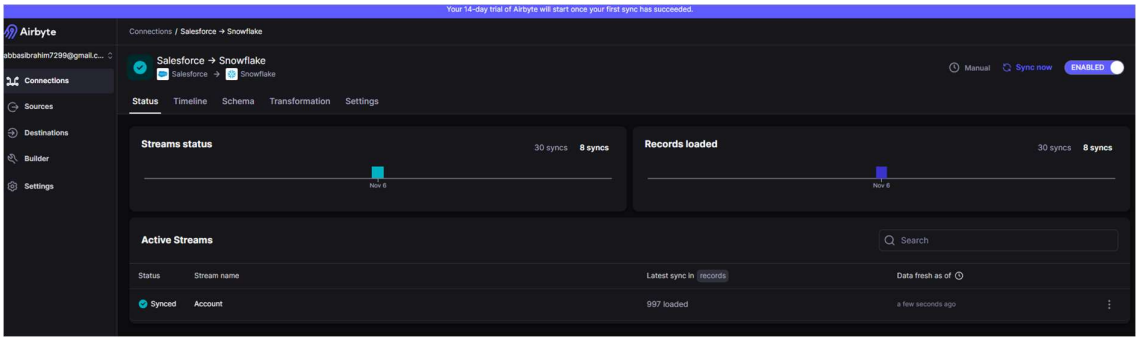



Your 14-day trial of Airbyte will start once your first sync has succeeded.
Airbyte
Connections / New connection
Salesforce → Snowflake
1 Define source › 2 Define destination › 3 Select streams › 4 Configure connection
Connections
Sources
Destinations
Builder
Settings
Help
Dark mode
Select sync mode
Tell us how you want your data moved and we'll select the right sync mode for your streams.
How do you want data to be delivered to the destination?
Replicate Source RECOMMENDED
Maintain an up-to-date copy of your source data in the destination.
Append Historical Changes
Track changes to your data over time. Changes are appended to the destination.
Schema
Search stream or field name
All
Enabled streams
Disabled streams
780 streams
Sync mode
Fields
AcceptedEventRelation
Account
AccountContactRole
AccountFeed
AccountHistory
AccountPartner
AccountShare

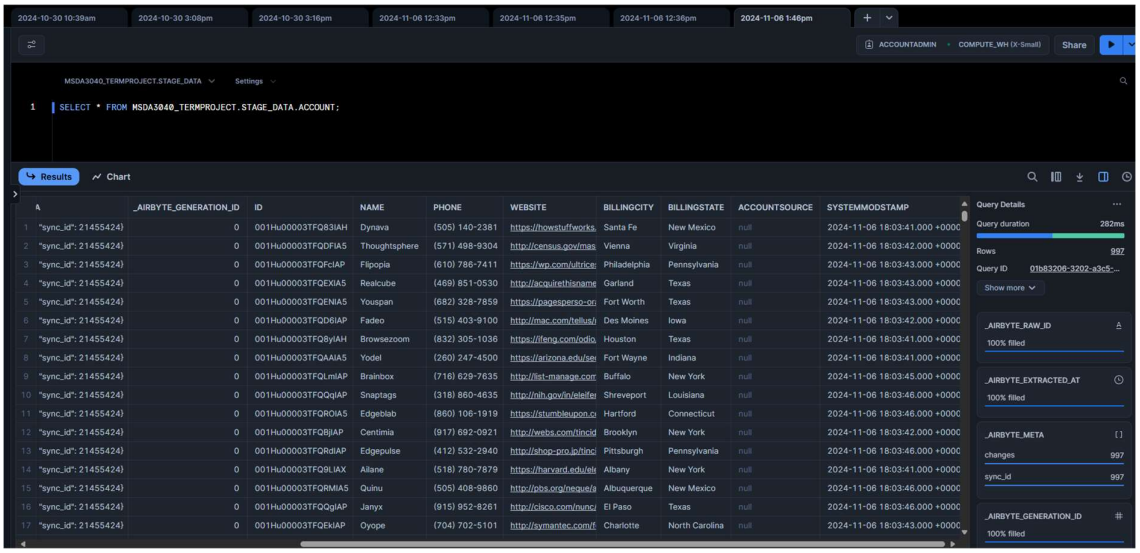

Schema
Search stream or field name
All
Enabled streams
Disabled streams
780 streams
Sync mode
Fields
AcceptedEventRelation
Account
Incremental | Append + Deduped
Id
SystemModstamp
43/43


Connections / New connection
Salesforce → Snowflake
1 Define source › 2 Define destination › 3 Select streams › 4 Configure connection
Configure connection
Connection name
Name for your connection
Salesforce → Snowflake
Schedule type
How you want your syncs to be triggered
Manual
Destination Namespace
The location where the replicated data will be stored in the destination
Learn more
Custom format
Sync all streams to a unique new schema. Useful when syncing from multiple sources.
Destination-defined
Sync all streams to STAGE_DATA, the schema defined in the destination's settings.
Source-defined
Use the schema(s) defined by the source.
Stream Prefix Optional
Prefix text to each stream name in the destination
no prefix set
Advanced settings ›
We've chosen defaults for these settings. Most users don't need to change them.
Back to Set up schema
Set up connection

After all the configuration, click setup connection and sync now to successfully extract the raw data from the salesforce and load the selective fields into the Snowflake data warehouse database object “MSDA3040_TERMPROJECT.STAGE_DATA”. Therefore, successfully leveraged airbyte to extract data from Salesforce and loaded it into the staging tables in Snowflake.



Finally, I used the SQL query below to validate whether the Salesforce records were successfully extracted and ingested into the Snowflake by the Airbyte tool for further transformation and data serving.



	A	_AIRBYTE_GENERATION_ID	ID	NAME	PHONE	WEBSITE	BILLINGCITY	BILLINGSTATE	ACCOUNTSOURCE	SYSTEMMODSTAMP
1	{"sync_id": 21455424}	0	001Hu00003TFQ83IAH	Dynava	(505) 140-2381	https://howstuffworks.	Santa Fe	New Mexico	null	2024-11-06 18:03:41.000 +0000
2	{"sync_id": 21455424}	0	001Hu00003TFQDFIA5	Thoughtsphere	(571) 498-9304	http://census.gov/mas	Vienna	Virginia	null	2024-11-06 18:03:42.000 +0000
3	{"sync_id": 21455424}	0	001Hu00003TFQFcIAP	Flipopia	(610) 786-7411	https://wp.com/ultrices	Philadelphia	Pennsylvania	null	2024-11-06 18:03:43.000 +0000
4	{"sync_id": 21455424}	0	001Hu00003TFQEXIA5	Realcube	(469) 851-0530	http://acquirethisname	Garland	Texas	null	2024-11-06 18:03:43.000 +0000
5	{"sync_id": 21455424}	0	001Hu00003TFQENIA5	Youspan	(682) 328-7859	https://pagesperso-or	Fort Worth	Texas	null	2024-11-06 18:03:43.000 +0000
6	{"sync_id": 21455424}	0	001Hu00003TFQD6IAP	Fadeo	(515) 403-9100	http://mac.com/tellus/i	Des Moines	Iowa	null	2024-11-06 18:03:42.000 +0000
7	{"sync_id": 21455424}	0	001Hu00003TFQ8yIAH	Browsezoom	(832) 305-1036	https://ifeng.com/odio	Houston	Texas	null	2024-11-06 18:03:41.000 +0000
8	{"sync_id": 21455424}	0	001Hu00003TFQAAIA5	Yodel	(260) 247-4500	https://arizona.edu/ser	Fort Wayne	Indiana	null	2024-11-06 18:03:41.000 +0000
9	{"sync_id": 21455424}	0	001Hu00003TFQLmIAP	Brainbox	(716) 629-7635	http://list-manage.com	Buffalo	New York	null	2024-11-06 18:03:45.000 +0000
10	{"sync_id": 21455424}	0	001Hu00003TFQQqIAP	Snaptags	(318) 860-4635	http://nih.gov/in/eleifer	Shreveport	Louisiana	null	2024-11-06 18:03:46.000 +0000
11	{"sync_id": 21455424}	0	001Hu00003TFQROIA5	Edgeblab	(860) 106-1919	https://stumbleupon.cı	Hartford	Connecticut	null	2024-11-06 18:03:46.000 +0000
12	{"sync_id": 21455424}	0	001Hu00003TFQBjIAP	Centimia	(917) 692-0921	http://webs.com/tincid	Brooklyn	New York	null	2024-11-06 18:03:42.000 +0000
13	{"sync_id": 21455424}	0	001Hu00003TFQRdIAP	Edgepulse	(412) 532-2940	http://shop-pro.jp/tinci	Pittsburgh	Pennsylvania	null	2024-11-06 18:03:46.000 +0000
14	{"sync_id": 21455424}	0	001Hu00003TFQ9LIAX	Ailane	(518) 780-7879	https://harvard.edu/ele	Albany	New York	null	2024-11-06 18:03:41.000 +0000
15	{"sync_id": 21455424}	0	001Hu00003TFQRMIA5	Quinu	(505) 408-9860	http://pbs.org/neque/a	Albuquerque	New Mexico	null	2024-11-06 18:03:46.000 +0000
16	{"sync_id": 21455424}	0	001Hu00003TFQQgIAP	Janyx	(915) 952-8261	http://cisco.com/nunc/	El Paso	Texas	null	2024-11-06 18:03:46.000 +0000
17	{"sync_id": 21455424}	0	001Hu00003TFQEkIAP	Oyope	(704) 702-5101	http://symantec.com/t	Charlotte	North Carolina	null	2024-11-06 18:03:43.000 +0000