

Implementation of Data Migration and Validation to Azure using Talend

¹ Niranjani V, ² Logapriya K, ³ Pavani R, ⁴ Kiruthika M and ⁵ Ranjith V

^{1,2,3,4,5}Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, India

¹niranjani.v@sece.ac.in, ²kathirvellogapriya@gmail.com, ³ rajakapavani80238@gmail.com,

⁴ kiruthi.01mk@gmail.com, ⁵ v.ranjith9095490240@gmail.com

Article Info

A. Haldorai et al. (eds.), *2nd International Conference on Materials Science and Sustainable Manufacturing Technology*, Advances in Computational Intelligence in Materials Science.

Doi: https://doi.org/10.53759/acims/978-9914-9946-9-8_13

©2023 The Authors. Published by AnaPub Publications.

This is an open access article under the CC BY-NC-ND license. (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Abstract—This project uses Talend and Azure. Before the source and target systems, data different types, and volume can be decided, the project's scope must be set the migration process must then be planned, and an Azure storage account must be created with the proper security settings. Talend is used to extract, transform, and load data from the source system into the target system utilising Azure components. To ensure that everything is correct, data validation checks are put up, and Talend is utilised to validate the data in the target system with any problems or errors being fixed as needed. Both user acceptability testing and post-migration tasks must be finished. In order to assure data's reliability and preciseness, monitoring is done last. This endeavor necessitates thorough preparation, careful oversight of details, and effective use of technologies and tools in hopes of successfully completing the transfer and requirements of a specification.

Keywords—ETL, TALEND, Data Migration, Data Validation, Database, Loading.

I. INTRODUCTION

Today's quicksilver digital world, businesses are creating and collecting enormous amounts of information. This data is ordinarily hosted across multiple websites, programmes, and systems, which can lead to data silos and limit the firm's capacity for sound judgments. To tackle this problem, data migration and validation processes are crucial for ensuring that data is sent seamlessly and consistently across numerous systems. Timestamping the authenticity, coherence, and entirety of information prior to being communicated is known as data validating. Switching documentation from a single system to another is commonly referred to as data migration. Executing these techniques involves a really well plan in addition to the implementation of trimming technology and instruments in hopes of guaranteeing information is of the highest standard. The Talend virtual components, which provide a comprehensive range of services for data integration, data quality, and big data integration, are an excellent example of one of these solutions. [1] By presenting an assortment of features for data acquisition, investigation, and warehousing, Microsoft is a robust cloud-based platform. **Fig 1** shows data migration.

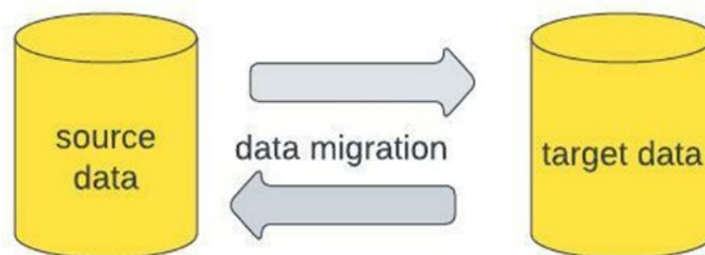


Fig 1. Data Migration.

II. LITERATURE SURVEY

ETL, or extract, transform, and load, is the acronym for this phrase. It assembles information from numerous sources and arranges it logically. This input is utilized in a database with the goal of being retained and eventually used. Reading information from a specific database that has been assembled from various sources is the process of extracting [2]. There are different techniques for data storage, including the use of metadata, Yaml, and the governance of interpersonal software

applications. The information is extracted and transformed to the desired format from its original one. The techniques used to modify data include duplicate removal, filtering, sorting, conversion, and translation. using accurate mappings on the obtained information. The corrected data is subsequently updated in the target database. Big data technologies are being embraced by various partners to keep their info, then utilize it for various kinds of research. Yet, periodically, for a wide range of reasons, the cloud migration process results in conflicts in the data, which could result in inaccurate data analysis.

ETL In Cloud

It is a means for acquiring data from various sources, interpreting it as applicable, and putting it into the target device [3]. ETL processes were executed on-site with the use of infrastructure and specialised ETL tools. But, since the arrival of cloud computing, ETL processes are feasible there. There are multiple perks to using the cloud over on-premises ETL processes. Businesses are able to swiftly and efficiently handle huge amounts of information due to the mobility of cloud-based ETL operations. They also provide greater agility and cost efficiency because no additional infrastructure or software licences are required. The simplicity with which cloud-based ETL processes can be integrated with other cloud services enables a seamless analysis and processing of information process.

Talend Data Integration

Integration is the skill of blending multiple sources of data together into singular point of view. starting with documentation, absorption, purifying, transformation to a downstream outlet, and making the data useful and accessible for the user [4]. For ETL processes, Talend offers powerful data integration capabilities. As integrating data is a challenging and time-consuming process, talent offers a 10x faster, 10x less expensive alternative to manual programming.

Future Scope of Talend

The organisations receive enormous amounts of data every day from enquiries, emails, and service requests. To achieve success, a firm makes handling the data effectively a top priority assignment. The organization's future hinges on how well they manage the data to preserve positive customer relationships. With the use of ETL solutions, which enhance data processing and boost productivity, managing data is made simpler [5]. The most sought-after Talend-related job profiles are those for Talend ETL developers, Talend developers, and Talend administrators. As a rewarding career path with the best chances in Big Data, talent has a wide range of job types available.

Existing Solution

Earlier to the arrival of cloud computing, organisations heavily depended on localized or on-premises file storage solutions. Underneath this case, each business constructed and operated its own dedicated server, and then each business saved all of its data locally. On-premises solutions have a past of providing enhanced confidentiality and safety of information albeit becoming extremely costly to establish, manage, and grow [6]. Businesses learned that they needed to regularly spend in hardware improvements and capacity expansions in order to archive and manage the growing volume of easily available data. As the online world of things (IoT), personal computing, and world wide web 2.0 took off, entrepreneurs consider that it was increasing increasingly difficult and costly to meet their own backup and recovery needs.

III. PROPOSED WORK

Planning our data migration will come next if we made the decision that it's time to transfer your data to a cloud storage option. Data transfer from nearby or on-premises servers doesn't have to be difficult, dangerous, or expensive. By avoiding the challenging and time-consuming process of hand coding your migration, data integration solutions offer a simple route to hosting your cloud data. Industry-leading data quality and data governance features are incorporated right into Talend Cloud's data integration tools, which help automate a large portion of the cloud migration process [7].

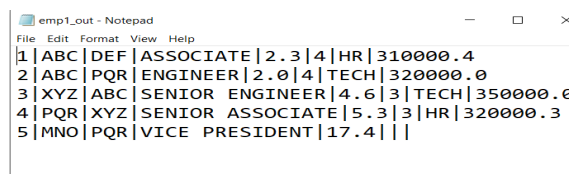
Implementation

Usecase

The on-premises data currently stored by Nexus solutions will be upgraded to cloud storage. For the migration, they wish to employ an ETL pipeline. Users who want to convert their data to cloud-based tables now store it in a file system **Fig 2**. There are some guidelines to follow [8-10].

The information files employe.txt and pay.txt are created here.

text file should use "|" as a separator.



```
emp1_out - Notepad
File Edit Format View Help
1|ABC|DEF|ASSOCIATE|2.3|4|HR|310000.4
2|ABC|PQR|ENGINEER|2.0|4|TECH|320000.0
3|XYZ|ABC|SENIOR ENGINEER|4.6|3|TECH|350000.6
4|PQR|XYZ|SENIOR ASSOCIATE|5.3|3|HR|320000.3
5|MNO|PQR|VICE PRESIDENT|17.4|||
```

Fig 2. File Data

JOB1-Several employee files (we consider 4 files like below) must be kept in a table called the employee table text. This is achieved by using tUnit component [8] in **Fig 3**.

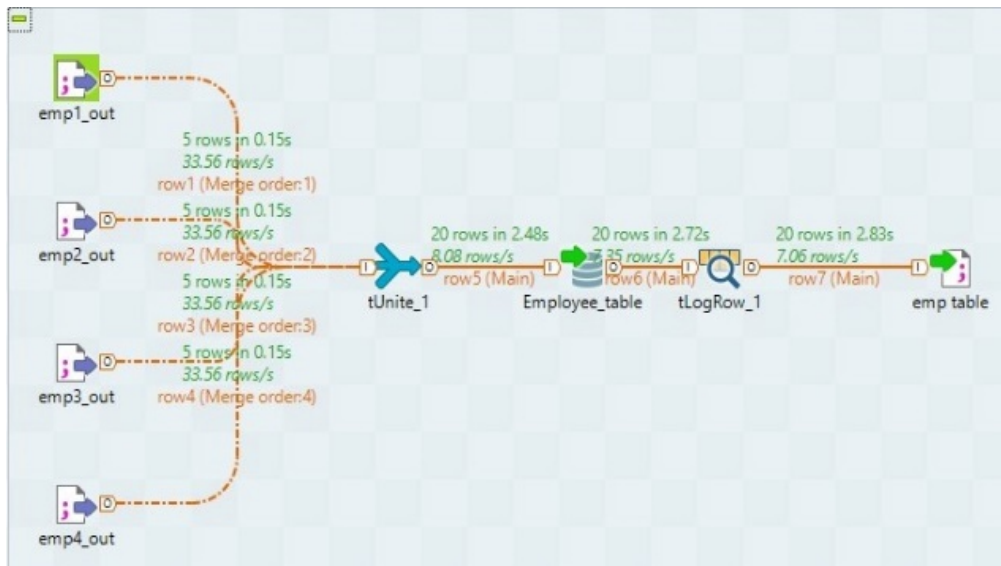


Fig 3. Job Service

JOB2-If the file is missing, the appropriate notice should be displayed. This can be done by tExist component in **Fig 4**.

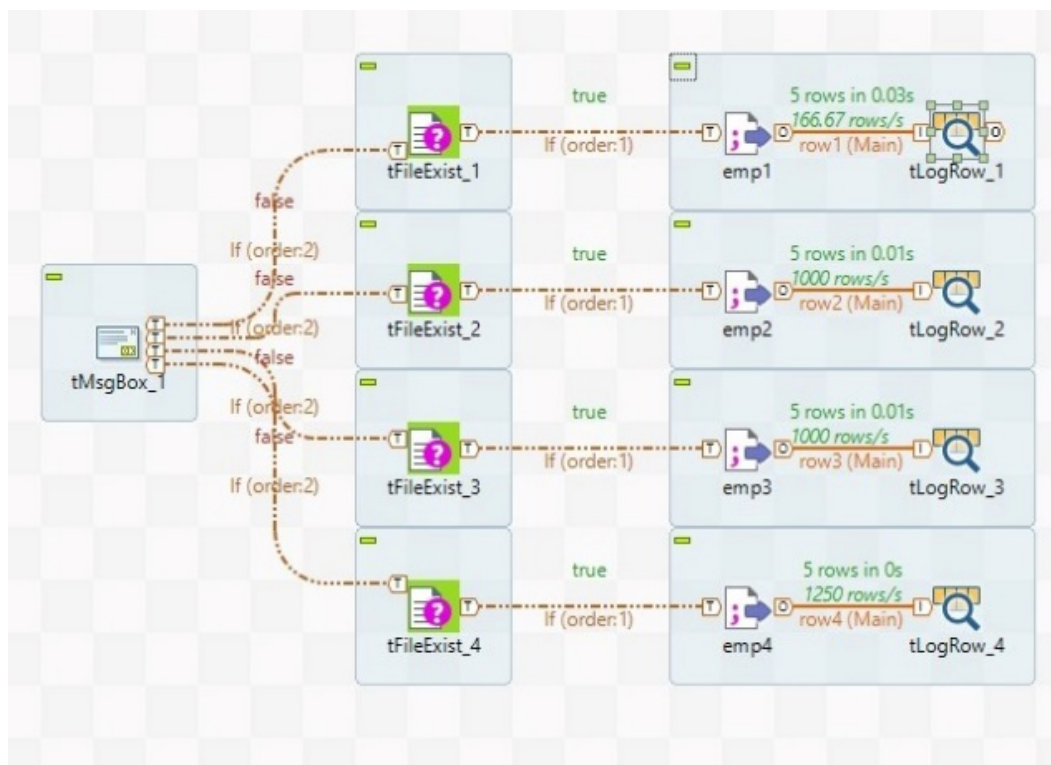


Fig 4. tExist Component

emp2,
emp4

JOB3-It need to be able to open the specific file.

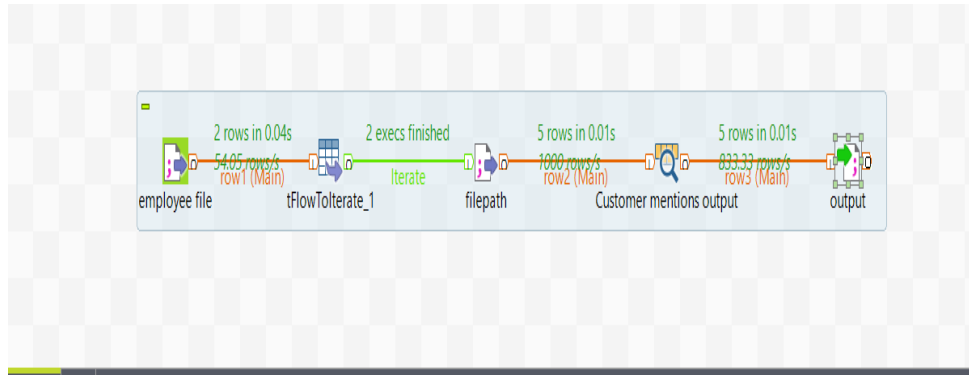


Fig 5. Job Communication

JOB4-Verify employee text files the specified structure in Fig 5 and Fig 6.
 if the employee IDs are identical.
 if any of the employee names contain characters that aren't valid.
 if the pay.txt file contains the necessary experience information about the compensation

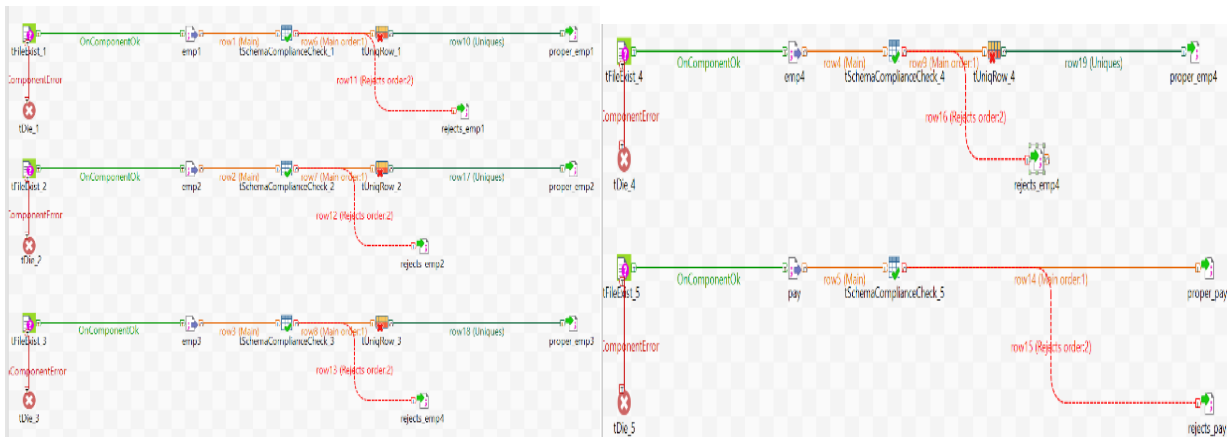


Fig 6. Validation

Connecting To an Azure Storage Account

The Azure holding login has already been created, enabling you to access the Azure Table storage service and store the supplied employee data in Fig 7.

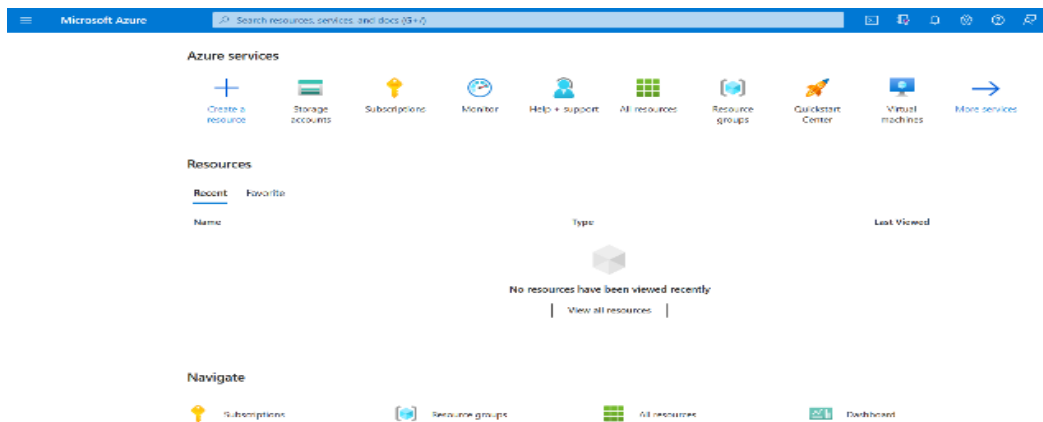


Fig 7. Azure Storage Account

Create A Storage Account

A storage account houses all of our material items from Cloud Applications, such as blobs, directories, threads, and tabular. Everything from your Cloud Applications is visible from anywhere in the globe through Http protocol thanks to the storage account, which offers a distinctive domain **Fig 8**.

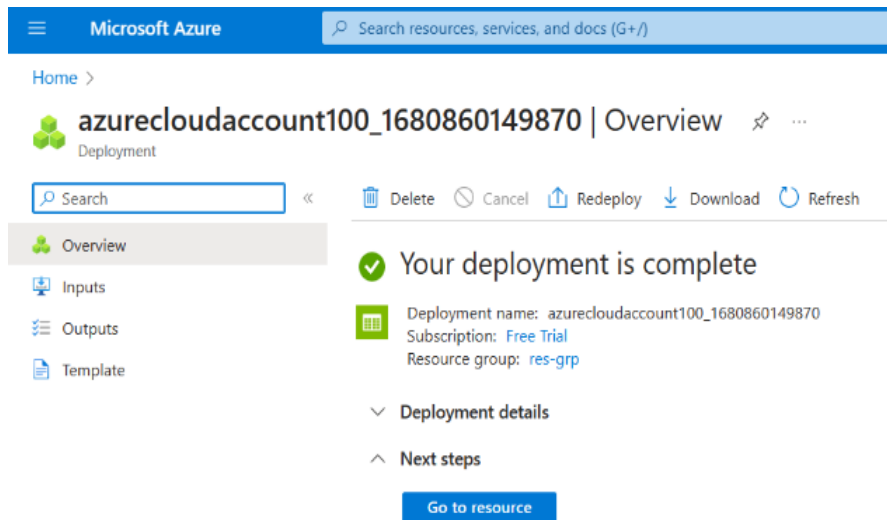


Fig 8. Storage Account Houses

Introduction Of Azure Blob Storage

Blob storage offers a cost-effective way to store big amounts of unorganised details. Because they don't correspond to a peculiar data representation else description, text or binary data are examples of unstructured data **Fig 9**.

For what Blob Storage is intended,

- directly contributing text or images to a website.
- archiving information for online access.
- streaming of music and video.
- saving information to log files.
- Data storage is necessary for incident management, backup and restoration, and store.
- Data storage for investigation performed on-site or by Cerulean services.

Blob Storage Resources

Three resource categories are provided by Blob Storage:

- Account that can be stored
- A container in the account that can be stored
- A spot in a container

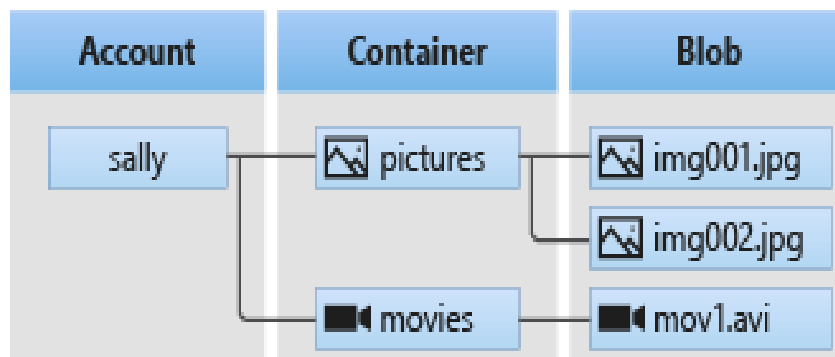


Fig 9. Blob Storage

Storage Account

Thanks to a storage account, your data is kept in Azure within a specific scope. Your personal screenname appears in the address of every item you store in Azure Storage.

Container Using Talend tool

In a way akin to a directory in a file system, a container groups a collection of blobs. Containers and blobs are stored in a single storage account **Fig 10**.

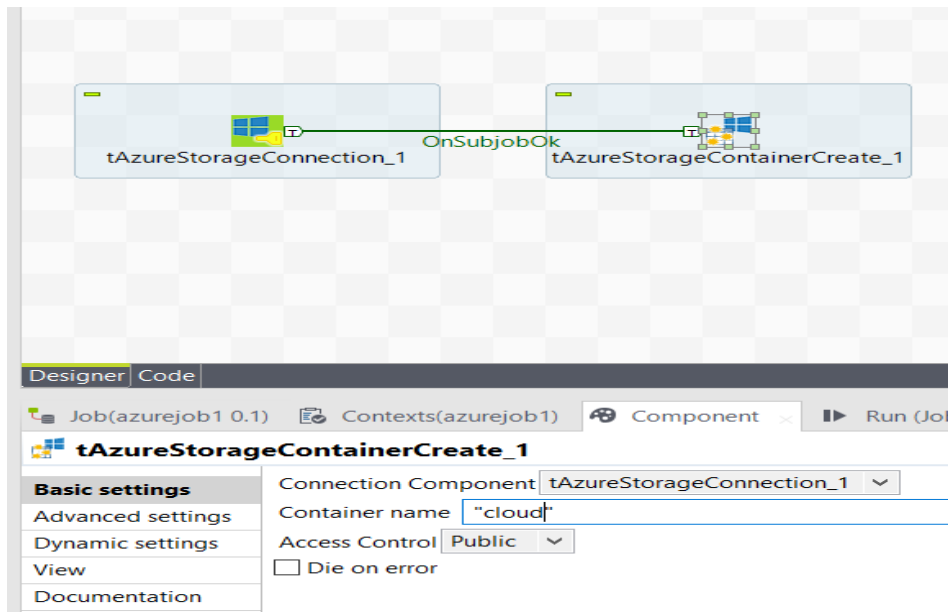


Fig 10. Talend tool

Blob

Three blob kinds are valid by Azure Storage:

- The data in 0's and 1's format and text format are occupied in spot. The memory of the blocks can store up to 190.7 TiB in blob.
- While inserting blob and block blob are made out of blocks, attach blobs are more suited for append operations. For entering data from practical tool (VM) append blobs are perfect.
- Page blobs are used to store randomly access files. It can store up to 8 TiB in size. For Azure virtual machines, page blobs serve as discs and host virtual hard drive (VHD) files.

Connecting To Azure Storage Account

Open the connection to an Azure Storage account by configuring the tAzureStorageConnection component in **Fig 10**.

- To open the Basic settings view on the Component tab, double-click tAzureStorageConnection component.

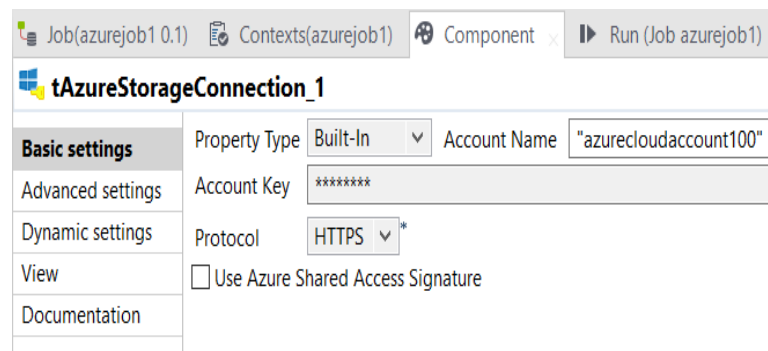


Fig 11. Connection Component

- Give name for the accessing storage account it should be typed in the Account Name field.
- The Account Key field needs to contain the key for the storage account you are accessing.

Adding Information To A Table In Azure Storage

Set up the components tFixedFlowInput and tAzureStorageOutputTable such that the data is written to an Azure Storage table.

- To open the Base settings vision on the Component tab double-click the current tFixedFlowInput component.
- To enter the schema dialogue box and specify the schema, click next to modify schema.
- After choosing Use Inline Content(delimited file) in the Mode section, enter the data that will be written into the Azure Storage database in the Content box that is supplied.
- Choose the component whose connection information will be used to establish the connection to the Azure Storage service it is available at the drop-down menu for the direct connection element.
- Enter the name of the table where the data will be stored in the "Table name" field.
- From the Action on table drop-down menu, select the action to perform on the specified table.
- To view it, select Advanced Settings.
- To add three rows and map the schema column name to the property name of each entity in the Azure database, click under the Name mappings table.

Data Recovery From The Azure Storage Table

Set up the tLogRow and tAzureStorageInputTable components to retrieve the data from the Azure Storage table.

- To open the Basic settings vision, double-click the current tFixedFlowInput component.
- In the Table name field,the name of the table can be typed from which the data will be retrieved.
- Click Edit schema to show the dialogue box for the schema.
- Select advanced options to reveal its view.
- The PartitionKey and RowKey columns for the tAzureStorageInputTable component have already been added to the schema automatically, so you do not need to specify the mapping relationship for them.
- To make it better, double-click the tLogRow component and then choose Table in the Mode section of the Basic settings view.

Executing The Job

You can run the Job and see the outcome of the Job execution after setting up the Job and configuring the components used in the Job for managing data with Azure Table storage.

- To save the Job, use Ctrl + S.
- To run the Job, press the F6 key.
-

Table 1. Components Used

Talend components	Description
Tfileinputdelimited	reads a delimited file row by row to divide it into fields, then delivers the fields to the following component according to the schema.
Tfileinputexcel	uses regular expressions to read an excel file row by row and divide each row into a field before sending each field as specified by the schema to the following component.
tAzureStoragePut	local files are uploaded into a container specified for an Azure storage account.
tAzureFSConfiguration	gives spark the necessary authentication details to connect to a specific Azure file system.
tmap	It is typically utilised for mapping one Schema to another, or input data to output data. aids in the organisation and explanation of your mental process.

IV. RESULT

JOB 1 RESULT Fig 12 to Fig 14.

```

emp table - Notepad
File Edit Format View Help
1|ABC|DEF|ASSOCIATE|2.3|4|HR|310000.4
2|ABC|PQR|ENGINEER|2.0|4|TECH|320000.0
3|XYZ|ABC|SENIOR ENGINEER|4.6|3|TECH|350000.0
4|PQR|XYZ|SENIOR ASSOCIATE|5.3|3|HR|320000.3
5|MNO|PQR|VICE PRESIDENT|17.4|||
6|REG|CEW|ENGINEER|1.3|3|TECH|350000.0
7|AEC|WEG|SENIOR ENGINEER|4.0|4|TECH|320000.0
8|WRZ|WFE|ASSOCIATE|1.6|||
9|CRG|NIU|VICE PRESIDENT|3.3|4|TECH|320000.0
10|VRE|VYD|SENIOR ASSOCIATE|7.2|||
11|KRE|BTU|ENGINEER|3.0|||
12|CEF|IJK|SENIOR ENGINEER|2.5|1|TECH|2400000.2
13|QTT|OPQ|VICE PRESIDENT|5.2|1|HR|340000.0
14|LKE|BCD|ASSOCIATE|1.8|3|HR|320000.3
15|CEP|HIJ|SENIOR ENGINEER|6.2|4|TECH|320000.0
16|CDE|RTE|SENIOR ENGINEER|2.0|1|TECH|2400000.2
17|FGH|LOP|SENIOR ASSOCIATE|1.5|3|HR|320000.3
18|JKL|QWE|ASSOCIATE|3.0|3|HR|320000.3
19|STU|RTY|ASSOCIATE|7.1|4|TECH|320000.0
20|XPO|UIO|VICE PRESIDENT|4.0|1|TECH|2400000.2
    
```

Fig 12. Result Job 1

JOB 2 RESULT:

tLogRow_1						
EMPLOYEE_ID	FIRST_NAME	LAST_NAME	DESIGNATION	TIER	DEPARTMENT	YR_OF_EXP
1	ABC	DEF	ASSOCIATE	4	HR	2.3
2	ABC	PQR	ENGINEER	4	TECH	2.0
3	XYZ	ABC	SENIOR ENGINEER	3	TECH	4.6
4	PQR	XYZ	SENIOR ASSOCIATE	3	HR	5.3
5	MNO	PQR	VICE PRESIDENT	0	TECH	17.4

tLogRow_2						
EMPLOYEE_ID	FIRST_NAME	LAST_NAME	DESIGNATION	TIER	DEPARTMENT	YR_OF_EXP
6	REG	CEW	ENGINEER	3	TECH	1.3
7	AEC	WEG	SENIOR ENGINEER	4	TECH	4.0
8	WRZ	WFE	ASSOCIATE	1	HR	1.6
9	CRG	NIU	VICE PRESIDENT	4	TECH	3.3
10	VRE	VYD	SENIOR ASSOCIATE	0	TECH	7.2

tLogRow_3						
EMPLOYEE_ID	FIRST_NAME	LAST_NAME	DESIGNATION	TIER	DEPARTMENT	YR_OF_EXP
11	KRE	BTU	ENGINEER	0	HR	3.0
12	CEF	IJK	SENIOR ENGINEER	1	TECH	2.5
13	QTT	OPQ	VICE PRESIDENT	1	HR	5.2
14	LKE	BCD	ASSOCIATE	3	HR	1.8
15	CEP	HIJ	SENIOR ENGINEER	4	TECH	6.2

tLogRow_4						
EMPLOYEE_ID	FIRST_NAME	LAST_NAME	DESIGNATION	TIER	DEPARTMENT	YR_OF_EXP
16	CDE	RTE	SENIOR ENGINEER	1	TECH	2.0
17	FGH	LOP	SENIOR ASSOCIATE	3	HR	1.5
18	JKL	QWE	ASSOCIATE	3	HR	3.0
19	STU	RTY	ASSOCIATE	4	TECH	7.1
20	XPO	UIO	VICE PRESIDENT	1	TECH	4.0

Fig 13. Result Job 2

IN REJECT FILE:

Name	Type	Compressed size	Password p...	Size	Ratio	Date modified
emp1r	Text Document	1 KB	No	1 KB	15%	07-04-2023 12:24
emp2r	Text Document	1 KB	No	1 KB	15%	07-04-2023 12:24
emp3r	Text Document	1 KB	No	1 KB	15%	07-04-2023 12:24
emp4r	Text Document	1 KB	No	1 KB	15%	07-04-2023 12:24
payr	Text Document	1 KB	No	1 KB	7%	07-04-2023 12:24

Fig 14. Reject file

V. CONCLUSION

Using Talend to migrate data to Azure can provide a multitude of benefits, including improving output, budgetary reductions, accurate data reliability, and heightened security. Organizations can make judgements regarding the project's success and impact when the migration and validation processes are finished. By transporting stuff to Azure, firms may leverage from the versatility and speed of the platform, which can strengthen the effectiveness of data processing and analysis. Organizations with a substantial amount of information to administer may notice efficiency improvements as a result of this. By utilising Talend's data integration capabilities, businesses can make sure that data is correctly structured, cleansed, and transformed before being transferred to Azure. This can aid in enhancing data quality and spotting flaws and discrepancies in the data.

References

- [1]. K. Sharma and V. Attar, "Generalized Big Data Test Framework for ETL migration," 2016 International Conference on Computing, Analytics and Security Trends (CAST), Dec. 2016, doi: 10.1109/cast.2016.7915025.
- [2]. N. Prasath and J. Sreemathy, "A New Approach for Cloud Data Migration Technique Using Talend ETL Tool," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2021, doi: 10.1109/icaccs51430.2021.9441898.
- [3]. N. Saranya, R. Brindha, N. Aishwariya, R. Kokila, P. Matheswaran, and P. Poongavi, "Data Migration using ETL Workflow," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2021, doi: 10.1109/icaccs51430.2021.9441840.
- [4]. J. Sreemathy, S. Priyadharshini, K. Radha, K. Sangeema, and G. Nivetha, "Data Validation in ETL Using TALEND," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Mar. 2019, doi: 10.1109/icaccs.2019.8728420.
- [5]. C. Shrinivasan, "Data Migration from a Product to a Data Warehouse Using ETL Tool," 2010 14th European Conference on Software Maintenance and Reengineering, Mar. 2010, doi: 10.1109/csmr.2010.25.
- [6]. H. Tahir and P. Brezillon, "A shared context approach for supporting experts in data ETL (Extraction, Transformation and Loading) processes," 2011 11th International Conference on Intelligent Systems Design and Applications, Nov. 2011, doi: 10.1109/isda.2011.6121741.
- [7]. Kamil, M. M. Inggriani, and Y. D. W. Asnar, "Data migration helper using domain information," 2014 International Conference on Data and Software Engineering (ICODSE), Nov. 2014, doi: 10.1109/icodse.2014.7062492.
- [8]. P. Pamami, A. Jain, and N. Sharma, "Cloud Migration Metamodel : A framework for legacy to cloud migration," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Jan. 2019, doi: 10.1109/confluence.2019.8776983.
- [9]. Y. S. Wijaya and A. A. Arman, "A Framework for Data Migration Between Different Datastore of NoSQL Database," 2018 International Conference on ICT for Smart Society (ICISS), Oct. 2018, doi: 10.1109/ictss.2018.8549944.
- [10]. H. Zou, M. Li, Z. Li, and J. Gao, "Design of multi-intelligent data migration strategy based on SDN secondary mode," 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), May 2018, doi: 10.1109/icaibd.2018.8396170.