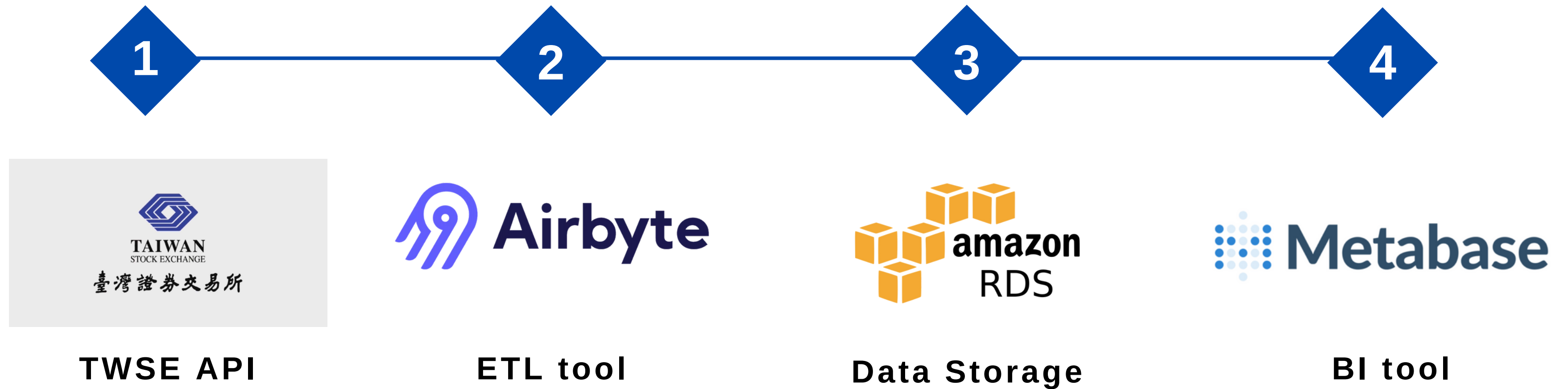


殷富期末專案

上市櫃公司在ESG社會責任和永續發展分析

資料流與架構



遇到的問題

Airbyte 在 ETL 中的 T (transform)

型態問題

```
{  
  "出表日期": "1121227",  
  "報告年度": "111",  
  "公司代號": "1101",  
  "公司名稱": "台泥",  
  "員工福利平均數(仟元/人)": "1189",  
  "員工薪資平均數(仟元/人)": "991",  
  "非擔任主管職務之全時員工薪資平均數(仟元/人)": "1030",  
  "非擔任主管之全時員工薪資中位數(仟元/人)": "922",  
  "管理職女性主管佔比": "27.00%",  
  "職業災害人數及比率-人數": "5人",  
  "職業災害人數及比率-比率": "0.43%"  
},
```

The screenshot shows the Airbyte interface for a stream named "ESG: Human Development". The "Declared schema" tab is active, displaying a JSON schema. A modal window is open, titled "Detected schema and declared schema are different", listing several issues:

- Property "公司代號" does not match schema. (#/properties)
- Instance type "string" is invalid. Expected "integer". (#/properties/%E5%85%AC%E5%8F%B8%E4%BB%A3%E8)
- Property "員工福利平均數(仟元/人)" does not match schema. (#/properties)
- Instance type "string" is invalid. Expected "number". (#/properties/%E5%93%A1%E5%B7%A5%E7%A6%8F%E5)
- Property "員工薪資平均數(仟元/人)" does not match schema. (#/properties)
- Instance type "string" is invalid. Expected "number". (#/properties/%E5%93%A1%E5%B7%A5%E8%96%AA%E8)
- Property "非擔任主管之全時員工薪資中位數(仟元/人)" does not match schema. (#/properties)
- Instance type "string" is invalid. Expected "number". (#/properties/%E9%9D%9E%E6%93%94%E4%BB%BB%E4)

The schema editor shows the following JSON:

```
{  
  "$schema": "http://json-sch  
  "properties": {  
    "公司代號": {  
      "type": "integer"  
    },  
    "公司名稱": {  
      "type": "string"  
    },  
    "出表日期": {  
      "type": "string"  
    },  
    "員工福利平均數(仟元/人)": {  
      "type": "number"  
    },  
    "員工薪資平均數(仟元/人)": {  
      "type": "number"  
    },  
    "報告年度": {  
      "type": "string"  
    },  
  }  
}
```

遇到的問題

Airbyte 在 ETL 中的 T (transform)

在METABASE中讀取型態為STRING，後續分析造成問題



在Esg Human Development中的欄位

欄位名稱	欄位類型	資料類型
公司代號	沒有欄位類型	type/Text
管理職女性主管佔比	沒有欄位類型	type/Text
職業災害人數及比率 比率	沒有欄位類型	type/Text
職業災害人數及比率 人數	沒有欄位類型	type/Text
報告年度	分類	type/Text
員工薪資平均數(千元/人)	沒有欄位類型	type/Text


遇到的問題

Airbyte 在 ETL 中的 T (transform)

TWSE → AWS RDS Postgres Enabled

 TWSE CUSTOM →  AWS RDS Postgres COMMUNITY

[Status](#) [Job History](#) [Replication](#) **[Transformation](#)** [Settings](#)


Normalization 

☐ Raw data (JSON)

☒ Normalized tabular data Map the JSON object to the types and format native to the destination. [Learn more](#)

Cancel

Save changes

Custom Transformations 

No custom transformation

+ Add transformation

Cancel

Save changes

遇到的問題

Airbyte 在 ETL 中的 T (transform)

僅提供使用DBT做額外轉換功能

Add

Transformation name *

My dbt transformations

Entrypoint arguments for dbt cli to run the project. *

run

Docker image URL with dbt installed *

fishtownanalytics/dbt:1.0.0

Git branch name (leave blank for default branch)

Git repository URL of the custom transformation project *

https://github.com/<organisation>/<git_repo>.git

Cancel

Save transformation

問題處理

從Docker裡面調，嘗試修改Airbyte自動產生的.sql檔案

```
13     jsonb_extract_path_text(_airbyte_data, '職業災害人數及比率-比率') as "職業災害人數及比率-比率",
14     jsonb_extract_path_text(_airbyte_data, '職業災害人數及比率-人數') as "職業災害人數及比率-人數",
15     jsonb_extract_path_text(_airbyte_data, '報告年度') as "報告年度",
16     jsonb_extract_path_text(_airbyte_data, '員工薪資平均數(仟元/人)') as "員工薪資平均數(仟元/人) ",
17     jsonb_extract_path_text(_airbyte_data, '非擔任主管職務之全時員工薪資平均數(仟元/人)') as "非擔任主管職務之全時員工薪資平均數(仟元/人) ",
18     jsonb_extract_path_text(_airbyte_data, '員工福利平均數(仟元/人)') as "員工福利平均數(仟元/人)",
19     jsonb_extract_path_text(_airbyte_data, '出表日期') as "出表日期",
20     jsonb_extract_path_text(_airbyte_data, '公司名稱') as "公司名稱",
21     jsonb_extract_path_text(_airbyte_data, '非擔任主管之全時員工薪資中位數(仟元/人)') as "非擔任主管之全時員工薪資中位數(仟元/人) ",
22     _airbyte_ab_id,
23     _airbyte_emitted_at,
24     now() as _airbyte_normalized_at
25 from "metabase_data_source".public._airbyte_raw_esg_human_development as table_alias
26 -- esg_human_development
27 where 1 = 1
28 ), __dbt__cte__esg_human_development_ab2 as (
29
30 -- SQL model to cast each column to its adequate SQL type converted from the JSON schema type
31 -- depends_on: __dbt__cte__esg_human_development_ab1
32 select
33     cast("公司代號" as INTEGER) as "公司代號",
34     cast(NULLIF(trim(replace("管理職女性主管佔比", '%', '')), '') as NUMERIC) / 100 as "管理職女性主管佔比",
35     cast(NULLIF(trim(replace("職業災害人數及比率-比率", '%', '')), '') as NUMERIC) / 100 as "職業災害人數及比率-比率",
36     cast(NULLIF(trim(replace("職業災害人數及比率-人數", '人', '')), '') as INTEGER) as "職業災害人數及比率-人數",
37     cast("報告年度" as INTEGER) as "報告年度",
38     cast(NULLIF(trim("員工薪資平均數(仟元/人)"), '') as NUMERIC) as "員工薪資平均數(仟元/人) ",
39     cast(NULLIF(trim("非擔任主管職務之全時員工薪資平均數(仟元/人)"), '') as NUMERIC) as "非擔任主管職務之全時員工薪資平均數(仟元/人) ",
40     cast(NULLIF(trim("員工福利平均數(仟元/人)"), '') as NUMERIC) as "員工福利平均數(仟元/人)",
41     to_date(NULLIF(trim("出表日期"), ''), 'YYYYMMDD') as "出表日期",
42     "公司名稱",
43     cast(NULLIF(trim("非擔任主管之全時員工薪資中位數(仟元/人)"), '') as NUMERIC) as "非擔任主管之全時員工薪資中位數(仟元/人) ",
44     _airbyte_ab_id,
45     _airbyte_emitted_at,
46     now() as _airbyte_normalized_at
47 from __dbt__cte__esg_human_development_ab1
48 -- esg_human_development
49 where 1 = 1
50 ), __dbt__cte__esg_human_development_ab3 as (
51
52 -- SQL model to build a hash column based on the values of this record
```


問題處理

結論:修改後運報錯

```
temporal.internal.activity.RootActivityInboundCallsInterceptor.execute(RootActivityInboundCallsInterceptor.java:43) ~[temporal-sdk-1.17.0.jar:?]
temporal.internal.activity.ActivityTaskExecutors$BaseActivityTaskExecutor.execute(ActivityTaskExecutors.java:95) ~[temporal-sdk-1.17.0.jar:?]
temporal.internal.activity.ActivityTaskHandlerImpl.handle(ActivityTaskHandlerImpl.java:92) ~[temporal-sdk-1.17.0.jar:?]
temporal.internal.worker.ActivityWorker$TaskHandlerImpl.handleActivity(ActivityWorker.java:241) ~[temporal-sdk-1.17.0.jar:?]
temporal.internal.worker.ActivityWorker$TaskHandlerImpl.handle(ActivityWorker.java:206) ~[temporal-sdk-1.17.0.jar:?]
temporal.internal.worker.ActivityWorker$TaskHandlerImpl.handle(ActivityWorker.java:179) ~[temporal-sdk-1.17.0.jar:?]
temporal.internal.worker.PollTaskExecutor.lambda$process$0(PollTaskExecutor.java:93) ~[temporal-sdk-1.17.0.jar:?]
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1144) ~[?:?]
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:642) ~[?:?]
java.lang.Thread.run(Thread.java:1589) ~[?:?]
17:39 INFO i.a.w.g.DefaultNormalizationWorker(run):102 - Normalization executed in 20 seconds for job 26.
17:39 INFO i.a.w.g.DefaultNormalizationWorker(run):114 - Normalization summary: io.airbyte.config.NormalizationSummary@5f7bf1fe[startTime=1703614638580]
creating table model public.esg__human_development..... [ERROR in 2.37s]
in model esg__human_development (models/generated/airbyte_tables/public/esg__human_development.sql)
syntax for type double precision: ""
at ../build/run/airbyte_utils/models/generated/airbyte_tables/public/esg__human_development.sql
creating table model public.esg__human_development..... [ERROR in 2.37s]
in model esg__human_development (models/generated/airbyte_tables/public/esg__human_development.sql)
syntax for type double precision: ""
at ../build/run/airbyte_utils/models/generated/airbyte_tables/public/esg__human_development.sql, retryable=<null>, timestamp=1703614659434, additionalProperties={}
creating table model public.esg__human_development..... [ERROR in 2.37s]
in model esg__human_development (models/generated/airbyte_tables/public/esg__human_development.sql)
syntax for type double precision: ""
at ../build/run/airbyte_utils/models/generated/airbyte_tables/public/esg__human_development.sql, retryable=<null>, timestamp=1703614659434, additionalProperties={}
creating table model public.esg__human_development..... [ERROR in 2.37s]
in model esg__human_development (models/generated/airbyte_tables/public/esg__human_development.sql)
syntax for type double precision: ""
at ../build/run/airbyte_utils/models/generated/airbyte_tables/public/esg__human_development.sql, retryable=<null>, timestamp=1703614659434, additionalProperties={}
17:39 INFO i.a.c.i.LineGobbler(voidCall):149 - 
17:39 INFO i.a.c.i.LineGobbler(voidCall):149 - ----- END DEFAULT NORMALIZATION -----
17:39 INFO i.a.c.i.LineGobbler(voidCall):149 - 
17:39 INFO i.a.w.t.s.a.AppendToAttemptLogActivityImpl(log):56 - Retry State: RetryManager(completeFailureBackoffPolicy=BackoffPolicy(minInterval=PT10S, maxInterval=PT1M, maxAttempts=10), next attempt: 0 seconds
```

Sync Succeeded

遇到的問題

因型態問題，導致拉表時無法讀取數據

STRING /NUMBER



資... > METABASE_DATA_... > ESG_HUMAN_DEVE...

在Esg Human Development中的欄位

欄位名稱	欄位類型	資料類型
公司代號 公司代號	T 沒有欄位類型	type/Text
管理職女性主管佔比 管理職女性主管佔比	T 沒有欄位類型	type/Text
職業災害人數及比率 比率	T 沒有欄位類型	type/Text
職業災害人數及比率 人數	T 沒有欄位類型	type/Text
報告年度 報告年度	T 分類	type/Text
員工薪資平均數(仟元/人) 員工薪資平均數(仟元/人)	T 沒有欄位類型	type/Text
非擔任主管職務之全時員工薪資平均數(仟元/人) 非擔任主管職務之全時員工薪資平均數(仟元/人)	T 沒有欄位類型	type/Text

遇到的問題

Data Warehouse選擇

在使用AIRBYTE時跳出此建議


Postgres

This page guides you through the process of setting up the Postgres destination connector.

- ⚠ Postgres, while an excellent relational database, is not a data warehouse. Postgres is likely to perform poorly with large data volumes. Even postgres-compatible destinations (e.g. AWS Aurora) are not immune to slowdowns when dealing with large writes or updates over ~500GB. Especially when using normalization with `destination-postgres`, be sure to monitor your database's memory and CPU usage during your syncs. It is possible for your destination to 'lock up', and incur high usage costs with large sync volumes.

問題處理 Data Warehouse選擇

嘗試轉為使用AWS S3儲存體

 **AWS_S3**
S3 v0.5.1 CERTIFIED

Settings Connections

Destination Settings

Destination name ⓘ
AWS_S3

S3 Bucket Name ⓘ
stock-analysis-pipeline-s3

S3 Bucket Path ⓘ
airbyte/destination

一般用途儲存貯體 (1) 資訊

儲存貯體是存放在 S3 中資料的容器。 [進一步了解](#)

🔍 依名稱尋找儲存貯體

	名稱	AWS 區域
<input type="radio"/>	stock-analysis-pipeline-s3	美國東部 (維吉尼亞北部) us-east-1

問題處理 Data Warehouse選擇

物件 (26) 資訊

物件是存放在 Amazon S3 中的基本實體。您可以使用 [Amazon S3 庫存](#) 取得儲存貯體中所有物件的清單。若要讓其

  複製 S3 URI  複製 URL  下載  開啟  刪除

🔍 依前綴尋找物件

<input type="checkbox"/>	名稱	類型
<input type="checkbox"/>	📁 _____/	資料夾
<input type="checkbox"/>	📁 _____2____/	資料夾
<input type="checkbox"/>	📄 0669b017-194a-401f-95d6-da50c39b1e4a.txt	txt
<input type="checkbox"/>	📄 0669b017-194a-401f-95d6-da50c39b1e4a.txt.metadata	metadata
<input type="checkbox"/>	📄 06892302-0602-4625-8c54-4bc4ffa8636d.txt	txt
<input type="checkbox"/>	📄 06892302-0602-4625-8c54-4bc4ffa8636d.txt.metadata	metadata
<input type="checkbox"/>	📄 3af6b766-dae3-4bdc-88ea-ab3a6e0d3acf.txt	txt
<input type="checkbox"/>	📄 3af6b766-dae3-4bdc-88ea-ab3a6e0d3acf.txt.metadata	metadata
<input type="checkbox"/>	📄 538eb34d-5b03-41f4-aa17-12815109e521.txt	txt
<input type="checkbox"/>	📄 538eb34d-5b03-41f4-aa17-12815109e521.txt.metadata	metadata

運作方式



問題處理 Data Warehouse選擇

Metabase支援存取Athena服務

DATABASES > ADD DATABASE

Database type

Amazon Athena

Display name

Our Amazon Athena

問題處理 Data Warehouse選擇

最終沒順利換回Postgre

資料

<

資料來源

AwsDataCatalog ▼

資料庫

default ▼

資料表和檢視

建立 ▼ ⚙️

🔍 篩選資料表和檢視

▶ 資料表 (0) < 1 >

▶ 檢視 (0) < 1 >

❌ 查詢 2 ⋮

1 SELECT * FROM "default"."airbyte_s3" limit 10

SQL 行 1，欄 1

再次執行

解釋 🔗

取消

清除

建立 ▼

查詢結果

查詢狀態

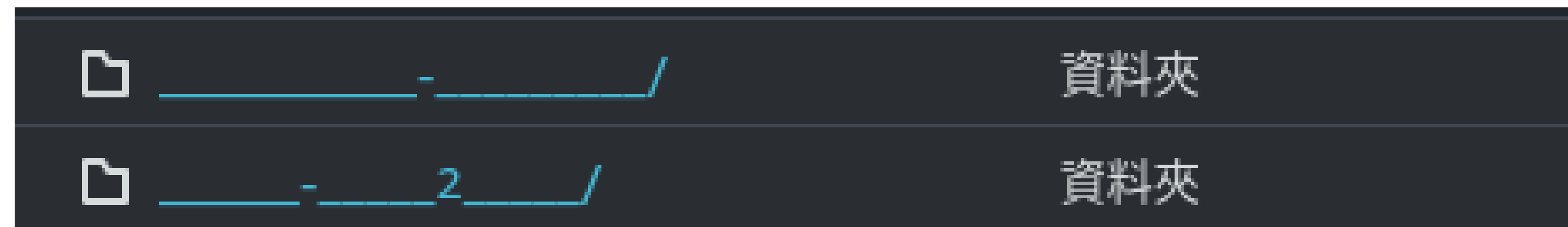
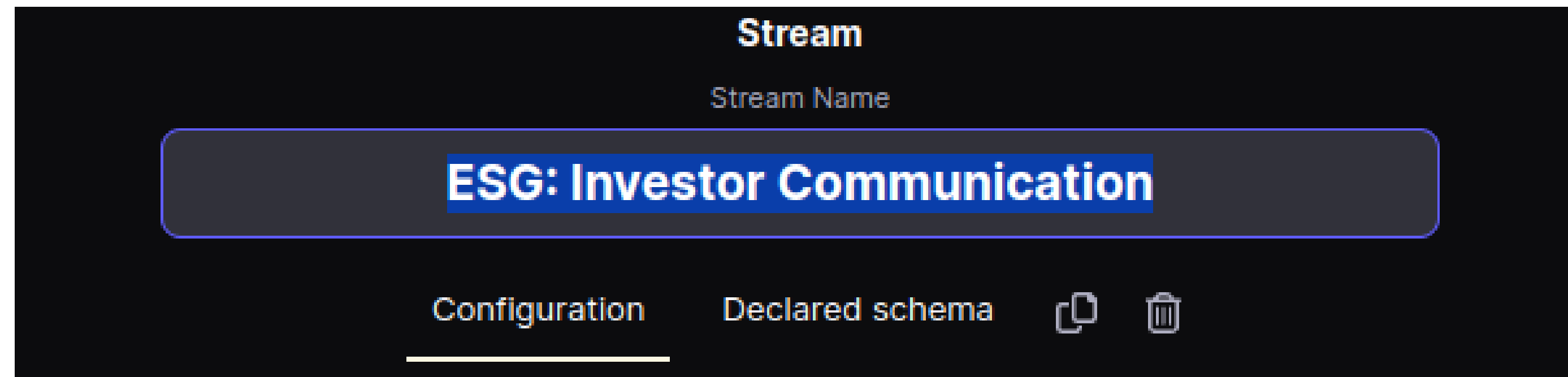
❌ 失敗

❌ HIVE_CANNOT_OPEN_SPLIT: Error opening Hive split s3://stock-analysis-pipeline-org.apache.hadoop.hive ql.io.avro.AvroContainerInputFormat: Not an Avro data
除非查詢認定為符合資格，否則會對「default」資料庫執行此查詢。請將錯誤訊息

遇到的問題

Airbyte部份不支援中文

單一stream名稱不得為中文



遇到的問題

TWSE API資料結構問題

metabase_data_source / Esg Emission Of Greenhouse Gases

篩選器匯總(Summarize)

	範疇三取得驗證	溫室氣體排放密集度(噸CO2e/單位)	溫室氣體排放密集度單位
24座製品廠)	ISO 14064	0.8033	產品
	ISO 14064	0.8390	產品
	ISO 14064	0.0043	百萬元營業額
	N	0.0123	新台幣千元產值
	N	0.5200	產品
	ISO 14064	75.1200	百萬元營業額
	N	0.0330	總生產量(噸)
i) 2.直接擁有或控制的排放源所直接排放的溫室氣體 含2-1營業所/工...	N	4.0000	百萬元營業額
	N	4.3950	百萬元營業額
	N		
	N		
	N		
	N		
	ISO 14064	0.4240	鋼胚生產量

可視化

☰☱

分析工具比較

功能豐富：

提供強大的數據分析功能，包括複雜的數據模型、關聯性探索和預測分析，能夠處理大型數據集和進行高級的數據操作。

交互式和探索性分析：

透過其“直覺式探索”功能，用戶可以輕鬆地進行交互式數據探索，發現數據之間的關係和模式。

商業級支持和安全性：

提供完善的商業級支持和安全性控制，適用於企業級使用。

學習曲線：

對於新手來說，學習需要一些時間，因為它的功能和操作方式較為複雜。



分析工具比較

易於使用：

設計用於非技術人員，界面直觀，易於上手

開源和免費：

可以自由使用滿足特定的需求，並且不需要支付許可費用。

多數據庫支持：

支持多種數據庫，包括MySQL、PostgreSQL、MongoDB等，這使得它可以連接到不同的數據源進行分析。

自助式查詢：

METABASE允許用戶進行自助式查詢，從而能夠快速地提取所需數據進行分析。

缺點：

功能限制：相較於其他商業數據分析工具，METABASE在功能和高級分析方面可能有所限制，例如進階的統計分析、機器學習整合等。



結論

- ◆ 資訊自動化、即時
- ◆ 數據邏輯性
- ◆ 資料庫基礎知識
- ◆ 滿足部分需求



The background features a series of overlapping rectangles in blue and orange. A large orange rectangle is centered behind the text. Other blue rectangles are positioned at the top left, top right, bottom left, and bottom right. Smaller orange rectangles are located at the top right and bottom center. The rectangles vary in size and are arranged in a way that creates a sense of depth and geometric complexity.

THANK YOU