**Totesys Data Engineering Project**

**Infrastructure**

Here is a rough overview of the infrastructure that will need to be in place.

A diagram of a cloud computing system

Description automatically generated

The S3 buckets are used for data storage. In this case we will need three S3 buckets:

1. Ingestion Data
2. Processed Data
3. Terraform Data

The Lambda functions are used to implement code and perform actions on our data. The three lambda functions will be:

1. Ingestion Lambda
   1. Continually ingests all tables from the ‘totesys’ database.
   2. Data to be saved in files in the ingestion bucket.
   3. Data must be of a suitable format.
   4. Operates on a schedule; every 30 minutes.
   5. Logs progress to Cloudwatch.
   6. Triggers email alerts in the event of failures.
   7. Should follow security practices (prevents SQL injection and maintain password security).
2. Transformation Lambda
   1. Remodels data into a predefined schema suitable for a data warehouse and store the data in Parquet format.
   2. Stores data in the processed S3 bucket.
   3. Triggers automatically when it detects the completion of an ingested data job.
   4. Logs progress to Cloudwatch.
   5. Populates the dimension and fact tables of a single ‘star’ schema in the warehouse.
3. Loading Lambda
   1. Loads data into a prepared data warehouse at defined intervals.
   2. Monitored with Cloudwatch.

**Infrastructure as code**

Terraform is an open source tool. It is a declarative tool that allows us to specify the infrastructure we want as code. Terraform is a command line tool rather than a service – all the API calls to create resources will come from where we run it. Terraform is useful for creating resources on AWS – but can also support other cloud infrastructure providers too!

Terraform templates contain the resources that we want to create and can optionally contain variables that can be passed in when we run the tool.

Resources are declared with a ‘resource’ block that is followed by an identifier, enclosed in quotes (“”). Resources represent entities such as virtual machines, storage buckets, networks, DNS records and many others. Anything that can be created, updated, or deleted within a provider’s infrastructure can be represented as a resource in Terraform.

Here is an example of a resource:

resource "aws\_s3\_bucket" "my\_bucket" {

bucket = "my-unique-bucket-name"

acl = "private"

}

* **Aws\_s3\_bucket**: is the resource type – in this instance representing an AWS S3 bucket.
* **My\_bucket**: is the local name assigned to this specific instance of the resource block.
* **{}**: defines the rest of the code that we want to run; for example the NAME of the S3 bucket that will be hosted on AWS.

**Create Provider**

The provider configuration file ensures that Terraform knows which provider to use for the resources you will be managing in the infrastructure. Once the provider has been configured, you can proceed to create the resources such as the S3 buckets.

A screenshot of a bucket settings

Description automatically generatedIt would be prudent to have a pre-existing bucket for Terraform to store its tfstate file in. In this case there is the bucket; ‘terraform-xrs’.

Here is our current provider.tf file:

provider "aws" {

region = "eu-west-2"

}

terraform {

required\_providers {

aws = {

*source* = "hashicorp/aws"

*version* = "5.39.0"

}

}

backend "s3" {

bucket = "terraform-xrs"

key = "tf-state"

region = "eu-west-2"

}

}

The provider specifies that we will be using Amazon Web Services. The region will specify that London is the region resources are to be saved.

TFSTATE

It is essential to be able store backend information. The S3 backend stores state data in an S3 object set by the *key* parameter. In this case it has been hard coded to be ‘terraform-xrs’.

Here is the useful documentation on backends:

<https://developer.hashicorp.com/terraform/language/settings/backends/s3>

**Create Permissions**

We will now need to implement an IAM policy that will allow writing to and from the S3 bucket. Terraform, presently can not perform any actions as it does not have permission.

Because this is the providers.tf file it is unlikely to change throughout the project. Therefore it might be prudent to directly implement the IAM policy into the file. Here are some possible reasons for this:

1. Proximity to Provider Configuration: Keeping IAM policies within the same file as the provider configuration ensures that they are closely associated with the provider settings, making it easier to understand and manage the relationship between the provider and its associated permissions.
2. Simplicity: For simpler permissions that are specific tot eh provider configuration and are unlikely to be reused across multiple configurations, storing them directly in the ‘provider.tf’ file avoids unnecessary complexity introduced by separate policy files.
3. Ease of maintenance: If the permissions are closely tied to the AWS provider and are unlikely to change frequently, storing them in the provider.tf file can simplify maintenance by keeping all relevant configuration in one place.

resource "aws\_iam\_policy" "tfstate\_policy" {

name = "terraform-tfstate-policy"

description = "IAM policy for managing TFSTATE files"

policy = jsonencode({

*Version* = "2012-10-17"

*Statement* = [

{

*Effect* = "Allow"

*Action* = [

"s3:GetObject",

"s3:PutObject",

"s3:DeleteObject",

"s3:ListBucket"

]

*Resource* = "arn:aws:s3:::terraform-xrs/\*"

}

]

})

Here is a breakdown of the previous code.

* **Name:** specifies the name of the IAM policy, in this case it will be; ‘terraform-tfstate-policy’.
* **Description:** provides a description of what the IAM policy does.
* **Policy:** Defines the policy document in JSON format using the jsonencode() function. The policy allows the specified actions s3:getObject, s3:PutObject…etc….

This policy allows the IAM entity (user, group, or role) associated with this policy to perform the specified actions on objects within the ‘terraform-xrs’ bucket.

**Initialising Terraform**

* **‘terraform init’:** initialises terraform.
* **‘terraform plan’:** starts a terraform plan.
* **‘terraform apply’:** starts to apply terraform plan on Aws.

**Setting Up the S3 Buckets**

Our first task will be setting up the S3 buckets on AWS. The buckets will need a unique name and we will also need to be able to recall the names of the buckets after they have been created!

resource "aws\_s3\_bucket" "ingestion" {

bucket\_prefix = "ingestion-"

force\_destroy = true

}

Here the resource we are using is the aws\_s3\_bucket – which will allow us to create an S3 bucket. We then specify the name of the bucket. In this instance we use the bucket\_prefix which means the name of the bucket will have a unique set of numbers after ‘ingestion-……’.

Force\_Destroy is optional, however it allows us to delete all objects from the bucket when the bucket is destroyed so that it may be destroyed without error.

resource "aws\_s3\_bucket" "ingestion" {

bucket\_prefix = "ingestion-"

force\_destroy = true

}

# Creates Process Bucket:

resource "aws\_s3\_bucket" "process" {

bucket\_prefix = "process-"

force\_destroy = true

}

# Creates Storage Bucket:

resource "aws\_s3\_bucket" "storage" {

bucket\_prefix = "storage-"

force\_destroy = true

}

resource "aws\_ssm\_parameter" "ingestion\_bucket\_name" {

name = "/ingestion"

type = "String"

value = aws\_s3\_bucket.ingestion.bucket

}

resource "aws\_ssm\_parameter" "process\_bucket\_name" {

name = "/process"

type = "String"

value = aws\_s3\_bucket.process.bucket

}

resource "aws\_ssm\_parameter" "storage\_bucket\_name" {

name = "/storage"

type = "String"

value = aws\_s3\_bucket.storage.bucket

}

This code will setup three uniquely named S3 buckets. Additionally it will save the names of the buckets to AWS Parameter Store. This could be useful later!

**Implementing CI/CD**

Continuous integration is the process of merging changes into the main repository whilst using automated tests to verify it performs as expected. Changes should be small and integrated rather than having larger launch events. By introducing small changes we can lower the risk of creating code breaking bugs into the main codebase. Tests are usually run on a dedicated server so that the integration environment is common for every engineer working on the application.

Continuous delivery is a practice where the software can be deployed easily at any time with no manual intervention from engineers. As soon as the tests have passed after integrating a change then it should be possible to deploy that change. The idea is that the faster the change can be in front of the end users of the software then the faster feedback can be gathered and acted upon – one of the central ideas of Agile.

Continuous deployment is the idea that every change in version control is automatically deployed all the way through to production without any intervention from the engineers.

GitHub Actions

To get GitHub to trigger actions on a repository, the repository needs a directory named .github at the root level, and inside that, a workflows directory. Workflow is the name given to what might be called a pipeline with other providers. There may be multiple pipelines for a project. The repo may be responsible for multiple applications deployed separately, for example, or one could separate the CI/CD parts of the pipeline. Each workflow needs a YAML file to configure.

To begin we should name our file ‘test-and-deploy.yml’. This must be installed in the root of the directory. A black background with red text

Description automatically generated

The pipeline can be triggered by various events in the GitHub architecture. Commonly for CI/CD pipelines, you would use push and pull\_requests to initiate the pipeline.

name: Test & Deploy

on: [push]

There are some checks that one may wish to run every time the code is pushed into the main repository. This is part of the continuous integration of the code into the repository. These may include unit tests, or using tools for static code analysis – tools that check code without running it – such as linting, assessing vulnerabilities and other code quality checks.

Other processes may only make sense to run once we are happy that the code has passed previous checks, plus any other manual code reviews that need to be undertaken anyway – this might involve a staging area so checks for things like UX and accessibility can be signed off, or it might mean straight into production as we will do – that’s continuous deployment.

As the workflow will handle all of the CI/CD, the workflow file can be configured to run whenever a specific event happens. In this case, some code being pushed to the main branch. This will allow us to work on other branches without the pipeline being initiated.

on:

push:

branches:

- main

Jobs

GitHub Actions allows you to define separate tasks, or jobs that you may wish to undertake in the pipeline. With this structure we can separate the processes from each other, and provide different contexts or environments for them to be undertaken in.

jobs:

run-tests:

name: run-tests

runs-on: ubuntu-latest

steps:

- name: Checkout Repo

uses: actions/checkout@v4

On this run-tests job property we have defined a human-readable name, used by the GitHub UI to inform us about the pipeline’s progress. Additionally an OS has been provided for the job to be run on. Typically we should be emulating the environment of our deployment. On GitHub Actions, Linux machines are the cheapest, so ubuntu-latest would be a sensible option.

For our project, initially we want to ‘just’ setup a basic terraform deployment – we can run checks later.

jobs:

deploy:

runs-on: ubuntu-latest

steps:

**Note:** There are common github actions provided by the community:

<https://github.com/actions/setup-python>

This is the ‘setup-python’ github action that we can use for this project.

Jobs can be further broken down into steps that can be undertaken sequentially. The first step of our code is specific to GitHub Actions – it is necessary to **checkout** the code that we are working with. In a GitHub CI/CD workflow, checkout out the code is necessary because the CI/CD pipeline needs access to the source code of the repository in order to perform various tasks such as running tests. The source code of a project refers to the human-readable files and documents that contain the instructions and logic necessary to build and run the software application.

Next we will need to setup a python interpreter. There is a common GitHub-made Action we can use. We can provide information to the action using the ‘with’ key. Here we can use the latest Python version (at time of writing).

- name: Use Python

uses: actions/setup-python@v5

with:

python-version: '3.12'

Make

The next task our workflow needs is to create a virtual environment, install the project requirements and run the tests with the correct PYTHONPATH. It is possible to simplify instructions using a Makefile. We can use makefiles to run a sequence of commands.

The Makefile is a repository for code at the root of the project that is accessed by the make command. Make is installed on MacOS along with the XCode utility.

- name: Application Requirements

run: make requirements

This command will run our makefile.

Deployment

Assuming our continuous integration checks work, we can rigger a deployment action. We can create a job which runs a deployment script file. The one thing we need to do differently is how we authenticate to AWS – the virtual machine will not have access to our AWS credential files.

In order to deploy, we will need to provide the deployment script with access to the AWS Access Key ID and Secret Key.

**Note:** we do not want these to be hard coded into our files! This would create a security vulnerability. The solution is to declare ‘Secrets’ from within the Github console. A better way is to make these specific.

We can use this github actions: https://github.com/aws-actions/configure-aws-credentials

In order to deploy to AWS we need the following secrets:

* AWS\_ACCESS\_KEY\_ID
* AWS\_SECRET\_ACCESS\_KEY
* AWS\_REGION

- name: Implement AWS Credentials

uses: aws-actions/configure-aws-credentials@v4

with:

aws-region: ${{ secrets.AWS\_REGION }}

aws-secret-access-key: ${{ secrets.AWS\_SECRET\_ACCESS\_KEY }}

aws-access-key-id: ${{ secrets.AWS\_ACCESS\_KEY\_ID }}

We now need to add our secrets to github. To do this, simply go to the repository you are working in and add the secrets:

A screenshot of a computer

Description automatically generated

Here is the final CI/CD file:

name: Test & Deploy

# Specify deployment intialisation.

on:

push:

branches:

- main

jobs:

# Tests

# Deploy

deploy:

runs-on: ubuntu-latest

steps:

- name: Checkout Repo

uses: actions/checkout@v4

- name: Use Python

uses: actions/setup-python@v5

with:

python-version: '3.12'

- name: Application Requirements

run: make requirements

- name: Implement AWS Credentials

uses: aws-actions/configure-aws-credentials@v4

with:

aws-region: ${{ secrets.AWS\_REGION }}

aws-secret-access-key: ${{ secrets.AWS\_SECRET\_ACCESS\_KEY }}

aws-access-key-id: ${{ secrets.AWS\_ACCESS\_KEY\_ID }}

- name: Deploy Terraform

uses: hashicorp/setup-terraform@v3

- name: Terraform Init

working-directory: terraform

run: terraform init

- name: Terraform Plan

working-directory: terraform

run: terraform plan

- name: Terraform Apply

working-directory: terraform

run: terraform apply -auto-approve

run: terraform init

- name: Terraform Plan

working-directory: terraform

run: terraform plan

- name: Terraform apply

working-directory: terraform

run: terraform apply -auto-approve

This ultimately creates a very basic setup. One of the advantages of this is that the tfstate is stored in a predefined file which can be read. The advantage of this, is that every time the CI/CD pipeline is initiated, terraform will read from the tfstate to see which jobs need to be completed. This means that if we already have said up an ‘ingestion’ bucket; it will not be setup again.

We now have a working CI/CD pipeline:

A screenshot of a computer

Description automatically generated

We can verify that three buckets have been created; ingestion, process and storage.

A screenshot of a computer

Description automatically generated

Now that we have a working CI/CD pipeline with a very basic terraform initialisation, we can begin to think about building our first Lambda and incorporating it into the CI/CD pipeline.

**Lambda 1 – Ingestion**

We know that the first lambda should;

1. Continually ingests all tables from the ‘totesys’ database.
2. Data to be saved in files in the ingestion bucket.
3. Data must be of a suitable format.
4. Operates on a schedule; every 30 minutes.
5. Logs progress to Cloudwatch.
6. Triggers email alerts in the event of failures.
7. Should follow security practices (prevents SQL injection and maintain password security).

**Creating Functions**

To implement the first lambda successfully, it would be useful to setup some helper functions. For basic functionality we need to consider:

1. A function that can access the time stored in AWS Parameter Store.
2. A function that could update the time stored in AWS Parameter Store.
3. A function that can scan the totesys database and extract information updated since last update.
4. A function that can take our data and store it in an S3 bucket.

Following the creation of the functions, they will need to be tested. All functions should be developed through Test Driven Development to ensure they function correctly.

**Function 1 – “**get\_bucket\_names()”

One of the very first functions we should consider, is a function that is able to read the tfstate file and is able to extract the names of the buckets that have been created.

Fortunately during CI/CD deployment, Terraform wrote the names of the buckets to an AWS Parameter Store file. Therefore it should be a simple matter of retrieving them.

*def* get\_bucket\_names():

"""

Args:

------

None.

Returns:

------

Object containing names of S3 buckets. Example:

{'ingestion': 'ingestion-20240304201826545600000001',

'process': 'process-20240304201826547200000003',

'storage': 'storage-20240304201826546800000002'}

"""

client = boto3.client('ssm')

bucket\_obj = {'ingestion': None, 'process': None,

'storage': None}

for name in bucket\_obj:

bucket\_obj[name] = client.get\_parameter(

*Name*=*f*"/{name}")['Parameter']['Value']

return bucket\_obj

**Function 2 –** ‘get\_aws\_time()’

In order to determine which data we need to lift out of the totesys database, we will need to know the time we last scanned the database. To do this we will need to extract the time out of the AWS Parameter Store.

*def* get\_aws\_time():

"""

## Args:

None.

---

## Returns:

---

- datetime object stored in aws parameters store.

Returns a message: "File path created: {date\_str}. File added!"

This function accesses the time stored in AWS Parameter Store.

This function will return a datetime OBJECT.

"""

client = boto3.client('ssm')

str = client.get\_parameter(*Name*='/time')['Parameter']['Value']

return datetime.strptime(str, '%Y-%m-%d %H:%M:%S.%f')

**Function 3 –** ‘update\_aws\_time()’

Assuming we can extract the time. Note – the default time is set during the Terraform Init phase (to around 1900). Setting the default time when Terraform starts should ensure that ALL data is pulled from the database the first time our Lambda function will run.

After the Lambda function runs, we then need to update the time in the AWS Parameter Store. To do this we need a new function.

*def* update\_aws\_time(*datetime*):

"""

## Args:

---

datetime object.

---

## Returns:

Updates time in aws parameter store. Does not return anything.

---

- `str`

Returns a message: "File path created: {date\_str}. File added!"

This function will update the time stored in the AWS Parameter Store.

The function should take a datetime object.

It does not return anything.

"""

to\_string = *datetime*.strftime('%Y-%m-%d %H:%M:%S.%f')

client = boto3.client('ssm')

client.put\_parameter(

*Name*='/time',

*Value*=to\_string,

*Type*='String',

*Overwrite*=True)

**Function 4 –** ‘all\_data()’

The purpose of this function is to scan the PSQL database and make a datetime comparison. If the last updated time is in the past compared to a more modern ‘last\_updated’ in the PSQL database, then the function will extract that information.

This function will assemble a json string, so that we can laster store the information in our first S3 bucket!

*def* all\_data():

"""

---

## Args:

---

No Arguments.

---

## Returns:

---

- `str`

Returns a json string with the latest data.\n

```

This function will compare 'last\_updated' fields to\n

time stored in aws\_parameter store. If the last\_updated time\n

is ahead of aws\_parameter store, it will pull\n

that information out of the database. \n

It is used to scan the database to find recently updated records.

When it has finished, it will update the time in the aws\_parameter store.

"""

# Returns datetime object from last query time.

last\_query\_time = get\_aws\_time().strftime('%Y-%m-%d %H:%M')

table\_dict = {'counterparty': None, 'currency': None,

'department': None, 'design': None,

'staff': None, 'sales\_order': None,

'address': None, 'payment': None,

'purchase\_order': None, 'payment\_type': None,

'transaction': None}

# Establish a connection to the PostgreSQL database

con = wr.postgresql.connect(*secret\_id*='new\_tote')

data = {}

# Iterate through table\_dict

for table in table\_dict:

# For each Table

cursor = con.cursor()

query = (*f*"SELECT \* FROM {table} "

*f*"WHERE last\_updated > '{last\_query\_time}';")

cursor.execute(query)

# Establish names of columns

column\_names = [col\_desc[0] for col\_desc in cursor.description]

rows = con.run(query)

list = []

# Iterate through each ROW.

for row in rows:

# Create temp dictionary

temp = {}

# Iterate through data in each row:

for i, x in enumerate(row):

# Add to my\_row dictionary: column name (i).

# Value is (x)

temp[column\_names[i]] = x

# Append dictionary to list.

list.append(temp)

# Add to dictionary. Table name is key. List is value.

data[table] = list

con.close()

# Convert into a JSON string:

json\_str = json.dumps(data, *default*=str, *indent*=2)

update\_aws\_time(datetime.now())

return json\_str

**Function 5** – ‘create\_path\_add\_file()’

The purpose of this function is to store our json string from all\_data(). The function will first check to ensure that a valid json string has been provided. If it has not, it will raise an error.

If a valid JSON file has been provided, it will then check to make sure that there is data for all the values.

Our previous function may have returned this:

A screen shot of a computer code

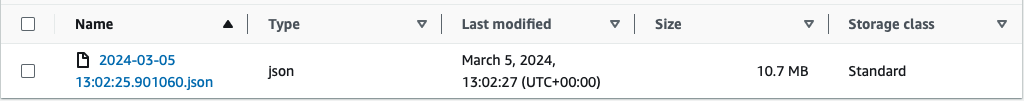
Description automatically generated

If there are no values, then the function will return ‘None’. If there are values, it will try to store the json file in the specified bucket. By default it will try to place items into the S3 Ingestion bucket.

The function will create a file path to keep data organised:



Year, Month, Day, Hour is the specified format. This way we select the exact json file we want if we need to extract data later.

Finally the function will upload at the specified file path. 

Here is the final code:

*def* create\_path\_add\_file(*file*, *bucket\_name*=bucket\_names['ingestion']):

"""

---

## Args:

---

JSON String.

bucket\_name - OPTIONAL

Default bucket name is 'ingestion' bucket name.

---

## Returns:

---

- `str`

Returns a message: "File path created: {date\_str}. File added!"

This function creates a file path in an S3 bucket and

adds a json to the created file path.

In order to maintain a clean file structure within

the S3 bucket, this function will try to create a unique file path.

The file path is set to the time the function is

run and will create a file path in this format:

Year/Month/Date/Hour

2023/2/16/12

This assumes that the function was run about

12pm on the 16th February 2023.

If the file path already exists, it will not create a new file path.

This function accepts two arguments; a json string and a bucket name.

The bucket name is set to a default value of 'ingestion-123'.

If the file is NOT a json file, it will raise a value error.

If the file IS a json file, it will create a file path

(assuming it does not exist already) and place

the json file in the final folder.

"""

# Check that File is a VALID JSON file.

try:

dict\_values = json.loads(*file*).values()

print("Valid JSON provided. Checking for values...")

except (json.JSONDecodeError, TypeError):

raise ValueError("File is not valid JSON format.")

# Check JSON not empty

# If JSON totaly empty, returns.

if all(len(value) == 0 for value in dict\_values):

print('No new values detected. Update unnecessary')

return None

s3 = boto3.client('s3')

current\_time = datetime.now()

# Create a date dictionary to iterate over.

date\_dict = {'year': current\_time.year, 'month': current\_time.month,

'day': current\_time.day, 'hour': current\_time.hour}

# Date string added to to create file path.

date\_str = ""

# Iterate through date dictionary:

for time in date\_dict:

date\_str += *f*"{date\_dict[time]}/"

try:

# Checks if file path exists.

s3.head\_object(*Bucket*=*bucket\_name*, *Key*=date\_str)

print(*f*"File path: {date\_str} already exists continuing...")

continue

except Exception as e:

print(e)

# If file path does not exist. Create file path.

try:

print(

*f*"File path: {date\_str} does not exist."

*f*"Creating file path now...")

s3.put\_object(*Key*=date\_str, *Bucket*=*bucket\_name*)

print(*f*"{date\_str} created successfully.")

except Exception as e:

print(e)

return 'Unable to create file path.'

# File path exists, will now attempt to place

# JSON object into the file path location.

# Create a unique file name.

file\_name = *f*"{date\_str}{current\_time}.json"

# Add file to the file location created earlier.

s3.put\_object(*Body*=*file*, *Bucket*=*bucket\_name*, *Key*=file\_name)

# Update AWS Time

# update\_aws\_time(datetime.now())

print(*f*"File path created: {date\_str}. File added!")

return *f*"File path created: {date\_str}. File added!"

**Lambda Layers**

Before we write the first ingestion Lambda. We need to take into consideration that the functions written previously, will not be included in the Lambda itself. This is because the functions, libraries and dependencies will need to be shared across multiple lambda files. The advantage of this, is that if a function needs to be modified or updated, it only needs to be updated in one place.

Here is a rough example of a terraform lambda layer:

resource "aws\_lambda\_layer\_version" "lambda\_layer" {

filename = "lambda\_layer\_payload.zip"

layer\_name = "lambda\_layer\_name"

compatible\_runtimes = ["nodejs16.x"]

}

Currently on the local machine the FOLDER that we want to package and zip up is called src. It will contain the file functions.py. In functions.py we can add all of the functions that will be utilised in our Lambdas. We need to zip this up, and implement it into AWS lambda layers.

A screenshot of a computer

Description automatically generated

Fortunately [Terraform](https://registry.terraform.io/providers/hashicorp/archive/latest/docs/data-sources/file) can assist us with this: data “archive\_file” is a Terraform utility that will generate an archive from content, a file or directory of files.

data "archive\_file" "init" {

type = "zip"

source\_file = "${path.module}/init.tpl"

output\_path = "${path.module}/files/init.zip"

}

This is the basic syntax. And here is what it looks like in the project:

data "archive\_file" "lambda\_layer\_functions\_zip"{

type = "zip"

output\_path = "${path.module}/lambda\_layer.zip"

source\_dir = "${path.module}/../src"

}

“lambda\_layer\_functions\_zip” is simply the name of this code. We will need to reference it later, because there is important information that we will need to extract; the file location for example.

Type simply specifies that we want a zip file – the type of file that AWS needs.

A screenshot of a computer

Description automatically generatedUnderstanding the file locations. On the local machine, the root folder is ‘totesys-de’. Terraform is a folder within the root directory (totesys-de/terraform). The purpose of {path.module} is to define the location of the terraform folder. Path.module effectively becomes the ‘terraform’ folder from our local machine. This means we can define the location of the files/folders that we now want to package up.

For our source\_dir we can see that we start from the ‘terraform’ folder and then jump up a directory before jumping back into the src folder, which is the folder we are going to package up and include as a layer.

This code will have created a zip file. We then need to specify this file when we are uploading to AWS Lambda/layers.

resource "aws\_lambda\_layer\_version" "layer\_one" {

filename = data.archive\_file.lambda\_layer\_functions\_zip.output\_path

layer\_name = "first\_layer"

compatible\_runtimes = ["python3.12"]

}

Notice that it references data.archive\_file….. This way we can extract the correct file name every time.

If successful we will see the first\_layer uploaded to AWS. Now we have access to functions.py from our other lambdas!

A screenshot of a computer

Description automatically generated

**Linking first Lambda to Lambda Layer**

After creating

**Attaching Policies to Lambda Ingestion**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedIn order to allow our Lambda function access to resources within AWS, we need to consider creating an IAM role. When we create an IAM role we can specify the name that we want; in this case it is ‘iam\_for\_ingestion’. We then can specify the policies that we want to attach to the role.

Here we can see that the ingestion lambda currently has three policies attached to it.

This means that when we call a function from our ingestion lambda, it will be able to access these specified resources.

Here is what it looks like in Terraform:

# Create IAM role.

resource "aws\_iam\_role" "iam\_for\_lambda" {

name = "iam\_for\_ingestion"

assume\_role\_policy = data.aws\_iam\_policy\_document.assume\_role.json

}

First we create the role. This is our ‘iam for ingestion’ role.

Note that while we have named the policy ‘iam\_for\_ingestion’ – this is only how it will be referenced within the AWS console. In order for us to reference it within terraform, we need to use the “aws\_iam\_role” – so that we can attach additional policies.

Now we need to attach the policy to our ‘iam\_for\_ingestion’ role.

# Attach Secrets Access policy to IAM role.

resource "aws\_iam\_role\_policy\_attachment" "secret\_access\_policy" {

role = aws\_iam\_role.iam\_for\_lambda.name

policy\_arn = "arn:aws:iam::aws:policy/SecretsManagerReadWrite"

}

**Deployment**

After successful deployment we should see our ingestion lambda uploaded into AWS:

A screenshot of a computer

Description automatically generated

It should have a layer attached:

A screenshot of a computer

Description automatically generated

Which links to our ‘first\_layer’:

A screenshot of a computer

Description automatically generated

And we should see a window with our code. Note that for now, some very simple rudimentary code has been implemented to demonstrate functionality.

A screenshot of a computer

Description automatically generated

If we run the code we can see the logs:

A screenshot of a computer

Description automatically generated

Notice that we have an interesting problem. We have a timeout message, but we also have some logs.

Interestingly our print(all\_data()) command, which was included within our ‘layer’ has printed *some* of the data.

A screenshot of a computer program

Description automatically generated

This confirms that our function is establishing a connection to our PSQL database and extracting some information. However, there is a lot of data and it takes time to extract the information!

What is likely happening is this:

*def* handler(*event*, *context*):

names = get\_bucket\_names()

date = get\_aws\_time()

data = all\_data()

return names

The lambda function is trying to save the results of calling these functions to variables. The ‘data’ variable is taking a long time to extract the information, and is therefore causing a timeout to occur. Because of our print() function in our layer – it is causing some of the data to print out! What we can do to solve this issue is to increase our timeout time and our memory allocated to our lambda to verify it is working:

Increase the timeout time and memory allocation to resolve this issue:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe result is this: Although we are returning our bucket names function, we are also getting function logs (from the print statement!). Now it is not timing out. For reference this took approximately 20-30 seconds to execute!

**Immutability**

<https://docs.aws.amazon.com/AmazonS3/latest/userguide/object-lock.html>

Immutability refers to the inability to modify data once it is written. When we are storing data in our S3 bucket, it is extremely important that this data is not edited in anyway.

Fortunately Amazon allows us to use a write-once-read-many (WORM) model to store objects. This means we can write to our S3 bucket and then prevent this data from being written over.

There are two main methods:

Retention period = specifying a fixed period of time in which an object remains locked.

Legal hold = same protected but without an expiration date.

Note that when using Terraform to remove infrastructure, it will be unable to delete the S3 bucket and it will report an error indicating that there is an object lock policy in place.

**Transformation**

In data engineering, transformation refers to the process of converting data from one format or structure into another. This typically involves manipulating, cleaning, aggregating, or otherwise modifying data to prepare it for analysis, storage or presentation.

Eventually we need to create a ‘data warehouse’. One of the fundamental differences of a data warehouse compared to a data lake is that a data warehouse is specifically designed for structured data within a predefined schema.

To this end, we need to consider how we want our structure to look like. We then need to consider how we can achieve this and transform our original data into our new schema. Here is the new schema:

**Sales Schema:**

A screenshot of a computer

Description automatically generated

**Purchases Schema:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated**Transaction Schema:**

Now that we know the structure of our schemas, we can begin to think about our transformation process. One method, would be to create SQL queries so that we are able to extract the correct data from our original database.

Another method, would be to read from the JSON file and build up our data that way. If we look at our transaction schema, we can see the structure that we would need to create.

Currently our JSON file is residing in our S3 ingestion bucket.

A screenshot of a computer

Description automatically generated

**Dealing with multiple Files**

When we write to our S3 bucket, it is worth thinking about our second Lambda. How can the Lambda access the file names? Which files should it access first if there are multiple? How can we stop reading the same file multiple times?

There are various methods to keep track of this, but perhaps a convenient one would be to use the AWS Parameter Store. We can use it to keep a JSON string.

A screenshot of a computer

Description automatically generated

**Building The Schema**

**Counterparty**

A screenshot of a computer

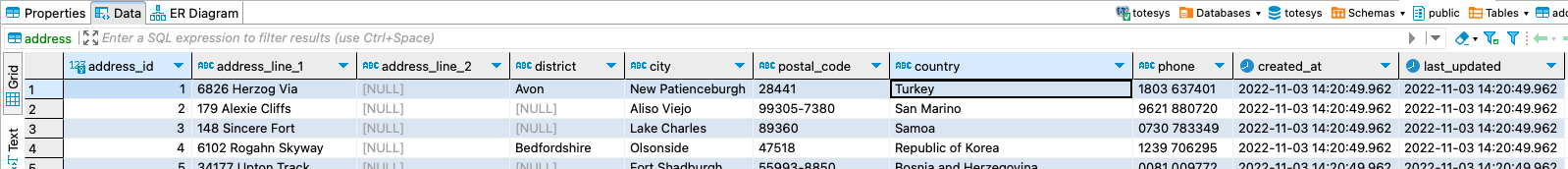
Description automatically generated

Here is what the dim\_counterparty table needs to look like.

A screenshot of a computer

Description automatically generatedIf we look at the current counterparty table:

It is evident that we do not have the necessary fields to build up our dim\_counterparty table. The remaining fields come from our address table:



A screenshot of a computer

Description automatically generated

Here is a visual diagram of the relationship between the tables.

The counterparty table will therefore is a conglomeration of two different tables. Here is roughly how it works:

|  |  |
| --- | --- |
| **Dim\_counterparty** | |
| Row Name | Table Origin |
| Counterparty\_id | Counterparty\_counterparty\_id |
| Counterparty\_legal\_name | Counterparty\_counterparty\_legal\_name |
| Counterparty\_legal\_address\_line\_1 | Address\_address\_line\_1 |
| Counterparty\_legal\_address\_line\_2 | Address\_address\_line\_2 |
| Counterparty\_legal\_district | Address\_ district |
| Counterparty\_legal\_city | Address\_ city |
| Counterparty\_legal\_postal\_code | Address\_ postal\_code |
| Counterparty\_legal\_country | Address\_country |
| Counterparty\_legal\_phone\_number | Address\_phone |

We now need to write a python file that will create dim\_counterparty.

*def* dim\_counterparty(*file*):

# Check there is data in 'counterparty'

try:

len(*file*['counterparty']) > 0

counterparty = *file*['counterparty']

print('Data in counterparty.')

except Exception as e:

print(e)

print('No data in counterparty.')

# Check there is data in 'address'

try:

len(*file*['address']) > 0

address = *file*['address']

print('Data in address.')

except Exception as e:

print(e)

print('No data in address.')

# Create 'dim\_counterparty' object.

dim\_counterparty\_obj = {'dim\_counterparty': []}

for record in counterparty:

schema = {'counterparty\_id': record['counterparty\_id'],

'counterparty\_legal\_name': record['counterparty\_legal\_name']}

for item in address:

if item['address\_id'] == record['legal\_address\_id']:

schema['counterparty\_legal\_address\_line\_1'] = item['address\_line\_1']

schema['counterparty\_legal\_address\_line\_2'] = item['address\_line\_2']

schema['counterparty\_legal\_district'] = item['district']

schema['counterparty\_legal\_city'] = item['city']

schema['counterparty\_legal\_postal\_code'] = item['postal\_code']

schema['counterparty\_legal\_country'] = item['country']

schema['counterparty\_legal\_phone\_number'] = item['phone']

break

dim\_counterparty\_obj['dim\_counterparty'].append(schema)

return dim\_counterparty\_obj

Here we iterate through each record of our counterparty data. Then we can use a separate for loop to iterate through the address table looking for a record where the address\_id matches the address id in our counterparty table. Once we have found this record, we can then combine the information into our dim\_table.

A computer screen shot of a computer code

Description automatically generated

Here is the result.

Finally we need to transform the data into a pandas dataframe.

<https://pandas.pydata.org/docs/user_guide/index.html>

A screenshot of a computer

Description automatically generatedFrom the user guide we can see that if we have a dictionary, for the value ‘A’, each value will become a separate row.

This is what our dim\_counterparty object will look like:

{'dim\_counterparty': []}

This means that each value will be an object within the list.

Which produces a data frame that looks like this:

A screen shot of a computer

Description automatically generated