

# CI-CD en Azure Data Factory

Por Santos Nicanor Oliva Escobar









# Que es CI-CD

Se utilizan las siglas CI-CD para hacer referencia a Integración Continua, Entrega Continua.

- La integración continua es el procedimiento de probar cada cambio realizado en el código base automáticamente y tan pronto como sea posible.
- La entrega continua sigue las pruebas realizadas durante la integración continua y envía los cambios a un sistema de ensayo o producción.

En Azure Data Factory, la integración y la entrega continuas (CI/CD) implican el traslado de pipelines de Data Factory de un entorno (desarrollo, prueba o producción) a otro.

Azure Data Factory usa las plantillas de Azure Resource Manager para almacenar la configuración de las distintas entidades de ADF (canalizaciones, conjuntos de datos, flujos de datos, etc.).

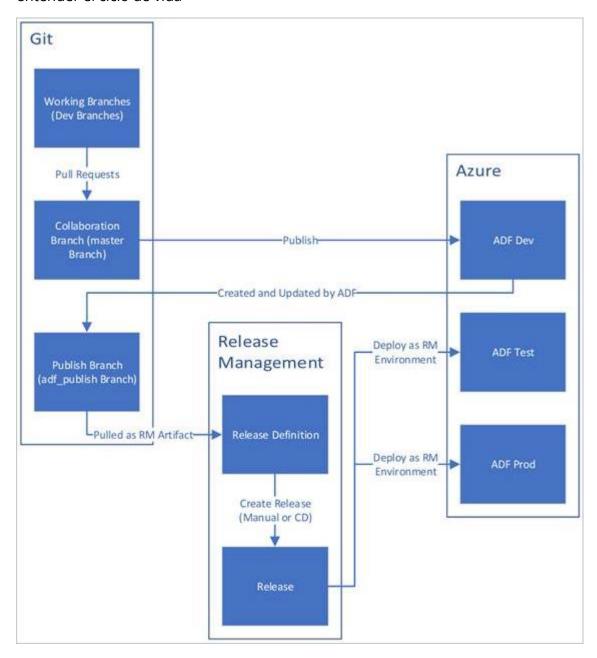






# Ciclo de vida

Para trabajar integración continua en Azure DataFactory primero es necesario entender el ciclo de vida

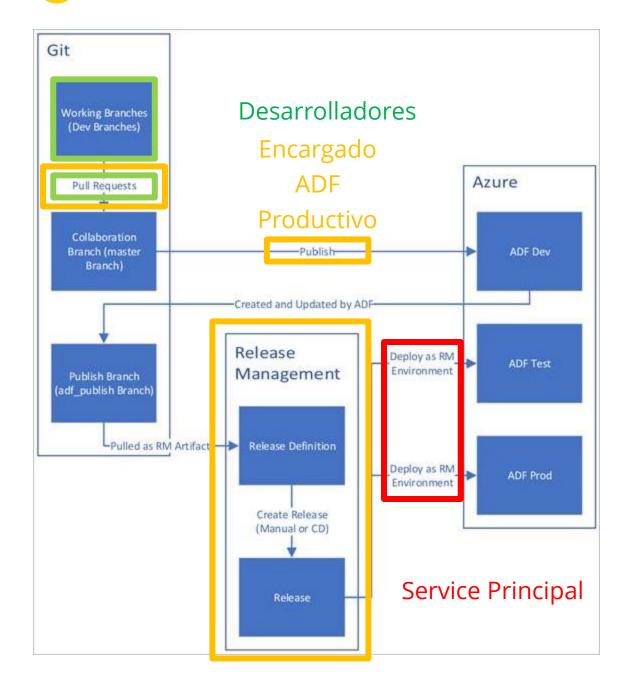




















#### Acá podemos observar:

- 3 ramas de GIT
  - Dev branches
  - Master branch
  - o Adf-publish branch
- 3 Recusos ADF
  - o Dev
  - Test
  - o Prod
- La herramienta de gestión de Releases
  - Los cuales se pueden gestionar manual o automáticamente

Antes de explicar el paso a paso es importante comprender la lógica con la cual se trabajará:

- En el recurso **ADF Dev** vamos a trabajar con Git, trabajando sobre distintas ramas,
  - La rama Master branch será la única que podra publicar en nuestro recurso ADF y no se podrán hacer cambios en ella sino a través de pull request.
  - Para realizar cambios se puede trabajar creando distintas ramas de master e ir eliminándolas a medida que se realiza merge o realizar todos los cambios en una rama **Dev branch.**
  - o **Adf-publish branch** es una rama creada automáticamente que guarda todo lo publicado en nuestro recurso ADF-dev.
- Una vez realizado el Publish en ADF-Dev, Devops se encargará de hacer un deploy de lo publicado en los entornos ADF-Test y ADF-Prod.
- Solo el Service Principal definido podra hacer modificaciones en el ADF productivo, a través de los sucesivos deploys

















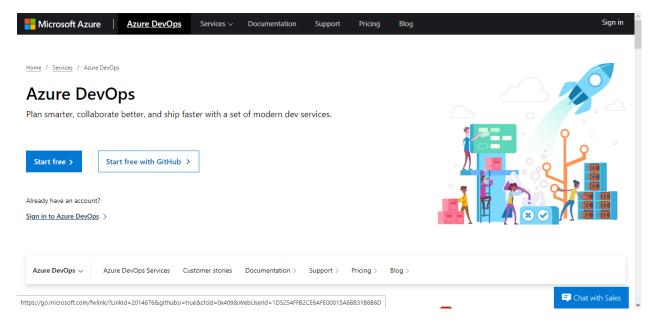
# Repositorios

Teniendo en cuenta esto, y ya teniendo los distintos recursos ADF trabajando, crearemos un repositorio GIT, tanto para la integración continua como para control de versión.

Esto se puede trabajar utilizando un repo de GITHUB, pero teniendo en cuenta la existencia de Azure Repos GIT que permite que el código sea privado de manera gratuita, utilizaremos dicha herramienta.

# Creación DevOps

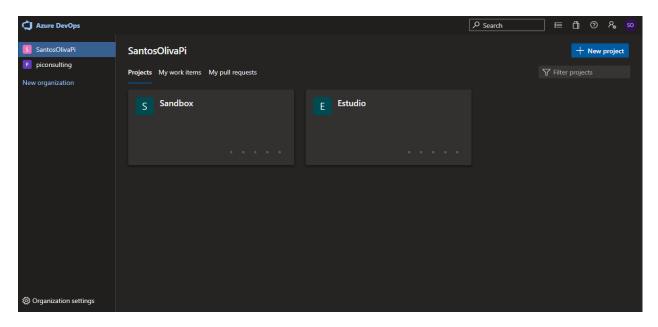
Entramos a dev.azure.com y logueamos con nuestra cuenta





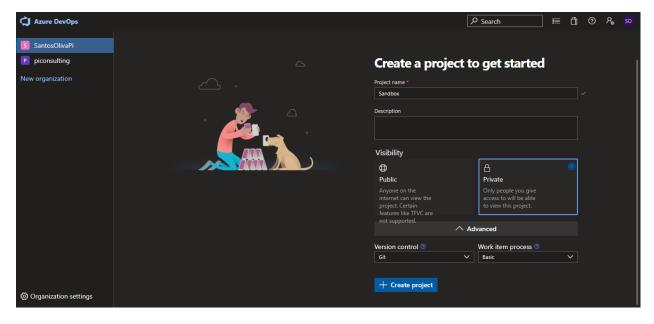






A la izquierda podemos seleccionar la organización, en mi caso, cree una organización propia con motivos de práctica.

En el cuadro central se observan los proyectos, y la opción Nuevo proyecto, donde se crea un proyecto nuevo, seleccionando su visibilidad.







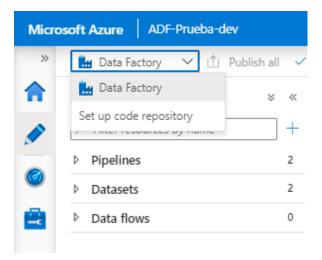




# Configuración ADF

El siguiente paso se trabaja desde Azure Data Factory.

Una vez dentro de la sección de edición. Se hace click en el combo box "Data Factory" y se selecciona Set up code repository.







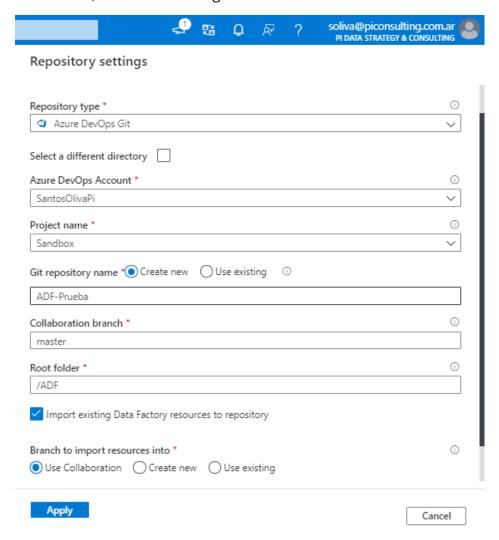




Aquí, configurando el repositorio, seleccionaremos como tipo Azure DevOps Git, seleccionamos el nombre del proyecto, en nuestro caso "Sandbox" y el repositorio Git.

Nosotros crearemos un repositorio desde aquí, dándole nombre y seleccionando la rama de colaboración, esta suele ser master. Esta rama será la única con posibilidad de realizar Publish en nuestro recurso ADF.

También seleccionaremos la opción de guardar lo trabajado en nuestro ADF en nuestro Git, en nuestro caso guardaremos en la rama de colaboración.



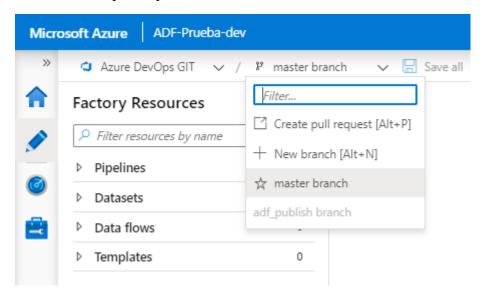








Una vez aplicadas estas configuraciones, haremos click en la rama master, en la cual estamos trabajando y seleccionaremos crear una nueva rama



Ahí creamos una rama Dev, donde realizaremos todos los cambios del ADF





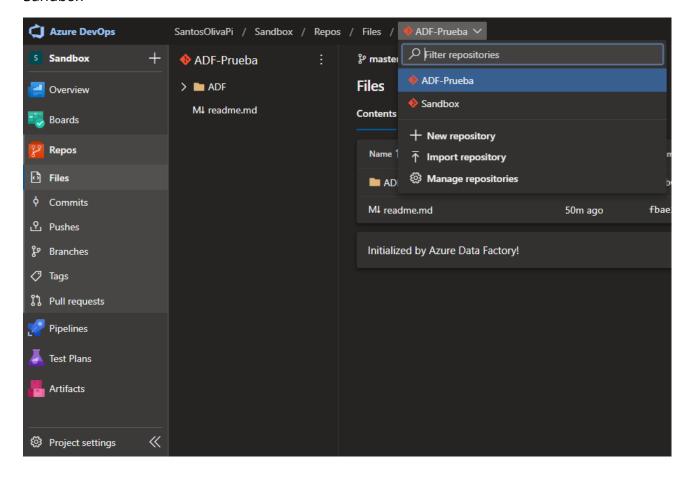


# Gestión de repositorios en Azure DevOps

Volviendo al entorno Azure DevOps, posicionándonos en

Repos - Files

Y luego haciendo click en el lugar que se indica en la imagen veremos el path donde guardamos nuestro código, como en nuestro caso llamamos al root ADF-Prueba por el recurso utilizado, toda nuestra información se encontrara allí, en lugar que en Sandbox





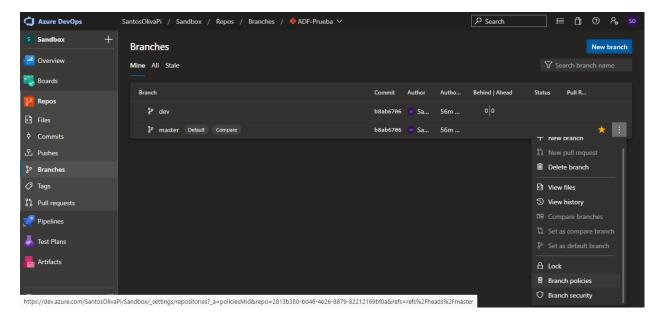






Haciendo click en la opción Branches, podremos gestionarlas.

Como no deseamos que se realicen cambios directamente sobre la rama master sino por pull requests, modificaremos sus políticas haciendo click en los 3 puntos situados a la derecha del mismo y seleccionando "Branch policies"



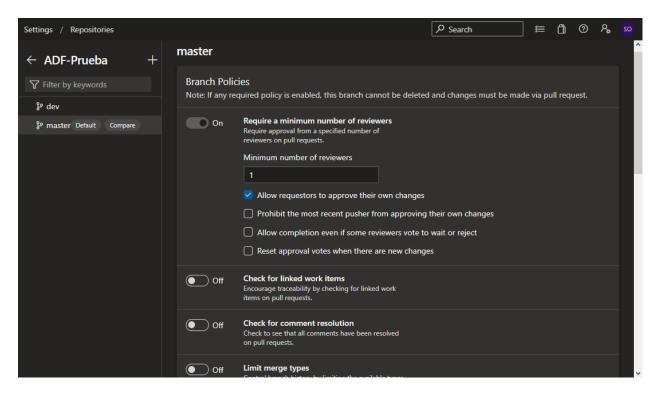
Aquí haremos las configuraciones correspondientes a las políticas deseadas, como este es un caso de ejemplo, seleccione las que se muestran a continuación.





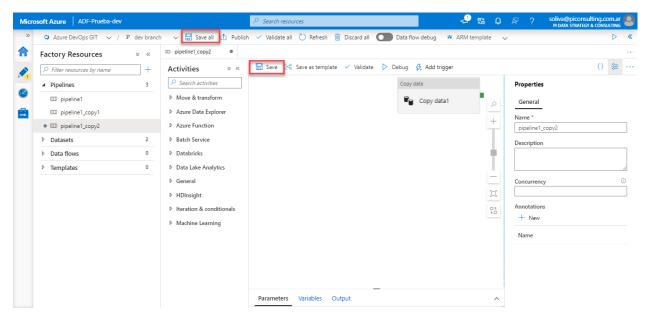


Autor: Santos Nicanor Oliva Escobar



Con esto seteado, probamos como es el flujo desde la Branch dev hasta la adf\_publish

Realizando cualquier cambio en ADF, en nuestro caso clonamos un pipeline, se guarda haciendo click en SAVE o SAVE ALL opciones no disponibles previamente











Si quisiéramos publicar estos cambios, encontraríamos que no esta disponible tal opción, nos aparece esta notificación

Publish is only allowed from collaboration ('master') branch.

Merge the changes to 'master'.

OK



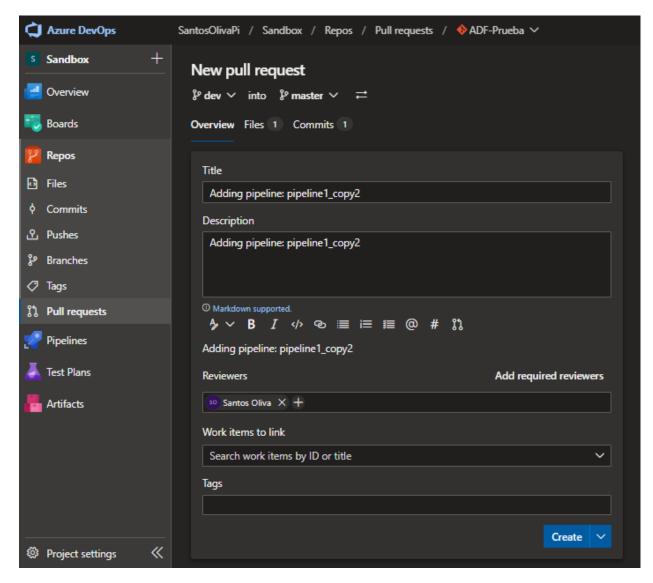








Haciendo clic en Merge, podremos observar la siguiente ventana, una vez seleccionadas las opciones deseadas, por ejemplo los Reviewers necesarios y haremos click en Create



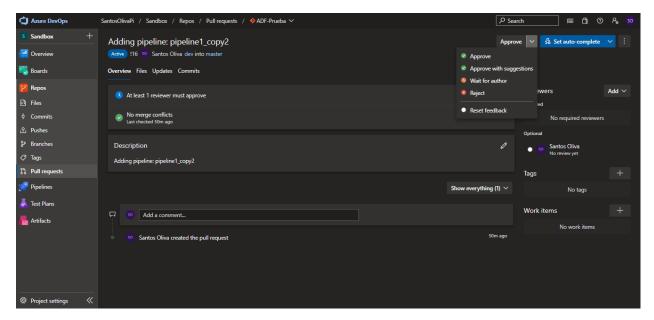




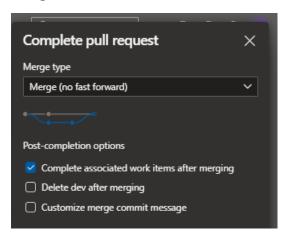




Aquí quien deba revisar, dará su feedback o aprobará el cambio, y hará click en complete



Se gestionara las opciones deseadas en cuanto a las ramas de Git, y luego Completar Merge



Una vez realizado este paso, los cambios se podrán ver en la rama Master, que ahora podremos observar desde nuestro ADF, con los cambios implementados en la rama Dev.

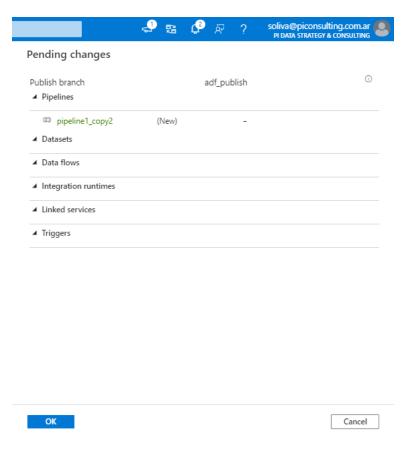
Aquí, podremos publicar los cambios trabajados, que impactaran en la rama adf\_publish, que representa lo publicado en el recurso ADF correspondiente.











Con esto, ya terminamos nuestro trabajo en ADF-Prueba-dev



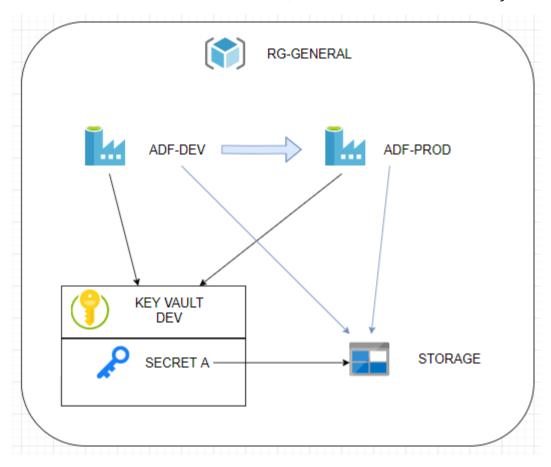






# Arquitectura

La arquitectura de datos utilizada en nuestro caso fue la siguiente, por lo tanto, lo único a cambiar en Override Parameters, era el nombre del Data Factory.



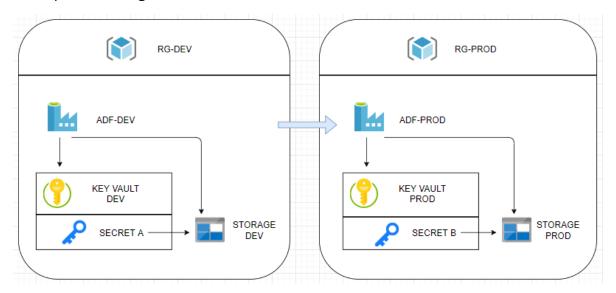








A modo de ejemplo, este es otro caso que puede suceder, donde cambien cuestiones de arquitectura según el ambiente



#### Donde:

- Los grupos de recursos son diferentes
  - o Que se gestiona en cual se realizara en la operación de Deploy
- No solo el nombre del ADF es diferente, sino también el de los recursos con los que se conecta, en este ejemplo,
  - o el Key Vault y el nombre de su secreto
  - o el storage donde se guarda información.

Este problema de arquitectura se estudia analizando los servicios que corresponde a cada ADF, y reemplazándolos via parametrización.



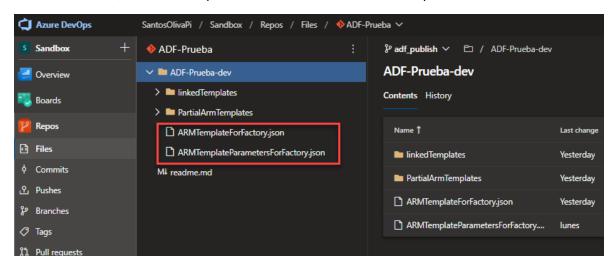




## Parametrización de ARM Templates

En nuestro proyecto de DevOps, Repos-Files, y posicionándonos dentro de la rama adf\_publish encontraremos los archivos:

- ARMTemplateForFactory.json
  - o Que tiene toda la información del Data Factory
- ARMTemplateParametersForFactory.json
  - o Que contiene parámetros utilizados en el template



Utilizaremos estos archivos para cubrir las diferencias entre ambos ADF y que se pueda realizar el paso de Dev a Prod de manera correcta.







Haciendo clic en cada uno de los archivos, tendremos acceso a su contenido, y también la posibilidad de editarlos

Acá se puede observar como se relacionan ambos archivos

Así es como se consumen dichos parametros

```
🞖 adf_publish ∨ 🗀 / ADF-Prueba-dev / ARMTemplateForFactory.json
ARMTemplateForFactory.json
                                                                                                                          Edit
Contents History Compare Blame
                "kv_sandpox_properties_typerroperties_paseuri": {
                   "type": "string",
"defaultValue": "https://kv-sandbox1.vault.azure.net/"
                "st_sandbox_properties_typeProperties_url": {
                   "type": "string",
"defaultValue": "https://stestudio.dfs.core.windows.net"
               },
"st_sandbox_properties_typeProperties_tenant": {
                   "type": "string",
"defaultValue": "cald08b0-9543-4bd9-8718-42becddc7786"
                "st_sandbox_properties_typeProperties_servicePrincipalId": {
                   "type": "string",
"defaultValue": "226958b0-7eba-49a4-9d49-2387b482d03a"
          "factoryId": "[concat('Microsoft.DataFactory/factories/', parameters('factoryName'))]"
          },
"resources": [
                   "name": "[concat(parameters('factoryName'), '/pipeline1')]",
"type": "Microsoft.DataFactory/factories/pipelines",
"apiVersion": "2018-06-01",
```











Teniendo en cuenta esta información, parametrizaremos los datos que se modificarán cuando se pase del entorno de desarrollo al de producción (tema trabajado en Releases).

En nuestro caso de arquitectura alternativo, observando los parámetros presentes en ARMTemplateParametersForFactory.json nos faltaría parametrizar el "secret name".

Para utilizar parámetros personalizados de manera correcta, se deberá trabajar desde la rama que corresponda, entrando a "Manage" y luego a "Parametrization template". Editando este .json, se establecerán parámetros que serán tenidos en cuenta al momento de realizar el Deploy.







### Plantilla ARMTemplateParametersForFactory

#### Sintaxis de los parámetros parametrizados

Trabajaremos la plantilla de la siguiente manera:

Se utilizará la nomenclatura <acción>:<nombre>:<tipoDato>

- Siendo las acciones
  - o "=" si se desea mantener el valor actual como valor por defecto
  - o "-" si no se desea mantener el valor actual como valor por defecto
  - o "|" como un caso especial para Connection strings o keys.
- En <nombre> se colocará el nombre que se desea utilizar como nombre del parámetro, por defecto tomara el nombre de la propiedad.
- <tipoDato> es el tipo del parametro:
  - o Por defecto será "string", si no se completa
  - Puede tomar los valores "string", "bool", "number", "object" y "secureString"

Es importante tener en cuenta que una definición no puede ser específica de una instancia de recurso. Cualquier definición se aplica a todos los recursos de ese tipo.









Dentro de cada parte del factory se especificará las propiedades que se quieran parametrizar, utilizando \* para indicar que se quieren parametrizar todas las propiedades internas en un primer nivel

Por ejemplo, aquí se quieren parametrizar las propiedades "waitTimeInSeconds" y "headers" de los pipelines, ubicadas en properties-Activities-typeProperties

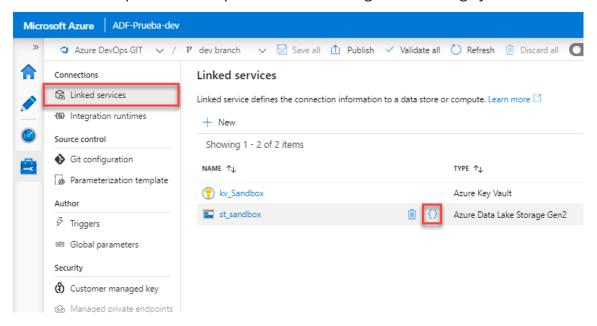
Mientras que aquí, trabajariamos todas las propiedades dentro de "parameters"







Considerando esta información, trabajaremos nuestro ADF, primero buscaremos la ubicación de lo que deseamos parametrizar, investigando su código json,



Aquí buscaremos la ubicación de lo que buscamos parametrizar, recordando el path del mismo

```
2
         "name": "st_sandbox",
         "type": "Microsoft.DataFactory/factories/linkedservices",
 3
         "properties" {
 4
 5
            annotations": [],
            "type": "AzureBlobFS",
 6
 7
           "typeProperties" {
 8
               'url": "https://stestudio.dfs.core.windows.net",
 9
              "tenant": "ca1d08b0-9543-4bd9-8718-42becddc7786",
              "servicePrincipalId": "226958b0-7eba-49a4-9d49-2387b482d03a",
10
               'servicePrincipalKey": {
11
12
                 "type": "AzureKeyVaultSecret",
13
                "store": {
                   "referenceName": "kv Sandbox",
14
                   "type": "LinkedServiceReference"
15
16
                 "secretName": "SP-sandbox-soliva"
17
18
19
20
            "connectVia": {
              "referenceName": "integrationRuntime2",
21
22
              "type": "IntegrationRuntimeReference"
23
24
25
```



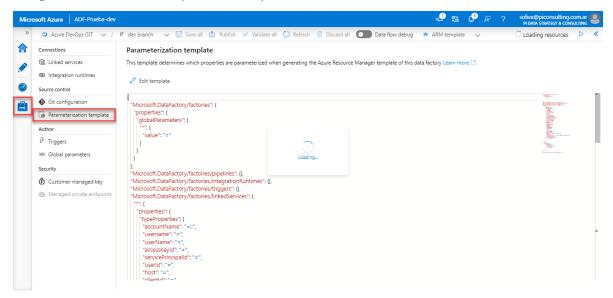








#### Luego modificaremos la plantilla de parametrización desde nuestro ADF



```
"Microsoft.DataFactory/factories/linkedServices": {
 <sup>™⊗™</sup>: {
  "properties": {
    "typeProperties": {
     "servicePrincipalKey": {
         "secretName": "="},
     "accountName": "=::".
     "username": "=",
     "userName": "=",
     "accessKeyId": "=",
     "servicePrincipalId": "=",
     "userId": "=",
     "host": "=".
     "clientId": "=",
     "clusterUserName": "=",
     "clusterSshUserName": "=",
     "hostSubscriptionId": "=",
     "clusterResourceGroup": "=",
```

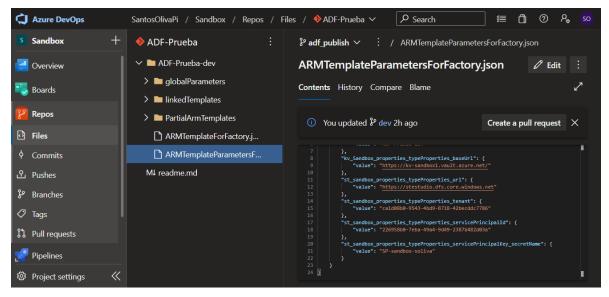
Con esto modificado, luego de publicar veremos los cambios desde DevOps









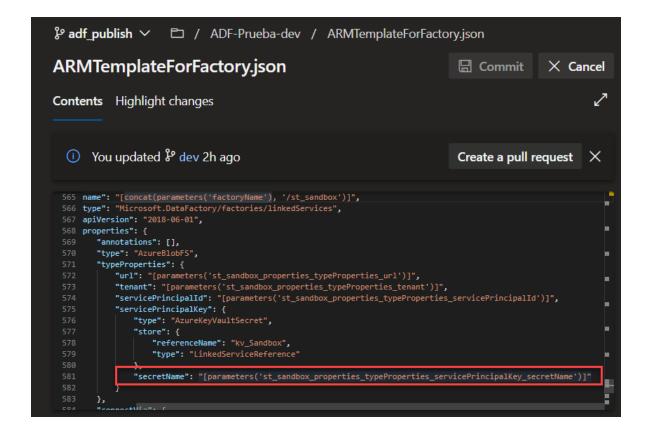


















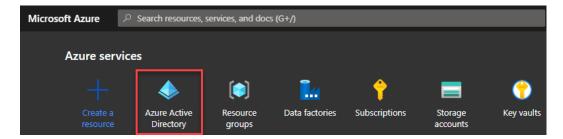


# Service Principal

Un Service Principal [SP] es una identidad creada para su uso con aplicaciones, servicios hospedados y herramientas automatizadas que acceden a los recursos de Azure.

Este acceso está restringido por los roles asignados a la entidad de servicio, lo que permite controlar a qué recursos pueden tener acceso y en qué nivel. Por motivos de seguridad, se recomienda usar siempre entidades de servicio con las herramientas automatizadas, en lugar de permitirles iniciar sesión con una identidad de usuario.

#### Creación de un SP

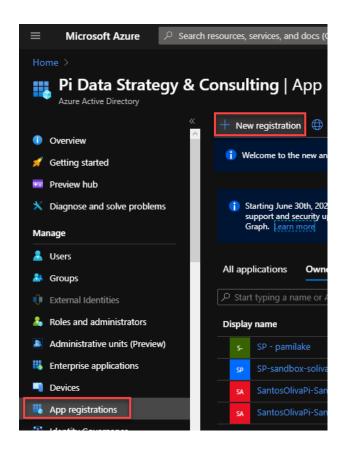








# PI DATA STRATEGY & CONSULTING

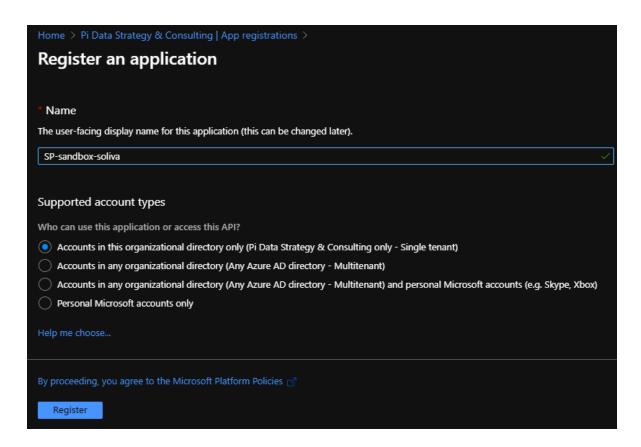
















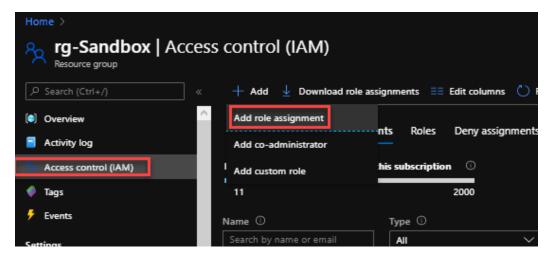


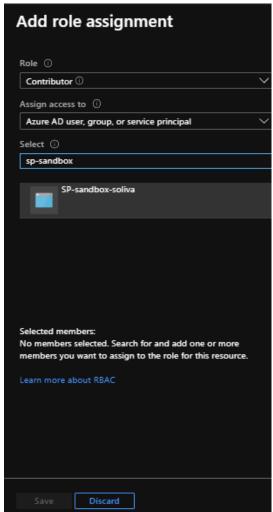




# Acceso al grupo de recursos

Accediendo al grupo de recursos desde el portal de Azure







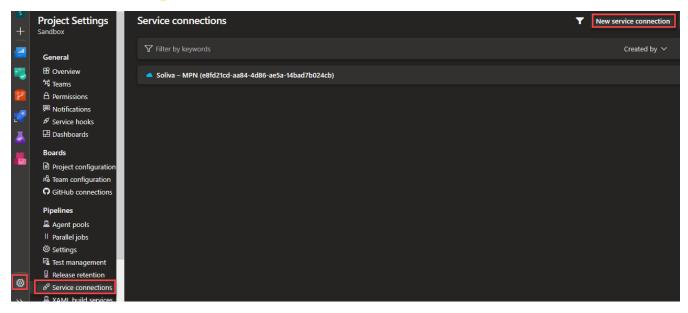


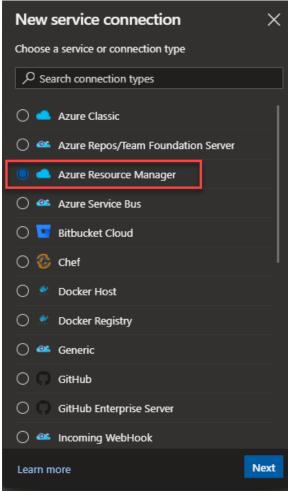






# Acceso a DevOps





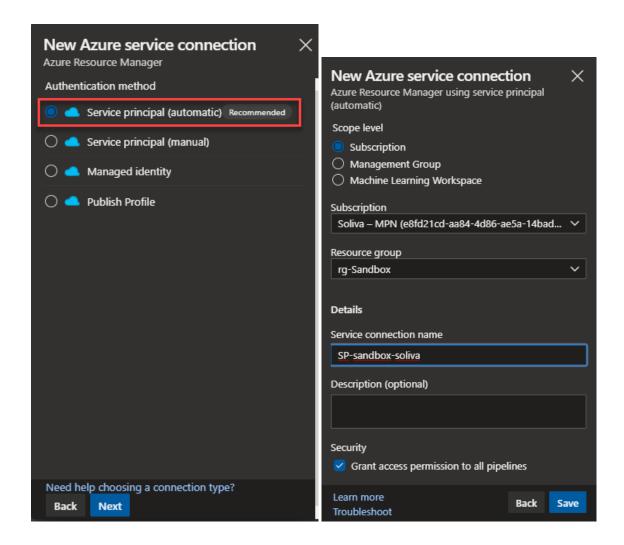








# PI DATA STRATEGY & CONSULTING







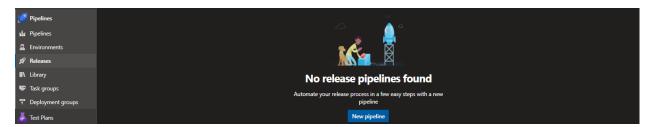




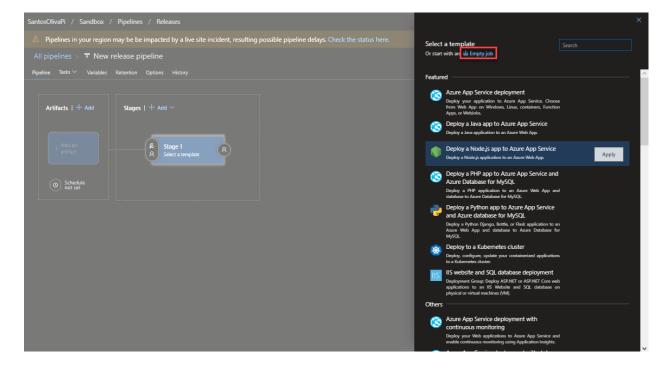
### Releases

Ahora gestionaremos los releases desde DevOps.

Nos posicionamos en Pipelines-Releases y crearemos un nuevo pipeline



#### Y haremos clic en Empty Job



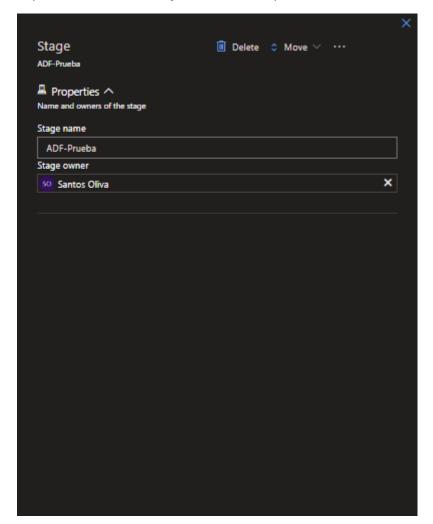








## Aquí lo nombraremos y definiremos quien es el owner













Aquí iremos a Add an artifact y completaremos con los datos correspondientes



Utilizaremos Azure Repos como fuente, seleccionamos nuestro proyecto y respectivo repositorio.

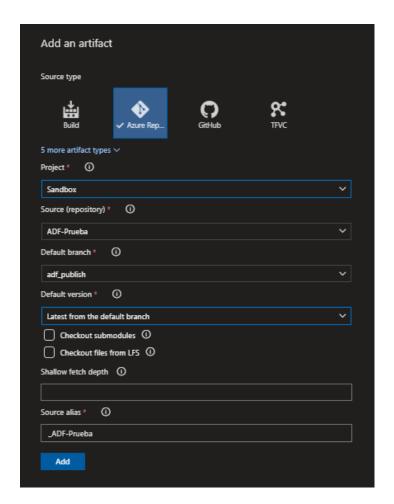
Es importante que la rama por default seleccionada aquí sea adf\_publish, en caso de no encontrarse disponible es porque no se ha realizado ningún Publish en el recurso ADF-prueba-dev.

Finalmente haremos clic en Add.











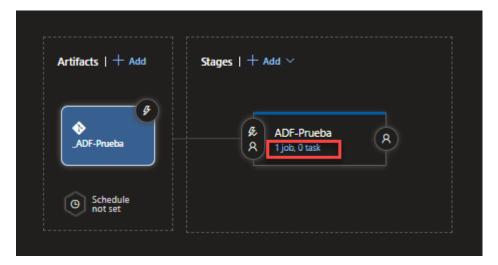




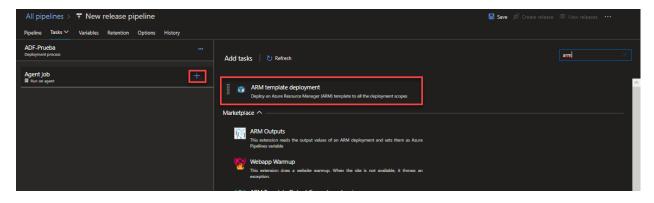




## Ahora haremos clic en 1 job, 0 task



Aquí se abrirá el Agent Job y agregamos la siguiente actividad.









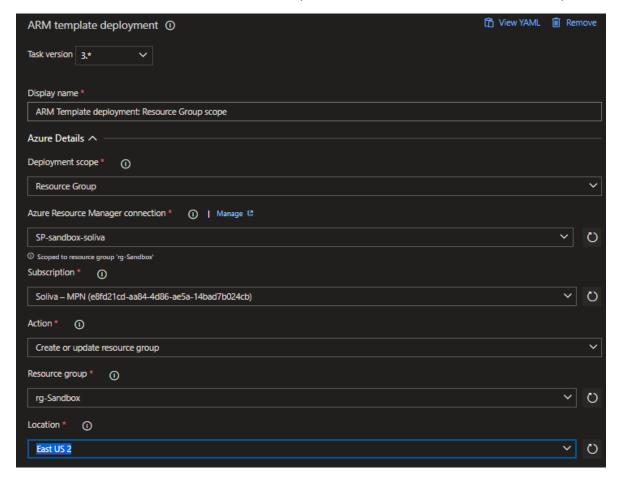


Configuro el deployment de la siguiente manera.

La acción utilizada será la creación o update del grupo de recursos destino, este puede llegar a ser el mismo que el grupo de recursos origen, pero de ser así, el nombre del ADF será distinto.

La localización utilizada deberá ser la misma en la que se alojan los diferentes recursos.

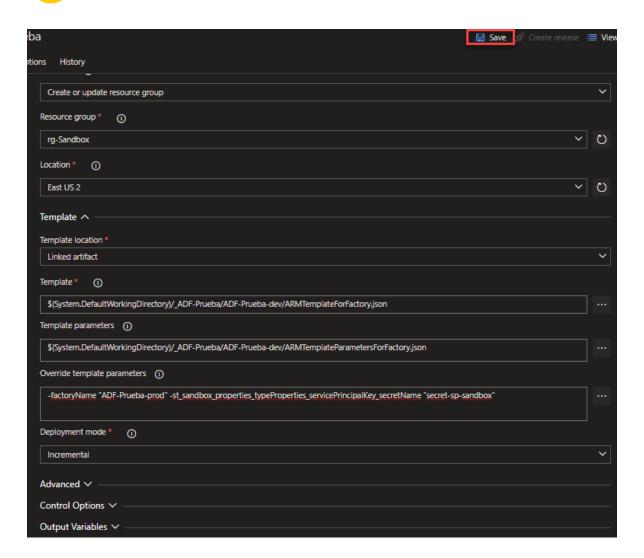
En Override parameters se sobreescriben parámetros que aplican en el recurso fuente, en nuestro caso ADF-Prueba-dev, pero no en el de destino (ADF-Prueba-prod).













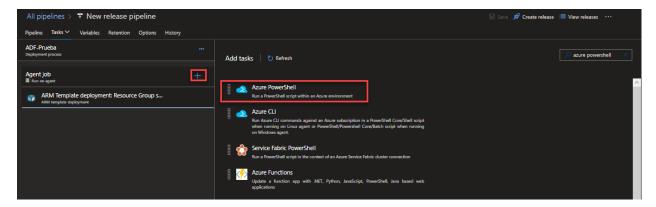








De manera opcional, también se puede gestionar el pausar y reanudar los Triggers de ADF, via powershell



Esto lo realizaremos 2 veces, una para pausar, otra para reanudar.

Los Scripts utilizados serán los siguientes, reemplazando lo resaltado por los nombres que correspondan al entorno trabajado:

En nuestro caso, DataFactoryName será ADF-Prueba-prod y nuestro ResourceGroupName será rg-Sandbox

#### **PAUSAR**

 $\label{thm:convex} $$ triggersADF = Get-AzDataFactoryV2Trigger - DataFactoryName $$ (DataFactoryName) - ResourceGroupName $$ (ResourceGroupName) $$$ 

\$triggersADF | ForEach-Object { Stop-AzDataFactoryV2Trigger -ResourceGroupName
\$(ResourceGroupName) -DataFactoryName \$(DataFactoryName) -Name \$\_.name -Force }

#### **REANUDAR**

\$triggersADF = Get-AzDataFactoryV2Trigger -DataFactoryName \$(DataFactoryName) ResourceGroupName \$(ResourceGroupName)

\$triggersADF | ForEach-Object { Start-AzDataFactoryV2Trigger -ResourceGroupName \$(ResourceGroupName) -DataFactoryName \$(DataFactoryName) -Name \$\_.name -Force }







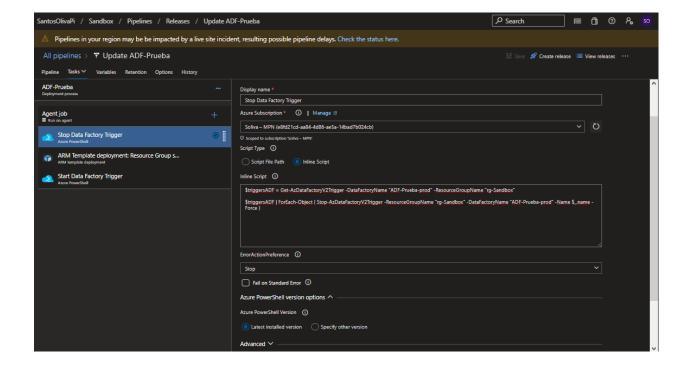
### Como ejemplos terminados...

\$triggersADF = Get-AzDataFactoryV2Trigger -DataFactoryName "ADF-Prueba-prod"-ResourceGroupName "rg-Sandbox"

 $\label{thm:continuous} $$ triggersADF \mid ForEach-Object \{ Stop-AzDataFactoryV2Trigger - ResourceGroupName "rg-Sandbox" - DataFactoryName "ADF-Prueba-prod"-Name $\_.name - Force \} $$$ 

\$triggersADF = Get-AzDataFactoryV2Trigger -DataFactoryName "ADF-Prueba-prod"-ResourceGroupName "rg-Sandbox"

\$triggersADF | ForEach-Object { Start-AzDataFactoryV2Trigger -ResourceGroupName "rg-Sandbox" - DataFactoryName "ADF-Prueba-prod" -Name \$\_.name -Force }





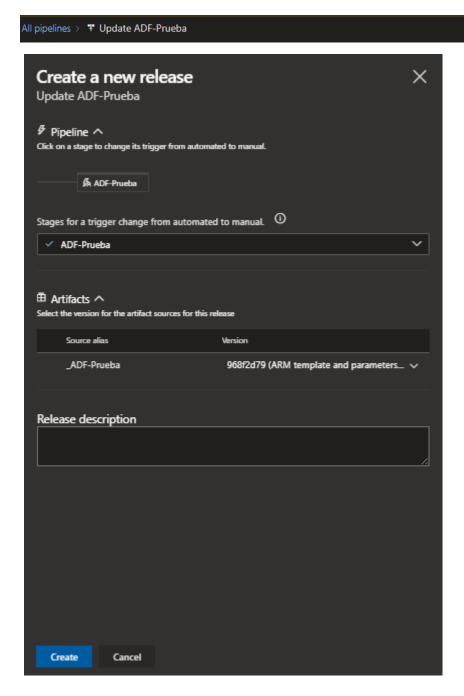








Una vez gestionadas las tareas a realizar por el Agent Job, se hará clic en Create Release y haremos el Deploy



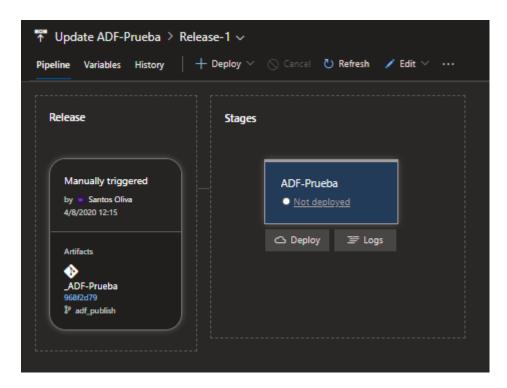








💋 Create release



Con eso realizado, se realizará el primer deploy, en este caso manual y se verificara que las tareas programadas se ejecutan correctamente.







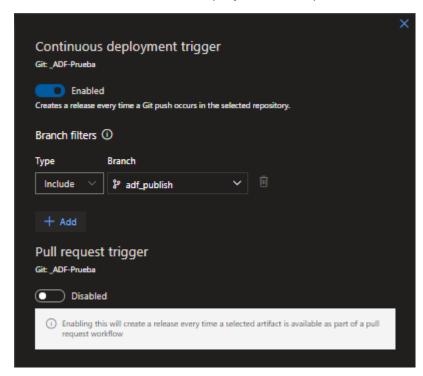


Editando el release automático, haremos clic en el rayo para gestionar el deploy continuo.



Aquí seleccionaremos las opciones deseadas,

- CD trigger
  - para que se realice un deploy cada vez que sucede un push en Git (cuando confirman el pull request)
- Pull request trigger
  - o Se realiza el deploy cada vez que se realiza un pull request.













En nuestro caso habilitaremos el CD trigger, y finalmente haremos clic en Save.











# Para mas información

https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment





