## Microsoft Azure Data Factory

|  |  |
| --- | --- |
| Revision | 01 |
| Issued | 14-Apr-2020 |
| Functional team | Cyber Security Data & Analytics |
| Document classification | GE Internal |

### **Contents**

1. Opening Data Factory
2. Trigger Runs
3. Stream Analytics (Umbrella)
4. Stream Inputs
5. Opal reference
6. Query
7. Output
8. Downloaded

# **Microsoft Azure Data**

* Home --> gw-dfactory-poc
* Monitoring indicating Pipeline runs - Activity runs - Trigger Runs Integration Runtime CPU - Integration Runtime Memory. Pulling Data from a random Cisco S3 Bucket

## **1.1 Data Factory**

* Open Data Factory

Machine generated alternative text:
Mkrosoft Azure 
Home Ed.ctory•poc 
Access 
md problems 
Started 
settmgs 
Supp«t • troubW.oote.g 
suwo't 
ActMryRuns 
Memory 

Machine generated alternative text:
Azure Data Factory 
Let's get started 
00000 
Create pipeline Create data Create pipeline Copy data Configure SSIS 
from template 
Integration 
Videos 
Overview Of Azure Data Factory 
Prepare data using Data 
• Arun • 
Execute Azure Machine Learning 
service pipelines 

Already integrated Data Factory with GitHub repo

* When setting up with GitHub you must set up with an API Key
* 3 different loads
* Bulk, DNS, Proxy
* Bulk: Copy entire bucket
* Both Deltas: Copies everything changed since the

last run

## **. Resources**

**Machine generated alternative text:
C) R.tr.sh 
Factory Resources 
p by **

## **1.3. Value**

* Within the parameters from the windows Start and end anything changed will be copied over from the Deltas.

**Machine generated alternative text:
Variables **

## **File Path**

* Its pulling data from the S3 bulk, connected from the umbrella\_cisco\_Manages S3.

Machine generated alternative text:
Parameters 
Add • PI 
type 

# **Trigger Runs**

* Every 10 min a file is being loaded to that S3.

Machine generated alternative text:
Microsoft Azure I 
OathbO*ds 
Trigger 
O & metrics 
Trigger runs 
: Int 24 hours AM • AMI 
St*uS v 
• 96 ite•ns 
trg• Id 
Tine zone : Eastan Tin* (US S C.vada) 
an,'2a AM 
AM 
AM 
Trine' runs : Latest runs 
Succeec*d 
Succee&d 
Succee&d 
Succee&d 
o 
122332CW3 
08*' swans2sn»a2861220txwa 
oesu 

## **1.1. Alerts and Metrics**

• Alerts and metrics

Machine generated alternative text:
Microsoft Azure I Data Factory qw.dfaaory•po•;: 
Dashbouds 
Trigger nns 
O Alerts & metrics 
Alerts & metrics 
C.) Refresh iiii Metrics 

## **1.2. Edit or Trigger**

○ Trigger is what you use to kick off a job

○ Trigger now - on demand

○ New/Edit - what you use to kick off the job

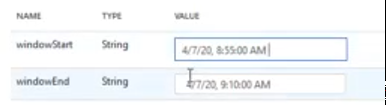
Machine generated alternative text:
trine' Machine generated alternative text:
Edit trigger 
Date 
• No Ena S' On Date 

## **1.3. Trigger Now**

* + If trigger now - add value

Machine generated alternative text:
Pipeline run 
Pararneters 

\*Example from properties to add in value textbox\*

Machine generated alternative text:
o 
4/7/20 
essm- 
nom— 
Add 
Add 

## **Pipeline**

* Pipeline name pops up within pipeline Run

Machine generated alternative text:
Pipeline runs 
: Int 24 hours AM • AM) 
Rerun R.•tr.•• 
• 97 
4/7/20 AM 
4/7/20 
Tine zone : Eastan Tin* (US S C.vad.) u. 

* Result on dashboard - Showing success or fail

Machine generated alternative text:
Dashboards 
: Int 24 hours 929 AM • an,'20929 AM) 
Pipeline 
100% 
Trigger 
Q rme zone : Eastan Tin* (US S 
Activity 
100% 

# **Stream Analytics Umbrella**

* On Microsoft Azure

Home > Storage Accounts > rawdata01 | Containers > feeds

Machine generated alternative text:
Microoft Azure 
Add C*edry t) Refresh e 
Access Oy Switch to Aa•e AD user ACCI 
feeds / 
C) prÄ09S 

## **1.1. DNS logs**

* + To look at what we just copied over look at cisco umbrella, click DNS logs

Every day there is a different folder. Within them have files.

* + Reason why there are csv.gz and text files is because the bulk load is doing a binary copy incremental loads are coping and reading in the email on to it.

Machine generated alternative text:
C] 2020-03-os 
2020-0306 
202003-07 
O 2020•03•« 
C) 202003-Ø Machine generated alternative text:
Lrpbad Add Refresh 
2020-03-03-22-S0-3Ma.bt 
2020-03-%-22-SO-Y7dcsvgz 
i? 202003-05-22•S0-mata 
{$ 202003-OS-23-X•4' 

## **1.2. Time**

* + Once data has landed into rawblob we have a bunch a crowd strike data.
  + Some static data given by the DAN team dates to JULY 2019
  + Structured by YEAR/Month/Day/Hour

Machine generated alternative text:
Add t) Refresh e 
Access Oy V."tch to Azure AD User kCOunt 
f 2020 / 03 / 18 
r_hsnvF •oer 
3/23/2020. e 3:25 PM 

# **Stream Inputs**

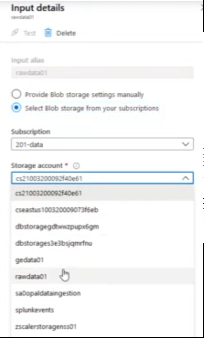
* Two types of inputs

1. Feeds container within rawdata01
2. SQL Database - opalrefdata

Machine generated alternative text:
Mkrown Azure 
I Inputs 
Access 
+ Add reference 
Source type 
Oobase 

○ If you select blob storage from your subscription and pick the right one.

○ You can change data, path and time format.

 Machine generated alternative text:
Input details 
O Cre*e 
Path patt«n O 
Oat. 
O 
Eroding C) 
Event type C) 
x 

## **File type**

All the data from cribble is "json.gz"

Machine generated alternative text:
Add 
csN•31DmZLO.jsmgz 
.json.gz 
Bong Machine generated alternative text:
O 
Ergodinq O 
O 
Event O 
The seeded reou•ce the 
Vou be move 

# **Opal reference data**

* A query to a database.
* Input alias is whatever name you would like to enter.

Machine generated alternative text:
Horne x, I 
Input details 
test Delete 
tus 
(S) SOL Oobase settings 
O Sek•a SOL from 
Subscription 
Srwsh« 
SELECT 
tsys_ld) 
. ted_t 

## **Storage**

* Storage account - creating a screenshot on a storage account
* You can choose either to provide settings manually or select SQL database from

Your subscriptions

Machine generated alternative text:
Input details 
Test Delete 
@ Prcne:ie SOL O•abase settings 
'tom 

○ Created a read only user and login on this server

○ Now we can only expect it to be changed every 6 hours

Machine generated alternative text:
Su iptior, 
Sever 
op.*erv. r 
Refresh O 

* And save
* Once we have opalrefdata and rawdata now treated as table objects that you can now query, join and do tsql functions on to.

# **Query**

* Under topology click query

Machine generated alternative text:
Mkrowft &ure 
- I Query 
I Query 
Access 
Cods 
topobgy 
Functons 
v Outputs (1) 
Bd001•cs 

* Collects from rawdata01
* Left outer join meaning that they keep everything to the left which is the rawdata01

Machine generated alternative text:
34 
WTER 

## **1.1. Computer name**

* Computer name is a rare field in crowdstrike,
* It is used to join to a host and to a user
* Computer name is equivalent in opal to the reference name of your computer
* Every computer GE has a name listed in opal

Machine generated alternative text:
G4RCOYY2E 
Filter Results 
Ag.t Nam. 
Home I 
Search for Assets 
Type 
External 
1 lassaa93 

* + In the actual select field, it is all the fields that the tiger team said would be useful
  + Pulling from rawdata and for reference data

Machine generated alternative text:
Save query X 
1 SELECT 
[raul. 
6 
8 
. ( cc—and). 
_ (event_pl 
traul. . 
(raul. 
(r" I. (R'SHashOatal. 
. (p.r•entAuthent1ationIdJ 

* + Looking through other business units its coming from

• Pulling from rawdata01 and for reference data

Machine generated alternative text:
39 
. conf ig_l te._10, 
' corp 
THEN • ey• 

* + If company is null: no information = unclassified (uc)

Ends up becoming BU field

* + Last modified time - date ingested
  + Name of the file

Machine generated alternative text:
e THEN 
ELSE | 
as (ingest_dttnl, 
as tingest_filel 

# **Output**

* + Setup the same way for input stream data
  + Once running output, cannot change the output details

Machine generated alternative text:
Horne I 
ge-stream•proc-crowdstrike I Outputs 
o 
to the 
Eb stor.* 
Access 
Md problems 
'oas 
Functions 

Machine generated alternative text:
Output details 
test Ltee'e 
• Vou cm 
o 
to 
Date 

## **Done**

* + Completed and kicked off the job

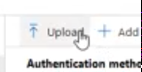
Machine generated alternative text:
p 
Mkro«.ft Azure 
ge-stream•proc-crowdstrike 
o 
Resource : 
Access 
Md problems 
Send 
Created 
Hog•nq 
: End", 27. 2020.2.23:35 PM 
: Tuesday. 7, 2020 
query 
2 
op*efd". 
1 
SELECT 
SQL 
topobgy 
Funcoons 
O St.-or... 
12 
. (Authenticationldl . 
. ( cc—and I , 
. sieleNWl. 
. (file_shal), 
. lid), 

# **Downloaded**

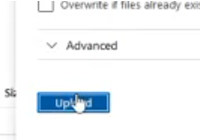
We downloaded the file earlier, so we upload it to test directory where its listening for new files

Machine generated alternative text:
'eees 
Add Refresh 
Access Oy (Sw•tch to AD user Account' 
'reds test 
Prepefties 
Meta"a 

* + Click upload



* + Select file

Machine generated alternative text:
Upload blob Machine generated alternative text:
GZ 
02.2m  

* Able to see monitoring

Machine generated alternative text:
Reguce 

* + To see real time go to job diagram
  + It is a view of what the job does
  + Down below has a live monitor of data being pulled in

Machine generated alternative text:
Hc«ne 
Job diagram 
Reset vie User 
o 