

01

Overview of HW 3



HW 3 – Transforming & Staging



Homework 3

Transforming & Staging



01

SQL Database +
Data Factory



02

Transform

Change 'DateA'
to Date



03

Staging

Copy Data to
SQL DB Table

HW3 Final Output

Most Recently Updated Rows/Observations from the SQL Table Data

Microsoft Azure

Search resources, services, and docs (G+)

SQL database

Query editor (preview)

Query 1

```
1 SELECT *
2 FROM dbo.CMPL
3 WHERE CONVERT(Date, DATEA) = CONVERT(Date, GETDATE() - 3)
```

Results

CMPLID	ODINO	MFR_NAME	MAKETXT	MODELTX	YEARTXT	CRASH	FAILDATE	FIRE	INJURED	DEATHS	COMPDESC	CITY	STATE	VIN	DATEA	L
1930425	115470...	Chrysler (FC...	JEEP	WRANGLER	2020	N	2023-03-...	N	0	0	POWER TR...	Blacklick	OH	1C4H...	2023-09-28T00:00:00.0000000	2
1930426	115470...	Kia America...	KIA	SORENTO	2014	N	2021-06-...	N	0	0	ENGINE	Jackson	GA	5XYK...	2023-09-28T00:00:00.0000000	2
1930427	115470...	Chrysler (FC...	CHRYSLER	PACIFICA	2023	N	2023-09-...	N	0	0	ELECTRICAL...	Scotts...	AZ	2C4R...	2023-09-28T00:00:00.0000000	2
1930428	115470...	Chrysler (FC...	CHRYSLER	PACIFICA	2023	N	2023-09-...	N	0	0	FUEL/PROP...	Scotts...	AZ	2C4R...	2023-09-28T00:00:00.0000000	2
1930429	115470...	Chrysler (FC...	RAM	2500	2017	N	2023-07-...	N	0	0	SERVICE BR...	El Dor...	CA	3C6U...	2023-09-28T00:00:00.0000000	2
1930430	115470...	Chrysler (FC...	RAM	2500	2017	N	2023-07-...	N	0	0	ELECTRICAL...	El Dor...	CA	3C6U...	2023-09-28T00:00:00.0000000	2
1930431	115470...	General Mot...	CHEVROL...	BOLT EUV	2023	N	2023-09-...	N	0	0	FORWARD ...	Los An...	CA	1G1F...	2023-09-28T00:00:00.0000000	2
1930432	115470...	Chrysler (FC...	CHRYSLER	PACIFICA	2017	N	2023-09-...	N	0	0	ENGINE	Heber ...	UT	2C4R...	2023-09-28T00:00:00.0000000	2
1930433	115470...	General Mot...	CADILLAC	LYRIQ	2024	N	2023-09-...	N	0	0	EXTERIOR I...	Wheat...	CO	1G1F...	2023-09-28T00:00:00.0000000	2

Query succeeded | 292s



Homework 3

Transforming & Staging



01

SQL Database



02

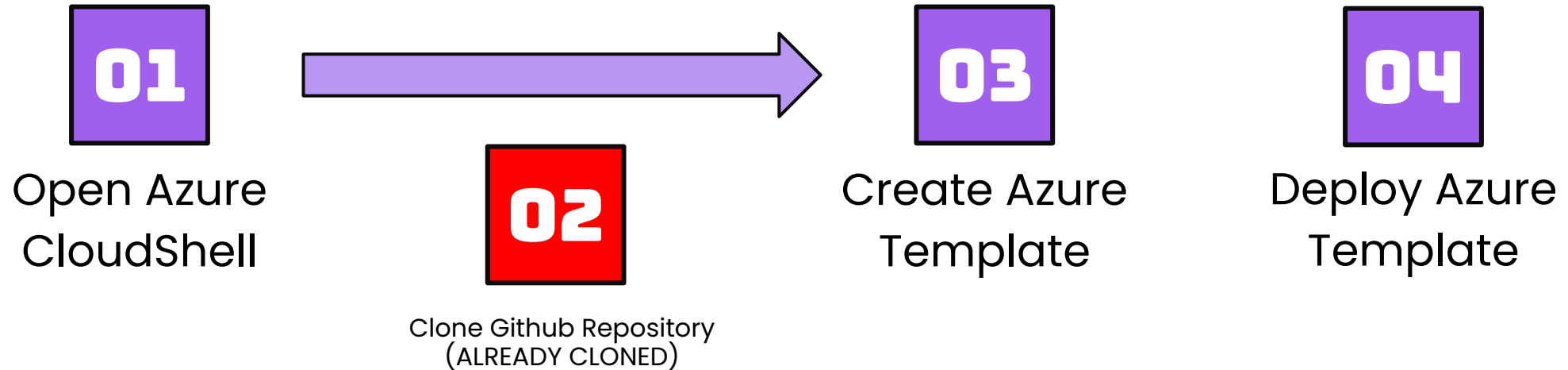
Data Factory



03

SQL Database

0. Deploy hw3 template



*Hint

```
PS /home/dk98> cd './ OMDMod4 /Homework 3/Transforming and Staging/'
PS /home/dk98/ OMDMod4 /Homework 3/Transforming and Staging> bash ./formTemplate.sh
Template and parameters created successfully: ./template/template.json and ./template/parameters.json
```

Deployment Result

<input type="checkbox"/>	 db47fdf4- <input type="text"/>	SQL server	East US
<input type="checkbox"/>	 dba09cc5- <input type="text"/>	SQL database	East US




Azure SQL Database



Azure SQL Server

1. Go to Azure SQL Database

<input type="checkbox"/>	 db47fdf4-		SQL server	East US
<input type="checkbox"/>	 dba09cc5-		SQL database	East US



Azure SQL Database



Azure SQL Server

1.1 Access Query Editor in SQL DB

Home > David >



dba09cc5-
SQL database

Query editor (preview) ☆ ...



Login



New Query



Open query



Feedback



Getting started

Overview

Activity log

Tags

Diagnose and solve problems

Query editor (preview)

Settings

Compute + storage

Connection strings

Properties

Locks

Data management

Replicas

Sync to other databases

Query editor (preview) is a tool to run SQL queries against Azure SQL Database in the Azure portal. It is designed for lightweight querying and object exploration in your database. For more information and troubleshooting, [Learn more](#)



Welcome to SQL Database Query Editor

SQL server authentication

Login *

Password *

OK

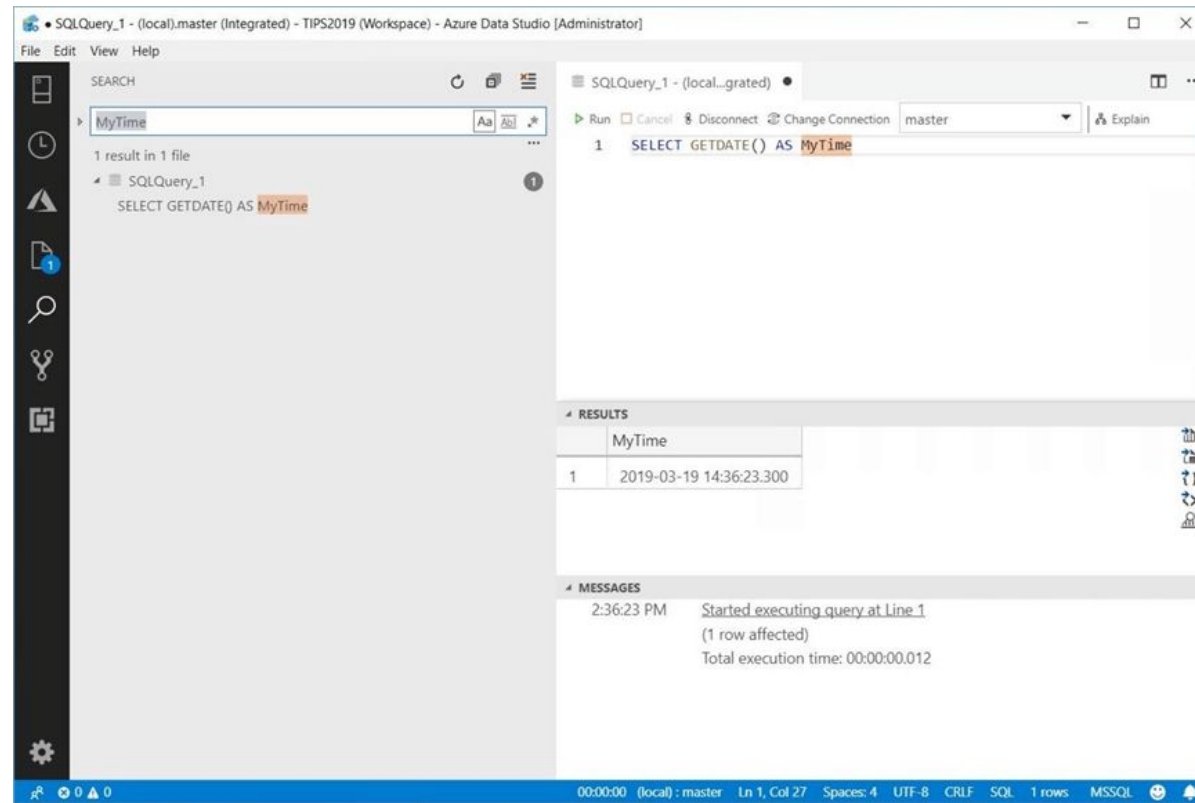
Microsoft Entra authentication

[Continue as dk98@bu.edu](#)

OR

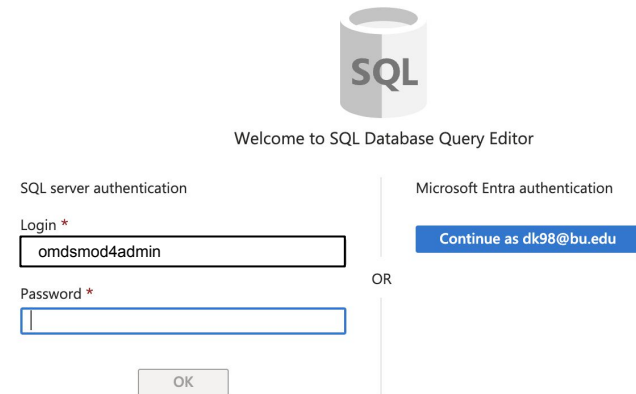
1.1 Or Azure Data Studio

[Link](#)



1.2 Authentication

1. Deployed our template:
 - a. Login: `omdsmod4admin`
 - b. Password: `omdsmod4password013!`
2. Created SQL Database Manually:
 - a. Own Credentials



SQL

Welcome to SQL Database Query Editor

SQL server authentication

Login *

omdsmod4admin

Password *

Microsoft Entra authentication

Continue as dk98@bu.edu

OR

OK

*Firewall Error

If Prompted,
Give Access to your own IP Address

1.3 Create Table in SQL Database

Look at [Reference](#) File to create the following query

```
CREATE TABLE Complaints
(
    CMPLID int Primary,
    *Hint: NVARCHAR(max),
    ...
    DATEA date,
)
```

Query 1 X

▶ Run ☐ Cancel query ⬇ Save query

```
1 CREATE TABLE complaints
2 (
3     CMPLID int primary key,
```

Appendix A. Complaints File Characteristics

Field#	Name	Type/Size	Description
1	CMPLID	CHAR(9)	NHTSA'S INTERNAL UNIQUE SEQUENCE NUMBER. IS AN UPDATEABLE FIELD, THUS DATA FOR A GIVEN RECORD POTENTIALLY COULD CHANGE FROM ONE DATA OUTPUT FILE TO THE NEXT.
2	ODINO	CHAR(9)	NHTSA'S INTERNAL REFERENCE NUMBER. THIS NUMBER MAY BE REPEATED FOR MULTIPLE COMPONENTS.
3	MFR_NAME	CHAR(40)	MANUFACTURER'S NAME
4	MAKE	CHAR(25)	VEHICLE/EQUIPMENT MAKE
5	MODEL	CHAR(256)	VEHICLE/EQUIPMENT MODEL
6	YEAR	CHAR(4)	MODEL YEAR, 9999 IF UNKNOWN or N/A
7	CRASH	CHAR(1)	WAS VEHICLE INVOLVED IN A CRASH, 'Y' OR 'N'
8	DATE	CHAR(8)	DATE OF INCIDENT (YYYYMMDD)
9	FIRE	CHAR(1)	WAS VEHICLE INVOLVED IN A FIRE 'Y' OR 'N'
10	INJURED	NUMBER(2)	NUMBER OF PERSONS INJURED
11	DEATHS	NUMBER(2)	NUMBER OF FATALITIES
12	COMPDESC	CHAR(128)	SPECIFIC COMPONENT'S DESCRIPTION
13	CITY	CHAR(30)	CONSUMER'S CITY
14	STATE	CHAR(2)	CONSUMER'S STATE CODE
15	VIN	CHAR(17)	VEHICLE'S VIN#
16	DATEA	CHAR(8)	DATE ADDED TO FILE (YYYYMMDD)
17	LDATE	CHAR(8)	DATE COMPLAINT RECEIVED BY NHTSA (YYYYMMDD)
18	MILES	NUMBER(7)	VEHICLE MILEAGE AT FAILURE
19	OCCURENCES	NUMBER(4)	NUMBER OF OCCURENCES
20	CDESCR	CHAR(2048)	DESCRIPTION OF THE COMPLAINT
21	CMPL_TYPE	CHAR(4)	SOURCE OF COMPLAINT CODE: ...

CREATE TABLE complaints (
CMPLID int primary key,

ODINO nvarchar(max),
MFR_NAME nvarchar(max),
MAKETXT nvarchar(max),
MODELTX nvarchar(max),
YEARTXT nvarchar(max),
CRASH nvarchar(max),
FAILDATE date,
FIRE nvarchar(max),
INJURED int,
DEATHS int,
COMPDESC nvarchar(max),
CITY nvarchar(max),
STATE nvarchar(max),
VIN nvarchar(max),
DATEA date,
LDATE date,
MILES int,
OCCURENCES int,
CDESCR nvarchar(max),

CMPL_TYPE nvarchar(max),
POLICE_RPT_YN nvarchar(max),
PURCH_DT date,
ORIG_OWNER_YN nvarchar(max),
ANTI_BRAKES_YN nvarchar(max),
CRUISE_CONT_YN nvarchar(max),
NUM_CYLS int,
DRIVE_TRAIN nvarchar(max),
FUEL_SYS nvarchar(max),
FUEL_TYPE nvarchar(max),
TRANS_TYPE nvarchar(max),
VEH_SPEED int,
DOT nvarchar(max),
TIRE_SIZE nvarchar(max),
LOC_OF_TIRE nvarchar(max),
TIRE_FAIL_TYPE nvarchar(max),
ORIG_EQUIP_YN nvarchar(max),
MANUF_DT date,
SEAT_TYPE nvarchar(max),
RESTRAINT_TYPE nvarchar(max),
DEALER_NAME nvarchar(max),

DEALER_TEL nvarchar(max),
DEALER_CITY nvarchar(max),
DEALER_STATE nvarchar(max),
DEALER_ZIP nvarchar(max),
PROD_TYPE nvarchar(max),
REPAIRED_YN nvarchar(max),
MEDICAL_ATTEN nvarchar(max),
VEHICLES_TOWED_YN nvarchar(max)

)

The decision to utilize nvarchar(max) instead of adhering to the reference file is due to the presence of human error in the data.

1.4 Check your 'complaints' Table

The screenshot shows the Azure Data Studio interface with the 'complaints' table schema defined in the query editor. The table has the following columns:

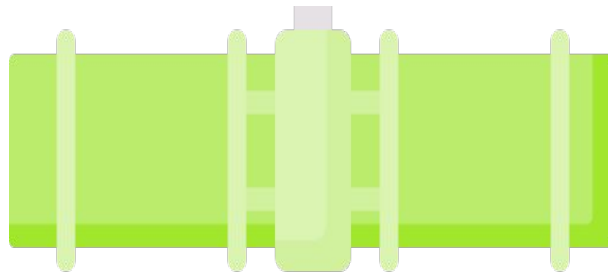
- CMPLID (PK, int, not null)
- ODINO (nvarchar, null)
- MFR_NAME (nvarchar, null)
- MAKETXT (nvarchar, null)
- MODELTX (nvarchar, null)
- YEARTXT (nvarchar, null)
- CRASH (nvarchar, null)
- FAILDATE (date, null)
- FIRE (nvarchar, null)
- INJURED (int, null)
- DEATHS (int, null)
- COMPDESC (nvarchar, null)
- CITY (nvarchar, null)
- STATE (nvarchar, null)
- VIN (nvarchar, null)
- DATEA (date, null)
- LDATE (date, null)
- MILES (int, null)
- OCCURENCES (int, null)
- CDESCR (nvarchar, null)

Tables	
dbo.BuildVersion	...
dbo.complaints	...
CMPLID (PK, int, not null)	
ODINO (nvarchar, null)	
MFR_NAME (nvarchar, null)	
MAKETXT (nvarchar, null)	
MODELTX (nvarchar, null)	
YEARTXT (nvarchar, null)	
CRASH (nvarchar, null)	
FAILDATE (date, null)	
FIRE (nvarchar, null)	
INJURED (int, null)	
DEATHS (int, null)	
COMPDESC (nvarchar, null)	
CITY (nvarchar, null)	
STATE (nvarchar, null)	

2. Data Factory

01

COPY DATA 1



Transform + Stage

Copy Data 1 – Source



Azure Blob Storage



(Output from HW 2)

complaints.txt

1.2GB



DelimitedText

Source

Azure Blob Storage

Delimited Text

complaints.txt

*Import Schema – 49 Columns

Delimited Text? Delimiter?

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

☒ Delimited - Characters such as commas or tabs separate each field.

☐ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: 437 : OEM United States

☐ My data has headers.

Preview of file C:\Complaints\FLAT_CMPL\FLAT_CMPL.txt

```
1 1958156TOYOTA MOTOR CORPORATIONTOYOTAPASEO1994Y19941226N00PARKING BRAK
2 2958164TOYOTA MOTOR CORPORATIONTOYOTALAND CRUISER1994Y19941223N00SERVICE
3 3958124TOYOTA MOTOR CORPORATIONTOYOTACOROLLA1994Y19941128N00AIR BAGS:F
4 4958173Ford Motor CompanyLINCOLNTOWN CAR1994Y19941222N00SERVICE BRAKES
5 5958146General Motors LLCGMCSONOMA1995Y19941215N00SERVICE BRAKES, HYDRA
6 6958127Ford Motor CompanyFORDRANGER1994N00ENGINE AND ENGINE COOLING:EX
7 7958170Ford Motor CompanyMERCURYCOUGAR1995Y19950101N00SERVICE BRAKES, H
```

Cancel < Back Next > Finish

Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☒ Tab

☐ Semicolon

☐ Comma

☐ Space

☐ Other:

☐ Treat consecutive delimiters as one

Text qualifier: " " >

Data preview

1	958156	TOYOTA MOTOR CORPORATION	TOYOTA	PASEO	1994	Y	19941226	N
2	958164	TOYOTA MOTOR CORPORATION	TOYOTA	LAND CRUISER	1994		19941223	N
3	958124	TOYOTA MOTOR CORPORATION	TOYOTA	COROLLA	1994	Y	19941128	N
4	958173	Ford Motor Company	LINCOLN	TOWN CAR	1994	Y	19941222	N
5	958146	General Motors LLC	GMC	SONOMA	1995		19941215	N
6	958127	Ford Motor Company	FORD	RANGER	1994			N
7	958170	Ford Motor Company	MERCURY	COUGAR	1995		19950101	N

Cancel < Back Next > Finish

Copy Data 1 – Sink



Azure SQL Database

Complaints Table

*Authentication:
Back to Slide 'Authentication'

Sink

SQL Database

Table dbo.complaints

*Import Schema – 49 Columns

Copy Data Overview

Source

(Output from HW 2)

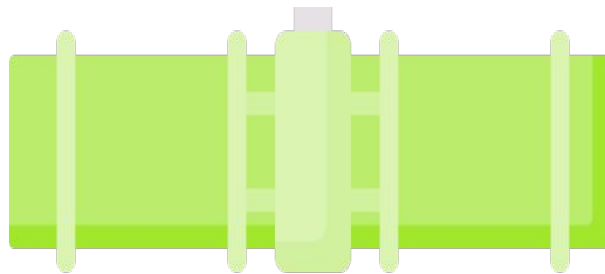
Azure Blob Storage

Delimited Text

.txt

01

COPY DATA 1



Transform + Stage

Sink

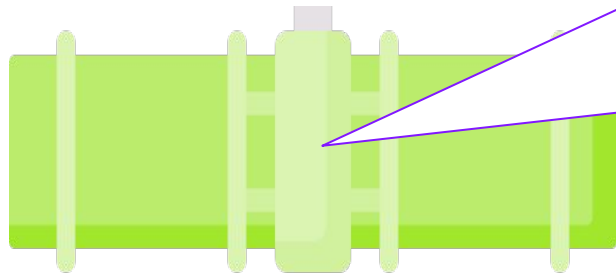
SQL Database

Table dbo.Complaints

Transform

01

COPY DATA 1



Transform

◁Mapping▷

'DateA' Column

From 'String' to 'Date' Type

***Type Conversion Setting:**

- Datetime Format: yyyyMMdd

* From Reference File

16	DATEA	CHAR (8)	DATE ADDED TO FILE (YYYYMMDD)
----	-------	----------	-------------------------------

IMPORTANT!

Create Table

CMPLID INT PRIMARY

- Optional: Add **'NOT NULL'**
- Do not add **IDENTITY(1,1)**

Copy Data 1 'Settings'

*Navigate to **'Fault Tolerance'**

- ☒ **Skip Incompatible Rows**
- ☒ **Logging**
- This is to skip rows affected by **human errors**

Running Pipeline

- Pipeline runs on **Azure Cloud**
 - Turning off your computer will not affect the run
- Approx. **2.0 hours** for completion
- **Monitor** to ensure the data is being read & written properly

*RUNNING A PIPELINE

- **Publish & Add Trigger** >> Cheaper than 'Debug' Action
- Monitor pipeline runs through the '**monitor**' section
 - Go to details (eyeglasses icon) for details on data read & written)

The image shows three screenshots from a software interface. The first is a dropdown menu for 'Add trigger' with options 'Trigger now' and 'New/Edit'. The second is a sidebar menu with 'Pipeline runs' selected. The third is a page titled 'Activity runs' showing a single run with status 'Succeeded'.

Add trigger

- Trigger now
- New/Edit

Dashboard

- Home
- 1 Pipeline runs
- Trigger runs
- Change Data Capture (previ...

Activity runs

Pipeline run ID 21c23e6d-83e4-4684-aa20-a0e5562acef8

All status ▾

Showing 1 - 1 items

Activity name	Activity status
Copy_4p9	Succeeded

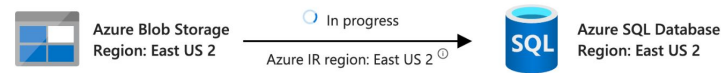
Details

Details Refresh

Performance tuning tips:
Writing to sink is slow, as 11 rows are found incompatible with source format and skipped. To achieve better performance, you are suggested to make sure that source and sink have compatible format. Refer to this [document](#).

[Learn more on copy performance details from here.](#)

Activity run id: 88ce5804-420e-4ea2-87a9-44248f283a87



Data read: ⓘ	1,347 GB	Data written: ⓘ	897,177 MB
Files read: ⓘ	1	Rows written: ⓘ	829,530
Rows read:	1,960,164	Rows skipped: ⓘ	12
Peak connections: ⓘ	10	Peak connections: ⓘ	2

Copy duration 00:47:54
Throughput: ⓘ 470.162 KB/s

▼ Azure Blob Storage → Azure SQL Database

Start time 2024-02-08T19:48:15.0101701Z
Used DIUs ⓘ 4
Used parallel copies ⓘ 1

▼ Duration	00:47:54		
Details		Working duration	Total duration
Queue ⓘ			00:00:07
Transfer ⓘ			00:47:45
	Listing source ⓘ	00:00:00	
	Reading from source ⓘ	00:00:09	
	Writing to sink ⓘ	00:47:34	

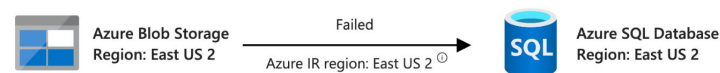
▼ Data consistency verification ⓘ Verified
Verification result 100% of data copied was verified to be consistent

Details Refresh

Performance tuning tips:
Writing to sink is slow, as 29 rows are found incompatible with source format and skipped. To achieve better performance, you are suggested to make sure that source and sink have compatible format. Refer to this [document](#).

[Learn more on copy performance details from here.](#)

Activity run id: 88ce5804-420e-4ea2-87a9-44248f283a87



Data read: ⓘ	1,347 GB	Data written: ⓘ	1,71 GB
Files read: ⓘ	1	Rows written: ⓘ	1,393,441
Rows read:	1,960,164	Rows skipped: ⓘ	30
Peak connections: ⓘ	10	Peak connections: ⓘ	2

Copy duration 01:33:55
Throughput: ⓘ 239.427 KB/s



▼ Azure Blob Storage → Azure SQL Database

Start time 2/8/2024, 2:48:15 PM
Used DIUs ⓘ 4
Used parallel copies ⓘ 1

▼ Duration	01:33:55		
Details		Working duration	Total duration
Queue ⓘ			00:00:07
Transfer ⓘ			01:33:46
	Listing source ⓘ	00:00:00	
	Reading from source ⓘ	00:00:09	
	Writing to sink ⓘ	01:33:42	

▼ Data consistency verification ⓘ Verified
Verification result 100% of data copied was verified to be consistent

3. Go back to Azure SQL Database

<input type="checkbox"/>	 db47fdf4-		SQL server	East US
<input type="checkbox"/>	 dba09cc5-		SQL database	East US



Azure SQL Database



Azure SQL Server

SQL Queries

Final Query:

SELECT *

FROM Complaints

WHERE DATEA = CONVERT(Date, GETDATE()-1)

- ^^ Get Data From Yesterday?
- But... No Data From Yesterday?
- **Change the 1** in GETDATE()-1 to find the most recent date in your data

Find Today's Date (Azure Time Zone)

SELECT GETDATE()

- Azure's Time Zone is not EST

Find Dates in Data:

SELECT DISTINCT CONVERT(VARCHAR, DATEA)

FROM Complaints

ORDER BY CONVERT(VARCHAR, DATEA) DESC

- List the dates in 'DateA' column
- Descending order (Most recent on Top)

HW3 Screenshot

Transforming & Staging

Microsoft Azure

Home > Resource groups > [Resource Group]

dbdcfbdb SQL database

Query editor (preview)

Run Cancel query Save query Export data as Show only Editor

Showing limited object explorer here. For full capability please click here to open Azure Data Studio.

Tables

- dbo.BuildVersion
- dbo.CMPL
 - COMPLAINTID (PK, int, not null)
 - CMPLID (nvarchar, null)
 - ODINO (nvarchar, null)
 - MFR_NAME (nvarchar, null)
 - MAKETXT (nvarchar, null)
 - MODELTXT (nvarchar, null)
 - YEARTXT (nvarchar, null)
 - CRASH (nvarchar, null)
 - FAILDATE (date, null)
 - FIRE (nvarchar, null)
 - INJURED (int, null)
 - DEATHS (int, null)
 - COMPDESC (nvarchar, null)
 - CITY (nvarchar, null)
 - STATE (nvarchar, null)
 - VIN (nvarchar, null)
 - DATEA (date, null)
 - MILES (int, null)
 - OCCURENCES (int, null)
 - CDESCR (nvarchar, null)

Query 1

```
1 SELECT *
2 FROM dbo.CMPL
3 WHERE CONVERT(Date, DATEA) = CONVERT(Date, GETDATE() - 3)
```

Results Messages

CMPLID	ODINO	MFR_NAME	MAKETXT	MODELTXT	YEARTXT	CRASH	FAILDATE	FIRE	INJURED	DEATHS	COMPDESC	CITY	STATE	VIN	DATEA	L
1930425	115470...	Chrysler (FC...	JEEP	WRANGLER	2020	N	2023-03-...	N	0	0	POWER TR...	Blacklick	OH	1C4H...	2023-09-28T00:00:00.0000000	2
1930426	115470...	Kia America...	KIA	SORENTO	2014	N	2021-06-...	N	0	0	ENGINE	Jackson	GA	5XYK...	2023-09-28T00:00:00.0000000	2
1930427	115470...	Chrysler (FC...	CHRYSLER	PACIFICA	2023	N	2023-09-...	N	0	0	ELECTRICAL...	Scotts...	AZ	2C4R...	2023-09-28T00:00:00.0000000	2
1930428	115470...	Chrysler (FC...	CHRYSLER	PACIFICA	2023	N	2023-09-...	N	0	0	FUEL/PROP...	Scotts...	AZ	2C4R...	2023-09-28T00:00:00.0000000	2
1930429	115470...	Chrysler (FC...	RAM	2500	2017	N	2023-07-...	N	0	0	SERVICE BR...	El Dor...	CA	3C6U...	2023-09-28T00:00:00.0000000	2
1930430	115470...	Chrysler (FC...	RAM	2500	2017	N	2023-07-...	N	0	0	ELECTRICAL...	El Dor...	CA	3C6U...	2023-09-28T00:00:00.0000000	2
1930431	115470...	General Mot...	CHEVROL...	BOLT EUV	2023	N	2023-09-...	N	0	0	FORWARD ...	Los An...	CA	1G1F...	2023-09-28T00:00:00.0000000	2
1930432	115470...	Chrysler (FC...	CHRYSLER	PACIFICA	2017	N	2023-09-...	N	0	0	ENGINE	Heber ...	UT	2C4R...	2023-09-28T00:00:00.0000000	2
1930433	115470...	General Mot...	CADILLAC	LYRIQ	2024	N	2023-09-...	N	0	0	EXTERIOR ...	Whitt ...	CO	1C6K...	2023-09-28T00:00:00.0000000	2

Query succeeded | 292s

Best of Luck!

