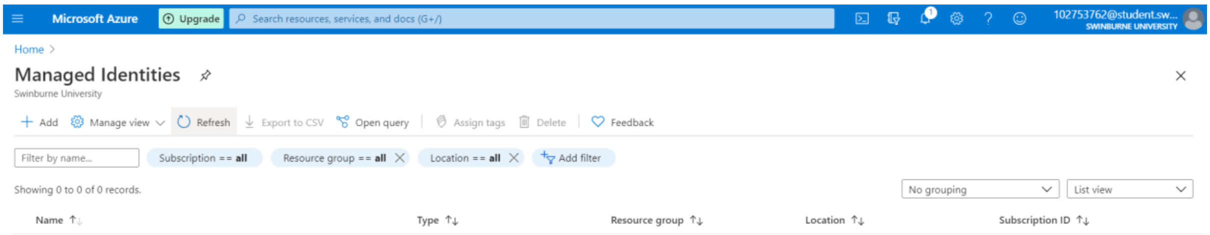
Task 3P

Step One: Creating a Managed Identity:



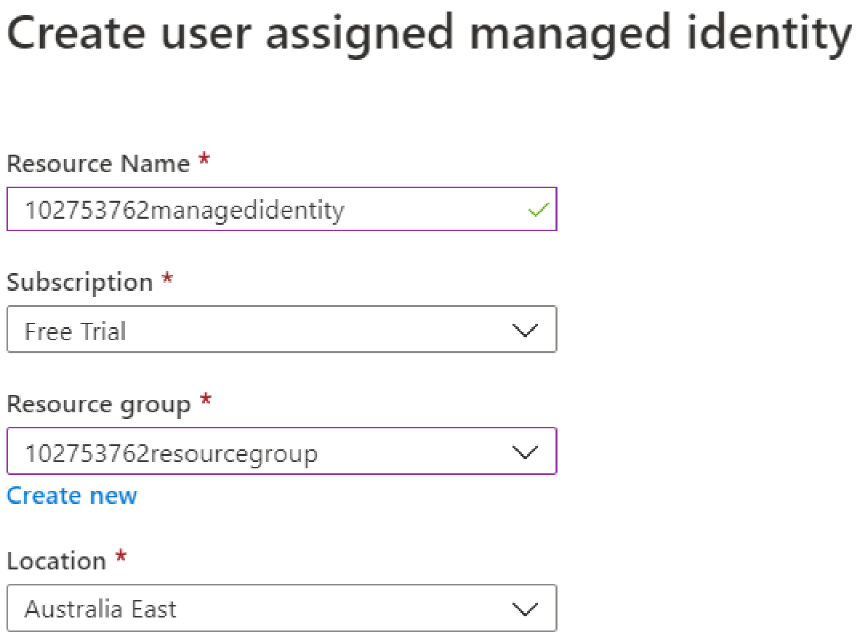
Managed Identities

Swinburne University

Subscription == all | Resource group == all | Location == all

Showing 0 to 0 of 0 records.

Name	Type	Resource group	Location	Subscription ID



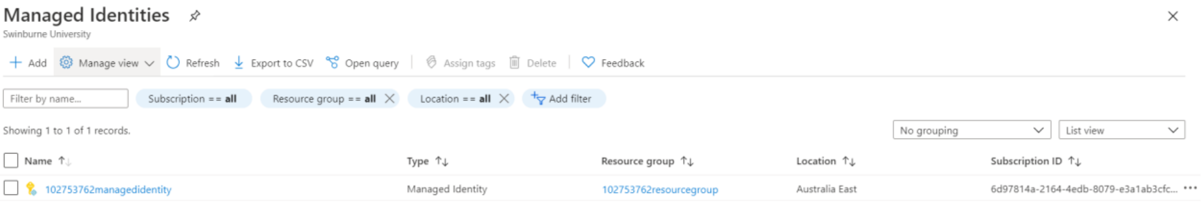
Create user assigned managed identity

Resource Name *: 102753762managedidentity

Subscription *: Free Trial

Resource group *: 102753762resourcegroup

Location *: Australia East



Managed Identities

Swinburne University

Subscription == all | Resource group == all | Location == all

Showing 1 to 1 of 1 records.

Name	Type	Resource group	Location	Subscription ID
102753762managedidentity	Managed Identity	102753762resourcegroup	Australia East	6d97814a-2164-4edb-8079-e3a1ab3cfc...

Question 1: What is a data warehouse, and how does a data lake differ from it?

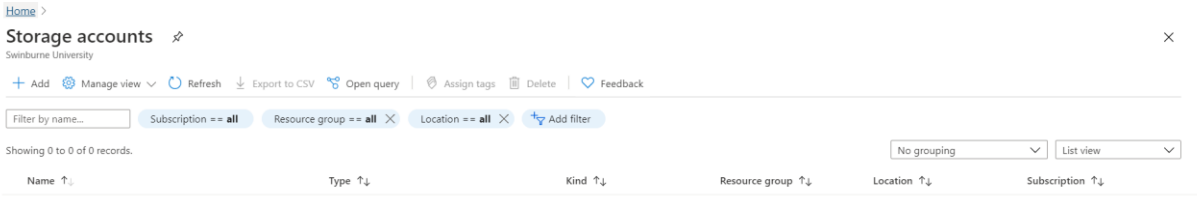
Answer: According to IBM, “A data warehouse is a system that aggregates data from different sources into a single, central, consistent data store to support business analytics, data mining, artificial intelligence (AI), and machine learning”.

On the other hand, according to redhat, “A data lake is a type of data repository that stores large and varied sets of raw data in its native format. Data lakes let you keep an unrefined view of your data”.

Difference between data warehouse and data lake:

01. When developing data warehouse, data source is analysed and it is a structured data model. Not all data from the source is stored in the data warehouse. So only data that will be used to answer a specific question or will be included in a report are stored in data warehouse. On the other hand, all data are stored in data lake not just the data that will be used for analysing or reporting.
02. All type of data can be stored in data lake including non-traditional data like, web server log, sensor data, social network activity etc. On the other hand, in data warehouse, data extracted from transactional system are stored.
03. 80% of the user, can use data warehouse as they need structured data. Structured data is easy to understand and work with. These users use these data to answer specific questions. 10% of the user, do more analysis. Sometimes they need more data then the one available in the data warehouse. Finally, there is user who does deep analysis. They need very large set of data to find out new questions which requires an answer. First two type of user can use data warehouse whereas for the next user they need data lake. However, data lake can support all user types.
04. It is time consuming to change a data warehouse. It requires developer resources. Where is data lake are more adaptable as there is no defined structure or the data is stored in raw format.
05. As all data is available in data lake, user can obtain their result faster than a data warehouse (Campbell 2015).

Step two: Creating a Data Lake Gen2 on Azure



Storage accounts

Swinburne University

Filter by name...

Subscription == **all**

Resource group == **all**

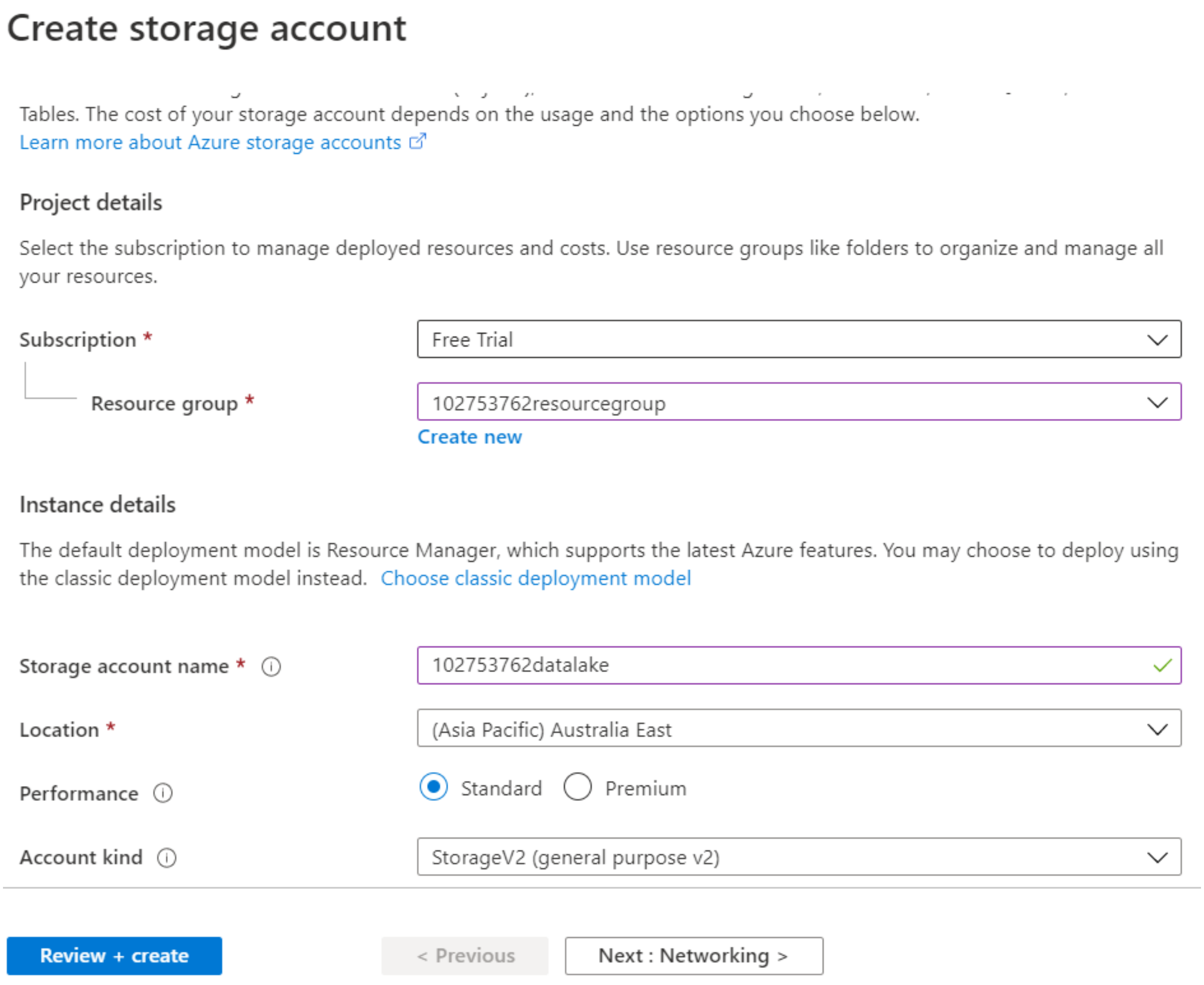
Location == **all**

Showing 0 to 0 of 0 records.

No grouping

List view

Name ↑↓	Type ↑↓	Kind ↑↓	Resource group ↑↓	Location ↑↓	Subscription ↑↓



Create storage account

Tables. The cost of your storage account depends on the usage and the options you choose below.
Learn more about Azure storage accounts

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * Free Trial

Resource group * 102753762resourcegroup

Create new

Instance details

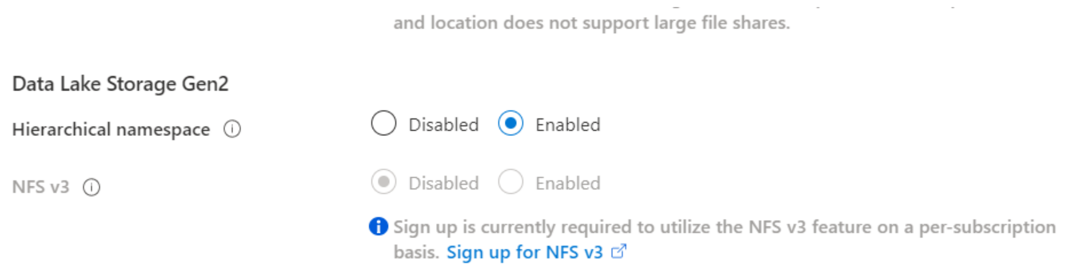
The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. Choose classic deployment model

Storage account name * 102753762datalake ✓

Location * (Asia Pacific) Australia East

Performance ● Standard ○ Premium

Account kind StorageV2 (general purpose v2)



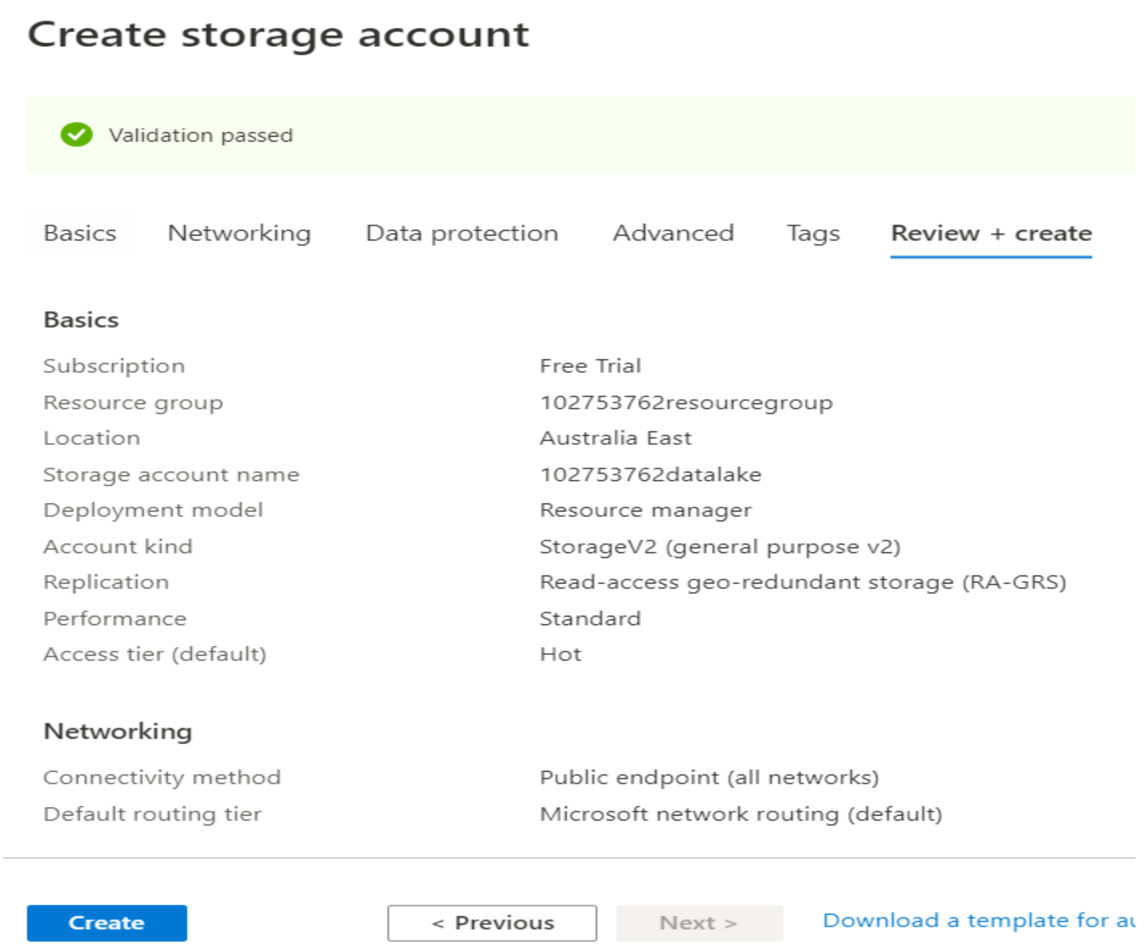
and location does not support large file shares.

Data Lake Storage Gen2

Hierarchical namespace ○ Disabled ● Enabled

NFS v3 ● Disabled ○ Enabled

Sign up is currently required to utilize the NFS v3 feature on a per-subscription basis. Sign up for NFS v3



Create storage account

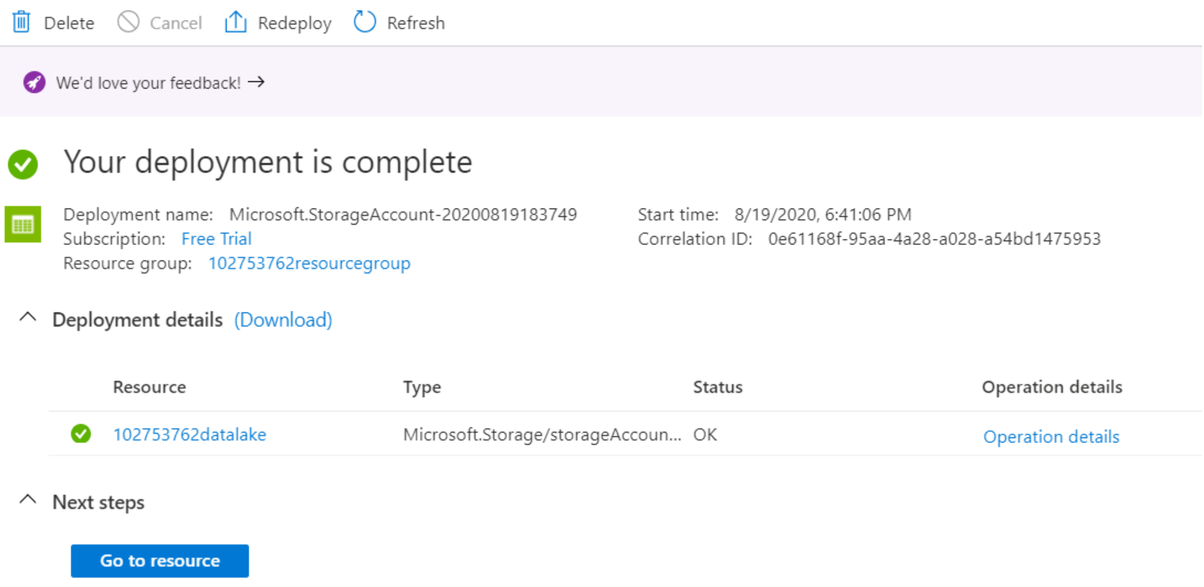
✓ Validation passed

Basics

Subscription	Free Trial
Resource group	102753762resourcegroup
Location	Australia East
Storage account name	102753762datalake
Deployment model	Resource manager
Account kind	StorageV2 (general purpose v2)
Replication	Read-access geo-redundant storage (RA-GRS)
Performance	Standard
Access tier (default)	Hot

Networking

Connectivity method	Public endpoint (all networks)
Default routing tier	Microsoft network routing (default)



✓ Your deployment is complete

Deployment name: Microsoft.StorageAccount-20200819183749
Subscription: Free Trial
Resource group: 102753762resourcegroup

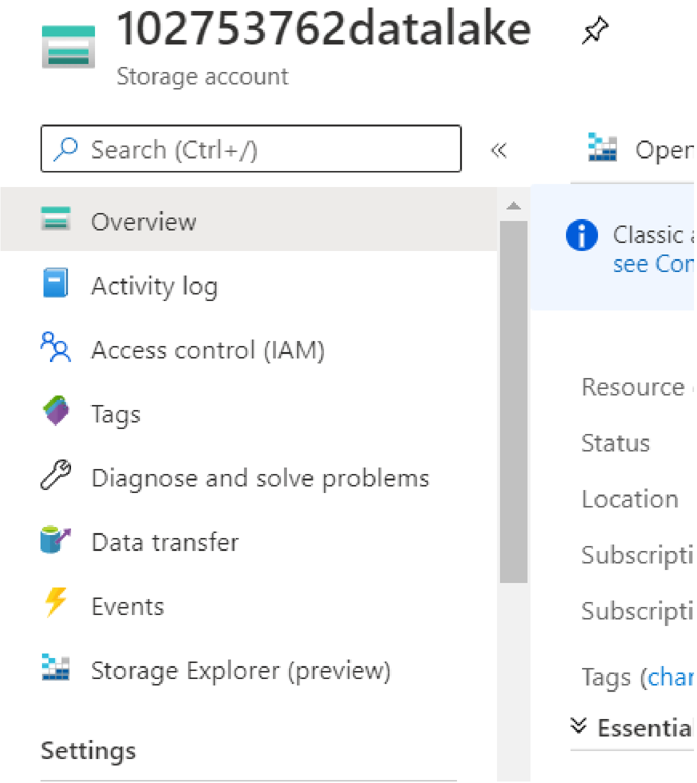
Start time: 8/19/2020, 6:41:06 PM
Correlation ID: 0e61168f-95aa-4a28-a028-a54bd1475953

Deployment details (Download)

Resource	Type	Status	Operation details
✓ 102753762datalake	Microsoft.Storage/storageAccoun...	OK	Operation details

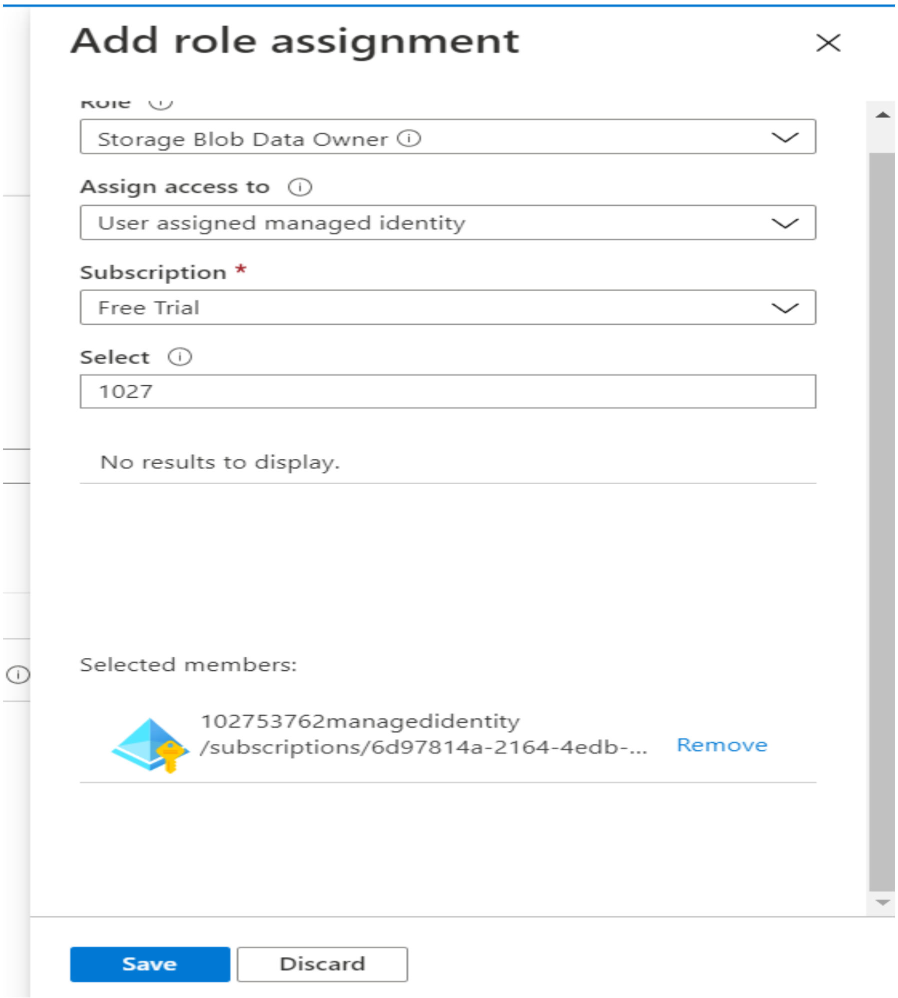
Next steps

Step Three: Set up permissions for the managed identity on the Data Lake Storage Gen2



102753762datalake

Storage account



Add role assignment

Role

Storage Blob Data Owner

Assign access to

User assigned managed identity

Subscription *

Free Trial

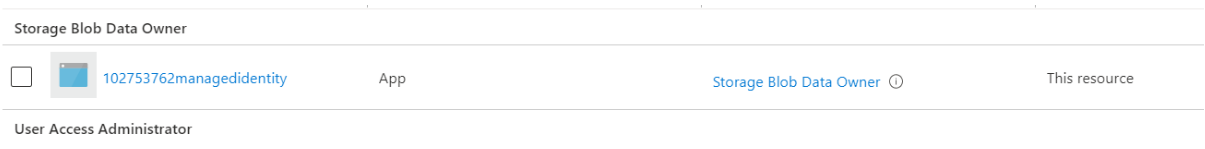
Select

1027

No results to display.

Selected members:

102753762managedidentity
/subscriptions/6d97814a-2164-4edb-...

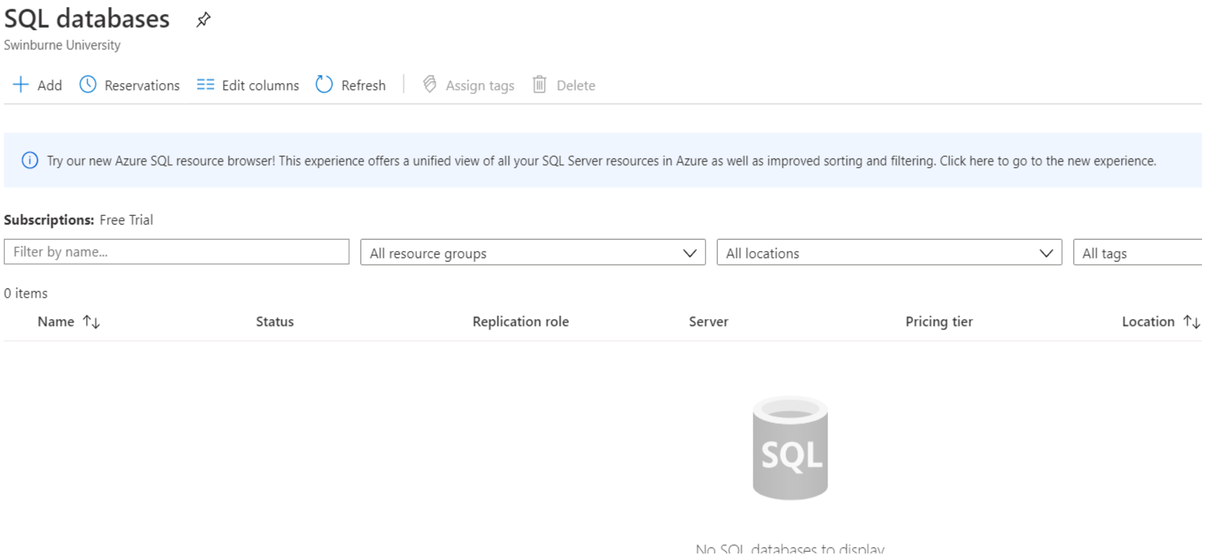


Storage Blob Data Owner

□	102753762managedidentity	App	Storage Blob Data Owner ⓘ	This resource

User Access Administrator

Step Four: Create an SQL Database



SQL databases

Swinburne University

ⓘ Try our new Azure SQL resource browser! This experience offers a unified view of all your SQL Server resources in Azure as well as improved sorting and filtering. Click here to go to the new experience.

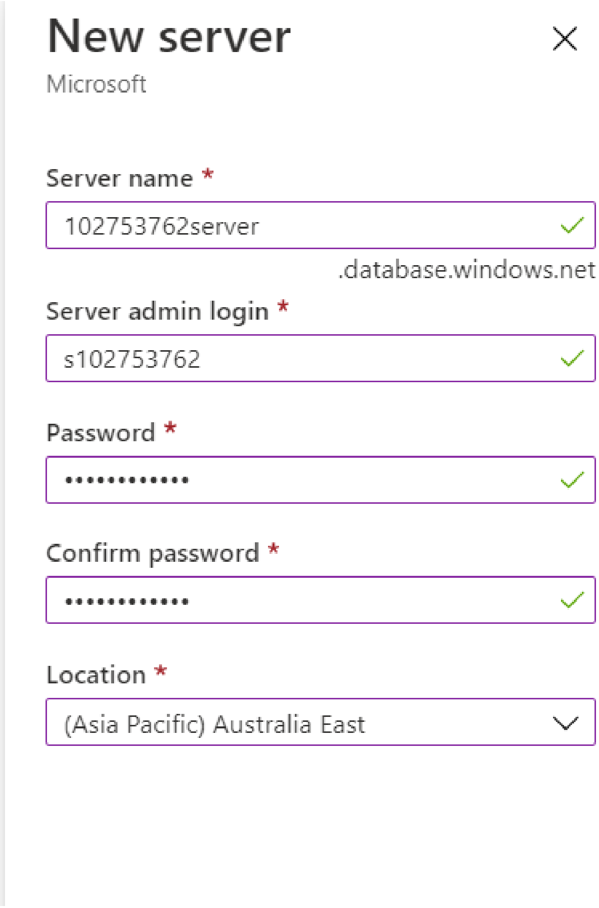
Subscriptions: Free Trial

Filter by name... | All resource groups | All locations | All tags

0 items

Name ↑↓	Status	Replication role	Server	Pricing tier	Location ↑↓

No SQL databases to display



New server

Microsoft

Server name *

102753762server ✓

.database.windows.net

Server admin login *

s102753762 ✓

Password *

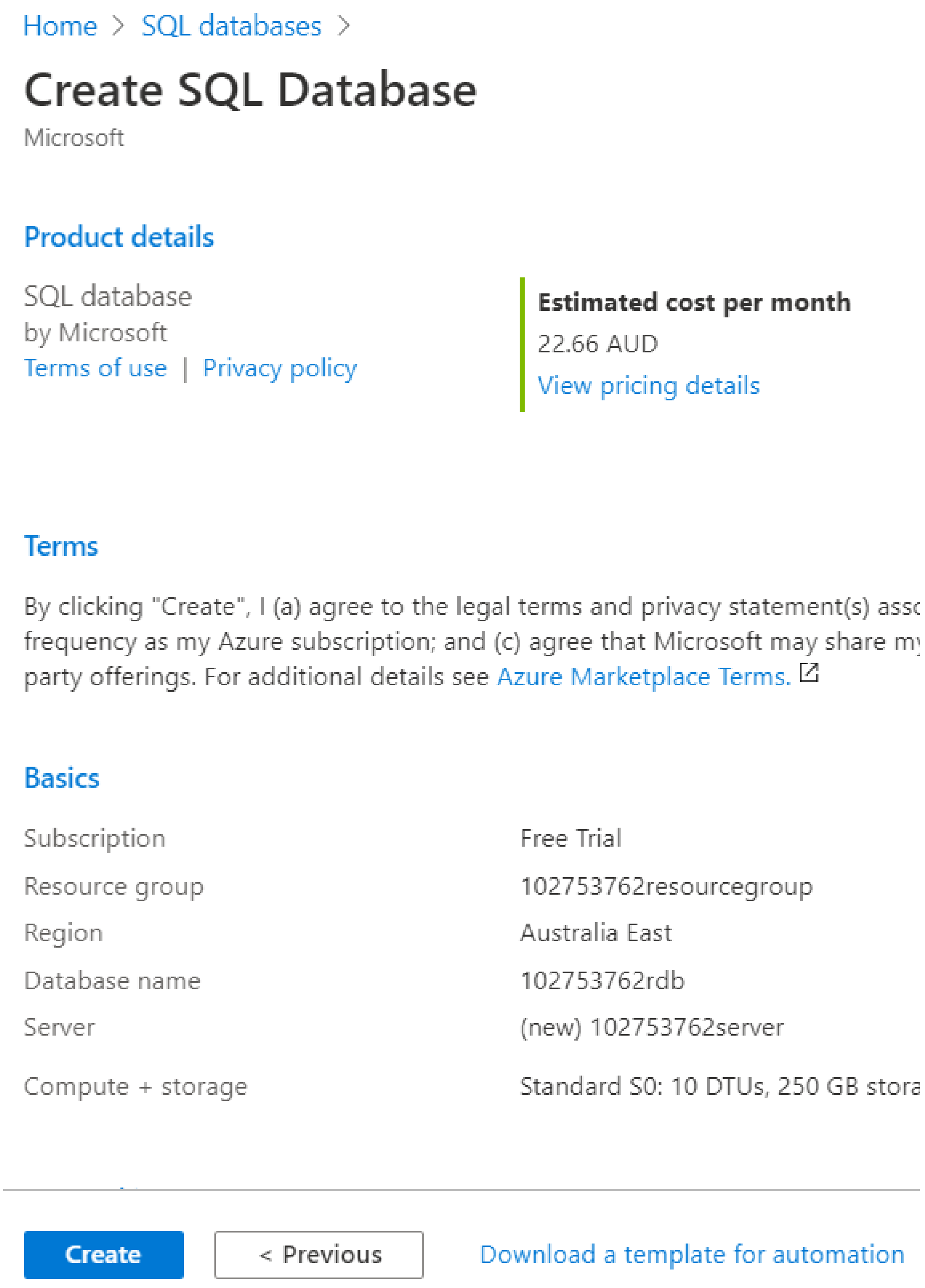
•••••••••••• ✓

Confirm password *

•••••••••••• ✓

Location *

(Asia Pacific) Australia East



Home > SQL databases >

Create SQL Database

Microsoft

Product details

SQL database
by Microsoft
Terms of use | Privacy policy

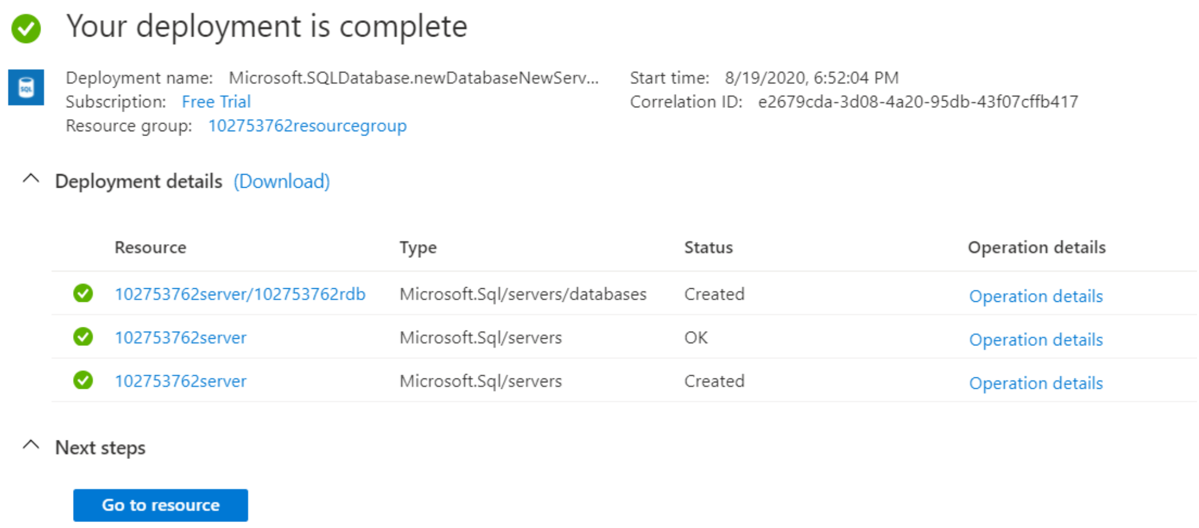
Estimated cost per month
22.66 AUD
View pricing details

Terms

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) assc
frequency as my Azure subscription; and (c) agree that Microsoft may share my
party offerings. For additional details see Azure Marketplace Terms.

Basics

Subscription	Free Trial
Resource group	102753762resourcegroup
Region	Australia East
Database name	102753762rdb
Server	(new) 102753762server
Compute + storage	Standard S0: 10 DTUs, 250 GB stora



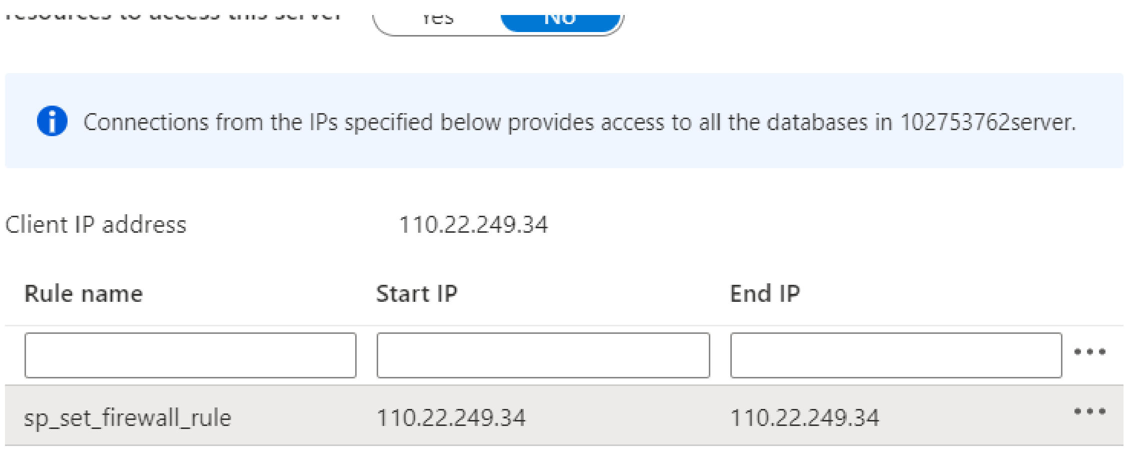
Your deployment is complete

Deployment name: Microsoft.SQLDatabase.newDatabaseNewServ...
Subscription: Free Trial
Resource group: 102753762resourcegroup
Start time: 8/19/2020, 6:52:04 PM
Correlation ID: e2679cda-3d08-4a20-95db-43f07cffb417

Deployment details (Download)

Resource	Type	Status	Operation details
✓ 102753762server/102753762rdb	Microsoft.Sql/servers/databases	Created	Operation details
✓ 102753762server	Microsoft.Sql/servers	OK	Operation details
✓ 102753762server	Microsoft.Sql/servers	Created	Operation details

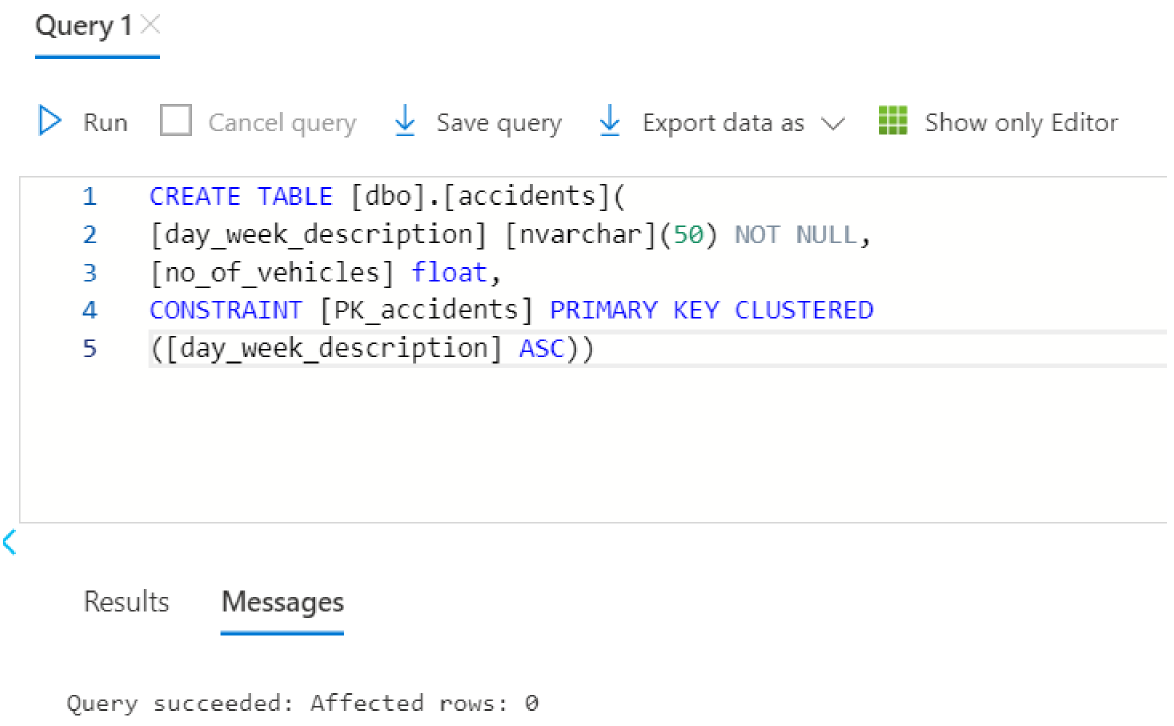
Next steps



Connections from the IPs specified below provides access to all the databases in 102753762server.

Client IP address 110.22.249.34

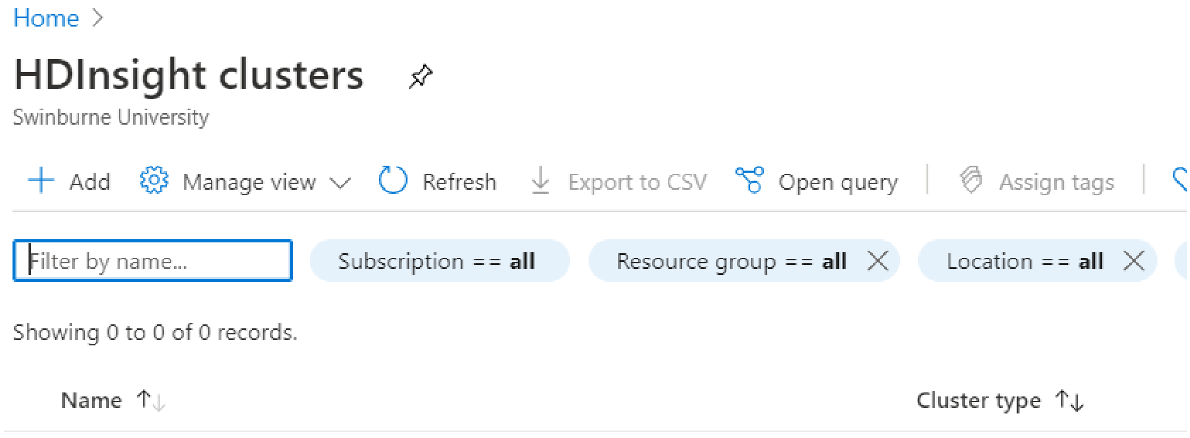
Rule name	Start IP	End IP
sp_set_firewall_rule	110.22.249.34	110.22.249.34



```
CREATE TABLE [dbo].[accidents](
[day_week_description] [nvarchar](50) NOT NULL,
[no_of_vehicles] float,
CONSTRAINT [PK_accidents] PRIMARY KEY CLUSTERED
([day_week_description] ASC))
```

```
Query succeeded: Affected rows: 0
```

Step Five: Create a HDInsight Cluster



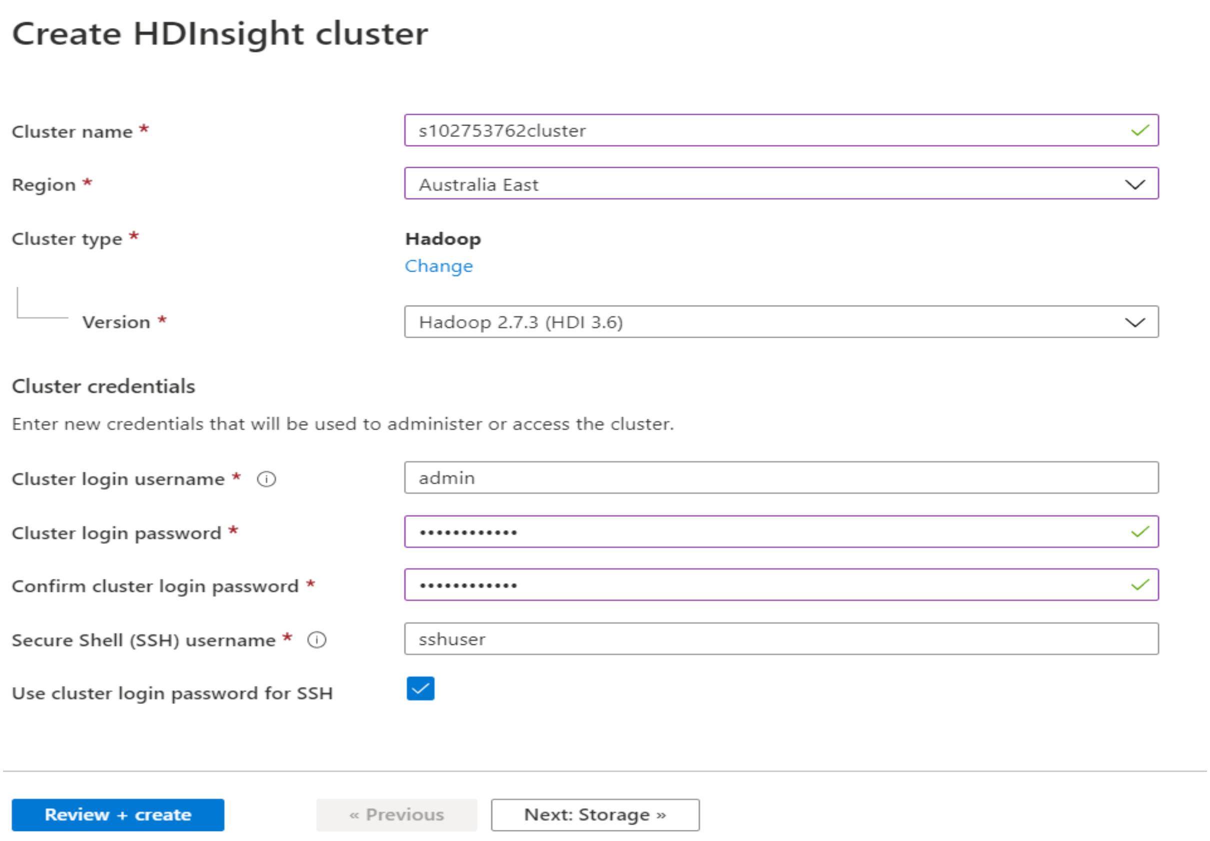
HDInsight clusters

Swinburne University

Subscription == all | Resource group == all | Location == all

Showing 0 to 0 of 0 records.

Name	Cluster type



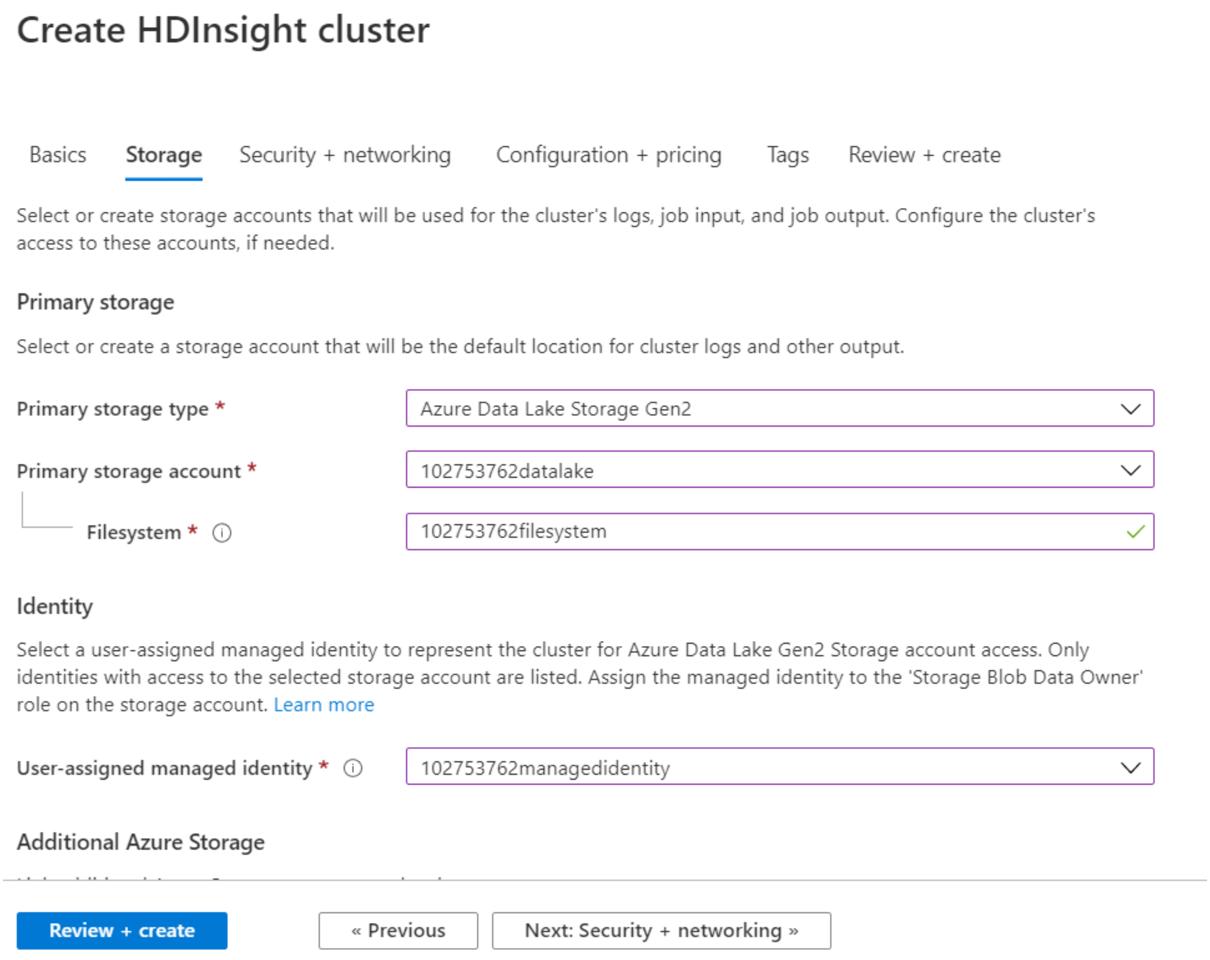
Create HDInsight cluster

Cluster name *	s102753762cluster ✓
Region *	Australia East
Cluster type *	**Hadoop** Change
Version *	Hadoop 2.7.3 (HDI 3.6)

Cluster credentials

Enter new credentials that will be used to administer or access the cluster.

Cluster login username * ⓘ	admin
Cluster login password *	•••••••••••• ✓
Confirm cluster login password *	•••••••••••• ✓
Secure Shell (SSH) username * ⓘ	sshuser
Use cluster login password for SSH	☑



Create HDInsight cluster

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

Primary storage

Select or create a storage account that will be the default location for cluster logs and other output.

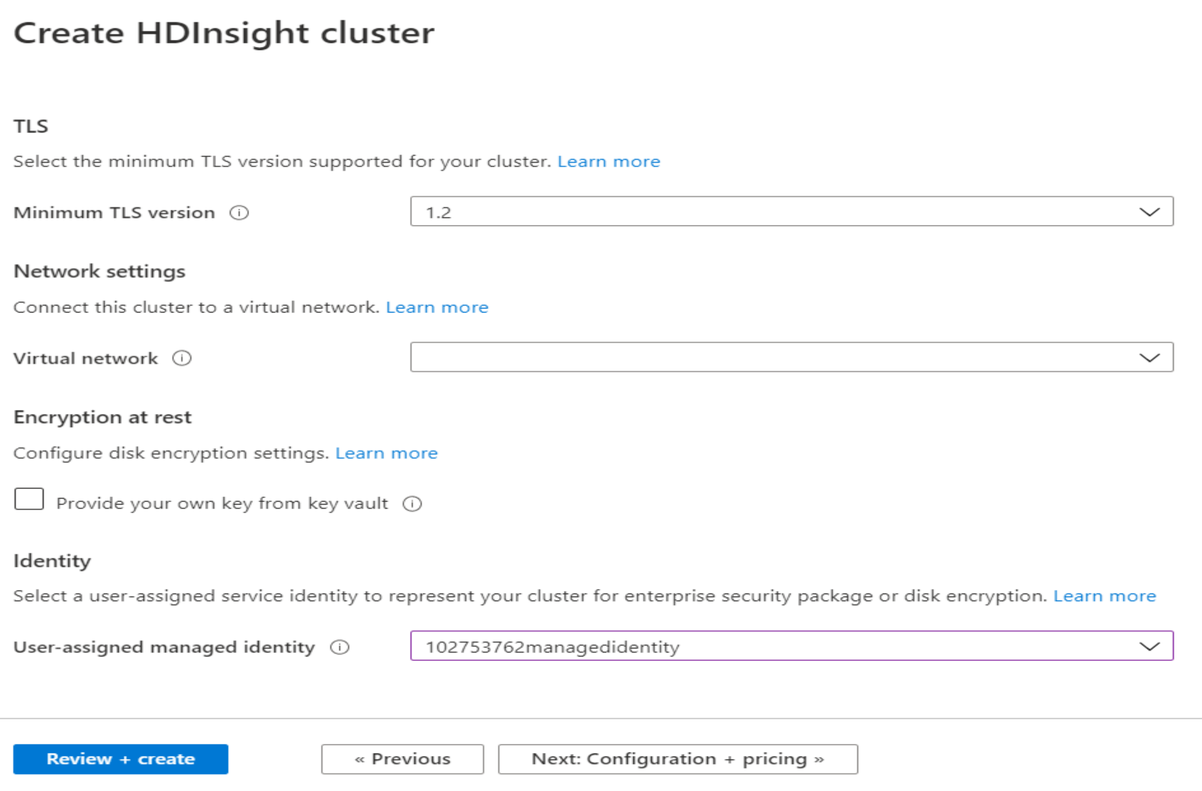
Primary storage type *	Azure Data Lake Storage Gen2
Primary storage account *	102753762datalake
Filesystem * ⓘ	102753762filesystem ✓

Identity

Select a user-assigned managed identity to represent the cluster for Azure Data Lake Gen2 Storage account access. Only identities with access to the selected storage account are listed. Assign the managed identity to the 'Storage Blob Data Owner' role on the storage account. Learn more

User-assigned managed identity * ⓘ	102753762managedidentity

Additional Azure Storage



Create HDInsight cluster

TLS

Select the minimum TLS version supported for your cluster. Learn more

Minimum TLS version ⓘ 1.2

Network settings

Connect this cluster to a virtual network. Learn more

Virtual network ⓘ

Encryption at rest

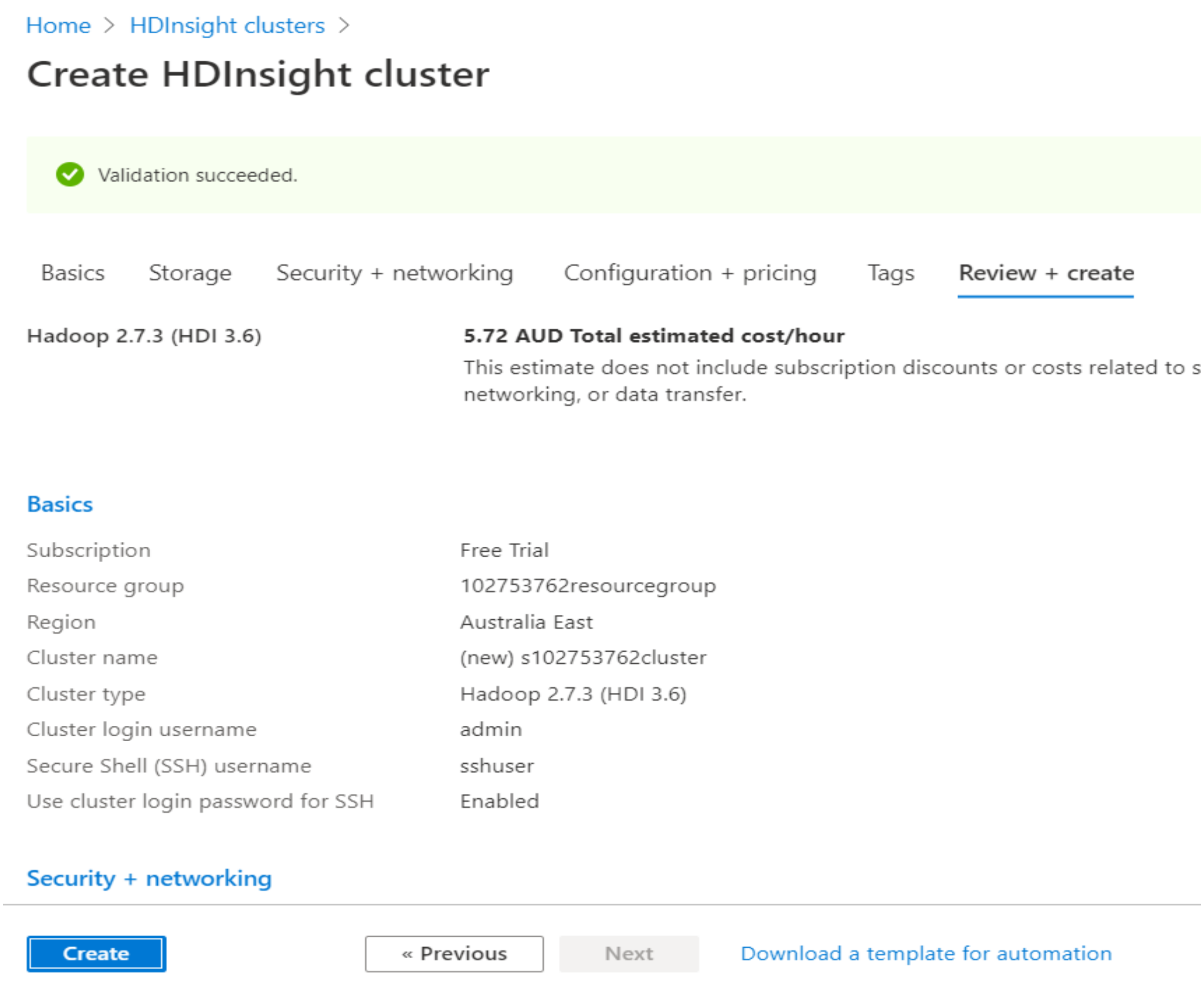
Configure disk encryption settings. Learn more

□ Provide your own key from key vault ⓘ

Identity

Select a user-assigned service identity to represent your cluster for enterprise security package or disk encryption. Learn more

User-assigned managed identity ⓘ 102753762managedidentity



Home > HDInsight clusters >

Create HDInsight cluster

✓ Validation succeeded.

Hadoop 2.7.3 (HDI 3.6)

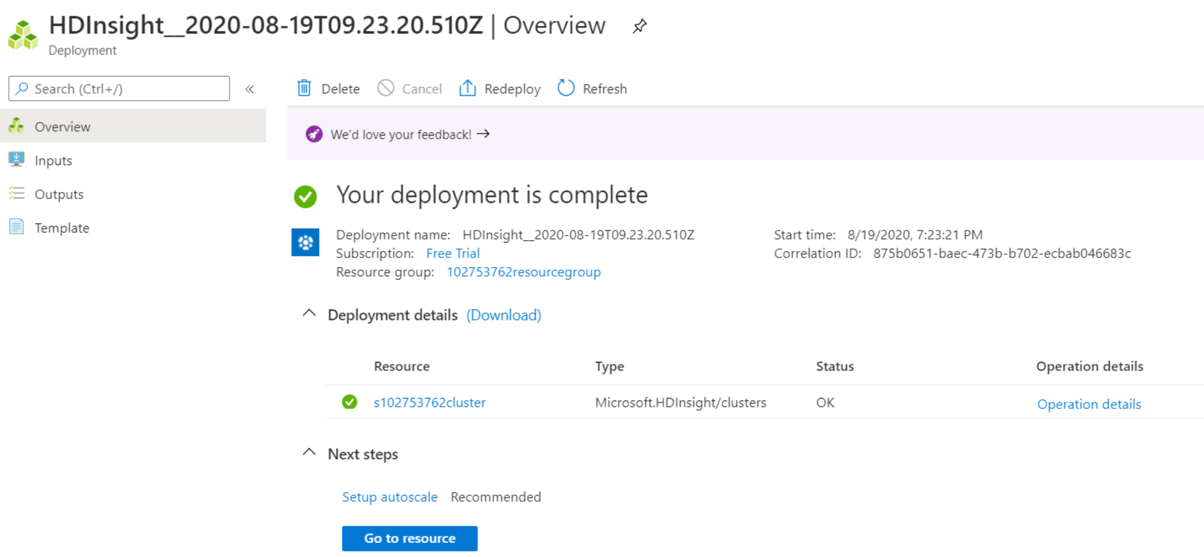
5.72 AUD Total estimated cost/hour

This estimate does not include subscription discounts or costs related to s networking, or data transfer.

Basics

Subscription	Free Trial
Resource group	102753762resourcegroup
Region	Australia East
Cluster name	(new) s102753762cluster
Cluster type	Hadoop 2.7.3 (HDI 3.6)
Cluster login username	admin
Secure Shell (SSH) username	sshuser
Use cluster login password for SSH	Enabled

Security + networking



HDInsight__2020-08-19T09.23.20.510Z | Overview

Deployment

Your deployment is complete

Deployment name: HDInsight__2020-08-19T09.23.20.510Z
Subscription: Free Trial
Resource group: 102753762resourcegroup

Start time: 8/19/2020, 7:23:21 PM
Correlation ID: 875b0651-baec-473b-b702-ecbab046683c

Deployment details (Download)

Resource	Type	Status	Operation details
s102753762cluster	Microsoft.HDInsight/clusters	OK	Operation details

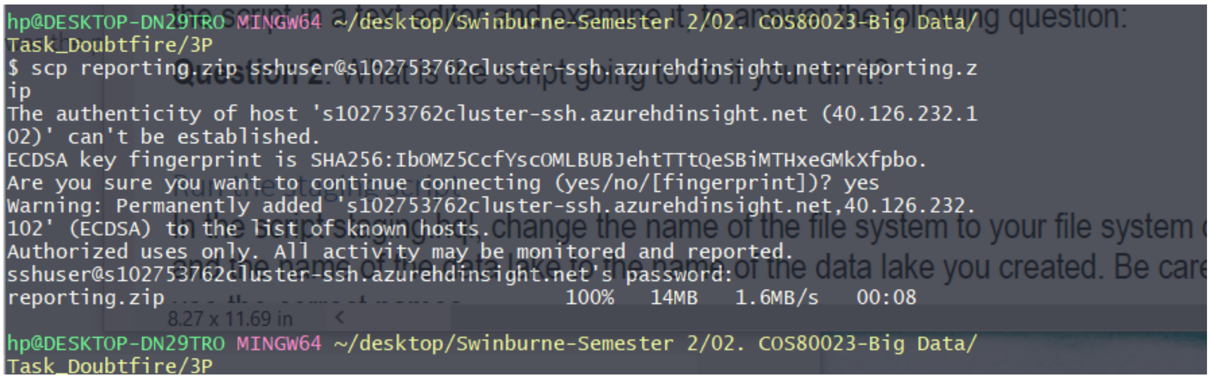
Next steps

Setup autoscale Recommended

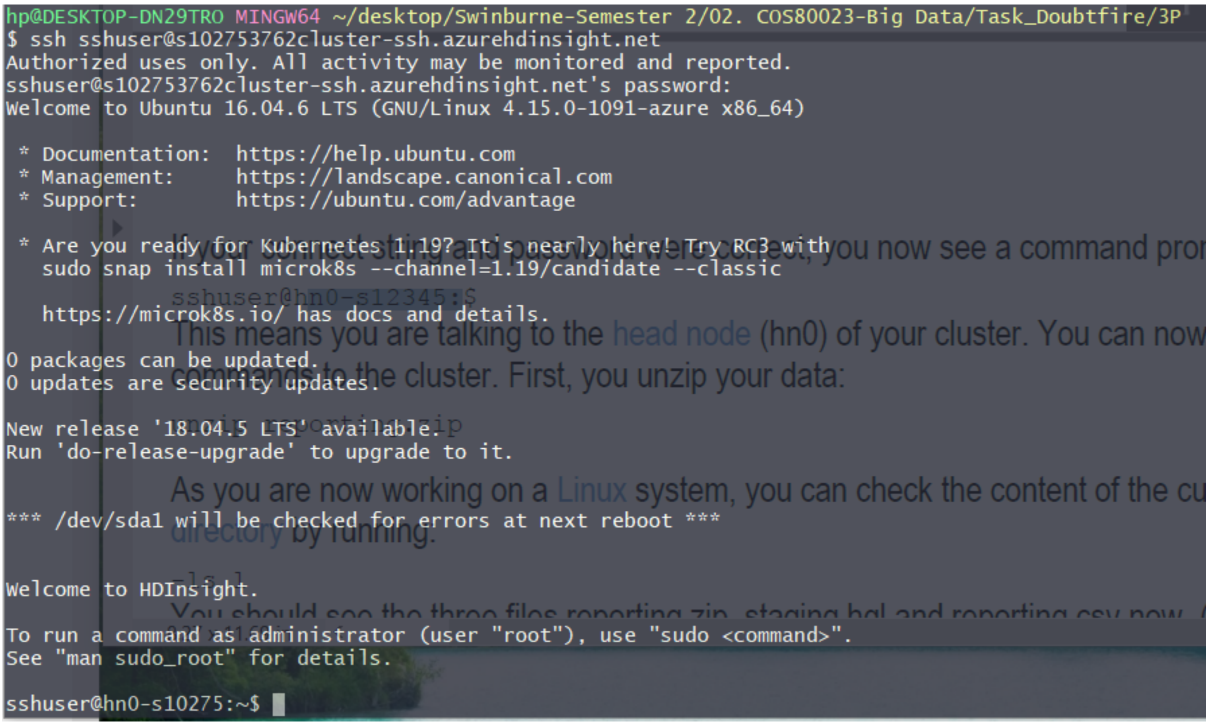
Step Six: Upload the data and staging code

Question 2: What is the script going to do if you run it?

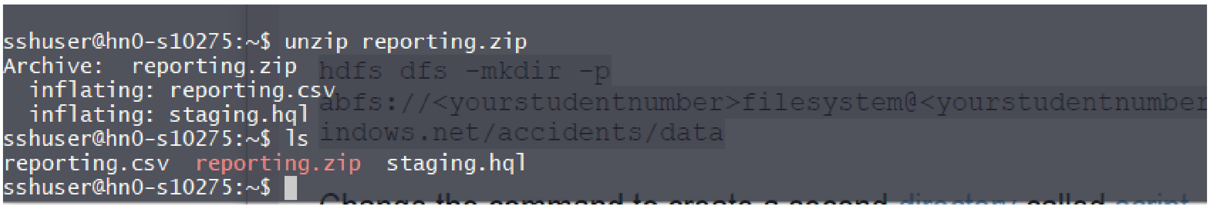
Answer: The script will create a table and store it in textfile format in “data” directory.



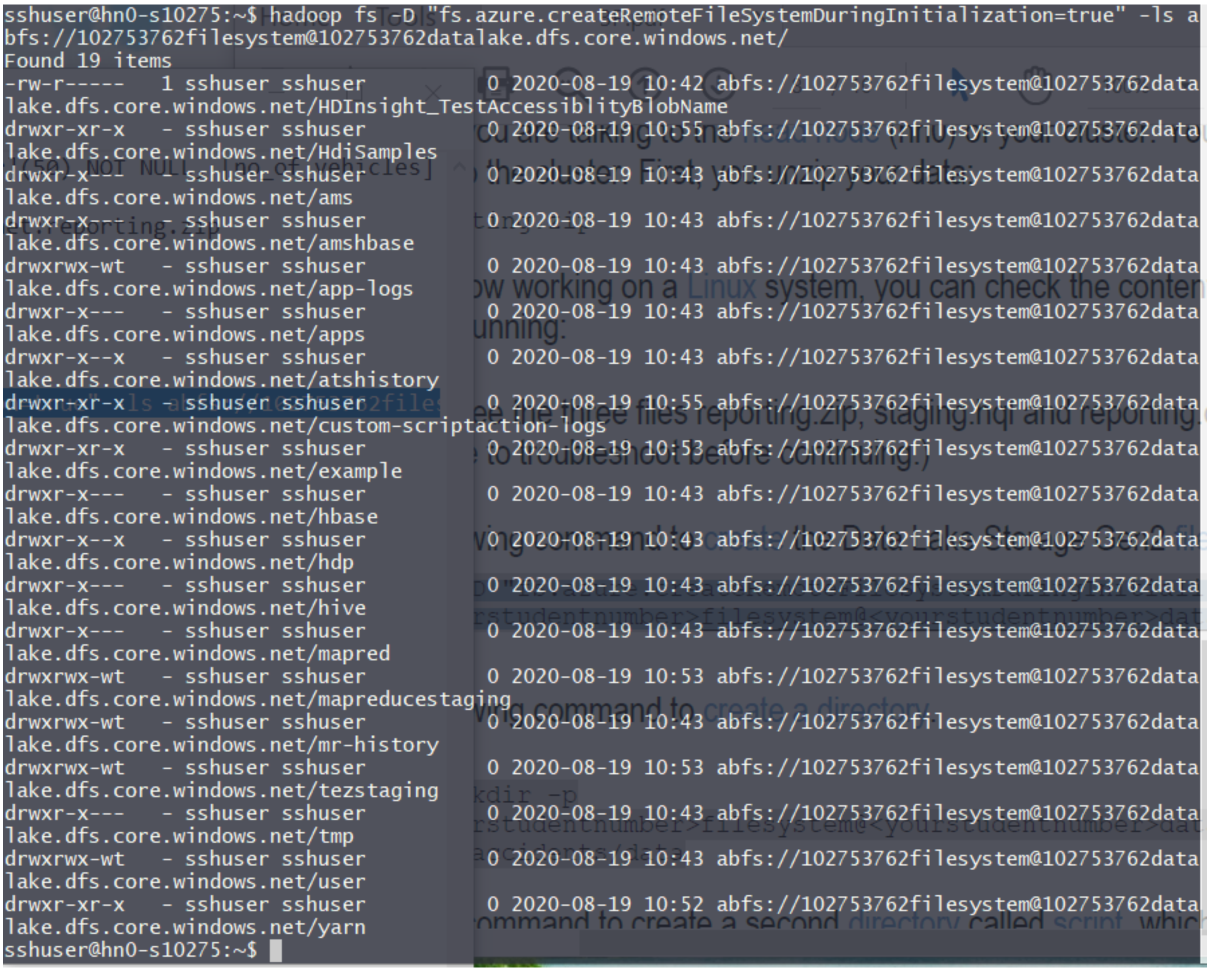
```
hp@DESKTOP-DN29TRO MINGW64 ~/desktop/Swinburne-Semester 2/02. COS80023-Big Data/
Task_Doubtfire/3P
$ scp reporting.zip sshuser@s102753762cluster-ssh.azurehdinsight.net:reporting.z
ip
The authenticity of host 's102753762cluster-ssh.azurehdinsight.net (40.126.232.1
02)' can't be established.
ECDSA key fingerprint is SHA256:IbOMZ5CcfYscOMLBUBJehtTTtQeSBiMTHxeGMkXfpbo.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 's102753762cluster-ssh.azurehdinsight.net,40.126.232.
102' (ECDSA) to the list of known hosts.
Authorized uses only. All activity may be monitored and reported.
sshuser@s102753762cluster-ssh.azurehdinsight.net's password:
reporting.zip                                   100%   14MB   1.6MB/s   00:08

hp@DESKTOP-DN29TRO MINGW64 ~/desktop/Swinburne-Semester 2/02. COS80023-Big Data/
Task_Doubtfire/3P
```



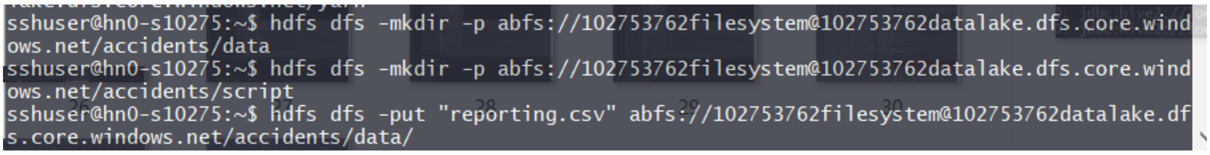
```
hp@DESKTOP-DN29TRO MINGW64 ~/desktop/Swinburne-Semester 2/02. COS80023-Big Data/Task_Doubtfire/3P
$ ssh sshuser@s102753762cluster-ssh.azurehdinsight.net
Authorized uses only. All activity may be monitored and reported.
sshuser@s102753762cluster-ssh.azurehdinsight.net's password:
Welcome to Ubuntu 16.04.6 LTS (GNU/Linux 4.15.0-1091-azure x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Are you ready for Kubernetes 1.19? It's nearly here! Try RC3 with
   sudo snap install microk8s --channel=1.19/candidate --classic

   https://microk8s.io/ has docs and details.

0 packages can be updated.
0 updates are security updates.

New release '18.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


*** /dev/sda1 will be checked for errors at next reboot ***


Welcome to HDInsight.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

sshuser@hn0-s10275:~$
```



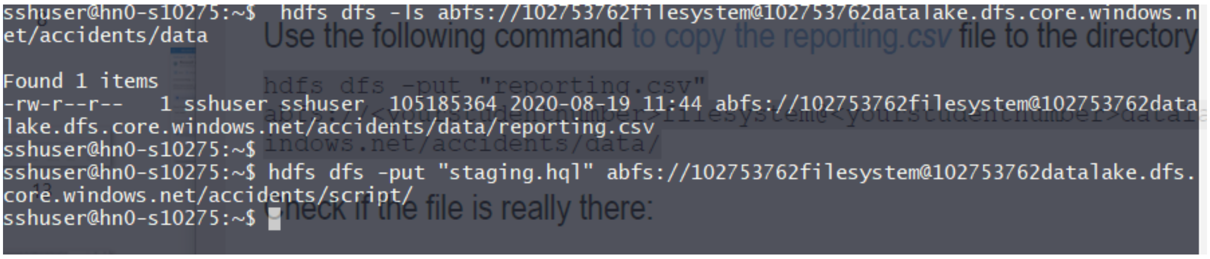
```
sshuser@hn0-s10275:~$ unzip reporting.zip
Archive:  reporting.zip
  inflating: reporting.csv
  inflating: staging.hql
sshuser@hn0-s10275:~$ ls
reporting.csv  reporting.zip  staging.hql
sshuser@hn0-s10275:~$
```



```
sshuser@hn0-s10275:~$ hadoop fs -D "fs.azure.createRemoteFileSystemDuringInitialization=true" -ls a
bfs://102753762filesystem@102753762datalake.dfs.core.windows.net/
Found 19 items
-rw-r-----   1 sshuser sshuser          0 2020-08-19 10:42 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/HDInsight_TestAccessiblityBlobName
drwxr-xr-x   - sshuser sshuser          0 2020-08-19 10:55 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/HdiSamples
drwxr-x---   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/ams
drwxr-x---   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/amshbase
drwxrwx-wt   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/app-logs
drwxr-x---   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/apps
drwxr-x--x   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/atshistory
drwxr-xr-x   - sshuser sshuser          0 2020-08-19 10:55 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/custom-scriptaction-logs
drwxr-xr-x   - sshuser sshuser          0 2020-08-19 10:53 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/example
drwxr-x---   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/hbase
drwxr-x--x   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/hdp
drwxr-x---   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/hive
drwxr-x---   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/mapred
drwxrwx-wt   - sshuser sshuser          0 2020-08-19 10:53 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/mapreducestaging
drwxrwx-wt   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/mr-history
drwxrwx-wt   - sshuser sshuser          0 2020-08-19 10:53 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/tezstaging
drwxr-x---   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/tmp
drwxrwx-wt   - sshuser sshuser          0 2020-08-19 10:43 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/user
drwxr-xr-x   - sshuser sshuser          0 2020-08-19 10:52 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/yarn
sshuser@hn0-s10275:~$
```

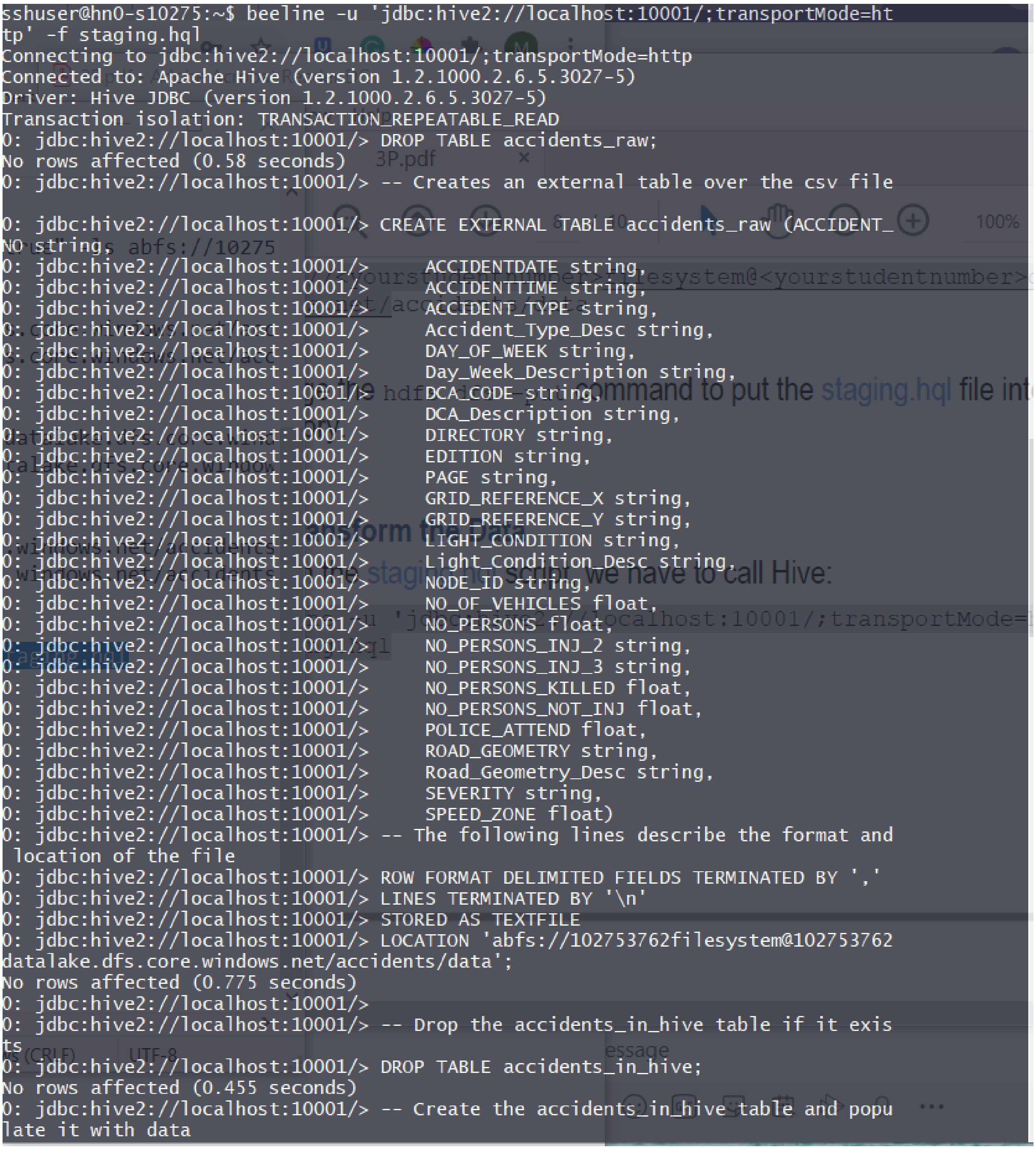


```
sshuser@hn0-s10275:~$ hdfs dfs -mkdir -p abfs://102753762filesystem@102753762datalake.dfs.core.wind
ows.net/accidents/data
sshuser@hn0-s10275:~$ hdfs dfs -mkdir -p abfs://102753762filesystem@102753762datalake.dfs.core.wind
ows.net/accidents/script
sshuser@hn0-s10275:~$ hdfs dfs -put "reporting.csv" abfs://102753762filesystem@102753762datalake.df
s.core.windows.net/accidents/data/
```

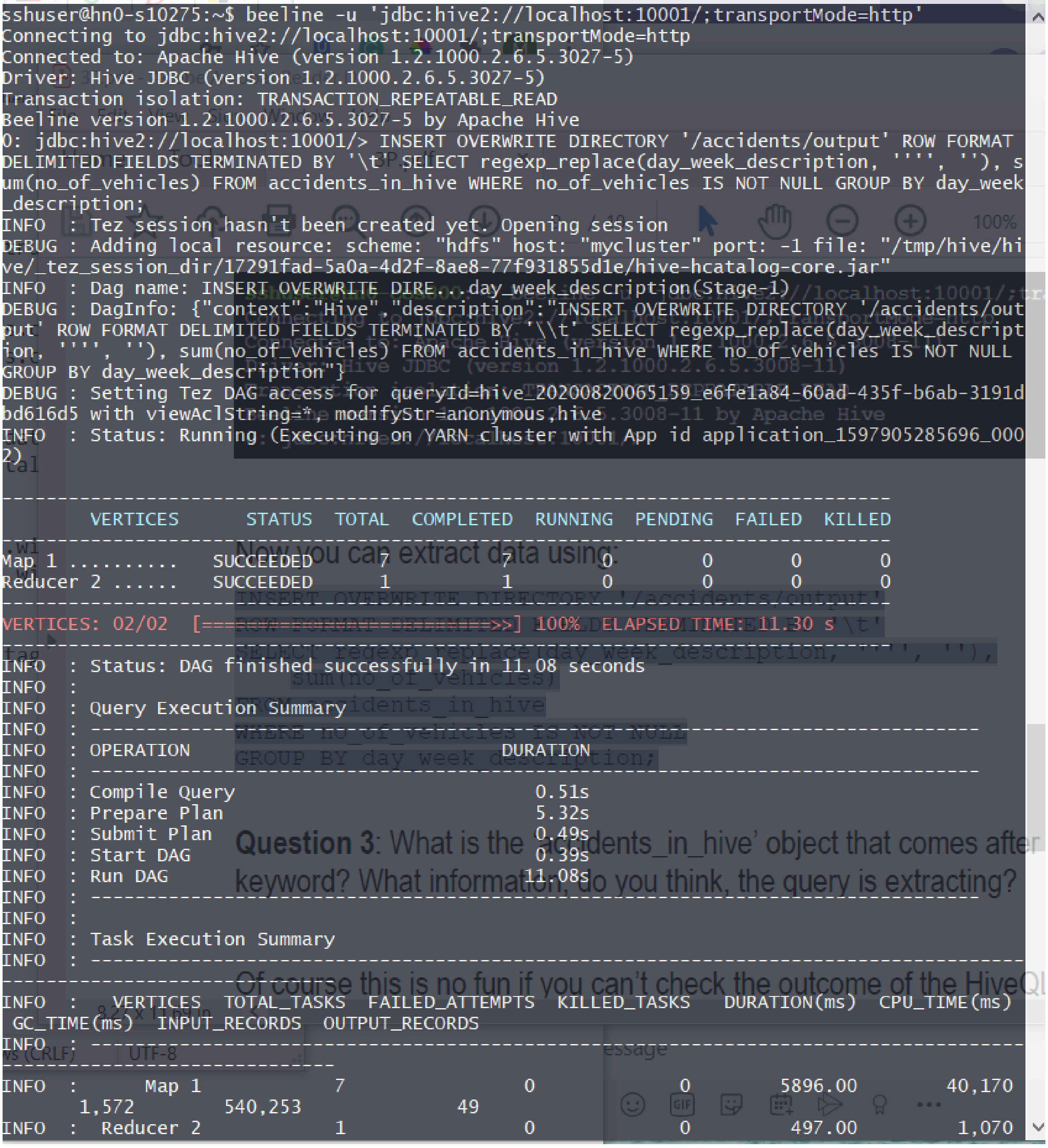


```
sshuser@hn0-s10275:~$  hdfs dfs -ls abfs://102753762filesystem@102753762datalake.dfs.core.windows.n
et/accidents/data

Found 1 items
-rw-r--r--   1 sshuser sshuser  105185364 2020-08-19 11:44 abfs://102753762filesystem@102753762data
lake.dfs.core.windows.net/accidents/data/reporting.csv
sshuser@hn0-s10275:~$
sshuser@hn0-s10275:~$ hdfs dfs -put "staging.hql" abfs://102753762filesystem@102753762datalake.dfs.
core.windows.net/accidents/script/
sshuser@hn0-s10275:~$
```

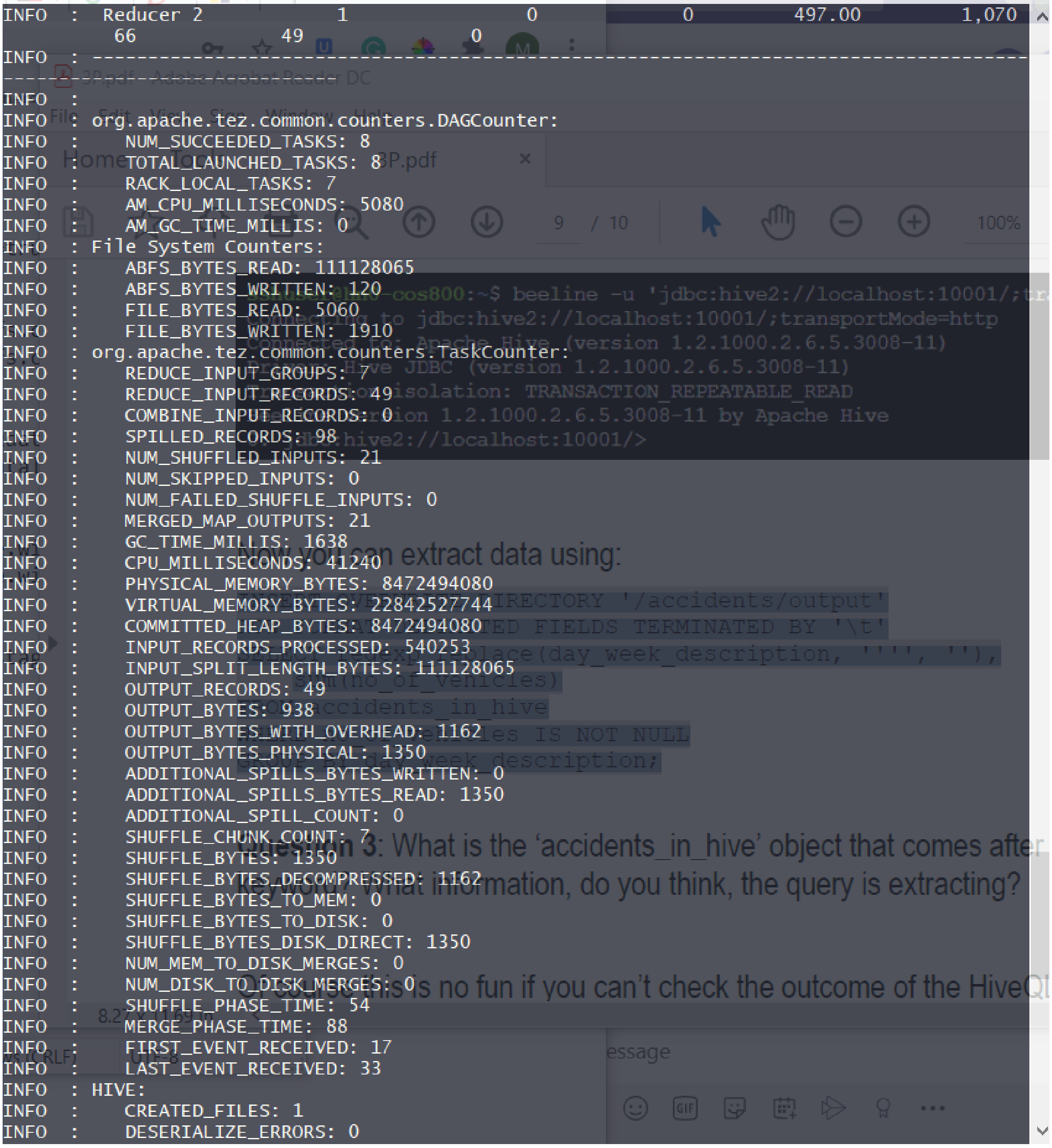
Step Seven: Transform the Data



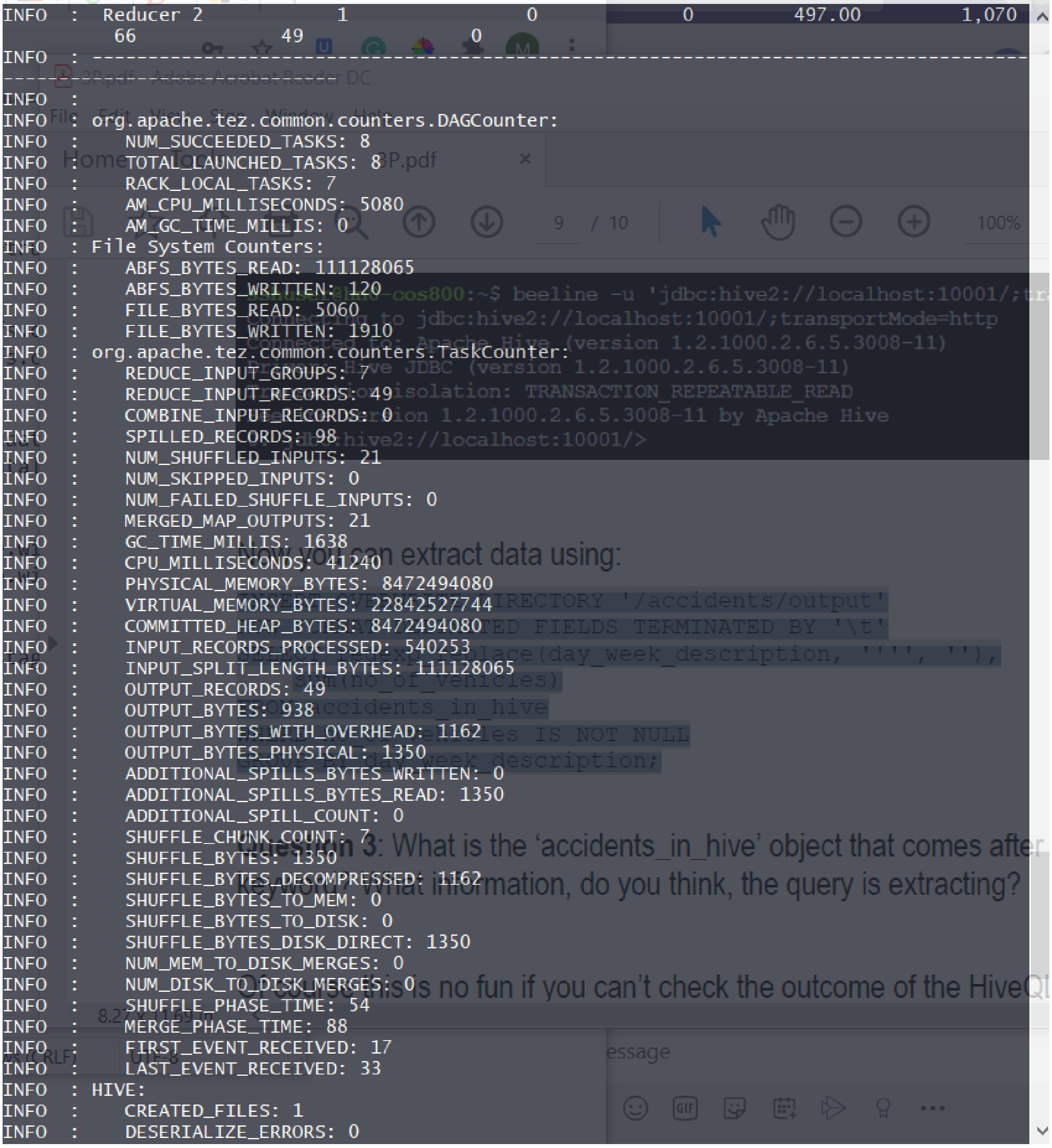
```
sshuser@hn0-s10275:~$ beeline -u 'jdbc:hive2://localhost:10001/;transportMode=ht
tp' -f staging.hql
Connecting to jdbc:hive2://localhost:10001/;transportMode=http
Connected to: Apache Hive (version 1.2.1000.2.6.5.3027-5)
Driver: Hive JDBC (version 1.2.1000.2.6.5.3027-5)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://localhost:10001/> DROP TABLE accidents_raw;
No rows affected (0.58 seconds)
0: jdbc:hive2://localhost:10001/> -- Creates an external table over the csv file

0: jdbc:hive2://localhost:10001/> CREATE EXTERNAL TABLE accidents_raw (ACCIDENT_
NO string,
0: jdbc:hive2://localhost:10001/>     ACCIDENTDATE string,
0: jdbc:hive2://localhost:10001/>     ACCIDENTTIME string,
0: jdbc:hive2://localhost:10001/>     ACCIDENT_TYPE string,
0: jdbc:hive2://localhost:10001/>     Accident_Type_Desc string,
0: jdbc:hive2://localhost:10001/>     DAY_OF_WEEK string,
0: jdbc:hive2://localhost:10001/>     Day_Week_Description string,
0: jdbc:hive2://localhost:10001/>     DCA_CODE string,
0: jdbc:hive2://localhost:10001/>     DCA_Description string,
0: jdbc:hive2://localhost:10001/>     DIRECTORY string,
0: jdbc:hive2://localhost:10001/>     EDITION string,
0: jdbc:hive2://localhost:10001/>     PAGE string,
0: jdbc:hive2://localhost:10001/>     GRID_REFERENCE_X string,
0: jdbc:hive2://localhost:10001/>     GRID_REFERENCE_Y string,
0: jdbc:hive2://localhost:10001/>     LIGHT_CONDITION string,
0: jdbc:hive2://localhost:10001/>     Light_Condition_Desc string,
0: jdbc:hive2://localhost:10001/>     NODE_ID string,
0: jdbc:hive2://localhost:10001/>     NO_OF_VEHICLES float,
0: jdbc:hive2://localhost:10001/>     NO_PERSONS float,
0: jdbc:hive2://localhost:10001/>     NO_PERSONS_INJ_2 string,
0: jdbc:hive2://localhost:10001/>     NO_PERSONS_INJ_3 string,
0: jdbc:hive2://localhost:10001/>     NO_PERSONS_KILLED float,
0: jdbc:hive2://localhost:10001/>     NO_PERSONS_NOT_INJ float,
0: jdbc:hive2://localhost:10001/>     POLICE_ATTEND float,
0: jdbc:hive2://localhost:10001/>     ROAD_GEOMETRY string,
0: jdbc:hive2://localhost:10001/>     Road_Geometry_Desc string,
0: jdbc:hive2://localhost:10001/>     SEVERITY string,
0: jdbc:hive2://localhost:10001/>     SPEED_ZONE float)
0: jdbc:hive2://localhost:10001/> -- The following lines describe the format and
 location of the file
0: jdbc:hive2://localhost:10001/> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
0: jdbc:hive2://localhost:10001/> LINES TERMINATED BY '\n'
0: jdbc:hive2://localhost:10001/> STORED AS TEXTFILE
0: jdbc:hive2://localhost:10001/> LOCATION 'abfs://102753762filesystem@102753762
datalake.dfs.core.windows.net/accidents/data';
No rows affected (0.775 seconds)
0: jdbc:hive2://localhost:10001/>
0: jdbc:hive2://localhost:10001/> -- Drop the accidents_in_hive table if it exis
ts
0: jdbc:hive2://localhost:10001/> DROP TABLE accidents_in_hive;
No rows affected (0.455 seconds)
0: jdbc:hive2://localhost:10001/> -- Create the accidents_in_hive table and popu
late it with data
```



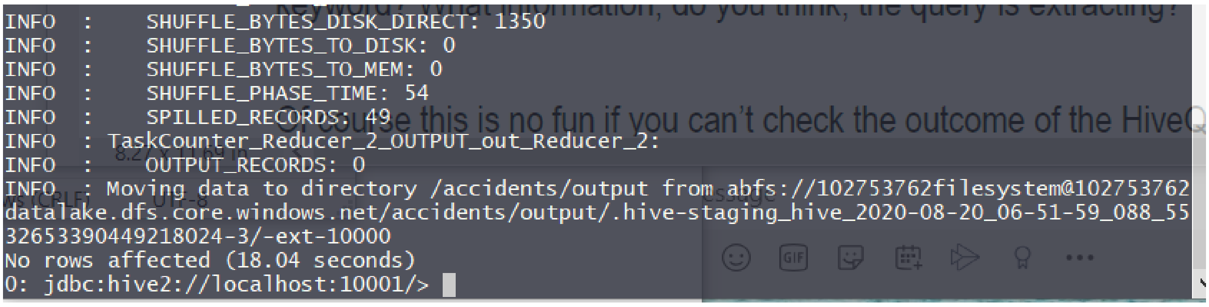
```
sshuser@hn0-s10275:~$ beeline -u 'jdbc:hive2://localhost:10001/;transportMode=http'
Connecting to jdbc:hive2://localhost:10001/;transportMode=http
Connected to: Apache Hive (version 1.2.1000.2.6.5.3027-5)
Driver: Hive JDBC (version 1.2.1000.2.6.5.3027-5)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 1.2.1000.2.6.5.3027-5 by Apache Hive
0: jdbc:hive2://localhost:10001/> INSERT OVERWRITE DIRECTORY '/accidents/output' ROW FORMAT
DELIMITED FIELDS TERMINATED BY '\t' SELECT regexp_replace(day_week_description, '''', ''), s
um(no_of_vehicles) FROM accidents_in_hive WHERE no_of_vehicles IS NOT NULL GROUP BY day_week
_description;
INFO  : Tez session hasn't been created yet. Opening session
DEBUG : Adding local resource: scheme: "hdfs" host: "mycluster" port: -1 file: "/tmp/hive/hi
ve/_tez_session_dir/17291fad-5a0a-4d2f-8ae8-77f931855d1e/hive-hcatalog-core.jar"
INFO  : Dag name: INSERT OVERWRITE DIRE...day_week_description(Stage-1)
DEBUG : DagInfo: {"context":"Hive","description":"INSERT OVERWRITE DIRECTORY '/accidents/out
put' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\\t' SELECT regexp_replace(day_week_descript
ion, '''', ''), sum(no_of_vehicles) FROM accidents_in_hive WHERE no_of_vehicles IS NOT NULL
GROUP BY day_week_description"}
DEBUG : Setting Tez DAG access for queryId=hive_20200820065159_e6fe1a84-60ad-435f-b6ab-3191d
bd616d5 with viewAclString=*, modifyStr=anonymous,hive
INFO  : Status: Running (Executing on YARN cluster with App id application_1597905285696_000
2)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      7          7        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.30 s
--------------------------------------------------------------------------------
INFO  : Status: DAG finished successfully in 11.08 seconds
INFO  :
INFO  : Query Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  : OPERATION                            DURATION
INFO  : ----------------------------------------------------------------------------------------------
INFO  : Compile Query                           0.51s
INFO  : Prepare Plan                            5.32s
INFO  : Submit Plan                             0.49s
INFO  : Start DAG                               0.39s
INFO  : Run DAG                                11.08s
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : Task Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
------------------------------
INFO  :   VERTICES  TOTAL_TASKS  FAILED_ATTEMPTS  KILLED_TASKS   DURATION(ms)  CPU_TIME(ms)
 GC_TIME(ms)  INPUT_RECORDS  OUTPUT_RECORDS
INFO  : ----------------------------------------------------------------------------------------------
------------------------------
INFO  :      Map 1            7                0             0        5896.00        40,170
      1,572        540,253               49
INFO  :   Reducer 2           1                0             0         497.00         1,070
```



```
INFO  :  Reducer 2          1          0           0       497.00       1,070
          66        49        0
INFO  : ---------------------------------------------------------------------------------
-------
INFO  :
INFO  : org.apache.tez.common.counters.DAGCounter:
INFO  :    NUM_SUCCEEDED_TASKS: 8
INFO  :    TOTAL_LAUNCHED_TASKS: 8
INFO  :    RACK_LOCAL_TASKS: 7
INFO  :    AM_CPU_MILLISECONDS: 5080
INFO  :    AM_GC_TIME_MILLIS: 0
INFO  : File System Counters:
INFO  :    ABFS_BYTES_READ: 111128065
INFO  :    ABFS_BYTES_WRITTEN: 120
INFO  :    FILE_BYTES_READ: 5060
INFO  :    FILE_BYTES_WRITTEN: 1910
INFO  : org.apache.tez.common.counters.TaskCounter:
INFO  :    REDUCE_INPUT_GROUPS: 7
INFO  :    REDUCE_INPUT_RECORDS: 49
INFO  :    COMBINE_INPUT_RECORDS: 0
INFO  :    SPILLED_RECORDS: 98
INFO  :    NUM_SHUFFLED_INPUTS: 21
INFO  :    NUM_SKIPPED_INPUTS: 0
INFO  :    NUM_FAILED_SHUFFLE_INPUTS: 0
INFO  :    MERGED_MAP_OUTPUTS: 21
INFO  :    GC_TIME_MILLIS: 1638
INFO  :    CPU_MILLISECONDS: 41240
INFO  :    PHYSICAL_MEMORY_BYTES: 8472494080
INFO  :    VIRTUAL_MEMORY_BYTES: 22842527744
INFO  :    COMMITTED_HEAP_BYTES: 8472494080
INFO  :    INPUT_RECORDS_PROCESSED: 540253
INFO  :    INPUT_SPLIT_LENGTH_BYTES: 111128065
INFO  :    OUTPUT_RECORDS: 49
INFO  :    OUTPUT_BYTES: 938
INFO  :    OUTPUT_BYTES_WITH_OVERHEAD: 1162
INFO  :    OUTPUT_BYTES_PHYSICAL: 1350
INFO  :    ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :    ADDITIONAL_SPILLS_BYTES_READ: 1350
INFO  :    ADDITIONAL_SPILL_COUNT: 0
INFO  :    SHUFFLE_CHUNK_COUNT: 7
INFO  :    SHUFFLE_BYTES: 1350
INFO  :    SHUFFLE_BYTES_DECOMPRESSED: 1162
INFO  :    SHUFFLE_BYTES_TO_MEM: 0
INFO  :    SHUFFLE_BYTES_TO_DISK: 0
INFO  :    SHUFFLE_BYTES_DISK_DIRECT: 1350
INFO  :    NUM_MEM_TO_DISK_MERGES: 0
INFO  :    NUM_DISK_TO_DISK_MERGES: 0
INFO  :    SHUFFLE_PHASE_TIME: 54
INFO  :    MERGE_PHASE_TIME: 88
INFO  :    FIRST_EVENT_RECEIVED: 17
INFO  :    LAST_EVENT_RECEIVED: 33
INFO  : HIVE:
INFO  :    CREATED_FILES: 1
INFO  :    DESERIALIZE_ERRORS: 0
```



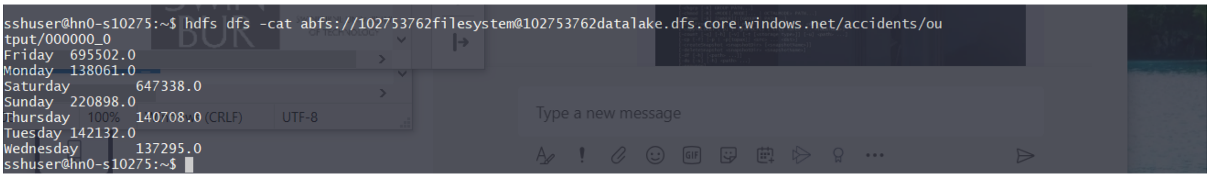
```
INFO  :  Reducer 2              1              0              0          497.00          1,070
   66              49              0
INFO  : ----------------------------------------------------------------------------------------
----------
INFO  :
INFO  : org.apache.tez.common.counters.DAGCounter:
INFO  :    NUM_SUCCEEDED_TASKS: 8
INFO  :    TOTAL_LAUNCHED_TASKS: 8
INFO  :    RACK_LOCAL_TASKS: 7
INFO  :    AM_CPU_MILLISECONDS: 5080
INFO  :    AM_GC_TIME_MILLIS: 0
INFO  : File System Counters:
INFO  :    ABFS_BYTES_READ: 111128065
INFO  :    ABFS_BYTES_WRITTEN: 120
INFO  :    FILE_BYTES_READ: 5060
INFO  :    FILE_BYTES_WRITTEN: 1910
INFO  : org.apache.tez.common.counters.TaskCounter:
INFO  :    REDUCE_INPUT_GROUPS: 7
INFO  :    REDUCE_INPUT_RECORDS: 49
INFO  :    COMBINE_INPUT_RECORDS: 0
INFO  :    SPILLED_RECORDS: 98
INFO  :    NUM_SHUFFLED_INPUTS: 21
INFO  :    NUM_SKIPPED_INPUTS: 0
INFO  :    NUM_FAILED_SHUFFLE_INPUTS: 0
INFO  :    MERGED_MAP_OUTPUTS: 21
INFO  :    GC_TIME_MILLIS: 1638
INFO  :    CPU_MILLISECONDS: 41240
INFO  :    PHYSICAL_MEMORY_BYTES: 8472494080
INFO  :    VIRTUAL_MEMORY_BYTES: 22842527744
INFO  :    COMMITTED_HEAP_BYTES: 8472494080
INFO  :    INPUT_RECORDS_PROCESSED: 540253
INFO  :    INPUT_SPLIT_LENGTH_BYTES: 111128065
INFO  :    OUTPUT_RECORDS: 49
INFO  :    OUTPUT_BYTES: 938
INFO  :    OUTPUT_BYTES_WITH_OVERHEAD: 1162
INFO  :    OUTPUT_BYTES_PHYSICAL: 1350
INFO  :    ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :    ADDITIONAL_SPILLS_BYTES_READ: 1350
INFO  :    ADDITIONAL_SPILL_COUNT: 0
INFO  :    SHUFFLE_CHUNK_COUNT: 7
INFO  :    SHUFFLE_BYTES: 1350
INFO  :    SHUFFLE_BYTES_DECOMPRESSED: 1162
INFO  :    SHUFFLE_BYTES_TO_MEM: 0
INFO  :    SHUFFLE_BYTES_TO_DISK: 0
INFO  :    SHUFFLE_BYTES_DISK_DIRECT: 1350
INFO  :    NUM_MEM_TO_DISK_MERGES: 0
INFO  :    NUM_DISK_TO_DISK_MERGES: 0
INFO  :    SHUFFLE_PHASE_TIME: 54
INFO  :    MERGE_PHASE_TIME: 88
INFO  :    FIRST_EVENT_RECEIVED: 17
INFO  :    LAST_EVENT_RECEIVED: 33
INFO  : HIVE:
INFO  :    CREATED_FILES: 1
INFO  :    DESERIALIZE_ERRORS: 0
```



```
INFO  :    SHUFFLE_BYTES_DISK_DIRECT: 1350
INFO  :    SHUFFLE_BYTES_TO_DISK: 0
INFO  :    SHUFFLE_BYTES_TO_MEM: 0
INFO  :    SHUFFLE_PHASE_TIME: 54
INFO  :    SPILLED_RECORDS: 49
INFO  : TaskCounter_Reducer_2_OUTPUT_out_Reducer_2:
INFO  :    OUTPUT_RECORDS: 0
INFO  : Moving data to directory /accidents/output from abfs://102753762filesystem@102753762
datalake.dfs.core.windows.net/accidents/output/.hive-staging_hive_2020-08-20_06-51-59_088_55
32653390449218024-3/-ext-10000
No rows affected (18.04 seconds)
0: jdbc:hive2://localhost:10001/>
```

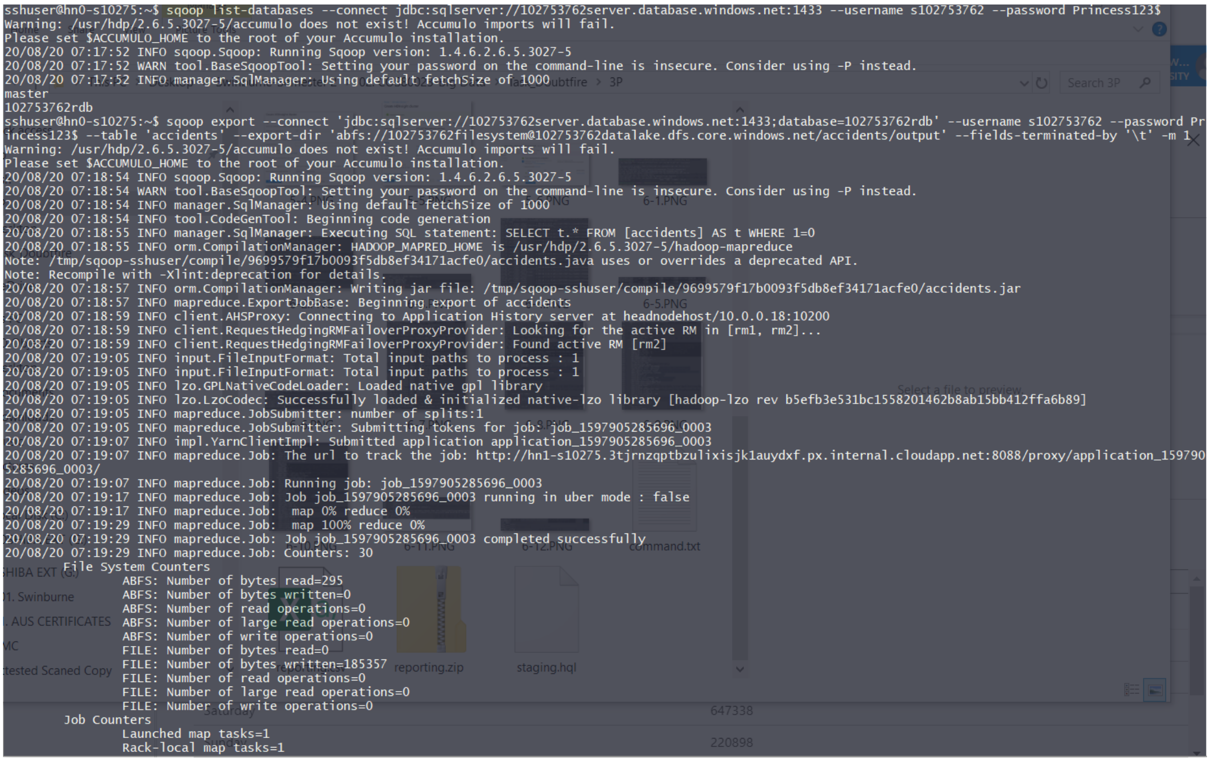
Question 3: What is the 'accidents_in_hive' object that comes after the 'FROM' keyword? What information, do you think, the query is extracting?

Answer: "accidents_in_hive" is the table created in hive by running the staging.hql script. The query is extracting data from "accident_in_hive" table to show no of vehicle involved in accident as per weekdays.

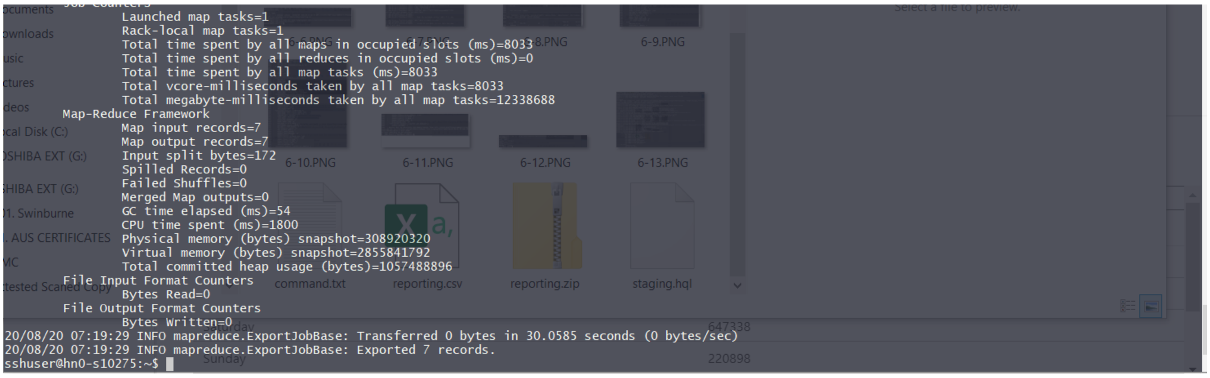


```
sshuser@hn0-s10275:~$ hdfs dfs -cat abfs://102753762filesystem@102753762datalake.dfs.core.windows.net/accidents/ou
tput/000000_0
Friday  695502.0
Monday  138061.0
Saturday        647338.0
Sunday  220898.0
Thursday        140708.0
Tuesday 142132.0
Wednesday       137295.0
sshuser@hn0-s10275:~$
```

Step Eight: Loading the data



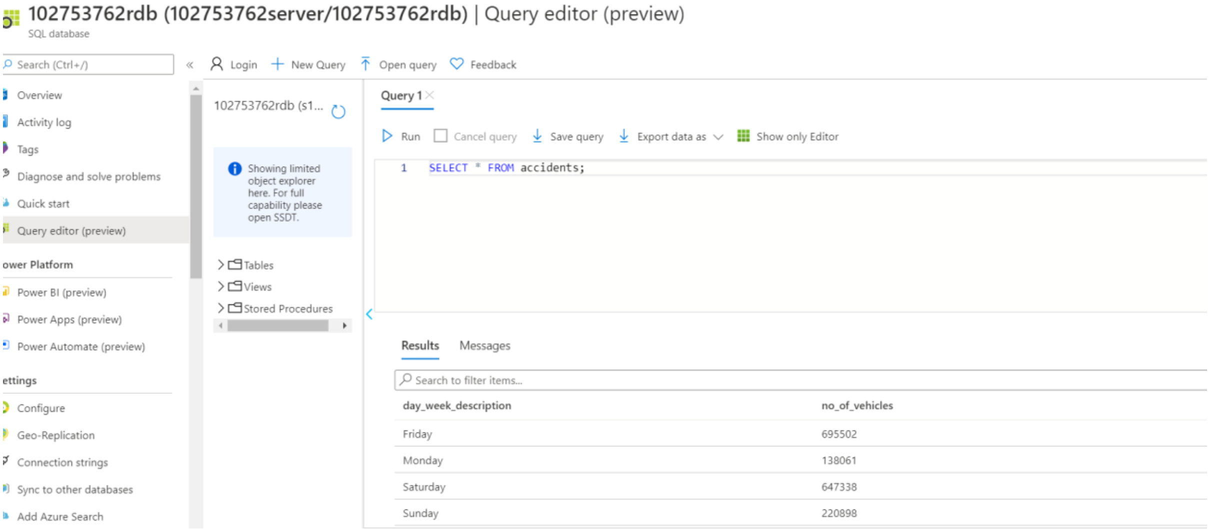
```
sshuser@hn0-s10275:~$ sqoop list-databases --connect jdbc:sqlserver://102753762server.database.windows.net:1433 --username s102753762 --password Princess123$
Warning: /usr/hdp/2.6.5.3027-5/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/08/20 07:17:52 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6.2.6.5.3027-5
20/08/20 07:17:52 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/08/20 07:17:52 INFO manager.SqlManager: Using default fetchSize of 1000
master
102753762rdb
sshuser@hn0-s10275:~$ sqoop export --connect 'jdbc:sqlserver://102753762server.database.windows.net:1433;database=102753762rdb' --username s102753762 --password Pr
incess123$ --table 'accidents' --export-dir 'abfs://102753762filesystem@102753762datalake.dfs.core.windows.net/accidents/output' --fields-terminated-by '\t' -m 1
Warning: /usr/hdp/2.6.5.3027-5/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/08/20 07:18:54 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6.2.6.5.3027-5
20/08/20 07:18:54 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/08/20 07:18:54 INFO manager.SqlManager: Using default fetchSize of 1000
20/08/20 07:18:54 INFO tool.CodeGenTool: Beginning code generation
20/08/20 07:18:55 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM [accidents] AS t WHERE 1=0
20/08/20 07:18:55 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/hdp/2.6.5.3027-5/hadoop-mapreduce
Note: /tmp/sqoop-sshuser/compile/9699579f17b0093f5db8ef34171acfe0/accidents.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/08/20 07:18:57 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-sshuser/compile/9699579f17b0093f5db8ef34171acfe0/accidents.jar
20/08/20 07:18:57 INFO mapreduce.ExportJobBase: Beginning export of accidents
20/08/20 07:18:59 INFO client.AHSProxy: Connecting to Application History server at headnodehost/10.0.0.18:10200
20/08/20 07:18:59 INFO client.RequestHedgingRMFailoverProxyProvider: Looking for the active RM in [rm1, rm2]...
20/08/20 07:18:59 INFO client.RequestHedgingRMFailoverProxyProvider: Found active RM [rm2]
20/08/20 07:19:05 INFO input.FileInputFormat: Total input paths to process : 1
20/08/20 07:19:05 INFO input.FileInputFormat: Total input paths to process : 1
20/08/20 07:19:05 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
20/08/20 07:19:05 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev b5efb3e531bc1558201462b8ab15bb412ffa6b89]
20/08/20 07:19:05 INFO mapreduce.JobSubmitter: number of splits:1
20/08/20 07:19:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1597905285696_0003
20/08/20 07:19:07 INFO impl.YarnClientImpl: Submitted application application_1597905285696_0003
20/08/20 07:19:07 INFO mapreduce.Job: The url to track the job: http://hn1-s10275.3tjrnzqptbzulixisjk1auydxf.px.internal.cloudapp.net:8088/proxy/application_159790
5285696_0003/
20/08/20 07:19:07 INFO mapreduce.Job: Running job: job_1597905285696_0003
20/08/20 07:19:17 INFO mapreduce.Job: Job job_1597905285696_0003 running in uber mode : false
20/08/20 07:19:17 INFO mapreduce.Job:  map 0% reduce 0%
20/08/20 07:19:29 INFO mapreduce.Job:  map 100% reduce 0%
20/08/20 07:19:29 INFO mapreduce.Job: Job job_1597905285696_0003 completed successfully
20/08/20 07:19:29 INFO mapreduce.Job: Counters: 30
        File System Counters
                ABFS: Number of bytes read=295
                ABFS: Number of bytes written=0
                ABFS: Number of read operations=0
                ABFS: Number of large read operations=0
                ABFS: Number of write operations=0
                FILE: Number of bytes read=0
                FILE: Number of bytes written=185357
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Rack-local map tasks=1
```



```
                Launched map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=8033
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=8033
                Total vcore-milliseconds taken by all map tasks=8033
                Total megabyte-milliseconds taken by all map tasks=12338688
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=172
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=54
                CPU time spent (ms)=1800
                Physical memory (bytes) snapshot=308920320
                Virtual memory (bytes) snapshot=2855841792
                Total committed heap usage (bytes)=1057488896
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
20/08/20 07:19:29 INFO mapreduce.ExportJobBase: Transferred 0 bytes in 30.0585 seconds (0 bytes/sec)
20/08/20 07:19:29 INFO mapreduce.ExportJobBase: Exported 7 records.
sshuser@hn0-s10275:~$
```

Question 4: On what weekday do the most accidents happen? How many vehicles are involved? Document the answer with a screenshot.

Answer: Most accident happened on Friday and number of vehicles involved were 695502.

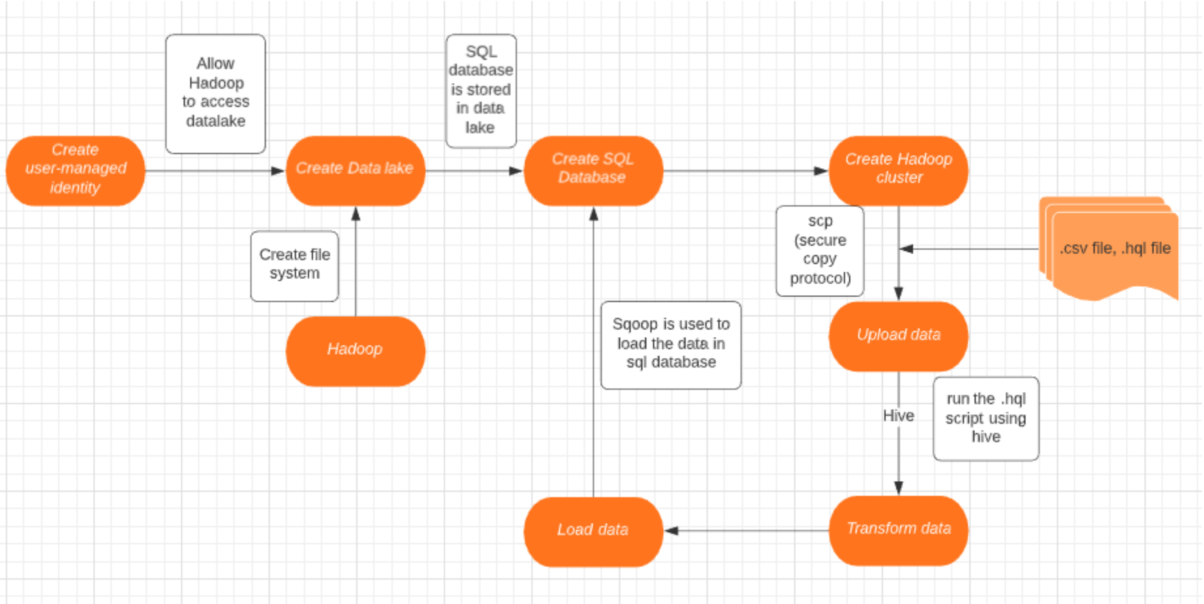


```
1   SELECT * FROM accidents;
```

day_week_description	no_of_vehicles
Friday	695502
Monday	138061
Saturday	647338
Sunday	220898

Question 5: What steps and tools were involved in this process? Draw a diagram that shows the tools and files used. You can draw in Powerpoint or by hand and take a photo to document the workflow.

Answer:


Create user-managed identity
Allow Hadoop to access datalake
Create Data lake
Create file system
Hadoop
SQL database is stored in data lake
Create SQL Database
Sqoop is used to load the data in sql database
Create Hadoop cluster
scp (secure copy protocol)
.csv file, .hql file
Upload data
Hive
run the .hql script using hive
Transform data
Load data

Reference

01. IBM 2020, *What is a Data Warehouse?*, viewed 22 August, 2020, <https://www.ibm.com/cloud/learn/data-warehouse#:~:text=A%20data%20warehouse%20is%20a,AI)%2C%20and%20machine%20learning.&text=Find%20out%20more%20about%20data%20warehouse%20solutions%20from%20IBM.>.
02. Redhat 2020, *What is a data lake?*, viewed 22 August, 2020, <https://www.redhat.com/en/topics/data-storage/what-is-a-data-lake>.

03. Campbell, C 2015, *Top Five Differences between Data Lakes and Data Warehouses*, viewed 22 August, 2020, <https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses>.