

# ETL in Hortonworks Sandbox on Azure

Lab Guide

March 2016





## Table of Contents

|  |    |
|--|----|
| ETL in Hortonworks Sandbox on Azure .....                                  | 1  |
| Overview.....  | 2  |
| Requirements.....  | 2  |
| Technical Support .....  | 2  |
| Exercise 1: Environment Setup .....  | 3  |
| Exercise 2: Create an Azure SQL Database .....                             | 4  |
| Exercise 3: Create a Virtual Machine with Hortonworks Sandbox.....         | 7  |
| Exercise 4: Configure your Azure SQL Database for remote connectivity..... | 18 |
| Exercise 5: Transfer data using Sqoop.....                                 | 21 |
| Validate Lab Completion.....   | 25 |
| Lab Summary .....  | 26 |



## Overview

In this lab, you will create a Hortonworks Sandbox virtual machine and an Azure SQL Database from the Azure Marketplace. You will then extract data from Azure SQL Database into the Hortonworks Sandbox using Sqoop. You will then load data into a Hive table in Hadoop.

### Requirements

- A Microsoft Azure Subscription, you can create a free trial here <http://Azure.microsoft.com/en-us/pricing/free-trial/>, or you can use a subscription from your organization's EA agreement.
- Windows client computers will need an SSH client to complete the lab. Alternatively you may use the web based SSH client built into the Hortonworks Sandbox.
  - Git Bash with SSH client from <http://www.git-scm.com/downloads>
  - [PuTTY](#), and [AnyConnect](#) amongst others

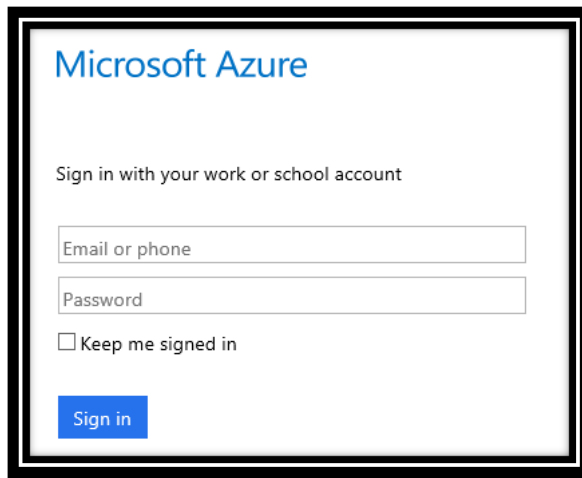
### Technical Support

Having trouble with this lab or have a question? Please contact [SuperHuman\\_Help@microsoft.com](mailto:SuperHuman_Help@microsoft.com) for technical assistance.

## Exercise 1: Environment Setup

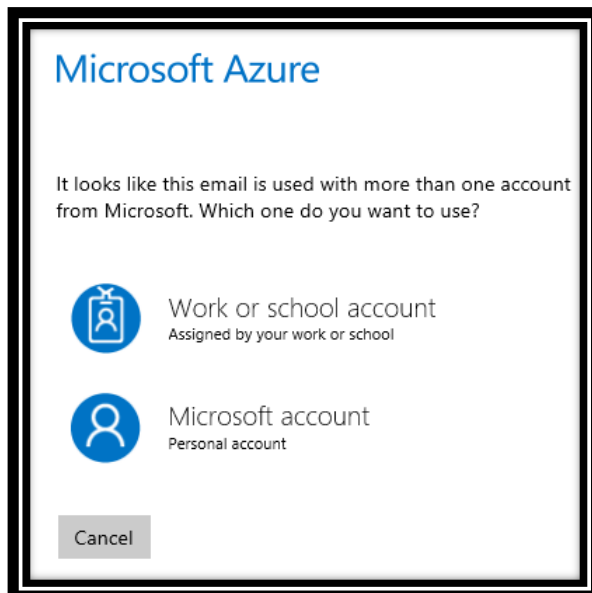
In this exercise, you will use your Microsoft or Organization account to login to the Azure preview portal to start the lab exercise.

1. Open your browser and navigate to <https://portal.azure.com/>
2. Enter the account associated with your Microsoft Azure subscription.



The screenshot shows the Microsoft Azure sign-in interface. At the top, the 'Microsoft Azure' logo is displayed. Below it, the text 'Sign in with your work or school account' is shown. There are two input fields: 'Email or phone' and 'Password'. Below these fields is a checkbox labeled 'Keep me signed in'. At the bottom left, there is a blue 'Sign in' button.

3. If your account is associated with an organization account and a Microsoft account, you may be prompted to choose which one to authenticate with for your Microsoft Azure account.

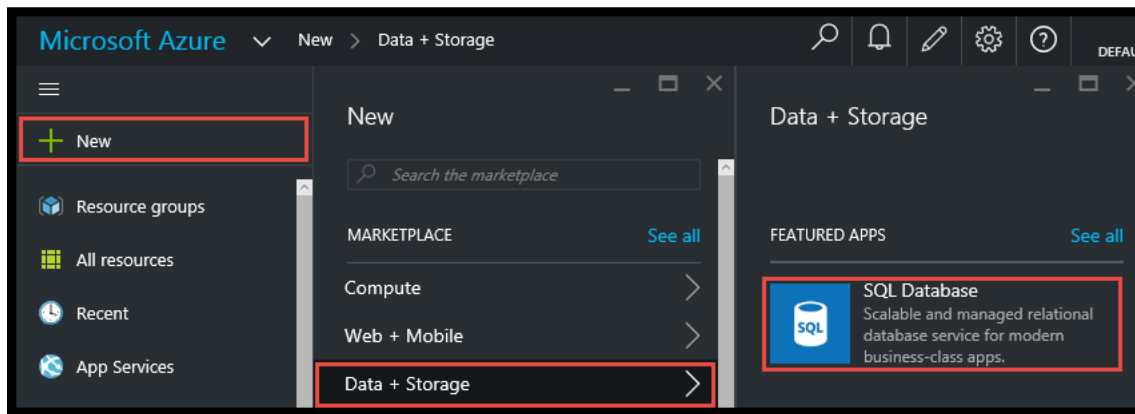


The screenshot shows the Microsoft Azure account selection interface. At the top, the 'Microsoft Azure' logo is displayed. Below it, the text 'It looks like this email is used with more than one account from Microsoft. Which one do you want to use?' is shown. There are two account options listed: 'Work or school account' (Assigned by your work or school) and 'Microsoft account' (Personal account). Each option has a corresponding icon. At the bottom left, there is a grey 'Cancel' button.

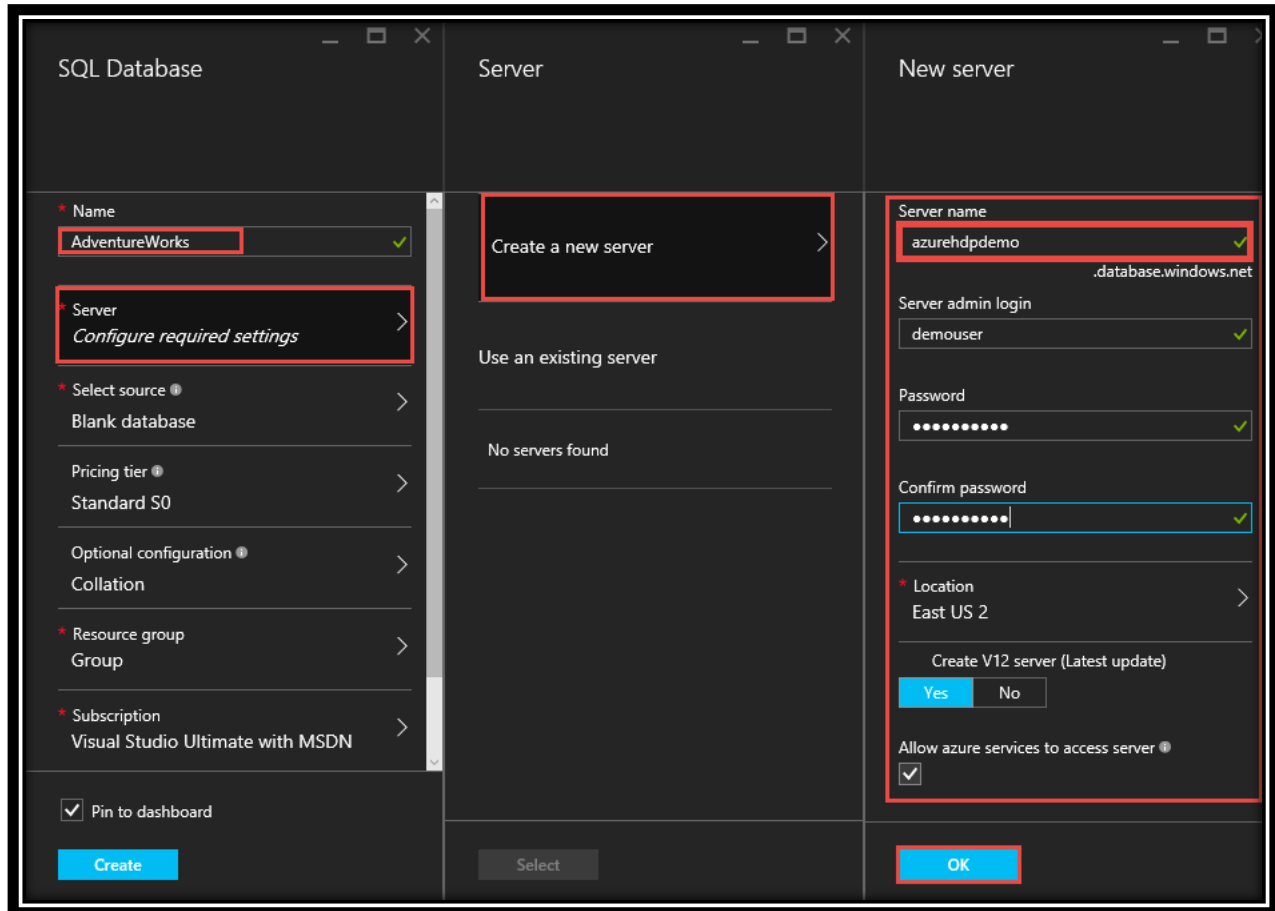
## Exercise 2: Create an Azure SQL Database

In this exercise, you will create an Azure SQL Database in Azure Marketplace.

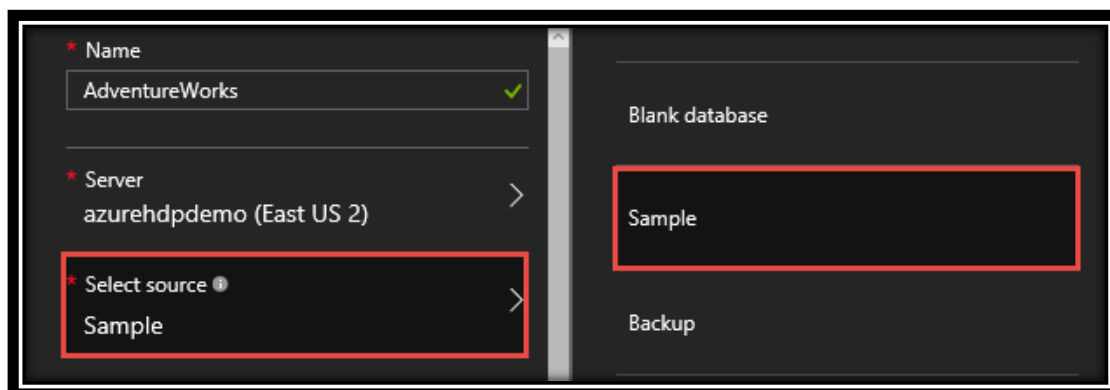
1. Click the **+New** button from the portal, then click **Data + Storage** and choose **SQL Database** from the Marketplace.

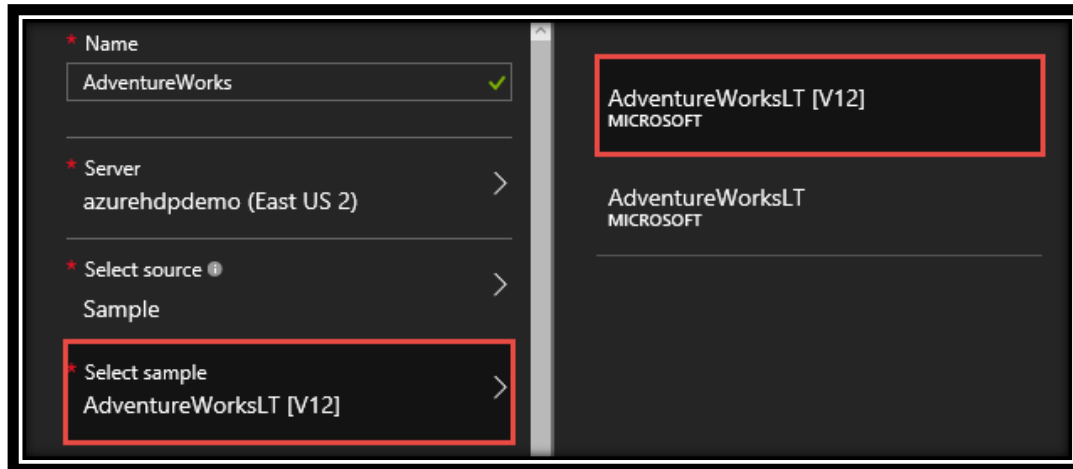


2. The Azure portal will open the **SQL Database** blade. Type *AdventureWorks* into the **Name** box. Choose **Server**, and then **Create a new server**. Specify the following **New server** configurations and click **OK**.
  - Server name: *<Unique server name for your Azure SQL Database>*
  - Server admin login: demouser
  - Password: demo@pass1
  - Location: *<Location nearest to you>*
  - Create V12 server: Yes
  - Allow Azure services to access server: Checked

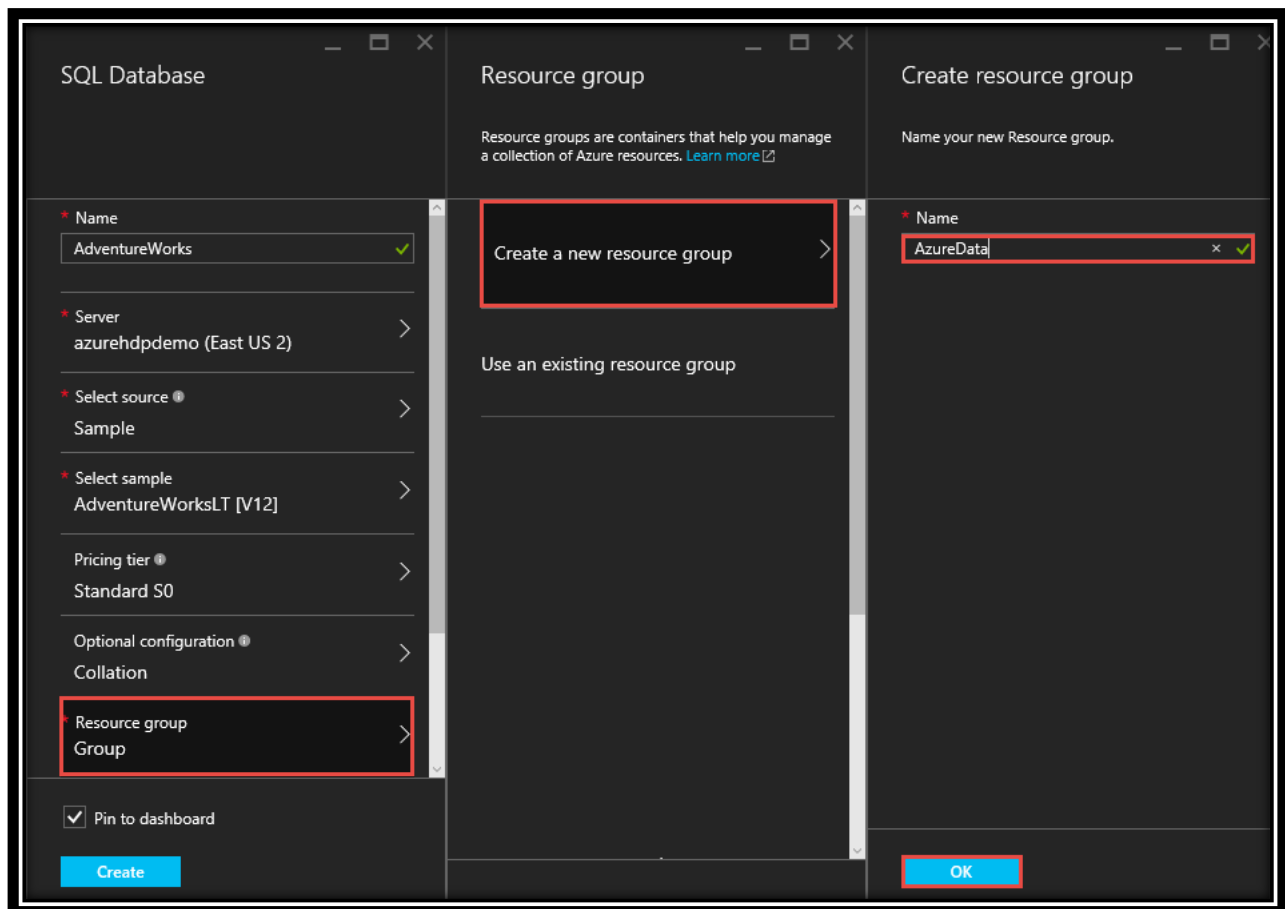


3. **Note the server name of the SQL Database and the Server name** for later reference. The server name will be used in connection strings later in the lab.
4. Choose **Select source**, then select **Sample**. Choose **Select sample**, then select **AdventureWorksLT [V12]**





5. Choose **Resource group**, then select **Create a new resource group**. Type **AzureData** for the resource group name and click OK.



6. Verify the following SQL Database configurations and click **Create**.



- Name: AdventureWorks
- Server: *<Unique server name for you Azure SQL Database>*
- Select source: Sample
- Select sample: AdventureWorksLT [V12]
- Pricing tier: Standard S0
- Resource group: AzureData
- Subscription: *<your subscription>*

7. Now you will see that the SQL Database is being created, as per the status on portal dashboard.



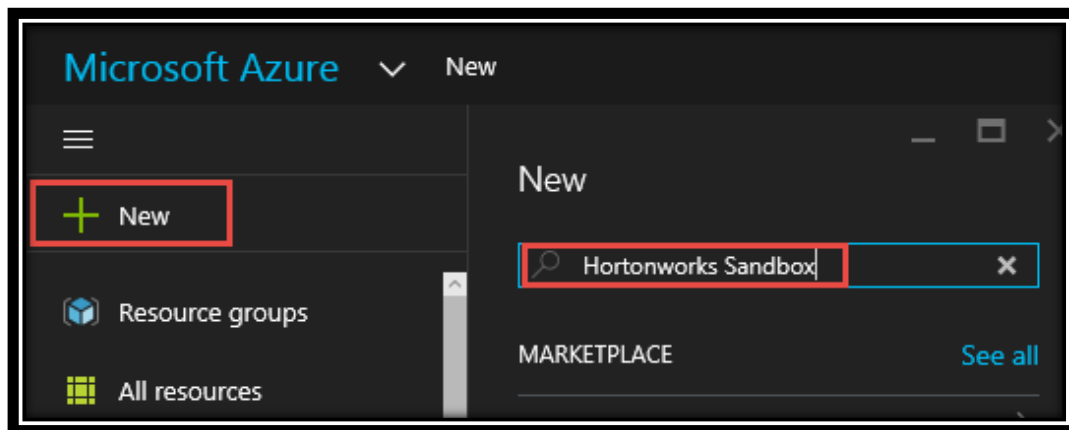
8. When the creation is complete, the status of Azure SQL Database is '**Online**'. The creation of the SQL Database will take a few minutes. You may continue to the next exercise while the SQL Database is being created.



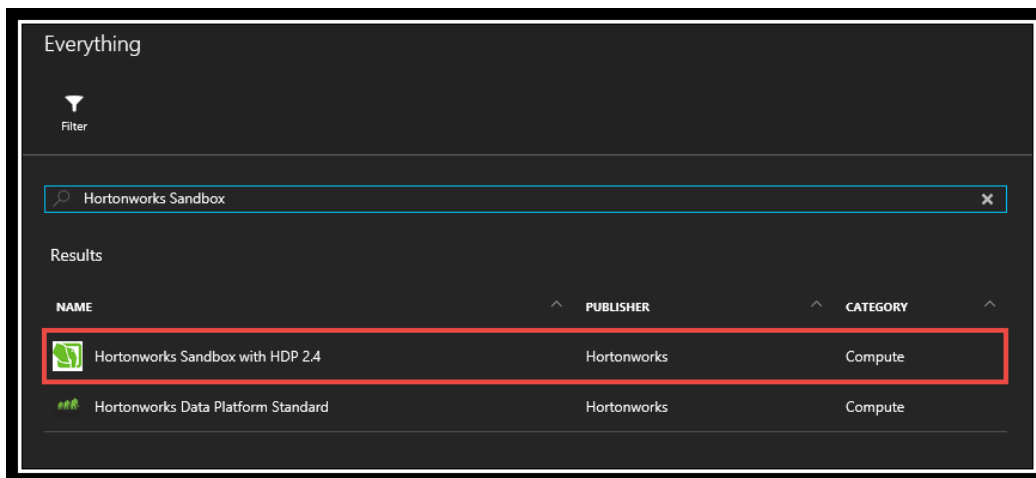
## Exercise 3: Create a Virtual Machine with Hortonworks Sandbox

In this exercise, you will create a virtual machine using the 'Hortonworks Sandbox with HDP' image available in Azure Marketplace.


1. Click the **+New** button from the portal. Type **Hortonworks Sandbox** in the search filter and press **Enter**.



2. From the search result, click **Hortonworks Sandbox with HDP**



3. In the **Select a deployment model** dropdown verify that **Resource Manager** is selected. Read the description then click the **Create** button.



## Hortonworks Sandbox with HDP 2.4

Hortonworks

**Bring Your Own License enabled.**

**Learn Hadoop**

Sandbox comes with over fifty hands-on [tutorials](#) that will guide you through the Hadoop, Spark, Storm, HBase, Kafka, Hive, Ambari and YARN; tutorials built on the experience gained from training thousands of people in our [Hortonworks University Training classes](#).

If you are new to Hadoop, HDP and the Sandbox we suggest sequence of tutorials to get started:

- [Deploying Hortonworks Sandbox on Microsoft Azure](#)
  - This tutorial will walk you through the process of setting up the Sandbox on Azure.
- [Learning the Ropes of the Hortonworks Sandbox](#) This tutorial will provide you an introduction on key user interfaces and how to get access to the various tools.
  - [Hadoop Tutorial – Getting Started with HDP](#) This tutorial will provide you an introduction to Ambari, Ambari Views, HDFS, Hive, Pig, Spark, Zeppelin and remote ODBC connectivity.

**Instructions**

Once you have deployed the Hortonworks Sandbox on Azure, navigate to `http://<hostname/ipaddress>:8888` to get started. You can get the DNS name or Ip address information by looking in the Azure portal for the newly created virtual machine and navigating to its network information.

**NOTE: If you want to log into the Ambari portal as an admin, you will need to do the following:**

1. ssh into the VM and run the following command as the root user:  
`ambari-admin-password-reset`
2. You will be prompted to enter the password. Ambari server will restart.
3. Then run command  
`ambari-agent restart`
4. You can then go to the `http://<hostname/ipaddress>:8888` which displays the splash page and click on the Ambari url. Enter the user "admin" and put the password you set previously in step 2. You

Select a deployment model ●

Resource Manager ▼

Create

4. Specify the following configuration options in the create virtual machine **Basics** blade then click .

- Name: hortonworks-sandbox-vm
- User Name: demouser
- Authentication Type: Password
- Password: demo@pass1
- Resource group: AzureData
- Location: Same location you used previously.

Basics

\* Name

hortonworks-sandbox-vm ✓

\* User name

demouser ✓

\* Authentication type

Password

SSH public key

\* Password

•••••••• ✓

Subscription

Visual Studio Ultimate with MSDN ▼

\* Resource group

AzureData ✓

Select existing

Location

East US 2 ▼

OK

- On the **Choose a size** blade, select **A5 Standard** and click **Select**.

### Choose a size

Browse the available sizes and their features

Prices presented below are estimated retail prices that include both Azure infrastructure and applicable third-party software costs. Prices do not reflect applicable discounts for your subscription and may include currency conversions.

★ Recommended | [View all](#)

| A4 Standard ★                   | A5 Standard ★                   | A6 Standard ★                   |
|---------------------------------|---------------------------------|---------------------------------|
| 8 Cores                         | 2 Cores                         | 4 Cores                         |
| 14 GB                           | 14 GB                           | 28 GB                           |
| 16 Data disks                   | 4 Data disks                    | 8 Data disks                    |
| 16x500 Max IOPS                 | 4x500 Max IOPS                  | 8x500 Max IOPS                  |
| Load balancing                  | Load balancing                  | Load balancing                  |
| Auto scale                      | Auto scale                      | Auto scale                      |
| 357.12<br>USD/MONTH (ESTIMATED) | 163.68<br>USD/MONTH (ESTIMATED) | 327.36<br>USD/MONTH (ESTIMATED) |

Select

- On the Settings blade accept the defaults and click the **OK** button.

Settings

Storage

Disk type ⓘ

Standard Premium (SSD)

\* Storage account ⓘ

(new) azuredata2937

Network

\* Virtual network ⓘ

(new) AzureData

\* Subnet ⓘ

default (10.2.0.0/24)

\* Public IP address ⓘ

(new) hortonworks-sandbox-vm

\* Network security group ⓘ

(new) hortonworks-sandbox-vm

OK

- On the Summary blade, verify your configuration, and click **OK**

Summary

*i* Validation passed

Basics

|                |                                  |
|----------------|----------------------------------|
| Subscription   | Visual Studio Ultimate with MSDN |
| Resource group | (new) AzureData                  |
| Location       | East US 2                        |

Settings

|                             |                              |
|-----------------------------|------------------------------|
| Computer name               | hortonworks-sandbox-vm       |
| User name                   | demouser                     |
| Size                        | Standard A5                  |
| Disk type                   | Standard                     |
| Storage account             | (new) azuredata2937          |
| Virtual network             | (new) AzureData              |
| Subnet                      | (new) default (10.2.0.0/24)  |
| Public IP address           | (new) hortonworks-sandbox-vm |
| Network security group      | (new) hortonworks-sandbox-vm |
| Availability set            | None                         |
| Diagnostics                 | Enabled                      |
| Diagnostics storage account | (new) azuredata2937          |

OK

- On the Purchase blade, review the details and click **Purchase**.

Purchase

### Offer details

|  |   |
|--|---|
| Hortonworks Sandbox<br>by Hortonworks<br><a href="#">Terms of use</a> and <a href="#">privacy policy</a> | 0.0000 USD/hr *   |
| Standard A5<br>by Microsoft<br><a href="#">Terms of use</a> and <a href="#">privacy policy</a>           | 0.2200 USD/hr +<br><a href="#">Pricing for other VM sizes</a> |

\* **Marketplace Offering:** May not be purchased using Microsoft subscription credits or monetary commitment funds and does not participate in discounts. These purchases are billed separately.

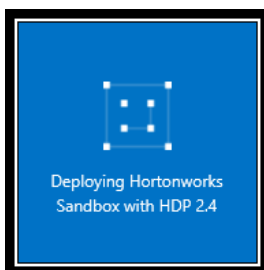
+ **Azure Resource:** May be purchased using Microsoft subscription credits or monetary commitment funds and participates in discounts. Prices presented are retail prices and may not reflect discounts associated with your subscription.

### Terms of use

By clicking "Purchase", I (a) agree to the legal terms and privacy statement(s) associated with each Marketplace offering above, (b) authorize Microsoft to charge or bill my current payment method for the fees associated with my use of the offering(s), including applicable taxes, with the same billing frequency as my Azure subscription, until I discontinue use of the offering(s), and (c) agree that Microsoft may share my contact information and transaction details with the seller(s) of the offering (s). Microsoft does not provide rights for third-party products or services. See the [Azure Marketplace Terms](#) for additional terms.

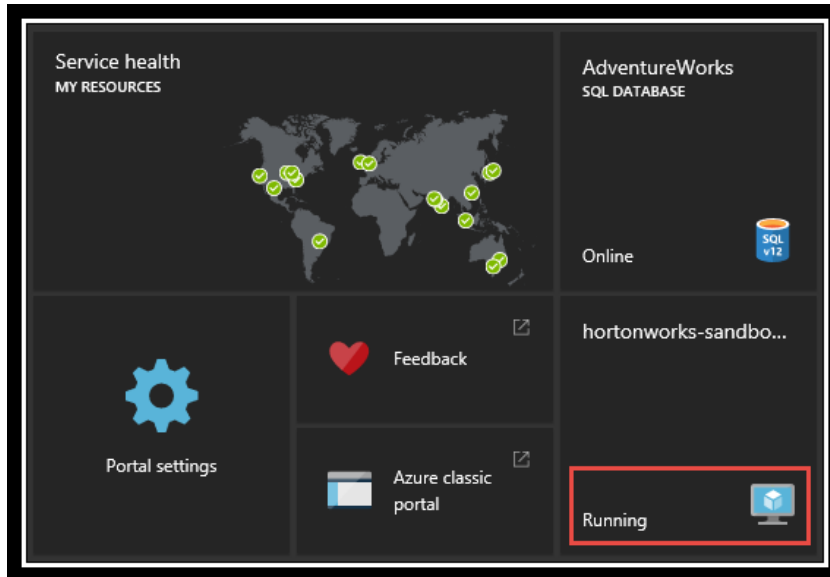
Purchase

- Now you will see that the new virtual machine is creating, as per the status on portal dashboard.

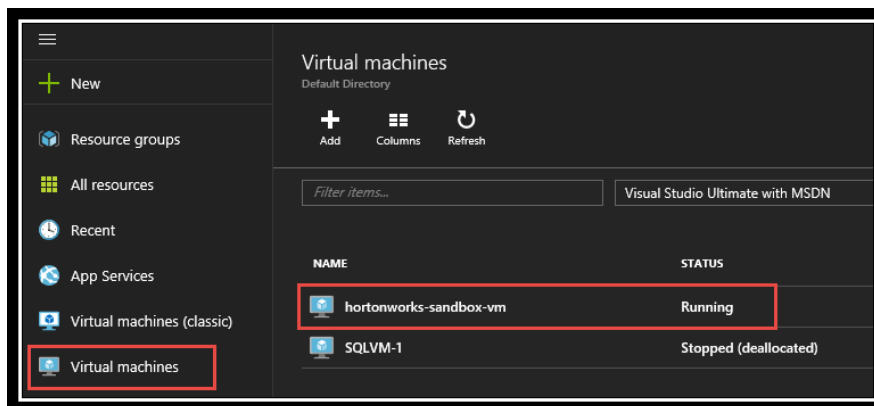


- Wait until the status of created virtual machine is '**Running**'. Note: It may take 10-15 minutes for the virtual machine to complete provisioning.

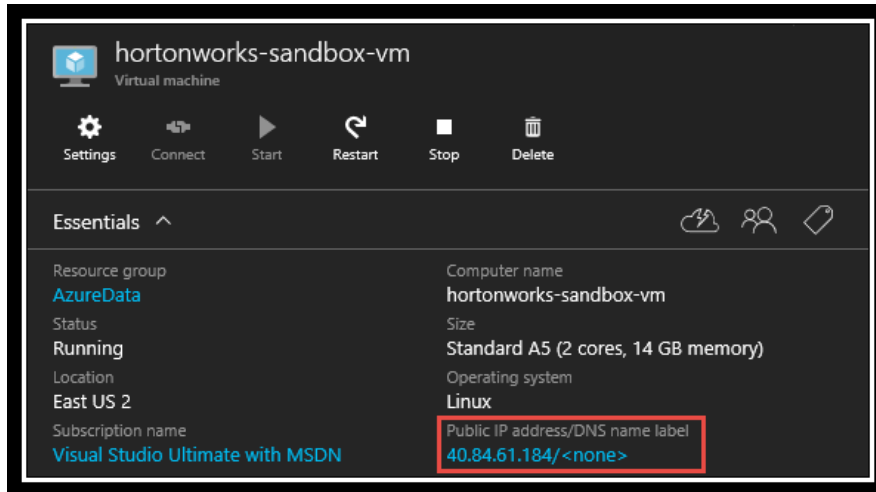




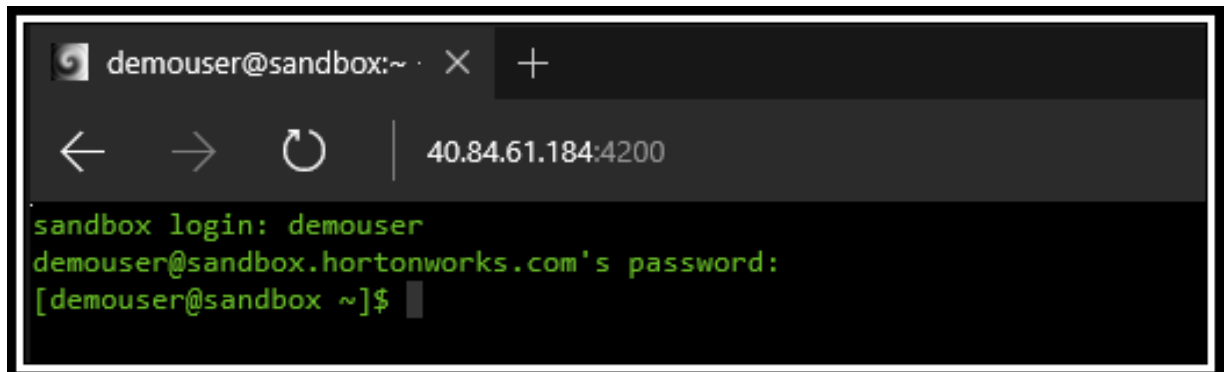
11. If the portal does not refresh, click **Virtual Machines** to see the latest status of the virtual machine. Wait until the status turns to '**Running**'.



12. Once the virtual machine is in status '**Running**', click on the **virtual machine name** to drill to the details.
13. **Note the Virtual IP Address of the Hortonworks** Sandbox virtual machine as you will reference it in the next step and future steps.



14. Launch a browser and navigate to the virtual IP address of the virtual machine using port 4200 ([http://\[Virtual IP of the Hortonworks VM\]:4200](http://[Virtual IP of the Hortonworks VM]:4200)). This will connect you to the built-in SSH client on your Hortonworks Sandbox VM. Login by entering **demouser** for the **login** and **demo@pass1** for the password.



15. Enter the following command download and extract the SQL Server JDBC driver.

```
curl -L 'https://download.microsoft.com/download/0/2/A/02AAE597-3865-456C-AE7F-613F99F850A8/sqljdbc_4.2.6420.100_enu.tar.gz' | tar xz
```

**NOTE: You can copy this command into your clipboard and then paste it into the terminal window open in your browser. Simply right-click anywhere in the browser terminal window and select Paste from Browser and then paste the copied text the box provided. This will copy the text into the terminal then click Enter on your keyboard to execute the command.**

16. In the next step, copy the extracted SQL JDBC drivers to /usr/hdp/current/sqoop-client/lib. Note that use of the sudo command will require you to reenter your password.

```
sudo cp sqljdbc_4.2/enu/*.jar /usr/hdp/current/sqoop-client/lib
```

The Shell will then resemble the next screen.

```

demouser@sandbox:~ · X +
104.208.155.191:4200
sandbox login: demouser
demouser@sandbox.hortonworks.com's password:
Last login: Tue Dec 22 00:08:38 2015 from 172.16.0.4
[demouser@sandbox ~]$ curl -L 'https://download.microsoft.com/download/0/2/A/02AAE597-386
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total     Spent    Left     Speed
100 2571k  100 2571k    0     0  3812k      0  --:--:-- --:--:-- --:--:--  7723k
[demouser@sandbox ~]$ sudo cp sqljdbc_4.2/enu/*.jar /usr/hdp/current/sqoop-client/lib
[sudo] password for demouser:
[demouser@sandbox ~]$

```

17. Execute the following two commands to navigate to /usr/hdp/current/sqoop-client/lib to verify that the JDBC drivers are installed.

```
cd /usr/hdp/current/sqoop-client/lib
ls -l sqljdbc*
```

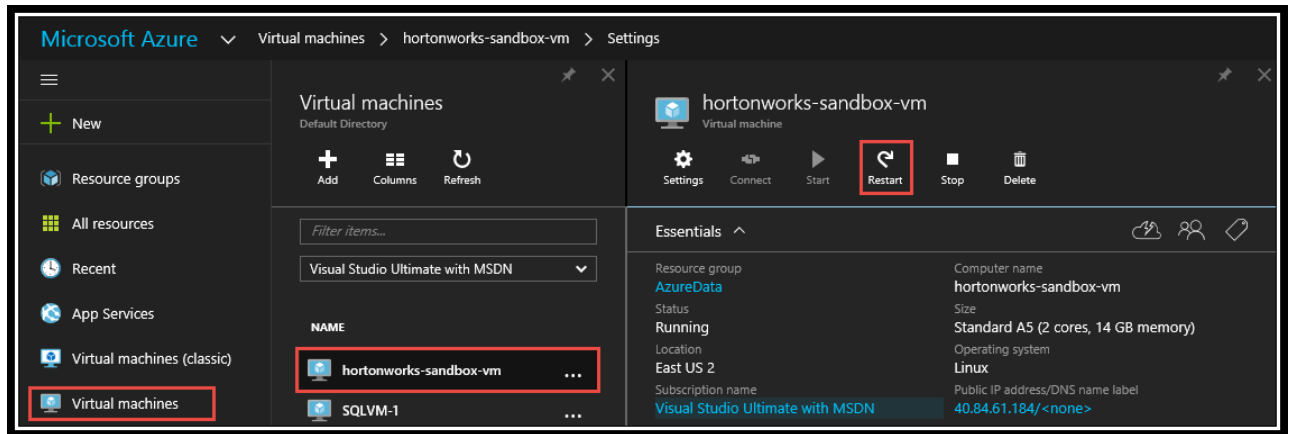
The screen will resemble below.

```

demouser@sandbox:/u: X +
104.208.155.191:4200
[demouser@sandbox ~]$ cd /usr/hdp/current/sqoop-client/lib
[demouser@sandbox lib]$ ls -l sqljdbc*
-rw-r--r-- 1 root root 654891 2015-12-22 00:15 sqljdbc41.jar
-rw-r--r-- 1 root root 653574 2015-12-22 00:15 sqljdbc42.jar
-rw-r--r-- 1 root root 585020 2015-12-22 00:15 sqljdbc4.jar
-rw-r--r-- 1 root root 563935 2015-12-22 00:15 sqljdbc.jar
[demouser@sandbox lib]$

```

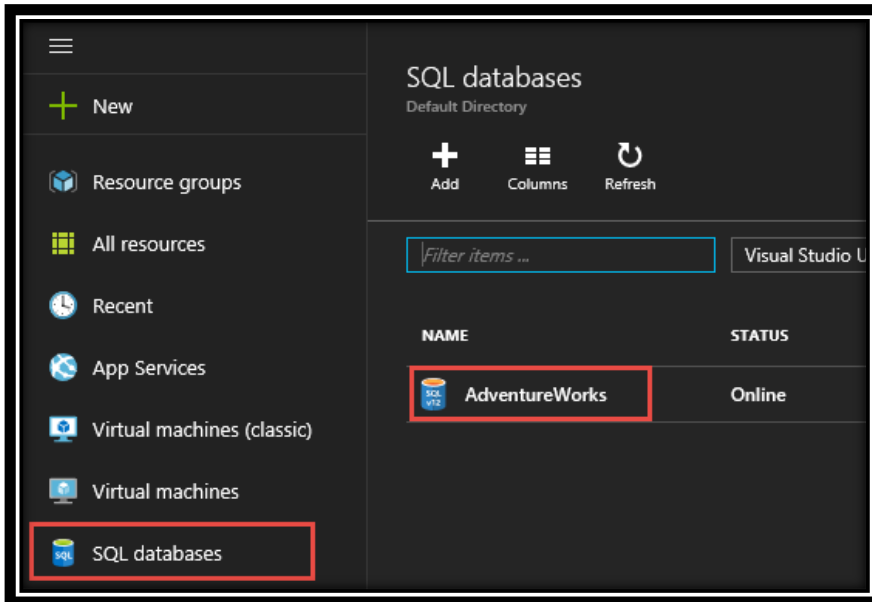
- Restart the Hortonworks Sandbox VM so that the new driver will become available. In the Portal click, **Virtual machines** and then click the **hortonworks-sandbox-vm**. When the Blade for the VM loads click **Restart**.



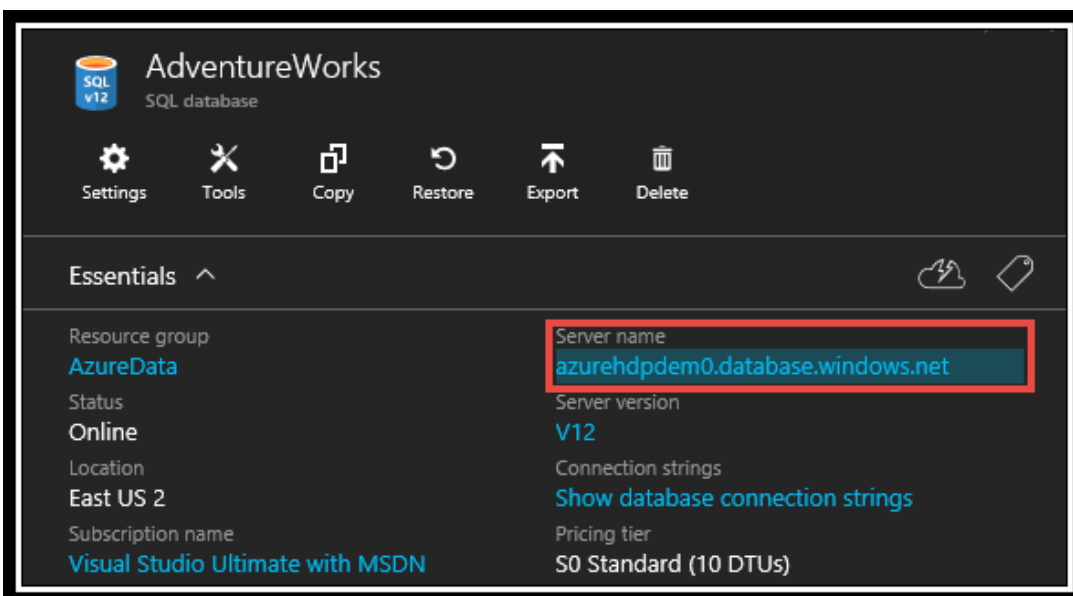
## Exercise 4: Configure your Azure SQL Database for remote connectivity.

In this exercise, you will configure the firewall for your Azure SQL Database through the Azure Portal.

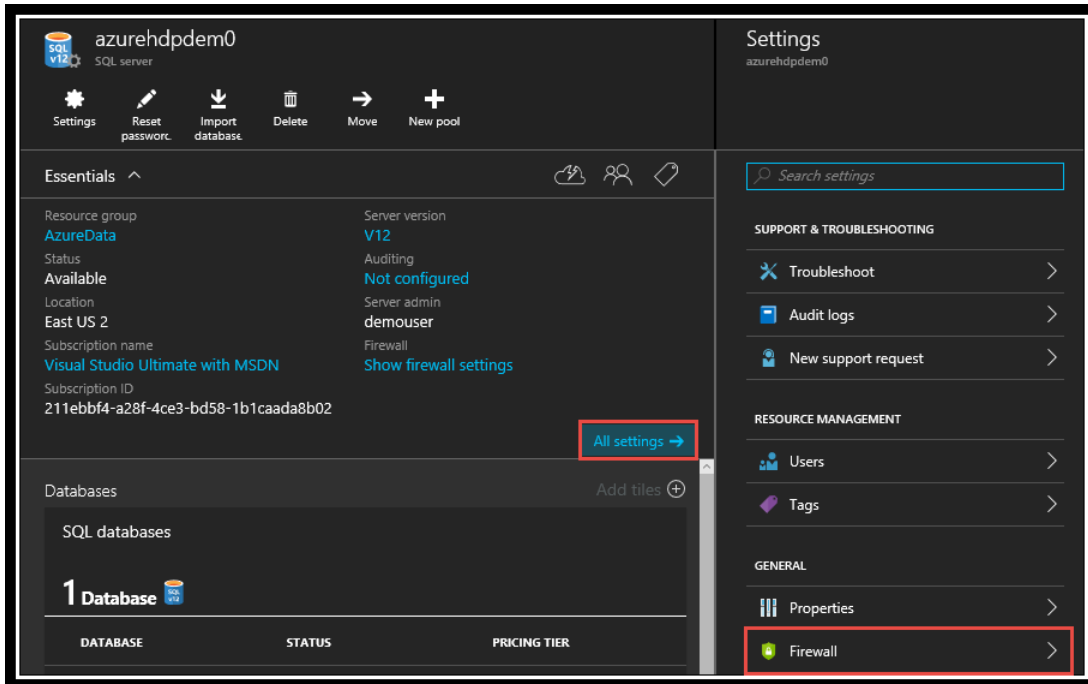
- Open the AdventureWorks Azure SQL Database you created in Exercise 2.



2. This will open the AdventureWorks details. Click on the **Server name** of the AdventureWorks database.

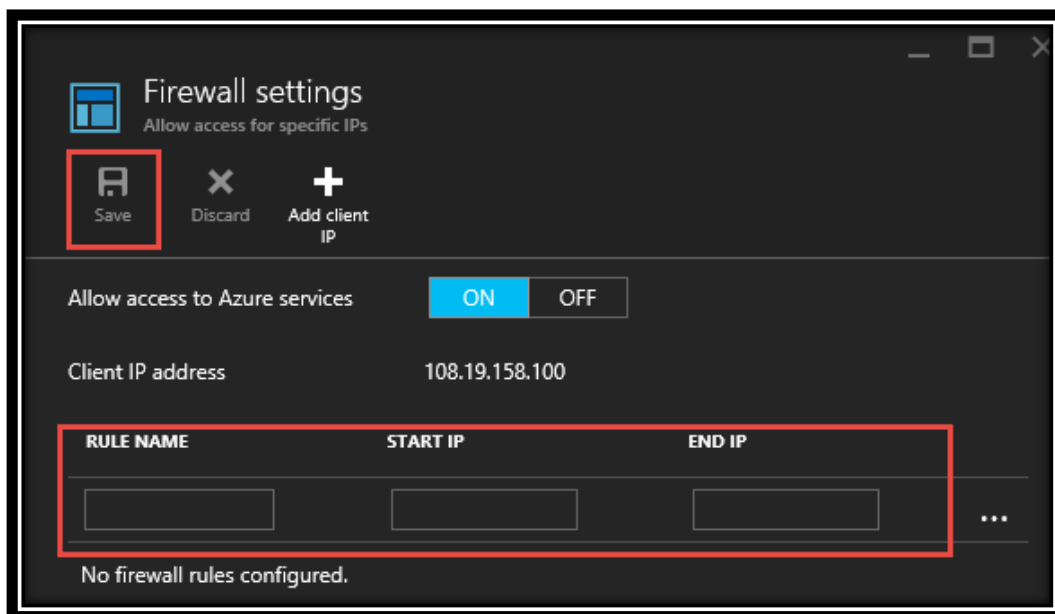


3. This will open the details blade of your server. Click **All settings**, then choose **Firewall**.



4. In the **Firewall settings** enter the following values then click **Save**:

- Rule Name: HDP
- Start IP: [Virtual IP of your Hortonworks VM]
- End IP: [Virtual IP of your Hortonworks VM]
- Note: The Start IP and End IP will be the same value.



**NOTE: If for some reason the Hortonworks VM is stopped (not just restarted) from within the portal and then started again it could have a different Virtual IP address as assigned dynamically. If this is the case, you will need to update this rule's Start IP and End IP.**

## Exercise 5: Transfer data using Sqoop

In this exercise, you will login to the Hortonworks Sandbox VM and transfer data from your Azure SQL Database into a Hive table.

1. Connect to the Hortonworks Sandbox VM's SSH session by clicking the Connect button within the browser that was connected before the VM restarted, or by launching your browser and navigating to `http://[Virtual IP of Hortonworks VM]:4200` (**replace the placeholder value** with the IP you saved earlier).
2. Execute the following command to view the available databases in your Azure SQL Database. **Replace the placeholder value** with the **Server name** of your Azure SQL Database Server that you created and noted in Exercise 2.

```
sqoop list-databases --connect jdbc:sqlserver://[Server
name].database.windows.net:1433 --username demouser --password
demo@pass1
```

Below you can see that we have the AdventureWorks database available.

```

demouser@sandbox:~ · × +
← → ↺ | 104.208.155.191:4200
sandbox login: demouser
demouser@sandbox.hortonworks.com's password:
Last login: Tue Dec 22 01:11:47 2015 from 108.19.158.100
[demouser@sandbox ~]$ sqoop list-databases --connect jdbc:sqlserver://azurehdpdemo.da
Warning: /usr/hdp/2.3.2.0-2950/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
15/12/22 04:31:32 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6.2.3.2.0-2950
15/12/22 04:31:32 WARN tool.BaseSqoopTool: Setting your password on the command-line
15/12/22 04:31:33 INFO manager.SqlManager: Using default fetchSize of 1000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/hadoop/lib/slf4j-log4j12-1.7.
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/zookeeper/lib/slf4j-log4j12-1
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
master
AdventureWorks ←
[demouser@sandbox ~]$

```

- Next, extract data from our AdventureWorks database into a Hive table by executing the following command. **Replace the placeholder value** using the SQL Database name you saved earlier.

```

sudo -u hdfs sqoop import --connect "jdbc:sqlserver://[Server
Name].database.windows.net:1433;database=AdventureWorks;user=demo
user;password=demo@pass1;encrypt=true;trustServerCertificate=fals
e;hostNameInCertificate=*.database.windows.net;loginTimeout=30;"
--table SalesOrderDetail --hive-import -- --schema SalesLT

```

The output from the above command should return output like this. If you scroll back through the output you will see job metrics, error and warning information, etc.

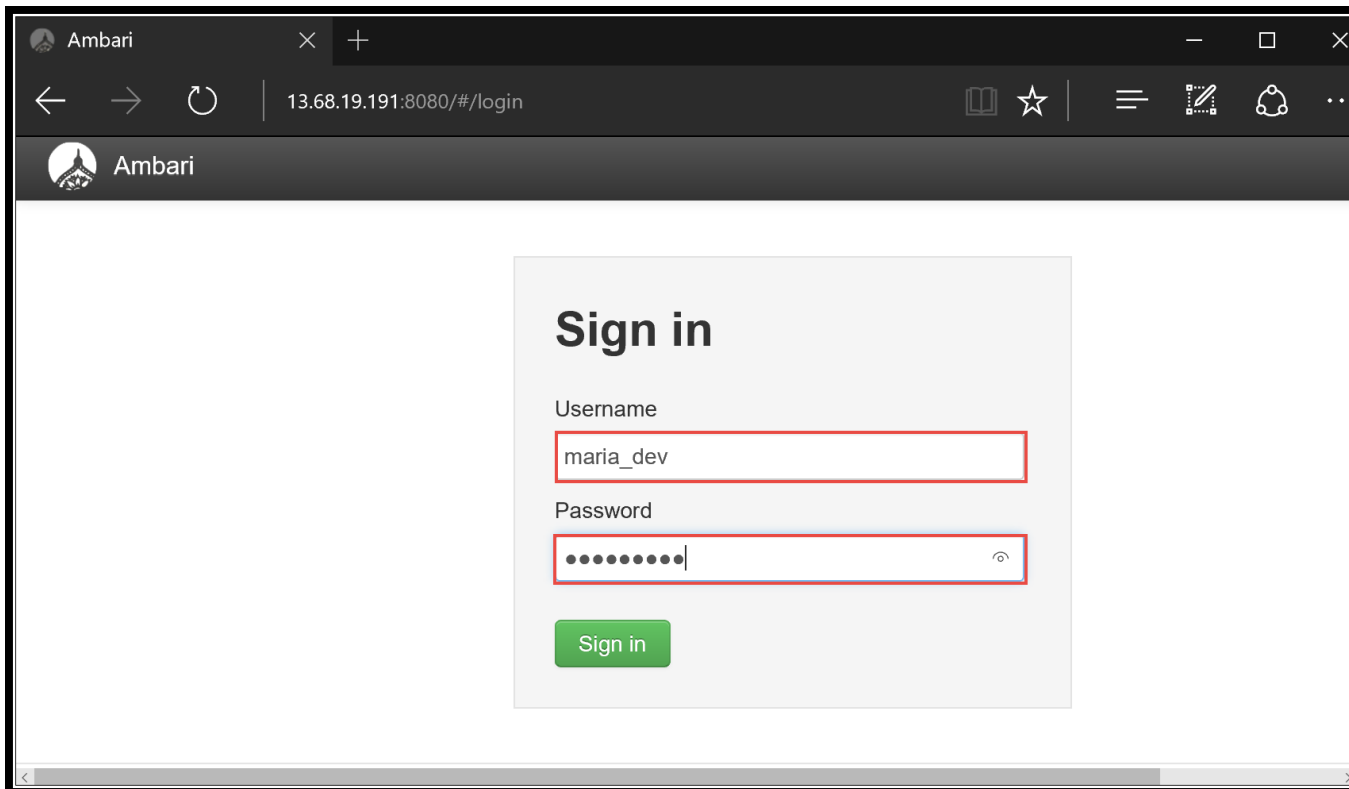
```

15/12/22 04:56:54 INFO mapreduce.ImportJobBase: Transferred 55.4287 KB in 126.1912 seconds (449.7856 bytes/sec)
15/12/22 04:56:54 INFO mapreduce.ImportJobBase: Retrieved 542 records.
15/12/22 04:56:54 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM [SalesLT].[SalesOrderDetail] AS t WHERE 1=0
15/12/22 04:56:54 WARN hive.TableDefWriter: Column UnitPrice had to be cast to a less precise type in Hive
15/12/22 04:56:54 WARN hive.TableDefWriter: Column UnitPriceDiscount had to be cast to a less precise type in Hive
15/12/22 04:56:54 WARN hive.TableDefWriter: Column LineTotal had to be cast to a less precise type in Hive
15/12/22 04:56:54 WARN hive.TableDefWriter: Column ModifiedDate had to be cast to a less precise type in Hive
15/12/22 04:56:54 INFO hive.HiveImport: Loading uploaded data into Hive
Logging initialized using configuration in jar:file:/usr/hdp/2.3.2.0-2950/hive/lib/hive-common-1.2.1.2.3.2.0-2950.jar!/hive-log
OK
Time taken: 19.576 seconds
Loading data to table default.salesorderdetail
Table default.salesorderdetail stats: [numFiles=4, totalSize=56759]
OK
Time taken: 3.792 seconds
[demouser@sandbox ~]$

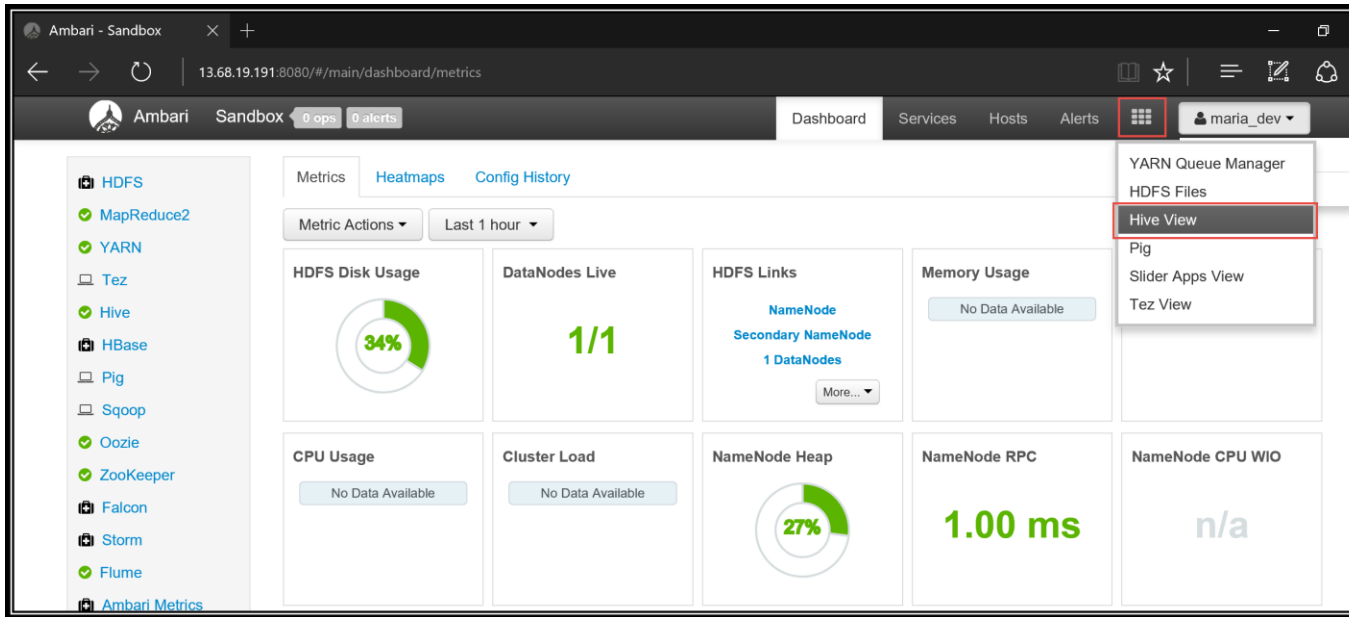
```



4. Query data from the SalesOrderDetail Hive table. Navigate to the Ambari Hive View interface of your Hortonworks Sandbox VM. This is located by browsing to the **Virtual IP** of the Hortonworks Sandbox VM using port 8080 `http://[Virtual IP of Hortonworks VM]:8080`.
5. On the Sign In screen, use **maria\_dev** for the Username and Password.

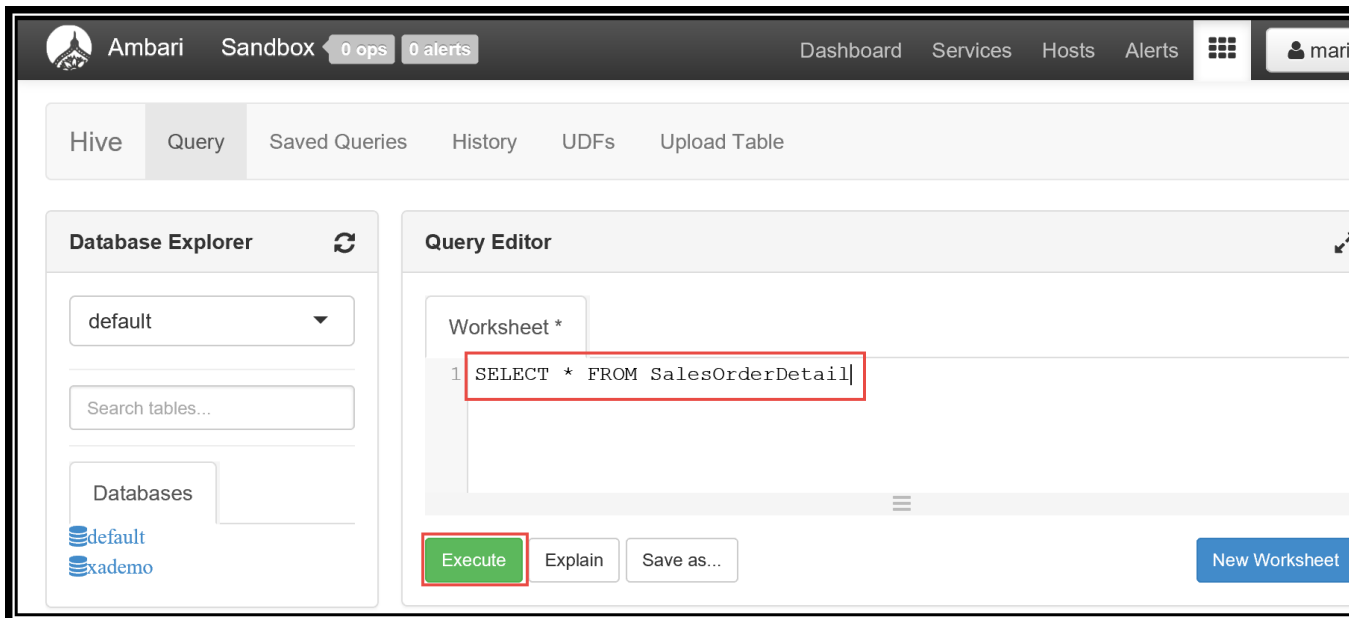


6. Next Click the menu icon in the upper right and choose **Hive View** from the dropdown. This will open the Query Editor window.

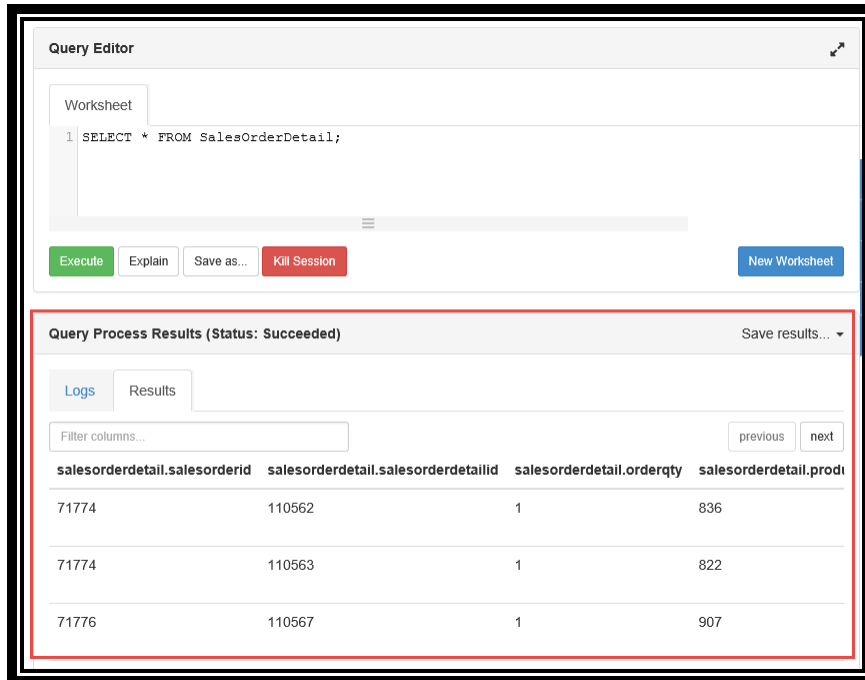


7. In the Query Editor type the following query in the space provided and click the Execute button

```
SELECT * FROM SalesOrderDetail
```



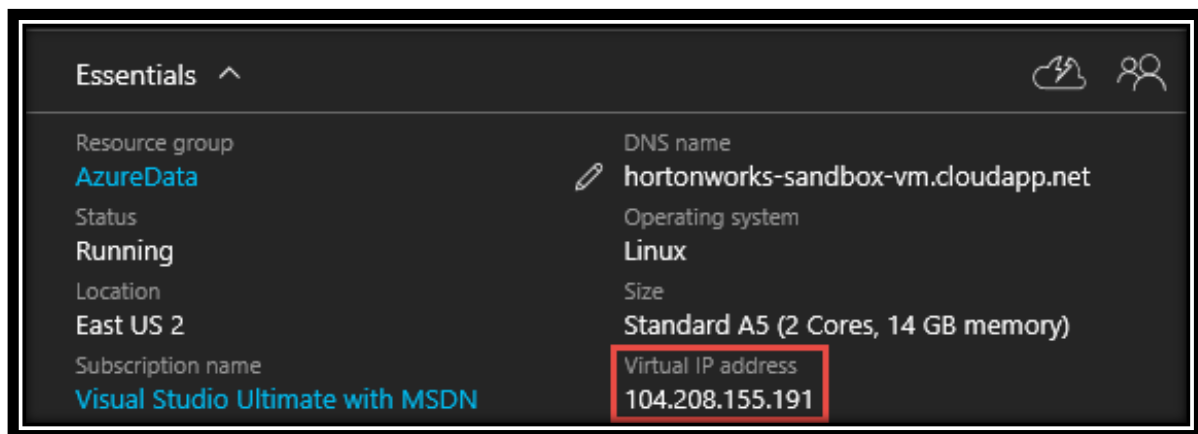
8. The results will be shown below the query.



## Validate Lab Completion

1. Create a screenshot that shows the essentials panel from within the Azure Portal of your Hortonworks Sandbox virtual machine instance.

Please save your lab screenshots as either a .jpeg or .png. Upload your screenshots in one .zip file [here](#).



- Next take a screenshot of the web-based SSH connection to the Hortonworks Sandbox instance using the same IP address as the previous screenshot.

```

104.208.155.191:4200
sandbox login: demouser
demouser@sandbox.hortonworks.com's password:
Last login: Tue Dec 22 05:50:22 2015 from 172.16.0.4
[demouser@sandbox ~]$

```

- A screenshot of the output of the following query from the Ambari Hive View interface using the same IP

```
SELECT * FROM SalesOrderDetail
```

The screenshot shows the Ambari Hive View interface. The URL bar displays `http://104.208.155.191:8080/#/main/views/HIVE/1.0.0/AUTO_HIVE_INSTANCE`. The interface is divided into two main sections: Database Explorer on the left and Query Editor on the right.

**Database Explorer:** Shows a dropdown menu set to 'default' and a search bar. Below, it lists two databases: 'default' and 'xademo'.

**Query Editor:** Contains a 'Worksheet' tab with the query `1 SELECT * FROM SalesOrderDetail;`. Below the query are buttons for 'Execute', 'Explain', 'Save as...', and 'Kill Session'. A 'New Worksheet' button is also present.

**Query Process Results (Status: Succeeded):** This section shows the results of the query. It includes a 'Logs' tab and a 'Results' tab. The 'Results' tab displays a table with the following data:

| salesorderdetail.salesorderid | salesorderdetail.salesorderid | salesorderdetail.orderqty | salesorderdetail.prod |
|-------------------------------|-------------------------------|---------------------------|-----------------------|
| 71774                         | 110562                        | 1                         | 836                   |
| 71774                         | 110563                        | 1                         | 822                   |
| 71776                         | 110567                        | 1                         | 907                   |

## Lab Summary

In this lab, you have created a Hortonworks Sandbox virtual machine from the Microsoft Azure Marketplace and an Azure SQL Database sample. You extracted data from the Azure SQL Database into a Hive table and queried the data from Hive.

