

ETL in Hortonworks Sandbox on Azure

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.

Prerequisites

- Client computer with Internet connectivity.
- A Microsoft Azure Subscription, you can create a free trial here [Azure Trail](#), 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.

Objectives

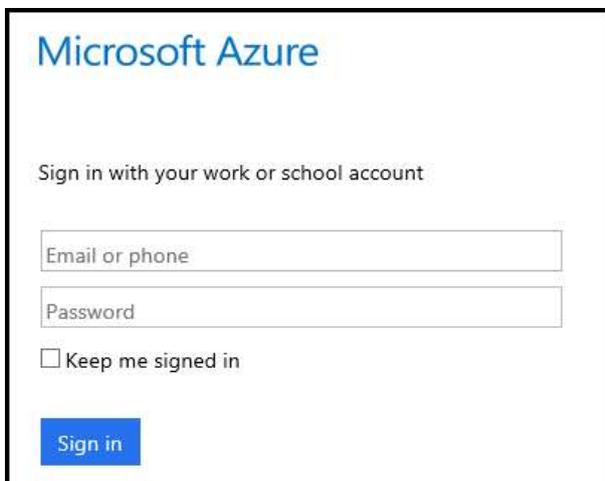
To create a credit risk assessment solution, we'll follow these steps through a series of tasks:

- Creating an Azure Environment
- Create an SQL Azure Database
- Create a Virtual Machine with Hortonworks Sandbox
- Configure your Azure SQL Database for remote connectivity.
- Transfer data using Sqoop
- Validate Lab Completion.

Task 1: Azure - 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. Launch an **In-Private Browser Window** and navigate to <https://portal.azure.com> . The following page should load.
2. Enter the account associated with your Microsoft Azure subscription.



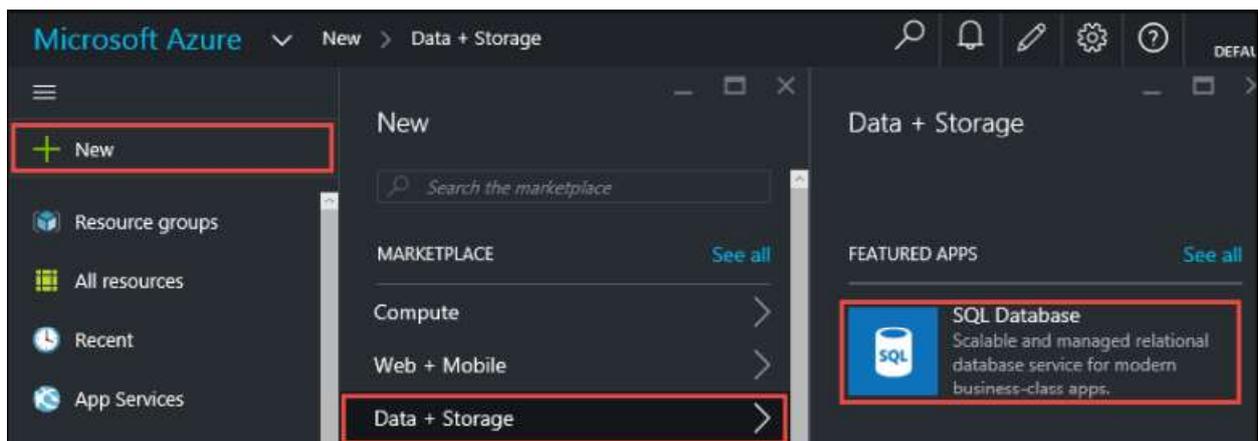
3. If your account is associated with an organization account and a Microsoft account, you may be prompted to choose which one to authenticate with for your Microsoft Azure account.



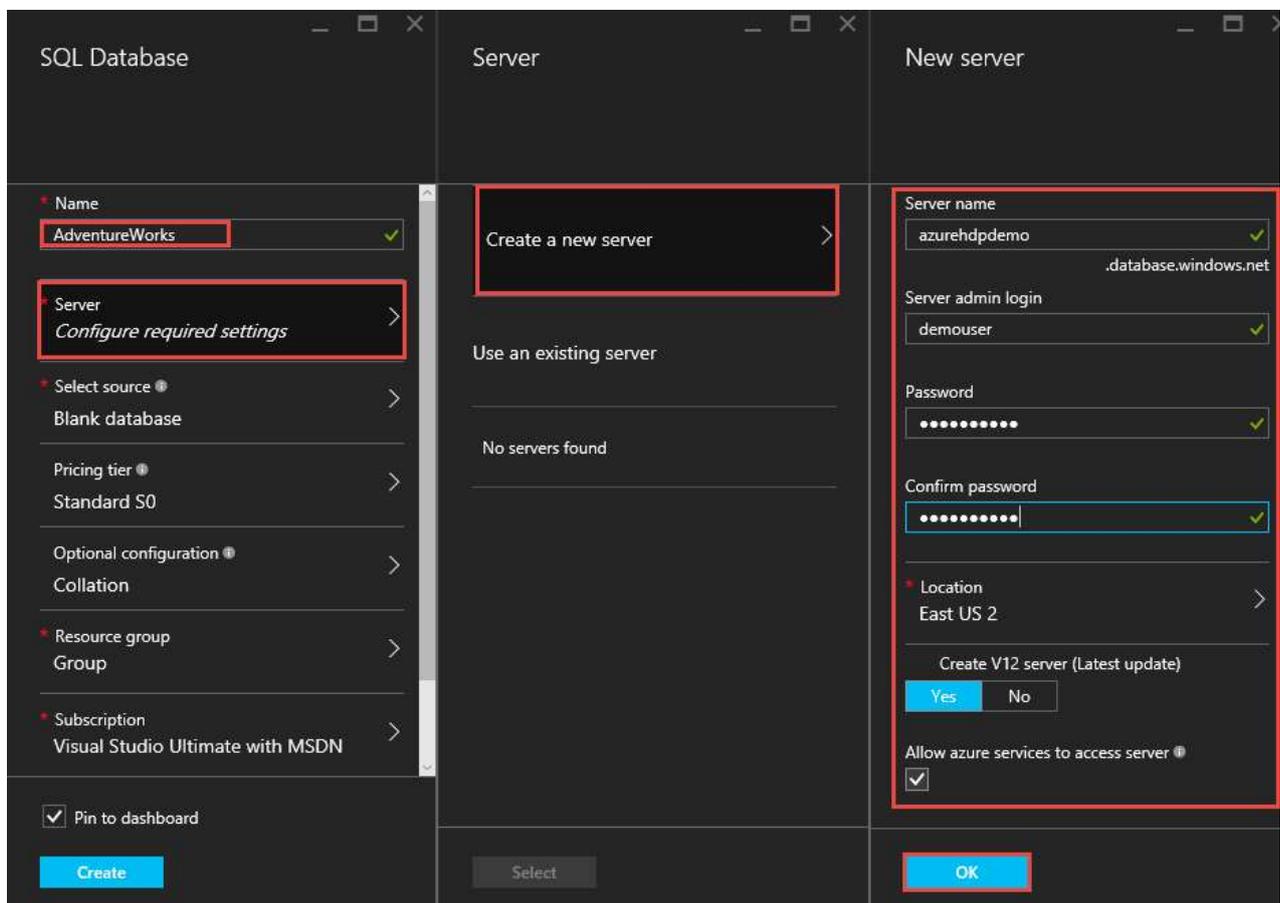
Task 2: Create an SQL Azure Database

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.

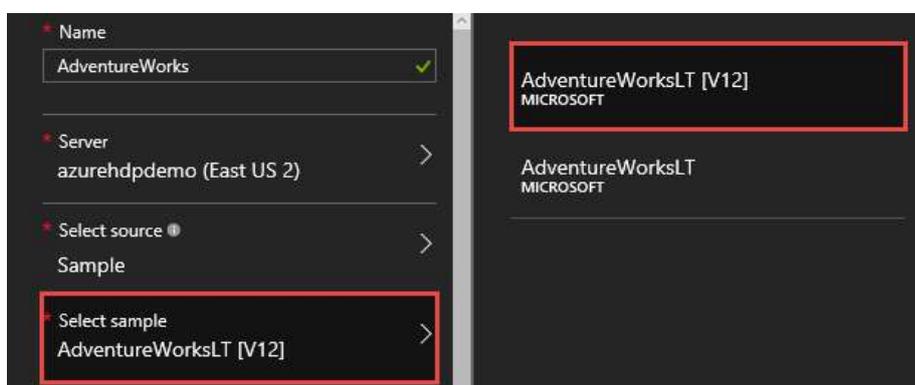
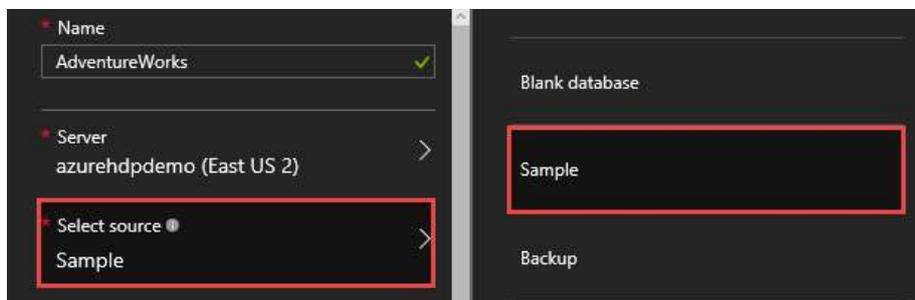


2. The Azure portal will open the **SQL Database** blade. Type *AdventureWorks* into the **Name** box. Choose **Server**, and then **Create a new server**. Specify the following **New Server** configurations and click **OK**.
 - **Server Name:**
 - **Server Admin Login:** demouser
 - **Password:** demo@pass1
 - **Location:**
 - **Create V12 Server:** Yes
 - **Allow Azure Services to Access Server:** Checked

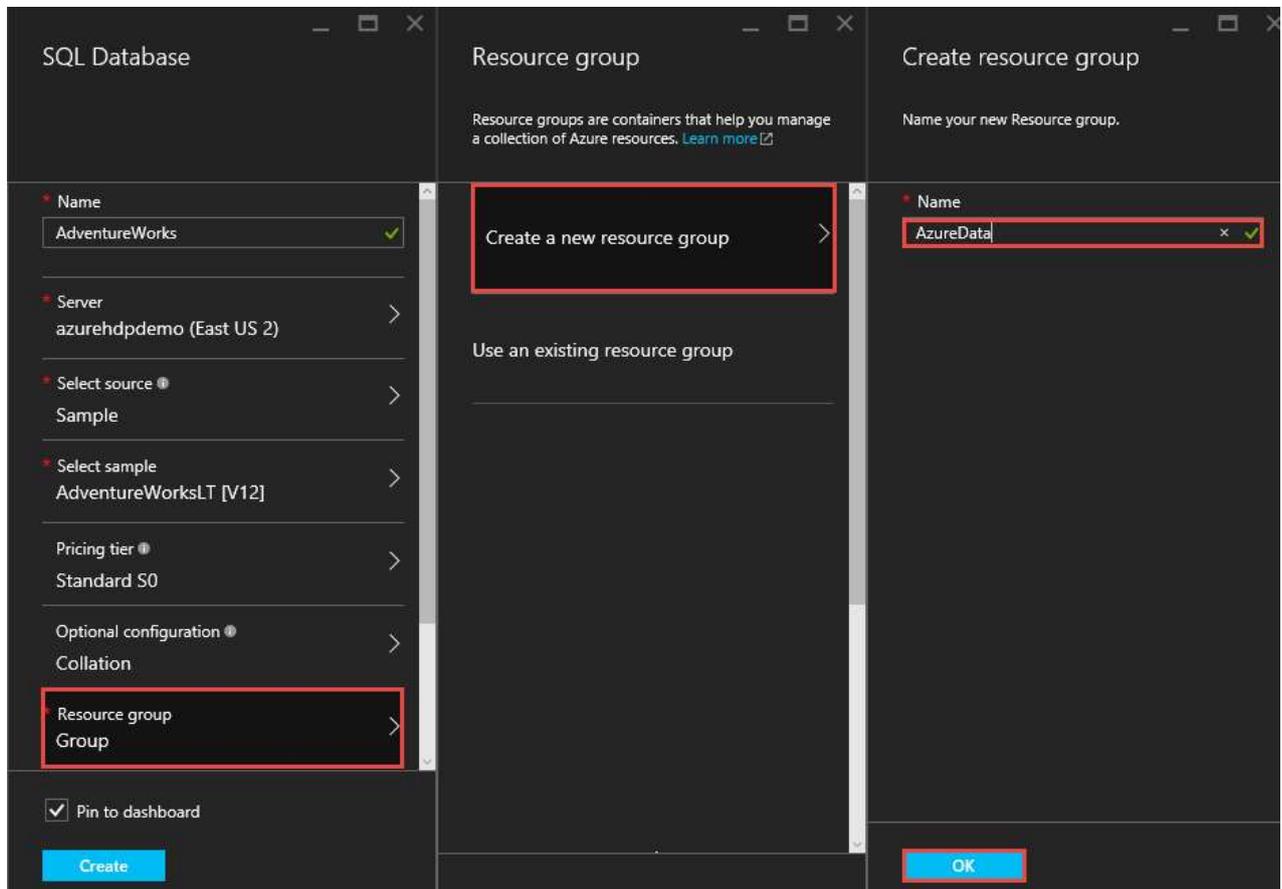


3. Note the **Server Name of the SQL Database** for later reference.

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:**
- **Select Source:** Sample
- **Select Sample:** AdventureWorksLT [V12]
- **Pricing Tier:** Standard S0
- **Resource Group:** AzureData
- **Subscription:**

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



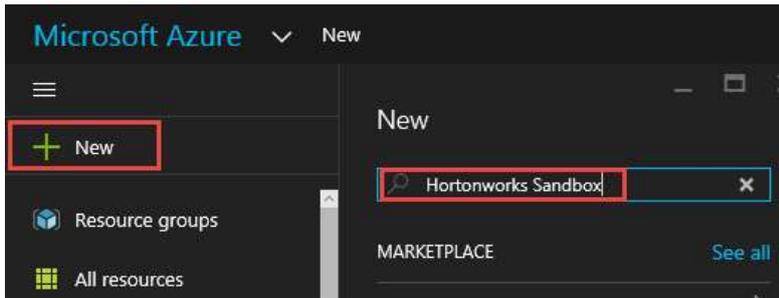
8. Wait until the status of Azure SQL Database is **'Online'**.



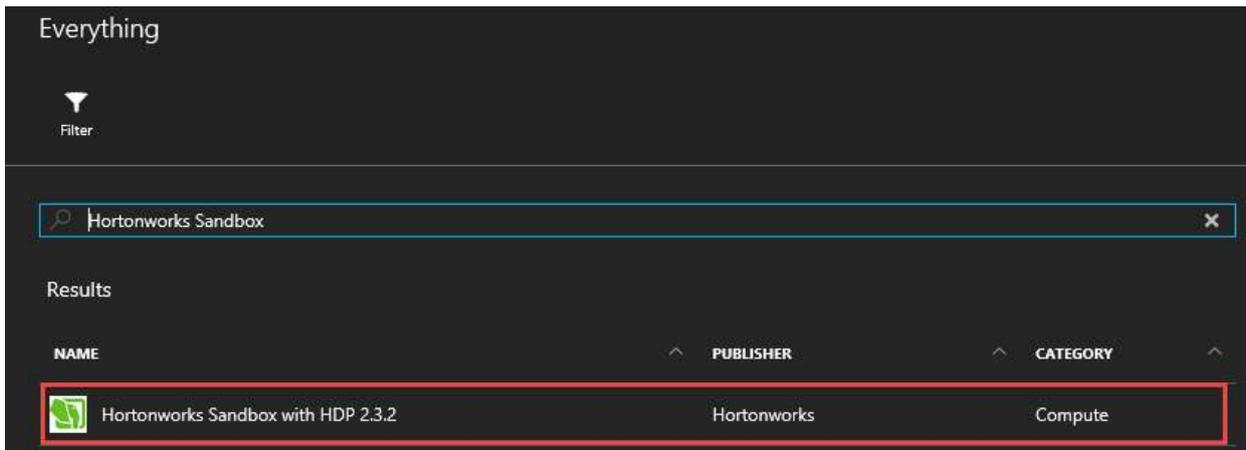
Task 3: Create a Virtual Machine with Hortonworks Sandbox

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 **Classic** is selected. Read the description then click the **Create** button.

Hortonworks Sandbox with HDP 2.3.2
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](#).

Build a Proof of Concept

The Sandbox includes the Hortonworks Data Platform in an easy to use form. You can add your own datasets, and connect it to your existing tools and applications. With this, you can prove out your use of Hadoop and plan the integration points for your first Hadoop project.

Test New Functionality

You can test new functionality with the Sandbox before you put it into production. Simply, easily and safely.

Once you have deployed the Hortonworks Sandbox on Azure, navigate to `http://<hostname>.cloudapp.net:8888` to get started. Replace `<hostname>` with what you enter in the "Hostname Name" field in next steps.

Sandbox is an enterprise Hadoop environment with many interactive tutorials to jumpstart your learning. Sandbox includes many of the most exciting developments from the latest HDP distribution, which you can get up and running in under 15 minutes. It comes with tutorials to help you learn Hadoop, Spark, Pig, Hive, Kafka, Storm, HBase, Ranger, Falcon, Ambari and YARN.

[Twitter](#)
[Facebook](#)
[LinkedIn](#)
[YouTube](#)
[G+](#)
[Email](#)

APPLICATIONS
 Business Analytics Custom Applications Packaged Applications

SYSTEM
 HDFS EMR HPP Batch Interactive Real-Time
 YARN: Data Operating System

Select a deployment model

Classic

Create

4. Specify the following configuration options in the Create VM blade.

- **Name:** hortonworks-sandbox-vm
- **User Name:** demouser
- **Authentication Type:** Password
- **Password:** demo@pass1
- **Resource Group:** AzureData
- **Location:** Will be auto-selected to match the resource group.

Create VM
Hortonworks Sandbox with HDP 2.3.2

* Host Name
hortonworks-sandbox-vm ✓

* User name
demouser ✓

Authentication type
Password SSH public key

* Password
•••••••• ✓

Pricing Tier
Standard A5 >

Optional Configuration
Network, storage, diagnostics 🔒

Resource Group
AzureData >

Subscription
Visual Studio Ultimate with MSDN 🔒

Location
East US 2 >

* Legal terms
Review legal terms >

Pin to dashboard

Create

5. For the **Pricing Tier** configuration, select **A5 Standard** and click **Select**.

Create VM

Hortonworks Sandbox with HDP 2.3.2

Host Name
hortonworks-sandbox-vm ✓

User name
demouser ✓

Authentication type
 Password SSH public key

Password
..... ✓

Pricing Tier
Standard A4 >

Optional Configuration
Network, storage, diagnostics >

Resource Group
Group >

Subscription
Visual Studio Ultimate with MSDN >

Location
East US 2 >

Legal terms
[Review legal terms](#) >

Pin to dashboard

Create

Choose your pricing tier

Browse the available pricing tiers 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 USD/MONTH (ESTIMATED)	163.68 USD/MONTH (ESTIMATED)	327.36 USD/MONTH (ESTIMATED)

Select

6. For **Legal terms**, review the terms and click **Purchase**.

Create VM
Hortonworks Sandbox with HDP 2.3.2

Host Name
hortonworks-sandbox-vm ✓

User name
demouser ✓

Authentication type
Password SSH public key

Password
•••••••• ✓

Pricing Tier
Standard A5 >

Optional Configuration
Network, storage, diagnostics 🔒

Resource Group
AzureData >

Subscription
Visual Studio Ultimate with MSDN 🔒

Location
East US 2 >

Legal terms
Review legal terms >

Pin to dashboard

Create

Purchase

Offer details

Hortonworks Sandbox
by Hortonworks 0.0000 USD/hr *
[Terms of use and privacy policy](#)

Standard A5
by Microsoft 0.2200 USD/hr +
[Terms of use and privacy policy](#) [Pricing for other VM sizes](#)

* **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.

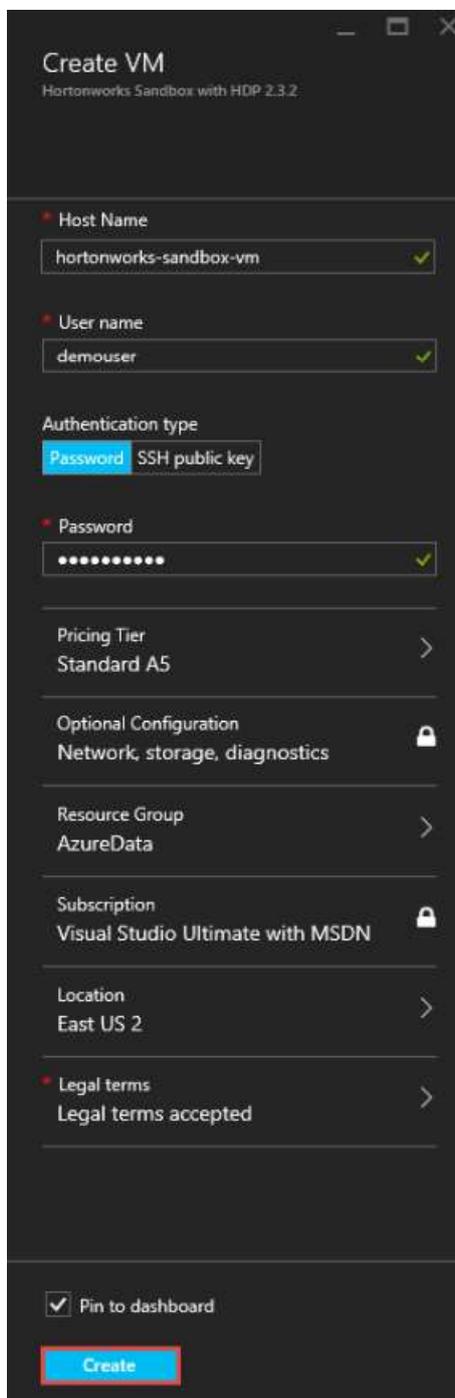
+ **Azure Resource:** May be purchased using Microsoft subscription credits or monetary commitment funds and participates in discounts. Prices presented are retail prices and may not reflect discounts associated with your subscription.

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

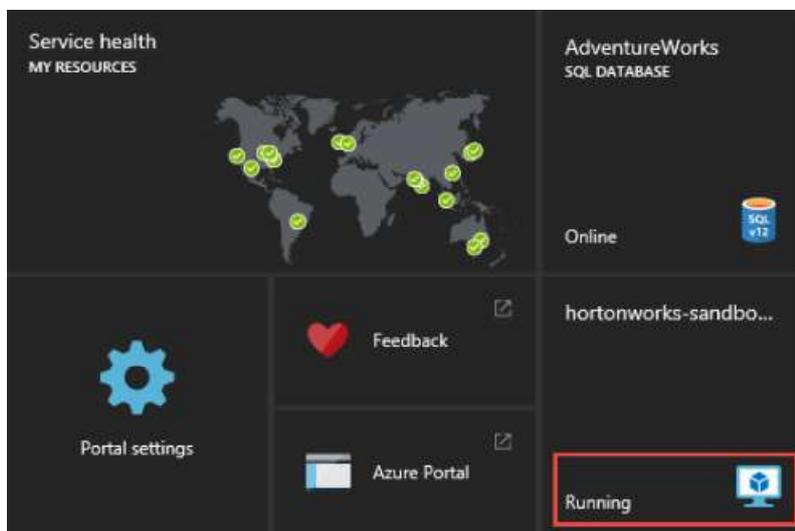
7. Verify your **Create VM Configuration**, and click **Create**.



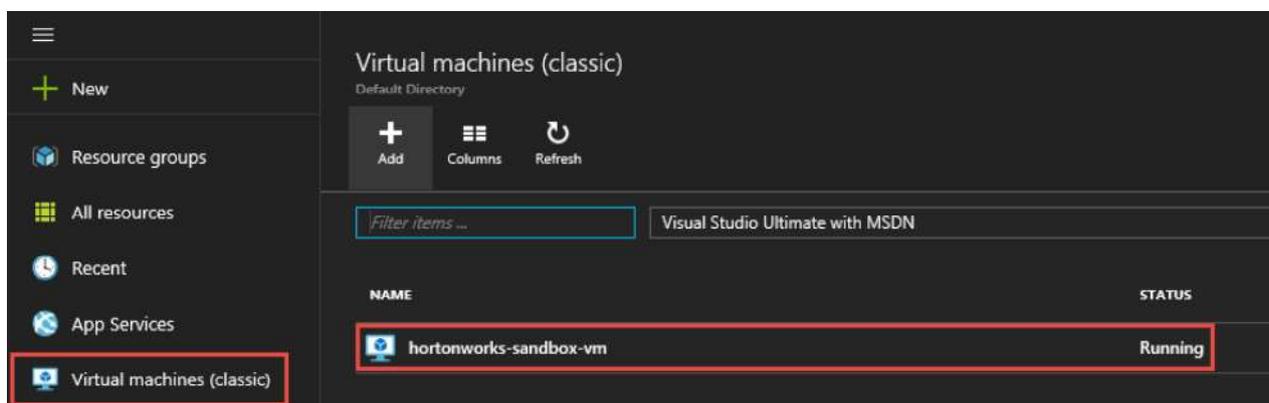
8. Now you will see that the new virtual machine is creating, as per the status on portal dashboard.



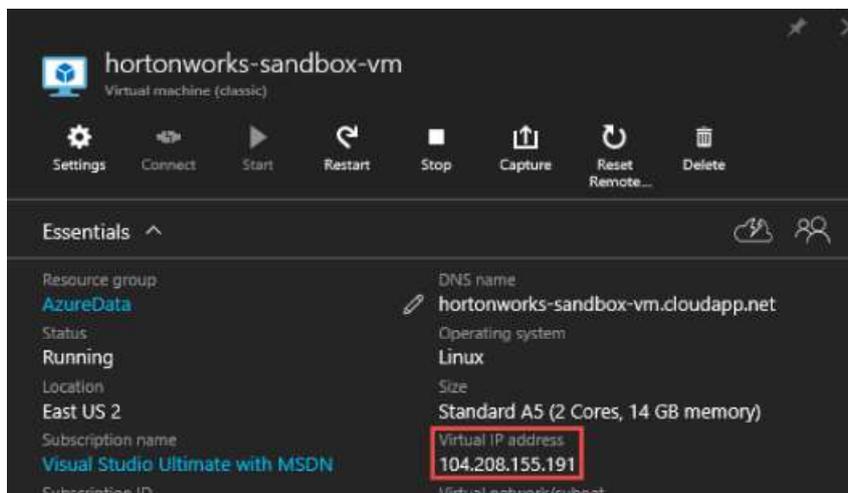
9. Wait until the status of created virtual machine is **'Running'**.
Note: It may take 10-15 minutes for the virtual machine to complete provisioning.



- If the portal does not refresh, click **Virtual Machines (Classic)** to see the latest status of the Virtual Machine. Wait until the status turns to **'Running'**.



- Once the virtual machine is in status **'Running'**, click on the **Virtual Machine Name** and go to details.



- Note the Virtual IP Address of the Hortonworks Sandbox Virtual Machine** as you will reference it in the next step and future steps.
- 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.

```

demouser@sandbox:~ x +
104.208.155.191:4200
sandbox login: demouser
demouser@sandbox.hortonworks.com's password:
[demouser@sandbox ~]$

```

- 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.

- 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/02AAE597-3865-456C-AE7F-613F99F850A8/sqljdbc_4.2.6420.100_enu.tar.gz' | tar xz
  % 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 ~]$

```

- Execute the following 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*

```

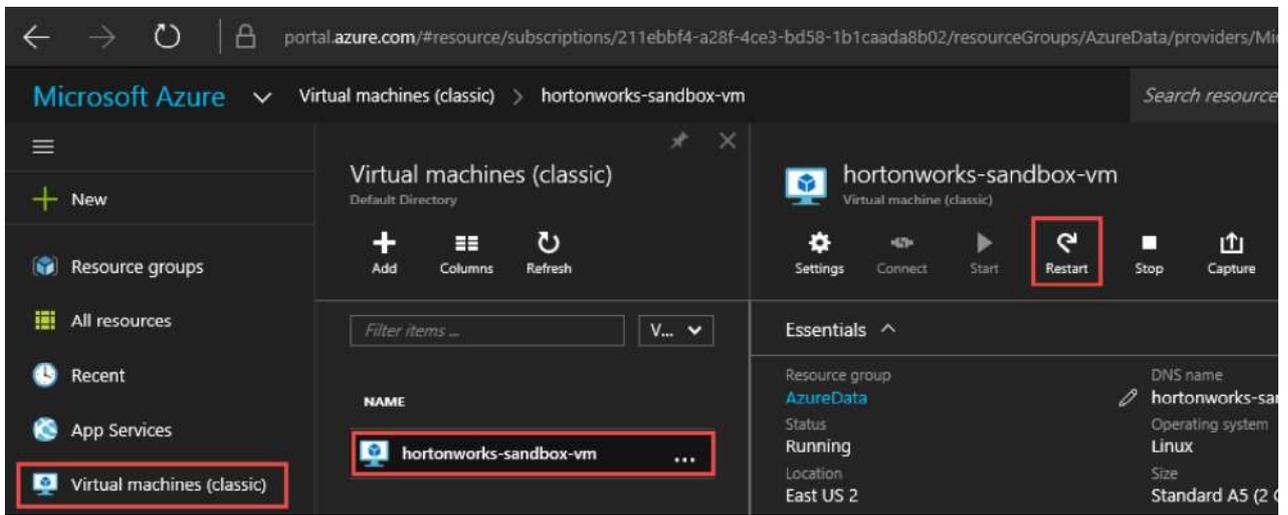
The screen will resemble below.

```

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]$

```

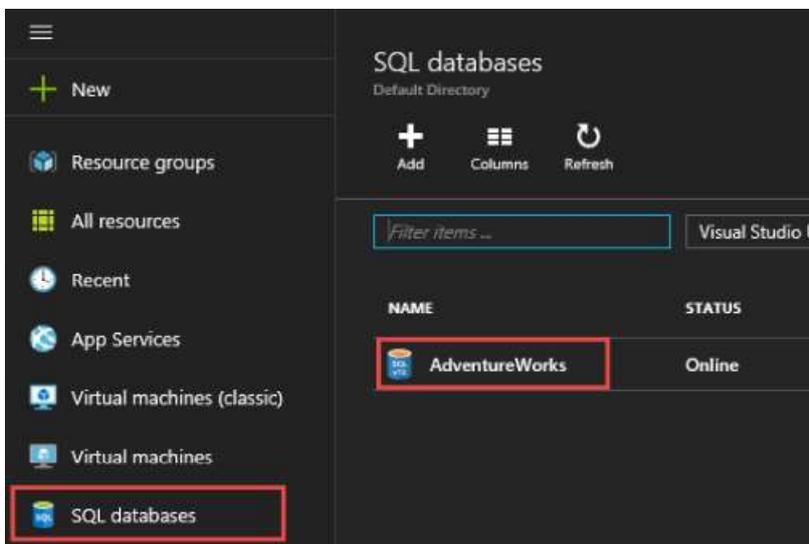
- Restart the Hortonworks Sandbox VM so that the new driver will become available. In the Portal click, **Virtual machines (Classic)** and then click the **hortonworks-sandbox-vm**. When the Blade for the VM loads click **Restart**.



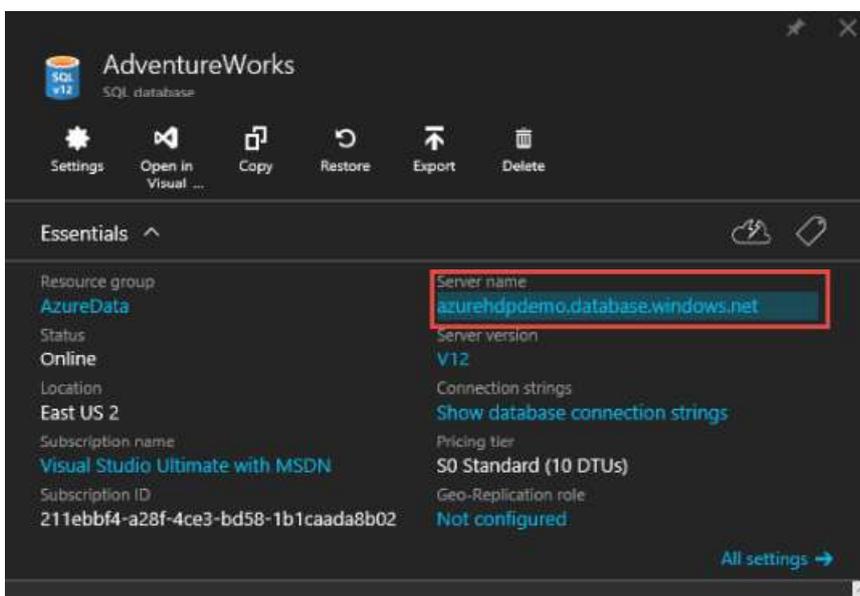
Task 4: Configure your Azure SQL Database for remote connectivity

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

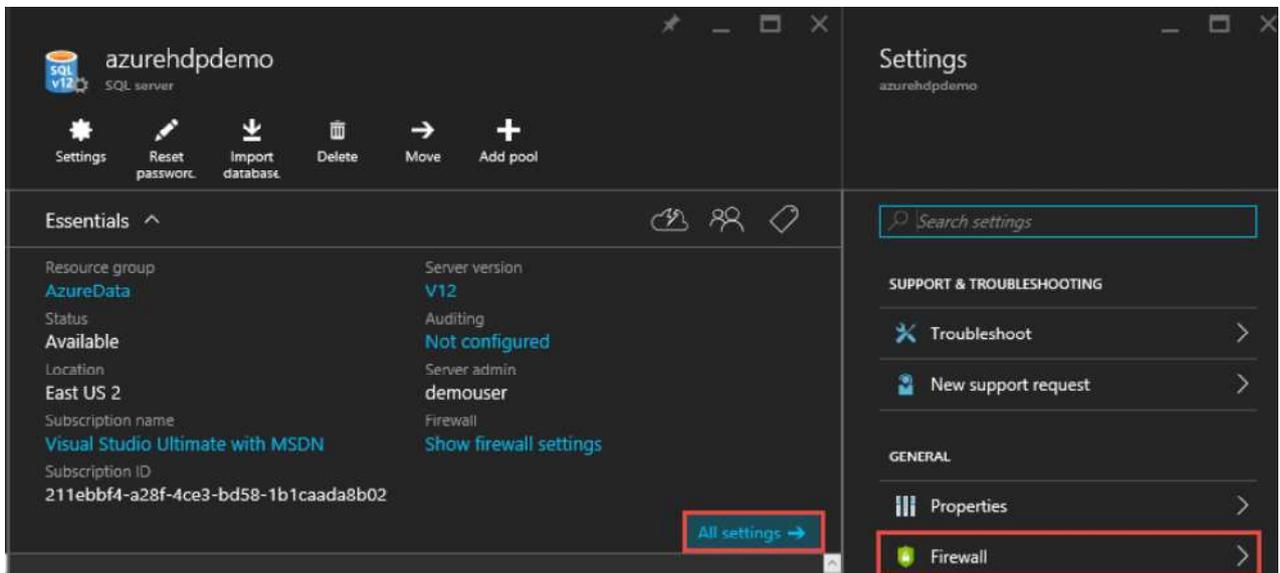
1. Open the **AdventureWorks Azure SQL Database** you created in Task 2.



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



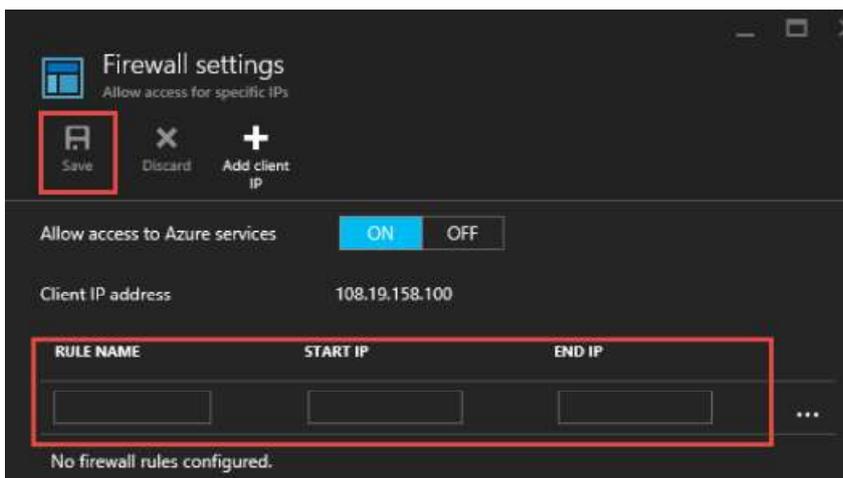
3. This will open the details blade of your server. Click **All Settings**, then choose **Firewall**.



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.

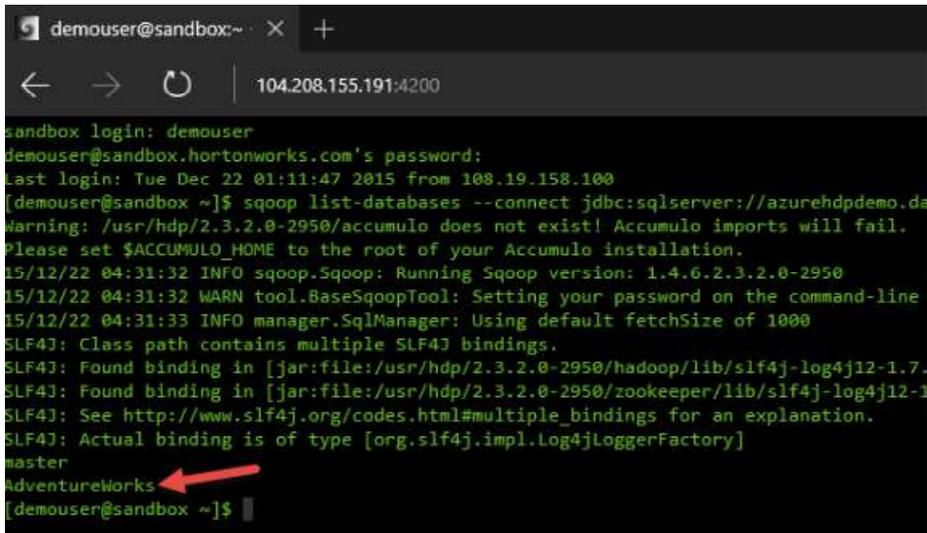
Task 5: Transfer data using Sqoop

In this task 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 name of your Azure SQL Database Server that you created and noted in Task 2.

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

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

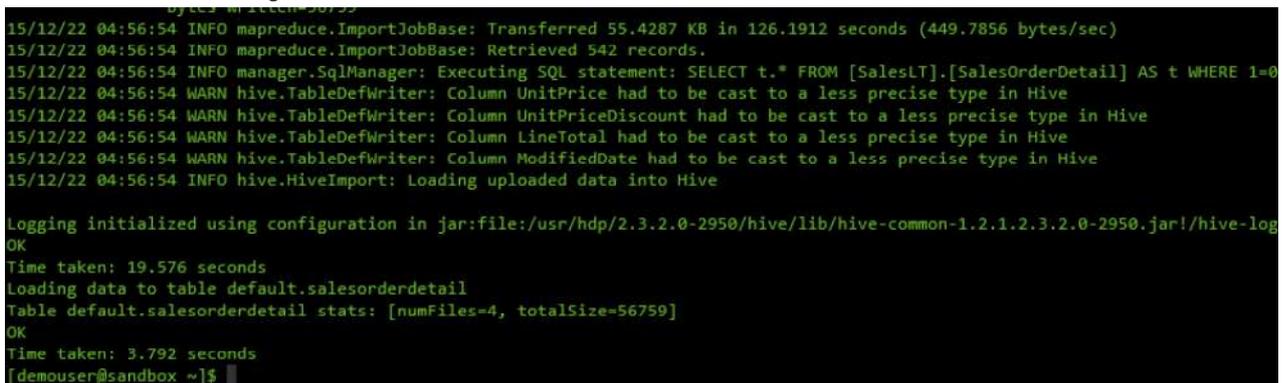


```
demouser@sandbox:~ · × +
104.208.155.191:4200
sandbox login: demouser
demouser@sandbox.hortonworks.com's password:
Last login: Tue Dec 22 01:11:47 2015 from 108.19.158.100
[demouser@sandbox ~]$ sqoop list-databases --connect jdbc:sqlserver://azurehdpdemo.d
Warning: /usr/hdp/2.3.2.0-2950/accumulo 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://[AdventureWorks SQL Database Server
Name].database.windows.net:1433;database=AdventureWorks;user=demouser;password=demo@pass1;encrypt=true;trustServer
Certificate=false;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 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

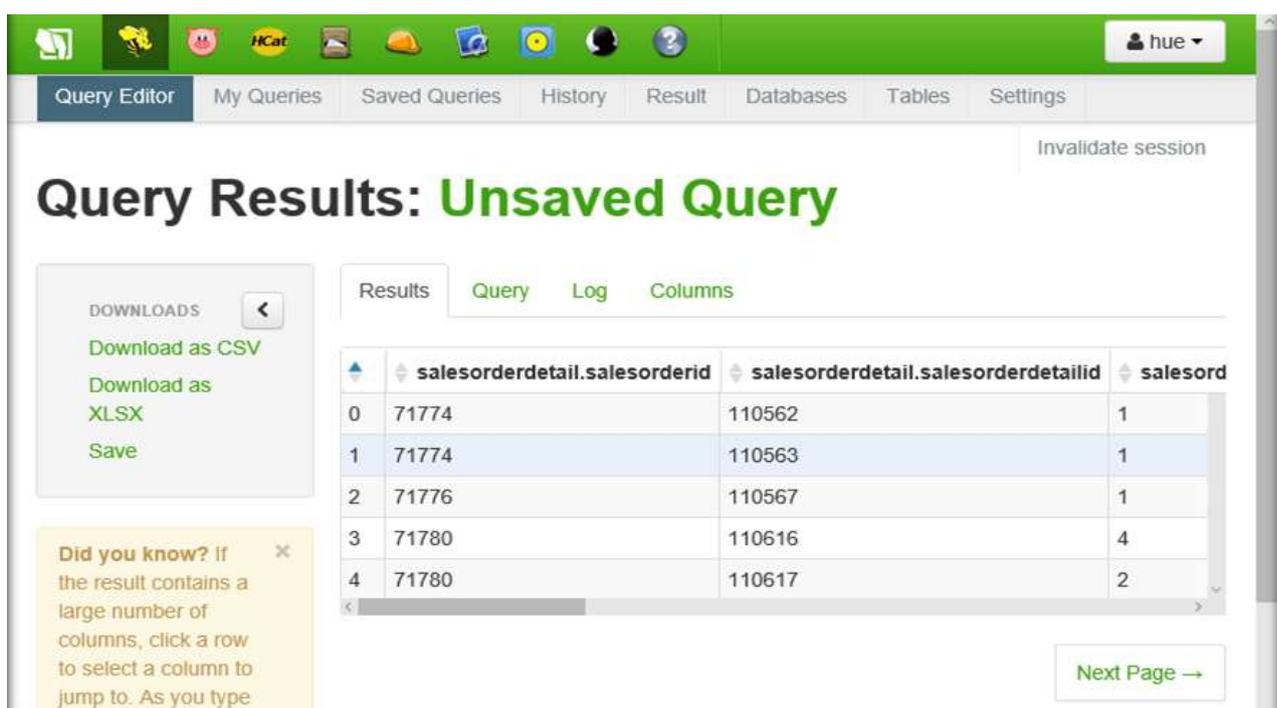
logging initialized using configuration in jar:file:/usr/hdp/2.3.2.0-2950/hive/lib/hive-common-1.2.1.2.3.2.0-2950.jar!/hive-log
OK
Time taken: 19.576 seconds
loading data to table default.salesorderdetail
Table default.salesorderdetail stats: [numFiles=4, totalSize=56759]
OK
Time taken: 3.792 seconds
[demouser@sandbox ~]$
```

- Query data from the SalesOrderDetail Hive table. Navigate to the Hue interface of your Hortonworks Sandbox VM. This is located by browsing to the **Virtual IP** of the Hortonworks Sandbox VM using port 8000 `http://[Virtual IP of Hortonworks VM]:8000`.
- Next Click the **Beeswax** icon. This will open the Query Editor window.



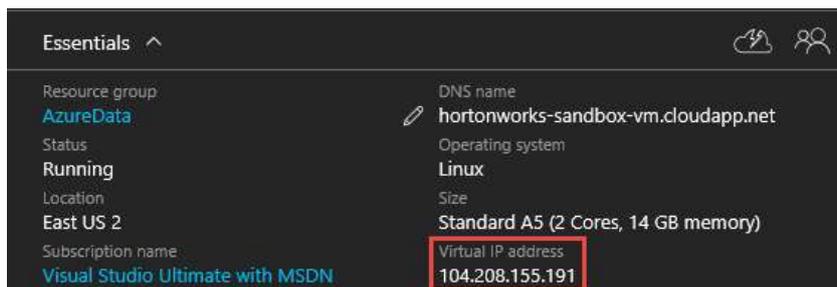
6. In the Query Editor type the following query in the space provided and click the Execute button.

```
SELECT * FROM SalesOrderDetail
```

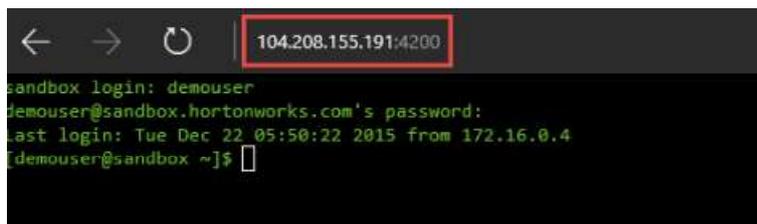


Task 6: Validate Lab Completion

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



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



3. A screenshot of the output of the following query from the Beeswax interface using the same IP.

```
SELECT * FROM SalesOrderDetail
```

104.208.155.191:8000 beeswax/results/11/0?context=design%3A15

Query Editor My Queries Saved Queries History Result Databases Tables Settings Invalidate session

Query Results: Unsaved Query

Downloads: Download as CSV, Download as XLSX, Save

Did you know? If the result contains a large number of columns, click a row to select a column to jump to. As you type into the field, a drop-down list displays column names that

	salesorderid	salesorderdetailid	orderqty	sales
0	71774	110562	1	836
1	71774	110563	1	822
2	71776	110567	1	907
3	71780	110616	4	905
4	71780	110617	2	983

Next Page →

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