

# **Cloudera + Tableau Quickstart Deployment and Usage Guide**



## Table of Contents

1	Introduction.....	3
2	Prerequisites.....	4
2.1	Tableau.....	4
2.2	Cloudera .....	4
3	Cloudera Topology .....	5
4	Deployment Steps .....	6
5	Connecting to the Cloudera Cluster .....	7
6	Loading Sample Data .....	8
7	Viewing Sample Data with Tableau .....	12

## 1 Introduction

This Cloudera + Tableau Quickstart ARM template will enable an organization to pilot the Cloudera version of Hadoop with Tableau Server all in one deployment. Hadoop clusters can be challenging to deploy without the requisite technical skills, and loading and consuming data from the cluster can be difficult to understand for those new to Hadoop. With these challenges in mind, scripts to load sample data into the Cloudera Hadoop cluster are included, as well as a sample Tableau dashboard to query and display the sample data.

## 2 Prerequisites

### 2.1 Tableau

This is a Bring Your Own License (BYOL) version of Tableau. Once the Tableau Server virtual machine is deployed, a user must log in to the server and enter the Tableau license key. The sample Tableau dashboard can only be deployed once the license key is entered. In addition, version 10.0 of Tableau Server is installed, so version 10.0 of Tableau Desktop must be used.

### 2.2 Cloudera

The ARM template will deploy Cloudera Express with a 60-day trial of Cloudera Enterprise features. Once the trial has concluded, the Cloudera Enterprise features will be disabled until you obtain and upload a license. Also, the default 20 cores in the Azure region should be increased to accommodate the number of cores needed for the Cloudera cluster.

This template creates a multi-server Cloudera CDH 5.4.x Apache Hadoop deployment on CentOS virtual machines, and configures the CDH installation for either POC or high availability production cluster.

The template also provisions storage accounts, virtual network, availability set, network interfaces, VMs, disks and other infrastructure and runtime resources required by the installation.

The template expects the following parameters:

Name	Description	Default Value
<b>adminUsername</b>	Administrator user name used when provisioning virtual machines	testuser
<b>adminPassword</b>	Administrator password used when provisioning virtual machines	Eur32#1e
<b>cmUsername</b>	Cloudera Manager username	cmadmin
<b>cmPassword</b>	Cloudera Manager password	cmpassword
<b>storageAccountPrefix</b>	Unique namespace for the Storage Account where the Virtual Machine's disks will be placed	defaultStorageAccountPrefix
<b>numberOfDataNodes</b>	Number of data nodes to provision in the cluster	3
<b>dnsNamePrefix</b>	Unique public dns name where the Virtual Machines will be exposed	defaultDnsNamePrefix
<b>region</b>	Azure data center location where resources will be provisioned	
<b>masterStorageAccountType</b>	The type of the Storage Account to be created for master nodes	Premium_LRS
<b>workerStorageAccountType</b>	The type of the Storage Account to be created for worker nodes	Standard_LRS
<b>virtualNetworkName</b>	The name of the virtual network provisioned for the deployment	clouderaVnet
<b>subnetName</b>	Subnet name for the virtual network where resources will be provisioned	clouderaSubnet
<b>subnet1Name</b>	Subnet name for the virtual network where resources will be provisioned	tableauSubnet
<b>tshirtSize</b>	T-shirt size of the Cloudera cluster (Eval, Prod)	Eval
<b>vmSize</b>	The size of the VMs deployed in the cluster (Defaults to Standard_DS14)	Standard_DS14

### 3 Cloudera Topology

The deployment topology is comprised of a predefined number (as per t-shirt sizing) Cloudera member nodes configured as a cluster, configured using a set number of manager, name and data nodes. Typical setup for Cloudera uses 3 master nodes with as many data nodes are needed for the size that has been chosen ranging from as few as 3 to thousands of data nodes. The current template will scale at the highest end to 200 data nodes when using the large t-shirt size.

The following table outlines the deployment topology characteristics for each supported t-shirt size:

<b>T-Shirt Size</b>	<b>Member Node VM Size</b>	<b>CPU Cores</b>	<b>Memory</b>	<b>Data Disks</b>	<b># of Master Node VMs</b>	<b>Services Placement of Master Node</b>
<b>Eval</b>	Standard_DS14	16	112 GB	10x1000 GB	1	1 (primary, secondary, cloudera manager)
<b>Prod</b>	Standard_DS14	10	112 GB	10x1000 GB	3	1 primary, 1 standby (HA), 1 cloudera manager

## 4 Deployment Steps

- 1) Navigate to <https://github.com/Azure/azure-quickstart-templates/tree/master/cloudera-tableau>.
- 2) Open the README.md file within the repo and click on the “Deploy to Azure” button.

## 5 Connecting to the Cloudera Cluster

The machines are named according to a specific pattern. The master node is named based on parameters and using the nomenclature:

```
[dnsNamePrefix]-mn0.[region].cloudapp.azure.com
```

If the `dnsNamePrefix` was `clouderatest` in the West US region, the machine will be located at:

```
clouderatest-mn0.westus.cloudapp.azure.com
```

The rest of the master nodes and data nodes of the cluster use the same pattern, with `-mn` and `-dn` extensions followed by their number. For example:

```
clouderatest-mn0.westus.cloudapp.azure.com
```

```
clouderatest-mn1.westus.cloudapp.azure.com
```

```
clouderatest-mn2.westus.cloudapp.azure.com
```

```
clouderatest-dn0.westus.cloudapp.azure.com
```

```
clouderatest-dn1.westus.cloudapp.azure.com
```

```
clouderatest-dn2.westus.cloudapp.azure.com
```

To connect to the master node via SSH, use the username and password used for deployment

```
ssh testuser@[dnsNamePrefix]-mn0.[region].cloudapp.azure.com
```

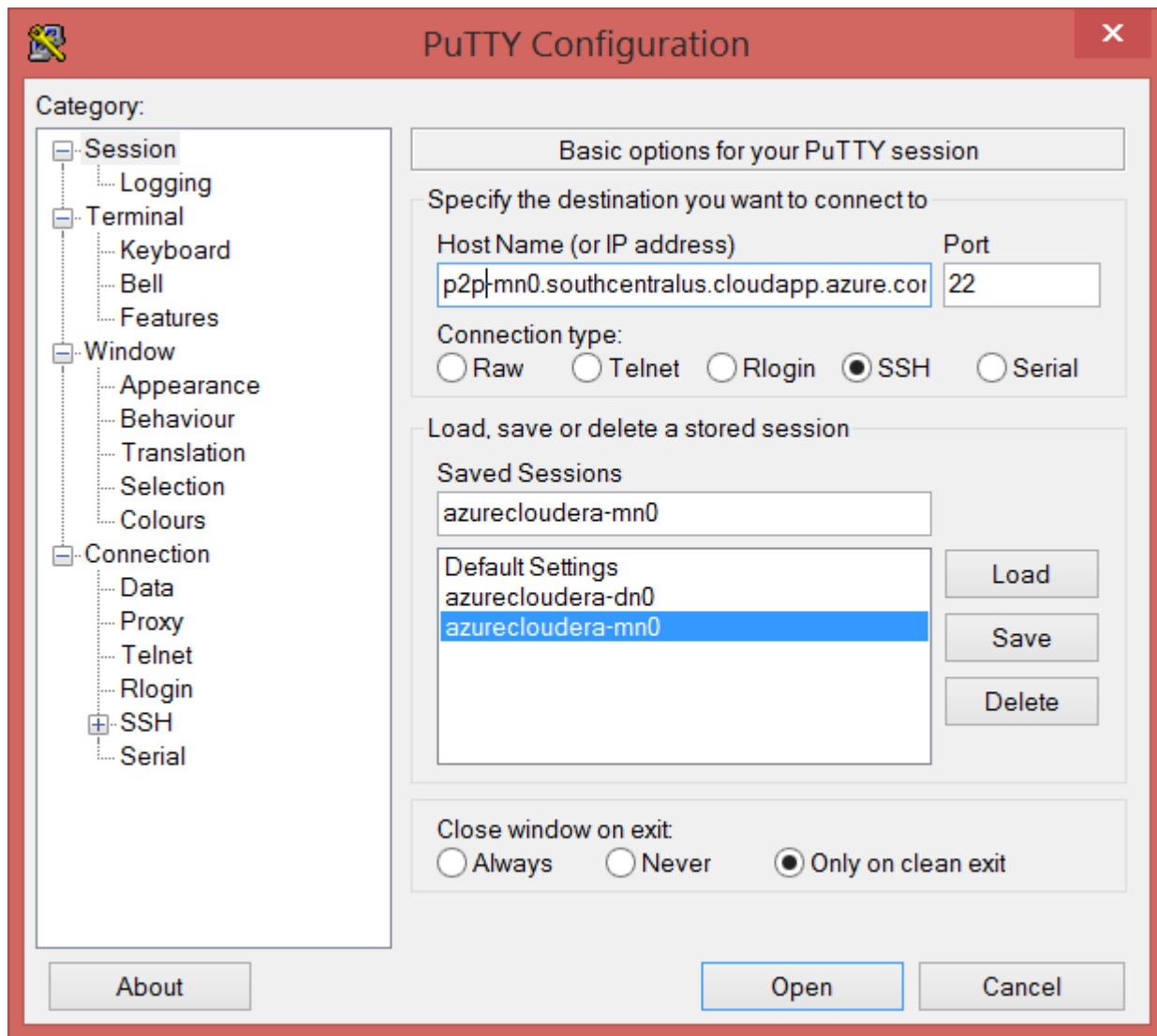
Once the deployment is complete, you can navigate to the Cloudera portal to watch the operation and track its status. Be aware that the portal dashboard will report alerts since the services are still being installed.

```
http://[dnsNamePrefix]-mn0.[region].cloudapp.azure.com:7180
```

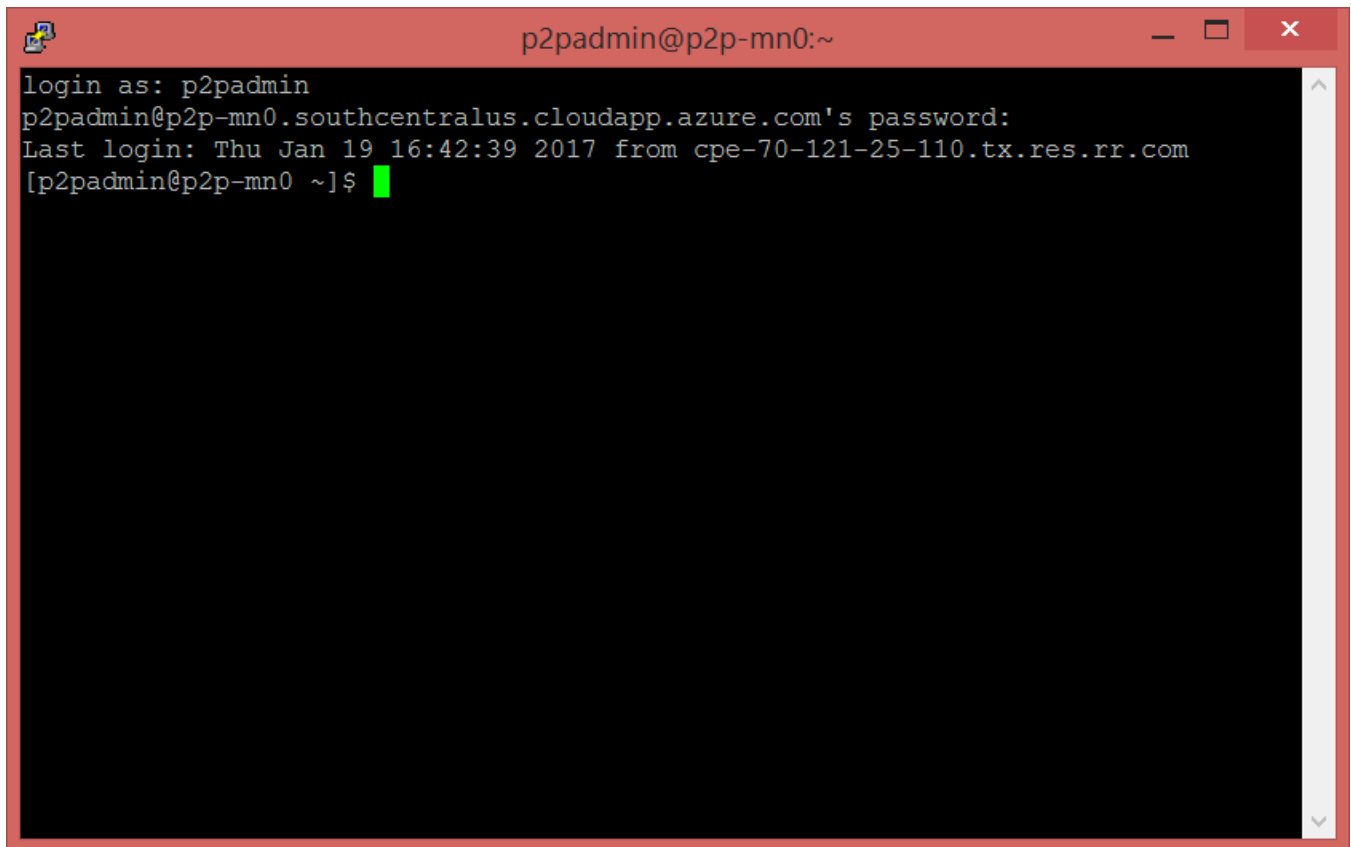
## 6 Loading Sample Data

Sample data can be loaded into Cloudera Impala and viewed via a Tableau dashboard. The following steps can only be executed after all Cloudera and Tableau servers have deployed successfully. In general, if all deployments are successful for the resource group, then Cloudera and Tableau should have deployed successfully.

To generate and load the sample data, connect to the "-mn0" Cloudera master node (referenced above) using PuTTY or another SSH client tool. The user name and password used when setting up the Cloudera deployment should be used.





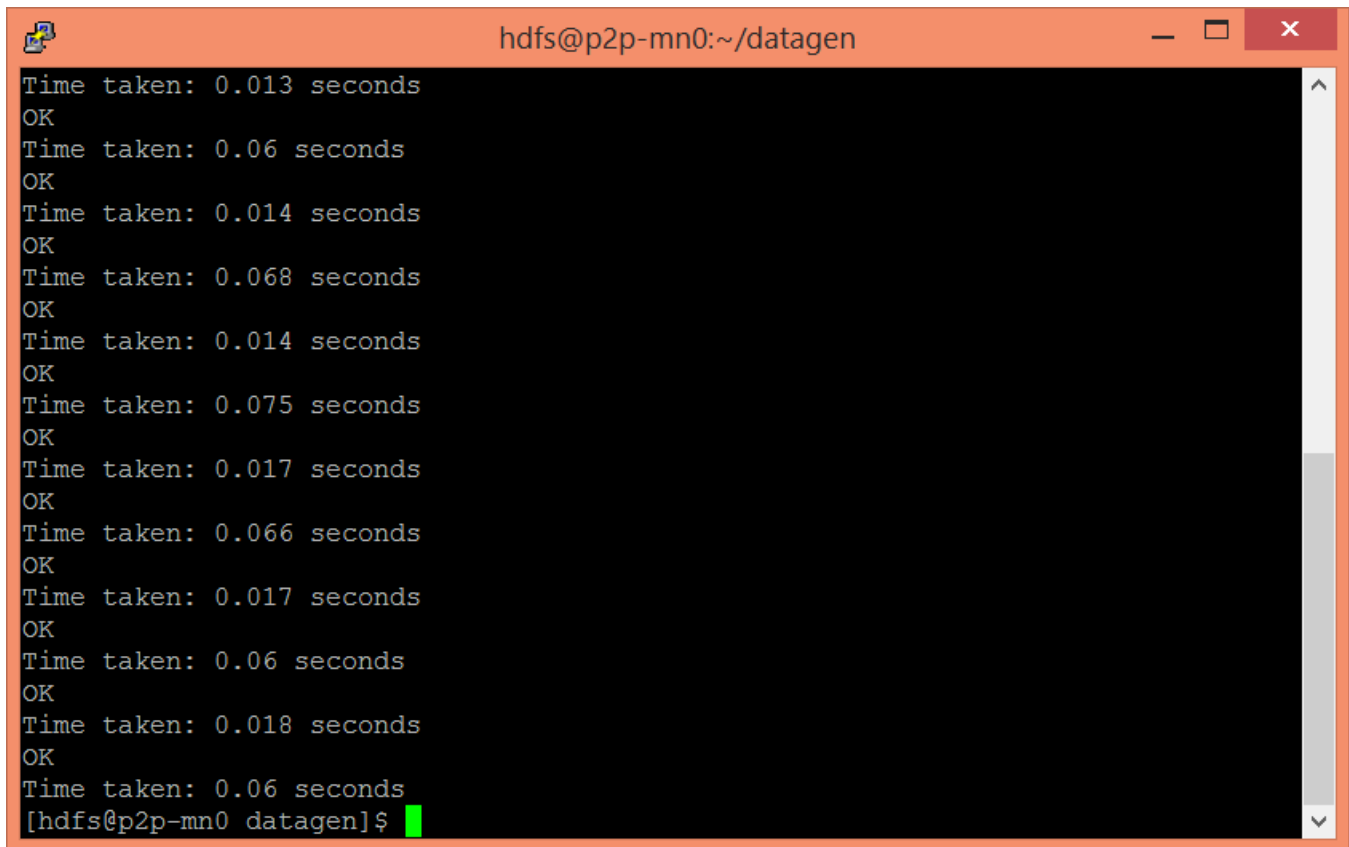


A terminal window titled 'p2padmin@p2p-mn0:~' with standard window controls. The terminal output shows a successful login for the 'p2padmin' user. The prompt is '[p2padmin@p2p-mn0 ~]\$' followed by a green cursor. The login details include the username, the host 'p2padmin@p2p-mn0.southcentralus.cloudapp.azure.com', and the last login time and IP address.

```
p2padmin@p2p-mn0:~  
login as: p2padmin  
p2padmin@p2p-mn0.southcentralus.cloudapp.azure.com's password:  
Last login: Thu Jan 19 16:42:39 2017 from cpe-70-121-25-110.tx.res.rr.com  
[p2padmin@p2p-mn0 ~]$
```

Execute the following commands via the command line:

- `sudo su - hdfs`
- `wget https://clouderatableau.blob.core.windows.net/datagen/datagen.tar.gz`
- `tar -xzf datagen.tar.gz`
- `cd datagen`
- `sh datagen.sh 2`

A terminal window with an orange title bar. The title bar contains the text 'hdfs@p2p-mn0:~/datagen' and standard window control buttons (minimize, maximize, close). The terminal area has a black background with white text. It displays a sequence of 15 lines, each consisting of 'Time taken: [value] seconds' followed by 'OK' on the next line. The values are: 0.013, 0.06, 0.014, 0.068, 0.014, 0.075, 0.017, 0.066, 0.017, 0.06, 0.018, and 0.06. The last line shows the prompt '[hdfs@p2p-mn0 datagen]\$' followed by a green cursor. A vertical scrollbar is on the right side of the terminal area.

```
hdfs@p2p-mn0:~/datagen
Time taken: 0.013 seconds
OK
Time taken: 0.06 seconds
OK
Time taken: 0.014 seconds
OK
Time taken: 0.068 seconds
OK
Time taken: 0.014 seconds
OK
Time taken: 0.075 seconds
OK
Time taken: 0.017 seconds
OK
Time taken: 0.066 seconds
OK
Time taken: 0.017 seconds
OK
Time taken: 0.06 seconds
OK
Time taken: 0.018 seconds
OK
Time taken: 0.06 seconds
[hdfs@p2p-mn0 datagen]$
```

Next, connect to the "-dn0" Cloudera worker node (referenced above) using PuTTY or another SSH client tool. Execute the following commands via the command line:

- `sudo su - hdfs`
- `wget https://clouderatableau.blob.core.windows.net/datagen/datagen.tar.gz`
- `tar -xzf datagen.tar.gz`
- `cd datagen`
- `sh load_data.sh`

```

hdfs@p2p-dn0:~/datagen
+-----+
Fetched 1 row(s) in 0.91s
Query: compute stats tpch_parquet.region
+-----+
| summary |
+-----+
| Updated 1 partition(s) and 3 column(s). |
+-----+
Fetched 1 row(s) in 0.61s
Query: compute stats tpch_parquet.supplier
+-----+
| summary |
+-----+
| Updated 1 partition(s) and 7 column(s). |
+-----+
Fetched 1 row(s) in 0.81s
Query: invalidate metadata
Query submitted at: 2017-01-19 23:23:31 (Coordinator: http://p2p-dn0.southcentralus.cloudapp.azure.com:25000)
Query progress can be monitored at: http://p2p-dn0.southcentralus.cloudapp.azure.com:25000/query_plan?query_id=a44876977032834c:9de57d7700000000
Fetched 0 row(s) in 4.37s
[hdfs@p2p-dn0 datagen]$

```

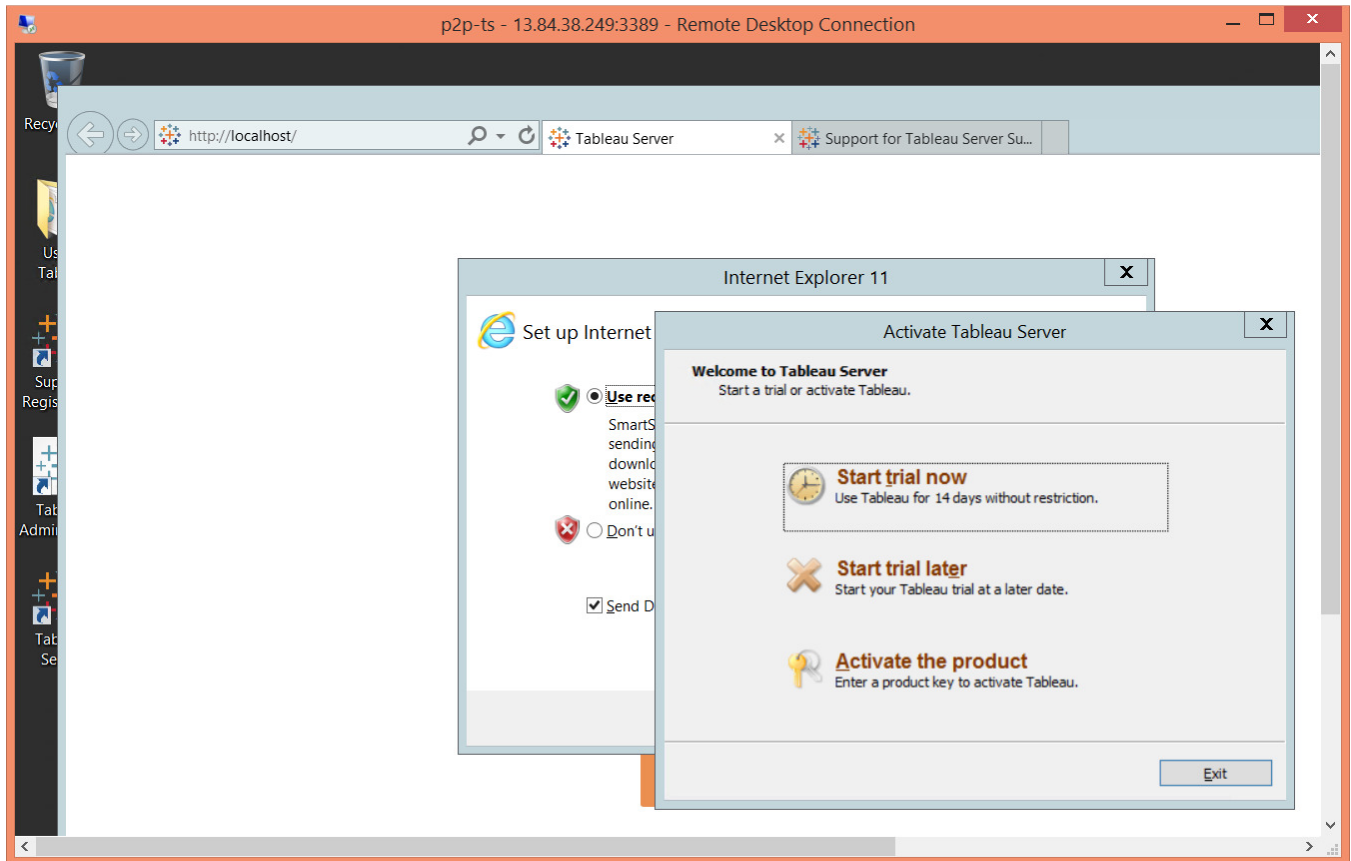
The sample data should now be accessible in Hadoop Hive (tpch\_text\_2 database) and Cloudera Impala (tpch\_parquet database). This can be validated using the Hue Query Editor.

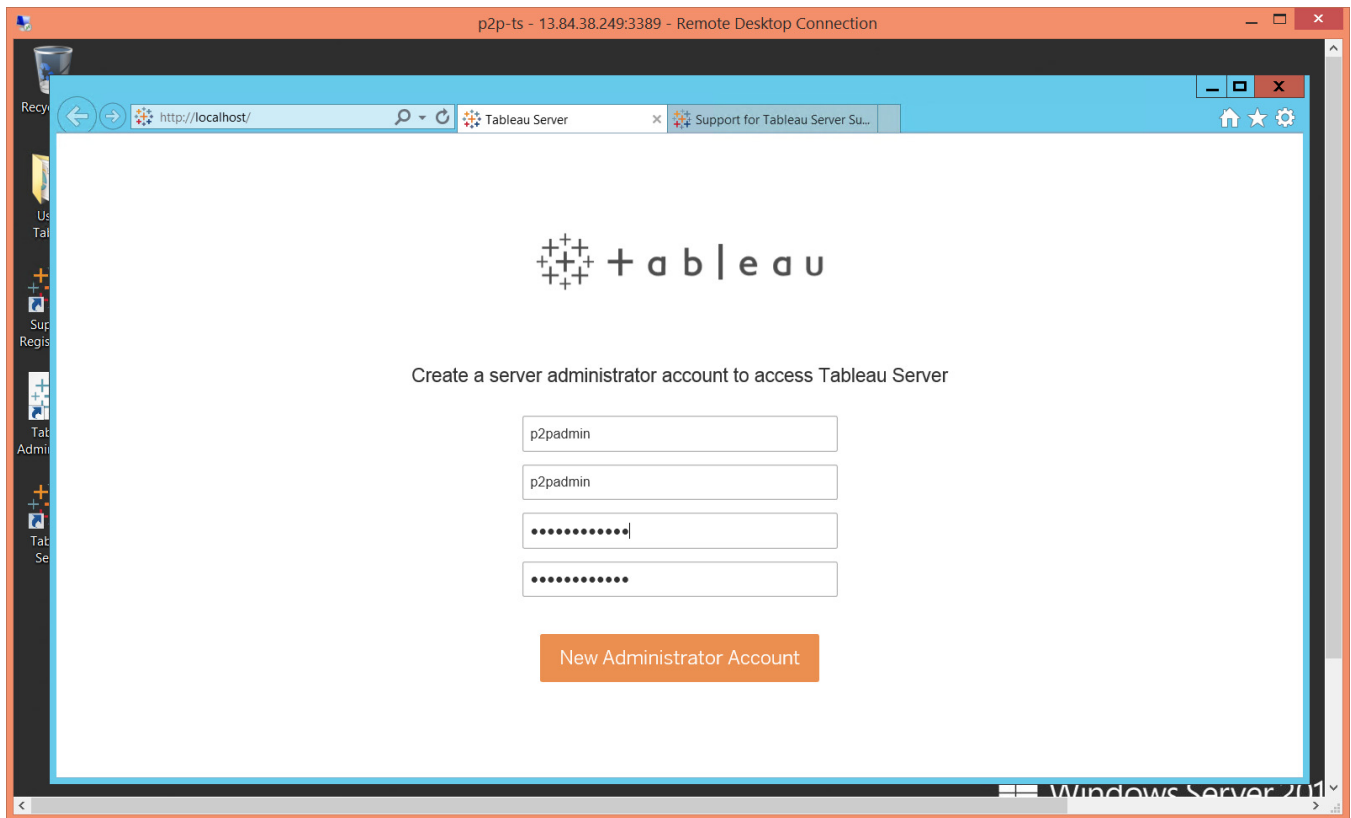
## 7 Viewing Sample Data with Tableau

Next, using the Microsoft Azure portal, remote into the Tableau server using the "Connect" button for the Tableau Virtual Machine (VM). This will establish an RDP session into the Tableau Windows Server. Make sure to use a different account than your default Windows account. Also, use this login format: <computer name>\<azureuser>

The screenshot displays the Microsoft Azure portal interface for a virtual machine named 'gicldtabdns3-ts'. The left sidebar shows navigation options like Overview, Activity log, Access control (IAM), Tags, and Diagnose and solve problems. The main pane shows the VM's status as 'Running' and its location as 'East US 2'. Below this, there are monitoring graphs for CPU percentage (showing a peak of 1.2%) and Network in and out. A Windows Security login dialog is overlaid on the right, prompting for credentials to connect to the VM. The dialog shows the computer name 'gicldtabdns3-ts' and the domain 'gicldtab3'. The user 'gicldtab3\azureuser' is entered, and the 'Remember me' checkbox is unchecked. The 'Use a different account' option is highlighted at the bottom of the dialog.

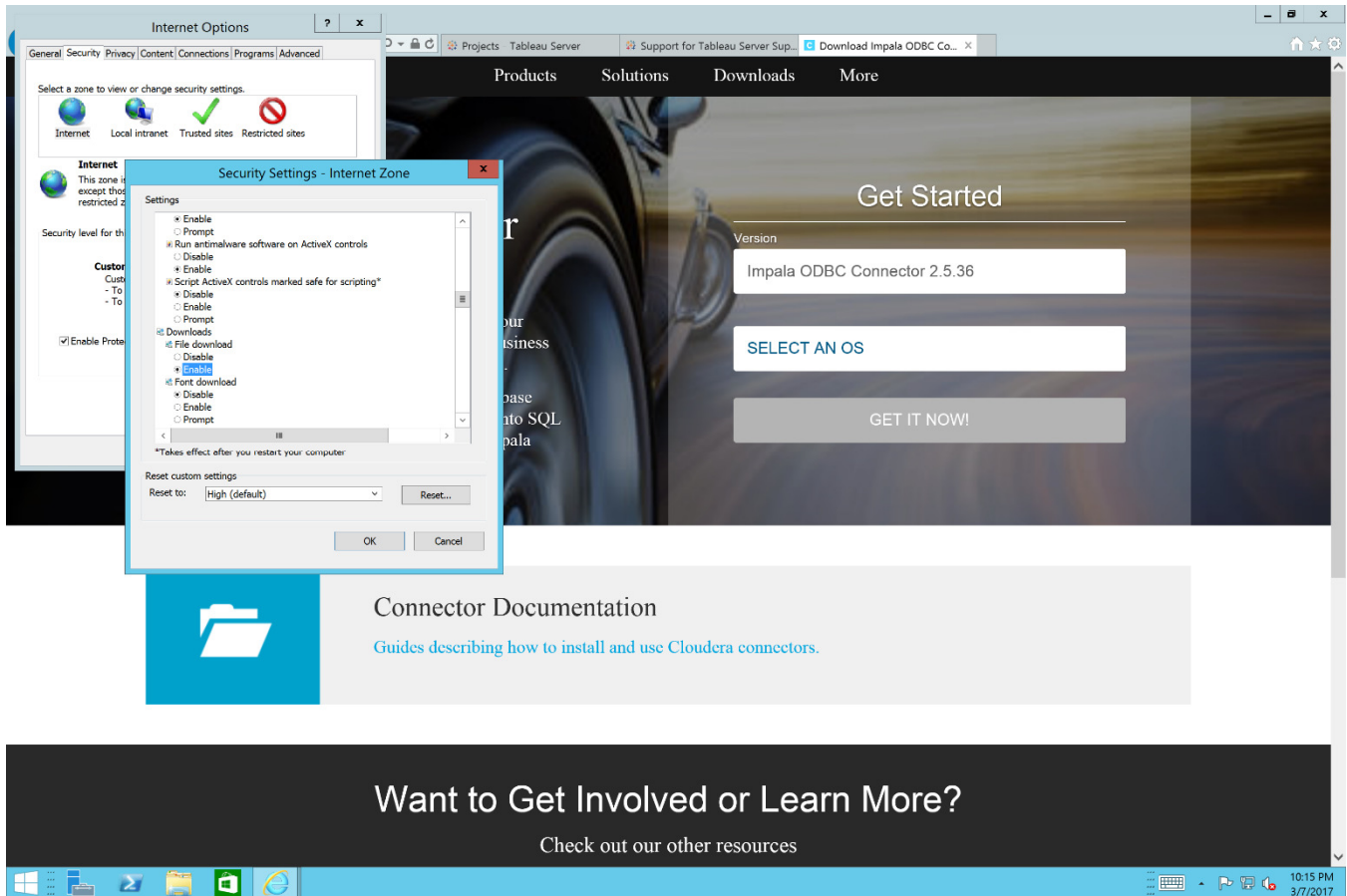
Follow the Tableau registration process and either start a trial or enter the appropriate Tableau license key. This step must be completed before the dashboard can be deployed.



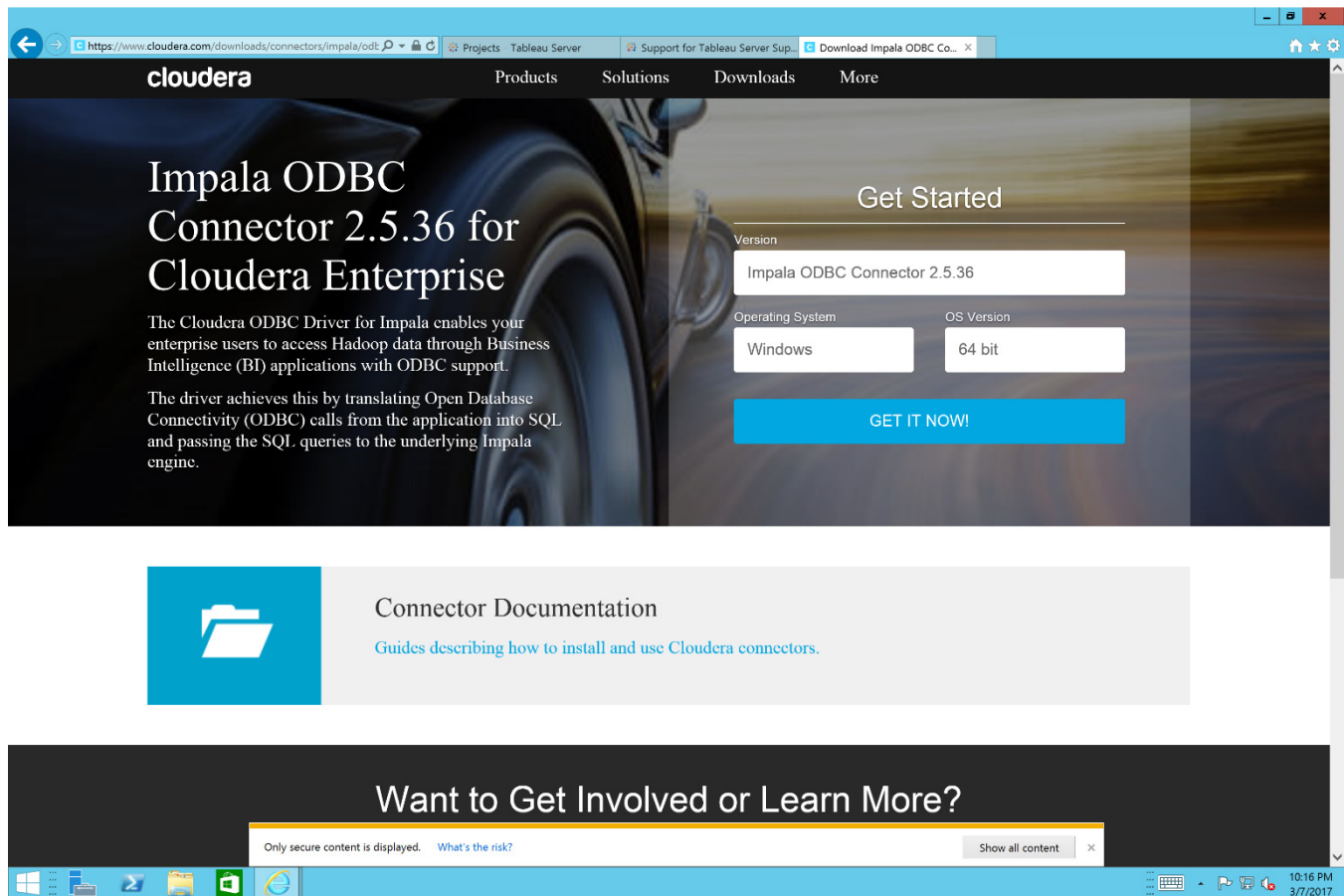


To complete the process of viewing the Cloudera Impala sample data with a Tableau dashboard, install the Cloudera Impala driver for Windows on the Tableau server:

- 1) Set IE to allow file downloads by selecting the following options in the browser: Internet Options → Security → Internet → Custom Level → Downloads → File Download → Enable.



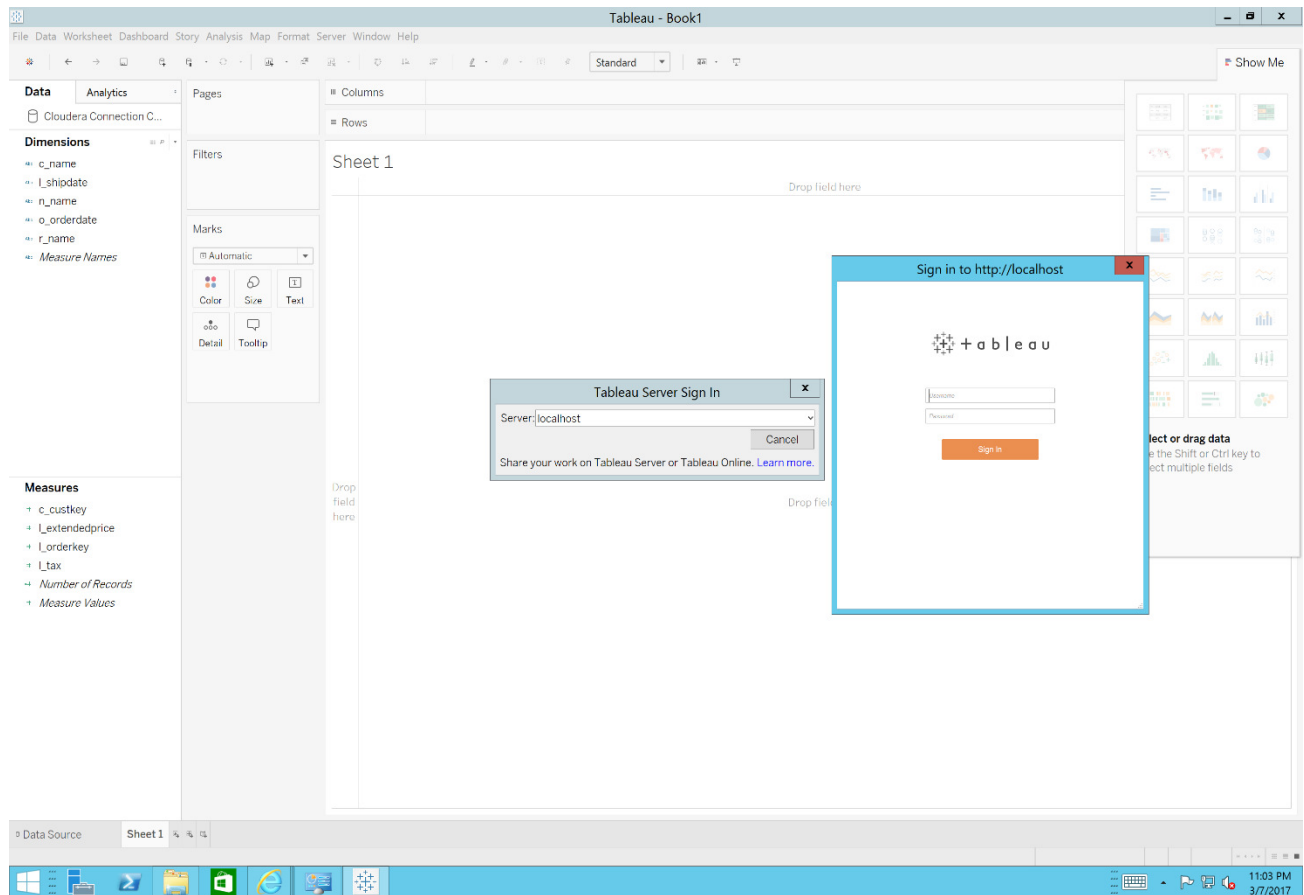
- 2) Navigate to <http://www.cloudera.com/downloads/connectors/impala/odbc.html> and select the Windows 64-bit driver.



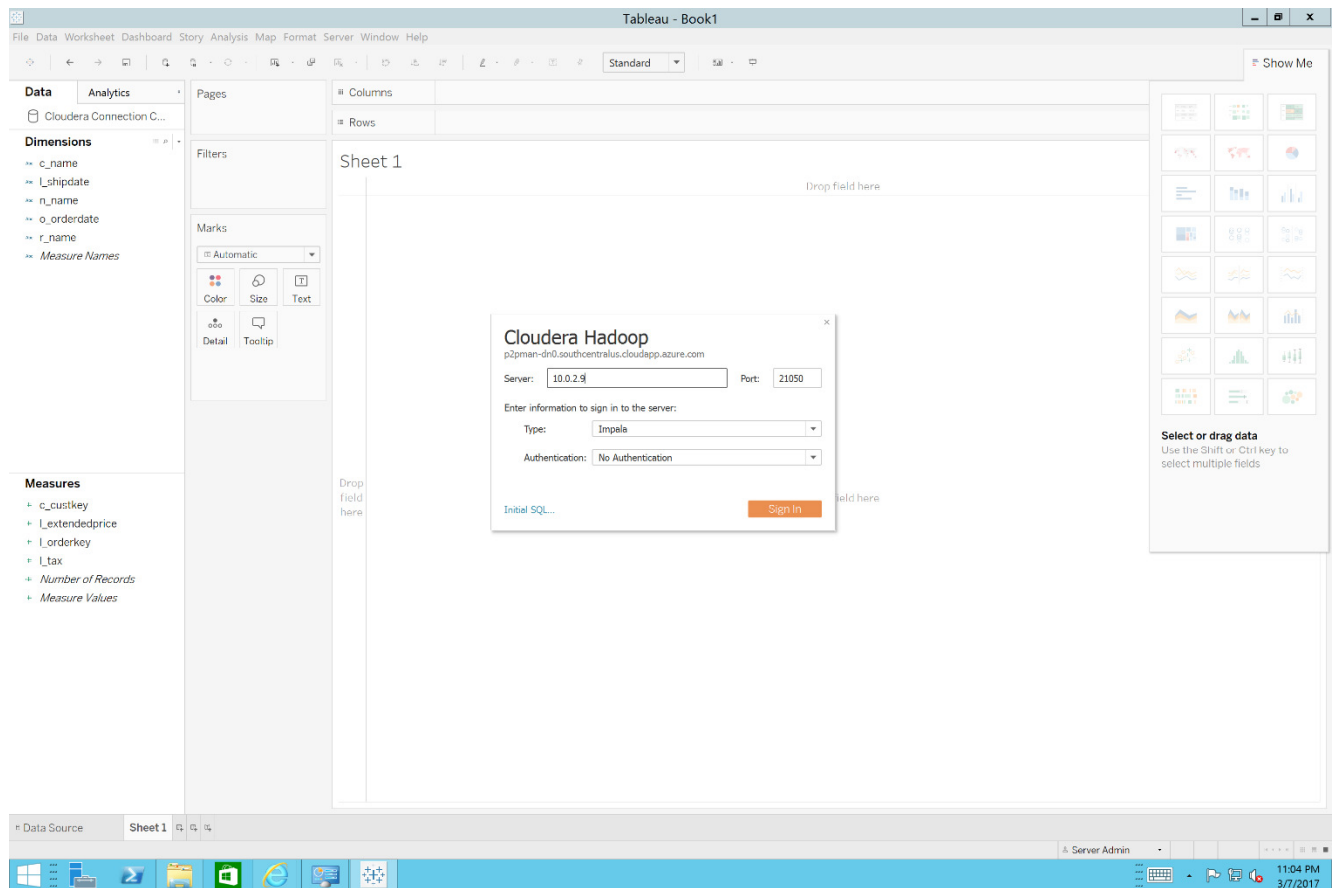
- 3) Follow the registration process and download the driver. Double-click on it to install.
- 4) Download the sample Tableau data source and workbook from <https://github.com/Azure/azure-quickstart-templates/tree/master/cloudera-tableau/tableau> and save them on the Tableau Server VM:
  - a. Cloudera Connection Custom SQL.tdsx
  - b. Cloudera Widget Dashboard.twbx



- 5) Install the latest version of Tableau Desktop 10.1 from <http://www.tableau.com/support/esdalt>.
- 6) Open Tableau Desktop and login to the Tableau Server.



- 7) Using Tableau Desktop, open the "Cloudera Connection Custom SQL" data source file.
  - a. In Tableau Desktop, choose Server -> Publish Data Source -> Cloudera Connection Custom SQL.
  - b. Use the instructions in Step #8 to get the Private IP Address for the Cloudera -dn0 VM.
  - c. Edit the data source to point to the Cloudera "-dn0" private IP address.
  - d. Publish the Custom SQL to the 'Samples' project on the Tableau Server.



- 8) To get the private IP address of the “-dn0” server:
  - a. In the Microsoft Azure Portal, open the Overview blade for the “dn0” server VM.
  - b. Click on the Network Interfaces blade and click on the machine name to view the Overview blade for the network interface.
  - c. Click on the “Click to copy” icon to copy the private IP address.

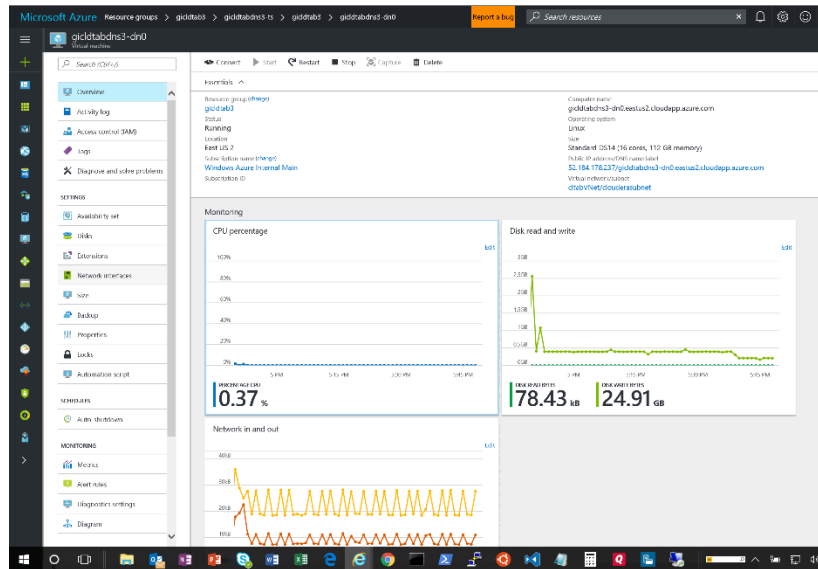


Figure 1 - VM Overview

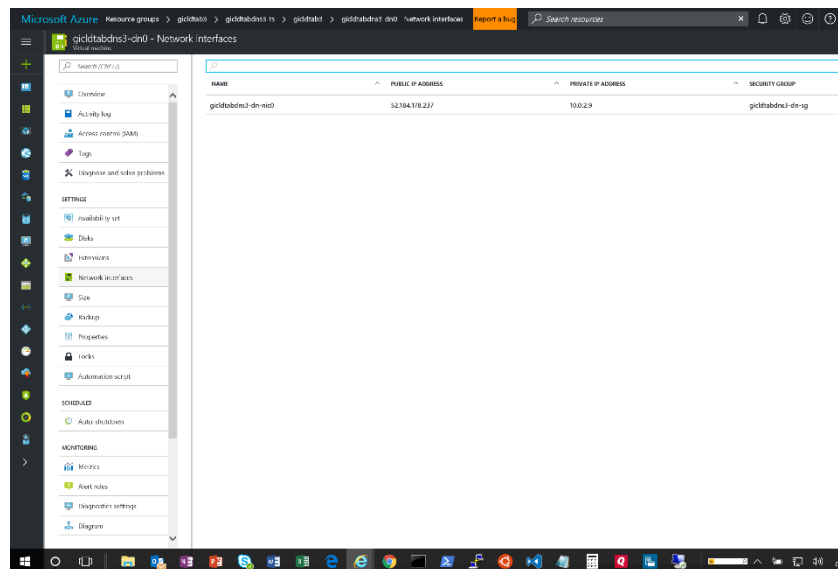
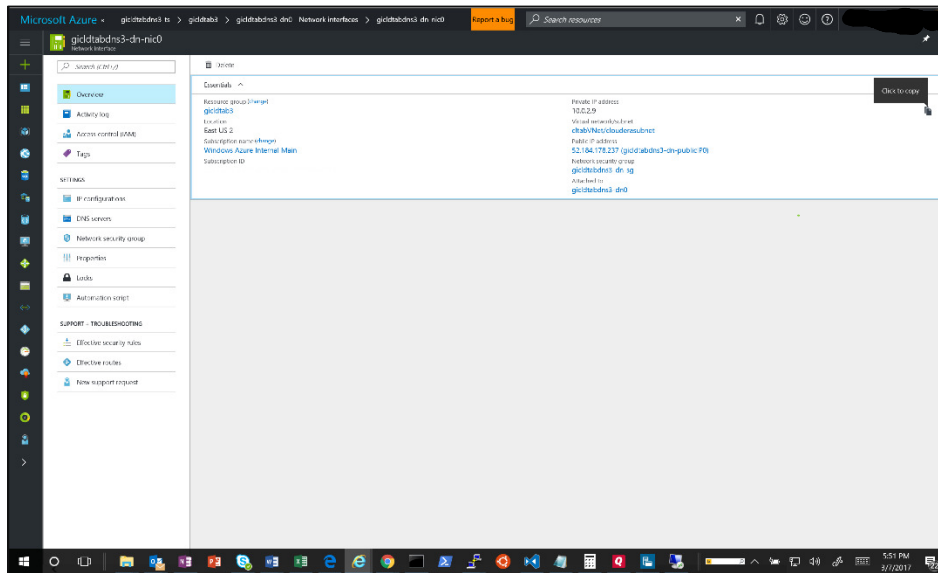
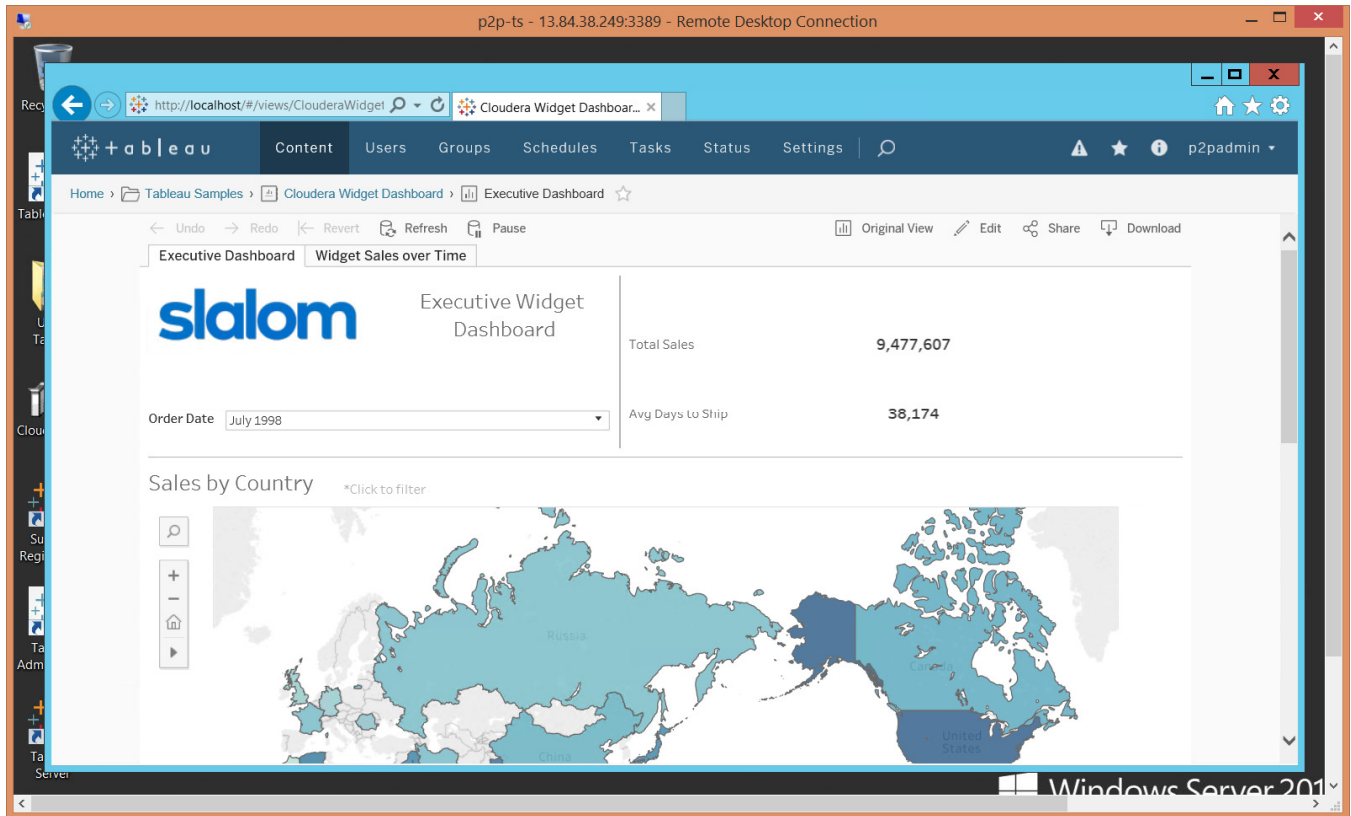


Figure 2 - VM Network Interfaces



**Figure 3 - VM NIC Overview**

- 9) Once the data source is deployed to the server, open the "Cloudera Widget Dashboard.twbx" file in Tableau Desktop. The "Executive Dashboard" and "Widget Sales over Time" worksheets should populate with data.
- 10) Choose Server -> Publish Workbook and deploy the workbook to the Tableau Server "Tableau Samples" project.
- 11) Click on the "Cloudera Widget Dashboard" under the "Tableau Samples" project to view the Cloudera Impala sample data.





If you would like more information or to discuss your project needs with a Slalom consultant, please contact us at - [AzureMarketplace@slalom.com](mailto:AzureMarketplace@slalom.com)