Accelerate Cloud Analytics on Azure with Paxata

INSTALLING PAXATA ON AZURE MARKETPLACE
Welcome!

Be it self-service analytics or exploring and profiling your data lake, you have taken the right first step towards democratizing data and information in your organization by choosing to install Paxata in your Azure environment. Paxata customers turn raw data into information using Paxata’s Self-service data preparation application – Information that is clean, complete, contextual and consumable. Now with one-click install on Azure, and 30-day free trial license, evaluating Paxata just got a lot easier.

This document walks you through the necessary steps to provision Paxata on Azure within your virtual network.

PRE-REQUISITE

To install Paxata on Azure, you need a Microsoft account and Azure subscription. It is simple to create your account and create a pay-as-you-go subscription. If you work for a larger company, you may want to ask your IT administrator to set it up for you and provision a user account for you with access to the subscription.

MINIMUM REQUIREMENTS

1. Cluster Type and Version: To install Paxata, you need a HDInsight Spark Cluster (Version Spark 2.1/ HDInsight 3.6)
2. Hardware Requirements:
   a. Worker CPU Cores:
i. To install Paxata, you will need to have **at least a 4 CPU Worker Cores**. We recommend getting a D or D-V2 Series hardware.

ii. If you plan to run larger interactive workloads (say 10 Million Rows/1 GB), you may increase the number of workers up to 32 worker cores. This will give you good interactive response time.

iii. If you are getting 32 cores, it is better to get four 8-core VMs. This allows for higher redundancy in case there are worker core failures. At the same time with 4 cores, you will not pay much of Spark *shuffle cost*.

b. Memory

i. Each worker node must have at least 14 GB of memory. If you attempt to install Paxata on a cluster with less than 14 GB worker memory, you will get an error message.

**INSTALLING PAXATA - THREE INSTALL PATHS**

There are three paths one can take on Azure portal to install Paxata. First visit Azure portal at [portal.azure.com](http://portal.azure.com) and login with your Microsoft Account.

1. First install a HDInsight spark cluster and then install Paxata. Or you can start from an existing HDInsight Cluster.
2. Search for Paxata on the portal and you will be guided to a wizard where you can install both a HDInsight Spark Cluster and Paxata
3. Start Installing a HDInsight Spark Cluster and you will be given the option to install Paxata along with it
INSTALL PAXATA ON TOP OF A NEW/ EXISTING CLUSTER

First ensure that the HDInsight cluster you have provisioned is suitable for Paxata. The image below shows the right cluster type you must have to install Paxata. If you have any other type of cluster, Paxata will be unavailable to be installed on the cluster.

1. Cluster name becomes part of the Paxata URL. For example, if you name the cluster as mycluster, the URL to access your Paxata installation will be https://mycluster-pax.apps.azurehdinsight.net/
2. In the cluster type, remember to select a Linux based Spark cluster with Spark version 2.1
3. Provide an administrator password that you remember. You will need this to access the edge node where Paxata will be installed.

The below image shows a cluster with 8 worker cores (2 x D12 v2). You could start with as little as 4 cores (1 x D14 v2). All HDInsight clusters come standard with two head nodes.
Once you step through the wizard and submit it, your cluster will be ready in 15-20 minutes.

Once you have provisioned a cluster, you can install Paxata. Click on the Applications link on your HDInsight Cluster. In the screenshot below, the applications link appears in two places. Clicking on either one will take you to the same place. This is called the “HDInsight Application Blade”.

HDInsight application blade is the place where you can see all the applications installed on top of the HDInsight cluster. In the below screenshot, you can see that there are no applications installed.

Click on + Add button on top and select “Self Service Data Preparation by Paxata”. This will launch the Paxata blade (screenshot below).
1. Once you are in Paxata blade, click on the “GET INSTALL KEY” link. This will take you to a Paxata web page with a registration form.

![Paxata registration form](image)

2. While register, be sure to provide a valid work email (e.g. joe.foo@microsoft.com).
   a. If your email is associated with company/organization that is a Paxata customer, prospect or a partner you will get an install key immediately (in less than 5 minutes).
   b. Else someone will have to manually review and approve your request. Once approved you will get an install key. This could take about 24 hours.

3. Check your email inbox. You will receive an install key along with a set of credentials from Paxata. Be sure to check the spam folder after 24 hours.
   a. Install key is a long string that typically looks like this: 3a6809fe-fcdf-4d8e-8ad8-bc7c48445a81

4. Once you have the install key, proceed back to the Azure portal > Paxata blade.
   a. Enter the key in the “License Key” field.
   b. Review the terms of use before accepting them. Click Purchase after you have reviewed the terms.
   c. Click OK in the Paxata blade.
d. Click **Next** on the application blade.

5. You should see a notification “... Installing apps to <cluster-name>”
6. Once the installation is complete you should see Paxata among the installed applications on your application blade.

7. Either the Portal Link next to the application or the web page link on the right-hand side will take you Self Service Data Preparation application from Paxata.
   a. If you are a technical user, you may want to note down the SSH URL. This URL is always listed in this page.
   b. Also, this may be a good time to check out some of the Useful Links, especially the Tip of the Day link.

8. Click on the above link to go to Paxata
9. Use the credentials sent to you in the first email (with install key) to log in to Paxata.

SUCCESSFUL INSTALLATION
Once the installation is complete you will receive a welcome email stating that your “30-day free trial starts today”. This email comes with YouTube Tip of the day video.

You will receive a few more emails articulating how to use, administer the product to successfully complete your functional evaluation of Paxata. We are also constantly in work to add more videos and bringing you access to our administration guides. Stay tuned.

ERROR CONDITIONS – DURING INSTALLATION
While installing Paxata you can run into four possible error conditions. If any of the error condition occurs, instead of seeing Paxata login screen (image above), you will see a static web page with an error message (image below).

1. Installation Key already used
   a. All install keys are valid only for one-time. Please register again to get another installation key. Paxata customers,
prospects and partners can get as many install keys as they want.

2. Installation Key Expired
   a. All install keys are valid for only 30 days from the day you receive them. Please register again to get another installation key.

3. Invalid Installation Key
   a. If you tried to enter any arbitrary text (instead of an install key that was generated and sent to you) then you will get an invalid install key error.
   b. Simply register in our web site and use the installation key you receive from Paxata.

4. Internal Error
   a. Most likely reason for this is Paxata installation files did not download properly.
   b. If you see this error, you can reuse the installation key. Installation key is not marked as used if the installation failed due to an internal error.

INSTALLING PAXATA ALONG WITH THE HDINSIGHT CLUSTER

Instead of installing Paxata on top on an existing cluster, you can install Paxata along with the HDInsight cluster at the same time. There are two ways to do it.
1. Start from the marketplace home (portal.azure.com).
2. Click on the ✪ New icon on the left.
3. Search for “Paxata”
4. Click on Paxata

This will take you to a wizard where you can build a HDInsight Cluster and install Paxata. The choices you will see are the same as listed above, except you will have one combined wizard to deploy the cluster and install Paxata.
1. Click on the + icon on the left. 
2. Select Data and Analytics 
3. Select HDInsight 
4. This will take you to the HDInsight installation Wizard. Expand the wizard by clicking on “Custom” 

5. Select a cluster size, machine type and follow Paxata installation instructions as listed above (under existing cluster deployment section).