



Welcome to BIGDATA 210: Introduction to Data Engineering

Saptak Sen [saptak@uw.edu]

Agenda

- Setup and Concept - 1 week
- Data at Rest - 3 weeks
 - Hive
 - Spark
- Data in Motion - 3 weeks
 - NiFi
 - Kafka
- Massive Scale - 2 weeks
 - HBase
 - Phoenix
- Project Presentation - 1 week

Week 1

What we will do?

- Introductions
- Logistics
 - Topics
- Setup
 - Virtual Machine
 - Cloud
 - YARN Queues
- Hadoop Concepts

Introduce yourself:

- Your name
- Job responsibilities
- Previous programming experiences
 - What languages?
 - Previous Hadoop experience (if any)
- Your expectation for this class
- Favorite hobby

Logistics

- Weekly schedule and breaks
 - 6pm – 9pm every Wednesday except 1/13
 - 5 minute breaks every hour or so
- Restrooms
- Computers and the VM/Azure Environment
- Assignments
- Team Project

Setup

Download Hortonworks Sandbox

<http://hortonworks.com/sandbox>

The screenshot shows a web browser window with the URL hortonworks.com/products/hortonworks-sandbox/#install in the address bar. The page has a dark header with the Hortonworks logo and navigation links for Why Hortonworks, Products, Customers, Solutions, Training, Services, Developers, and Get Started. Below the header, a green banner features the title "Hortonworks Sandbox" and the subtitle "The easiest way to get started with Enterprise Hadoop". A laptop screen displays the Ambari interface. The main content area includes sections for "Download & Install", "Hortonworks Sandbox on a VM", and "HDP 2.3 on Hortonworks Sandbox".

Hortonworks

Why Hortonworks | Products | Customers | Solutions | Training | Services | Developers | Get Started

hortonworks.com/products/hortonworks-sandbox/#install

Blog | Partners | Contact | | | Login

PRODUCTS » SANDBOX

Hortonworks Sandbox

The easiest way to get started with Enterprise Hadoop

Overview Download & Install Tutorials Archive Q&A Forum

Download & Install

The Hortonworks Sandbox provides an easy way to get started to learn and develop with the Hortonworks Data Platform (HDP) anywhere. You can either run it in the cloud or your personal machine.

Hortonworks Sandbox on a VM

No data center, no cloud service and no internet connection needed! Full control of the environment. Easily extend with additional components or try the various Hortonworks technical previews. Always updated with latest edition.

HDP 2.3 on Hortonworks Sandbox

Runs on VirtualBox, VMware or Hyper-V

Try out the very latest features and functionality in Hadoop and its' ecosystem of projects with [HDP 2.3](#). Follow the [Step by Step Tutorials](#).

[System Requirements](#) | [Installation Steps](#) | [Release Notes](#)

INSTALL GUIDES

for VirtualBox
[Mac & Windows](#)

for VMware
[Mac & Windows](#)

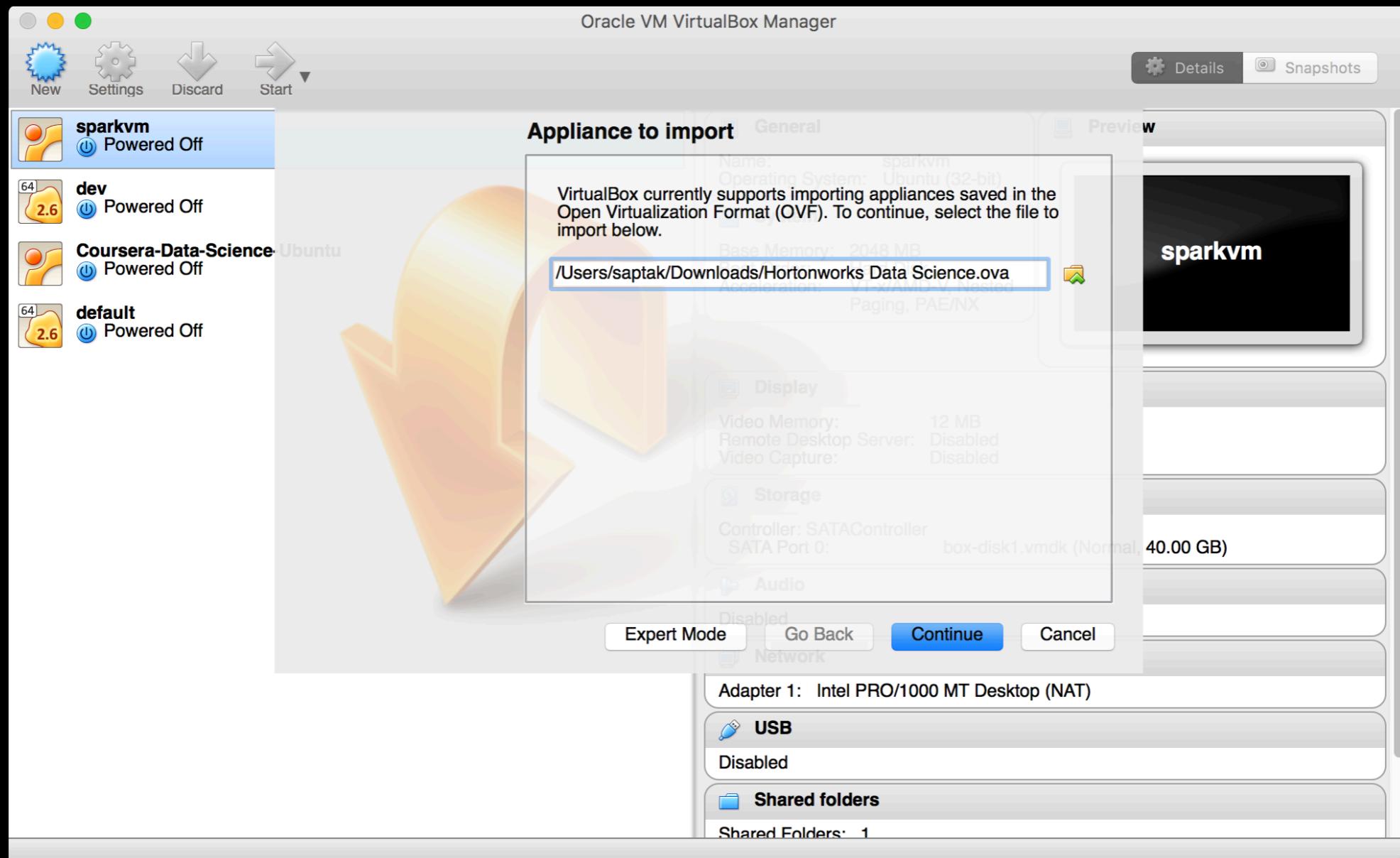
for Hyper-V
[Windows](#)

for VirtualBox
(HDP 2.3 - 6.9GB)

for VMware
(HDP 2.3 - 7.1 GB)

for Hyper-V
(HDP 2.3 - 6.8 GB)

Configure Hortonworks Sandbox

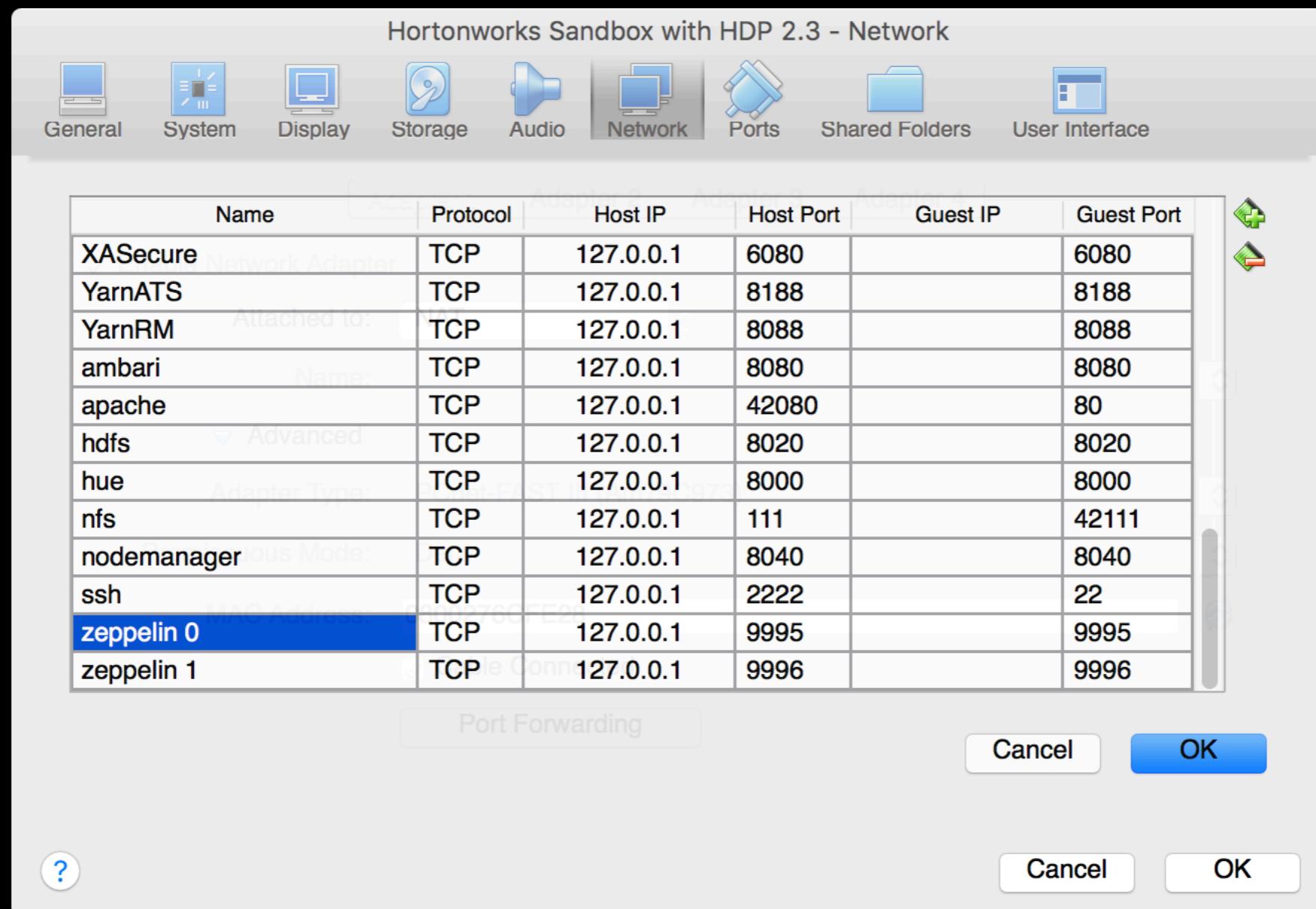


Open network settings for VM

Name	Protocol	Host IP	Host Port	Guest IP	Guest Port
WebHcat	TCP	127.0.0.1	50111		50111
WebHdfs	TCP	127.0.0.1	50070		50070
XASecure	TCP	127.0.0.1	6080		6080
YarnATS	TCP	127.0.0.1	8188		8188
YarnRM	TCP	127.0.0.1	8088		8088
ambari	TCP	127.0.0.1	8080		8080
apache	TCP	127.0.0.1	42080		80
hdfs	TCP	127.0.0.1	8020		8020
hue	TCP	127.0.0.1	8000		8000
nfs	TCP	127.0.0.1	111		42111
nodemanager	TCP	127.0.0.1	8040		8040
ssh	TCP	127.0.0.1	2222		22

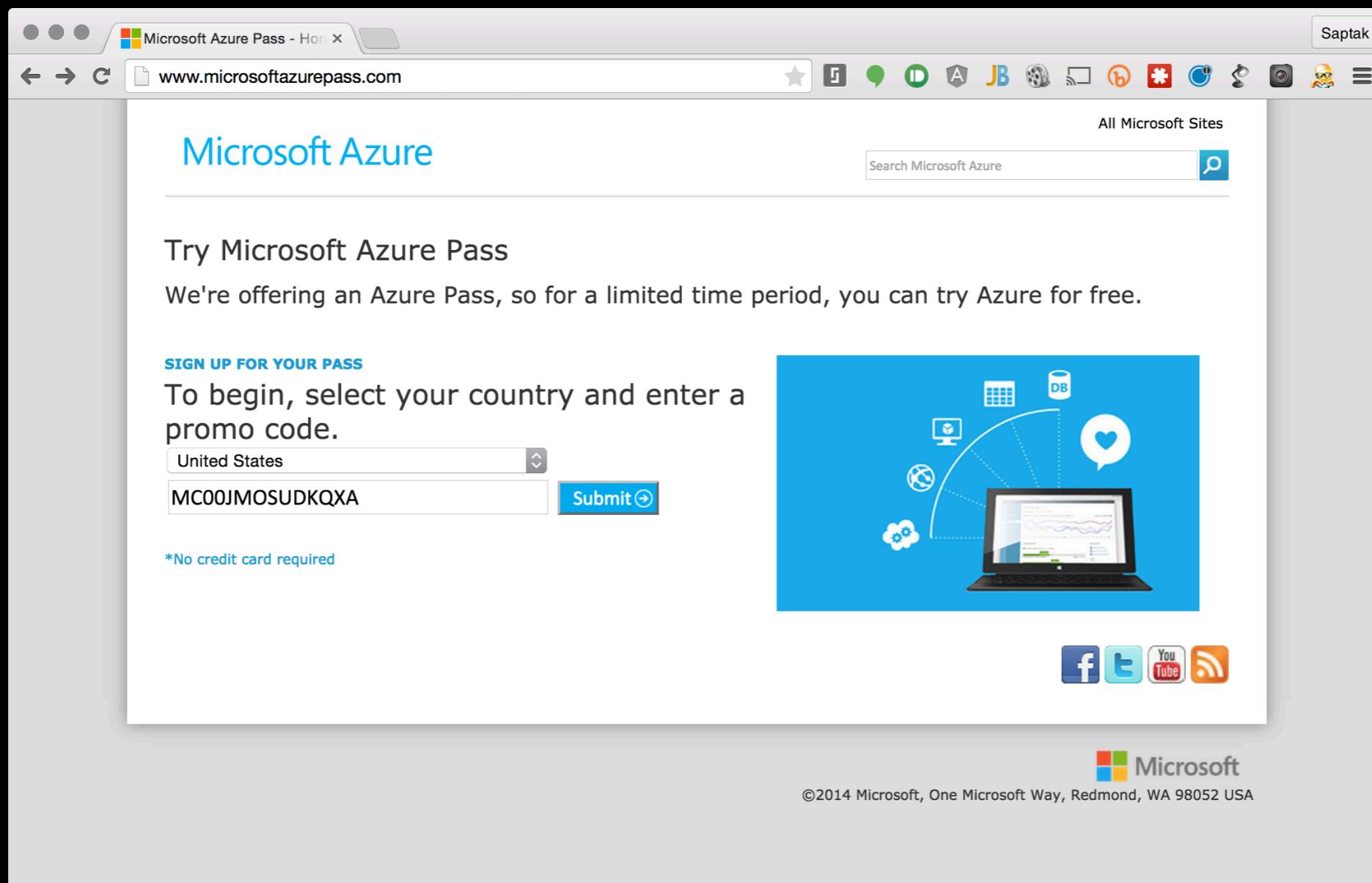
Cancel OK

Open port for Zeppelin

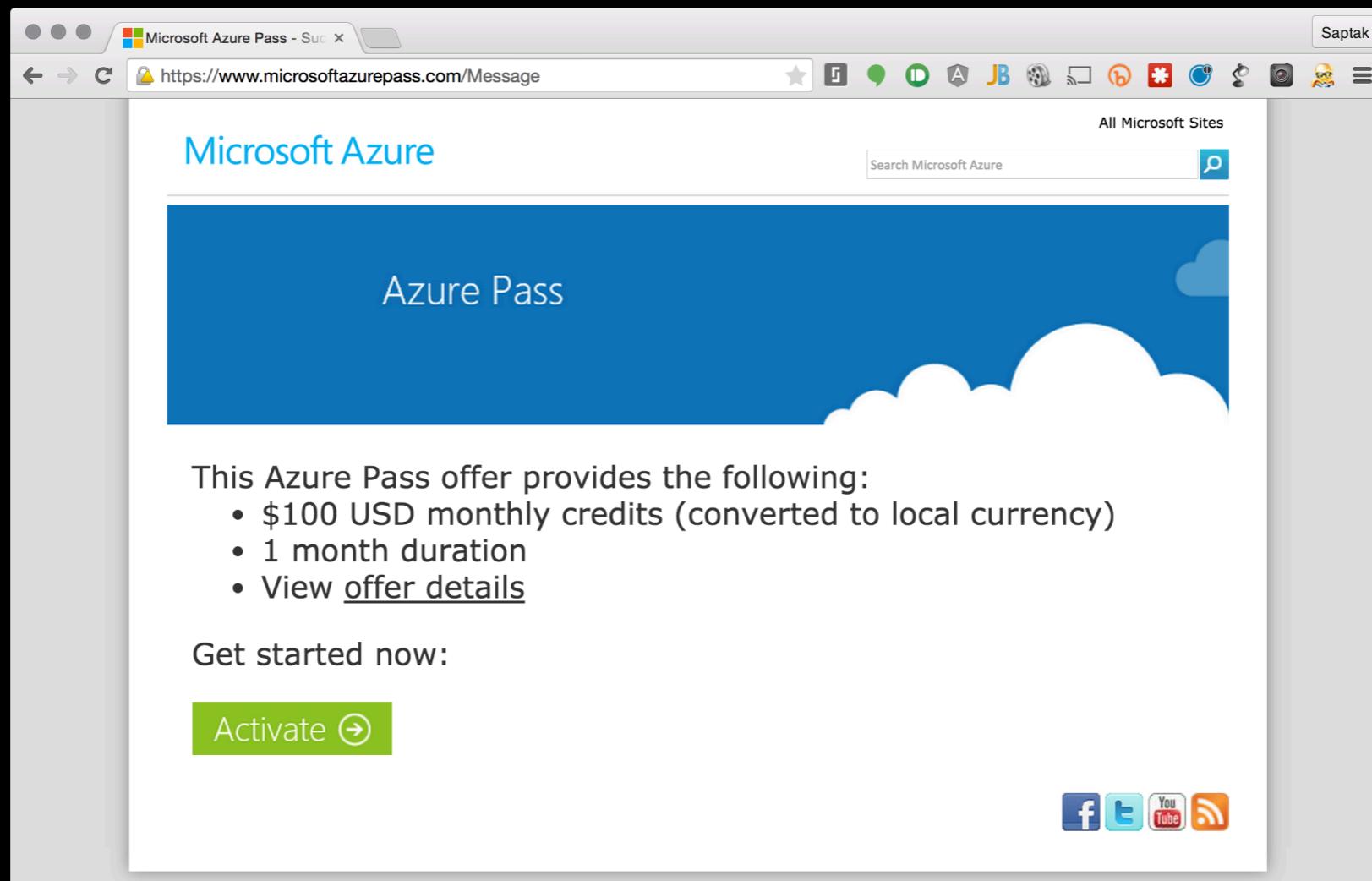


Configuring Hortonworks Sandbox on Azure

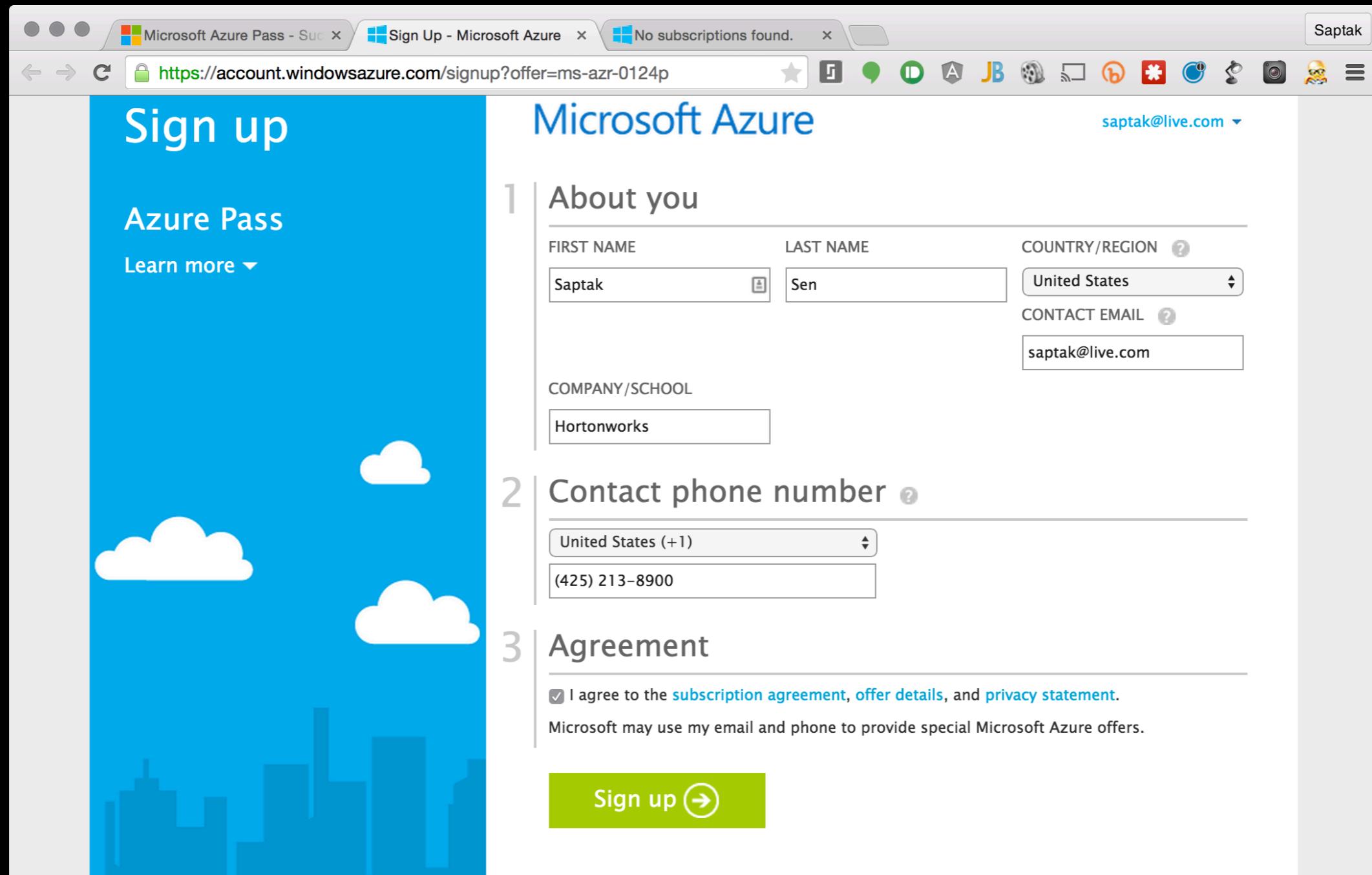
To use the Microsoft Azure Pass navigate to <http://www.microsoftazurepass.com>



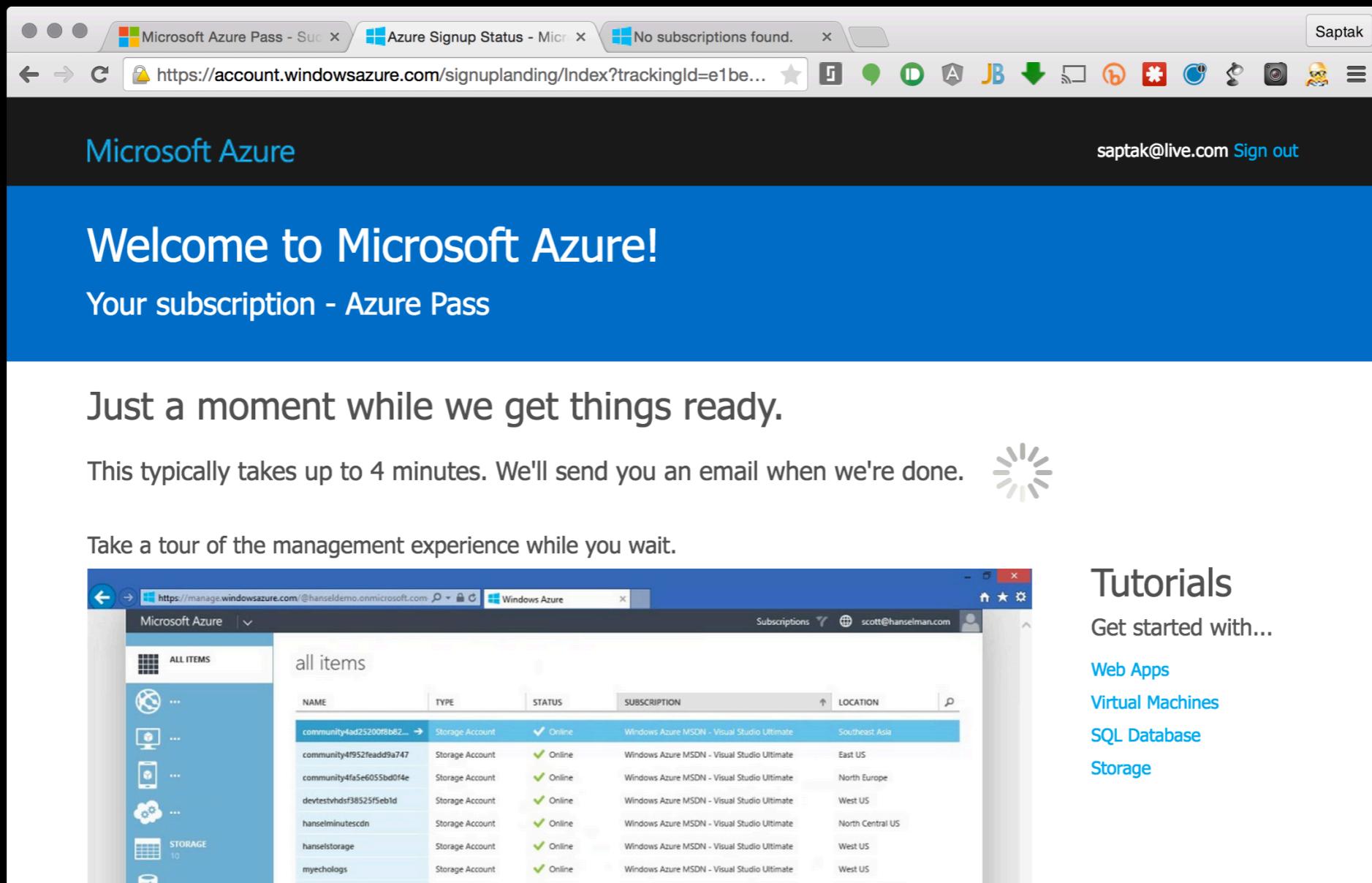
\$100 credit/month <-> Hortonworks
Sandbox node running on a A3 size VM for
about ~ 240 hours/month



Complete the signup:



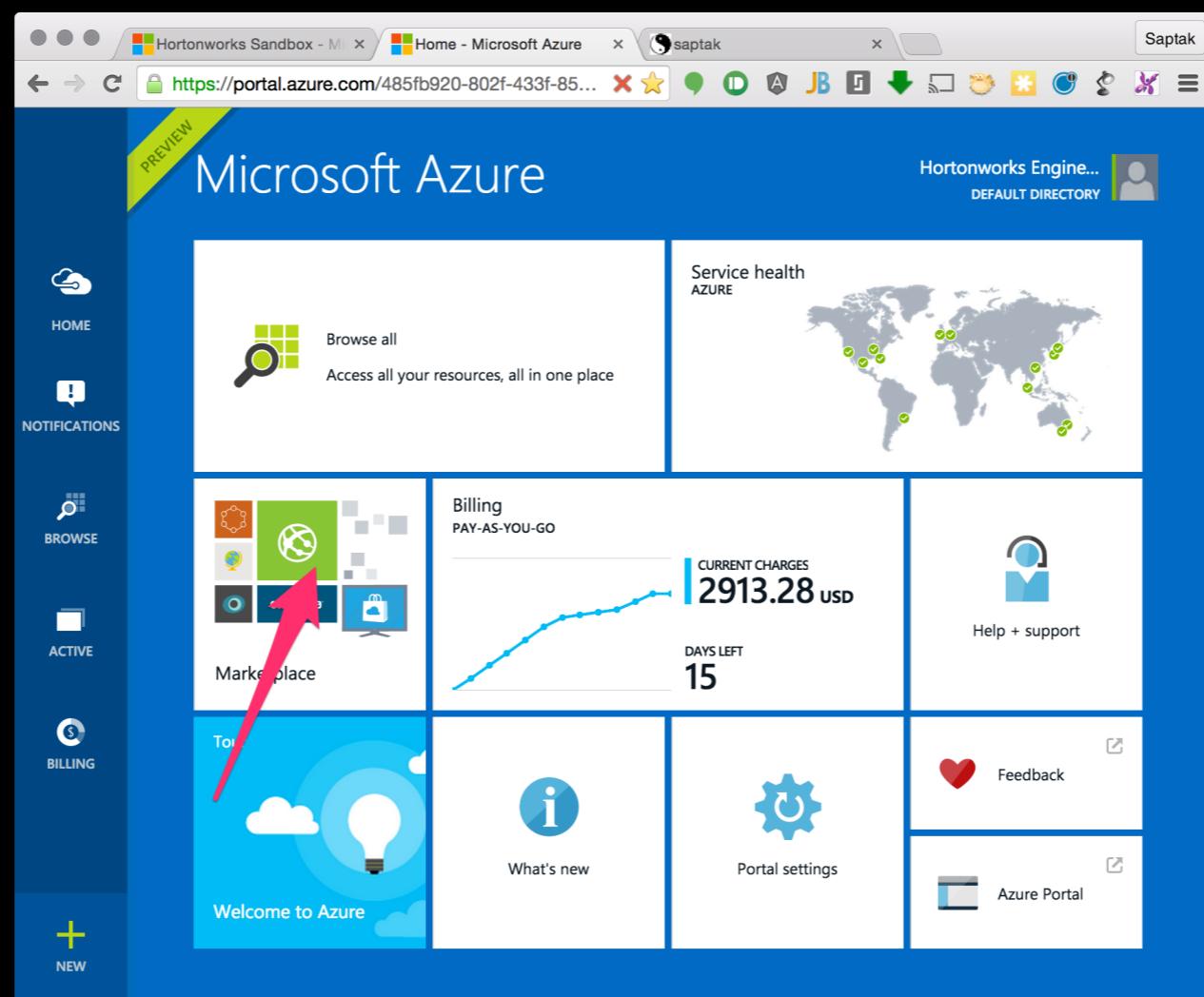
Wait, your Azure account will be ready in a few minutes



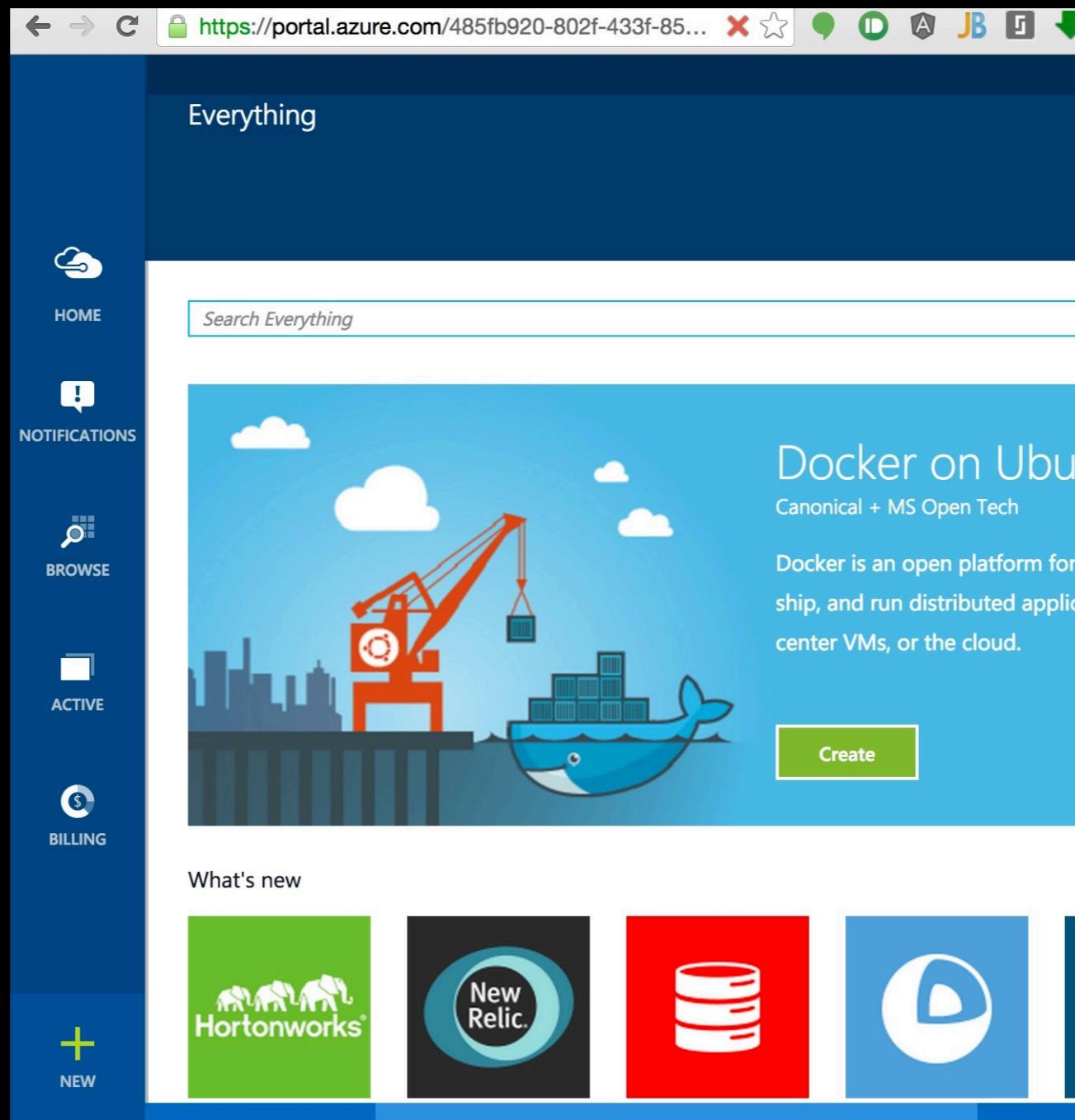
The screenshot shows a web browser window with three tabs at the top: "Microsoft Azure Pass - Suc...", "Azure Signup Status - Micr...", and "No subscriptions found.". The main content area is titled "Welcome to Microsoft Azure!" and "Your subscription - Azure Pass". It displays the message "Just a moment while we get things ready." followed by "This typically takes up to 4 minutes. We'll send you an email when we're done." A small loading icon is shown next to the message. Below this, there's a link to "Take a tour of the management experience while you wait." On the right side, there's a sidebar titled "Tutorials" with links to "Get started with...", "Web Apps", "Virtual Machines", "SQL Database", and "Storage". At the bottom left, there's a screenshot of the Windows Azure Management Portal showing a list of storage accounts.

Name	Type	Status	Subscription	Location
community4ad25200fb6b...	Storage Account	✓ Online	Windows Azure MSDN - Visual Studio Ultimate	Southeast Asia
community4f952feadd9a747	Storage Account	✓ Online	Windows Azure MSDN - Visual Studio Ultimate	East US
community4fa5e6055bd0f4e	Storage Account	✓ Online	Windows Azure MSDN - Visual Studio Ultimate	North Europe
devtestvhdf38525f5eb1d	Storage Account	✓ Online	Windows Azure MSDN - Visual Studio Ultimate	West US
hanselminutescdn	Storage Account	✓ Online	Windows Azure MSDN - Visual Studio Ultimate	North Central US
hanselstorage	Storage Account	✓ Online	Windows Azure MSDN - Visual Studio Ultimate	West US
myechologs	Storage Account	✓ Online	Windows Azure MSDN - Visual Studio Ultimate	West US

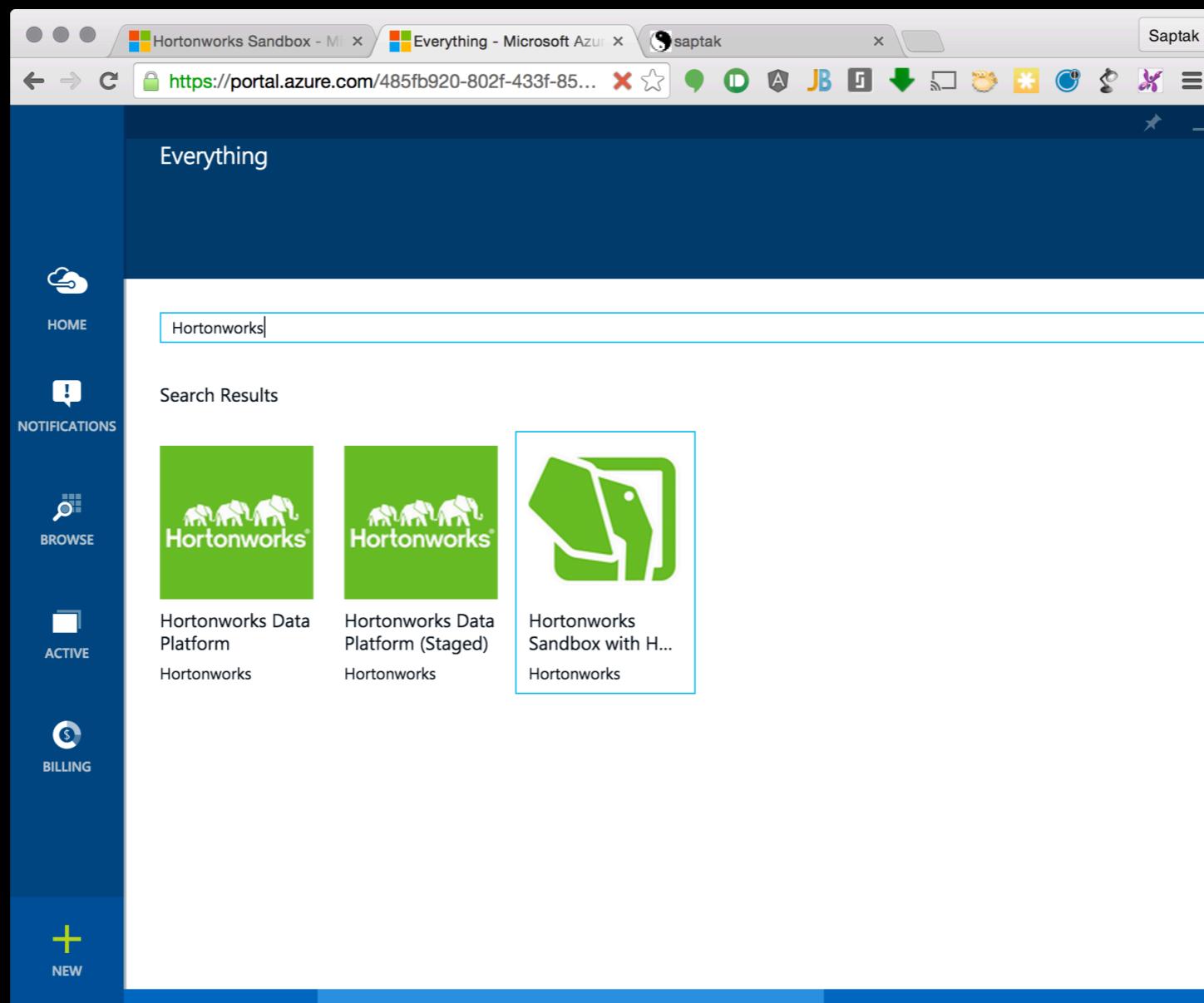
Start by logging into the Azure Portal with your Azure account: [https://
portal.azure.com/](https://portal.azure.com/)



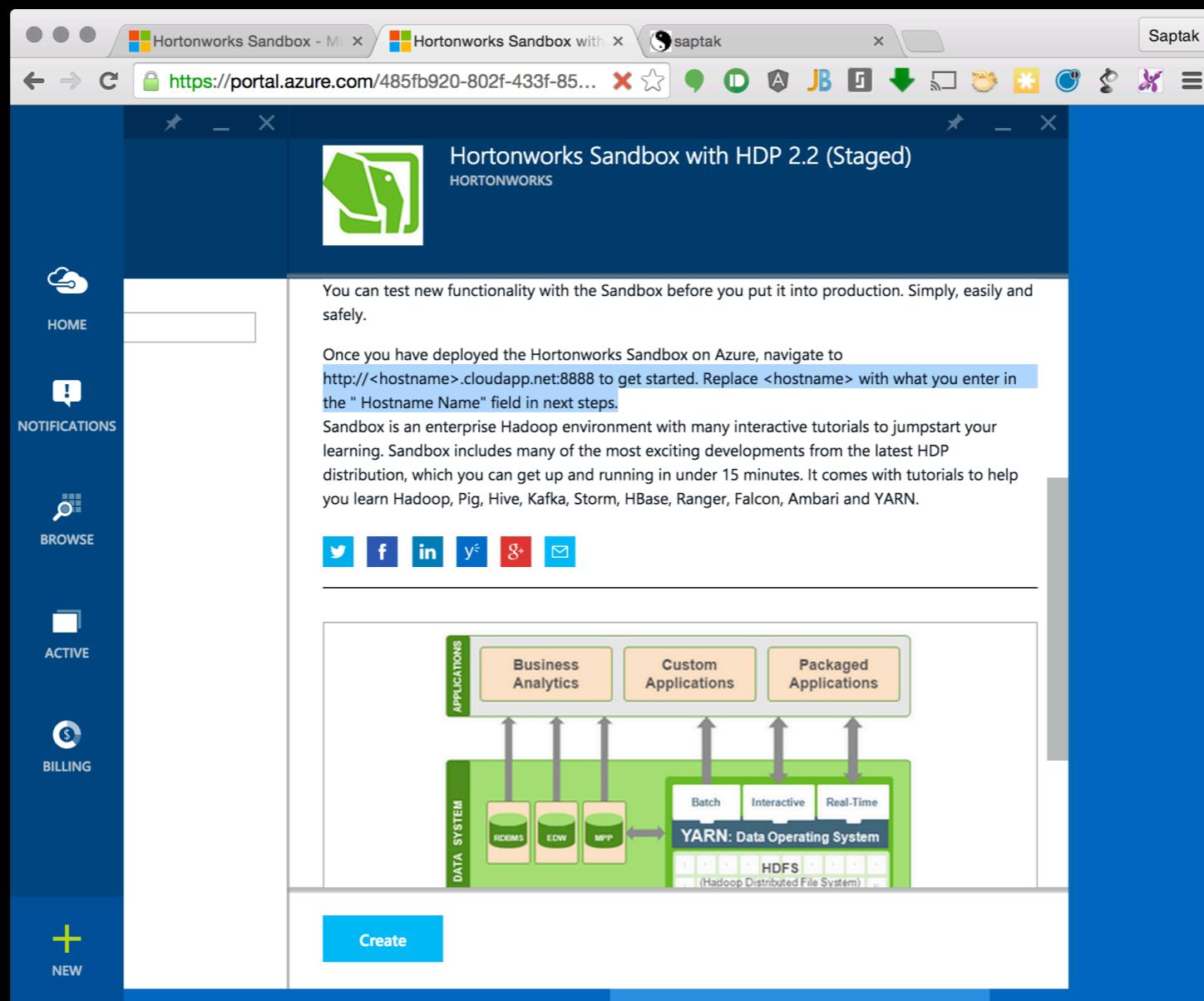
Navigate to the MarketPlace



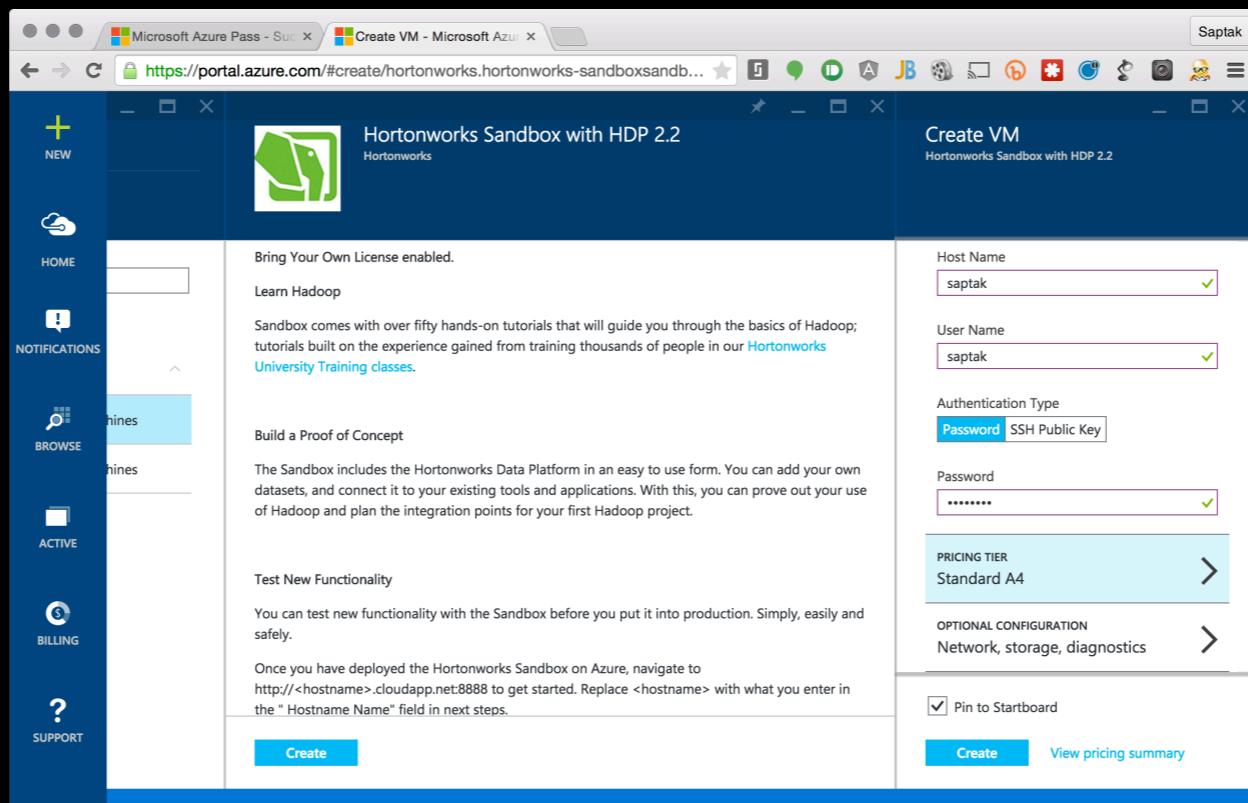
Search for Hortonworks. Click on the Hortonworks Sandbox icon.



This will launch the wizard to configure Hortonworks Sandbox for deployment.

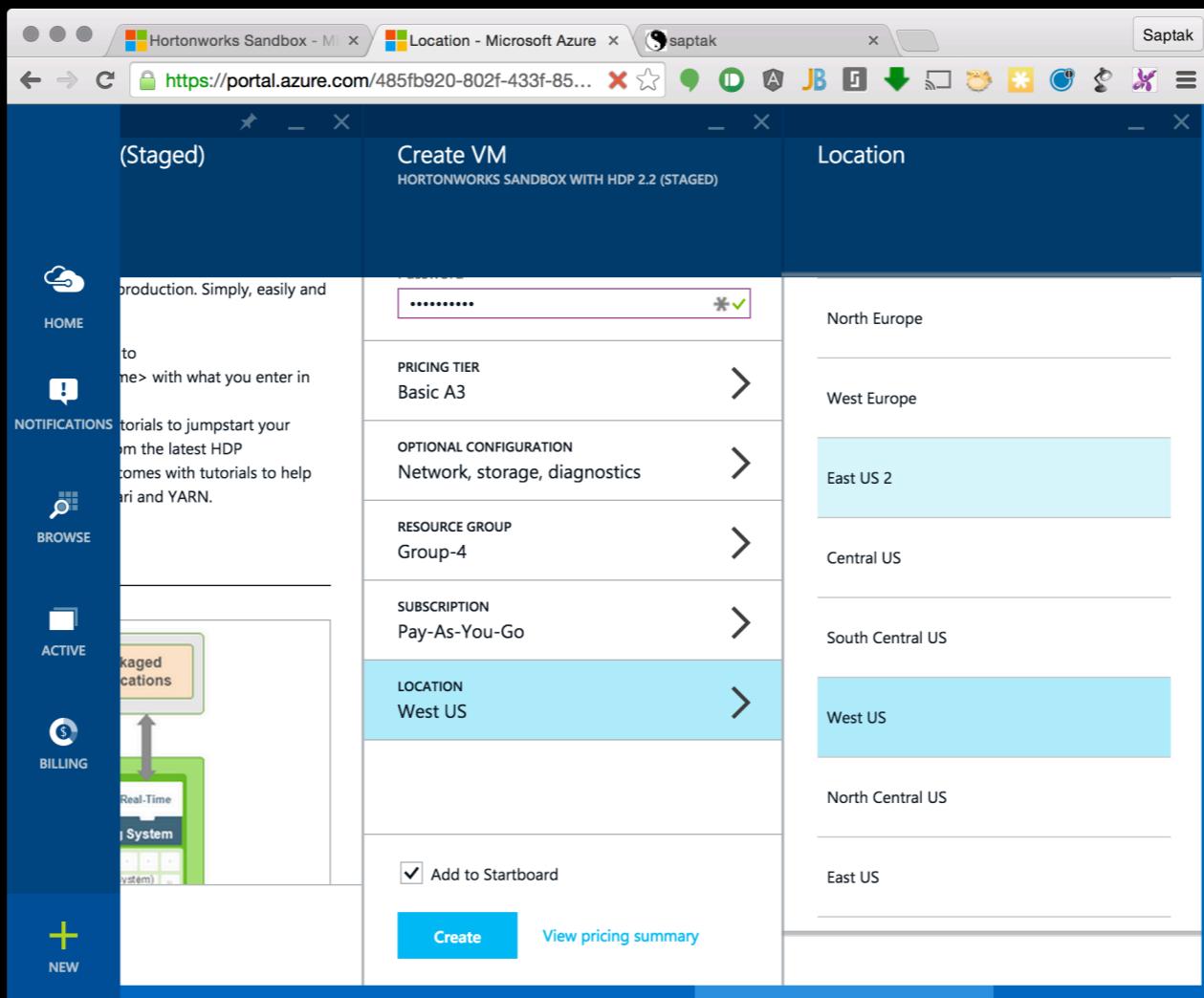


Note down the hostname and the username/password that you enter

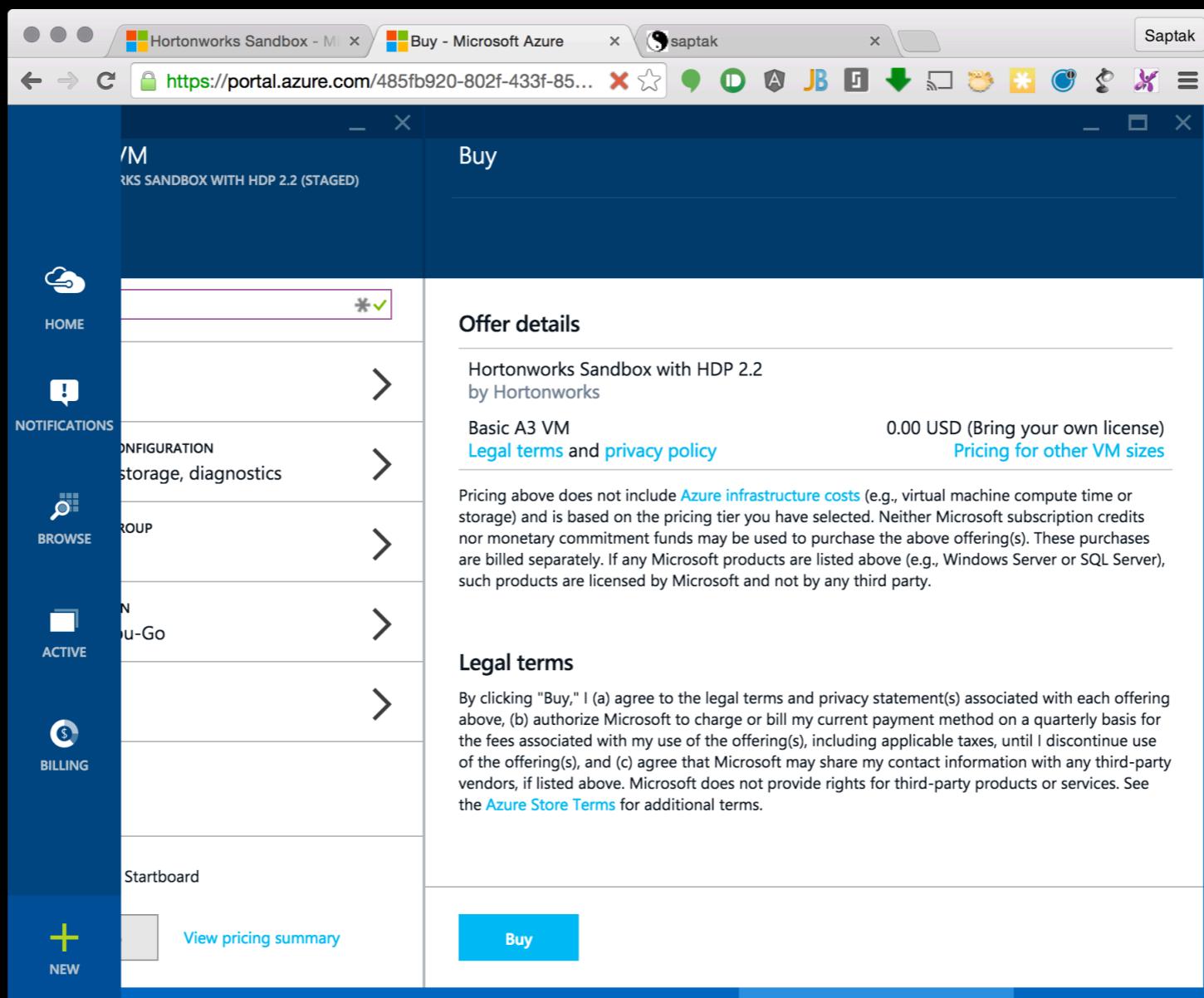


I recommend you select Standard A4 for the pricing tier

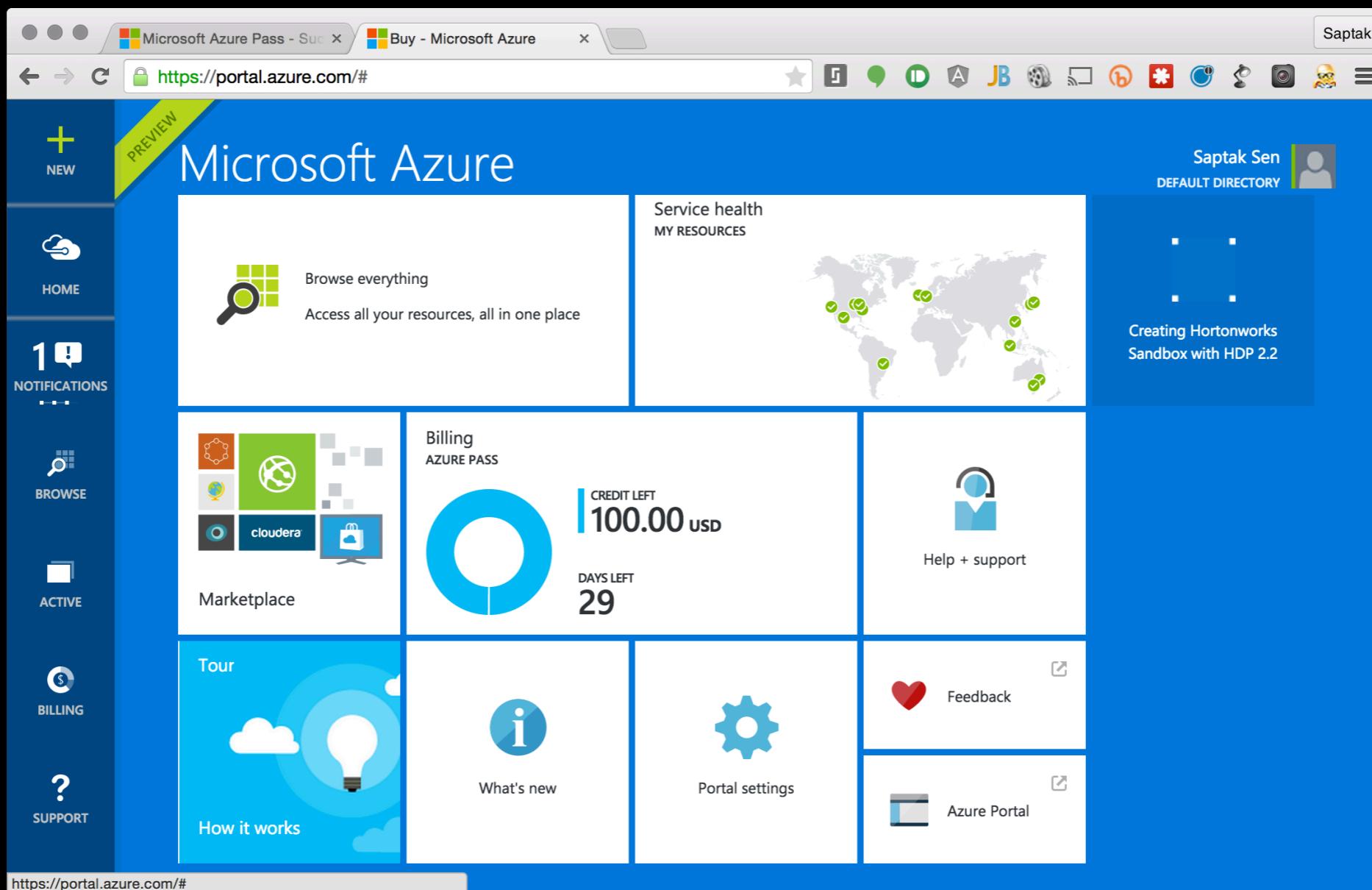
Change the location to the datacenter to where your preexisting Azure resources are or the one closest to you



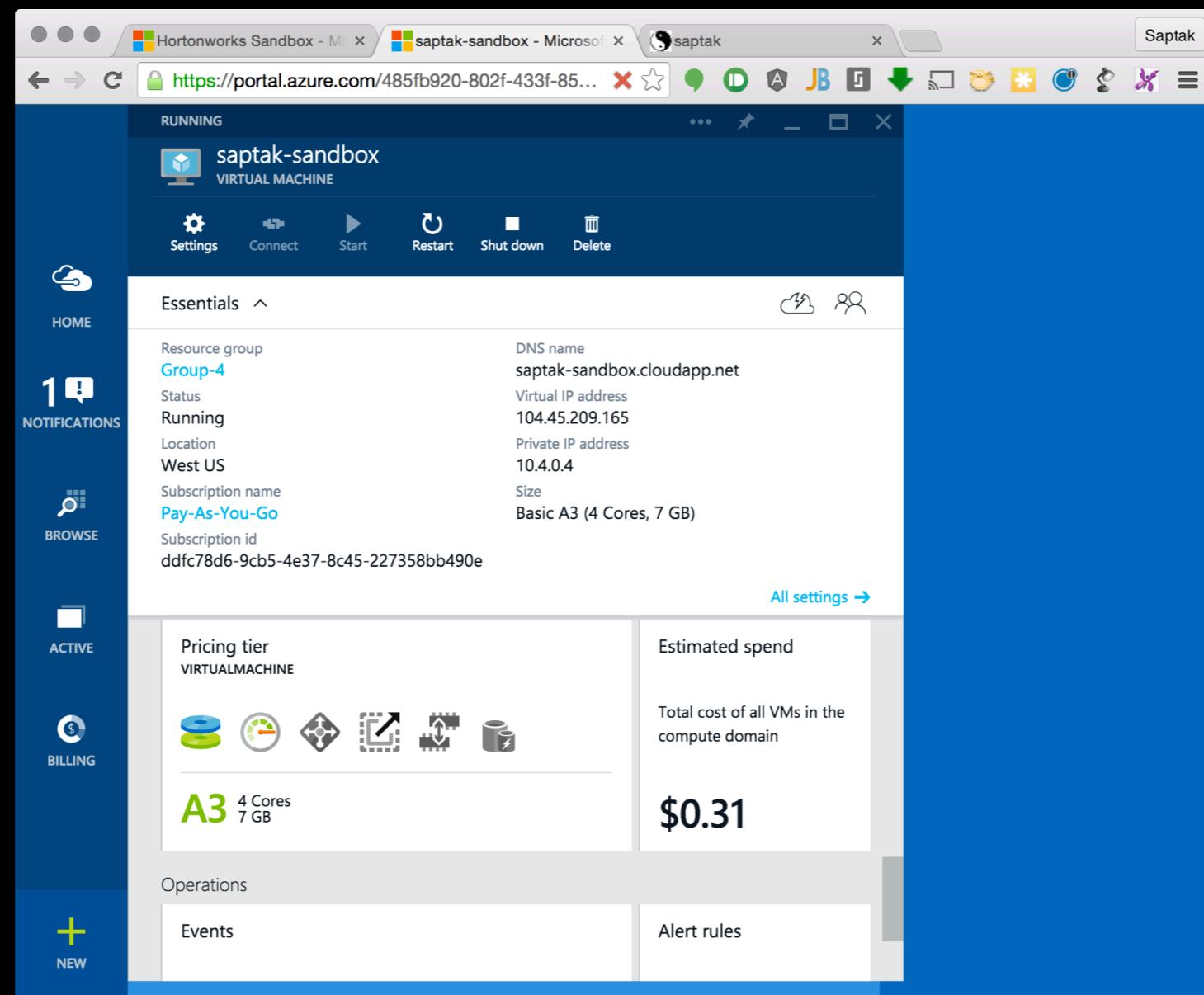
Click Buy if you agree with everything on this page.



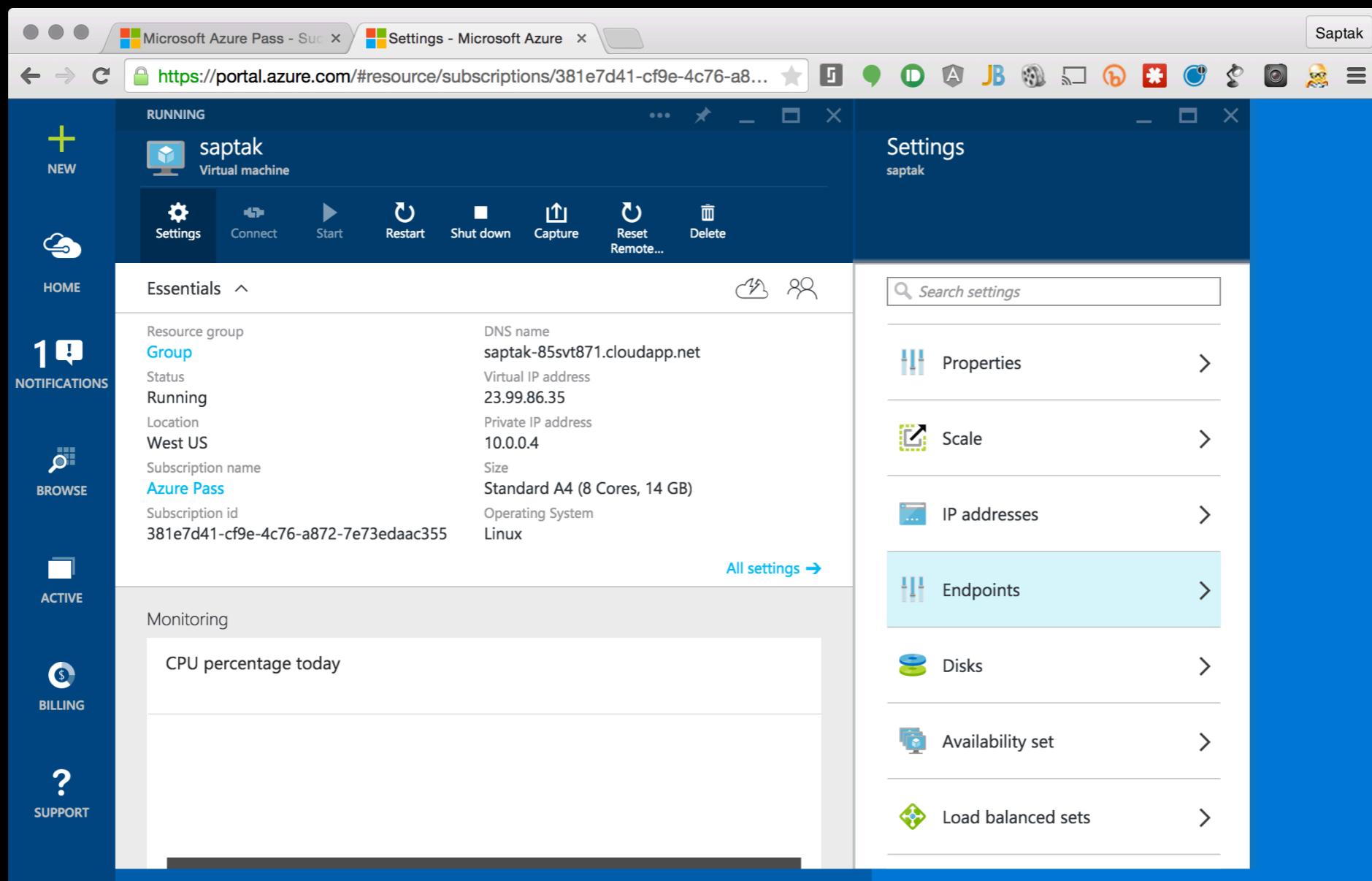
On the Azure portal home page, you can see the deployment in progress



If you scroll down you can see the Estimated spend and other metrics for your VM



To look up the SSH port click on the Settings icon on the top panel

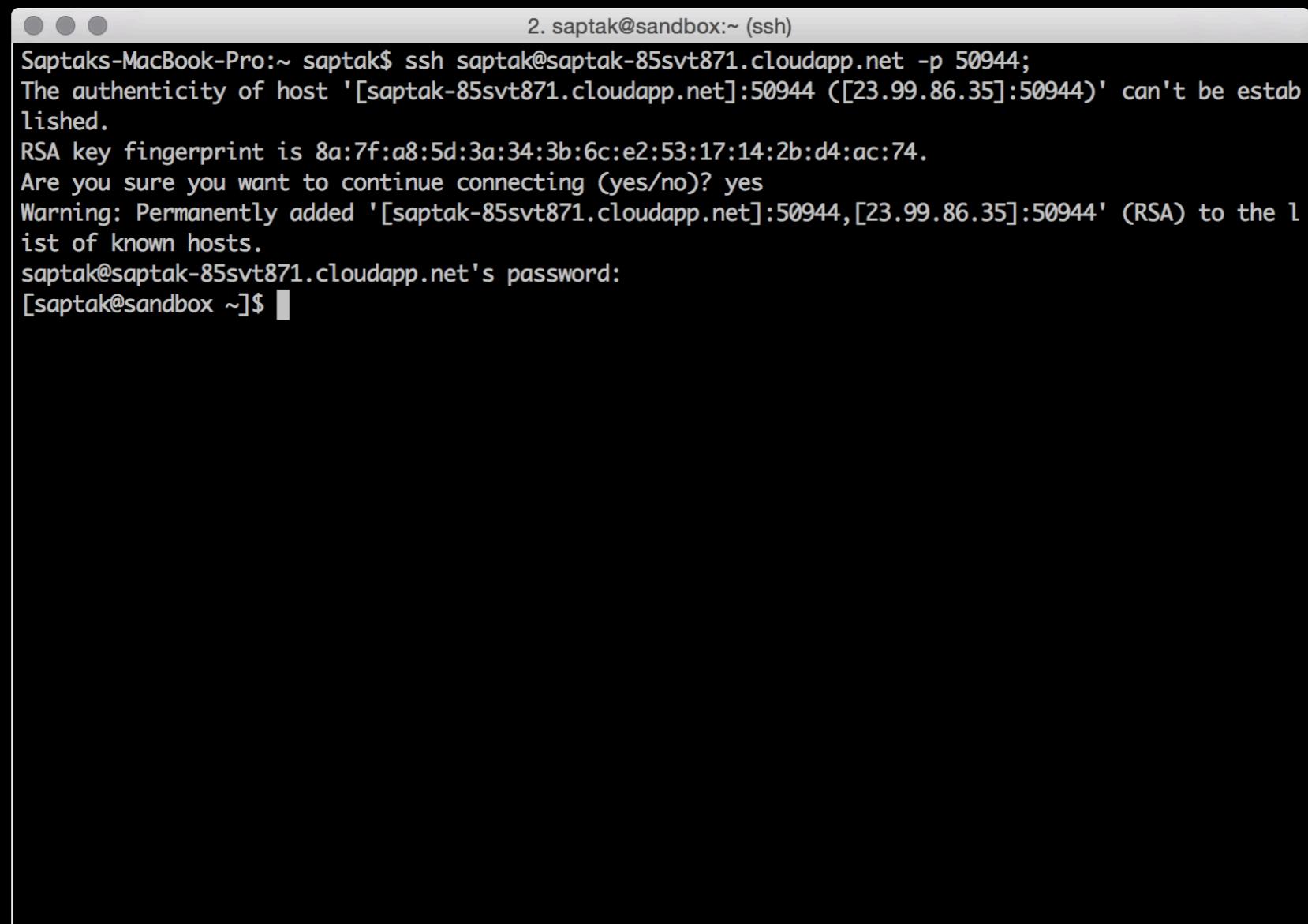


Click on Endpoints and scroll down to note the public SSH port for your VM

Name	Protocol	External Port	Internal Port	Count	...
HS2Http	TCP	10001	10001	0	
Hue	TCP	8000	8000	0	
JobHistory	TCP	19888	19888	0	
Knox	TCP	8443	8443	0	
nfs	TCP	111	111	0	
nodemanager	TCP	8040	8040	0	
Oozie	TCP	11000	11000	0	
ResourceManager	TCP	8050	8050	0	
SOLR	TCP	8993	8993	0	
SSH	TCP	50944	22	0	...
StormUI	TCP	8744	8744	0	
Tutorials	TCP	8888	8888	0	
WebHBase	TCP	60080	60080	0	

Now we can use the command below to login

```
ssh <username>@<hostname>.cloudapp.net -p <port>;
```

A screenshot of a terminal window titled "2. saptak@sandbox:~ (ssh)". The window shows the following text:

```
Saptaks-MacBook-Pro:~ saptak$ ssh saptak@saptak-85svt871.cloudapp.net -p 50944;
The authenticity of host '[saptak-85svt871.cloudapp.net]:50944 ([23.99.86.35]:50944)' can't be established.
RSA key fingerprint is 8a:7f:a8:5d:3a:34:3b:6c:e2:53:17:14:2b:d4:ac:74.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '[saptak-85svt871.cloudapp.net]:50944,[23.99.86.35]:50944' (RSA) to the l
ist of known hosts.
saptak@saptak-85svt871.cloudapp.net's password:
[saptak@sandbox ~]$
```

change the root password to a known password using the command `sudo passwd root`

```
1. saptak@saptak:~ (ssh)
Last login: Sun Jun  7 15:27:41 on ttys000
Saptaks-MacBook-Pro:~ saptak$ ssh saptak@DNS name
ssh: Could not resolve hostname DNS: nodename nor servname provided, or not known
Saptaks-MacBook-Pro:~ saptak$ ssh saptak@saptak-168bmf0.cloudapp.net -p 54340
The authenticity of host '[saptak-168bmf0.cloudapp.net]:54340 ([23.99.92.106]:54340)' can't be established.
RSA key fingerprint is d7:c4:91:e4:79:6a:a4:c4:a3:31:49:35:dc:d8:7d:ac.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '[saptak-168bmf0.cloudapp.net]:54340,[23.99.92.106]:54340' (RSA) to the list of known hosts.
saptak@saptak-168bmf0.cloudapp.net's password:
[saptak@saptak ~]$ sudo passwd root

We trust you have received the usual lecture from the local System
Administrator. It usually boils down to these three things:

#1) Respect the privacy of others.
#2) Think before you type.
#3) With great power comes great responsibility.

[sudo] password for saptak: [REDACTED]
```

Now we can login as root with the command su

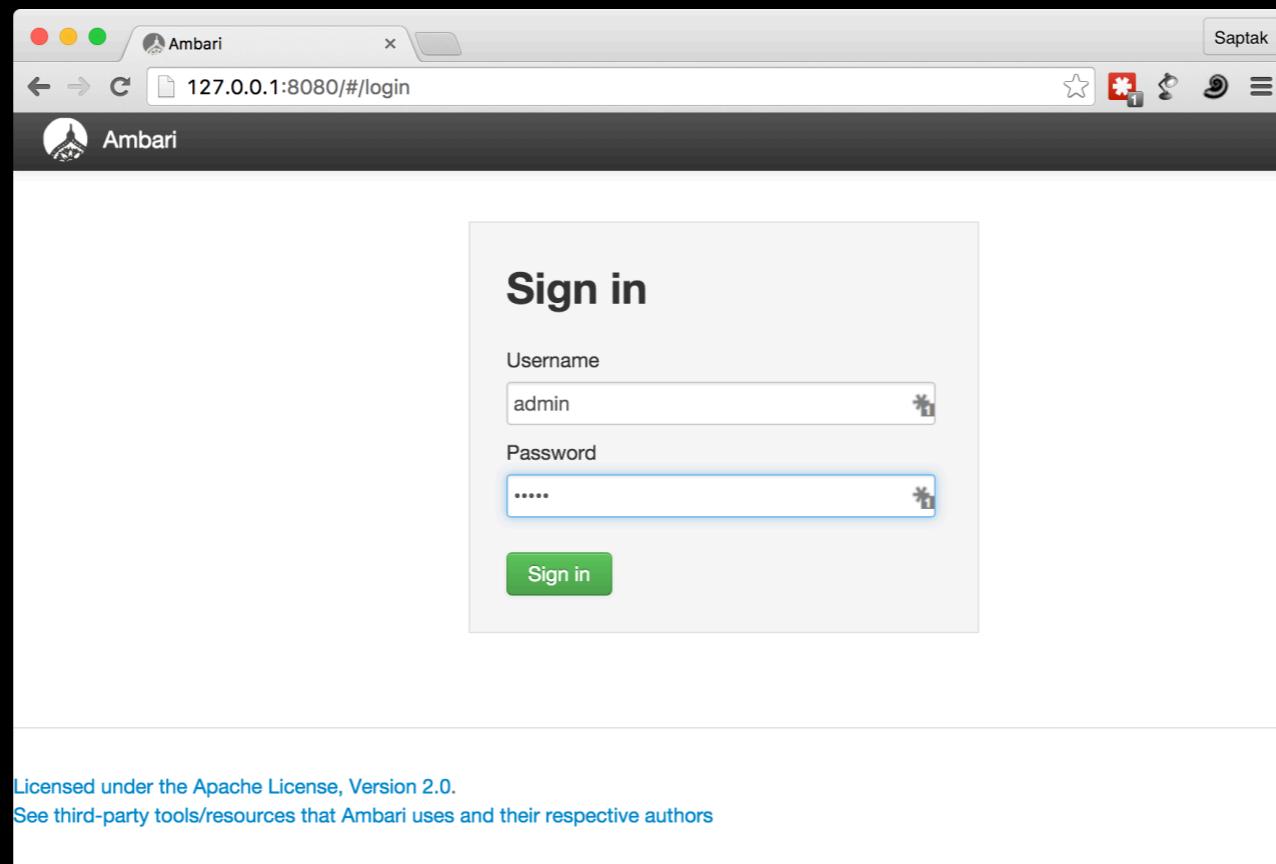
```
1. saptak@saptak:~ (ssh)
The authenticity of host '[saptak-168bmf0.cloudapp.net]:54340 ([23.99.92.106]:54340)' can't be established.
RSA key fingerprint is d7:c4:91:e4:79:6a:a4:c4:a3:31:49:35:dc:d8:7d:ac.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '[saptak-168bmf0.cloudapp.net]:54340,[23.99.92.106]:54340' (RSA) to the list of known hosts.
saptak@saptak-168bmf0.cloudapp.net's password:
[saptak@saptak ~]$ sudo passwd root

We trust you have received the usual lecture from the local System Administrator. It usually boils down to these three things:

#1) Respect the privacy of others.
#2) Think before you type.
#3) With great power comes great responsibility.

[sudo] password for saptak:
Changing password for user root.
New password:
Retype new password:
passwd: all authentication tokens updated successfully.
[saptak@saptak ~]$ su
Password:
[root@saptak saptak]# cd
[root@saptak ~]# █
```

Login to Ambari



Ambari is at port 8080. Username and password is admin/admin

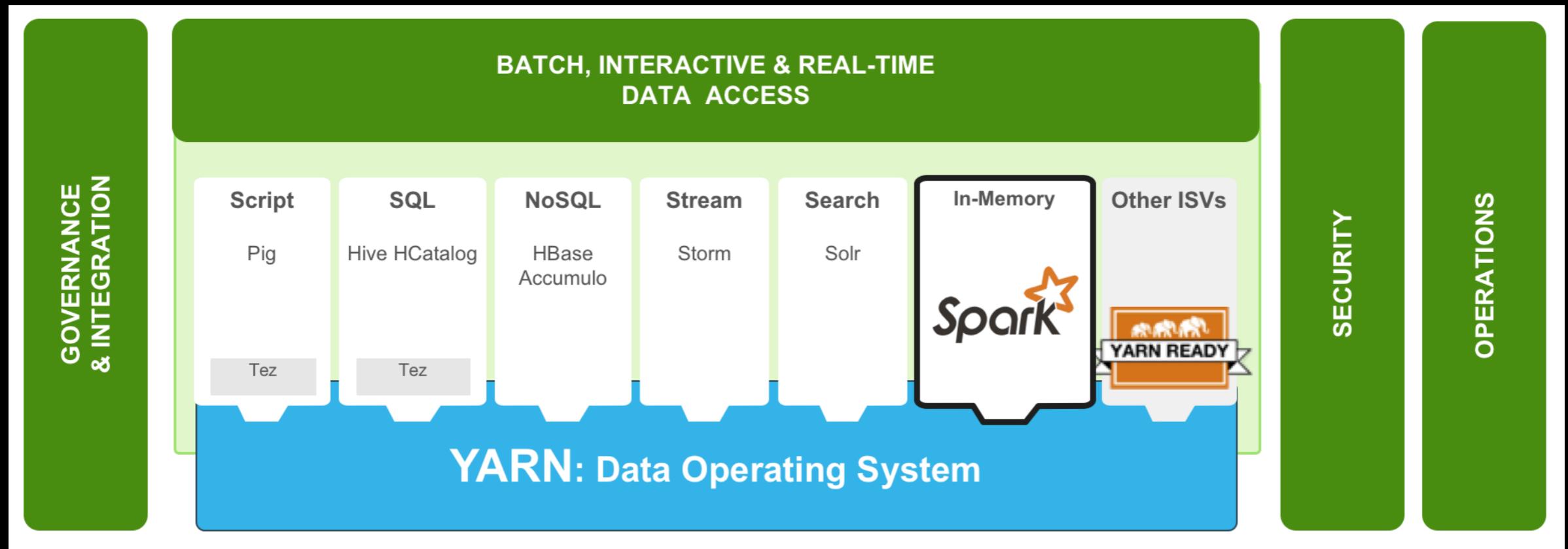
What is Hadoop?

Apache Hadoop is a scalable, fault tolerant, open source framework for the distributed storing and processing of large sets of data on commodity hardware.

The goals of Hadoop are:

- To use inexpensive hardware to create very large clusters of servers
- To distribute data and processing across many servers to achieve massive scalability
- Each server provides CPU, memory, network, and internal disk resources to the cluster.

Hadoop



Data

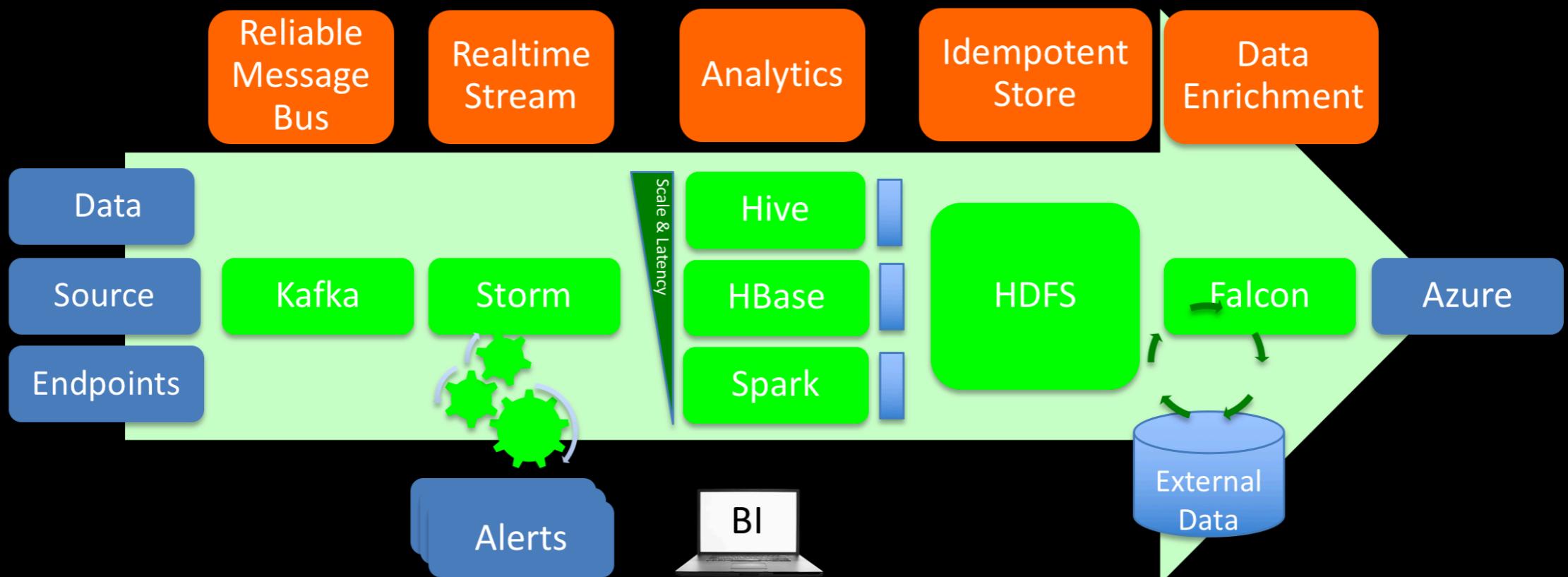
Structured Data

- Structured data resides in a fixed field within a record or file
- Structured data depends on creating a data model called a schema
 - The schema defines the types of data that will be recorded and how that data will be stored, accessed, and processed
 - This includes defining the fields of data to be stored along with the type of data in each field
 - Types include strings, integers, floating point numbers, dates, and others

Unstructured Data

- Data that does not reside in fixed fields or records
- There is no data model, contained in a schema or tags, that defines how to store, access, or process the data
- Unstructured data has irregularities and ambiguities that make it difficult to understand and analyze using traditional computer programs.
 - This data is often ignored or deleted which limits its business value.
 - Experts estimate that 80-90 percent of the data in any enterprise is unstructured.
- Hadoop's primary contribution has been the capability to extract value from this unstructured data.

Where is my data?



HDFS

- Hadoop stores files using the Hadoop distributed file system (HDFS).
- HDFS is the basis for Hadoop's storage scalability and availability. HDFS:
 - Splits large data files into smaller chunks called blocks
 - Spreads those blocks across different slave nodes
 - Tracks data block location
 - Automatically replicates data for high availability

Using the HDFS Command-Line Interface

- To display command-line syntax and options:

```
hadoop fs
```

- To list HDFS directory contents:

```
hadoop fs -ls HDFS_dir_name/
```

- To make an HDFS directory:

```
hadoop fs -mkdir HDFS_dir_name/new_dir
```

- To copy a local file to an HDFS directory:

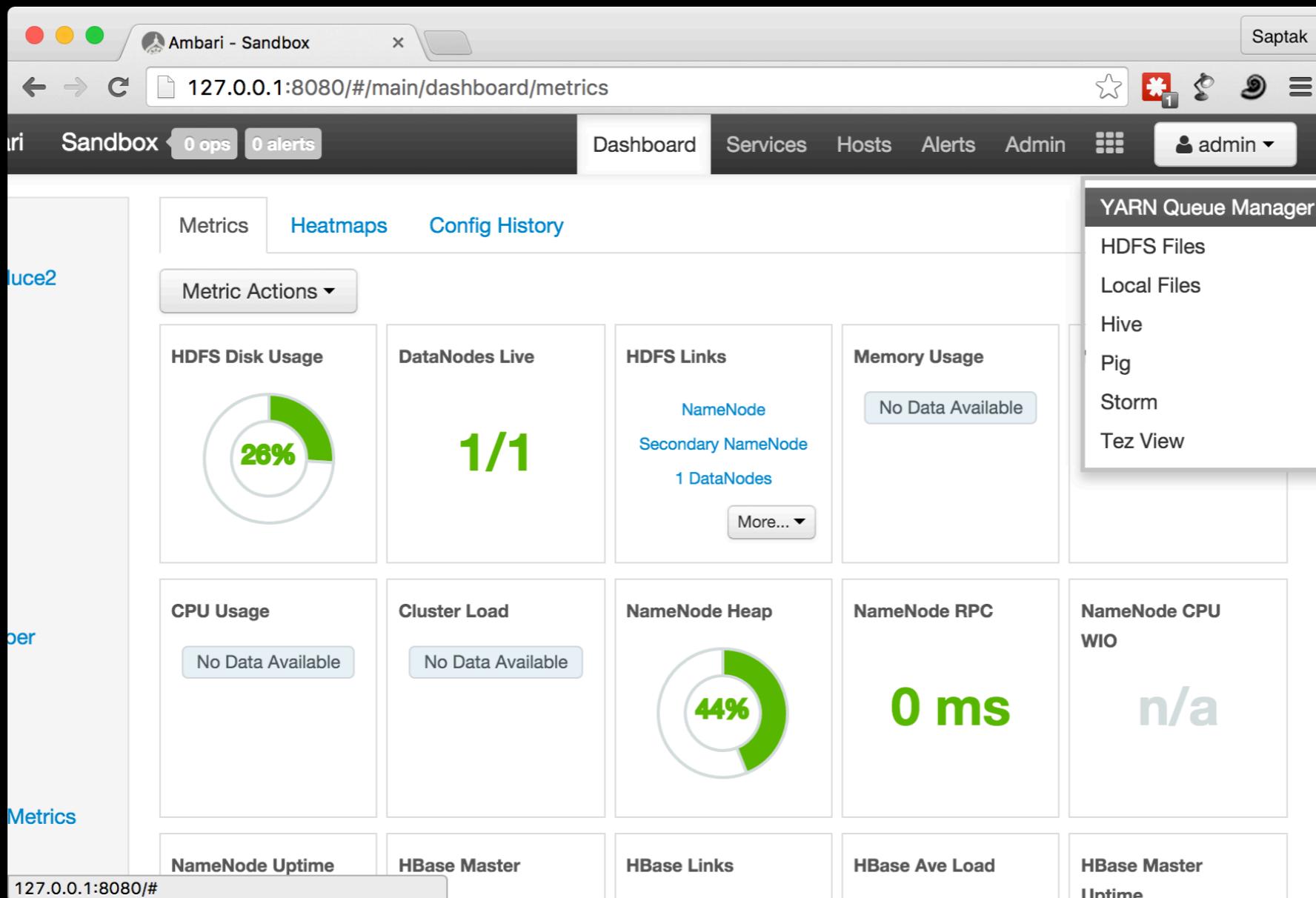
```
hadoop fs -put local_file HDFS_dir_name/new_file
```

- To copy an HDFS file to a local file system directory:

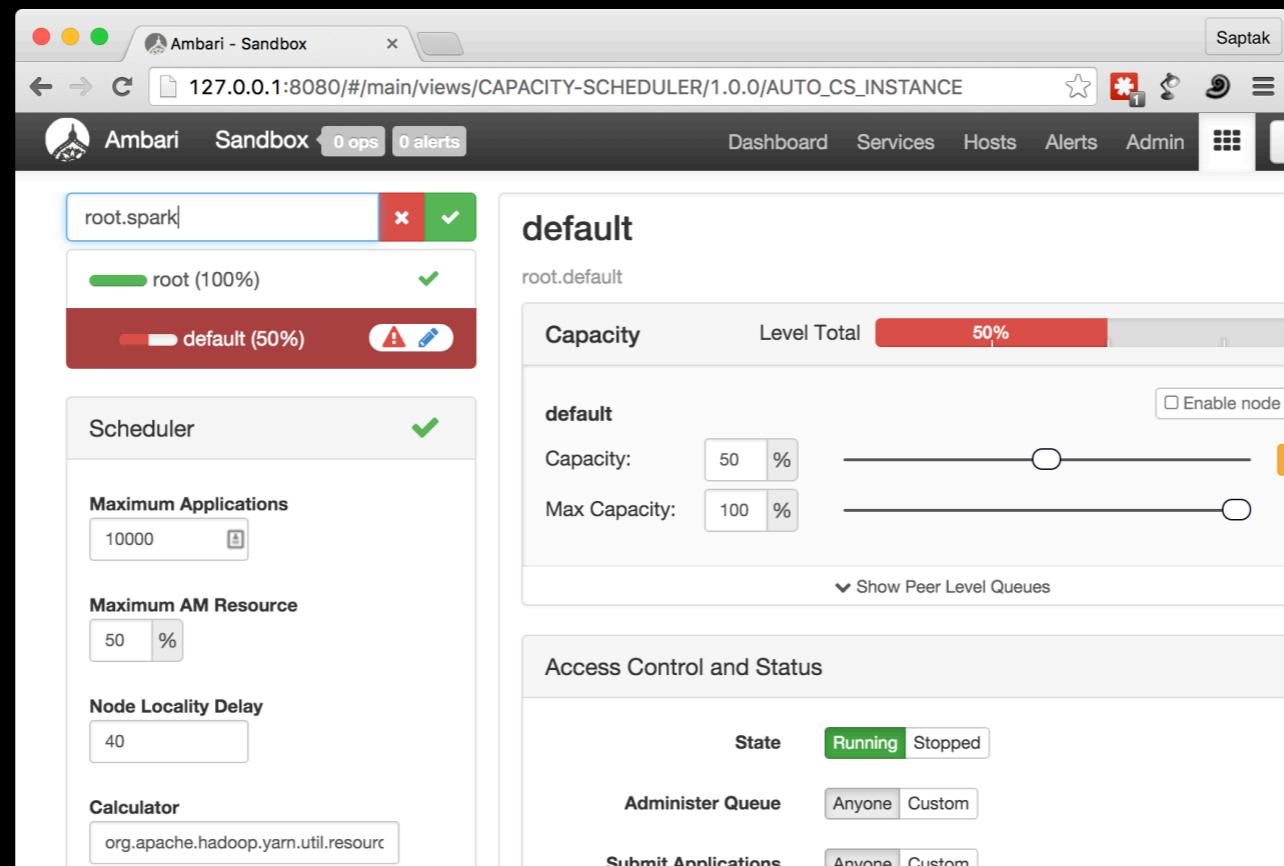
```
hadoop fs -get HDFS_dir_name/file /local_dir/new_file
```

Lab

Open YARN Queue Manager



Create a dedicated queue for Spark



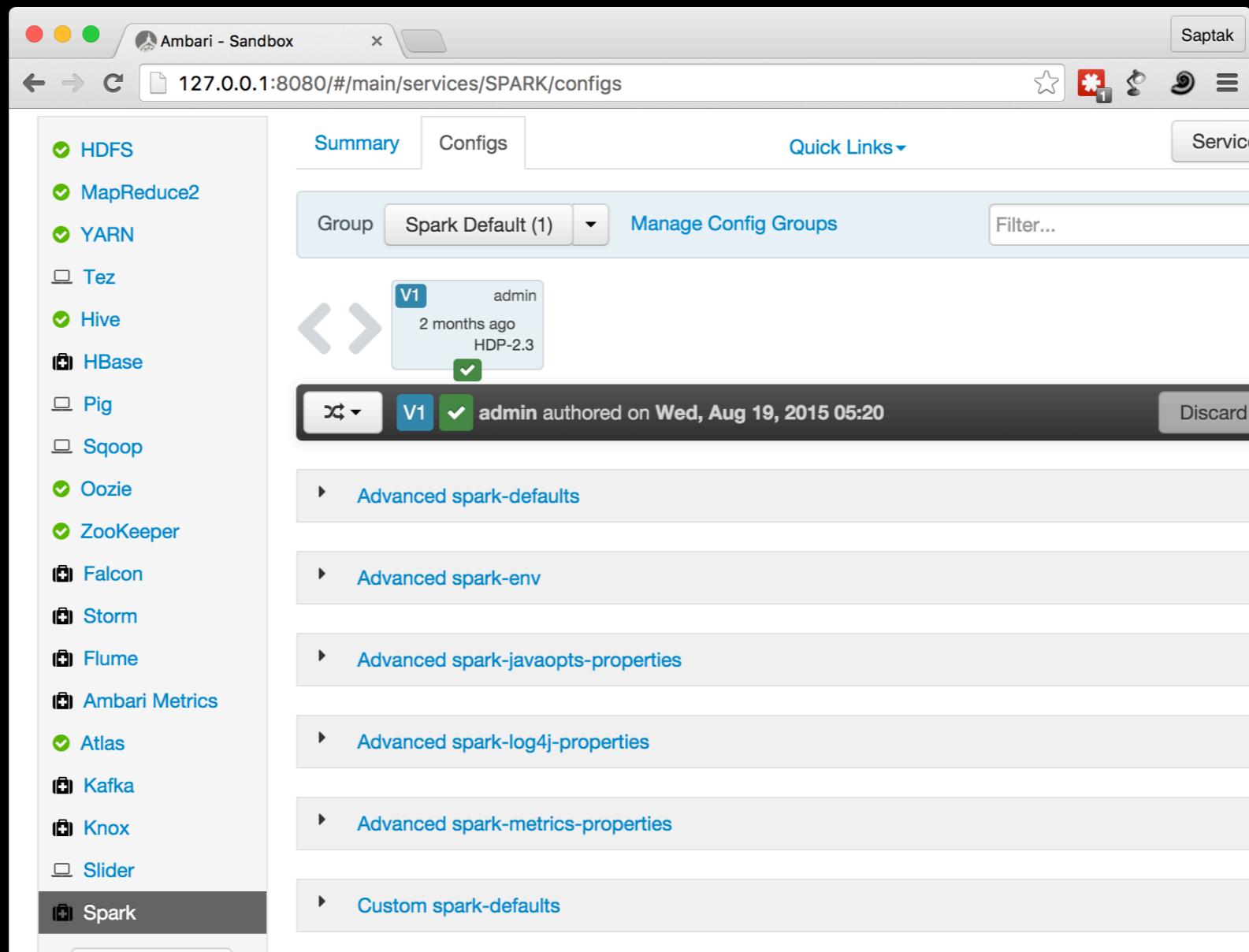
Set name to `root.spark`, Capacity to 50% and Max Capacity to 95%

Save and Refresh Queues

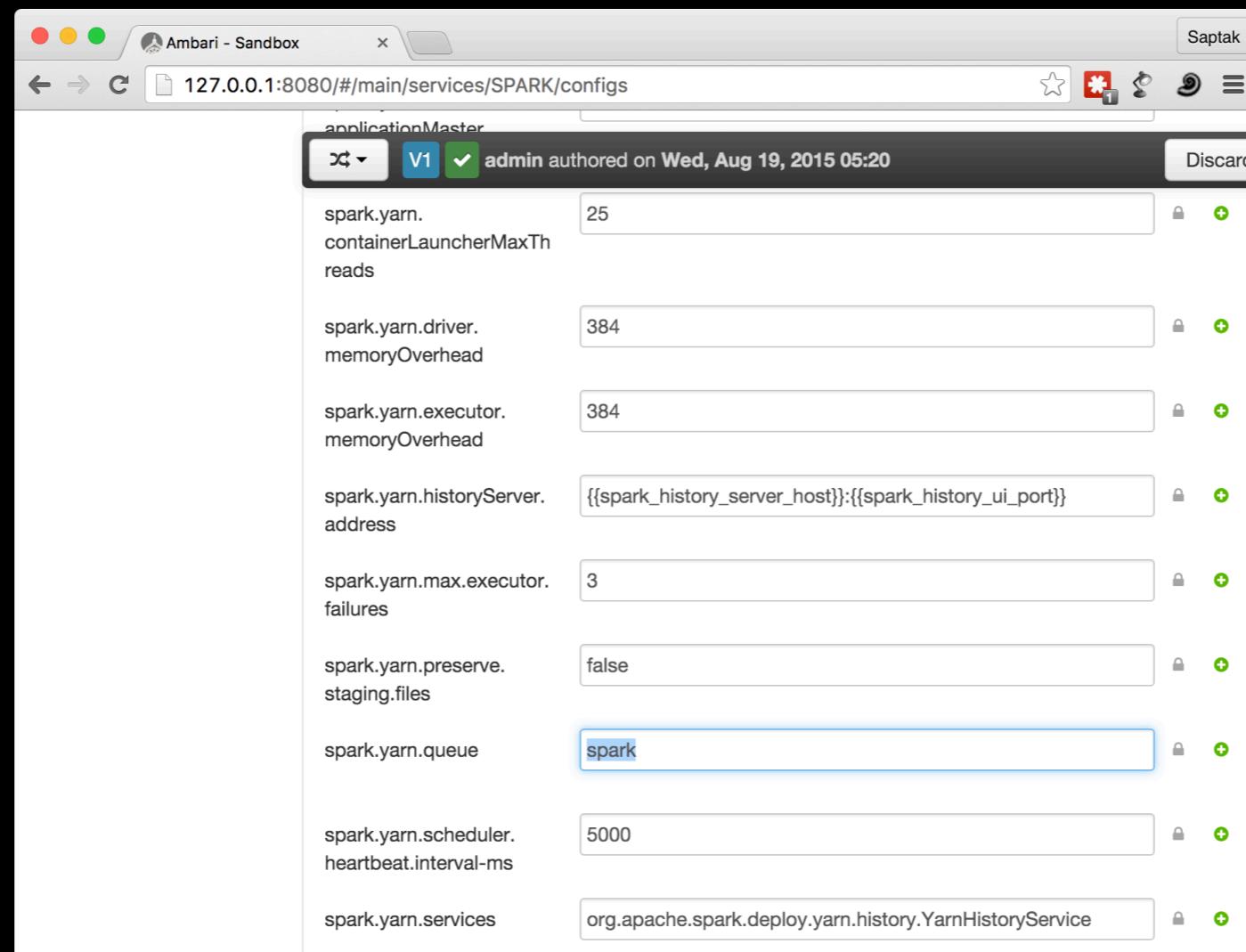
The screenshot shows the Ambari Capacity Scheduler interface. On the left, there's a sidebar with options like 'Add Queue', 'Actions', 'Scheduler', 'Maximum Applications' (set to 10000), 'Maximum AM Resource' (set to 50%), and 'Node Locality Delay'. The main area displays three queues: 'root (100%)', 'default (50%)', and 'spark (50%)'. The 'spark' queue is currently selected. A modal window for the 'spark' queue is open, showing its configuration details. The 'Capacity' section has 'Level Total' set to 100% and 'Capacity' set to 50%. The 'Max Capacity' is set to 95%. There's also a checkbox for 'Enable node labels'. Below the capacity section, there's a link to 'Show Peer Level Queues'. At the bottom of the modal, there's an 'Access Control and Status' section with a 'State' dropdown set to 'Running'.

Adjust capacity of the default queue to 50% if required before saving

Open Apache Spark configuration



Set the 'spark' queue as the default queue for Spark



Restart All Affected Services

