UNIVERSITY of WASHINGTON

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

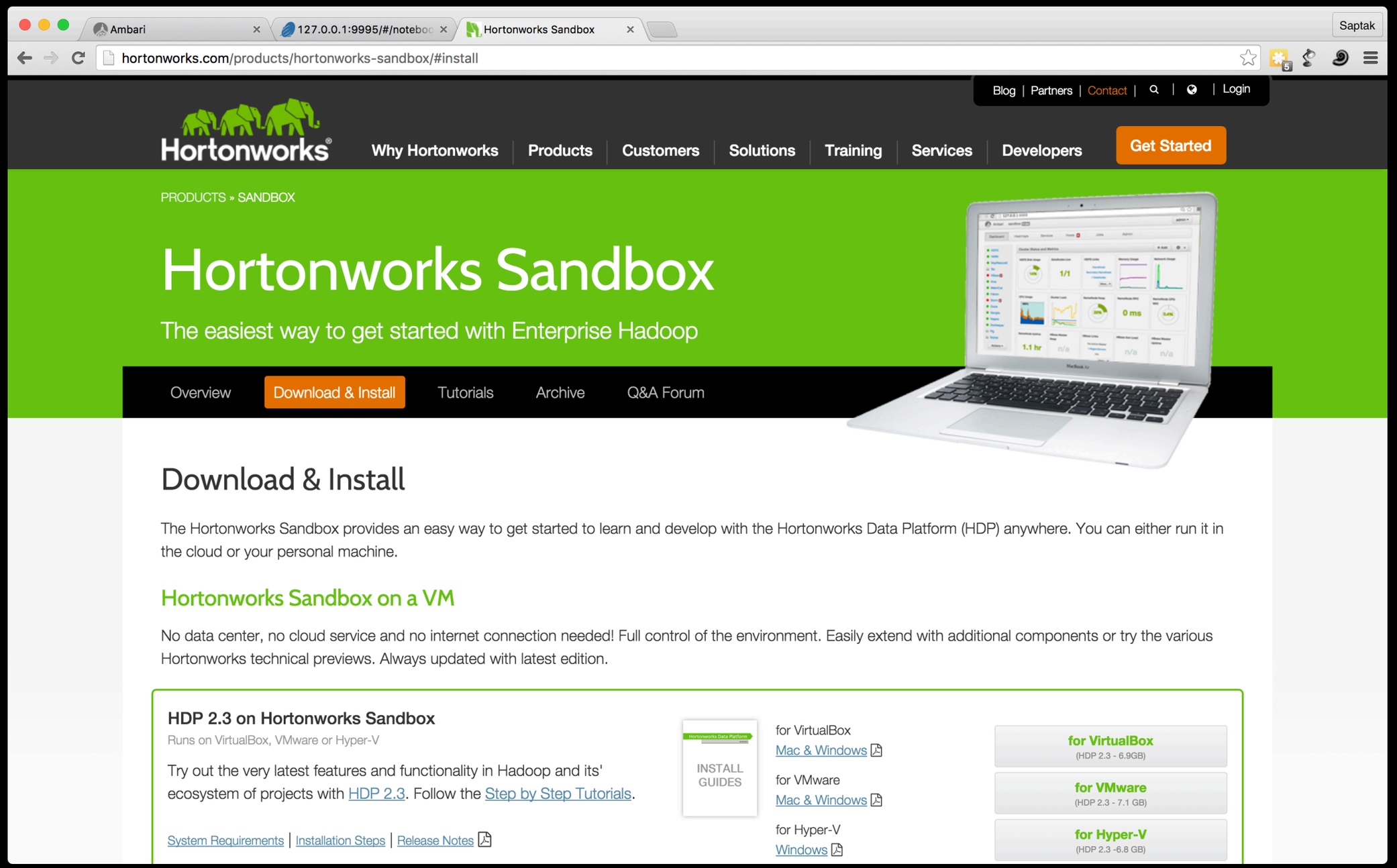
Logistics

- 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

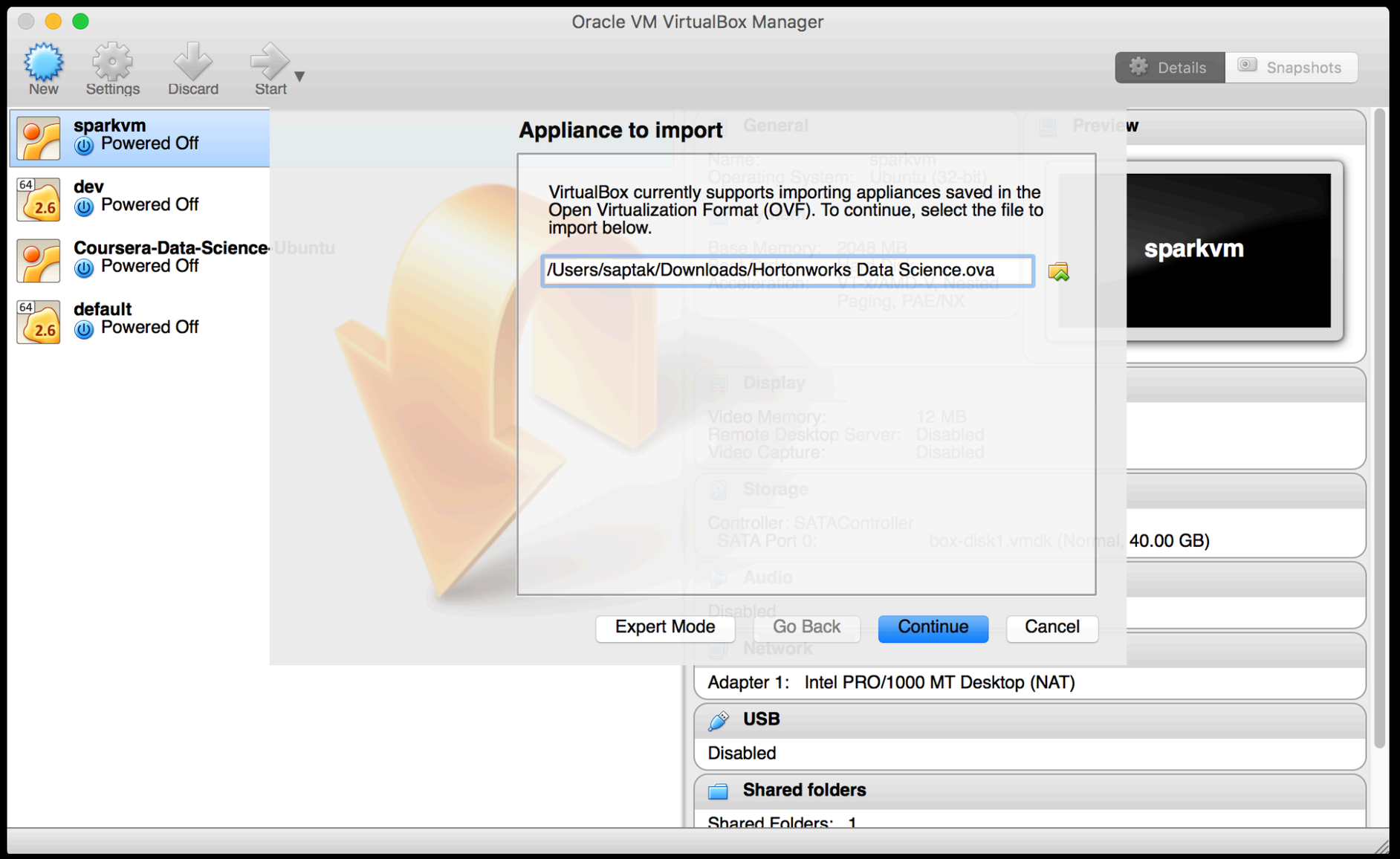
Setup

Download Hortonworks Sandbox

http://hortonworks.com/sandbox



Configure Hortonworks Sandbox

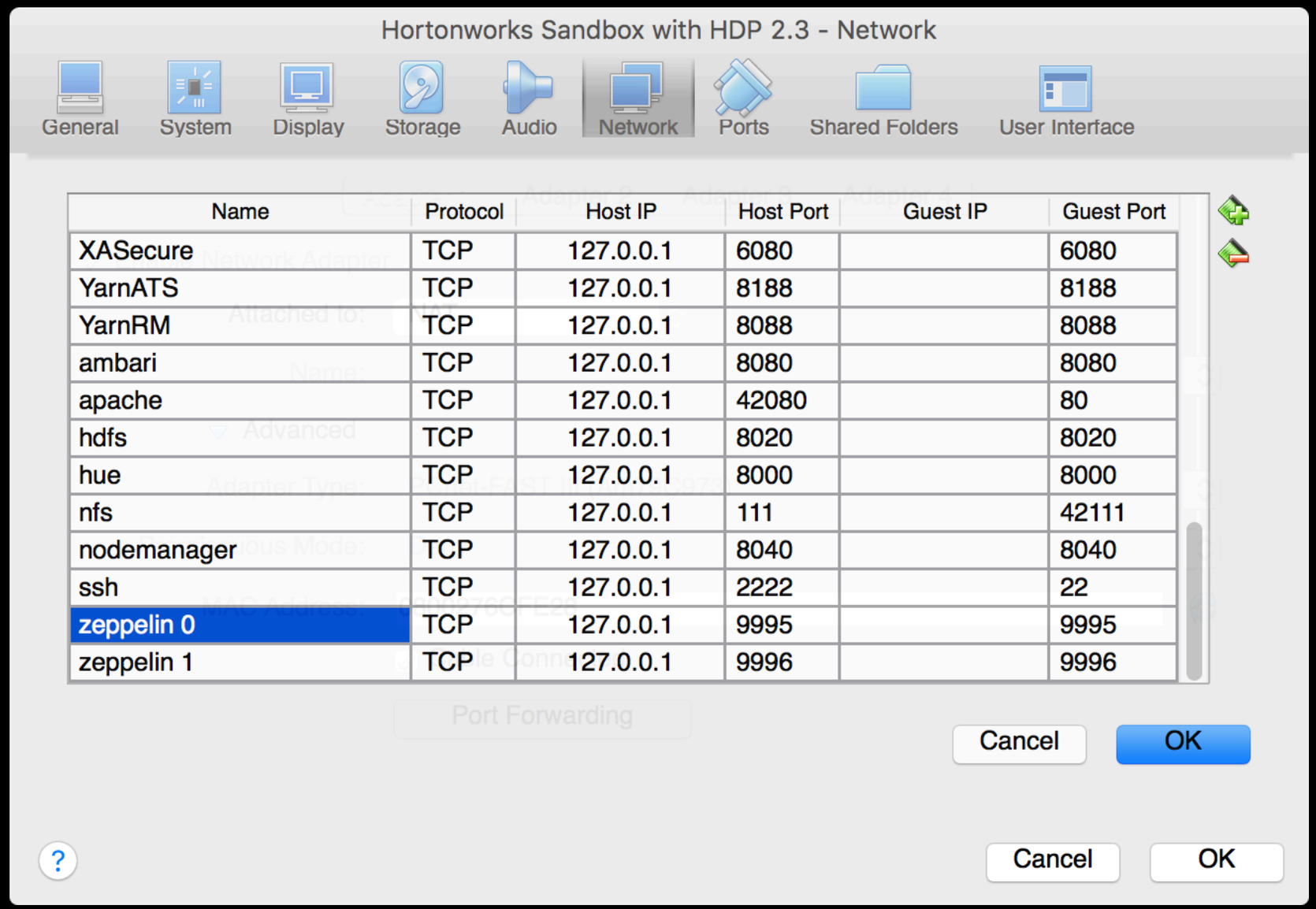


Open network settings for VM

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

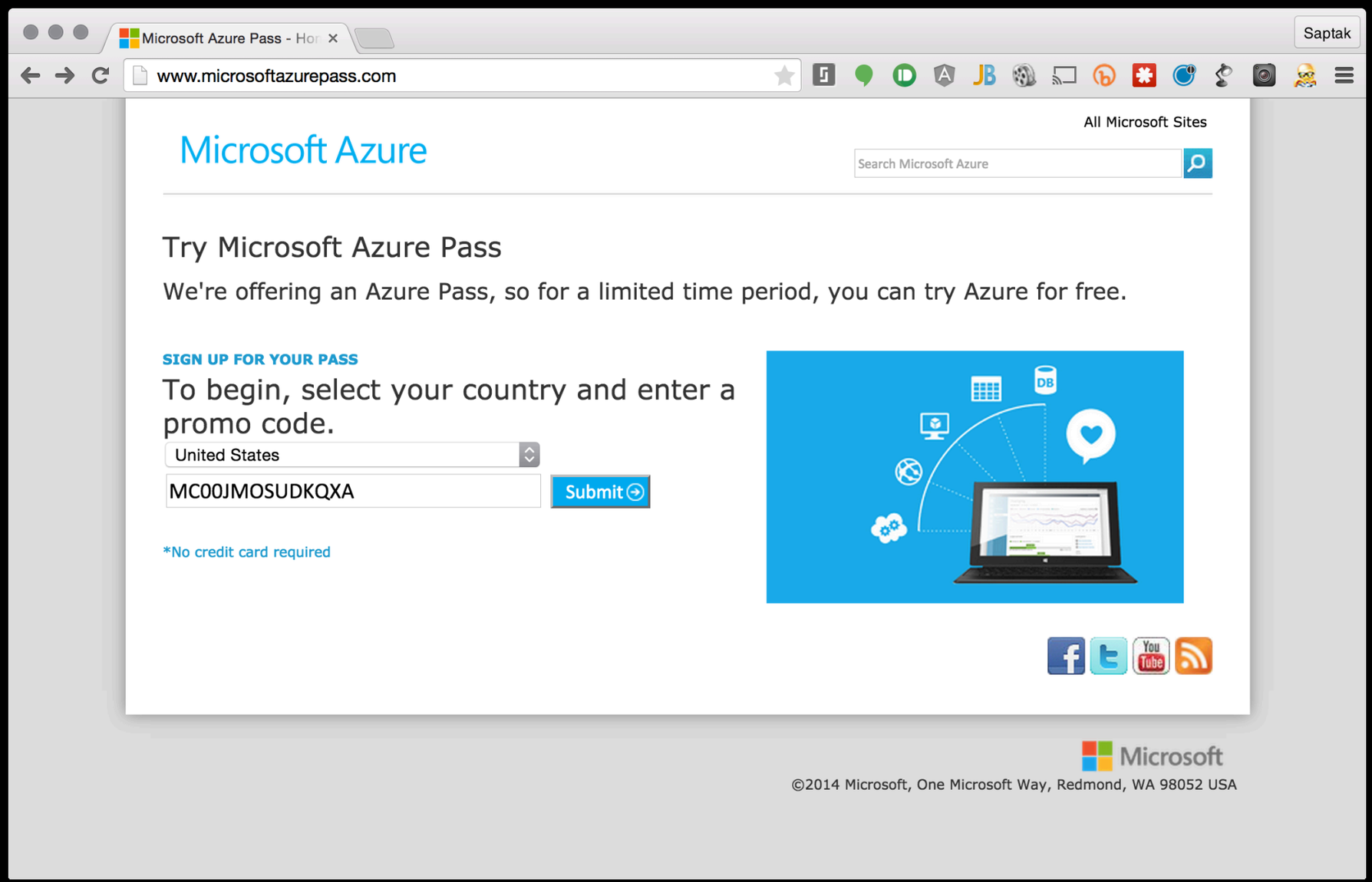
Cancel OK

Open port for Zeppelin

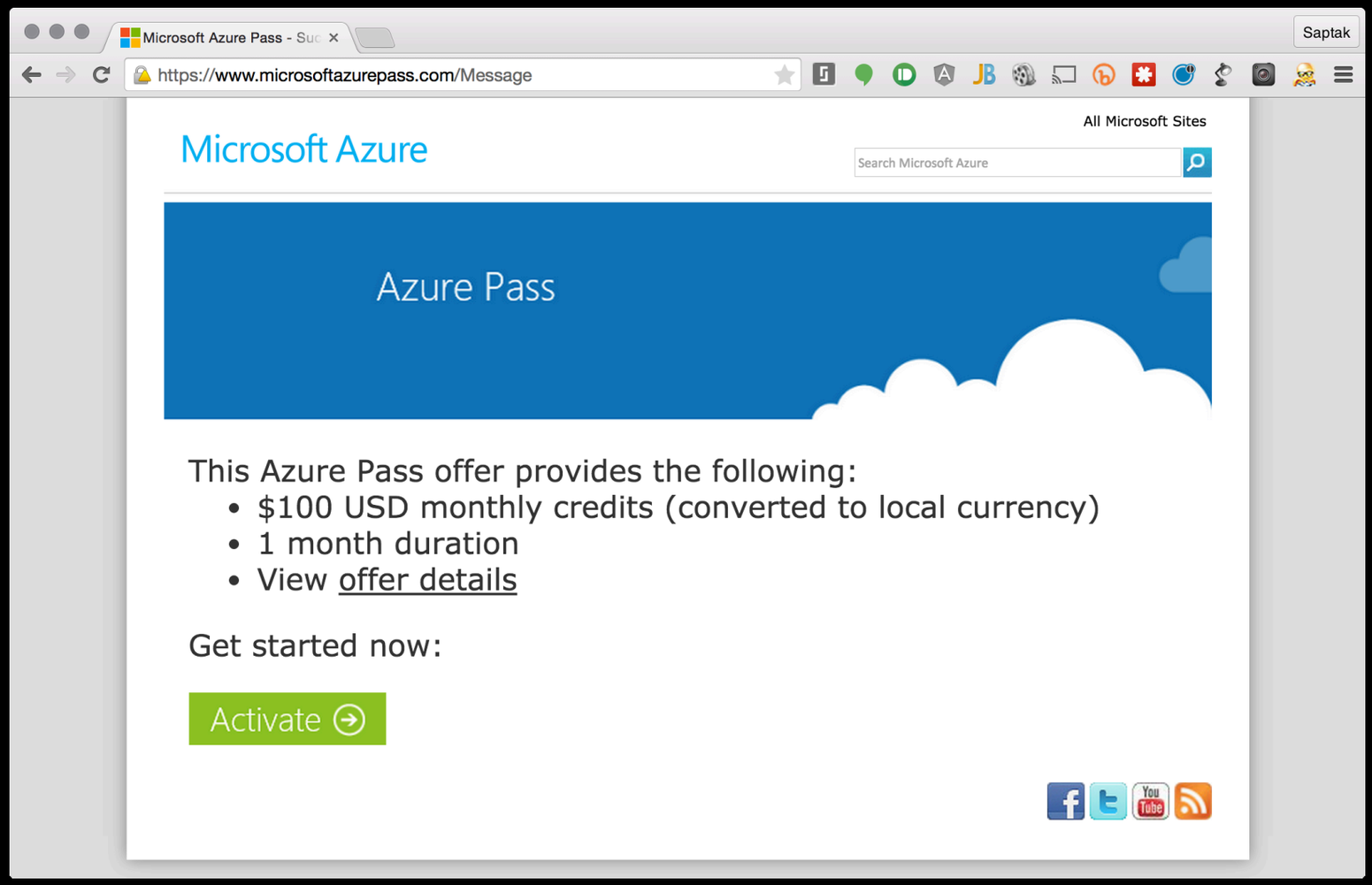


Configuring Hortonworks Sandbox on Azure

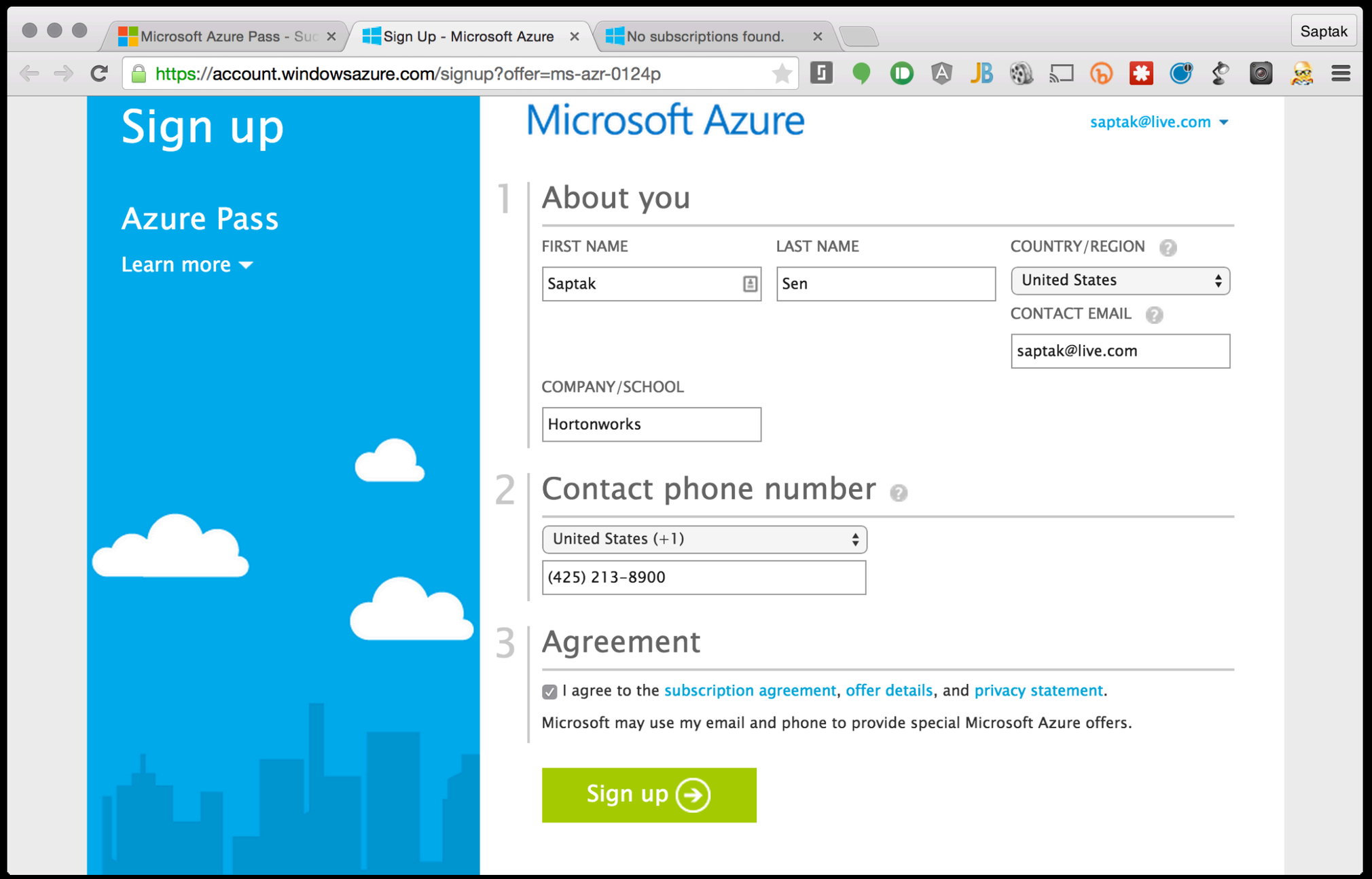
To use the Microsoft Azure Pass navigate to http://www.microsoftazurepass.com



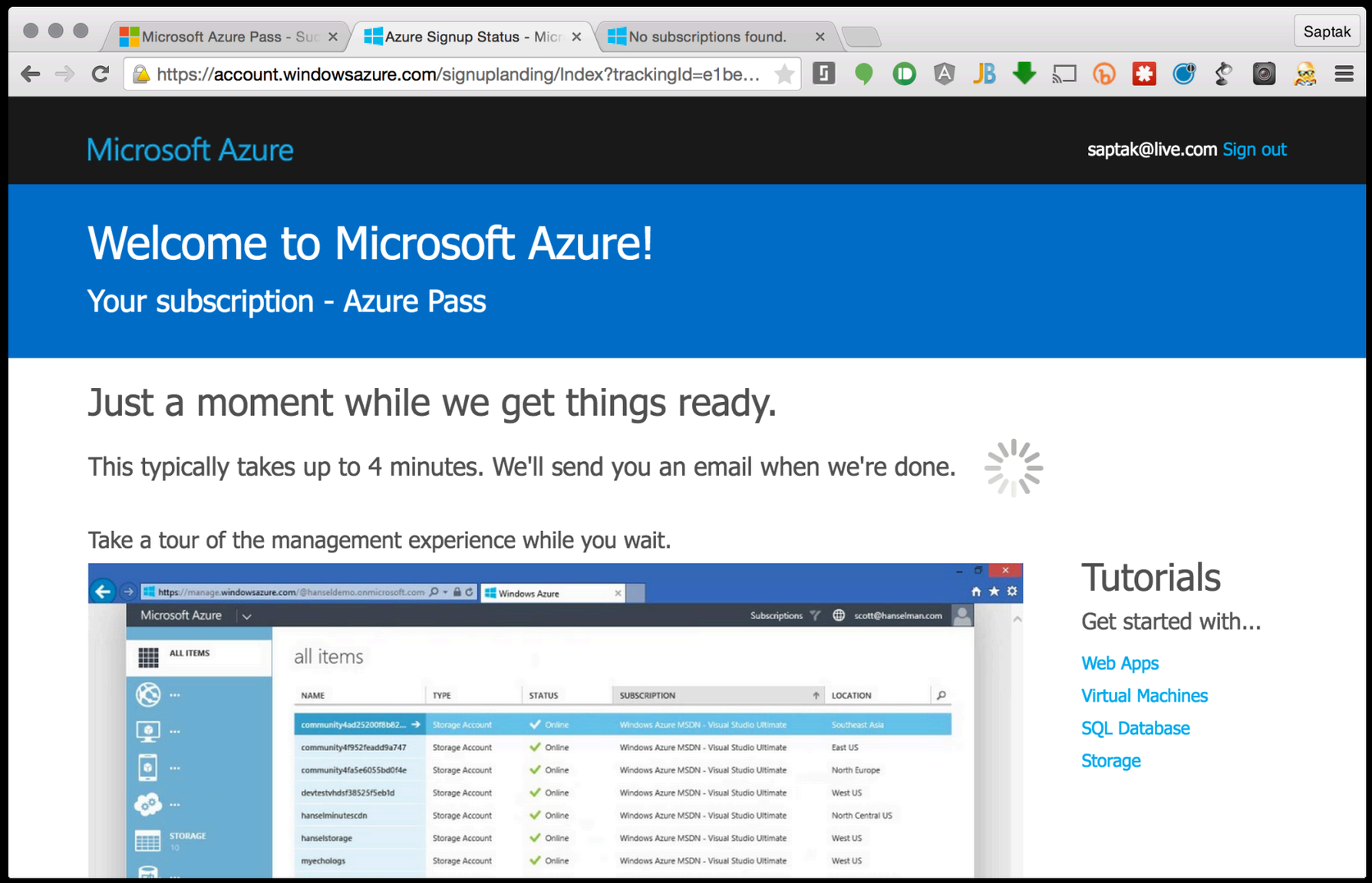
$100 credit/month <-> Hortonworks Sandbox node running on a A3 size VM for about ~ 240 hours/month



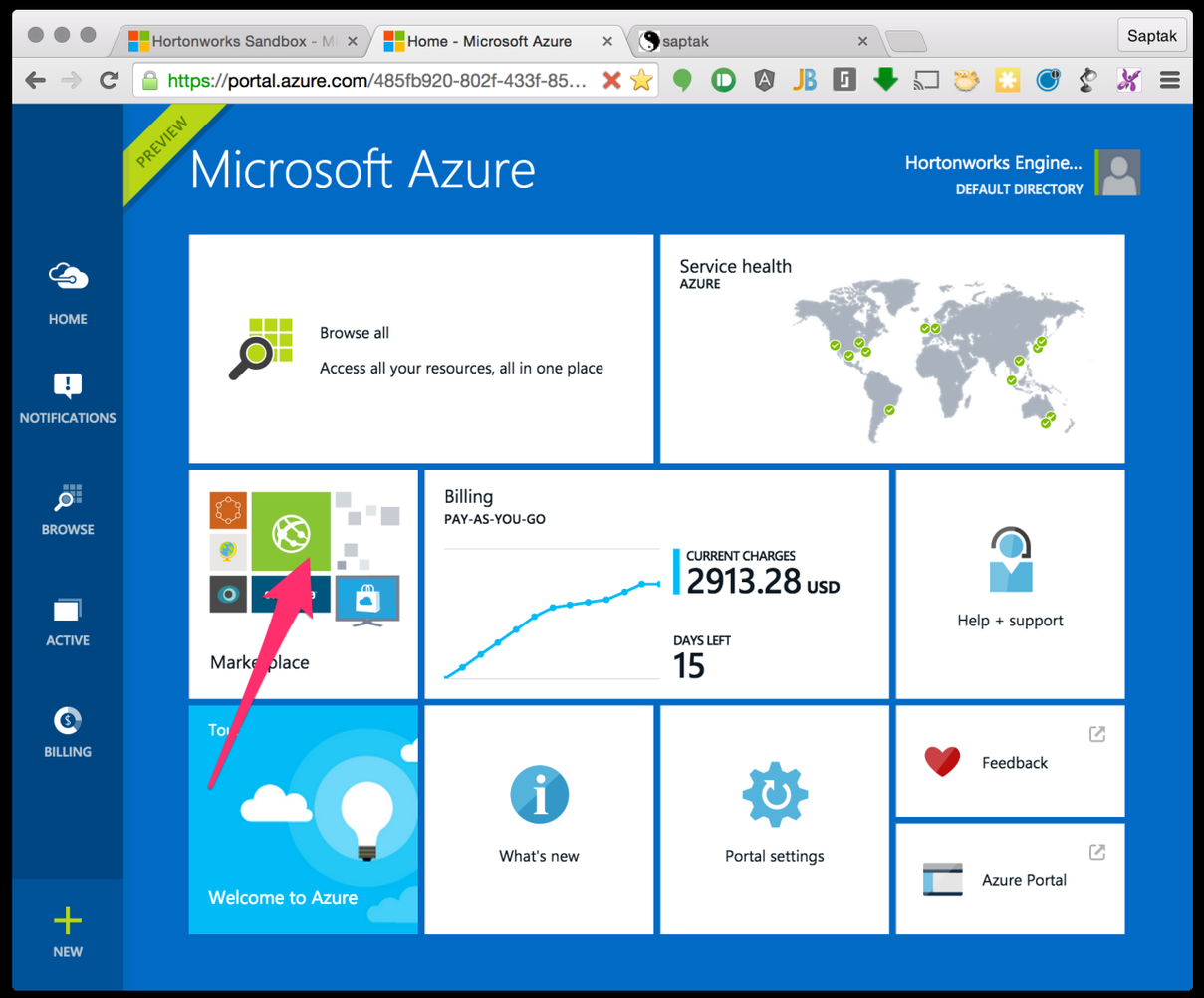
Complete the signup:



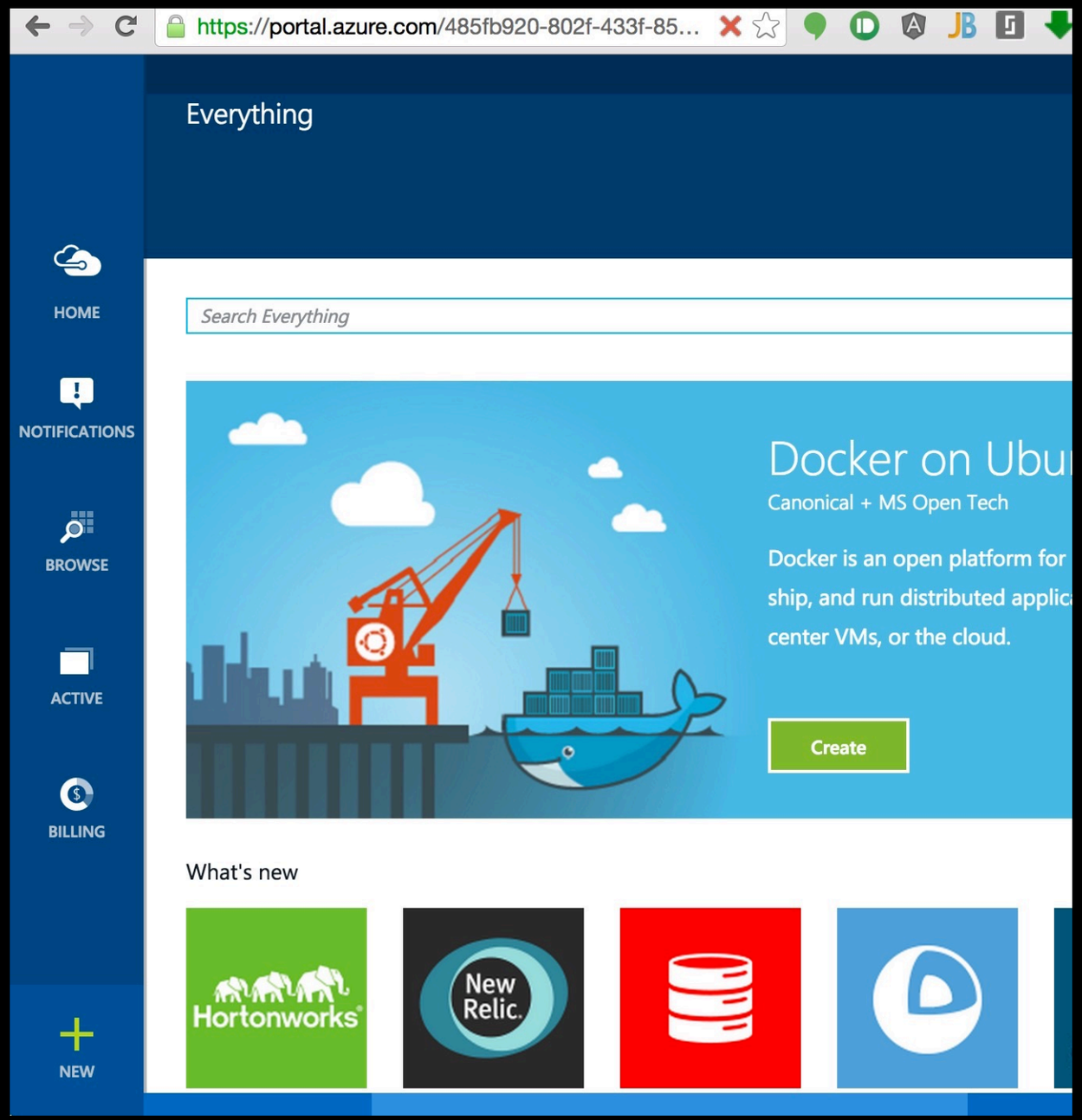
Wait, your Azure account will be ready in a few minutes



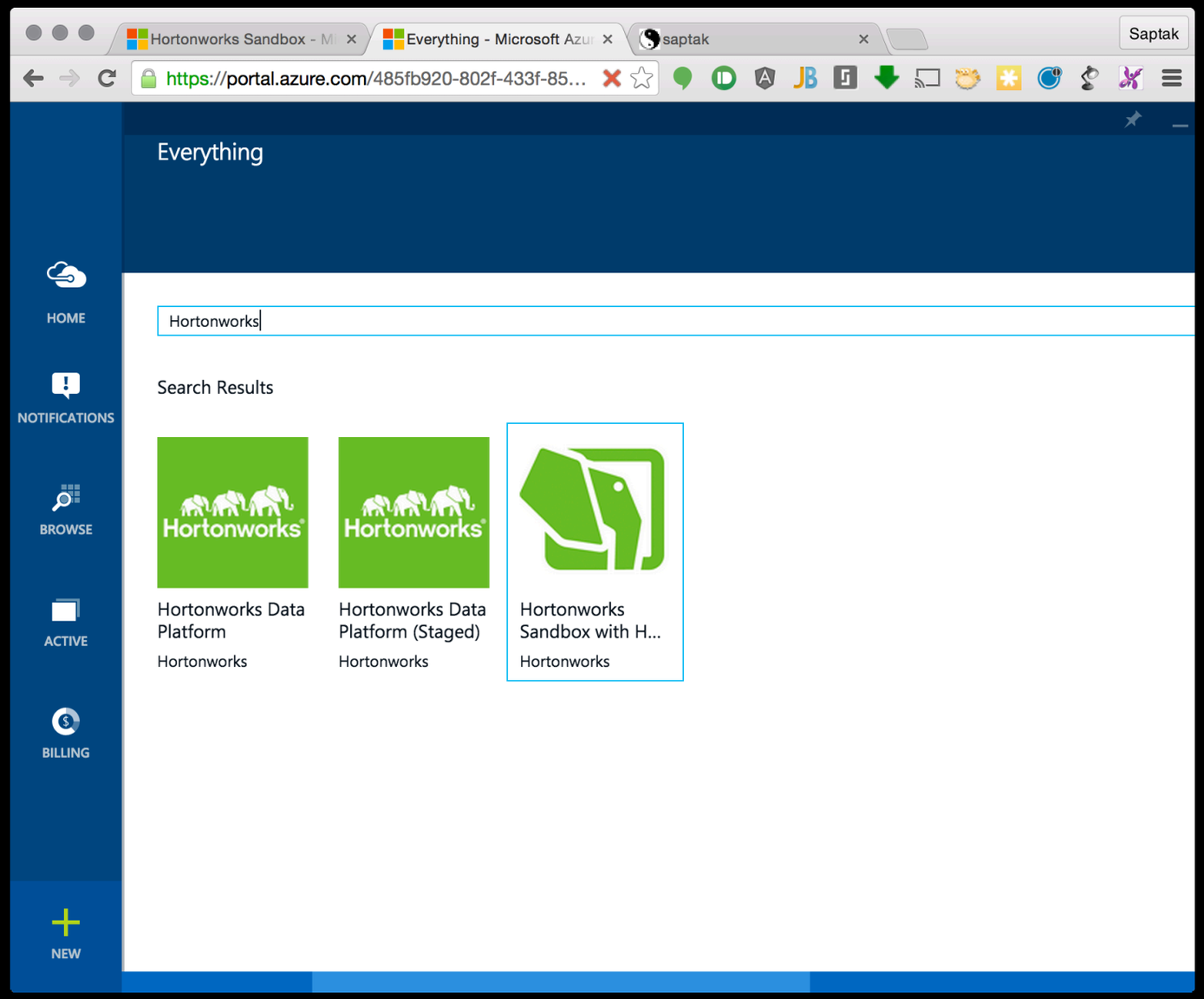
Start by logging into the Azure Portal with your Azure account: https://portal.azure.com/



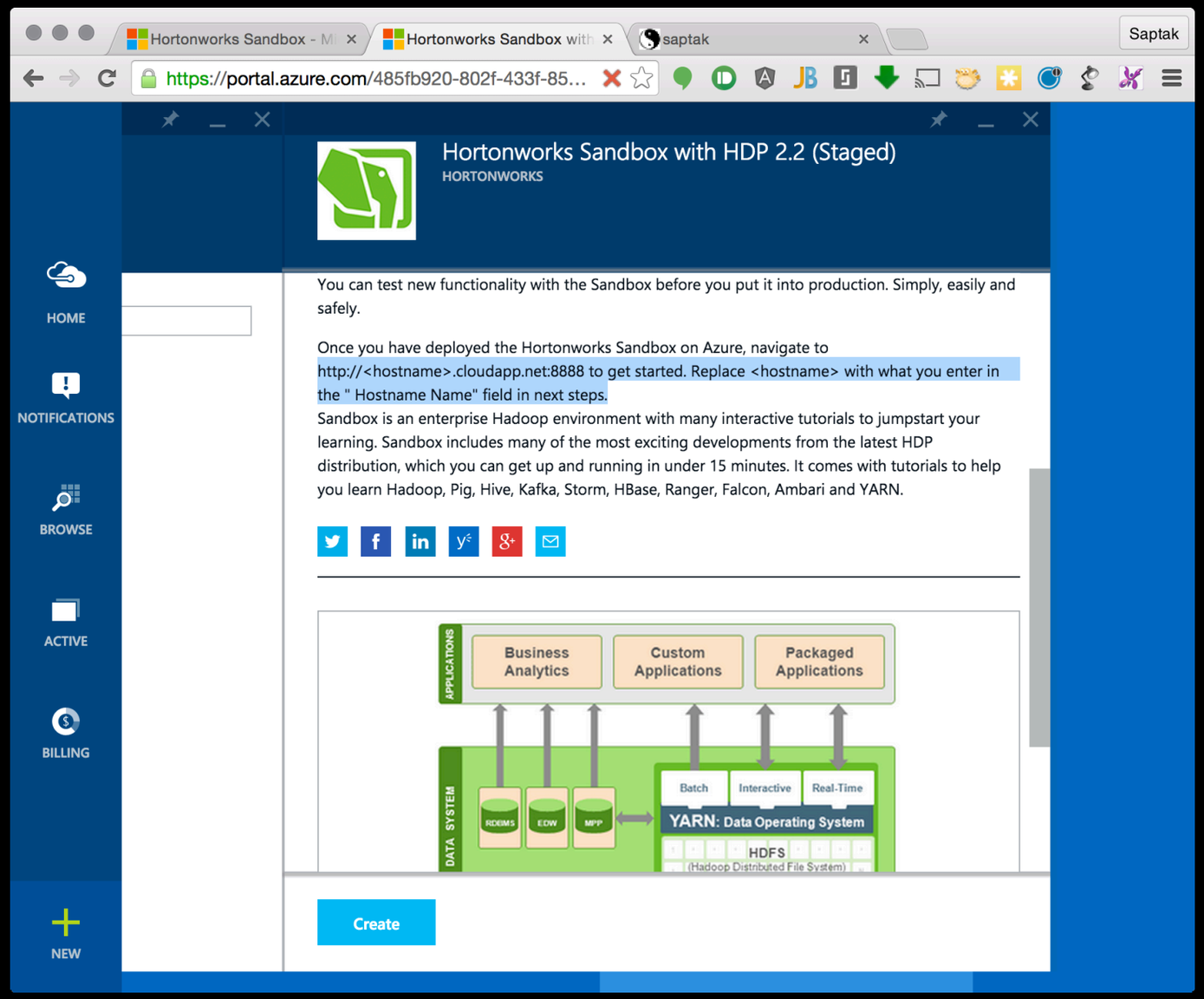
Navigate to the MarketPlace



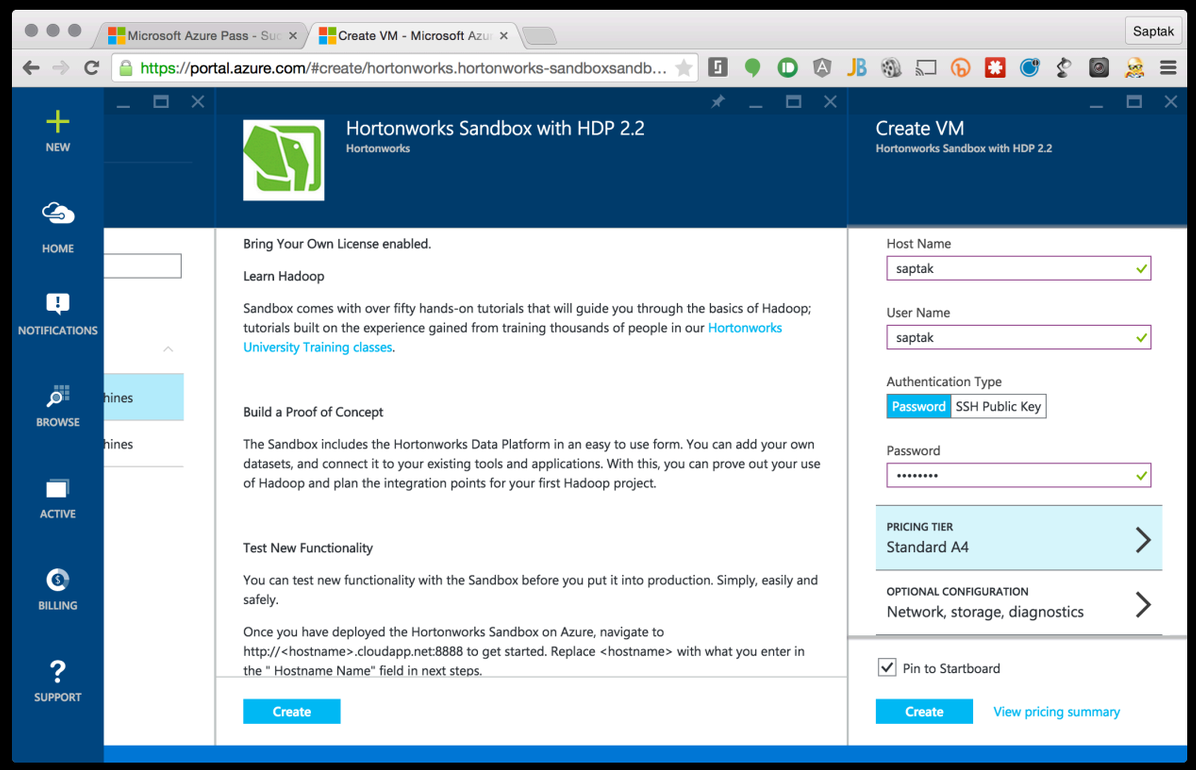
Search for Hortonworks. Click on the Hortonworks Sandbox icon.



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

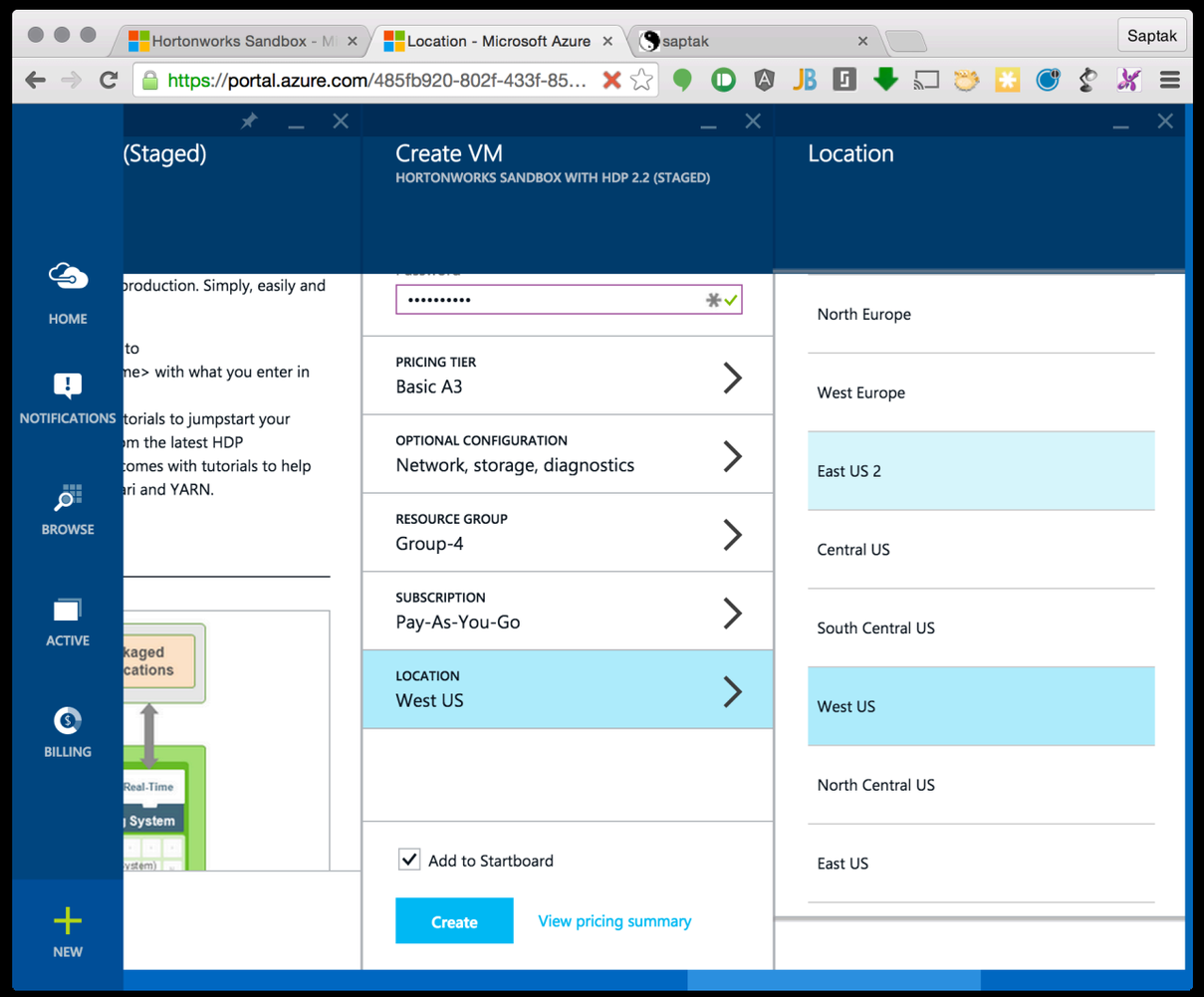


Note down the hostname and the username/ password that you enter

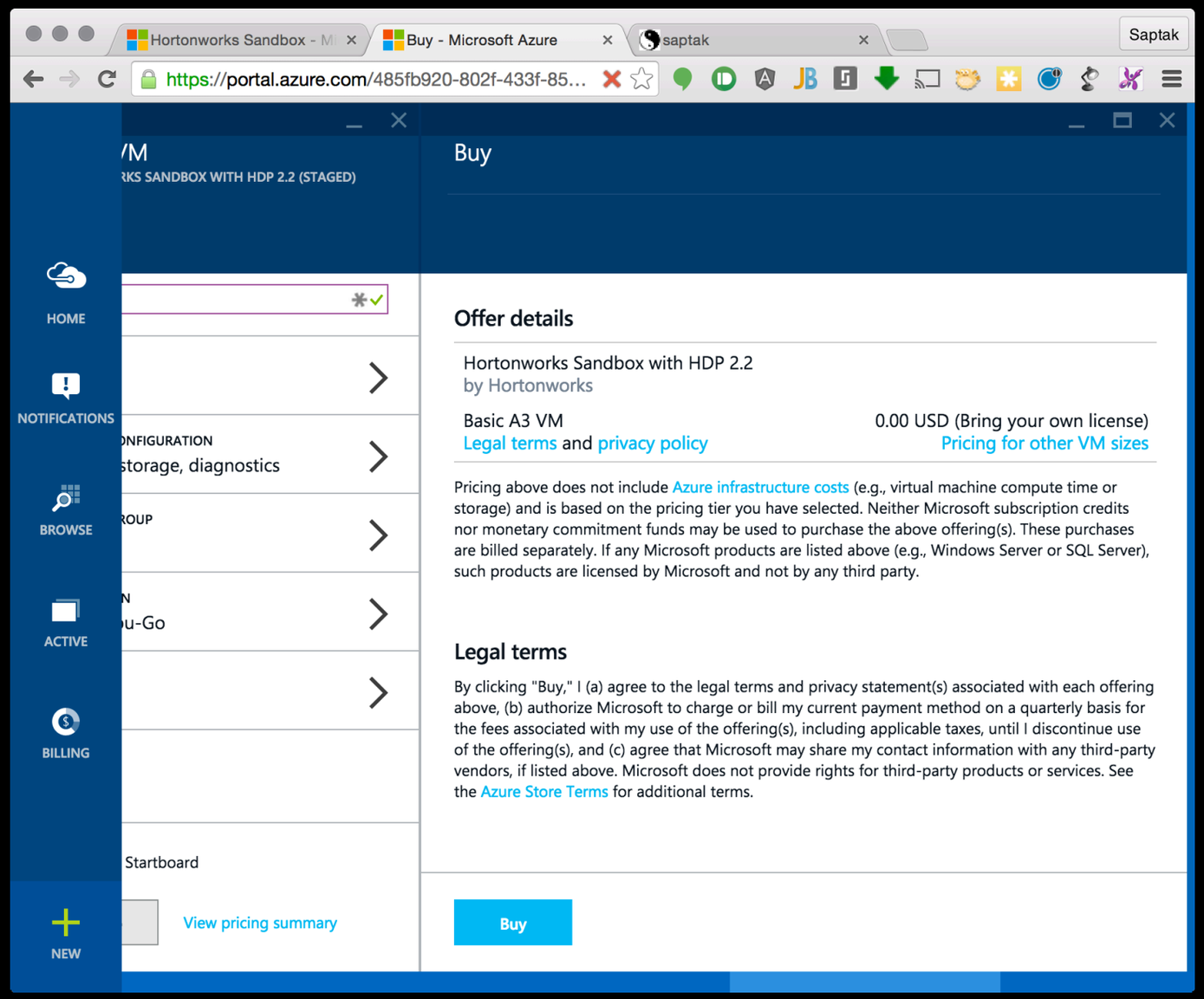


I recommend you select Standard A4 for the pricing tier

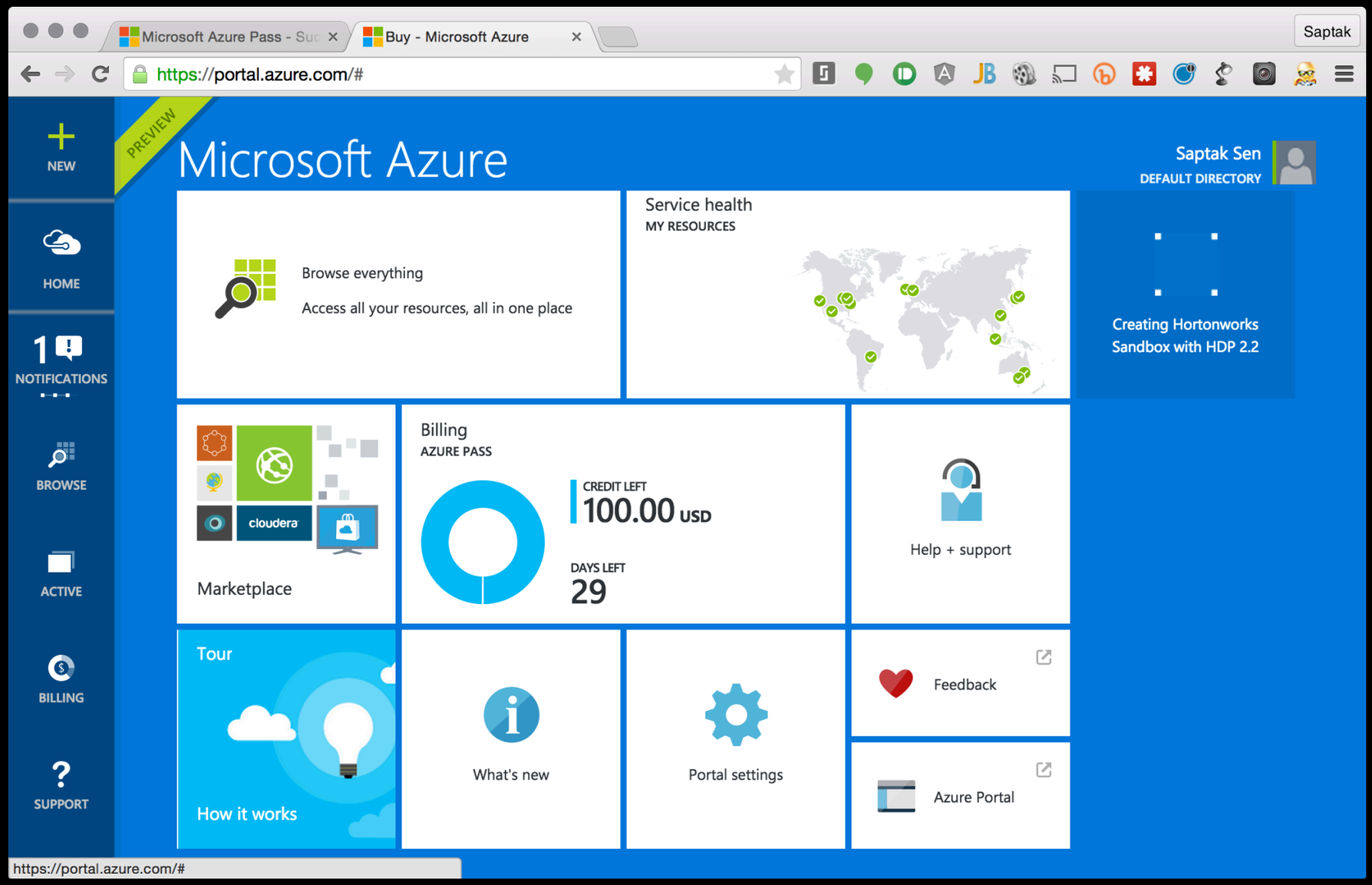
Change the location to the datacenter to where your preexisting Azure resources are or the one closest to you



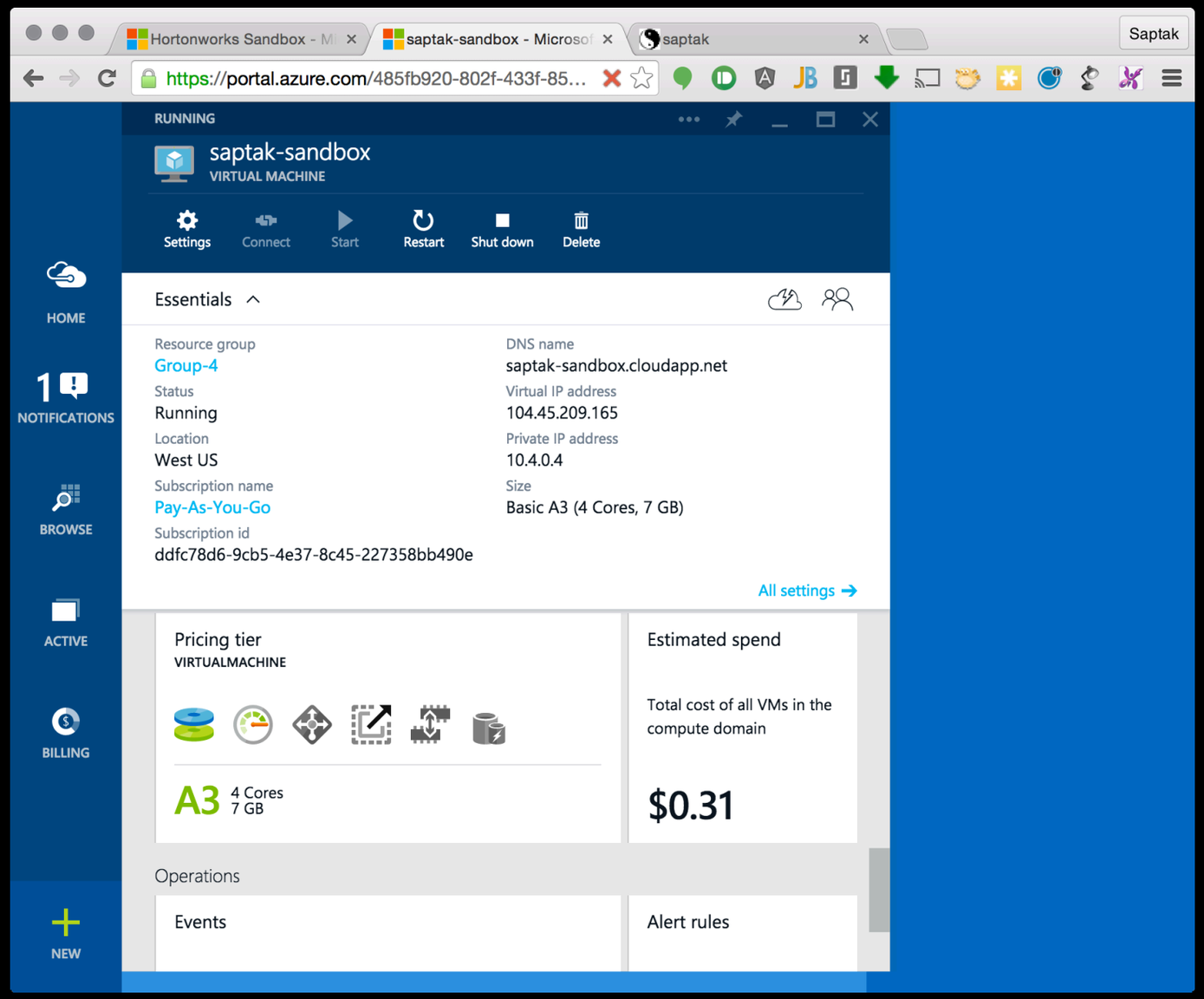
Click Buy if you agree with everything on this page.



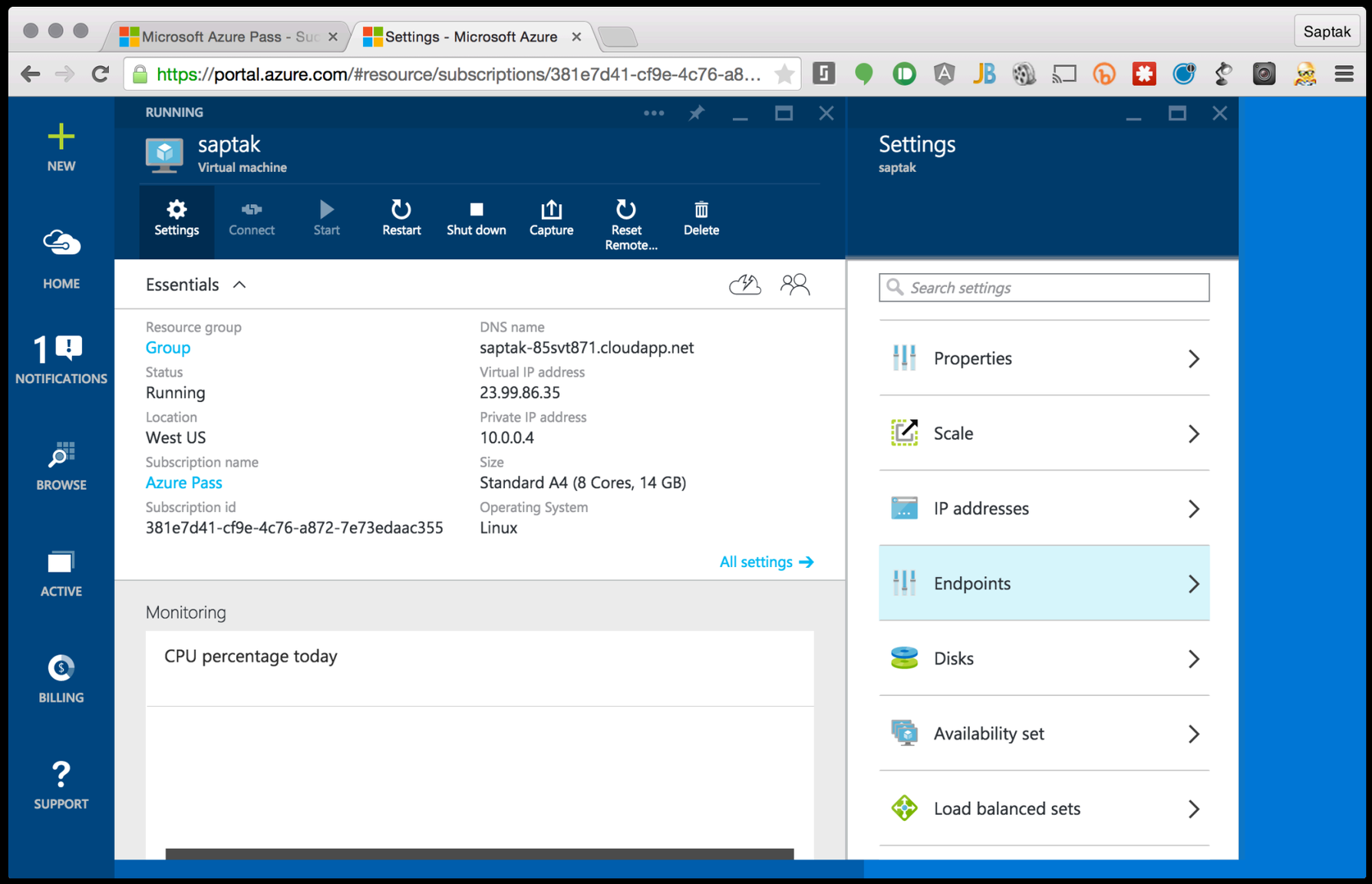
On the Azure portal home page, you can see the deployment in progress



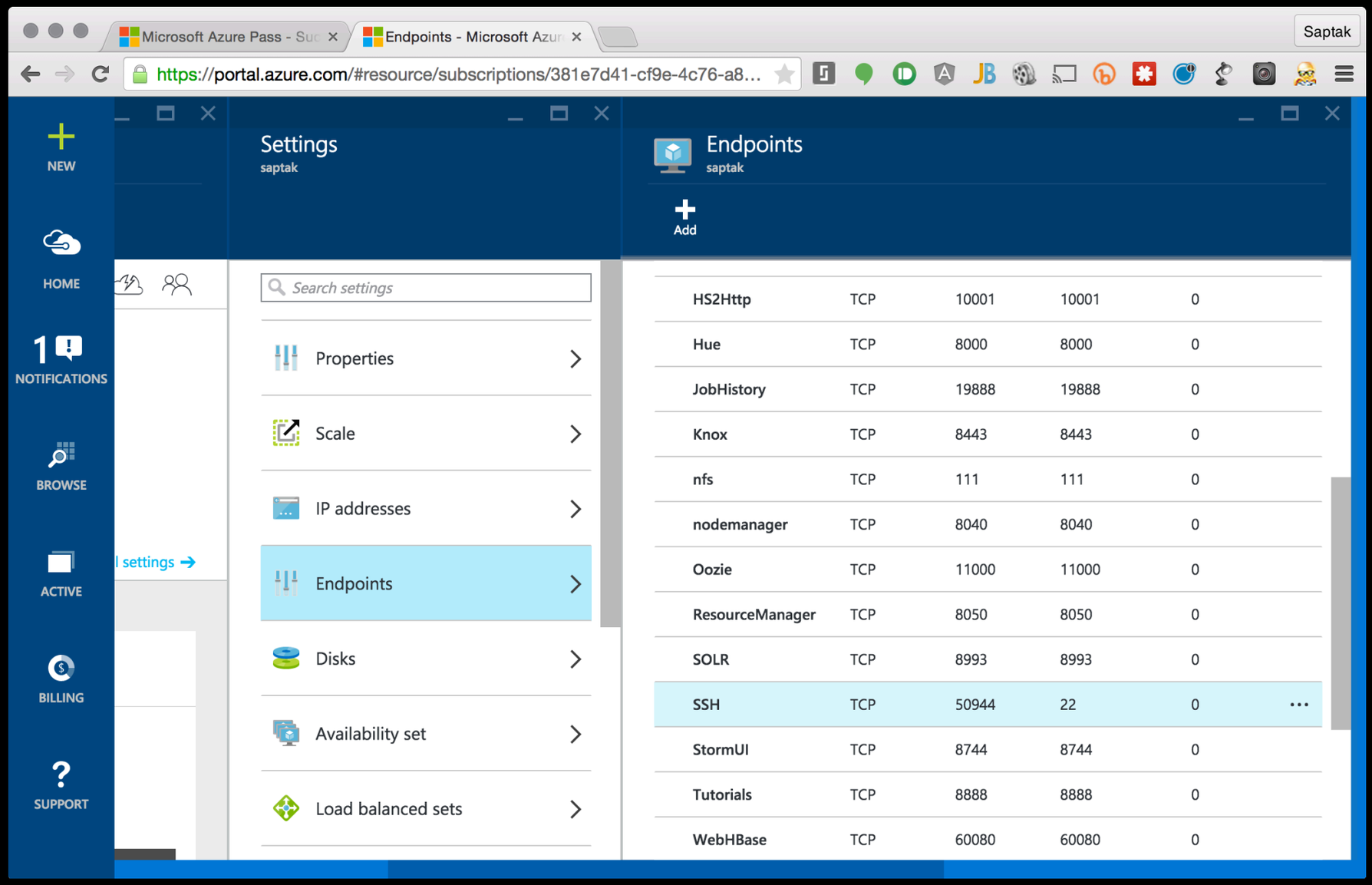
If you scroll down you can see the Estimated spend and other metrics for your VM



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

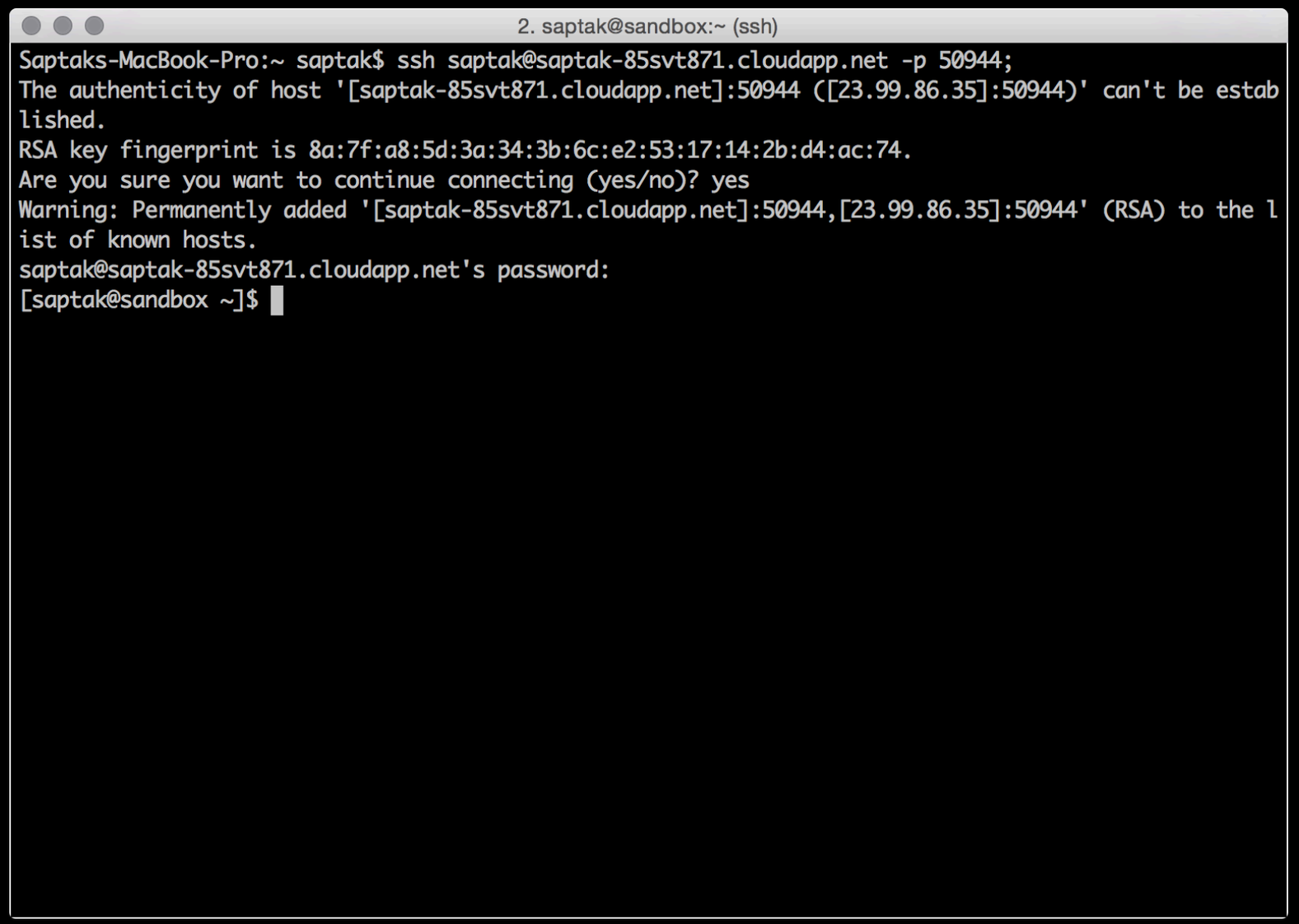


Click on Endpoints and scroll down to note the public SSH port for your VM

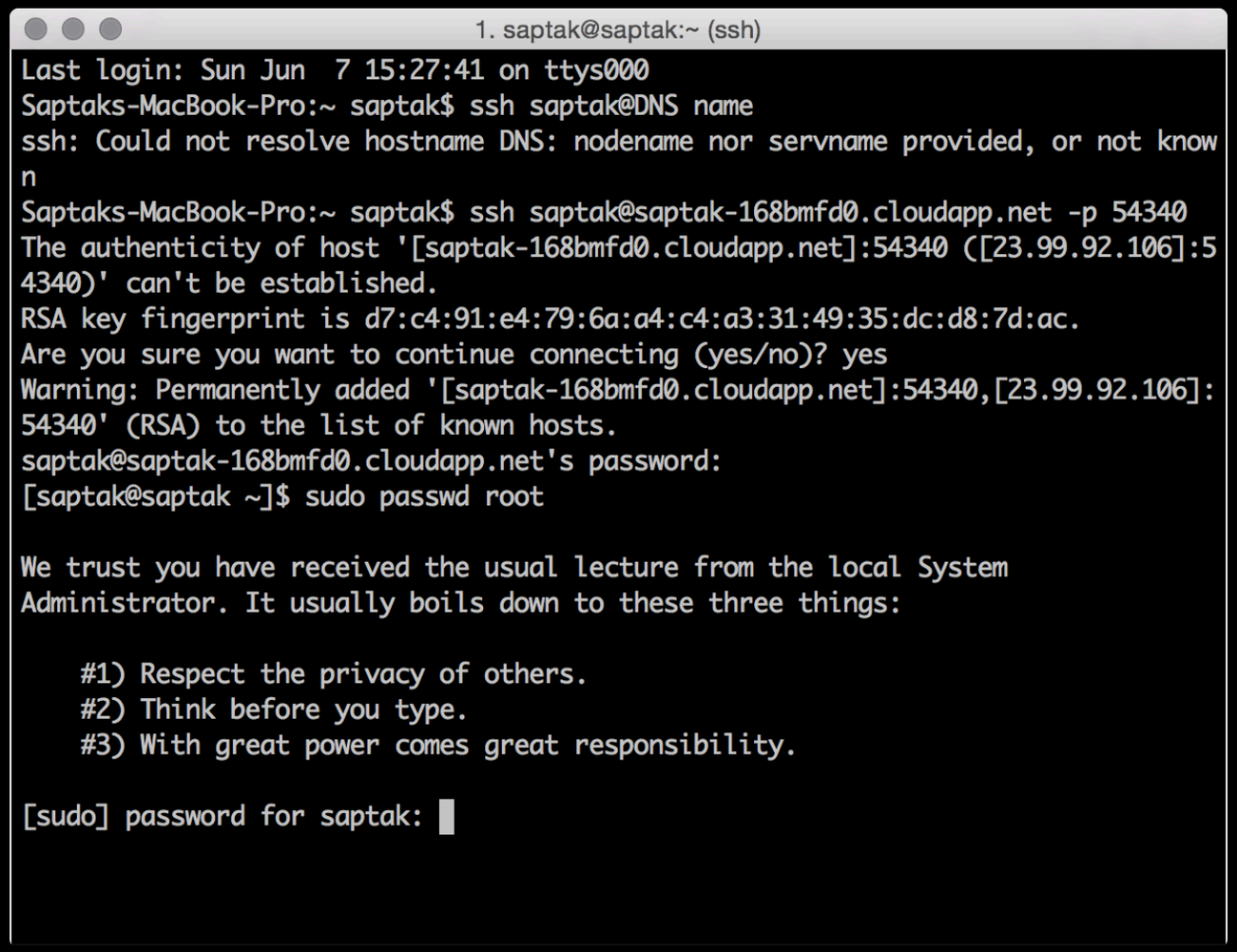


Now we can use the command below to login

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



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



Now we can login as root with the command su

1. saptak@saptak:~ (ssh)

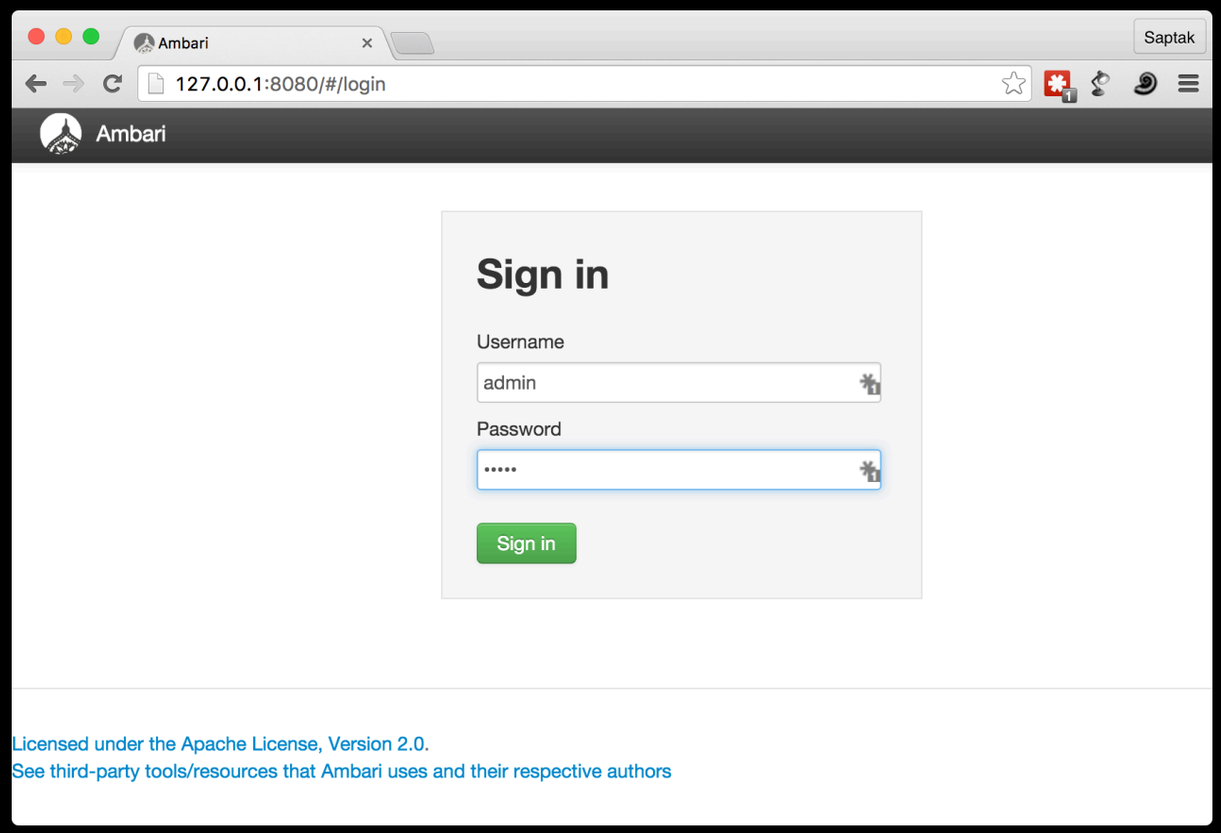
```
The authenticity of host '[saptak-168bmfd0.cloudapp.net]:54340 ([23.99.92.106]:5
4340)' can't be established.
RSA key fingerprint is d7:c4:91:e4:79:6a:a4:c4:a3:31:49:35:dc:d8:7d:ac.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '[saptak-168bmfd0.cloudapp.net]:54340,[23.99.92.106]:
54340' (RSA) to the list of known hosts.
saptak@saptak-168bmfd0.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

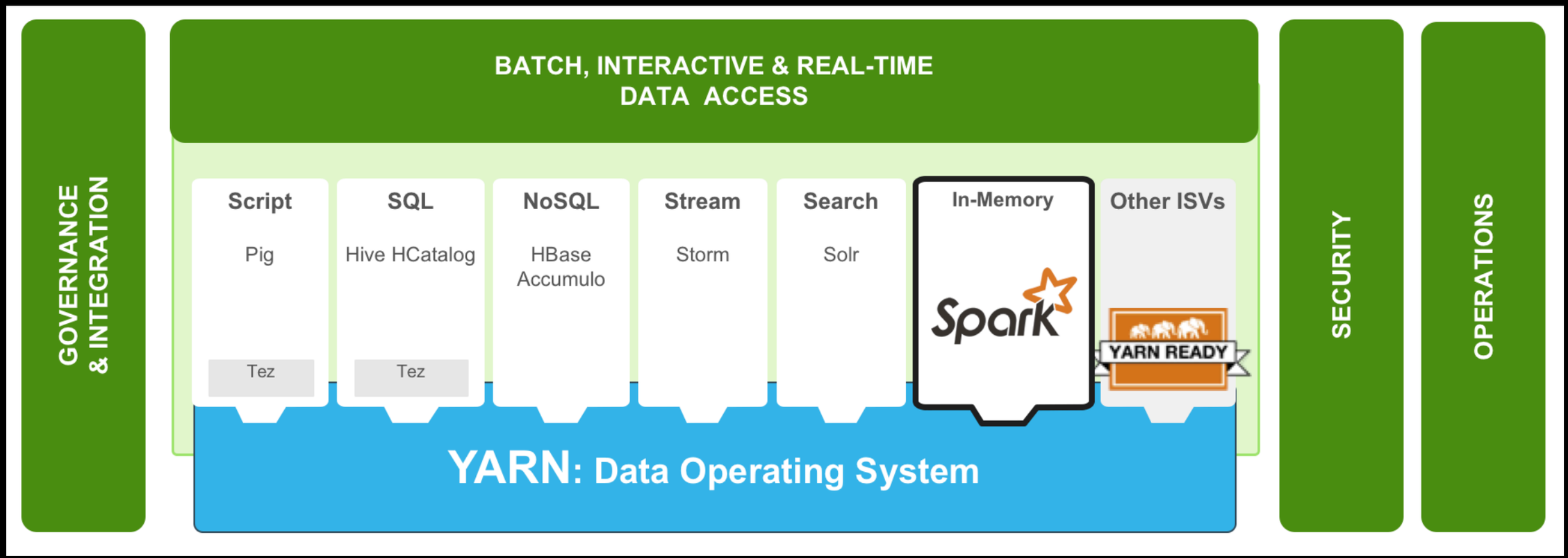
What is Hadoop?

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

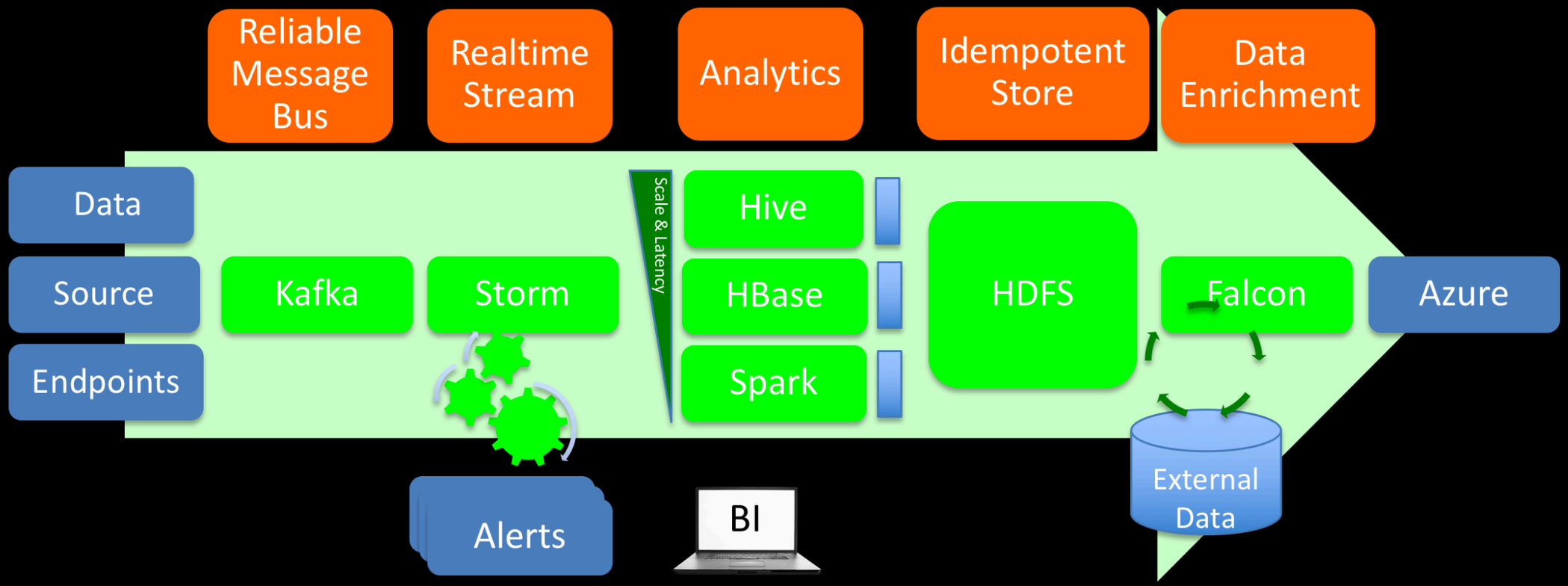
Structured Data

- 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:

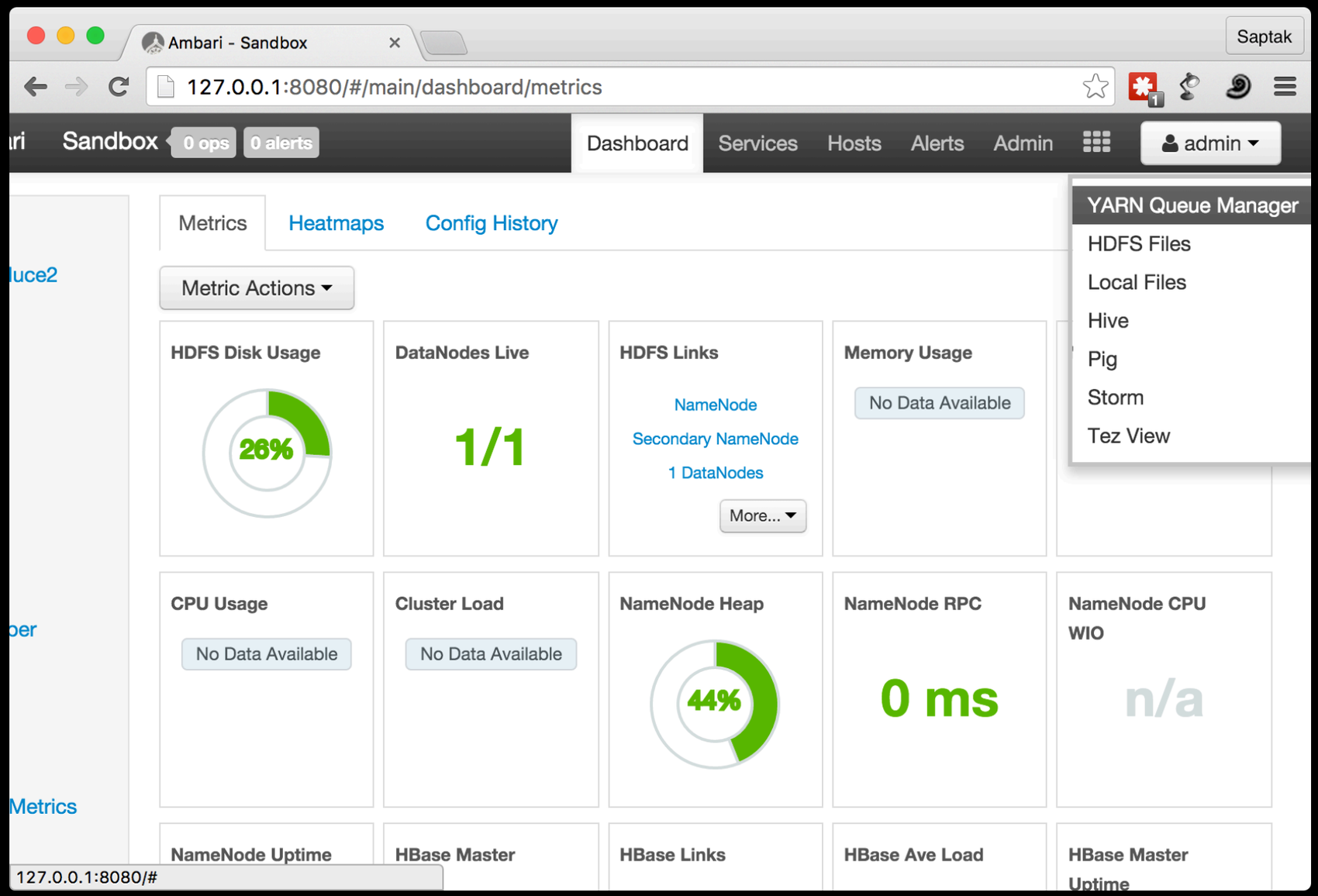
  ```
  hadoop fs -put local_file HDFS_dir_name/new_file
  ```

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

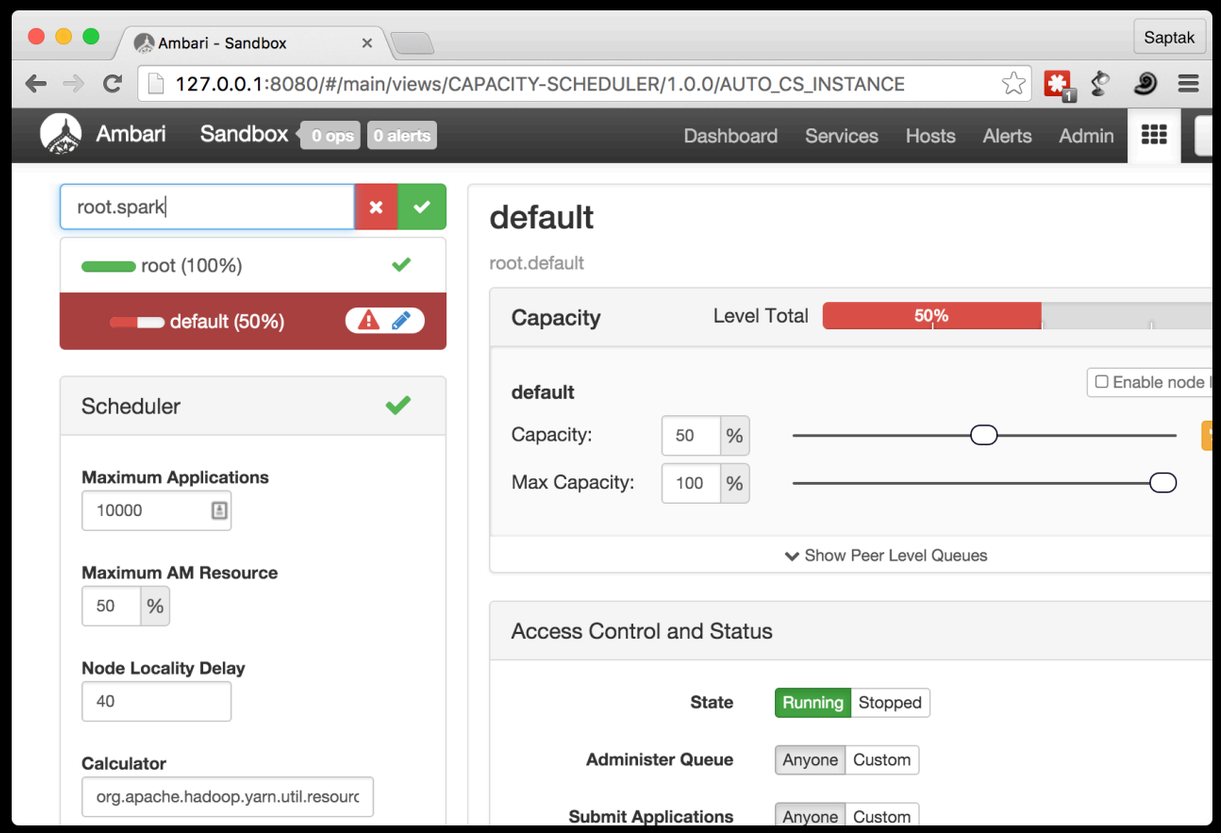
  ```
  hadoop fs -get HDFS_dir_name/file /local_dir/new_file
  ```

Lab

Open YARN Queue Manager

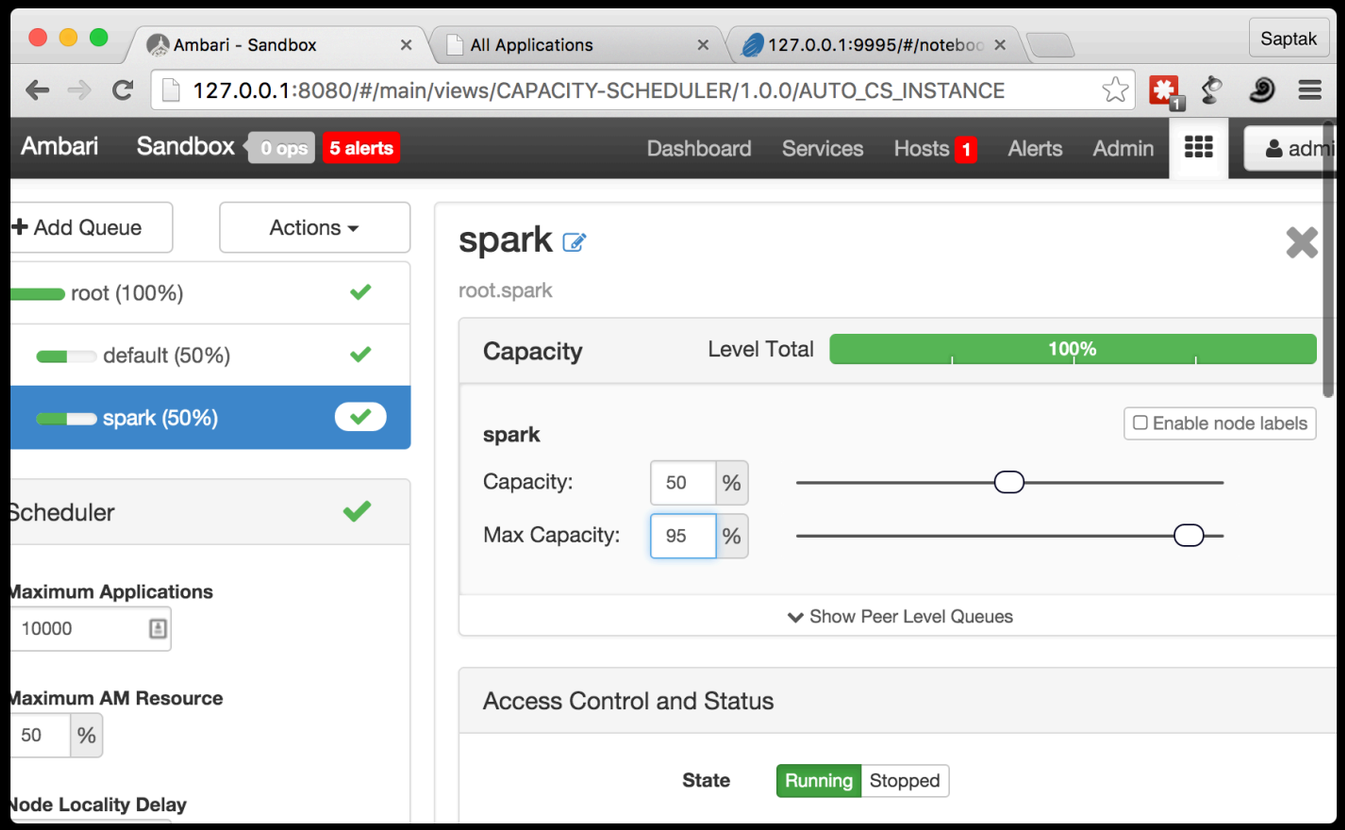


Create a dedicated queue for Spark



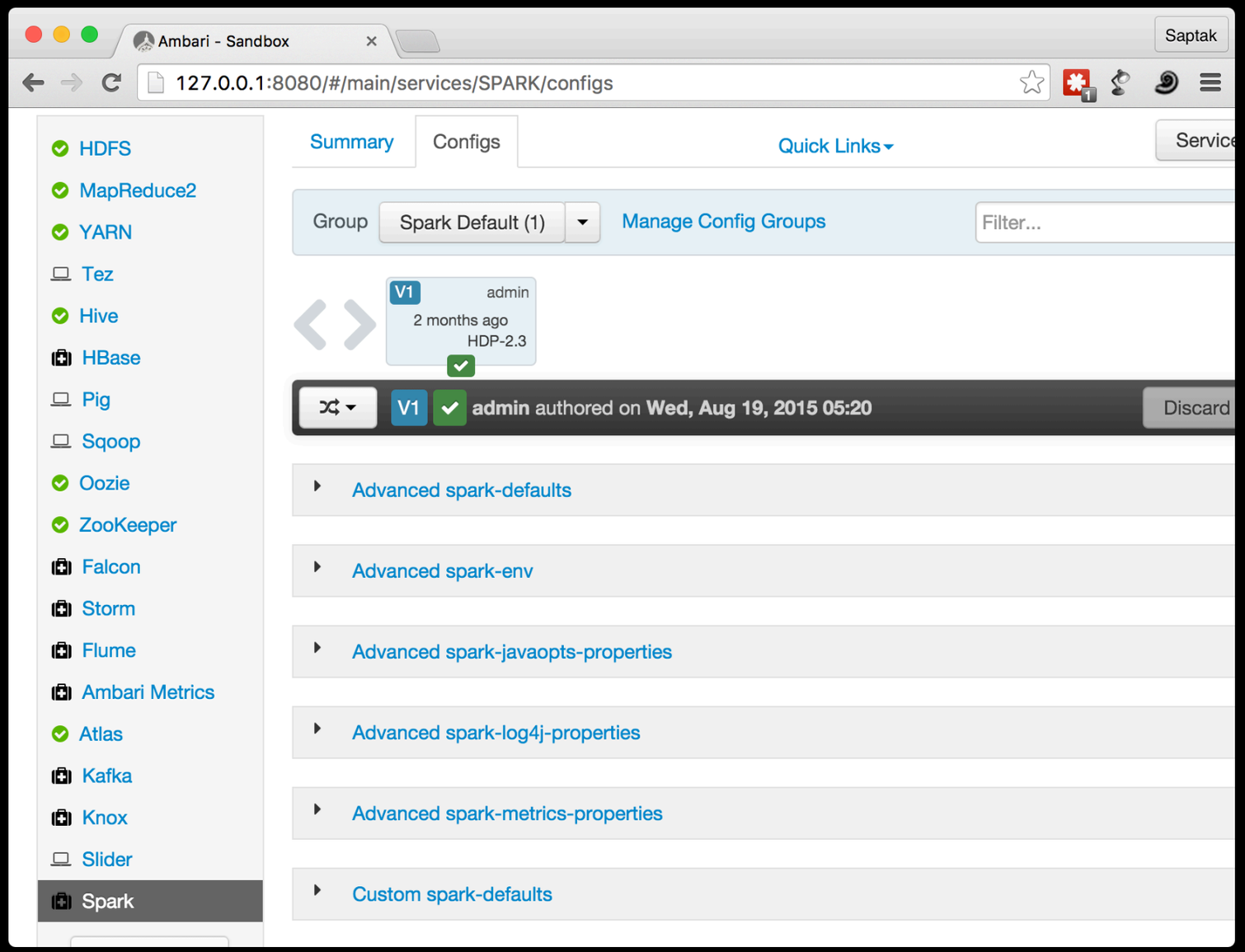
Set name to `root.spark`, Capacity to 50% and Max Capacity to 95%

Save and Refresh Queues

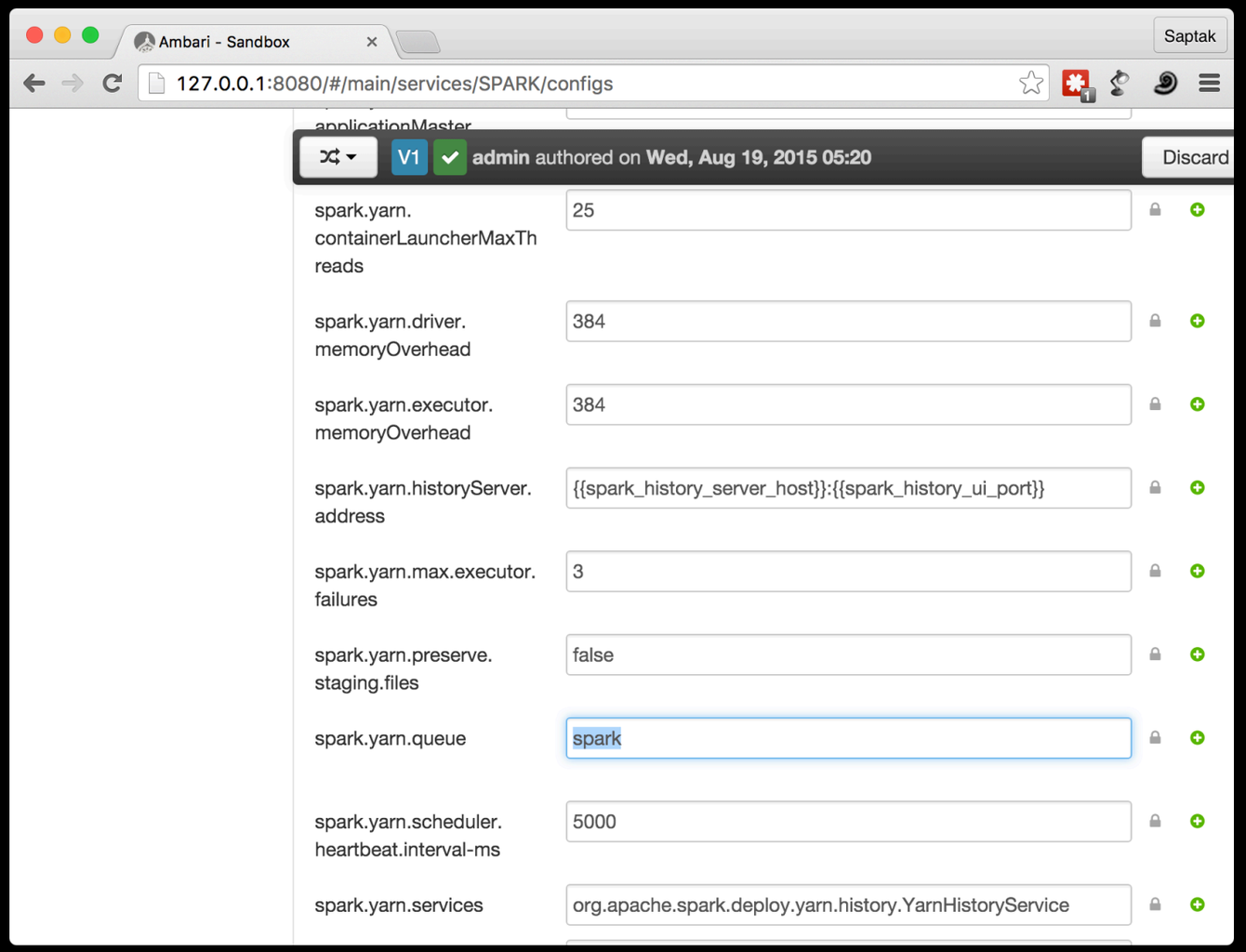


Adjust capacity of the default queue to 50% if required before saving

Open Apache Spark configuration



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



Restart All Affected Services