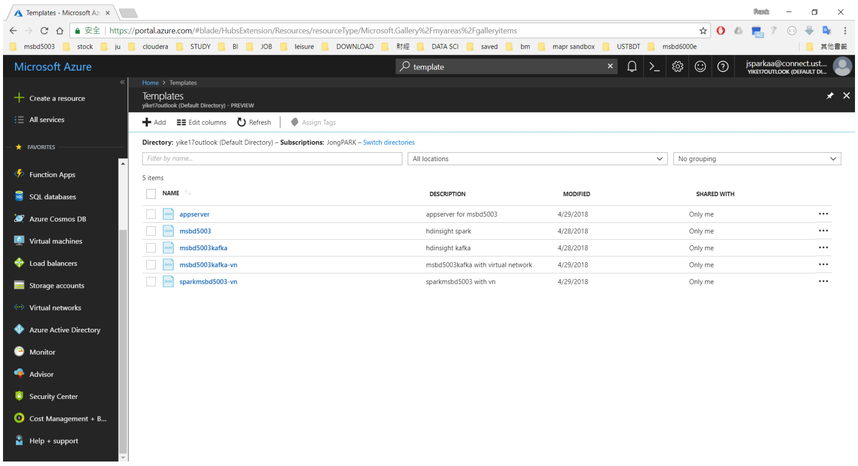


Templates
yike17outlook (Default Directory) - PREVIEW

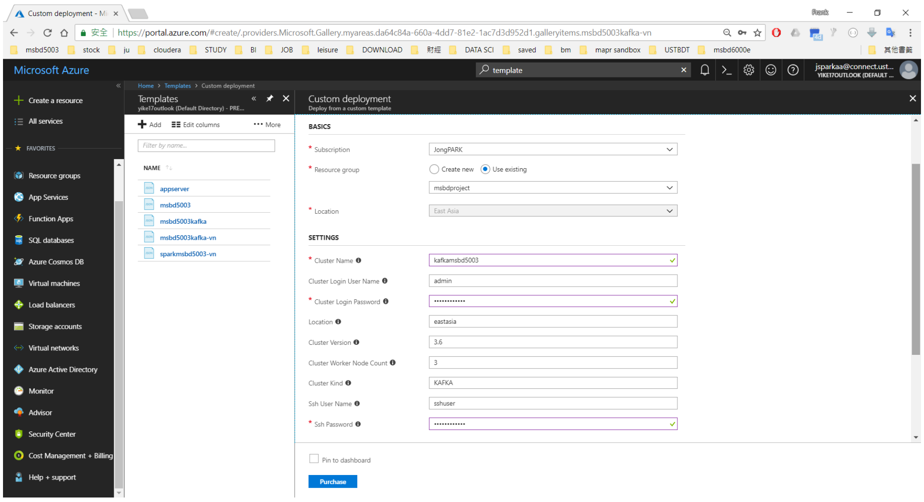
Directory: yike17outlook (Default Directory) – **Subscriptions:** JongPARK – Switch directories

Filter by name... | All locations | No grouping

5 items

NAME	DESCRIPTION	MODIFIED	SHARED WITH
appserver	appserver for msbd5003	4/29/2018	Only me
msbd5003	hdinsight spark	4/28/2018	Only me
msbd5003kafka	hdinsight kafka	4/28/2018	Only me
msbd5003kafka-vn	msbd5003kafka with virtual network	4/29/2018	Only me
sparkmsbd5003-vn	sparkmsbd5003 with vn	4/29/2018	Only me

Startup KAFKA cluster using msbd5003kafka-vn template. Cluster Name must be called “kafkamsbd5003”.



Custom deployment
Deploy from a custom template

BASICS

* Subscription	JongPARK
* Resource group	○ Create new ● Use existing
	msbdproject
* Location	East Asia

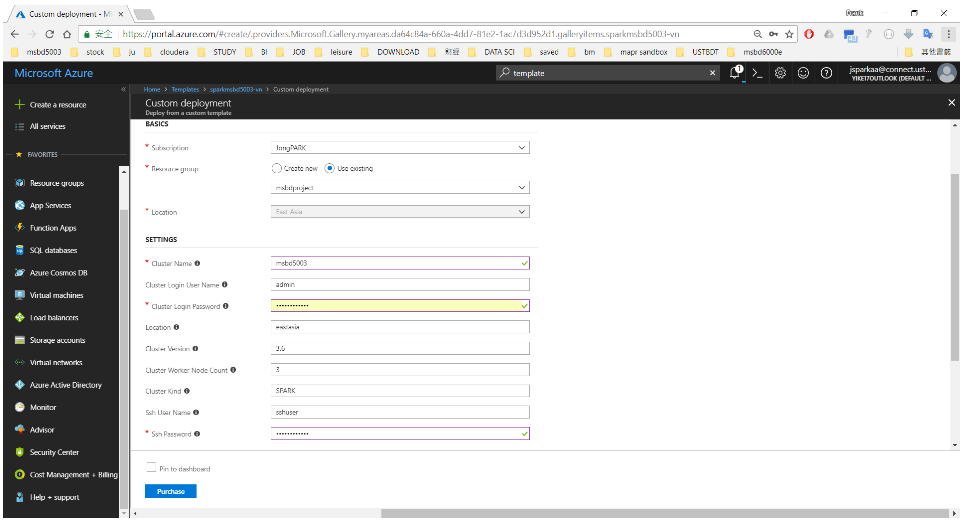
SETTINGS

* Cluster Name	kafkamsbd5003 ✓
Cluster Login User Name	admin
* Cluster Login Password	•••••••••••• ✓
Location	eastasia
Cluster Version	3.6
Cluster Worker Node Count	3
Cluster Kind	KAFKA
Ssh User Name	sshuser
* Ssh Password	•••••••••••• ✓

□ Pin to dashboard

Purchase

Startup SPARK cluster using sparkmsbd5003-vn template. Cluster name must be called “msbd5003”.



Custom deployment
Deploy from a custom template

BASICS

* Subscription	JongPARK
* Resource group	○ Create new ● Use existing
	msbdproject
* Location	East Asia

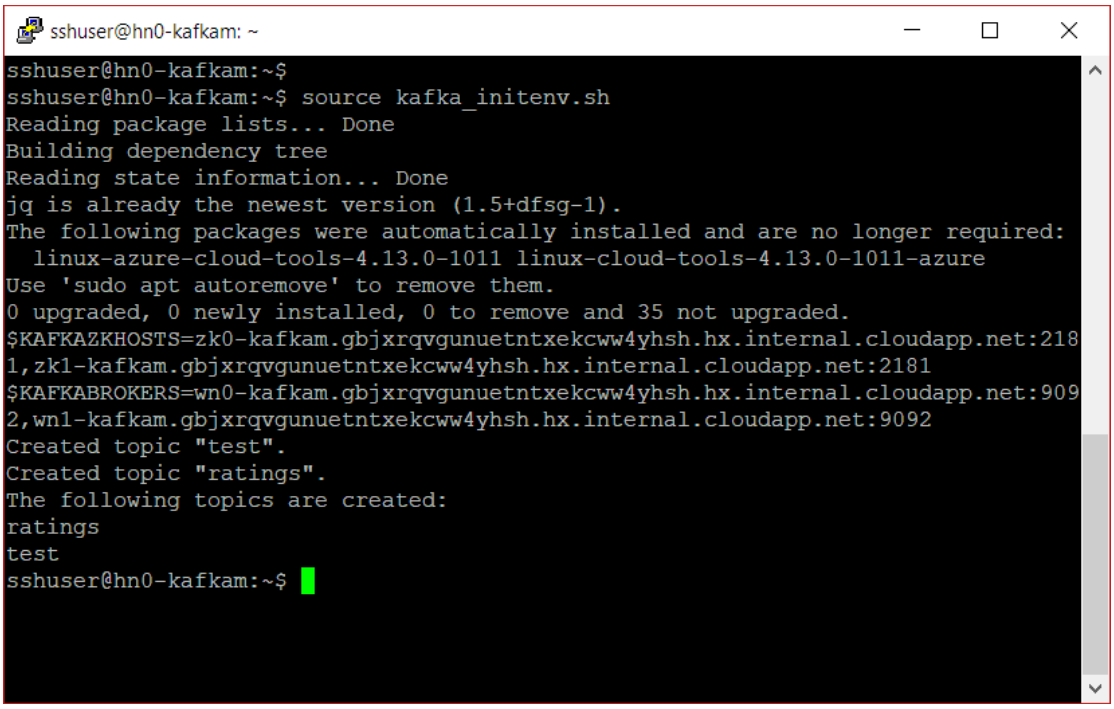
SETTINGS

* Cluster Name	msbd5003 ✓
Cluster Login User Name	admin
* Cluster Login Password	•••••••••••• ✓
Location	eastasia
Cluster Version	3.6
Cluster Worker Node Count	3
Cluster Kind	SPARK
Ssh User Name	sshuser
* Ssh Password	•••••••••••• ✓

□ Pin to dashboard

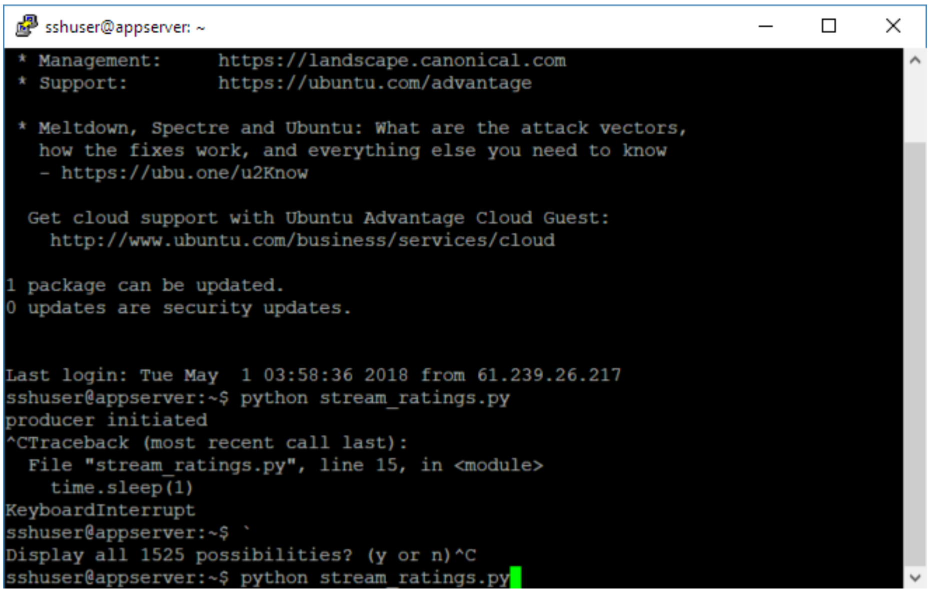
Purchase

Run source kafka_initenv.sh on kafkamsbd5003 to create the kafka topic "ratings".



```
sshuser@hn0-kafkam:~$
sshuser@hn0-kafkam:~$ source kafka_initenv.sh
Reading package lists... Done
Building dependency tree
Reading state information... Done
jq is already the newest version (1.5+dfsg-1).
The following packages were automatically installed and are no longer required:
  linux-azure-cloud-tools-4.13.0-1011 linux-cloud-tools-4.13.0-1011-azure
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 35 not upgraded.
$KAFKAZKHOSTS=zk0-kafkam.gbjxrqvgunuetntxekcww4yhsh.hx.internal.cloudapp.net:218
1,zk1-kafkam.gbjxrqvgunuetntxekcww4yhsh.hx.internal.cloudapp.net:2181
$KAFKABROKERS=wn0-kafkam.gbjxrqvgunuetntxekcww4yhsh.hx.internal.cloudapp.net:909
2,wn1-kafkam.gbjxrqvgunuetntxekcww4yhsh.hx.internal.cloudapp.net:9092
Created topic "test".
Created topic "ratings".
The following topics are created:
ratings
test
sshuser@hn0-kafkam:~$
```

Run "python stream_rating.py" on appserver VM. the program will put a rating message to the queue every 1 second:

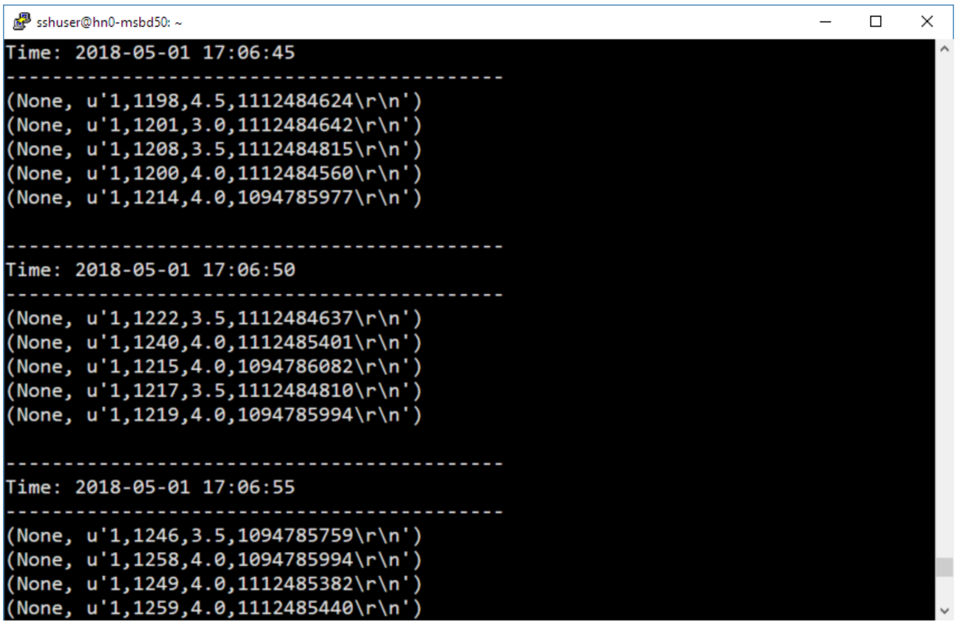


```
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Meltdown, Spectre and Ubuntu: What are the attack vectors,
   how the fixes work, and everything else you need to know
   - https://ubu.one/u2Know

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

1 package can be updated.
0 updates are security updates.


Last login: Tue May  1 03:58:36 2018 from 61.239.26.217
sshuser@appserver:~$ python stream_ratings.py
producer initiated
^CTraceback (most recent call last):
  File "stream_ratings.py", line 15, in <module>
    time.sleep(1)
KeyboardInterrupt
sshuser@appserver:~$ `
Display all 1525 possibilities? (y or n)^C
sshuser@appserver:~$ python stream_ratings.py
```

Run sample kafka consumer program on SPARK cluster “msbd5003”:

DirectStream (real time):
spark-submit --packages org.apache.spark:spark-streaming-kafka-0-8_2.11:2.2.0 ratings_direct_stream.py 2>log_err

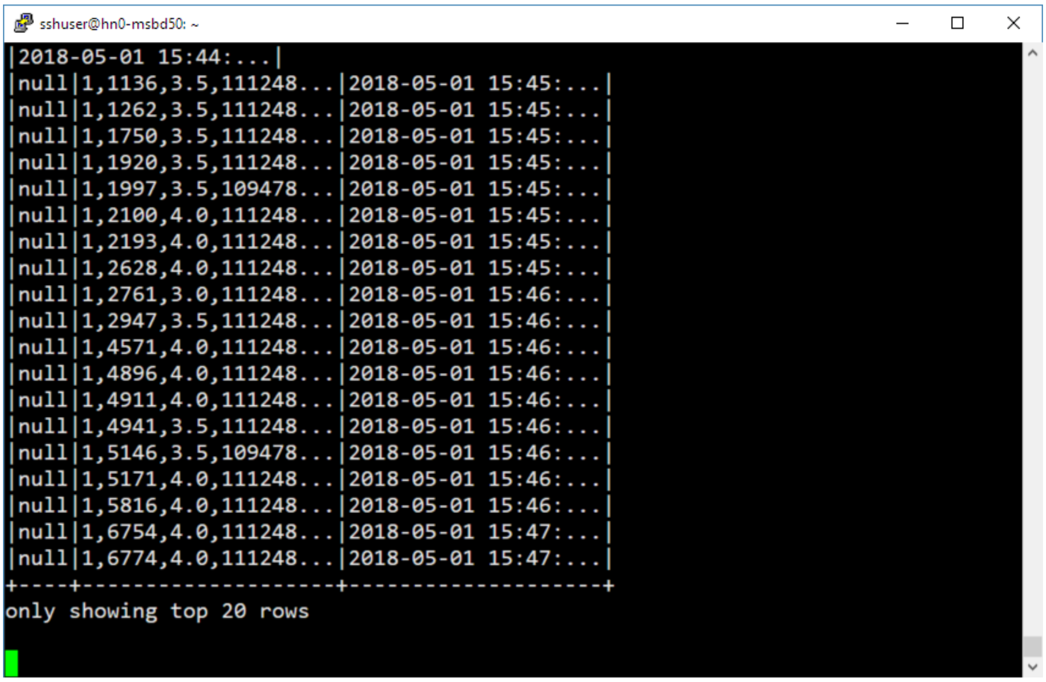
OR open jupyter notebook ratings_direct_stream.ipynb



```
Time: 2018-05-01 17:06:45
-------------------------------------------
(None, u'1,1198,4.5,1112484624\r\n')
(None, u'1,1201,3.0,1112484642\r\n')
(None, u'1,1208,3.5,1112484815\r\n')
(None, u'1,1200,4.0,1112484560\r\n')
(None, u'1,1214,4.0,1094785977\r\n')

-------------------------------------------
Time: 2018-05-01 17:06:50
-------------------------------------------
(None, u'1,1222,3.5,1112484637\r\n')
(None, u'1,1240,4.0,1112485401\r\n')
(None, u'1,1215,4.0,1094786082\r\n')
(None, u'1,1217,3.5,1112484810\r\n')
(None, u'1,1219,4.0,1094785994\r\n')

-------------------------------------------
Time: 2018-05-01 17:06:55
-------------------------------------------
(None, u'1,1246,3.5,1094785759\r\n')
(None, u'1,1258,4.0,1094785994\r\n')
(None, u'1,1249,4.0,1112485382\r\n')
(None, u'1,1259,4.0,1112485440\r\n')
```

Structured Streaming (accumulated from beginning):
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.2.0 ratings_structured_streaming.py 2>log_err

OR open ratings_structured_streaming.ipynb (however query.writeStream.format("console") will not print as jupyter output)



```
|2018-05-01 15:44:...|
|null|1,1136,3.5,111248...|2018-05-01 15:45:...|
|null|1,1262,3.5,111248...|2018-05-01 15:45:...|
|null|1,1750,3.5,111248...|2018-05-01 15:45:...|
|null|1,1920,3.5,111248...|2018-05-01 15:45:...|
|null|1,1997,3.5,109478...|2018-05-01 15:45:...|
|null|1,2100,4.0,111248...|2018-05-01 15:45:...|
|null|1,2193,4.0,111248...|2018-05-01 15:45:...|
|null|1,2628,4.0,111248...|2018-05-01 15:45:...|
|null|1,2761,3.0,111248...|2018-05-01 15:46:...|
|null|1,2947,3.5,111248...|2018-05-01 15:46:...|
|null|1,4571,4.0,111248...|2018-05-01 15:46:...|
|null|1,4896,4.0,111248...|2018-05-01 15:46:...|
|null|1,4911,4.0,111248...|2018-05-01 15:46:...|
|null|1,4941,3.5,111248...|2018-05-01 15:46:...|
|null|1,5146,3.5,109478...|2018-05-01 15:46:...|
|null|1,5171,4.0,111248...|2018-05-01 15:46:...|
|null|1,5816,4.0,111248...|2018-05-01 15:46:...|
|null|1,6754,4.0,111248...|2018-05-01 15:47:...|
|null|1,6774,4.0,111248...|2018-05-01 15:47:...|
+----+--------------------+--------------------+
only showing top 20 rows
```