**Windows HDInsight 群集向Linux 环境迁移示例之HBase篇**

Azure HDInsight作为Azure上的大数据分布式处理平台，目前支持基于Windows和Linux的群集。目前Windows环境中的HDInsight最高版本HDI 3.3即将在2017年7月31日停止支持，因此需要将其上的Windows群集迁移到Linux环境中更高HDI版本的群集。Global Azure官网上的文章(以下简称官网文章) <https://docs.azure.cn/zh-cn/hdinsight/hdinsight-hbase-tutorial-get-started-linux> 介绍了此项迁移任务的大致操作步骤，但并未给出具体示例。本文旨在介绍HDInsight迁移的一个简单示例，以供读者参考。

截至2017年6月30日，Windows平台上支持的HDInsight群集类型及最高版本为Hadoop 2.7.0(HDI 3.3)，HBase 1.1.2(HDI 3.3)和Storm 0.9.3(HDI 3.3)。本次迁移示例是将Windows HDInsight上的HBase 1.1.2 (HDI 3.3) 迁移到Linux HDInsight上的HBase 1.1.2 (HDI 3.5) 。

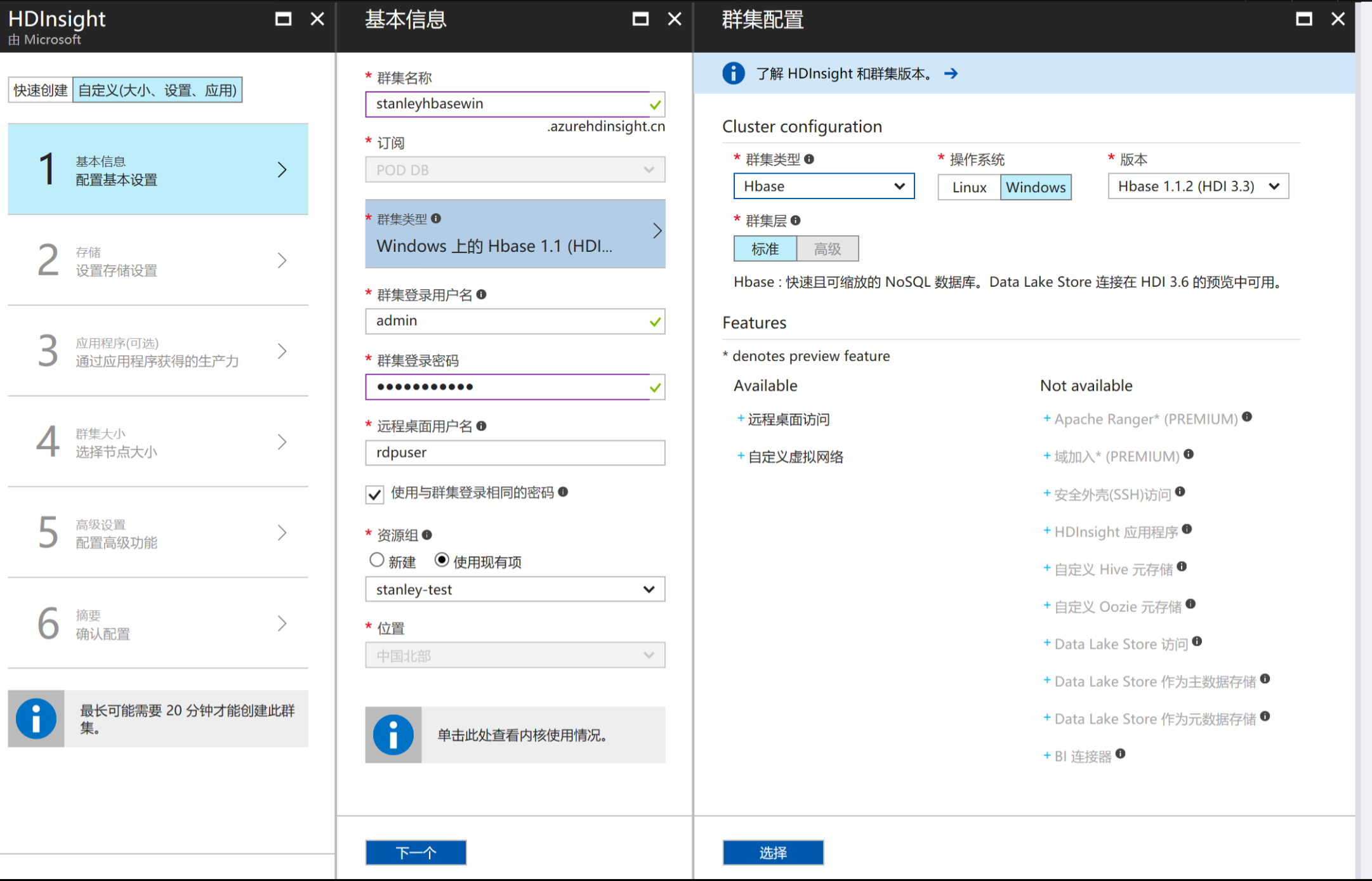
以下为本次HBase迁移示例的具体步骤：

1. 在Ibiza Portal上创建Windows HDInsight的HBase群集；
2. 用HBase Shell创建示例表格；
3. 用Microsoft Azure Storage Tool复制HBase群集所用的容器；
4. 基于所复制的容器创建Linux HDInsight的HBase群集。

下面将就这几步进行具体介绍，其中将Windows HDInsight上的HBase群集简称为Windows HBase群集，将Linux HDInsight上的HBase群集简称为Linux HBase群集。

**在Ibiza Portal上创建Windows HDInsight的HBase群集**

登陆Ibiza Portal后，点击**“新建”**→**HDInsight**，开启HDInsight创建页面，选择自定义创建方式：



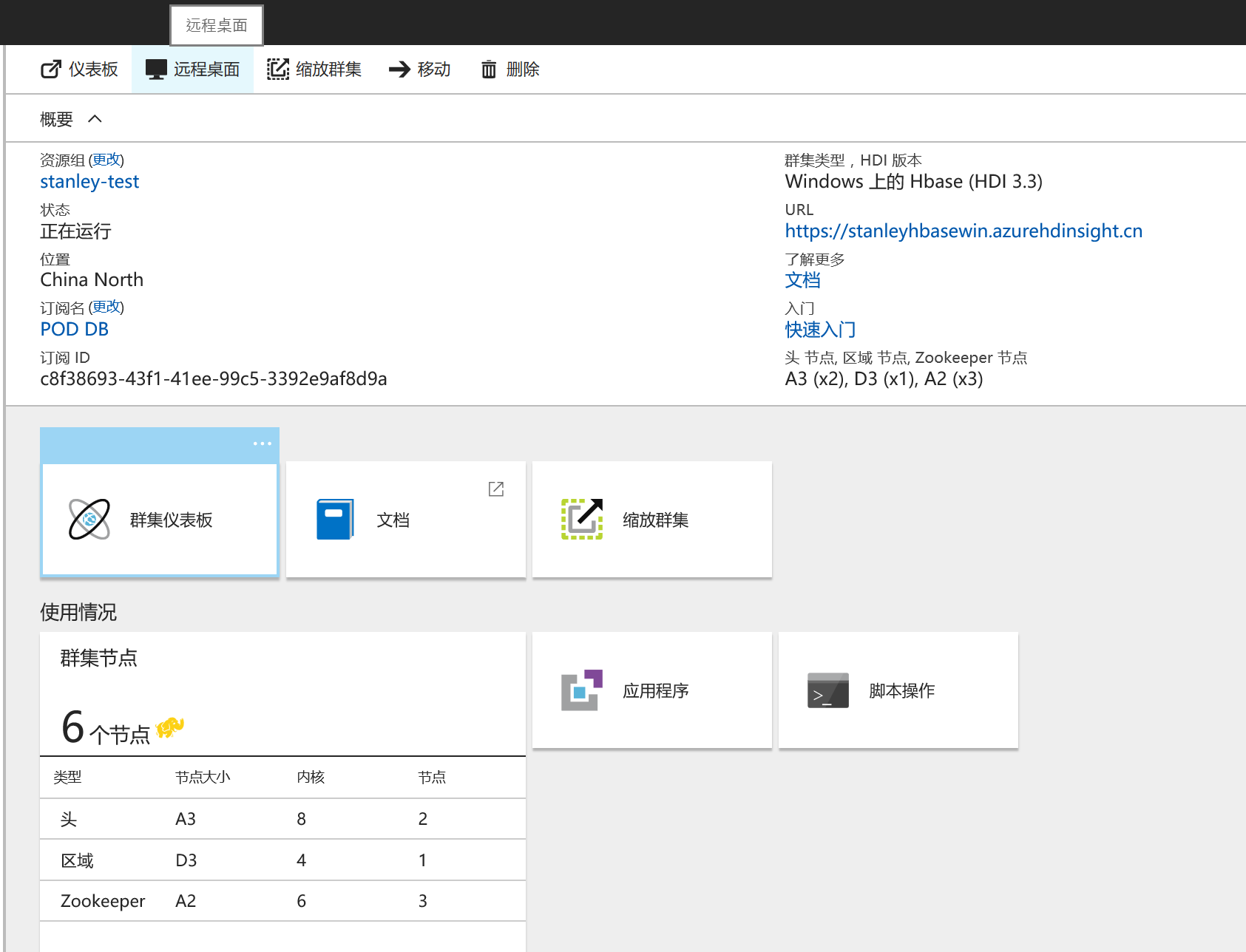
之后可参考官网文章进行参数设置。本示例所用的容器是位于中国北部的存储账户stanleynorth下的stanleyhbasewin(容器名可自行设定)，区域节点为D3，头结点为A3，Zookeeper节点为A2。在创建群集前显示的摘要如图所示：



所有参数设定完后，点击“创建”按钮，便开启了HDInsight群集部署，此过程可能耗时二十多分钟甚至更长时间。



待HDInsight群集创建完毕，便可通过RDP进行远程桌面登陆：



**用HBase Shell创建示例表格**

远程登陆机器之后，双击Hadoop Command Line图标打开命令行界面，输入以下命令打开hbase shell界面：

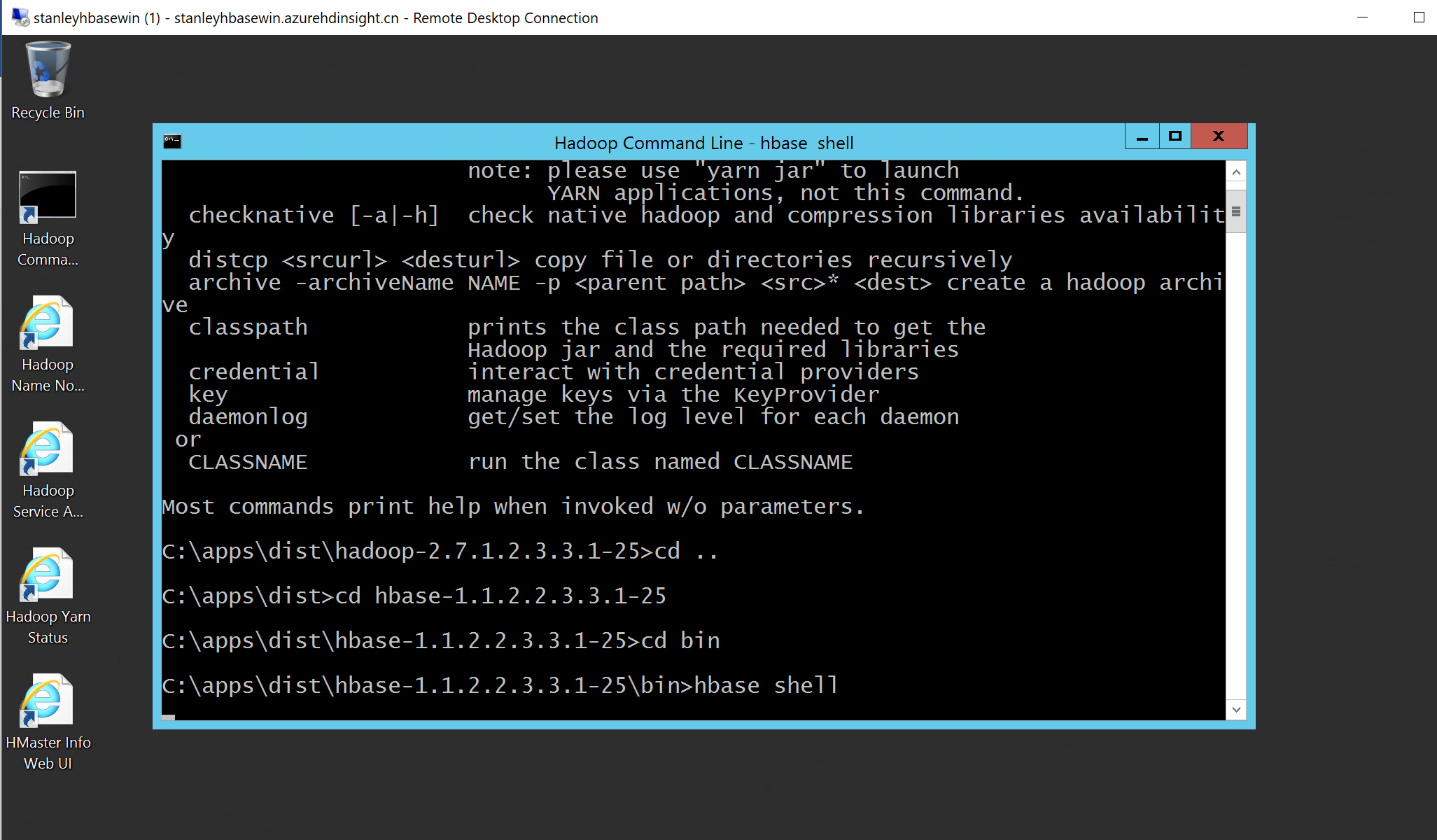
cd ..

cd hbase-1.1.2.2.3.3.1-25

cd bin

hbase shell

注意：第二步跳转到HBase目录，所用的目录名视具体HBase版本而定。



之后创建测试表‘Contacts’，该表有‘Personal’和’Office‘两个Column Family，共包含11行数据。以下为操作命令，具体内容可参考<https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hbase-tutorial-get-started-linux> 中的”Create Table and Insert Data”部分：

create 'Contacts', 'Personal', 'Office'

list

put 'Contacts', '1000', 'Personal:Name', 'John Dole'

put 'Contacts', '1000', 'Personal:Phone', '1-425-000-0001'

put 'Contacts', '1000', 'Office:Phone', '1-425-000-0002'

put 'Contacts', '1000', 'Office:Address', '1111 San Gabriel Dr.'

scan 'Contacts'

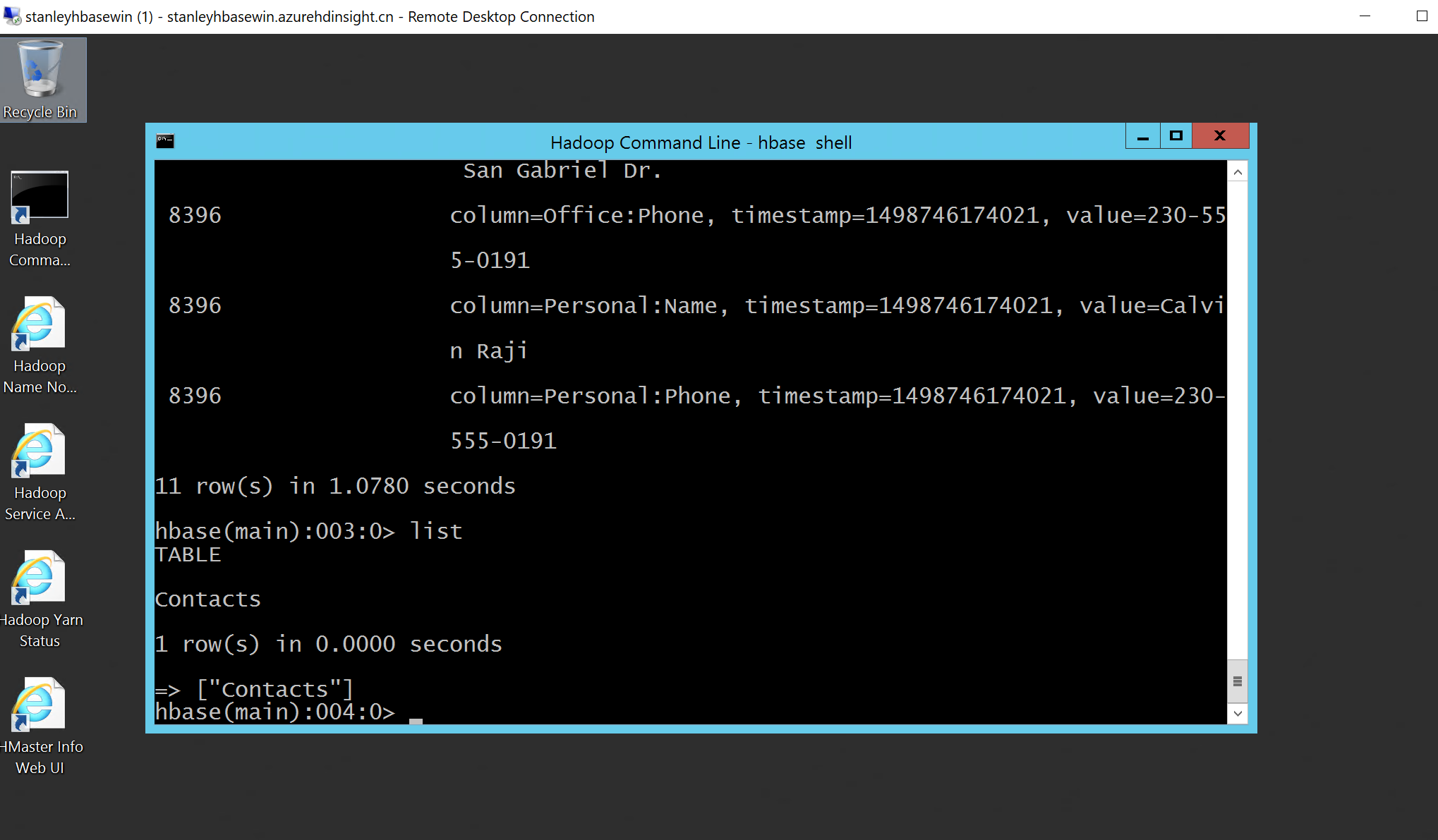
get 'Contacts', '1000'

exit

hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.columns="HBASE\_ROW\_KEY,Personal:Name,Personal:Phone,Office:Phone,Office:Address" -Dimporttsv.bulk.output="/example/data/storeDataFileOutput" Contacts wasbs://hbasecontacts@hditutorialdata.blob.core.windows.net/contacts.txt

hbase org.apache.hadoop.hbase.mapreduce.LoadIncrementalHFiles /example/data/storeDataFileOutput Contacts

表格创建完毕，如下所示：



注意：实例表格创建完毕后，需要将表’Contacts’给disable，这样做的目的是为了将HBase内存中的数据完整地写到存储中，确保数据的完整性。

disable ’Contacts’

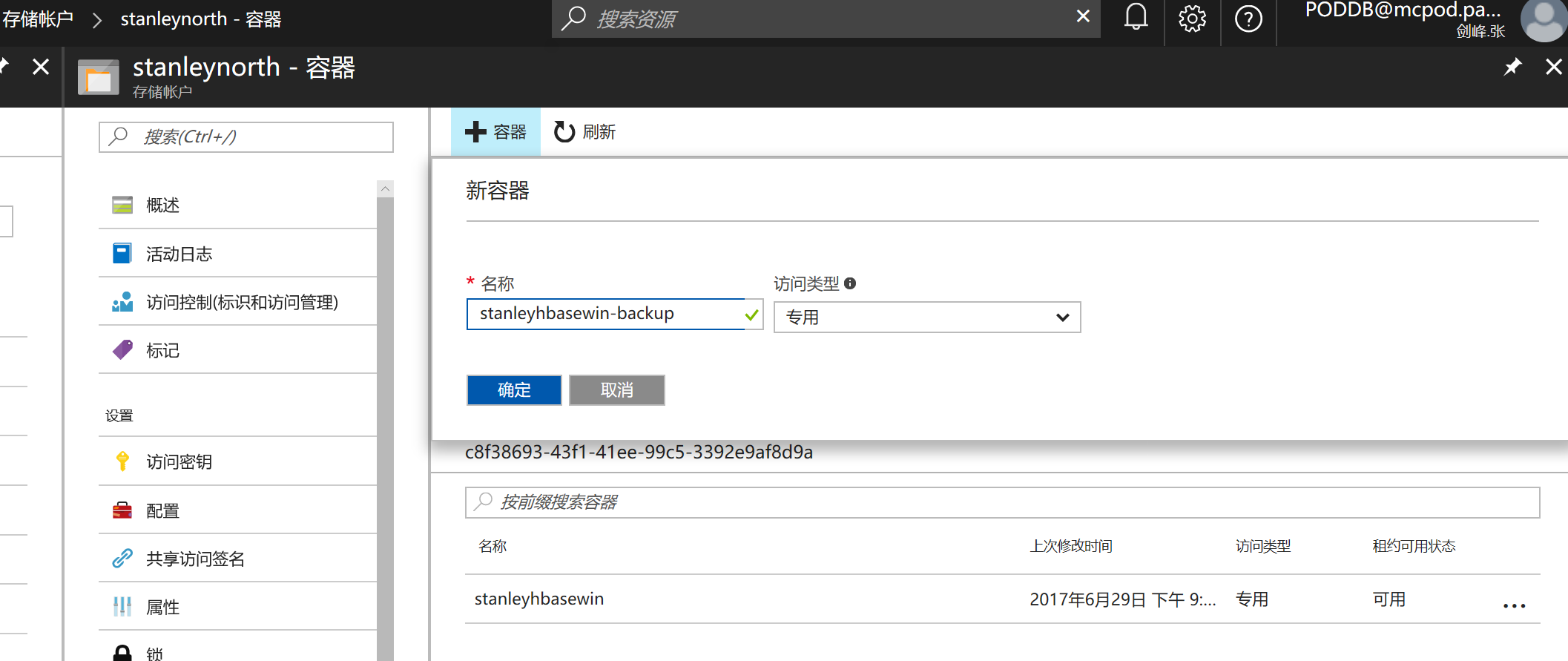
**用Microsoft Azure Storage Tool复制HBase群集所用的容器**

上述操作完成后，便可将Windows HBase群集的数据迁移到Linux HBase群集中。官网文章介绍的方法是在部署好Linux HBase群集后，用SSH工具登陆并输入以下命令进行数据迁移：

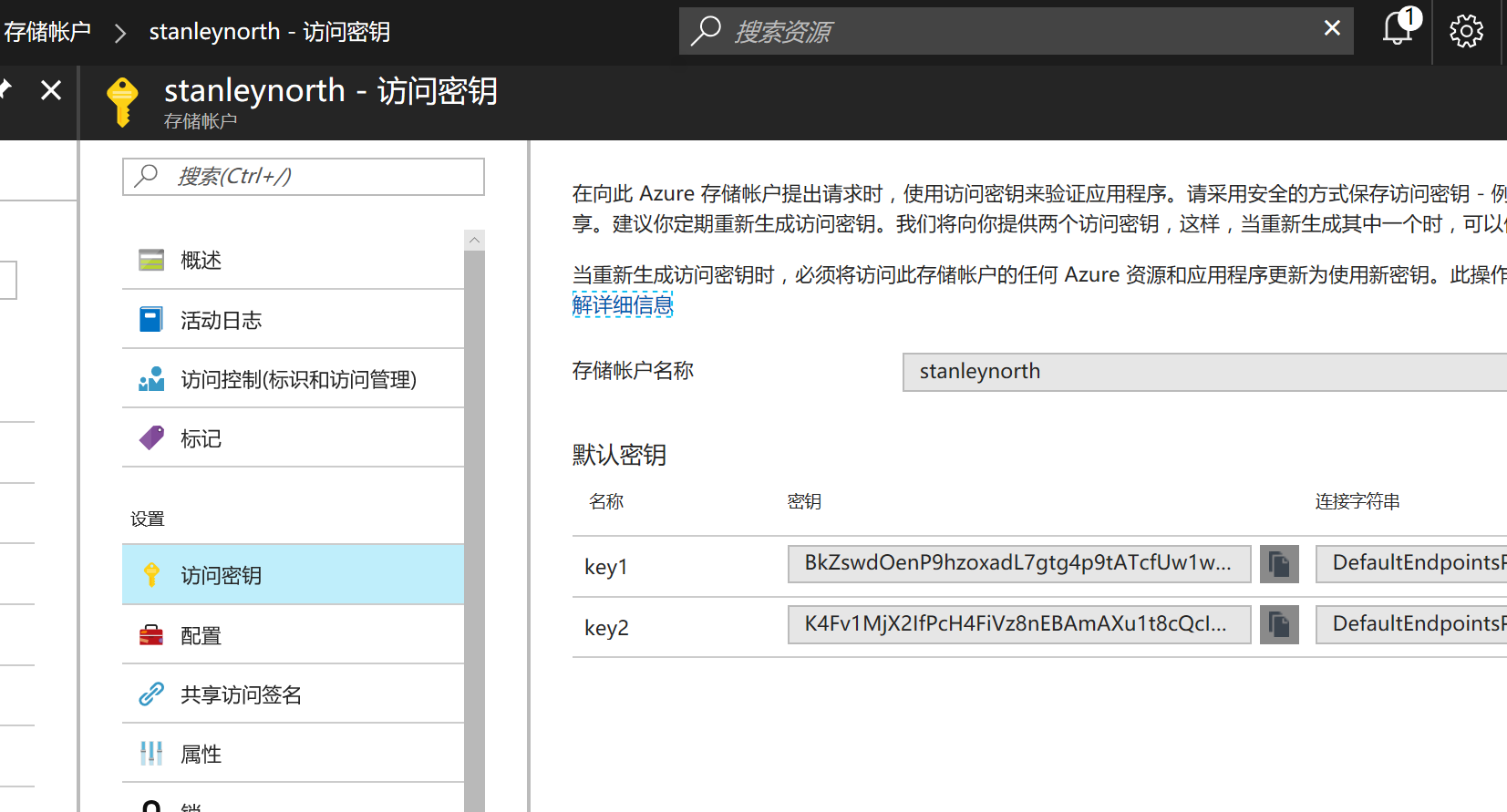
hdfs dfs -cp wasbs://CONTAINER@ACCOUNT.blob.core.windows.net/path/to/old/data /path/to/new/location

但经本地测试发现，也可通过Microsoft Azure存储工具命令行(参考<https://docs.microsoft.com/zh-cn/azure/storage/storage-use-azcopy>)直接将Windows HBase群集所用容器的内容复制到一个新容器中，并在后者的基础上新建Linux HDInsight的HBase群集。以下是具体步骤：

首先，在Ibiza Portal上创建新容器。为方便起见，本示例选择在Windows HBase群集所用的存储账户stanleynorth下新建名为stanleyhbasewin-backup的容器。当然，也可根据需要而在其他存储账户下创建新容器。



待容器创建完成后，点击“访问密钥”，查看存储账户stanleynorth所用的密钥，key1或key2互为备份，均可使用。

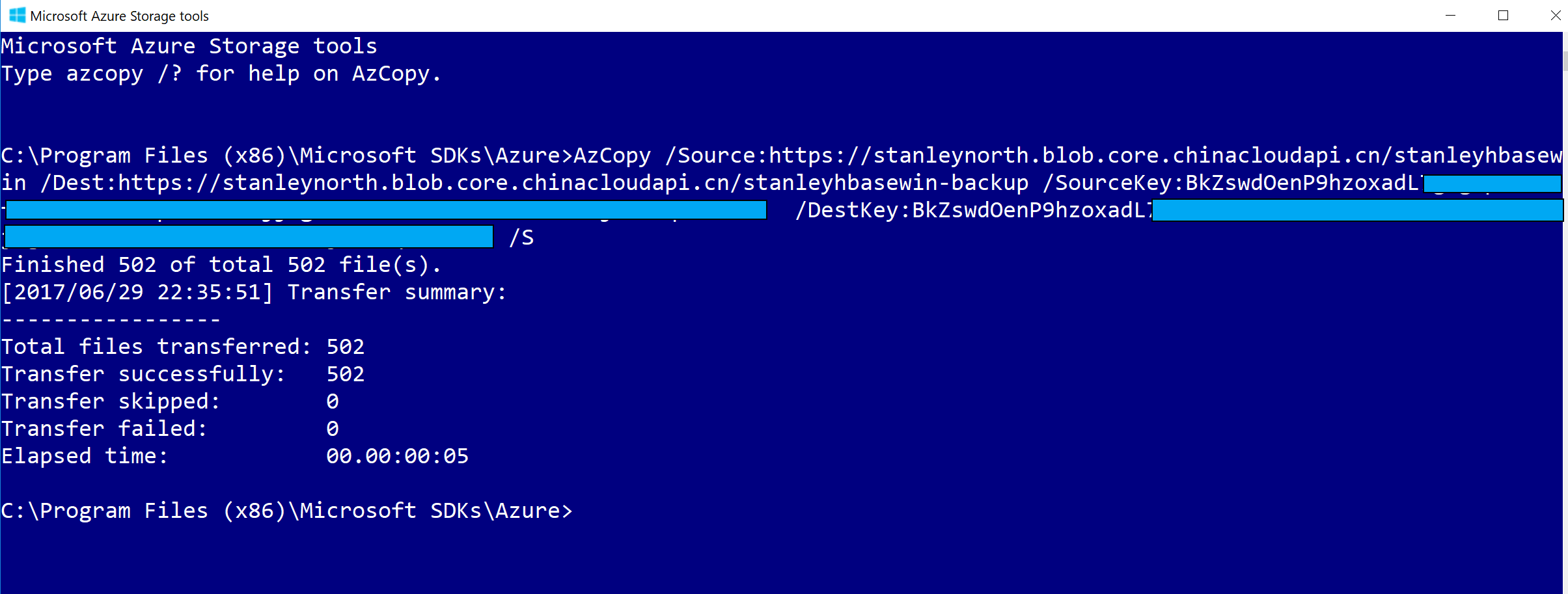


然后打开Microsoft Azure Storage命令行窗口，输入以下命令将Windows HBase群集所用容器stanleyhbasewin中的数据全部复制到新建容器stanleyhbasewin-backup中：

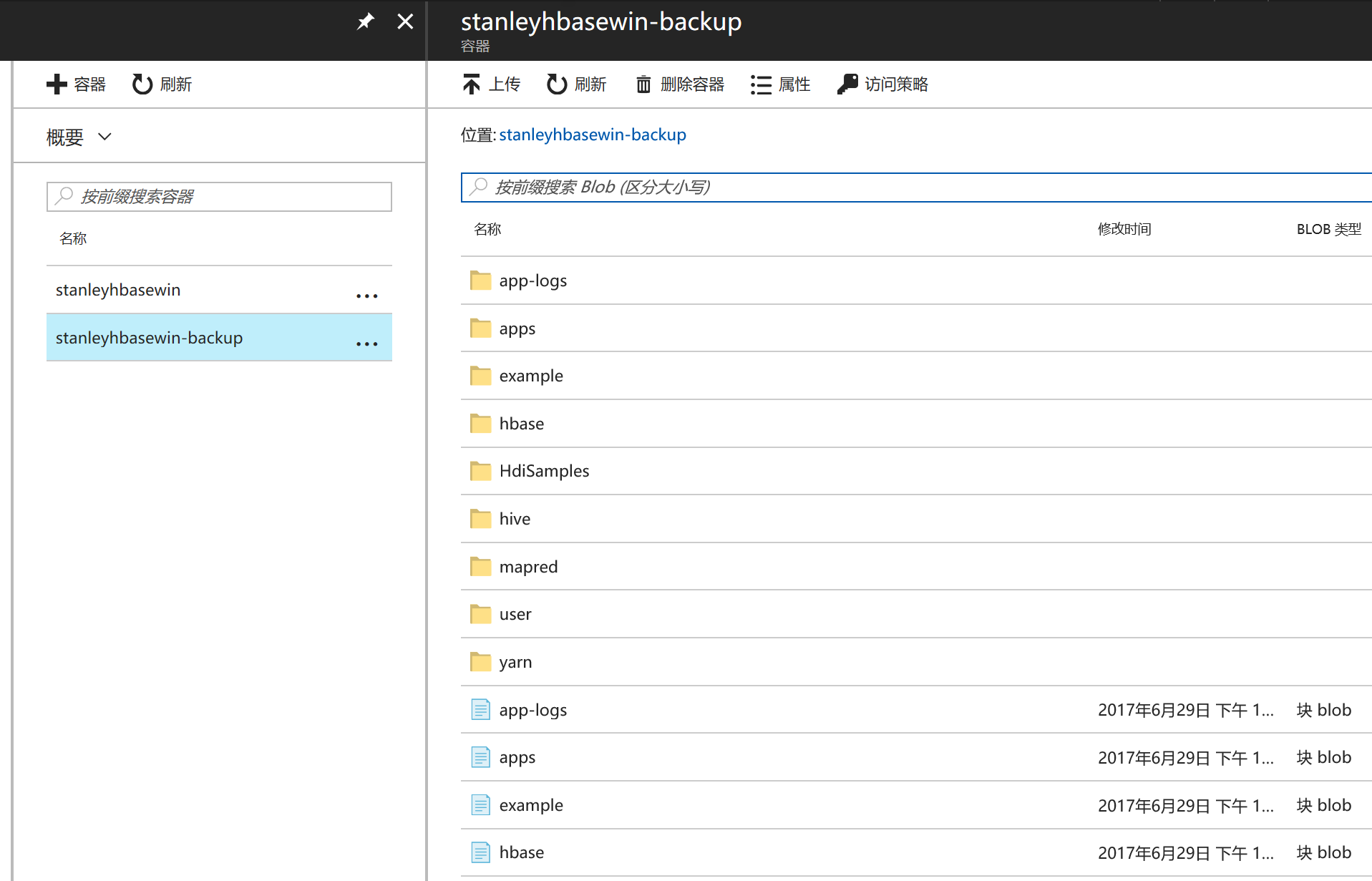
AzCopy /Source:https://stanleynorth.blob.core.chinacloudapi.cn/stanleyhbasewin /Dest:https://stanleynorth.blob.core.chinacloudapi.cn/stanleyhbasewin-backup /SourceKey:<key\_source> /DestKey:<key\_dest> /S

其中<key\_source>表示源存储账户的密钥，<key\_dest>表示目的存储账户的密钥，在本示例中，二者均为存储账户stanleynorth的密钥。

容器复制完成后，会显示复制文件数，所花时间等信息。



从Ibiza Portal上看，容器stanleyhbasewin中的内容已复制到stanleybasewin-backup中。

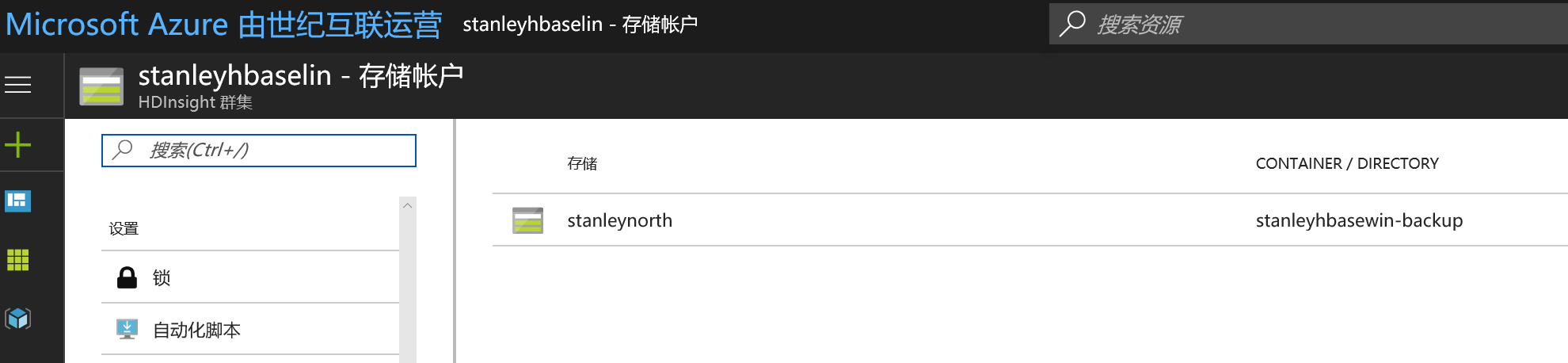


**基于所复制的容器创建Linux HDInsight的HBase群集**

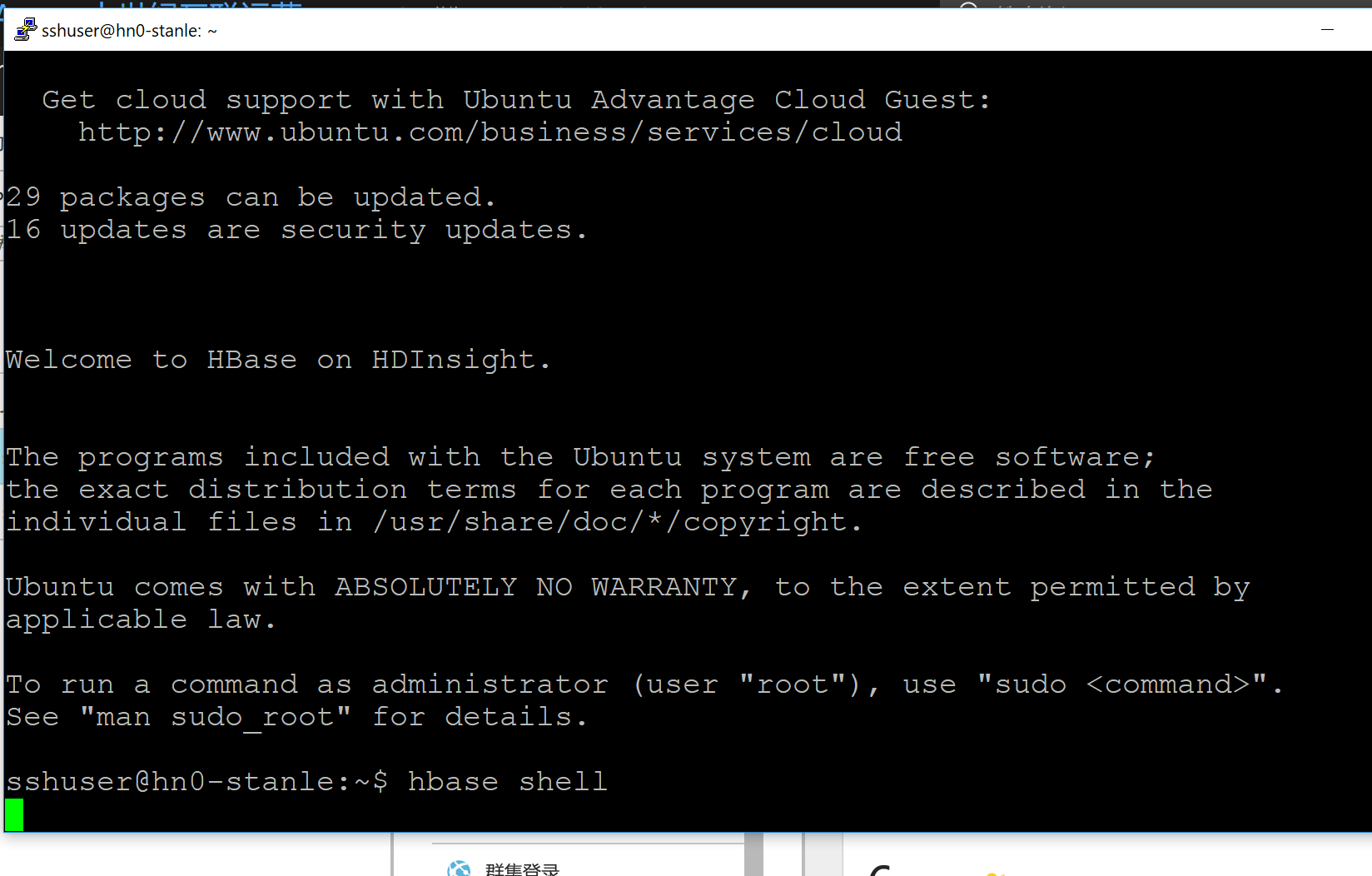
最后创建Linux HBase群集，群集名为stanleyhbaselin，存储账户和默认容器分别选择stanleynorth和stanleyhbasewin-backup，区域节点为D3，头结点为D2 v2，Zookeeper节点为A2。



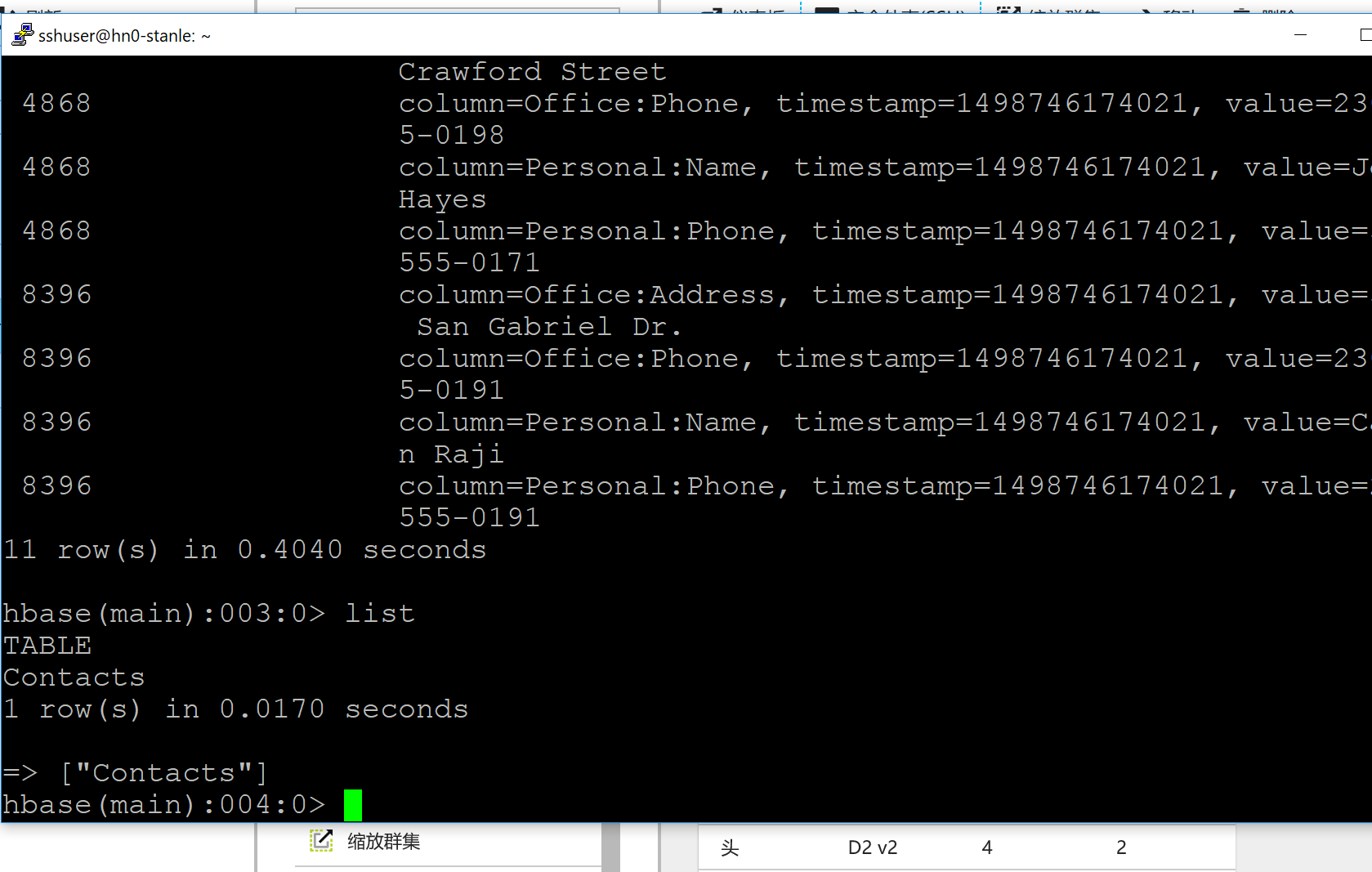
同之前情况类似，此次创建也需耗时二十多分钟以上。待群集创建完毕后，点击存储账户，可确认所使用的容器为stanleyhbasewin-backup。



然后用PuTTY登陆stanleyhbaselin群集，直接输入hbase shell进入hbase环境。



执行list和scan ‘Contacts’，可看到之前在Windows HBase上创建的有11行数据的Contacts表，这表明从Windows HDInsight到Linux HDInsight的数据迁移是成功的。



**结语**

本文介绍了一个从Windows HBase群集向Linux HBase群集迁移的示例，所用的方法是在Microsoft Azure Storage命令行界面中用AzCopy命令将Windows HBase群集所用存储账户容器stanleyhbasewin中的内容复制到一个新的容器stanleyhbasewin-backup中，并基于stanleyhbasewin-backup创建Linux HBase群集，之后登陆Linux HBase群集便可看见之前在Windows HBase群集中创建的表Contacts。

如前文所言，目前Windows HDInsight平台支持Hadoop，HBase和Storm三个群集类型，本文所介绍的HBase群集的迁移只涉及到表格，相对比较简单。之后我会继续研究Hadoop和Storm群集所涉及的数据和作业方面的迁移，并另外撰文进行总结。