Hadoop with Microsoft Azure

https://azure.microsoft.com/ru-ru/

Регистрация

Каждому студенту необходимо получить у преподавателя код для регистрации на Microsoft Azure. Далее необходимо зарегистрироваться на данном сайте.

Следует помнить, что коды для регистрации выдаются факультету на коммерческой основе, поэтому следует пройти процедуру регистрации внимательно, а также помнить, что за использование hadoop-кластеров от Microsoft (Продукт HDInsight) взимается поминутная оплата. Каждому студенту выдаются на месяц виртуальные 100$, поэтому стоит выполнять задания внимательно и не оставлять зависшие задачи и работающие кластеры, так как деньги не снимаются только тогда, когда все кластеры УДАЛЕНЫ. Деньги начинают сниматься сразу же после создания кластера.

Более подробную информацию о том, за что снимаются деньги, Вы можете посмотреть на данном сайте:

https://azure.microsoft.com/ru-ru/pricing/details/hdinsight/

На каждой лекции будет рассмотрено по одной задаче, а также по одной задаче каждую лекцию будет выдаваться в качестве домашнего задания.

Инструкцию по регистрации на сайте Microsoft Azure Вы можете найти на данном сайте: https://www.microsoftazurepass.com/howto

Для регистрации необходимо иметь учетную запись Microsoft.

После регистрации Вы будете направлены непосредственно на сайт рабочей среды https://portal.azure.com/

Обратите внимание, что в данный момент Microsoft Azure переходит на новый портал (адрес старого портала: https://manage.windowsazure.com). Работать мы будем исключительно на новом портале: https://portal.azure.com/ (если не оговорено обратное).

Главное: не забудьте, что после работы кластеры необходимо УДАЛЯТЬ!!!

Компиляция и формирование .jar-файла

Технология HDInsight подразумевает, что Ваша программа компилируется на Вашем компьютере, Вами формируется .jar-файл, и далее данный файл загружается на Ваш удаленный кластер для выполнения.

Для того чтобы получить .jar-файл, можно использовать различные технологии, но крайне желательно использовать для этого консоль во избежание различных дополнительных проблем.

О том, как это можно сделать, Вы можете прочитать по данной ссылке (пока останавливаемся на получении .jar-файла и дальше по этой ссылке не читаем):

https://azure.microsoft.com/da-dk/documentation/articles/hdinsight-develop-deploy-java-mapreduce-linux/

Если у Вас не получилось скомпилировать программу и получить .jar-файл в консоли, Вы можете воспользоваться средой разработки **Intellij IDEA**. Для этого необходимо:

1. Загрузить и установить Intellij IDEA – https://www.jetbrains.com/idea/
2. Загрузить и установить JAVA SE Development Kit 8 Downloads – http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html (для Win64 могут возникнуть проблемы при установке 64-битной версии, тогда следует установить 32-битную).
3. Задать значение переменной среды JAVA_HOME
4. Скачать собранный Hadoop: http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.1/hadoop-2.7.1.tar.gz
   Обратите внимание, что данные исходники с официального сайта – это исходники под Linux, для того, чтобы ими пользоваться под Windows, их необходимо перекомпилировать (или взять более старую версию отсюда: http://static.barik.net/software/hadoop-2.6.0-dist/hadoop-2.6.0.tar.gz)
5. Задать значение переменной среды HADOOP_HOME
6. При открытии Intellij IDEA выбираем SDK – установленный перед этим JDK.
7. Добавить в проект библиотеки в Intellij IDEA: https://mrchief2015.wordpress.com/2015/02/09/compiling-and-debugging-hadoop-applications-with-intellij-idea-for-windows/
8. Компилируем программу (Например, простейший Word Count с MapReduce, код данной программы легко можно найти в Интернете)
9. Собираем JAR-файл: http://blog.jetbrains.com/idea/2010/08/quickly-create-jar-artifact/

Создание кластера и отправка заданий на счет с HDInsight

Собранный .jar-файл необходимо перенести на созданный кластер и запускать его уже на нем. Поэтому Вам необходимо создать кластер и настроить ssh-соединение. Рассмотрим данный процесс подробнее.

Поскольку кластеры на Windows значительно дороже, мы будем использовать кластеры на Linux.

Будем использовать удаленный доступ. Для ОС Linux подробную инструкцию о настройке ssh можно найти по ссылке:

https://azure.microsoft.com/da-dk/documentation/articles/hdinsight-hadoop-linux-use-ssh-unix/

Для ОС Windows:

https://azure.microsoft.com/da-dk/documentation/articles/hdinsight-hadoop-linux-use-ssh-windows/

Основные шаги, которые Вам необходимо при этом проделать:

1) Необходимо через puttygen сгенерировать ключи для удаленного подключения. Не забываем сохранить публичный и приватный ключи (публичный – в формате .txt, приватный – в формате .ppk)
2) Далее необходимо создать HDInsight–кластер, это можно сделать прямо на портале Microsoft Azure. Предварительно Вам необходимо будет создать хранилище данных. О том, как это сделать, Вы сможете прочитать ниже.
3) Далее Вам необходимо будет перекинуть Ваш .jar-файл на созданный Вами кластер.
4) После этого необходимо через ssh-соединение запустить Вашу программу на Вашем кластере и получить результат.

Теперь опишем данные шаги подробнее.

Шаг 1. Генерация ключей для удаленного доступа.

Программу puttygen Вы сможете скачать по этой ссылке:

http://the.earth.li/~sgtatham/putty/latest/x86/puttygen.exe

О том, как сгенерировать ключи, подробно описано в ссылках, данных выше:

Linux: https://azure.microsoft.com/da-dk/documentation/articles/hdinsight-hadoop-linux-use-ssh-unix/

Windows: https://azure.microsoft.com/da-dk/documentation/articles/hdinsight-hadoop-linux-use-ssh-windows/

Шаг 2. Создание хранилища данных и HDInsight-кластера.

На данных шагах создания хранилища и HDInsight-кластера преподаватели настойчиво советуют делать скриншоты всех своих шагов, чтобы не забыть названия созданных главных и промежуточных объектов, а также различные настройки.

Вначале создадим хранилище данных. Этот шаг является обязательным, так как без него невозможно создать HDInsight-кластер.

Выбираем на портале: Создать -> Данные + хранилища -> Хранилище.

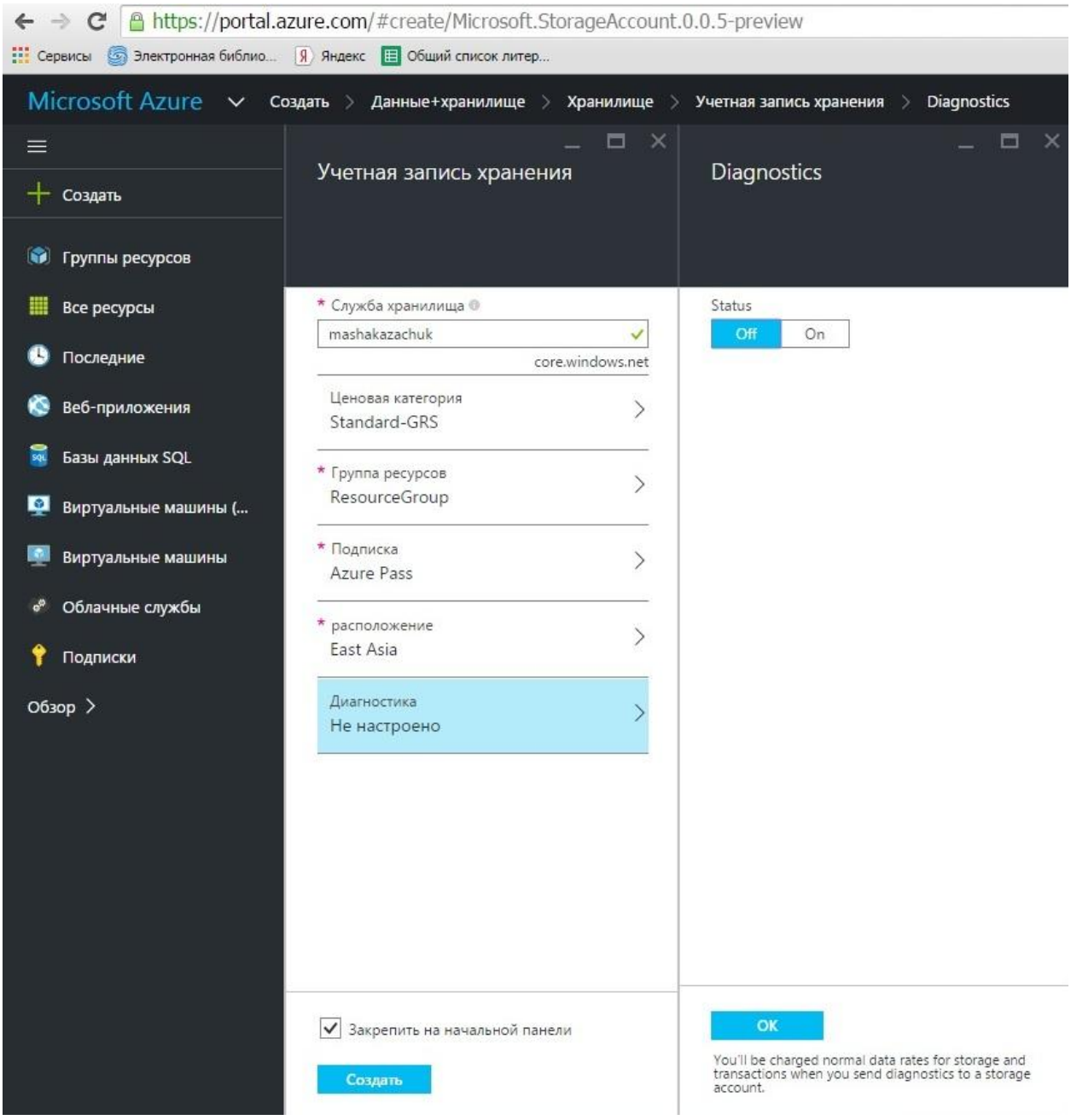
Называем службу хранилища незанятым именем.

Поскольку объем наших средств ограничен, можно ограничиться ценовой категорией Standard-GRS или Standard-LRS (всегда обращаем внимание на стоимость ресурсов!).

Создаем группу ресурсов.

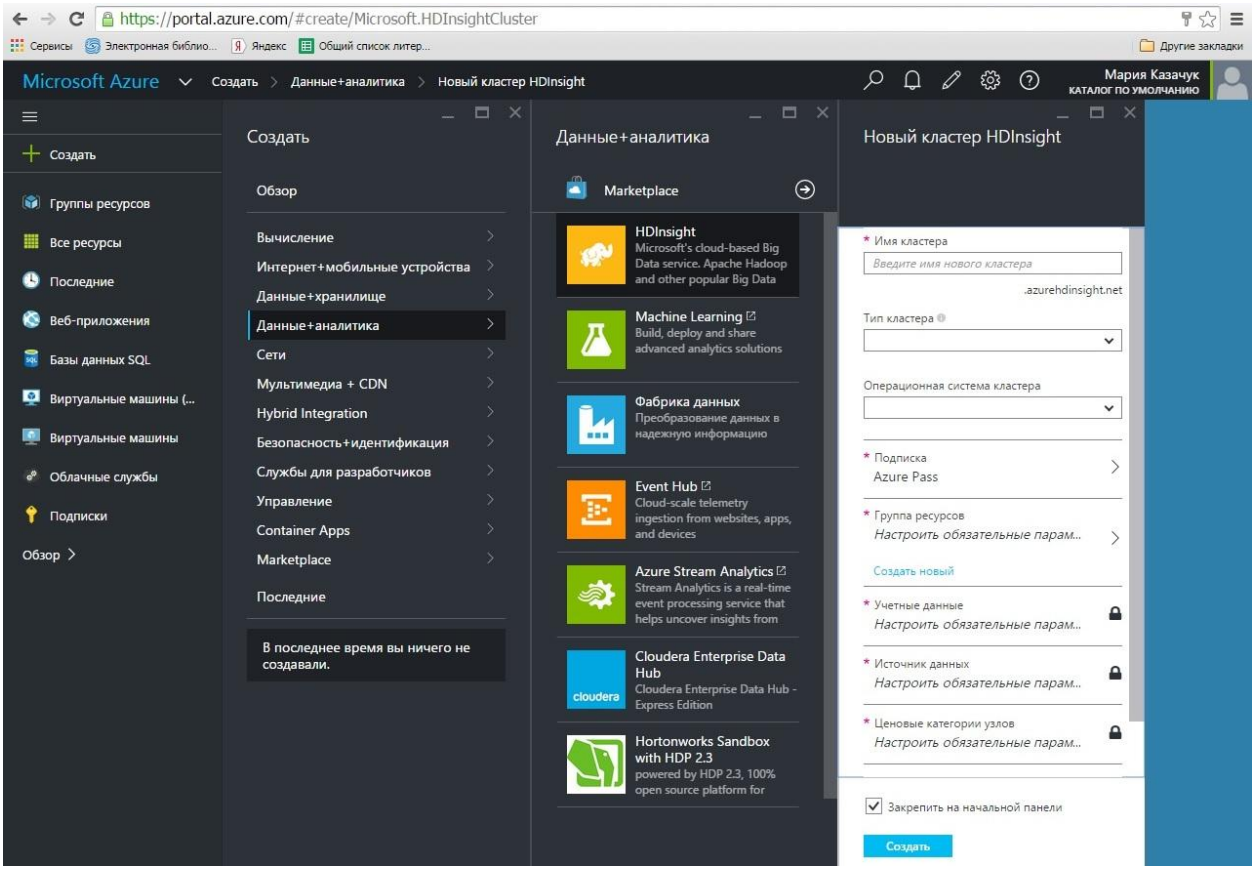
Проверяем, что в качестве подписки стоит Ваша подписка Azure Pass.

Далее необходимо выбрать расположение хранилища. Стоит его выбирать рядом с нами (так как чем дальше, тем дороже хранилище), поэтому можно оставить Восточную Азию, стоимость будет минимальной.

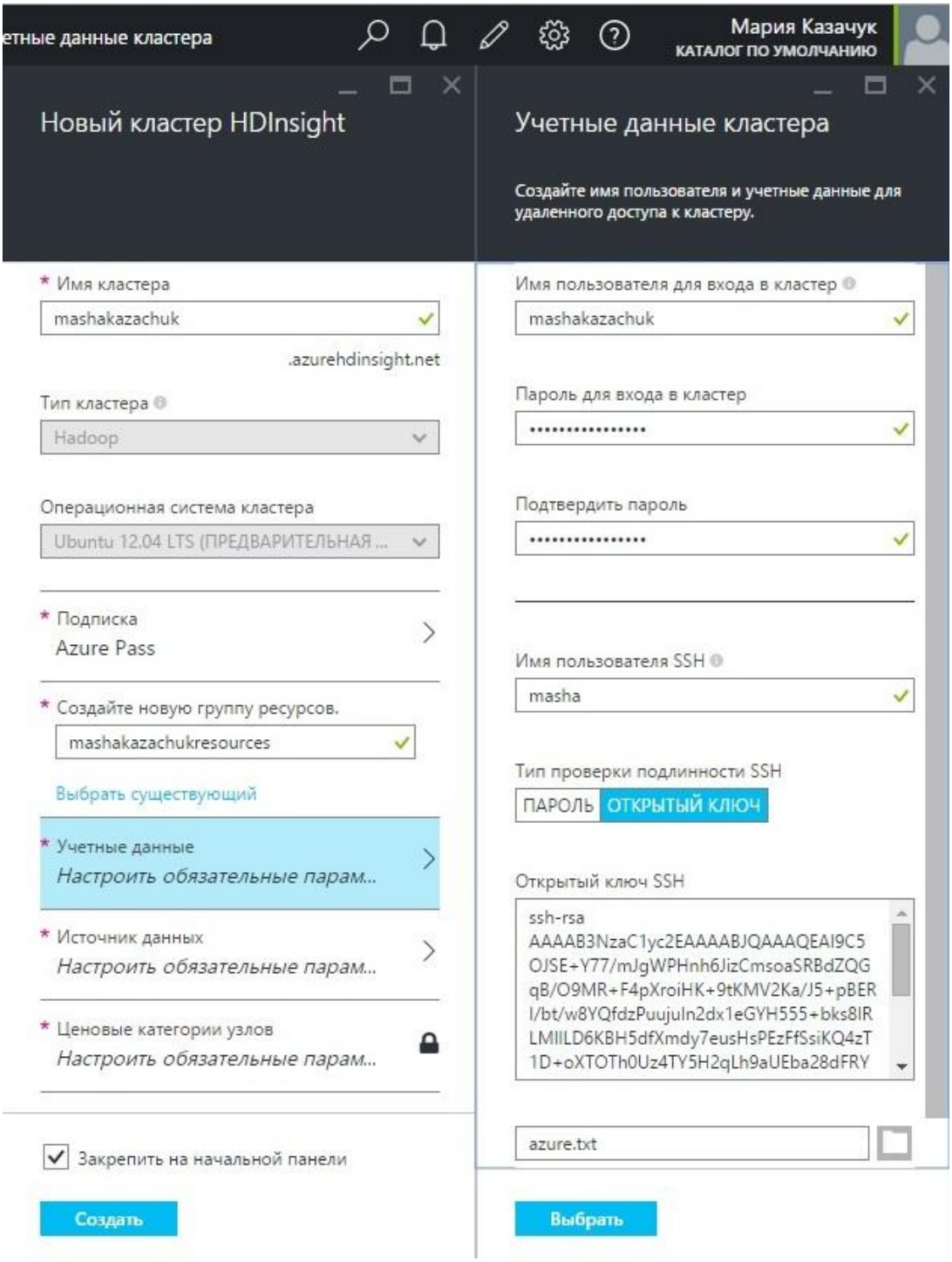


Теперь перейдем непосредственно к созданию HDInsight-кластера. Вам необходимо будет выполнить следующую последовательность шагов:

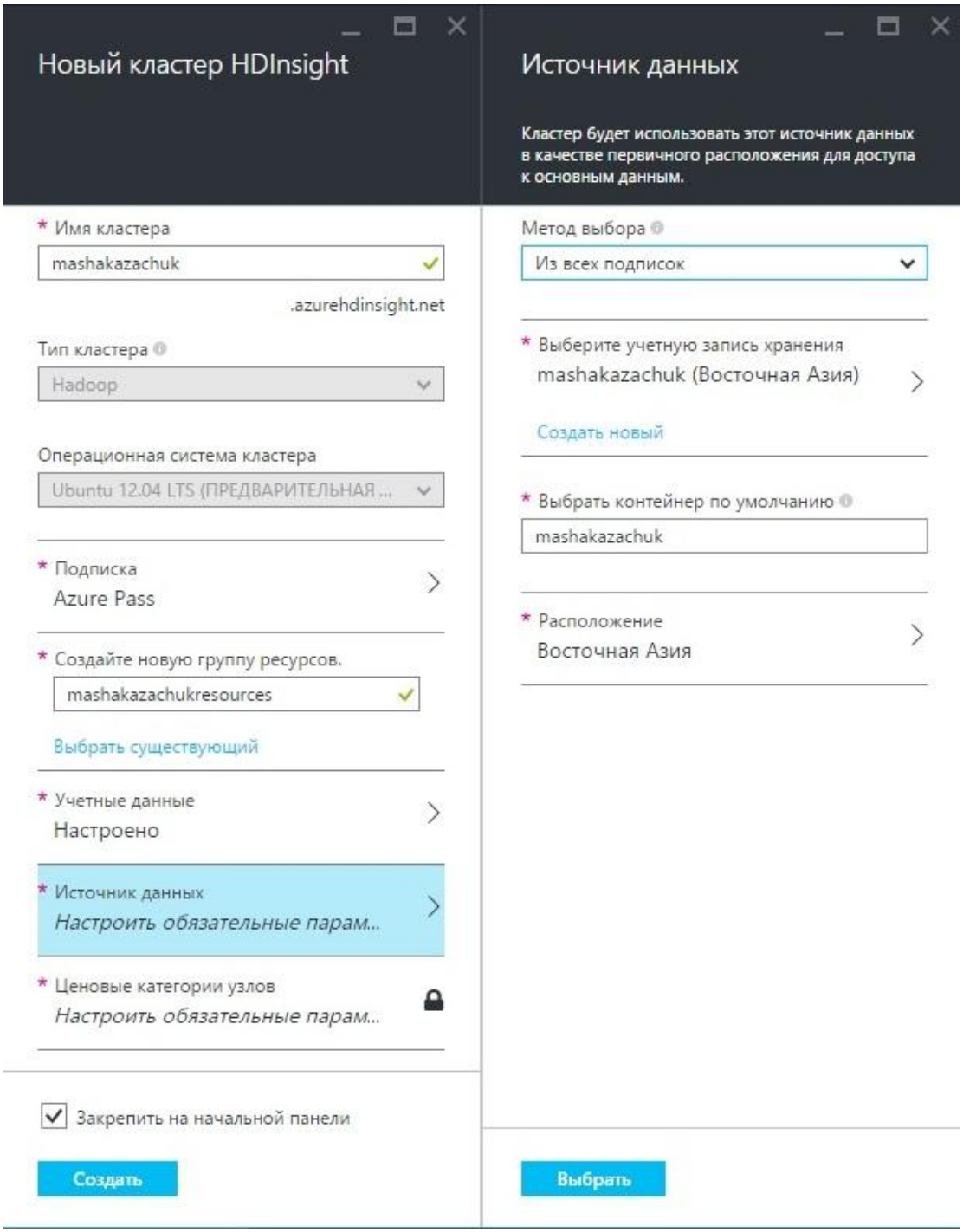
Выбираем: Создать -> Данные + аналитика -> HDInsight.



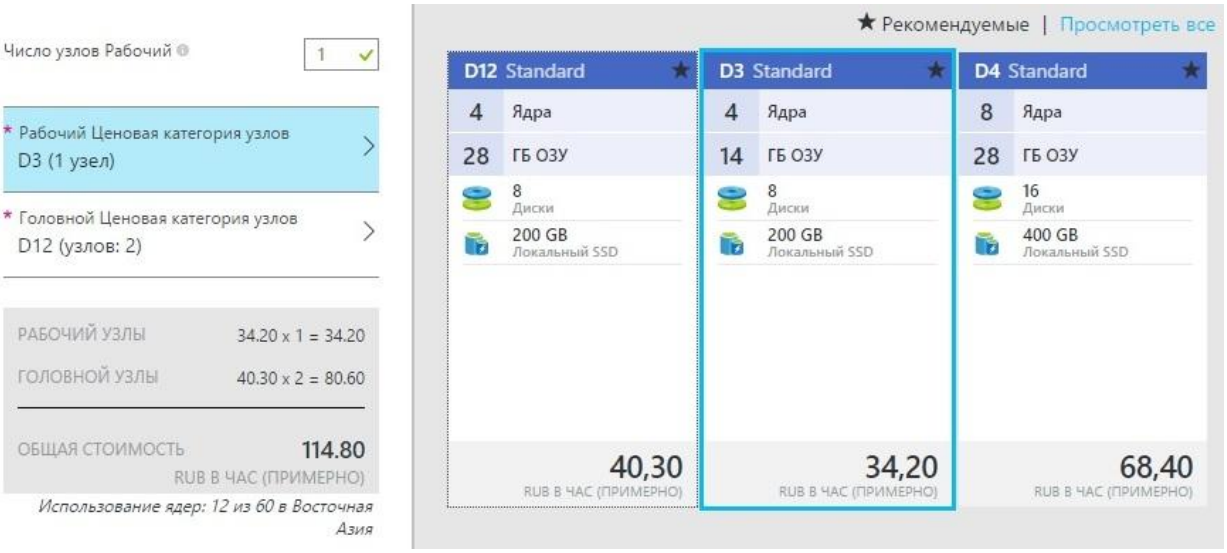
Заполняем поля. Выбираем тип кластера Hadoop, ОС Ubuntu (Windows использовать значительно дороже, мы не уложимся в лимит средств). В графе «Учетные данные» выбираем тип проверки подлинности – открытый ключ и загружаем его.

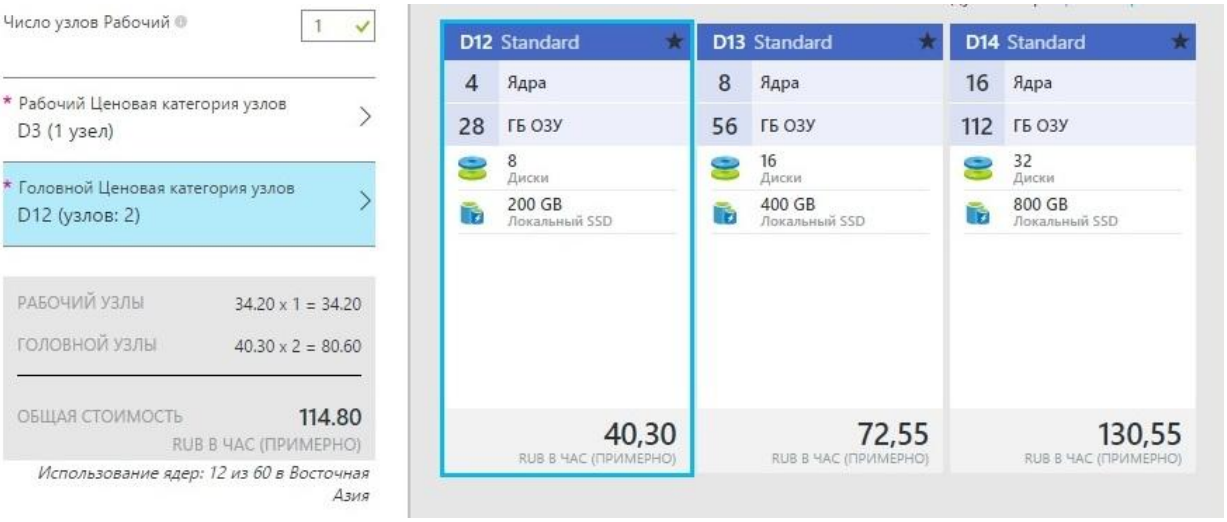

етные данные кластера
Мария Казачук
КАТАЛОГ ПО УМОЛЧАНИЮ
Новый кластер HDInsight
* Имя кластера
mashakazachuk
.azurehdinsight.net
Тип кластера
Hadoop
Операционная система кластера
Ubuntu 12.04 LTS (ПРЕДВАРИТЕЛЬНАЯ ...
* Подписка
Azure Pass
* Создайте новую группу ресурсов.
mashakazachukresources
Выбрать существующий
* Учетные данные
Настроить обязательные парам...
* Источник данных
Настроить обязательные парам...
* Ценовые категории узлов
Настроить обязательные парам...
Закрепить на начальной панели
Создать
Учетные данные кластера
Создайте имя пользователя и учетные данные для удаленного доступа к кластеру.
Имя пользователя для входа в кластер
mashakazachuk
Пароль для входа в кластер
Подтвердить пароль
Имя пользователя SSH
masha
Тип проверки подлинности SSH
ПАРОЛЬ
ОТКРЫТЫЙ КЛЮЧ
Открытый ключ SSH
ssh-rsa
AAAAB3NzaC1yc2EAAAABJQAAAQEAI9C5
OJSE+Y77/mJgWPHnh6JizCmsoaSRBdZQG
qB/O9MR+F4pXroiHK+9tKMV2Ka/J5+pBER
l/bt/w8YQfdzPuujuln2dx1eGYH555+bks8IR
LMIILD6KBH5dfXmdy7eusHsPEzFfSsiKQ4zT
1D+oXTOTh0Uz4TY5H2qLh9aUEba28dFRY
azure.txt
Выбрать

Далее выбираем источник данных – созданное ранее нами хранилище.



Далее необходимо ввести ограничения на ценовые категории узлов. Для того чтобы уложиться по имеющимся средствам, выбираем: один рабочий узел, два головных узла (подробнее – на скриншотах). Обратите внимание, что нам подойдут и более дешевые узлы, чем D3 Standard (например, A3 Standard). Для того чтобы увидеть все возможные узлы, нажимаем на «Просмотреть все»:





Число узлов Рабочий 1

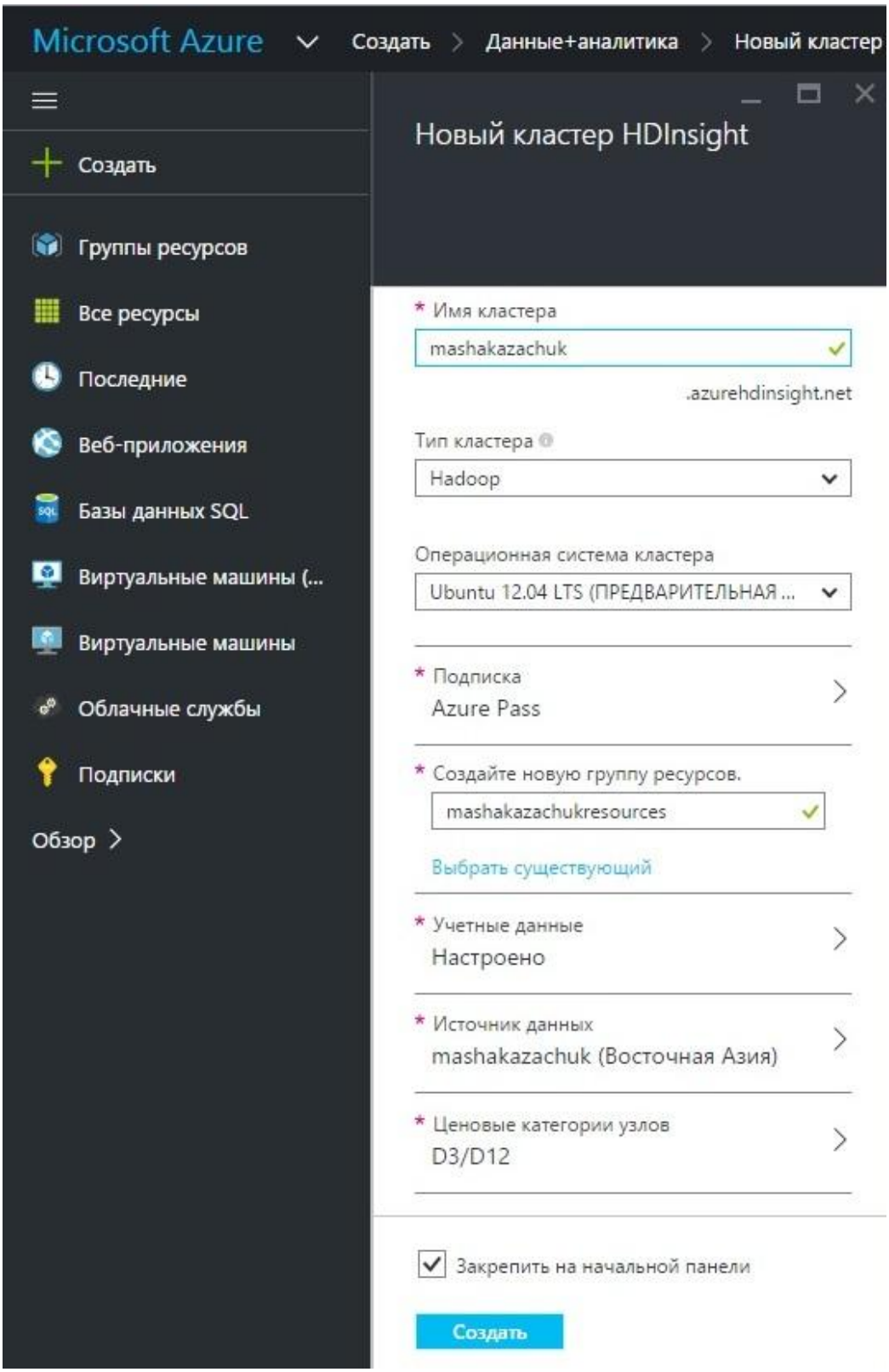
* Рабочий Ценовая категория узлов
D3 (1 узел)

* Головной Ценовая категория узлов
D12 (узлов: 2)

РАБОЧИЙ УЗЛЫ	34.20 x 1 = 34.20
ГОЛОВНОЙ УЗЛЫ	40.30 x 2 = 80.60
ОБЩАЯ СТОИМОСТЬ	**114.80** RUB В ЧАС (ПРИМЕРНО)

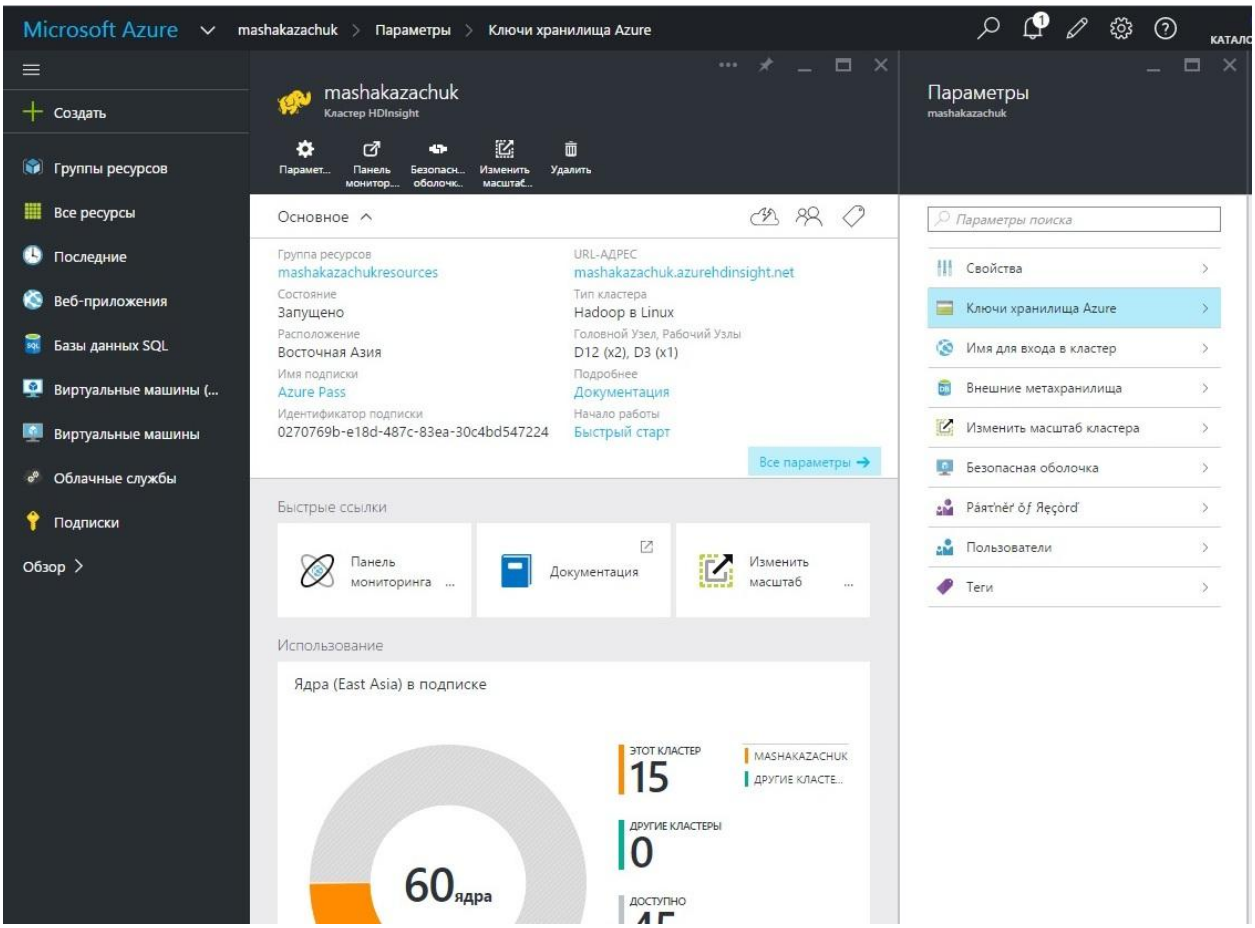
Использование ядер: 12 из 60 в Восточная Азия

Далее нажимаем на «Создать»:



После этого ждем некоторое время (порядка 20 минут), и мы получаем кластер, который можно использовать. В дальнейшем необходимо быть внимательными и не путать имя кластера с именем пользователя ssh! Также рекомендуем Вам запомнить особенности паролей при создании кластера: они должны содержать буквы разных регистров, цифру и

специальный символ. Запомните свои логины и пароли, несмотря на то, что после работы кластер необходимо будет удалить, Вы сможете их использовать в дальнейшем.



Шаг 3. Перенос .jar-файла на кластер.

Данный этап подробно описан по ссылке (пункт «Upload the jar»):

https://azure.microsoft.com/da-dk/documentation/articles/hdinsight-develop-deploy-java-mapreduce-linux/

Команда scp работает в ОС Linux. Для того чтобы перенести .jar в файл на кластер в случае, если вы используете Windows, необходимо воспользоваться программой WinSCP. Скачать ее можно по адресу:

http://winscp.net/download/winscp575setup.exe

Необходимо помнить про то, что следует приложить приватный ключ. В программе WinSCP это делается при помощи Pageant:

http://the.earth.li/~sgtatham/putty/latest/x86/pageant.exe

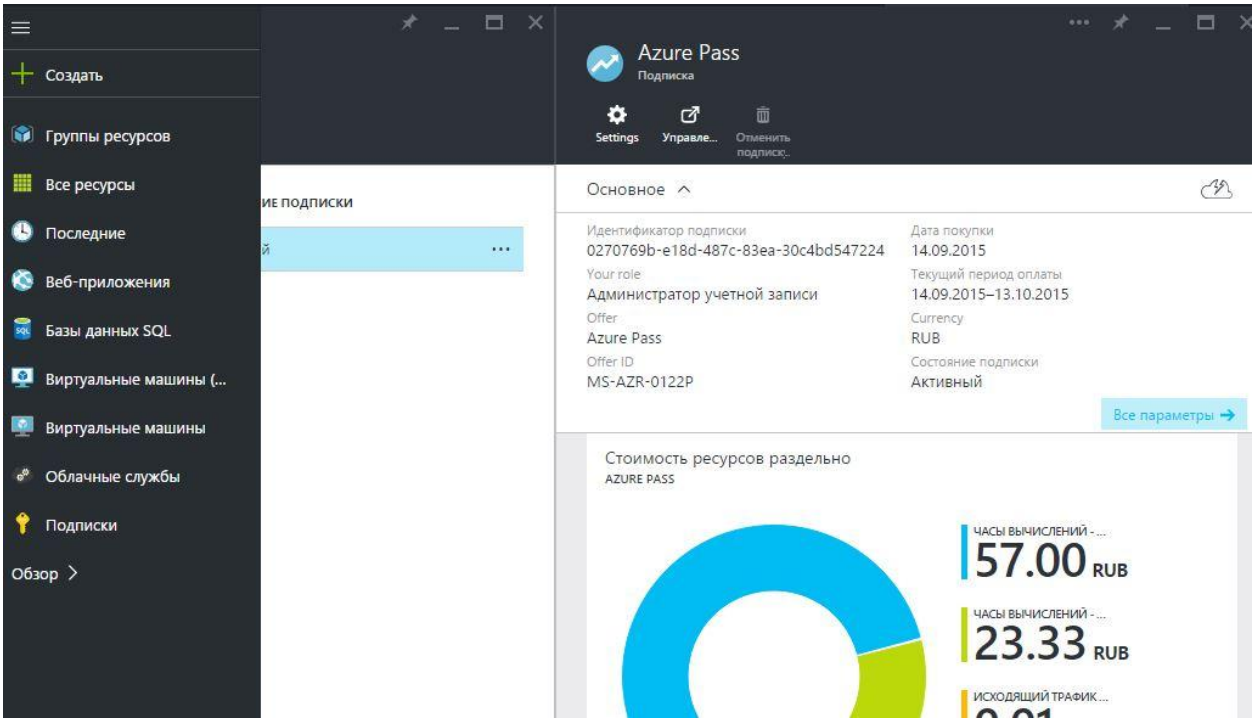
Теперь Вы получили файл wordcountjava-1.0-SNAPSHOT.jar на Вашем удаленном кластере.

Шаг 4. Запуск программы на Вашем удаленном кластере.

Данный шаг подробно описан в пункте «Run the MapReduce job» по вышеприведенной ссылке. Перед выполнением данного шага следует вручную переименовать ваш .jar-файл на кластере в «wordcountjava.jar».

После успешного вывода результата, не забудьте, что созданный кластер и хранилище Вам необходимо удалить, иначе с Вашего счета будут продолжать сниматься деньги.

Обратите внимание, что в связи с тем, что сейчас происходит обновление портала Microsoft Azure, деньги за использование кластеров снимаются не моментально, а через несколько дней. Остаток средств Вы можете проверить, зайдя в раздел «Подписки» и выбрав свою подписку:



Более детальный отчет по средствам Вы можете просмотреть на старой версии портала:

https://manage.windowsazure.com (ждем, пока портал догрузится и нажимаем на «Состояние кредита» в верхушке экрана).

Создание кластера при помощи командной строки и PowerShell

Поскольку после каждого сеанса работы кластеры необходимо удалять, удобным будет создавать кластеры не вручную на портале, а с помощью скриптов. Рассмотрим данный подход с использованием Azure PowerShell и Azure Cli.

Использование Azure PowerShell

Данный метод создания кластеров подойдет пользователям Windows. Подробно о том, как пользоваться данным продуктом, Вы можете прочитать по ссылкам:

https://azure.microsoft.com/en-us/documentation/articles/powershell-install-configure/

https://azure.microsoft.com/zh-cn/documentation/articles/hdinsight-administer-use-powershell/

Установщик Azure PowerShell Вы можете скачать по адресу:

http://go.microsoft.com/fwlink/p/?linkid=320376&clcid=0x409

После того, как Вы установили Azure PowerShell, Вам необходимо будет запустить его и перейти в режим работы AzureResourceManager (существуют два режима работы: AzureResourceManager и AzureServiceManager. В скором времени поддержка второго режима будет отключена, поэтому мы будем пользоваться исключительно первым режимом. По умолчанию же Azure PowerShell запускается в режиме AzureServiceManager).

После этого Вам необходимо будет скачать скрипт для создания необходимых ресурсов: http://bigdata.cs.msu.ru/images/b/ba/Provision.zip.

Открываем данный скрипт с помощью текстового редактора и заменяем “gerasimov” на свою фамилию (либо другое незанятое название).

Запускаем данный скрипт в Azure PowerShell (для этого необходимо просто прописать абсолютный путь до данного скрипта).

Далее у Вас появится окно для ввода логина и пароля от Microsoft Azure. Вводим их. Далее, спустя некоторое время, Вам последовательно выведутся другие окошки для ввода логина и пароля. В первом окошке запрашиваются логин и пароль от кластера, во втором – для удаленного доступа (ssh). Обратите внимание, что правила создания пароля сохраняются (иначе говоря, Вы можете ввести такие же логины и пароли, как при создании кластеров на портале). Также здесь мы пользуемся доступом к ssh по паролю (при создании кластеров на портале Вы также можете использовать пароль вместо механизма публичных и приватных ключей).

Если все Ваши логины и пароли удовлетворяют необходимым требованиям (которые, как уже оговаривалось, полностью совпадают с требованиями к логинам и паролям при создании кластеров на портале), через некоторое время у Вас создадутся необходимые ресурсы (кластер и хранилище). Они также будут отображены на портале.

Для того чтобы удалить кластер после работы, необходимо воспользоваться командой Remove-AzureHDInsightCluster -Name <ClusterName>

Также созданные ресурсы можно удалить вручную на портале.

Использование Azure Cli

Для того чтобы установить Azure Cli, Вы можете воспользоваться несколькими способами (https://azure.microsoft.com/ru-ru/documentation/articles/xplat-cli-install/).

Рекомендуется скачать установщик с официального сайта по вышеприведенной ссылке.

Однако если у Вас не получилось установить Azure Cli, Вы можете воспользоваться вторым способом (Установка файла Node.js и выполнение команды npm install). При работе на Windows необходимо установить Python версии >= 2.5 и < 3, Visual Studio 2013 и запустить команду

npm install --msvs_version=2013

После этого Azure Cli должен установиться на Ваш компьютер.

О том, как пользоваться Azure Cli, Вы можете прочитать по адресу: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-provision-linux-clusters/#cli

Запускаем Azure Cli.

Для начала необходимо залогиниться:

azure login

Обратите внимание, что обычные логин и пароль не подходят, при работе с Azure Cli необходимо создать новую учетную запись организации. Для этого необходимо зайти на старую версию портала: https://manage.windowsazure.com/ и выполнить несколько простых действий:

https://azure.microsoft.com/en-us/documentation/articles/xplat-cli-connect/#create-an-organizational-account

После этого необходимо перейти в режим Azure Resource Manager:

azure config mode arm

Далее скачиваем с сайта

https://github.com/matt1883/azure-quickstart-templates/tree/master/hdinsight-linux-ssh-publickey (обратите внимание, что на вышеприведенном сайте Azure Cli ссылки битые) скрипты: azuredeploy.json и azuredeploy.parameters.json

Открываем файл azuredeploy.parameters.json и редактируем параметры.

Далее создаем группу ресурсов:

azure group create RESOURCEGROUPNAME LOCATION

После этого непосредственно создаем хранилище и кластер командой

azure group deployment create -f PATHTOTEMPLATE -e PATHTOPARAMETERSFILE -g RESOURCEGROUPNAME -n InitialDeployment

(подробнее о данной операции Вы можете прочитать по вышеприведенной ссылке

https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-provision-linux-clusters/#cli)

Обратите внимание, что создать менее дорогой кластер с использованием Azure Cli достаточно проблематично (но уменьшить количество узлов в данном кластере достаточно просто).