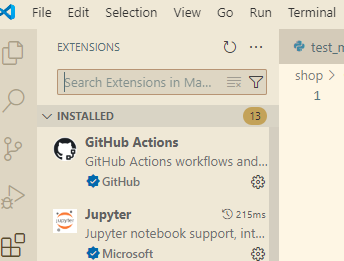
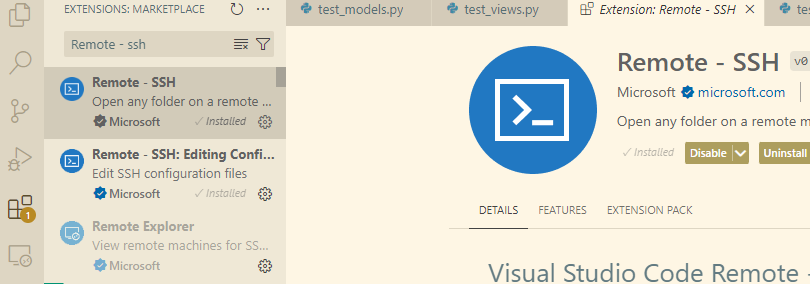
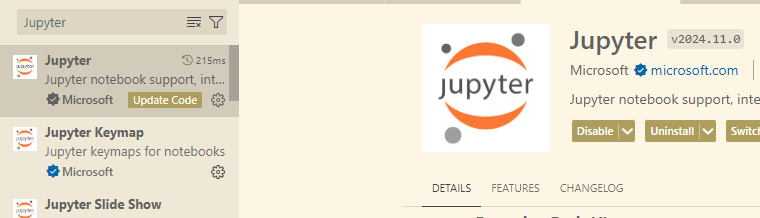
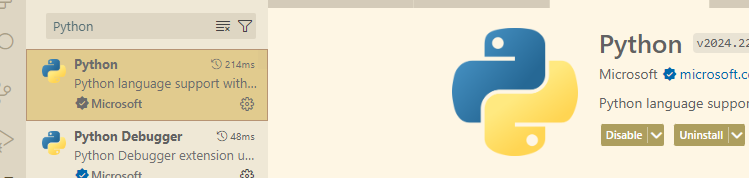
**Порядок выполнения работы:**

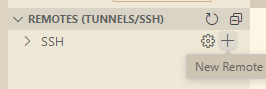
1. Установите среду разработки [Microsoft Visual Studio Code](https://code.visualstudio.com/).



1. Запустите VS Code, на левой панели откройте вкладку Extensions и установите расширения Python, Jupyter, Remote - SSH.



1. На левой панели откройте вкладку Remote Explorer, наведите мышкой на строку SSH и щелкните знак +.



1. В открывшейся строке наберите следующую команду, заменив user, ip и port на параметры учетной записи. Учетные записи на кластере доступны в ЭИОС. Если система сообщает, что порт 12345 занят, вместо него используйте любой свободный.

ssh -D 12345 -J user@ip:port user@hadoop-node  
(ssh -D 12345 -J user6@ip:57322 user6@hadoop-node)  


После введения строки укажите файл ~/.ssh/config в качестве хранилища записи.

После этого в файле ~/.ssh/config появится запись вида:

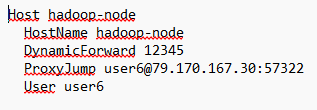
Host hadoop-node

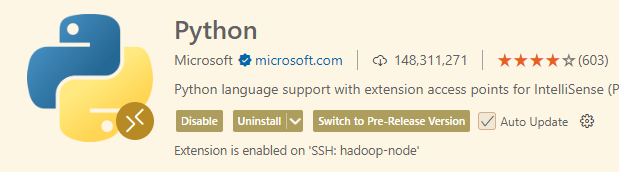
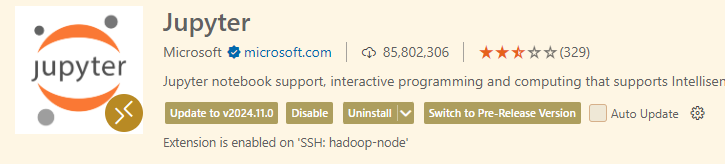
HostName hadoop-node

DynamicForward 12345

ProxyJump user@ip:port

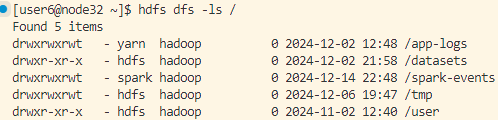
User user



1. Щелкните по кнопке с изображением стрелки на появившемся хосте. Если VS Code запросит платформу удалённого узла, укажите Linux.
2. При запросе пароля от учетной введите его (возможно, дважды). Дождитесь соединения с кластером и настройки удалённого сервера.
3. Установите на удаленный сервер расширения Python и Jupyter тем же способом, что и ранее (убедитесь, что рядом с именем расширения указано Install in SSH: hadoop-node)  
    
4. Нажмите на вкладку Terminal на панели инструментов VS Code и откройте новый терминал.

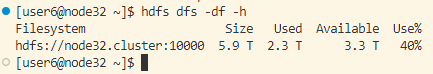


1. Проверьте доступность файловой системы HDFS, набрав команду:

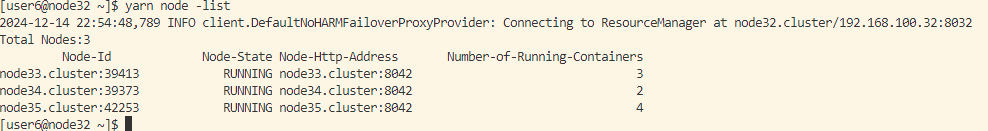
hdfs dfs -ls /  


Побродите по файловой системе HDFS, обратите внимание на права доступа к файлам и каталогам.

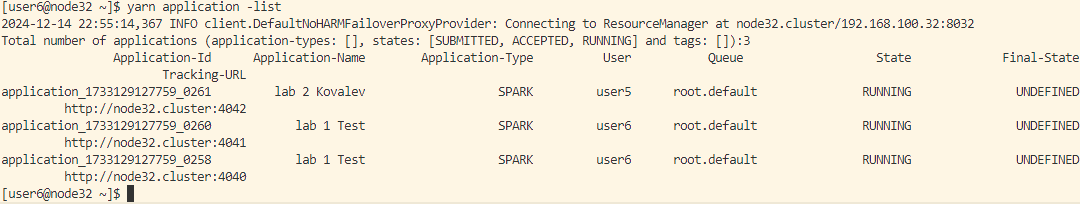
Проверьте объём оставшегося дискового пространства HDFS, набрав команду:

hdfs dfs -df -h  


1. Проверьте доступность системы YARN командой:

yarn node -list  


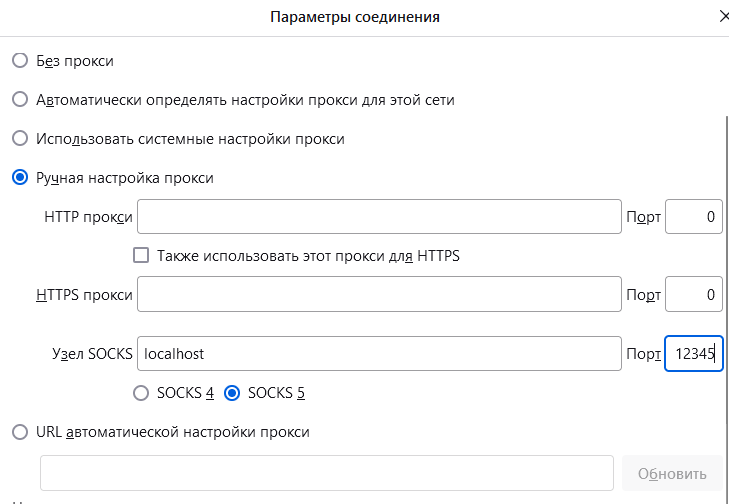
1. Проверьте загрузку очереди YARN, набрав команду:

yarn application -list  


Ненужное приложение можно остановить командой:

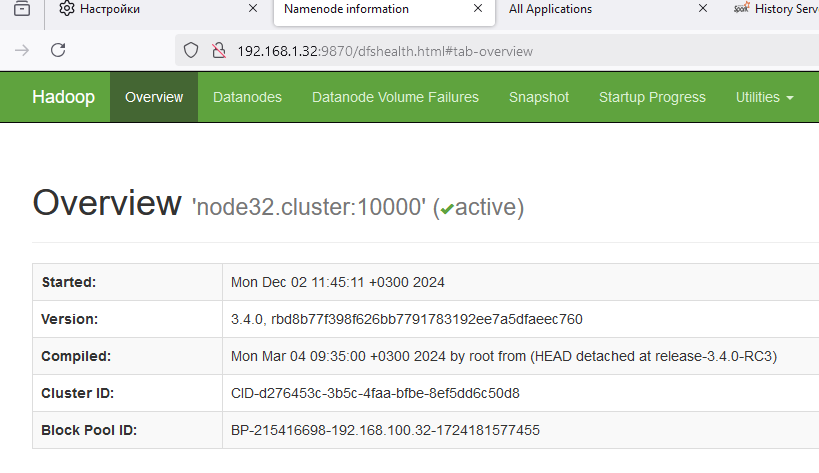
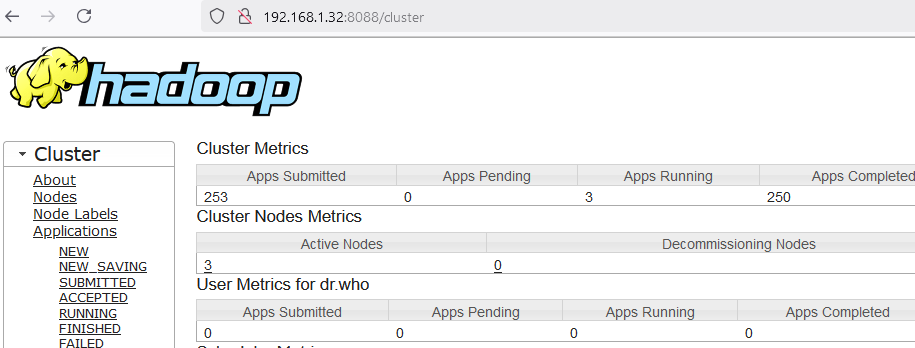
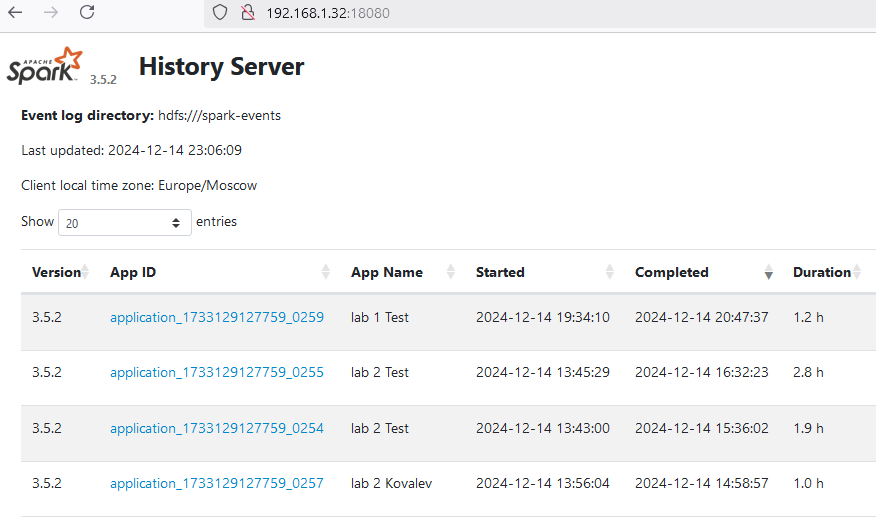
yarn application -kill <app-id>

1. Настройте браузер для использования SOCKS5 Proxy. Проще всего это сделать в браузере [Mozilla Firefox](https://www.mozilla.org/ru/firefox/new/), у которого есть настройки прокси, отличные от системных. В настройках укажите:

* Тип прокси для доступа в Интернет: Ручная настройка прокси.
* Узел SOCKS: localhost.
* Порт: 12345 -- это порт, который Вы указали в опции -D при настройке соединения с кластером.
* Тип прокси: SOCKS 5.  
  

1. Теперь в браузере должны быть доступны веб-страницы сервисов Hadoop и Spark:

| **Служба** | **Адрес** |
| --- | --- |
| Name Node | [http://192.168.1.32:9870](http://192.168.1.32:9870/) |
| Resource Manager | [http://192.168.1.32:8088](http://192.168.1.32:8088/) |
| Spark History Server | [http://192.168.1.32:18080](http://192.168.1.32:18080/) |

1. Исследуйте содержимое представленных страниц, оцените информацию, которую удалось получить.
2. Создайте **свою директорию** в локальной директории пользователя на кластере:

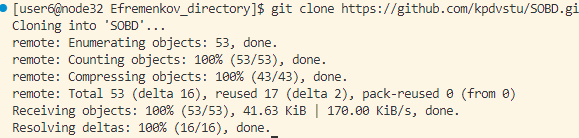
mkdir ivanov\_directory

и перейдите в неё:

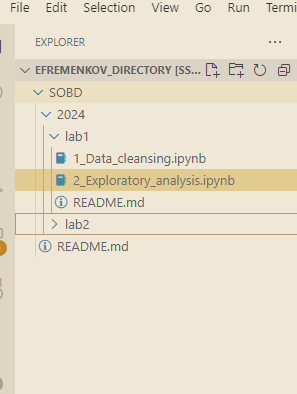
cd ivanov\_directory



1. Склонируйте репозиторий с лабораторными работами:

git clone <https://github.com/kpdvstu/SOBD.git>  


1. На левой панели VS Code нажмите на кнопку Open Folder и перейдите в директорию с первой лабораторной работой.



1. Откройте файлы Jupyter Notebook (\*.ipynb), выберите виртуальное окружение pyspark-env и поочерёдно запустите инструкции в файлах на исполнение. Некоторые инструкции потребуют модификации информации о запустившем их пользователе, не пропустите их.



* **Совет**. В процессе работы с ноутбуком удобно запустить в терминале команду, позволяющую отслеживать в реальном времени все работающие задачи в YARN:

watch yarn application -list

1. **Выполните аналогичным образом разведочный анализ датасета по Вашему выбору** с определением:

* типов признаков в датасете;
* пропущенных значений и их устранением;
* выбросов и их устранением;
* расчетом статистических показателей признаков (средних, квартилей и т.д.);
* визуализацией распределения наиболее важных признаков;
* корреляций между признаками.

**Обратите внимание**:

* Объём датасета -- не менее **нескольких гигабайт**.
* Датасеты можно выбрать на [Kaggle](https://www.kaggle.com/datasets) (для скачивания датасета требуется регистрация).
* Перемещать файлы между локальной файловой системой Вашего компьютера и локальной файловой системой удаленного узла кластера, к которому Вы подключились, можно с помощью программы [WinSCP](https://winscp.net/eng/download.php). Также можно просто перетащить файлы мышкой в нужную папку в VS Code.
* Загрузить файлы с локальной файловой системой удаленного узла кластера в распределенную файловую систему HDFS можно с помощью следующей команды:

hdfs dfs -copyFromLocal /home/ivanov/dataset/data/ data  
hdfs dfs -copyFromLocal /home/user6/Efremenkov\_directory/dataset/data/epc\_cut.csv Efremenkov\_directory

ЕСЛИ не грузит load, удалить файл hdfs dfs -rm Efremenkov\_directory/epc\_cut.csv и звгрузить с помощью copyFromLocal

если \_processed\_ грохнулись hdfs dfs -rm -r /user/user6/warehouse/Efremenkov\_database и заново загрузить

* А загрузить файлы с HDFS в локальную директорию удаленного узла -- командой:

hdfs dfs -copyToLocal /home/user6/Efremenkov\_directory/dataset/data/

1. Сделайте выводы по работе.
2. Напишите **главу курсовой работы** в пределах **15-20 страниц**, в которой опишите постановку задачи, описание Вашего датасета со ссылкой на него, проведенный разведовательный анализ и выводы.
3. Сохраните ноутбук с проведенным анализом и написанную главу курсовой в GitHub / GitLab.

**К отчету** следует представить репозиторий на GitHub / GitLab с выполненным разведочным анализом и главой курсовой работы, а также быть готовым продемонстрировать работоспособность кода и пояснить спорные моменты.

