

## Демонстрация Yandex DataSphere

1. Авторизация в Yandex.Cloud.
2. Создание нового проекта в DataSphere.
3. Клонирование git-репозитория.
4. Концепция DataSphere, работа с File Manager, установка нужного пути.
5. Использование сниппетов.
6. Установка необходимых пакетов.
7. Запуск поочередно нескольких ячеек, выделение и запуск нескольких ячеек, просмотр результата и обсуждение продукта.
8. Запуск оставшихся ячеек на выполнение.
9. Проверка работы DataSphere при закрытии вкладки браузера (обновлении страницы).
10. Экспорт ноутбука и проекта.
11. Управление вычислительными ресурсами.

1. Авторизация в Yandex.Cloud
  - a. Открыть Яндекс Браузер.
  - b. Открыть Консоль:  
<https://console.cloud.yandex.ru/>
  - c. Войти в Yandex.Cloud.
  - d. Войти в свой каталог для лабораторной работы.
2. Создание нового проекта в DataSphere
  - a. Открыть DataSphere в нижнем меню.
  - b. Нажать кнопку «Создать проект».
  - c. Ввести любое название проекта (только строчными латинскими буквами и цифрами, без спецсимволов).
  - d. Ввести описание проекта (до 50 символов).
  - e. Нажать кнопку «Создать».
  - f. Открыть созданный нами проект (нажать его название).
3. Клонирование git-репозитория
  - a. Выбрать в меню «Git» -> «Clone».
  - b. Скопировать адрес в строке:  
[https://github.com/dalyona/Yandex\\_Scale\\_DataSphere\\_demo](https://github.com/dalyona/Yandex_Scale_DataSphere_demo)
  - c. Нажать кнопку «Clone».
  - d. Дождаться, пока слева в меню появится каталог Yandex\_Scale\_DataSphere\_demo, открыть его двойным щелчком.
  - e. Ознакомиться с содержимым ноутбука (опрос пользователей Kaggle 2017 года).
4. Концепция DataSphere, работа с File Manager, установка нужного пути
  - a. DataSphere работает как сервис (слайды архитектуры).

- b. Доступ к файловой системе осуществляется через оболочку Python. Терминал отключен. При переключении машины данные в рабочей директории прозрачно переносятся.
  - c. Запустить выполнение первой ячейки с кодом:  

```
import os  
os.getcwd()
```
  - d. Пронаблюдать, как виртуальная машина запустила выполнение ячейки, отслеживать статус выполнения.
  - e. Изменить рабочий каталог на  

```
%cd Yandex_Scale_DataSphere_demo
```
5. Использование сниппетов
- a. Разархивировать архив с данными для модели при помощи сниппета.
  - b. Выбрать в меню «Snippets» -> «Extract ZIP file.py».
  - c. Команда добавит новую ячейку с кодом для разархивации файла.
  - d. Изменить в имени файла `fname = './file.zip'` на `fname = './input.zip'`
  - e. Запустить ячейку на выполнение, в файловом менеджере разархивируется каталог с данными.
6. Установка необходимых пакетов:
- a. Часть необходимых библиотек и пакетов уже установлена в DataSphere. Для их импорта следует использовать стандартную команду `import`. Список предустановленных библиотек можно посмотреть в [документации](#) или с помощью команды:  

```
%pip list
```
  - b. Установить пакеты, необходимые для работы, но не включенные в перечень уже установленных, с помощью команды:  

```
%pip install <Имя пакета>
```
  - c. Запустить ячейку с установкой пакетов и библиотек (треугольник «Run» на панели).
7. Поочередный запуск нескольких ячеек, выделение и запуск нескольких ячеек, просмотр результата, обсуждение продукта
- a. Запустить поочередно несколько ячеек командой Run (треугольник «Run» на панели).
  - b. Чтобы выделить несколько ячеек, удерживать клавишу «shift» и нажать левую кнопку мыши слева от ячейки.
  - c. Запустить несколько выделенных ячеек на выполнение в меню «Run» -> «Run Selected Cells».
  - d. Просмотреть результаты.
8. Запуск всех оставшихся ячеек на выполнение
- a. Запустить на выполнение выделенную ячейку и все следующие. Для этого выбрать в меню «Run» -> «Run Selected Cell and All Below».

- b. Дождаться, пока завершатся все вычисления, пролистать ноутбук до конца и посмотреть статистику опроса в различных разрезах.
- 9. Проверка работы DataSphere при закрытии вкладки браузера (обновлении страницы)
  - a. Закрыть вкладку браузера, в которой запущен ноутбук.
  - b. Вернуться к списку проектов.
  - c. Открыть наш проект еще раз, дождаться его загрузки.
  - d. Все состояния ноутбука, все данные и переменные, все вычисления сохранились.
- 10. Экспорт ноутбука и проекта
  - a. Готовым ноутбуком можно поделиться.
  - b. Варианты экспорта в виде отчета (HTML-страницы): выбрать в меню «File» -> «Export Notebook as HTML».
  - c. Получить ссылку, скопировать ее.
  - d. Открыть новую вкладку в браузере, скопировать в нее полученную ссылку, посмотреть отчет.
  - e. Чтобы скачать весь проект целиком в виде zip-файла, выбрать в меню «File» -> «Export project as zip».
  - f. Чтобы скачать произвольный файл проекта, выделить его и в контекстном меню выбрать пункт «Download».
- 11. Управление вычислительными ресурсами
  - a. Вычислительные ресурсы в DataSphere можно переключать прямо внутри ноутбука, из ячейки, с полным сохранением данных, переменных, состояния.
  - b. Изменить тип виртуальной машины, на которой выполняется ячейка, можно на панели управления:  
«S (4 cores) - default»  $\leftrightarrow$  «M (8 cores)».  
В превью доступны три типа виртуальных машин, в дальнейшем их количество будет расширено.
  - c. Переключить в последней ячейке тип машины на «M»:  
«S (4 cores) - default»  $\rightarrow$  «M (8 cores)».  
В ячейку добавится служебная команда «#!/M», которая показывает, что данная ячейка будет выполняться на машине типа M.
  - d. Запустить ячейку на выполнение еще раз, понаблюдать запуски виртуальной машины типа M. На ней повторно производятся все вычисления, при этом полностью сохраняются данные, переменные, состояние ноутбука на момент «до переключения».
  - e. Новые ячейки (добавляются кнопкой «+» на панели) по умолчанию будут создаваться с тем типом машины, на которой в последний раз велись расчеты.
  - f. Весь уже существующий ноутбук можно запустить на другом типе машин. Для этого выбрать все ячейки ноутбука (меню «Edit» -> «Select All Cells»), а затем — нужный тип машины на инструментальной панели.

- g. Аналогичным образом можно запустить параллельные вычисления на кластере SPARK в Data Proc (показываем на слайде, в лабораторной работе демонстрация данной функциональности не предусмотрена).