

## Демонстрация Yandex DataSphere

1. Создание проекта в DataSphere
2. Клонирование Git-репозитория
3. Концепция DataSphere, работа с File Manager, установка пути
4. Использование сниппетов
5. Установка (обновление на примере TF) пакетов
6. Запуск поочередно нескольких ячеек, выделение и запуск нескольких ячеек, просмотр результата и обсуждение продукта
7. Управление вычислительными ресурсами
8. Версионирование, контрольные точки
9. Запуск оставшихся ячеек на выполнение
10. Проверка работы DataSphere при закрытии вкладки браузера (обновлении страницы)
11. Экспорт ноутбука и проекта
12. Работа с Data Proc, настройка проекта
13. Работа с Data Proc, распределенные вычисления

1. Создание проекта в DataSphere
  - a. Откройте Яндекс.Браузер (рекомендуется).
  - b. Откройте консоль: [console.cloud.yandex.ru](https://console.cloud.yandex.ru).
  - c. Войдите в Yandex.Cloud.
  - d. Войдите в свой каталог.
  - e. В нижнем меню откройте DataSphere.
  - f. Нажмите кнопку **Создать проект**.
  - g. Введите название (строчными латинскими буквами и цифрами, без спецсимволов).
  - h. Введите описание проекта (до 50 символов).
  - i. Нажмите кнопку **Создать**.
  - j. Нажмите название проекта, чтобы открыть его.
2. Клонирование Git-репозитория
  - a. В меню выберите Git — Clone.
  - b. Вставьте в строку адрес репозитория:  
[https://github.com/dalyona/DataSphere\\_demo\\_Nov20](https://github.com/dalyona/DataSphere_demo_Nov20)
  - c. Нажмите кнопку Clone.
  - d. Слева в меню появится каталог *DataSphere\_demo\_Nov20*. Откройте его.
  - e. Ознакомьтесь с содержимым ноутбука *Demo\_kaggle2017.ipynb* (опрос пользователей Kaggle 2017 года).
3. Концепция DataSphere, работа с File Manager, установка пути

DataSphere работает как сервис (слайды архитектуры).  
Для доступа к файловой системе используется оболочка Python. Терминал отключен. При переключении машины данные в рабочей директории прозрачно переносятся.

- a. Запустите выполнение первой ячейки с кодом:  
`import os`  
`os.getcwd()`
- b. Понаблюдайте, как виртуальная машина запустила выполнение ячейки, отслеживайте статус выполнения.
- c. Измените рабочий каталог на  
`%cd DataSphere_demo_Nov20`

#### 4. Использование сниппетов

- a. Разархивируйте архив с данными для модели при помощи сниппета. Для этого выберите в меню Snippets — Extract ZIP file.py. Команда добавит ячейку с кодом для разархивации файла.
- b. Измените название файла `fname = './file.zip'` на `fname = './input.zip'`.
- c. Запустите выполнение ячейки. В файловом менеджере разархивируется каталог с данными.

#### 5. Установка пакетов

Часть библиотек и пакетов уже установлена в DataSphere. Импортируйте их стандартной командой `import`. Список предустановленных библиотек можно посмотреть в [документации](#) или с помощью команды

```
%pip list
```

- a. Установите дополнительные пакеты с помощью команды  
`%pip install <Имя пакета>`
- b. Запустите ячейку с установкой пакетов и библиотек (треугольник Run на панели).
- c. В новом релизе мы добавили возможность обновить предустановленные библиотеки до любой версии. Проверьте версию библиотеки и обновите ее, выполнив команды:  
`%pip show urllib3`  
`%pip install urllib3==1.24`  
*Примечание:* для обновления библиотек до последней версии используйте команду с аргументом `-U`:  
`%pip install urllib3-U`

#### 6. Поочередный запуск нескольких ячеек, выделение и запуск нескольких ячеек, просмотр результата, обсуждение продукта

- a. Выберите команду Run (треугольник **Run** на панели), чтобы запустить несколько ячеек поочередно.
- b. Удерживайте Shift и нажимайте левую кнопку мыши слева от ячеек, чтобы выделить несколько ячеек.
- c. В меню выберите Run — Run Selected Cells, чтобы запустить выполнение выделенных ячеек.
- d. Просмотрите результаты.

#### 7. 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, в выпадающем списке.
  - c. Выберите в следующей ячейке в выпадающем списке тип машины **M (8 cores)**, чтобы в этой ячейке переключить тип машины на M. В ячейку добавится служебная команда `#!/M`. Она показывает, что ячейка будет выполняться на машине типа M.
  - d. Запустите выполнение ячейки, наблюдайте, как запустится виртуальная машина типа M. На ней производятся все вычисления, при этом сохраняются все данные, переменные, состояние ноутбука на момент до переключения.
  - e. Важно: состояние машины отражается именно для текущей ячейки, поэтому у новой ячейки тип виртуальной машины опять будет S. Обратите внимание: сейчас после переезда на новую машину не сохраняется путь для чтения файла, если мы его меняли. Поэтому еще раз смените путь:  
`%cd DataSphere_demo_Nov20`
  - f. Выполните следующую ячейку.
  - g. По умолчанию для нового ноутбука и новой ячейки используется тип S. Весь ноутбук можно запустить на другом типе машин. Для этого выберите все ячейки ноутбука (Edit — Select All Cells), а затем на панели инструментов — тип машины.
8. Версионирование, контрольные точки
- a. В DataSphere появились контрольные точки, или Checkpoints. Найдите их на панели слева.
  - b. Откройте панель контрольных точек, выберите последнее состояние, зафиксируйте его кнопкой **Pin**.
  - c. Вернитесь к предпоследнему состоянию, проверьте, что откат произошел.
9. Запуск всех оставшихся ячеек на выполнение
- a. В меню выберите Run — Run Selected Cell and All Below, чтобы запустить выполнение выделенной ячейки и всех следующих.
  - b. Дождитесь, пока завершатся вычисления, пролистайте ноутбук до конца и посмотрите статистику опроса в различных разрезах.
10. Проверка работы DataSphere при закрытии вкладки браузера (обновлении страницы)
- a. Закройте вкладку браузера, в которой запущен ноутбук.
  - b. Вернитесь к списку проектов.
  - c. Откройте наш проект еще раз, дождитесь, пока он загрузится. Убедитесь, что состояние ноутбука, данные и переменные, вычисления сохранились.
11. Экспорт ноутбука и проекта

Готовым ноутбуком можно поделиться несколькими способами:

- a. *Экспортировать ноутбук в виде HTML-страницы*
  - В меню выберите File — Export Notebook as HTML.
  - Получите ссылку, скопируйте ее.
  - Откройте вкладку в браузере, вставьте туда ссылку и посмотрите отчет.
- b. *Скачать проект целиком в формате ZIP*
  - В меню выберите File — Export Project as ZIP.
- c. *Скачать файл проекта*
  - Выделите файл и в контекстном меню выберите Download.
- d. *Экспортировать ноутбук с состоянием*
  - Из меню контрольных точек сделайте Pin для контрольной точки.
  - В меню выберите Share, чтобы поделиться контрольной точкой.
  - Чтобы импортировать сохраненное состояние, выберите в меню File — Import Notebook from Checkpoints.

Возможностью экспорта ноутбука с состоянием может воспользоваться любой пользователь DataSphere.

## 12. Работа с Data Proc, настройка проекта

В DataSphere можно:

— работать с кластерами Data Proc, созданными в сервисе Data Proc;

— создать временный кластер Data Proc непосредственно из DataSphere.

Мы создадим временный кластер Data Proc из DataSphere и выполним на нем простые вычисления. Для этого необходимо указать в дополнительных настройках проекта сервисный аккаунт и подсеть, которые будут использоваться для кластера.

- a. Закройте вкладку браузера с проектом.
- b. На вкладке Консоль выберите наш проект, рядом с ним нажмите многоточие (...) и выберите **Изменить**.
- c. На вкладке **Изменение проекта** выберите **Дополнительные настройки**.
- d. Для пункта меню **Сервисный аккаунт** создайте аккаунт, нажмите кнопку **Создать новый**. Задайте имя, оставьте обязательные роли в каталоге vpc.user, mdb.admin и mdb.dataproc.agent, назначенные аккаунту по умолчанию.
- e. **Важно:** в пункте меню «Подсеть» выберите подсеть в зоне доступности ru-central1-a.
- f. Сохраните свойства проекта.  
**Обратите внимание:** проекты, у которых в настройках указана кастомная подсеть, требуют больше времени на выделение и переключение машин.
- g. Нажмите название измененного нами проекта, чтобы открыть его.

## 13. Работа с Data Proc, распределенные вычисления.

- a. Создайте временный кластер Data Proc из DataSphere.  
В меню выберите File — Data Proc Clusters.

- b. В форме создания кластеров создайте новый кластер, выберите минимальную конфигурацию из списка (более мощные могут потребовать расширения стандартных квот).  
Для этого в разделе Create new cluster задайте имя кластера и оставьте тип кластера по умолчанию XS.
- c. Нажмите кнопку создания **Create**.
- d. Кластер появится в списке создаваемых со статусом Starting. Дождитесь, пока он поднимется и статус изменится на Up. Возможно, для обновления статуса понадобится закрыть и еще раз открыть окно Data Proc Clusters.
- e. В File Manager слева выберите файл Spark\_example.ipynb и откройте его. Вы увидите этот файл в открывшейся дополнительной вкладке.
- f. В первой строке добавьте название созданного кластера:  
`#!/spark --cluster test`
- g. Запустите выполнение ячейки.