

Настройка облачной инфраструктуры для проекта по определению мошеннических транзакций

Домашнее задание No 2

Цель работы. В данном домашнем задании Вы познакомитесь с облачным провайдером Yandex Cloud, поработаете с сервисами Object Storage и Data Proc, создадите свой Spark-кластер и скопируете в него данные, научитесь оценивать затраты при проектировании облачной инфраструктуры.

Уважаемый слушатель!

...Наконец-то! Оформив трудоустройство и уладив все формальности юристами компании, Вы можете приступить к работе. В первую очередь, Вас интересуют данные о транзакциях, которые были собраны за последние несколько лет. Пообщавшись с системным администратором, Вы узнаете, что необходимая Вам информация расположена в озере данных компании в объектном хранилище, которое на время проведения работ по проекту будет Вам доступно по адресу:

s3://otus-mlops-source-data/

Поскольку данное хранилище предоставляется в режиме «только для чтения», то Вам потребуется перенести данные в другое, уже Ваше, хранилище для дальнейшей работы с ними, которое нужно будет предварительно создать. В качестве такого хранилища обычно используют либо S3-bucket, либо HDFS.

Также, поскольку заказчик не имеет возможности предоставить собственную вычислительную инфраструктуру для выполнения проекта, для этой цели придется использовать облачные ресурсы. Поскольку данные для анализа расположены на серверах Yandex Cloud, то логично развернуть там и кластер для их обработки.

Обратите внимание!

Перед выполнением работы Вы можете запросить у менеджеров Otus промокод к Yandex Cloud, позволяющий работать с ресурсами облака в течение рамках данного курса.

Вам предлагается:

Вам предлагается на основе представленной информации:

1. Создать новый bucket в Yandex Cloud Object Storage с использованием terraform скрипта. Примеры вы сможете найти в материалах занятия или на странице документации Yandex cloud (<https://yandex.cloud/ru/docs/storage/tf-ref>). Выложить созданный вами скрипт в github репозиторий с заданием.

2. Скопировать в него содержимое предоставленного Вам хранилища с использованием инструмента s3cmd. Для проверки преподавателем данный basket необходимо сделать общедоступным, а точку доступа к нему привести в README-файле Вашего GitHub-репозитория.

3. Создать Spark-кластер в Data Proc с двумя подкластерами со следующими характеристиками:

а) Мастер-подкластер: класс хоста s3-c2-m8, размер хранилища 40 ГБ.

б) Data-подкластер: класс хоста s3-c4-m16, 3 хоста, размер хранилища 128 ГБ.

Для экономии ресурсов необходимо использовать terraform скрипт для создания и удаления кластера. Примеры вы сможете найти в материалах занятия или на странице документации Yandex cloud (<https://yandex.cloud/ru/docs/data-proc/tf-ref>). Выложить созданный вами скрипт в github репозиторий с заданием.

4. Соединиться по SSH с мастер-узлом и выполнить на нём команду копирования содержимого хранилища в файловую систему HDFS с использованием инструмента hadoop distcp. Для проверки преподавателем необходимо вывести содержимое HDFS-директории в консоль, а снимок экрана с этой информацией привести в README-файле Вашего GitHub-репозитория.

5. Пользуясь тарифным калькулятором Yandex Cloud, оценить месячные затраты для поддержания работоспособности созданного кластера. Оценить, насколько использование HDFS-хранилища дороже, чем объектного.

Указание. Кроме тарифного калькулятора, позволяющего делать оценку требуемых средств, на странице платежного аккаунта есть раздел с детализацией биллинга за произвольный период времени. С его помощью можно определить сумму уже потраченных средств на каждый из используемых облачных сервисов в процессе работы.

6. Предложить способы для оптимизации затрат на содержание Spark- кластера в облаке и попробовать их реализовать.

7. В соответствии с достигнутыми результатами, изменить статус ранее созданных задач на Kanban-доске в GitHub Projects. Возможно, некоторые задачи нужно будет скорректировать, разделить на подзадачи или объединить друг с другом.

8. Полностью удалить созданный кластер, чтобы избежать оплаты ресурсов в период его простаивания. Воспользуйтесь для этого командой terraform destroy.

Обратите внимание!

Даже если кластер находится в выключенном состоянии, все равно облачный провайдер списывает денежные средства со счета клиента за резервирование части ресурсов. Это может привести к быстрому израсходованию денежных средств на счете!

Для получения положительной оценки за работу необходимо выполнить 1 – 4 и 8 пункты вышеприведенные задания.

Желаем успехов!