Обзор Databrick. Что облачный продукт может дать начинающим специалистам.

*Добрый день, уважаемые читатели! Данная публикация посвящена Databricks и она получилась не совсем обычный по двум причинам.*

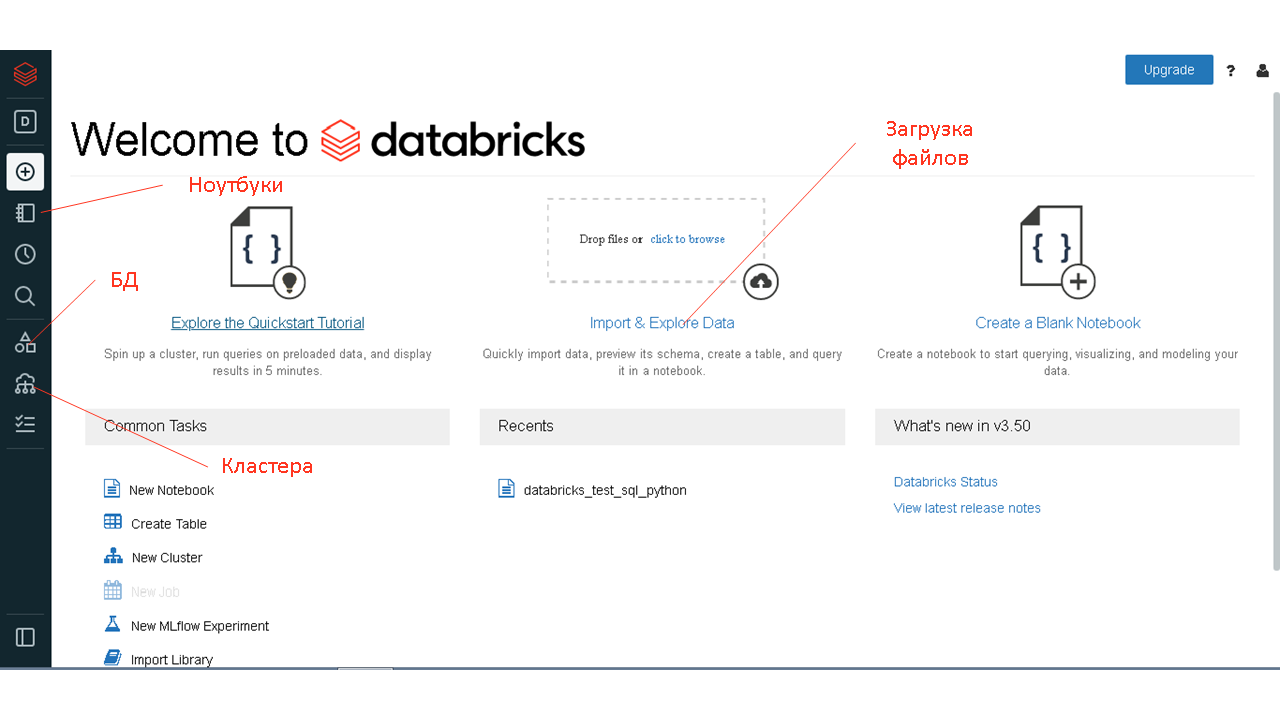
*Во-первых, в ней не будут обсуждаться типичные вопросы: целесообразность перевода всех сервисов и служб компаний в облако, возможные ошибки при такой миграции, стоимость использования сервиса, правовые аспекты хранения данных не на территории РФ. Это темы сложные, дискуссионные и без указания конкретных цифр и бизнес-факторов их сложно полноценно раскрыть в статье. Поэтому мы решили посмотреть на Databricks под другим углом и заострить внимание читателей на другом аспекте: что может дать изучение инструмента специалисту по работе с данными.*

*Во-вторых, чтобы взгляд на продукт был максимально комплексным и всесторонним я решил написать материал в соавторстве с* ***ФИО (должность, компания, ник на Хабр).*** *Таким образом, я постараюсь раскрыть начинающим аналитикам базовые возможности Databricks по анализу накопленной информации, а* ***ФИО*** *расскажет об основных моментах, которые могут заинтересовать инженеров данных. Это была преамбула, а теперь настало время познакомиться поближе с Databricks.*

Взгляд аналитика.

Для начала работы с платформой нужно пройти регистрацию на сайте. Сразу стоит оговориться, что лучше указывать о себе корректные данные, которые позволят без проблем пройти спам-фильтр. Если с предоставленными реквизитами все в порядке, далее предоставляется возможность выбрать один из двух вариантов взаимодействия с облачным сервисом: trial-версия и community. Рекомендую выбирать последний вариант, так как хотя он и несколько уступает в плане функциональности, но в отличие от trial-версии не имеет временных ограничений и полностью бесплатен. Это позволит основательно и постепенно разобраться с со всеми базовыми компонентами.

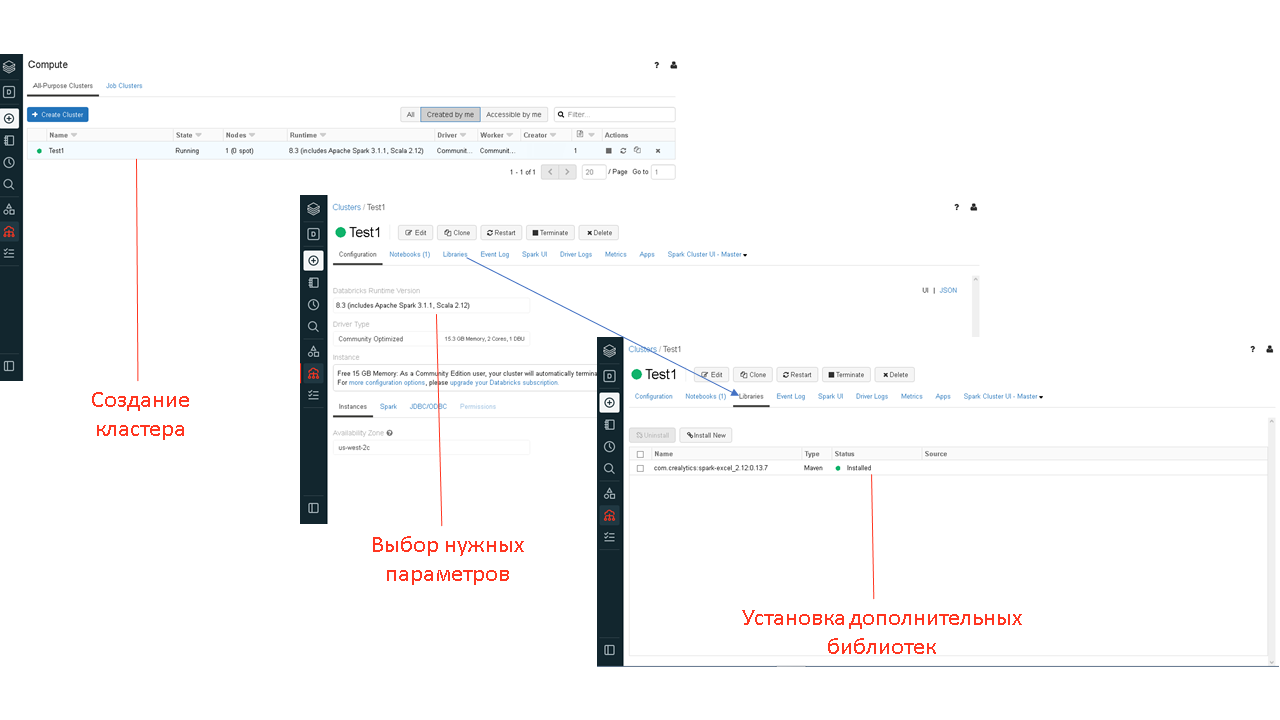
Начало работы с инструментом начинается со стартового рабочего поля. Для аналитика данных основными здесь будут 4 раздела: кластера, базы данных, ноутбуки и пункт по загрузке локальных файлов в облачное хранилище. Остальные компоненты, на мой взгляд, относятся больше к сфере data engineering и data science.



Теперь на образном примере я постараюсь объяснить, что такое Databricks. Для многих читателей данное сравнение покажется совершенно неразумным, и я знаю, что сравнивать столь разные программные продукты совершенно некорректно. Но апеллирование к уже знакомым технологиям всегда упрощает понимание нового.

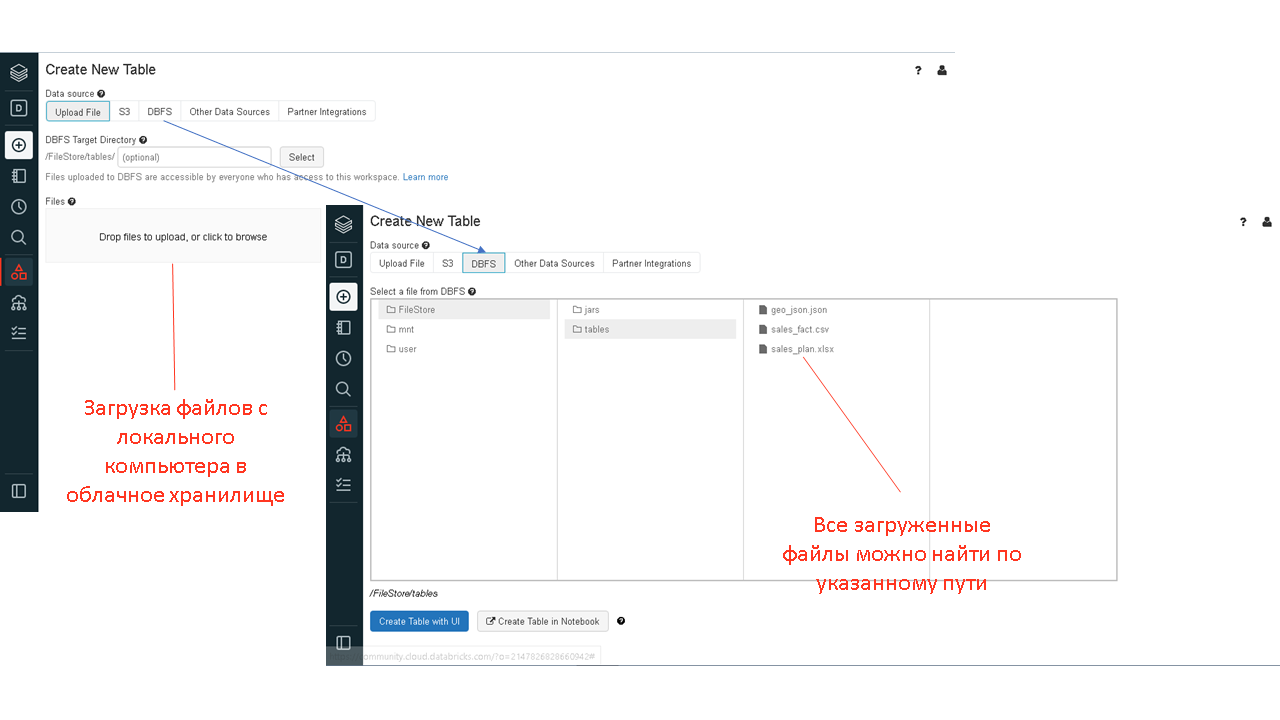
Итак, представьте себе Google Colab. Теперь мы запускаем ноутбук и устанавливаем Spark. Затем выбираем любую реляционную базу данных и инсталлируем ее. Самое интересное, что эти манипуляции абсолютно реальны и их без труда может выполнить любой специалист. В финале мы берем данные, заливаем их в БД и с помощью Spark начинаем проводить необходимые расчеты в ноутбуке. Представьте себе, что описанный функционал уже по умолчанию развернут на отдельно взятой облачной платформе и вам не нужно что-то дополнительно устанавливать. Прибавьте суда поддержку из коробки четырех языков работы с данными: SQL, Python, R, Scala. На выходе мы получаем рассматриваемый нами инструмент. Надеюсь, что такой метафорой я никого не сбил с толку. Пойдем дальше и посмотрим на основные разделы более предметно.

Перед тем, как открывать новый ноутбук, вам обязательно нужно запустить кластер.



При его старте возможно задать вручную некоторые параметры, но можно ставить все пункты по умолчанию. Если вам потребуется функционал какой-то специфической библиотеки, то ее легко можно инсталлировать в одноименном разделе. В рамках этого демонстрационного примера я захотел задействовать файл формата xlsx. Наверное, многих это может удивить (так как обработка данных с помощью Spark и файлы табличного редактора плохо сочетаются), но в этом нет ничего особенного. И вот вам самый простой пример из жизни. Факты хозяйственной деятельности предприятия занимают много места и традиционно хранятся в БД. А вот плановые показатели это всегда агрегированные величины, и в большинстве случаев в департамент по работе с данными они попадают именно в виде файлов xlsx. Отводить место в БД под плохо структурированные данные ситуационной аналитики в большинстве случаев нет никакого смысла, но обрабатывать их конечно нужно. Поэтому столкнувшись с нетиповой ситуацией, начинаем искать решение. В данном примере ответом на вопрос был пакет – com.crealytics:spark-excel.

Когда кластер создан и запущен, можно подумать об источниках информации. В реальной ситуации вам не придется генерировать тестовой массив, так как у компании уже будет какая-то база данных. Но в рамках демо-примера я взял за основу три файла с данными (csv, json, xlsx), чтобы затем на их основе строит расчеты.



Добавление файлов в систему происходит настолько просто, что нет особого смысла останавливаться на данном моменте. Единственное замечание, которое я хотел бы сделать касательно данного раздела. Если таблица данных плохо подготовлена к заливке в БД, то лучше не использовать пользовательский интерфейс, а прописать все моменты кодом.

Изображение выглядит как текст

Автоматически созданное описание

Вся подготовительная работа выполнена, можно запускать первый ноутбук. Чтобы посмотреть на все файлы, которые мы можем задействовать в работе, используем встроенную утилиту. По сути она помогает использовать команды Linux. Вручную манипулировать локальными файлами в системе не получиться.

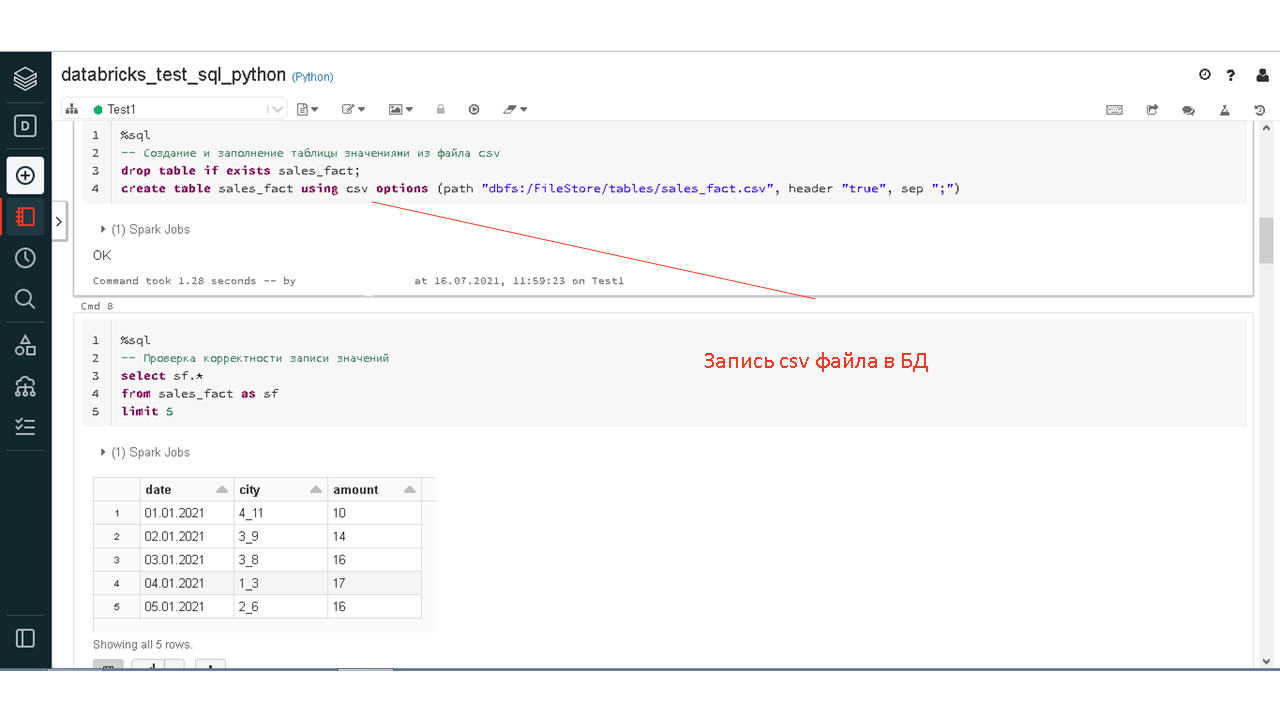
На старте в системе отсутствуют пользовательские базы данных. Исправить это “недоразумение” поможет стандартный язык запросов SQL, который хорошо знаком всем, кто хоть раз работал с реляционной БД.

Изображение выглядит как текст

Автоматически созданное описание

Чтобы упростить нам задачу я поместил информацию о фактических продажах в стандартный файл формата csv, который не требует каких-либо предварительных трансформаций перед заливкой в БД. Единственный недочет это разделитель (точка с запятой), и его мы укажем прямо по ходу. Проверяем результат азбучным запросом. Писать его можно так, как будто вы работаете в окне менеджера базы данных.

Единственный нюанс, который следует запомнить. При создании ноутбука вам предлагается выбрать язык манипулирования данными, который будет поддерживаться по умолчанию. Я выбрал Python, поэтому при написании скриптов SQL потребовалось указывать в начале кода специальную метку - %sql. Данное замечание справедливо также в отношении инъекций на языке R и Scala.

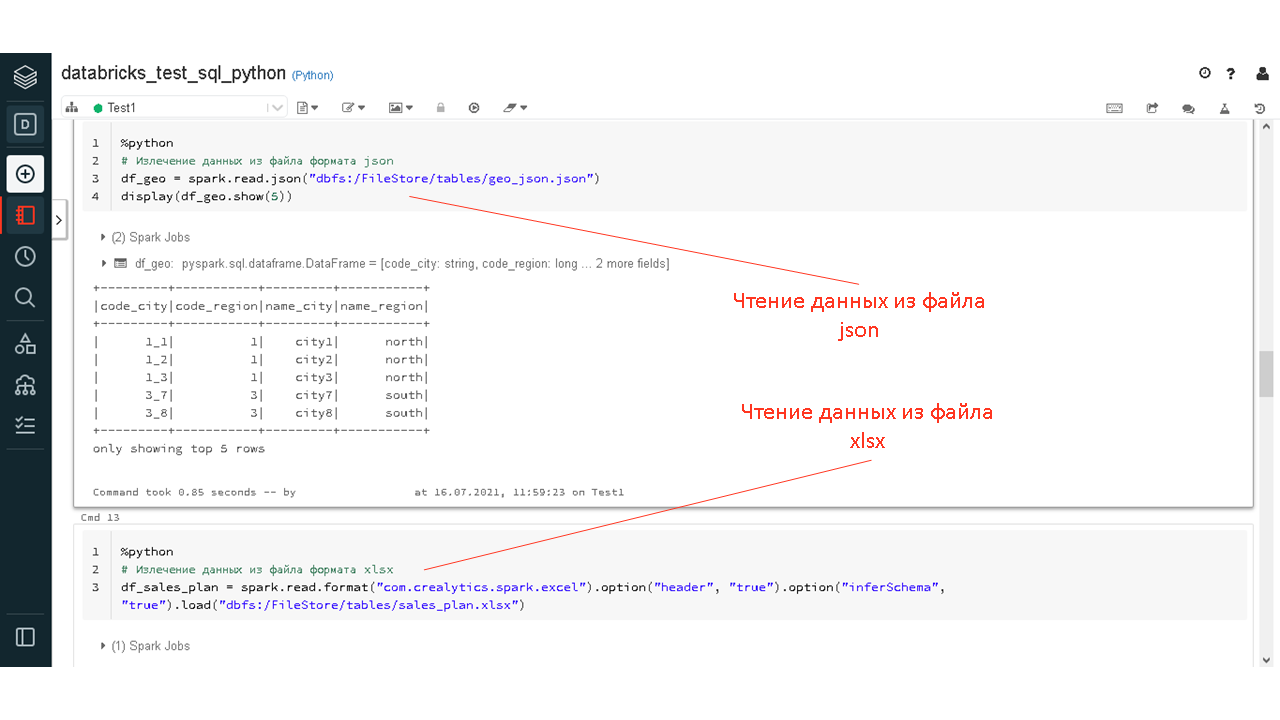


Результат sql-запроса легко можно сохранить в датафрейм Spark. Тема работы с данными посредством Spark довольно обширна. Помимо pandas-подобных манипуляций с массивами информации возможно также практиковать машинное обучение. Тема “разрастается” еще больше, так как взаимодействовать со Spark можно посредством SQL, R, Python, Scala, Java (последний язык облачный сервис не поддерживает).

Изображение выглядит как текст

Автоматически созданное описание

Если вы решили самостоятельно разобраться как обрабатывать информацию с помощью Spark, могу порекомендовать книги по заявленному вопросу двух издательств O’Reilly и Apress. В качестве основной книги для комплексной подготовки назову следующую рукопись “Learning Spark. Lightning-Fast Data Analytics”. Но просто прочитать книгу это одно, а вот запомнить все многообразие синтаксических конструкций это другое. Поэтому на первом этапе вам не помешает шпаргалка. Относительно удачный вариант формата pdf вы можете найти на портале Datacamp.



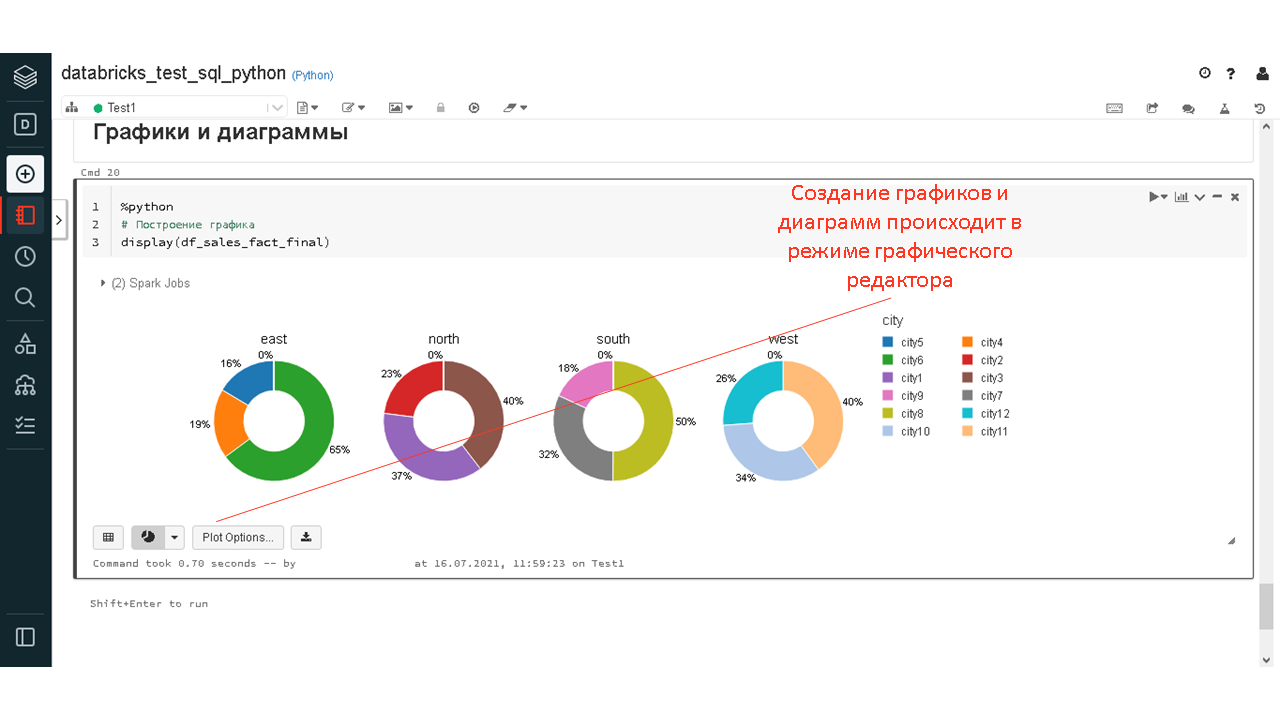
Если вам захочется самостоятельно повторить все этапы расчетов, приведенные на скриншотах, то демонстрационный ноутбук и необходимые файлы можно скачать в **репозитории** на Гитхаб.

Важно также понимать, что целью таких скриптов может быть не только построение отчетов по запросу руководства, но и расчет витрин данных на регулярной основе.

Изображение выглядит как текст

Автоматически созданное описание

Поговорим немного о функционале системы по построению графиков и диаграмм. Для того, чтобы отобразить данные на графике в ноутбуках Google Colab используются библиотеки matplotlib и seaborn. Прорисовка элементов диаграмм осуществляется посредством кода. Что же касается разработчиков Databricks, то они пошли другим путем и дали возможность пользователям создавать визуализации в режиме графического редактора. В комплект системы входят самые распространенные графики, которые по желанию аналитика могут быть собраны на отдельный дашборд. Сразу стоит оговориться, что данный функционал по возможностям не может тягаться с профессиональными и узкоспециализированными BI-решениями. Визуализации в данном случае носят вспомогательный характер, позволяя инженеру или аналитику глазами контролировать ситуацию с данными в системе.



**Краткие выводы по данному разделу.**

*1). Для аналитика данных облачный сервис представляет собой сплав уже знакомых технологий и инструментов, поэтому для самостоятельного изучения подойдет как документация по продукту, так и сторонняя литература.*

*2). Платформа хорошо подойдет тем специалистам, кто хочет разобраться со Spark, при этом ничего не устанавливая на свой ПК.*

*3). Наработанные подходы и практики в результате взаимодействия с Databricks могут быть с успехом применяться в сторонних решениях.*

Взгляд инженера данных.

На этом все. Всем здоровья, удачи и профессиональных успехов!