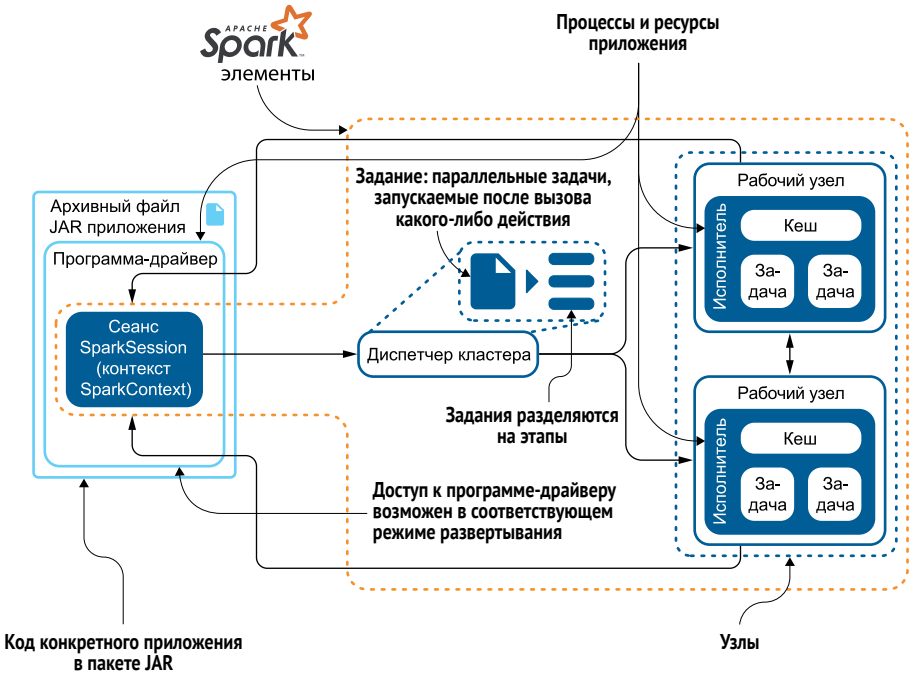
Оглавление

Содержание

Словарь терминов

Краткое описание терминов Spark, используемых в процессе развертывания.

Термин на английском языке	Термин на русском языке	Определение
Application	Приложение	Программа, которая создается в рабочей среде Spark и для функционирования в Spark. Состоит из программы-драйвера и исполнителей в кластере
Application JAR	JAR-приложение	Архивный файл Java (JAR), содержащий приложение Spark. Это может быть файл uber-JAR, включающий все зависимости
Cluster manager	Диспетчер кластера	Внешний сервис для распределения ресурсов в кластере. Возможно использование встроенного в Spark диспетчера кластера
Deploy mode	Режим развертывания	Определяет, где запускается и работает процесс драйвера. В режиме кластера (cluster mode) фреймворк запускает драйвер внутри кластера. В режиме клиента (client mode) ведомый (подчиненный) запускает драйвер за пределами кластера. Можно определить, в каком режиме вы находитесь, выполнив метод `deployMode()`. Метод возвращает свойство, защищенное от изменения (только для чтения)
Driver program	Программа-драйвер	Процесс, выполняющий основную функцию `main()` приложения и создающий контекст `SparkContext`. Все начинается именно здесь
Executor	Исполнитель	Процесс, запускаемый для приложения на рабочем узле. Исполнитель выполняет задачи и сохраняет промежуточные данные в памяти или на диске. У каждого приложения имеются собственные исполнители
Job	Задание	Параллельное вычисление (выполнение), состоящее из нескольких задач, которые порождаются в ответ на некоторое действие Spark (например, `save()` или `collect()`, подробнее см. приложение I)
Stage	Этап	Каждое задание разделяется на более мелкие наборы задач, называемые этапами, зависящими друг от друга (подобно этапам отображения и сокращения в MapReduce)
Task	Задача	Минимальная единица работы, которая передается одному исполнителю
Worker node	Рабочий узел	Любой узел, который может выполнять код приложения в кластере


APACHE Spark
элементы
Процессы и ресурсы приложения
Задание: параллельные задачи, запускаемые после вызова какого-либо действия
Архивный файл JAR приложения
Программа-драйвер
Сеанс SparkSession (контекст SparkContext)
Диспетчер кластера
Задания разделяются на этапы
Доступ к программе-драйверу возможен в соответствующем режиме развертывания
Рабочий узел
Исполнитель
Кеш
За-дача
За-дача
Рабочий узел
Исполнитель
Кеш
За-дача
За-дача
Код конкретного приложения в пакете JAR
Узлы

Вступительное слово

Аналитическая операционная система

В XX веке масштабы деятельности в деловой сфере значительно укрупнились благодаря глобальному расширению и распространению. Любая компания, выполняющая бизнес-операции по всему миру, получила вытекающее из этого преимущество в стоимости и степени распространения, что привело к появлению более конкурентоспособной продукции. Предприятия розничной торговли с глобальной сетью магазинов получили преимущество в распространении товаров, несравнимое с деятельностью более мелких компаний. Результаты этого глобального укрупнения определили преимущества в конкуренции на несколько десятилетий вперед.

Всю ситуацию в целом изменил интернет. В настоящее время существует три главных результата глобального укрупнения:

- сеть – замкнутость, управляемая надежной сетью (Facebook, Twitter, Etsy и т. д.);
- экономика глобального укрупнения – более низкая цена единицы товара, управляемая объемом (Apple, TSMC и т. д.);
- данные – всеобъемлющее машинное обучение и интуиция на основе динамически изменяющихся совокупностей данных.

В книге Big Data Revolution (Wiley, 2015 г.) я исследовал несколько компаний, которые извлекали реальную выгоду из данных в результате глобального укрупнения бизнес-деятельности. Но сейчас, в 2019 году, большие данные все еще остаются в основном неисследованной областью в организациях по всему миру. Spark – аналитическая операционная система – представляет собой средство, способное изменить существующее положение.

Spark стал инструментом, изменившим инновационную политику компании IBM. Spark – это аналитическая операционная система, обеспечивающая стандартизацию и унификацию источников данных и средства доступа к данным. Стандартизированная программная мо-

дель Spark делает эту систему самым лучшим вариантом выбора для разработчиков, создающих приложения анализа огромных объемов данных. Spark сокращает время и снижает сложность создания аналитических рабочих потоков, позволяя разработчикам сосредоточиться на задачах машинного обучения и создания экосистемы на основе Spark. Мы наблюдали ранее и видим сейчас, как проект с открытым исходным кодом быстро становится источником инноваций в широком масштабе.

Эта книга погружает вас в мир Spark. В ней рассматривается мощь этой технологии и гибкость ее экосистемы, а также практические приложения Spark для работы в конкретной современной компании. Вне зависимости от того, являетесь ли вы инженером по обработке данных, научным исследователем в области обработки данных, разработчиком прикладных программ или инженером поддержки ИТ-операций, эта книга расскажет об инструментальных средствах и откроет секреты, которые вы должны знать, чтобы управлять инновациями в своей компании или сообществе.

Наша стратегия в IBM – создание собственных проектов на основе и с использованием успешных открытых платформ, но эти проекты должны быть значительными и должны выделяться на общем фоне. Spark является именно такой платформой. В IBM можно найти множество примеров применения этой стратегии, и вы получите такой же результат в своей компании, если выберете этот путь.

Spark можно считать инновацией – это аналитическая операционная система, в которой будут успешно развиваться новые решения, устраняющие сдерживающий фактор обработки больших данных. Кроме того, Spark – это сообщество высококвалифицированных ученых и аналитиков в области обработки данных, хорошо знающих Spark, которые могут быстро превратить любые сегодняшние задачи в завтрашние решения. Spark представляет собой один из самых быстро развивающихся проектов с открытым исходным кодом в истории ИТ. Присоединяйтесь к этому движению.

– Роб Томас (Rob Thomas),
первый вице-президент,
подразделение Cloud and Data Platform, IBM

Предисловие

Не думаю, чтобы для Apache Spark требовалось какое-то предварительное представление. Если вы читаете эти строки, то, вероятнее всего, уже кое-что знаете о теме этой книги: инженерии данных и науки о данных, применяемой в крупных масштабах с использованием распределенной обработки. Но Spark – это нечто большее, и об этом вы скоро узнаете, прочитав вступительное слово Роба Томаса и главу 1.

Как Обеликс упал в котел с магическим зельем[1], я «упал в котел» Spark в 2015 году. В то время я работал во французской компании, производящей аппаратные компоненты для компьютеров, где участвовал в проектировании высокопроизводительных систем для анализа данных. Вполне естественно, первое впечатление от Spark было скептическим. Затем я начал постоянно работать со Spark и понял, что результат зависит от самого пользователя. Первоначальный скептицизм превратился в настоящее глубокое увлечение превосходным инструментом, который позволяет обрабатывать данные чрезвычайно простым способом, – это мое искреннее убеждение.

Я начал работу над несколькими проектами с использованием Spark, и это позволило мне выступить с сообщениями на Spark Summit, IBM Think, а также сблизиться с инициативами All Things Open, Open Source 101. Кроме того, через местную пользовательскую группу Spark я принял участие в анимации региона Роли-Дурэм в Северной Каролине. Это позволило мне встретиться с замечательными людьми и наблюдать за многочисленными проектами, связанными со Spark. В результате я увлекся Spark еще больше.

В этой книге я попытался поделиться с вами своим увлечением.

---

[1] Обеликс (Obelix) – известный персонаж комиксов и мультфильмов. Обеликс – неизменный спутник Астерикса (Asterix). Когда галл Астерикс выпивает магическое зелье, он получает суперсилу, которая помогает ему постоянно побеждать римлян (и пиратов). В раннем детстве Обеликс упал в котел, в котором варилось магическое зелье, воздействие которого на Обеликса стало постоянным. В Европе комиксы (и мультфильмы) про Астерикса (и Обеликса) весьма популярны. Больше информации см. на www.asterix.com/en/.

Примеры (или домашние задания) в этой книге основаны на использовании языка Java, но в репозитории книги содержится также код на языках Scala и Python. Сразу после выпуска версии Spark 3.0 сотрудники издательства Manning и я решили убедиться в том, что книга соответствует самым свежим версиям, а не отстает от действительности.

Как вы могли догадаться, мне очень нравятся комиксы. Я вырос на комиксах. Мне нравится такой способ передачи информации, в чем вы можете убедиться сами, читая эту книгу. Это не книжка комиксов, но около 200 иллюстраций в ней должны помочь вам понять великолепный инструмент, который называется Apache Spark.

Спутником Астерикса является Обеликс, точно так же и у книги «Spark в действии» (второе издание) есть спутник – постоянные средства поддержки, которые можно загрузить бесплатно из раздела ресурсов сайта издательства Manning по быстрой ссылке: http://jgp.net/sia. Эти средства поддержки содержат справочную информацию о статических функциях Spark и в итоге должны превратиться в более полезные справочные ресурсы.

Неважно, понравится вам книга или нет, в любом случае отправьте мне сообщение в твиттер на @jgperrin. Если книга понравилась, напишите отзыв на Amazon. Если книга не понравилась, то, как говорят при бракосочетании, (скажите сейчас) или храните молчание вечно. И все же я надеюсь, что книга вам понравится.

Alea iacta est – Жребий брошен[1].

[1] По-английски это высказывание звучит так: «The die is cast». Эту фразу приписывают Юлию Цезарю (заклятому врагу Астерикса), так как именно Цезарь перевел свою армию через Рубикон: событие произошло, и отменить его невозможно – как в случае, когда эта книга вышла из печати и стала доступна читателям.

Благодарности

В этом разделе я выражаю благодарность всем, кто помог мне в работе над этой книгой. Возможно, в этом разделе я забыл упомянуть некоторых людей, и если кто-то не обнаружил здесь своего имени, то я очень сожалею об этом. Искренне сожалею. Эта книга потребовала огромных усилий, и работа над ней в одиночку, вероятнее всего, была бы оценена в две или три звезды на Amazon вместо пятизвездочного рейтинга, который мы с вами получим очень скоро (это награда и благодарность за труд!).

Я хотел бы начать с благодарности тем рабочим группам, которые доверили мне этот проект, начиная с Zaloni (Анупам Ракшит (Anupam Rakshit) и Туфейл Хан (Tufail Khan)), Lumeris (Йон Фарн (Jon Farn), Сурья Кодуру (Surya Koduru), Ноэль Фостер (Noel Foster), Дивья Пенметса (Divya Penmetsa), Срини Гаддам (Srini Gaddam) и Брайс Татт (Bryce Tutt); всем, кто почти вслепую следовал за мной в массовом движении Spark, всем людям из Veracity Solutions и моей новой команде в Advance Auto Parts.

Спасибо Мэри Паркер (Mary Parker) из отдела статистики (Department of Statistics) Техасского университета (University of Texas) в Остине (Austin) и Кристиане Страччиалана Парада (Cristiana Straccialana Parada). Их вклад помог уточнить и сделать более ясным содержимое некоторых разделов.

Я хочу поблагодарить все сообщество в целом, в том числе Джима Хьюджеса (Jim Hughes), Майкла Бен-Давида (Michael Ben-David), Марселя-Яна Крейгсмана (Marcel-Jan Krijgsman), Жана-Франсуа Морена (Jean-Francois Morin) и всех неизвестных участников, присылавших запросы на включение изменений в репозиторий GitHub. Также хочу выразить искреннюю благодарность сотрудникам Databricks, IBM, Netflix, Uber, Intel, Apple, Alluxio, Oracle, Microsoft, Cloudera, NVIDIA, Facebook, Google, Alibaba, многочисленных университетов и всем, кто сделал Spark таким, какой он есть. В частности, спасибо за работу, вдохновение и поддержку Холдену Карау (Holden Karau), Яцеку Ласковски (Jacek Laskowski), Шону Оуэну (Sean Owen), Матею Захариа (Matei Zaharia) и Жюлю Дамжи (Jules Damji).

Во время работы над этим проектом я участвовал в нескольких подкастах. Благодарю Тобиаса Мейси (Tobias Macey) за подкаст «Data Engineering Podcast» (http://mng.bz/WPjX), Эла Мартина (Al Martin) из IBM за подкаст «Making Data Simple» (http://mng.bz/8p7g) и за подкаст «Roaring Elephant» Джона Масскелейна (Jhon Masschelein) и Дэйва Рассела (Dave Russell) (http://mng.bz/EdRr).

В качестве члена команды IBM я был счастлив работать со многими сотрудниками IBM на протяжении всего этого проекта. Они помогали прямо, косвенно или оказывали вдохновляющее воздействие: Роб Томас Rob Thomas (мы должны работать вместе и в дальнейшем), Мариус Чиортеа (Marius Ciortea), Алберт Мартин (Albert Martin) (он, помимо всего прочего, выпускает великолепный подкаст «Make Data Simple»), Стив Мур (Steve Moore), Сурав Мазумдер (Sourav Mazumder), Стейси Ронэхэн (Stacey Ronaghan), Мей-Мей Фу (Mei-Mei Fu), Виджай Боммиреддипалли (Vijay Bommireddipalli) (сохраняй присущее тебе равновесие на всех этих холмах Сан-Франциско!), Сунита Камбампати (Sunitha Kambhampati), Сахдев Зала (Sahdev Zala) и мой брат Стюарт Лайтел (Stuart Litel).

Хочу поблагодарить сотрудников издательства Manning, которые приняли в работу этот сумасшедший проект. Как в хороших кинофильмах, перечисляю «в порядке их появления»: рецензент издательства Майкл Стивенс (Michael Stephens), издатель Марьян Бэйс (Marjan Bace), редакторы-консультанты Марина Майклз (Marina Michaels) и Тони Арритола (Toni Arritola); производственный персонал: Ирин Туохи (Erin Twohey), Ребекка Райнхарт (Rebecca Rinehart), Берт Бейтс (Bert Bates), Кэндэс Джиллхулли (Candace Gillhoolley), Радмила Эрцеговац (Radmila Ercegovac), Алекс Драгосавлевич (Aleks Dragosavljevic), Матко Хрватин (Matko Hrvatin), Кристофер Кауфманн (Christopher Kaufmann), Ана Ромак (Ana Romac), Шерил Вайсман (Cheryl Weisman), Лори Уэйдерт (Lori Weidert), Шэрон Уилки (Sharon Wilkey) и Мелоди Долаб (Melody Dolab).

Также хочу выразить признательность и поблагодарить всех рецензентов издательства Manning. Это Анупам Сенгупта (Anupam Sengupta), Арун Лаккакулам (Arun Lakkakulam), Кристиан Кройтцер-Бек (Christian Kreutzer-Beck), Кристофер Карделл (Christopher Kardell), Конор Редмонд (Conor Redmond), Эзра Шрёдер (Ezra Schroeder), Габор Ласло Хайба (Gábor László Hajba), Гэри А. Стаффорд (Gary A. Stafford), Джордж Томас (George Thomas), Джулиано Араухо Бертоти (Giuliano Araujo Bertoti), Игор Франка (Igor Franca), Игор Карп (Igor Karp), Йерун Бенкхейсен (Jeroen Benckhuijsen), Хуан Руфес (Juan Rufes), Келвин Джонсон (Kelvin Johnson), Келвин Роулз (Kelvin Rawls), Марио-Леандер Раймер (Mario-Leander Reimer), Маркус Бройер (Markus Breuer), Массимо Далла Ровере (Massimo Dalla Rovere), Паван Мадхира (Pavan Madhira), Самбаран Хазра (Sambaran Hazra), Шоба Айвер (Shobha Iyer), Убальдо Пескаторе (Ubaldo Pescatore), Виктор Дюран (Victor Durán) и Уильям И. Уилер (William E. Wheeler). С помощью всей этой компании удалось написать (надеюсь) хорошую книгу. Также хочу поблагодарить Петара Зечевича (Petar Zečević) и Марко Банаси (Marco Banaći), которые написали первую редакцию этой книги. Спасибо Томасу Локни (Thomas Lockney) за подробную техническую

рецензию, а также Рамбабу Поса (Rambabu Posa) за включение исходного кода в книгу. Спасибо Йону Риу (Jon Rioux) (merci, Йонатан!) за первоначальный вариант проекта PySpark in Action. Йон подал идею «создания группы Spark в издательстве Manning».

И еще раз спасибо Марине. Марина (Marina Michaels) была моим редактором-консультантом в течение большей части времени работы над книгой. Она всегда была рядом, когда я выдавал предварительные версии, когда нужны были ее советы, она была строга со мной (да, твой контроль никогда не ослабевал!), но ее участие было весьма важным в этом проекте. Я всегда буду помнить наши долгие споры об этой книге (вне зависимости от того, существовали или нет предлоги для разговоров о чем-то еще). Я буду скучать по тебе, старшая сестра (до того момента, когда начнется работа над другой книгой).

Наконец, хочу сказать спасибо своим родителям, которые поддерживали меня больше, чем я того заслуживал, а также тем, кому я посвятил эту книгу: моей жене Лиз, которая помогала всем, чем могла, на любом уровне, в том числе и в достижении взаимопонимания с редакторами, и нашим детям Пьеру-Николя, Джеку, Натаниелю и Руби, у которых я украл так много времени, работая над этой книгой.

О чем эта книга

Когда я начинал работу над этим проектом, который стал книгой «Spark в действии», второе издание, моими целями были:

- помощь сообществу Java в использовании Apache Spark с наглядной демонстрации того, что нет необходимости дополнительно изучать языки Scala и/или Python;
- описание главных концепций, лежащих в основе Apache Spark, инженерии (больших) данных и науке о данных, требующих только знаний о реляционных базах данных и основ языка SQL, и ничего больше;
- тщательное разъяснение того, что Spark – это операционная система, специально предназначенная для распределенных вычислений и анализа.

Я верю в метод обучения по любой теме из области информационных технологий с использованием большого количества примеров. Примеры в этой книге являются чрезвычайно важной частью процесса обучения. Я создавал примеры так, чтобы они были как можно ближе к ситуациям, возникающим в настоящей профессиональной деятельности. Предлагаемые здесь наборы данных взяты из реальных ситуаций, со всеми присущими им качественными недостатками. Это не идеализированные учебные наборы данных, которые «работают всегда». Именно поэтому, объединяя примеры и такие наборы данных, вы будете работать и учиться более практическим способом по сравнению со «стерилизованным» подходом к обучению. Я называю эти примеры лабораторными работами (labs) с надеждой, что они окажутся интересными для вас и вы захотите поэкспериментировать с ними.

В книге очень много иллюстраций. Основываясь на широко известном высказывании «Рисунок заменяет тысячу слов», я избавил вас от чтения 183 000 дополнительных слов.

Для кого предназначена эта книга

Трудно связать название профессии с названием книги, поэтому если ваша профессия называется инженер по обработке данных, ученый-ис-