**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

(национальный исследовательский университет)»

**Факультет (институт, филиал) №4 Кафедра 406**

**Направление подготовки Радиотехника Группа М4В-302Б**

**Квалификация (степень) \_\_\_\_\_Бакалавриат\_\_**

**РЕФЕРАТ**

На тему: Что такое DataOps?­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­

Реферат сдал \_Николаев Алексей Андреевич\_\_\_\_\_\_\_ \_\_\_\_\_ (\_\_\_\_\_\_\_\_\_\_\_\_)

(фамилия, имя, отчество)

Реферат принял Терехин Алексей Геннадиевич\_\_\_\_\_\_\_\_\_\_\_ (\_\_\_\_\_\_\_\_\_\_\_\_)

(фамилия, имя, отчество)

Москва 2020

**ОГЛАВЛЕНИЕ**

[Введение 3](#_Toc59103447)

[1 Понятие и принципы DataOps 4](#_Toc59103448)

[2 Формирование команды DataOps 9](#_Toc59103449)

[Заключение 13](#_Toc59103450)

[Список использованных источников 14](#_Toc59103451)

ВВЕДЕНИЕ

Объём данных, которые генерируются и собираются компаниями разных мастей, продолжает расти. Предприятия, которые хотят ускорить свои сквозные процессы и улучшить понимание бизнеса, не могут продолжать использовать ручные процессы управления данными, на которые они полагались в течение десятилетий. И в ближайшие годы ситуация будет только усугубляться.

DataOps — это относительно новая и пока ещё развивающаяся дисциплина, возникшая около пяти лет назад. Она построена на фундаменте, созданном гибкой разработкой и DevOps. Цель состоит в том, чтобы привнести аналогичные принципы в аналитику данных и науку о данных, чтобы улучшить качество данных и сократить время, необходимое для получения качественной бизнес-аналитики.

В течение следующих нескольких лет предприятия, которые сопротивляются внедрению DataOps, будут тратить ещё больше времени на реагирование на ошибки в данных и в ручных процессах, и еще больше отставать в своей способности предоставлять своевременную и точную информацию бизнесу. В то же время те, кто использует DataOps, создадут оптимизированные автоматизированные конвейеры данных, которые позволят администраторам данных и учёным-исследователям оптимизировать бизнес-процессы, сосредоточиться на более важных задачах и поддерживать принятие эффективных решений.

Таким образом, целью данной работы выступает исследование понятия DataOps.

В соответствии с поставленной целью в работе предстоит решить задачи:

- описать понятие и принципы DataOps;

- охарактеризовать формирование команды DataOps;

- сделать выводы по работе в целом.

1 ПОНЯТИЕ И ПРИНЦИПЫ DATAOPS

DataOps (DATA Operations, датаопс), по аналогии с DevOps (DEVelopment Operations, девопс) — это концепция и набор практик непрерывной интеграции данных между процессами, командами и системами для повышения эффективности корпоративного управления или отраслевого взаимодействия за счет распределенного сбора, централизованной аналитики и гибкой политики доступа к информации с учетом ее конфиденциальности, ограничений на использование и соблюдения целостности.

Впервые термин DataOps (от Data Operations) прозвучал в 2015 году в блоге Энди Палмера, сооснователя компании Tamr, примечательной тем, что другим ее сооснователем был Майкл Стоунбрейкер. В марте 2017 года на конференции Strata+Hadoop World, организуемой O'Reilly, о DataOps заговорили снова, а затем термин был зафиксирован в книге, авторы которой четыре года проработали в команде Facebook Data Service Team над проектом Apache Hive. Движение SQL для Hadoop, начатое Hive, было главным образом нацелено на предоставление аналитикам удобных средств демократизации больших данных. Несмотря на то что на рынке уже имеется пул инструментов обработки больших данных, Hadoop был и до сих пор остается трудным даже для инженеров по данным, не говоря уже о бизнесе, который не желает тратить время на его освоение, а в результате в цепочке «потоки гетерогенных данных — инженеры по данным — конечные пользователи» образуется «бутылочное горло». Однако одного только средства доступа к данным еще недостаточно для устранения узкого места — требуется обеспечить сбор и подготовку данных, управлять эксплуатацией и своевременно реагировать на рост и падение нагрузок и объемов[[1]](#footnote-1).

На основе происхождения и целевой направленности DataOps и DevOps можно выявить несколько характеристик, которые их объединяют:

1. Сокращение сроков разработки и поставки готового продукта (программного приложения или актуальных данных) за счет принципов [Agile](https://www.bigdataschool.ru/wiki/agile" \t "_blank), в частности, [гибкости и самоорганизации](https://www.bigdataschool.ru/bigdata/%D0%BF%D1%80%D0%B8%D0%BD%D1%86%D0%B8%D0%BF%D1%8B-agile-%D0%B2-big-data.html).
2. Непрерывность процессов интеграции (Continuous Integration) и развертывания (Continuous Deployment).
3. Автоматизация процессов тестирования, развертывания и мониторинга с помощью технологий контейнеризации и виртуализации (Jenkins, Docker, Rocket, [Kubernetes](https://www.bigdataschool.ru/wiki/kubernetes" \t "_blank) и т.д.).
4. [T-образная модель компетенций](https://www.bigdataschool.ru/bigdata/devops-data-science-agile-big-data.html), когда необходимо иметь широчайший технический кругозор, будучи экспертом в конкретной прикладной области.
5. Работа с инфраструктурными решениями для Big Data (локальными кластерами [Hadoop](https://www.bigdataschool.ru/wiki/hadoop" \t "_blank) и [облачными платформами](https://www.bigdataschool.ru/bigdata/hadoop-%D0%B2-%D0%BE%D0%B1%D0%BB%D0%B0%D0%BA%D0%B0%D1%85.html)), включая подготовку, настройку и [администрирование](https://www.bigdataschool.ru/bigdata/tooltips-4-big-data-administrator.html) сред развертывания.

Итак, DataOps и [DevOps](https://www.bigdataschool.ru/wiki/devops" \t "_blank) являются практическим продолжением принципов [Agile](https://www.bigdataschool.ru/wiki/agile" \t "_blank) и нужны для повышения эффективности ИТ-процессов, а также цифровой трансформации бизнеса. Однако, при схожести многих моментов, их локальные цели и, соответственно, средства достижения, отличаются.

Если [DevOps](https://www.bigdataschool.ru/wiki/devops" \t "_blank) концентрируется на быстрой и непрерывной поставке работающего программного обеспечения, то DataOps сосредоточен на демократизации и актуализации данных, а также возможностях оперативного и безопасного доступа к ним.

Поэтому процессы, в которых задействованы DataOps- и [DevOps](https://www.bigdataschool.ru/wiki/devops)-инженеры, тоже отличаются: [DevOps](https://www.bigdataschool.ru/wiki/devops" \t "_blank) -процесс имеет меньше шагов и больше концентрируется на непрерывности цепочки разработка-тестирование-развертывание. В DataOps  используется гораздо большее количество инструментов, т.к. приходится иметь дело с множеством источников и моделей данных, управляя информационными потоками (оркестрация, [orc](https://www.bigdataschool.ru/wiki/orc" \t "_blank)hestrate), корпоративными озерами ([Data Lake](https://www.bigdataschool.ru/wiki/data-lake" \t "_blank)s) и «песочницами» данных (Sandbox Management), формируя воспроизводимые среды для работы с [Big Data](https://www.bigdataschool.ru/wiki/big-data" \t "_blank)[[2]](#footnote-2).

Если главной целью [DevOps](https://www.bigdataschool.ru/wiki/devops" \t "_blank) считается быстрая и непрерывная поставка бизнесу работающего программного обеспечения [2], то DataOps предполагает оперативное и безбарьерное предоставление актуальных и рабочих данных каждому участнику корпоративных процессов. Это означает устранение когнитивных, временных и организационных разрывов между исследователями данных (data scientist’ы), бизнес-аналитиками, разработчиками, руководителями и пользователями [Big Data](https://www.bigdataschool.ru/wiki/big-data" \t "_blank).

Такая демократизация данных повышает скорость реакции на любые изменения, что весьма актуально для современного бизнеса и соответствует принципам [Agile](https://www.bigdataschool.ru/wiki/agile" \t "_blank)[[3]](#footnote-3).

Архитектура DataOps должна быть в высокой степени адаптивной, поскольку требования к данным и способы их использования меняются быстро и непрерывно. У ваших потребителей данных — аналитиков, исследователей и руководителей — по мере изменения их бизнес-приоритетов и конъюнктуры рынка возникают разнообразные новые потребности. Адаптивная архитектура принимает эти изменения и подстраивается под них, позволяя улучшить потоки данных и качество полезной информации на каждом этапе.

Успешная архитектура DataOps требует совместной работы в масштабах всей компании и поддерживает такое сотрудничество. Потребители данных извлекают данные и получают информацию для своих бизнес-инициатив, и им нужна возможность быстро создавать и структурировать свои данные и конвейеры, через которые эти данные проходят. Архитектура должна максимально упрощать эти операции с данными, способствуя адаптации и повышению эффективности бизнеса.

 Инфраструктура для поддержания бесперебойной работы конвейера данных цифрового предприятия должна обеспечивать выполнение операций нескольких категорий.

Оркестровка конвейера данных. Для формирования потоков данных требуется маршрутная карта с описанием всех источников данных, моделей их представления и интеграции, а также шагов процесса анализа. Для этого могут использоваться следующие инструменты: Apache Oozie — планировщик процессов заданий Apache Hadoop; BMC Control-M — решение по автоматизации пакетной обработки; DataKitchen — платформа DataOps поддержки всего цикла аналитической обработки, минимизирующая время на подготовку и доставку данных нужного качества; Reflow — система инкрементальной обработки данных в облаке с помощью произвольных программ, упакованных в контейнеры Docker.

Тестирование и обеспечение качества. В DataOps важно обеспечить автоматическую проверку качества данных, их очистку на всех этапах обработки. Возможные инструменты: ICEDQ — ПО для автоматизации тестирования при работе с хранилищами ETL и миграции данных; Naveego — облачная платформа для построения информационных панелей и витрин с целью мониторинга состояния данных и управления исключениями.

Автоматическое распределение. В DataOps осуществляется непрерывное перемещение кода и конфигураций между этапами разработки приложений, сбора данных, стадиями анализа и извлечения из данных полезных сведений. Типичным инструментом здесь может быть Jenkins, осуществляющий непрерывную доставку ПО с автоматическим контролем всех этапов жизненного цикла приложений от написания кода к сборке, автотестированию и развертыванию в эксплуатационных средах.

Развертывание моделей данных и управление «песочницами». Команды по данным, работающие в соответствии с DataOps, формируют воспроизводимые во всех подразделениях компании операционные среды. Инструменты: Domino — ускорение процессов разработки приложений работы с данными, доставка моделей, бесшовная интеграция; Open Data Group — программное решение по развертыванию систем аналитики на основе моделей; DSFlow — ускорение процессов извлечения данных для бизнеса.

Виртуализация данных и управление тестовыми данными. Инструменты: Delphix — платформа виртуализации, защиты и управления данными; Redgate — SQL-инструменты, помогающие внедрять DataOps, управлять производительностью баз и подключать новые базы.

Интеграция и унификация данных. Tamr — унифицированное решение для работы с корпоративными базами данных с привлечением методов машинного обучения; Switchboard Software — аутсорсинг и интеграция данных.

Управление производительностью и облачные платформы. Инструменты: SelectStar — мониторинг баз данных; Unravel — управление производительностью и работой с приложениями и платформами больших данных; MapR — конвергентная платформа работы с большими данными, объединяющая средства аналитики реального времени и операционные бизнес-приложения; Quobole — облачный сервис класса Вig Data as a Service для работы с разнородными структурированными и неструктурированными данными[[4]](#footnote-4).

# **2 ФОРМИРОВАНИЕ КОМАНДЫ DATAOPS**

DataOps — подход, направленный на поддержку процессов машинного обучения. Например, исследователи данных, работая по схеме DataOps, могут более продуктивно взаимодействовать с инженерами ПО в процессе передачи отделу эксплуатации моделей для развертывания. Но DataOps не ограничивается машинным обучением. Этот подход полезен для любых работ, связанных с анализом данных, так как он облегчает использование преимуществ, предоставляемых распределенными платформами управления данными.

Кроме того, DataOps хорошо сочетается с архитектурами на основе микросервисов.

Новые технологии требуют от предприятий перемен, способностей работать с большими объемами данных и оперативно реагировать на события.

Традиционные, изолированные друг от друга отделы — слишком негибкие и медлительные для организаций, осваивающих работу с Большими Данными и проходящих через цифровую трансформацию. Именно в таких ситуациях способны помочь принципы DataOps.

Методология DevOps сближает разработчиков и специалистов по эксплуатации ПО, помогает лучше согласовать разработку с целями бизнеса, сократить циклы выпуска и увеличить частоту развертываний. Главная характерная черта DevOps — кросс-функциональные группы, включающие специалистов по эксплуатации, программной инженерии, разработке архитектуры, планированию и управлению продуктами. DataOps добавляет к этому перечню исследователей данных и специалистов по их подготовке, дополнительно улучшая взаимодействие между соответствующими участниками.

чтобы группа DataOps стала максимально эффективной, в ней должны быть специалисты по исследованию данных.

От традиционного принципа организации работы, когда аналитики изолированы от разработчиков, необходимо отказываться: реальная координация возможна лишь при тесном взаимодействии, вплоть до совместного выслушивания жалоб заказчиков. Исследователи данных должны проходить те же процедуры отбора в команду, что и разработчики, вместе с ними рассматривать комментарии пользователей и рекомендовать решения проблем.

При этом не требуется, чтобы специалисты в области Data Science работали в группе DataOps на постоянной основе: обычно исследователя данных включают в команду на определенное время, впоследствии он может перейти в другую команду, а его роль в группе берет на себя другой участник, например, инженер по подготовке данных, иными словами, происходит постоянная ротация.

Для создания команды DataOps не обязательно нанимать новых специалистов. Во многих организациях уже есть костяк группы DataOps — участники имеющихся команд DevOps. Следующий этап — выбрать проекты, сопряженные с разработкой механизмов анализа больших объемов данных, и найти специалиста, имеющего опыт исследования данных. Это может быть инженер по подготовке данных, а не обязательно специалист в области Data Science.

В крупномасштабных проектах какие-то роли DataOps могут быть распределены между несколькими участниками. Часто встречается и ситуация, когда один специалист берет на себя более одной роли. Например, член команды может иметь и навыки эксплуатации, и навыки программной инженерии, а участники с опытом разработки могут дополнительно иметь квалификацию инженеров по подготовке данных. Часто соответствующий опыт есть и у исследователей данных. Реже всего встречается совмещение навыков исследования данных и эксплуатации ПО[[5]](#footnote-5).

Когда речь идет об инженерных командах, достаточно бывает четко обозначить цели, после чего, видя общую цель, группа самоорганизуется для решения поставленной задачи. В командах DataOps ситуация сложнее, поскольку разные участники рассматривают задачу с разных точек зрения. Например, специалисты по эксплуатации будут беспокоиться о надежности, о том, чтобы отклик обязательно поступал в течение определенного времени. Исследователи данных в свою очередь фокусируются на обеспечении точности поступающих ответов. Но когда все участники стремятся решить общую задачу и готовы к компромиссам, выстроить необходимую организационную структуру будет достаточно просто.

В общем случае в команду работы с данными должны входить:

- аналитик данных (Data Analyst) — специалист по данным, работающий в каждом линейном бизнес- или операционном подразделении компании;

- инженер данных (Data Engineer) — специалист, отвечающий за поставку данных в платформу, из которой их черпают исследователи данных и аналитики;

- исследователь данных (Data Scientist) — специалист по статистике и машинному обучению, предлагающий модели для изучения данных в соответствии с целями бизнеса;

- директор по данным (Chief Data Officer) — человек, контролирующий работу команды по данным, непосредственно подчиняющийся генеральному или техническому директору компании.

Кроме того, применительно к DataOps, по крайней мере на первых порах, в компании должен быть еще инженер по эксплуатации (DataOps Engineer), отвечающий за непосредственное применение технических средств Agile и DevOps ко всему процессу работы с данными. Его основная задача — устранение барьеров между операционной деятельностью и аналитикой. Он же на первых порах может занять место, традиционно отводимое в командах DevOps инженеру по качеству (Quality Engineer), создавая инфраструктуру для автотестирования, через которую должны проходить все новые наработки, и обеспечивая таким образом «неразрушающий контроль», а также постепенно включая в культуру непрерывного управления качеством всех остальных участников команды по данным[[6]](#footnote-6).

В целом компании ожидают от практики DataOps несколько весомых преимуществ:

* улучшение сотрудничества между различными командами;
* возможность быстро и эффективно реагировать на новые запросы;
* обеспечение более качественной работы и поддержки;
* мониторинг достижения целей компании в режиме реального времени;
* возможность избежать аварий и неполадок, заранее прогнозируя их с помощью анализа данных;
* повышение эффективности работы благодаря статистическому контролю процессов;
* сокращение времени исправления ошибок[[7]](#footnote-7).

# **ЗАКЛЮЧЕНИЕ**

Таким образом, в заключении следует сделать выводы по рассмотренной теме.

DataOps (акроним от data и operations) — это новая методология, предполагающая совместную работу объединенных команд разработки и эксплуатации ИТ-систем (DevOps), инженеров по подготовке данных (data engineer) и исследователей данных (data scientist) для создания инструментов, процессов и организационных структур, необходимых предприятию, опирающемуся в работе на данные.

Подобно DevOps, DataOps опирается на принципы скорой (agile) разработки. В рамках DataOps приоритетами являются непрерывная доставка аналитических знаний и ориентация на удовлетворенность клиента.

Группы DataOps выдают аналитику, имеющую практическую ценность, которая измеряется пользой от предоставляемых знаний. Группы DataOps готовы работать в условиях постоянного изменения потребностей клиента и всегда стремятся их учитывать. Такие группы обычно представляют собой коллективы, самоорганизующиеся для достижения определенных целей и отдающие предпочтение стабильности и масштабируемости команды и процессов, а не «трудовым рекордам».

Одна из главных задач группы DataOps — обеспечивать оркестровку данных, инструментов, кода и сред в масштабах всего проекта. Аналитические конвейеры работают по аналогии с линиями бережливого производства; при этом для DataOps большое значение имеет воспроизводимость результатов.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Белов В.В. Проектирование информационных систем: Учебник / В.В. Белов. - М.: Академия, 2018. - 144 c.
2. Вичугова А. Что такое DataOps: зачем Big Data свой DevOps с блокчейном и данными // Школа больших данных. – 2019. - [Электронный ресурс]. – Режим доступа: <https://www.bigdataschool.ru/blog/dataops-devops-4-big-data.html>
3. Волков Д., Николаенко А. DataOps: данные в стиле Agile // OSP – Гид по технологиям цифровой трансформации. 2019. - [Электронный ресурс]. – Режим доступа: <https://www.osp.ru/os/2018/2/13054175>
4. Ипатова Э. Р. Методологии и технологии системного проектирования информационных систем / Э.Р. Ипатова, Ю.В. Ипатов. - М.: Флинта, 2017. - 256 c.
5. Рафалович В. Data mining, или интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович. - М.: SmartBook, 2018. - 352 c.

1. Волков Д., Николаенко А. DataOps: данные в стиле Agile // OSP – Гид по технологиям цифровой трансформации. [Электронный ресурс]. – Режим доступа: <https://www.osp.ru/os/2018/2/13054175>. [↑](#footnote-ref-1)
2. Белов В.В. Проектирование информационных систем: Учебник / В.В. Белов. - М.: Академия, 2018. – С. 68. [↑](#footnote-ref-2)
3. Вичугова А. Что такое DataOps: зачем Big Data свой DevOps с блокчейном и данными // Школа больших данных. – 2019. - [Электронный ресурс]. – Режим доступа: <https://www.bigdataschool.ru/blog/dataops-devops-4-big-data.html>. [↑](#footnote-ref-3)
4. Волков Д., Николаенко А. DataOps: данные в стиле Agile // OSP – Гид по технологиям цифровой трансформации. - [Электронный ресурс]. – Режим доступа: <https://www.osp.ru/os/2018/2/13054175>. [↑](#footnote-ref-4)
5. Ипатова Э. Р. Методологии и технологии системного проектирования информационных систем / Э.Р. Ипатова, Ю.В. Ипатов. - М.: Флинта, 2017. – С. 142. [↑](#footnote-ref-5)
6. Волков Д., Николаенко А. DataOps: данные в стиле Agile // OSP – Гид по технологиям цифровой трансформации. - [Электронный ресурс]. – Режим доступа: <https://www.osp.ru/os/2018/2/13054175>. [↑](#footnote-ref-6)
7. Рафалович В. Data mining, или интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович. - М.: SmartBook, 2018. – С. 192. [↑](#footnote-ref-7)