SRE — обеспечение надёжности систем

01 Кому подойдёт курс

DevOps-инженерам и инженерам по эксплуатации
Освоите принципы SRE и станете главным экспертом по надёжности в вашей команде

Системным администраторам и архитекторам ПО
Сделаете инфраструктуру предсказуемой и сократите число «необъяснимых» сбоев

Разработчикам
Снизите отток пользователей за счёт стабильности, а не только новых фич

02 Чему научитесь на курсе

Управлять надёжностью, основываясь на данных

- Мыслить как SRE, перейдя от интуиции к точным метрикам
- Принимать решения, которые балансируют скорость разработки и стабильность продукта
- Внедрять SLO/SLI — это общий язык для команды и бизнеса, чтобы все понимали, что значит «система работает хорошо»

Быстро восстанавливать системы без хаоса

- Создавать инфраструктуру как код (IaC) для безопасных и предсказуемых развёртываний
- Настраивать мониторинг, который предупреждает о реальных проблемах, а не захламляет систему уведомлениями
- Организовывать чёткие процессы, чтобы быстро находить и устранять сбои, а не тушить «пожары» вручную

Оптимизировать затраты и повышать производительность

- Превращать инциденты в возможности для роста: от обнаружения до postmortem — без поиска виноватых, только полезные выводы
- Анализировать производительность систем и внедрять решения, которые сокращают расходы и ускоряют работу

SRE — обеспечение надёжности систем

03 Как проходит обучение

- Сопровождение кураторами
- Обратная связь от опытных наставников
- Воркшопы с экспертами
- Теория на платформе Практикума
- Практические задания с ревью на готовой инфраструктуре в облаке

Что вас ждёт на обучении

Полный цикл SRE: от быстрого обнаружения до устранения и анализа «пожара»

2 практических воркшопа по расследованию инцидента, созданию postmortem и просчёту времени простоя

Диплом о профессиональной переподготовке

SRE — обеспечение надёжности систем

4 месяца

продолжительность курса

1 НЕДЕЛЯ

01

Введение в SRE

- История возникновения SRE
- Отличие SRE от DevOps
- SLI / SLO / SLA
- Golden Signals
- Error Budget

2 НЕДЕЛИ

02

Симптомы «пожара»

- Способы получения информации об инцидентах
- Кросс-системные метрики
- Анализ и разбор инцидентов
- SLI/SLO/SLA
- Технические метрики
- Бизнес-метрики
- Golden Signals и RED

2 НЕДЕЛИ

03

Как узнать о «пожаре»: наблюдаемость

- Логи, метрики и трейсы
- Визуализация логов, метрик и трейсов в Grafana

2 НЕДЕЛИ

04

Как узнать о «пожаре»: метрики

- Корреляция метрик, логов и трейсов
- Алерты
- Типы метрик
- Функции агрегации метрик
- Дашборды

2 НЕДЕЛИ

05

Что делать во время «пожара»

- Инциденты
- Действия при инциденте
- Эскалация
- Best practice
- Коммуникации во время инцидента
- Траблшутинг

2 НЕДЕЛИ

06

Что делать после «пожара»

- Postmortem
- Blameless-культура
- Экшн-планы
- Ретроспектива действий на инциденте
- Сбор информации и анализ инцидента
- Исследование проблемы
- Подсчёт убытков от «пожара»

2 НЕДЕЛИ

07

Как снизить риски будущих «пожаров»: отказоустойчивость

- RTO/RPO
- Лучшие практики для увеличения доступности приложения
- Лучшие практики по организации отказоустойчивости
- Лучшие практики для повышения отказоустойчивости в Kubernetes

2 НЕДЕЛИ

08

Как снизить риски будущих «пожаров»: надёжность

- Disaster Recovery Plan
- Тестирование отказоустойчивости инфраструктуры
- Техдолг
- Требования к отказоустойчивости

2 НЕДЕЛИ

09

Как снизить время устранения «пожара»

- Дежурства
- Передача знаний и опыта
- Автоматизация реакции на алерты
- Шум в алертах

Введение в SRE 01

1 неделя

- Возникновение SRE в Google, исходные причины и проблемы, которые решает методология. SRE Books. Концепции, подход, культура, роли и навыки специалиста по SRE.
- Отличие SRE от DevOps: узнаете, в чём между ними разница. DevOps — это набор практик и изменений в культуре, а SRE — это вариант имплементации.
- SLI / SLO / SLA: познакомитесь с терминами, их назначением, принципами и задействованными лицами.
- Golden Signals: история возникновения, что означают, почему они важны и как их использовать.
- Error Budget: подход, принцип, использование, важность, идеализированный вариант в сравнении с реальностью.

Инструменты и технологии

- SLI
- SLO
- SLA
- Error Budget

Практика

Пройдёте тесты и квизы, чтобы закрепить свои знания о практиках и подходах SRE

Симптомы «пожара» 02

2 недели
1 проект

- Способы получения информации об инцидентах: взаимодействие с пользователями, сбор клиентских ошибок и серверных метрик, анализ логов, метрики, основанные на логах, синтетические тесты, тесты на продакшне, real user monitoring.
- Кросс-системные метрики: важность, мониторинг интеграций между системами и бизнес-метриками, мониторинг продукта в целом и каждого компонента.
- Анализ и разбор инцидентов: предиктивный анализ, разбор различных ситуаций, последствия бесконечных обновлений без backoff-тайм-аутов, которые увеличивают размер очереди. Расчёт количества реплик и допустимого количества «умерших».
- SLI/SLO/SLA: способы и важность правильного определения SLI, хорошие и плохие примеры, советы. Отличие SLO от SLA, способы определения значения SLO и его подсчёта. Описание дозволенного времени даунтайма при различных значениях SLO и влияние количества девяток на стоимость инфраструктуры. Целесообразность большого количества девяток.
- SLO: зависимость SLO компонента от SLO вышестоящих и нижестоящих компонентов, подсчёт суммарного SLO продукта.
- Технические метрики: описание распространённых технических метрик для бэкендов, фронтендов, баз, очередей, веб-серверов и балансировщиков. Примеры использования и аргументация, почему нужны именно эти метрики.
- Бизнес-метрики: важность бизнес-метрик и их отличие от технических метрик.
- Golden Signals и RED: различия, применимость и примеры использования.

Инструменты и технологии

- SLI
- SLO
- SLA
- Golden Signals
- RED

Практика

Определите SLI и максимальный SLO с учётом SLO описанных зависимостей приложения. Посчитаете допустимое время даунтайма приложения

Как узнать о «пожаре»: наблюдаемость 03

2 недели
1 проект

- Логи, метрики и трейсы: виды систем хранения, агенты для сбора, их различия и особенности. Разбор примера использования, установка и настройка, важность структуры данных, влияние на производительность и стоимость, риски и предостережения, советы по внедрению.
- Визуализация логов, метрик и трейсов в Grafana: установка и настройка Grafana, настройка датасорсов и демонстрация собранных логов/метрик/трейсов.

Инструменты и технологии

- Grafana
- Loki
- Mimir
- Prometheus
- Tempo
- OpenTelemetry

Практика

Настроите локальный экземпляр Grafana Loki, отправите тестовые логи через API Loki и визуализируете их в Grafana

Как узнать о «пожаре»: метрики 04

2 недели
1 проект

- Корреляция метрик, логов и трейсов: почему это важно и чем полезно, как настроить и какие нужны пререквизиты для этого, как использовать.
- Алерты: назначение, трешхолды, severity алертов, их критичность, разница между severity и критичностью. Обзор систем доставки алертов. Пример конфигурации Alertmanager, интеграция с источниками алертов и способом доставки: при помощи почты, мессенджеров или СМС.
- Типы метрик: гайд по типам метрик (counter, gauge, histogram и другим). Различия, применимость, примеры разных типов метрик из популярных опенсорс-продуктов, их особенности, визуализация показателей.
- Функции агрегации метрик: в чём их полезность и важность, зачем эти функции нужны и какие есть особенности, лучшие практики. Влияние функций агрегации метрик на производительность систем хранения метрик и оптимизацию сложных запросов.
- Дашборды: подходы к созданию дашбордов, импорт и экспорт готовых дашбордов. Конфигурации Grafana. Отображение различных типов метрик в разных вариантах визуализации. Виджеты, влияние сложных комплексных дашбордов на производительность систем хранения. Лучшие практики по созданию и оптимизации дашбордов.

Инструменты и технологии

- Loki
- Mimir
- Tempo
- Grafana
- AlertManager

Практика

Развернёте локально экземпляр Prometheus или Grafana Mimir и запустите тестовое приложение. Соберёте метрики в Prometheus или Grafana Mimir, визуализируете их в дашборде Grafana и настроите тестовый алерт на стороне Prometheus

Что делать во время «пожара» 05

2 недели
1 практикоориентированная ролевая игра

- Инциденты: что считать инцидентом, а что нет. Различия между обычным алертом и реальным инцидентом, их влияние на бизнес.
- Действия при инциденте: как себя вести при появлении инцидента, что делать в первую очередь, что отложить на потом. Разбор ситуации при классическом подходе и сравнение с SRE-подходом. Разные роли во время инцидентов, зоны ответственности этих ролей, их скрипт поведения и порядок действий.
- Эскалация: важность эскалационной модели, наличие эскалации в рабочее и нерабочее время, рассказ про инструменты, пример установки и настройки. Описание условных уровней (L1/L2/L3) on-call-инженера, дежурства, ротации, подмены.
- Best practice: важность хронологии событий, кто, что и когда сделал, что получилось. Рассказ про war-румы, лидерство в инциденте и назначении ролей, описание порядка действий.
- Коммуникации во время инцидента: внутри команды, между отделами, внутри компании и с клиентами. Различные способы коммуникации и Statuspage.
- Траблшутинг: как решать проблемы, с чего начинать, на что обращать внимание. Разбор примера инцидента и поиска проблемы. Митигация проблемы, отличие от полноценного решения, цель и способы.

Инструменты и технологии

- Alertmanager
- GoAlert
- Loki
- Mimir
- Tempo
- Grafana

Практика

Поучаствуете в ролевой игре по расследованию инцидента

Что делать после «пожара» 06

2 недели
1 практикоориентированная ролевая игра

- Postmortem: цель, важность и суть процедуры, структура, назначение и алгоритм использования. Лучшие практики по написанию.
- Blameless-культура: почему важна, какие проблемы решает. Описание подхода, примеры плохих и хороших ситуаций.
- Экшн-планы: зачем нужны планы действий, почему после инцидента надо их писать и искать конкретных исполнителей, ставить дедлайны и проверять статус.
- Ретроспектива действий на инциденте: как разобрать, что было хорошо или плохо, что нужно исправить в управлении инцидентами на будущее.
- Сбор информации и анализ инцидента: почему важно собирать информацию по горячим следам, даже если инцидент случился ночью. Первичный анализ инцидента, определение списка возможных первопричин.
- Исследование проблемы: сбор дополнительных данных, метрик и логов, проверка гипотез, тестирование, воспроизведение на непродакшн-окружениях.
- Подсчёт убытков от «пожара»: узнаете, почему важно определять размер потерь бизнеса за время простоя, какие есть способы их подсчёта, рекомендации. Трекинг потерь на повторяющихся инцидентах с аналогичными первопричинами.

Инструменты и технологии

- GitLab
- Postmortem

Практика

Поучаствуете в ролевой игре по созданию postmortem и подсчёту потерь за время простоя

Как снизить риски будущих «пожаров»: отказоустойчивость 07

2 недели
1 проект

- RTO/RPO: чем они различаются, зачем нужны, где используются и на что влияют. Способы определения и подсчёта, влияние на SLO/SLA. Лучшие практики по выбору RTO/RPO для разных типов данных. Зависимость RTO/RPO от бизнеса и законов. Влияние RTO/RPO на стоимость инфраструтктуры и продукта. Лучшие практики при удовлетворении различных RTO/RPO.
- Лучшие практики для увеличения доступности приложения: резервные реплики, отказоустойчивость внутри одного дата-центра и внутри нескольких, в одном регионе или в нескольких, на уровне нескольких облаков. Важность удалённости дата-центров друг от друга и от пользователя. Квоты, лимиты разных провайдеров, советы по организации отказоустойчивости для распределённых систем. Хранение данных, кеши, CDN.
- Лучшие практики по организации отказоустойчивости: описание топологий инфраструктуры, балансировка трафика, управление трафиком, blue/green, canary, балансировка сессий, описание балансировщиков L4/L7 и их различий. Лучшие практики по настройке на примере опенсорсных решений.
- Лучшие практики для повышения отказоустойчивости в Kubernetes: Replicas, HPA, VPA, requests/limits, PDB, node afinity, pod afinity, taints/tolerations. Мультизональные Kubernetes-кластеры. Возможные схемы реализации на уровне нескольких Kubernetes-кластеров в разных зонах доступности.

Инструменты и технологии

- Kubernetes
- GitLab
- Terraform
- Ansible
- DNS
- nginx

Практика

Определите RTO/RPO, предложите свой вариант развёртывания и примените лучшие практики для него

Как снизить риски будущих «пожаров»: надёжность

08

2 недели
1 проект

- Disaster Recovery Plan: что это и зачем он нужен, от чего спасает и как приводится в действие. Зависимость плана от RTO/RPO. Схема описания, документация, принятие решения о введении в действие. Ручная или автоматическая активация.
- Тестирование отказоустойчивости инфраструктуры: ручное согласованное выведение из строя различных компонентов на продакшн. Исследование результатов, влияние на бизнес, проверка работоспособности DRP и достижения значения RTO/RPO. Автоматизированное согласованное и несогласованное тестирование отказоустойчивости, хаос-инжиниринг. Теория и применимость, примеры инструментов и реализации.
- Техдолг: в приложениях, инфраструктуре и конфигурациях. Влияние техдолга на отказоустойчивость, доступность систем и потери компании. Необходимость постоянной работы по устранению техдолга.
- Требования к отказоустойчивости: внедрение требований к коду, чтобы обеспечить необходимую отказоустойчивость. Внедрение скриптов проверки, формирование требований, практика Quality Gates, которым должны удовлетворять архитектура приложений и инфраструктура.

Инструменты и технологии

- План реакции на инциденты
- Методы продакшн-тестирования

Практика

Составите план реакции на инциденты, примените методы продакшн-тестирования

Как снизить время устранения «пожара»

09

2 недели
1 проект

- Дежурства: организация, их важность и цель. Зоны ответственности, учёт часовых поясов пользователей и членов команды, учёт норм рабочего времени. Лучшие практики по подсчёту необходимого количества людей для дежурства.
- Передача знаний и опыта: ведение документации, передача знаний, ротация в команде, обучающие тренинги. Подготовка Run-Books, где описан чёткий порядок действий при возникновении типовой проблемы.
- Автоматизация реакции на алерты: написание скриптов для устранения типовых проблем. Возможные инструменты и схемы реализации.
- Шум в алертах: наличие и влияние на команду, причины возникновения, советы по уменьшению шума, примеры различных видов шума в алертах. Флапающие алерты и работа с ними, реакция на них. False-positive-алерты: что это такое, почему возникают, как влияют на команду и как с ними бороться.

Инструменты и технологии

- Alertmanager
- GoAlert
- Terraform
- Ansible

Практика

Предложите решения для флапающего и false-positive-алерта, а также для алерта о заканчивающемся месте на диске.
Для предложенных алертов определите критичность и severity

Медведев Кирилл
Директор АНО ДПО «Образовательные технологии Яндекса»