



Урок 14 (бонусный) Spark и Kubernetes

Запускаем Spark в Kubernetes с помощью Spark Operator

Александр Волынский

Technical Product Manager

ML Platform

VK Cloud Solutions

Для кого

- В первую очередь будет интересно для Data Engineer, Data Scientist
- Для тех, кто разворачивает и поддерживает инфраструктуру для DE и DS команд
- Для всех, кто интересуется большими данными и ролью Kubernetes при работе с данными



Александр Волынский

Technical Product Manager
ML Platform, VK Cloud Solutions

- Архитектор VK Cloud Solutions
- Специалист по Big Data
- Участвовал в создании хранилищ данных в «Платформе ОФД», Mail.ru Group (теперь VK), X5 Group и других компаниях
- Энтузиаст использования Kubernetes для построения Data Lake, DWH и ML-платформ в облаках

Коротко о Spark

- Apache Spark - is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters
- Умеет работать с Yarn, Mesos, Kubernetes и в Standalone режиме
- Пришел на смену Hadoop MapReduce
- Чаще всего идет в комплекте с Hadoop кластером

Spark и Kubernetes

- Spark можно запускать в Kubernetes, начиная с версии 2.3 (2018). Production-ready в версии 3.1
- Есть два способа запуска: spark-submit и Kubernetes Operator for Spark
- Kubernetes Operator for Spark – Kubernetes native way
(<https://www.lightbend.com/blog/how-to-manage-monitor-spark-on-kubernetes-introduction-spark-submit-kubernetes-operator>)
- Spark operator позволяет решить проблемы с доступом к логам, получением текущего статуса и состояния джобы
(<https://github.com/GoogleCloudPlatform/spark-on-k8s-operator/blob/master/docs/quick-start-guide.md>)

Почему стоит запускать Spark в Kubernetes



Изоляция сред (контейнеризация и dependency management)



Управление ресурсами



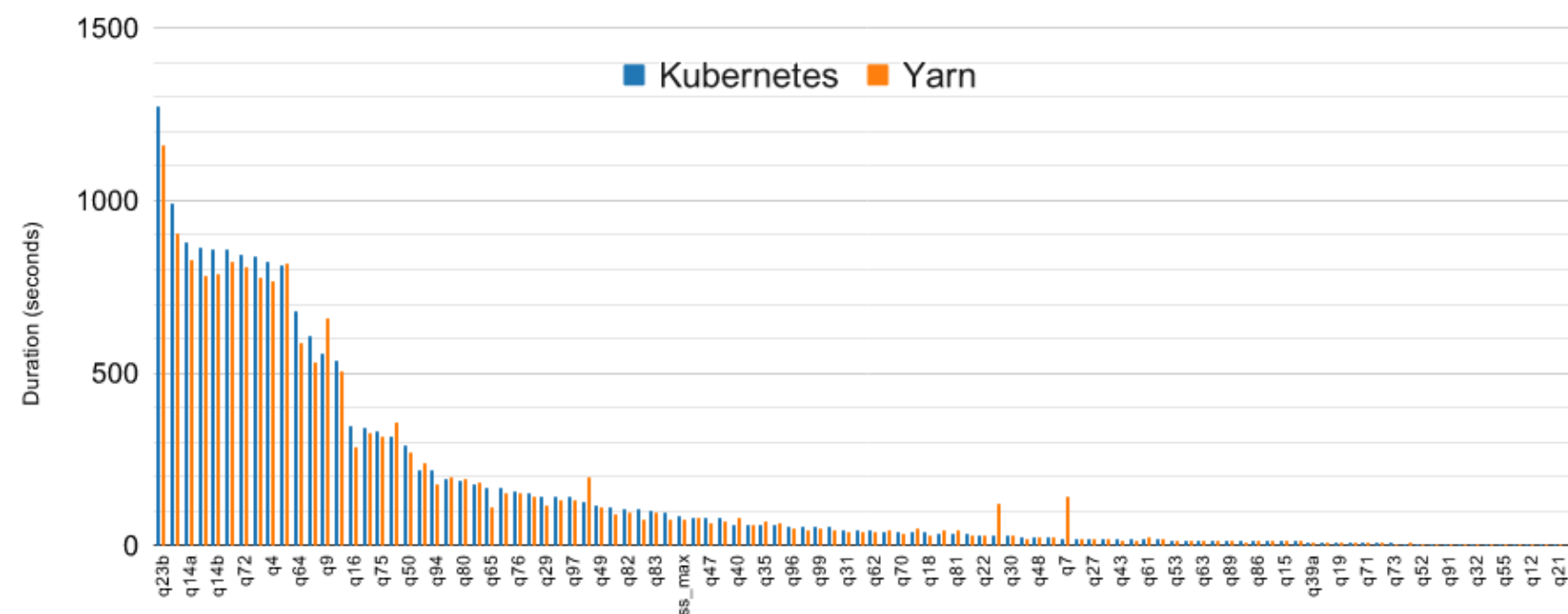
Гибкое масштабирование



Разделение storage и compute слоёв

Сравнение быстродействия

Local SSDs, Kubernetes versus Yarn



Yarn в среднем быстрее на 4-5%

<https://www.datamechanics.co/blog-post/apache-spark-performance-benchmarks-show-kubernetes-has-caught-up-with-yarn>

Этапы

01

Запустим k8s кластер

02

Установим Spark Operator

03

Запустим стандартную тестовую джобу для проверки работоспособности Spark Operator

04

Соберём кастомный образ с нашим приложением

05

Запустим приложение, проверим логи, результаты работы

Репозиторий с инструкцией и необходимыми шаблонами

https://github.com/stockblog/webinar_spark_k8s

ИТОГИ

01

Познакомились со Spark: преимуществами и ограничениями при запуске в K8s

02

Получили базовые навыки работы со Spark Operator

03

Протестировали Spark в Kubernetes

04

Собрали собственный Docker образ со Spark приложением

05

Протестировали Spark History Server в K8s



СЛЕПМ при поддержке



VK Cloud Solutions



intel.



Спасибо!

Александр Волынский

Почта: a.volinsky@corp.mail.ru

Моб. тел: +7 (960) 115-76-94

Вконтакте: <https://vk.com/volinski>



VK Cloud Solutions

Полезные статьи по теме:



Памятка по Spark
в Kubernetes



MLOps-1.
Как развернуть
Kubeflow
в Kubernetes
в продакшен-
варианте



MLOps-2.
MLflow — это еще один инструмент
для построения MLOps, для работы
с которым не обязателен Kubernetes

Подписывайтесь на канал t.me/k8s_vk



Скоро анонсируем
конференцию
Kubernetes
9 декабря
!

