Remote Shuffle Services for Apache Spark on K8S

kt NexR 민영근

목차

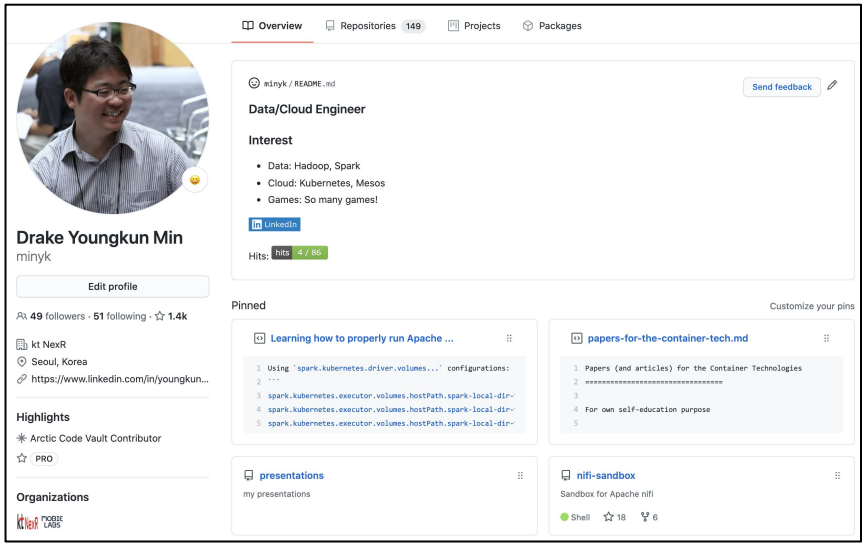
- 자기소개/ 회사소개
- Spark on Kubernetes
  - Spark 소개
  - Spark on Kubernetes
- Data Shuffling
  - On-Prem/Cloud 구조적 차이
- Remote Shuffle Services
  - Apache Uniffle/ Apache Celeborn
- 마무리
- 참고

자기소개

- 2019.4 ~ kt NexR R&D 센터
- 2016.6 ~ 2019.4 AJ 네트웍스 IT센터
- 2013.8 ~ 2016.6 kt NexR TA팀
- 2011.8 ~ 2013.7 단국대 연구전담교수
- 2011.2 단국대학교 공학박사

- 관심사: 데이터 처리, 컨테이너

https://github.com/minyk



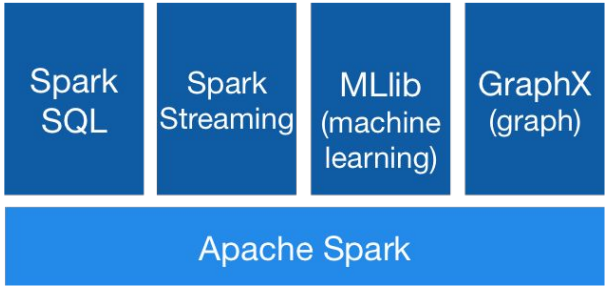
kt NexR

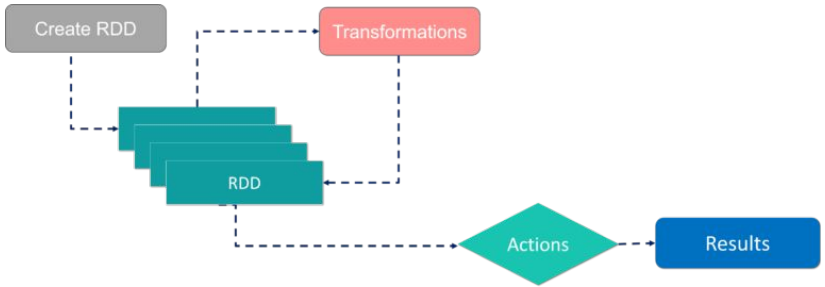
- 2023: K8S 기반의 데이터레이브 플랫폼 NDAP 7.0 서비스 출시
- 2021: Hadoop 3.0 적용된 NDAP 5.1 출시
- 2020: K8S 기반의 분석 플랫폼 NexR Enterprise v1.2.0 출시
- 2016: 실시간 처리 플랫폼 Lean Stream 1.0 출시
- 2014: Hadoop 2.0 적용된 NDAP 4.0 출시
- 2012: Hadoop 기반의 빅데이터 플랫폼 NDAP 출시, kt 빅데이터 플랫폼 개발
- 2011: kt 그룹사 편입
- 2007: NexR 설립

Spark on Kubernetes

Apache Spark: 소개

- Unified Analytics Engine for Large-scale Data Processing
  - 2009년 시작, Hadoop MapReduce 프로그래밍 모델의 대안
  - API(scala/java/python/R), SQL, Streaming, GraphX, ML 지원
- Hadoop Mapreduce의 뒤를 잇는 대용량 데이터 처리 엔진
  - RDD(Resilient Distributed DataSet): 메모리에 저장되는 변경되지 않는 데이터집합
  - Lazy execution이 특징

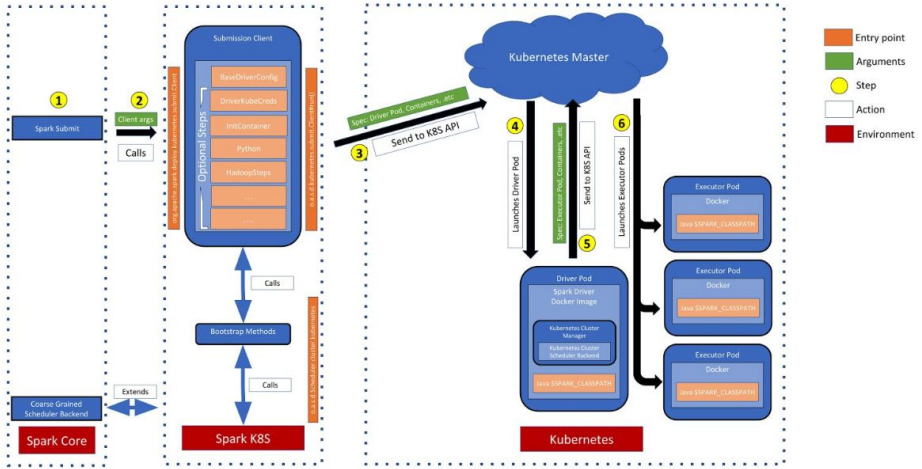




Spark on Kubernetes

- 2016.11~2018.3 진행
  - SPARK-18278 SPIP: Support native submission of spark jobs to a kubernetes cluster
- 별도의 repo에서 개발되어 2.3.0에 병합
- 현재(3.4.0) GA
  - 커스텀 K8S 스케쥴러 지원
  - Dynamic Allocation 지원
  - 외부 Shuffle 서비스 미지원

Summary Architecture Diagram



From Design Proposal Doc.

Data Shuffling

Data shuffling

- 분산 데이터 처리 구조에서 중요한 역할
- MapReduce, Spark, Presto 등의 분산 데이터 처리 프로그램에서 사용
- 데이터를 균등하게 분산; 분산된 데이터 처리 연산의 성능 향상
- 재분배 과정에서 CPU, 네트워크, 저장소 자원을 광범위하게 사용

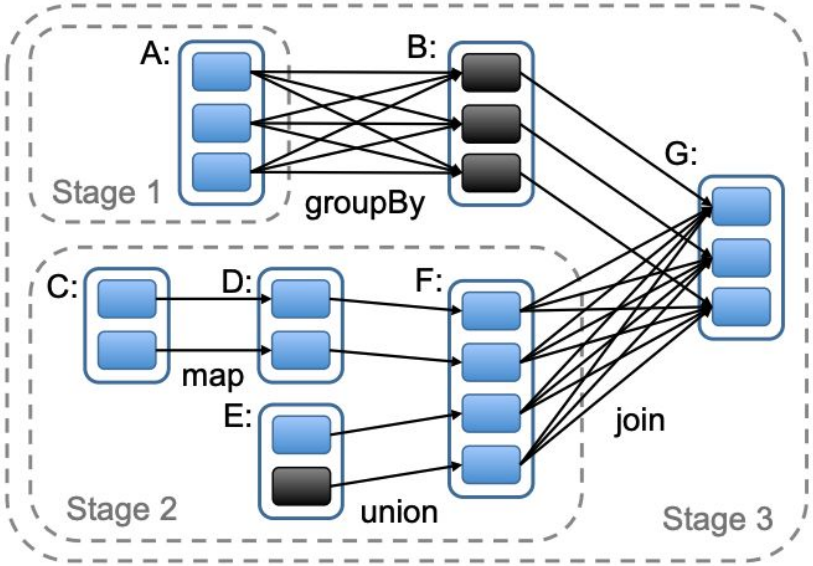
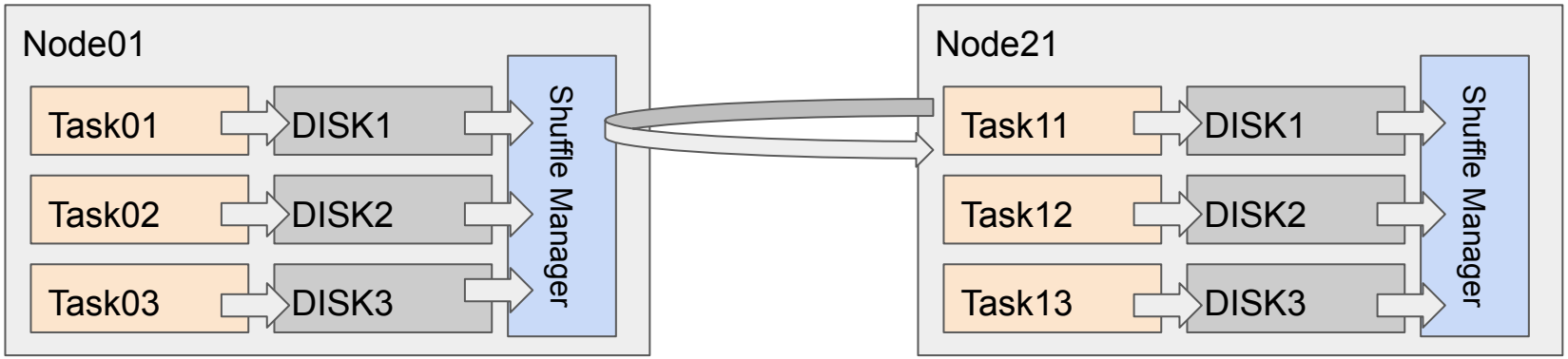


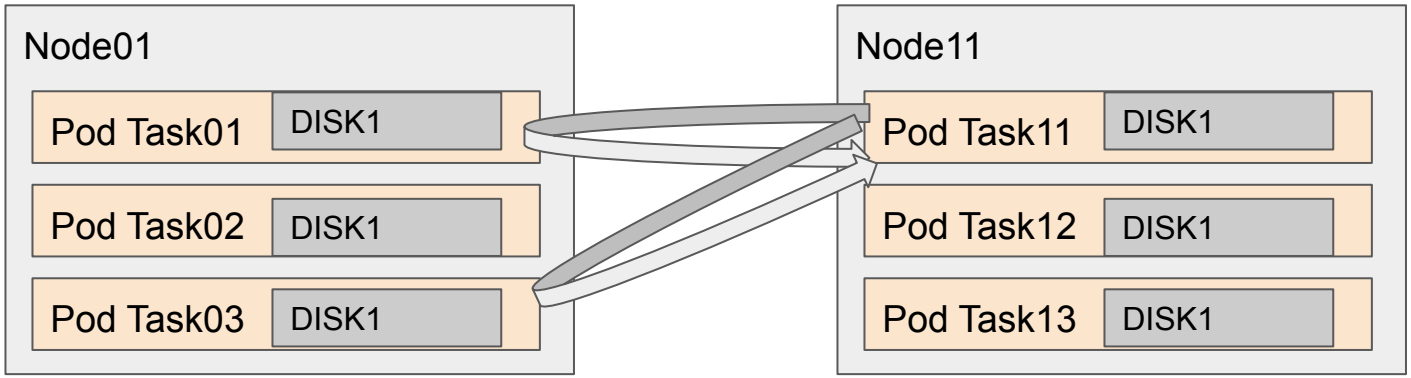
Fig 5. Example of how Spark computes job stages, from RDD paper, M. Zaharia, et al.

구축형 구조(Dedicated Architecture)



- 별도의 Shuffle Manager
- 연산을 담당하는 Task는 연산 종료 즉시 중지; 자원 해제
- Hadoop YARN

컨테이너 클라우드 구조(Container Cloud Arch.)

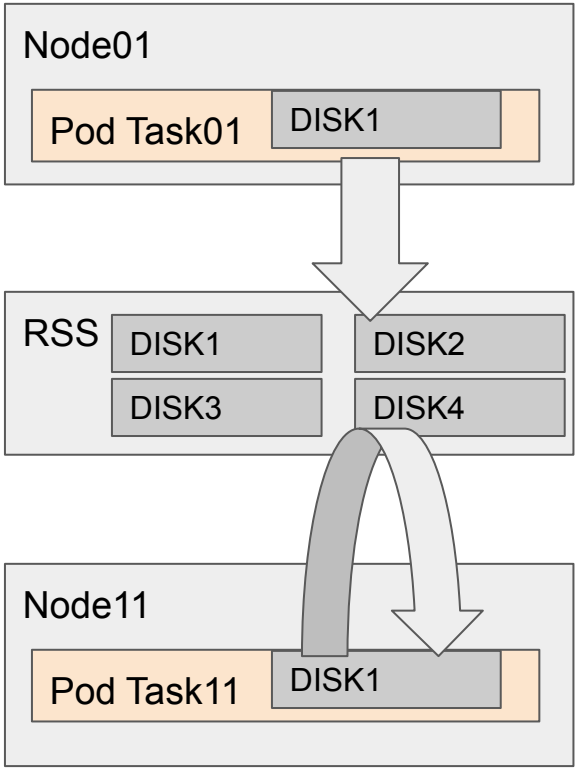


- 별도의 Shuffle Manager 없음
- 연산을 담당하는 Task가 직접 재분배 데이터 제공(ShuffleTracking 필요)
- 재분배 데이터가 다 사용되는 것을 기다려서 종료

Remote Shuffle Service

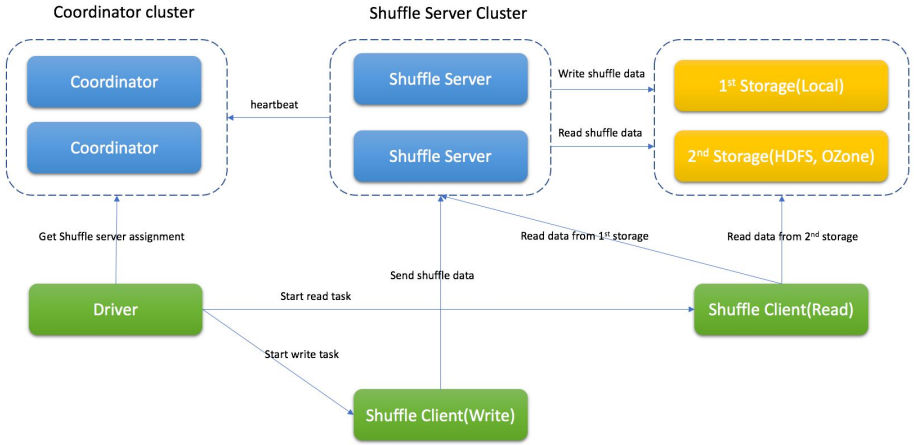
원격 데이터 재분배 서비스 (Remote Shuffle Service)

- 재분배 되는 데이터를 원격지에 저장
  - 컨테이너의 로컬 저장소를 사용하지 않음
  - 클라우드 환경에서 발생하는 문제 해소
- 비용이 높은 자원(특히 NVMe)의 효율적 사용
  - 별도 전용 노드 구성 가능
- 저장소 형태에 따른 계층화 가능
  - Memory, SSD, Disk
- 다양한 클라우드/ 서비스 회사에서 개발 중
  - Tencent, Aliyun, Uber, IBM 등



Apache Uniffle

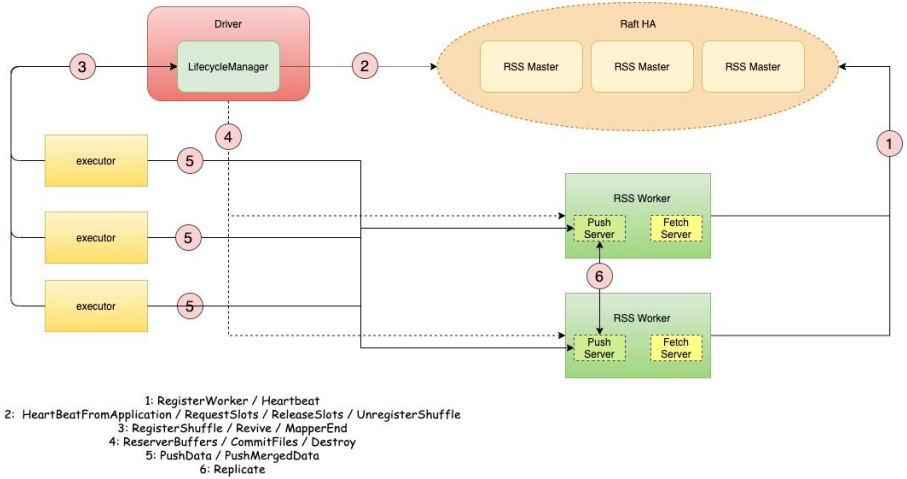
- Tencent 개발/ 기부
- Coordinator, Shuffle Server, 외부 데이터 저장소
- 3단계 저장소 구성
  - 메모리, 디스크, 외부 저장소
- Kubernetes Operator 제공
  - K8S에서 운용이 용이함



Uniffle Architecture, https://github.com/apache/incubator-uniffle#architecture

Apache Celeborn

- Aliyun 클라우드에서 개발/ 기부
- LifecycleManager, RSS Master, RSS Worker
- 부하 분산
  - 사용율이 낮은 워커에 할당
- 워커간에 데이터 복제
- Shuffle 데이터 병합 기능



Celeborn Architecture, https://github.com/apache/incubator-celeborn#architecture

비교

기능	Uniffle	Celeborn	비고
클러스터링 구조	자체 구성	Raft 알고리즘	
저장소 사용 제한	사용자별 프로그램 수	사용자별 디스크 사용량	
저장소 티어 구성	2, 3 티어	2, 3 티어	
확장 가능성	가능	어려움	Uniffle은 다양한 Interface 클래스 제공
최신 버전	0.7	0.2	둘다 incubating 상태
개발 언어	Java	Scala/Java	Celeborn은 5:5 비율
Kubernetes 지원	Operator 제공, 설정 리로드 가능	Helm Chart 제공	

비교

- Uniffle
  - 티어 구조: MEMORY_LOCALFILE, MEMORY_HDFS, MEMORY_LOCALFILE_HDFS
  - on-heap 메모리 사용
  - K8S 컨테이너 클라우드 환경에 유리
- Celeborn
  - 티어 구조: Memory-Local, Memory-Local-HDFS
  - off-heap 메모리 사용
  - 나눠진 재분배 데이터를 병합하여 한번에 읽기 가능
  - 재분배 데이터의 복제가 특징적(최신 버전에서는 off)

마무리

마무리

- Spark on Kubernetes
  - 현재 GA 상태, 아직 발전해야 할 부분이 있음
- Data Shuffle
  - 분산 데이터 처리에서 중요한 요소
  - Spark는 현재 직접 제공되는 k8s shuffle service가 없음
- Remote Shuffle Service
  - 데이터를 원격지에 보관, 자원을 좀 더 알뜰히 사용
  - Uniffle, Celeborn 등의 Opensource Project 들이 있음
- Production 적용 여부
  - in-house 가능, 외부 및 제품은 …

References

- Spark: Cluster Computing with Working Sets, Matei Zaharia, et. al., HotCloud 2010
- Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI 2012
- Optimizing Data Shuffling in Data-Parallel Computation by Understanding User-Defined Functions, Jiaxing Zhang, et. al., NSDI 2012
- Running Spark on Kubernetes, https://spark.apache.org/docs/latest/running-on-kubernetes.html
- Apache Uniffle, https://uniffle.apache.org/
- Apache Celeborn, https://celeborn.apache.org/
- LinkedIn Magnet, https://engineering.linkedin.com/blog/2020/introducing-magnet
- Uber RSS, https://github.com/uber/RemoteShuffleService