

Spark

용어 정리

- Application
 - Spark에서 작동하는 프로그램. driver와 cluster의 executors로 구성
- Driver
 - 프로그램의 main() 함수를 실행하고 SparkContext를 만드는 프로세스
- Cluster manager
 - 클러스터에서 자원을 확보하기 위한 외부 서비스
(독립형 관리자, Mesos, YARN, Kubernetes)
- Worker node
 - 클러스터에서 응용 프로그램 코드를 실행할 수 있는 모든 노드

용어 정리

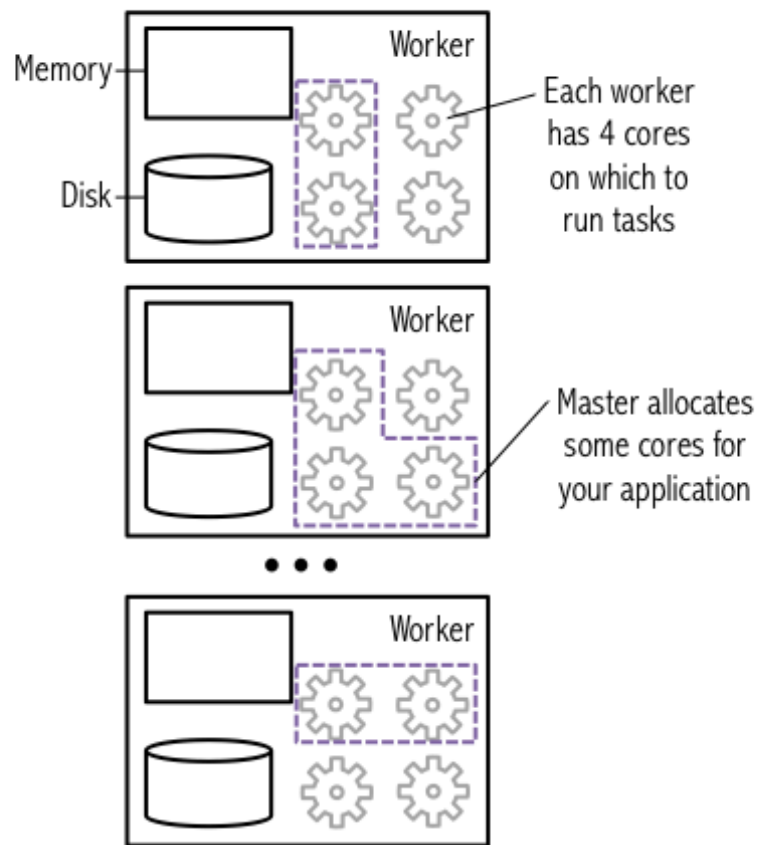
- Executor
 - 작업을 실행하고 데이터를 메모리 또는 디스크 스토리지에 보관하는 프로세스
- Job
 - Spark에서 처리해야할 작업
- Task
 - 한 executor에 보내지는 job 단위
- Stage
 - 의존성이 있는 task
 - Each job gets divided into smaller sets of tasks called stages that depend on each other (similar to the map and reduce stages in MapReduce)

Spark 동작 방식

Spark 동작 방식

- spark간 driver과 executor 사이에서 발생
- driver은 실행에 필요한 spark job들을 가지고 spark job들은 executor에서 실행되기 위해 task단위로 쪼개짐
- spark와 API를 사용하기 위해서는 SparkContext 사용이 필요
- sparkContext가 생성되면 마스터에게 동작 가능한 core들을 요청
- 마스터가 동작 가능한 core들을 설정하면 설정된 core들은 다른 application에서는 사용되지 않음

Spark 동작 방식



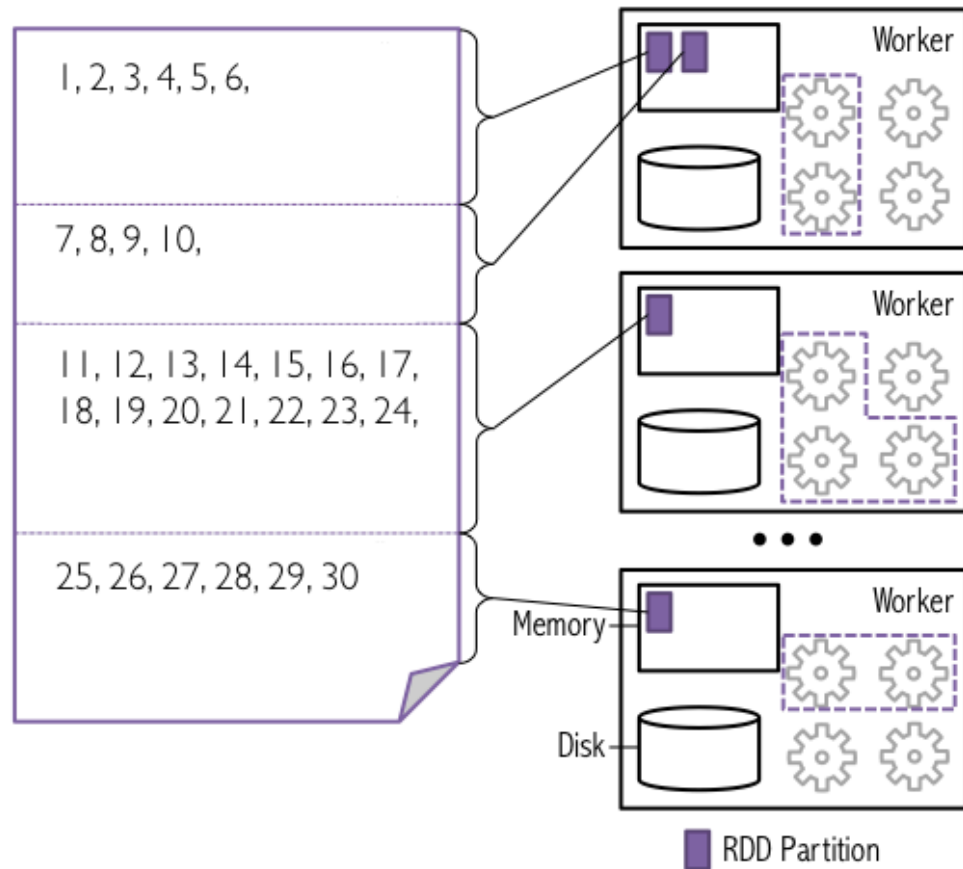
Spark 동작 방식

- worker에서 사용할 core가 정해진 이후 immutable(불변) 형태의 base RDD를 생성
- base RDD에 여러 transformation을 통해 사용할 RDD를 만듦
- 이렇게 만들어진 RDD는 각기 다른 worker에 저장되기 위해 각각의 파티션으로 분할
- 각각의 파티션은 리스트 내의 유일한 집합 서브셋을 가짐
(중복 x)

Spark 동작 방식

Dataset is broken into partitions

Partitions are each stored in a worker's memory



Python - Spark (Pyspark)

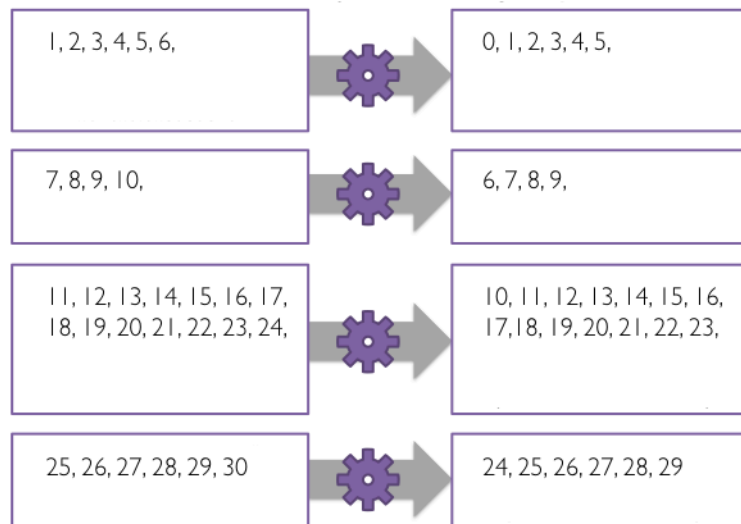
Python - Spark (Pyspark) - RDD 생성

- RDD를 생성하기 위해 pyspark에서 `sc.parallelize()`를 사용
ex) `sc.parallelize(data, 8)` //data가 메모리에 저장될 때 8조각으로 쪼개서 메모리에 저장 - base RDD

Python - Spark (Pyspark) - map

- map(f)
 - 가장 기본적인 transformation
 - dataset에 있는 각각의 item에 함수 f가 적용되고 task 단계가 시작
 - task는 각각의 파티션에서 시작되고 서로 다른 input data에 대해 동일한 로직이 실행
 - task가 완료되면 새로운 파티션이 출력

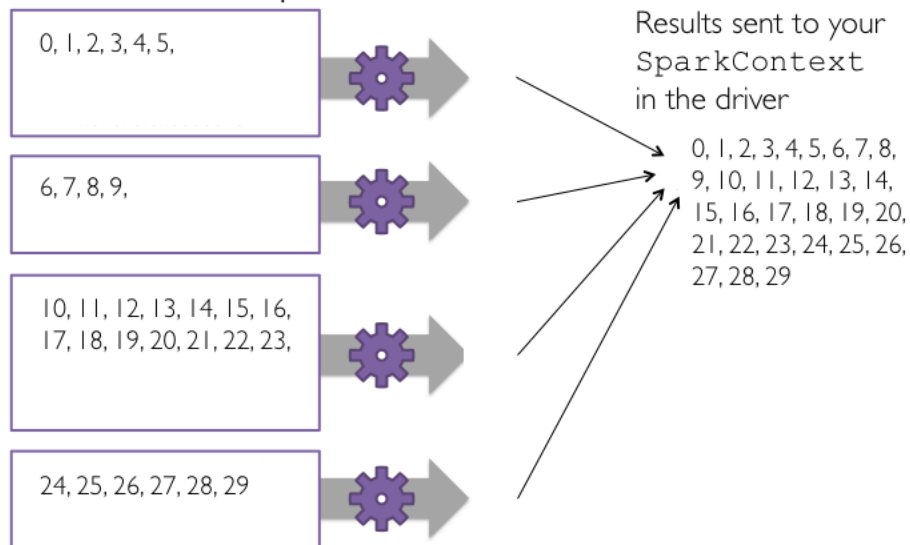
map (f) : Each task makes a new partition by calling $f(e)$ on each entry e in the original partition



Python - Spark (Pyspark) - collect

- Collect()
 - 분배된 데이터들을 새로운 list로 합치고 transformation 적용 - action
 - collect() 함수가 호출되면 RDD가 메모리에 올려져서 계산이 이루어짐

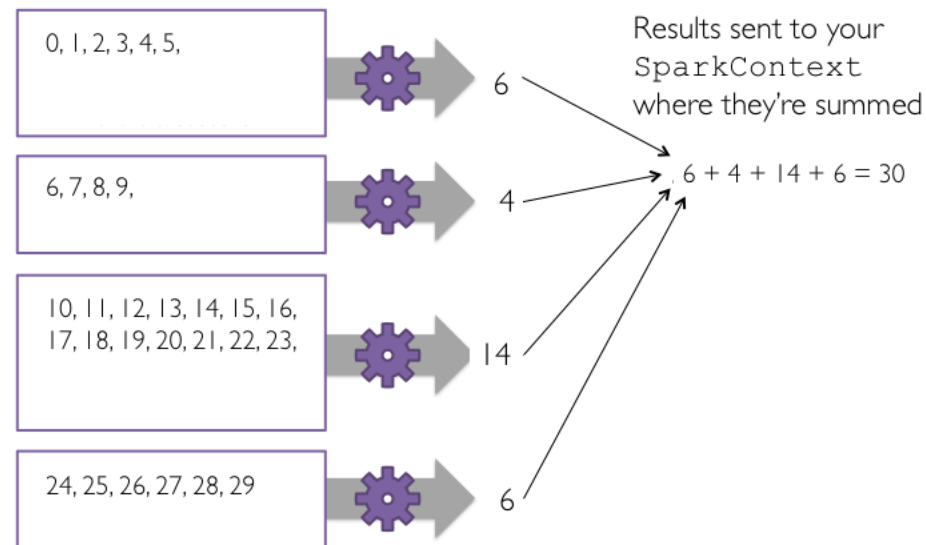
collect() : Gathers the entries from all partitions into the driver



Python - Spark (Pyspark) - count

- Count()
 - RDD 내에 있는 요소들의 개수를 세는 함수로 action에 속함
 - 각각의 task가 자신의 파티션에 있는 entry의 개수를 센 후 그 결과를 SparkContext에 전송

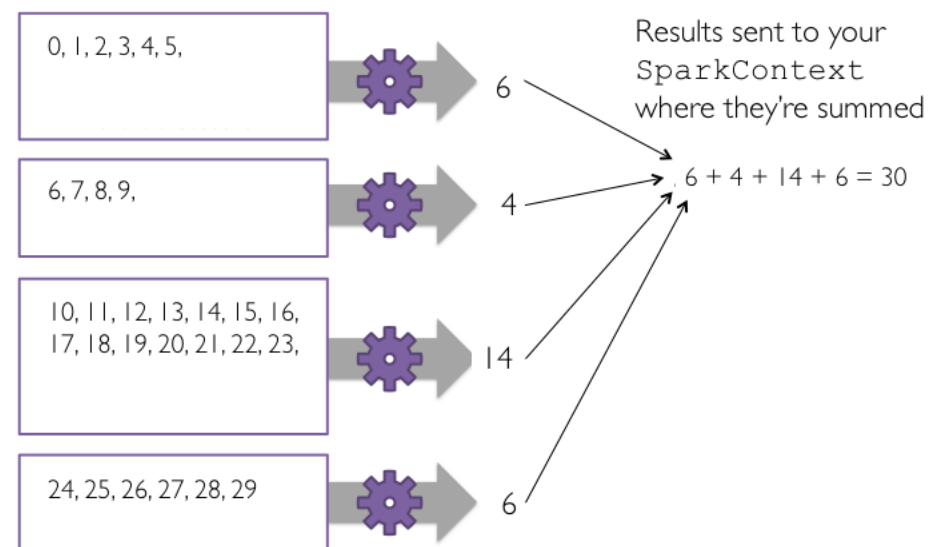
count () : Each task counts the entries in one partition



Python - Spark (Pyspark) - filter

- Filter()
 - 함수 결과가 참인 경우에만 요소들을 통과시키는 함수로 새로운 RDD를 생성 - transformation

count () : Each task counts the entries in one partition



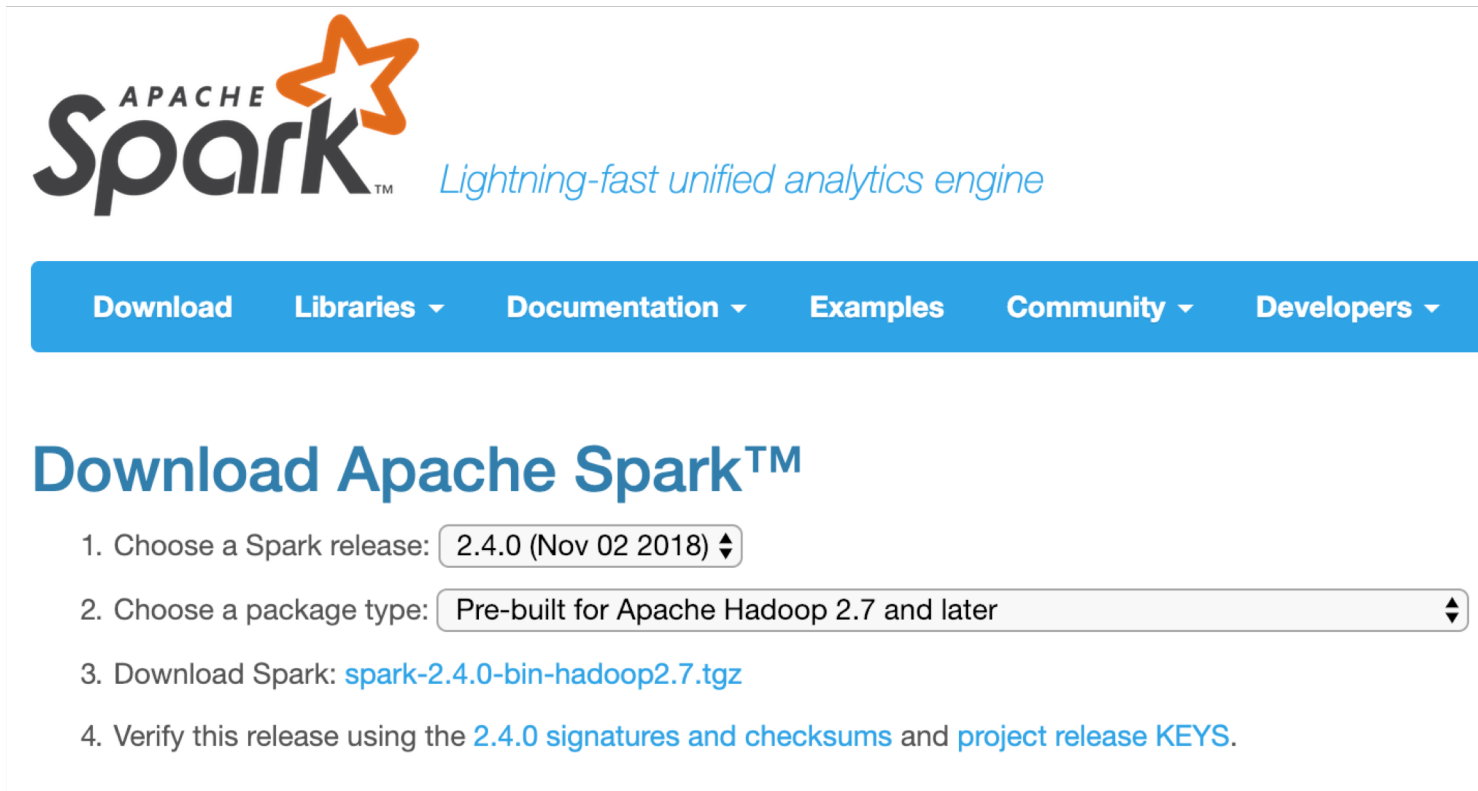
Spark - Python 설치 및 예제

Spark standalone(싱글 노드) 설치

- JDK가 설치되어 있어야 함(1.8 버전이 설치되어 있는 것이 좋음)
- JDK 11 버전으로 실행하는 경우 오류가 발생
- stack overflow Link 참고
- <https://stackoverflow.com/questions/52524112/how-do-i-install-java-on-mac-osx-allowing-version-switching>
- <https://stackoverflow.com/questions/44914144/error-sparkcontext-error-initializing-sparkcontext-java-net-bindexception-can/44916827>

Spark standalone(싱글 노드) 설치

- <http://spark.apache.org/downloads.html> 에서 다운로드



The screenshot shows the Apache Spark website's download section. At the top is the Apache Spark logo with the tagline "Lightning-fast unified analytics engine". Below the logo is a blue navigation bar with links: "Download", "Libraries", "Documentation", "Examples", "Community", and "Developers". The main heading is "Download Apache Spark™". Below this, there are four steps for downloading:

1. Choose a Spark release: 2.4.0 (Nov 02 2018) (dropdown menu)
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later (dropdown menu)
3. Download Spark: [spark-2.4.0-bin-hadoop2.7.tgz](#)
4. Verify this release using the [2.4.0 signatures and checksums](#) and [project release KEYS](#).

Spark standalone(싱글 노드) 설치

- 다운 받은 spark~~.tgz 파일 압축 해제
- tar -zxvf spark~~.tgz

```
yjhanui-MacBook-Pro:spark yj.han$ ls
spark-2.4.0-bin-hadoop2.7      spark-2.4.0-bin-hadoop2.7.tgz
```

- scala를 사용하는 경우 spark~/bin/spark-shell 실행
- python을 사용하는 경우 spark~/bin/pyspark 실행

```
yjhanui-MacBook-Pro:bin yj.han$ ./pyspark
Python 2.7.15 (default, Aug 22 2018, 16:36:18)
[GCC 4.2.1 Compatible Apple LLVM 9.1.0 (clang-902.0.39.2)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/Users/yj.han/spark/spark-2.4.0-bin-hadoop2.7/jars/had
oop-auth-2.7.3.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
2019-02-10 23:02:31 WARN NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

      ____
     /___\   _ __| |__ \
    / ___ \| '_ \| |___| |
   /_/ ___| |_) | |___|_|
  /_____\| .__/|_|___|_|
         |_|

version 2.4.0

Using Python version 2.7.15 (default, Aug 22 2018 16:36:18)
SparkSession available as 'spark'.
>>>
```

Spark - Python 설치 및 예제

- 'a' 혹은 'b'가 포함된 Line 갯수를 찾는 예제 프로그램 작성
- README.md 파일에서 'a'와 'b'가 포함된 line수를 반환하는 예제
- spark가 설치된 경로에서
spark-2.4.0-bin-hadoop2.7/bin/spark-submit --master local[4]
spark_test.py로 실행

```
1 ab
2 ab
3 ab
4 b
5 b
```

```
from pyspark import SparkContext

logFile = '/Users/yj.han/spark/README.md'
sc = SparkContext('local', 'Simple App')
logData = sc.textFile(logFile).cache()

numAs = logData.filter(lambda s: 'a' in s).count()
numBs = logData.filter(lambda s: 'b' in s).count()

print ("Lines with a: %i: lines whth b: %i" % (numAs, numBs))
```

```
2019-02-17 22:24:04 INFO DAGScheduler
2019-02-17 22:24:04 INFO DAGScheduler
Lines with a: 3: lines whth b: 5
2019-02-17 22:24:04 INFO SparkContext
```

Spark 동작 방식

- <http://hellowuniverse.com/2017/03/08/spark-standalone-%EC%84%A4%EC%B9%98%EB%B6%80%ED%84%B0-%EC%98%88%EC%A0%9C-%ED%85%8C%EC%8A%A4%ED%8A%B8%EA%B9%8C%EC%A7%80/>
- 예제를 통해 확인 가능

참고 자료

- <http://spark.apache.org/docs/latest/cluster-overview.html>
- <https://yujuwon.tistory.com/entry/spark-tutorial>
- <http://blog.naver.com/PostView.nhn?blogId=gyrbsd118&logNo=220880041737&categoryNo=3&parentCategoryNo=0&viewDate=¤tPage=1&postListTopCurrentPage=1&from=postView>
- <https://stackoverflow.com/questions/52524112/how-do-i-install-java-on-mac-osx-allowing-version-switching>
- <http://hellowuniverse.com/2017/03/08/spark-standalone-%EC%84%A4%EC%B9%98%EB%B6%80%ED%84%B0-%EC%98%88%EC%A0%9C-%ED%85%8C%EC%8A%A4%ED%8A%B8%EA%B9%8C%EC%A7%80/>