KOSA
Korea Software Industry Association

빅데이터 & 딥러닝

PySpark Install

정 준 수 PhD

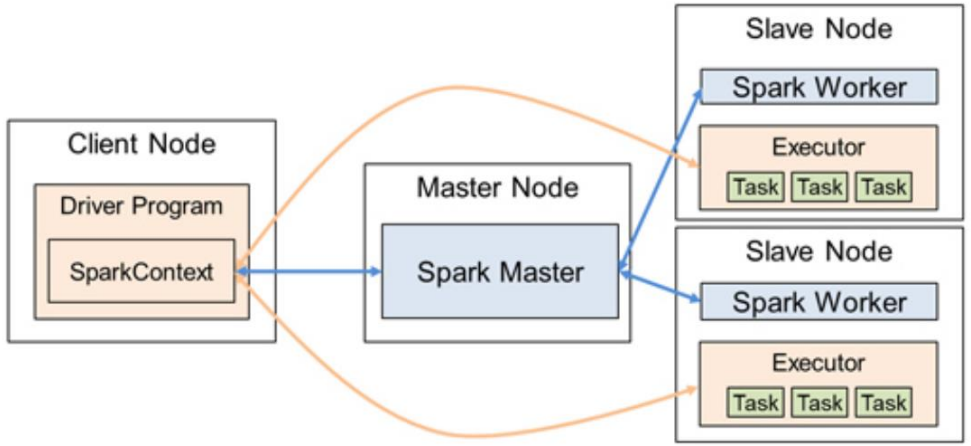
Hadoop을 통해 저렴한 PC로도 빅데이터 처리가 가능하게 함

Spark은 빠른 속도와 강력한 기능을 제공하는 오픈소스가 되었고,
이제는 빅데이터 Ecosystem의 중심

Spark의 시작

Standalone Cluster Manager는 Master와 Worker로 구성

Master 노드에는 Spark Master 프로세스가 실행되고, 나머지 노드에는 Spark Worker 프로세스가 실행



Spark Cluster Architecture

[출처] https://www.samsungsds.com/kr/insights/Spark-Cluster-job-server.html?moreCnt=1&backTypeId=undefined&category=undefined

◎ (Spark) Application: Spark에서 수행되는 사용자 프로그램으로, Driver Program과 여러 Executor 프로세스로 구성

◎ Driver Program: main 함수를 실행시키고 그 안에서 SparkContext를 생성하는 메인 프로세스

◎ SparkContext: Driver Program에서 Job을 Executor에 실행하기 위한 Endpoint

◎ Cluster Manager: Application 자원을 할당, 제거하는 등 Cluster 자원을 관리하는 서비스

◎ Master: Standalone의 Master 프로세스 또는 Cluster의 Master 역할로서 Worker를 관리하는 컴퓨터 노드

◎ Worker: Standalone의 Worker 프로세스 또는 Cluster에서 Slave 역할로서 실제 연산 작업을 수행하는 컴퓨터 노드

◎ Executor: Application에서 Driver Program이 요청한 Task들의 연산을 실제로 수행하는 프로세스

◎ Job: Application에서 Spark에 요청하는 일련의 작업. 여러 개의 Task로 나뉘어 실행됨

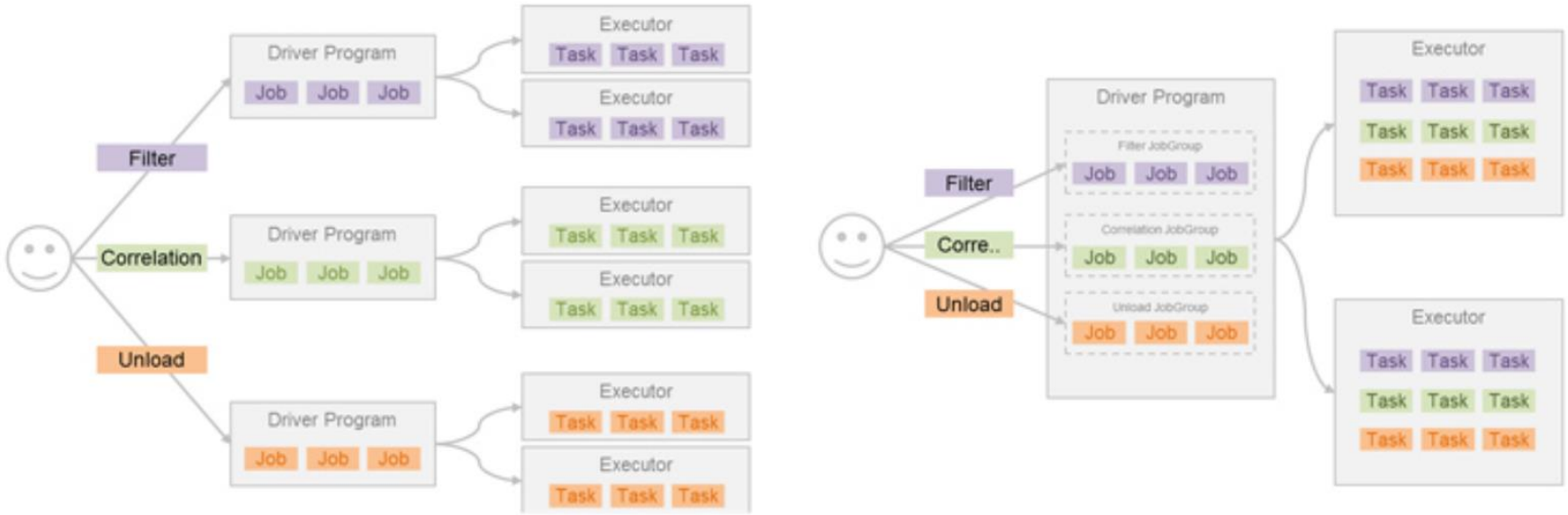
◎ Task: Spark Executor에서 수행되는 최소 작업 단위

Application은 기본적으로 Driver Program 프로세스로 실행됩니다. 이 프로그램 안에서는 'SparkContext'라는 아주 중요한 객체를 생성시켜야 합니다. 이를 통해 Spark Cluster와 커뮤니케이션할 수 있기 때문입니다. SparkContext를 통해 Application에서 요구하는 리소스를 요청하면, Spark Master는 Worker들에게 요청받은 리소스만큼의 Executor 프로세스를 실행하도록 요구

2개의 Executor 프로세스에서 각각 3개 작업을 동시 실행하는 총 6 CPU cores를 할당받음

Application은 1개 이상의 Job을 실행시키고, 이 Job은 여러 개의 Task로 나누어서 Executor에게 요청하고 결과를 받으면서 Cluster 컴퓨팅 작업을 수행

Spark Context 관리



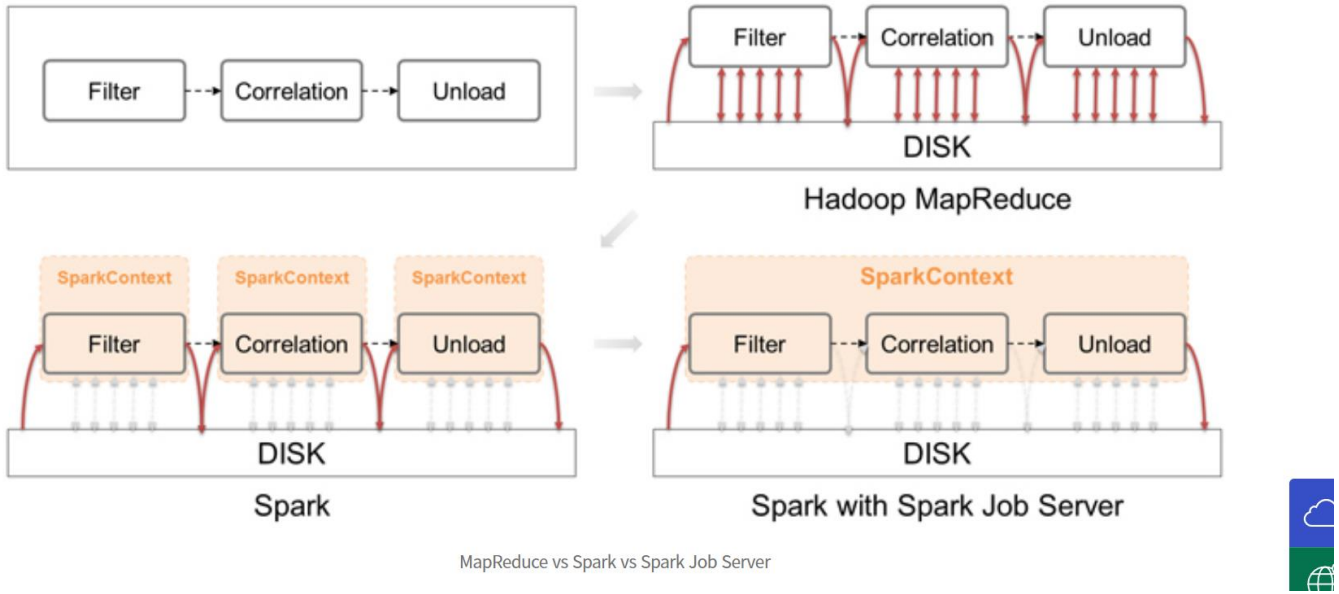
Spark Applications vs Spark Job Groups

Filter → Correlation → Unload(DB) 작업을 수행하는 분석 모델이 있다고 가정

왼쪽의 일반적인 상황에서는 각 Application을 하나씩 수행한 후, 다음 Application을 수행하는 방식으로 모델을 수행합니다. 반면 오른쪽은 하나의 Application을 Daemon 형태로 실행한 후, 이곳에 Job Group들을 실행하는 방식

왼쪽은 Application의 실행/중지가 세 차례 이루어지는 반면, 오른쪽은 단 한 번도 이루어지지 않음을 확인

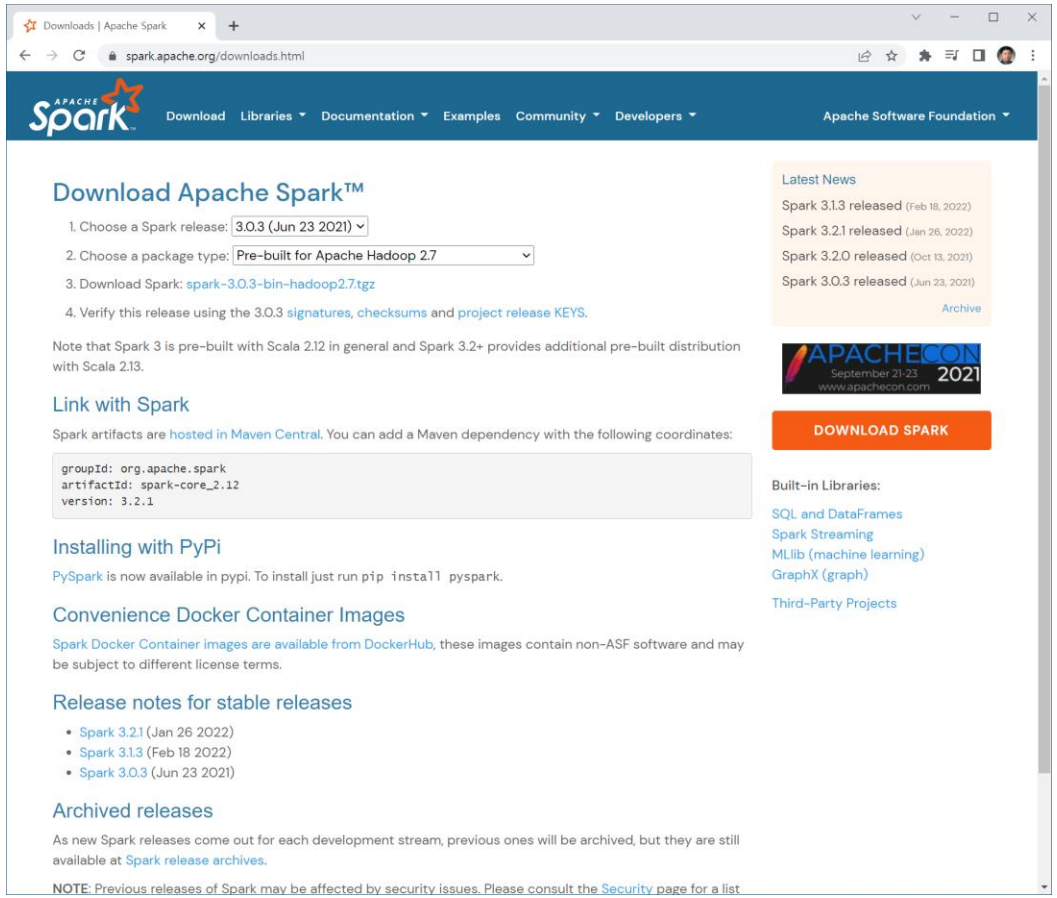
결과적으로 순수한 분석 시간에는 차이가 없지만 Application 실행/종료에 따른 시간, 각 Application에서 공유하는 데이터를 읽고 쓰는 시간을 줄여서 전체적인 분석 시간을 단축



MapReduce vs Spark vs Spark Job Server

이렇게 Spark Context를 독립적으로 관리하면 하나의 Spark Cluster에서 Multi-tenancy를 지원할 수도 있고, Application의 리소스 확보도 안정적으로 이룰 수 있습니다. 더불어 Disk I/O도 많이 줄여서 성능 향상도 꾀할 수 있습니다.

Spark Download



Release: 3.0.3

Package type: Apache Hadoop 2.7

URL: download 받지 마세요!!!

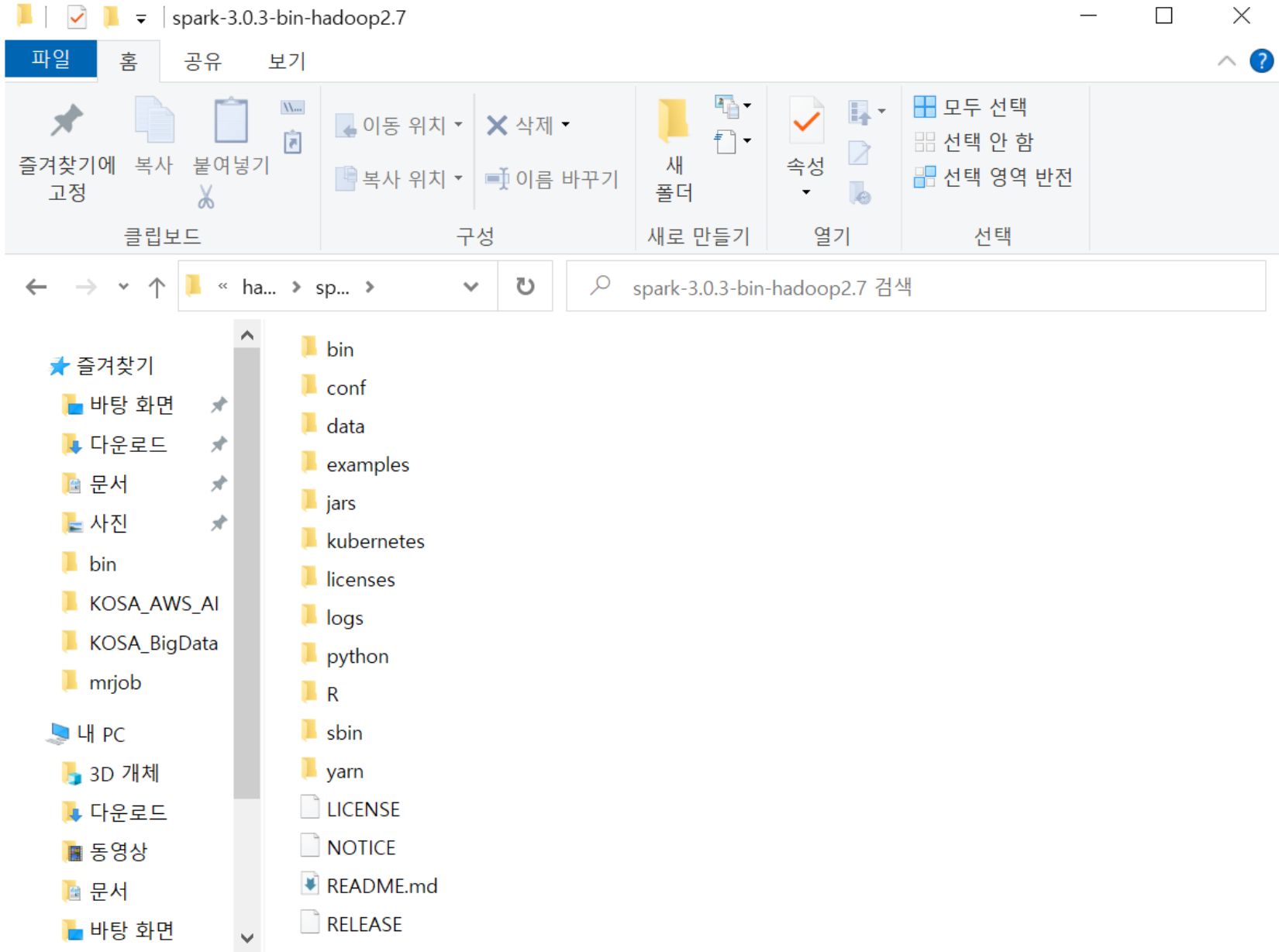
$ https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz

$ tar -xzf spark-3.0.3-bin-hadoop2.7.tgz

https://spark.apache.org/downloads.html

Pyspark Install

```
$ pip install pyspark pandas numpy
```


spark-3.0.3-bin-hadoop2.7
파일
홈
공유
보기
즐겨찾기에 고정
복사
붙여넣기
이동 위치
복사 위치
삭제
이름 바꾸기
새 폴더
속성
모두 선택
선택 안 함
선택 영역 반전
클립보드
구성
새로 만들기
열기
선택
ha...
sp...
spark-3.0.3-bin-hadoop2.7 검색
즐겨찾기
바탕 화면
다운로드
문서
사진
bin
KOSA_AWS_AI
KOSA_BigData
mrjob
내 PC
3D 개체
다운로드
동영상
문서
바탕 화면
bin
conf
data
examples
jars
kubernetes
licenses
logs
python
R
sbin
yarn
LICENSE
NOTICE
README.md
RELEASE
16개 항목

Python Path 지정

$ ₩home₩hadoop₩**spark-3.0.3-bin-hadoop2.7**₩conf₩ **cp spark-env.sh.template spark-env.sh**

$ hadoop@DESKTOP-HAAI0JO:~/spark-3.0.3-bin-hadoop2.7/conf$ vi spark-env.sh

export PYSPARK_PYTHON=/home/hadoop/virtualenv/bin/python

spark standalone 클러스터 모드에서 어플리케이션 실행

1. conf/slaves에 slave 서버의 호스트 네임 또는 아이피주소를 입력해준다.

2. sbin/start-master.sh 로 시작

3. master 아이피의 8080포트로 이동하면 web UI 확인 가능

4. 최상단 제목 부분에서 spark://spark-master ... 로 시작하는 주소를 가지고 애플리케이션을 실행해야 한다.

<아래는 실행 Code>

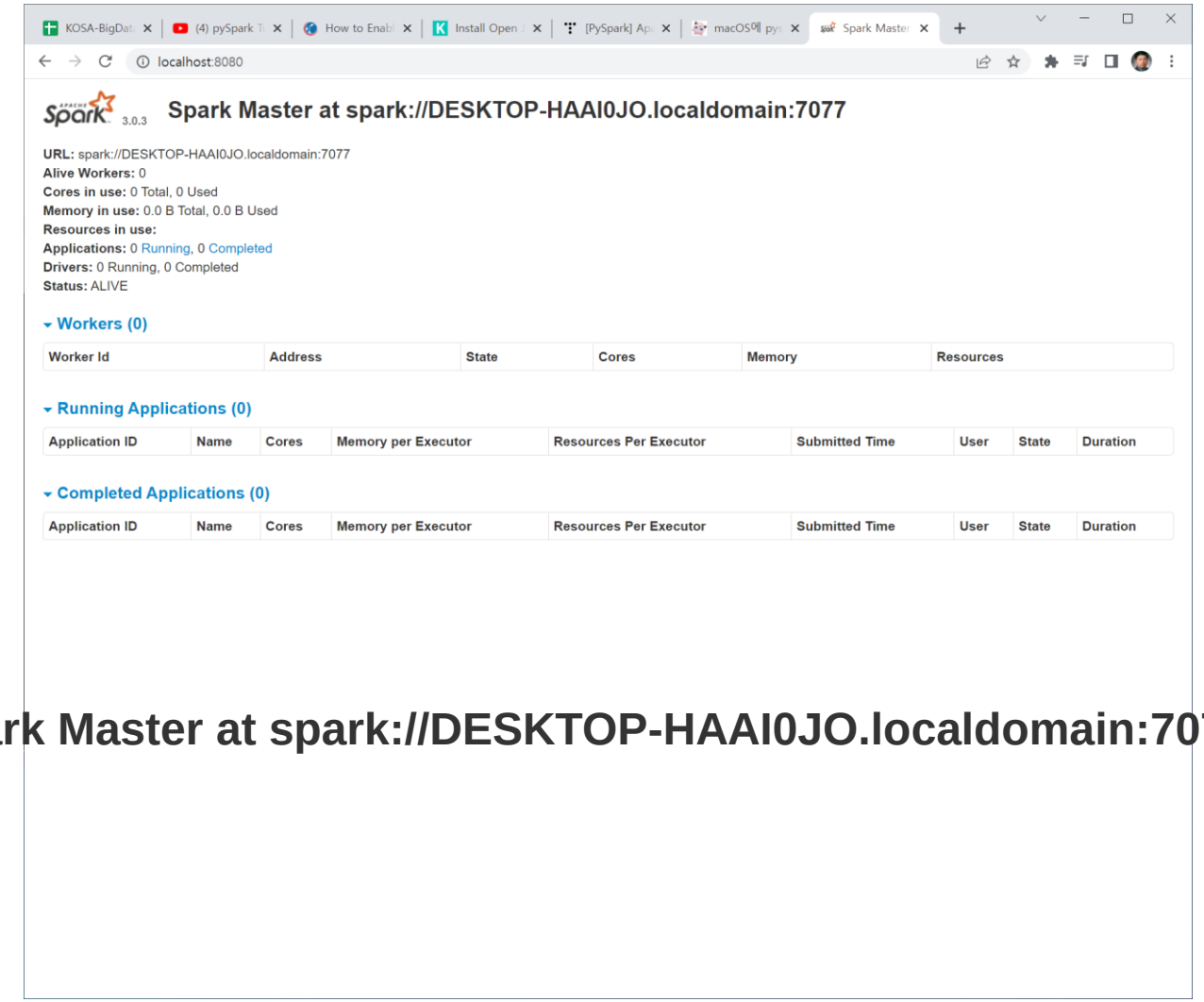
hadoop@DESKTOP-HAAI0JO:~/spark-3.0.3-bin-hadoop2.7/conf$ cp slaves.template slaves

hadoop@DESKTOP-HAAI0JO:~/spark-3.0.3-bin-hadoop2.7/sbin$ ./start-all.sh

Ssh connect refused 나오면 : $ sudo service ssh start

https://batterflyyin.tistory.com/6

http://localhost:8080/ 로 Spark Master URL을 확인



Spark Master at spark://DESKTOP-HAAI0JO.localdomain:7077

hadoop@DESKTOP-HAAI0JO:~/spark-3.0.3-bin-hadoop2.7/bin$ ./pyspark --master spark://DESKTOP-HAAI0JO.localdomain:7077

PySpark - jupyterlab 실행 방법

$ sudo pip install jupyterlab

Spark-env.sh 2 line 추가

```
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
```

$ source spark-env.sh

$ spark-3.0.3-bin-hadoop2.7/bin/pyspark --master spark://DESKTOP-HAAI0JO.localdomain:7077

URL을 Browser 에 복사

PySpark cli 실행

hadoop@DESKTOP-HAAI0JO:~/spark-3.0.3-bin-hadoop2.7/bin$ ./spark-submit ~/prj2/line_count.py

https://bab-dev-study.tistory.com/22

Spark RDD (Resilient Distributed DataSet)

RDD 는 여러 분산 노드에 걸쳐서 저장되는 변경이 불가능한 데이타(객체)의 집합으로 각각의 RDD는 여러 개의 파티션으로 분리가 된다. 즉, 스파크 내에 저장된 데이타를 RDD라고 하고, 변경이 불가능하다. 변경을 하려면 새로운 데이타 셋을 생성해야 한다.

RDD에서는 딱 두 가지 오퍼레이션만 지원한다.
Transformation : 기존의 RDD 데이타를 변경하여 새로운 RDD 데이타를 생성해내는 것. 흔한 케이스는 filter와 같이 특정 데이타만 뽑아 내거나 map 함수 처럼, 데이타를 분산 배치 하는 것 등을 들 수 있다.
Action : RDD 값을 기반으로 무엇인가를 계산해서(computation) 결과를 (셋이 아닌) 생성해 내는것으로 가장 쉬운 예로는 count()와 같은 operation들을 들 수 있다.

Spark RDD 실습 예제

https://github.com/JSJeong-me/KOSA_BIGDATA_DEEPLEARNING/blob/main/PySpark/rdd.py

클러스터에서 각 CPU 당 2~4개의 파티션을 가지고 수행한다. 일반적으로 스파크는 자동적으로 파티션의 개수를 적정하게 세팅해주지만 사용자가 직접 수동적으로 파라미터를 집어 넣어 파티션 개수를 커스터마이징 해줄 수 있다.

```
data = [1, 2, 3, 4, 5]
distData = sc.parallelize(data, 10) # 10개의 파티션으로 수행!
res = distData.reduce(lambda a, b: a + b)
print(res)
```

Spark 외부 파일 불러 오기

타이타닉 Dataset : A Titanic Probability

https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html

$ ./pyspark --master spark://DESKTOP-HAAI0JO.localdomain:7077

https://towardsdatascience.com/spark-essentials-how-to-read-and-write-data-with-pyspark-5c45e29227cd

Reading CSV using InferSchema

df=spark.read.format("csv").option("inferSchema","true").load(filePath)

Example:

Welcome to

```
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.0.3
      /_/
```

Using Python version 3.9.2 (default, Feb 28 2021 17:03:44)
SparkSession available as 'spark'.

```
>>> titanic =
spark.read.format("csv").option("seperator",",").option("header",
True).option("inferSchema", True).
load("file:////home/hadoop/prj2/titanic.csv")
```

```
>>> titanic.show()
+--------+------+--------------------+------+----+--------------------+----------------------+-------+
|Survived|Pclass|                Name|   Sex| Age|Siblings/Spouses Aboard|Parents/Children Aboard|   Fare|
+--------+------+--------------------+------+----+--------------------+----------------------+-------+
|       0|     3|Mr. Owen Harris B...|  male|22.0|                   1|                     0|   7.25|
|       1|     1|Mrs. John Bradley...|female|38.0|                   1|                     0|71.2833|
|       1|     3|Miss. Laina Heikk...|female|26.0|                   0|                     0|  7.925|
|       1|     1|Mrs. Jacques Heat...|female|35.0|                   1|                     0|   53.1|
|       0|     3|Mr. William Henry...|  male|35.0|                   0|                     0|   8.05|
|       0|     3|     Mr. James Moran|  male|27.0|                   0|                     0| 8.4583|
|       0|     1|Mr. Timothy J McC...|  male|54.0|                   0|                     0|51.8625|
|       0|     3|Master. Gosta Leo...|  male| 2.0|                   3|                     1| 21.075|
|       1|     3|Mrs. Oscar W (Eli...|female|27.0|                   0|                     2|11.1333|
|       1|     2|Mrs. Nicholas (Ad...|female|14.0|                   1|                     0|30.0708|
|       1|     3|Miss. Marguerite ...|female| 4.0|                   1|                     1|   16.7|
|       1|     1|Miss. Elizabeth B...|female|58.0|                   0|                     0|  26.55|
|       0|     3|Mr. William Henry...|  male|20.0|                   0|                     0|   8.05|
|       0|     3|Mr. Anders Johan ...|  male|39.0|                   1|                     5| 31.275|
|       0|     3|Miss. Hulda Amand...|female|14.0|                   0|                     0| 7.8542|
|       1|     2|Mrs. (Mary D King...|female|55.0|                   0|                     0|   16.0|
|       0|     3| Master. Eugene Rice|  male| 2.0|                   4|                     1| 29.125|
|       1|     2|Mr. Charles Eugen...|  male|23.0|                   0|                     0|   13.0|
|       0|     3|Mrs. Julius (Emel...|female|31.0|                   1|                     0|   18.0|
|       1|     3|Mrs. Fatima Masse...|female|22.0|                   0|                     0|  7.225|
+--------+------+--------------------+------+----+--------------------+----------------------+-------+
only showing top 20 rows
```

```
>>> titanic.printSchema()
root
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- Siblings/Spouses Aboard: integer (nullable = true)
 |-- Parents/Children Aboard: integer (nullable = true)
 |-- Fare: double (nullable = true)

>>>
```

정 준 수 / Ph.D （ jsjeong@hansung.ac.kr ）

- 前) 삼성전자 연구원
- 前) 삼성의료원 (삼성생명과학연구소)
- 前) 삼성SDS (정보기술연구소)
- 現) (사)한국인공지능협회, AI, 머신러닝 강의
- 現) 한국소프트웨어산업협회, AI, 머신러닝 강의
- 現) 서울디지털재단, AI 자문위원
- 現) 한성대학교 교수(겸)
- 전문분야: Computer Vision, 머신러닝(ML), RPA
- https://github.com/JSJeong-me/