KOSA
Korea Software Industry Association

빅데이터 & 딥러닝

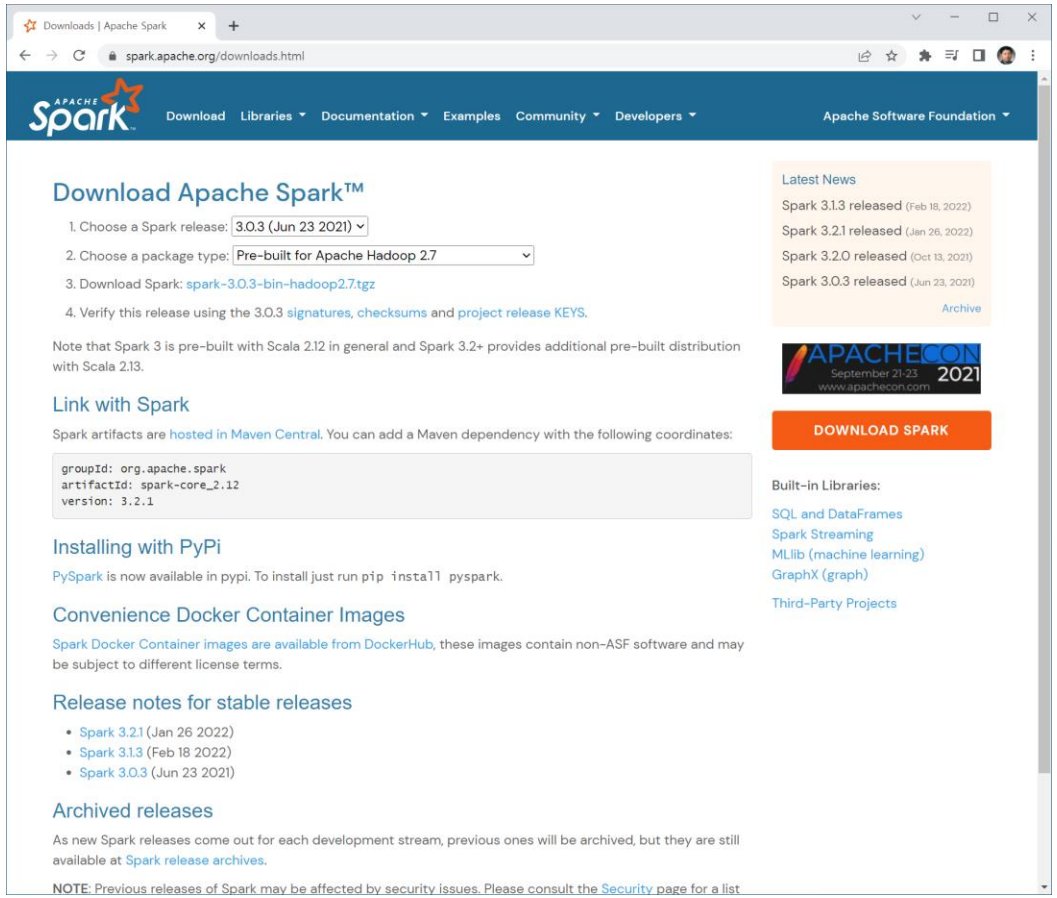
PySpark Install

정 준 수 PhD

Hadoop을 통해 저렴한 PC로도 빅데이터 처리가 가능하게 함

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

Spark Download



Release: 3.0.3

Package type: Apache Hadoop 2.7

URL: download 받지 마세요!!!

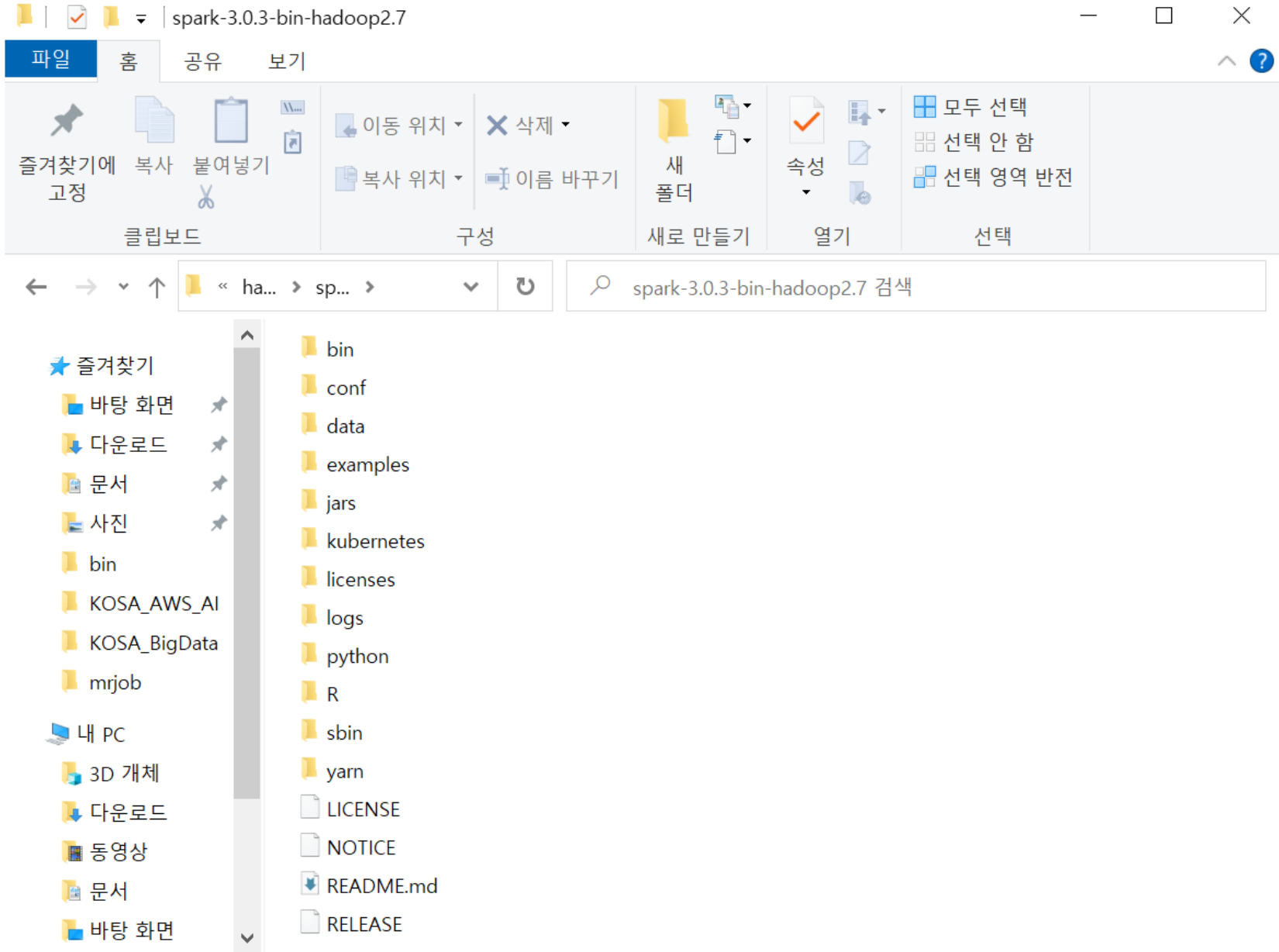
$ https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz

$ 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

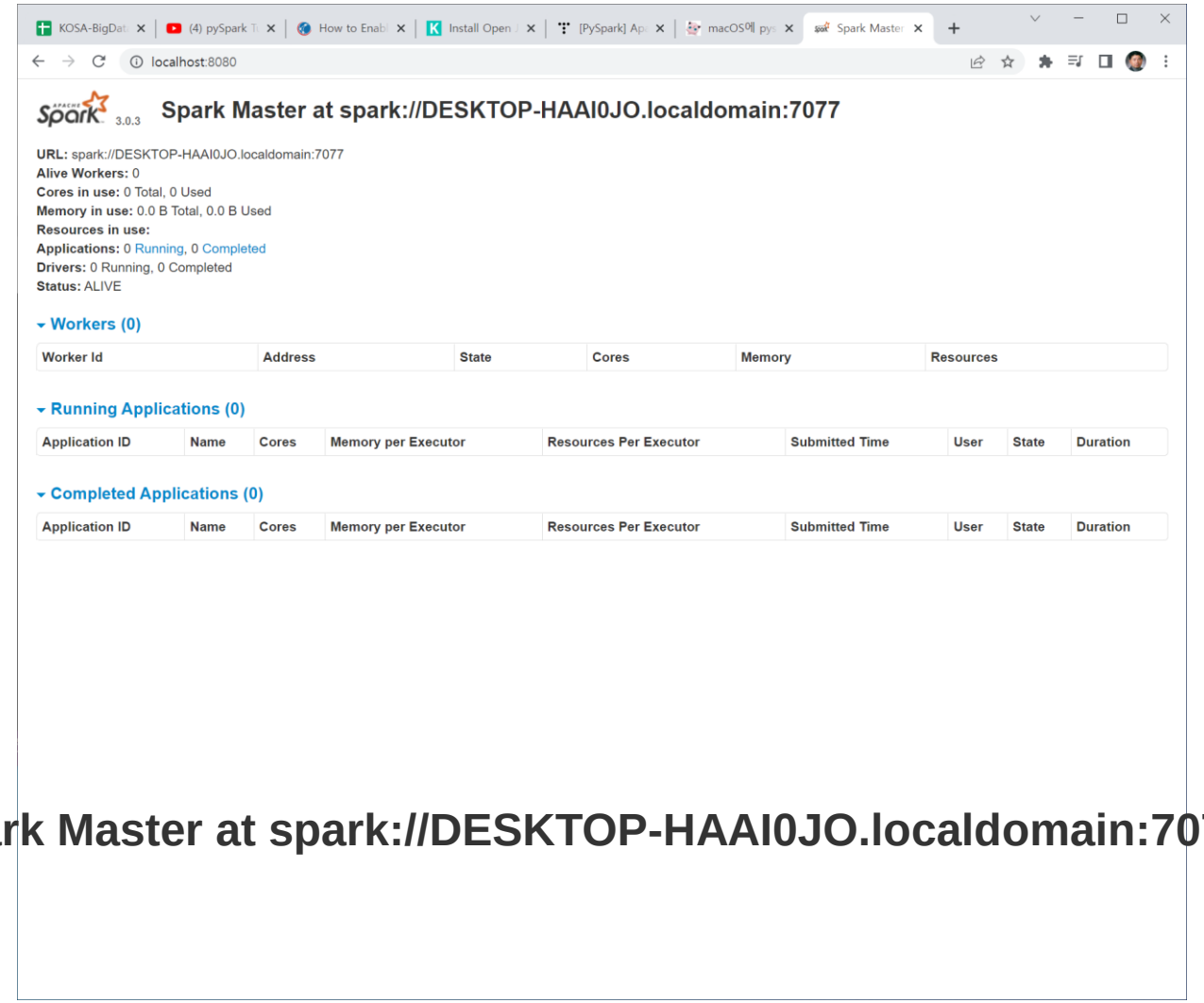
spark standalone 클러스터 모드에서 어플리케이션 실행

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

~/spark-3.0.3-bin-hadoop2.7/sbin$ ./start-master.sh

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

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/