KOSSCON 2018

기술의 발전,
그리고
기술 커뮤니티의 발전

Apache Spark과
스사모(스파크 사용자 모임)

김상우, 스사모 운영진
(dk@socar.kr)


APACHE
Spark™

Apache Spark

예전: "매우 빠른 속도의 분산 데이터 분석 시스템"

얼마 전: "다목적 분산 컴퓨팅 프레임워크"

현재: "대규모 데이터 처리를 위한 통합 분석 엔진"

현재의 Apache Spark은?

대규모 데이터를 처리하길 원하는 회사들의 기본 선택지

빠르고, 사용하기 쉬운 클러스터 컴퓨팅

기본적인 데이터 분석부터 복잡한 분석, 프로그램을 통한 데이터 처리, 통계, 스트리밍, 머신러닝까지 모두 가능

클라우드에서도 손쉽게 사용 가능

DataFrame API

Scala, Java, Python, R 로 사용 가능

```
df = spark.read.json("examples/src/main/resources/people.json")

df.filter(df["age"] > 21).show()

df.groupBy("age").count().show()
```

다양한 입출력 가능

Streaming도 똑같은 API로 처리 가능 (Structured Streaming)

MLlib (Machine Learning Library)

High Level 머신러닝 API

```
from pyspark.ml.regression import LinearRegression

lr = LinearRegression(maxIter=10, regParam=0.3)

lrModel = lr.fit(training)
```

Linear, Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Gradient-Boosted Trees 등 많은 알고리즘들이 구현되어 있음

Feature Engineering 을 위한 여러가지 기능들이 구현되어 있어 편리하게 사용

앞으로의 전망?

현재로써는 대용량 데이터 처리 시,
Spark을 대체할 만한 대체제는 별로 없음

(Apache Flink 정도 관심 가질 만 함)

머신러닝, 딥러닝 연구 및 적용시에도
대용량 데이터 처리는 단연 Apache Spark

스사모
(스파크 사용자 모임)

“우리도 제대로된 테크커뮤니티가 있었으면”

기술적 리더십

비영리, 멤버들에게 혜택을

어려운 기술을 누구나 쉽게

스사모

스파크 사용자 모임

기술적 리더십

앞서나가는 기술, 그리고 깨어있는 테크리더들

비영리, 멤버들에게 혜택을

철저한 비영리 운영

어려운 기술을 누구나 쉽게

열려있는 질답의 장, 세미나와 교육

스사모

스파크 사용자 모임

facebook.com/groups/sparkkoreauser

5,000명의 회원님들이 활동 중

Spark Day 2017

트레이닝 세션

발표세션

토론 & 네트워킹

트레이닝 세션

Spark Bootstrap

업계 최고의 전문가가 진행하는,

2시간 30분간의 실습이 포함된 Spark A to Z 트레이닝

발표 세션

Machine Learning Talks

- Spark의 과거, 현재, 미래
- Spark을 이용한 분산 머신러닝 & 딥러닝
- Deep Learning Text NLP 그리고 Spark Collaboration
- Spark Summit 2017 ML Session Top 3
- 한국어 Text Classification with Spark
- Apache Spark on Kubernetes

토론 & 네트워킹

스사모 전통의 피맥시간

Spark Day 2018 ?

TBD...

최근의 활동

온라인 활동은 여전히 꽤 활발한 편

오프라인 활동은 소규모 밋업 위주

오프라인에서 주로 이야기되는 주제는 머신러닝, 딥러닝과 스트리밍

커뮤니티!

기술은 흐름에 따라 변하지만, 사람은 남는다

비슷한 일을 하며 비슷한 목표를 가진 사람들끼리 서로 돕기

서로 배워가며 성장하는 소중한 경험을 남기기

감사합니다!