클라우드 환경 Spark으로 분산 컴퓨팅

Spark은 다양한 일을 할 수 있습니다.

Spark이 할 수 있는 일들

- 분산 데이터 분석
- 분산 기계학습
- 분산 ETL
- 분산 강화 학습
- 등, 분산 컴퓨팅이 필요한 모든 분야에 활용 가능

잠깐, 광고

클라우드 환경에서 Spark 쉽게 사용하기

Wadal - https://github.com/haje01/wadal

AWS EMR 클러스터 환경에서 Python/R을 편리하게 사용

```
> bin/create_cluster mypro
> bin/wait_ready mypro
> bin/jupyter mypro
```

사례 1 - Spark을 활용한 로그 Cleansing

상황

- 다루기 까다로운 원본 파일 형태
  - RAR 압축
  - 200여개의 서버당 하나의 압축 파일 안에 50여개의 서로 다른 로그 파일이 함께 압축 (모두)
- 대용량과 많은 파일들
  - 75일분, 50TB 분량의 분석 대상 로그
  - 15,000여개의 원본 파일들

Spark을 활용해 해결

- 20대(160코어, 320GB 램)로 구성된 EMR 클러스터 생성
- 각 장비에서 다음과 같은 작업 수행
  - 원본 파일 다운로드 -> RAR 압축 해제 -> 클렌징 후 개별 파일 Snappy로 압축 -> S3 업로드
- 12시간 동안 모든 파일에 대한 Cleansing 완료

코드

좋아진 것

- 원하는 기간, 원하는 타입의 로그들만 Spark에서 볼 수 있게 됨
- 매번 RAR파일 압축 해제 없이 Snappy 포맷으로 바로 읽어들임
- 빠른 작업 시간으로 Iterative한 ETL이 가능
- 급박한 분석 이슈를 50,000원 정도 비용으로 해결

팁

- 각 Executor 노드에 필요한 프로그램들이 미리 설치되어야 함 (Bootstrapping)
- Driver에 너무 많은 결과를 리턴하지 말것
- 분산 처리되는 스크립트 내에서는 Transform이나 Action 명령을 내릴 수 없다. (Driver에서만 가능)

사례 2 - 분산 기계학습

분산 기계학습 사례

- edX CS120X의 Lab3
  - https://courses.edx.org/courses/course-v1:BerkeleyX+CS120x+2T2016/info)
- Click-through Rate 예측
  - Kaggle에도 올라온 챌린지(Hashing 된 피쳐들)
  - https://www.kaggle.com/c/criteo-display-ad-challenge
- One-Hot Encoding과 Logistic Regression을 Spark으로 분산 처리

코드

좋아진 것

- 많은 데이터에 대한 반복적인 Feature engineering 작업에 장점
- 분산 학습과 예측이 가능

팁

- `ml` 라이브러리의 다양한 모듈을 사용하면, 저수준 작업을 많이 줄일 수 있을 듯

앞으로는..

강화 학습을 Spark에서 해보고 싶습니다~

- Spark에서 분산 강화학습으로 전력량 예측
- http://www.slideshare.net/ImpetusInfo/spark-use-case-distributed-reinforcement-learning-for-electricity-market-bidding-with-spark

감사합니다.