데이터 엔지니어로 6개월 살아보기

박두호

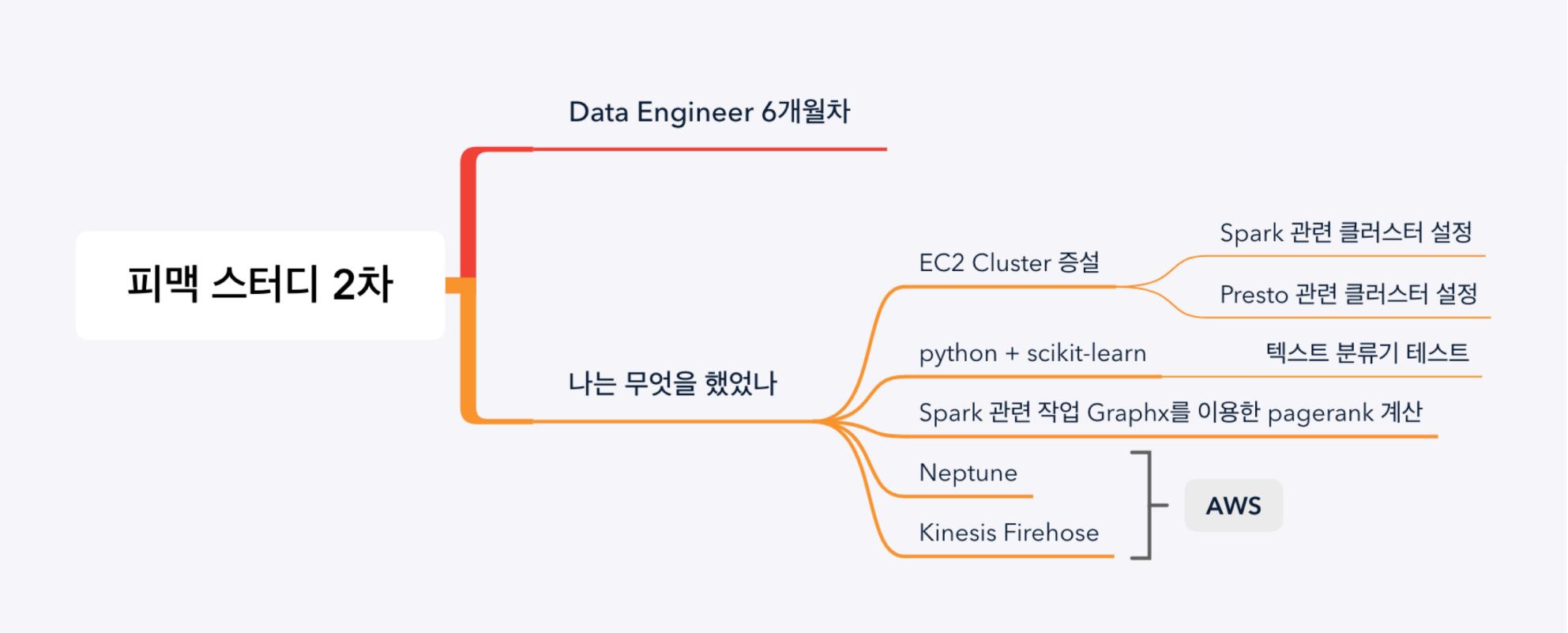
May, 18th 2019

exit2dream

What did I do as a ***Data Engineer*** for six months?

Agenda

- ✓ Data Engineer?
- ✓ What did I do?



Data Engineer

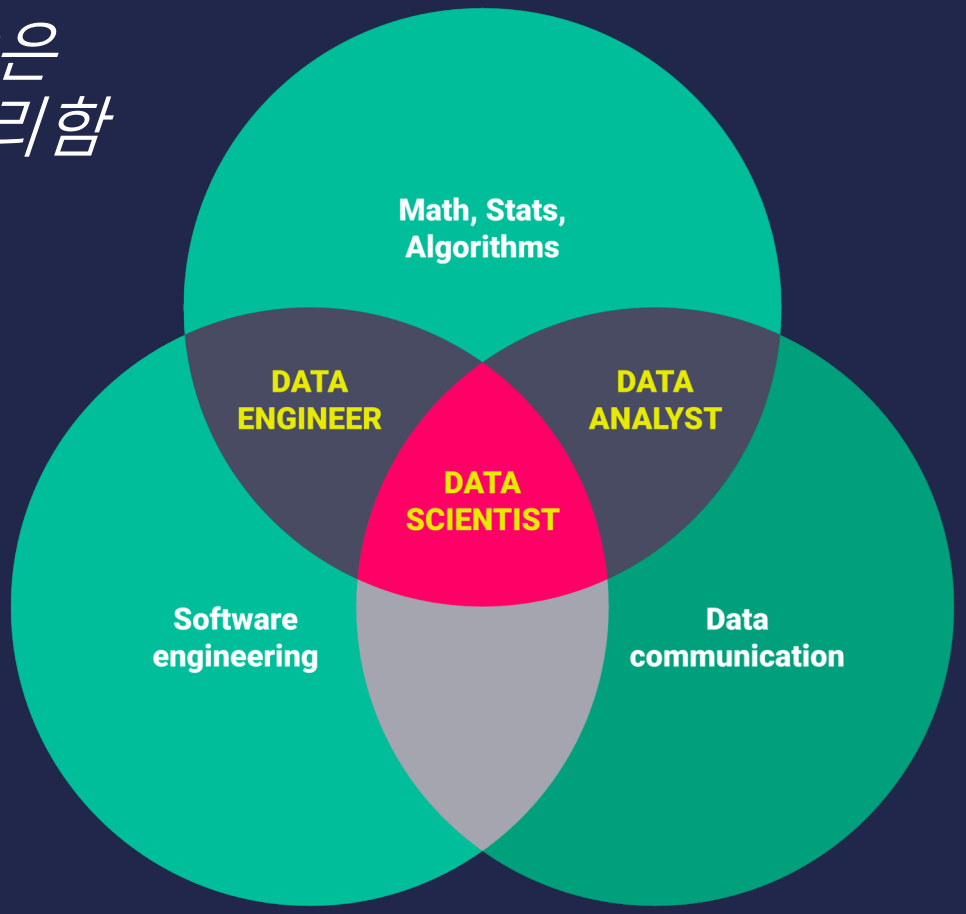
= Data **E**xtract
Transformation
Loading

관련 모든 업무를 담당하는 사람

Ref : https://www.slideshare.net/zzsza/data-science-intro

Develops, constructs, tests, and maintains architectures, Such as databases and large-scale processing systems

데이터베이스 및 대규모 처리 시스템과 같은 아키텍처를 개발, 구성, 테스트 및 유지 관리함



img ref: https://www.datacamp.com/community/blog/data-scientist-vs-data-engineer
https://www.quora.com/What-are-the-roles-and-responsibilities-of-a-data-scientist-And-why-the-term-Scientist

As a Data Engineer

Graph Theory & Network Analysis

Spark Cluster

Graph DB

Text Classification
Data Warehouse/Lake
AWS kinesis firehose

Graph Theory & Network Analysis

✓정의

- 여러 엔티티 간의 상관 관계를 그래프 형태의 네트워크로 나타내어 분석하는 기법

✓문제

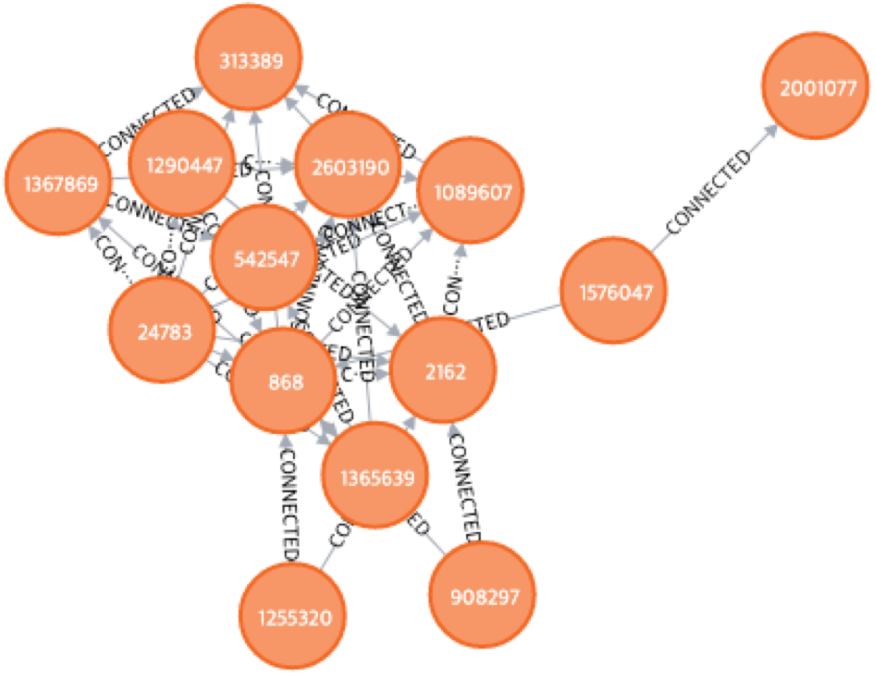
- 데이터 분석량이 많음

✓해결방안

- Graph DB, Spark Graph 라이브러리

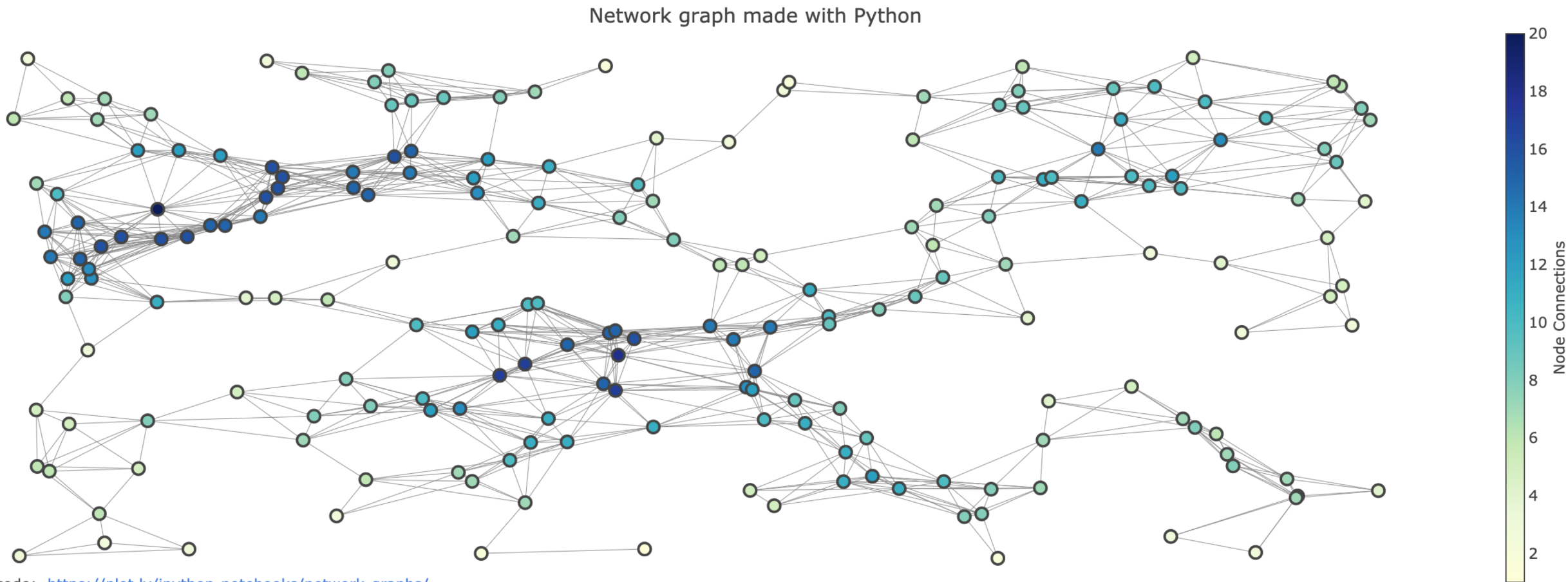
✓사용 기술

- Spark GraphX, GraphFrames
- Neo4j, AWS Neptune, S2Graph

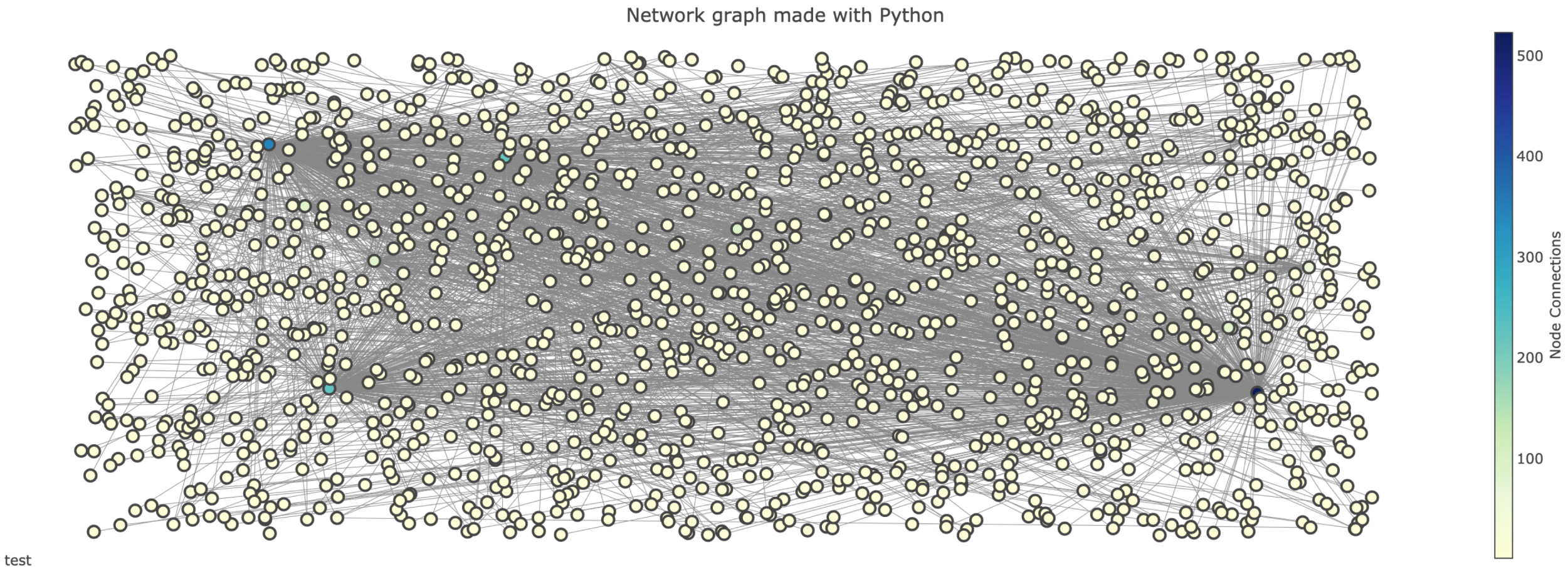


S2Graph

- *카카오 오픈 소스 그래프 DB*
- *http://s2graph.incubator.apache.org/*
- *http://tech.kakao.com/2016/01/29/opensource-1-s2graph/*



Python code: https://plot.ly/ipython-notebooks/network-graphs/



https://plot.ly/python/network-graphs/
https://www.datacamp.com/community/tutorials/social-network-analysis-python

Spark Graph 라이브러리

✓ 정의

- Apache Spark's API for graphs and graph-parallel computation https://spark.apache.org/graphx/

✓ 종류

- GraphX
- GraphFrame

✓ 제공되는 기능

- GraphX (RDD Base)
- GraphFrame (DataFrame Base)

https://docs.databricks.com/spark/latest/graph-analysis/graphframes/graph-analysis-tutorial.html

→ *계산량이 많아서… 우리 클러스터에서 안돌아가네..어쩌지.. ??*

Spark Cluster 증설

✓ 목적

- 기존 클러스터에서 처리가 불가능한 대용량 데이터 작업 처리를 위함

Amazon EC2

구분	사양
기존	r4.2xlarge * 1 || m4.xlarge * 2
변경 후	r4.2xlarge * 1 || r5.2xlarge * 2

✓ 방법

- EC2 사양 변경 후 노드별 yarn-site.xml 수정

```
<property>
<name>yarn.nodemanager.resource.memory-mb</name>
<value>61440</value>
</property>
<property>
<name>yarn.scheduler.maximum-allocation-mb</name>
<value>61440</value>
</property>
```

모델	vCPU	메모리 (GiB)	스토리지	네트워킹 성능 (Gbps)
r4.large	2	15.25	EBS 전용	최대 10
r4.xlarge	4	30.5	EBS 전용	최대 10
r4.2xlarge	8	61	EBS 전용	최대 10
r4.4xlarge	16	122	EBS 전용	최대 10
r4.8xlarge	32	244	EBS 전용	10
r4.16xlarge	64	488	EBS 전용	25

모델	vCPU	메모리 (GiB)	스토리지 (GB)	전용 EBS 대역폭(Mbps)	네트워킹 성능 (Gbps)
r5.large	2	16	EBS 전용	최대 3,500	최대 10
r5.xlarge	4	32	EBS 전용	최대 3,500	최대 10
r5.2xlarge	8	64	EBS 전용	최대 3,500	최대 10
r5.4xlarge	16	128	EBS 전용	3,500	최대 10
r5.12xlarge	48	384	EBS 전용	7,000	10
r5.24xlarge	96	768	EBS 전용	14,000	25

✓ 결과

- 총 120G 메모리 할당

Spark Cluster 증설

✓ 테스트 데이터량

- 사용자간의 연결 상태를 기록한 데이터를 대상으로 user간의 pagerank 를 계산

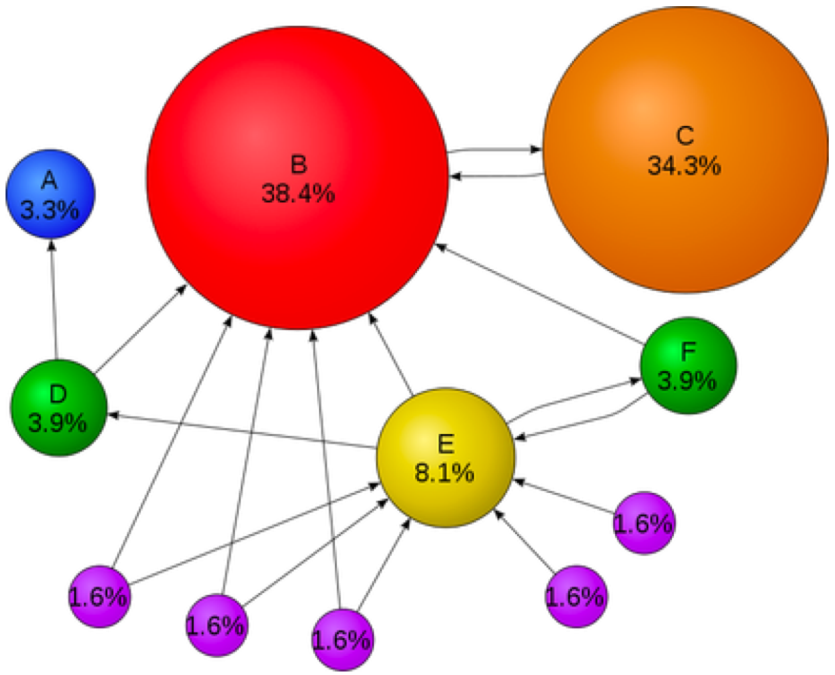
✓ 데이터량

- 96만 유저 데이터 대상, 연결 상태를 기록한 데이터 3,192만

✓ 결과요약

구분	데이터량	실행시간
변경 전	• 200만 건 대상 • 1,000만 건 대상 • 2,000만 건 대상	약 216초 소요 약 809초 소요 실행 불가
변경 후	• 2,000만 건 대상 • 3,200만 건 대상	약 327초 소요 약 521초 소요

→ *메모리는 많은데… 왜 100% 안쓰지???*

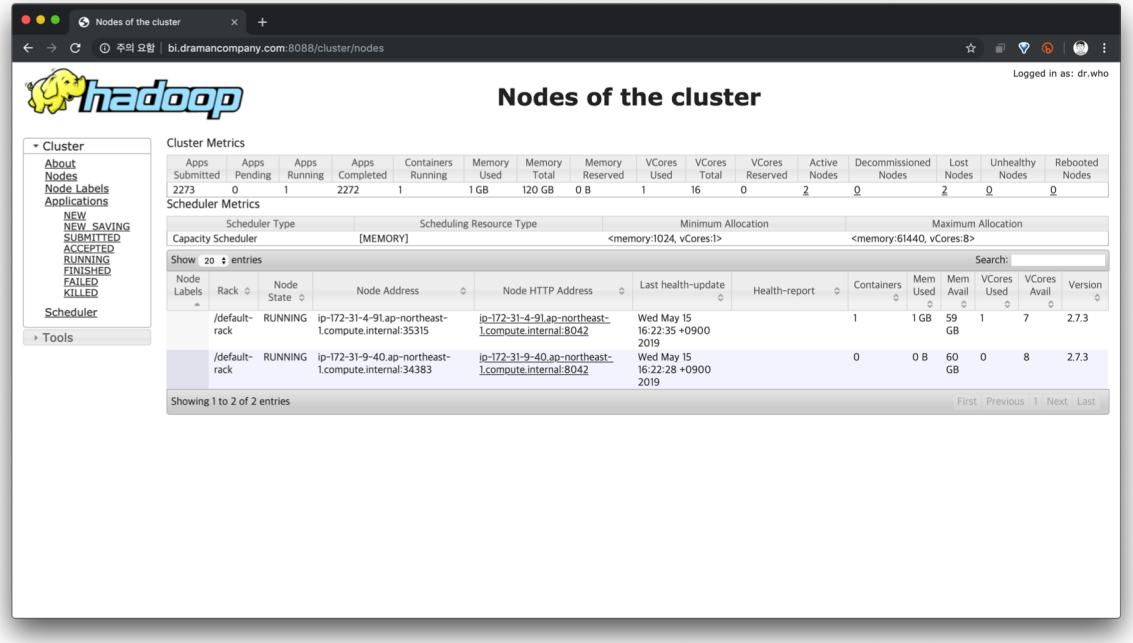


페이지랭크와 그 관계를 도식화한 그림. A, B, C 등은 페이지를 나타내고, 숫자는 PageRank를 의미한다. C의 경우 B에서 링크를 걸었다는 것만으로도 PageRank값이 높게 책정됨을 볼 수 있다.
(출처: https://sungmooncho.com/2012/08/26/pagerank/)

Spark Cluster 증설

✓ Spark Dynamic Allocation을 활성화

- Yarn 설정 변경
  - Map-reduce알고리즘을 spark으로 실행하도록



✓ Spark Default 옵션 변경

- Spark-default 파일 수정
  - 기존 애플리케이션도 동작할때 옵션이 적용될 수 있도록

	Property	Value
0	spark.app.id	application_1556267238281_0453
1	spark.app.name	Jupyter
2	spark.cores.max	2
3	spark.driver.appUIAddress	http://ec2-52-193-244-0.ap-northeast-1.compute...
4	spark.driver.host	172.31.2.146
5	spark.driver.memory	16g
6	spark.driver.port	45531
7	spark.dynamicAllocation.enabled	True
8	spark.executor.id	driver

→ *근데… 그래프 분석 결과는 어떻게 서비스할 수 있지??*

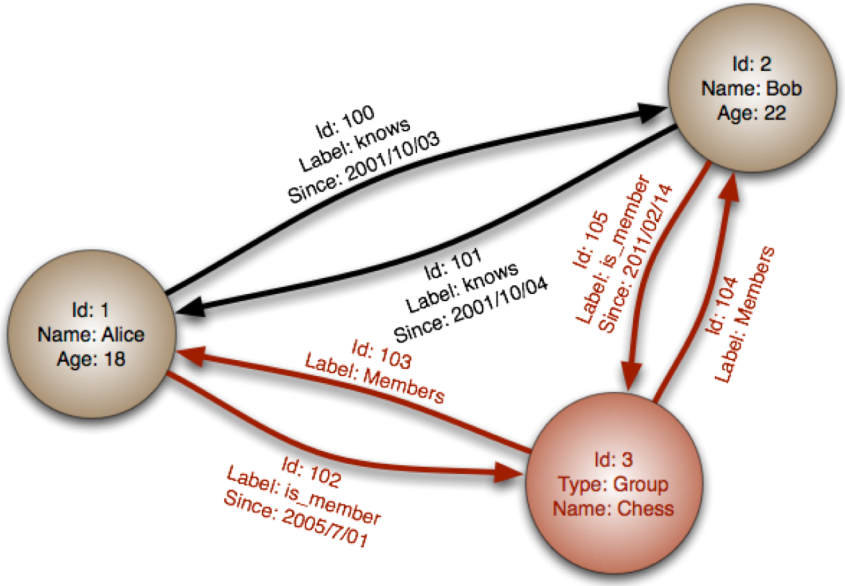
Graph DB

✓정의

- 시맨틱 쿼리를 위해 노드, 엣지, 프로퍼티와 함께 **그래프** 구조를 사용하여 데이터를 표현하고 저장하는 데이터베이스

✓관련 솔루션

- Neo4j, AWS Neptune



Neo4J

✓ 개요

- 가장 대중적인 GraphDB
- 자바로 구현되어 있으며 Cypher Query Language를 사용하여 다른 언어로 작성된 소프트웨어로부터 접근이 가능

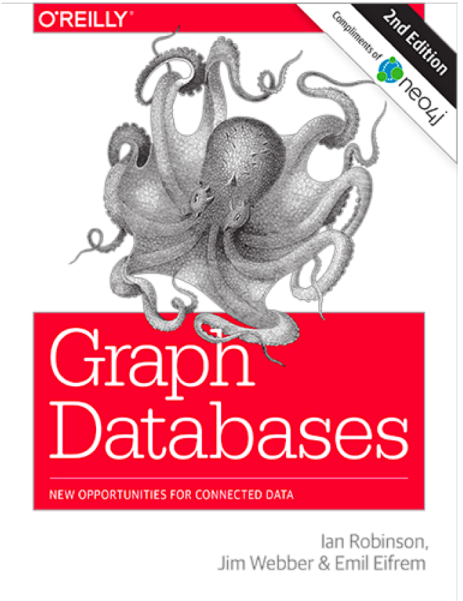
https://neo4j.com

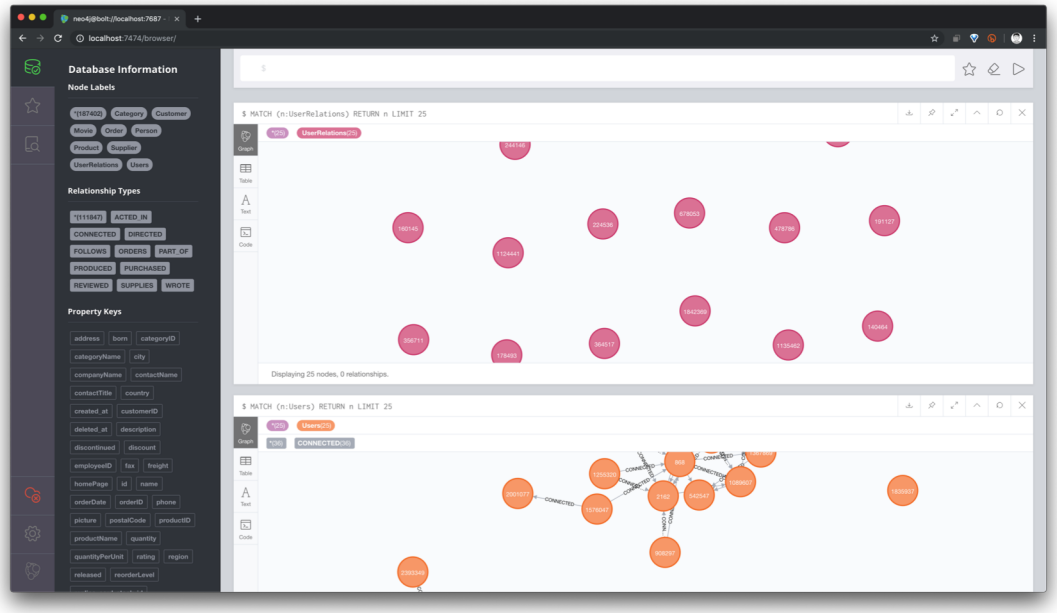
✓ 장점

- 시각화가 쉽고 접근하기 쉬움
- 참고자료가 다양함

✓ 단점

- 비싸다...?





AWS Neptune

✓ 개요

- AWS에서 제공하는 managed GraphDB 서비스
- Apache TinkerPop Gremlin 및 SPARQL을 지원

https://aws.amazon.com/ko/neptune/

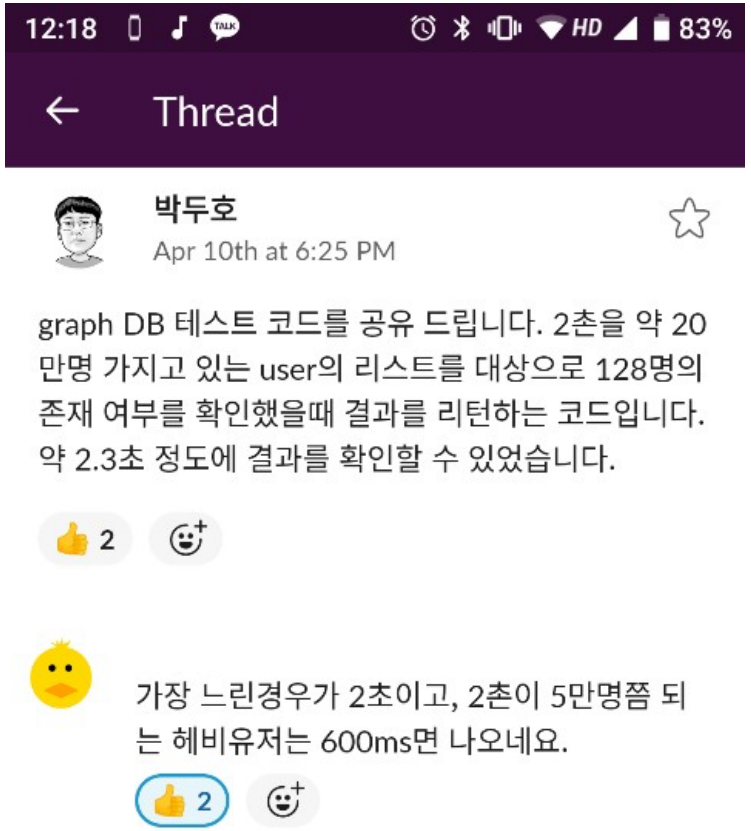
Amazon Neptune

✓ 장점

- AWS서비스와 연동이 쉬움
- 자동 확장 및 다양한 클라우드 기능을 포함

✓ 단점

- 시각화 따로 없음....



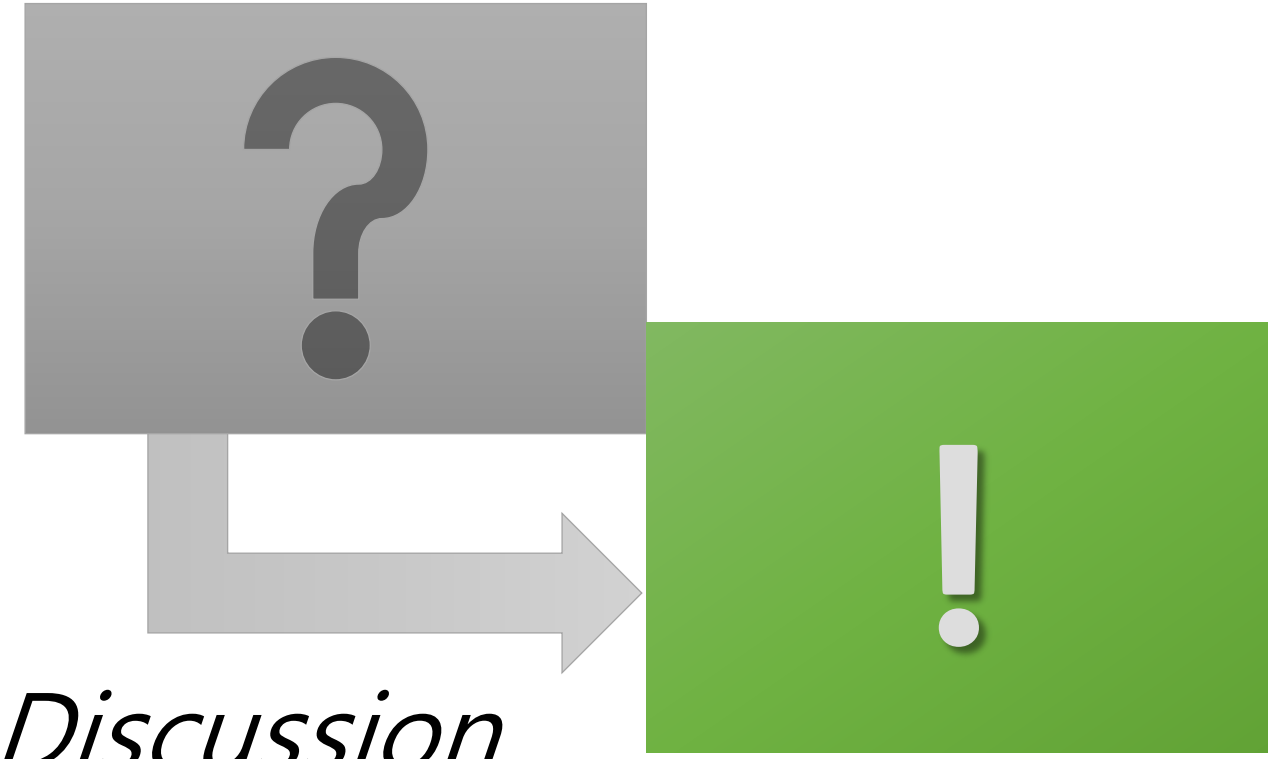
Data Engineer

Graph Theory & Network Analysis

Spark Cluster

Graph DB

Text Classification
Data Warehouse/Lake
AWS kinesis firehose



Discussion

discussion 미국·영국 [dɪˈskʌʃn] 영국식 ★★

1. 논의, 상의
2. (어떤 주제를 말이나 글로) 논한 것, 논고