How to become a Googler ?

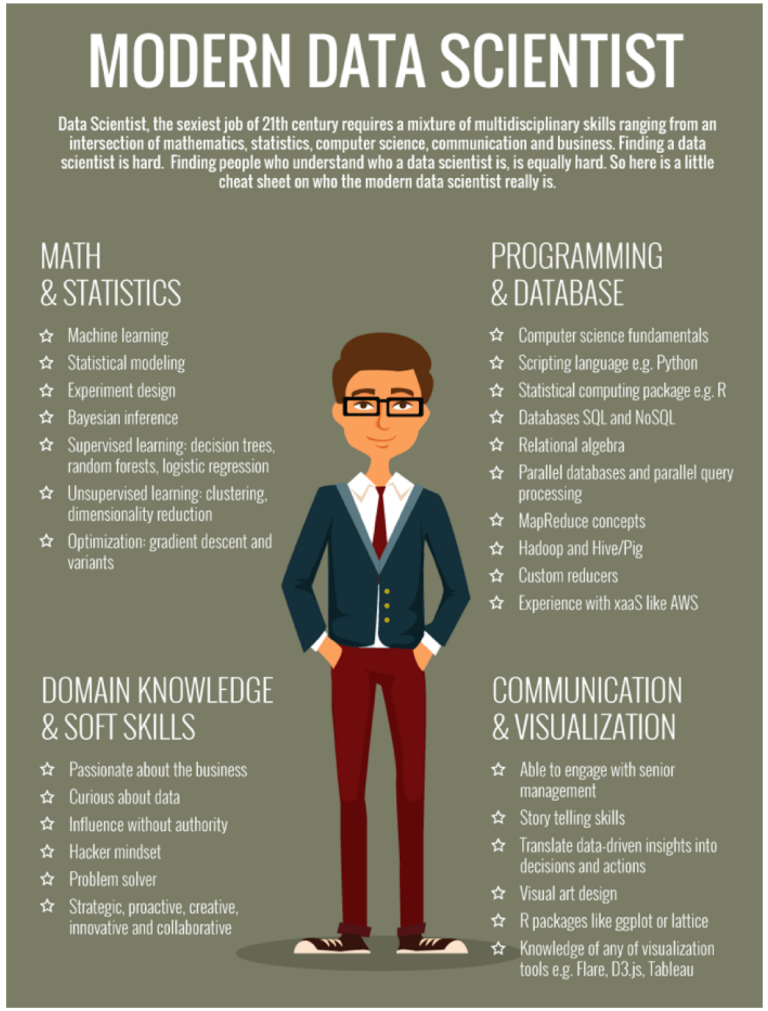
NYC Data Science BootCamp

R Shiny Project

SangYeon Choi

Why we are here?

- **Why do we study data science here?**
- **There are different reasons, but the greatest goal is to get a job as a Data Scientist.**



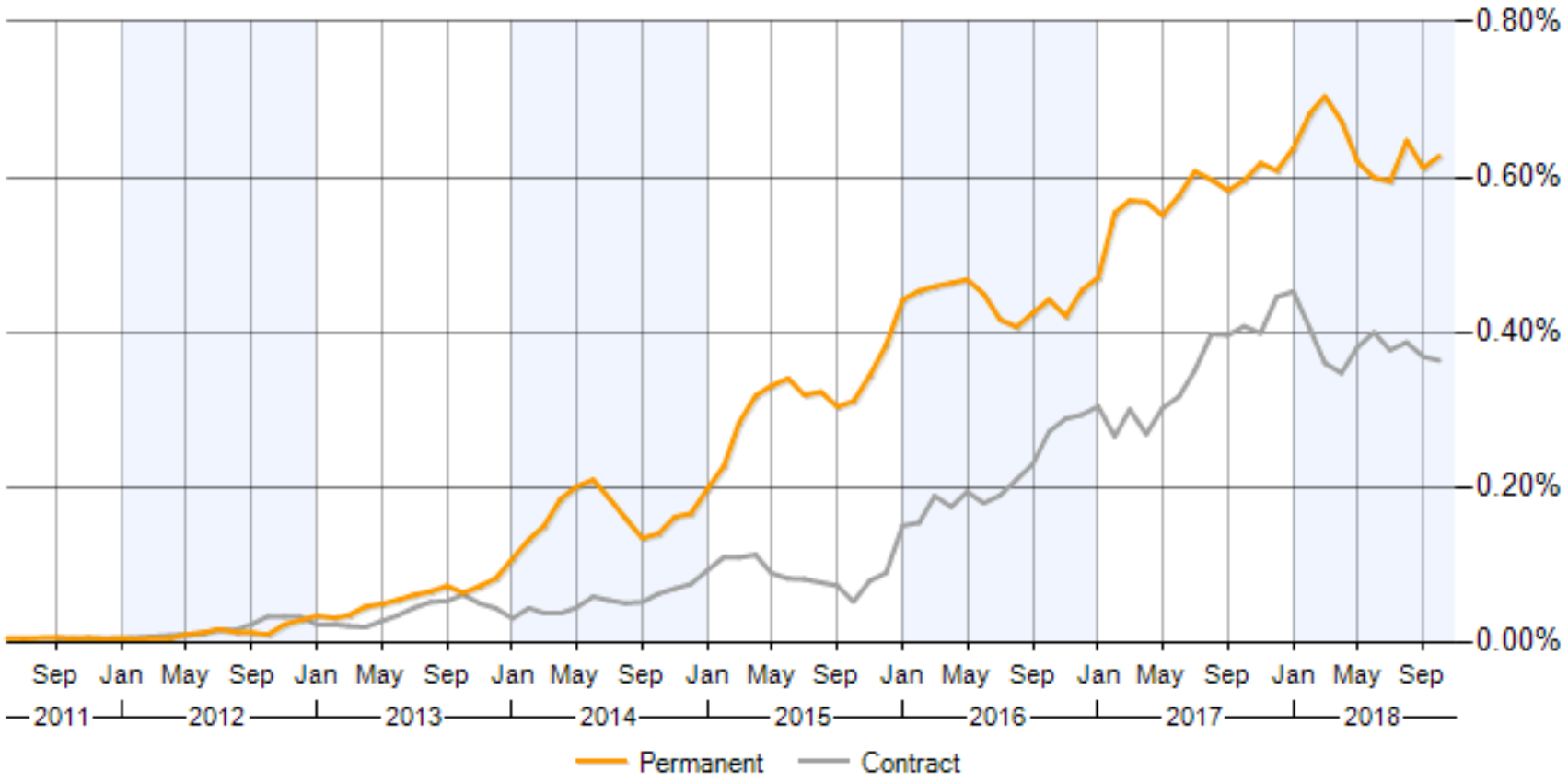
For Job?

- To-do list

1. **Skill up! – Ability of Data Science**
2. **Interview Skill**
3. **Network**
4. **Find Job opportunity**

Yes! We Know

- **Data Scientist is a prospective job**
- **Since interest in Big Data has increased, There is demand in all areas.**



<Data Scientist Job demand Trend >

What skills should I have? Which company can I join?

Male
70% of Data Scientists in our research were male

2 Languages
Data scientists speak at least 1 foreign language on average

2 years
This is a new profession. The median experience as data scientists of professionals in our research was 2 years

4.5 years
People who work as data scientists currently have a median work experience of 4.5 years (including previous positions)

R and/or Python
More than 50% of the data scientists in our research work in R and/or Python

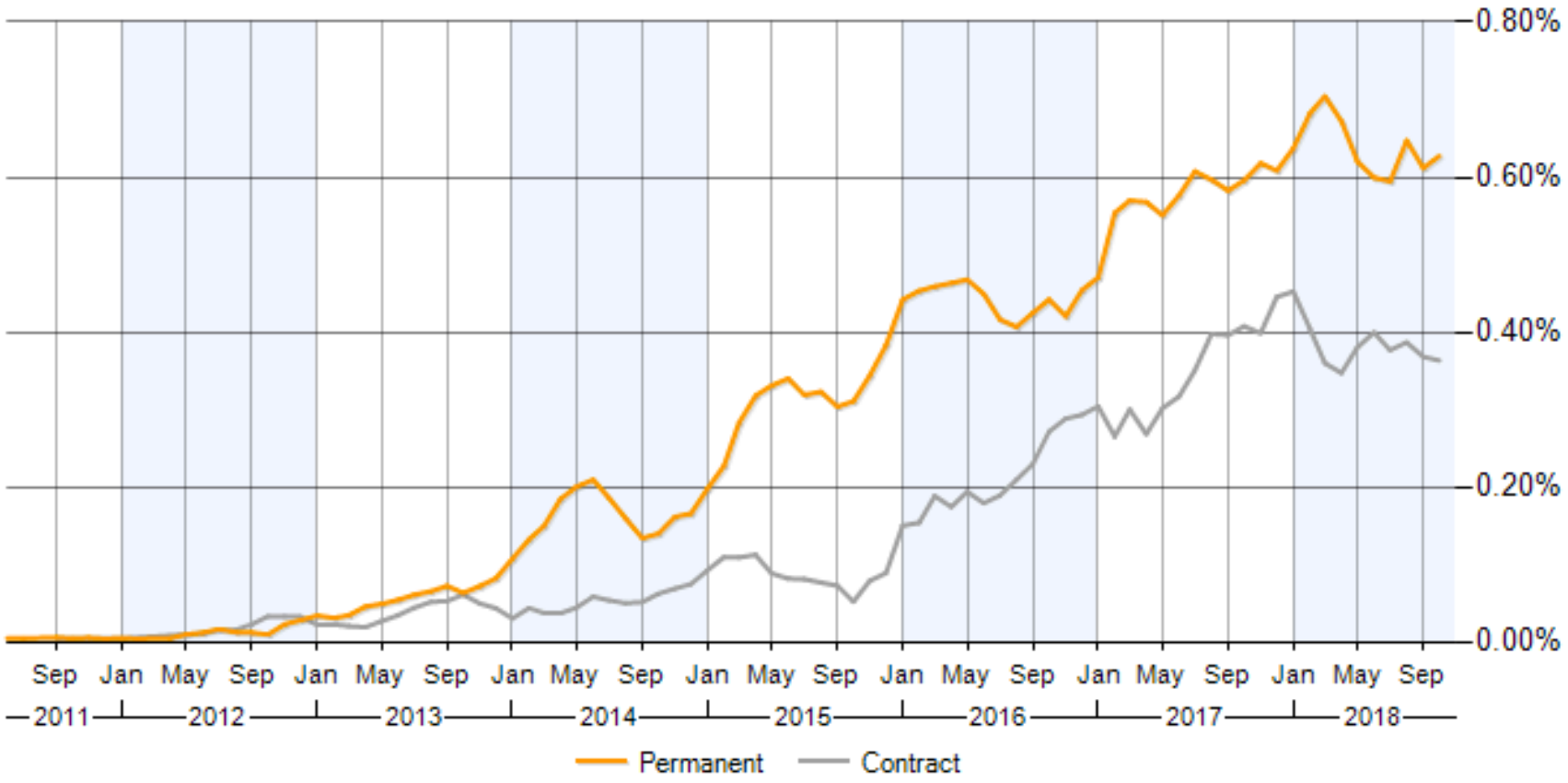
Master or PhD
75% of data scientists have a PhD (27%) or a Master (48%) degree

amazon NETFLIX Google airbnb intel Uber 365DataScience Tencent

OK! Let's find out from data!

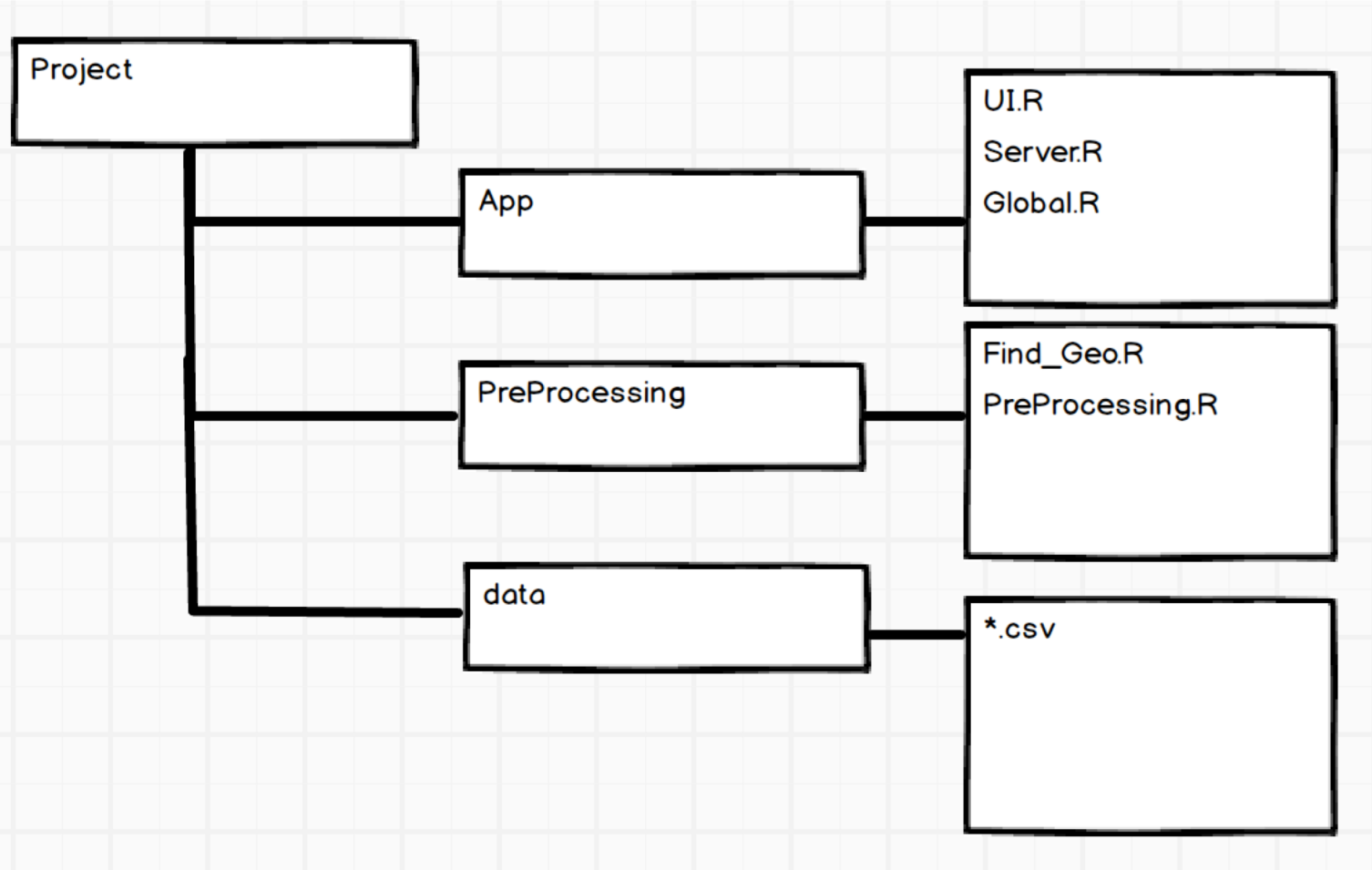
Yes! We Know

- **Data Scientist is a prospective job**
- **Since interest in Big Data has increased, There is demand in all areas.**



<Data Scientist Job demand Trend >

Project Structure



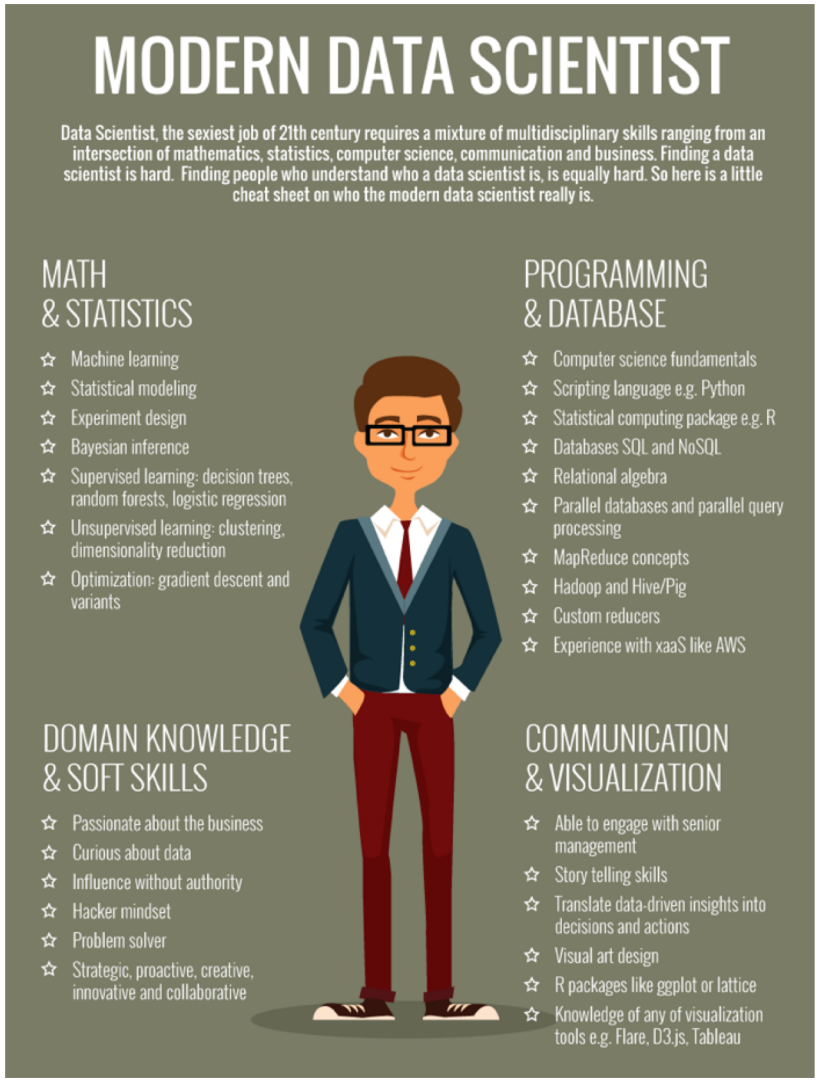
- **R Shiny Component**
  : **Shiny Dashboard / ggplot / leaflet**

- **Find_Geo.R**
  : **Transform Geo lat/lot**

- **PreProcessing.R**
  : **Add New Feature / ETL / Missing Value**

Why we are here?

- 우리가 여기서 데이터 사이언스를 공부하는 이유는 무엇인가?
- 각자 다른 전공과 다른 이유가 있지만 최대의 목적은 Data Scientist로 전직하여 Job을 구하는 일이다.



1) Data Science Skill up

- 급증하는 데이터 처리에서 발생하는 Issue 해결 필요
- HTTP/OLTP/ETL 등 다양한 Data Interface에 대응이 필요함

2) Job

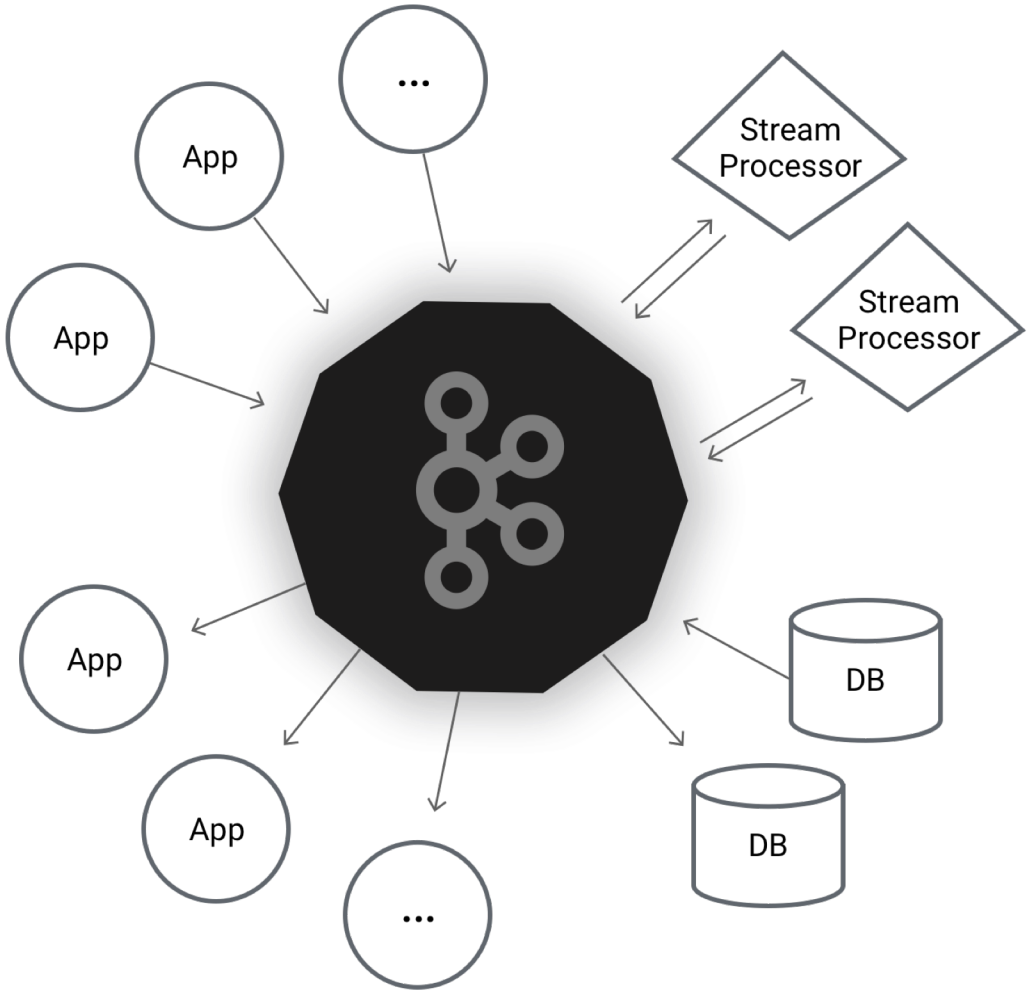
- 높은 처리량을 위한 메세지 설계
- Scale-out 및 클러스터링이 가능한 시스템 구축 가능

3) Data Pipeline에 대한 유연성 강화

- 복잡도를 줄이고 다양한 Pipe line을 구성할 수 있음

1. What is the Kafka?

- 높은 throughput을 가진 대용량 Massage Queue
- Pub / Sub 구조로 이루어져 서로 관계없이 대용량의 데이터를 IF하고 관리할 수 있다.



1) 11’ Linkedin에서 출발

- 급증하는 데이터 처리에서 발생하는 Issue 해결 필요
- HTTP/OLTP/ETL 등 다양한 Data Interface에 대응이 필요함

2) Disk기반의 데이터 처리량 강화

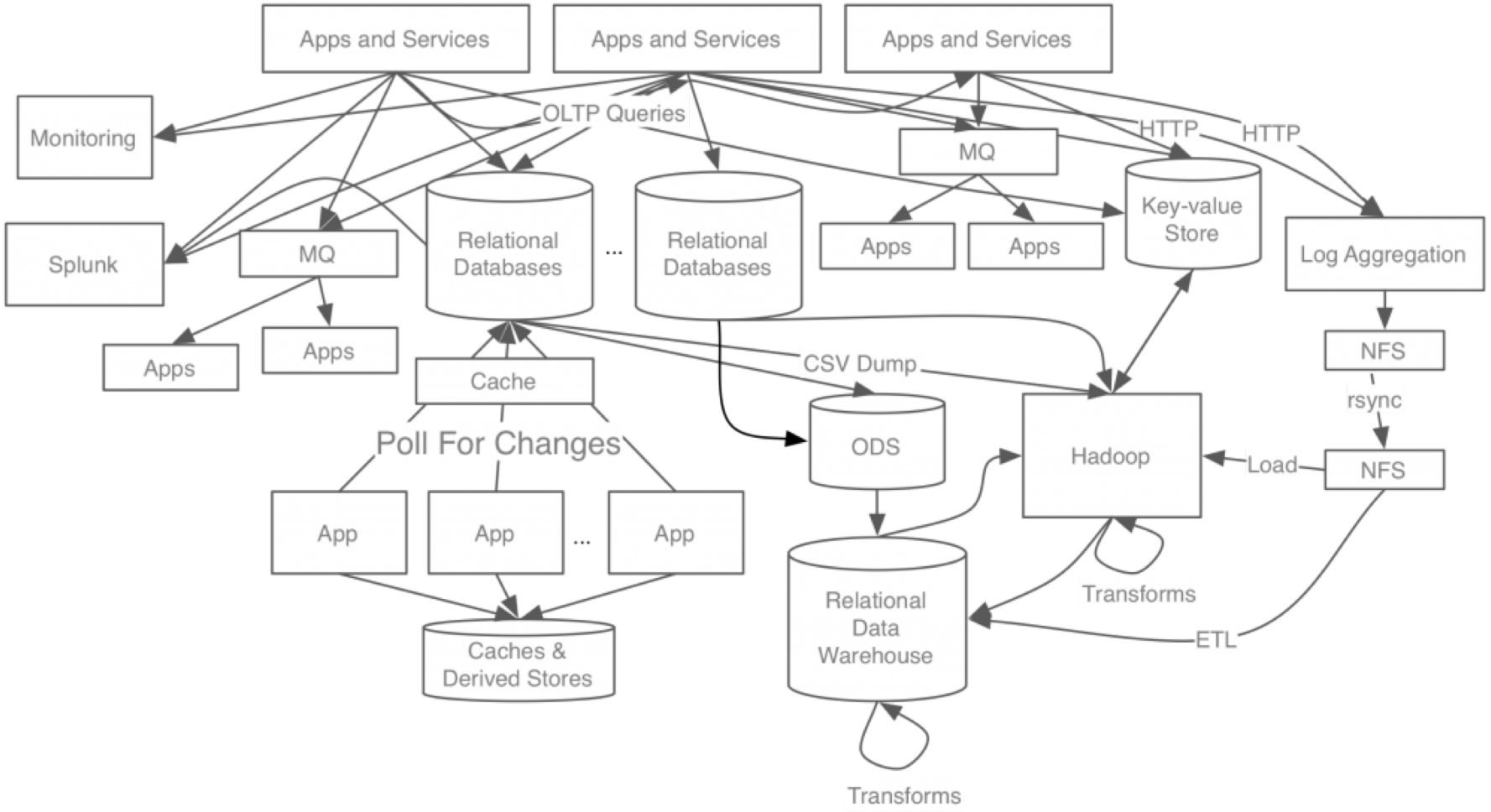
- 높은 처리량을 위한 메세지 설계
- Scale-out 및 클러스터링이 가능한 시스템 구축 가능

3) Data Pipeline에 대한 유연성 강화

- 복잡도를 줄이고 다양한 Pipe line을 구성할 수 있음

2. Kafka In LinkedIn - (1)

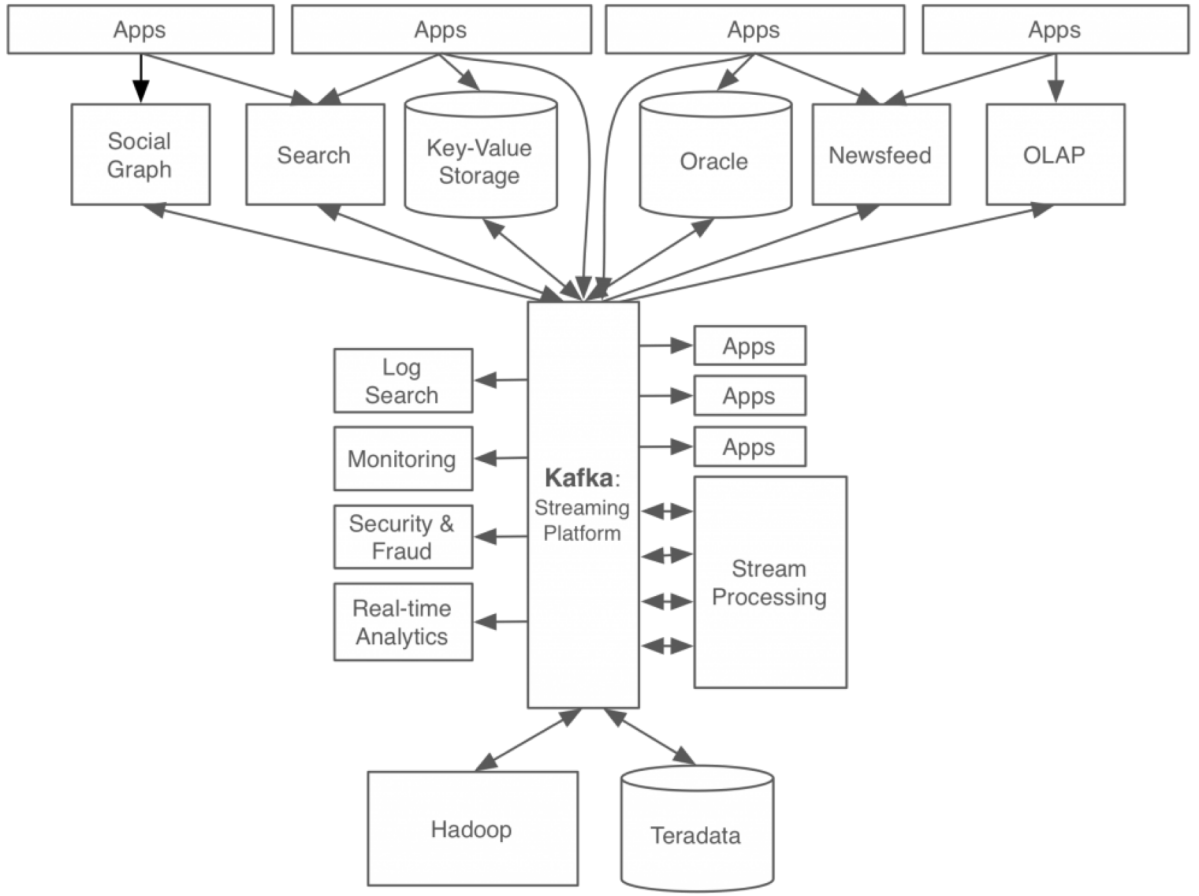
- Kafka 도입 이전의 Linkedin Data Stream Architecture
- 다양한 OSS 및 System간의 연동으로 운영하기 어려우며 복잡하여 이슈가 많음



< https://www.confluent.io/blog/stream-data-platform-1/>

2. Kafka In LinkedIn - (2)

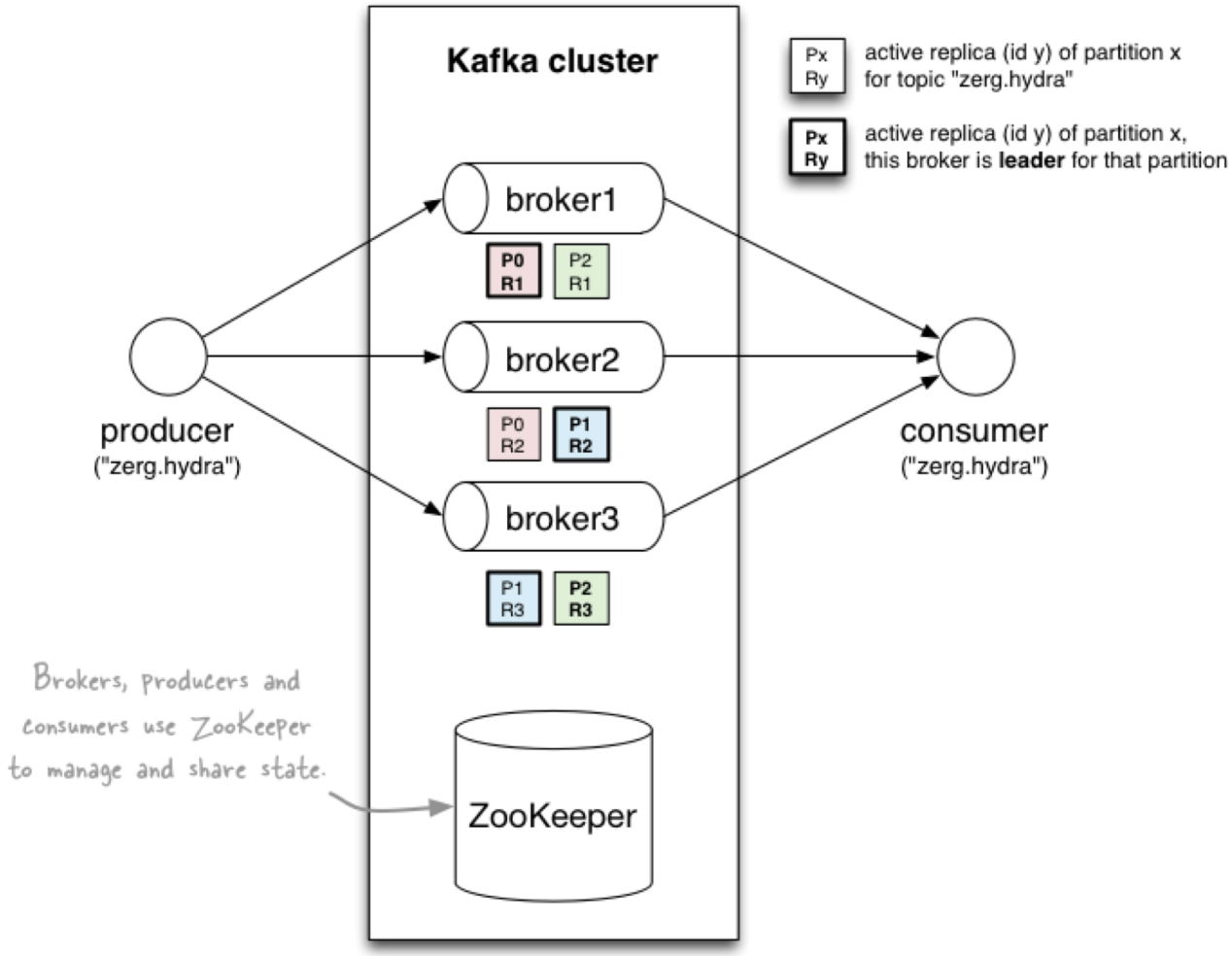
- Kafka 도입 이후의 Linkedin Data Stream Architecture
- 다양한 Data Source 및 Consumer를 한번에 정리할 수 있어 Simple하고 scalable한 Platform을 구축할 수 있다.



< https://www.confluent.io/blog/stream-data-platform-1/>

3. 구성요소

- Kafka에도 다양한 구성 요소 및 개념이 존재
- 구성요소에 대한 정확한 개념이 있어야 Operation이 가능하다.



< https://www.confluent.io/blog/stream-data-platform-1/>

- **Topic**
  - Data를 저장하는 단위
- **Producer**
  - 데이터를 하여 Broker에게 전달하는 주체
- **Broker Server**
  - Topic을 관리하여 데이터를 저장하는 Scale-out이 가능한 Server
- **Consumer (group)**
  - 데이터를 제공 받는 주체 (단위)
- **Zookeeper**
  - Offset 관리 (0.9 이하)
  - Broker Server Cluster management

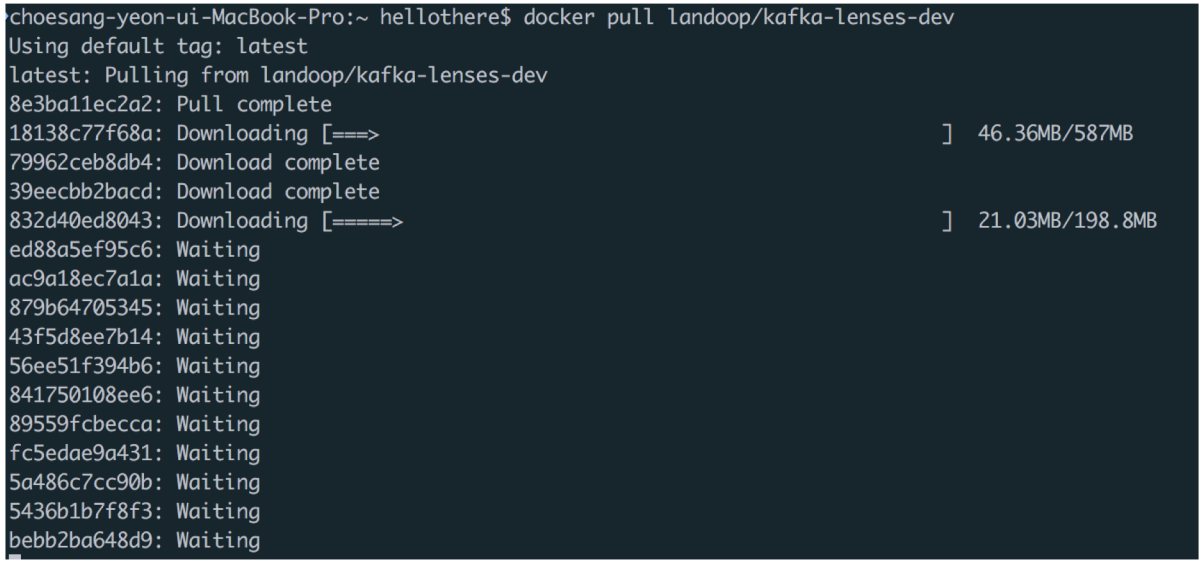
7. Install Lenses – (1)

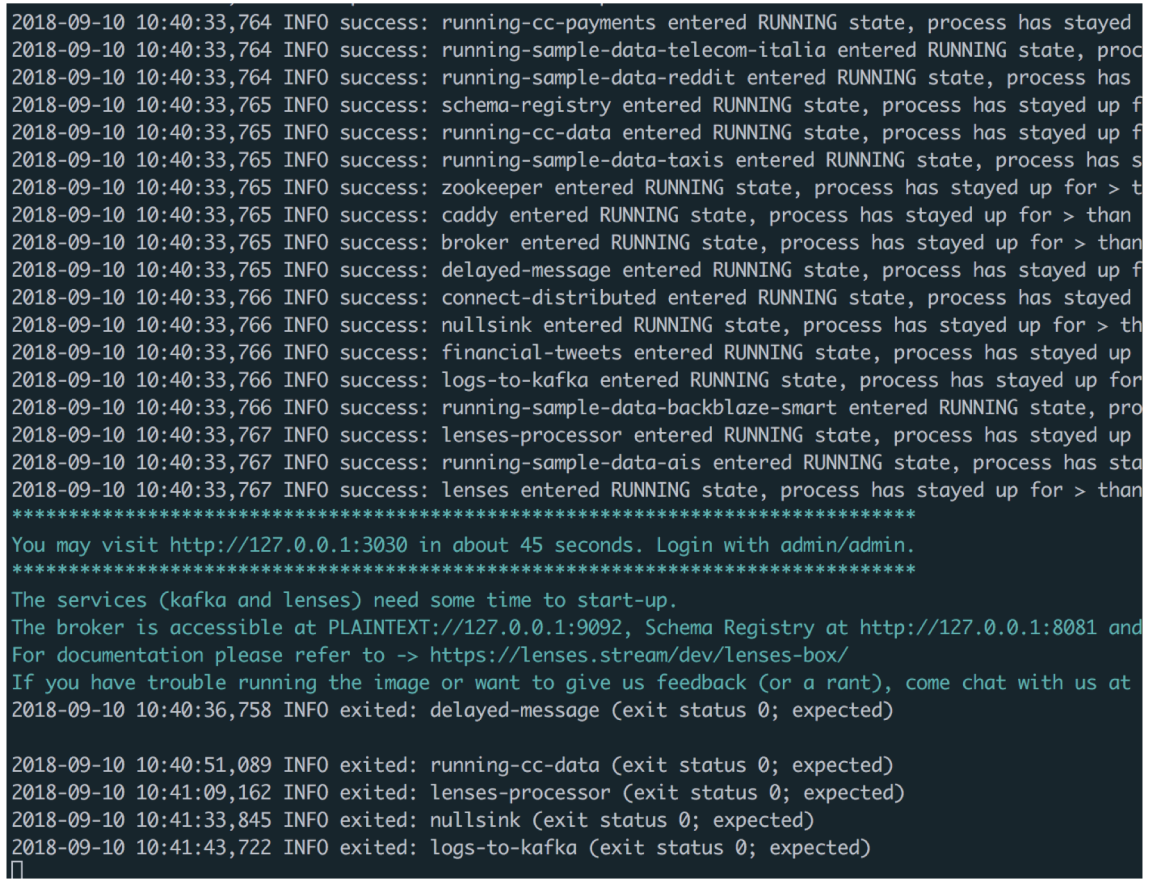
- **Lenses Box Docker Install**

$> docker pull landoop/kafka-lenses-dev

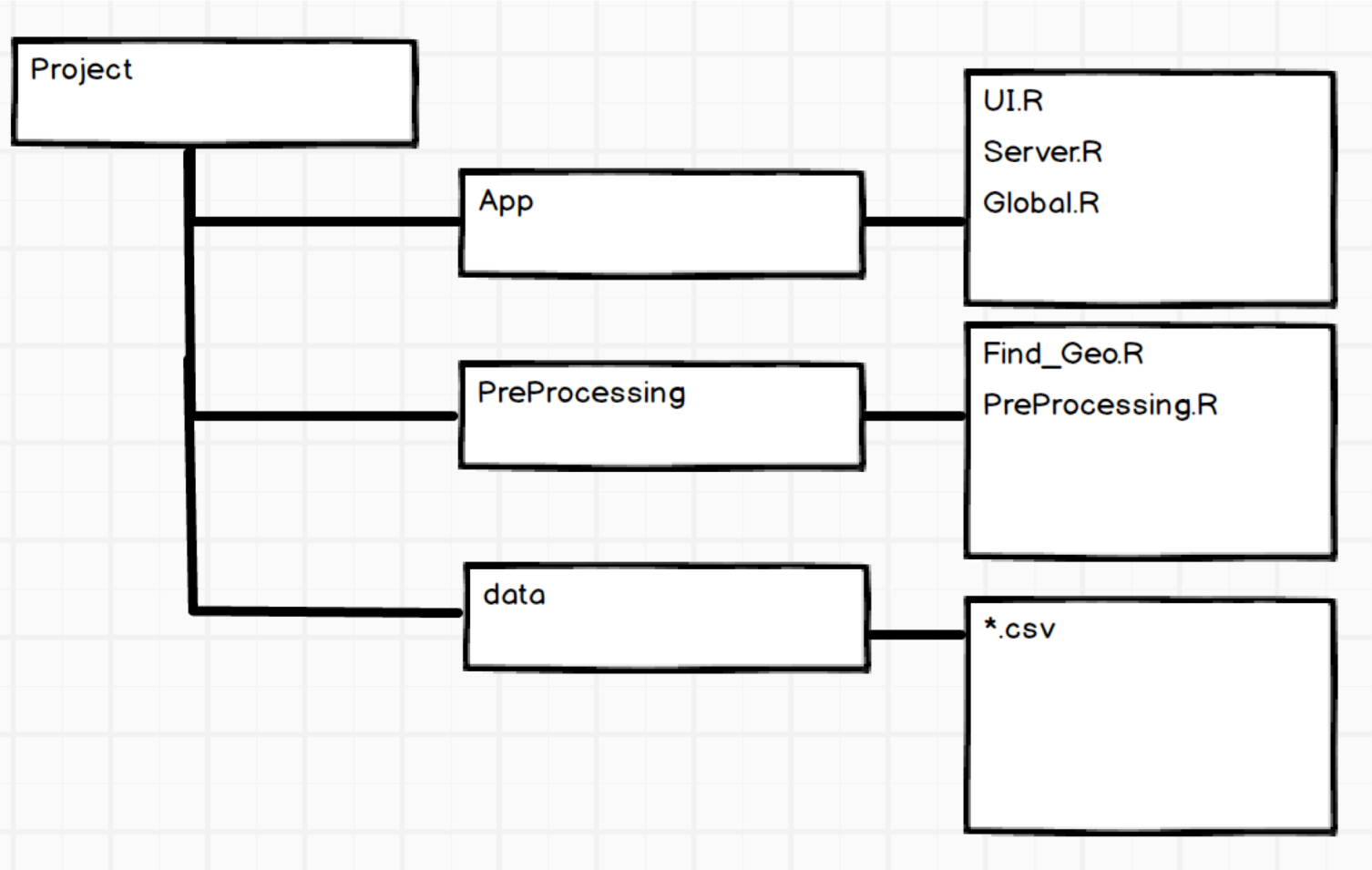
$> docker run -p 3030:3030 --rm --net=host ₩ -e EULA="https://dl.lenses.stream/d/?id=CHECK_YOUR_EMAIL_FOR_KEY" ₩ landoop/kafka-lenses-dev

→ 사전에 Email로 인증을 받아야 설치가 가능함.





Project Structure



- **R Shiny Component**

- **Find_Geo.R**
  : Transform Geo lat/lot

- **PreProcessing.R**
  : Add New Feature / ETL / Missing Value

7. Install Lenses – (3)

- https://www.youtube.com/watch?v=8nBLF46vqK4

  → Lenses Overview

Thank you!