

# data-engineering-study

---

데이터 엔지니어링 학습 포트폴리오

패스트 캠프 데이터 엔지니어링 강의를 학습하며 정리, 공부

## 학습 목표

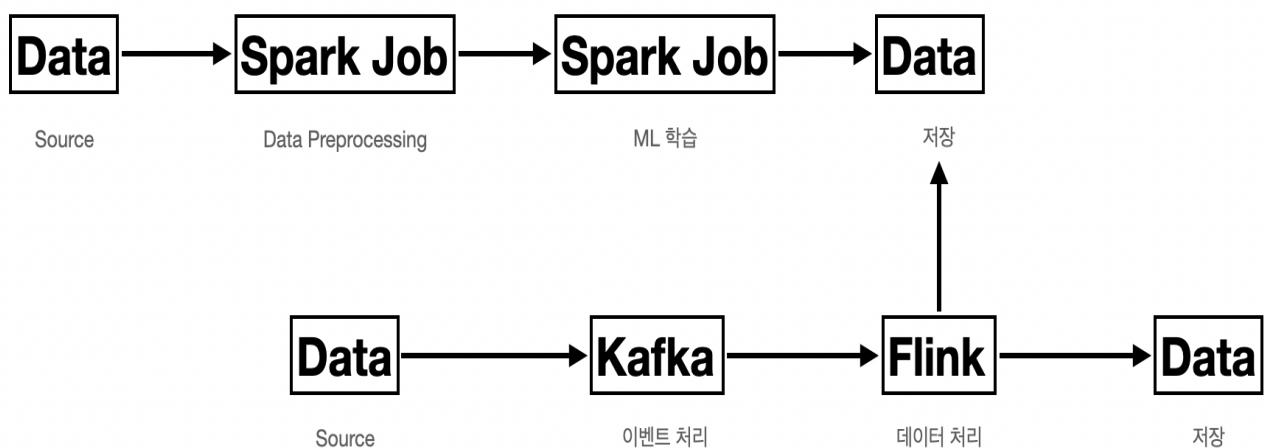
- ETL 방식에서 ELT 흐름으로 넘어가는 모던 데이터 엔지니어링 아키텍처 이해
- 과거 데이터와 실시간 데이터를 기반으로 배치 파이프라인과 스트림 파이프라인을 동시에 사용하는 ML 데이터 학습 & 서빙 파이프라인 설계

## 학습 내용

- **Spark** : 데이터 병렬-분산 처리
- **Airflow** : 데이터 오케스트레이션
- **Kafka** : 이벤트 스트리밍
- **Flink** : 분산 스트림 프로세싱

## 학습 아키텍처

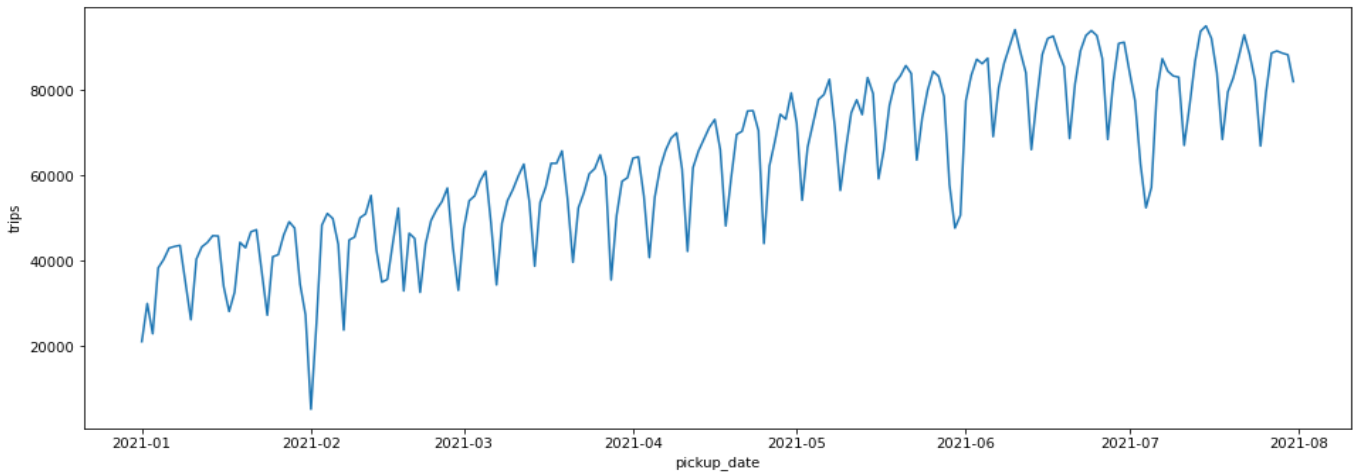
# 배치 + 스트림 파이프라인



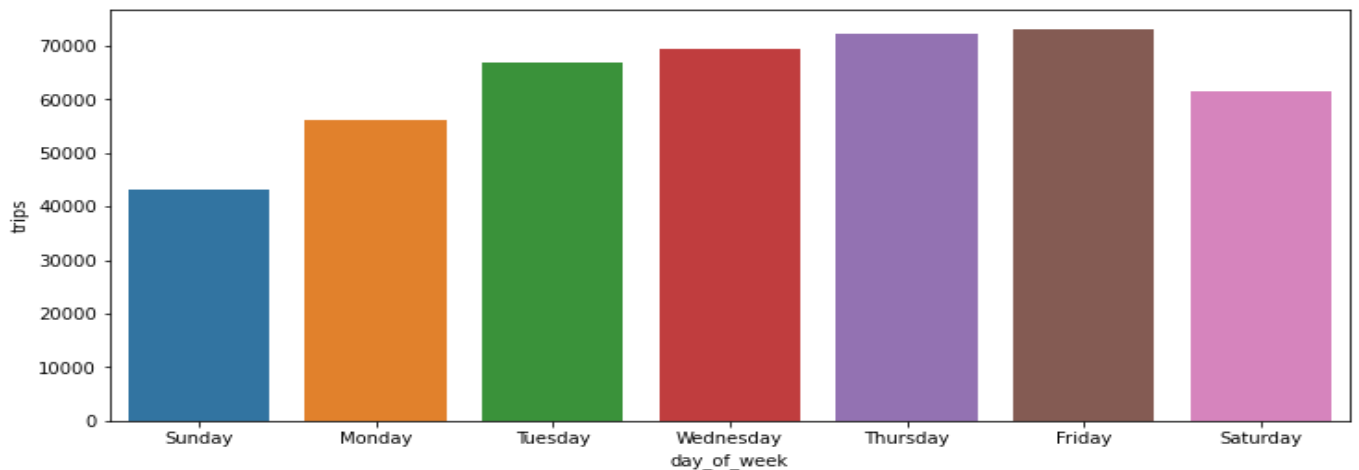
# Spark

- 배치 데이터 분석
- Data Preprocessing
- Hyper Parmeter 파이프라인
- ML 예측 모델 학습 파이프라인

데이터 분석 - 날짜별 택시 이용



데이터 분석 - 요일별 택시 이용



학습 모델 예측 결과 값

day_of_week	trip_distance	total_amount	prediction
Thursday	1.6	12.3	14.905032148096899
Saturday	3.3	23.15	20.335917682834957
Wednesday	4.1	16.3	16.11540300827887
Thursday	0.4	5.8	8.136376385141844
Thursday	15.4	65.3	47.26759901943578
Friday	3.8	13.3	46.6271718097099
Friday	4.6	17.8	48.1340994114685
Wednesday	15.2	76.3	66.94334757308079
Sunday	3.5	17.3	18.84901665167066
Monday	6.3	24.3	26.543356292003317
Saturday	5.6	27.35	24.55224258636062
Tuesday	7.6	32.75	30.046050392502785
Wednesday	0.1	8.8	12.37799176709074
Wednesday	2.0	12.8	16.751342623248892
Monday	2.0	15.8	16.368626347518443
Friday	3.6	20.75	20.64708625351806
Saturday	4.5	20.3	22.293420211923724
Saturday	3.5	19.56	19.06859748558028
Saturday	4.9	24.3	23.13131594298647
Tuesday	0.8	8.3	13.204988549955665

only showing top 20 rows

## Airflow

- 배치 데이터 가공, 저장 파이프라인 DAG 설계
  - Data Preprocessing -> Train/Test 데이터 저장
  - Hyper Parameter 학습 -> 파라미터 csv 파일로 저장
  - Train Model -> 학습 된 모델을 저장
  - 위 과정을 에어플로우 DAG의 작업화(Task) 하고 의존성 추가
- Airflow DAG tag 관리

## DAGs

All 3Active 0Paused 3

\* spark

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks
<div><div></div>air-pollution-pipeline<div>spark</div></div>	airflow	<div><div>1</div></div>	@once	2022-01-01, 00:00:00		<div><div>3</div></div>
<div><div></div>spark-example<div>spark</div></div>	airflow	<div><div>1</div></div>	@daily	2022-04-17, 00:00:00	2022-05-03, 00:00:00	<div><div>1</div></div>
<div><div></div>taxi-price-pipeline<div>spark</div></div>	airflow	<div><div>1</div></div>	@once	2021-01-01, 00:00:00		<div><div>3</div></div>

«

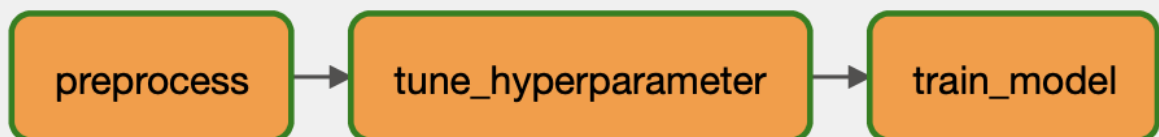
<

1

>

»

Airflow DAG Dependency Graph



# Kafka

- 카프카 실시간 빅데이터 처리 실습
- 카프카를 활용한 Fraud Detection Sub Project

## 실시간 결재 정보 스트림

```

[Base] devkhk@devkhk-MacBook-Air review % python fraud_processor.py
거래 승인 {'date': '2022-05-31', 'time': '16:56:44', 'method': 'TOSS', 'to': 'mom', 'amount': 254} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:36 / 거래 승인 BITCOIN /friend->117
거래 승인 {'date': '2022-05-31', 'time': '16:56:45', 'method': 'KAKAOPAY', 'to': 'stranger', 'amount': 39} | 의심 징상 목록: 비트코인 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:56:46', 'method': 'ACCOUNT', 'to': 'stranger', 'amount': 135} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:37 / 거래 승인 KAKAOPAY /stranger->155
거래 승인 {'date': '2022-05-31', 'time': '16:56:47', 'method': 'ACCOUNT', 'to': 'stranger', 'amount': 60} | 의심 징상 목록: 미등록 계좌와 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:56:48', 'method': 'NAVERPAY', 'to': 'mom', 'amount': 133} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:39 / 거래 승인 ACCOUNT /stranger->227
거래 승인 {'date': '2022-05-31', 'time': '16:56:49', 'method': 'NAVERPAY', 'to': 'dad', 'amount': 157} | 의심 징상 목록: 미등록 계좌와 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:56:50', 'method': 'NAVERPAY', 'to': 'stranger', 'amount': 29} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:45 / 거래 승인 KAKAOPAY /stranger->39
5) 거래 승인 {'date': '2022-05-31', 'time': '16:56:51', 'method': 'NAVERPAY', 'to': 'friend', 'amount': 119} | 의심 징상 목록: 미등록 계좌와 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:56:52', 'method': 'TOSS', 'to': 'dad', 'amount': 56} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:46 / 거래 승인 ACCOUNT /stranger->135
거래 승인 {'date': '2022-05-31', 'time': '16:56:53', 'method': 'CREDITCARD', 'to': 'dad', 'amount': 124} | 의심 징상 목록: 미등록 계좌와 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:56:54', 'method': 'NAVERPAY', 'to': 'friend', 'amount': 99} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:47 / 거래 승인 ACCOUNT /stranger->60
거래 승인 {'date': '2022-05-31', 'time': '16:56:55', 'method': 'ACCOUNT', 'to': 'dad', 'amount': 196} | 의심 징상 목록: 미등록 계좌와 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:56:56', 'method': 'TOSS', 'to': 'friend', 'amount': 99} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:50 / 거래 승인 NAVERPAY /stranger->295
거래 승인 {'date': '2022-05-31', 'time': '16:56:57', 'method': 'BITCOIN', 'to': 'stranger', 'amount': 195} | 의심 징상 목록: 미등록 계좌와 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:56:58', 'method': 'KAKAOPAY', 'to': 'stranger', 'amount': 21} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:57 / 거래 승인 BITCOIN /mom->271
1) 거래 승인 {'date': '2022-05-31', 'time': '16:56:59', 'method': 'KAKAOPAY', 'to': 'friend', 'amount': 80} | 의심 징상 목록: 비트코인 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:57:00', 'method': 'KAKAOPAY', 'to': 'friend', 'amount': 78} | 비정상 거래로 의심됩니다. / 2022-05-31 16:57:02 / 거래 승인 KAKAOPAY /stranger->211
거래 승인 {'date': '2022-05-31', 'time': '16:57:01', 'method': 'KAKAOPAY', 'to': 'mom', 'amount': 95} | 의심 징상 목록: 미등록 계좌와 거래 /
거래 승인 {'date': '2022-05-31', 'time': '16:57:02', 'method': 'BITCOIN', 'to': 'mom', 'amount': 271} | 비정상 거래로 의심됩니다. / 2022-05-31 16:56:58 / 거래 승인 KAKAOPAY /stranger->211
[Base] devkhk@devkhk-MacBook-Air review %
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:49', 'method': 'NAVERPAY', 'to': 'dad', 'amount': 157} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:44', 'method': 'TOSS', 'to': 'mom', 'amount': 254}
의심스러운 거래 확인 {'date': '2022-05-31', 'time': '16:56:50', 'method': 'NAVERPAY', 'to': 'stranger', 'amount': 39} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:48', 'method': 'NAVERPAY', 'to': 'mom', 'amount': 133}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:51', 'method': 'NAVERPAY', 'to': 'friend', 'amount': 119} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:49', 'method': 'NAVERPAY', 'to': 'dad', 'amount': 157}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:52', 'method': 'TOSS', 'to': 'dad', 'amount': 56} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:51', 'method': 'NAVERPAY', 'to': 'friend', 'amount': 99}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:53', 'method': 'CREDITCARD', 'to': 'dad', 'amount': 124} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:52', 'method': 'TOSS', 'to': 'dad', 'amount': 56}
24) 정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:54', 'method': 'NAVERPAY', 'to': 'friend', 'amount': 99} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:54', 'method': 'NAVERPAY', 'to': 'friend', 'amount': 99}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:55', 'method': 'ACCOUNT', 'to': 'dad', 'amount': 196} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:55', 'method': 'ACCOUNT', 'to': 'dad', 'amount': 196}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:56', 'method': 'TOSS', 'to': 'friend', 'amount': 99} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:56', 'method': 'TOSS', 'to': 'friend', 'amount': 99}
의심스러운 거래 확인 {'date': '2022-05-31', 'time': '16:56:57', 'method': 'BITCOIN', 'to': 'stranger', 'amount': 195} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:57', 'method': 'BITCOIN', 'to': 'stranger', 'amount': 195}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:58', 'method': 'KAKAOPAY', 'to': 'stranger', 'amount': 21} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:58', 'method': 'KAKAOPAY', 'to': 'stranger', 'amount': 21}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:56:59', 'method': 'KAKAOPAY', 'to': 'friend', 'amount': 80} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:56:59', 'method': 'KAKAOPAY', 'to': 'friend', 'amount': 80}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:57:00', 'method': 'KAKAOPAY', 'to': 'friend', 'amount': 78} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:57:00', 'method': 'KAKAOPAY', 'to': 'friend', 'amount': 78}
정상 거래 확인 {'date': '2022-05-31', 'time': '16:57:01', 'method': 'KAKAOPAY', 'to': 'mom', 'amount': 95} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:57:01', 'method': 'KAKAOPAY', 'to': 'mom', 'amount': 95}
의심스러운 거래 확인 {'date': '2022-05-31', 'time': '16:57:02', 'method': 'BITCOIN', 'to': 'mom', 'amount': 271} | 정상 거래로 등록 {'date': '2022-05-31', 'time': '16:57:02', 'method': 'BITCOIN', 'to': 'mom', 'amount': 271}

```

## 의심 거래 정황 Slack 알람봇

Slack Connect
 

Slack 찾아보기

채널

# 랜덤

# 이상거래-알림봇

# 일반

+ 채널 추가

다이렉트 메시지

devkhk 나

+ 팀원 추가

앱

Fraud-Detector

+ 앱 추가

이상거래-알림봇

이상 거래 탐지

오늘

4:56

Fraud-Detector

비정상 거래로 의심되는 알림 / 2022-05-31 16:56:36 / 거래승인 BITCOIN /friend->117

의심 정황 목록: 비트코인 거래 /

비정상 거래로 의심되는 알림 / 2022-05-31 16:56:37 / 거래승인 KAKAOPAY /stranger->155

의심 정황 목록: 미등록 계좌와 거래 /

비정상 거래로 의심되는 알림 / 2022-05-31 16:56:39 / 거래승인 ACCOUNT /stranger->227

의심 정황 목록: 미등록 계좌와 거래 /

4:56

비정상 거래로 의심되는 알림 / 2022-05-31 16:56:43 / 거래승인 CREDITCARD /stranger->73

의심 정황 목록: 미등록 계좌와 거래 /

비정상 거래로 의심되는 알림 / 2022-05-31 16:56:45 / 거래승인 KAKAOPAY /stranger->39

의심 정황 목록: 미등록 계좌와 거래 /

비정상 거래로 의심되는 알림 / 2022-05-31 16:56:46 / 거래승인 ACCOUNT /stranger->135

의심 정황 목록: 미등록 계좌와 거래 /

비정상 거래로 의심되는 알림 / 2022-05-31 16:56:47 / 거래승인 ACCOUNT /stranger->60

의심 정황 목록: 미등록 계좌와 거래 /

B I U | 🔗 | 📌 | 📋 | 📄 | </> | 🗑️

#이상거래-알림봇에 메시지 보내기

+ | 📎 | 🗣️ | 😊 | @ | Aa

▶ | ▼

# Flink

- 스트림 데이터 프로세싱 실습
- 플링크를 활용한 Twitter API RealTime Stream Process Sub Project
- 배치 데이터 모델을 실시간 택시비 예측에 사용하기

## 카프카 + 플링크를 활용한 실시간 피드 단어 수 세기 Sub Project

```
{'text': '@cutelovejimin @BTS_twt 공금금 🍷👉🏻\n\nbst ost #WithYou by JIIMIN #BTSJIMIN of #BTS\x00 (@BTS_twt) and Sungwoon', 'created_at': '2022-06-02T13:06:55.000'}
{'text': '! 사실 페스타도 종친한대 핼트버가 더 기다려져\n\nThe most beautiful song I've ever listened to is #WithYou by JIIMIN #BTSJIMIN of #BTS\x00 (@BTS_twt) and Sungwoon', 'created_at': '2022-06-02T13:06:56.000'}
{'text': '@Neojimi1n @BTS_twt 음서서 달님네다 \n\nMy heart beats listening to the most lovely OST #WithYou by JIIMIN #BTSJIMIN of #BTS (@BTS_twt) and Sungwoon', 'created_at': '2022-06-02T13:06:56.000'}
{'text': '@sososoc7 ππ 출 ππ 다들엔 꼭!!! π\n\n후회하더라도 갔다고 후회를 해야해요!! π\n\n그때도 지인덕분에 밤이 사진이 π\n\nπ\n\n받으면 풍소니이지만 올라볼게요 ππ', 'created_at': '2022-06-02T13:06:56.000'}
{'text': '*방탄소년단 (BTS) *, 다양한 하트로 여심 홀릭 *설립식 폭발...' *브로커* VIPA사자! [하루의] #BTS #V #결정화 https://t.co/Zz8Aouu14a via YouTube', 'created_at': '2022-06-02T13:06:58.000'}
{'text': '@Ermsussudsud 그레서 총 가격은 얼마만니까?', 'created_at': '2022-06-02T13:06:59.000'}
{'text': '@Efineday_v @BTS_twt 브로커 기사에 우리태형이 지분이 너무많네요 오늘 작성하러 해매 다 안봐먹어\n\n\nSOUTH KOREA'S PRIDE TAEHYUNG \n\nlistening to the O ST chart toppler #ChristmasTree by #V of @BTS_twt from the Kdrama Our Beloved Summer #V_ChristmasTree \n\n\nKimTaehyung #BTSV *방탄소년단원 https://t.co/dSY8nQ0tC*', 'created_at': '2022-06-02T13:06:59.000'}
{'text': '@Esokjin_vocal 이틀 뵈는데 건을아래서 급,토요일 연차내고 가려고 했는데... 아깝네요 🙄\n\n\nyour's_jin \n\n\nsuperturna_jin \n\n\nmoon_jin \n\n\n방탄소년단원 #JIN \n\n\n@bts_twt', 'created_at': '2022-06-02T13:07:01.000'}
{'text': '@Yeomnot_bts @Just_BTS_Army_7 @_KING_BTARSARMY @BT_Yubury @tess3821 @60_02eni @l0qk4FRbZpfDwaf @JoYorBTS7 @jk_forever9901 @kwon9866 @newlaif_army @Dontd @78786 @BTS_twt #프드 예쵸! 주먹박 박고 그치지 마요 \n\n\nToCome BTS_Proof #BTS_BTSA_TWT', 'created_at': '2022-06-02T13:07:03.000'}
{'text': '@ijchG23 @pleroma217 @slfnis @haven4800daum1 @ZINY_JINY @junsung3523 @CutiePeachJin @BTS_twt 첫전이는 살아있는게 지가 할정지 키키는거야 🍷👉🏻\n\n\nCONGRATULATIONS JIN\n\nBST DROMA OST ARTIST JIN\n\n하고의 OST 전직의 YOURS\n\n\n@Jin @BTS_twt ht tps://t.co/UWHM4H3zet', 'created_at': '2022-06-02T13:07:03.000'}
{'text': '@BTS_twt TAEHYUNG 좋은 아침!', 'created_at': '2022-06-02T13:07:04.000'}
{'text': '너무 귀여워 ~ 🍷👉🏻\n\n\n영하야아아아 사랑해 U000f1979U0001faf 🍷 ht tps://t.co/FzaJss94h8 https://t.co/Umh0mh42ed', 'created_at': '2022-06-02T13:07:04.000'}
{'text': '@Bts_backarmy13 그니까요 우리 지민이 최송하다는말 금지!!', 'created_at': '2022-06-02T13:07:06.000'}
{'text': '그것도 1위랑 막상막하 얼치락뒤치락하는 존나 선 2위 시켜주고 싶어.그렇게 트롤 놀이를 올려주조고 싶어 .1위 타이틀을,2위 보노,3위 무조건 쓸수있을 속히 좀 시연해달라게? .....후우 .....일단 지금은 해서 더 열심히 달려보게 🍷👉🏻\n\n\nWithYou
```

[illegible]

## 배치 모델을 활용한 실시간 택시비 예측

```
(base) devkhk@devkhk-MacBook-Air taxi-pricing % python flink_taxi_pricing.py
+I[2021-01-01T00:30:10, 2.1, 0, 13.269387184252423]
+I[2021-01-01T00:51:20, 0.4, 0, 8.792177072267071]
+I[2021-01-01T00:43:30, 0.2, 0, 42.9603595058395]
+I[2021-01-01T00:15:48, 10.6, 0, 33.299011369458005]
+I[2021-01-01T00:31:49, 4.94, 0, 19.961638088483163]
+I[2021-01-01T00:16:29, 1.6, 0, 12.091173996878577]
+I[2021-01-01T00:00:28, 4.1, 0, 17.982239937310687]
+I[2021-01-01T00:12:29, 5.7, 0, 21.752522133277303]
+I[2021-01-01T00:39:16, 9.1, 0, 29.764371887356355]
+I[2021-01-01T00:26:12, 2.7, 0, 14.683243009808903]
+I[2021-01-01T00:15:52, 6.11, 0, 22.718656914916248]
+I[2021-01-01T00:46:36, 1.21, 0, 11.172167710743494]
+I[2021-01-01T00:10:46, 7.4, 0, 25.758446970316825]
+I[2021-01-01T00:31:06, 1.7, 0, 12.32681663436077]
+I[2021-01-01T00:42:11, 0.81, 0, 10.22959716085184]
+I[2021-01-01T00:17:48, 1.01, 0, 10.700882435797668]
+I[2021-01-01T00:33:38, 0.73, 0, 10.041083085087351]
+I[2021-01-01T00:47:56, 1.17, 0, 11.077910655754328]
+I[2021-01-01T00:04:21, 0.78, 0, 10.158904369609967]
+I[2021-01-01T00:18:36, 1.66, 0, 12.23255979371603]
+I[2021-01-01T00:43:41, 0.93, 0, 10.512368235819338]
+I[2021-01-01T00:56:30, 1.16, 0, 11.054346392007037]
+I[2021-01-01T00:16:27, 2.2, 0, 13.505029821725337]
+I[2021-01-01T00:37:59, 3.6, 0, 16.804026746346125]
```

```
(base) devkhk@devkhk-MacBook-Air taxi-pricing % python producer.py
1,2021-01-01 00:30:10,2021-01-01 00:36:12,1,2.10,1,N,142,43,2,8,3,0.5,0,0,0,0.3,11.8,2.5

1,2021-01-01 00:51:20,2021-01-01 00:52:19,1,.20,1,N,238,151,2,3,0.5,0.5,0,0,0.3,4.3,0

1,2021-01-01 00:43:30,2021-01-01 01:11:06,1,14.70,1,N,132,165,1,42,0.5,0.5,8.65,0,0.3,51.95,0

1,2021-01-01 00:15:48,2021-01-01 00:31:01,0,10.60,1,N,138,132,1,29,0.5,0.5,6.05,0,0.3,36.35,0

2,2021-01-01 00:31:49,2021-01-01 00:48:21,1,4.94,1,N,68,33,1,16,5.0,0.5,0.5,4.06,0,0.3,24.36,2.5

1,2021-01-01 00:16:29,2021-01-01 00:24:30,1,1.60,1,N,224,68,1,8,3,0.5,2.35,0.3,14.15,2.5

1,2021-01-01 00:00:28,2021-01-01 00:17:28,1,4.10,1,N,95,157,2,16,0.5,0.5,0,0.3,17.3,0

1,2021-01-01 00:12:29,2021-01-01 00:30:34,1,5.70,1,N,90,40,2,18,3,0.5,0,0.3,21.8,2.5

1,2021-01-01 00:39:16,2021-01-01 01:00:13,1,9.10,1,N,97,129,4,27,5.0,5.0,5.0,0,0.3,28.8,0

1,2021-01-01 00:26:12,2021-01-01 00:39:46,2,2.70,1,N,263,142,1,12,3,0.5,3.15,0,0.3,18.95,2.5

2,2021-01-01 00:15:52,2021-01-01 00:38:07,3,6.11,1,N,164,255,1,20,5.0,5.0,5.0,0,0.3,24.3,2.5

2,2021-01-01 00:46:36,2021-01-01 00:53:45,2,1.21,1,N,255,80,1,7,0.5,0.5,2.49,0,0.3,3.79,0

1,2021-01-01 00:10:46,2021-01-01 00:32:58,2,7.40,1,N,138,166,2,24,5.2,5.0,5.0,6.12,0,0.3,33.92,0

2,2021-01-01 00:31:06,2021-01-01 00:38:52,5,1.70,1,N,142,50,1,8,0.5,0.5,2.36,0,0.3,14.16,2.5

2,2021-01-01 00:42:11,2021-01-01 00:44:24,5,.81,1,N,50,142,2,4,5.0,5.0,5.0,0,0.3,14.16,2.5
```

## 수료 링크 & 기타

[수료증명서](#)

[학습 과정 기록 블로그](#)

[깃허브 링크](#)

[이력 링크](#)