data-engineering-study

데이터 엔지니어링 학습 포트폴리오

데이터 엔지니어링을 학습하며 정리, 공부

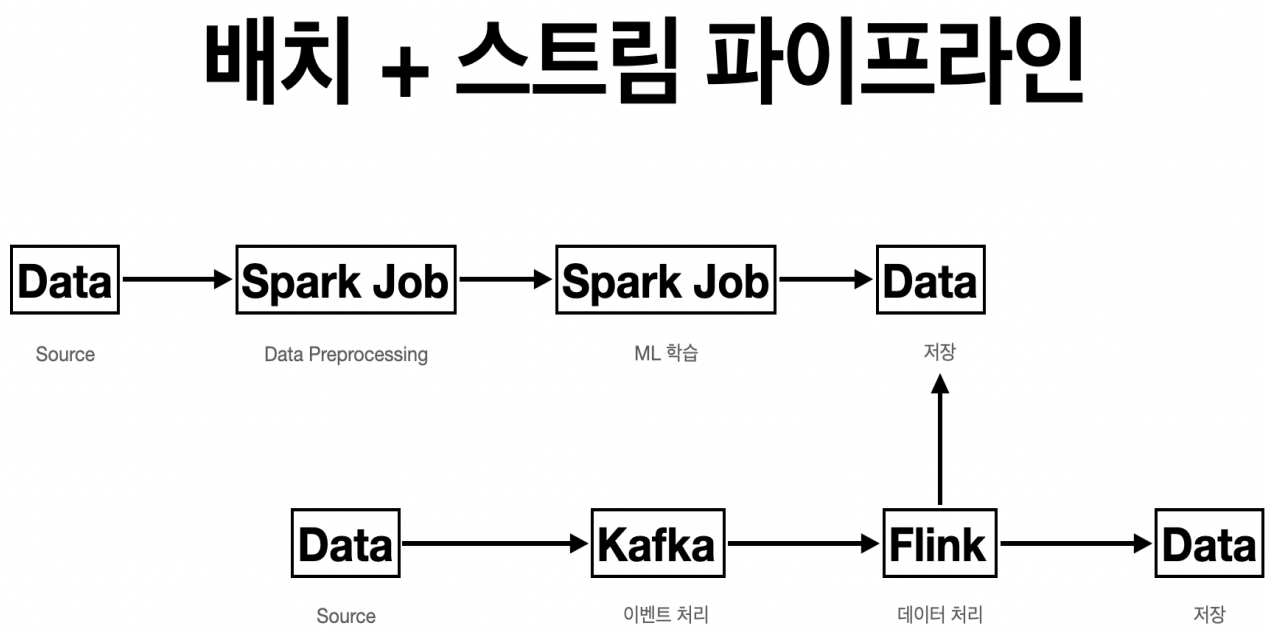
학습 목표

- ETL 방식에서 ELT 흐름으로 넘어가는 모던 데이터 엔지니어링 아키텍쳐 이해
- 과거 데이터와 실시간 데이터를 기반으로 배치 파이프라인과 스트림 파이프라인을 동시에 사용하는 ML 데이터 학습 & 서빙 파이프라인 설계
- 우버 택시 통계 데이터를 기반으로 실시간 거리에 따른 택시비 예측 출력

학습 내용

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

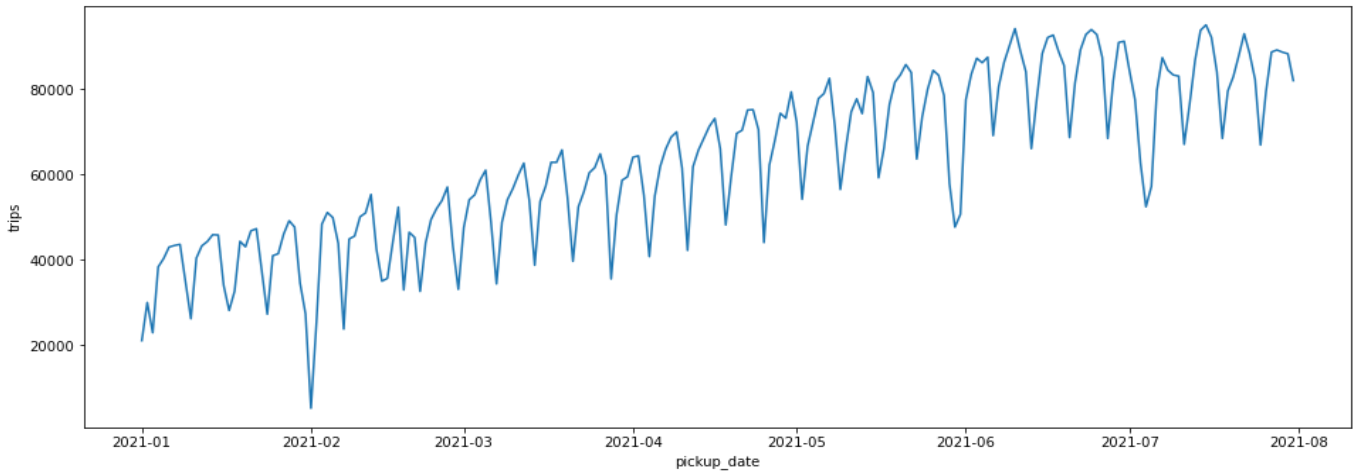
학습 아키텍쳐



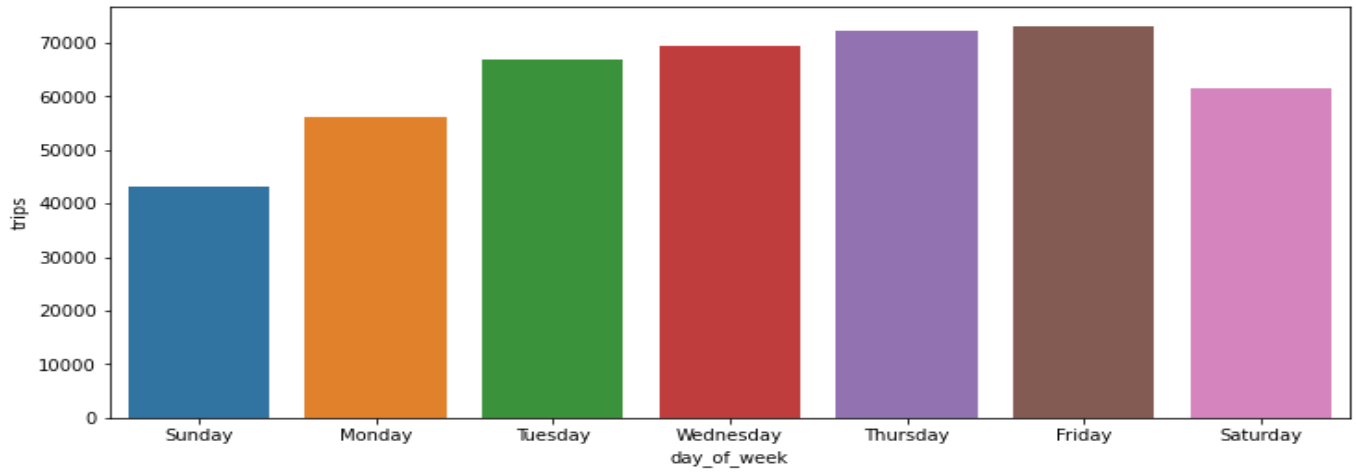
Spark

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

데이터 분석 - 날짜별 택시 이용



데이터 분석 - 요일별 택시 이용



학습 모델 예측 결과 값

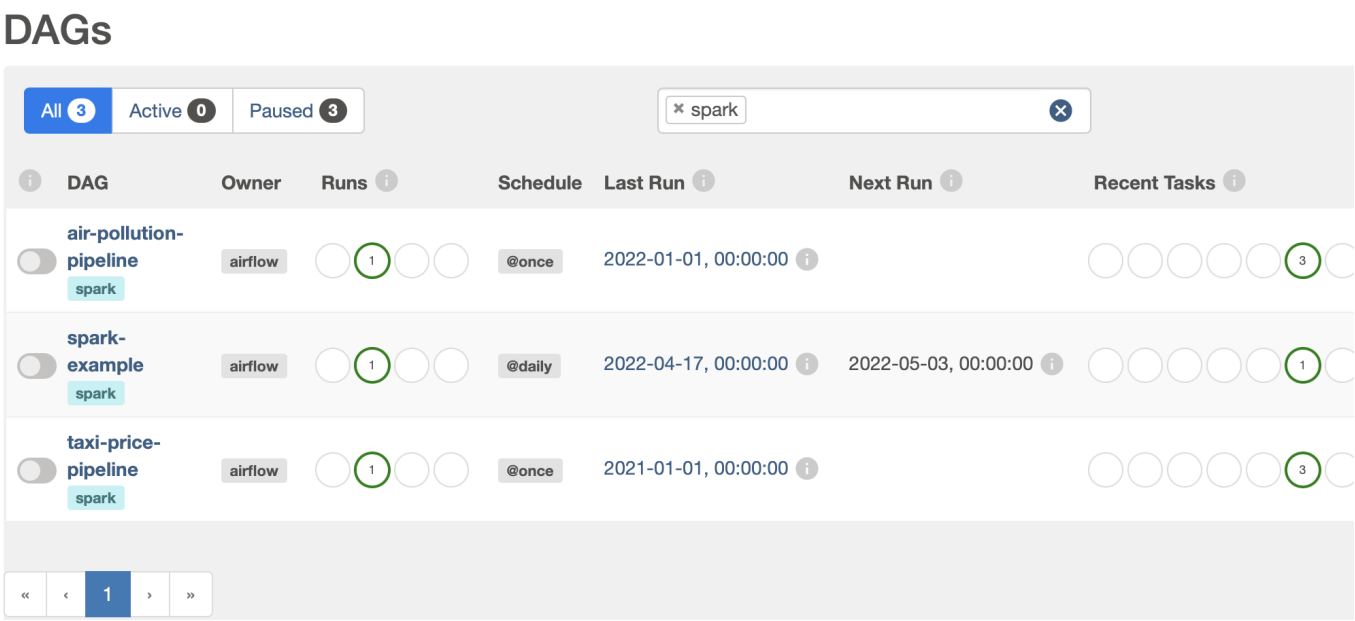
day_of_week	trip_distance	total_amount	prediction
Thursday	1.6	12.3	14.905032148096899
Saturday	3.3	23.15	20.335917682834957
Wednesday	4.1	16.3	16.11540300827887
Thursday	0.4	5.8	8.136376385141844
Thursday	15.4	65.3	47.26759901943578
Friday	3.8	13.3	46.6271718097099
Friday	4.6	17.8	48.1340994114685
Wednesday	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

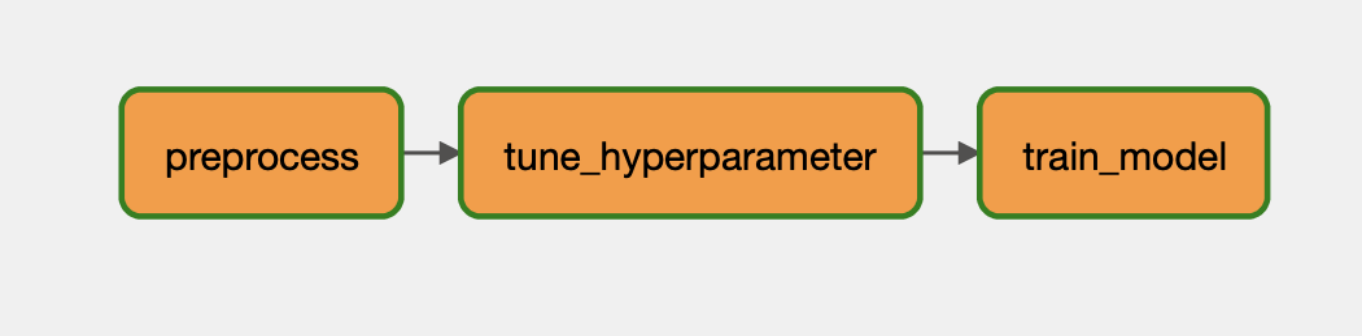
Airflow

- 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



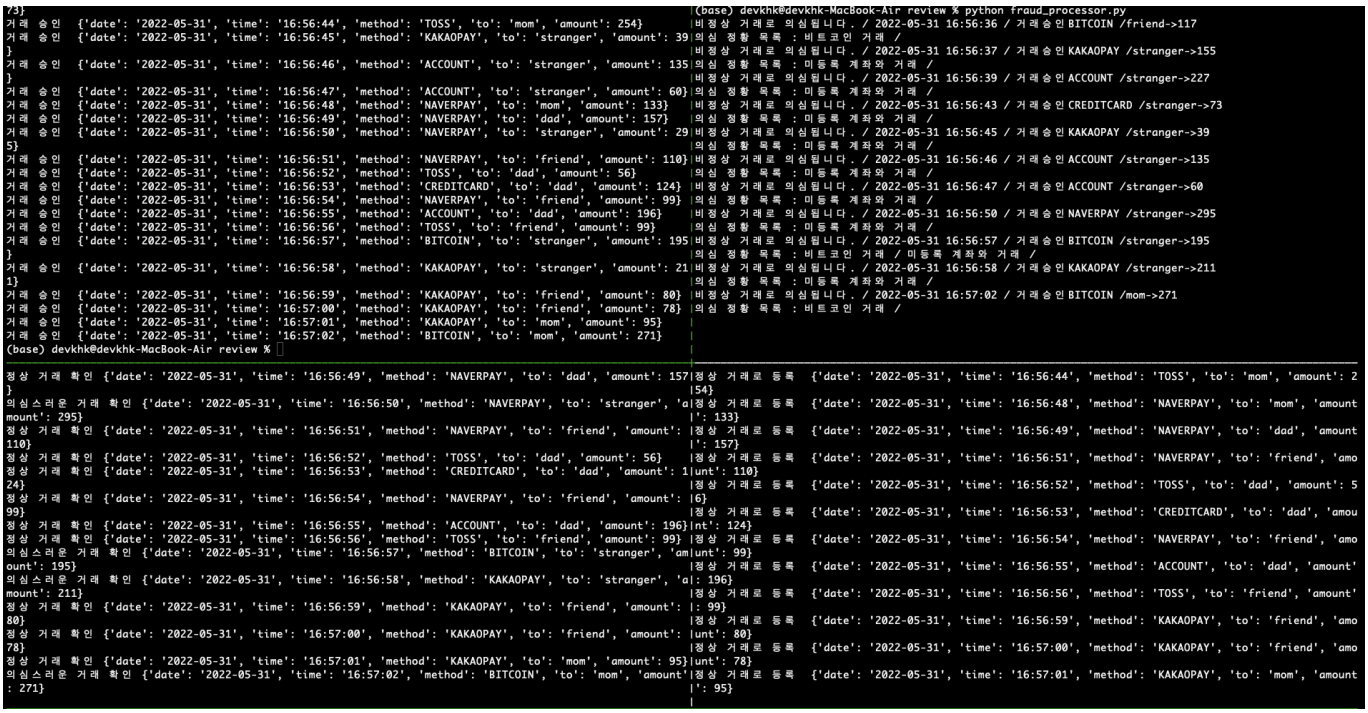
Airflow DAG Depedency Grpah



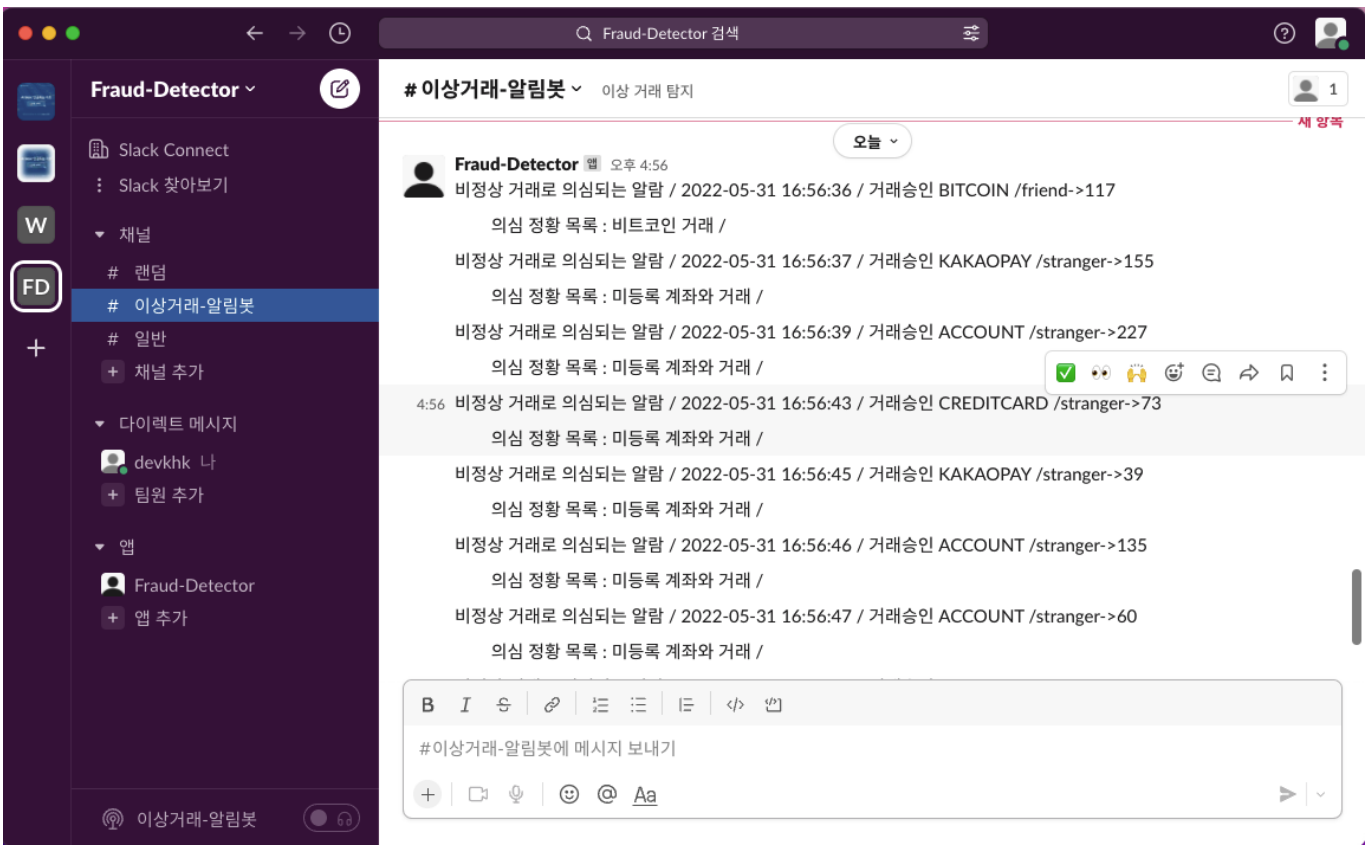
Kafka

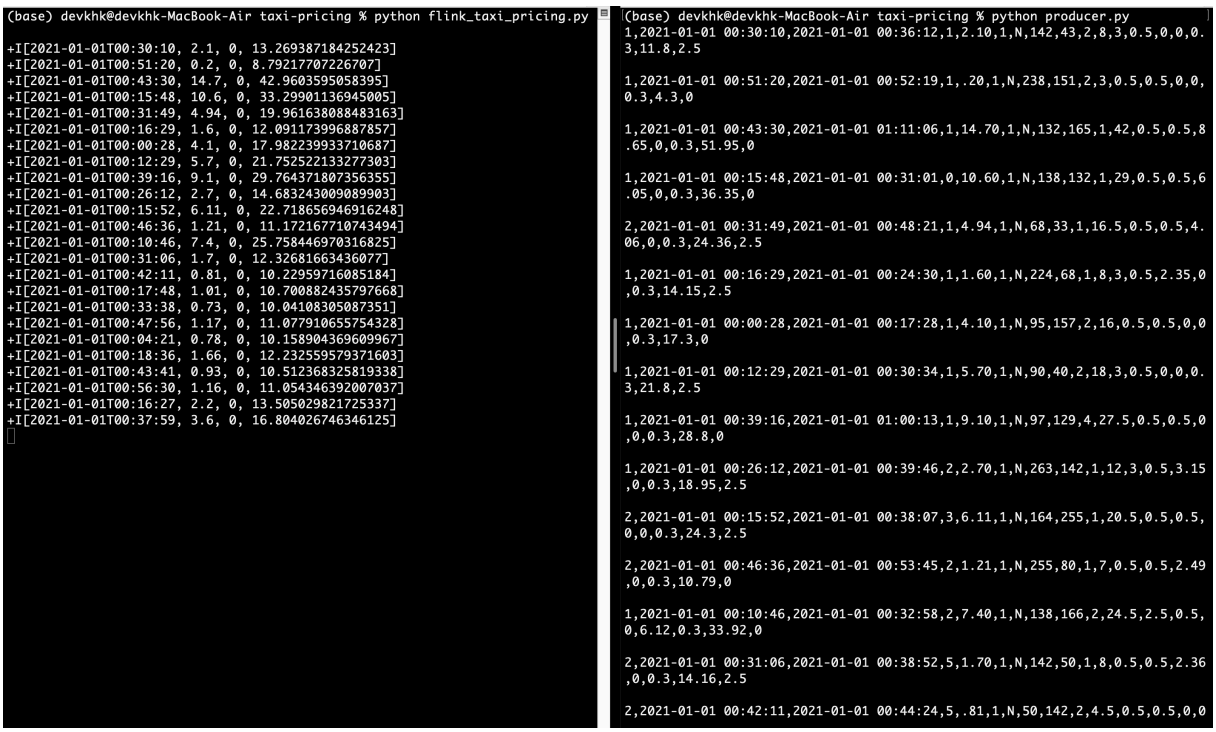
- 카프카 실시간 빅데이터 처리 실습
- 카프카를 활용한 Fraud Detection Sub Project

실시간 결재 정보 스트림



의심 거래 정황 Slack 알람봇





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 궁금궁금 🤔😗\n\n♡♡\n\nbest ost #WithYou by #JI
MIN #BTSJIMIN of #BTS\xa0 ( @BTS_twt ) and Sungwoon', 'created_at': '2022-06-02T
13:06:55.000'}
{'text': '나 사실 페스타도 좋긴한데 찜토버가 더 기다려져\n\nThe most beautiful s
ong I’ve ever listened to is  #WithYou by #JIMIN #BTSJIMIN of #BTS\xa0 (@BTS_twt
) and Sungwoon', 'created_at': '2022-06-02T13:06:56.000'}
{'text': '@Neojimin1 @BTS_twt 울면서 달립니다 \nMy heart beats listening to the
most lovely OST #WithYou by #JIMIN #BTSJIMIN of #BTS (@BTS_twt) and Sungwoon', '
created_at': '2022-06-02T13:06:56.000'}
{'text': '@sososoc_7 ㅠㅠ 흑 ㅠㅠ 다음엔 꼭!!! ㅠ \n후회하더라도 갔다오고 후회를
 해야겠어요! ㅠ\n그래도 지인덕분에 텽이 사진이 ㅠㅠ\n\n원본받으면 똥소니이지만
올려볼게요 ㅠㅠ', 'created_at': '2022-06-02T13:06:56.000'}
{'text': "방탄소년단 (BTS) 뷔, 다양한 하트로 여심 홀릭 '잘생김 폭발'... '브로커'
 VIP시사회 [비하인드] #BTS #V #김태형 https://t.co/Zz8Aouul4a via @YouTube", 'cr
eated_at': '2022-06-02T13:06:58.000'}
{'text': '@rmssudsudsud 그래서 총 가격은 얼마입니까?', 'created_at': '2022-06-02
T13:06:59.000'}
{'text': '@fineday_v @BTS_twt 브로커 기사에 우리태형이 지분이 너무많네요 오늘 착
장하며 헤메 다 완벽했어요\n\nSOUTH KOREA’S PRIDE TAEHYUNG \n\nListening to the O
ST chart topper #ChristmasTree by #V of @BTS_twt from the KDrama Our Beloved Sum
mer #V_ChristmasTree \n#KimTaehyung #BTSV #방탄소년단뷔 https://t.co/D5YBnG0t0C'
, 'created_at': '2022-06-02T13:06:59.000'}
{'text': '@seokjin_vocal 이제 봤는데 건물이래서 금,토요일 연차내고 가려고 했는데
... 아깝네요 🤣🤣\n#yours_jin \n#supertuna_jin \n#moon_jin \n#방탄소년단진 #JIN \
n @bts_twt', 'created_at': '2022-06-02T13:07:01.000'}
{'text': '@Yeonmot_bts @Just_BTS_Army_7 @_KING_BTSARMY @BT_Yubury @mtess3821 @60
02Eni @1Oqk4fRbZfgDWaf @JoYforBTS7 @jk_forevr0901 @kwon9866 @newlaif_army @Dontd
oit78768 @BTS_twt 쪼매  죄송..주먹꽉 쥐고. 그러지 마요\n#YetToCome_BTS #BTS_Proo
f  #BTS  @BTS_twt', 'created_at': '2022-06-02T13:07:03.000'}
{'text': '@jch623 @pleroma217 @slfnsid @haven4800daumn1 @ZZINY_JINY @junsung3523
 @CutiePeachJin @BTS_twt 석진이는 살아있는게 지구환경 지키는거야 🦄✨😎\n\nCONGRA
TULATIONS JIN\nBEST KDRAMA OST ARTIST JIN\n최고의 OST 석진이의 YOURS\n\n#Yours_J
in #SuperTuna_Jin 존재만으로도 공기청정 #Moon_Jin #방탄소년단진 #JIN @BTS_twt ht
tps://t.co/UWH4lt36tt', 'created_at': '2022-06-02T13:07:03.000'}
{'text': '@BTS_twt_TAEHYNG 좋은 아침', 'created_at': '2022-06-02T13:07:04.000'}
{'text': '너무 귀여워 … 😍😍😍\n태형아아아아아 사랑해\U0001f979\U0001faf6🏻🐻 ht
tps://t.co/FzaJss94h8 https://t.co/Umh0mh42ed', 'created_at': '2022-06-02T13:07:
04.000'}
{'text': '@bts_backarmy13 그니까요 우리 지민이 죄송하다는말 금지!!', 'created_at
': '2022-06-02T13:07:06.000'}
{'text': '그것도 1위랑 막상막하 엎치락뒤치락하는 존나 센 2위 시켜주고 싶어.그렇
게 ㅍㅌ를 높이 올려주고 싶어.1위 타이틀곡,2위 ㅍㅌ,3위 위드유 줄세우면 속이 좀
시원해질까?아.....후우......일단 지금은 해시 더 열심히 달려보자 🔥🔥🔥\n\n#WithYo
```

```
+I[[('@BTS_twt', 29), ('of', 25), ('#BTS', 23), ('by', 20), ('#WithYou',
 18)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 29), ('of', 25), ('#BTS', 23), ('by', 20), ('#WithYou',
 18)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 31), ('of', 25), ('#BTS', 23), ('by', 20), ('#WithYou',
 18)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 32), ('of', 26), ('#BTS', 24), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 32), ('of', 26), ('#BTS', 24), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 32), ('of', 26), ('#BTS', 24), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 32), ('of', 26), ('#BTS', 24), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 32), ('of', 26), ('#BTS', 24), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07, 2022-06-02T13:07:10]
+I[[('@BTS_twt', 33), ('of', 26), ('#BTS', 24), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 34), ('of', 26), ('#BTS', 25), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 36), ('#BTS', 26), ('of', 26), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 36), ('#BTS', 26), ('of', 26), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 36), ('#BTS', 27), ('of', 26), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 36), ('#BTS', 27), ('of', 26), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 36), ('#BTS', 27), ('of', 26), ('by', 21), ('#WithYou',
 19)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 37), ('#BTS', 28), ('of', 27), ('by', 22), ('#WithYou',
 20)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 38), ('#BTS', 28), ('of', 27), ('by', 22), ('#WithYou',
 20)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 38), ('#BTS', 28), ('of', 28), ('by', 23), ('#BTSJIMIN'
, 22)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 39), ('#BTS', 29), ('of', 29), ('by', 24), ('#BTSJIMIN'
, 23)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 39), ('#BTS', 29), ('of', 29), ('by', 24), ('#BTSJIMIN'
, 23)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 39), ('#BTS', 30), ('of', 30), ('by', 25), ('#BTSJIMIN'
, 24)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
+I[[('@BTS_twt', 39), ('#BTS', 31), ('of', 31), ('by', 26), ('#BTSJIMIN'
, 25)], 2022-06-02T13:07:10, 2022-06-02T13:07:20]
```

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, 0.2, 0, 8.79217707226707]
+I[2021-01-01T00:43:30, 14.7, 0, 42.9603595058395]
+I[2021-01-01T00:15:48, 10.6, 0, 33.29901136945005]
+I[2021-01-01T00:31:49, 4.94, 0, 19.961638088483163]
+I[2021-01-01T00:16:29, 1.6, 0, 12.091173996887857]
+I[2021-01-01T00:00:28, 4.1, 0, 17.982239933710687]
+I[2021-01-01T00:12:29, 5.7, 0, 21.752522133277303]
+I[2021-01-01T00:39:16, 9.1, 0, 29.764371807356355]
+I[2021-01-01T00:26:12, 2.7, 0, 14.683243009089903]
+I[2021-01-01T00:15:52, 6.11, 0, 22.718656946916248]
+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, 10.04108305087351]
+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, 12.232559579371603]
+I[2021-01-01T00:43:41, 0.93, 0, 10.512368325819338]
+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.
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,68,33,1,16.5,0.5,0.5,4.
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
,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
,0.3,17.3,0

1,2021-01-01 00:12:29,2021-01-01 00:30:34,1,5.70,1,N,90,40,2,18,3,0.5,0,0,0.
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
,0,0.3,10.79,0

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,
0,6.12,0.3,33.92,0

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
```

수료 링크 & 기타

수료증명서

학습 과정 기록 블로그

깃허브 링크

이력 링크