

Elastic Stack을 활용한 Data Dashboard 만들기

Week 5 - Dashboard 만들기 최종실습



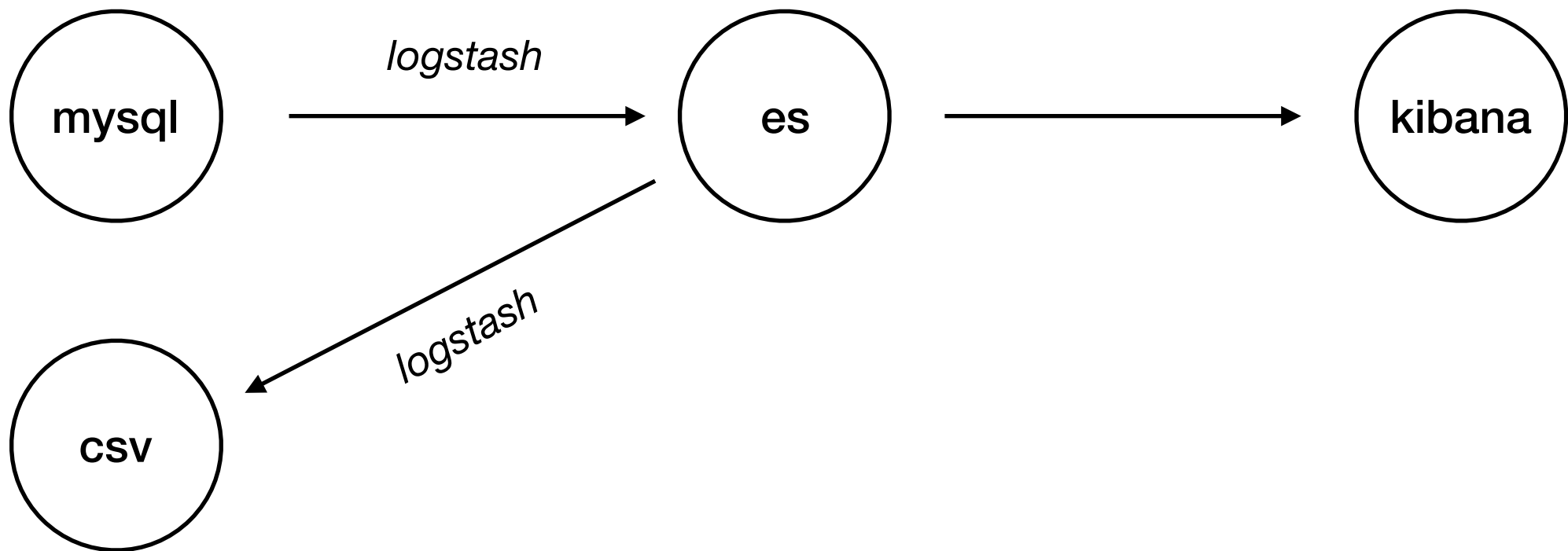
Fast Campus

Guide

- 실제로 대시보드 구축 프로젝트를 가정하고 실습 진행
- 그러하기에 자세한 안내보다는 **요구사항 위주의 문서** 스타일
- 최소한의 정보는 제공하지만 **직접 데이터를 조회하면서 작업**해야 하는 흐름
- 그 과정에서 지금까지 배웠던 내용을 사용하기를 권장
- 특히 **시각화의 경우 정해진 답이 없으므로**, 사용자의 니즈를 충족하는 범위 내에서 최대한 **자유롭게** 제작

개요

- mysql에서 elasticsearch로 분석하려는 데이터 전송
- kibana를 통해 대시보드 구축
- Filter, Query DSL, Lucene Query Syntax 등을 이용해 원하는 정보 검색
- elasticsearch 데이터 csv 추출



1. AWS에서 최소한 t2.micro 이상으로 EC2 Instance를 생성하자 🏰

2. 생성한 EC2 Instance로 접속하자 👑

3. Elastic Stack (5.6.4) 을 설치하자 🏰

jdbc driver도 함께 설치하자 🏰

4. **Elastics Stack** 환경 설정을 하자 👑

5. Elasticsearch와 Kibana를 실행하자 🏰

6. week5_exercise_{id} 라는 Index를 생성하자 👑

7. 아래와 같이 mapping을 설정하자 👑

Field	Data Type
customer_age	integer
customer_card	keyword
customer_location	keyword
customer_sex	keyword
date_delivery	date
date_order	date
product_gps	geo_point
product_item	keyword
product_price	integer
product_quantity	integer
seller_rating	integer
seller_site	keyword

8. Logstash를 이용해서 mysql 데이터를 elasticsearch로 전송하자 👑

(단, @timestamp와 @version field 제거)

customer_age	customer_sex	customer_location	customer_card	product_item	product_gps	product_price	product_quantity	date_order	date_delivery	seller_rating	seller_site
52	여성	광주광역시	삼성	셔츠	36.288199573715545, 128.13057513572724	25000	3	2018-01-12 15:02:34	2018-01-16 00:41:34	3	옥션
29	남성	광주광역시	삼성	티셔츠	36.0869495071031, 127.10481983759166	14000	4	2018-01-09 11:57:47	2018-01-11 16:46:47	5	GS샵
36	여성	전라남도	신한	자켓	35.90825189035691, 128.0671312045079	20000	2	2018-01-02 09:24:16	2018-01-05 01:20:16	5	GS샵
38	여성	충청북도	삼성	코트	35.68831118876786, 128.133195592593	25000	1	2018-02-04 14:35:27	2018-02-07 00:57:27	3	티몬
49	남성	전라남도	신한	셔츠	36.513374488580546, 127.30635747850515	23000	3	2018-01-11 22:14:48	2018-01-14 11:31:48	3	티몬

기본정보

- host : 13.124.230.195:3306
- database : week5
- user/password : week5/week5
- table : week5_test

9. Kibana에서 Index Patterns를 등록하자 👑

Time filter field : date_order

11. Scripted Field를 생성하자 👑

Field	Value	type	format	detail
<i>dayofweek</i>	“평일” 또는 “주말”	string	string	<i>date_order</i> 를 기준으로 “평일” 또는 “주말” 중 하나의 값을 갖도록 설정
<i>hourofday</i>	“새벽” 또는 “아침” 또는 “점심” 또는 “오후” 또는 “저녁” 또는 “밤”	string	string	<i>date_order</i> 의 시간대를 기준으로 다음과 같은 기준으로 설정 - 0 ~ 5 : “새벽” - 5 ~ 10 : “아침” - 10 ~ 13 : “점심” - 13 ~ 17 : “오후” - 17 ~ 20 : “저녁” - 20 ~ 23 : “밤”

12. Discover에서 다음과 같은 질문에 답해보자

Time Range : Year-to-Date

- 1) 전체 건 수는?
- 2) 주말 동안 신한 카드 또는 국민 카드 결제 건수는?
- 3) 주별로 봤을 때 건 수가 가장 많았던 주는 언제인가? 그 주의 건 수는?
- 4) date_order를 ascending으로 Documents를 정렬했을 때, 가장 인기 있었던 product_item과 그 비율은?

13. Heat Map Visualization을 만들자

- 목적 : date_order를 기준으로 요일별 시간대별 건 수 시각화해서 요일/시간에 따른 판매 현황 파악
- 데이터 : 요일별, 시간대별 정보는 13페이지에서 생성한 Scripted Field 이용

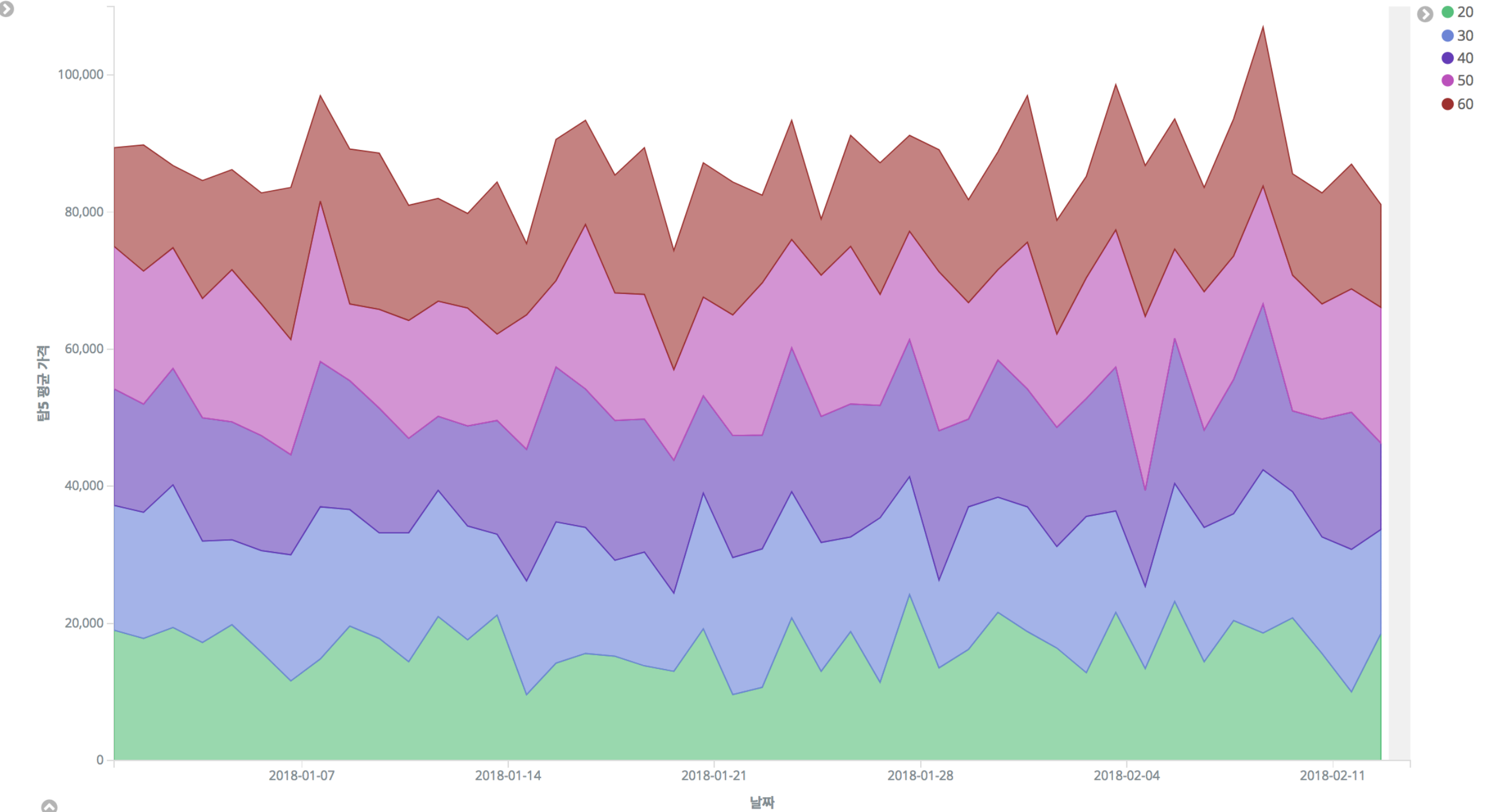
Time Range : Year-to-Date



14. Area Chart Visualization을 만들자

- 목적 : 일 별로 각 연령대가 구매한 상품중 seller_rating이 가장 높은 5개 상품의 product_price의 평균을 각각 추적
- 데이터
 - 일 별 : date_order를 기준으로 daily로 구분
 - 연령대 : customer_age를 10살 단위로 구분

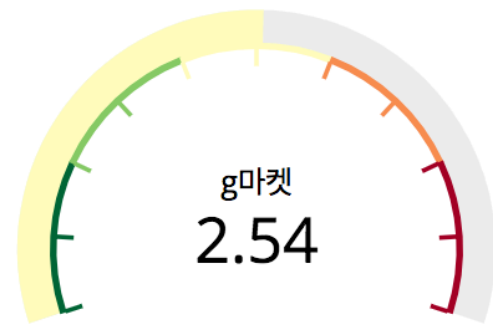
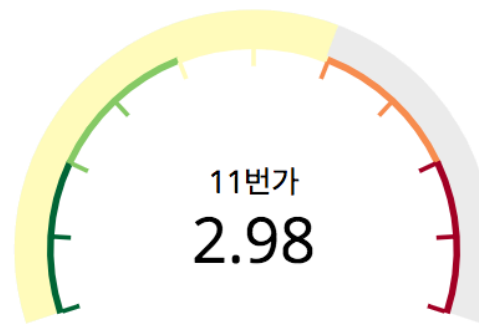
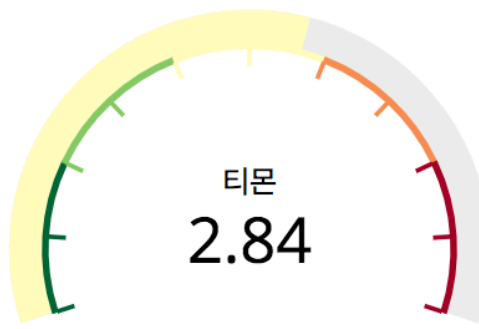
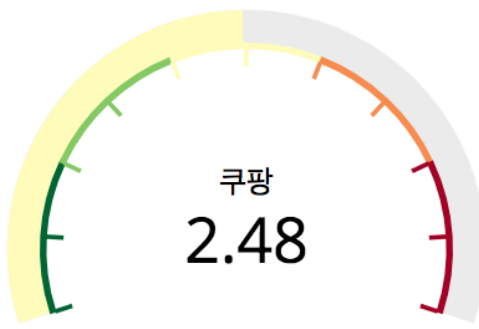
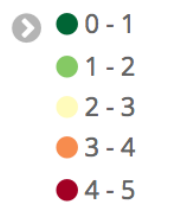
Time Range : Year-to-Date



15. Gauge Visualization을 만들자

- 목적 : 판매개수가 가장 많았던 5개 seller_site 별로 하위 50개 product_price의 seller_rating의 평균 추적
- 목표 구간
 - 1단계 : 0 - 1
 - 2단계 : 1 - 2
 - 3단계 : 2 - 3
 - 4단계 : 3 - 4
 - 5단계 : 4 - 5

Time Range : Year-to-Date



16. Data Table Visualization을 만들자

- 목적 : 요일 별로 product_price의 합이 가장 큰 3개 customer_location에 대해서 판매개수, 매출, 탑3 상품 시각화
- 데이터
 - 요일 별 : date_order가 “평일”인지 “주말”인지 구분
 - 매출 : product_price의 합
 - 탑3 상품 : product_price가 가장 큰 3개의 product_item

Time Range : Year-to-Date

평일: 요일

지역 ◆	판매수 ◆	매출 ◆	탑3 상품 ◆
울산광역시	976	16,640,000	티셔츠, 팬츠, 셔츠
경상북도	917	16,105,000	팬츠, 자켓, 코트
광주광역시	960	16,054,000	코트, 코트, 자켓

Export: [Raw](#)  [Formatted](#) 

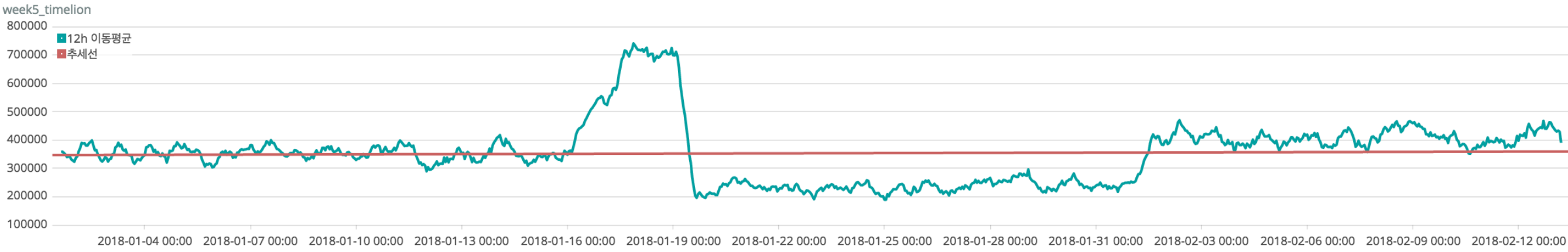
주말: 요일

지역 ◆	판매수 ◆	매출 ◆	탑3 상품 ◆
울산광역시	358	6,270,000	점퍼, 팬츠, 팬츠
서울특별시	345	5,941,000	니트, 자켓, 스커트
경상북도	337	5,752,000	청바지, 자켓, 니트

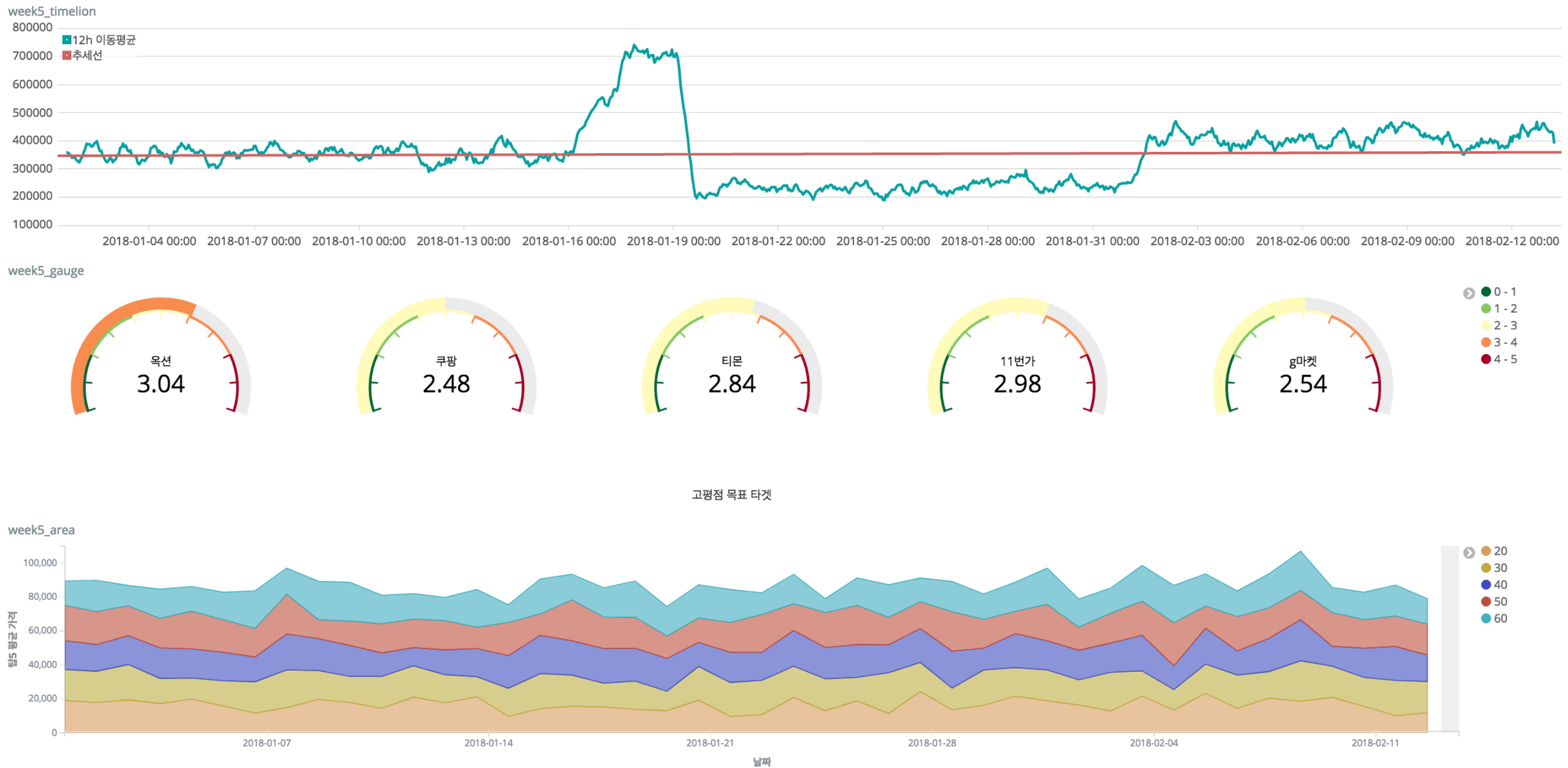
17. Timelion을 만들자 👑

- 목적 : 12시간 별로 매출의 이동평균을 구하고, 그 이동평균의 추세선도 함께 시각화
- 데이터
 - time field : date_order
 - 매출의 이동평균 : product_price의 sum의 moving average
 - 그 이동평균의 추세선 : (product_price의 sum의 moving average)에 대한 trend

Time Range : Year-to-Date



18. 앞서 만든 Visualizations을 적절히 배치해서 Dashboard를 만들자 🏰



19. Query DSL로 Dashboard에서 다음 조건에 맞는 데이터만 보여주자 👑

- 1) 25세 이상 경상도 사람 중에서 “신한”으로 결제했거나 18시 이후 구매한 사람들의 데이터
- 2) 75세 이상 서울시민이거나, 35세 이하 부산시민이거나, 30~35세 충청도민인 사람들의 데이터

20. Logstash를 이용해서 elasticsearch 데이터를 csv로 추출하자 👑

1) 다음 모든 조건을 만족하는 Documents만 출력

- product_item : 셔츠 또는 니트
- $2 \leq \text{seller_rating} \leq 4$
- customer_location : “부산”으로 시작

2) 출력 Field

- product_item
- product_price
- seller_rating
- seller_site
- customer_location

질문 및 Feedback은
gshock94@gmail.com로 주세요