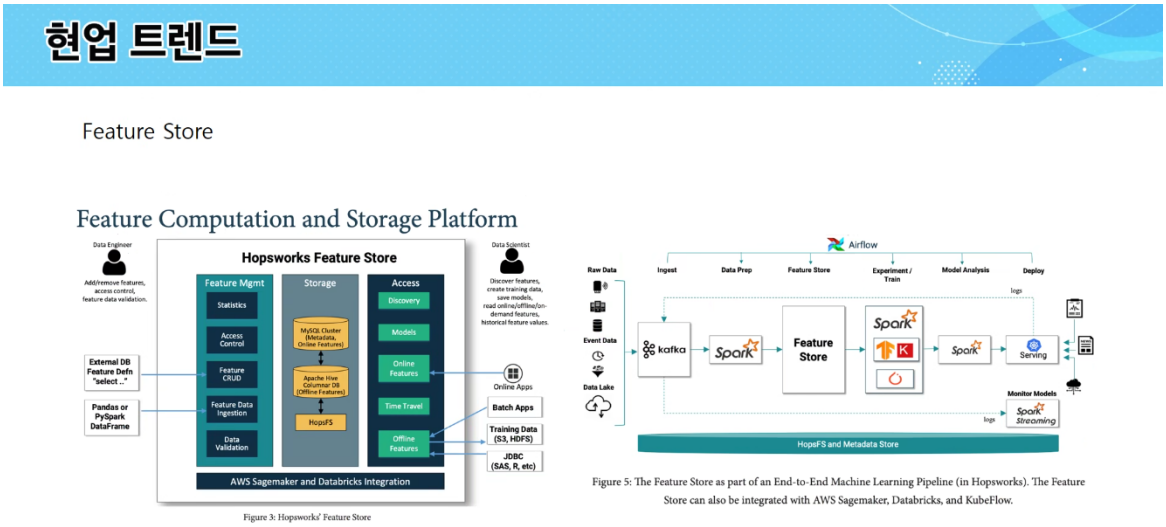
전문가 미팅 기록

▼ 현업 트렌드



▼ 멘토님 질문

1. 분석할만한 데이터가 안 나올수도 있다

   ⇒ 빨리 데이터를 모아서 분석해봐라

2. 멜론과 같이 실시간 화두가 되는 자료를 보여주는 것이 좋을 것 같아
3. 실시간으로 보여줄 거면 어느정도의 텀인지 정해놓고 테스트 해서 성능을 체크해라

▼ 사전 질문

1. 데이터 엔지니어링 분야의 동향 또는 최근 분위기를 알고 싶다.

```
- 엔지니어는 예전이나 지금이나 하둡 생태계 이용하고 있어
- 현재는 머신러닝 엔지니어들이 갑이라 데이터 엔지니어들은 데이터를 머신러닝 엔지니어 들의
  입맛에 맞게 줄 수 있어야 한다.
  (머신러닝 엔지니어는 정형화 되지 않은데이터를 원한다.but 데이터 엔지니어는 반대.)
- 예전에는 스파크 많이 사용했으나, 최근에는 flink (다람쥐)를 많이 사용함
```

```
  => 장애 복구 능력이 뛰어남
- 한국은 슬로우 스티터라 서서히 바뀌는 중
```

2. 주어지는 EC2 사양으로 수만~수백만에 이르는 뉴스 기사, 트위터 글을 저장할 수 있을지 궁금합니다. 또, 데이터 처리에 문제가 없을지 궁금합니다.

```
- EC2 하나로는 힘들 수 있어
  => batch process로는 어려울 것 같다. for loop 돌려야함
- 추가적으로 EC2를 쓰거나 노트북 같은 새로운 worker를 추가해
```

3. 뉴스, 트위터로부터 데이터를 수집한 후 순차적으로 여러 모델(혐오 필터링, 키워드 분석 등)에 사용하고 싶은데 Spark MLlib으로 분산처리가 가능한지 궁금합니다.

```
- MLlib을 사용하면 충분히 가능해
```

4. 실시간 데이터 동기화를 어떻게 수행하는지 알고 싶습니다.사용자의 요청에 대한 응답이 발생한 직후 실시간 데이터가 업데이트되는 경우 (사용자가 얻은 데이터와 데이터베이스의 `데이터 간 불일치`가 발생할 경우)

```
- 현업에서도 자주 발생하는 이슈
- 실시간에서 보여주기 위해서 redis를 사용하여 캐싱을 하는 방법을 써 봐
- 도출될 결과에 대한 성질에 따라 데이터의 정확성 이슈를 받아드리는 게 달라
  => 금융, 배송, 광고와 같이 데이터의 누락이 치명적인 경우, 주기적으로 정합성 테스트를 거쳐 방지하고 손해를 줄여야함
  => 그외의 경우라면 적당히 이상있는 데이터를 처리하는 rule만 정해주면 됨
```

5. 수집한 데이터의 이상/결측이 많아서, 혹은 그 수가 적어서 서비스 개발(분석 및 추천 등)이 어려우면 현업에서는 어떤 방법을 적용하는지 궁금합니다.

```
- 현업에서 흔히 발생하는 'cold start' 이슈
- 데이터가 부족한 경우, 현업에서는 비슷한 모델을 사용하여 일정 데이터를 쌓으면 모델을 갈아치움
  => 혹은 목업 데이터를 만들어주는 infra도 존재하기에 이것을 이용
- 수집한 데이터의 이상 및 결측이 많다는 것을 인지했다는 것은 이미 비상 (장애)
  => stop하고 장애를 우선적으로 해결해야함
  => data validation을 통해 null값이 일정치 이상 감지
  => 지우거나 default 값을 넣어줌
```

▼ 그 외 질문

1. 시각적으로 실시간 데이터를 어떻게 보여줄 수 있을까?

```
- 우선, 실시간이면 streaming으로 데이터를 받는 게 좋아
  => socket 기반으로 프로그램을 짤 수 있지만 힘들거야
  => spark streaming을 이용해서 소켓을 읽어오는 방식을 사용
  => 카프카 투 카프카 방식으로 카프카에 넣어서 데이터를 처리
  => 해당 데이터와 새로운 데이터를 다시 카프카로 넣어서 데이터 처리
```

2. DB를 NoSQL(MongDB)을 사용할지 RDB(MySQL)을 사용할지 모르겠습니다.

```
- MySQL을 사용해야 트랜젝션을 완벽히 관리할 수 있다.
- 단, 시간적 여유가 없고, 데이터로 도출된 값의 오차가 민감한 부분이 아니라면 MongDB를 사용해!
```