Data를 추출하는 자세

최수경 @ NBT

왠 자세?

- 데이터를 추출하는 기술적인 방법에 대한 이야기는 넘치도록 많이 있습니다.
- 기술이나 기법이 충분함에도 추출작업은 종종 실패하는데요. 이걸 줄일 수 있는 방법은 "자세 : Attitude"가 아닌가 생각합니다.
- 그래서 한번 정리 해봤습니다.

하나 : 문제를 뚫어져라 자세히 본다.

- 모든 문제 상황을 면밀히! 파악한다.
- 이때, "내맘대로 해석하기"를 주의해야 한다. ----->
- 나도 모르게 그런걸 어떡하라구요!
  - 새로운 시도가 도움이 됩니다. 소리내서 읽어본다든지, 손으로 써본다든지...
- 세세하게 쪼개고 뜯어서 문제를 분석해 봅니다.

캠릿브지 연결구과에 따르면 한
단어 안에서 글자가 어떤 순서로
배되열어있는가 하것는은
중하요지 않고 첫째번와 마지막
글자가 올바른 위치에 있것는이
중하요다고 한다. 나머지
글들자은 완전히 엉진찬망의
순서로 되어있지을라도 당신은
아무 문없제이 이것을 읽을 수
있다. 왜하냐면 인간의 두뇌는
모든 글자를 하나 하나 읽것는이
아니라 단어 하나를 전체로
인하식기 때이문다.

하나 : 문제를 뚫어져라 자세히 본다.

- 세세하게 쪼개고 뜯어서 문제를 분석해 봅니다. (cont...)
  - 예를들어 "8월 가입자의 일자별 클릭율"이 할일 이라고 하면
  - '8월가입자'의 기준은 가입시작/가입완료 중에 무엇일까?
  - '일별'이라면 8월 인가? 아니면 8월 1일부터 현재까지 인가?
  - '클릭'의 기준은 광고를 누른 것 / 최종페이지 랜딩 중에 무엇일까?
  - '평균'의 계산 방식은 어떤식이어야 할까? 아래 둘은 결과가 달라요. 그죠?

    1. 개인별 일별 클릭율을 계산해서 클릭율의 평균을 낸다.

    2. 일별로 노출과 클릭 숫자를 합계내서 그 결과로 클릭율을 뽑는다.

둘 : 결과를 예상해 본다.

- 일단 최종 추출 결과를 머릿속으로 그려보고,
  - 이걸 요청한 분과 얼굴을 맞대고 확인하는것이 BEST
- 대략 어느정도의 값/건수가 나올지도 생각해보고,

날짜	클릭율
8월1일	11%
8월2일	13%

- 그걸 모르니까 뽑는거 아니냐구욧. 왜 나더러 생각해 보라고 하죠?
  - 여러분의 '촉'을 믿으세요. 결과가 예상과 많이 다르다면 추출 방법(SQL, raw data, prcessing 과정...등등)에 문제가 있을 확률 99% (는 과장....--;;)

셋 : 작게 시작한다.

- 처음부터 다섯개 테이블을 조인하여 20개 컬럼을 나열하고...(X)

- 가장 중요한 테이블 한 개만 가져와서 결과를 뽑고, 점점 살을 붙여 나가죠.

- 저라고 한방에 20줄 쓰고 'go' 하진 않습니다. 그럴 때 원숭이 나무에서 뚝! 인거에요.

넷 : 모든 수단과 방법을 동원해 확인한다.

- 우리는 기계가 아닌데 데이터베이스와 대화하다보니 오해를 사는일은 부지기수!
- 개발좀 한다면 Test case가 대세... 우리도 먼가 이런 방향으로 가야할것 같아요!
- 예1) 제한된 조건을 주고 출력해서 결과 데이터를 눈으로 확인
    - 사용자별 일별 클릭율이면 내 아이디를 조건으로 해서 클릭율이 맞게 나오는지 보고 전체 유저 대상으로 확대
- 예2) 오늘의 총매출을 뽑는다면 어제의 값, 지난주 오늘의 값과 비교
- 예3) 완전히 새로운 방식으로 query를 짜서 비교.... (고수들의 방식!)
    - SQL은 한 문제를 풀 때 정말 다양한 방법이 가능하죠. 결과 추출 후 완전히 새로운 방식으로 짜서 두 결과를 비교하면 배우는 것도 많고 실력이 부쩍 늘꺼에요.

요약

- 하나. 문제를 뚫어져라 자세히 본다.

- 둘. 결과를 예상해본다.

- 셋. 작은 곳에서 시작한다.

- 넷. 모든 수단과 방법을 동원 해 확인한다.

화이팅 ^^/