목차

1. 빅데이터 개괄
2. 빅데이터 직무
3. 빅데이터 지식
4. 빅데이터 취업
5. **빅데이터 개괄**

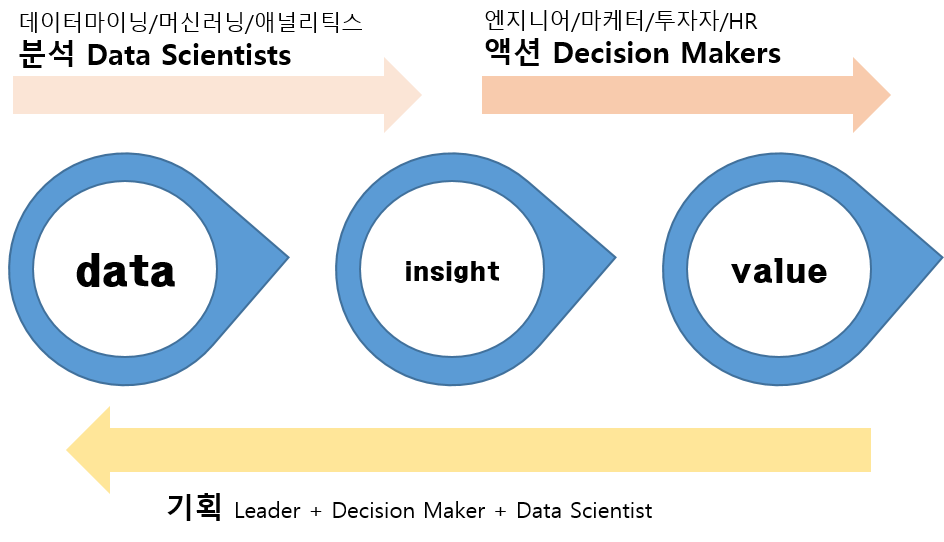
* 공개 데이터를 찾을 수 있는 곳(국내)

: data.or.kr , 네이버 데이터랩(<https://datalab.naver.com/>), 중소중견기업 데이터 유통 플랫폼([www.bigdata-sme.kr/](http://www.bigdata-sme.kr/))

* 해외 – 미국연방정부 및 세계 53개국 데이터 (<https://www.data.gov/>) 영국 데이터 (<https://data.gov.uk/>) OECD (<https://data.oecd.org/>)
* 해외민간 – <https://registry.opendata.aws/> AWS 데이터셋

**[빅데이터 실행]**

데이터에서 인사이트를 도출함으로써 가치를 창출해내는 과정 (빅데이터에서 사업적 가치를 창출해 냄)



* 빅데이터 실행을 가능하게 하려면 : 사업적 가치에 대한 비전을 세우고, 비전을 실현할 수 있는 인사이트를 상정하고, 그 인사이트를 도출할 수 있는 데이터를 찾는 과정이 필요

이러한 과정을 ‘빅데이터 기획’이라고 부르고, 빅데이터 기획을 위해서는 의사결정자와 빅데이터 전문가가 함께 전략을 세우는 것이 유리함.

* 디지털트랜스포메이션: 정보통신기술을 활용해 새로운 해결책을 제시하거나 운영 혁신, 사업기반 재구축 등으로 경쟁력을 확보하고 새로운 성장 발판을 마련하는 행동. => 기업이 기존에 보유하고 있거나 보유할 수 있는 데이터를 디지털화하고 재가공해서 기존에는 몰랐던 새로운 인사이트를 얻을 기회를 만듦
* 미래 기업에서 데이터의 중요성: 클라우드 플랫폼 인프라나 분석기술은 오픈소스 형태로 점점 더 저렴하게 배포되고 있지만 고객 데이터는 데이터를 수집할 고객 창구와 장비, 수집할 시간이 필요하다. 그래서 미래 기업 경쟁력은 차별화된 데이터 수집량에 달려있다.

**[빅데이터 직무]**

* 기업의 빅데이터 분석인력 직무명이 다르지만 역할은 비슷한 이유: 데이터 조직이 아직 성숙되지 않았거나(핵심사업이 저조해 디지털 트랜스포메이션에 대한 투자가 낮음) 유기적인 협력을 중요시하기 때문에
* 명확하게 빅데이터 직무를 나누면 ‘데이터엔지니어’, ‘데이터 애널리스트’, ‘데이터 사이언티스트’, ‘데이터 리서처’, ‘데이터 기획자(data project manager)’로 구분할 수 있음. 특별히 빅데이터 직무로 분류하지는 않지만 본연의 빅데이터 업무를 수행하는 사람은 ‘시티즌 데이터 사이언티스트’로 분류
* 데이터 엔지니어: 데이터와 데이터를 둘러싼 시스템을 책임, 데이터의 관리 수집 보관 처리 시스템의 개발, 구조설계, 유지보수를 담당함. 사내 데이터 분석 도구나 시각화 도구를 직접 개발하기도 함.
* 데이터 애널리스트: 데이터에서 기업의 현재 상태와 관련된 인사이트를 도출해 경영진에게 효율적으로 전달하는 업무를 담당. 여기서 인사이트란 단순한 통찰력이 아니라, 개선해야 할 문제를 찾아내고 이를 해결할 아이디어를 생각해내는 것. 주로 간단한 데이터분석 도구나 대시보드 등과 같은 다양한 시각화 도구를 활용해 데이터에서 인사이트를 발견함
* 데이터 사이언티스트: 통계방법론, 머신러닝, 데이터마이닝 지식을 바탕으로 정량적이고 과학적인 데이터 분석을 통해 데이터에서 인사이트를 발굴함. 미래 상황을 예측하거나 미래에 발생할 수도 있는 문제를 예방. 새로운 알고리즘을 습득할 수 있는 이론적인 이해도와 더불어 분석하고자 하는 업종에 대한 전반적인 지식, 가치 창출과 인사이트 발굴 능력 등을 골고루 갖추고 있어야 함.
* 데이터 리서처: 데이터를 분석하기 위한 새로운 알고리즘과 방법론을 연구,개발하는 일을 담당
* 데이터 기획자(데이터 프로젝트 매니저): 빅데이터 비즈니스 사이클의 계획과 원활한 운영을 책임. 목표 설계, 목표를 실행하기 위한 데이터, 시스템, 인력을 효율적으로 운영함. 데이터 시각화나 간단한 분석을 통해 프로젝트 방향을 정하고, 데이터 기술의 동향을 파악해 적절한 기술이 프로젝트에 활용되고 있는지 살펴야 함
* 제너럴리스트형 인재 : 스페셜리스트의 반대로, 특정 분야에 대한 고도의 전문지식은 부족하더라도 빠르게 변하는 스타트업에서 다양한 역할을 수행하는 데 거부감이 없고 새로운 비즈니스 기회를 찾는 능력이 뛰어난 사람
* ETL파이프라인을 구성하기 위한 애플리케이션의 집합, 분석과 시각화를 위한 플랫폼을 ‘데이터 플랫폼’이라고 부름. 데이터 플랫폼은 데이터를 수집하기 위한 Logstash, fluentD나 배달 성능 최적화나 안정성 향상을 위해 사용하는 kafka, Redis 등의 메시지 큐 서비스, 데이터 저장과 분석을 위한 RDB, NoSQL 데이터베이스, 시각화를 위한 스팟파이어, 태블로, 키바나 등의 애플리케이션 집합으로 구성됨.

**예시) 번개장터 데이터팀 소속 ㅇㅇㅇ 인터뷰**

인프라팀 소속 업무: 데이터 웨어하우스, 데이터 마트 구축, 사용자 로그 등 스트리밍 데이터 수집 시스템 구축/ 머신러닝: 추천시스템 모델 및 API 개발, 광고 시스템 예측 모델 개발 등.

수집데이터는 e-commerce 특성상 페이지 이동 로그, 상품 클릭 로그, 검색어 로그, 구매 목록, 사용자 정보, 유입 경로, 퍼널 데이터 등으로 하루에 몇 억 건 생성

데이터 저장에 이전에는 많은 비용이 소모되었지만, 저장 및 프로세상 관련 서비스가 개발되어 현재는 stream 데이터를 처리하기 위해 카프카나 키네시스 등을 이용.

OLAP과 OLTP의 차이점: 번개장터는 OLAP 데이터베이스를 사용하는데, 이러한 데이터베이스는 데이터를 일정 주기로 처리하여 분석용 데이터로 사용한다. OLTP는 사용자 구매, 판매와 같은 이벤트를 실시간으로 저장하여 풀스캔(데이터베이스에 포함된 데이터를 모두 확인해 원하는 데이터 추출)이나 쿼리(DB에서 특정 데이터를 요청) 작업이 불가능함.

엔지니어에 필요한 능력: 코딩, 수학, 영어 / 데이터 로그를 어떻게 수집할지, 수명은 어떤지, 오류 처리는 어떻게 할지 등 모델링과 관련한 부분을 고민할 때 데이터 엔지니어 지식이 필요

다양한 경험보다는 어떤 문제를 얼마나 깊이 고민했는지, 하나의 문제를 처음부터 끝까지 고민했는지가 주요하다. 이력서에 적힌 기술이 있다면 그 기술의 원리는 무엇인지 알아야 한다. (기술의 의미, 기술로 할 수 있는 것)

**데이터 애널리스트 업무**

* 데이터를 통해 조직의 현재 상태를 진단. 데이터 준비, 비즈니스 인사이트 도출, 데이터 시각화 등의 업무 수행. 정형 데이터 외의 새로운 데이터 수집 업무도 담당함. (예를 들어 브랜드 평판을 알고 싶을 때 SNS나 블로그, 커뮤니티 등 온라인 데이터를 수집)
* 근데 뭔가… 데이터분석에 가장 중요한 건 끈기인 듯.. 뭔가 변화가 생겼다면 시기별, 소비군별 등 다양한 카테고리로 변수를 넣어보면서 규칙을 발견해야 하니까 끈기있게 집중력을 유지하는 게 중요한 것 같음.
* 데이터 애널리스트에 필요한 지식 및 스킬: 모델링을 완벽하게 이해하고 있지 않더라도 데이터 분석 기법과 통계에 관련된 기초적인 지식을 갖춰야 함. 이런 지식을 갖춘 데이터 애널리스트는 전문가 집단과 비전문가 집단을 모두 이해하고 있기 때문에 두 집단 간의 소통에 도움을 줄 수 있음. (또 데이터분석가는 빅데이터팀과 경영진의 교두보 역할이라고도 볼 수 있음) , 데이터 준비와 기초적 분석을 위해 SQL, 엑셀, 기초 코딩 능력, 분석 소프트웨어와 시각화 툴을 자유롭게 다룰 수 있는 능력도 필요. 또한 쉽게 이해할 수 있는 시각화 자료를 활용해 설득력 있게 발표할 수 있는 능력도 중요.
* 다양한 팀과 소통해야 하기 때문에 불필요한 오해와 마찰을 최소화할 수 있는 소통 능력도 중요. 시각화에 대한 미적 감각과 색상,배치 등에 대한 고민도 필요함
* 성향적으로는 여러 집단 사이에서 중개자 역할을 한 경험, 중재자 역할을 즐기고 다양한 사람과 교류하고 소통하며, 예기치 못한 상황에 유연하게 대처할 수 있는 자질도 필요. 마지막으로 기업의 목표나 전체적인 프로세스 등을 염두에 두고 있으면 효과적인 인사이트 도출이 가능함.

**예시) 게임회사 데이터 애널리스트**

* 게임회사는 광고를 많이 하고, 광고 마케팅 효과를 최대로 올리기 위해 퍼포먼스 마케팅을 수행함. 같은 광고라도 타깃팅한 그룹별로 광고 성과가 다르기 때문에 마케터는 가장 성과가 좋은 타깃을 알아내기 위해 분석함. 광고집행 대비 매출 효율 (Return On AdSpend, ROAS), 광고를 클릭한 수, 게임으로 직접 유입된 수 등의 지표를 kpi로 활용. 이때 마케터들이 광고 성과를 분석하기 위한 데이터를 제공해주고 시각화 리포트를 전달하는 것이 데이터분석가의 역할임. 마케터와 상호작용 하여 어떤 광고가 가장 효과적인지 파악한 후 리포트를 유지보수함.
* 데이터 애널리스트는 도메인 지식이 중요한데 게임회사의 통계지표나 예측모델의 주요 목표가 실무 마케터에게 인사이트를 제대로 전달하는 것이기 때문임.
* 권장능력: 도메인 지식, 시각화 도구(오픈 데이터 등을 이용해 다양한 형태의 그래프를 그려보거나 대시보드를 만드는 연습이 필요) – 같은 데이터라도 어떻게 표현하느냐에 따라 뽑을 수 있는 인사이트가 다름 ex) 매출 지표를 시간에 따라 정렬하면 분기에 따른 트렌드를 알 수 있고, 연령대에 따라 시간순으로 정렬하면 연령대별로 특징적인 트렌드가 드러남….

혼자 하기 어렵다면 태블로 사의 ‘메이크오버 먼데이’를 추천(매주 특정 데이터가 공개되고, 가장 효율적인 시각화 아이디어를 생각해냄)

**데이터 사이언티스트의 업무**: 포뮬레이션(Formulation) – 당면한 비즈니스 문제를 데이터 분석 문제로 바꾸는 과정. 예를 들어 현 비즈니스 문제가 상황을 묘사하고 진단하는 것이라면 보유 데이터에서 통계치를 내거나 여러 종류의 데이터 간 관계를 파악하기 위해 회귀분석 등의 기법을 적용. 미래 목표 예측이 목표라면 회귀예측 뿐 아니라 머신러닝 또는 딥러닝의 예측 모델을 활용

또한 엔지니어가 구축한 파이프라인에서 분석에 사용할 데이터를 수집하고 데이터가 적합하지 않으면 조직 밖에서 데이터를 수집(웹크롤러 또는 api), 수집한 데이터는 필터링, 정제, 전처리의 과정을 거침. DB에서 가져오고 싶은 데이터만 가져오는 것(필터링), 더티데이터-오류 또는 누락, 모호하거나 비표준화된 데이터 정제, 변수의 활용을 바꾸는 전처리 과정

* 필요스킬: 데이터마이닝, 머신러닝, 통계기법 분야에 대한 수학적/통계학적 지식, 해결하고자 하는 문제에 가장 적합한 모델을 선정하고 결과를 정확하게 이해하기 위해 모델 작동방식과 알고리즘을 알아야 함. SQL, Python/R과 구축된 파이프라인을 이애할 수 있는 컴퓨터공학적 지식이 있으면 좋음.
* 필요성향: 스스로 질문할 수 있는 호기심(분석결과에 순응하기보다 왜 이런 결과가 나왔는지 탐구), 발전기술에 관심을 갖고 스스로 탐색해 적용하는 능력, 학습력과 의사소통능력

**DS 예시 – 네이버 지역 및 장소 추천 서비스 ‘스마트어라운드’ 추천서비스 개발자**

* 이용자의 성별, 연령대, 검색시간대, 현재 위치 주변 행사 등 요소를 고려해 상황에 적합한 장소 추천. ‘POI(Point of Interest)’- 정식 지리명은 아니지만 대부분의 사람이 알고 있는 주요 지명과 시설의 명칭. 네이버의 경우 보유한 데이터의 양이 방대하고 질이 좋기 때문에 추가로 데이터 수집을 할 필요가 없음.
* 기술적이론을 실제로 적용해보는 것이 중요. 국외의 기술이나 관련 보고서를 번역해 개인 홈페이지에 업데이트하거나 글쓰기..
* 실무에서 일어날 수 있는 문제: 실무는 수억건의 데이터 활용, 데이터가 정제되어 있지 않아 학계 논문보다 좋은 성능이 나오기 어려움. 따라서 새 논문보다 오히려 전통적이고 간단한 통계 기법을 선택할 때가 많음
* 업무에 필요한 능력과 성향: 학교에서 배운 방법을 이해하면서 실제 상황은 다를 수 있다는 열린 마음,유연한 태도, 무조건 코딩이나 모델링만 잘 하는 것보다 관련 분야에 대한 인사이트도 중요.
* 취업에는 데이터 전문가 업무가 주가 되는 회사인지, 배울 점이 많은 회사인지 파악하는 것이 중요. 데이터 업무가 보조 역할이 되는 경우에는 성장하기 힘들 수 있음.

**DS 예시2 – 삼성전자 메모리 사업부 데이터사이언티스트**

* 양산품질관리팀 데이터분석 프로젝트 팀에서 근무하면서 알고리즘/모델링 중 데이터베이스 구축 등의 업무를 맡음. 프로젝트: 센서 데이터가 특정 수치를 초과하거나 수치에 미치지 못할 때 알림을 제공하느 ㄴ모델, 불량 여부를 예측하는 모델, CNN 활용 불량여부 예측.
* 데이터사이언티스트로서 어려운 점: 이미 잘 작동하고 있는 프로세스의 일부분에 머신러닝 등의 새로운 모델을 도입하는 경우 어려움이 존재 – 안정적인 프로세스를 굳이 바꿔가면서 제대로 검증되지 않은 기술을 적용해야 하는지에 대한 반발이 있음. 복잡하고 정교한 아주 거대한 프로세스의 나머지는 건드리지 않고 일부분을 새로운 기술로 교체하는 것의 어려움. / 또 근무하고 싶은 분야의 전문지식 습득이 중요함 – 인터뷰이는 산학프로젝트를 통해 제조 데이터로 머신러닝을 진행한 경험이 있음.

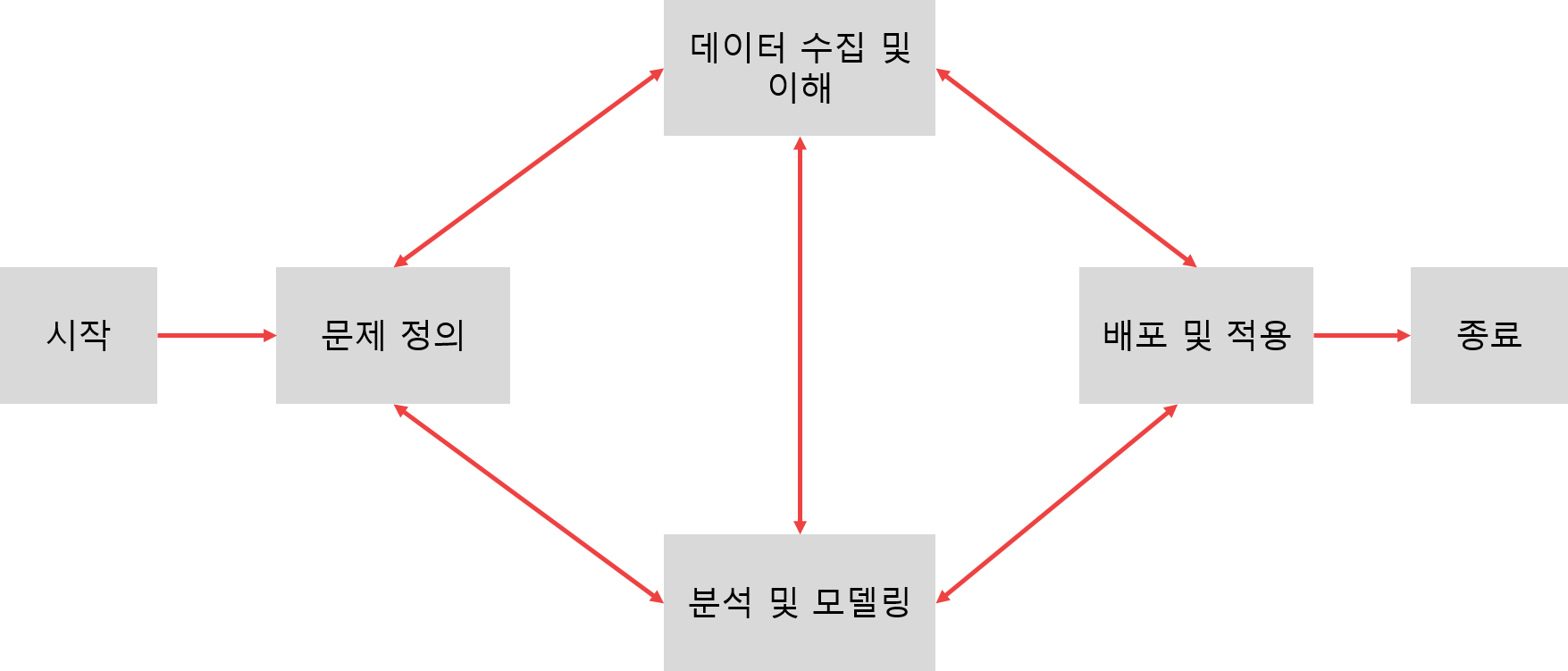
**시티즌 데이터 사이언티스트의 업무:** 빅데이터 지식과 기술이 발전하면서 데이터 분석 수요가 증가하는 데 비해 실제 데이터 전문가는 부족함. 이런 상황에서 데이터 분석이 불가능하기 때문에 빅데이터 전문가의 수요와 공급 차이를 빠르고 효율적으로 줄일 수 있는 해결책 중 하나가 ‘시티즌 데이터 사이언티스트’. 본연의 업무가 있고, 빅데이터에 관련된 지식을 능동적으로 습득함으로써 기초적인 분석 능력을 갖춘 인력을 Gatner에서 시티즌 데이터 사이언티스트로 부름. 이들은 자신의 분야 속 전문가이기 때문에 자신의 분야에서 어떤 데이터를 확보할 수 있는지, 확보한 데이터를 어떻게 사용하면 좋은지 확실히 알 수 있음

데이터를 분석하고 처리하는 전반적인 과정을 이해해야 함. 사내 데이터의 구성 및 저장이 궁금하면 DB와 쿼리에 대한 공부를, 데이터로 어떤 현상을 증명하고 설명하고자 한다면 데이터 시각화와 대시보드 제작을, 또 데이터 분석을 수행 시에는 자동화된 소프트웨어 툴을 사용함 (KNIME, Orange, RapidMiner, SAS 등)

* 한마디로 현업 부서에 섞여 있으면서 기초적인 데이터 지식을 가지고 간단한 분석을 통해 비즈니스 의미를 도출하는 사람 / 현업 전문가는 제공한 분석 결과를 보고 비즈니스적 의미를 제공하는 사람

**데이터 기획자의 업무**: 클라이언트와 소통하면서 요구사항을 데이터 전문 영역의 언어로 구체화해야 함. 이후 기획자와 클라이언트가 서로 같은 것을 생각하고 이해했는지 확인한 뒤 요구사항을 재설정함. 이후 프로젝트 수행을 위한 인적/전산/가용예산 등을 파악하고, 팀별 가용인력의 역량을 확인해 외부 인력을 고용, 또 프로젝트 종료 뒤에도 유지보수에 필요한 비용이 적절하게 잡혔는지 확인한다. 각 프로젝트 단계가 어떻게 진행되는지 확인하여 프로젝트의 병목이 없도록 함

1. **빅데이터 지식**



**[빅데이터 프로젝트 프로세스]**

1. 문제정의

문제 정의를 잘 하려면 다음 네 가지 단계가 필요하다.

* 문제를 명확하게 표현했는가?

더 이상 문제에 꼬리를 달 수 없는 상태로 만들어야 명확하다고 볼 수 있음

지원자 중 누구를 채용할까? -> 지원자 중 누가 이직할 확률이 낮을까? -> 지원자 중 누가 1년 내에 퇴사할 확률이 낮을까?

* 문제를 얼마나 잘 풀어야 하는지에 대한 목표를 설정했는가?

빅데이터 분야는 성능이 높아질수록 추가로 성능을 개선하기 위한 시간과 예산이 기하급수적으로 늘어남 따라서 이런 비용을 고려해 합리적인 성능 목표를 정해야 함.

* 결과물을 전달받아 사용하는 사람은 누구인가?

사용하는 사람에 따라 결과물의 형태가 달라져야 함

* 정의한 문제를 해결할 수 있는 자원이 있는가?

데이터가 있는지, 분석할 IT 환경이 갖추어져 있는지, 빅데이터 전문가가 있는지

1. 데이터의 수집 및 이해

* 가장 많은 시간이 소요되는 단계. 데이터 엔지니어는 IT환경 조성, 데이터 수집 및 관리 업무, 데이터 사이언티스트와 애널리스트는 데이터 수집하고 이해하는 업무.
* 모델링과 상호 관계에 있음. 모델링 단계에서 미흡한 점이 발견되면 다시 데이터 준비 단계로 돌아가서 데이터를 만들어야 함. 이를 방지하기 위해서 데이터 수집 및 이해 업무를 체계화하고 관련 IT 환경을 적절히 조성해야 함.

1. 분석과 모델링
2. 배포 및 적용

* 결과물을 사용자에게 적합한 형태 (코드, 소프트웨어, 발표자료 등)

**[시각화 공부]**

Python 지도 시각화는 folium 패키지를 이용

대시보드 활용 툴: Tableau, PowerBI, Google Analytics 등

시각화 공부는 크게 ‘구상’과 ‘구현’으로 나눌 수 있음. 구상은 어떤 종류 그래프로 시각화할지, 정보를 어떻게 배치해야 할지, 어떤 색상을 선택해야 할지 등을 스케치하는 것이다.

구상에 좋은 책: Information is Beautiful

구현은 상상한 시각화를 코딩으로 그리는 것이다

**구현을 위한 온라인 자료**

* Coursera, Data Visualization with Python (IBM)
* Coursera, Data Visualization & Dashboarding with R (Johns Hopkins University)
* Coursera, Data Visualization with Tableau (UC Davis)
* Coursera, Using Python to Access Web Data (University of Michigan)
* 데이터 시각화 교과서 (책만)
* 데이터 스토리텔링 (에이콘 출판)
* 대시보드 설계와 데이터 시각화 (책만)

**[데이터 분석 소프트웨어]**

* 깊은 코딩은 필요없지만 빠르고 간단한 데이터분석이 필요한 데이터 분석가들에게 필요

**[웹크롤링]**

접근 방식에 따라 세 가지로 나누어서 볼 수 있음

1) HTTP와 HTML을 사용하는 방법. Html 소스코드를 다운로드 한 후 이를 파싱해 데이터 추출

: requests – get/post 함수 , BeautifulSoup (문서 파싱에 사용)

2) 웹페이지의 URL을 알기 어려운 경우에 웹 브라우저를 이용해 사람이 직접 데이터를 가져오는 모습을 흉내내는 프로그램을 만들기

3) API를 이용

크롤러를 직접 구현하지 않고 대신해주는 웹 크롤링 서비스 : selenium

**공부자료**

* <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-046j-design-and-analysis-of-algorithms-spring-2015/>
* 파이썬으로 웹 크롤러 만들기 2판 (한빛미디어)
* 파이썬 웹 스크래핑 (에이콘출판)
* 파이썬을 이용한 웹 크롤링과 스크래핑 (위키북스)

**[프로그래밍]**

어떤 언어로 프로그래밍을 해야할까? 데이터 전문가가 가장 선호하고 앞으로 유망한 언어는 파이썬, R, 자바스크립트, 자바에 해당

1) 파이썬

인터프리터식 언어이기 때문에 컴파일 없이 바로 실행할 수 있고, 결과를 바로 확인할 수 있으며, 가독성이 좋아 개발시간을 단축할 수 있는 장점. 확장성을 고려한 디자인 덕분에 개발자가 오픈소스 패키지를 자유롭게 활용하는 것이 가능

2) 자바

컴파일식 언어로 코드를 실행시킨 뒤 ‘컴파일(프로그래밍 언어를 기계어로 번역하는 과정)’을 거쳐야 함. 이때 최적화를 할 수 있어서 인터프리터 언어에 비해 실행 속도가 빠름. 객체지향 언어이고 C C++ 보다 고나리가 쉽기 때문에 시스템 구축 언어로 많이 사용됨

3) 코틀린

안드로이드 app 개발 공식 언어 중 하나로 자바를 대체하기 위해 만들어짐. 자바가 실행되는 환경에서 항상 동작하는 프로그래밍 언어이기 때문에 기존에 개발돼 있는 자바 코드와 함께 실행되는 프로그래밍을 만들기 수월함.

4) 자바스크립트

서버와 웹 브라우저 모두에서 실행하는 코드를 만들 수 있는 장점. Node.Js라는 런타임 실행환경의 등장으로 서버 시스템에도 쓰이기 시작했으며, 리액트나 뷰 같은 프런트엔드 프레임워크가 유행하면서, 시스템 전반을 한 가지 언어로 구축할 수 있는 자바스크립트의 인기가 상승함. JSON이라는 표기법으로 객체를 쉽게 표현할 수 있고, 비동기 프로그래밍에도 유리함.

5) 알고리즘

* 시간복잡도: 알고리즘이 실행되는 데 걸리는 시간 (처리하는 연산의 양에 비례)
* 공간복잡도: 알고리즘이 사용하는 메모리의 양 (대용량 데이터를 불러올 때 용량부족에 대비하기 위한 지표)

**공부자료**

* MIT 온라인 강의 18.410J
* 애드위드, 인공지능을 위한 선형대수(카이스트 주재걸 교수)
* 코딩테스트: 생활코딩, 백준 온라인 저지, 리트코드, 프로그래머스, 해커랭크, 삼성 SW 엑스퍼트아카데미
* 모두의 알고리즘 with 파이썬 (길벗)
* Do it! 자료 구조와 함께 배우는 알고리즘 입문(이지스퍼블리싱)
* Introduction to Algorithms (MIT Press)
* 알고리즘 트레이닝 – 프로그래밍 대회 입문 가이드 (인사이트)
* 프로그래밍 대회에서 배우는 알고리즘 문제 해결 전략 (인사이트)
* 코딩 인터뷰 완전 분석 (인사이트)

**[수학 및 통계학]**

논문을 이해하거나 작성하기 위해 손실함수 설계, 최적화 알고리즘을 이해하기 위해 수학 기본기 필요

필요 단원 : 선형대수, 해석기하학(벡터 공간과 선형 사상의 개념에 직관성과 엄밀성을 부여해줌), 최적화(특정 제약 조건에서 주어진 목적함수의 최댓값이나 최솟값을 찾는 방법을 탐구, 모델의 파라미터 값을 찾을 때 활용함 – SVM은 Quadratic Programming 이차계획법을 사용하는 등 머신러닝 기법은 최적화로 정의할 수 있음), 통계학, 정보이론(머신러닝이 푸는 문제의 목적함수나 손실함수를 정의하는 데 적용)

공부자료

* <https://m.edwith.org/ai151?validMobileApp=true>
* <https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/>
* <https://www.coursera.org/specializations/mathematics-machine-learning>
* <https://youtube.com/playlist?list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo>
* <https://online.stanford.edu/courses/xfds110-introduction-statistics>
* <http://cs109.github.io/2015/>
* Linear Algebra and its Applications (Cengage Learning)
* Linear Algebra (Pearson)
* Fundamentals of Linear Algebra and Optimization with Applications to Machine Learning (WSPC) 난이도 상
* Convex Optimization (Cambridge University Press) – 강의자료와 무료 pdf (<https://web.stanford.edu/~boyd/cvxboox/>)
* Information Theory, Inference and Learning Algorithms (Cambridge University Press) – 무료 pdf (<https://www.inference.org.uk/mackay/itprnn/book.html>)
* 일반통계학(영지문화사)
* 확률의 개념 및 응용 (자유아카데미)
* 수리통계학(민영사)
* Introduction to Linear Regression Analysis (Wiley)
* 범주형 자료 분석 개론 (자유아카데미)
* 프로그래머를 위한 확률과 통계 (길벗)
* 프로그래머를 위한 선형대수(길벗)

**[머신러닝]**

* 연구용 오픈 데이터: UCI Machine Learning Repository, IMDB-Wiki, MNIST, ImageNet, Cifar10, VQA Dataset

하이퍼파라미터 탐색기법: 그리드서치, 랜덤서치, 베이지안 최적화

모델배포관리: 시간에 따라 모델의 정확도가 떨어질 수 있음. 시간 변화에 따라 데이터 분포가 달라지는 것을 ‘시간일관성 Time Consistency’ 부재가 발생했다 또는 ‘개념변화 Concept Drift’가 일어났다고 표현

문장임베딩: Dense Representation으로 분류되는 임베딩 방법을 주로 사용함. 밀집표현방식 임베딩은 희소표현(Sparse Representation – n개 단어면 n개 차원으로 만드는 원핫인코딩식 임베딩) 과 다르게 사용자가 지정한 차원으로 임베딩 결과를 얻을 수 있어 메모리 사용이 효율적임. 또한 임베딩 과정에서 단어 간의 연관관계 등 의미 및 문법적 관계 파악 가능. TF-IDF, skip-gram, Word2Vec, FastText, ELMO, GPT, BERT 등 방법이 있음.

**머신러닝 공부방법**

* MNIST, IRIS Dataset 등의 데이터셋에 머신러닝 알고리즘을 적용해 실전감각 쌓기.
* 머신러닝 공부에 가장 많이 쓰이는 자료는 Andrew Ng 교수의 머신러닝 수업. Coursera, Youtube, 하버드대학 CS109 : 데이터사이언스 에서 무료 수업 듣기 가능
* 책으로는 Mathematical for Machine (Deisenroth) 의 두번째 파트인 Central Machine Learning Probblems를 추천. 모델선택부터 회귀, 차원 감축, 분포 추정, 군집까지를 다룸. 코세라에 임페리얼칼리지런던에서 제공하는 동일 이름의 강의가 있음 이 강의보다 난이도가 낮은 책은 An Introduction to Statistical Learning (Tibshirani), 난이도가 더 높은 책은 Elements of Statistical Learning (Tibshirani), Pattern Recognition and Machine Learning(Bishop) 등을 추천
* 우리나라 책으로는 단단한 머신러닝(제이펍)이 입문서로 좋고, 원서 번역본으로 ‘패턴 인식과 머신러닝’(제이펍), ‘통계학으로 배우는 머신러닝’, An Introduction to Statistical Learning은 ‘가볍게 시작하는 통계학습’으로 번역되어 있음.
* 처음부터 프로그래밍을 이용해 머신러닝을 공부하면 핸즈온 머신러닝을 추천.
* 딥러닝 공부는 입문자에게는 김성훈 교수의 ‘모두를 위한 딥러닝’ , 중급 수준 딥러닝에서 이미지 처리 분야는 스탠퍼드대학 CS231n, 자연어 처리는 CS224n 추천. 딥러닝 프레임워크를 익히고 싶다면 부스트코스 강의를 추천하며, 관련 교과서는 Neural Networks and Deep Learning(Springer)가 추천됨. 한국어 교재는 기계학습(한빛미디어)를 추천.
* 강화학습은 UCL(University College Londeon) 교수 David Silver의 강화학습 온라인 강의를 추천. 유튜브 채널 ‘팡요랩’에서 해당 수업을 한국말로 풀어 설명하는 영상이 있음.
* 강화학습에 딥러닝을 적용한 심층 강화학습 분야는 UC 버클리 Sergey Levine 교수의 CS285 수업, 스탠퍼드 대학 Emma Brunskill 교수의 CS234 수업 추천. 최적 제어 관점이 궁금하면 MIT Dimitri Bertsekas 교수의 ‘Reinforcement Learning Course at ASU’를 추천
* 강화학습 교과서로는 Reinforcement Learning : An Introduction (Sutton)이 널리 활용됨. 강화학습 역사, 문제 정의, 전통적 강화학습 해결법을 공부 가능. 번역본: 단단한 강화학습(제이펍). 이 책보다 더 난이도가 높은 책은 Reinforcement Learning and Optimal Control(Bertsekas), Reinforcement Learning: Theory and Algorithms(Alekh Agarwal)
* 서울대 최병선 교수가 마이클 조던 교수의 책 목록 기반으로 머신러닝 공부용 책 목록을 업데이트하고 있기 때문에 데이터 리서처를 꿈꾸면 해당 커리로 공부 추천

공부자료

* DeepMind 유튜브 채널 RL Course by David Silver
* 팡요랩 유튜브 채널 강화학습의 기초 이론
* <http://rail.eecs.berkeley.edu/deeprlcourse/>
* <http://cs229.stanford.edu/>
* <https://youtube.com/playlist?list=PLKs7xpqpX1beJ5-EOFDXTVckBQFFyTxUH>
* <http://seslab.kaist.ac.kr/xe2/page_GBex27>
* <http://gilbut.co/c/21092880Eu>
* <http://web.stanford.edu/class/cs224n/>
* <http://web.mit.edu/dimitrib/www/RLbook.html>
* <https://www.coursera.org/specializations/deep-learning>
* <https://sites.google.com/view/cbsdataly/essays>

위에서 언급되지 않은 책들 중 추천 목록

Deep Learning(MIT Press), Reinforcement Learning and Optimal Control(Athena Scientific), Machine Learning: A Bayesian and Optimization Perspective (Academic Press), 케라스 창시자에게 배우는 딥러닝(길벗), 밑바닥부터 시작하는 딥러닝(한빛미디어), 기계학습(한빛아카데미)

기타자료

* OpenAI Spinning Up Official Documentation – <https://spinningup.openai.com/en/latest/>
* Reinforcement Learning: Theory and Algorithms – <https://rltheorybook.github.io/>

**[데이터파이프라인]**

데이터의 이동이 체계화, 자동화된 것으로 수집, 가공, 처리, 저장이 함께 일어남. 예를 들어 인터넷쇼핑몰에서 구매가 발생하면 구매기록 데이터가 생성, 수집된 후 1차 클렌징되어 데이터 웨어하우스에 저장되는데 이 경로를 파이프라인이라고 부름.(ETL)

* 데이터 아키텍쳐 설계가 중요한 이유: 데이터의 유입 및 변동 가능성 (데이터는 바운디드데이터와 언바운디드 데이터로 나뉨. 바운디드는 데이터 수가 바뀌지 않는 데이터, 언바운디드 데이터는 추가 데이터가 유입될 가능성이 있는 데이터) , 데이터 처리 방식 (배치-일정주기 한방에 처리와 스트리밍-연속적 데이터 처리, 람다 아키텍처 – 배치와 스트리밍 둘 다에 적용 가능), 데이터 저장방식 등을 설정해야 해서

**4. 빅데이터 취업**

**빅데이터 직업시장 이해**

- 직무 정하기 : 스스로 숫자에 강하고 풀리지 않는 문제의 답을 찾기 위해 몇주, 몇 달 동안 끊임없이 고민하는지, 배운 것을 응용하는 능력이 뛰어난지, 스토리텔링을 이용해 설명이나 설득을 잘하는지 질문하기.

데이터와 관련된 전체적인 시스템 설계를 좋아하면 데이터 엔지니어, 스토리텔링을 좋아하면 데이터 애널리스트, 숫자에 강하다면 데이터 사이언티스트, 깊게 연구하고 이해하는 것이 즐겁다면 데이터 리서처를 선택.

무엇을 잘하는지, 어떤 일을 할 때 즐거움이나 성취감을 느끼는지 고민하기. 지원하는 빅데이터 직업의 특성에 맞는 강점을 구별해 이력서를 작성하면 좋은 인상을 심어줄 수 잇음

* 산업 도메인 정하기: 도메인 지식이 부족하면 다른 것을 강점으로 내세워야 하고, 빅데이터 고나련 전무 지식을 갖췄더라도 신입과 같은 대우를 받을 수도 있음. 최대한 많은 데이터를 다뤄보면서 구체적인 경험을 쌓아보면 관심 도메인이 구체화될 수 있음.

학습으로 전문성을 쌓을 수 있는 분야가 아니더라도 평소 친숙한 분야라면 이를 바탕으로 다루게 될 데이터와 발생 가능한 문제, 그리고 해결책을 예상 가능. ㄷx) 게임 시스템의 허점을 이용해 게임 내 재화를 획득하는 어뷰징 행위나 마우스, 키보드와 같은 하드웨어 또는 소프트웨어를 조작하는 불법 매크로 등의 문제를 예상하고, 어떤 데이터를 수집해 무엇을 살펴봐야 할지, 어떤 방법으로 문제를 해결할 수 있을지 구체적으로 고민해 보는 것이 좋음.

도메인 지식이 부족한 분야에 지원하더라도 면접 전 관련 분야에 관련된 키워드와 발생할 수 있는 문제와 해결책, 진행 가능한 프로젝트와 이를 통해 해당 기업이 얻을 수 있는 가치 등을 고민하고 자신의 이력이나 경력과 연관시켜 어필하는 것이 좋음.

* 공공기관 중 새롭게 조직을 신설한 경우 데이터 엔지니어의 업무에서 데이터 사이언티스트의 업무까지 광범위한 역할을 요구함 => 위계질서가 엄격하고 보수적인 문화인 점은 단점이 될 수 있음. 딥러닝과 같은 최근 기술들은 성능이 뛰어나지면 모델의 결과를 설명하기가 어렵고 안정성이 검증되지 않아 실무에 적용하기가 어려움. 따라서 검ㅈ으되고 안정적인 방식으로 문제를 해결하기 때문에 도전적이기 어렵
* 좋은 회사의 기준: 데이터팀에 투자를 많이 하는 회사가 성장에 도움이 됨. 데이터 관련 인력을 얼마나 많이 채용하고 홍보하는지 보기, 면접관 중 데이터팀 실무자가 없다면 입사 후 데이터 업무 비중이 낮은 것, 채용 과정 중 별도의 코딩 테스트나 데이터 분석 프로젝트가 포함되어 있으면 빅데이터 전문가로서의 역할을 기대하고 있는 것. 스타트업에 취직한다면 빅데이터 기술력만큼은 대기업에 뒤쳐지지 않는 회사에 취업해야 함.
* 면접 마지막에 궁금한 것에 대해 질문 받을 때는, 궁금한 것이 없더라도 질문을 많이 하면 정보를 얻을 수 있음. 빅데이터 프로젝트를 진행하면서 문제가 된 점 등을 물어보면 실제 취업 중 겪는 문제를 간접적으로 알 수 있음. 장비부족(데이터팀에 대한 투자 정도를 알 수 있음), 프로젝트 마감시간 부족(근무 강도), 데이터 팀의 규모, 회사에 도움을 준 데이터 팀의 프로젝트(이전에 진행한 프로젝트, 앞으로 어떤 업무를 맡게될지 알 수 있음)

**경쟁력 있는 신입이 되기 위해 필요한 것:**

1) 인턴십: 해외인턴십 Top 50 Data Science Internship Program으로 검색, 그 중 Infosys의 Instep이라는 인턴십 프로그램은 매년 100명 내외의 인턴들을 뽑음(서류,전화면접, 짧게는 6줒 길게는 16주까지 원하는 길이의 인턴십 진행),

2) 연구경험:

대학 연구 또는 기업 연구 인턴십 지원

\*\* 대학 연구실 인턴십을 위한 콜드메일 가이드: 콜드메일은 최대한 간결하고 공손하게 작성해야 함.

메일 제목에 자기 소속과 이름, 콜드메일 목적을 분명히 밝히기(~연구실 겨울 인턴 공고를 확인하고 연락드립니다), 어떤 계기로 연구실 인턴에 관심을 가지게 되었는지 설명(어떤 경험을 하고 싶고, 왜 연구실과 연구에 관심을 갖게 되었는지 전공과 경험에 연관지어 설명), 첨부파일로 발신자의 이력인 CV와 학부성적증명서를 첨부했다는 것을 알리고 파일 첨부하기.

3) 공모전: 캐글(외국), 빅콘테스트, 데이콘, 카카오 아레나, 공공데이터 활용 빅데이터 분석 공모전 등의 데이터 분석 플랫폼 – 이런 대회에서는 높은 점수를 얻기 위해 모델링하기 보다는 대회에 참가해 배운 내용을 정리하는 것이 좋음. 대회 경험을 통해 공부한 내용을 기술 블로그에 정리하고, 이력서와 함께 제출하면 좋음. 기술 블로그에 정리한 내용을 바탕으로 해결법을 도출하기 위해 얼마나 많이 고민했고, 얼마나 성장할 수 있었는지 답변하면 좋은 인상을 줄 수 있음.

4) 개인프로젝트: 자기 주도적이고 적극적인 면을 강조할 수 있는 스펙. 그러나 꼭 끝까지 완성하는 것이 중요함.

5) 인사이트 도출 훈련: 데이터 분석 목적은 인사이트 도출임. 다양한 데이터 분석 방법을 연구하는 것도 문제 해결에 상상력을 더해 넓게 이해하기 위해서임. 인사이트 도출 훈련으로는 Makeover Monday 프로젝트가 있음

6) 링크드인과 깃허브로 온라인 프로필 관리하기.

링크드인- 현업 전문가의 명함을 받았다면 링크드인에서 1촌 신청하기. 자연스럽게 연락을 이어갈 수 있는 수단으로 활용 가능함. 또 예비 면접관이 깃허브에 찾아오는 경우가 있으므로 프로필 관리가 필요

7) 헤드헌터 – 헤드헌터에게 연락이 오면 일단은 이직할 생각이 없더라도 최대한 좋은 관계를 이어가는 것이 좋음. 언제 다시 제안을 받을지 모르기 때문에. 또 먼저 헤드헌터에게 연락하고 싶다면 인크루트, 잡코리아와 같은 취업 사이트를 이용해 직접 헤드헌터에게 연락하거나 자신의 정보를 작성. 제안을 받았을 때는 헤드헌터가 어떻게 자신을 찾았는지(확인한 경로 중심으로 정보를 업데이트해나가면 더 좋은 제안을 받을 수 있음), Job Description과 전임자가 급하게 퇴사했는지 혹은 신설업무인지 등을 확인해야 함. 또 2명 이상의 헤드헌터와 교류하면서 업계 동향을 파악하고 다양한 조건을 받는 것이 좋음

**취업절차**

서류준비는 기업의 공고에 맞게, 면접준비는 인성과 직무(데이터 관련 경험에서 1차원적인 경험담 보다는 왜 그 알고리즘을 선택했고 해당 알고리즘을 적용하는 과정에서 겪었던 애로사항, 해결법, 결과와 한계를 논리있게 설명하기, 데이터 분야에서는 기초 수학 및 통계학에 관련된 지식 질문을 받는 경우도 많음) 임원면접은 열정, 관심도, 살아오면서 경험한 일, 특별했던 일, 앞으로 회사에서 이루고 싶은 일 등에 대한 답변을 준비

해외취업의 경우 팀 단위 상시채용이 많음. 각 회사의 공고를 전부 확인하는 것은 어렵기 때문에 Glassdoor Job Search, Indeed 등의 해외 인사/채용 사이트나 링크드인을 이용해 확인하기. 모든 회사가 채용공고를 등록하지는 않기 때문에 평소 관심있는 기업 리스트를 정리해서 자주 확인할 필요가 있음. 빅데이터 직무는 명확한 요건이 분류되지 않았기 때문에 직무 Responsibility와 요건Qualifications 항목을 잘 살펴야 함. 또 외국인 취업비자를 후원할 여력이 되는 회사인지 확인해야 함.

해외기업은 이력서 양식이 정해져있지 않기 때문에 이력서 내용 뿐만 아니라 디자인에도 신경을 써야 함. 개인정보를 많이 요구하지 않으므로 이름, 연락처, 깃허브 블로그 등 개인 페이지 주소만 입력함. 본인 사진도 첨부할 필요가 없음. 보통 이력서 확인에 20초를 할애한다고 하므로 A4 한페이지나 두 페이지 안에 필요한 내용을 정리하는 것이 좋음.

이력서 양식을 보는 웹 사이트로는 Canva, Resume, Overleaf 가 좋음. 이력서 글씨체는 명조체(Serif)-가독성 좋은 Georgia, Bell MT, Garamond, 고딕체(san-serif) – 현대적이고 도시적인 느낌의 Arial, Tahoma, Century Gothic, Lucia Sans 등.

이력서는 네 부분으로 나눠서 기본사항, 학력사항, 이력사항, 기타사항을 정리.

1) 기본사항: 이름과 연락처 정도만 기재

2) 학력사항: 최종학력 or 취득예정 학력을 기재. 학교명, 학위명, 재학기간(졸업예정일 가능), 학위 과정 중 지원하는 직무와 관련있는 수업명을 기재하면 장점으로 강조 가능. 만약 수업을 썼을 시에는 수업 관려 질문에 잘 대답해야 함.

3) 이력사항: 근무했던 회사, 직책명, 근무기간과 직무 세부사항 기재. 무엇을 ‘했는지’ 보다는 무엇을 ‘이뤘는지’를 강조해야 함. ‘Action verb for CV’를 검색해 영어 동사 리스트를 참고하고 수동태를 삼가야 함.

Ex) 홈페이지 유저의 접속 데이터를 분석해 새로운 마케팅 전략을 수립했다는 것을 이력서에 쓰고 싶다면 아래와 같이 실제로 팀에 기여한 바를 중심으로 기술하고 성과를 숫자로 표현할 수 있다면 숫자로 표현해야 함.

* 무엇을 했는지 관점: Responsible for monitoring Google Analytics to propose new marketing strategies
* 무엇을 이뤘는지 관점: Delivered a 15% exposure boost by promoting novel ads strategies through audience analysis with Google Analytics

만약 채용공고에 특정 언어나 프레임워크에 대한 언급이 있다면 해당 기술을 이용해 산출한 결과물을 기재하는 것이 좋다. 이력사항은 한 직무당 3~4줄, 맡은 기간에 따라 더 길거나 짧게 쓴다.

4) 기타사항: 학력과 이력 외에 채용에 도움이 될만한 정보를 정리

- 대학교 동아리 임원, 대외활동, 온라인 공개수업(MOOC) 수강 코스, 자격증, 수상경력 등을 기재. 또 흥미롭거나 특이한 내용을 기재하면 면접관의 관심을 끌어 자연스러운 대화 가능

- 이력사항과 별개의 연구 경력이나 연구 성과 또한 기재. 개인 연구라면 연구 내용, 팀 연구라면 팀에서 맡은 직책이나 소속 기관을 기재. 연구내용을 컨퍼런스나 저널에 등재했다면 해당 내용까지 기재.

별도의 CV를 요구하는 경우

* A4 한페이지 이내에 3~4 문단으로 이력서에 담지 못한 자신의 이야기, 경력 목표, 장점, 지원하는 회사와 부서, 직무에 대한 이해도를 하나의 이야기로 엮어 작성함. 에세이는 주로 회사와 직무에 대한 이해도와 관련 산업에서 지원하는 회사의 위치나 역할에 대해 제대로 이해하고 있는지 질문을 제시함

**면접준비**

* 면접 전 꼭 Glassdoor Job Search에서 해당 업무의 면접 질문을 검색해 미리 확인. 회사에 지원한 이유, 지원한 업계에서 일하고 싶은 이유, 자신의 경력 계획 등 기본적인 사항에 대한 답변을 미리 준비.
* 보통 면접은 세 단계로 인사담당자의 fit interview(지원자의 준비 정도를 확인하는 인터뷰, Fit Interview Questions라는 검색어로 예상 질문 확인 가능. 약 50개 정도의 질문에 답할 수 있을 정도로 준비, 핏인터뷰에서 좋은 점수를 받으려면 최대한 지원 직무 관련 질문에 답변할 수 있도록 질문함. 간혹 핏 인터뷰에 실무자도 참여하기 때문에 본인이 지원 직무에 적합하다는 것을 보여주는 게 중요함) Technical Interview(지원 업무에 필요한 필수 지식 뿐 아니라 이력서 기재 경험을 해당 업무와 연관지어 설명하는 연습 필요, 간단한 코딩테스트도 진행될 수 있음. 데이터 애널리스트의 경우 즉석에서 발표 자료를 만들게 하거나, 리서처는 학위논문이나 실무 지식과 관련한 질문이 들어옴. 기술질문에 자신이 진행한 프로젝트나 분석사례를 연결해 답변할 수 있다면 좋음. 테크니컬 인터뷰 관련 질문으로는 “데이터 과학자와 데이터 엔지니어를 위한 인터뷰 문답집(제이펍)” 을 추천) 임원급 면접관이 주관하는 최종 면접(개괄적인 경력, 회사에서 이루고자 하는 개인 목표, 첫번째 인터뷰와 같은 질문을 하기도 함. 지원한 회사와 비슷한 비전을 가지고 있는지 평가하기 때문에 회사 홈페이지를 꼼꼼히 살피거나 ‘Glassdoor Job Search’의 후기를 통해 회사에 대해 많이 알아둬야 함)
* 면접 전 인사담당자에게 면접 성격이나 준비 사항에 대해 체크하는 것도 좋음. 예정된 면접이 테크니컬인지 핏인터뷰인지, 면접관에는 어떤 사람이 포함되어 있는지 실례되지 않는 선에서 최대한 질문하기. 면접관으로 포함된 사람을 링크드인에서 검색할 수 있다면 경력이나 전공, 관심 분야를 미리 파악하여 자신과의 교집합을 만들고 면접에서 부각하는 것도 좋은 전략.
* 끝으로 롤모델을 만들고 링크드인, 깃허브 등의 SNS 프로필을 활용해 여러 전문가의 경력을 살펴보아, 그들의 선택을 공부해보는 것이 취업에 도움이 됨. 목표로 하는 직책명, 산업군, 회사명 등을 검색해 최대한 많은 프로필을 확인하다보면 링크드인 프로필이 해당 롤모델과 비슷한 사람을 자주 추천하게 됨. 연관된 사람들이 어떤 과정을 거쳐 지금 위치에 올랐는지, 무엇을 공부해야 하는지 정리하고 학부전공과 지금 하는 일의 연관성을 살펴보는 것도 도움이 됨. 경력 중간에 MBA나 다른 대학원을 진학했는지도 참고사항임. 기억에 남는 프로필이 있다면 따로 연락해 조언을 구하는 것도 인맥 관리 방법임. 링크드인 서비스 목적이 직장 네트워킹이기 때문에 일면식 없는 사람과 메시지를 주고받아도 괜찮음. 능동성이 중요!