6
랭킹이 관건이다.

난데없이 "랭킹"이란 말을 꺼내면 월드컵 관련 FIFA 랭킹부터 떠올리는 이들이 많을 것이다. 전 세계의 축구 팀들을 일등서부터 꼴찌까지 나열해 놓고 번호를 매긴다는 것은 사실 쉬운 일이 아니다. 그런 작업에는 많은 변수가 요구될 것이며, 그 모델을 짜는 방법이나 어떤 데이터에 더 중점을 두었느냐에 따라 순서가 완전히 뒤바뀌게 나올 수도 있다. 게다가 축구란 경기 당일 선수들의 컨디션이나 다른 많은 요소들이 작용하기 때문에 변방의 작은 나라가 난데없이 유럽의 강호를 혼쭐나게 할 수도 있다. 그래서 월드컵이 늘 몇몇 나라가 메달을 쓸어가는 올림픽보다 더 흥미로운 국제 스포츠행사로 꼽히는 것이기도 하다. 그런 이유로, 등수를 맹신하는 것은 무의미하고 또 바람직하지도 않아 보인다. 미국의 대학들을 일렬로 나열해 놓고 모든 면에서 반드시 하버드(Harvard)가 스탠포드(Stanford)보다 낫다고 주장하는 식의 논리와 같다고도 할 수 있다. 다만, 랭킹을 단지 참고로 사용한다는 가정 아래 FIFA 랭킹 11위와 57위가 맞붙었을 때 그 게임의 결과를 국적과 감정을 빼고 객관적으로 예측한다면 어느 팀에 베팅을 해야 할 것인지는 비교적 쉬운 선택이 될 것이다.

의사결정의 끝은 랭킹에 근거한 선택

사실 의사결정이란 여러 옵션을 바람직한 순서대로 나열하고 고르는 것에 다름 아니다.

마케팅을 한다면 어떤 상품을 팔고자 할 때 어떤 대상부터 상대할 것인가? 사원을 뽑을 때 그 많은 지원자들 중 누구를 우선적으로 채용할 것인가? 투자를 할 때 어떤 펀드나 주식을 고를 것인가? 휴가를 간다면 이 넓은 세상에서 어디부터 갈 것인가? 그러한 의사결정의 끝은 선택이다.

그리고 모든 결정이 그렇듯 그러한 선택은 한가지 변수로만 이루어지지 않는다. 간단해 보일 수도 있는 휴가지의 선택도 문화, 예술, 음식, 날씨, 유적지, 명소, 호텔, 항공료, 할인 여부, 가이드 여부, 목적지까지의 거리, 교통수단의 편리, 언어소통 등 수많은 요소들이 어우러져 있다.

게다가 그 모든 변수들이 같은 가치를 지니지도 않는다. 예를 들어 어떤 사람에게는 먹을 거리가 박물관 등 문화적 요소보다 더 중요할 수도 있으며, 그런 선호도는 개개인마다 다 다르기 때문이다.

예를 들어, 자동차를 구입할 때 고려해야 할 변수들은 가격, 브랜드, 모델, 인지도, 연료비, 연료타입, 섀시, 인테리어, 좌석, 감촉, 색깔, 파워, 가속력, 코너링, 트렁크 크기, 안전도, 뒷좌석 크기, 할인 여부, 애프터 서비스, 수리비용, 스테레오 성능, 기타 옵션 등 더욱 다양해지겠지만, 결국은 어느 변수들에 가장 중점을 두느냐에 따라 개개인의 선택이 달라지는 것이다.

그 모든 면에서 우월한 차종이란 없는 것이고, 브랜드는 좋은데 가격이 안 맞는다던가 가속력은 우수한데 연비가 너무 많이 든다는 등 변수들 간의 조합도 결정에 영향을 미치게 마련이다.

민주사회의 시민으로서 후보자에게 표를 주는 것도 랭킹에 근거한 선택이다. 사실 정치적 관점의 차이란 수많은 정치적, 때로는 비정치적 요소들 중 어떠한 변수에 더 가중치를 주느냐의 차이라고 볼 수 있다.

즉, 경제, 외교, 안보, 교육, 세금 제도, 복지, 환경 문제, 지역 이익, 그리고(미국의 경우) 종교관이나 총기 휴대에 관한 이슈까지도 포함한 다양한 변수들에 유권자들은 각기 다른 가중치를 주는 것이고, 그 결과는 각 후보에 대한 '점수'로 나타나게 되는 것이다. 예를 들어, 교육이 중요하지 않다고 생각하는 사람은 없겠지만 투표에서 그것을 얼마나 중요시 하느냐는 사람마다 다 다른 것이다.

데이터베이스의 구조는 '랭킹'이 가능하게 짜여 있어야

여기서 이런 예를 드는 것은 정치적 토론을 시작하려는 것이 아니라 이러한 의사결정 과정이 랭킹에 의해서 이루어진다는 것이고,

그러한 랭킹을 정하는 것은 많은 요소들과 그에 연관된 가중치의 합계인 총점이라는 점을 말하기 위해서다.

그리고 그것은 큰 그림으로 볼 때 바로 통계적 모델이 만들어지는 과정과 매우 흡사하다. 즉 많은 변수들과 그것들의 가중치의 합산의 결과가 모델 점수이며, 그것이 바로 많은 데이터를 질문에 대한 대답의 형식으로 줄여 나가는 과정이다.

필자가 늘 주장해 왔듯이 데이터는 추려져서 작아져야만 의사결정에 도움이 되는 것이고, 그 작아지는 과정의 중심에 통계적 모델이 있으며, 그래서 데이터베이스는 그러한 고등적 분석을 위해 최적화되어 있어야 하는 것이다. 그리고 그런 데이터베이스의 구조란 '랭킹'이 가능하게 짜여 있어야 한다.

데이터로 돈을 버는 방법을 얘기하는 것이 많은 사람들에게 유용하겠기에 다시 마케팅의 예를 들자면, 사업에 유익한 'Sales Lead', 즉 영업에 유용하게 사용할 수 있는 '고객이 될만한 회사들의 명단'을 엑셀 파일로 얻었다고 치자.

거기에 수천 명의 전화번호가 있다면 과연 누구에게 먼저 전화를 할 것인가? 그냥 랜덤으로 아무에게나 먼저 전화를 할 것인가, 아니면 가나다 순으로 할 것인가? 아니면 동네 이름을 봐서 '괜찮아 보이는' 곳에 있는 회사부터 접촉을 할 것인가? 아무 다른 정보가 없다고 해도 영업부 직원들은 어떻게든 그 명단을 효과적으로 정렬해보려고 시도할 것이다.

그런데 그 명단에 회사이름, 담당자 이름, 그리고 전화번호나 이메일 이외에 그 회사의 연도별 수익 총액, 사업연도, 직원 수, 산업별 구분 등이 포함되어 있다면 그러한 정렬화의 시도는 훨씬 수월해질 것이다. 일단 산업별로 대상이 아닌 곳은 추려내고, 단순하지만 수익 총액이 많거나 직원 숫자가 많은 곳부터 접촉하는 것도 한 방법일 것이다.

하지만 팔고 있는 상품이 오히려 대기업보다는 중소기업에 더 잘 먹히는 제품이라면? 세상에 간단하고 쉬운 일이란 없는 법이다. 그리고 배짱과 직관에 의존해서 성공하는 경우도 있지만 항상 그럴 수는 없는 것이다. 게다가 그 명단이 수천 명이 아니라 수십, 혹은 수백만의 상대를 다루는 것이라면 더더욱 의사결정자의 직관에만 의지할 수 없게 된다.

만약에 그 리스트가 현재 상대하고 있는 고객의 명단이라면 일은 더욱 복잡해진다. 즉 CRM (Customer Relationship Management)과 관련된 경우란 것인데, 그럴 경우 얻을 수 있는 데이터의 목록은 그야말로 눈이 돌아갈 정도로 많을 수도 있다.

엑셀 파일의 칼럼 숫자가 수백 개에 육박하는 것도 순식간이다. 그간의 거래기록, 구매 내 력, 마케팅과 영업을 통한 접촉의 시도와 그에 대한 반응의 기록 등을 액수, 날짜, 상품, 채널 별로 제대로 구분 관리하면 그 변수의 숫자는 눈덩이처럼 순식간에 불어난다.

그렇게 되면 아무리 머리가 좋은 사람이라도 그 리스트를 효과적으로 정렬하기가 어려워지며, 그래서 많은 사람들은 그 좋은 데이터를 두고도 한두 가지 변수만 반복적으로 사용하게

되는 것이다.

데이터는 잘 집적 시켜 줄여 나가야 하는 것

그런데 이런 상황에서 '접촉에 대해 호의적 반응을 할 확률'과 '고객의 잠재적 가치'에 대한 딱 두 가지 모델 점수가 그 리스트에 포함되어 있다면 얼마나 일이 수월 해질지 상상하기 어렵지 않다.

그 두 가지의 성향은 역함수 관계(inversely related)에 있을 수도 있겠지만(호의적으로 반응할 대상의 가치가 그리 높지 않은 경우가 많을 때), 이미 그 점수들은 수많은 다양한 변수들을 포함하고 있는 것이며, 사용자는 그저 목적에 따라 목록의 순서를 바꿔가며 사용할 수 있게 되고, 두 모델을 동시에 사용하여 '잠재적 가치도 높고 호의적인 고객'부터 우선적으로 접촉할 수도 있는 것이다.

대개 모델 점수들은 쓰기 편하게 1-10이나 1-20 단위로 나누어져 있으며, 수학 전공자가 아니더라도 누구나 모델 그룹 1번부터 사용하는 것이 유익하다는 것을 쉽게 터득하게 된다.

그래서 데이터는 잘 집적시켜 줄여나가야지 모아놓은 데이터만 많다고 자랑할게 아니라는 것이다. 의사결정에 도움이 되려면 데이터는 사람들이 질문을 하는 방식대로 추려져 있어야지 다양한 변수는 오히려 일반 사용자에게 혼란만 가져올 수 있다.

여기서 다시 모델 점수가 부여되지 않는 리스트로 돌아가보도록 하자. 많은 이들은 그간 통계 전문가의 도움 없이 오랫동안 이런 파일들을 다루어 온 경험을 토대로 가치가 높은 순서대로 명단을 재정렬하는 것이 그리 어려운 일이 아니라고 여길지도 모른다.

그런데 여기에다 미국식 표현으로 커브 볼을 던져 보자면, 그 주어진 명단에 한 사람의 이름이 여러 번 나타날 수도 있다는 것이다. 고객 명단을 가치 별로 정렬하라는데 중복된 라인이 곳곳에 있다면 그것을 어떻게 할 것인가?

당장 중복된 엔트리(entry)를 한 줄로 만드는 작업부터 하는 게 순서일 것이다. 그리고 그것은 거래 내역이 담긴 데이터베이스, 특히 관계형 DB를 상대할 때 반드시 거쳐야 할 작업이다.

관계형 DB나 거래 내역(transaction data)를 포함하고 있는 많은 데이터베이스들은 모든 내역을 효과적으로 저장하고 또 꺼내볼 수 있도록 최적화되어 있다. 우리가 온라인 쇼핑을 하면서 흔히 접하는 'Shopping Basket'을 들여다보면 그 관계를 쉽게 이해할 수 있다.

온라인 쇼핑몰의 관점에서 보자면, 고객이라는 각 개체가 고객의 이름, 번호, 주소, 이메일, 고객등급(Status) 등을 포함한 고객명단 테이블에서 한 줄을 차지하게 되고, 만약 그 고객이

여러 번 구매를 했다면 구매 때마다 Transaction Table에 거래 액수, 날짜 및 결재방법 등을 포함한 새로운 엔트리가 생길 것이며, 더 나아가 그 고객이 한번의 결재 시 많은 아이템을 구입했다면 Item Table에 상품명, 상품 카테고리, 가격 및 개수를 포함한 새로운 여러 개의 줄이 필요하게 된다.

즉 고객, 거래, 아이템은 데이터베이스 디자인에서 흔히 언급되는 1대 다중 관계(1-to-many relationship)에 있게 된다. 요즘 새로운 데이터베이스 디자인 테크닉의 발달로 그 'relation ship'이란 개념이 없이 개개의 엔트리가 독립적으로 입력되고 관리되는 경우도 많지만, 개념적으로 Shopping Basket을 보자면 그렇다는 말이고, 그 외 다른 종류의 고객 활동도 다 기록하자면 고객 하나 당 여러 줄이 필요하다는 것이다.

그리고 그런 기록을 바탕으로 고객에게 그 물품들을 보내야 하는 담당자의 입장에서 보면 그 중 하나도 버릴 게 없는 귀중한 데이터이다. 누구에게 어디로 어떤 물건들을 어떤 가격에 어떤 운송방법으로 보낼지를 다 기록하고 꺼내봐야 그 간단하다고 볼 수 있는 배송이 가능해지는 것이다. 데이터베이스를 디자인 하는 사람들은 그래서 이러한 완성형 구조를 '정상적 상태의 구조'(normal state)라 부르기도 한다.

영업 위해서는 데이터구조 자체를 고객 중심으로 요약/집적해야

하지만 고객을 중심으로 마케팅이나 영업을 담당하는 사람들의 입장에서 보면

그러한 관계형 구조는 그것이 거래(transaction)나 아이템(item)을 중심으로 이루어져 있기 때문에 구매자 중심(buyer-centric)의 관점에서 데이터를 이해하기가 어렵다는 단점이 있다.

물론 모든 데이터 라인을 다 검색하여 "지난 12개월 간 건당 20만원 이상의 거래가 있었던 고객을 모두 고를 것"이라는 식의 초보적인 필터링(filtering)은 가능할 수 있겠다. 그리고 그런 검색이 분석의 거의 전부라고 생각하는 사용자들도 많다.

하지만 약간만 질문이 복잡해져도 데이터베이스의 구조 자체를 '임시적으로라도' 바꿔야만 한다. 예를 들자면 "지난 12개월간 아웃도어 스포츠 카테고리 안에서 개인별 평균 거래액이 얼마였나?", 혹은 "온라인 채널을 통한 각 고객의 미래 가치는 원화로 얼마인가?" 라는 질문에 대답을 하자면 데이터베이스를 '고객중심'으로 집약해야 하는 것이다.

한 걸음 더 나아가 고객 전체의 명단을 미래 가치의 역순으로 정렬하려면 거래나 아이템 중심의 기록을 가지고는 시도도 하기 어렵다. 그것은 바로 먼저 예를 든 중복된 고객의 기록이 다수 포함되어 있는 리스트와 비슷한 경우이기 때문이다.

거래 기록 등을 ‘고객을 묘사하는 데이터’로 바꾸려면 그 구조 자체를 고객 중심으로 요약/집적하여야 한다. 개인별로 랭킹을 매기자면 개인별로 추려야 하는 것이고, 이메일을 정렬하자면 이메일 주소별로 요약이 되어야 한다. 그러한 단계에는 주소, 가구, 이메일, 회사, 모회사 등 여러 종류가 있을 수 있겠지만 요는 상대하려는 대상의 레벨로 모든 데이터가 재정리되어야 한다는 말이다.

예를 들자면 단순 거래별 액수의 나열이 아닌 “대상 별 지난 12개월 및 24개월의 거래 액수 총액”, 게다가 그런 개인 총액들이 상품이나 채널 별로 구분되어야 비로서 그 대상을 묘사하는 데이터로 탈바꿈하게 되는 것이다.

이러한 요약/집적을 데이터베이스 설계자들은 ‘De-normalization Process’, 즉 ‘Normal State에 반하는 비정상화 과정’이라고 부르기도 한다. 그것은 기존의 관계형 구조를 거스르는 작업이란 뜻인데, 다르게 표현하자면 이것은 같은 데이터를 90도로 돌려보는 과정이기도 하다.

그리고 그 결과는 깊이(number of records)는 줄어들고 너비(number of columns or variables)는 늘어나는 형태를 갖추게 된다. 어찌 보면 단순하다고도 할 수 있는 이러한 구조가 모델링과 그 결과인 점수를 이용한 랭킹에는 반드시 필요한 것이다.

이러한 ‘대상을 위주로 한’ 집적 과정의 첫 단계는 그 대상을 제대로 표현할 수 있는 아이디(ID) 체계를 제대로 갖추는 것이다. 놀랍게도 소위 말하는 예술의 경지(State of the Art)에 들어섰다는 데이터베이스도 개인이나 가구 별, 혹은 사업체 및 사업장 단위의 아이디가 아예 없는 경우가 많다.

그저 이름과 주소를 이용하거나 이메일 주소를 이용하는 경우도 많은데, 우리 모두가 여러 개의 이메일 주소를 동시에 사용하고 있는 마당에 그것을 ‘개인 아이디’라고 부르는 것은 어불성설이다. 게다가 거주지 주소를 사용한다 하여도 특정 인물이 이사를 갈 때마다 새로운 아이디를 부여할 수 있게 되니 아주 조심해야 한다.

대상 중심의 ‘묘사적 데이터’를 제대로 모으려면 여기저기 흩어져 있는 정보를 한 곳으로 제대로 모으는 것이 시작이며, 그러려면 그 개인이나 가구, 혹은 사업장 단위를 확실히 정해놓고 데이터를 집적하는 것이 순서다. 한 데이터베이스에 얼마나 많은 사람들이 기록되어 있는지도 정확하게 파악하지 못한다는 것은 그 개개인의 아이디 체계가 흐트러져 있다는 뜻이다.

즉 데이터베이스 안에 고객이 얼마나 되냐는 질문에 “한 이백만 명이 좀 안될 겁니다”라는 것은 대답이 될 수도 없다는 말이다. 고객이 제대로 파악이 되어있지 않으면 ‘고객 당 평균 구매액’, ‘고객 별 거래 간의 평균 날짜’ 등 아주 기본적인 측정 기준도 마련할 수 없게 된다.

분석전용의 데이터마트는 임시적으로라도 필요

데이터베이스의 구조 자체가 더욱 많은 다양한 데이터의 수집과 관리를 위해 진화 발전해왔지만 분석에 관한 기본 틀은 인간이 세상을 보고 이해하는 관점에 맞춰져 있기 때문에 쉽게 버릴 수 없다.

분석방법과 그 관련 소프트웨어의 발전도 분석에 필요한 인간의 노력과 시간을 줄여주는 것이지 분석의 틀까지 바꾸는 것은 아니다.

그렇기 때문에 '임시적'이라도 주 데이터베이스에 연결이 되어 있는 분석전용 데이터 마트가 필요한 것인데, 필자는 그것을 'Analytical Sandbox'(분석 샌드박스)라고 부르고자 한다.

즉 분석전문가들의 모래밭 놀이터라는 개념인데, 그것이 필요한 이유는 위에서 많은 예를 든 것과 같이 분석가들과 의사결정자들에게는 데이터베이스의 목적 자체가 효과적인 수집과 관리가 아닌 대상에 대한 분석과 의사결정을 위한 랭킹이기 때문이다.

그리고 정보의 저장과 처리속도, 그리고 제반 비용이 계속 내려가고 있는 시절에 그러한 Analytical Sandbox가 꼭 임시적일 필요도 없다.

그것이 제대로 운영되면 그 모래밭 안에서는 분석과 통계의 전문가들이 그들에게 유용한 SAS, SPSS, R 등 분석 전용 프로그램 언어로 소통이 편리하게 되며, 분석에 필요한 샘플링과 모델을 만드는 과정, 모델이 완성된 뒤에 반드시 따르는 적용과정(score application) 등이 다른 사용자들의 눈치를 볼 필요 없이(scoring은 컴퓨터의 연산 능력을 많이 요구하는 작업이기 때문이다), 또 새로운 데이터 처리과정 없이 진행될 수 있게 된다.

대부분의 오류는 모델이 시작되기 전, 그리고 완성된 모델을 적용시키는 과정에서 발생하며, 그래서 이러한 분석과 모델링 전용의 장소가 필요한 것이다.

분석과 통계전문가에게 이런 환경이 주어진다면 그들은 허구한날 오류로 가득 차있고 정리도 제대로 되어 있지 않으며 빈 곳 또한 많은 '남의' 데이터나 고치며 세월을 보내지 않고, 그들의 전문성을 발휘하여 타깃과 정의와 분석방법 등에 관하여 주로 시간과 노력을 쓰게 될 것이다.

게다가 그러한 모델링을 통해 데이터가 질문에 대한 대답의 형태를 갖추게 되면 다른 데이터베이스나 사용자들과 정보를 공유하는 것 또한 수월해지게 된다.

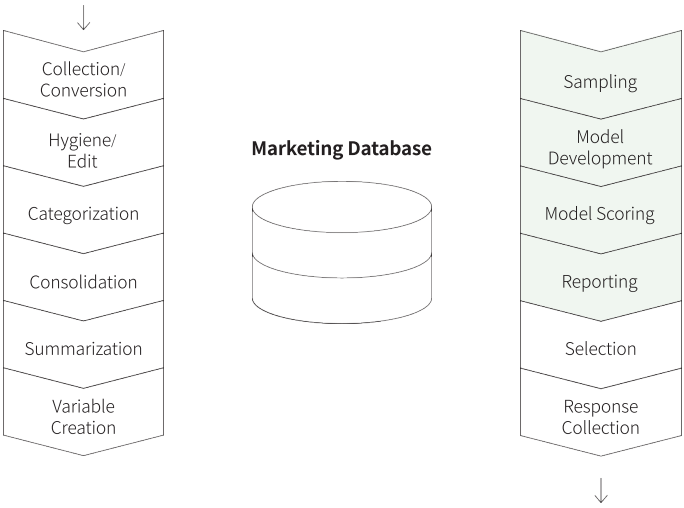
일단 공유해야 하는 변수의 숫자가 획기적으로 줄어들게 되기 때문인데, 산더미 같은 데이터를 늘 이리저리 옮기는 대신 많은 정보가 집약되어 있는 모델 점수만 사용자나 다른 기존의 데이터베이스로 되돌아가게 되는 것이며, 한 대상이 수백 가지의 성향이 모델 점수로 표현되어 있다고 해도 그 데이터의 크기는 요즘 기준으로 작은 데이터라고 보이게 된다.

단순 계산으로 수백만 명에게 수백 가지의 성향적 점수를 부여한다 하더라도 메가바이트 단위

의 데이터 밖에 창출되지 않는다.

첨부된 도표는 분석(Analytics)과 모델을 위주로 한 Analytical Sandbox, 즉 분석 전용 데이터 마트가 어떠한 과정을 아우르고 있어야 하는지를 보여준다.

Exhibit: Analytical Sandbox – "Analytics-Ready" Environment



이 도표에서 데이터베이스를 나타내는 원통의 오른쪽에 우리가 흔히 말하는 모델을 만드는 과정들이 나열되어 있는데, 모델을 단시간 내에 오류 없이 만들어 데이터베이스 안에 존재하는 모든 대상에 적용시키려면 원통의 왼쪽에 열거된 과정들도 반드시 필요하다.

즉 수집된 데이터(혹은 주 데이터베이스에서 연결, 배달되어 오는 데이터)는 정해진 룰에 따라 걸러지고 수정되고 보완되어야 하며, 많은 자유형태의 변수들은 카테고리 별로 구분되어야 하고, 숫자로 표현되는 데이터 또한 규격화되어야 한다.

더욱이 여러 종류의 데이터 또한 위에서 설명한 대로 대상을 중심으로 한 아이디 별로 집적되어야 분석에 쉽게 사용될 수 있다.

모델이 잘 짜여져야 의사결정자 원하는 대답 얻을 수 있어

그 결과로 '대상을 묘사하는 형태의' 많은 변수(variable)가 새로이 창출되는 것인데,

많은 면에서 그러한 변수의 창출(variable creation)은 통계적 방법이나 이론보다도 궁극적으로 마케팅과 영업의 결과에 훨씬 더 중대한 영향을 미치기 마련이다.

흔히 모델 경시대회에서 수상을 하는 모델들을 잘 살펴보면, 창조적인 변수들을 많이 만들어 사용하고 여러 통계적 방법을 혼합하여 사용한 팀들이 우수한 성적을 거두는 것을 볼 수 있다. 이론이나 방법론에 관련된 차이는 소수점 단위이지만, 제대로 정리되어 있지 않은 데이터는 프로젝트 자체를 파탄에 빠뜨릴 수도 있다.

즉 모델이 짜여지기 전과 후의 과정이 잘 되어 있어야 결과적으로 의사결정자들이 원하는 대답을 얻을 수 있다는 말이다.

이 전 과정이 제대로 된 공정을 통하여 관리가 되어야 모델이나 분석의 결과가 일관되고 빠르게 나온다는 것이고, 그런 공정을 만드는 것은 단지 통계 전문가나 분석가, 혹은 요즘 유행하는 타이틀인 데이터 사이언티스트들에게 미루어서 가능한 게 아니라 계획수립 단계에서부터 마케팅 중역들과 IT 관련부서들의 협조를 받아 제대로 투자, 관리되어야 할 독립적 프로젝트이어야 한다.

아무리 비싸고 유명한 분석도구를 구매하여 사용하고, 날고 긴다는 분석 전문가들로 팀을 꾸린다 해도, 전체적 환경이 제대로 되어 있지 않으면 좋은 결과가 나오지 않는다.

데이터를 분석에 최적화된 형태로 저절로 만들어주는 소프트웨어란 존재하지 않으며, 그건 애초에 분석 전문 도구들이 고전적 정의의 분석과 모델링(첨부된 도표의 오른쪽에 있는 연두색 부분)만을 효과적으로 할 수 있도록 디자인되어 있기 때문이다.

빅 데이터는 데이터를 '작게 만들어' 사람에게 대답을 주는 것

앞으로 이 도표에 있는 과정 하나 하나를 더 깊이 다루며 어떻게 단순한 숫자나 문자로 기록된 데이터를 모델링을 포함한

고등적 분석과정에 효과적으로 사용할 수 있도록 탈바꿈하게 만들며, 동시에 비 통계전문가들에게도 도움이 되게 하는 방법들을 세밀히 다룰 것이다.

그 과정이 바로 데이터를 점점 작게 만들어 사람들이 쉽게 이해할 수 있도록 하는 것이며, 그것이 데이터를 '인본화'하는 길이다.

빅 데이터는 데이터를 작게 만들어 사람들에게 대답을 주는 것이어야 하지, 크기와 다양성만 강조하는 것은 인간의 지식과 효율성에 대한 갈망을 간과하는 것이다. 의사결정과정은 여러 옵션을 점수를 통한 랭킹을 바탕으로 최선을 선택하는 것에 다름 아니며, 데이터는 의사결정 과정을 도와주는 형태로 존재해야 한다.

그러므로 많은 데이터베이스들이 수집과 저장만을 위해 최적화되어 있는 환경에서 임시적으로나마 그러한 랭킹이 수월하고 대상을 위주로 한 고등적 분석이 가능한 analytical sandbox가 요구되는 것이다. Sandbox가 너무 아이들 놀이터 같다는 느낌을 준다면 analytical haven, 즉 '분석의 안식처' 등 다른 이름으로 부르더라도 말이다.