빅데이터란 무엇인가?

글쓴이 : 김수보, 알앤비소프트웨어 빅데이터연구소 소장, 2013.10.10

최근 빅데이터가 화두입니다.

Google, Facebook 등 글로벌 서비스 업체들이 활용사례를 발표하면서 그동안 볼 수 없었던 Insight들이 공개되기 시작했고, 이것은 업계에 새로운 기대감을 갖게하기에 충분했습니다.

하지만, 빅데이터가 초기이다보니 용어가 생소한 사람들도 많을 뿐더러 일반 산업의 적용사례를 찾기도 쉽지 않습니다. 해외 사례를 참조해보아도 아직 우리 환경에 얼마나 맞을지 속단하기도 힘듭니다.

▶ (그림 1) 국내 주요 업체들의 플랫폼 구축사례 (2013년 상반기)

반면, 플랫폼 구축을 통해 빅데이터 성과를 얻어내려던 선행업체들이 만족할만한 성과를 얻어내지 못함으로써, 투자를 준비하던 업체들도 갈피를 잡지 못하고 있습니다.

그동안 대량 트래픽 처리환경과 대량 데이터 분석 및 처리에 참여해본 경험을 바탕으로 느낀 바를 몇가지 공유해보고자 정리해 보았습니다.

1. 분석 플랫폼으로서 빅데이터

데이터가 모이고 쌓이면, 당연히 분석하고자 하는 욕구가 생깁니다.

빅데이터의 기반기술인 Hadoop 조차 검색의 기반기술인 만큼 기술의 연관성이 매우 높습니다.

그래서, 최근에는 비정형 데이터(텍스트, 오디오, 비디오) 검색 및 분석을 위한 인프라로서 접근하고자 하는 시도가 많습니다.

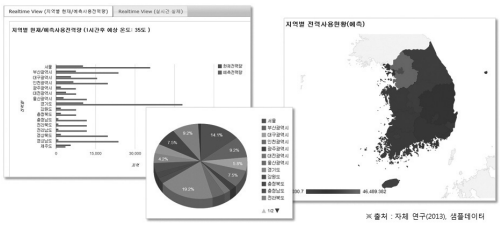
하지만, 분석의 본질은 기술이 아닙니다.

즉, 빅데이터 기반의 훌륭한 검색시스템을 구축했다고 해서 새로운 분석과 해석이 나오지 않는다는 뜻입니다.

분석의 본질은 호기심과 탐구능력이라고 보아야 합니다.

즉, 사람의 영역이며 분석기술은 통계적 기법과 도메인지식이 기본이 됩니다.

분석 플랫폼으로 빅데이터를 도입하기 위해서는 다음 사항들을 충분히 고려해보시기 바랍니다.



▶ (그림 2) 빅데이터 플랫폼을 활용한 전력사용량 분석 및 예측 연구 (샘플데이터)

1) 만드는 이가 왜 만드는지를 알아야 합니다.

데이터는 육감을 숫자로 바꾸어주는 역할을 합니다.

대부분, 경험자의 육감은 맞습니다.

하지만, 육감은 공유되기 힘듭니다. 공유되지 않으면, 다른 분야와 시너지가 나지 않습니다.

지난 기록으로 남지도 않습니다. 육감은 그 때 그 사람 만이 가지는 의견이기 때문입니다.

데이터는 디테일한 탐색을 가능하게 합니다.

우리가 궁금한 것들을 해소하려면, 오랜 시간 그 증거들을 잘 살펴보아야 합니다.

그러나, 육감은 살펴보거나 심층 분석하기 힘든 대상입니다.

그러나, 육감이 맞다는 것을 확인하기 위해 시스템을 만들지는 마십시요.

분석을 위한 모든 도구들을 다 갖출 수는 없습니다.

데이터 탐구활동이 어떤 가치를 만드는지 제작자들이 알아야 합니다.

그래야, 필요한 제대로 된 도구들을 갖출 수 있습니다.

2) RDB와 하드디스크로 시작하십시요.

먼저, 관심있는 서버의 데이터부터 모두 하나의 서버에 모아봅니다.

1 TB의 Disk 와 PostgreSQL이나 MySQL 정도면 충분합니다.

시간이 오래 걸리겠지만, 이것 저것 데이터를 가공하고 뽑아봅니다.

가공하고 뽑아보는 절차나 패턴 등을 엑셀에 꼼꼼히 적어둡니다.

중요한 건 결과보다는 과정입니다.

이 과정이 빅데이터의 필요성을 인지하는 과정일 뿐더러, 개발해야할 항목입니다.

3) 처음부터 완벽하려고 애쓰지 마십시요.

탐구적 분석은 헤맨다는 뜻과 동일합니다.

처음부터 완벽하게 구현하면, 업무변화에 대처할 수 없게 됩니다.

만들어진 모습이 원하던 최종 이미지도 아닐 수 있습니다.

완벽할수록 기민하게 변경할 수 없습니다.

업무가 정형화되고 패턴화될 때까지는 데이터 사이언티스트 1명과 개발자 1명을 붙여놓으십시요. 새로운 업무라면 가치를 알아낼 때까지는 마땅히 그래야 합니다.

4) 가설을 세우는 것이 핵심입니다.

데이터 분석업무는 '요구사항'을 받는 것이 핵심이 아닙니다.

예측’ 과 ‘분석’ 이란 가설을 세우고, 데이터를 통해 가설을 증명하는 과정입니다.

가설을 세우는 사람은 스스로 탐구하는 사람입니다.

스스로 탐구하는 사람이 없으면, 데이터 기반의 사후 분석은 성공할 수 없습니다.

요구사항을 내세요.’ 라고 말하는 사람을 초기에 투입하지 마십시요.

초기에는 요구사항을 내는 사람, 받는 사람, 구현하는 사람을 나누면 안됩니다.

아래의 과정은 나눌 수 없습니다. 처음에는 일을 가리지 않는 1~2명으로 시작하십시요.

- 가설을 수립합니다. (궁금증과 호기심을 가집니다.)
- 데이터를 탐색합니다. 또는 새로 수집합니다.
- 가설을 검증할 수 있게 가공, 변형처리 합니다.
- 가설을 검증해봅니다.
- 가설이 맞지 않으면 다시 시작해봅니다.
- 가설이 맞으면 현장에 바로 적용해봅니다.

5) 데이터웨어하우스와는 무엇이 다른가?

데이터웨어하우스DW도 데이터를 정제하고 분석하는 분야입니다.

니즈의 출발점은 동일합니다. 사용패턴도 동일합니다.

다만, 기술적으로는 아래와 같이 몇가지 차이점이 존재합니다.

첫째, DW는 비용 때문에 원시데이터를 보관하지 않습니다. 그러나, 빅데이터는 사후분석을 위해 원시데이터를 남겨 둡니다. 그래서 값싼 스토리지를 이용합니다.

둘째, DW는 해당시점의 스냅샷을 저장합니다. 따라서, 정보를 Historical 하게 쌓지 않아서 지난 통계를 재집계하기 힘듭니다. 그래서, DW는 대부분 변환룰을 바꾼 이후에야 분석이 가능합니다. 반면, 빅데이터는 데이터를 Historical 하게 저장합니다. 따라서, 소급시점을 기준으로 여러가지 재처리나 시뮬레이션 분석이 가능합니다.

비싼 스토리지에 선별된 데이터가 많다고 해서, 빅데이터라고 부르기 힘들 것 같습니다.

그것은 그냥 큰 데이터웨어하우스 입니다.

Hadoop은 x86기반의 저렴한 하드웨어를 활용할 수 있게 했습니다.

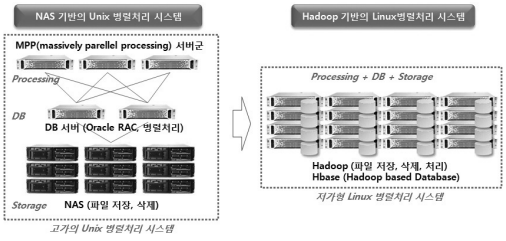
비용부담이 적어 값싼 로그성 데이터 적재도 겁나지 않게 되었습니다.

빅데이터는 다룰 수 있는 데이터 양이나 종류를 크게 확장시킴으로써 분석플랫폼의 가용성을 확장시켰다고 볼 수 있습니다.

2. 병렬처리 플랫폼의 빅데이터

빅데이터 기술의 시작점은 Hadoop 입니다.

Hadoop은 2005년 더그커팅에 의해 Nutch라는 검색엔진을 지원하기 위해 만들어진 대용량 분산스토리지 시스템이었습니다. Apache 재단으로 넘어가 공개소프트웨어가 되면서 급격하게 각광받기 시작했습니다.



▶ (그림 3) 기존 병렬처리 시스템과의 차이점

Hadoop 이전에는 큰 데이터를 처리하기 위한 병렬처리 시스템이 비쌌습니다.

그리고, 병렬처리를 위한 별도의 프로그램 개발이 필요했습니다. 벤더마다 달랐고, OS마다 달랐습니다.

Hadoop은 이 부분의 비용을 혁신적으로 낮추어주었습니다.

Map Reduce 작업이 아직 개발 부담이 있지만, 과거에 비해서 개발 편의성이 혁신적으로 낮아졌다고 볼 수 있습니다. 최근에는 SQL on Hadoop 을 통해 RDB 개발 수준까지 병렬프로그램의 개발비용이 낮아지려 하고 있습니다.

병렬처리가 쉬워지면 어떤 효과들이 나타날까요?

1) 전수데이터 처리가 쉬워지게 되었습니다.

병렬처리를 이용하면 이론적으로 무한한 데이터를 처리할 수 있습니다.

Hadoop 이전에는 수집.적재는 병렬처리를 했으나, 통계처리 및 다차원 분석 등은 OLAP 엔진을 이용했습니다. 하지만, OLAP은 구조적으로 CPU Cost 가 높은 독립형 작업입니다.

따라서, 데이터를 수집단계에서 실시간으로 처리하고 버릴 수 밖에 없었습니다.

그러나, Hadoop은 적재 이후의 데이터 병렬처리를 손쉽게 만들었습니다.

대용량 적재에 대한 부담이 적어지면서, 전수데이터 처리의 문턱이 많이 낮아지게 되었습니다.

인구조사, 범죄자 정보, 통화이력 정보 등은 모든 데이터가 다 있을 때만 의미를 가지는 대표적인 데이터라고 할 수 있습니다.

2) 개인화된 실시간 추천 서비스가 가능하게 되었습니다.

이전에는 나이, 지역, 성별 등 사전 정의된 룰을 기반으로 추천 서비스를 제공했습니다.

구글 애드센스는 사용자가 클릭하거나 검색한 걸 기반으로 유사제품을 추천합니다.

이는 사용자가 광고로 유입될 확률을 크게 높여주었습니다.

기존에는 사용자별 검색패턴을 실시간으로 처리할 수 없었습니다.
물리적 공간도 부족했을 뿐더러, 처리하기 위한 소프트웨어 알고리즘도 복잡했기 때문입니다. 이제는 Storm, Esper 등 오픈소스를 이용하여 대용량 실시간 분석 시스템을 구축할 수 있게 되었습니다.

이것은 글로벌한 대형 서비스들도 더욱 인간적일 수 있게 되었다는 뜻입니다.
보다 즉흥적이고, 보다 복잡한 지능적 패턴을 멀티로 처리할 수 있게 되었습니다.

개인화된 실시간 추천 서비스' 는 '실시간 대량 데이터 병렬처리 능력' 을 의미합니다.
이 능력은, 다른 분야에서도 관심이 높고 응용분야가 넓어서 앞으로 발전가능성이 높다고 볼 수 있습니다.

병렬처리 플랫폼으로서 가장 효과적인 분야가 제조분야일 것 같습니다.
공장은 투입 대비 결과를 측정하는 수율이 매우 중요합니다.
그리고, 이미 수율분석을 위해 계속 로봇장비의 데이터를 축적시키고 있습니다.
하지만, 데이터 병목현상으로 인해 대부분 샘플링 분석기법을 쓰고 있습니다.

하지만, 센서 오류란 것이 특정 오류는 소량발생이라도 민감한 경우가 있어 병렬처리를 통해 전수데이터 분석이 가능해진다면, 보다 넓은 오류 대응이 가능할 것으로 예상됩니다.

3. 빅데이터는 기술이 아니라 '데이터 비즈니스' 입니다.
빅데이터는 데이터를 다루는 기술입니다.포털과 검색에서 시작했죠.
아직 다른 분야에 응용하기에는 부족한 기술 측면이 많습니다.

Map Reduce 작업이 이유없이 갑자기 다운되기도 합니다.
여러개의 Map Reduce 작업을 동시에 수행시키기 어렵습니다.
작업요청의 진입채널인 Name Node는 아직 이중화되지 않습니다.

시간이 흐르면 이런 기술 한계들은 자연스레 극복될 것입니다.

그러나, 이런 기술한계들이 빅데이터 도입을 막거나, 사업적 실패를 초래하거나 하지는 않습니다.

빅데이터를 이끄는 것은 비즈니스 현장의 Insight 입니다.
기술을 이해하고 어떻게 현장에 적용해볼까를 고민하지 않는 이상, 훌륭한 사업가치는 만들어지지 않습니다.

빅데이터는 외주를 통해 성과를 얻기 힘든 분야입니다.
자사의 핵심 정보와 Insight는 당사자가 아니면 가질 수 없기 때문입니다.

컨설팅이나 플랫폼 제공 업체는 기술적 솔루션을 제공할 뿐입니다.

사업주체가 스스로 연구하고 고민하는 것.
그것이 빅데이터 도입의 시작점이라고 할 수 있습니다.

사업주체가 스스로 행동하는 것
그것이 빅데이터가 일으키는 변화의 시작점이라고 할 수 있습니다.

데이터를 통해 할 수 있는 일의 가능성에 관심을 가지십시오.

저/자/소/개

김 수 보

- 빅데이터 SW 품질평가모델 사업 전문위원(TTA)
- 알앤비소프트 빅데이터 연구소 소장
- kth 개발실 수석팀장
- 지어소프트 플랫폼 사업팀장
- 삼성SDS