데이터 사이언스 기반의 디지털 트랜스포메이션

□ 이진수 / SK Telecom

언제부턴가 빅데이터라는 키워드가 모든 기업의 관심을 받으며 등장하더니, 최근에는 데이터 사이언스, 인공지능, 디지털 트랜스포메이션, 4차 산업혁명 등과 같은 다양한 키워드가 거의 유행처럼 모든 산업에 걸쳐 뜨거운 관심을 받고 있다. 이러한 높은 관심을 반영하듯, 데이터 사이언스 혹은 빅데이터에 대한 다양하고 풍부한 Article과 정보들이 넘쳐나고 있고, 데이터 사이언티스트를 양성하기 위한 각종 교육과 기술 교류를 위한 컨퍼런스 등의 활동도 매년 증가하고 있다. 하지만 이러한 폭발적인 인기에도 불구하고, 많은 기업들은 데이터 사이언스 기반의 혁신을 추구하는 과정에서 현실적인 어려움에 직면하게 된다. 데이터 사이언티스트들을 영입하고, 많은 비용을 데이터 사이언스 활동을 위한 IT 환경 구축에 투자함에도 불구하고 레퍼런스로 등장하는 성공 사례와는 달리, 현실 속에서는 단순히 IT 환경 개선 수준에 머무르는 등 대부분 성공적인 혁신으로 연결하지 못하는 것은 무엇 때문일까? 본 고에서는 데이터 사이언스 기반의 디지털 트랜스포메이션을 추구하는 과정에서 만나게 되는 실제 문제점들은 어떠한 것들이 있으며 이를 극복하기 위해 필요한 부분은 무엇인지, 현실적인 실행 관점에 좀 더 집중하여 살펴 보고자 한다.

Ⅰ. 데이터 사이언스와 디지털 트랜스포메이션

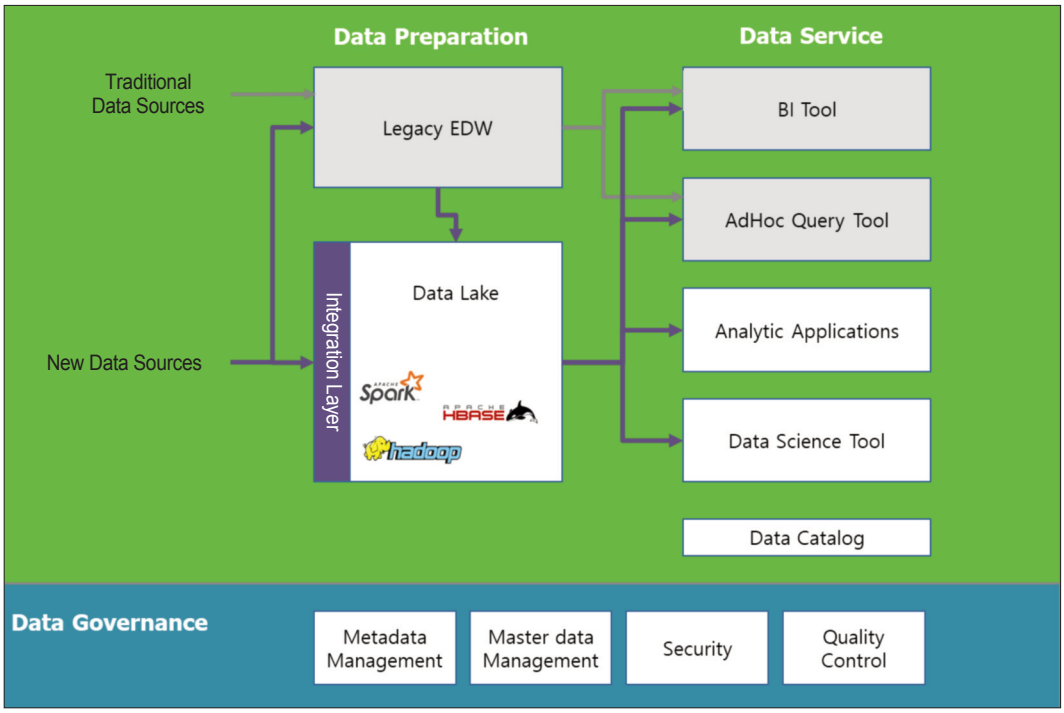
흔히 디지털 트랜스포메이션이라고 하면, 아마존과 같이 기존 비즈니스 모델의 한계를 벗어나 새로운 비즈니스 모델로의 전환 혹은 확장한 경우와, Airbnb와 같이 플랫폼 기반의 새로운 비즈니스 모델을 창출한 경우, 그리고 스타벅스, 나이키와 같이 기존 비즈니스 모델을 획기적으로 혁신한 사례를

생각하게 된다. 이들 모두의 공통점은 바로 "디지털 기술과 역량을 도입"하여 이를 기반으로 고객과 시장, 경영환경의 "파괴적인 변화에 적응하거나 선제적으로 대응"함으로써, "비즈니스의 경쟁력을 혁신적인 방향으로 변화"시켰다는 데 있다. 여기서 디지털 기술과 역량의 도입이란, 사례마다 다소 차이는 있겠지만 대부분은 필요한 데이터를 생산하고 이를 활용할 수 있는 디지털 환경과 역량의 확보를 의미하고 있다. 즉, 다양하고 풍부한 데이터가 생산되는 환경적인 변화와 이를 수용하고 처리할 수 있는 기술의 발전이 빅데이터, 혹은 데이터 사이언스라는 키워드를 만들어냈다면, 이러한 환경에 일찍부터 적응한 혁신 기업들의 사례가 증가하면서, 이들을 디지털 마스터라 일컬으며 디지털 트랜스포메이션이라는 키워드가 자연스럽게 연결될 수 있었다. 때문에 많은 기업들이 이러한 레퍼런스 사례를 참고하여 디지털 트랜스포메이션을 추구할 때, 우선은 데이터 사이언스 기반의 경영 혁신을 위해 먼저 기존 비즈니스 과정을 디지털화함으로써 데이터를 생성, 수집, 통제할 수 있도록 하고, 이를 적극적으로 비즈니스 혁신에 활용함으로써 새로이 도약하는 기업으로 Re-positioning을 시도하게 된다.

II. 데이터 레이크 기반의 데이터 수집/저장/활용

일반적으로 기업에서 데이터를 활용하는 초기 단계에는, 각 서비스별로 독립된 데이터 생성/관리 환경을 구축하고 이를 각 서비스 개별 목적으로 활용하게 된다. 하지만 해당 기업이 보유한 모든 데이터를 통합적으로 자산화하고 이를 활용하여 데이터 기반의 혁신을 이루고자 할 경우에는, 상이한 서비스 별로 관리되어온 데이터를 통합해서 볼 수 있도록 데이터 설계에 대한 수정이 필요할 수 있고, 흩어진 데이터와 새롭게 추가될 데이터를 물리적, 혹은 논리적으로 통합 저장/분석할 수 있는 환경 구축이 필요하다. 즉 기존에 관리되어온 Legacy 데이터베이스, 새롭게 생성되는 서비스 데이터, 고객센터의 VOC 데이터 등 다양한 데이터 소스로부터, 분석할 가치를 갖는 데이터를 한 곳에 모아 저장 관리하고, 통합적으로 분석 활용하기 위한 데이터 레이크 환경을 구축할 수 있다. 데이터 레이크를 구축하게 되면, 기존의 개별 서비스는 영향 받지 않으면서 별도의 통합 분석 환경을 확보할 수 있는 장점이 있다. 데이터 레이크는 기존 관계형 데이터베이스로부터 수집되는 데이터뿐 아니라 고객의 VOC 콘텐츠 등과 같은 비정형 데이터까지 다양하고 방대한 데이터를 저장, 처리하고 분석할 수 있어야 하므로, 일반적으로 이러한 저장 처리에 적합하고 확장성이 좋은 Hadoop[1] 오픈소스 기반으로 구성되며, 높은 저장 효율과 빠른 가공 처리 등을 위해 HBASE[2], SPARK[3] 등 다양한 오픈소스 기술들이 함께 적용된다.

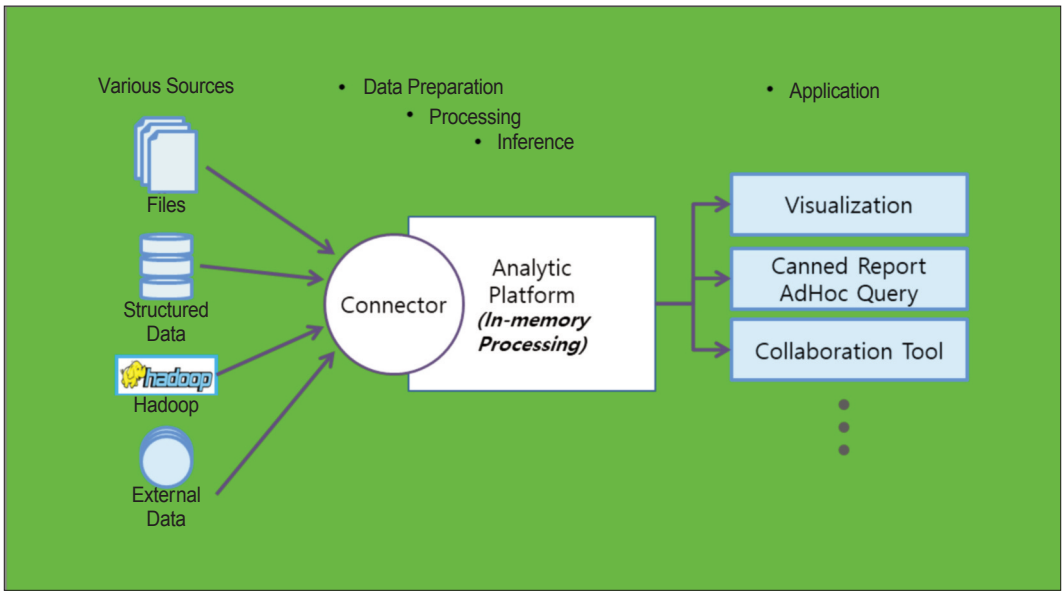
데이터 레이크는 기본적으로 기업의 데이터 리파지토리 역할을 하지만 분석에 필요한 Summary, Machine Learning 등의 데이터 처리도 수행한다. 이를 통해 새롭게 만들어진 통계, 인사이트 정보를 포함한 가공된 데이터는 정통적인 Business Intelligence Tool이나 OLAP, Canned Report 형태뿐 아니라, 웹 형태의 인사이트 제공 서비스, Machine Learning 기법을 통한 분석이 가능한 Tool(R Studio 혹은 Tensorflow[4] 등)과 연결되어, 각 비즈니스에서 데이터를 활용해야 하는 사람들이 직접 분석할 수 있는 Self-Discovery 환경을



〈그림 1〉 데이터 레이크 기반 데이터 분석 시스템 구조

제공한다. 데이터 레이크 기반의 Self-Discovery 환경에서는 자신이 담당하지 않는 다른 서비스의 데이터까지 통합적으로 분석해야 하므로, 여러 비즈니스 도메인 데이터를 모든 사람이 쉽게 이해하고 접근할 수 있어야 한다. 하지만 테이블 명세서나 ERD와 같은 정통적인 메타 데이터로는 다양한 서비스로부터 생성된 수많은 테이블을 이해하기 힘들 뿐 아니라, 많은 기업에서 주로 IT 운영 목적으로 관리되는 테크니컬 메타 데이터 정도만 관리할 뿐 비즈니스 메타 데이터를 잘 관리하지 않기 때문에, 이러한 데이터의 이해를 위한 메타데이터 환경을 마련하고 콘텐츠를 관리하는 것은 더욱 중요해진다. 과거 논리적 메타 데이터나 물리적 메타데이터 외에, 비즈니스와 연계한 상위레벨의 데이터 설명에서부터 시작해 논리적, 물리적 메타데이터로 이어지는 메타데이터 콘텐츠 서비스가 데이터 카탈로그로서 제공되어야 한다. 많은 기업들이 기존에는 이러한 콘텐츠를 구축하거나 최신 내용으로 유지 관리하는데 관심이 적어왔기 때문에, IT 환경 구축 대비 콘텐츠를 생성하고 현행화하기 위해 필요한 비용을 투자하는 데는 익숙하지 않으므로 이 부분을 소홀히 하지 않도록 주의할 필요가 있다. [그림 1]

이와 같이 큰 규모의 기업에서는 다양한 형태와 기술을 접목한 데이터 분석을 신속하게 진행하기 위해 데이터 레이크에 데이터를 물리적으로 수집/통합 저장해 놓는 방식을 일반적으로 사용하는데, 만일 데이터 소스가 지속적으로 변하고 정기적인 분석 모델보다는 Ad-Hoc 분석의 빈도가 많은 경우에는 데이터 레이크에 모든 소스 데이터를 수집 관리하는 것 자체가 Cost가 많이 필요하기 때문에



〈그림 2〉 논리적 데이터 연결 기반 데이터 분석 시스템 구조

분석 설계를 진행하는 시점에 각 데이터 소스로부터 Connector에 의해 필요한 데이터를 연결한 후 바로 필요한 데이터를 가져와 분석 모델을 개발하고 Visualization 및 Publish까지 수행하는 논리적 데이터 연결 구조를 지원하는 솔루션들도 최근에 사용되고 있다. [그림 2]

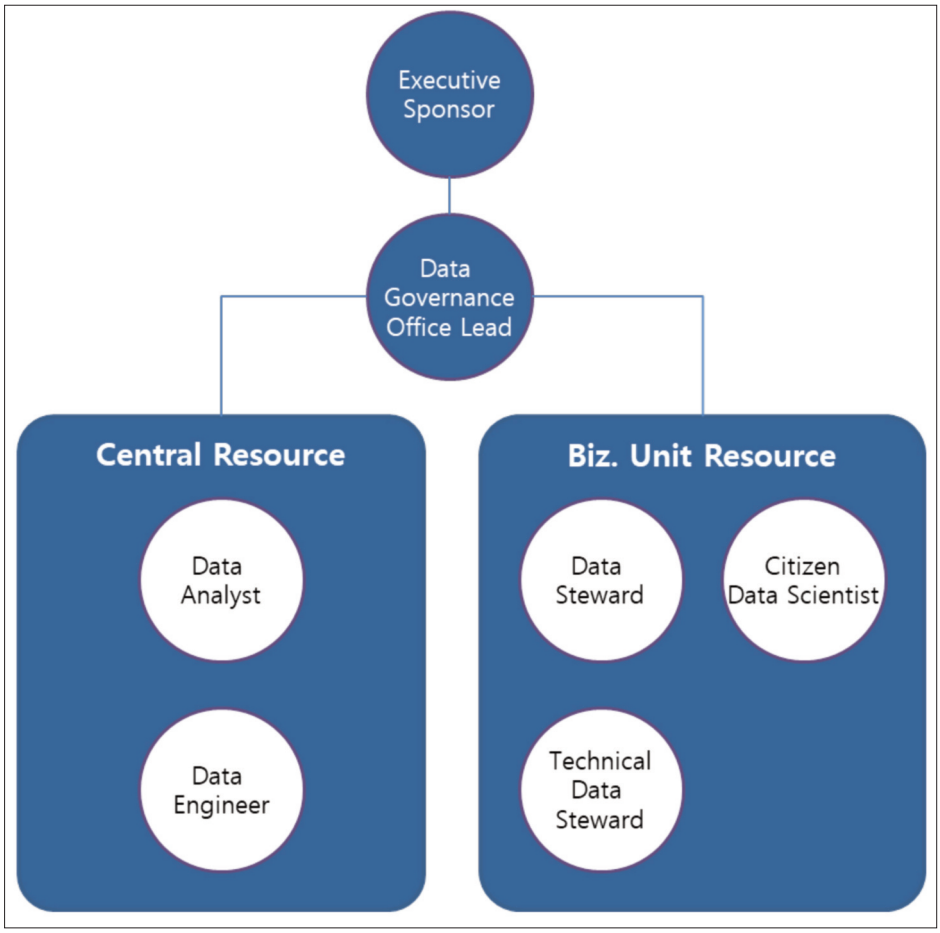
앞서 설명한 두 가지 방식 모두, 데이터의 품질 관리와, 데이터 접근 권한 관리, 그리고 데이터를 잘 이해하기 위한 메타 데이터나 마스터 데이터 관리 등이 체계적으로 관리되어야만 하는데, 바로 이러한 데이터의 통제 관리가 데이터 거버넌스 영역이다. 데이터 거버넌스는 단순히 메타데이터 관리 시스템과 같은 IT환경 구축의 문제가 아니라 프로세스, 원칙, 조직과 사람 등 비즈니스를 유지하는 전체 요소에 대한 변화 관리를 요구한다. 예를 들어 새로운 비즈니스나 서비스가 기획되는 시점부터, 해당 비즈니스/서비스가 기업에 어떠한 가치의 데이터를 공급하고 활용함으로써 비즈니스 가치를 높일 수 있을 지가 반영되어야 하며, 기업에서 정한 표준에 따라 데이터를 생성함으로써 품질관리가 용이하고 통합적 분석이 가능하도록 해야 한다. 또한 표준 프로세스에 의해 이러한 필수 요건이 기업 차원에서 강제화되고 이를 감수하는 역할이 존재하며, 각 비즈니스는 데이터를 기반으로 의사 결정하는 비즈니스 체질을 키워나가야 한다. 이러한 변화 관리를 좀 더 체계적으로 수행할 수 있도록 도와주는 환경적 요소가 데이터 거버넌스에 필요한 IT 환경이며, 기업 내 다양한 조직과 사람들이 같은 기준과 컨센서스를 가지고 변화를 추진하기 위한 원칙과 조직 체계 등이 마련되어야 한다.

이러한 변화 관리를 위해서는 기업 내 특정 전문가 집단이나 한 조직이 리드하기에는 한계가 있다. 이러한 시행착오를 거쳐온 기업들과 데이터 거버넌

스의 리더들은 그 동안의 경험을 기반으로 다음과 같이 성공적인 데이터 거버넌스 확보를 위해 필요한 기업 내 조직 및 개인의 역할을 정의/분류하였다.[5] [그림 3]

[Data Governance 운영 조직]

- **Data Governance Steering Committee**: 전략적인 비전을 제시하고, 실질적으로 예산 및 전사의 리소스 할당에 대한 결정권을 가지고 있다.
- **Data Governance Office**: Data Governance를 위한 변화 관리 업무를 전담하는 인력으로 구성되며, Data Governance Steering Committee와 Data Governance Working Group간의 연결고리 역할과, 사내 표준, 정책, 프로세스를 정하고 교육 및 커뮤니케이션 Planning과 감사 업무 등을 담당한다.
- **Data Governance Working Group**: 각자의 원래 고유 Role을 수행하면서, Data Governance



〈그림 3〉 데이터 거버넌스 역할

역할을 병행하도록 업무를 부여 받은 인력으로 구성된다. 주로 각 비즈니스 영역 담당자와 IT 영역 담당자가 함께 구성되며, 실질적인 데이터 역량의 전파와 사업 조직별 실행을 담당하게 된다.

[Data Governance 운영을 위한 역할]

- **Executive Sponsor**: 전사 차원의 변화관리가 체계적이고 일관적으로 이루어지려면, 강력한 스폰서십은 매우 중요하다. 또한 이러한 스폰서십은 상황에 따라 다르지만 일정기간 지속성이 보장되어야 일관되게 추진될 수 있다. 비즈니스 영역과 IT영역에 대해 고른 이해와 추진이 가능한 Position이 Executive Sponsor를 맡는 것이 가장 이상적인데, 실질적으로 CEO가 그 예가 될 수 있으며, 만일 비즈니스 영역과 IT 영역 중 하나를 선택해야만 한다면 비즈니스 영역에서 Executive Sponsor를 맡는 것이 Data Governance 변화관리를 성공적으로 지속시킬 가능성이 높다.
- **Data Governance Office Lead (DGO)**: 기업 내 데이터 거버넌스 활동에 있어 실질적인 의사 결정권자이다.
- **Data Steward**: Data Governance를 도입하는 시점에는, 개별 비즈니스 조직의 적극적인 참여와 이해가 매우 중요한데, 각 비즈니스를 대표하여 이러한 역할을 직접 수행, 리드하기 위해 해당 비즈니스 조직에서 선발된 역할이다. 이들은 전사에 데이터 기반의 변화 관리 문화를 전파하고 실행하는 데 있어 실질적이고 핵심적인 역할을 수행한다. 비즈니스 관점에서 어떤 데이터가 필요하고 생성되는 지 이해하고, 데이터 기준 정보, 품질, 보안 Life cycle 등을 관리 감독한다. 또한 해당 비즈니스 조직과 전사 차원의 Data Governance 조직, 데이터 IT 조직과의 연결 역할을 담당한다.
- **Technical Data Steward**: 각 비즈니스의 IT를 담당한 부서에서, 해당 비즈니스 데이터의 Technical 품질 관리와 시스템 활용 측면 전반을 관리한다. Data Steward가 기술적인 측면까지 이해하고 접근하는 데는 한계가 있기 때문에 Technical Data Steward와 긴밀하게 협업을 수행한다.
- **Data Scientist – Data Analyst**: 비즈니스 조직 내에 이론적 Analysis 역량이 부족할 수 있기 때문에, 이를 Support하기 위한 통계/머신러닝 전문가 집단이다. 최근에는 사람마다 Data Scientist를 조금씩 다르게 정의하는 것을 볼 수 있는데, Data Analyst를 Data Scientist라고 칭하고, Data Engineer와 구분짓기도 하지만, 보통은 Data Scientist란 Data Analyst와 Data Engineer를 모두 칭하거나, 이 둘 모두를 아우를 수 있는 고급 전문가를 칭한다.
- **Data Scientist – Data Engineer**: 전통적인 관계형 데이터베이스뿐 아니라, Hadoop, SPARK 등 Data 와 관련한 오픈 소스 기술 등을 활용하여 데이터를 수집/저장/처리/공급하는 IT 전문가 집단이다. 기업이 필요로 하는 전사 차원의 데이터 IT환경의 아키텍쳐를 설계하고 구축, 운영하는 역할을 맡는다.
- **Citizen Data Scientist**: 각 비즈니스 조직에서 Domain에 대한 깊은 이해와 기본적인 이론적 분석역량을 모두 가지고 있어, 해당 비즈니스의 데이터 기반 업무 혁신을 리드하는 분석가이다. 전문가 집단인 Data Analyst만으로는

각 비즈니스 특징을 깊게 이해하고 데이터를 접목시키는 데 한계가 있기 때문에, 가장 이상적인 데이터 활용 구조는 각 비즈니스 전문가가 이론적 분석 역량을 확보하여 데이터를 직접 활용하는 것이라는 기본 생각에서 최근 주목받고 있는 역할이다.

Ⅲ. 데이터 기반의 성공적인 디지털 트랜스포메이션을 위한 요건

많은 기업들은 데이터 기반의 디지털 트랜스포메이션을 시도하기 위해 우선 외부에서 데이터 사이언티스트를 영입하거나 내부 직원의 전문가 교육을 통한 역량 확보를 시행하고, 이후에 데이터 수집/저장/분석 관리가 가능한 IT환경을 구축하거나 강화하는 활동을 한다. 하지만 이러한 활동들이 바로 혁신으로 이어지지 못하고 단순히 기존 IT환경의 개선 수준에 머무르는 경우를 자주 볼 수 있다. 이는 많은 기업이 기존의 대규모 프로젝트를 진행하는 방식과 유사하게 접근하기 때문인데, 데이터 기반의 디지털 트랜스포메이션은 한 기업이 오랜동안 익숙해진 문화의 트랜스포메이션, 인적 자원의 트랜스포메이션, IT의 트랜스포메이션 전반에 걸친 변화 관리 관점으로 접근해야만 성공할 수 있다.

그렇다면 어떠한 관점의 변화 관리를 통해 데이터 기반의 비즈니스 혁신을 가져올 수 있을까? 처음 데이터 사이언스 체계를 도입하게 되는 단계에서 흔히 만나는 어려움은, 기존 전통적인 프로세스와 조직, 사람들과의 충돌이다. 기존에 이미 검증되고 숙련된 프로세스와 IT환경으로부터, 익숙하지 않은 새로운 체계를 받아들이는 것은 매우 어렵다. 각 비즈니스 조직 입장에서는 새로운 체계 변환 과정에서 나타날 수 있는 리스크의 감수가 부담스럽고, 익숙하지 않은 새로운 프로세스와 역량의 수용도 부담스러울 수 있다. 이를 극복하려면 데이터 기반의 변화 목적과 예상 결과물이 각 비즈니스 조직에도 리스크를 감수할만큼 유용한 것임에 대한 Top-Down 컨센서스가 있어야 하며, 지속적인 협업과 컨센서스를 유지하려면, 앞서 설명된 것과 같은 강력한 스폰서십이 필요하다.

컨센서스가 마련된다 하더라도, 기존 비즈니스를 유지하고 있는 IT환경을 버리고 새로운 데이터 사이언스에 적합한 IT환경으로 전환하는 것은 매우 리스크가 크다. 앞서 [그림 1]에서 설명된 것과 같이 기존 IT환경을 그대로 유지하면서, 새로운 데이터 사이언스 IT 환경에서 이를 연계/수용하도록 하는 기술적, 전략적 접근이 필요할 수 있으며, 어느 정도 IT 환경의 트랜스포메이션이 진행되고 난 이후에 점진적으로 불필요한 기존 IT 환경을 Fade Out 하는 전략이 바람직하다.

비즈니스 프로세스 혁신 과정에서도 어려움이 존재한다. 외부에서 영입된 데이터 사이언스 전문가는 해당 비즈니스 도메인에 대한 이해도가 낮고, 기존 비즈니스 전문가는 데이터 사이언스에 대한 이론적 지식이 부족하기 쉽다. 이들간에 깊은 공감대가 형성되지 않을 경우, 데이터 사이언스 전문가는 초반 이론적 근거를 토대로 강한 추진을 진행하다가 결국 비즈니스 전문가들의 지속적인 반발로 더 이상 과정이 진행되지 않을 수도 있다. 서로의 이해와 배려를 기반으로 트랜스포메이션 초기에는 데이터 사이언스 전문가의 보다 적극적인 개입과 지원으로 비즈니스 전문가와 협업하며 강력한 변화를

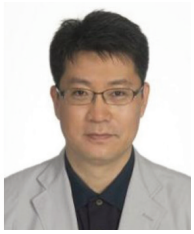
추진하다가, 일정 수준 도달 이후로 갈수록 데이터 사이언스 전문가의 역할은 줄이고, 궁극적으로는 비즈니스 전문가가 데이터 사이언스 역량을 키워 스스로 판단하고 변화를 추진할 수 있도록 유도하는 것이 필요하다. 특히 최근에는 이와같이 각 비즈니스 전문가 중 데이터 사이언스 전문 역량을 키워 스스로 데이터 사이언티스트 역할을 수행하는 것이 중요하여 Citizen Data Scientist라는 역할이 새롭게 강조되고 있다.

기업이 생존함에 있어, 데이터 사이언스를 기반으로 디지털 트랜스포메이션을 통한 기업의 혁신은 더 이상 선택이 아니라 필수가 되었다. 이러한 새로운 혁신 트랜드는, 과거의 전통적인 개선 활동처럼 IT 환경의 개선이나 특정 소수 전문가 집단에 의존하는 접근방법으로는 근본적으로 한계가 있으며, Top Manager에서부터 각 비즈니스 담당자와 기존 IT 담당자까지, 일관된 목표 아래 각자의 Position에서의 역할과 역량을 요구받는다. 따라서 기업이 추구하고자 하는 방향성과 목적이 명확해야 하며, 규모가 큰 기업일수록 이러한 변화 관리는 보다 체계화된 프로세스와 시스템을 요구한다. 성공적인 디지털 트랜스포메이션은 한 사이클로 완성되는 프로젝트 성격이 아니라, 지속적으로 발전하고 있는 데이터 사이언스 기술의 도입과 비즈니스의 접목이 반복적으로 이루어지는 변화 관리 과정을 통해 이루어질 것이다.

참 고 문 헌

[1] https://hadoop.apache.org/
[2] https://spark.apache.org/
[3] https://hbase.apache.org
[4] https://www.tensorflow.org/
[5] “The Best Place to Start is HERE: Getting Started Governing Data” Kelle O’Neal, DGIQ (Data Governance & Information Quality Conference) 2016

필 자 소 개



이 진 수

– 현재 : SK Telecom, Data서비스 팀장
– POSTECH 석사 졸업 Neural Network 전공
– (전) 삼성전자 VD사업부 빅데이터 담당 파트장
(전) NHN 솔루션개발실장 및 검색모델링 담당
(전) LG전자 MPEG표준화, 멀티미디어/AI 담당
– 주관심분야 : Data Science, Data Engineering, Deep Learning, Machine Learning, Data Governance