빅데이터의 스마트한 활용

– 데이터 사이언스와 데이터 가시화

김 상 수 부장 / ㈜한컴MDS IoT사업부
sangsu@hancommds.com

데이터의 홍수 속에서

우리는 데이터의 홍수 속에서 살고 있다. 웹사이트는 모든 사용자의 클릭을 추적하고 있고, 스마트폰은 매초 자신의 이동 경로와 속도를 기록한다. 웨어러블 센서들은 사람의 심박수, 운동 습관, 수면 패턴을 기록하며, 스마트 자동차는 운전자의 운전 습관을 기록하고, 각국의 정부는 사회적으로 쓸만한 통계 정보를 주기적으로 생산해 내고 있다. 특히, 인터넷은 그 자체로 거대한 데이터의 묶음으로 연결된 모든 것이 상호 참조될 수 있는 데이터베이스이며, 백과사전이자 데이터의 재생산자이다.

산업 현장에서도 같은 현상들이 일어나고 있다. 생산설비는 단 1초의 중단도 허용되지 않은 채 동작하며, 제품을 생산하는 동시에 그 제품이 만들어지는 동안의 설비 상태에 대한 데이터를 쏟아내고 있으며, 생산된 제품을 대상으로 측정된 품질과 포장 유통에 필요한 데이터가 새로 발생되는 일이 반복되고 있는 것이다.

기업은 새로운 가치를 창출하고, 이 데이터를 활용하고자, 무의미해 보이는 데이터 더미에서 가치를 채굴하기 위해 데이터를 모으고, 분석하여 가시화함으로써 숨어있는 의미를 찾아내고 기업의 이익으로 전환할 수 있기를 기대한다. 이러한 활동은 최근 강조되고 있는 4차 산업혁명을 대비하여 기업의 생존을 도모한다는 것과도 일맥 상통한다.

데이터를 모으고 분석하여 가치를 창출하는 것은 최근에 일어난 변화가 아니다. 이미 수많은 기업에서는 자사의 데이터를 잘 정리하여 분석하고 있고, 훌륭한 보고서를 자동으로 출력해 내기도 한다. 분야에 따라 다르기는 하지만, 비즈니스 인텔리전스(BI) 도구를 적극 활용하는 기업도 있고, DW를 구축하여 자사의 데이터를 여러 각도에서 분석하여 마케팅에 활용하는 기업도 상당수 있다. 하지만 데이터를 바라보는 이러한 접근 방식은 최근의 빅데이터 활용에 대한 요구사항을 만족시키기에는 분석의 목표 범위나 분석의 기술적 구현 방식 등에서 볼 때 여러 가지 차이점을 가지고 있다.

데이터 과학(Data Science)

"21세기의 가장 섹시한 직업은 데이터 과학자(Data Scientist)가 될 것이다."라는 말이 하버드 비즈니스 리뷰에서 언급된 이래 유행처럼 퍼지고 있다. 데이터 과학자는 통계학이나 수학에 능통하며, 데이터를 해킹하며, 특정한 분야에 대해 충분한 지식을 가지고 있어 지저분한 데이터로부터 통찰력을 이끌어 내는 사람으로 일컬어지고 있다. 누구도 하지 못했던 질문을 데이터를 대상으로 던지고, 소스 데이터를 이리저리 탐색하며, 새로운 데이터로 변형을 가해, 보이지 않던 가치를 다른 사람이 이해할 수 있는 언어 또는 다양한 매체로 설명하는 역할을 한다.

어떤 데이터 과학자는 정부를 보다 효율적으로 만들고, 노숙자를 돕고, 공중 보건을 개선하기 위해 데이터를 사용하며, 또 다른 어떤 데이터 과학자는 기업의 광고를 효율적으로 하기 위해 유저의 웹 활동 로그를 다양한 차원으로 분석하기도 한다.

때로는 통계학자와 데이터 과학자를 구분하기가 어렵고, 머신러닝 전문가와 데이터 과학자를 구분하는 것도 어려운 경우가 있다. 빅데이터를 잘 다루는 데이터 분석가가 자신을 데이터 과학자로 명명하는 경우도 있다. 여기에서는 데이터 과학자를 엄격하게 구분하여 특정 짓지 않고, 다만 데이터 과학이 무엇인지 살펴보고자 한다.

데이터 과학이란, 데이터 기반 과학이라고도 하는데, 데이터 마이닝과 유사하게 구조적 또는 비구조적 다양한 형태의 데이터에서 지식이나 통찰력을 추출하는 과학적 방법, 프로세스 및 시스템에 대한 학문 분야로 데이터로 실제 현상을 이해하고 분석하기 위해 통계, 데이터 분석 및 관련 방법을 통합하는 개념이다. 수학, 통계, 정보 과학 및 컴퓨터 과학의 다양한 영역, 특히 기계학습(머신러닝 ; Machine Learning), 분류(Classification), 클러스터링(Clustering), 데이터 마이닝(Data Mining), 데이터베이스 및 시각화(Visualization) 등의 많은 분야에서 가져온 기술과 이론을 사용한다.

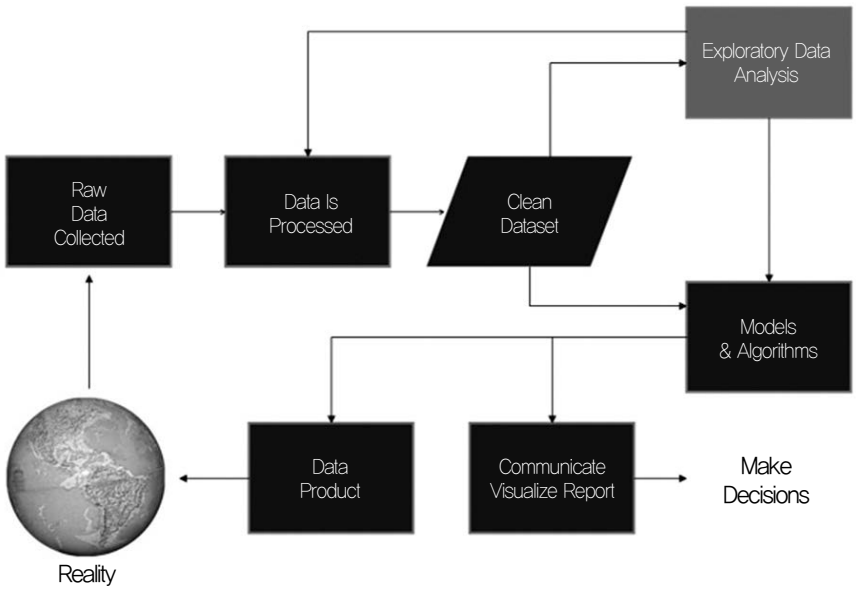
- 통계학(Statistics)
- 수학(Mathematics)
- 정보과학(Information Science)
- 컴퓨터과학(Computer Science)
- 기계학습(Machine Learning)
- 분류(Classification)
- 클러스터 분석(Clustering Analytics)
- 데이터베이스(Database)
- 가시화(Data Visualization)

[표 1] Data Science를 이루는 여러 학문 분야

데이터 과학자들은 고도로 발달한 학문과 기술들을 토대로 데이터에 생명을 불어넣는 작업을 진행한다. 이 작업에는 데이터를 수집하는 일, 데이터를 탐색하는 일, 다양하게 분류하는 일, 목적을 위해 새로운 파라미터의 패턴을 찾아내는 일, 필요없는 데이터를 제거하는 일, 변환을 위한 알고리즘과 모델을 설계하는 일 등이 포함되며, 밝혀진 의미를 다른 사람들에게 효과적으로 보여주는 일도 중요하다.

[그림 1]은 데이터 과학자가 하는 일들을 간략히 도식한 다이어그램이다.

빅데이터와 IoT 기술의 중요성에 비해 데이터 과학은 최근에야 그 중요성이 강조되고 있는데, 우리는 혼란스럽고 위험하기 그지없는 데이터 홍수에서 살아남기 위해 정교한 방법을 선택해야 하며, 이 도구가 바로 '데이터 과학'이 될 것이라는데 공감대가 형성되어 가는 추세이다.



[그림 1] Data science process from "Doing Data Science", Cathy O'Neil and Rachel Schutt, 2013

데이터 분석 셀프 서비스

데이터로부터 인사이트를 얻는 과정에는 데이터를 '수집'하고 '탐색'하여 '표현'하는 과정이 필연적으로 따른다.

데이터 과학자 혹은 데이터 분석가들은 '데이터 탐색'에 대한 오랜 경험을 통해 숨어있는 가치를 찾아내는 직관력을 가지게 되는데, 이에는 데이터를 분석하고 처리하는 이론적 토대가 필요한 동시에, 더 자유롭게 데이터를 조작할 수 있는 소프트웨어 도구들이 필요하다.

데이터 과학자나 데이터 분석가가 아니더라도 특정 목적을 위해 업무를 처리하는 과정에도 데이터를 자유롭게 변형하고 표현해 보는 도구는 필요하다. 아직도 일반적인 데이터 표현의 도구로 '엑셀'을 많이 사용하는 것을 보면, 스스로 데이터를 해석하고 표현하여 타인에게 전달하는 것이 아주 일반적이라는 것을 알 수 있다.

과거부터 지금까지도 대량의 데이터를 사용자 스스로 분석하게 하는 소프트웨어 솔루션들은 계속하여 발전해 왔다. 대표적으로 비즈니스 인텔리전스(BI) 도구는 비즈니스 분석가 스스로 데이터를 조회하고 계산하여 리포트를 만들 수 있도록 다양한 사용자 인터페이스를 제공하고 있다.

그런데, 이런 전통적인 도구들과 대비되는 현재의 데이터 분석 셀프 서비스의 차이점은 무엇일까? 그것은 다음 두 가지에서 크게 차이가 난다. (물론 전통적인 BI 진영도 데이터 과학을 위해 어느 분야보다도 빠르게 진화하고 있지만)

첫 번째는, 셀프 서비스의 대상 '범위'에 해당한다. 과거의 데이터 분석 소프트웨어들은 이미 정제되어 수집되어 있는 데이터를 대상으로 스스로 데이터 질의(쿼리)을 입력하여 대상 데이터 세트를 도출하는 방식을

취하고 있다. 비록, 전통적인 BI 도구들 역시 최신의 데이터 소스에 연결 할 수 있도록 다양한 커넥터를 추가 지원하고 있지만, 이 방식을 통해, 최신의 IoT 세상에서 발생되는 비정형 빅데이터를 스스로 취득하여 정제하거나 데이터 과학자의 관점에서 스스로 다양한 방식으로 데이터를 탐색하는 데는 한계가 있어 보인다.

두 번째는, 사용자가 데이터에 변형을 가하는 알고리즘의 접근 방식이 다름을 들 수 있다. 데이터 과학자는 통계학에 능통함을 가정해 보자. 이는 과거의 데이터 분석가나 유능한 통계학자와 마찬가지로 데이터를 바라보는 시각이 유사할 수 있음을 뜻한다. 그렇다면, 더 차이가 나는 것은 무엇일까? 차이는 많겠지만 대표적인 예로는 데이터 과학자는 머신러닝이 비록 통계학에 기반하였다 하더라도, 스스로 이 도구를 이용하여 데이터를 기반한 기계학습을 고려할 필요가 있음을 예로 들 때 과거의 BI 도구들로는 부족함을 느끼게 된다.

데이터 과학자 도구로서의 노트북

데이터 과학자는 데이터에 기반한 '스토리 텔링'의 역할을 한다. 소설가가 하얀 백지 원고지에 이야기를 창작하고, 화가가 캔버스에 그림을 그리며, 그래픽 디자이너가 컴퓨터 그래픽 도구로 이미지를 디자인하는 것처럼, 데이터 과학자는 데이터 시각화(Data Visualization) 도구를 이용하여 데이터에 대한 스토리를 창작한다.

그런데, 데이터 과학자의 상상력을 표현할 수 있으며, 동시에 숨어있는 가치를 발견할 수 있도록 영감을 불어넣어주는 시각화 도구라는 것이 무엇일까 궁금해진다. 참고로, 요즘의 데이터 과학자들은 빅데이터를 빈번하게 다루며, 기계학습 알고리즘도 설계하고, 가장 효과적인 방법으로 데이터를 표현하여 외부와 공유를 한다.

이렇게 빅데이터를 다루고, 머신러닝 알고리즘도 개발하고, 사람이 알 수 있는 방식으로 표현해 줄 수 있는 소프트웨어 도구 중 하나가 '노트북'이다. 현재 많은 분석가와 데이터 과학자들이 좋아하는 대표적인 노트북은 3가지가 있다.

Notebook	Originated
Jeppelin	Spark
Jupyter	iPython
Kibana	Elastic Stack

[표 2] 대표적인 노트북

노트북마다 강점을 가진 분야가 있어 데이터 과학자는 데이터의 특성 또는 데이터가 들어있는 컨테이너 종류 등에 따라 알맞은 노트북을 골라 사용하게 된다. 예를 들어, 주어진 데이터가 스팍(Spark) 혹은 하붑 계열의 컨테이너를 가진다면 Kibana는 좀처럼 이용하기가 쉽지 않고, Jeppelin의 쉬운 연동성을 선택하여 사용할 것이며, 통계 해석에 강한 R언어보다는 수치해석 패키지가 강한 Python 패키지가 필요한 경우, iPython의 진보한 형태의 Jupyter 노트북을 선택할 것이다.

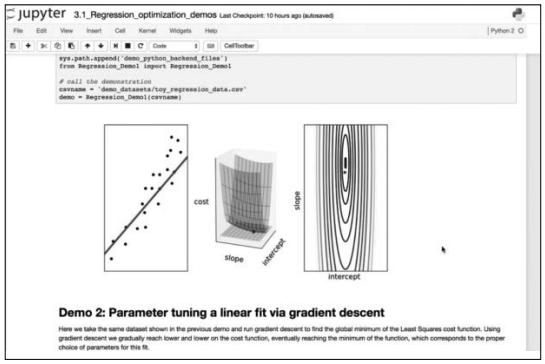
최근 들어 각 노트북들이 지원하는 패키지와 프로그래밍 언어는 점점 더 다양해지고 있어 개인의 취향이나 숙련도를 제외한다면 어떤 것이 더 좋은 노트북이라고 말하기란 쉽지 않다.

데이터 가시화(Data Visualization)

아래의 예시 화면들을 보자. 전달하려는 의미에 따라 데이터의 표현 방식은 천차만별로 다양하다.

보통 사람들의 인지능력은 3차원 이상에서는 현저히 저하된다고 한다. 하지만 실제 '현상'은 3차원 공간만

[그림 2] Jeppelin Notebook

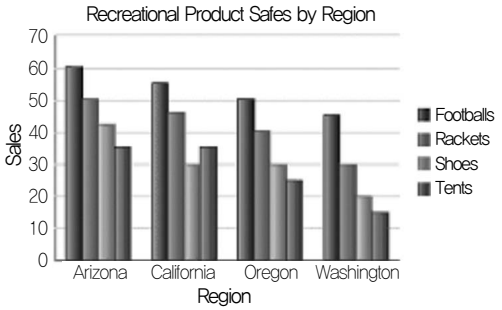


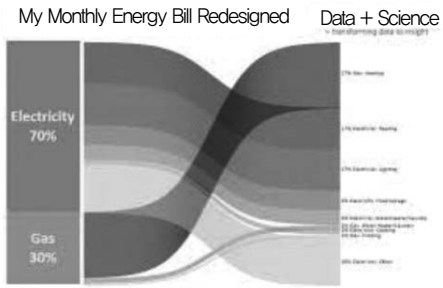
[그림 3] Jupyter Notebook

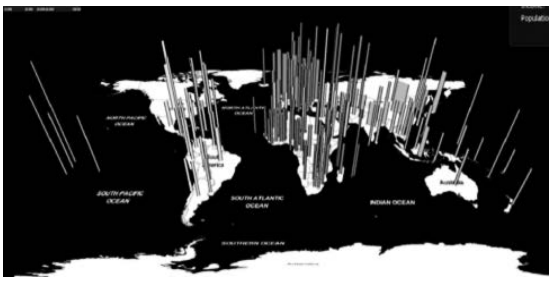
으로는 표현의 한계가 있다. 아주 간단한 예로, 시계열 데이터와 공간 사물의 변화를 동시에 표현하는 것이다. 특히, 종이나 컴퓨터 화면처럼 2차원 평면으로 실제 현상을 다양한 파라미터를 고려하여 표현한다는 것은 말처럼 쉽지는 않다. 데이터 과학자는 최대한 많은 의미를 직관적이며 효율적으로 타인에게 전달할 수 있는 방법을 계속 고려하지 않을 수 없다. 수많은 데이터 표현 방법이 있겠지만, 공통적으로 추구하는 바는 아래와 같다.

- 가능한 많은 컨텍스트를 담는다
- 데이터를 보는 사람의 의도에 반응한다

ID	Name	E-mail	Investments
231	Albert Master	albert.master@gmail.com	Bonds
210	Alfred Alan	aalan@gmail.com	Stocks
256	Alison Smart	asmart@biztalk.com	Residential Property
211	Ally Emery	allye@easymail.com	Stocks
248	Andrew Phips	andyp@mycorp.com	Stocks
234	Andy Mitchel	andym@hotmail.com	Bonds
226	Angus Robins	arobins@robins.com	Bonds
241	Ann Melan	ann_melan@iinet.com	Residential Property
225	Ben Bessel	benb@hotmail.com	Stocks







[그림 4] 데이터의 여러 가지 표현 방법

• 직관적 표현으로 데이터 해독을 돕는다

데이터의 보다 정확한 의미를 분석하기 위해서는 단편적인 결과보다는 데이터가 어떤 경로를 통해 변형이 되어 왔는지, 그 과정에 데이터 변형에 영향을 준 요소는 어떤 데이터 인지를 밝혀 최종으로는 '표현'을 해 주어야 한다. 이는 데이터 분석을 책임지는 데이터 분석가 혹은 데이터 과학자의 최종 결과물이 될 것이다. 아래의 목록은 데이터를 '표현'하는 도구로서 요즘 많이 거론되고 있는 오픈소스 또는 기술 요소들 중 일부다.

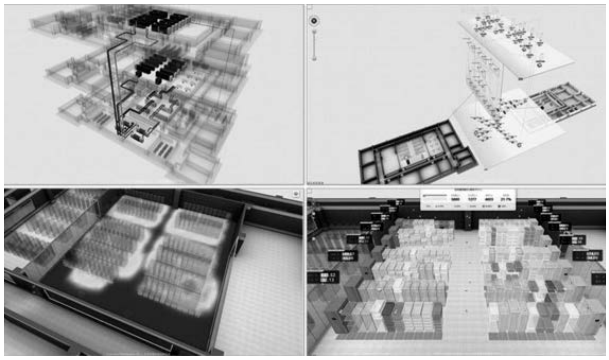
- D3.js
- SVG
- Chart.js
- Notebook
- Cesium(Data oriented globe system)
- Unity

이외에도 셀 수 없이 많은 기술과 오픈소스 프로젝트가 있지만 몇 가지만 나열한 것이며, 이들 중 대부분은 Web을 통한 인터렉티브를 기반한 표현과 공유가 가능함에 주의를 기울일 필요가 있고, '실제'를 '가상' 공간에 투영하기 위해, 3D 또는 2.5D 오브젝트에 데이터(실세계의 컨텍스트)를 기반한 에니메이션 등이 직관적인 표현을 돕는 도구로 활발히 이용되는 경우가 증가함에 주목할 필요가 있다.

이렇게 직관적인 표현은 최근에 강조되고 있는 CPS (Cyber Physical System)의 분야에도 관계된다.

위의 화면을 보자.

CPS에 꼭 3D를 이용해야 한다는 의미는 아니다. 하지만, 실제 세계를 가장 잘 표현해 주는 방법 중 하나는 실세계와 동일한 모양을 하고 있는 가상의 오브젝트를 이용하고, 실 세계로부터 센싱되는 데이터를 그 오브젝트에 적절히 표현하는 것은 사람의 인지력을 고려한 아주 효과적인 방법임에는 틀림 없다. 이에 더해, 가상의 오브젝트와 실세계의 조작자(사람 또는 환경)가 상호작용하고 반응하면 데이터의 좋은 표현 방법이 될 수 있다.



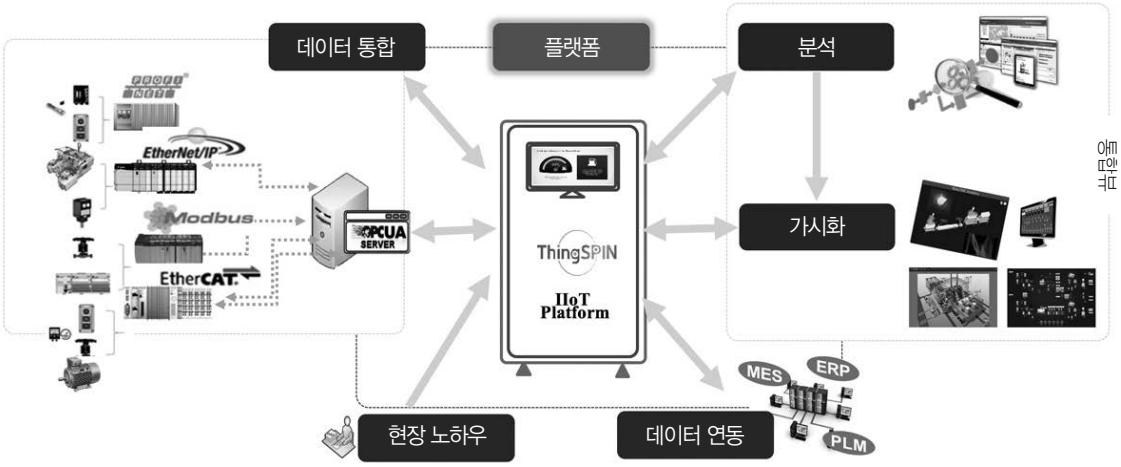
[그림 5] 3D 기반 데이터 센터 관제의 예

ThingSPIN®의 데이터 수집/분석, 가시화

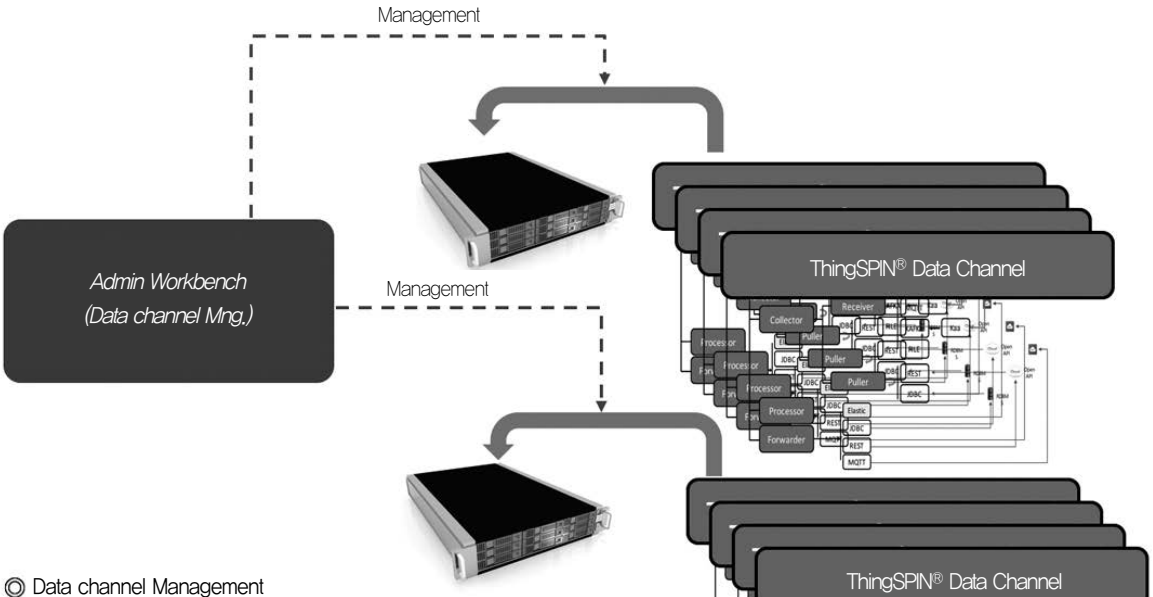
산업용 IoT 플랫폼 ThingSPIN®은 스마트 팩토리, 에너지 분야에서 제조/생산 및 전력 설비로부터 발생하는 데이터를 손쉽게 수집하고 분석하여 모니터링할 수 있도록 가시화해 주는 웹 서비스 플랫폼으로 데이터 사이언티스트를 위한 다양한 도구를 아래의 영역에서 제공한다.

- 다양한 데이터의 취득 및 연결
- 유연한 질의를 통한 데이터 탐색
- 데이터 재처리 및 분석 알고리즘 적용
- 다양한 사용자 정의 시각화 위젯

스마트 팩토리 구현의 시작은 산재한 데이터 소스로부터 데이터를 수집하는 것. 데이터는 생산설비, PLC, 센서 등의 장치에서도 발생하고, ERP·MES와 같은 레거시 시스템에서도 발생한다.



[그림 6] ThingSPIN® 플랫폼을 통한 데이터 수집/분배/분석 및 가시화



◎ Data channel Management

Property	Management Goal
Data Channel Server	Service Where? host : port (client should know)
Data Channel Spec	ID, protocol data format, data forwarding spec
Control	Enable, Disable, Create, Delete

[그림 7] ThingSPIN®의 데이터 채널 관리

이렇게 다양한 데이터를 항해하기 위해서는 먼저 데이터의 소스를 이해하고 발생하는 데이터를 통합하여 관리할 수 있어야 한다.

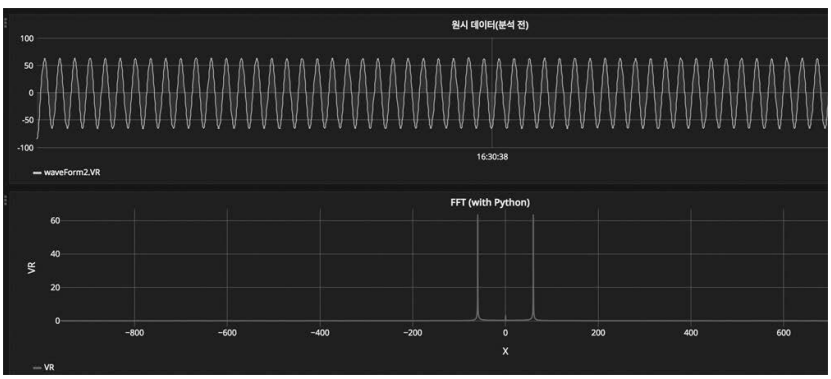
ThingSPIN에서는 데이터의 부류를 3가지 시각으로 바라본다. 하나는 데이터 필드값들이 잘 정리되어 구조화되어 있는 RDB 형식의 데이터, 하나는 서로 개연성을 가지고 있지 않아 보이는 아주 다양한 형식의 비정형 데이터들이고, 나머지 하나는 데이터의 시간에 따른 변화량이 중요한 시계열 데이터이다. 다양한 데이터 소스는 각기 다른 프로토콜로 외부에 데이터를 내어 준다. 이와 같은 데이터 연결에 대한 프로토콜의 다양성은 데이터 취득 과정에 어려움을 야기하는데, 이를 극복하기 위한 노력은 여러 분야에서 다양한 표준 제정의 시도로 나타나고 있으며, 산업계에서는 OPC UA(IEC 62541)가 그 대표적인 예라고 할 수 있다.

물론, 표준 프로토콜을 이용한다고 해서 각양각색으로 발생되는 데이터를 모두 연결할 수는 없을 것이다. 오픈 소프트웨어 생태계는 표준(Standards)의 지원뿐만 아니라 소프트웨어의 긴 역사 속에서 활발히 이용되어온 개발자 친화적인 기법들을 이용한 오픈소스 소프트웨어를 발전시킴으로써 데이터의 취득을 쉽게 할 수 있도록 다양한 솔루션들을 소개하고 있다. ThingSPIN은 이들을 최대한 이용함으로써 데이터 수집과 통합을 위한 기능을 사용자에게 제공한다.

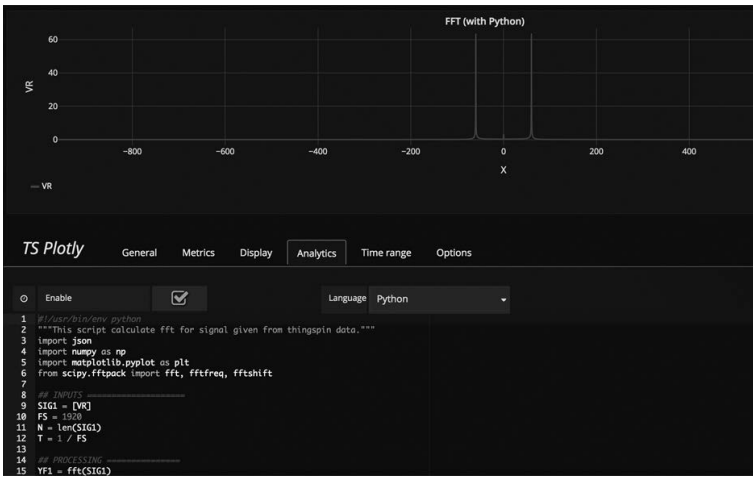
다양한 채널로 수집된 데이터는 RDBMS, NoSQL Big Data, Time-series DB 등 적절한 컨테이너에 저장될 필요가 있다. 데이터의 특성에 따라 가장 적절한 컨테이너를 선택하는 것도 중요하다.

수집된 데이터를 탐색할 차례다. 전통적으로 데이터에 대한 탐색은 데이터베이스 질의문(Database Query Language ; SQL문 등)을 통해 이루어졌다. 이는 현재까지도 유효하며 강력한 방법이다. 전통적인 데이터 질의 방법에 더해 필요한 것이 있다면, RDB빅데이터(Big Data)에 특화된 NoSQL을 대상으로 한 질의 역시 지원되어야 함을 들 수 있다.

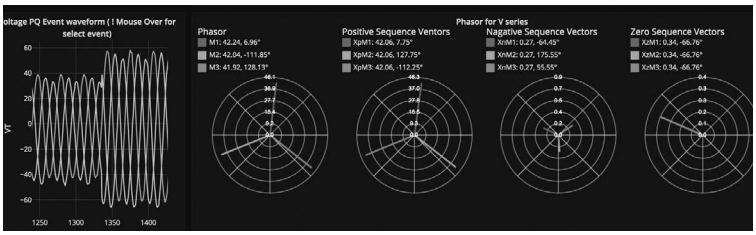
데이터의 탐색은 '질의(Query)'를 통해 소스데이터(Raw data)를 구하는 것에서 시작한다. 이제는 이 소스데이터를 이리저리 조작하거나 다른 소스데이터와 조합하면서 숨어있는 의미를 찾아내는 과정을 밟아야 한다. ThingSPIN®은 이 과정을 위해 데이터 사이언티스트가 사용할 수 있도록 앞에서 언급했던 '노트북(Notebook)' 기



[그림 8] 데이터 질의와 분석 결과 예시



[그림 9] R, Python 언어로 DSP 알고리즘 입력 예시



[그림 10] 전력 시그널 분석 결과의 예

능을 제공한다. R과 Python 언어로 데이터 분석 패키지를 사용할 수 있는 프로그램 입력기를 제공하여 데이터 질의를 통해 탐색된 소스데이터에 대한 분석을 수행하고, 분석된 데이터는 최종 가시화를 위해 다양한 UI 위젯들에 바인드(Bind) 되는 방식을 취하고 있다. [그림 8]의 예시 화면은 R 및 Python 언어로 DSP(Digital Signal Processing) 알고리즘을 적용하는 과정을 보여준다.

이렇듯, ThingSPIN® 플랫폼은 다양한 형식의 데이터 취득, 분석 및 가시화를 위한 노트북 기능 외 다양한 모니터링 기능을 제공하여 스마트 팩토리 또는 에너지 분야의 설비들을 효율적으로 분석할 수 있게 함으로써 실시간으로 현장 상태를 모니터링하는 동시에, 과거 데이터의 히스토리를 추적, 분석함으로써 문제가 발생했을 때 근본적인 원인을 찾아 개선하는데 도움이 될 것이다.

결 언

스마트 팩토리 분야의 '설비 예지 진단' 등에서도 데이터를 기반한 '현상 예지'와 '기계의 자율 의사 결정' 필요성이 대두되고 있는 요즘, 기업은 데이터 더미 속에 숨어있는 의미를 '채굴'하여 '가치'로 환산할 수 있는 새로운 직무역량을 갖출 필요가 있다. 데이터 과학자로 불리는 이들은 통계, 정보공학, 데이터베이스, 머신러닝, 가시화 기술 등의 백그라운드 지식과 도구를 가지고 데이터를 탐색하며, 경영 전반에 통찰력을 줄 것이다.