IBM Data Science Experience에서 데이터 사이언스의 진수를 배우다

주요 내용

- 데이터 사이언티스트와 그들이 이끄는 데이터 사이언스 팀은 첨단 분석 기능을 사용하여 원시 데이터를 유의미한 인사이트로 변환해야 합니다. 이를 위해서는 오픈소스 혁신을 포함한 최상의 툴을 협업과 공유의 소셜 기능과 연계하여 활용할 필요가 있습니다.
- IBM Data Science Experience는 데이터 사이언스 팀이 새로운 툴 및 동향에 대해 학습하고 최상의 오픈소스 및 IBM 기술을 활용해 가치를 창출하며 팀원뿐 아니라 더 광범위한 데이터 사이언스 커뮤니티와 공동으로 프로젝트를 수행할 수 있는 단일 통합 환경을 제공합니다.

데이터는 우리 주변 어느 곳에나 있지만 이 데이터를 유의미하고 가치 있는 자산으로 바꾸는 것은 데이터 사이언티스트와 데이터 사이언스 팀의 몫입니다. 그리고 이러한 목표를 이루기 위해서는 무엇보다도 적합한 툴을 제대로 사용할 수 있어야 합니다.

현재 데이터 사이언티스트들은 필요한 툴을 얻기 위해 오픈소스 기술을 찾는 경우가 많습니다. 오픈소스가 중요한 혁신 및 가치의 원천이 될 수 있으나 데이터 사이언티스트가 각종 오픈소스 기능을 모아 통합적으로 기능하는 단일 환경을 구성하기란 쉽지 않습니다. 대개는 사일로 및 병목 지점이 많은 상호 단절된 툴의 집합에 머무를 뿐입니다. 이와 같이 단절된 환경은 협업 및 생산성을 저하시킬 수 있습니다.

현재 데이터 사이언티스트 및 데이터 사이언스 팀에게 필요한 것은 데이터 사이언스 플랫폼입니다. 최근 Gartner 보고서에 따르면, 진정한 데이터 사이언스 플랫폼은 일관성 및 상호 운용성을 갖춘 빌딩 블록을 제대로 통합한, 응집력 있는 솔루션이어야 합니다.[1] 성급하게 끌어모은 오픈소스 툴의 집합은 이러한 정의와 거리가 멉니다.

IBM® Data Science Experience(DSX)는 데이터 사이언티스트 및 데이터 사이언스 팀이 성공을 거두는 데 필요한 협업 플랫폼입니다. DSX를 선택한 데이터 사이언티스트는 오픈소스 및 IBM 기술을 망라하여 업무 수행에 필요한 모든 툴을 확보하고 비즈니스를 위한 가치를 창출할 수 있습니다. 뿐만 아니라 데이터 사이언스 팀은 데이터 사이언스 전문가 커뮤니티와 함께 데이터 세트, 노트북, 기사 등 다양한 공유 리소스를 활용하면서 협업할 수 있습니다.

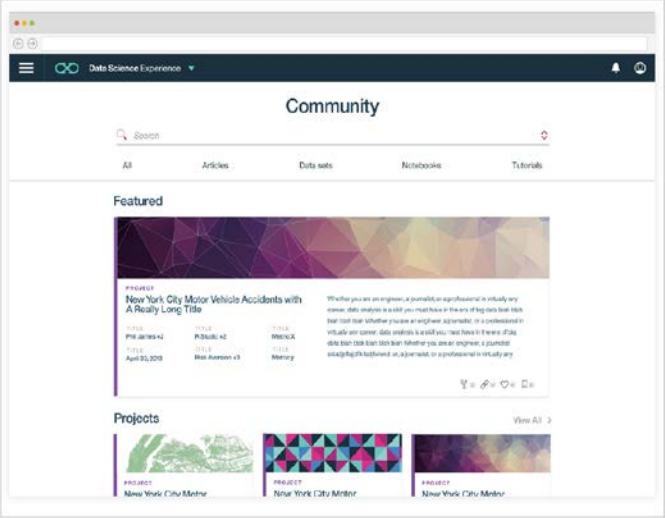

그림 1: IBM Data Science Experience의 커뮤니티 기능을 활용하여 전문가들의 도움으로 쉽게 시작하고 새로운 방식을 학습하거나 최신 데이터를 수집할 수 있습니다.

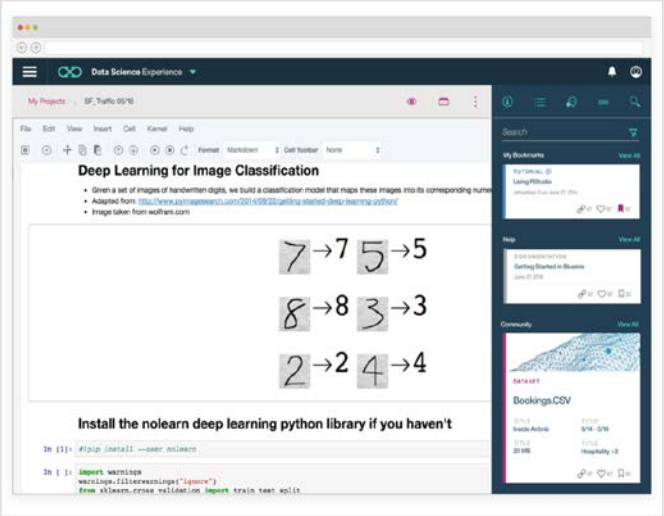

그림 2: 노트북은 다양한 언어(Scala, Python, R)를 지원하는 대화형 분석 및 시각화 툴이며 IBM Data Science Experience에 통합된 가이던스 기능을 활용합니다.

사전 구성된 데이터 사이언스 환경을 위한 오픈소스 툴 확대

현재 DSX는 Apache Spark, Jupyter 노트북, RStudio로 구성되어 있습니다. 이러한 초기의 툴 세트는 이미 데이터 사이언스를 위한 강력한 기초 환경을 제공하고 있지만 앞으로도 계속 발전할 것입니다.

대규모 데이터 처리를 위한 고속 범용 엔진인 Apache Spark는 80가지 이상의 고급 연산자를 제공하므로 손쉽게 병렬 애플리케이션을 개발할 수 있습니다. 데이터 사이언티스트는 Scala, Python, R shells에서 대화형 방식으로 사용할 수 있습니다. Spark의 모든 머신 러닝 라이브러리뿐 아니라 SparkR도 포함되어 있습니다. SparkR은 R에서 Spark를 사용하기 위한 경량의 프론트엔드로서 대형 데이터 세트에서 선택, 필터링, 집계 등의 연산을 지원하는 분산형 데이터 프레임워크를 구현합니다.

데이터 사이언티스트는 분석에 노트북 또는 RStudio를 사용할 수 있으며 하나의 프로젝트 내에서 두 가지 방식 모두 활용할 수도 있습니다. Jupyter 노트북은 코드 및 시각화를 포함한 Python, R, Scala 노트북을 생성하고 이를 기반으로 한 협업을 지원합니다. DSX에서 RStudio에도 액세스할 수 있습니다. 이 인기 높은 오픈소스 통합 개발 환경(IDE)은 신속한 R 스크립트 개발을 지원하므로 생산성이 더욱 향상됩니다.

DSX는 이러한 각각의 언어와 다양한 기법을 구사할 수 있다는 점에서 특별합니다. 다음 섹션에서 설명하겠지만, 이렇게 널리 사용되는 툴을 사전 구성된 환경으로 제공하여 시간과 노력을 줄일 수 있으므로 데이터 사이언티스트는 실제 데이터 분석과 같은 더 가치 있는 작업에 전념할 수 있습니다.

다양한 툴 및 데이터 소스 지원

DSX의 가장 매력적인 기능은 수많은 오픈소스 구성 요소를 통합하여 데이터 사이언티스트 및 데이터 사이언스 팀이 더 효과적으로 일할 수 있는 플랫폼을 제공한다는 것입니다. 보통 데이터 사이언스 툴을 사용할 때 설치, 설정, 유지 보수 등에서 어려움을 겪곤 하는데, DSX 사용자는 사전 구성된 데이터 사이언스 플랫폼을 즉시 사용할 수 있으므로 그러한 문제를 고민할 필요가 없습니다.

또한 DSX는 다음과 같이 다양한 데이터 소스를 직접 가져오거나 연결하여 사용할 수 있습니다.

- Amazon Redshift
- Apache Hive
- Cloudera Impala
- IBMDB2®
- IBM Informix®
- IBM Netezza®
- IBM dashDB™
- IBM Watson™ Analytics
- Microsoft Azure
- Microsoft SQL Server
- MySQL
- Oracle
- Pivotal Greenplum
- PostgresSQL
- Salesforce.com
- Sybase
- Sybase IQ

더 명석하고 빠른 팀워크

DSX를 활용하는 데이터 사이언티스트는 동료들과 협업하면서 프로젝트를 위한 더 나은 해결책을 찾습니다. 지식과 코드를 공유하면서 다른 데이터 사이언티스트의 연구에 도움을 주고 자신의 연구에 동료의 조언을 받을 수도 있습니다.

데이터 사이언티스트는 자신의 노트북을 전체 커뮤니티와 공유하여 성공적인 접근 방식을 다른 사이언티스트에게 시연하거나 연구에 대한 피드백을 받을 수 있습니다. 뿐만 아니라 DSX는 공유 데이터 세트, 다양한 튜토리얼, 실질적인 가이드를 제공하므로 새로 시작하는 데이터 사이언티스트 및 데이터 사이언스 팀은 필요한 모든 것을 갖춘 상태에서 시작할 수 있습니다. 경험 많은 데이터 사이언티스트 역시 여러 리소스를 활용하여 새로운 접근 방식을 시도해볼 수 있습니다.

나에게 맞는 구축 옵션

IBM은 DSX를 시작하는 사용자를 위해 다양한 옵션을 제공합니다.

퍼블릭 클라우드

데이터 사이언스에 대해 배우고 싶은 개인 사용자 또는 성능, 안정성, 확장성을 두루 갖춘 데이터 사이언스 툴 세트를 구축하려는 기업 모두 퍼블릭 클라우드 기반 DSX가 해결책이 될 수 있습니다. 퍼블릭 클라우드에 구축하면 직접 인프라를 구현하거나 관리할 필요 없이 간단하게 시작하면서 앞서 설명한 모든 툴 및 데이터 소스를 활용할 수 있습니다.

프라이빗 클라우드

프라이빗 기반 DSX는 협업을 활성화하고 자주 쓰이는 툴에 대한 간편한 액세스를 지원하는 등 퍼블릭 클라우드 옵션과 동일한 기능을 제공합니다. 게다가 기업의 자체 방화벽 내에서 구축할 수 있는데, 이는 구체적인 보안 요구 사항을 가진 기업에게 매우 중요한 조건입니다. 프라이빗 클라우드에서 DSX를 구축하는 기업은 직접 인프라를 구현하고 관리하면서 모든 보안 요구 사항을 확실하게 충족할 수 있습니다.

데스크탑

개인 사용자는 전체 플랫폼의 핵심 기능으로 구성된 더 작은 버전의 DSX를 다운로드하여 사용할 수 있습니다.

추가 정보

datascience.ibm.com에서 DSX를 시작해보십시오.

IBM 및 데이터 사이언스에 대한 자세한 내용은 ibm.com/datascience를 참조하시기 바랍니다.

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
2017년 4월

1 "Magic Quadrant for Data Science Platforms", Gartner, Inc. 2017년 2월 https://www.gartner.com/doc/3606026/magic-quadrant-data-science-platforms.

재활용하십시오.

XID12354-KRKO-01