2017년도 상위 10가지

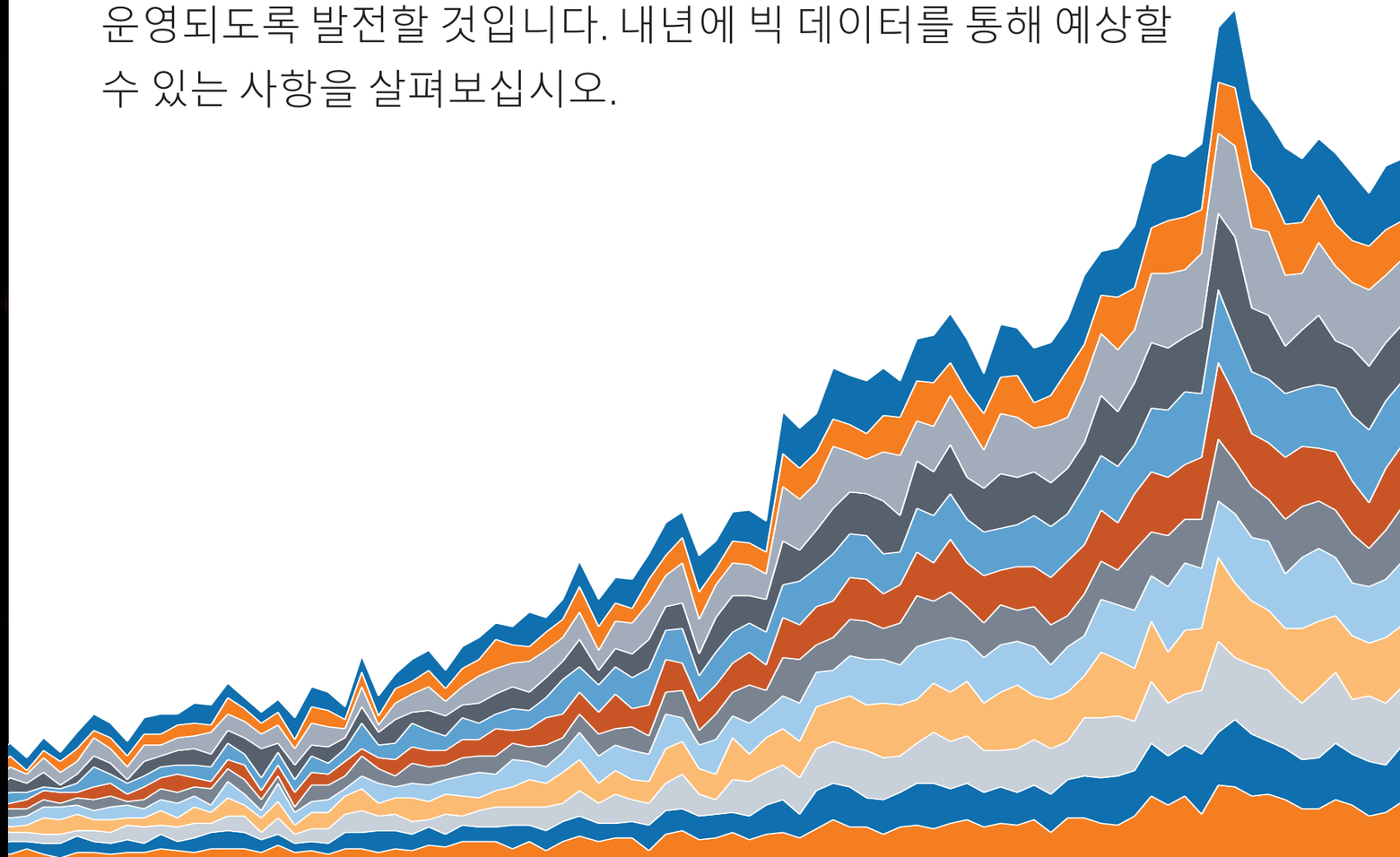
빅 데이터

동향

해마다 Tableau에서는 업계에서 관심을 끌만한 이슈로 토론을 이끌고 있습니다. 이러한 논의를 통해 다음 연도의 빅 데이터 상위 동향 목록을 만듭니다. Tableau가 예측한 동향은 다음과 같습니다.

상위 10가지 빅 데이터 동향

더욱 많은 조직이 모든 형태 및 규모의 데이터에서 가치를 저장하고 처리하며 추출하고 있습니다. 대량의 구조화된 데이터와 구조화되지 않은 데이터를 지원하는 시스템이 계속 증가하고 있습니다. 시장에서는 최종 사용자가 데이터를 분석할 수 있는 기능을 제공하면서 데이터 관리자가 빅 데이터를 관리 및 보호하도록 지원하는 플랫폼을 요구할 것입니다. 이러한 시스템은 엔터프라이즈 IT 시스템과 표준에 따라 원활하게 운영되도록 발전할 것입니다. 내년에 빅 데이터를 통해 예상할 수 있는 사항을 살펴보십시오.



1

데이터: Hadoop 속도 향상을 위해 확장되는 옵션

물론 Hadoop에서 기계 학습과 정서 분석을 수행할 수 있지만, 사람들은 일반적으로 대화형 SQL이 얼마나 빠른지 가장 먼저 질문합니다. 결국 더 빠르고 반복적으로 KPI 대시보드뿐 아니라 탐색적 분석을 위해 Hadoop 데이터를 사용하려는 비즈니스 사용자에게 SQL이 핵심 역할을 하기 때문입니다.

이러한 속도에 대한 요구는 Exasol 및 MemSQL 같은 더 빠른 데이터베이스와 Kudu 같은 Hadoop 기반 저장소, 더 빨리 쿼리할 수 있는 기술의 채택을 촉진했습니다. Hadoop 기반 SQL 엔진(Apache Impala, Hive LLAP, Presto, Phoenix 및 Drill)과 Hadoop 기반 OLAP 기술(AtScale, Jethro Data 및 Kyvos Insights)의 사용으로 인해 이러한 쿼리 가속기는 기존 웨어하우스와 빅 데이터의 경계를 더욱 모호하게 만듭니다.

2

Hadoop 이상의 빅 데이터: 사라져가는 Hadoop 전용 도구

지난 해에도 빅 데이터 물결과 함께 Hadoop에서의 분석 요구를 충족시키는 여러 기술의 부상을 목격했습니다. 하지만 복잡하고 이질적인 환경을 가진 기업에서는 더 이상 단 하나의 데이터 원본(Hadoop)을 위한 격리된 BI 솔루션을 도입하고 싶어하지 않습니다. 기업이 갖는 질문에 대한 답이 레코드 시스템에서 클라우드 웨어하우스까지, Hadoop 또는 Hadoop이 아닌 원본에서 온 구조화된 데이터와 구조화되지 않은 데이터에 이르는 수많은 원본에 매몰되어 있습니다. (그런데 관계형 데이터베이스조차도 빅 데이터를 지원하도록 준비하고 있습니다. 예를 들어 SQL Server는 최근에 JSON 지원을 추가했습니다.)

고객은 모든 데이터에 대한 분석을 요구할 것입니다. 데이터나 원본에 구속되지 않는 플랫폼이 번성하고, Hadoop 전용 플랫폼과 다양한 사용 사례 확보에 실패한 플랫폼은 사라질 것입니다. Platfora의 퇴장은 이러한 동향의 초기 지표로 여겨집니다.

3

처음부터 데이터 레이크를 활용하여 가치를 창출하는 조직

데이터 레이크는 인공 저수지와 유사합니다. 즉, 먼저 끝을 막은 다음(클러스터 구축) 물(데이터)을 채웁니다. 레이크를 구축하고 나면 전기를 생산하거나, 식수로 사용하거나, 레크레이션을 즐기는 등(예측 분석, ML, 사이버 보안 등) 다양한 용도에 맞게 물(데이터)을 사용하기 시작합니다.

지금까지는 레이크에 물을 공급하는 것이 그 자체로 중요했지만, Hadoop에 대한 비즈니스 요구가 커짐에 따라 점차 바뀌어 갈 것입니다. 조직에서는 더 빠른 답을 찾기 위해 레이크의 반복적이고 민첩한 사용을 요구하고 인력, 데이터 및 인프라에 투자하기 전에 비즈니스 성과를 신중하게 고려할 것입니다. 이로 인해 비즈니스와 IT 사이에 강력한 파트너십이 조성되며, 셀프 서비스 플랫폼이 빅 데이터 자산을 활용하기 위한 도구로 더 크게 인정받을 것입니다.

4

프레임워크에 획일적 적용 불가

Hadoop은 더 이상 데이터 과학에 사용하기 위한 일괄 처리 플랫폼이 아니며, 애드혹 분석을 위한 다용도 엔진이 되었습니다. 기존에 데이터 웨어하우스에서 처리하던 일상 워크로드에 대한 운영 보고에도 사용되고 있습니다.

조직은 사용 사례별 아키텍처 디자인을 통해 이러한 하이브리드 요구에 부응할 것입니다. 또한 사용자 특성, 질문, 양, 액세스 빈도, 데이터 속도, 집계 수준을 비롯한 다양한 요소를 조사한 뒤에 데이터 전략을 결정할 것입니다. 이러한 최신 참조 아키텍처에는 요구 변화에 따라 재구성되는 방식으로 최고의 셀프 서비스 데이터 준비 도구, Hadoop Core 및 최종 사용자 분석 플랫폼이 결합될 것입니다. 이러한 아키텍처의 유연성을 통해 궁극적으로 다양한 기술을 선택할 수 있게 됩니다.

5

규모나 속도가 아닌 다양성이 빅 데이터 투자 주도

Gartner에서는 빅 데이터를 많은 양, 빠른 속도, 고다양성 정보 자산의 3가지 속성으로 정의합니다. 규모, 속도 및 다양성 모두가 증가하고 있지만 New Vantage Partners에서 실시한 최신 설문조사에 따르면 다양성이 빅 데이터 투자에 가장 큰 영향을 미치는 것으로 파악되었습니다. 이러한 동향은 기업에서 더 많은 원본을 통합하고 빅 데이터의 '롱테일'에 집중함에 따라 계속 확대될 것입니다. 스키마가 없는 JSON에서 다른 데이터베이스(관계형 및 NoSQL)의 중첩 유형, 비플랫 데이터(Avro, Parquet, XML)까지 데이터 형식이 크게 증가하고 커넥터는 더욱 중요해지고 있습니다. 기업에서는 이러한 이질적인 데이터 원본에 직접 라이브 연결을 제공할 수 있는 능력에 따라 분석 플랫폼을 평가할 것입니다.

6

빅 데이터의 가치를 제고하는 Spark와 기계 학습

Apache Spark는 한때 Hadoop 환경의 구성 요소에 불과했지만 이제는 기업에서 선택하는 빅 데이터 플랫폼이 되고 있습니다. 데이터 설계자를 대상으로 한 설문조사에 따르면 IT 관리자와 BI 분석가의 70%가 이미 사용 중인 MapReduce보다 Spark를 선호했습니다. MapReduce는 일괄 처리 중심이며 대화형 응용 프로그램이나 실시간 스트림 처리에는 적합하지 않습니다.

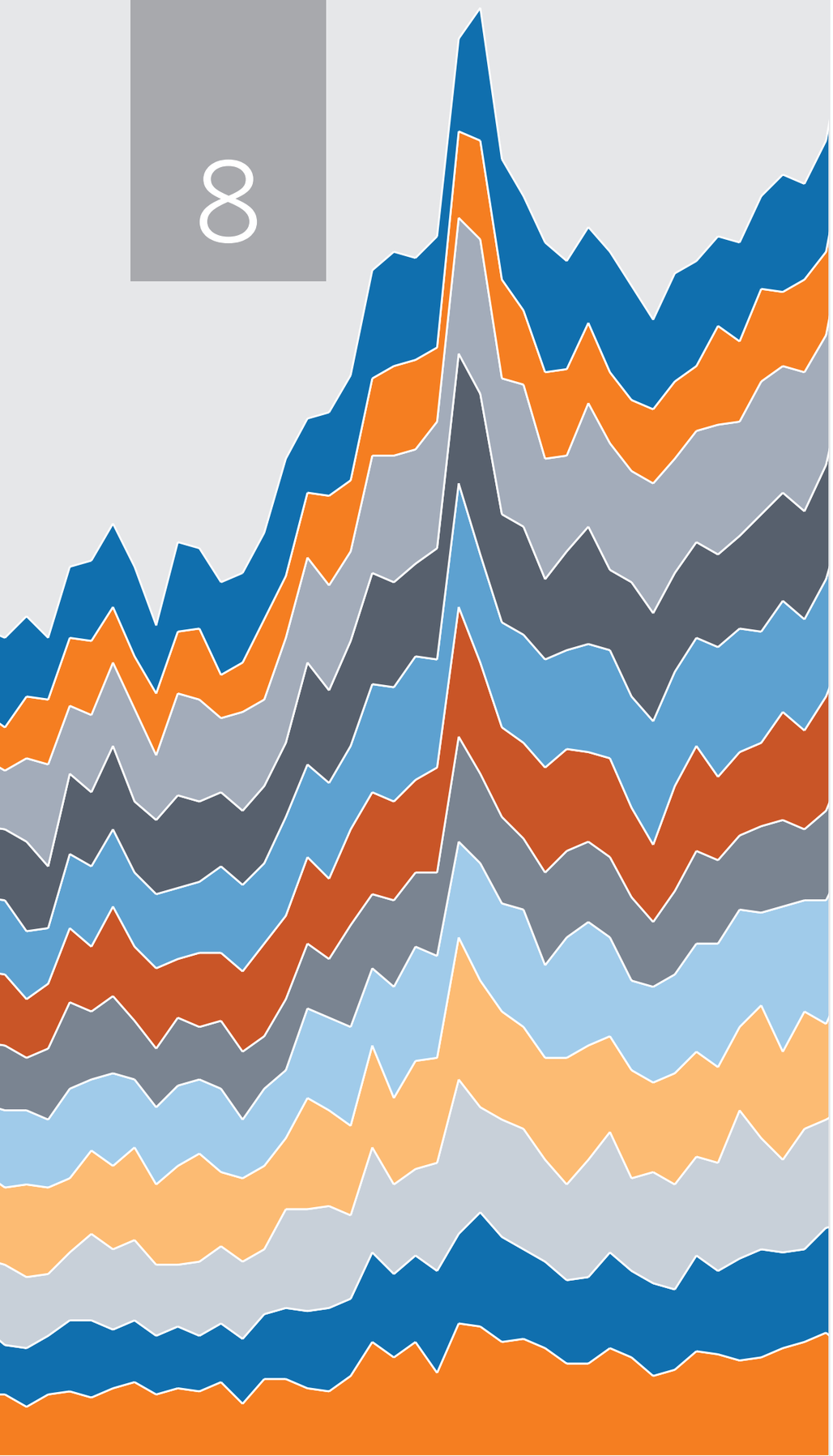
이러한 빅 데이터에 대한 컴퓨팅 기능은 컴퓨팅 집약적인 기계 학습, AI, 그래픽 알고리즘이 포함된 플랫폼의 수준을 높였습니다. 특히 Microsoft Azure ML의 경우 초보자도 쉽게 사용할 수 있고 기존 Microsoft 플랫폼과의 통합이 용이하여 인기가 급격히 높아졌습니다. ML이 대중에 공개됨에 따라 페타바이트 수준의 데이터를 처리하는 모델 및 응용 프로그램의 개발이 증가할 것입니다. 기계 학습과 스마트 시스템이 증가함에 따라 이러한 데이터에 대한 최종 사용자의 접근성을 높여주는 셀프 서비스 소프트웨어에 이목이 집중될 것입니다.

7

셀프 서비스 분석에 대한 새로운 기회를 만드는 IoT, 클라우드 및 빅 데이터의 융합

중앙으로 정보를 전송하는 센서가 모든 사물에 부착됩니다. IoT는 엄청난 양의 구조화된 데이터와 구조화되지 않은 데이터를 생성하고 있으며, 이러한 데이터가 클라우드 서비스에 제공되는 비중이 계속 증가하고 있습니다. 이러한 데이터는 대개 이질적이며, Hadoop 클러스터에서 NoSQL 데이터베이스까지 다양한 관계형 및 비관계형 시스템에 저장됩니다. 저장소와 관리형 서비스의 혁신으로 수집 프로세스의 속도가 높아졌지만, 데이터 액세스 및 파악은 여전히 해결해야 할 마지막 숙제로 남아 있습니다. 이에 따라 클라우드에 호스팅된 광범위한 데이터 원본에 원활하게 연결하고 이를 결합하는 분석 도구에 대한 요구가 증가하고 있습니다. 기업에서는 이러한 도구를 사용하여 데이터가 어디에 저장되어 있든지 모든 유형의 데이터를 탐색하고 시각화하며 IoT 투자에서 숨겨진 기회를 발견할 수 있습니다.

8



최종 사용자가 빅 데이터를 활용하기 시작하면서 부상한 셀프 서비스 데이터 준비

비즈니스 사용자가 Hadoop 데이터에 액세스할 수 있도록 하는 일은 현재 가장 큰 과제 중 하나입니다. 셀프 서비스 분석 플랫폼의 증가로 이러한 환경이 개선되고 있습니다. 하지만 비즈니스 사용자는 분석을 위한 데이터 준비에 소요되는 시간과 복잡성을 더 줄이고 싶어하며, 이는 다양한 데이터 유형과 형식을 처리할 때 특히 중요합니다.

민첩한 셀프 서비스 데이터 준비 도구를 사용하면 원본 수준에서 Hadoop 데이터를 준비할 수 있을 뿐 아니라 스냅샷 형태의 데이터로 더욱 빠르고 쉽게 탐색할 수 있습니다. 우리는 Alteryx, Trifacta, Paxata 등 빅 데이터용 최종 사용자 데이터 준비에 집중한 기업에서 많은 혁신을 보아 왔습니다. 이러한 도구는 Hadoop을 늦게 도입한 업체 및 후발 업체에 대한 진입 장벽을 낮추고 있으며, 이러한 동향은 계속 이어질 전망입니다.

9

빅 데이터 증가: 엔터프라이즈 표준이 되어가는 Hadoop

Hadoop이 점차 엔터프라이즈 IT 환경의 핵심 부분이 되어가고 있습니다. 보안 및 엔터프라이즈 시스템 관련 구성요소에 대한 투자가 늘어나고 있습니다. Apache Sentry는 Hadoop 클러스터에 저장된 데이터와 메타데이터에 매우 세부적인 역할 기반 인증을 시행할 수 있는 시스템을 제공합니다. 데이터 거버넌스 전략의 일환으로 개발된 Apache Atlas는 조직 전체 데이터 환경에서 일관성 있게 데이터를 분류할 수 있도록 지원합니다. Apache Ranger는 Hadoop에 대한 중앙 집중화된 보안 관리 기능을 제공합니다.

고객은 엔터프라이즈급 RDBMS 플랫폼에서 이러한 유형의 기능을 기대하기 시작했습니다. 이러한 기능들이 새로운 빅 데이터 기술의 중심으로 옮겨감에 따라 지속적으로 기업의 도입 장벽이 사라지고 있습니다.

10

메타데이터 카탈로그로 사용자가 분석 가치가 있는 빅 데이터를 찾을 수 있도록 지원

기업에서는 오랫동안 데이터를 버렸습니다. 처리할 것이 너무 많았기 때문입니다. Hadoop을 사용하여 많은 데이터를 처리할 수 있지만 데이터는 통상 찾기 쉬운 방식으로 정리되어 있지 않습니다.

메타데이터 카탈로그를 사용하면 사용자가 셀프 서비스 도구를 사용하여 분석 가치가 있는 관련 데이터를 탐색하고 이해할 수 있습니다. 고객 요구와 데이터 간의 이러한 차이는 Alation, Waterline 같은 회사의 기계 학습을 사용하여 Hadoop 데이터 검색 작업을 자동화하면 해결될 수 있습니다. 이러한 회사에서는 태그를 사용하여 파일을 분류하고, 데이터 자산 간의 관계를 파악하며, 검색 가능한 UI를 통해 쿼리 제안도 제공합니다. 이를 통해 데이터 사용자와 데이터 관리자 모두 신뢰할 수 있는 데이터를 찾고 정확하게 쿼리하는 데 걸리는 시간을 줄일 수 있습니다. 내년에는 셀프 서비스 탐색에 대한 인식과 요구가 증가하여 셀프 서비스 분석이 확대될 것입니다.

Tableau 정보

데이터 시각화를 소매 프로그램과 프로세스에 통합하는 것은 생각보다 쉽습니다.

Tableau Software는 데이터의 크기 및 데이터를 저장한 시스템의 수에 상관없이 사람들이 데이터를 보고 이해할 수 있도록 지원합니다. PC에서 iPad에 이르기까지 편리한 환경에서 데이터 대시보드를 신속하게 연결, 통합, 시각화 및 공유할 수 있습니다. 자동 데이터 업데이트 기능이 있는 대시보드를 만들고 게시하여 동료, 파트너 또는 고객과 공유할 수 있으며, 프로그래밍 기술이 필요하지 않습니다. 지금 바로 무료 평가판을 시작해 보십시오.

TABLEAU.COM/KO-KR/TRIAL