BESPIN GLOBAL

HELPING YOU ADOPT CLOUD.

CLOUD INSIGHT

빅데이터 분석 환경의 핵심, 데이터레이크 구축하기 Part 2

Contents

빅데이터 분석 환경의 핵심, 데이터레이크 구축하기 Part 2

1. 데이터전처리

워크로드별로 가장 효과적인 전처리 엔진 선택하기

단순히 데이터 저장을 위해 데이터레이크를 사용하는 경우를 제외하고는, 데이터레이크는 수집되고 저장된 데이터가 데이터드리븐 어플리케이션을 만들기 위해 **전처리**를 거칠 때에만 진정한 가치를 제공할 수 있습니다. 데이터레이크가 얼마나 다용도로 사용될 수 있는지는 아래 처리 가능한 워크로드를 보면 알 수 있습니다.

- 배치
- 스트림 워크로드
- 머신러닝 워크로드
- 그래프 프로세싱 워크로드
- BI 워크로드 - SQL 쿼리 프로세싱

그러나 데이터 전처리 과정의 어려움은 사용 가능한 데이터 프로세싱 엔진 종류가 다양하다는 것입니다. 각각의 데이터 프로세싱 엔진은 목적과 용도가 서로 다릅니다. 프로세싱 엔진을 선택할 때는 각 엔진의 내부 아키텍처, 가장 적합한 사용 용도, 비즈니스 SLA, 팀의 전문성과 데이터레이크 내 다른 구성요소들을 모두 고려해야 합니다.

▣ 체크 포인트

데이터레이크의 데이터 전처리 레이어를 설계할 때는, 프로세싱 플랫폼이나 엔진을 선택하기 전에 다음 질문을 던져보세요.

- 엔진이 어떻게 데이터 전처리를 수행하고 계획을 실행하는가?
  - 비순환 그래프(Directed acyclic graphs, DAGs)와 실행 계획은 분산 엔진이 수행하는 작업과 종속성을 드러냅니다. DAGs는 가장 간단한 방법으로는 실행 엔진 외부에서 관리할 수도 있고(하둡 MapReduce 등), 실행 엔진과 긴밀히 커플링 될수도 있습니다. 요즘은 대부분의 프로세싱 엔진이 외부 모델을 사용하지 않습니다. MapReduce는 하둡 생태계 내의 일부 핵심 프로덕트에서는 많이 사용되지만, Apache Spark과 같은 더 효과적인 모듈 형의 엔진들에 의해 대체되고 있습니다.
- 엔진의 시스템 아키텍처가 어떻게 여러 명의 사용자를 위한 전처리를 분산시키고 동시 실행과 계산을 위한 리소스를 배분하는가?
- 엔진이 처리량과 레이턴시 측면에서 얼마나 빠르게 서로 다른 작업을 실행하는가?
- 전처리 엔진이 어떻게 작업 및 노드 장애에 대응하는가?
  - 전처리 엔진별로 하드웨어나 소프트웨어 장애에 다르게 반응합니다. 전처리 엔진을 선택하기 전에 이것을 이해하고 분석하세요.

- 전처리 엔진이 어떻게 설계되어 있는가? 다량의 레코드로 되어 있는가, 아니면 이벤트의 그룹으로 되어 있는가?
- 조직이 어떤 워크로드를 데이터레이크에서 운영하려고 하는가?

▣ 핵심 노하우

아래는 데이터 전처리 프레임워크와 엔진을 선택할 때 추천하는 베스트 프랙티스입니다.

MapReduce를 언제, 어떻게 사용할 것인가: MapReduce는 개발자가 운영과 데이터 흐름에 대한 세부 항목을 모두 관리해야 하는 낮은 단계의 프레임워크입니다. 많은 설정과 상용구 코드 작성을 직접 해주어야 합니다. 파일 압축이나 분산 파일 복제와 같은 작업은 MapReduce 패러다임과 아주 잘 맞습니다. MapReduce는 이러한 패러다임에 익숙한 경험 많은 개발자들이 사용해야 하며, 세부 항목을 일일히 관리하는 것의 장점이 압도적으로 클 때 사용하는 게 좋습니다.

많은 기업이 Lambda나 Kappa 중심의 아키텍처를 사용하고 있습니다. 요즘은 전처리 패러다임으로 Kappa의 스트리밍 중심 아키텍처를 많이 사용하는 추세입니다.

대용량 데이터레이크를 위해 설계된 수많은 SQL 프로세싱 엔진이 SQL 데이터 전처리, BI, 특수 목적의 워크로드 등을 지원합니다. 그러나 대부분의 SQL 엔진은 대규모 데이터셋에 대한 복잡한 분석 쿼리가 동시다발적으로 실행될 때, 속도와 성능이 저하됩니다. 따라서 규모가 큰 SQL이나 BI 워크로드에 복잡한 분석 쿼리를 동시 실행해야 할 경우에는 SQL 엔진 사용을 추천하지 않습니다.

2. 사용

모든 사람이 비즈니스 트랜스포메이션의 주체가 되는 사용 레이어 설계하기

데이터레이크는 대용량 데이터를 수집하고 처리할 수 있지만, 실제 가치는 처리된 데이터가 데이터드리븐 어플리케이션을 위해 사용될 때 드러납니다. 다운스트림 어플리케이션과 이용자가 데이터레이크의 데이터를 사용할 수 있게 하는 것은 혁신과 데이터드리븐 어플리케이션 개발의 토대입니다. 데이터레이크 접근 프로세스를 완화하면 데이터 사이언티스트 뿐만 아니라 조직 내 모든 사람이 비즈니스 트랜스포메이션의 주체가 될 수 있습니다.

데이터레이크의 요건 중 하나는 다양한 이용자의 요구사항을 만족하는 것입니다. 데이터레이크 사용은 크게 데이터레이크 외부와 내부로 나눌 수 있습니다.

- 데이터레이크 내의 프로세스나 워크로드. 예를 들어:
  - ETL 또는 ELT 워크로드
  - ML 알고리즘
  - 데이터레이크 기반의 데이터 마트
  - 데이터 처리 및 분석 툴
  - 데이터레이크 내 데이터 거버넌스, 데이터 카탈로그, 보안, 메타데이터, MDM 툴
- 데이터레이크 외부의 프로세스나 워크로드. 예를 들어:
  - 데이터를 데이터 웨어하우스나 데이터 마트로 밀어내기
  - 데이터 시각화 툴이나 BI 툴에 접근
  - 데이터 분석 툴
  - 데이터레이크 외부의 데이터 거버넌스, 데이터 카탈로그, 보안, 메타데이터, MDM 툴

데이터레이크의 데이터에 아래 두 가지 방법으로 접근할 수 있습니다.

- 데이터 푸시 (밀어내기)
  - 데이터 내보내기
  - 다운스트림 프로세스를 위해 메세지 대기열로 내보내기
- 데이터 풀 (뽑아내기)
  - 데이터 서비스
  - 데이터 보기
  - SQL 접근
  - REST API
  - GraphQL
  - 데이터 시각화 툴
  - 데이터 마켓플레이스

- 데이터를 사용하는 어플리케이션에서 프로토콜에 따라 데이터를 뽑아낼 수 있는 서비스

▣ 체크 포인트

데이터레이크 아키텍트와 엔지니어는 데이터레이크의 사용 레이어를 설계하기 전에 아래를 고려해야 합니다:

- 데이터를 사용하는 어플리케이션의 데이터 접근 속도 및 처리량 관련 요구사항은 무엇인가?
- 데이터를 사용하는 어플리케이션은 어떤 종류의 접근을 요구하는가(랜덤 읽기, 순차적 읽기, 랜덤 쓰기, 순차적 쓰기)?
  - 색인 기능은 랜덤 읽기를 가능하게 하고 방대한 데이터셋에서 작은 부분을 빠르게 찾게 해줍니다.
- 데이터 사용 레이어를 설계할 때, 데이터 푸시와 풀의 확장성과 동시 실행을 어떻게 지원할 것인가?
- 내고장성과 하드웨어/소프트웨어 결함을 어떻게 처리할 것인가?
- 어떻게 데이터레이크가 여러 개의 데이터 모델을 제공하도록 할 것인가?
- 쉽고, 빠르고, 확장가능하고, 퍼포먼스가 좋은 데이터 접근을 가능하게 하는 데이터 구조와 스토리지는 무엇인가?
- 이용자가 마치 데이터 마켓플레이스처럼 데이터를 쉽게 구할 수 있는가?
- 이용자가 직접, 혹은 카탈로그를 사용해 데이터레이크를 전체적으로 검색할 수 있는가?
- 데이터 접근에 대한 자동화된 거버넌스가 이뤄지고 있는가?
- 데이터 접근이 기록, 검사, 모니터링되고 있는가? 이러한 기록에 대한 지표가 저장되어 있는가?
- 이용자들에게 데이터 품질과 개요를 어떻게 확인하게 할 것인가?

▣ 어려움

데이터레이크의 사용 레이어를 설계할 때 겪는 일반적인 어려움은 다음과 같습니다:

- 동시다발적인 접근이 가능하게 하는 것
- 데이터 모델과 스키마를 관리하고, 비일관성을 없애는 것
- 인덱싱, 태깅, 검색 기능을 통해 데이터를 잘 찾을 수 있게 하는 것
- 서로 다른 이용자의 셀프 서비스를 지원하는 것
- 내부와 외부의 이용자가 접근하는 데이터의 보안을 유지하는 것
- 접근되는 데이터의 품질을 지속적으로 유지하는 것
- 툴을 사용해 데이터 접근에 대한 기록, 트래킹, 감시를 수행하는 것

▣ 핵심 노하우

데이터레이크의 사용 레이어를 설계하기 위한 베스트 프랙티스는 다음과 같습니다:

- 사용 레이어를 데이터 색인/데이터 거버넌스/보안 툴과 연동해 접근 정책/권한 허가/사용자 식별을 통제하세요. 데이터레이크의 데이터를 성공적으로 사용하기 위해서는 보안 정책 설정과 실행이 필수적입니다. 데이터 소유자가 데이터의 보안 접근 요구사항을 설정하고, 누가 데이터에 접근할 수 있는지 정해야 합니다.
- 보안에 관한 세부사항은 메타데이터로 저장되며, 데이터 보안 담당자가 정책 수행을 위해 사용합니다.
- SQL, API, 검색, 내보내기, 대용량 접근과 같은 가장 일반적인 데이터 접근 방식을 허용하세요.
- 성공적인 데이터레이크를 위해서는 셀프 서비스가 필수적입니다. 서로 다른 사용자들이 데이터레이크를 사용하고, 요구사항도 천차만별이지만, 이들은 모두 IT 부서의 도움 없이 셀프 서비스 방식으로 데이터에 접근하기를 원합니다.
- 누가 어떤 데이터를 언제, 어디서, 어떻게 사용하는지 기록을 남기고 추적, 감독해야 합니다.
- 데이터 내용, 정의, 그 외 데이터가 수집, 보강, 통합, 전환, 사용을 거치는 데이터 파이프라인을 따라 이동하면서 생성되는 모든 메타데이터를 데이터 카탈로그에 기록하세요. 이것은 자동화된 메커니즘에 따라 업데이트되어야 합니다.

3. 운영

데이터레이크 운영을 자동화하는 "데이터옵스" 도입하기

데이터레이크를 만드는 것도 어렵지만, 유지하는 것은 더 어렵습니다. 적합한 툴과 효율적인 모니터링 및 알림 프레임워크가 없다면 프로덕션 데이터레이크 관리는 그만큼 복잡하고 어려워집니다.

아이데이션부터 프로덕션까지 데이터레이크 구축 절차는 매우 복잡합니다. 그래서 컴포넌트 중심의 운영관리보다는 어플리케이션 중심의 접근법이 효율적입니다. 이 접근법은 배포 시간, 시장 진출 시간을 단축하고, 반복적인 프로세스를 수립할 수 있습니다. 아래는 데이터레이크 운영을 위한 노하우입니다:

▣ 핵심 노하우

- 반복적인 워크플로우를 사용하고, 코드 기반이 아닌 설정 기반으로 사용하여 효율성을 높이세요.
- 자동화된 배포를 사용해 한 번의 클릭으로 배포가 가능하게 하고, 롤백도 쉽게 할 수 있도록 하세요.
- 플랫폼 가용성과 중요 지표를 체크하기 위해 사전적인 모니터링과 알림 기능을 사용하세요. 가용성, 트렌드, SLA 위반 등을 트래킹할 수 있는 지표를 설정해, 인텔리전스를 제공하세요.
- 자동화된 인시던트 분석과 헬스체크를 수행하세요.
- 데이터레이크를 구축하기 전에, 데이터레이크에서 사용할 수 있는 리소스를 툴, 소프트웨어, 프레임워크와 매핑해 보세요. 클러스터 계획, 용량 계획, 클러스터 설계, 컴포넌트 레이아웃을 작성해 보세요.
- 마스터-슬레이브 아키텍처를 사용하는 컴포넌트의 마스터 프로세스가 여러 랙에 분산되도록 하세요. 랙 결함으로 인한 리스크를 줄일 수 있습니다. 로드밸런싱과 클라이언트 오퍼레이션을 위해 여러 개의 게이트웨이 노드를 배포하세요.
- Chef, Puppet, Jenkins, 레드햇 앤시블과 같은 툴을 사용해 프로비저닝과 배포 프로세스를 자동화하세요.
- 모든 보안 컴포넌트는 HA모드로 가동되어야 합니다.
- 데브옵스와 "데이터 옵스" 팀을 활용해 지속적 통합, 지속적 딜리버리, 지속적 배포 원칙을 따르고 끊김없는 데이터레이크 배포 프로세스를 따를 수 있게 하세요.
- 컨테이너화된 워크로드를 사용하고 쿠버네티스와 같은 컨테이너 오케스트레이션 툴을 사용하세요. 리소스 사용 개선, 멀티테넌시, 리소스 분리, 오토스케일링(자동 확장), 자기 회복 기능을 활용할 수 있습니다.

결론

오늘날의 급변하는 비즈니스 환경에서, 기업은 모을 수 있는 데이터는 모두 모아 활용해야 합니다. 데이터를 최대한 많이 모으고, 비즈니스 목적에 맞게 인공지능 기술을 사용해야 합니다. 경영진들은 데이터레이크 구축에 점점 관심을 가질 것입니다. 데이터레이크는 빅데이터 분석 환경의 핵심이기 때문입니다.

데이터레이크를 성공적으로 구축하기 위해서는, 비즈니스를 먼저 이해해야 합니다. 비즈니스에서 어떤 종류의 데이터를 가지고 있는지, 누가 이러한 데이터를 사용하며 이들의 요구사항은 무엇인지 파악해야 합니다. 이를 바탕으로 각각의 단계에서 의사결정을 내리는 것이 좋습니다. 메타데이터 관리, 접근 통제 등 거버넌스에 대해서도 미리 계획을 세워둬야 합니다. 아래와 같이 결론을 정리할 수 있습니다.

- "끝을 염두에 두고 시작하라." 특정한 요구사항이나 비즈니스 그룹, 사용자 및 분석 집단이나 사용 사례를 염두에 두고 데이터레이크를 구축해야 성공 확률을 높일 수 있습니다. 엔터프라이즈 전체를 관통하는 "빅뱅" 전략은 추천하지 않습니다.
- 데이터 웨어하우스 경험에서 얻은 교훈을 학습하세요. 사용법은 차차 발견할 것으로 희망하면서 데이터레이크를 구축하지 마세요.
- 데이터레이크의 리스크를 낮추려면, 초기 단계부터 거버넌스를 염두에 두어야 합니다. 그렇지 않으면 데이터 접근성이 낮아지고, 데이터 퀄리티가 낮아지며, 데이터 간 연동성이 낮아지고 보안이 취약해질 수 있습니다. 또한 메티데이터 관리도 부실해집니다.
- 메타데이터 레이어를 구축하고 통제된 데이터 검색이 가능하도록 설계하세요. 이는 데이터 접근성, 상황에 맞는 인사이트 추출, 분석을 위해 꼭 필요합니다. 데이터 자산을 색인하는 것도 필수적입니다.

데이터레이크를 구축하기 가장 좋은 환경은, 유연성과 확장성이 높은 클라우드 환경입니다. 베스핀글로벌의 빅데이터 전문 팀은 고객의 IT 환경 뿐만 아니라 비즈니스 환경까지 분석하여 최적의 데이터레이크 아키텍처를 설계합니다. 데이터레이크 설계부터 구축, 운영까지 클라우드 전문가 베스핀글로벌에 문의하세요.

※ 함께 읽어 볼만 한 연관 콘텐츠

- 머신러닝을 똑똑하게 활용하기 위해 준비해야 하는 것
- 머신러닝 그리고 AI, 완벽해질 때까지 기다리지 말고 지금 바로 시작하세요!
- 건설기계사 도입 사례를 통해 보는 AI가 적용된 수요 예측 관리
- 엔터프라이즈의 인공지능(AI)과 머신러닝(ML) 적용은 왜 어려울까?

“Helping You Adopt Cloud.”

클라우드 전문 인력 및 클라우드 관리 어떻게 시작해야 하나요?
클라우드 IT를 대한민국에서 가장 잘 아는 400여 명의 클라우드 전문가를 베스핀글로벌에서 만나실 수 있습니다.

Meet 베스핀글로벌 전문 컨설팅팀
클라우드의 도입을 고민하신다면 베스핀글로벌 컨설팅팀이 최적의 클라우드 컨설팅을 제공해 드립니다.

Meet 아시아 최고의 클라우드 운영팀, 베스핀글로벌 매니지드 서비스팀
클라우드 운영에 대한 노하우가 부족해 걱정이시라면 클라우드 서비스의 베테랑이 여러분의 클라우드를 관리해드립니다.

Meet OpsNow
클라우드 비용을 최대 80% 절감하고, 장애 대응 시간을 절반으로 줄일 수 있는 멀티-클라우드 관리 플랫폼 OpsNow를 만나보세요.

About Bespin Global

국내 최다 클라우드 인증 자격을 보유한 MSP
국내 유일 ISO 인증, ISMS 인증, GS인증을 확보한 MSP
가트너가 인정한 한중일 유일한 MSP

베스핀글로벌은 클라우드 IT를 위해 태어난 클라우드 매니지드 서비스 기업입니다. 클라우드 도입을 위한 전략컨설팅, 클라우드 상에서 수많은 고객의 IT 자산을 안정적으로 운영하고 관리할 수 있는 서비스와 솔루션, 클라우드를 기반한 머신러닝과 빅데이터와 같은 최신 기술의 빠른 도입까지 클라우드로 할 수 있는 전 영역을 End - to - end로 제공합니다. 3년 연속 가트너 매직 쿼드런트 퍼블릭 클라우드 MSP 부문에 한·중·일 최초로 등재되었고, 포브스로부터 한국의 주목할 만한 유니콘으로 선정되기도 했습니다.

클라우드로 가기로 결정하였다면 누구와 함께 갈지를 선택해야 합니다.
처음부터 끝까지 믿을 만한 파트너를 찾는다면 베스핀글로벌이 정답입니다.

클라우드로 가기로 결정했다면
누구와 함께 갈지를 선택해야 합니다.

처음부터 끝까지 믿을만한 파트너를 찾는다면
베스핀글로벌이 정답입니다.

베스핀글로벌 웹사이트
서비스 문의

베스핀글로벌 소셜미디어

BESPIN GLOBAL
HELPING YOU ADOPT CLOUD.