



Informatica™

최고의 빅 데이터 워크북

첫 번째 빅 데이터 프로젝트를
시작하기 위한 실무 가이드

목차

빅 데이터로 전환	03	3장 린(Lean) 빅 데이터 공급망	28
1장 준비	04	팀	29
알아야 할 사항	05	- 팀 구축을 위한 5가지 주요 교훈	30
- 대부분의 조직에서 빅 데이터 프로젝트를 구현하는 이유	06	- 데이터 거버넌스 설정	33
- 빅 데이터 프로젝트가 실패하는 이유	07	- 필요한 기술과 소유한 기술	35
- 빅 데이터 프로젝트를 실현하는 방법	08	- 빅 데이터 톨 이해	37
올바른 프로젝트 선택	10	프로세스	38
- 올바른 프로젝트의 형태	11	- 빅 데이터 8단계	39
- 영향 고려	13	아키텍처	41
- 전술적 빅 데이터 프로젝트: 몇 가지 예	15	- 첫 번째 단계: 샌드박스(sandbox)	42
2장 전략	16	- 이상적인 빅 데이터 아키텍처	43
목표 정의	17	프로젝트 계획	44
- 현업 부서 목표	18	- 시작하기	47
- IT 목표	20	다음 단계	48
데이터 요구 사항 정의	22	Informatica 정보	49
- 필요한 데이터가 무엇입니까?	23		
- 데이터에 대한 5가지 주요 고려 사항	26		

각 단원으로 바로 들어 가려면 제목을 클릭하시기 바랍니다.

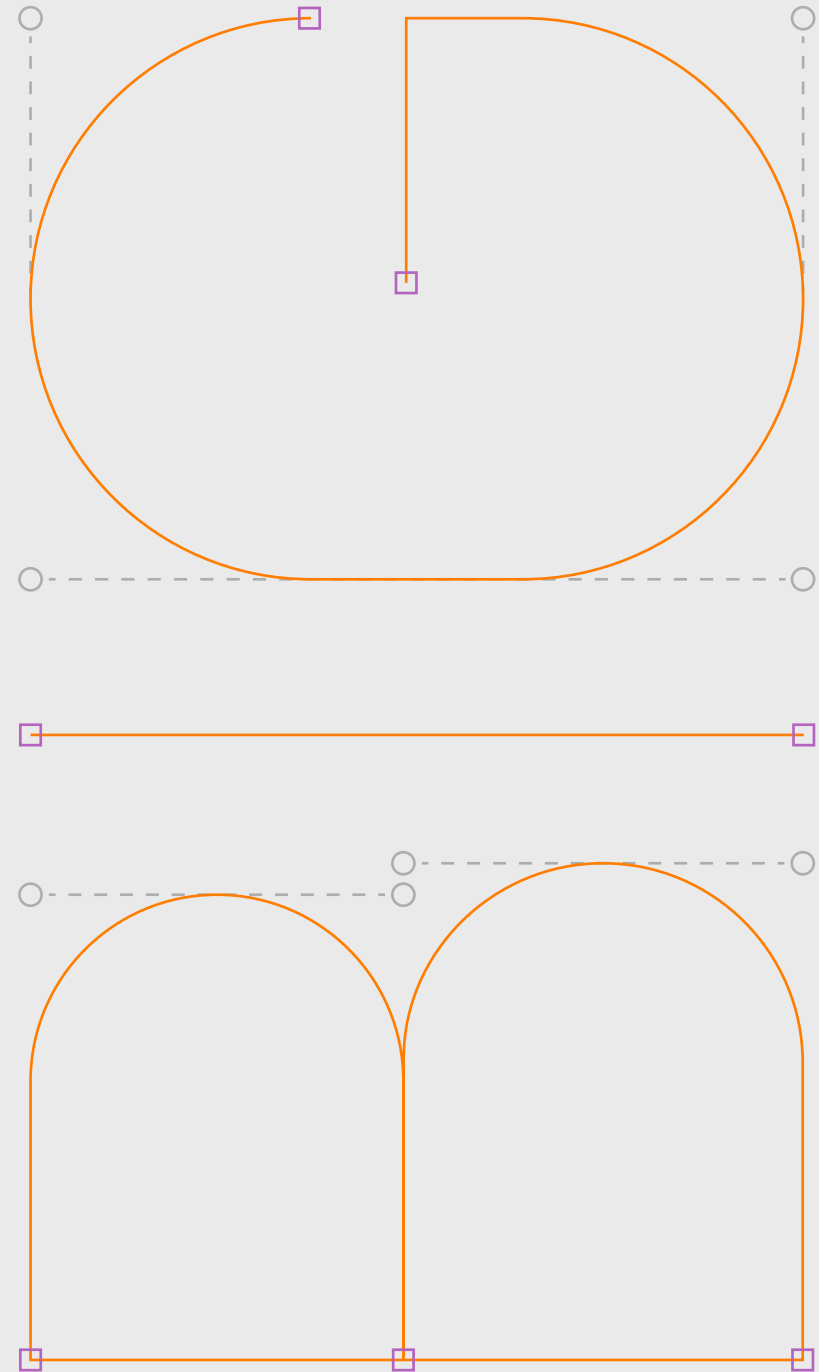
빅 데이터로 전환

빅 데이터 만큼 부풀려진 측면이
많은 기술 추세는 없었습니다.

반면 빅 데이터 만큼 많은 혁신의
잠재력을 기업에 제공한 기술 추세도
없었습니다. 20세기 초에 소프트웨어가
전체 비즈니스 프로세스에 적용되기
시작한 이래로 데이터가 작업 방식을
바꾼다는 사실은 명확해졌습니다.

빈틈 없는 전략을 수립하여 현실에
맞게 실천하는 것과 관련된 내용을
담고 있습니다. 국지화된 전술적
이니셔티브를 시작하든 아니면 전체
기업을 위한 보다 기초적인 노력을
계획하든 이 워크북은 여러분의 전환을
위한 실무 가이드 역할을 할 것입니다.

지금부터 자세히 살펴보겠습니다.



1장 준비

이 워크북은 세 단원으로 나누어져 있습니다. 이 첫 번째 단원에서는 올바른 프로젝트를 선택할 수 있도록 비전을 명확히 하는 데 중점을 두고 있습니다.

알아야 할 사항

여러분의 고유한 프로젝트에 대한 특성을
살펴보기에 앞서, 대부분의 빅 데이터
실무자들이 프로젝트를 시작하기 전에
알아야 할 몇 가지 교훈을 알아보겠습니다.

대부분의 조직에서 빅 데이터 프로젝트를 구현하는 이유

기업에서 빅 데이터를 다루기로 결정한 경우 그 이유는 일반적으로 다음 중 하나입니다.

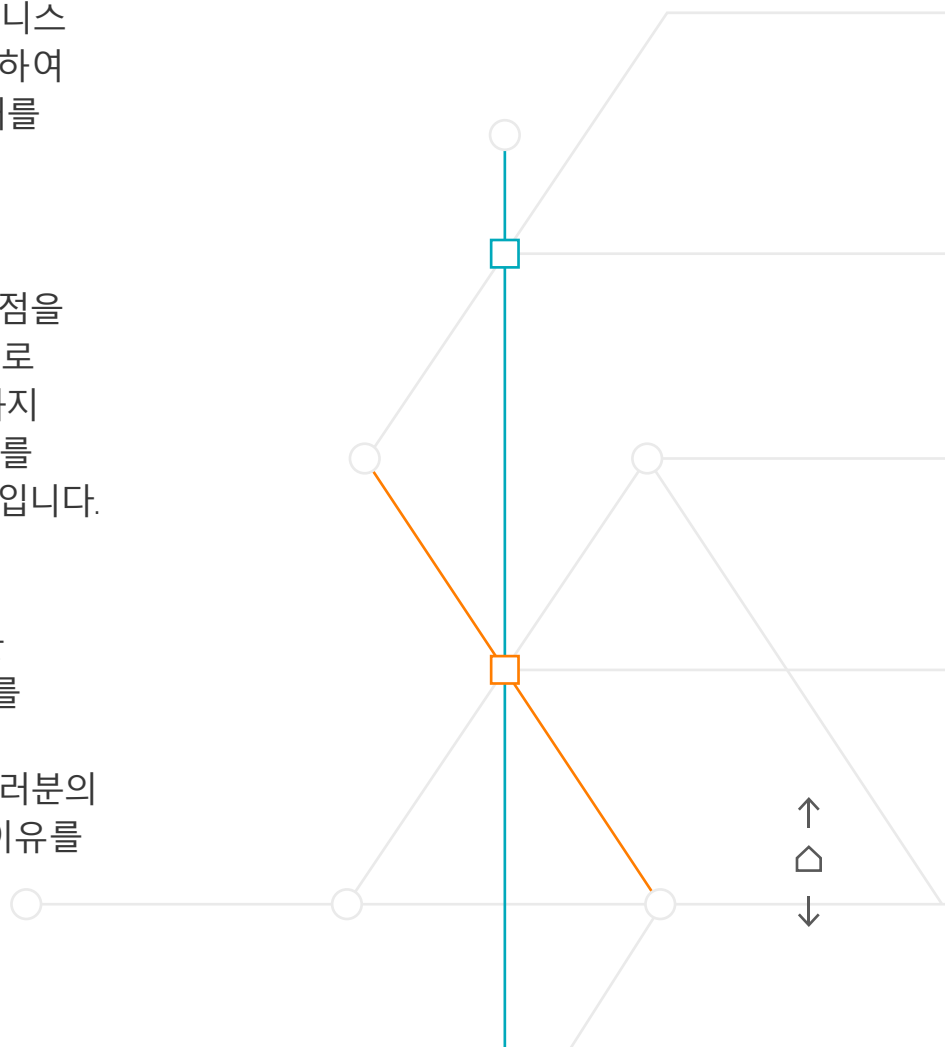
더 나은 분석을 수행하고 분석할 데이터의 양이 크게 증가할 것임을 인식하고 있기 때문입니다. 일반적으로 현업 부서(예: 마케팅)가 이러한 이니셔티브를 시작합니다.

고객이 제품을 보다 효율적이고 효과적으로 사용하도록 도와줄 수 있는 분석을 제공(종종 실시간으로)하여 제품을 서비스 계층에서 보완할 수 있다는 것을 인식하기 때문입니다. 빅 데이터를 사용하여 특정 비즈니스 단위 또는 프로세스에 대한 모든 의사 결정을 알려 작업을 보다 빠르고 효율적이며 저렴하게 수행하고 싶기 때문입니다.

빅 데이터가 조직의 모든 비즈니스 단위에 중요하다는 것을 인식하여 데이터 중심 관점을 위한 토대를 마련하려고 하기 때문입니다.

너무 늦지 않게 빅 데이터에 익숙해지기 시작해야 한다는 점을 알고 있지만 실제로 빅 데이터로 무엇을 해야 할지 아직 파악하지 못했기 때문입니다. 빅 데이터를 학습하고 실험하는 것이 목표입니다.

이 모든 이유는 빅 데이터에 대한 관심을 가지는 데 유익한 이유입니다. 하지만 프로젝트를 지속적으로 테스트하고 여러 부서의 관심을 유지하려면 여러분의 관심사를 가장 잘 나타내는 이유를 매우 명확히 알아야 합니다.



빅 데이터 프로젝트가 실패하는 이유

설문 조사¹ 결과, 모든 빅 데이터 프로젝트의 55%가 완료되지 않았으며, 나머지 45% 중 대부분은 목표에 미달한 것으로 나타났습니다.

이 충격적인 결과는 기술 추세의 초기 단계에서는 그다지 이례적이지 않지만 이러한 프로젝트가 가르쳐 줄 수 있는 교훈을 배우지 않는 것은 어리석은 짓입니다.

빅 데이터 프로젝트가 실패하는 네 가지 주요 원인을 살펴보겠습니다.

1. 모호한 목표

설문 조사에서 드러난 가장 일반적인 실패의 원인은 프로젝트의 '부정확한 범위'였습니다. 명확한 목표 없이 의욕적인(한데 합치면 과도하게 모호한) 프로젝트를 목표로 한 다음 중요한 사항을 가려내는 힘든 선택을 해야 할 때 실패하는 회사가 너무 많았습니다. 빅 데이터 프로젝트를 추진하기 위해 빅 데이터를 추구하는 것은 재앙을 불러옵니다. 이러한 프로젝트의 복잡성은 특정 결과를 얻기 위한 확고한 신념을 필요로 합니다. 이는 명확한 목표 없이는 불가능합니다.

2. 잘못 관리되는 기대치

빅 데이터와 관련된 모든 부풀려진 측면은 프로젝트가 제공할 수 있는 가치에 대한 몇 가지 매우 위험한 가정으로 이어집니다. 이는 단기간에 대담한 약속을 할 수 있을 정도로 유혹적이지만 프로젝트에서 기대할 수 있는 결과, 소요 기간 및 노력에 대한 사실적 관점을 유지하는 것이 중요합니다. 영향과 통찰력에 대한 기대가 지나치게 높으면 미지의 금광을 찾고 있는 자신을 발견하게 될 것입니다. 기대치가 비현실적이면 기한 및 예산이 전혀 타당하지 않게 됩니다.

3. 프로젝트의 과도한 비용 지출 및 지연

이 분야가 기업에 얼마나 새로운지를 고려하면 대부분의 빅 데이터 프로젝트가 과도한 비용으로 종결되거나

시간이 너무 오래 걸린다는 것이 그리 놀라운 일은 아닙니다. 이는 일반적으로 관리되지 않는 기대치와 확장 가능한 아키텍처를 구축하는 방법에 대한 오해가 혼합된 결과입니다. 희소하고 비싼 Hadoop/Java 개발자를 채용하여 핸드 코딩 구현을 맡긴다면 오류 없이 샌드박스 환경에서 벗어나는 것이 불가능하다는 것을 곧 깨닫게 될 것입니다. 그 결과 데이터 프로젝트는 과학 실험에 그치고 실험실 밖으로 나오지 못하게 될 것입니다.

4. 확장 불가능

5명의 뛰어난 Hadoop/Java 개발자를 찾는 것도 어려운데, 프로젝트가 성장하여 1년 후 30명의 Java 개발자로 확장해야 한다면 벽에 부딪힐 수 있습니다. 사용하지 않은 Hadoop 클러스터의 기회 비용을 날리게 되며, 최악의 경우 추진력과 시간의 기회 비용을 잃어버게 됩니다. 기업은 장기적인 지속 가능성보다 단기적인 편의를 목표로 하는 경우가 너무 많습니다. 이러한 상충 관계를 피하는 방법을 제안해 드릴 수 없지만, 장기적인 관점의 중요성을 재차 강조하는 바입니다. 데이터를 적절히 관리하고 보안을 유지하려면 프로젝트의 장기적인 영향을 계속 지켜보아야 합니다.

빅 데이터 실패의 네 가지 원인은 너무나도 자주 발생합니다. 따라서 이를 방지하고 지속되는 구현을 구축할 수 있는 방법을 알아보겠습니다.

¹ InformationWeek, "[Vague Goals Seed Big Data Failures](#)."

빅 데이터 프로젝트를 실현하는 방법

대부분의 빅 데이터 프로젝트가 명확성이 부족하고 이니셔티브의 실용성을 입증할 수 없어 실패한 경우 프로젝트에 중점 사항과 증거를 추가해야 합니다. 다음은 프로젝트에 착수하고 이를 본격적으로 추진하는데 유용한 세 가지 팁입니다.

1. 명확한 목표를 설정하고 기대치를 관리하십시오.

프로젝트의 목표를 잘 모를 경우 기존 데이터 인프라에 대해 설정한 목표를 고려하십시오.

귀사에서 특정 비즈니스 프로세스(예: 사기 감지 또는 시장 분석)에 대한 데이터의 필요성을 이미 느끼는 경우 빅 데이터를 통해 이러한 프로세스를 보다 효율적으로 수행하거나 더 가치 있게 수행하는 방법을 고려하십시오. 완전히 새로운 문제로 고민하기보다는 기존 프로세스 또는 프로젝트를 개선하는 것을 목표로 삼아야 합니다.

현업 부서 사용자에게 대한 명확한 초점과 증명 가능한 가치가 없으면 프로젝트는 실패합니다.

2. 프로젝트의 가치를 증명하는 지표를 정의하십시오.

명확하게 정의되고 목표와 연관된 지표는 많은 고민을 덜어 줄 수 있습니다. 측정할 수 있는 현실적인 목표를 설정하면 주변의 모든 사람이 진행 과정을 지켜볼 수 있습니다.

특히 이들은 여러분의 장기적인 목표를 알게 될 것입니다. 목표의 맥락에서 프로젝트의 영향을 어떻게 측정할 수 있는지에 대한 답을 구하십시오.

이는 현업 부서 사용자의 경우 지표를 합리적으로 만드는 것을 도와야 하는 단기적인 의무가 있고, 여러분의 경우 측정 가능한 목표를 통해 이들이 인식하는 것보다 더 큰 가치를 전달한다는 것을 증명할 수 있기 때문에 매우 중요합니다.

3. 툴 및 핸드 코딩을 전략적으로 고려하십시오.

모든 것을 Hadoop에서 수동으로 핸드 코딩하려는 유혹에 빠지지 마십시오. 여기에서 목표는 처음부터 맨손으로 구축하는 것이 아니라 조직에 빅 데이터의 가치를 전달하는 것입니다.

모든 통합을 핸드 코딩하고, 모든 데이터 세트를 정제하며, 모든 분석을 핸드 코딩하는 대신 이러한 프로세스를 보다 신속하게 수행하는 데 도움이 되는 툴 및 자동화를 찾아야 합니다.

특히, 다른 직원으로 확장 또는 전수할 수 없다는 측면에 희소하고 값비싼 Java 개발 인력을 낭비하는 것에 걸리지 마십시오. 여러분의 역할은 목표를 달성하기 위해 희소한 리소스의 개발에 대해 전략적인 의사를 결정하는 것입니다.

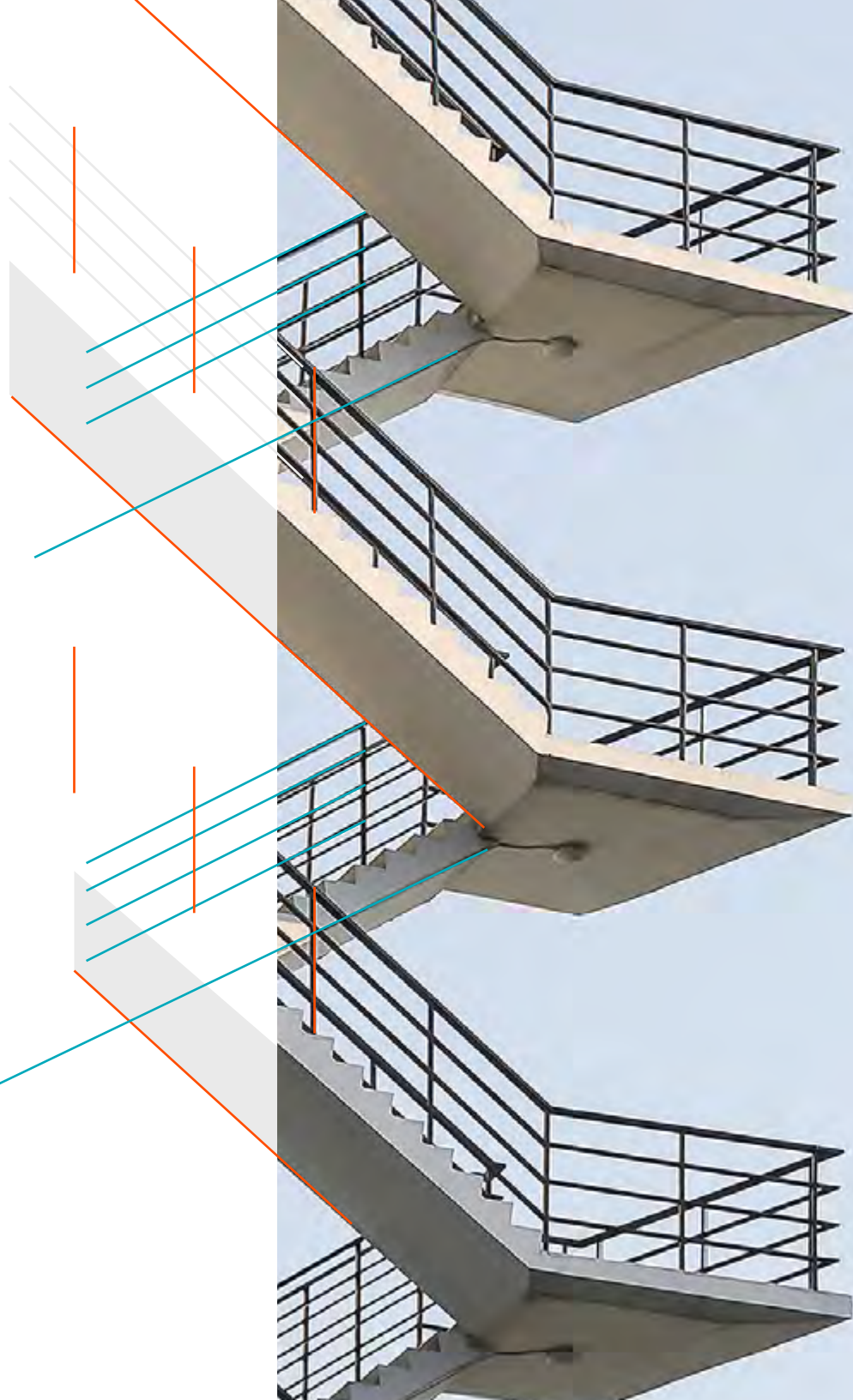
Java 전문가가 툴을 사용할 수 없는 특정 로직에 자유롭게 몰두하게 하는 한편, 기존 ETL, 데이터 품질 및 비즈니스 인텔리전스 전문가의 기술 및 지식을 활용하여 개발 팀의 생산성을 높일 수 있는 툴을 도입하십시오.

또한 Hadoop과 같은 기술은 매일 진화하기 때문에 기본 기술의 꾸준히 변화하는 사양으로부터 보호할 수 있는 추상화 계층을 고려하는 것이 좋습니다.

특히, 필요한 기술은 희소하지만 툴은 항상 사용할 수 있다는 점을 기억하십시오.

올바른 프로젝트 선택

직면한 과제를 바탕으로 이제 조직에
올바른 프로젝트를 선택하는 방법에 대해
알아보겠습니다.



올바른 프로젝트의 형태

귀사가 변화를 갈망하고 업무 방식을 개선하기 위해 포괄적인 데이터 거버넌스 프레임워크를 도입하기로 이미 결정했다면 이 섹션을 건너뛰어도 됩니다.

반면 나중에 사내에서 보다 폭넓게 도입할 수 있는 국지적인 전술적 프로젝트를 고려하고 있다면 계속 읽어보십시오.

올바른 프로젝트에는 다음 네 가지 구성 요소가 있습니다.

1. 증명 가능한 가치

올바른 프로젝트는 IT와 지원하려는 현업 부서 간에 가치를 동일하게 공유할 수 있는 프로젝트입니다. 즉 부서, 비즈니스 단위 또는 그룹에 그들이 확인할 수 있는 명확한 가치를 전달해야 합니다.

2. 후원

프로젝트의 성공에는 비전에 동의하는 임원이 반드시 필요합니다. 빅 데이터 프로젝트에는 여러분의 일을 보호해 줄 고위 후원자와 챔피언이 필요합니다.

따라서 물류 팀을 위한 뛰어난 분석을 구축할 수 있다는 것을 알고 있지만 이에 동의하는 임원이 CMO뿐인 경우에는 다시 생각해야 합니다. 마케팅 팀이 챔피언이면 마케팅 팀의 분석 요구 사항에 맞게 구축해야 합니다. 누구에게도 변화를 강요할 수 없습니다. 영향을 따르고 거기에서 창출할 수 있는 가치를 극대화하십시오.

3. 볼링 핀 효과

첫 번째 기술적 프로젝트의 전략적 중요성이 매우 중요합니다.

빅 데이터가 현업 부서를 도와줄 수 있다는 데 대한 의심의 여지가 없음을 증명할 뿐만 아니라 그 가치를 전사적으로 쉽게 전달할 수 있음을 확인할 수 있기 때문입니다.

따라서 첫 번째 프로젝트를 선택할 때 전략적으로 선택해야 합니다.

마케팅 부서에 대한 빅 데이터의 가치를 증명한 후에는 이러한 증명이 없었으면 동의하지 않았을 수 있는 물류 팀의 동의를 쉽게 얻을 수 있습니다.

4. 전수 가능한 기술

마지막 요점에서 설명한 바와 같이 첫 번째 프로젝트의 가치를 통해 회사의 다른 부서를 설득할 수 있습니다. 이를 위해서는 첫 번째 프로젝트에서 올바른 기술, 기능 및 교훈을 얻을 수 있어야 합니다. 특히 다음 프로젝트로 전수할 수 있도록 이 모든 사항을 문서화해야 합니다. 성공을 목표로 하는 경우 향후 프로젝트를 목표로 삼아야 합니다.

따라서 향후 더 많은 프로젝트를 처리할 수 있도록 확장을 준비해야 합니다. 이는 단순히 클러스터를 확장하는 문제가 아닙니다. 기술 및 운영의 확장을 의미합니다. 더 많은 Java/Hadoop 전문가를 찾거나 이미 보유한 리소스의 활용을 극대화할 방법을 찾아야 합니다.

영향 고려

다음 프로젝트를 선택할 때에도 조직에 미치는 영향을 고려해야 합니다. 추진 중인 빅 데이터 프로젝트가 올바른지 결정하는 데 중요한 역할을 하는 세 가지 광범위한 측면이 있습니다.

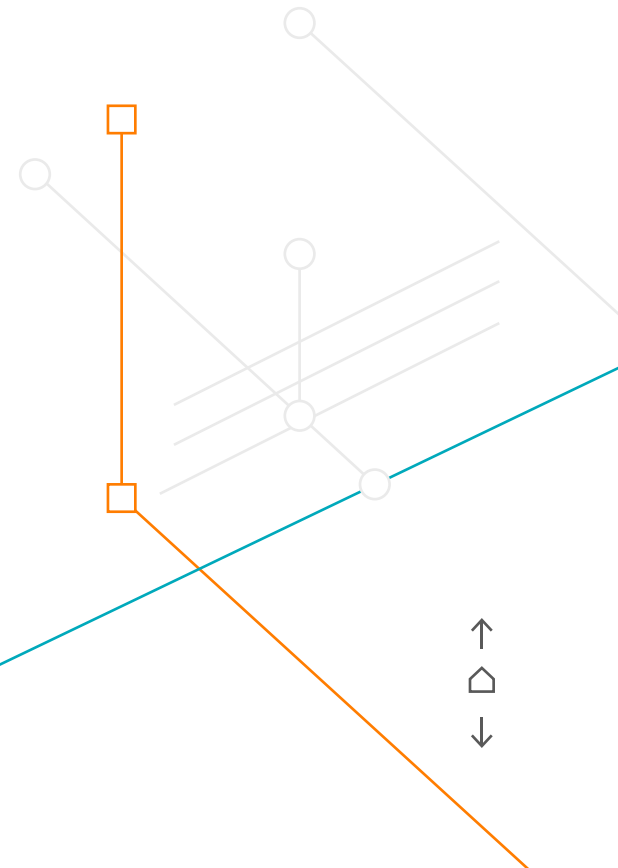
1. 비용 및 분열

가장 기본적인 수준에서 프로젝트 비용은 궤도에 오를 때까지 소요되는 시간 및 금액을 기반으로 합니다. 실제로는 이로 인해 발생할 잠재적 분열도 고려해야 합니다.

경우에 따라 분열은 절차와 관련이 있습니다. 예를 들어 현업에서 해당 데이터를 소유하는 경우 이들은 중앙 집중식 데이터 거버넌스 프레임워크에 제어 권한을 넘기는 것을 달가와 하지 않습니다.

또 다른 경우 분열은 기술과 관련이 있습니다. 예를 들어 새로운 기술을 기존 인프라에 통합하고 이를 위해 기술을 재구성하거나 업그레이드해야 하는 경우가 여기에 해당합니다.

어떤 경우든 분열을 예상하고 인정하며 최소화하거나 불가피한 이유를 설명해야 합니다.



2. 이점 및 영향의 시기

여러 가지 시작 프로젝트를 고려할 때는 자연스럽게 비즈니스 영향 및 개선 효과가 가장 큰 프로젝트 쪽으로 마음이 기울게 되어 있습니다. 하지만 비즈니스 영향의 특성도 고려해야 합니다. 대부분의 가치를 단기간에 전달할지 아니면 장기간에 전달할지 결정해야 합니다.

특히 현업 부서 사용자가 비즈니스 영향을 언제 느끼게 되는지를 고려해야 합니다. 예를 들어 데이터 웨어하우스에 마스터 데이터 관리를 도입하여 비즈니스 인텔리전스의 효율성을 크게 개선할 수 있습니다. 하지만 이 가치는 비즈니스 분석가가 재무 데이터를 다시 정제하지 않아도 된다는 점을 인식한 후에만 느낄 수 있습니다.

3. 리소스 및 제한

이전 두 요소의 분석을 고려하면, 귀하의 재량으로 리소스를 고려해야 합니다. 여기에 대해서는 나중에 좀 더 자세히 알아볼 것이므로 지금은 당연히 투자 효과가 더 큰 프로젝트를 원할 것이라는 점만 기억해 두십시오.

이 목표는 두 가지 방법으로 달성할 수 있습니다. 한편으로는 최대의 비즈니스 영향을 목표로 할 수 있습니다. 하지만 예산을 어떻게 사용할지를 전략적으로 결정해야 합니다. Google과 같은 수준의 데이터 과학자 팀을 구축하고 싶겠지만 정말 그럴 수 있는 여유가 있을까요? 톨과 인력 간의 현명한 선택은 프로젝트의 성공에 매우 중요합니다.

전술적 빅 데이터 프로젝트: 몇 가지 예

빅 데이터는 광범위한 분야에 적용됩니다. 이는 도전할 가치를 느끼게도 하지만 시작할 프로젝트를 잘 모르는 사람에게는 약간의 위협이 되기도 합니다. 다음은 Informatica 고객이 수행한 전술적 빅 데이터 프로젝트의 목록입니다.

조직에서 어떤 프로젝트를 시작해야 하는지 여전히 잘 모를 경우 다음 예를 통해 빅 데이터가 회사에 제공할 수 있는 이점을 생각해 보십시오.

금융

- 위험 및 포트폴리오 분석
- 투자 권유

제조

- 연결된 차량 프로그램
- 예방적 정비

소매

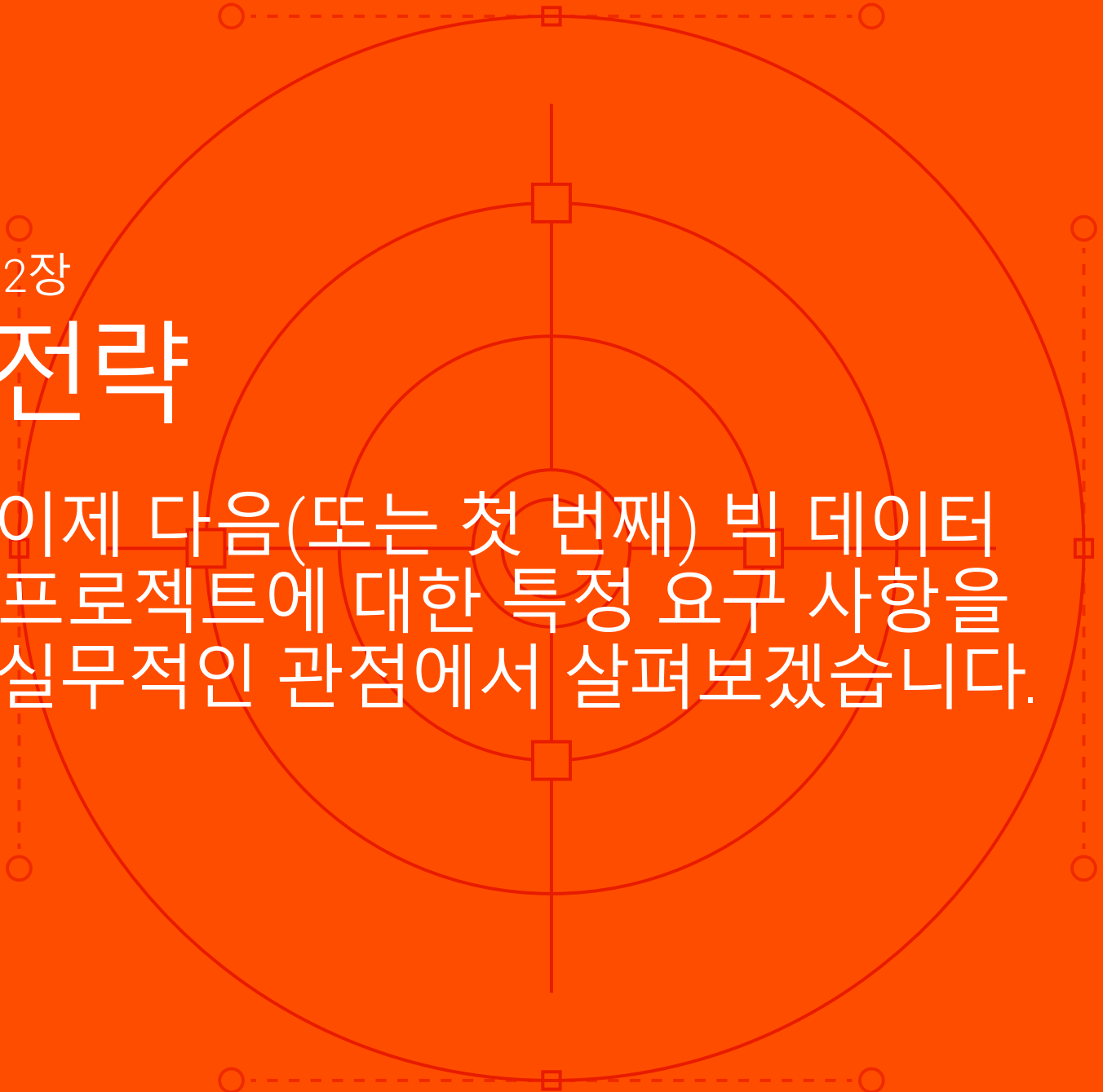
- 사전 예방적 고객 참여
- 위치 기반 서비스

공공 부문

- 의료 보험
- 외환
- 세금 최적화
- 사기 감지

의료

- 환자 결과 예측
- 총 진료 비용
- 약품 조사



2장 전략

이제 다음(또는 첫 번째) 빅 데이터 프로젝트에 대한 특정 요구 사항을 실무적인 관점에서 살펴보겠습니다.

목표 정의

연필을 꺼내십시오. 앞서 살펴본 바와 같이 빅 데이터 프로젝트가 실패하는 가장 큰 원인은 명확한 목표가 없기 때문입니다. 이제 여러분이 염두에 둔 프로젝트에서 모호함을 걷어내 보겠습니다.

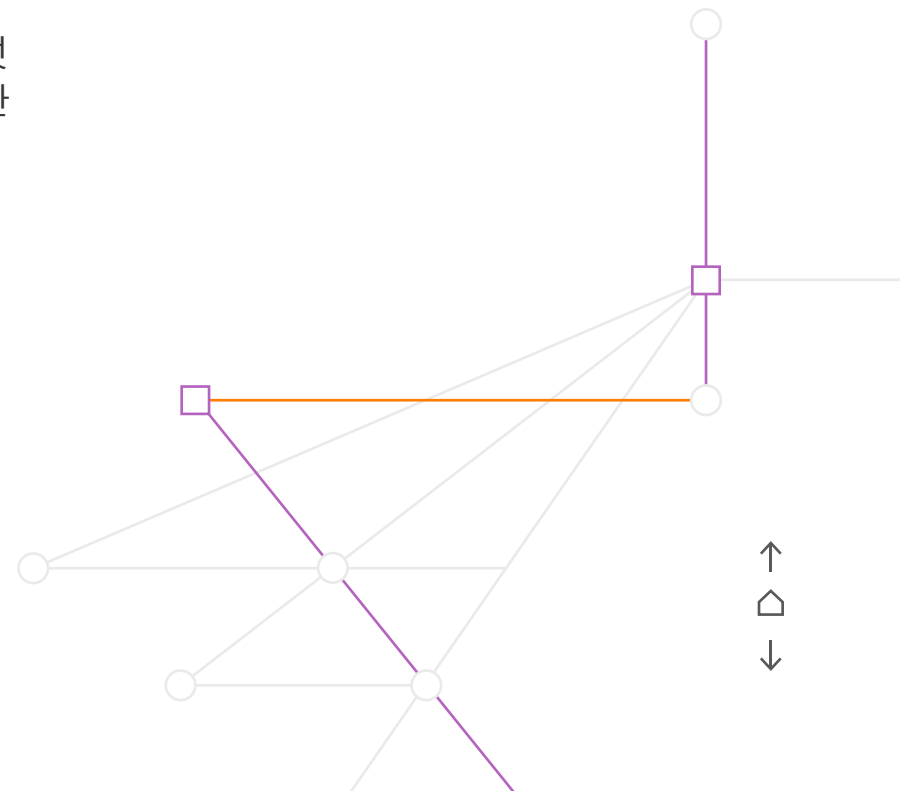
현업 부서 목표

프로젝트가 충분히 인정받으려면
비즈니스 목표가 IT 목표보다
우선해야 하므로 비즈니스부터
시작해 보겠습니다.

비즈니스를 위해 달성할 프로젝트
목표를 적을 때는 가급적 구체적으로
적으십시오. 또한 결과를 측정할
수 있는 목표를 정해야 합니다.

예를 들어 고객 이탈을 예측하는 고객
서비스 부서의 경우 해당 프로젝트의
목표를 '고객 경험 개선'과 같이
모호하게 적어서는 안 됩니다.

목표가 명확할수록 목표에 더
가까이 다가갈 수 있습니다. 다섯
가지 중점 목표가 한 가지 모호한
목표보다 더 가치 있습니다.



현업 부서 목표

현업 부서 및 현업 부서 사용자와 관련된 빅
데이터 프로젝트의 목표를 중요도 순으로
나열하십시오 (더 많거나 적은 목표를 자유롭게
입력).

예: 고객 이탈 감소

각 목표를 달성할 최소 및 최대 기간을 적으십시오.

예: 3~6개월

이제 각 목표에 대해 목표 달성 여부를 결정하는 데 사용할 수 있는 성공의 척도를 적으십시오. 사용 가능한 지표 또는 그 계산인 경우가 가장 좋습니다.

예: 월 평균 이탈률을 X% 만큼 감소

빅 데이터 프로젝트에 소요되는 시간

빅 데이터 프로젝트는 완전한 가치를 전달하는 데 시간이 필요한 만큼 오래 걸립니다. 경험상, 프로젝트 범위는 기간을 나타냅니다.

Informatica는 3개월 이내에 전문적인 프로젝트를 제공한 고객과 함께 일한 적이 있으며, 기본 프로그램을 제공하는 데에만 3년이 걸린 고객과 함께 일한 적도 있습니다.

장기적인 프로젝트의 경우 6개월마다 프로젝트의 가치를 입증하는 것을 목표로 삼아야 합니다. 프로젝트에 대한 민첩한 접근 방식을 도입하면 여러 단계 및 주요 일정을 보다 작은 규모의 프로젝트로 제공하는 데 도움이 됩니다.

한 가지 분명한 사실은 기간을 추측해서는 안 된다는 점입니다. 여러분의 경험과 이전에 유사한 프로젝트를 수행한 다른 사람의 경험을 바탕으로 완료 기간을 계산하십시오. 누구에게 도움을 요청할지 모를 경우 언제든지 당사료 문의하십시오.

IT 목표

이제 프로젝트와 관련된 IT 목표에 대해 살펴보겠습니다.

(프로젝트가 IT의 효율성 또는 속도를 개선하기 위한 것이라면 현업 부서 사용자에게 이러한 점을 설득시키기 어렵습니다. 따라서 현업 부서 사용자가 이미 관심을 가지고 있는 목표와 함께 IT의 목표를 전달해야 합니다.)

IT와 관련된 빅 데이터 프로젝트의 목표를 중요도 순으로 나열하십시오 (더 많거나 적은 목표를 자유롭게 입력).

예: 집계된 고객 데이터, 신용 카드 사용 데이터, 소셜 그래프 데이터 및 이탈 지표의 실시간 수집, 정제, 마스터 및 저장을 위한 프로세스 수립

증지, 협업 및 경청

이 워크북은 현업에서 일하든 IT에서 일하든 상관없이 빅 데이터 프로젝트를 시작할 수 있도록 도와주기 위해 작성되었습니다. 어느 분야에서 일하든 목표를 어림짐작으로 설정하지 마십시오. 목표 설정에 대한 특정 지침이 필요한 경우 필요한 전문성을 가진 파트너와 함께 지금 협력하십시오.

전략적 협업 없이는 프로젝트를 성공으로 이끌 수 없습니다.

IT 목표

각 목표를 달성할 최소 및 최대 기간을
적으십시오.

예: 2~4개월

이제 각 목표에 대해 목표 달성 여부를 결정하는
데 사용할 수 있는 성공의 척도를 적으십시오.
사용 가능한 지표 또는 그 계산인 경우가 가장
좋습니다.

예: 정확한 예상 이탈률 X%

데이터 요구 사항 정의

빅 데이터 노력의 구체적인 목표를 살펴보았으므로 이제 프로젝트의 핵심 요소인 데이터 자체에 대해 알아보겠습니다. 어떤 프로젝트든 필요한 정보, 이러한 요구에 적합한 데이터 세트, 데이터를 가져올 방법 및 데이터를 사용할 방법에 대해 전략적으로 생각해야 합니다.

필요한 데이터가 무엇입니까?

먼저 빅 데이터 프로젝트의 가장 기본적인 목적(조직에 제공하려는 정보)을 살펴보겠습니다. 다음 질문에 가능한 한 구체적으로 답변하십시오.

앞서 설명한 현업 부서 목표를 달성하기 위해
현업 부서 사용자가 정보를 바탕으로 한 의사
결정을 위해 알아야 할 사항은 무엇입니까?

예: 이탈할 가능성이 있는 가장 중요한 고객 및 이탈과 연관이 있는 행동

이러한 지식을 전달하기 위해 사용할 수 있는 데이터는 무엇입니까?

예: 고객 구매 기록, 리뷰 데이터, 구매율, 해지율, 이탈률, 고객 서비스 품질

이러한 데이터 세트는 어떤 소스 시스템에 포함되어 있습니까?

예: 고객 서비스 레코드, 제품 성능 지표, 고객 활동 데이터베이스, 고객 마스터 데이터 관리

이미 언급된 데이터 외에 분석을 위한 컨텍스트 또는 추가 가치를 제공할 수 있는 다른 정보가 있습니까?

예: 고객 서비스 설문 조사 데이터, 경쟁업체 분석, 기상 데이터, 소셜 데이터

현재 액세스 권한이 없는 데이터 세트 중 추가 컨텍스트 데이터를 포함하고 있을 수 있는 데이터 세트는 무엇입니까?

예: 타사 소셜 데이터, 타사 시장 데이터, 기상 데이터

[illegible]

다크 데이터(Dark Data) 찾기

액세스할 수 없는 데이터 세트를 고려할 때 조직 외부의 데이터로 한정하지 마십시오. Gartner에서는 대부분의 기업이 조직 내 데이터의 15%만 사용한다는 사실을 발견했습니다. 또한 데이터 웨어하우스 활용도에 대한 통계 분석을 수행하는 회사인 Appfluent는 데이터 웨어하우스에 있는 30~70%의 데이터가 유틸 상태에 있다는 사실을 발견했습니다. 나머지는 액세스하기 힘들거나, 사용하는 데 비용이 많이 들거나, 찾기 어려운 사일로, 레거시 아카이브 및 데이터 저장소에 숨어 있습니다. 이는 이 모든 데이터를 저장하기 위해 이미 비용을 지불하고 있기 때문에 문제가 되는 것입니다.

따라서 필요한 데이터를 찾을 때 조직에 이미 있는 데이터부터 찾는 것이 좋습니다.

데이터에 대한 5가지 주요 고려 사항

원하는 데이터를 정의하고 나면
특정 빅 데이터 과제를 보다
명확하게 이해할 수 있게 됩니다.

특히 프로젝트를 진행하기 전에
고려해야 하는 5가지 주요 요소가
있습니다. 이러한 요소는 각
데이터 세트뿐만 아니라 빅 데이터
데이터 세트에 대해 수행해야
하는 작업을 알려 줍니다.

1. 볼륨 준비

필요한 데이터의 큰 규모를 처리할 준비를
갖추어야 합니다. 다양한 차원에 걸쳐 값(즉,
고객 트랜잭션), 사용량(액세스 빈도), 크기(GB,
TB), 복잡성(기기 데이터, 관계형 데이터, 동영상
등), 액세스할 수 있는 사람(데이터 과학자만
또는 일반 현업 부서 사용자) 등에 따라 데이터를
분류하십시오.

철저히 구성된 데이터 저장소는 모든 데이터를
관리할 방법을 결정하는 데 도움이 됩니다. 현재
스토리지 및 처리 용량을 평가하고 이를 확장
가능하도록 만들 수 있는 가장 비용 효율적인
방법을 찾아보십시오.

2. 다양한 고려

빅 데이터의 가장 까다로운 문제는 분석 시
조정해야 하는 형식 및 구조의 다양성입니다.
새로운 데이터 유형 및 구조(소셜, 센서, 동영상)를
이미 사용 중인 소스(관계형, 레거시 메인프레임)와
합치려면 여러 소스를 통합해야 합니다.

모든 단일 통합을 수동으로 코딩하는 작업은
모든 시간과 리소스를 소비할 수 있을 정도로
번거로운 작업입니다. 데이터 통합 및 데이터 품질
툴을 최대한 활용하여 보다 중요한 작업에 대한
프로세스를 가속화하십시오.

3. 속도 처리

실시간 스트리밍 데이터와 과거 데이터의 조합은 일반적으로 분석의 예측력을 강화합니다. 따라서 원하는 데이터 중 일부는 시스템에 지속적으로 유입될 경우에만 가치가 있을 수 있습니다.

실제로 대부분의 실시간 분석은 다양한 소스에서 여러 형식으로 제공되는 스트리밍 데이터를 필요로 합니다. 스트리밍 분석 기술 및 모든 데이터를 관리할 수 있는 논리적 인프라로 프로젝트를 준비하십시오.

4. 정확성 보장

분석이 아무리 중요해도 데이터의 신뢰성을 합당하게 기대할 수 없으면 아무 소용 없습니다. 분석하는 데이터가 많을수록 높은 수준의 데이터 품질을 유지해야 중요성이 커집니다.

데이터를 목적에 맞게 활용하려면 데이터가 사용되는 목적을 알아야 합니다. 데이터 과학자가 집계된 고객 데이터에서 패턴을 찾고자 한다면, 필요한 준비가 많지 않을 것입니다. 반면, 재무 보고 및 공급망 데이터는 정확성 및 규정 준수를 위해 고도로 정제되고 인증되어야 합니다.

원시 데이터부터 정제되고 신뢰할 수 있는 데이터의 고도로 마스터 데이터화된 데이터 저장소에 이르기까지 필요한 준비 작업의 양에 따라 카테고리를 만드십시오.

5. 규정 준수 고려

여러분이 다루는 다양한 데이터 세트는 여러 보안 규정 및 요구 사항을 필요로 합니다. 각 데이터 세트에 대한 보안 정책에 따라 데이터를 익명화하기 위해 수행할 작업을 고려해야 합니다.

데이터의 양은 수백 개의 데이터 저장소에서 전사적으로 급증할 것입니다.

민감한 데이터가 어디에 있는지를 이해하고 암호화를 통해 소스에서 보안을 유지한 다음 데이터에 액세스할 수 있는 사람을 제어해야 합니다.

보안을 뛰어넘어, 민감한 데이터의 스마트 아카이빙은 데이터가 마이그레이션되거나 개발 및 테스트 환경에 적용될 때마다 미리 정의된 규칙으로 데이터를 마스킹합니다.

이러한 5가지 고려 사항을 여러분이 다루는 모든 데이터 세트에 적용하면 보다 현실적인 방법으로 빅 데이터 문제에 대비할 수 있게 됩니다.

3장

린(Lean) 빅 데이터 공급망

기존의 비즈니스 인텔리전스 및 데이터 웨어하우스 방법은 빅 데이터 이니셔티브의 요구 사항을 충족하도록 확장되지 않습니다. 이제 팀, 프로세스 및 인프라를 확장할 수 있는 방법을 살펴보겠습니다.

팀

빅 데이터 팀은 가장 큰 도전과 가장 큰 기회를 앞두고 있습니다. 비즈니스 목표를 이해하는 사람과 기술적 요구 사항을 이행할 수 있는 사람 간에 균형을 잘 유지해야 합니다.



1. 직원을 채용한 기술을 사용하십시오.

데이터 과학자 및 계량 분석가를 채용할 때 기업이 저지르는 가장 중대한 실수 중 하나는 이들에게 곳은 일을 시킨다는 것입니다. 역량이 가장 뛰어난 리소스가 핸드 코딩 데이터 통합 및 데이터 정제에 모든 시간을 허비한다면 단순히 이들을 실망시키는 것에 그치지 않고 찾기 어려운 기술을 활용하지 못하는 결과가 발생합니다.

찾기 힘든 기술력을 가진 사람은 해당 기술이 실제로 필요한 업무에 주력하도록 해야 합니다. 여러분은 가장 뛰어난 인재가 퇴사하는 것을 원치 않으며, 이들이 대신 툴을 통해 쉽게 수행할 수 있는 작업에 시간을 허비하는 것을 원치 않을 것입니다.

2. 팀 구성에 대해 전략적으로 생각하십시오.

제대로 진행된다면 프로젝트는 범위와 리소스의 두 가지 면에서 모두 성장할 것입니다. 필요한 기술을 가진 사람은 한정되어 있으므로 이제 전략적으로 생각하여 특정 프로세스를 신속하게 확장할 수 없는 냉엄한 현실을 깨달아야 합니다.

프로젝트의 범위가 커질 경우 요구 사항을 충족하기 위해 적시에 찾아야 하는 기술은 무엇입니까? 예를 들어 데이터 과학자는 개발자보다 찾고, 교육하고, 채용하기가 훨씬 더 어렵습니다.³

팀의 균형이 매우 중요합니다. 힘들게 얻은 데이터 관리 경험과 새로운 툴의 학습에 대한 열정이 적절히 조화되어야 합니다. 또한 기술적 역량이 뛰어난 사람과 올바른 모델을 구축하는 데 필요한 분야의 전문 지식을 가진 사람 간의 균형도 유지해야 합니다.

3. 조기에 프로젝트 목표를 조정한 다음 전달하십시오.

기업이 신입 직원을 채용할 때 저지르는 가장 일반적인 실수 중 하나는 프로젝트의 진정한 목표를 전달하는 것을 잊어버리는 것입니다. 첫 면접부터 실제 업무까지 현업 사용자에게 제공하려는 것이 무엇인지 명확히 이해해야 합니다. 경영진 후원을 활용하여 사명을 전파하고 성공 사례 및 문제점을 공유하십시오.

프로젝트의 비즈니스 가치를 명확히 이해하지 못하면, 신입 직원이 프로젝트의 IT 목표에 대해서만 생각하면 되는 것으로 착각할 수 있습니다.

³ ["Big Data's High-Priests of Algorithms,"](#) Wall Street Journal, 2014년 8월 8일

4. 팀이 성장하면 이들을 관리해야 하는 필요성도 증가합니다.

구축 및 구현한 다음 객관적인 방식으로 통합할 수 있는 새로운 기술과 달리, 신입 직원은 그들이 어디에서 어떤 일을 왜 하고 있는지 잘 알아야 합니다. 여러분이든 다른 사람이든 누군가는 새로운 팀에 필요한 관리 문제를 담당해야 합니다.

문화 및 화합력과 같은 요소를 과소평가해서는 안 됩니다. 신입 직원을 프로세스에 어떻게 통합할지 장시간 진지하게 고민해야 합니다. 여러분은 이들에게 기술을 가르치지 못할 수 있겠지만 분명 더 나은 팀 구성원이 되도록 도와줄 수는 있습니다.

5. 팀이 정체해 있을 여유가 없습니다.

빅 데이터 기술은 매일 새롭게 등장하고 있습니다. 또한 기존 기술도 급속히 진화하고 있습니다. 이는 베스트 프랙티스를 조기에 도입할 용기가 있는 기업에는 엄청난 기회입니다. 또한 경쟁업체보다 앞서 가기 위해 도전 과제를 정의해야 한다는 것을 의미합니다.

직원들은 세상이 변하는 속도 만큼 빠르게 기술을 진화시켜야 합니다. 좋은 소식은 앞서 도전하는 것보다 큰 동기를 부여하는 것은 없다는 것입니다. 문제는 역량을 지속적으로 개발하는 데 필요한 교육과 토론을 어떻게 제공하느냐입니다.

전략적 사고의 중요성

반복해서 선택해야 하는 중요한 사항은 기능을 구축할 때 자동화된 툴을 사용할지 또는 수동 통합을 사용할지입니다.

핸드 코딩은 구축할 대상에 대한 완전하고 정확한 제어 권한을 제공합니다. 이는 아직 가능하지 않은 방식으로 메타데이터를 추출하기 위해 복잡한 스크립트를 작성하는 경우 등에 중요하고 필요한 경우가 많습니다.

반면, 툴은 뛰어난 민첩성과 동일한 프로세스를 지속적으로 반복할 수 있는 기능을 제공합니다. 데이터 통합 및 데이터 품질과 같은 작업의 경우 이는 고급 분석가 및 과학자가 굳은 일을 하지 않도록 해주므로 매우 유용합니다.

리소스에 대한 현실을 직시해야 합니다. Google처럼 크고 뛰어난 팀을 구축할 수 없다면 부족한 리소스를 낭비하지 마십시오.

데이터 거버넌스 설정

빅 데이터에 대한 보다 기초적인 노력을 기울인다면 데이터 거버넌스에 대한 절차적 프레임워크를 수립해야 합니다. 실제로 빅 데이터 프로젝트가 단일 부서에 가치를 제공하는 것을 목표로 하는 경우에도 소규모 데이터 거버넌스 위원회를 구성하여 이러한 기구에서 겪는 고유한 문제에 대처하는 방법을 습득할 수 있도록 하는 것이 좋습니다.

기본적으로 데이터 거버넌스 위원회는 기업의 데이터 접근 방식을 감독하기 위해 임원들로 구성되는 정식 기구입니다. 하지만 여기에는 특정 비즈니스 단위의 데이터를 관리하는 직능 또는 부서별 담당자인 데이터 관리자도 포함됩니다.

실제로 일부 고객은 데이터 도메인에 따라 데이터 관리 역할을 할당합니다. 즉 제품 데이터, 고객 데이터 등에 대한 담당자를 별도로 두고 있습니다.

데이터 거버넌스 프레임워크가 단점보다 장점이 많은지 확인하는 프로세스를 생성하는 것을 목표로 삼아야 합니다. 모든 사람이 같은 기간 내에 동일한 목표를 달성하는 데 전념하도록 하여 관리 부담이 발생하지 않도록 적극적으로 노력해야 합니다.

데이터 거버넌스 프레임워크는 다음 5가지 특성을 목표로 해야 합니다.

1. 직능 간

유사한 역할을 가진 여러 사람으로 구성된 데이터 거버넌스 위원회는 비효율적입니다. 빅 데이터 프로젝트에 참여하는 각 현업 부서 단위의 고유한 관점 및 요구 사항을 대표할 수 있는 기구를 만들어야 합니다.

2. 소통적

직능, 부서 및 도메인 간의 원활한 커뮤니케이션 없이는 프로젝트가 관료주의에 빠지고 오해를 불러일으킬 수 있습니다. 이러한 일은 굉장히 자주 발생합니다. 모든 우려 사항을 불식시키거나 적절히 해결해야 합니다.

3. 효율적

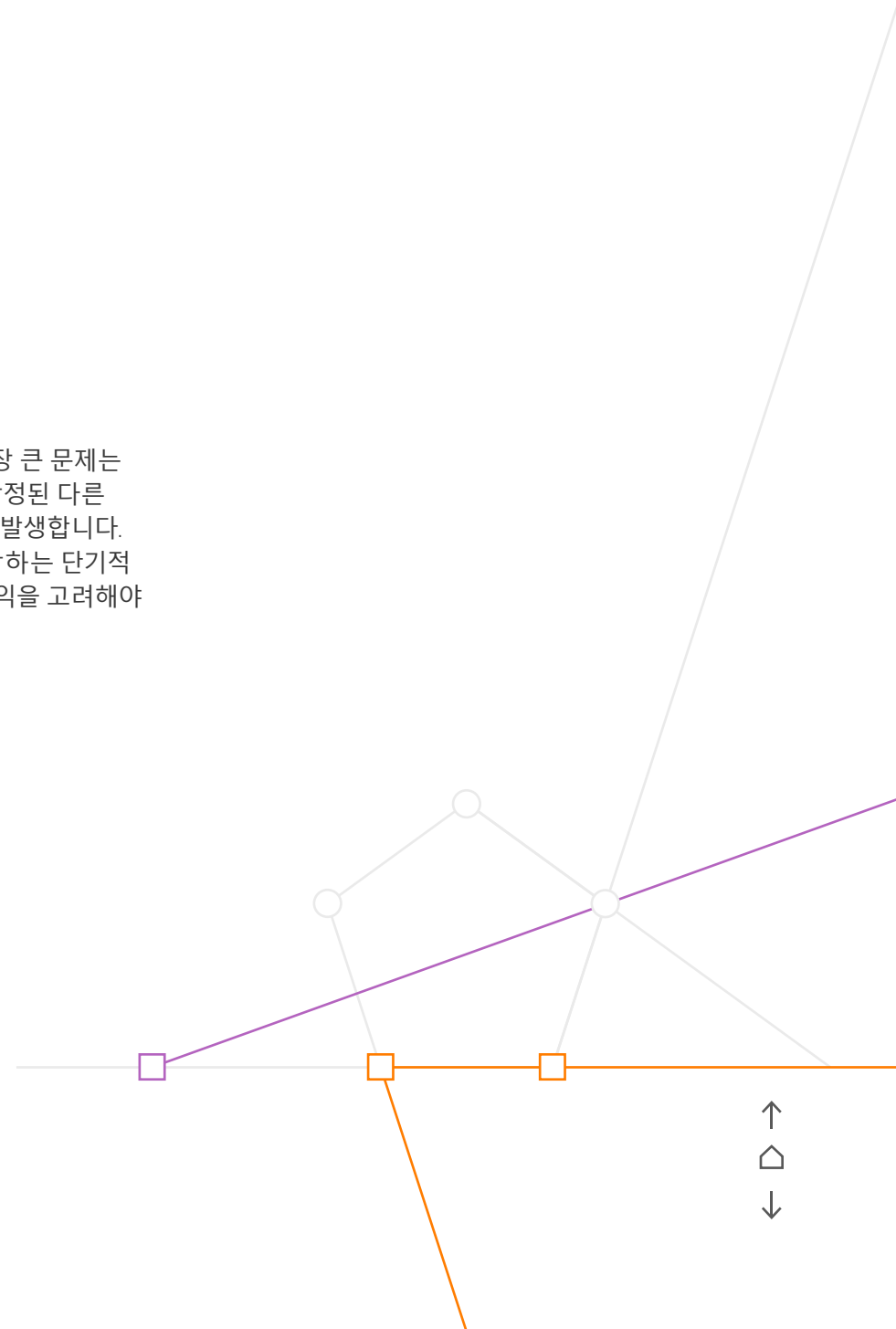
직능 간 프로세스는 장벽처럼 느껴져서는 안 됩니다. 빅 데이터 프로젝트의 성공은 뛰어난 민첩성을 요구합니다. 따라서 가능한 경우 자동화 및 예외 보고 규칙을 만들고 협업 툴을 채택하여 커뮤니케이션을 개방적이고 편리하게 유지해야 합니다.

4. 헌신적

프로젝트의 주요 목표를 효과적으로 전달하고 데이터 거버넌스 프레임워크에 참여한 모든 사람이 이러한 목표에 전념하도록 해야 합니다. 공동의 목표는 거버넌스 사고 및 의사 결정을 이끌어야 합니다.

5. 중앙 집중식

데이터 거버넌스 프레임워크의 가장 큰 문제는 한 부서 단위의 목표를 위원회에 상정된 다른 목표보다 우선시해야 하는 경우에 발생합니다. 의사 결정은 한 부서 단위에서 생각하는 단기적 이익보다 전체 위원회의 장기적 이익을 고려해야 합니다.

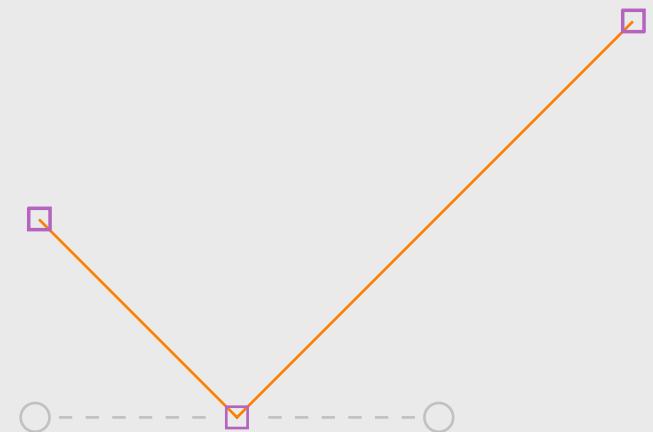
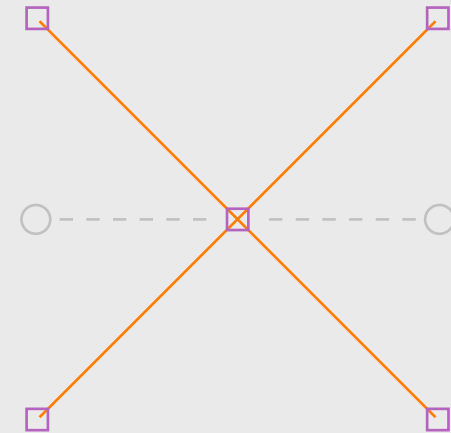


필요한 기술과 소유한 기술

다시 연필을 꺼낼 시간입니다. 이제 새 팀이 겪게 될 여러 가지 주관적 위험 및 기회를 확인했으므로 이 팀이 실제로 어떤 모습일지 알아보겠습니다.

다음 페이지에는 고객이 실제 채용했었던 직무에 따라 빅 데이터 역할이 나열되어 있습니다. 현재 가용할 수 있는 직원 및 프로젝트에 소요될 것으로 예상되는

시간([23페이지부터 시작되는 섹션에 입력한 시간](#))에 따라 채용해야 하는 인원을 적어 보십시오.



역할	이미 이 역할을 수행할 수 있는 사람이 있습니까?	이 역할을 위한 채용 필요성	주어진 시간에 따라 채용해야 하는 인원
데이터 과학자			
도메인 전문가			
비즈니스 분석가			
데이터 분석가			
데이터 엔지니어			
데이터베이스 관리자			
엔터프라이즈 아키텍트			
비즈니스 솔루션 아키텍트			
데이터 아키텍트			
데이터 관리자			
ETL(데이터 통합) 개발자			
애플리케이션 개발자			
대시보드 개발자			
통계 모델러			
기타			
기타			
기타			
기타			
기타			

통합적 사고의 필요성

새로운 팀 구성원을 찾을 때 적절한 자격을 갖춘 사람으로 제한하지 마십시오. 적절한 자격을 갖춘 사람을 찾는 것 자체가 분명히 하나의 도전 과제입니다. 비즈니스 목표와 기술적 역량을 조합할 의지가 있는 사람도 찾아야 합니다.

빅 데이터 프로젝트에 참여하는 사람들이 비즈니스의 현실을 이해하고 복잡한 데이터 과학을 이행할 수 있는 것이 얼마나 중요한지를 강조하는 고객이 점점 늘어나고 있습니다. 이러한 유형의 통합적 사고는 거대하고 찾기 어렵습니다. 이는 교육하고 보상할 가치가 있습니다.

빅 데이터 툴 이해

경험상, 다음 툴은 빅 데이터 프로젝트에 필요한 아키텍처에 필수적입니다(아키텍처에 대한 자세한 내용은 뒷부분에서 설명). 물론 특정 프로젝트에 필요한 기술 혼합은 목표 및 리소스에 따라 결정됩니다.

분석

분석할 도메인에 대한 통찰력, 패턴, 예측 및 계산으로 원시 데이터를 전환하는 툴 및 프로세스입니다.

시각화

데이터 및 결과물을 습득하고 이해하기 쉬운 방식으로 제공할 수 있습니까?

고급 분석

데이터 세트에 첨단 분석 알고리즘을 적용하여 복잡한 계산을 수행할 수 있습니까?

기계 학습

정교한 기계 학습 알고리즘을 적용하여 수동 작업으로 커버할 수 없는 수준의 패턴을 식별하고 예측도 할 수 있습니까?

이러한 툴 및 기술 중에서 데이터 통합, 데이터 품질 및 마스터 데이터 관리와 같은 일부 툴은 다시 구축할 가치가 없을 정도로 빅 데이터 프로젝트에 완전 기본적인 것입니다. 툴을 사용하지 않고 이러한 기능을 직접 구축하기 위해 소중한 기술과 인력을 사용하지 마시기 바랍니다.

프로젝트의 목표를 기억해 보십시오. 사용자 지정 구축은 여기에 포함되지 않습니다.

프로세스

빅 데이터를 처리해야 하는 실제 프로세스에 대해 살펴보겠습니다. 프로세스는 목표 및 요구 사항에 고유해야 하지만 이 섹션에서는 기대할 사항과 학습할 사항에 대한 개요를 제공합니다.



빅 데이터 8단계

경험상, 애자일 방법론이 빅 데이터 프로젝트에 매우 적합한 접근 방식이라고 할 수 있습니다. 이는 기대치를 관리하고, 실수로부터 학습하며, 최상의 프로세스를 얻기 위한 과정을 반복할 수 있도록 해줍니다. 즉 프로젝트의 접근 방식은 전적으로 여러분과 여러분의 상황에 달려 있습니다.

어떤 경우든 다음 8단계는 빅 데이터 공급망에 매우 중요합니다. 그러나 이러한 단계에 대한 효과적인 프로세스를 수립해야 합니다.

1. 데이터 액세스

첫 번째 과제는 필요한 모든 데이터를 획득하는 것입니다. 경우에 따라 여기에는 스트리밍 데이터 캡처가 수반될 수도 있고, 데이터베이스에서 데이터를 추출해야 하는 경우도 있습니다. 이 데이터를 사용하려는 방식에 따라 저장하려면 반복 및 관리 가능한 프로세스를 수립해야 합니다.

2. 데이터 통합

빅 데이터의 가장 복잡한 과제는 다양한 데이터 구조 및 형식과 관련이 있습니다. 분석을 지속적으로 수행하려면 이 모든 데이터를 통합하고 정규화하는 프로세스를 수립해야 합니다. 가급적 수동 처리를 최소화하는 것이 가장 좋습니다.

3. 데이터 정제

분석의 신뢰도를 높이려면 데이터를 정제하여 중복, 오류, 부정확성 및 불완전한 데이터를 제거해야 합니다. 프로세스를 통해 가장 역량 있는 분석가 및 과학자가 귀중한 일에 모든 시간을 투입하지 않도록 해야 합니다.

4. 마스터 데이터 관리

정제된 통합 데이터의 신뢰할 수 있는 소스를 유지 관리하는 한 가지 방법은 데이터 마스터 프로세스를 수립하는 것입니다. 이는 도메인(예: 제품, 고객 등)별로 구성되고 빅 데이터 통찰력으로 보완된 통합 데이터 컬렉션을 생성하여 다른 모든 시스템에 제공하기 위한 것입니다.

5. 데이터 보안

이 단계에서는 두 가지 기본 프로세스를 수립합니다. 첫 번째는 각 데이터 세트에 필요한 보안 규칙 및 프랙티스를 정의하는 것입니다. 두 번째는 민감한 데이터를 감지하여 영구적이거나 동적인 방식으로 마스킹함으로써 이러한 규칙 및 프랙티스가 지속적으로 적용되도록 하는 것입니다.

6. 데이터 분석

분석 프로세스는 분석가, 분석 툴 및 목표와 관련된 요구 사항에 따라 달라집니다. 시간이 지나고 경험이 쌓이면서 이 프로세스의 효율, 속도, 비용 및 확장성이 점점 개선되기를 원하므로 여기에서는 반복적 조사 및 지속적 개선에 대한 사고 방식이 매우 중요한 역할을 합니다.

7. 비즈니스 요구 사항 분석

이 단계는 매우 중요하지만 거의 항상 간과됩니다. 데이터를 분석하는 동안에도 현업 부서 요구 사항 분석에 대한 명확한 프로세스를 수립하십시오. 이는 현업 부서의 맥박을 느끼지 못하면 노력이 헛되고 비즈니스 영향이 최소화되는 위험이 발생하기 때문에 중요합니다.

8. 통찰력 운용

이 워크북의 앞부분에서 설명한 바와 같이 빅 데이터 프로젝트의 현업 부서 영향력을 느껴야 합니다. 해답을 찾은 경우 이에 대한 자동화된 파이프라인을 생성하여 이를 가장 필요로 하는 현업 부서 사용자에게 제공하십시오. 예를 들어 이탈할 가능성이 높은 고객에 대한 데이터는 대시보드를 통해 고객 서비스 상담원에게 제공해야 합니다. 또한 피드백 루프를 통합해야 하며, 통찰력을 받는 방법을 확인해야 합니다.

문서화의 중요성

이러한 8단계를 마스터하면 빅 데이터 프로젝트가 올바른 방향으로 추진됩니다. 목표는 명확하고, 반복 가능하며, 확장 가능하고, 지속적으로 개선되는 프로세스를 수립하는 것입니다. 이를 위해서는 이러한 프로세스와 이후의 개선 사항을 문서화하는 것이 팀에 매우 중요합니다.

빅 데이터 프로젝트의 기술, 가능성 및 교훈은 전수 가능하고 자주 커뮤니케이션되어야 합니다.

아키텍처

효과적인 린(Lean) 빅 데이터 공급망을 유지하려면 아키텍처를 탄탄하고 전략적으로 구축해야 합니다. 이 섹션에서는 이상적인 빅 데이터 아키텍처의 형태 및 단계별 접근 방식으로 아키텍처를 구축하는 방법에 대해 알아봅니다.



첫 번째 단계: 샌드박스(sandbox)

빅 데이터 프로젝트를 위한 아키텍처 구축을 시작할 때 가장 논리적인 출발점은 테스트 데이터를 사용하여 아키텍처의 실현 가능성을 확인할 수 있는 샌드박스 개발 환경을 설정하는 것입니다. 이 과정에서 고려해야 하는 교훈은 다음과 같습니다.

작게 시작하십시오

완벽하게 제어할 수 있는 잘 정의된 샌드박스로 시작하면 가장 성공적인 구현으로 가는 과정을 반복할 수 있습니다. 가능한 한 빨리 가동 및 실행하고 각 반복을 통해 얻은 교훈을 문서화하십시오.

규모가 중요합니다

샌드박스과 실제 구현 간의 중요한 차이점은 운영계 환경이 훨씬 더 커질 것이라는 사실입니다. 따라서 성과를 수집, 통합, 정제 및 배포하기 위한 자동화된 처리가 필요합니다. 또한 실시간 운영계 환경을 신뢰할 수 있고 유연하게 유지하려면 훨씬 더 강력한 구조와 검증된 구성 요소 및 프로세스가 필요합니다.

테스트 전에 마스킹하십시오

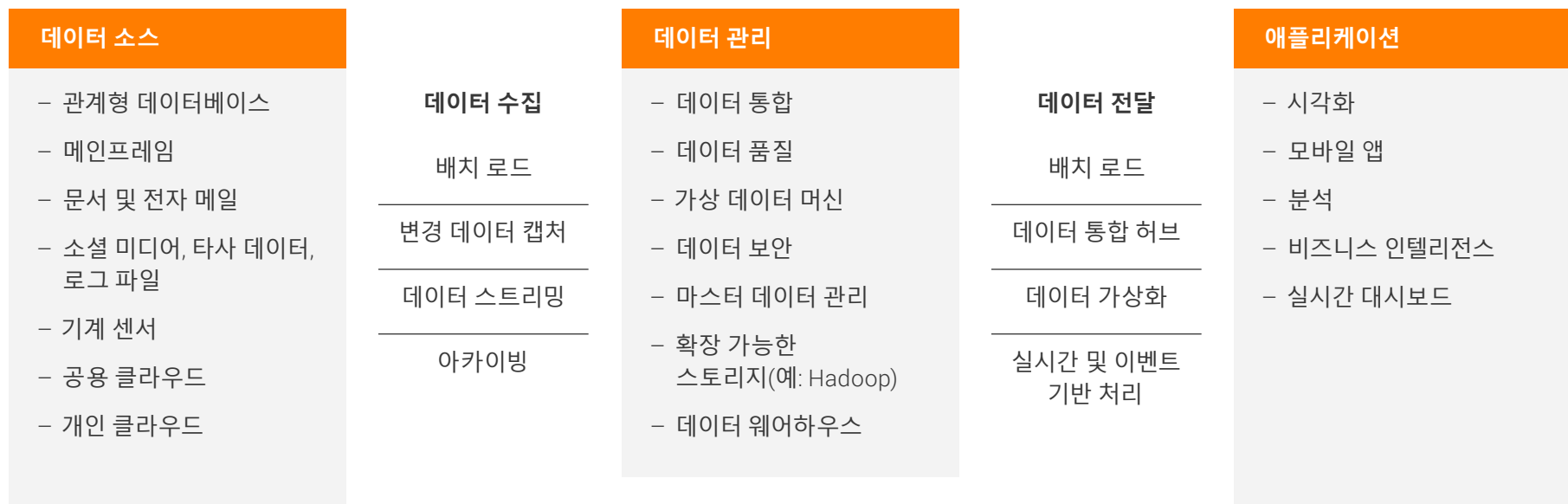
조직에서 테스트 데이터를 사용하는 경우 일반적으로 실시간 운영계 데이터의 변형을 사용하여 형식 및 구조가 실제 환경을 반영하게 합니다. 하지만 이 데이터를 적절히 마스킹하지 못하면 완전히 불안정한 테스트 환경에 민감한 데이터가 노출된 상태로 남겨질 수 있습니다.

변환 중에 손실되지 않도록 하십시오

빅 데이터 프로젝트의 과도한 비용 지출 및 소모적 지연의 가장 일반적인 원인 중 하나는 샌드박스에서 놓친 핸드 코딩 오류가 아키텍처 운영을 시작할 때 다시 발생한다는 사실에서 비롯됩니다. 따라서 아키텍처의 중요한 부분을 핸드 코딩할 경우 운영계 수준의 요구 사항을 충족하려면 많은 코드의 구조를 재조정해야 한다는 점을 예상하고 그에 따라 기대치를 관리해야 합니다. 하지만 생산성 및 자동화 툴을 사용하면, 코드의 구조를 재조정해야 하는 필요성 및 오류를 사전에 방지할 수 있습니다.

이상적인 빅 데이터 아키텍처

다음 다이어그램에서는 이상적인 빅 데이터 기술 및 프로세스 아키텍처를 구축하는 데 권장되는 방법을 보여 줍니다.



프로젝트 계획

이제 빅 데이터 전환의 모든 면을 분석했습니다. 다음 단계에서는 이 프로젝트를 계획을 기본 가이드로 사용하여 처음부터 구현까지 빅 데이터 프로젝트의 모든 과정을 관리해야 합니다.



프로젝트 계획

이 프로젝트 계획 템플릿을 프레임워크로 사용하여 빅 데이터 프로젝트의 세부 사항 및 여러 요소를 문서화하십시오. 그런 다음 편집된 문서를 사용하여 조직의 나머지 구성원으로부터 필요한 동의를 얻을 수 있습니다. 이는 외부 파트너에 접근할 때도 도움이 됩니다.

1단계: 전략

현업 및 IT의 목표 확인
성공의 척도 정의

2단계: 데이터

필요한 정보 확인
데이터 및 데이터를 전달할 소스 확인

3단계: 공급망

인력

- 필요한 기술 평가
- 소유한 기술 평가

프로세스

- 데이터 액세스
- 데이터 통합
- 데이터 정제
- 마스터 데이터
- 데이터 보안
- 데이터 분석
- 현업 부서 요구 사항 분석

툴

- 분산 컴퓨팅(예: Hadoop)
- 데이터 품질
- 데이터 통합
- 마스터 데이터 관리
- 데이터 마스킹
- 시각화
- 스트리밍 분석
- 분석
- 기계 학습

4단계: 통찰력 활용

대시보드 개발

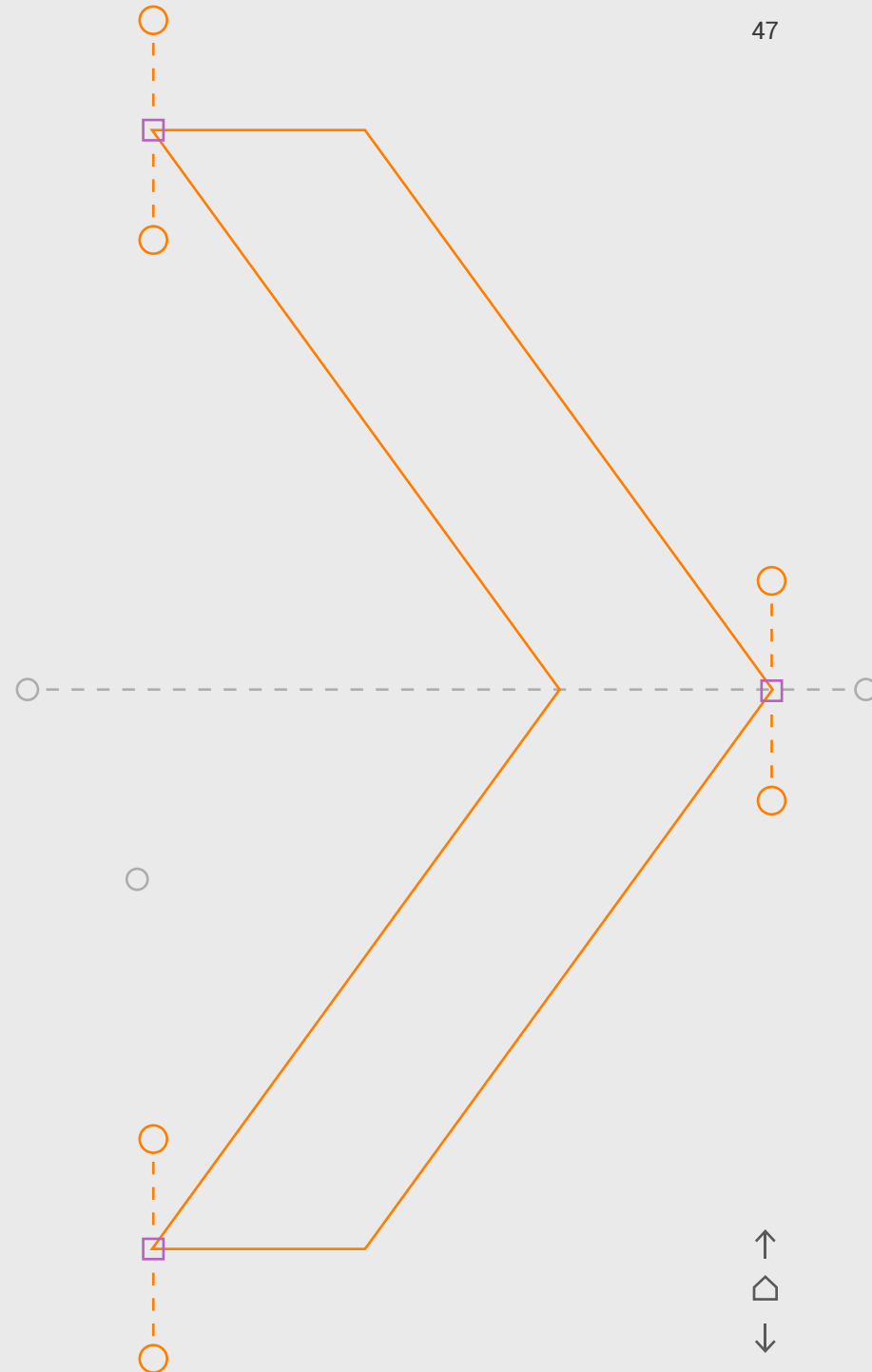
데이터 전달 프로세스 자동화

피드백 프로세스 확립

시작하기

이 워크북에 나와 있는 체크 리스트, 원칙 및 지침을 사용하여 조직에서 빅 데이터의 잠재력을 실현해 보십시오. 프로젝트의 규모에 상관없이 지금은 이 프로젝트에서 발생하는 많은 과제를 해결하기 위해 보다 철저한 준비를 갖추어야 합니다.

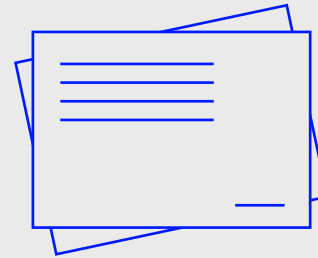
이는 대단한 성과로 이어질 것입니다.



다음 단계

배운 내용을 적용할 준비가 되셨습니까?

Informatica 개발자라면 Hadoop 개발자가 될 수 있습니다.
Informatica의 빅 데이터 소프트웨어 평가판, 커넥터 및
서비스가 여러분을 올바른 길로 안내해 드릴 것입니다.



[지금 다운로드](#)

Informatica 정보

디지털 변혁은 우리의 기대치를 바꿔 놓았습니다. 이제 더 적은 비용으로 더 나은 서비스를 더 빠르게 제공할 수 있어야 합니다. 이러한 상황에 부응하기 위해서는 기업은 변화해야 하며, 데이터가 이에 대한 해답을 쥐고 있습니다.

엔터프라이즈 클라우드 데이터 관리 분야의 세계적인 선도 기업인 Informatica는 모든 부문, 카테고리, 틈새시장에서 지능적인 방식으로 고객을 지원할 준비가 되어 있습니다. Informatica는 더욱 민첩하게 운영하고, 새로운 성장 기회를 발견하거나 새로운 혁신을 이룰 수 있는 통찰력을 기업에 제공합니다.

모든 종류의 데이터에 100% 집중하여 성공에 필요한 다양한 서비스를 제공하고 있습니다.

Informatica가 제공하는 모든 것에 대해 알아보고 다음에 올 지능형 혼란을 타개하기 위해 데이터의 힘을 활용해 보시길 바랍니다.

한국인포매티카,

서울시 서초구 서초동 강남대로 465 교보타워빌딩 B동 13층

대표전화: 02-6293-5000

팩스: 02-6293-5001

수신지 부담(미국 내): 1.800.653.3871

informatica.com/kr

[linkedin.com/company/informatica](https://www.linkedin.com/company/informatica)

twitter.com/Informaticakr

문의하기

IN18-0917-2730

© Copyright Informatica LLC 2015, 2017. Informatica 및 Informatica 로고는 미국 및 기타 국가에서 Informatica LLC의 상표 또는 등록 상표입니다. 기타 회사 및 제품 이름은 해당 소유주의 상품명 또는 등록 상표일 수 있습니다.



Informatica™