Enterprise Big Data Lake

https://product.kyobobook.co.kr/detail/S000001804948

빅데이터 by 가트너의 데그 러니 - 용량, 다양성, 속도, 진실성

GIGO (Garbage in = Garbage out)

데이터 레이크 by 타호의 제임스 딕슨 - 데이터는 원래의 형태와 포맷을 유지한다. 다양한 사용자가 데이터를 사용한다.

데이터 웨어하우스

데이터 마트

데이터 웅덩이 : 빅데이터 기술을 활용해서 구축한 단일의 목적 및 프로젝트를 위한 데이터 마트 ~ ETL... , 샌드박스

데이터 연못 : 데이터 웅덩이 여러개, 대상 프로젝트 데이터로 한정, 높은 IT 비용과 제한된 데이터 가용성으로 활용성 떨어짐

데이터 레이크 : 데이터 연못과 다르게 셀프서비스 지원

데이터 오션 : 데이터가 어디있든 기업의 모든 데이터가 셀프서비스와 데이터 주도 결정 과정에 활용

올바른 플랫폼 - 용량, 비용, 다양성 ~ 읽는 시점 스키마 적용, 미래 대비

올바른 데이터

올바른 인터페이스

전문 지식수준에 맞는 데이터 제공

데이터 취합 - 익숙한 인터페이스, 필터 검색, 평가와 분류, 맥락 기반 검색

데이터 늪

영역별 기대 관리 수준

ㄴ 민감 : 데이터 관리자 - 적절한 거버넌스, 제한적인 접근

ㄴ 직업 : 데이터 과학자 : 최소 거버넌스, 민감한 데이터가 없는지 확인

ㄴ 진입 : 데이터 엔지니어 - 최소 거버넌스, 민감한 데이터가 없는지 확인

ㄴ 골드 : 비즈니스 분석가 - 철저한 거버넌스, 신뢰 가능한 정화된 데이터, 이력 및 데이터 품질

데이터의 분석 4단계 : 데이터 찾아 이해하기 (60% 시간 소요) > 데이터 확보 > 데이터 전처리 > 데이터 사용

부족 지식 ~ 클라우드소싱 분석

워터라인 데이터 ~ 핑거 프린팅

메타데이터

데이터 준비 : 정형, 정화, 혼합

상용 클라우드 데이터 레이크 / 논리 데이터 레이크

시각화, 연방, 기업 정보 통합

데이터베이스 관리 시스템 (DBMS) - 관계형 데이터 베이스 관리 시스템

느린 변경 차원

고도 병렬 처리 (MPP) 시스템

ETL과 ELT (Extract, Load, Transform)

EII

노동 집약적 수동 과정 - 스키마와 논리 변경, 성능, 빈도

데이터 품질 도구 - 스칼라, 필드 수준, 기대 수준, 데이터 세트 수준, 복수 데이터 세트 수준

MDM 시스템 (마스터 데이터 관리 시스템)

엔티티 결의

데이터 사용 : 크리스털 리포트, 제스퍼 리포트, 코그노스

데이터 과학

맵리듀스 = 병렬로 연결된 매퍼 + 결과를 받아 처리하는 리듀서

하둡 파일 시스템 (HDFS)

A/B 스플릿 테스팅

머신러닝

특성 엔지니어링

모델 드리프트

하둡 - 대규모 병렬 저장소이자 확장성과 가용성이 높은 클러스터를 구축할 때 나타나는 여러 어려운 과정을 자동화하는 실행 플랫폼 ~ 극단적인 확장성, 비용 효과성, 모듈 방식, 약한 스키마 결함, 읽는 시점 스키마 사용

데이터 웅덩이 확산 방지

빅데이터 기술 활용 전략 : 기존 기능 가져오기, 신규 프로젝트를 위한 데이터 레이크, 일원화된 거버넌스 확립

분석용 차원 모델링

스타 스키마

미가공 데이터 - 데이터 폭, 원본이나 미가공 데이터, 표가 아닌 형태

외부 데이터 - 데이터 품질, 라이선스 비용, 지적 재산권

람다 아키텍처 ~ “Big Data: Principles and Best Practices of Scalable Realtime Data Systems - Nathan Marz”

데이터 변환 - 조화, 엔티티 구별과 결합, 성능 최적화

실시간 애플리케이션과 데이터 제품 - 대시보드, 자율 행동 ~ 복합 이벤트 처리 (CEP), 경보와 알림

셀프 서비스

비즈니스 분석가

태블러 파워, BI, 클릭, 엑셀, 트리팩타, 팍사타, 워터라인 데이터, IBM 왓슨 카탈로그

구글, 엘프, 위키피디아

신뢰 구축 - 범위, 이력, 관리, 포맷, 집합의 크기, 선택성, 참조 무결성, 데이터 품질 = 완전성, 데이터 유형

데이터 프로파일링

데이터 수준 밀도, 필드 수준 밀도

정규화된 표현 방법, 원래대로 표현 방법

프로비저닝

분석용 데이토 - 알터릭스, 데이터미어, 팍사타, 트리팍타, 인포매티카, 탈렌트,

IT 팀의 역할 : 문지기에서 가게 주인으로 데이터 관리

데이터 레이크 구조화

데이터 거버넌스 - 바이모달 데이터 거버넌스

작업 영역, 민감 영역

데이터 레이크 유지할 떄 장점 - 규범적 제약, 조직의 장벽, 예측 가능성

데이터 레이크 합쳤을 때 장점 - 자원 사용 최적화, 관리와 운영 비용, 데이터 중복 감소, 재사용, 전사적 프로젝트

클라우드 데이터 레이크

단순 저장 서비스 (S3)

데이터 연방

용어집, 분류 체계, 온톨로지

기술 메타 데이터

포크소노미

태깅

데이터 품질 - 주석 품질, 큐레이션 품질, 데이터 세트 품질

이질적 데이터 연관 짓기

민감 정보 비식별화 - 투명 암호화, 명시적 암호화, 비식별화

데이터 자주권과 규제 준수

고객 디지털화