1과목.빅데이터 분석 기획

(Ch_01. 빅데이터의 이해 - SEC 02. 빅데이터 기술 및 제도)

빅데이터 분석 기사(1과목. 빅데이터 분석 기획)

빅데이터 이해

빅데이터 이해 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 빅데이터 개요 및 활용
2. **빅데이터 기술 및 제도**

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

01 빅데이터 플랫폼

빅데이터 플랫폼은 빅데이터에서 가치를 찾아내기 위한 과정을 규격화한 기술이다.

1) 빅데이터 플랫폼 계층 구조

; 빅데이터 플랫폼은 소프트웨어 계층, 플랫폼 계층, 인프라 스트럭처 계층 구조로 구성된다.

① 소프트웨어 계층(Software Layer)

▶ 빅데이터 처리 및 분석과 이를 위한 데이터 수집 및 정제 등을 수행한다.

▶ 데이터 처리 및 분석 엔진, 데이터 수집 및 정제 모듈, 서비스 관리 모듈, 사용자 관리 모듈, 모니터링 모듈, 보안 모듈로 구성된다.

② 플랫폼 계층(Platform Layer)

▶ 빅데이터를 응용하는 기반을 제공하며, 데이터 처리 및 분석과 이를 위한 데이터 수집 및 정제 등을 수행한다.

▶ 작업 스케줄링 모듈, 데이터 자원 및 할당 모듈, 프로파일링 모듈, 데이터 관리 모듈, 자원 관리 모듈, 서비스 관리 모듈, 사용자 관리 모듈, 모니터링 모듈, 보안 모듈로 구성된다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

1) 빅데이터 플랫폼 계층 구조

③ 인프라 스트럭처 계층(Infrastructure Layer)

▶ 빅데이터 처리 및 분석에 필요한 자원을 제공한다.

▶ 자원 배치 모듈, 노드 관리 모듈, 데이터 관리 모듈, 자원 관리 모듈, 서비스 관리 모듈, 사용자 관리 모듈, 모니터링 모듈, 보안 모듈로 구성된다.

2) 빅데이터 플랫폼 구성 요소

구성 요소	주요 기능
데이터 수집	원천 데이터의 정형, 반정형, 비정형 데이터의 수집 기술 예) ETL, 크롤러 등
데이터 저장	정형 데이터, 반정형 데이터, 비정형 데이터의 저장 기술 예) RDBMS, NoSQL 등
데이터 분석	텍스트 분석, 머신러닝, 통계, 데이터 마이닝 기술 예) 자연어 처리, 예측 분석 등
데이터 활용	데이터 가시화 및 Open API 연계 예) 박스플롯, 인포그래픽 등

ETL(Extract, Transform, Load) : 원천 데이터로부터 필요한 데이터를 추출하여 적재하고자 하는 데이터 웨어하우스에 맞게 변환하여 적재하는 과정
크롤러(Crawler) : 웹 에이전트를 이용하여 인터넷 링크를 따라다니며 방문한 사이트의 웹 페이지나 소셜 데이터 등 공개되어 있는 데이터를 수집
NoSQL(Not-only SQL) : 전통적인 관계형 데이터베이스와는 다르게 데이터 모델을 단순화하여 설계된 비 관계형 데이터베이스로 SQL을 사용하지 않는 DBMS와 데이터 저장장치이다.
인포그래픽(Infographics) : 정보를 빠르고 분명하게 표현하기 위해 정보, 자료, 지식을 그래픽 시각적으로 표현한 것을 말한다.
박스 플롯(box plot) : 박스앤위스커 플롯(상자수염플롯; box-and-whisker plot)은 데이터의 대략적인 분포와 개별적인 이상치들을 동시에 보여줄 수 있으며 서로 다른 데이터 뭉치를 쉽게 비교할 수 있도록 도와주는 시각화 기법으로 가장 널리 쓰이는 시각화 형태 중 하나이다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

3) 아파치 하둡(Apache Hadoop)

- 아파치 하둡은 고가용성 분산형 객체 지향적 플랫폼(High Availability Distributed Object Oriented Platform)의 약자로 대용량의 데이터를 적은 비용으로 빠르게 분석할 수 있는 플랫폼을 의미한다.
- 객체 지향적 작업을 병렬 분산하여 고가용성을 확보할 수 있고, 구조적, 비구조적 데이터를 처리할 수 있다.

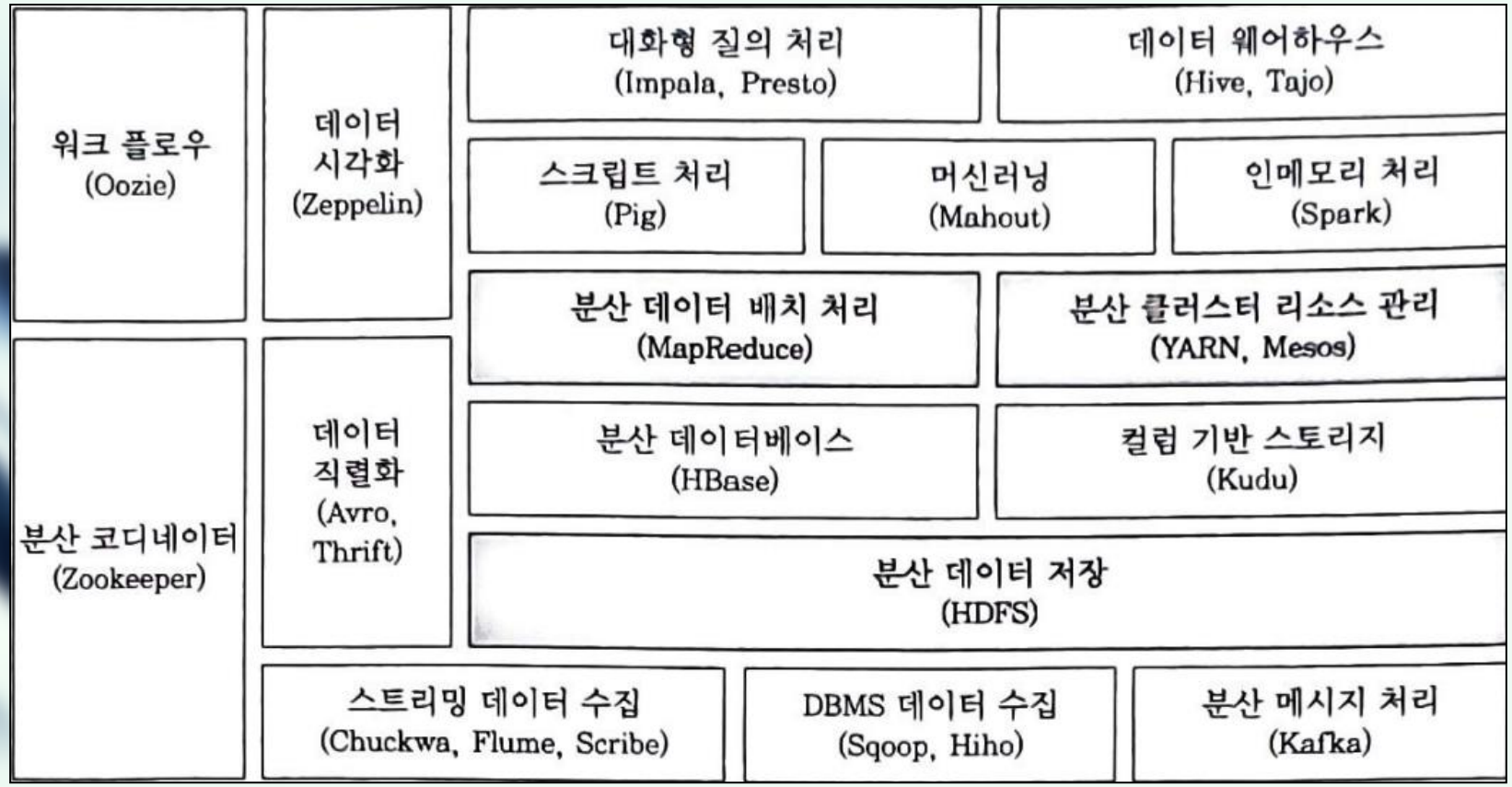
4) 하둡 에코 시스템(Hadoop Ecosystem)

- 하둡 프레임워크를 이루고 있는 다양한 서브 프로젝트들의 모임이다.
- 데이터 수집, 저장, 처리, 가공, 리소스 관리, 실시간 SQL 질의 등의 기능을 갖는다.
- 하둡의 코어 프로젝트는 분산 데이터 저장(HDFS)과 분산 데이터 처리(MapReduce)이고, 서브 프로젝트는 이를 제외한 워크플로우 관리, 데이터 마이닝, 분석, 수집, 직렬화 등이다.

분산 데이터 처리(DDP: Distributed Data Processing) : 다수의 컴퓨터를 네트워크로 연결하여 사용자가 여러 컴퓨터에 있는 데이터를 한 대의 컴퓨터 시스템에 저장된 것처럼 데이터를 처리하는 기술
MapReduce : 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크이다.
데이터 마이닝(Data Mining) : 대용량의 데이터로부터 인사이트를 도출할 수 있는 방법론이다.
마이닝 : 데이터로부터 통계적인 의미가 있는 개념이나 특성을 추출하고 패턴이나 추세 등의 정보를 끌어내는 과정
직렬화(serialization) : 데이터를 네트워크로 전송하기 위해 구조화된 객체를 바이트 스트림으로 전환하는 과정

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

4) 하둡 에코 시스템(Hadoop Ecosystem)



1. 빅데이터의 이해 - 빅데이터 기술 및 제도

개념 체크

1. 다음 중 빅데이터 플랫폼에 대한 설명으로 틀린 것은?

① 빅데이터 플랫폼은 빅데이터 수집, 저장, 처리, 분석 등 전 과정을 통합적으로 제공한다.

② 빅데이터를 처리하는 과정에서 발생하는 컴퓨팅 부하, 저장 부하, 네트워크 부하들을 해소하는 기능을 한다.

③ 빅데이터 플랫폼은 소프트웨어 계층과 하드웨어 계층 으로 구성되어 있다.

④ 데이터 구조의 변화와 신속성 요구, 데이터 분석 유연성 증대 등으로 인해 빅데이터 플랫폼이 등장하였다.

빅데이터 플랫폼은 빅데이터 수집부터 저장, 처리, 분석, 등 전 과정을 통합적으로 제공하여 그 기술들을 잘 사용할 수 있도록 준비된 환경이다.

빅데이터를 처리하는 과정에서 부하 발생은 불가피하며, 빅데이터 플랫폼은 이러한 부하들을 기술적인 요소들을 결합하여 해소를 한다.

빅데이터 처리 과정별 요소 기술을 고려한 3개의 계층이 존재하는데 소프트웨어 계층, 플랫폼 계층, 인프라스트럭처

2. 다음 중 하둡 에코 시스템(Hadoop Ecosystem)에 대한 설명으로 틀린 것은?

① 하둡 프레임워크를 이루고 있는 다양한 서브 프로젝트들의 모임이다.

② 데이터 수집, 저장, 처리 등의 기능을 한다.

③ SQL 질의 등의 기능은 없다.

④ 하둡의 코어 프로젝트는 분산 데이터 저장(HDFS)과 분산 데이터 처리(MapReduce)이다.

하둡 에코 시스템

하둡 프레임워크를 이루고 있는 다양한 서브 프로젝트들의 모임이다.

데이터 수집, 저장, 처리, 가공, 리소스 관리, 실시간 SQL 질의 등의 기능을 갖는다.

하둡의 코어 프로젝트는 분산 데이터 저장(HDFS)과 분산 데이터 처리(MapReduce)이고, 서브 프로젝트는 이를 제외한 워크플로우 관리, 데이터 마이닝, 분석, 수집, 직렬화 등이다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

① 비정형 데이터 수집

㉮ 척와(Chuckwa)

- ▶ 분산된 환경에서 생성되는 데이터를 HDFS에 안정적으로 저장시키는 플랫폼이다.
- ▶ 분산된 각 서버에서 에이전트(agent)를 실행하고, 콜렉터(collector)가 에이전트로부터 데이터를 받은 뒤, HDFS에 저장한다.

㉯ 플럼(Flume)

- ▶ 척와와 비슷하지만 전체 데이터의 흐름을 관리하는 마스터 서버가 존재하여 데이터 수집 방식 및 저장 위치에 대한 효율적 작업이 가능한 플랫폼이다.

㉰ 스크라이브(Scribe)

- ▶ 페이스북에서 개발한 대용량 실시간 로그 수집 플랫폼으로 척와와 달리 데이터를 중앙 집중 서버로 전송하는 방식이다.
- ▶ 최종 데이터는 HDFS 외에 다양한 저장소를 활용할 수 있고, 설치와 구성이 쉬우며 다양한 프로그램 언어를 지원한다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

② 정형 데이터 수집

㉮ 스쿱(Sqoop)

- ▶ 대용량 데이터 전송 솔루션이다.
- ▶ HDFS, RDBMS 등 다양한 저장소에 대용량 데이터를 신속하게 전송할 수 있는 방법을 제공한다.

㉯ 히호(Hiho)

- ▶ 대용량 데이터 전송 솔루션이다.
- ▶ 하둡에서 데이터를 가져오기 위한 SQL을 지정할 수 있고, JDBC 인터페이스를 지원한다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

③ 분산 데이터 저장

HDFS(Hadoop Distributed File System)

- 대용량 파일들을 분산된 서버에 저장하고, 그 저장된 데이터를 빠르게 처리할 수 있도록 설계된 하둡 분산 파일 시스템이다.
- 네임노드(Master)와, 데이터노드(Slave)로 구성된다.
- 네임노드(NameNode)는 HDFS의 메타 데이터를 관리하고 클라이언트가 HDFS에 저장된 파일에 접근할 수 있도록 한다.
- 데이터노드(DataNode)는 주기적으로 네임노드에게 하트 비트(Heart beat) 블록의 목록 리포트(Block Report)를 보낸다.

하트 비트(Heart beat) : 데이터노드가 네임노드에게 3초마다 보내는 정보로 하트 비트에는 디스크 가용 공간정보, 데이터 이동, 적재량 등의 정보가 들어 있다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

④ 분산 데이터 처리

맵리듀스(MapReduce)

- ▶ 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크이다.
- ▶ 맵리듀스는 맵(Map) 작업과 리듀스(Reduce) 작업의 결합이다.
- ▶ 맵(Map)작업은 여러 데이터를 Key-Value의 형태로 연관성 있는 데이터로 분류하여 묶는 작업이다.
- ▶ 리듀스(Reduce) 작업은 맵 작업한 데이터 중 중복 데이터를 제거하고 원하는 데이터를 추출하는 작업이다.
- ▶ 과정 : Input → Splitting → Mapping → Shuffling → Reducing → Final Result

분할(Splitting) : 입력한 파일 값을 라인 단위로 분할한다.
매핑(Mapping) : 분할된 라인 단위 문장을 맵(Map)함수로 전달하면 맵 함수는 공백을 기준으로 문자를 분리, 단어 개수를 확인한다.
셔플링(Shuffling) : 메모리에 저장되어 있는 맵 함수의 출력 데이터를 파티셔닝과 정렬하여 로컬 디스크에 저장, 네트워크를 통해서 리듀서의 입력 데이터로 전달한다.
리듀싱(Reducing) : 단어 목록들을 반복적으로 수행하고 합을 계산하여 표시한다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

⑤ 분산 데이터베이스

HBase

▶ HDFS의 분산 컬럼 기반 데이터베이스이다.

▶ 실시간 랜덤 조회 및 업데이트를 할 수 있으며, 각각의 프로세스는 개인의 데이터를 비동기적으로 업데이트 할 수 있다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

개념 체크

다음 중, 하둡 에코 시스템 기술 중에서 대용량 파일들을 분산된 서버에 저장하고, 그 저장된 데이터를 빠르게 처리 할 수 있도록 설계된 것은?

① HBase ② 맵리듀스

③ HDFS ④ 스크라이브(Scribe)

HDFS(Hadoop Distributed File System)

▶ 대용량 파일들을 분산된 서버에 저장하고, 그 저장된 데이터를 빠르게 처리할 수 있도록 설계된 하둡 분산 파일 시스템이다.

▶ 네임노드(Master)와, 데이터노트(Slave)로 구성된다.

▶ 네임노드(NameNode)는 HDFS의 메타 데이터를 관리하고 클라이언트가 HDFS에 저장된 파일에 접근을 할 수 있도록 한다.

▶ 데이터노드(DataNode)는 주기적으로 네임노드에게 하트비트(Heart Beat)블록의 목록 리포트(Block Report)를 보낸다.

HBase

▶ HDFS의 분산 컬럼 기반 데이터베이스이다.

▶ 실시간 랜덤 조회 및 업데이트를 할 수 있으며, 각각의 프로세스는 개인의 데이터를 비동기적으로 업데이트를 할 수 있어 매우 빠른 속도를 지닌다.

맵리듀스(MapReduce)

▶ 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크이다.

▶ 맵리듀스는 맵(Map) 작업과 리듀스(Reduce) 작업의 결합된 용어이다.

▶ 맵 작업은 여러 데이터를 Key-Value의 형태로 연관성 있는 데이터로 분류하여 묶는 작업이다.

▶ 리듀스 작업은 맵 작업을 한 데이터 중 중복 데이터를 제거하고 원하는 데이터를 추출하는 작업이다.

스크라이브(Scribe) – 비정형 데이터 수집 기술

▶ 페이스북에서 개발한 대용량 실시간 로그 수집 플랫폼으로 척와와 달리 데이터를 중앙 집중 서버로 전송하는

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

⑥ 리소스 관리

얀(Yet Another Resource Negotiator)

- 리소스 관리와 컴포넌트 처리를 분리한 아파치 소프트웨어 재단의 서브 프로젝트이다.
- 맵리듀스의 확장성과 속도 문제를 해소하기 위해 새롭게 만든 자원 관리 플랫폼이다.
- 얀의 구성 요소에는 리소스 매니저, 노드 매니저, 애플리케이션 마스터, 컨테이너가 있다.

구성 요소	설명
리소스 매니저	모든 시스템 자원을 관리하고, 효율적으로 자원 분배
노드 매니저	노드의 자원 관리 및 리소스 매니저에게 현재 자원 상태 보고
애플리케이션 마스터	컨테이너를 사용하여 작업 모니터링 및 실행 관리
컨테이너	프로그램 구동을 위한 다양한 시스템 자원

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

⑦ 인메모리 처리

아파치 스파크(Apache Spark)

- SQL, 스트리밍, 머신러닝 및 그래프 처리를 위한 기본 제공 모듈이 있는 대규모 데이터 처리용 통합 분석 엔진이다.
- 하둡 기반 대규모 데이터 분산처리 시스템이다.
- 스트리밍 데이터, 온라인 머신러닝 등 실시간 데이터 처리를 한다.
- 스칼라, 자바, 파이썬, R 등에 사용 가능하다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

⑧ 데이터 가공

㉮ 피그(Pig)

- ▶ 대용량 데이터를 고차원으로 분석하기 위한 플랫폼이다.
- ▶ 맵리듀스의 API를 단순화시켜, SQL과 유사한 형태로 설계된다.
- ▶ JOIN 연산 지원

㉯ 하이브(Hive)

- ▶ 하둡 기반의 DW(데이터 웨어하우스) 솔루션이다.
- ▶ SQL과 매우 유사한 HiveQL 쿼리를 제공한다.

데이터 웨어하우스 : 여러 소스에서 가져온 구조화된 데이터와 반 구조화된 데이터를 분석하고 보고하는데 사용되는 엔터프라이즈 시스템
엔터프라이즈 시스템(Enterprise System) : 대규모 조직의 각기 다른 기능과 조직 수준, 비즈니스 프로세스를 대상으로 개발된 다양한 정보 시스템을 연결해 상호간에 정보 교환이 수월해지고, 조직의 효율성과 경영성과를 증진시킬 수 있게 하는 시스템
HiveQL : 하둡에 저장된 데이터를 쉽게 처리할 수 있는 데이터웨어하우스 패키지이다. 하이브를 사용하면 하둡에 저장된 데이터를 SQL과 유사한 HiveQL로 처리할 수 있기 때문에 파이썬을 배우지 않아도 쉽게 데이터를 조회/분석할 수 있다는 장점이 있다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

⑨ 데이터 마이닝

머하웃(Mahout)

- ▶ 하둡 기반의 데이터 마이닝 알고리즘을 구현한 오픈소스이다.
- ▶ 분류, 클러스터링, 추천 및 협업 필터링, 패턴 마이닝, 회귀 분석, 진화 알고리즘 등 주요한 알고리즘 자원이다.

클러스터링 : 분류 기준이 없는 상태에서 데이터 속성을 고려해 스스로 전체 데이터를 N개의 소그룹으로 묶어내는 분석법
회귀 분석 : 독립변수가 종속변수에 미치는 영향을 분석할 때 사용하는 알고리즘
진화 알고리즘 (EA, evolutionary computation, artificial evolution) : 생식 (reproduction), 돌연변이 (mutation), 재조합 (recombination) 같은 생물학에서의 진화를 본뜬 메커니즘을 사용하는 어떤 개체군 기반의 조합 최적화 알고리즘 (population-based combinatorial optimization algorithm)을 나타내는 일반적인 용어이다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

⑩ 실시간 SQL 질의

㉮ 임팔라(Impala)

- 하둡 기반의 실시간 SQL 질의 시스템이다.
- 데이터 조회 시 HiveQL를 사용한다.
- 맵리듀스를 사용하지 않고, 자체 개발한 엔진을 사용해 빠른 성능을 가진다.

㉯ 타조(Tajo)

- 하둡파일시스템(HDFS)의 데이터에 SQL 형태의 명령을 통해 분산 분석 작업을 지원하는 대용량 데이터 웨어하우스(DW)이다.
- 2010년 고려대학교 컴퓨터학과 데이터베이스 연구실에서 처음 개발되어 2014년 3월에 아파치 재단의 최상위 프로젝트로 승격되었다.
- 기존 하둡 빅데이터 처리 엔진인 하이브(Hive)와 기능이 유사하나 하이브보다 데이터 처리 속도가 빠르다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 하둡 에코 시스템 기술

⑪ 워크플로우 관리

우지(Oozie)

- 하둡 작업을 관리하는 워크플로우 및 코디네이터 시스템이다.
- 자바 서블릿 컨테이너에서 실행되는 자바 웹 애플리케이션 서버로 맵리듀스 혹은 피그와 같은 특화된 액션들로 구성된 워크플로우를 제어한다.

⑫ 분산 코디네이션

주키퍼(Zookeeper)

- 분산 애플리케이션을 위한 코디네이션 시스템이다.
- 분산 애플리케이션이 안정적인 서비스를 할 수 있도록 분산되어 있는 각 애플리케이션의 정보를 중앙에 집중하여 구성 관리, 그룹 관리 네이밍, 동기화 등의 서비스를 제공한다.

자바 서블릿(Java Servlet) : 웹 페이지를 동적으로 생성하는 서버 측 프로그램 혹은 그 사양을 말하며, 흔히 "서블릿" 이라 불린다. 쉽게 말해 서블릿은 클라이언트의 요청에 맞춰 동적인 결과를 만들어 주는 자바 웹 프로그래밍 기술이라고 할 수 있다. 이러한 서블릿은 WAS(Web Application Server)의 서블릿 컨테이너 안에서 동작하게 된다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

개념 체크

다음 중, 리소스 관리와 컴포넌트 처리를 분리한 아파치 소프트웨어 재단의 서브 프로젝트인 기술은 무엇인가?

① 우지　　② 주키퍼

③ 임팔라　　④ 얀

얀(Yet Another Resource Negotiator)

▶ 리소스 관리와 컴포넌트 처리를 분리한 아파치 소프트웨어 재단의 서브 프로젝트이다.

▶ 맵리듀스의 확장성과 속도 문제를 해소하기 위해 새롭게 만든 자원 관리 플랫폼이다.

▶ 얀의 구성 요소에는 리소스 매니저, 노드 매니저, 애플리케이션 마스터, 컨테이너가 있다.₩

우지(Oozie) - 워크플로우 관리

▶ 하둡 작업을 관리하는 워크플로우 및 코디네이터 시스템이다.

▶ 자바 서블릿 컨테이너에서 실행되는 자바 웹 애플리케이션 서버로 맵리듀스 혹은 피그와 같은 특화된 액션들로 구성된 워크플로우를 제어한다.

주키퍼(Zookeeper) - 분산 코디네이션

▶ 분산 애플리케이션을 위한 코디네이션 시스템이다.

▶ 분산 애플리케이션이 안정적인 서비스를 할 수 있도록 분산되어 있는 각 애플리케이션의 정보를 중앙에 집중하여 구성 관리, 그룹 관리 네이밍, 동기화 등의 서비스를 제공한다.

임팔라(Impala) - 실시간 SQL 질의

▶ 하둡 기반의 실시간 SQL 질의 시스템이다.

▶ 데이터 조회 시 HiveQL를 사용한다.

▶ 맵리듀스를 사용하지 않고, 자체 개발한 엔진을 사용해 빠른 성능을 가진다.

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

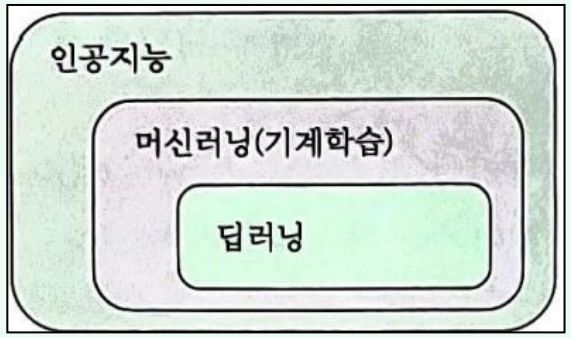
02 빅데이터와 인공지능

1) 인공지능의 정의

- 인공지능이란 인간의 학습능력, 인지능력을 인공적으로 학습시켜 일정 수준의 능력을 갖출 수 있도록 만든 소프트웨어이다.
- 인간과 비슷한 수준의 지능을 구사하기 위해서는 많은 양의 데이터가 수집, 분석, 학습되어야 한다.

2) 인공지능의 범위

- 인공지능의 범위는 작은 범위를 기준으로 딥러닝, 머신러닝(기계학습), 인공지능 순이다.
- 이는 범위를 표현하기 위한 도식화로 단순히 분야별 크기가 크고 작음을 의미하는 것이 아니라는 점을 기억할 수 있도록 한다.
- 각 단계별 기술은 지속적인 상호작용을 통해 최종적인 인공지능 기술을 구현할 수 있기 때문이다.



1. 빅데이터의 이해 - 빅데이터 기술 및 제도

2) 인공지능의 범위

구성 요소	설명
인공지능	사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 구현하는 기술
머신러닝	컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술
딥러닝	인간의 뉴런과 비슷한 인공신경망 방식으로 정보를 처리하는 기술

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

개념 체크

다음 중, 인공지능에 대한 설명 중 틀린 것은?

① 인공지능이란 인간의 학습능력, 인지능력을 인공적으로 학습시키는 것이다.

② 많은 양의 데이터가 수집, 분석, 학습되어야 한다.

③ 인공지능의 범위는 작은 범위를 기준으로 딥러닝, 머신러닝(기계학습), 인공지능 순이다.

④ 딥러닝은 컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술이다.

● 인공지능이란 인간의 학습능력, 인지능력을 인공적으로 학습시켜 일정 수준의 능력을 갖출 수 있도록 만든 소프트웨어이다.

● 인간과 비슷한 수준의 지능을 구사하기 위해서는 많은 양의 데이터가 수집, 분석, 학습되어야 한다.

● 인공지능의 범위는 작은 범위를 기준으로 딥러닝, 머신러닝(기계학습), 인공지능 순이다.

● 이는 범위를 표현하기 위한 도식화로 단순히 분야별 크기가 크고 작음을 의미하는 것이 아니라는 점을 기억할 수 있도록 한다.

● 각 단계별 기술은 지속적인 상호작용을 통해 최종적인 인공지능 기술을 구현할 수 있기 때문이다.

인공지능 : 사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 구현하는 기술

머신러닝 : 컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술

딥러닝 : 인간의 뉴런과 비슷한 인공신경망 방식으로 정보를 처리하는 기술

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

03 개인정보 법, 제도

1) 개인정보보호의 정의

; 개인정보보호는 정보주체자의 개인정보 자기결정권을 철저히 보장하는 활동을 의미한다.

2) 개인정보보호의 필요성

; 개인을 식별할 수 있는 정보가 대부분 개인정보로 구분되어 활용되기 때문에 개인정보가 유출되면 그 피해가 막심하다.

개인정보 자기결정권 : 자신에 대한 정보가 언제, 어떻게, 어떠한 범위까지 사용될 수 있는지를 정보 주체자가 스스로 결정할 수 있는 권리

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

3) 빅데이터 개인정보보호 가이드라인

- 2014년 12월 23일 한국방송통신위원회, 한국인터넷진흥원에서 제정하였다.
- 빅데이터 개인정보보호 가이드라인은 공개된 또는 이용내역정보 등을 전자적으로 설정된 체계에 의해 수집·저장·조합·분석 등을 처리하여 새로운 정보를 생성함에 있어서 이용자의 프라이버시 등을 보호하고, 안전한 이용환경을 조성하는 것을 목적으로 한다.
- 빅데이터 개인정보보호 가이드라인의 주요 내용은 다음과 같다.

구성 요소	설명
개인정보 비식별화 조치	개인정보가 포함된 공개된 정보 및 이용내역정보는 비식별화 조치를 취한 후 수집 · 저장 · 조합 · 분석 및 제 3자 제공 등 가능
투명성 확보	개인정보 취급 방침을 통해 비식별화 조치 후 빅데이터 처리 사실 · 목적 · 수집 · 출처 및 정보 활용 거부권 행사 방법 등을 이용자에게 투명하게 공개 • (개인정보 취급 방침) 비식별화 조치 후 빅데이터 처리 사실 · 목적 등을 이용자 등에게 공개하고 '정보 활용 거부 페이지 링크'를 제공하여 이용자가 거부권을 행사할 수 있도록 조치 • (수집 출처 고지) 이용자 이외의 자로부터 수집한 개인정보 처리 시 '수집 출처 · 목적, 개인정보 처리 정지 요구권'을 이용자에게 고지

개인정보 재식별 조치	빅데이터 처리 과정 및 생성정보에 개인정보가 재식별 될 경우, 즉시 파기하거나 추가적인 비식별화 조치토록 함
민감정보 처리	• 특정 개인의 사상 · 신념, 정치적 견해 등 민감정보의 생성을 목적으로 정보의 수집 · 이용 · 저장 · 조합 · 분석 등 처리 금지 • 이메일, 문자 메시지 등 통신 내용의 수집 · 이용 · 저장 · 조합 · 분석 등 처리 금지
수집정보 보호조치	비식별화 조치가 취해진 정보를 저장 · 관리하고 있는 정보 처리 시스템에 대한 기술적 · 관리적 보호조치 적용 ※ (보호조치) 침입차단시스템 등 접근 통제장치 설치, 접속 기록에 대한 위 · 변조 방지 조치, 백신 소프트웨어 설치 · 운영 등 악성 프로그램에 의한 침해 방지 조치

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

4) 데이터 3법

; 개인정보보호법, 정보통신망 이용 촉진 및 정보보호 등에 관한 법률(정보통신망법), 신용정보의 이용 및 보호에 관한 법률(신용정보법)을 일컫는다.

5) 개인정보보호법[개인정보의 수집·이용(제15조)]

① 개인정보처리자는 다음 각 호의 어느 하나에 해당하는 경우에는 개인 정보를 수집할 수 있으며 그 수집 목적의 범위에서 이용할 수 있다.

〈개정 2023. 3. 14.〉

1. 정보주체의 동의를 받은 경우
2. 법률에 특별한 규정이 있거나 법령상 의무를 준수하기 위하여 불가피한 경우
3. 공공기관이 법령 등에서 정하는 소관 업무의 수행을 위하여 불가피한 경우
4. 정보주체와 체결한 계약을 이행하거나 계약을 체결하는 과정에서 정보주체의 요청에 따른 조치를 이행하기 위하여 필요한 경우
5. 명백히 정보주체 또는 제3자의 급박한 생명, 신체, 재산의 이익을 위하여 필요하다고 인정되는 경우
6. 개인정보처리자의 정당한 이익을 달성하기 위하여 필요한 경우로서 명백하게 정보주체의 권리보다 우선하는 경우. 이 경우 개인정보처리자의 정당한 이익과 상당한 관련이 있고 합리적인 범위를 초과하지 아니하는 경우에 한한다.
7. 공중위생 등 공공의 안전과 안녕을 위하여 긴급히 필요한 경우

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

5) 개인정보보호법[개인정보의 수집·이용(제15조)]

② 개인정보처리자는 제1항 제1호에 따른 동의를 받을 때에는 다음 각 호의 사항을 정보주체에게 알려야 한다. 다음 각 호의 어느 하나의 사항을 변경하는 경우에도 이를 알리고 동의를 받아야 한다.

1. 개인정보의 수집 · 이용 목적
2. 수집하려는 개인정보의 항목
3. 개인정보의 보유 및 이용 기간
4. 동의를 거부할 권리가 있다는 사실 및 동의 거부에 따른 불이익이 있는 경우에는 그 불이익의 내용

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

6) 개인정보보호법[개인정보 유출 통지 등(제34조)]

①항 개인정보처리자는 개인정보가 유출되었음을 알게 되었을 때에는 지체 없이 해당 정보주체에게 다음의 사실을 알려야 한다.

1. 유출된 개인정보의 항목
2. 유출된 시점과 그 경위
3. 유출로 인하여 발생할 수 있는 피해를 최소화하기 위하여 정보주체가 할 수 있는 방법 등에 관한 정보
4. 개인정보처리자의 대응조치 및 피해 구제절차
5. 정보주체에게 피해가 발생한 경우 신고 등을 접수할 수 있는 담당부서 및 연락처

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

7) 개인정보보호법[개인정보보호원칙(제3조)]

① 개인정보처리자는 개인정보의 처리 목적을 명확하게 하여야 하고, 그 목적에 필요한 범위에서 최소한의 개인정보만을 적법하고 정당하게 수집하여야 한다.

② 개인정보처리자는 개인정보의 처리 목적에 필요한 범위에서 적합하게 개인정보를 처리하여야 하며, 그 목적 외의 용도로 활용하여서는 아니 된다.

③ 개인정보처리자는 개인정보의 처리 목적에 필요한 범위에서 개인정보의 정확성, 완전성 및 최신성이 보장되도록 하여야 한다.

④ 개인정보처리자는 개인정보의 처리 방법 및 종류 등에 따라 정보주체의 권리가 침해받을 가능성과 그 위험 정도를 고려하여 개인정보를 안전하게 관리하여야 한다.

⑤ 개인정보처리자는 제30조에 따른 개인정보 처리방침 등 개인정보의 처리에 관한 사항을 공개하여야 하며, 열람청구권 등 정보주체의 권리를 보장하여야 한다. 〈개정 2023. 3. 14.〉

⑥ 개인정보처리자는 정보주체의 사생활 침해를 최소화하는 방법으로 개인정보를 처리하여야 한다.

⑦ 개인정보처리자는 개인정보를 익명 또는 가명으로 처리하여도 개인정보 수집목적을 달성할 수 있는 경우 익명처리가 가능한 경우에는 익명에 의하여, 익명처리로 목적을 달성할 수 없는 경우에는 가명에 의하여 처리될 수 있도록 하여야 한다. 〈개정 2020. 2. 4.〉

⑧ 개인정보처리자는 이 법 및 관계 법령에서 규정하고 있는 책임과 의무를 준수하고 실천함으로써 정보주체의 신뢰를 얻기 위하여 노력하여야 한다.

[시행일 : 2023. 9. 15.] 제3조

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

8) GDPR(General Data Protection Regulation) 유럽 연합 일반 데이터 보호규칙

2018년 5월 25일부터 시행된 EU(유럽연합)의 개인정보보호 법령으로 정보주체의 권리, 기업의 책임성 강화, 개인정보의 EU 역외이전(onward transfer) 요건을 명확화한 규칙이다.

역외이전 : 개인정보가 EU경계를 넘어 제3국이나 국제기구로 이전되는 경우

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

04 개인정보 활용

1) 가명정보 활용 범위

- 가명정보는 특정인을 식별할 수 없도록 조치한 정보이다.
- 가명정보는 통계작성(상업적 목적 포함, 시장조사), 연구(산업적 연구포함), 공익적 기록 보존 목적 등에 사용 가능하다.
- 가명정보 처리 절차는 사전준비 -> 가명처리 -> 적정성 검토 및 추가처리 → 사후관리 순이다.

2) 프라이버시 보호 모델

① **k-익명성(k-Anonymity)**

▶ 주어진 데이터 집합에서 같은 값이 적어도 k개 이상 존재하도록 하여 쉽게 다른 정보와 결합할 수 없도록 한 모델

▶ 공개된 데이터의 연결공격 취약점을 보완하기 위한 모델

② **I-다양성(I-Diversity)**

▶ 주어진 데이터 집합에서 함께 비식별 되는 레코드들은 동질 집합에서 적어도 1개의 서로 다른 민감한 정보를 가져야 하는 모델

▶ k-의명성에 대한 두 가지 취약점 공격인 동질성 공격, 배경지식에 의한 공격을 방어하기 위한 모델

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

2) 프라이버시 보호 모델

③ t-근접성(t-Closeness)

- 동질 집합에서 특정 정보의 분포와 전체 데이터 집합에서 정보의 분포가 t 이하의 차이를 보여야 하는 모델
- l-다양성의 쏠림 공격, 유사성 공격을 보완하기 위한 모델

④ m-유일성(m-Uniqueness)

- 원본 데이터와 동일한 속성의 값 조합이 비식별 결과 데이터에 최소 m개 이상 존재하도록 만들어 재식별 가능성의 위험을 낮춘 모델

1. 빅데이터의 이해 - 빅데이터 기술 및 제도

3) 마이 데이터(My data)

- 마이 데이터란 개인이 데이터를 주체적으로 관리하는 것을 넘어 능동적으로 활용하는 일련의 과정을 의미한다.
- 2020년 8월부터 신용정보법 개정안을 비롯한 데이터 3법이 시행되면서 개인 데이터의 주인은 본인이라는 주장이 가능해졌다.
- 개인의 데이터 주권인 자기 정보결정권으로 개인 데이터의 활용과 관리에 대한 통제권을 개인이 갖는다는 것이 핵심이다.
- 마이 데이터 활용 예시 : 금융 정보 통합 관리 가능, 신용 및 자산 분석 가능
- 마이 데이터 원칙 : 데이터 권한, 데이터 제공, 데이터 활용

1. 빅데이터의 이해 - 빅데이터 기술 및 제도 예상 문제

예상 문제

1. 다음 중, 인공지능에 대한 설명 중 틀린 것은?

① 인공지능이란 인간의 학습능력, 인지능력을 인공적으로 학습시키는 것이다.

② 많은 양의 데이터가 수집, 분석, 학습되어야 한다.

③ 인공지능의 범위는 작은 범위를 기준으로 딥러닝, 머신러닝(기계학습), 인공지능 순이다.

④ 딥러닝은 컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술이다.

● 인공지능이란 인간의 학습능력, 인지능력을 인공적으로 학습시켜 일정 수준의 능력을 갖출 수 있도록 만든 소프트웨어이다.

● 인간과 비슷한 수준의 지능을 구사하기 위해서는 많은 양의 데이터가 수집, 분석, 학습되어야 한다.

● 인공지능의 범위는 작은 범위를 기준으로 딥러닝, 머신러닝(기계학습), 인공지능 순이다.

● 이는 범위를 표현하기 위한 도식화로 단순히 분야별 크기가 크고 작음을 의미하는 것이 아니라는 점을 기억할 수 있도록 한다.

● 각 단계별 기술은 지속적인 상호작용을 통해 최종적인 인공지능 기술을 구현할 수 있기 때문이다.

인공지능 : 사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 구현하는 기술

머신러닝 : 컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술

딥러닝 : 인간의 뉴런과 비슷한 인공신경망 방식으로 정보를 처리하는 기술

감사합니다.