CSE545
빅데이터처리 및 실습
(Big Data Processing and Practice)
오리엔테이션

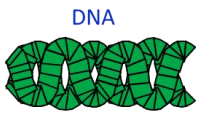
담당교수: 전강욱(컴퓨터공학부)
kw.chon@koreatech.ac.kr

담당교수 및 담당조교

- **담당교수: 전강욱**
  - 연구분야: 빅데이터(병렬 및 분산시스템 기반 데이터처리)
    - 대규모 시스템 운영(+1,000노드 구성 Data Lake 운영 경험)
    - 분산시스템/멀티코어/GPU 기반 SW 병렬화
    - 오픈소스 기반 빅데이터 SW기술(주로, 데이터처리)
  - 연락처: 041-560-1658
  - 연구실: 제2공학관 425호
  - E-mail: kw.chon@koreatech.ac.kr

- **담당조교: 김민형**
  - 연락처: 010-8886-0203
  - E-mail: kimexcel2@koreatech.ac.kr

주요 연구 분야



High Throughput
NGS Systems

High Throughput
Mass Spectrometry

10B Web Page

2.7B Users

1M Video / day

How can we design algorithms and platforms for large-scale data?

Big Data Engineering

Data Mining / Machine Learning

- Frequent Pattern Mining
  [INS'18][Cluster'18][Access'22] [ESWA (under review)]
- Tucker Decomposition
  [TKDE (in preparation)]
- Deep Learning
  [ESWA (under review)] [Access (under review)]

ETC (Ongoing Projects)

- Erasure Coding
  acceleration in erasure coding using heterogenous processors
- Scientific Applications
  acceleration in astrophysical systems using GPUs

NVIDIA A100
(8,192 Cores)

The 3rd Gen Xeon
(40 Cores)

Conventional Cluster

Supercomputer

Bioinformatics

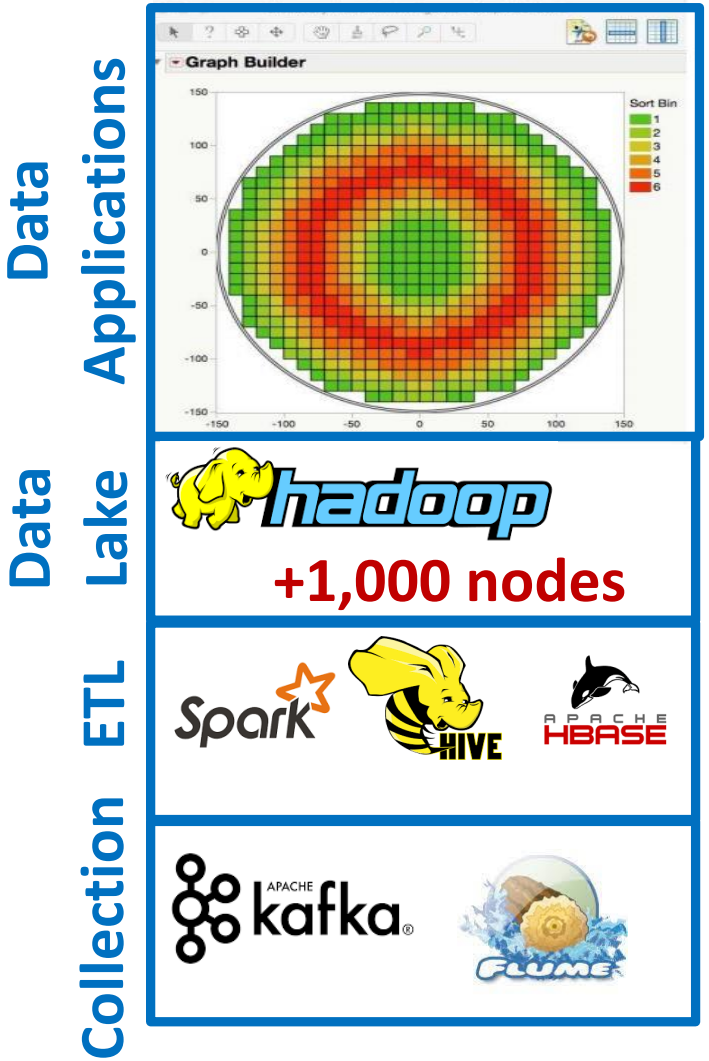
- Primer Design
  [NAR'15][IJDBM'17]
- Protein Identification
  [KSC'22][Bioinformatics (in preparation)]

Work Experience

- Data Lake
  (+1000 nodes at SK Telecom)
- Big Data Systems for Smart Factory
  (SK Hynix)

빅 데이터 처리 관련 회사 경험

- **Real-time data pipeline based on distributed systems**
  - Apache Hadoop, Hbase, Phoenix
    - Real time data ingestion
  - Kafka ecosystems
    - Apache Kafka
    - Kafka Rest Proxy
    - Apache Avro
    - Kafka Schema Registry
  - Distributed computing
    - Apache Spark
- **Data Lake**
  - Next-generation DW for call billing
  - Data collection (real-time)
  - Data processing
  - Data extraction and analysis



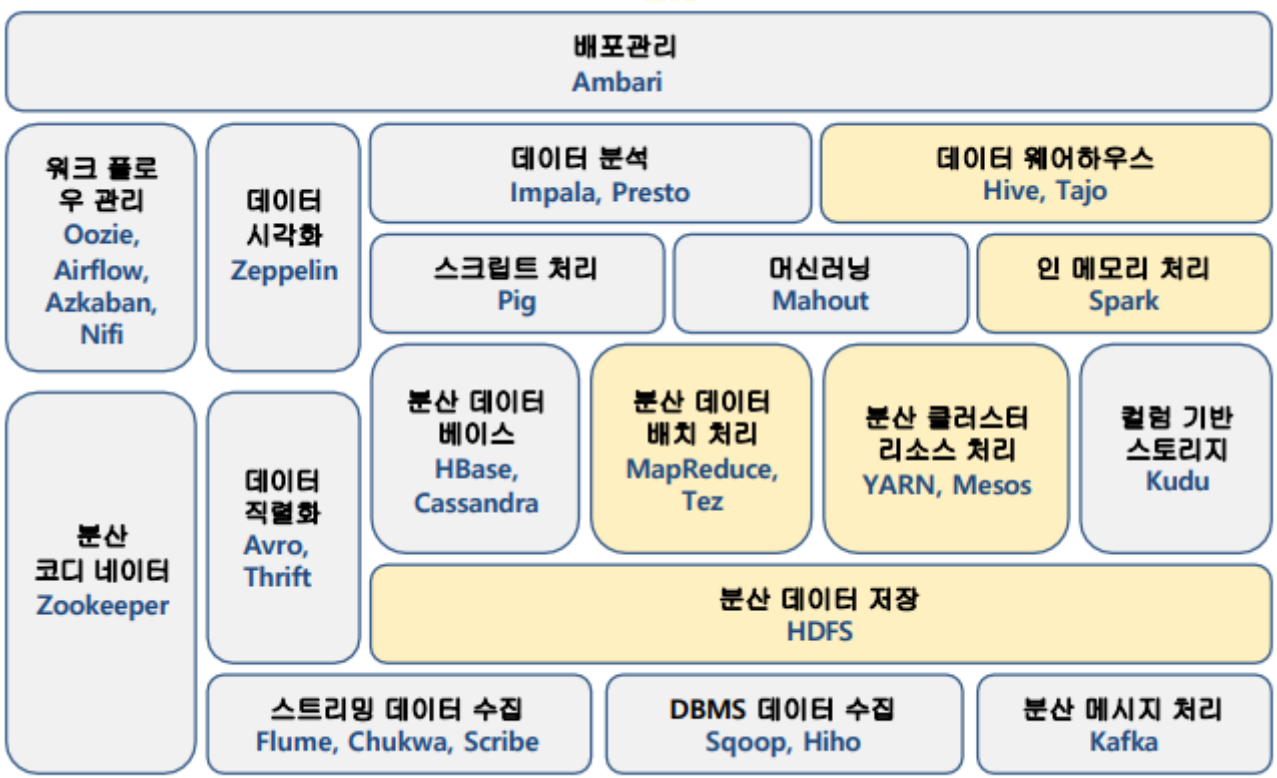
Hadoop Ecosystem

- **많은 수의 기업에서 Hadoop Ecosystem 기반의 빅 데이터 플랫폼 구축**
  - Hadoop 기반의 서비스 플랫폼 구축을 위한 오픈소스 기반의 세부 프로젝트들 (+100 프로젝트)

Distributed Filesystem	Distributed Programming		NoSQL Databases			NewSQL	Categorize Pending ...
			Colume	Document	Key-Value		
• Apache HDFS • GlusterFS • Quantcast File System QFS • Ceph Filesystem • Lustre file system • Tachyon • GridGain	• MapReduce • Apache Pig • JAQL • Apache Spark • Apache Flink • Netflix PigPen • AMPLab SIMR • Facebook Corona • Apache Twill	• Damballa Parkour • Apache Hama • Datasalt Pangool • Apache Tez • Apache DataFu • Pydoop • Kangaroo • TinkerPop • Pachyderm	• Apache HBase • Apache Cassandra • Hypertable • Apache Accumulo	• MongoDB • RethinkDB • ArangoDB **Graph** • ArangoDB • Neo4j • TitanDB	• Redis DataBase • Linkedin Voldemort • RocksDB • OpenTSDB **Stream** • EventStore	• TokuDB • HandlerSocket • Akiban Server • Drizzle • Haeinsa • SenseiDB • Sky • BayesDB • InfluxDB	• Twitter Summingbird • Apache Kiji • S4 Yahoo • Metamarkers Druid • Concurrent Cascading • Concurrent Lingual • Concurrent Pattern • Apache Giraph • Talend • Akka Toolkit • Eclipse BIRT • Spango BI • Jedox Palo • Twitter Finagle • Intel GraphBuilder • Apache Tika
SQL-On-Hadoop	**Data Ingestion**	**Service Programming**	**Scheduling**	**Machine Learning**	**Benchmark**	**System Deployment**	
• Apache Hive • HCatalog • Trafodion: SQL-on-HBase • Apache Drill • Cloudera Impala • Facebook Presto • Splout SQL • Apache Tajo • Apache Phoenix • Apache MRQL • Kylin	• Apache Flume • Apache Sqoop • Facebook Scribe • Apache Chukwa • Apache Storm • Apache Kafka • Netflix Suro • Apache Samza • Cloudera Morphline • HIHO • Apache NiFi	• Apache Thrift • Apache Zookeeper • Apache Avro • Apache Curator • Apache karaf • Twitter Elephant Bird • Linkedin Norbert	• Oozie • Azkaban • Apache Falcon **Security** • Sentry • Knox Gateway • Ranger	• Apache Mahout • WEKA • Cloudera Oryx • MADlib • H2O • Sparkling Water	• Apache Hadoop Benchmarking • Yahoo Gridmix3 • PUMA Benchmarking • Berkeley SWIM Benchmark • Intel HiBench	• Ambari • HUE • Whirr • Mesos • Myriad • Marathon • Brooklyn • HOYA • Helix • Bigtop • Buildoop • Deploop	

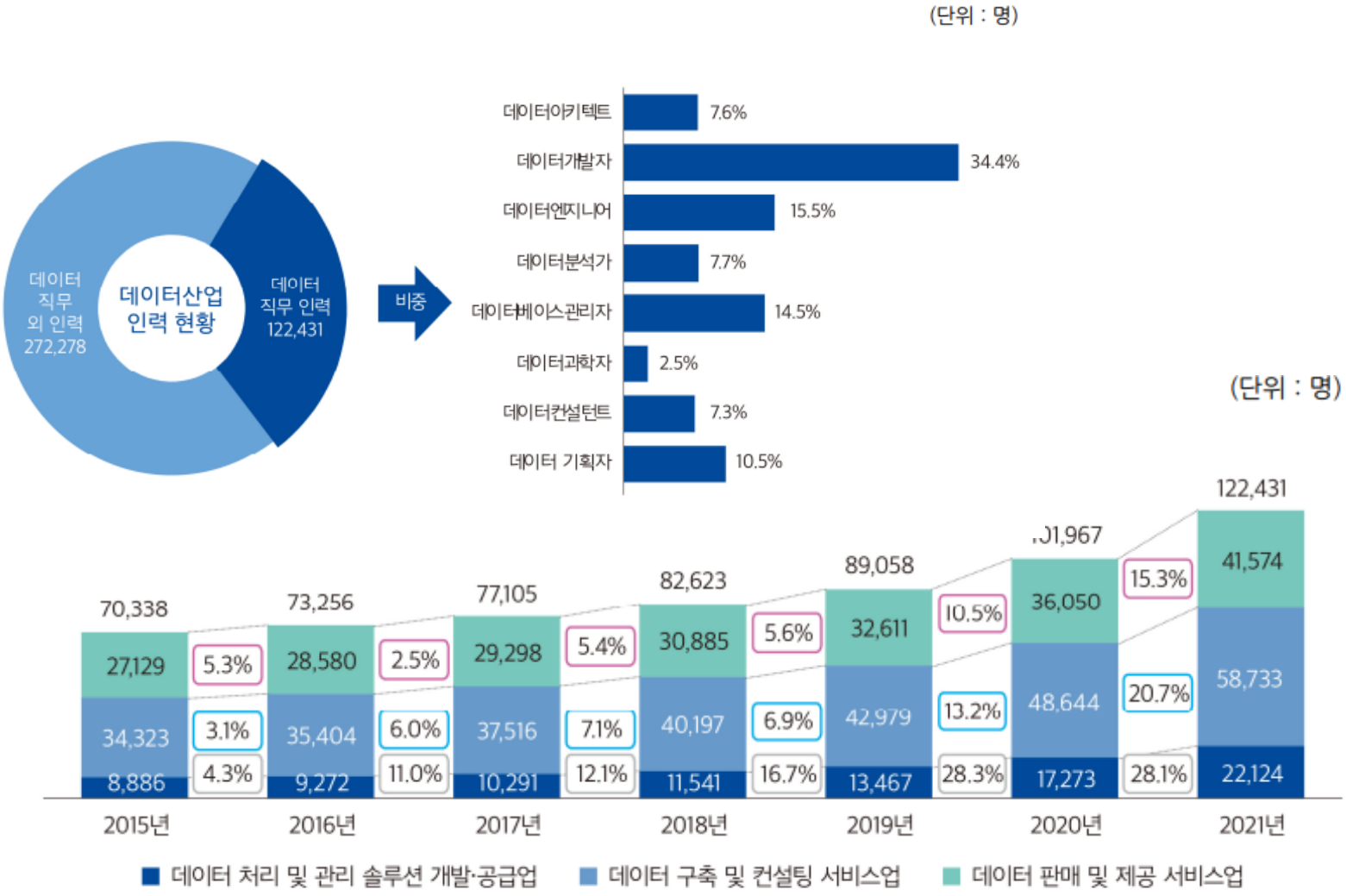
빅데이터 처리 및 실습 강의 개요

- **이번 강의에서는 주로 아래 framework를 다룸**
  - Data Ingestion: Sqoop, Flume, Kafka
  - Processing Engine: MapReduce, Spark
  - SQL on Hadoop: Hive



Major Hadoop Ecosystem (based on Hadoop 2.0).

데이터 관련 몇몇 통계



출처: 2021 데이터산업 현황조사, 한국데이터산업진흥원, 2022. 3.

데이터 관련 채용


185개의 포지션
DBA
핀다(FINDA) 평균 1주 이내 응답
3 ~ 4년 대한민국 서울특별시
DBA
자연어처리 엔지니어 (Natural Languag
로민 평균 3일 이내 응답
경력 무관 서울 서초구
머신러닝 인공지능(AI) 데이터 엔지니어 외 6개
Data Analyst (데이터 분석가)
띵스플로우 평균 6일 이내 응답
3 ~ 4년 대한민국 서울특별시
데이터 엔지니어
[강남/신사]병원컨설팅 DA담당
리팅랩스 평균 1주 이내 응답
2 ~ 5년 서울특별시 강남구
DBA 데브옵스 SQL 외 2개
데이터 엔지니어 채용
지원하기: 비즈니
5일 전
수행업무
AI/빅데이터 인프라
Data Mart 구축
내/외부 데이터 수
SQL 데이터처리 (
필수사항
데이터엔지니어 관
Hadoop 에코시스템
대용량 데이터 처리
실시간 데이터 분석
클라우드 환경(AW
우대사항
컴퓨터 공학계열 또
Hadoop 운영 및 퍼포먼스 튜닝에 대한 이해와 경험
Java, Python중 하나 이상을 이용한 개발 경험
머신러닝 모델 구현 및 실제 서비스 적용에 대한 이해와 경험
AICC 구축 경험자
SK텔레콤
Cloud기반 Data Platform Engineer
지원하기
이런 일을 합니다.
주요 수행업무 및 역할
SKT 전사 Cloud 기반 DataLake 개발 및 운영
Cloud 기반 Spark 수행 환경 구축 및 잡 진단/최적화
이런 분을 찾습니다.
필요 역량 및 경험
Cloud기반 Spark 수행 환경 구축 및 운영 경험
Spark 잡 진단/최적화 경험
Cloud 보안, 네트워크, 거버넌스 등 공통 기능 설계 및 구축 경험
Container 기반의 데이터 플랫폼 구축 및 운영 경험
Java, Scala, Python 프로그래밍 및 SQL 활용 능력
대용량 분산처리 환경기반에서의 논리적인 문제 해결 능력
자격요건
총 경력 : Data Platform 개발/운영 경력 5년 이상
학력/전공 : 학사 이상 / 전공 무관
이런 경험이 있다면 더욱 좋습니다.
우대사항
대규모 On-Premise Hive에서 Cloud Spark 잡으로 전환 경험
Spark 기반 데이터 솔루션 개발 경험
Public Cloud 기반 인프라 솔루션 개발 경험
다양한 Data Platform 컴포넌트(hadoop, spark, hive, kafka, flink, presto,..)에 대한 경험

빅 데이터 처리 및 실습 강의

- **강의 방식**
  - 이론과 실습을 병행
  - 실습
    - 맵리듀스 프로그램 작성
    - 스파크 프로그램 작성
    - Hive 또는 Pig 활용 대규모 batch 처리 스크립트 작성
    - Sqoop(또는 Flume, Kafka)을 이용한 데이터 수집기 작성

- **팀 프로젝트**
  - 팀 구성: 4명 한조로 구성
  - 방식: 11주차부터 대규모 데이터 처리 알고리즘 관련 논문 리뷰 및 재현
  - 최종발표: 최종산출물 관련 발표 및 보고서 작성
    - 전체적인 개발 개요
    - 성능을 높이기 위한 시도
    - 성능 벤치마킹(1 PC vs 10 PC)

강의 일정

주차	이론	실습	비고
1	교과목 소개 및 빅 데이터 개요	안전교육	
2	빅 데이터 주요 처리 과정	VM 환경 구축 및 리눅스 실습	
3	빅 데이터 처리 기술 (여러 프레임워크 간략하게 리뷰)	Local 버전 하둡 설치	
4	빅 데이터 처리 기술 (하둡 분산 파일 시스템 상세)	맵리듀스 알고리즘 (Database Join)	
5	빅 데이터 처리 기술 (맵리듀스 상세)	맵리듀스 알고리즘 (반복적인 방법)	
6	빅 데이터 처리 기술 심화 (맵리듀스 알고리즘 I)	클러스터 버전 하둡 설치	변경가능
7	빅 데이터 처리 기술 심화 (맵리듀스 알고리즘 II)	클러스터 상 맵리듀스 알고리즘 (Database Join)	변경가능
8	중간고사	중간고사	
9	SQL-Like 시스템 기반 빅 데이터 처리 기술 (Hive I)	Hive 설치 및 Hive 실습	
10	SQL-Like 시스템 기반 빅 데이터 처리 기술 (Hive II)	Hive 실습	
11	빅 데이터 수집 기술 (Flume, Kafka, Sqoop)	Hive 실습	프로젝트 시작
12	빅 데이터 수집 기술 (Flume, Kafka, Sqoop)	수집기(Flume, Kafka, Sqoop) 실습	
13	빅 데이터 프레임워크 활용 데이터 처리 사례 소개	프로젝트 진행	
14	프로젝트 발표	프로젝트 진행	
15	기말고사	기말고사	

평가 방법

- **평가 비율**
  - 출석(5%)
  - 과제(15%)
    - 문제풀이, 프로그래밍 등 다양한 유형
    - 마감일 넘기는 경우 1일마다 20% 감점
  - 중간 시험(30%)
  - 기말 시험(30%)
  - 팀 프로젝트(20%)

- **기타사항**
  - 1회 결석 OK
  - 2회, 3회 결석 시 3점 씩 감점
  - 학칙에 의거 4회 이상 결석 시 F

이 수업이 끝날 때 얻을 것들

- **빅 데이터 시스템의 구성과 동작방식 (이론)**
- **빅 데이터 시스템의 활용방법 (실습)**
  - 맵리듀스 및 스파크 프로그래밍
  - 대규모 데이터 처리를 위한 스크립트 프로그래밍 (Hive 또는 Pig)
- **맵리듀스 기반 데이터처리 알고리즘 구현 (프로젝트)**
- **한 학기 동안 즐거운 빅 데이터 처리 및 실습 강의가 되길 바라겠습니다!**

감사합니다!

담당교수: 전강욱(컴퓨터공학부)
kw.chon@koreatech.ac.kr