교육
과정
소개서.

Spark & Elastic Stack을 활용한 데이터 분산 처리

Fast campus

강의정보

강의장	온라인 강의	데스크탑, 노트북, 모바일 등
수강 기간	평생 소장
상세페이지	https://fastcampus.co.kr/data_online_ddp
강의시간	33시간 35분
문의	고객센터

강의특징

나만의 속도로	낮이나 새벽이나 내가 원하는 시간대에 나의 스케줄대로 수강
원하는 곳 어디서나	시간을 쪼개 먼 거리를 오가며 오프라인 강의장을 찾을 필요 없이 어디서든 수강
무제한 복습	무엇이든 반복적으로 학습해야 내것이 되기에 이해가 안가는 구간 몇번이고 재생

강의목표

- 분산 시스템 기본 개념부터 데이터 파이프라인 구축까지 분산 처리의 처음과 끝을 모두 배웁니다.
- Spark 기반으로 한 배치, 스트리밍 파이프라인과 Elastic Stack & Kafka 기반으로 한 실시간 처리 파이프라인을 직접 구축해 봅니다.
- 다양한 도메인 및 실시간 & 배치 환경에 따른 파이프라인 개요와 구성에 대해 학습하고 직접 실습해 봅니다.

강의요약

- 빅데이터 분산 처리 전문가분들과 함께 데이터 분산 처리의 기초 이론부터 실습 및 실무에서 활용되는 기술셋과 핵심 툴을 다루어 봅니다.
- Spark 기반으로 한 배치, 스트리밍 파이프라인과 Elastic Stack, Kafka 기반으로 한 실시간 처리 파이프라인을 직접 구축해 봅니다.
- 각각의 툴들이 가지고 있는 장점을 파악하여 데이터 파이프라인에 어떻게 활용되는지 배워봅니다.
- 질의응답 채널을 통해 강사님, 다른 수강생분들과 함께 문제를 해결할 수 있어요.다양한 꿀팁과 실무 이야기까지 공유해 보세요! (질의응답 채널은 2023.03.03 ~ 2025.03.03 까지 운영됩니다.)

강사

엄현호	**과목**	-	Spark & Elastic Stack을 활용한 데이터 분산 처리
	약력	-	현) 쿠팡 데이터 엔지니어
Jake	**과목**	-	Spark & Elastic Stack을 활용한 데이터 분산 처리
	약력	-	현) 유니콘 커머스 스타트업 데이터 엔지니어

CURRICULUM

01.

데이터 엔지니어링과 분산처리 기본 개념

파트별 수강시간 00:47:30

데이터 엔지니어링과 분산 처리 기본 개념
강의 소개 및 데이터 엔지니어링의 이해
데이터 엔지니어가 하는일
분산 시스템에 대한 이해 1
분산 시스템에 대한 이해 2
데이터 파이프라인의 종류
Data Observability

CURRICULUM

02.

Kafka 이해하기

파트별 수강시간 03:34:26

Kafka 란?
Kafka의 역사, 기본 구조 소개
매니지드 or 온프레미스 카프카 어떻게 결정할까
Kafka 의 구조
카프카 브로커의 이해
카프카 클러스터와 주키퍼
카프카의 토픽과 파티션
카프카의 레코드와 세그먼트
카프카에서 복제란
ISR에 대한 이해
Kafka 프로듀서
카프카 프로듀서의 이해
카프카 프로듀서 옵션들
카프카 프로듀서 acks
Kafka 컨슈머
카프카 컨슈머의 이해
카프카 컨슈머 그룹의 이해
컨슈머 랙
카프카 컨슈머 옵션들
Kafka 설치
로컬 환경 구성
로컬에 Kafka를 설치하기
Kafka 모니터링 툴 설치하기
로컬에 Kafka Producer, Consumer 설정하기
브로커 실행하기
AWS EC2에 카프카 설정하기1
AWS EC2에 카프카 설정하기2
AWS EC2에 Kafka 모니터링 툴 설치하기
Kafka Producer 설정
실시간 데이터를 카프카로 전송하기

CURRICULUM

03.

ElasticSearch, ELK스택

파트별 수강시간 05:11:45

ElasticSearch란?
ElasticSearch에 대한 이해
ELK 스택에 대한 이해 (1)
ELK 스택에 대한 이해 (2)
Elasticsearch의 기본 구성
ElasticSearch 아키텍쳐
ElasticSearch 인덱스, 샤드, 레플리카
역색인, Inverted Index
AWS EC2에 ElasticSearch 설치하기
Elasticsearch와 Opensearch 그리고 버전
EC2 설정, ElasticSearch설치
ES둘러보기, elasticsearch.yaml 설정
ElasticSearch 클러스터 구성 1
ElasticSearch 클러스터 구성 2
ElasticSearch 모니터링
간단한 모니터링 설치
ElasticSearch TLS 적용하기
TLS적용을 통한 보안설정 1
TLS적용을 통한 보안설정 2
AWS EC2에 Kibana 설치하기
Kibana 설치
Kibana 둘러보기
Kibana로 Stack Monitoring설정
ElasticSearch API
DocumentAPI
SearchAPI, Query DSL 1
SearchAPI, Query DSL 2
ElasticSearch Aggregation
Metrics, Bucket Aggregation
AggregationAPI 1
AggregationAPI 2

CURRICULUM

03.

ElasticSearch, ELK스택

파트별 수강시간 05:11:45

ElasticSearch Mapping API
Mapping
다양한 데이터 타입 소개-
AWS EC2에 Logstash 설치하기
로컬에 Logstash 설치
EC2에 Logstash 설치
Filebeat와 Logstash로 데이터 전송하기
Filebeat 설치 및 설정
apache log 전송하기
Lifecycle설정 적용하기
Text Tokenizer, Nori
Nori란, Nor설치
Nori 실습, 데이터 준비
Tag Cloud 생성하기

CURRICULUM

04.

Workflow Orchestration, Airflow

파트별 수강시간 03:47:39

Airflow란?
워크플로우 오케스트레이션 이해
Airflow의 역사, 특성
Airflow의 구성요소
싱글 노드 구조, 멀티 노드 구조
DAG
Operators 살펴보기
Cloud Managed Service of Airflow
Luigi, Prefect, Oozie 비교
AWS EC2에 Airflow 설치하기
Airflow 기본적인 설치, 싱글구조 살펴보기
Airflow Webserver 살펴보기
airflow.cfg 살펴보기
클러스터 구성하기1
클러스터 구성하기2 master node 설정 끝내기
클러스터 구성하기3 DB설정
클러스터 구성하기4 Worker 설정
Kubernetes에 Airflow 설치하기
Rancher란
EKS 생성마무리
Rancher 둘러보기
Airflow on k8s 살펴보기
sidecar 패턴, gitsync 설정
KubernetesPodOperator 사용 설정

CURRICULUM

05.

실전 프로젝트

파트별 수강시간 01:28:23

스트리밍 데이터 파이프라인 구성하기
스트리밍 파이프라인 설계하기
스트리밍 데이터 파이프라인 구성1 - Kafka
스트리밍 데이터 파이프라인 구성2 - ELK
스트리밍 데이터 파이프라인 구성3 - Kibana
배치 데이터 파이프라인 구성하기
배치 파이프라인 설계하기
배치 데이터 파이프라인 구성 - 환경구성
배치 데이터 파이프라인 구성 - Airflow
배치 데이터 파이프라인 구성 - MySQL
배치 데이터 파이프라인 구성 - DAG run

CURRICULUM

06.

Kubernetes

파트별 수강시간 01:19:39

Kubernetes란?
Kubernetse 소개mp4
Kubernetse를 위한 Docker
Kubernetes의 기본 개념과 구성
클러스터
워크로드 - Pod
워크로드 - Deployment
워크로드 - ReplicaSet, StatefulSet, DaemonSet
Service
Storage
Kubernetes 기타 개념
Helm Chart
Yaml

CURRICULUM

07.

강의 소개 및 개발 환경 구성

파트별 수강시간 00:08:45

강의소개 - 강의 소개 및 목차 안내
분산 시스템(빅데이터 시스템)의 요구 사항 정리

CURRICULUM

08.

Hadoop이란?

파트별 수강시간 01:15:11

Hadoop 개요
Hadoop이란?
Hadoop의 핵심 구성 요소
Hadoop의 핵심 구성 요소 1) - HDFS (파일 시스템)
Hadoop의 핵심 구성 요소 2) - MapReduce (배치 처리 알고리즘)
Hadoop 에코시스템 프레임워크 둘러보기
Hadoop 에코시스템 프레임워크 둘러보기 1) - Hbase (데이터베이스)
Hadoop 에코시스템 프레임워크 둘러보기 2) - Yarn (리소스 매니저)
Hadoop 에코시스템 프레임워크 둘러보기 3) - Zookeeper (분산 코디네이터)
Hadoop 에코시스템 프레임워크 둘러보 - Avro(데이터 직렬화)
Hadoop 에코시스템 프레임워크 둘러보기 5) - Hive (데이터 분석)

CURRICULUM

09.

Apache Spark

파트별 수강시간 11:07:50

스파크 개요
Apache Spark란?
로컬 환경에 스파크 설치 및 워드 카운트 예제 실행
스파크 애플리케이션의 기본 구성
Transformation, Action, Lazy Evaluation 의 개념
RDD
스파크 RDD란?
RDD 실습 - 로그 집계 파이프라인 만들기 - map, filter, reduce
RDD - RDD 실습 - join
RDD - RDD 실습 - 실전 예제(F사의 강의별 랭킹 데이터 만들기) - Part 1
RDD - RDD 실습 - 실전 예제(F사의 강의별 랭킹 데이터 만들기) - Part 2
SparkSQL, DataFrame, Dataset
스파크 SQL,Dataframe,Dataset이란
SparkSQL,Dataframe,Dataset - SQL 실습 - 로그 집계 파이프라인 만들기 - Dataframe API
SparkSQL,Dataframe,Dataset - 스파크 DataFrame, Dataset, SQL 실습 - 로그 집계 파이프라인 만들기 - SQL API
SparkSQL,Dataframe,Dataset - 스파크 DataFrame, Dataset, SQL 실습-로그 집계 파이프라인 만들기 - join
SparkSQL,Dataframe,Dataset SQL 실습-로그 집계 파이프라인 만들기-실전 예제 1
SparkSQL,Dataframe,Dataset - 스파크 DataFrame, Dataset, SQL 실습-로그 집계 파이프라인 만들기-udf(사용자 정의 함수)
RDD, SparkSQL, DataFrame 비교
RDD, SparkSQL, DataFrame 비교 - RDD, DataFrame 언제 사용해야 하는가
스파크 심층분석
스파크 클러스터, 런타임 아키텍처에 대한 이해
spark0submit 주요 파라미터 확인
Deploy mode -cluster, client mode
스파크 - action, stage, shuffle, task, slot 확인 실습
Join의 종류
스파크에서의 메모리 할당
스파크 메모리 관리
Repartition, Coalesce에 대한 이해
Caching, Persistence에 대한 이해
Shared variable(Accumulator, Broadcast variable)에 대한 이해
스파크 Query Plan
Dynamic resource allocation
Spark Scheduler에 대한 이해
(Spark 3.0) AQE - Adaptive Query Execution에 대한 이해_1
(Spark 3.0) DPP - Dynamic Partition Pruning에 대한 이해

CURRICULUM

09.

Apache Spark

파트별 수강시간 11:07:50

Partitioning
Partitioning 개요, 중요성
Custom Partitioner
partition by vs bucket by
Input Partitions From Data Files
Partitioning during Spark Transformations
Partitioning to Output Files
스파크 실무 팁
데이터 파이프라인 운영시 발생했던 스파크 에러들에 관한 case study
스파크 튜닝 팁
Spark Unit test 작성
Spark Streaming
Spark Structured Streaming이란
Spark Structured Streaming 실시간 로그 집계 파이프라인 만들기 Part1
Structured Streaming 실습 - 실시간 로그 집계 파이프라인 만들기 Part 2)
Dstream이란?
Dstream 실습 - 로그 집계 파이프라인 만들기 Part 1)
Dstream 실습 - 로그 집계 파이프라인 만들기 Part 2)
Event Time windows, Processing Time Windows 실습
Watermarking 개념 및 실습

CURRICULUM

10.

컬럼 기반의 NoSQL

파트별 수강시간 02:42:30

컬럼 기반 NoSQL 개요
컬럼 기반 NoSQL 을 사용하는 이유
Cassandra
Cassandra 개요
Cassandra Data Model
CQL (Cassandra Query Language) 소개
CQL에서 Partition Key, Clustering Key의 제약 조건
CQL에서 Secondary index의 제약 조건
Consistency
Storage Component의 종류
데이터 쓰기 과정
데이터 읽기 과정

CURRICULUM

11.

File system

파트별 수강시간 00:23:21

파일 시스템 개요
컬럼 기반의 파일 포맷 - Apache parquet
Data warehouse, lake, lakehouse 개념 소개 및 비교

CURRICULUM

12.

이커머스 상품 데이터 파이프라인 구축 실습

파트별 수강시간 01:48:22

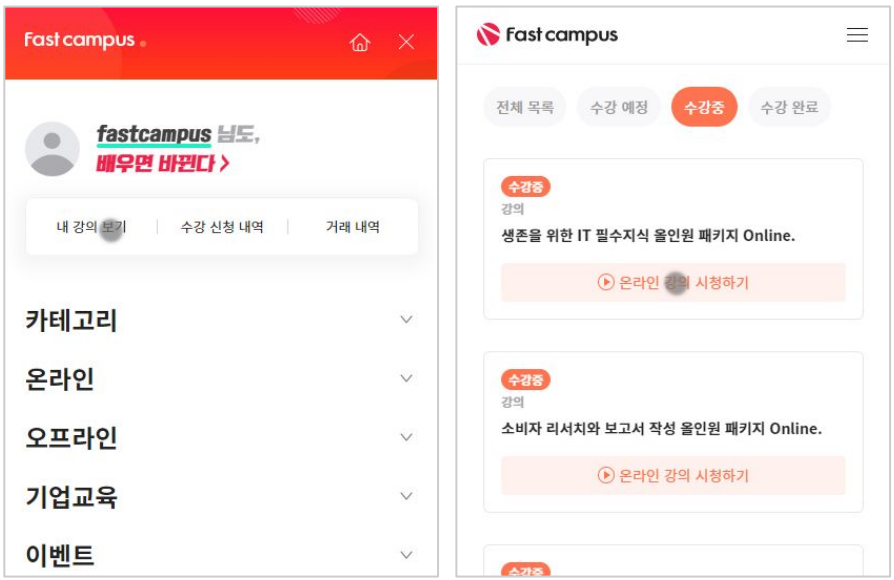
로컬 파이프라인 구축
요구 조건 정의
Spark 코드 작성 Part 1)
Spark 코드 작성Part 2)
Spark 코드 작성 Part 3)
Cassandra에 데이터 쓰기
Apache Iceberg에 데이터 쓰기
AWS EMR에 배포
AWS EMR에서 Spark application 실행

주의 사항

- 상황에 따라 사전 공지 없이 할인이 조기 마감되거나 연장될 수 있습니다.
- 패스트캠퍼스의 모든 온라인 강의는 **아이디 공유를 금지**하고 있으며 1개의 아이디로 여러 명이 수강하실 수 없습니다.
- 별도의 주의사항은 각 강의 상세페이지에서 확인하실 수 있습니다.

수강 방법

- 패스트캠퍼스는 크롬 브라우저에 최적화 되어있습니다.
- 사전 예약 판매 중인 강의의 경우 1차 공개일정에 맞춰 '온라인 강의 시청하기'가 활성화됩니다.



환불 규정

- 온라인 강의는 각 과정 별 '정상 수강기간(유료수강기간)'과 정상 수강기간 이후의 '복습 수강기간(무료수강기간)'으로 구성됩니다.
- 환불금액은 실제 결제금액을 기준으로 계산됩니다.

수강 시작 후 7일 이내	100% 환불 가능 (단, 수강하셨다면 수강 분량만큼 차감)
수강 시작 후 7일 경과	정상(유료) 수강기간 대비 잔여일에 대해 환불규정에 따라 환불 가능

※ 강의별 환불규정이 상이할 수 있으므로 각 강의 상세페이지를 확인해 주세요.