연구 내용 요약 및 계획

|  |  |
| --- | --- |
| 날짜 | 2020년 01월 17일 |
| **1. 이 번주 연구 관련 학습 내용 정리**  1. 버전관리 시스템 : 깃을 통한 저장소 생성 및 환경 구축 방법 학습  2. 빅데이터 : 빅데이터의 개념, HDFS의 개념과 역할, 하둡의 역할 학습  \*매일매일의 영상내용정리는 밑에 있습니다. | |
| **2. 다음 주 학습 계획**  - 데이터 사이언스 기초 강의  (<https://www.edwith.org/datascience/joinLectures/19265>)  -kaggle 을 이용한 데이터 사이언스 실습  <https://www.edwith.org/boostcourse-ds-kaggle/joinLectures/28015>  -위에 진행한 내용 깃에 commit하고 버전 관리하는 방법 습득 및 실습 | |
| **3. 일주일 간의 학습 내용 정리**  1. 깃  깃과 깃허브는 다르다 깃은 파일에 관한 모든 내용을 기록하는 것이고 깃허브는 그 내용을 클라우드에 저장하여 다른사람과 공유하기 위한 것이다.  2. 빅데이터  빅데이터를 저장과 분산을 위해 하둡파일 시스템을 사용한다. HDFS라고부르는데 이는 많은 데이터를 여러 컴퓨터에 나눠서 저장해주고 명령을 내리면 각자의 컴퓨터에서 분석해서 그 결과를 합쳐 최종결과를 내준다. 하둡외에도 cassandra, spark등이 있다.  HDFS를 사용하기 위해서 도와주는 툴들이 많이 있다. mapreduce, yarn등이 분산 컴퓨팅, 분산 프로그램을 도와주고, hive, impala, pig등은 sql을 사용하여 hdfs를 사용할 수 있게 도와준다. | |

20.01.09

깃 기초 공부

깃과 깃허브는 다른것이다. 깃은 파일에 관한 모든 내용을 기록하는 것이고

깃허브는 그 내용을 클라우드에 저장하여 다른 사람들과 공유할 수 있도록 하는 것이다

출처: <https://www.youtube.com/watch?v=YFNQwo7iTNc>

깃을 통한 저장소 생성 및 환경 구축, commit, add, branch, merge, rebase, revert등의 명령어

출처 : <https://www.youtube.com/watch?v=FXDjmsiv8fI>

빅데이터

빅데이터란 무엇이며 현실세계에서 어떻게 사용되고 있는지를 알아봄.

현재 빅데이터는 석유와 같다라고 불리는 자산이 되어가고 있다. 또한 우리가 실생활에서 생각보다 많은 부분에서 사용되고 있다. 이를 더욱더 많이 쉽게 사용하도록 하기 위해서 구글은 많은 개발을 하고 배포를 하는 중이라고 한다. 대표적인 예가 tensorflow이며 이를 이용한 고래 추적, 안과검사 등의 프로그램을 만들어서 사용중이라고 한다.

초기 빅데이터는 컴퓨터에게 세상 모든 내용을 익히도록 하여 컴퓨터가 똑똑해지도록 하는 거 였다면 현재는 컴퓨터가 주제를 이용하여 스스로 학습하여 똑똑해지도록하는 머신러닝을 사용한다.

출처 : <https://www.youtube.com/watch?v=ZG49_wOBmqg>

<https://www.youtube.com/watch?v=Spp39BgpsNk>

<https://www.youtube.com/watch?v=WAWTSavxvAE>

<https://www.youtube.com/watch?v=BKj3fnPSUIQ>

20.01.13

<https://www.youtube.com/watch?v=j-0cUmUyb-Y> 빅데이터 탐구 – 팀 스미스

정보의 변화 : 정적->유동적

의자에 센서를 달아서 자동차 운전자를 구별

머신러닝 : 문제와 데이터를 주고 학습하도록 하는 것

빅데이터의 문제점

우리가 예측에 의한 처벌을 받을 수도 있다.

일자리가 줄어든다. 화이트칼라 직원들이 필요없어진다. 연구자들과 같은

빅데이터를 우리 인간의 필요에 맞게 조정해야 한다.

<https://www.youtube.com/watch?v=8pHzROP1D-w> Kenneth Cukier :Bic data is better data

big data framework : Cassandra, Hadoop, spark

HDFS(Hadoop Distributed File System) is the storage layer of Hadoop

Data get stored in a distributed manner in HDFS

Smaller chunks of data are stored on multiple data nodes in HDFS

<https://www.youtube.com/watch?v=bAyrObl7TYE\>

Big Data In 5 Minutes | What Is Big Data?| Introduction To Big Data |Big Data Explained |Simplilearn

<https://www.youtube.com/watch?v=qXyzDd2heK8>

What is Big Data and why does it matter? | Donna Green | TEDxSouthamptonUniversity

20.01.14

<https://www.youtube.com/watch?v=lU9OLSVyIuw&t=113s> 빅데이터 개념 정리, 하둡 파일 시스템 (HDFS)의 탄생 배경

HDFS : 빅데이터의 저장과 분산을 위한 분산 컴퓨팅 시스템

거대한 데이터를 분산되어 있는 여러 컴퓨터에 나눠서 저장. 약 3대정도의 컴퓨터에 같은 내용의 데이터를 중복하여 저장.

각각의 데이터가 저장된 컴퓨터에게 명령을 내리면 각자 분석해서 그 결과를 합쳐 최종결과를 냄

아마존 클라우딩 컴퓨터를 이용해 여러 컴퓨터로 분석을하고 저장을 하지 않으면 비용 발생 x

<https://www.youtube.com/watch?v=HCR1ILMROfI> 생활코딩-하둡소개

가상머신을 이용해 여러대의 컴퓨터를 사용

네임노드는 데이터가 어느 데이터노드에 들어있는지를 저장하는 컴퓨터

Hadoop fs (파일 시스템 제어) -ls = 하둡에 저장된 파일 리스트

Hadoop fs -rm(r) ‘filename’ = 하둡에 저장된 파일 삭제(폴더)

Hadoop fs -put ‘filename’ = 하둡에 파일 저장

<https://www.youtube.com/watch?v=AZ136kwcsz8> 생활코딩-하둡-소개hello world

하둡 클라이언트는 파일을 잘게 조갠다.

클라이언트가 네임노드가 정해준 데이터 노드 중 하나에게 파일을 전송한다. 그러면 클라이언트로부터 파일을 받은 데이터 노드는 그 파일을 다른 지정된 데이터 노드에게 전송해준다.

전송받은 데이터노드가 다른 데이터노드에게 전송한 후 전송완료 메시지를 받는다. 다 받으면 그 데이터노드는 전송완료를 클라이언트에게 전송.

네임노드는 계속해서 데이터노드의 상태를 전송받는다. = heartbeat

<https://www.youtube.com/watch?v=xacUzaFYQCI> 하둡(HDFS) 의 파일 읽기, 쓰기 과정 쉽게 요약

20.01.15

파이썬이 R보다 범용적이고 파이썬이 더 배우기 쉽기 때문에 파이썬을 더 많이 사용함.

<https://www.youtube.com/watch?v=y25-gyUczUc> [데이터 분석]왜 파이썬인가

데이터 시각화는 데이터 분석결과를 이미지나 그래프로 표현하는 것

<https://www.youtube.com/watch?v=IZO5wq6qikg>빅데이터 분석 결과 시각화

20.01.16

MapReduce is processing paradigm that pairs with HDFS and is a distributed computation algorithm that compute down to each of the x86 servers

Mapreduce 는 대용량 데이터 처리를 위한 분산 프로그래밍 모델 타고난 병행성.

흩어져 있는 데이터를 key, value의 형태로 연관성 있는 데이터 분류로 묶는 작업=>map

map화한 작업 중 중복 데이터를 제거하고 원하는 데이터를 추출하는 작업=>reduce

Yarn is a resource manager for different workloads that you plug in on top of a do it manages computer resources and clusters and works as scheduled users applications

Yarn은 리소스 관리 플랫폼이다. 각 어플리케이션에서 필요한 리소스를 할당하고 모니터링함.

<contatiner = 리소스를 객체화한 것>

HDFS는 파일을 저장할 수 있는 분산파일 시스템이고 YARN은 분산 컴퓨팅 환경을 제공.

다수의 서버를 사용하는 경우 MAPRESURCE하나에만 할당하는 것이 비효율적이어서 SPARK. STORM과 같은 다른 컴퓨팅 시스템에 할당했으나 각자 다른 시스템을 할당하므로서 리소스 부족, 클러스터 사용 효율 저하등의 문제가 발생하여 yarn이라는 서브 프로젝트를 만들었다.

<https://www.popit.kr/what-is-hadoop-yarn/> (나중에 다시한번 보기)\

https://wikidocs.net/27357 (커멘드 정리-YARN)

HIVE는 HDFS의 데이터를 다룰 때 SQL쿼리로 쓸 수 있게 해주는 툴 데이터 처리 엔진에서 작업을 생산

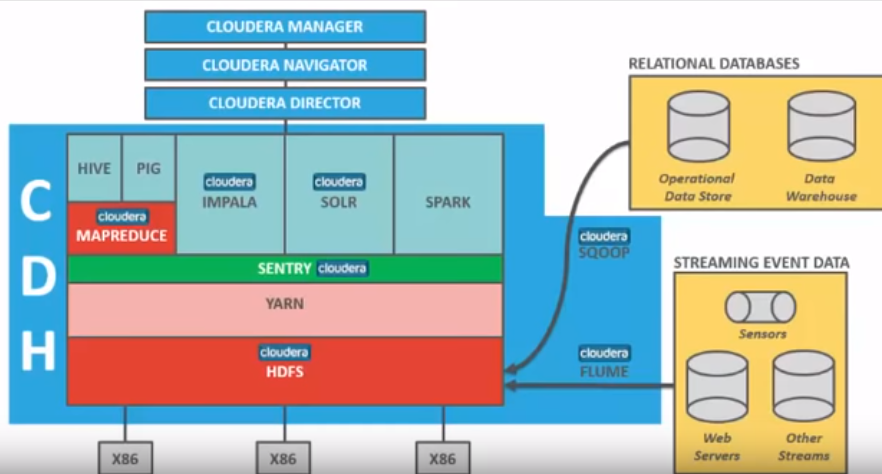
IMPALA는 HDFS의 데이터를 다룰 때 SQL쿼리로 쓸 수 있게 해주는 툴, 하툽 클러스터에서 직접 쿼리를 실행

=>거대한 양의 데이터 분석 가능, 생산적이며 다른 시스템과 호환 가능,

Pig는 대용량 데이터 분석 플랫폼.피그라틴이라는 텍스트기반의 언어 사용.(기반은 자바)

solr은 apache lucene를 기반으로 만든 빠른 검색플랫폼. 데이터 색인, 데이터 검색 프로세스를 lucene엔진을 사용하면서 http료청에 대한 처리와 응답을 하는 웹 기반 검색엔진.

sqoop(sql to Hadoop)rdb와 hadoop간 대량 데이터 전송이 가능. 이를 이용하면 데이터를 관계형 데이터베이스로부터 hdfs로 가져올 수 있다. 데이터 테이블 입력시 sqoop이 테이블을 행단위로 hdfs로 읽어온다.

flume은 로그 데이터 수집 기술. 여러 서버에서 발생한 로그 데이터를 모아서 hdfs에 전송하는 기능.

\*클라우데라는 아파치 하둡과 아파치 스파크 기반 소프트웨어, 지원 및 서비스, 그리고 트레이닝을 제공하는 미국의 소프트웨어 기업이다.

<https://www.youtube.com/watch?v=AZovvBgRLIY> Apache Hadoop & Big Data 101:The Basics