자바쌤~!

https://cafe.naver.com/javassem

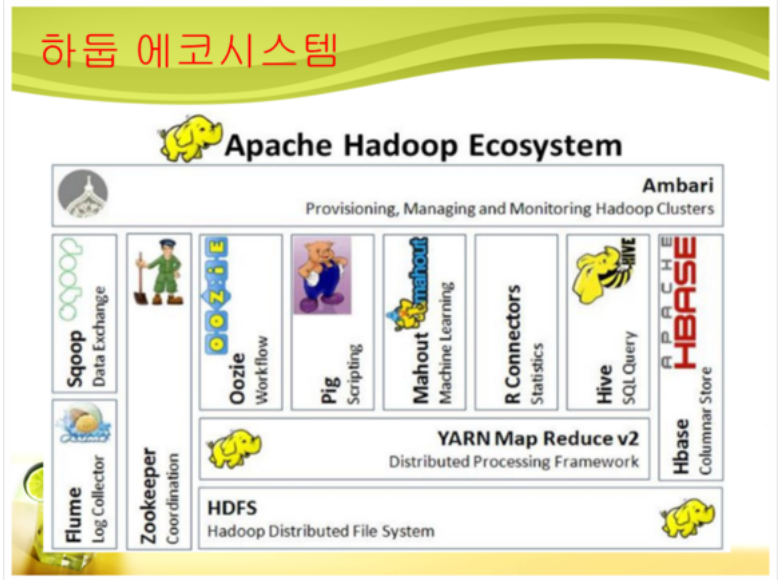
검색

63기-자바 & 빅데이타 >

[개념] 에코시스템

자바쌤 1:1 채팅
2020.06.21. 12:12 조회 12

댓글 0 URL 복사



카페정보 나의활동

매니저 자바쌤
since 2013.03.15.
카페소개

새싹2단계
437 초대하기
즐겨찾는 멤버 64명
게시판 구독수 3회
우리카페앱 수 0회

카페 글쓰기

카페 채팅

즐겨찾는 게시판
전체글보기 4,032

KOSMO
63기-자바 & 빅데이타 N

팁
심심문제
기타등등
JAVA
Oracle/Mysql
JSP
WebUI
파이썬
Spring
리눅스 N
하둡+eco
프로젝트

공유해요
알고리즘도전
기술면접 준비하기
실존 면접
기회는 준비된자에게
프로젝트 팁

쉬어가기

청년아카데미

전문가과정

공지사항
카페회칙

카페탈퇴하기

1- 하나의 서버에만 서비스가 집중되 않도록 분산해 동시에 처리
2- 하나의 서버에서 처리한 결과를 다른 서버와도 동기화해서 데이터의 안정성을 보장
3- 운영(active) 서버에 문제가 발생해서 서비스를 제공할 수 없는 경우, 다른 대기 중인 서버를 운영서버로 바꿔서 서비스 중지 없도록 해줌
4- 분산 환경을 구성하는 서버의 환경설정을 통합적으로 관리

하둡 에코시스템

2. 리소스관리

- Yarn

(http://hadoop.apache.org)

기존 하둡의 데이터 처리 프레임워크인 맵리듀스의 단점을 극복하기 위해 시작된 프로젝트로 하둡 2.0 부터 사용한다.

즉, Yarn은 데이터 처리 작업을 실행하기 위한 클러스트 자원(CPU, 메모리, 디스크 등)과 스케쥴링을 위한 프레임워크이다.

맵리듀스, 하이브, 임팔라, 타조, 스파트 등 여러 어플리케이션들은 Yarn 에서 리소스를 할당 받아 작업을 실행한다.

하둡 에코시스템

2. 리소스관리

- Mesos

(http://mesos.apache.org)

클라우드 인프라스트럭쳐 및 컴퓨팅 엔진의 다양한 자원(CPU, 메모리, 디스크 등)을 통합적으로 관리할 수 있도록 만든 자원 관리 프로젝트이다.

Facebook, Ebay, Twitter, AirBnB 등 글로벌 기업들이 메소드로 클러스터 자원을 관리하고 있다. 1만대 이상의 노드에서도 대응 가능하며,

Hadoop, Spark, Storm, Elastic Search, Cassandra, Jenkins 등 다양항 어플리케이션을 메소스에서 실행할 수 있다.

하둡 에코시스템

3. 데이터 저장

- Hbase

(http://hbase.apache.org)

구글의 빅테이블(big table) 논문을 기반으로 개발된 HDFS 기반의 컬럼 기반 데이터베이스이다.

실시간 랜덤 조회 및 업데이트가 가능하며, 각 프로세스는 개인의 데이터를 비동기적으로 업데이트 할 수 있다. 단, 맵리듀스는 일괄 처리 방식으로 수행된다. Twitter, Yahoo, Adobe 등 해외 업체에서 사용하고 있으며 ,

국내에서는 네이버가 모바일 메신저 라인에 Hbase를 적용한 시스템 아키텍쳐를 2012년도에 발표하였다.

하둡 에코시스템

3. 데이터 저장

- Kudu

(http://getkudu.io) - 클라우데라에서 시작한 프로젝트

컬럼 기반의 스토리지로서, 특정 컬럼에 대한 데이터 읽기를 고속화 할 수 있다. 기존에 HDFS에서도 Parquet(파케이), RC, ORC와 같은 파일 포맷을 사용하면 컬럼 기반으로 데이터를 저장할 수 있지만 HDFS 자체가 온라인 데이터 처리에 적합하지 않다는 약점이 있었다.

HDFS 기반으로 온라인 처리가 가능한 Hbase의 경우 데이터 분석 처리가 느리다는 단점이 있다.

하둡 에코시스템

4. 데이터 수집

- Chukwa (척와)

(http://chukwa.apache.org)

척와는 분산 환경에서 생성되는 데이터를 HDFS에 안정적으로 저장하는 플랫폼이다.

분산된 각 서버에서 Agent를 실행하고 Collector가 Agent로부터 데이터를 받아 HDFS에 저장한다. 콜렉터는 100개의 에이전트당 하나씩 구동되며, 데이터 중복 제거 등의 작업은 맵리듀스로 처리한다.

하둡 에코시스템

4. 데이터 수집

- Flume

(http://flume.apache.org)

플롬은 척와처럼 분산된 서버에 에이전트가 설치되고, 에이전트로부터 데이터를 전달받는 콜렉터로 구성된다.

전체 데이터의 흐름을 관리하는 마스터 서버가 있어서 데이터를 어디서 수지하고, 어떤 방식으로 전송하고, 어디에 저장할 지를 동적으로 변경할 수 있다.

하둡 에코시스템

4. 데이터 수집

- Sqoop

(http://sqoop.apache.org)

스쿱은 대용량 데이터 전송 솔루션으로 HDFS, RDBMS, DW, NoSQL 등 다양항 저장소에 대용량 데이터를 신속하게 전송하는 방법을 제공한다.

- Hiho (http://github.com/sonalgoyal/hiho)

스쿱과 같은 대용량 데이터 전송 솔루션으로, 하둡에서 데이터를 가져오기 위한 SQL을 지정할 수 있으며, JDBC 인터페이스를 지원한다. (현재 Oracle과 Mysql만 지원)

하둡 에코시스템

4. 데이터 수집

- Kafka

(http://kafka.apache.org)

카프카는 데이터 스크림을 실시간으로 관리하기 위한 분산 메세징 시스템이다.

발행(publish) - 구독(subscribe) 모델로 구성되어 있으며, 데이터 손실을 막기 위하여 디스크에 데이터를 저장한다.

카프카는 링크드인에서 자사의 대용량 이벤트를 처리하기 위해 개발된 이후 하루에 1조 1천억건의 이상의 메시지를 카프카에서 처리하고 있다.

하둡 에코시스템

5. 데이터 처리

- Pig

(http://pig.apache.org)

피그는 야휴에서 개발하여 현재 아파치 프로젝트이다.

복잡한 맵리듀스 프로그래밍을 대처할 피그 라틴 (Pig Latin) 자체 언어를 제공한다.

맵리듀스 API를 매우 단순화한 형태로 SQL 유사한 형태로 설계되었지만 기존 SQL 과는 동일하지는 않다.

하둡 에코시스템

5. 데이터 처리

- Mahout

(http://mahout.apache.org)

머하웃은 하둡 기반으로 데이터 마이닝 알고리즘을 구현한 오픈 소스 프로젝트이다.

현재 Classification, Clustering, Recommenders/Collaborative filtering, Pattern Mining, Regression, Dimension reduction, Evolutionary Algorithms 등 주요 알로리즘을 지원한다.

하둡 에코시스템

5. 데이터 처리

- Spark

(http://spark.apache.org)

스파크는 인메모리 기반의 데이터 처리 플랫폼이다.

배치 처리, 머신 러닝, SQL 질의 처리, 스트리밍 데이터 처리, 그래프 라이브러리 처리와 같은 다양한 작업을 할 수 있다.

현재 가장 빠르게 성장하고 있는 오픈소스 프로젝트이다

하둡 에코시스템

5. 데이터 처리

- Impala

(http://impala.io)

임팔라는 클라우데라에서 개발한 하둡 기반의 분산 쿼리 엔진이다.

맵리듀스를 사용하지 않고, C++로 개발한 인메모리 엔진을 사용해 성능이 빠르다.

임팔라는 데이터 조회를 위한 인터페이스로 HiveQL을 사용하여 SQL 질의 결과를 빠르게 확인한다.

하둡 에코시스템

5. 데이터 처리

- Hive

(http://hive.apache.org)

하이브는 페이스북에서 개발된 하둡 기반의 데이터웨어하우징용 솔루션이다.

SQL과 유사한 HiveSQL이라는 쿼리 언어를 사용하는데, 이는 내부적으로 맵리듀스 잡으로 변환되어 실행한다.

하둡 에코시스템

5. 데이터 처리

- Tajo

(http://tajo.apache.org)

타조는 고려대학교 박사 과정 학생들이 주도해서 개발한 하둡 기반의 데이터웨어하우스 시스템으로 아파치 프로젝트로 선정된 이후 2014년도 최상의 프로젝트로 승격되었다.

맵리듀스 엔진이 아닌 자체 분산 처리 엔진을 사용하고,

다른 시스템과는 다르게 HiveQL이 아닌 표준 SQL을 지원하는 것이 큰 특징이다.

하둡 에코시스템

6. 워크플로우

- Nifi

(http://nifi.apache.org)

나이파이는 데이터 흐름을 모니터링하기 위한 프레임워크이다.
여러 네트워크를 통과하는 데이터 흐름을 웹UI에서 그래프로 표현하며,
프로토콜과 데이터 형식이 다르더라도 분석이 가능한다.
원래 미국국가안보국(NSA)에서 개발한 기술로 공개된 오픈소스 기술이다.

하둡 에코시스템

6. 워크플로우

- Oozie

(http://oozie.apache.org)

우지는 하둡 작업을 관리하는 워크플로우 및 코디네이터 시스템이다.
자바 서블릿 컨테이너에서 실행되는 자바 웹 어플리케이션 서버이며, 맵리듀스 작업이나 피그 작업 같은 특화된 액션으로 구성된 워크플로우를 제어한다.

하둡 에코시스템

6. 워크플로우

- Airflow

(http://nerds.aribnb.com/airflow)

에어플로우는 에어비앤비에서 개발한 워크플로우 플랫폼이다.
데이터 흐름의 시각화, 스케쥴링, 모니터링이 가능하다
Hive, Presto, DBMS 엔진과 결합해서 사용할 수 있다

하둡 에코시스템

6. 워크플로우

- Azkaban

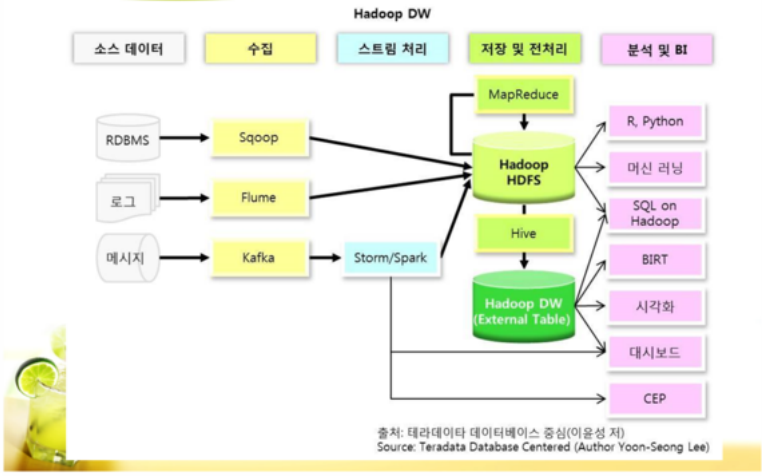
(http://Azkaban.github.io)

아크카반은 링크드인에서 개발한 워크플로우 플랫폼이다.

링크드인은 자사의 복잡한 데이터 파이프라인을 관리하기 위해 아즈카반을 개발하고 오픈하였다.

아즈카반은 워크플로우 스케쥴러, 시각화된 절차, 인증 및 권한 관리, 작업 모니터링 및 알람 등 다양한 기능은 웹 UI로 제공한다.

하둡 에코시스템



주요사이트

- http://hadoop.apache.org
- http://chukwa.apache.org
- http://flume.apache.org
- http://oozie.apache.org
- http://sqoop.apache.org
- http://hive.apache.org
- http://pig.apache.org
- http://mahout.apache.org
- http://zookeeper.apache.org

댓글 댓글알림

김용선

댓글을 남겨보세요

등록

자바쌤~! https://cafe.naver.com/javassem