if(kakao)2021

카카오 공용 하둡 운영 사례

설계 시 고려사항 및 장애 대응 사례 공유

이재영 Jace.Beleren
카카오

카카오 공용 하둡 소개

설계 시 고려사항

장애 대응 사례

공용 하둡

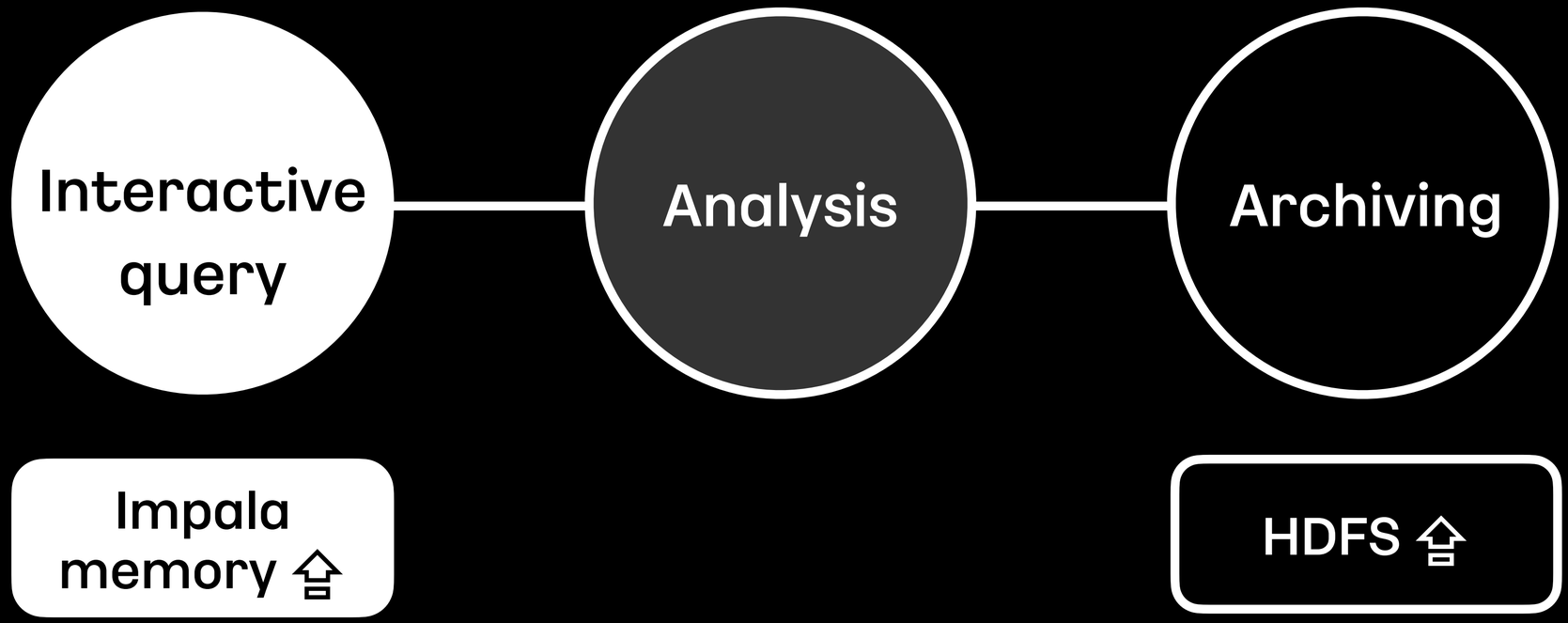
Multi-tenant hadoop cluster

카카오에서 서비스하는 공용 하둡

카카오에서 서비스하는 공용 하둡 (2)

Analysis

카카오에서 서비스하는 공용 하둡 (3)



카카오 공용 하둡 소개

설계 시 고려사항

장애 대응 사례

공용 하둡의 장단점

- 장점
  - 사용자간 데이터 공유 용이
  - 인프라 비용 절감
  - 사용자에게 동일한 표준 환경을 제공

- 단점
  - 장애 시 모든 사용자에게 영향
  - 리소스 격리를 위한 장치 필요

인증

Kerberos

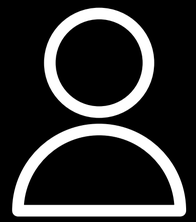
권한

LDAP

공용 하둡의 인증과 권한

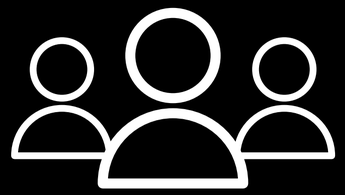
- Kerberos
  - 사용자 단위 인증
  - DC 별로 설치, GSLB 구성
  - password 와 keytab 둘 다 사용 가능
- LDAP
  - 사용자/그룹 단위 권한 관리
  - DC 별로 설치, GSLB 구성
  - posixGroup & groupOfNames 동시 사용
  - sssd를 활용하여 Linux user 연동

공용 하둡의 권한 관리 예시



사용자

[HDFS] /user/ryan
[Kerberos] ryan@REALM
[LDAP] cn=ryan,ou=People,dc=kakao,dc=com



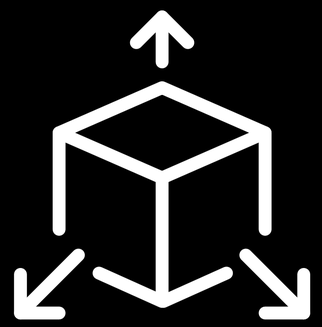
그룹

[HDFS] /friends
[LDAP] cn=friends,ou=Groups,dc=kakao,dc=com
[LDAP] cn=friends,ou=Gnames,dc=kakao,dc=com
[YARN] friends (queue)

안정적인 서비스를 위한 여러가지 정책

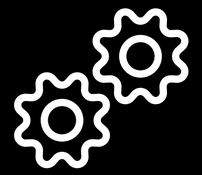
- [HDFS] Space & Name quota
  - 사용자/그룹 별 논리적으로 분리된 공간 사용
  - 권장 평균 파일 사이즈 가이드
- [YARN] queue 최대 리소스 제한
- [YARN] queue 최대 동시 실행 앱 수 제한
- [YARN] preemtion 활성화
- [Hive, Oozie, Hue] L3DSR을 활용한 HA 구성
- [Hive] 조직 전용 Hiveserver 군 제공

사용 편의성 증대를 위한 도구



통합 하둡 클라이언트

필수 요구 사항

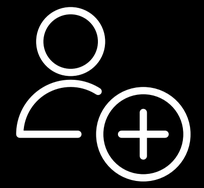


최신 하둡 설정

하둡 바이너리

사용 편의성 증대를 위한 도구 (2)

하둡 권한 관리 시스템



신규 사용자/그룹 생성

그룹 멤버 관리

HDFS 파일 권한 관리

HDFS 파일 권한 관리 상세

- HDFS facl 기능 활용
- dfs.namenode.acls.enabled = true
- hdfs command로 관리 가능

```
$ hdfs dfs -setfacl -m user:ryan:r-x /hadoopeng
$ hdfs dfs -getfacl /hadoopeng
# file: /hadoopeng
# owner: hdfs
# group: supergroup
user::rwx
user:ryan:r-x
group::rwx
mask::rwx
other::r-x
```

HDFS 파일 권한 관리 상세 (2)

- HDFS facl entry는 최대 32개
- 별도의 논리 acl 그룹을 만들어서 관리

```
$ hdfs dfs -setfacl -m group:0000001A:r-x /hadoopeng
$ hdfs dfs -getfacl /hadoopeng
# file: /hadoopeng
# owner: hdfs
# group: supergroup
user::rwx
group::rwx
group:0000001A:r-x
mask::rwx
other::r-x
```

카카오 공용 하둡 소개

설계 시 고려사항

장애 대응 사례

Kerberos KDC 인증 지연 및 간헐적 실패

- 간헐적 잡 실패와 인증 지연이 제보됨
- 모니터링 잡은 정상

- 초당 100회 이상의 인증이 확인
- 다수의 클라우드 컨테이너에서 특정 유저가 커맨드마다 새로운 인증을 시도

- 모니터링 잡에 ip ban 추가
- 시간당 1회의 빈도로 인증하도록 가이드

Hive 성능 저하

- 모니터링 소요 시간 증가 및 실패
- 다수의 사용자 제보가 잇따름

- Namenode 8020 포트에 syn flooding 공격을 확인
- Go 언어로 구현한 클라이언트에서 비정상적으로 접근

- 주요 포트에 대해 SYN_RECV 모니터링
- OS SYN ACK 튜닝

Namenode 반복적 failover

- 수차례 namenode가 exit 되면서 failover
- 다시 실행한 지 3시간 정도 후 다시 exit

- 로그 확인 결과 getGroups 메소드에서 상당히 많은 시간이 소요됨
- 그룹 조회 성능에 문제가 있다고 판단

- 시스템 계정에 static group mapping 적용
- 백그라운드에서 캐시를 reload 하도록 적용

Namenode 반복적 failover (2)

```
<property>
    <name>hadoop.user.group.static.mapping.overrides</name>
    <value>
spark=spark;hdfs=hdfs,hadoop;zookeeper=zookeeper;mapred=mapred,hadoop
;yarn=yarn,hadoop;hive=hive;oozie=oozie;hue=hue;
    </value>
</property>
<property>
    <name>hadoop.security.groups.cache.background.reload</name>
    <value>true</value>
</property>
```

YARN 잡 스케쥴링 지연

- 잡 스케쥴링 속도가 급격하게 떨어짐
- 잡이 평소보다 3배 이상 오래 걸림

- 지표를 통해 스케쥴링 성능 저하 시점 확인
- 동시 실행 잡 수가 900개 이상 증가하면 성능 저하

- root 큐 동시 실행 잡 수 제한
- RM NM 간 heartbeat 간격을 길게 조정
- 주요 조직들의 잡 부하 분산

YARN 잡 스케줄링 지연 (2)

```
<queue name="root">
    <maxRunningApps>880</maxRunningApps>
</queue>

<property>
    <name>yarn.resourcemanager.nodemanagers.heartbeat-interval-ms</name>
    <value>5000</value>
</property>
```

요약

- 카카오에서는 확장된 형태의 공용 하둡 군을 서비스

- 안정성과 사용성을 높일 수 있는 방법을 고민

- 더 폭넓은 모니터링 체계와 장애 대응

E.O.D