PlatformDay 2013:

공공 빅데이터 분석의 가치와 향후 전망

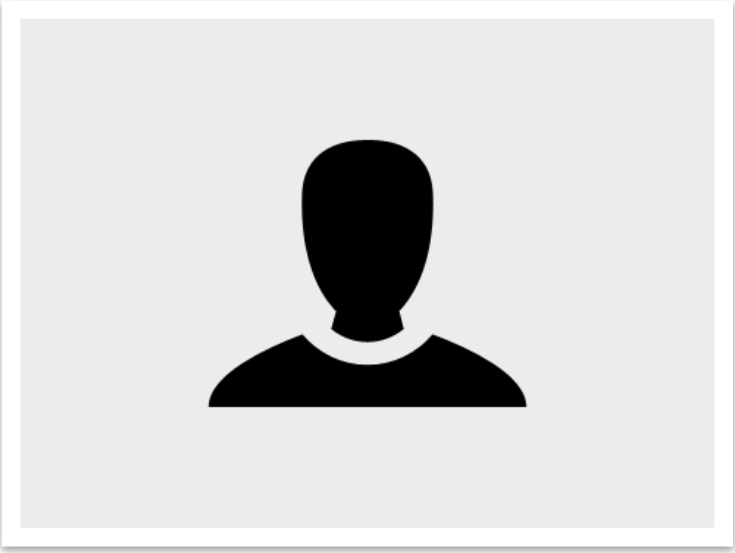
kt NexR

BigData Technical Architect

김영우, 민영근

kt NexR

발표자



- 김영우
  - kt NexR PS본부 TA팀장

- 민영근
  - kt NexR PS본부 TA팀 Technical Architect
  - 단국대학교 연구전담 교수
  - 단국대 대학원 졸업(공학박사)
- 관심 분야
  - 분산 처리, Hadoop
  - 지식 표현 및 추론, 시맨틱 웹

kt NexR

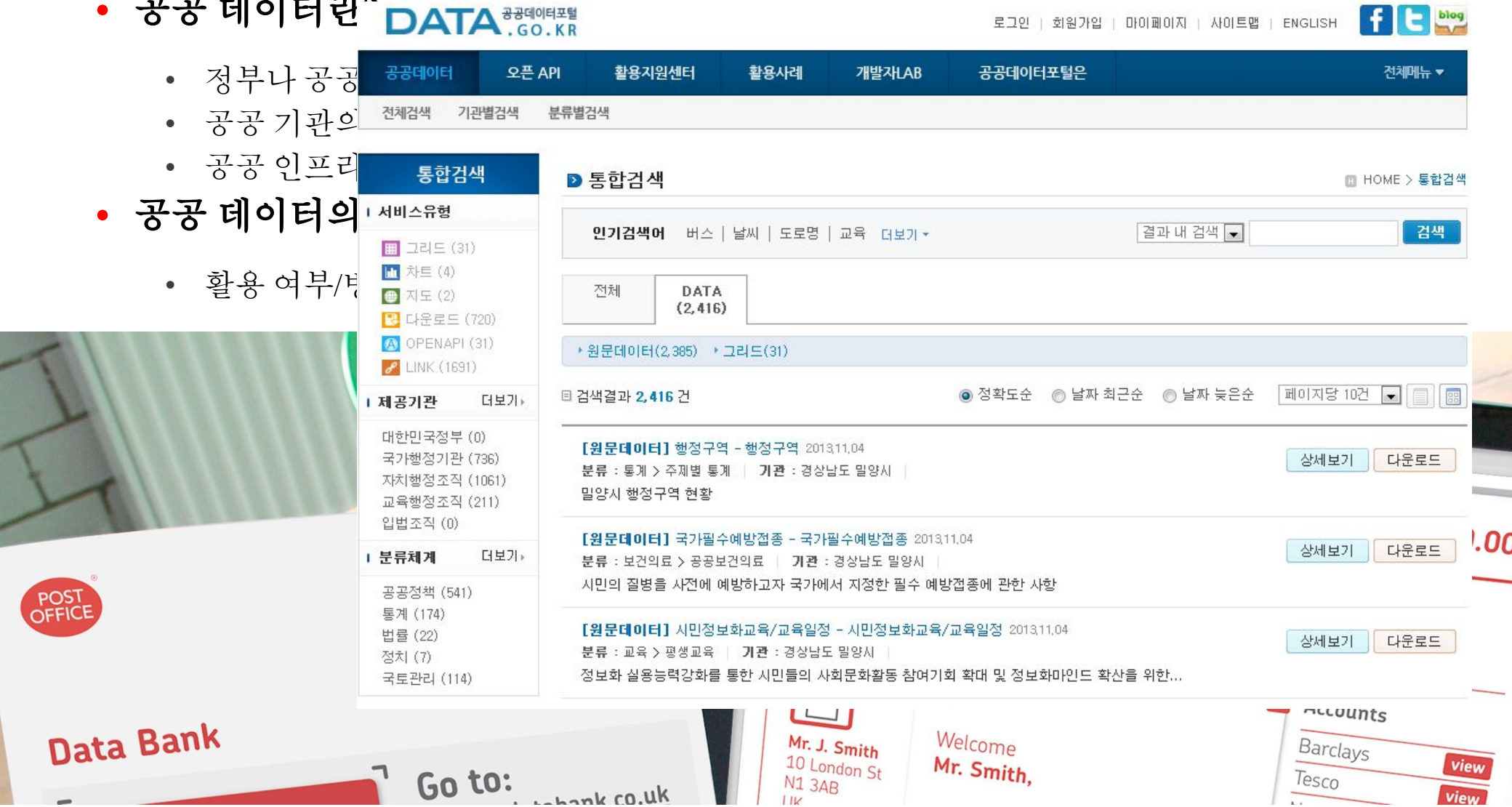
목차

공공 데이터의 의미와 가치

- **공공 데이터란?**
  - 정부나 공공기관이 보유하고 있는 데이터
  - 공공 기관의 업무와 밀접한 데이터
  - 공공 인프라스트럭처에서 생성한 데이터
- **공공 데이터의 가치**
  - 활용 여부/방법에 따라 국민 생활 향상에 밀접한 연관

공공 데이터의 의미와 가치

- 공공 데이터란
  - 정부나 공공
  - 공공 기관의
  - 공공 인프라
- 공공 데이터의
  - 활용 여부/



kt NexR

국외 공공 빅 데이터 사례 연구 (1)

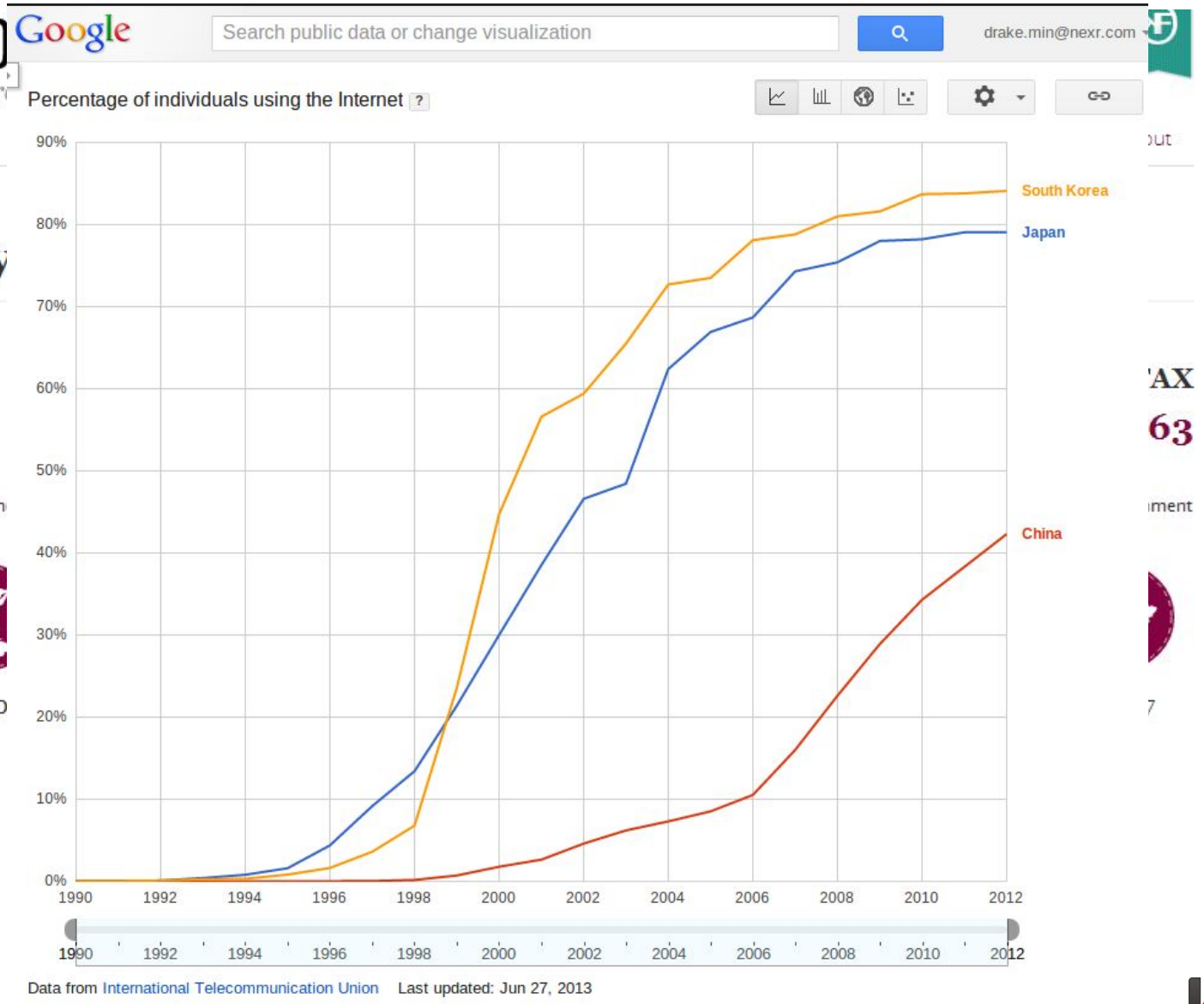
- **NASA(National Aeronautics and Space Administration)**
  - 사용 기술: Apache Hadoop
  - 데이터 크기: 테라바이트
  - 목표: 기후 데이터 분석
- **NARA(National Archive and Records Administration)**
  - 사용 기술: 대용량 매체 저장을 위한 메타데이터, 검색, 분류체계
  - 데이터 크기: 페타바이트, 테라바이트/초
  - 목표: 미국의 기록물에 대한 전자 기록물 보관 및 공개 시스템
- **KTH(Royal Institute of Technology of Sweden)**
  - 사용 기술: 스트리밍 분석 및 예측 분석
  - 데이터 크기: 기가바이트/초 (교통 정보)
  - 목표: 교통 혼잡 및 사고 비율 감소에 따른 교통 상황 향상

국외 공공 빅 데이터 사례 연구 (2)

- **Vestas Wind Energy**
  - 사용 기술: Apache Hadoop
  - 데이터 크기: 페타바이트
  - 목표: 전력 생산을 최대화 할 수 있는 풍력 발전기의 최적 위치 도출
- **CMS(Centers for Medicare & Medicaid Services)**
  - 사용 기술: 열 기반 NoSQL 데이터베이스, Hadoop 고려 중
  - 데이터 크기: 페타바이트, 테라바이트/일
  - 목표: 국민들의 건강 보호 및 보험 청구 절차 준수

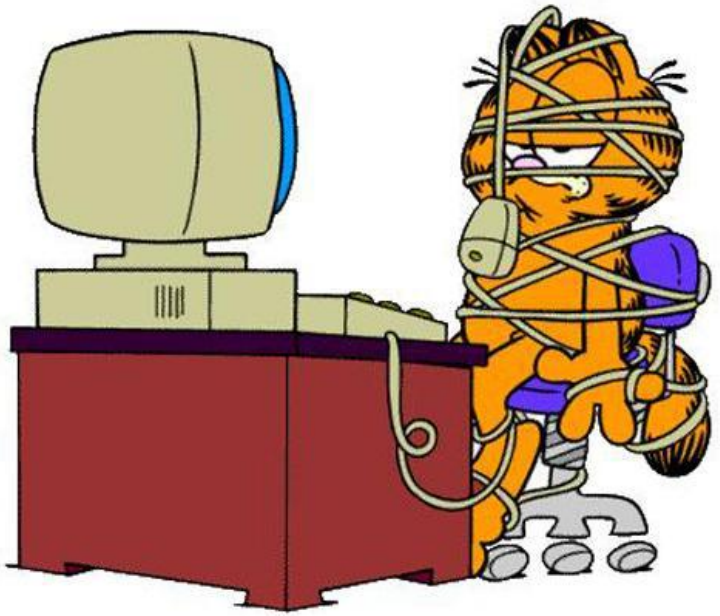
국외 공공 빅 데이터 사례 연구 (3)

국외 공공 빅 데이터 사례 연구 (3)



사례를 통해 본 기술적인 문제들

- 데이터 ACL
- 다중 클러스터 간 데이터 연계
- 레거시 데이터 연계
- Hadoop Balancer

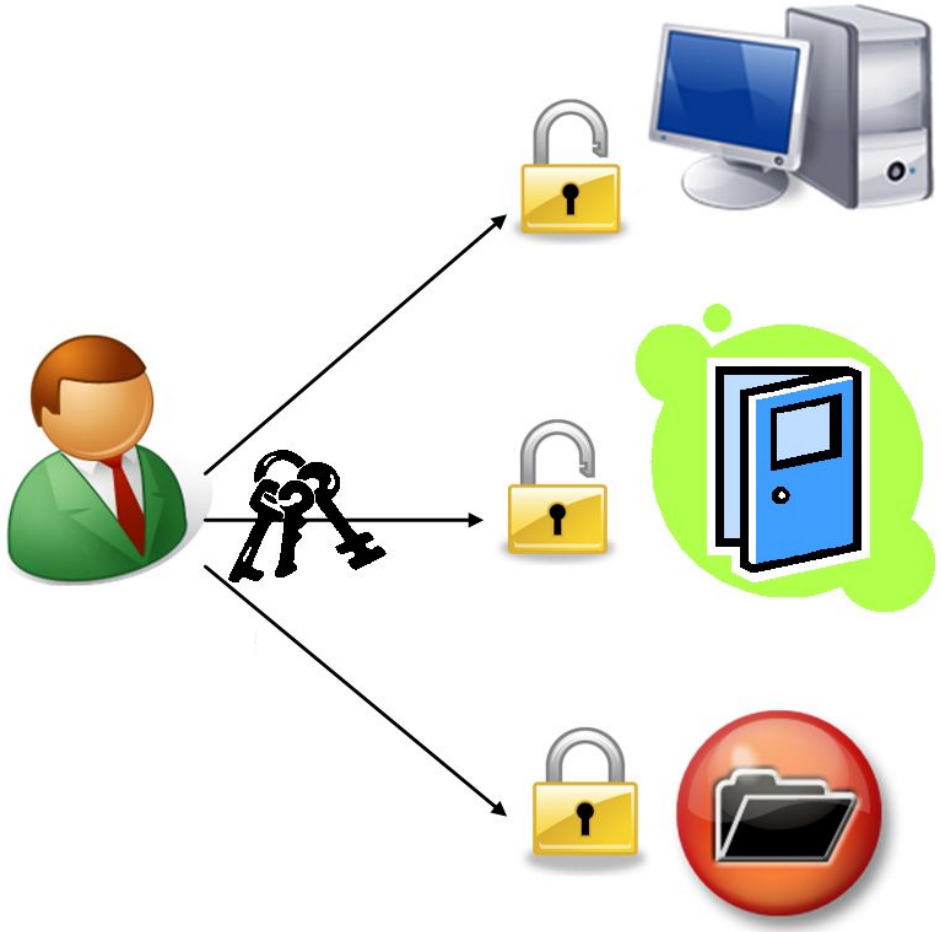


We are having some technical problems

kt NexR

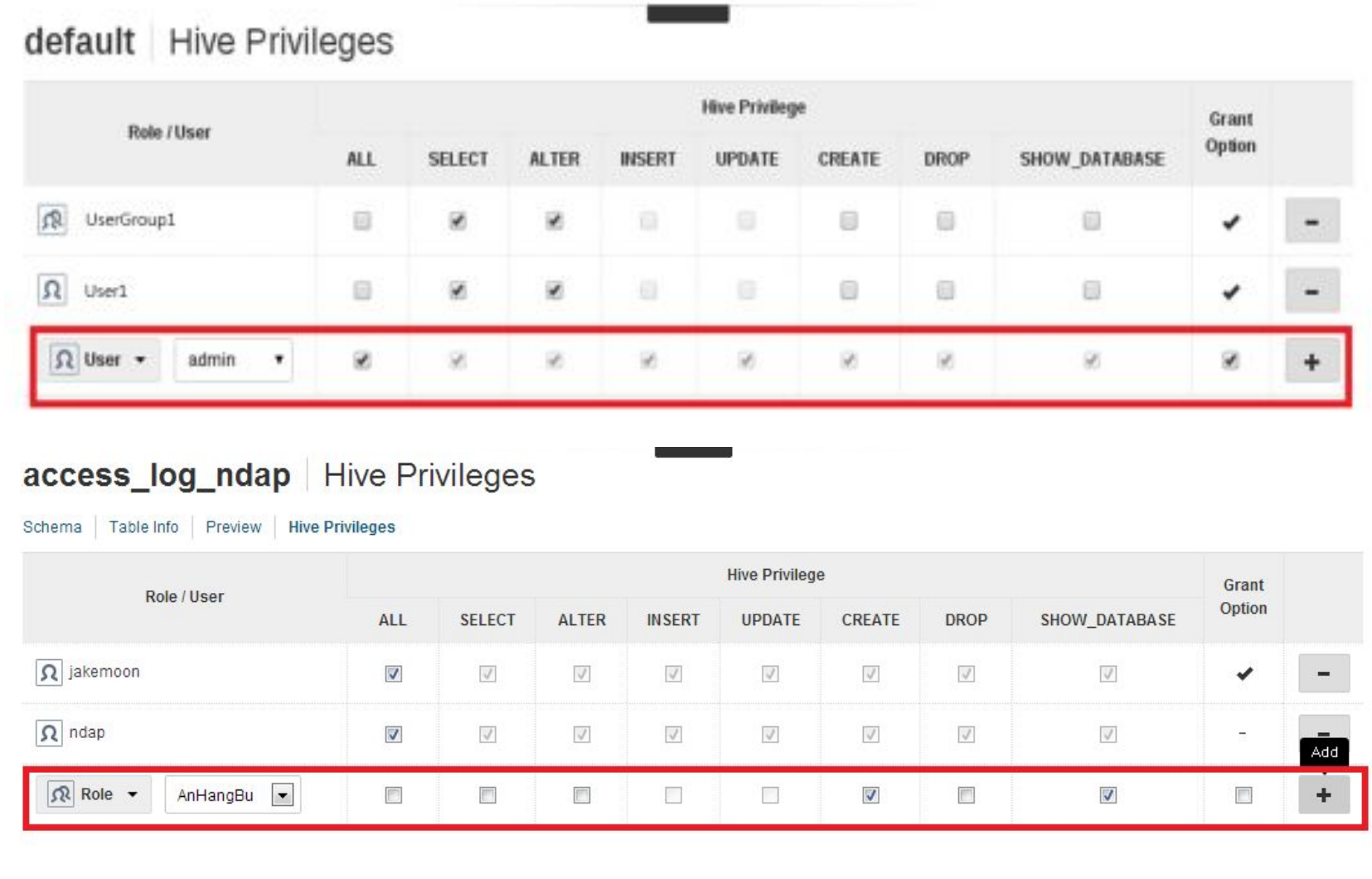
사례를 통해 본 기술적인 문제들: 데이터 ACL

- **부처간 데이터 연계**
  - 여전히 권한 통제는 필요
- **데이터 ACL**
  - 사용자 인증
  - 사용자, 역할
  - 데이터베이스/테이블 권한 관리



kt NexR

사례를 통해 본 기술적인 문제들: 데이터 ACL



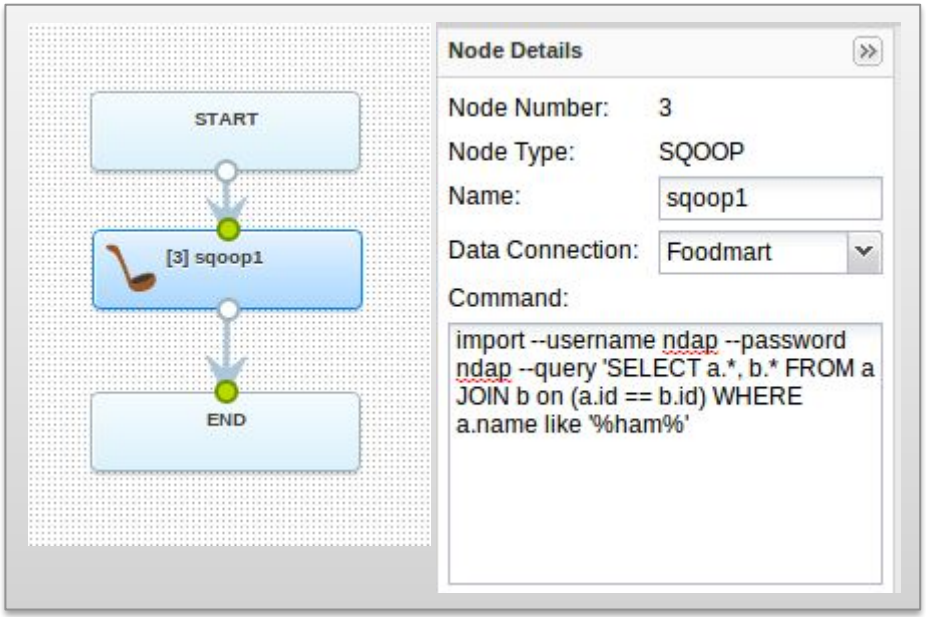
사례를 통해 본 기술적인 문제들: 클러스터 간 데이터 연계

- **DistCp**
  - 대규모 인터/인트라 클러스터 복사 도구
  - 맵리듀스(정확히는 맵 태스크) 프레임워크 사용

- **DistCp v1 vs. DistCp v2**
  - DistCp v1
    - 프로그래밍 방식으로 사용이 어려움: 비동기 호출 불가
    - 준비시간이 오래 걸림: 파일목록 생성, 체크섬 비교
  - DistCp v2
    - 비동기 호출 가능
    - 준비시간 감소: 준비 작업을 맵 태스크로 이동
    - 대역폭 제한 가능, 복사 전략 지정 가능(Dynamic|Uniform)

사례를 통해 본 기술적인 문제들: 레거시 데이터 연계

- **레거시 RDBMS 대상 데이터 가져오기/내보내기**
  - JDBC 표준 기반: Oracle, MS SQL Server, Tibero, IBM DB2
  - Hadoop MapReduce 이용
- **Apache Sqoop 지원**
  - Workflow, ‘Sqoop’ 노드



kt NexR

사례를 통해 본 기술적인 문제들: Hadoop Balancer

- **운영 중 "균형"이 맞지 않는 상태가 될 수 있다.**
  - 노드의 추가/제거 등
- **Balancer는 노드의 디스크 사용량을 기반.**
  - 평균 +- 문턱값의 범위 내
  - 하둡의 블록 위치 정책 준수
- **Balancer 부하 감소 방법 포함**
  - 네임노드에게 부분 블록 맵을 요청하여 사용
  - 가까운 곳에 있는 프록시 원본 노드 선택
  - 밸런싱 중 사용할 네트워크 대역폭 제한 가능

사례를 통해 본 빅데이터 분석

- **Hadoop/Hive 플랫폼 기반(NDAP)으로 RHive를 활용하여 분석 수행**
  - Hadoop – Hive – RHive – R 연계
- **데이터 탐색을 통한 분석 주제 및 범위 선정**
  - 주어진 문제 해결이 아닌 문제 발견
- **시각화**
  - R기반의 차트 및 지도 API 연계

<시각화 예시>

kt NexR

공공 빅데이터 프로젝트의 시사점과 향후 방향

- **빅데이터 인프라스트럭처와 아키텍처 그리고 데이터 분석을 체계적으로 수행한 사례**
  - 인프라 구축부터 분석까지.
- **'분석'에 초점을 맞춘, 공공 데이터에서 '가치'를 찾기 위한 프로젝트**
  - 데이터 공개 중심이 아닌
- **체계적인 데이터 접근 제어를 통한 데이터 보안 해결**
- **정책/법률적인 문제 선결 필요**
  - 데이터의 공개 및 활용에 대한 부처간 협업의 필요성 확인
- **표준 아키텍처 수립을 통한 검증된 인프라스트럭처 구현 필요**


kt NexR