

PlatformDay 2013:

공공 빅데이터 분석의 가치와 향후 전망

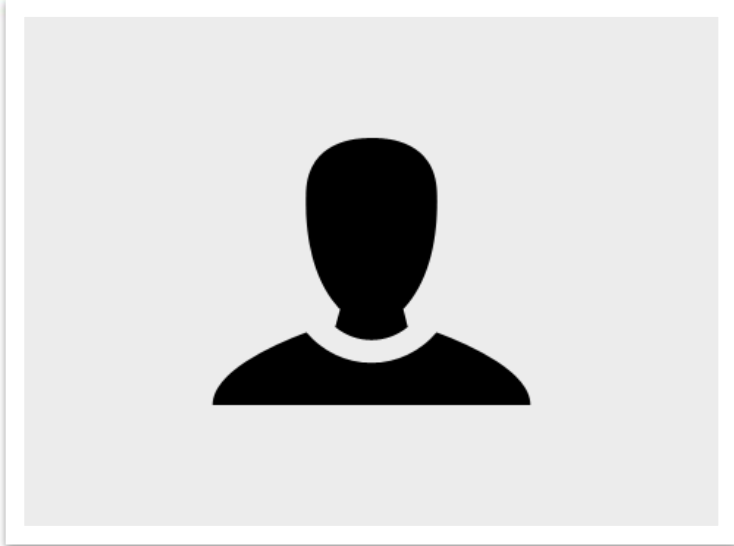
kt NexR

BigData Technical Architect

김영우, 민영근

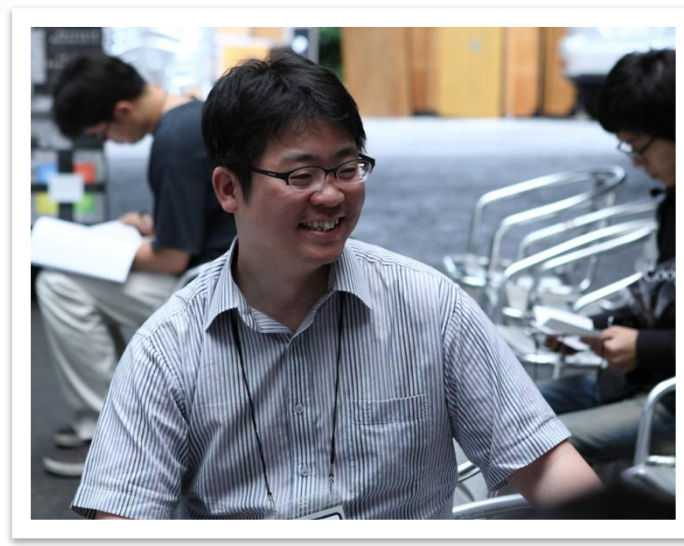


발표자



- 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

목차

- 1 공공 데이터의 의미와 가치
- 2 공공 데이터 활용 현황
- 3 사례를 통해 본 기술적인 문제들
- 4 사례를 통해 본 빅데이터 분석
- 5 프로젝트의 시사점과 향후 방향

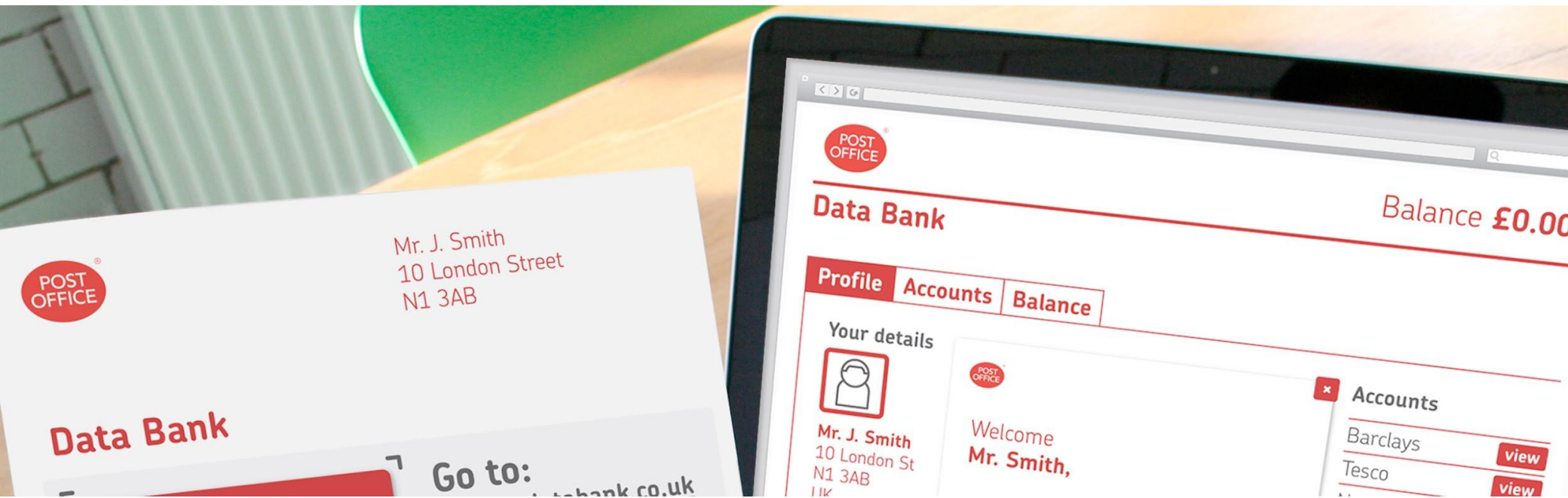
공공 데이터의 의미와 가치

- 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 공공기관이 보유한
- 공공 기관의 업무
- 공공 인프라

• 공공 데이터의

- 활용 여부/방식

DATA.GO.KR 공공데이터포털

로그인 | 회원가입 | 마이페이지 | 사이트맵 | ENGLISH

공공데이터 | 오픈 API | 활용지원센터 | 활용사례 | 개발자LAB | 공공데이터포털은 | 전체메뉴

전체검색 | 기관별검색 | 분류별검색

통합검색

서비스유형

- 그리드 (31)
- 차트 (4)
- 지도 (2)
- 다운로드 (720)
- OPENAPI (31)
- LINK (1691)

제공기관 더보기

- 대한민국정부 (0)
- 국가행정기관 (736)
- 자치행정조직 (1061)
- 교육행정조직 (211)
- 입법조직 (0)

분류체계 더보기

- 공공정책 (541)
- 통계 (174)
- 법률 (22)
- 정치 (7)
- 국토관리 (114)

통합검색

인기검색어 버스 | 날씨 | 도로명 | 교육 더보기

결과 내 검색

검색

전체 DATA (2,416)

원문데이터(2,385) | 그리드(31)

검색결과 2,416 건

정확도순 | 날짜 최근순 | 날짜 늦은순 | 페이지당 10건

[원문데이터] 행정구역 - 행정구역 2013.11.04

분류 : 통계 > 주제별 통계 | 기관 : 경상남도 밀양시 | 밀양시 행정구역 현황

상세보기 | 다운로드

[원문데이터] 국가필수예방접종 - 국가필수예방접종 2013.11.04

분류 : 보건 의료 > 공공보건 의료 | 기관 : 경상남도 밀양시 | 시민의 질병을 사전에 예방하고자 국가에서 지정한 필수 예방접종에 관한 사항

상세보기 | 다운로드

[원문데이터] 시민정보화교육/교육일정 - 시민정보화교육/교육일정 2013.11.04

분류 : 교육 > 평생교육 | 기관 : 경상남도 밀양시 | 정보화 실용능력강화를 통한 시민들의 사회문화활동 참여기회 확대 및 정보화마인드 확산을 위한...

상세보기 | 다운로드

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)

WHERE DOES MY MONEY GO?

Showing you where your taxes get spent



The Daily Bread

Country & Regional Analysis

Departmental Spending

About

The Daily Bread Costs for the British Taxpayer per Day

SALARY

£50,060

SELECT YOUR SALARY

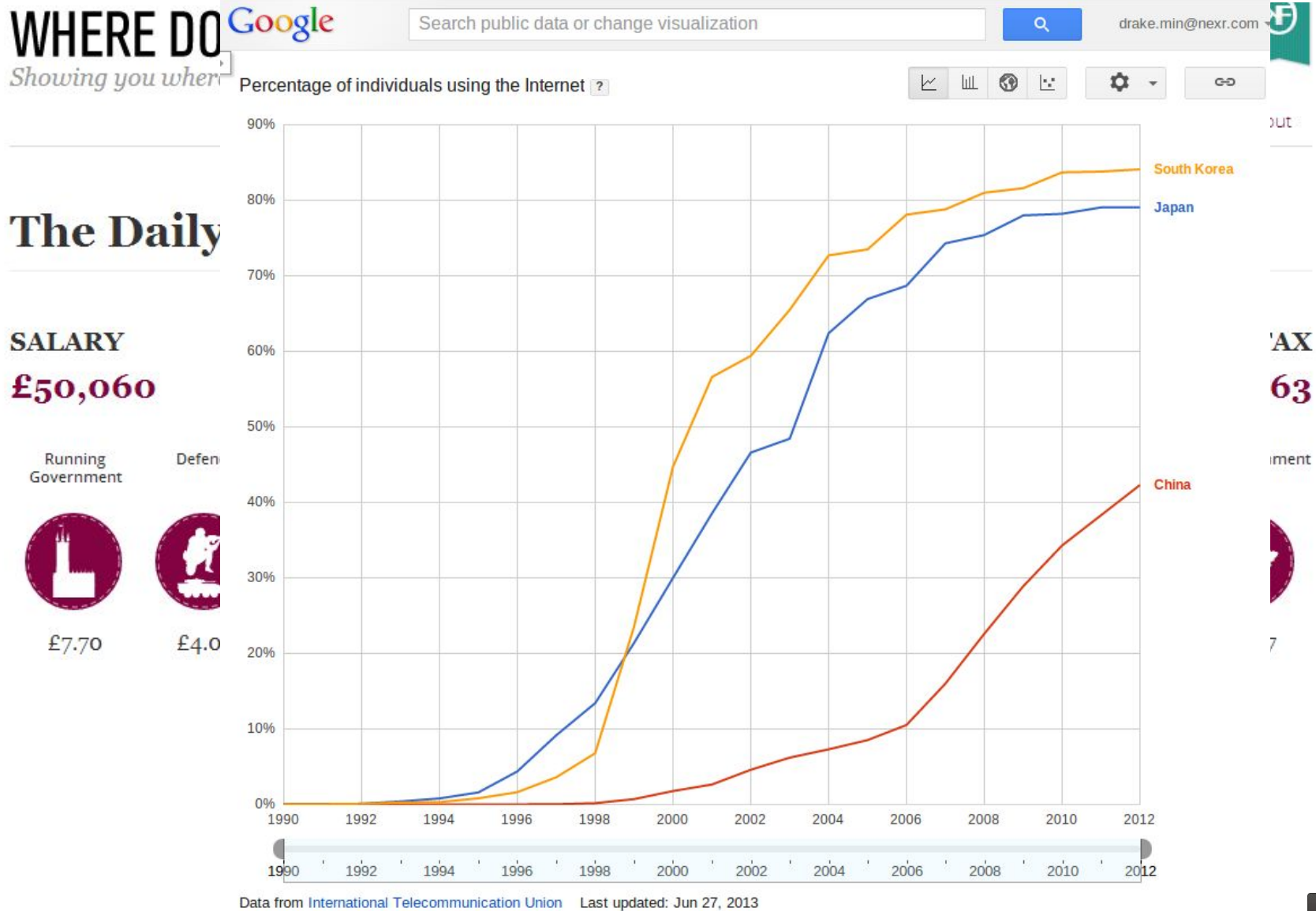


YOUR TAX

£20,763

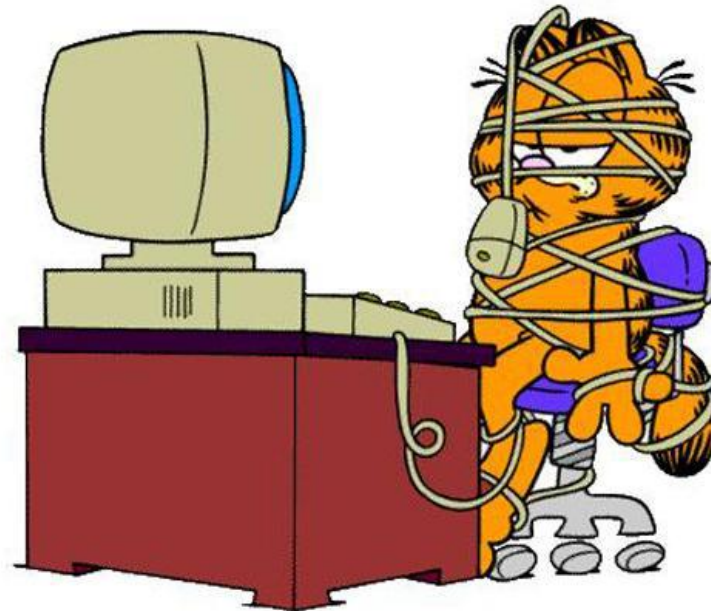


국외 공공 빅 데이터 사례 연구 (3)



사례를 통해 본 기술적인 문제들

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



We are having some technical problems

_____.

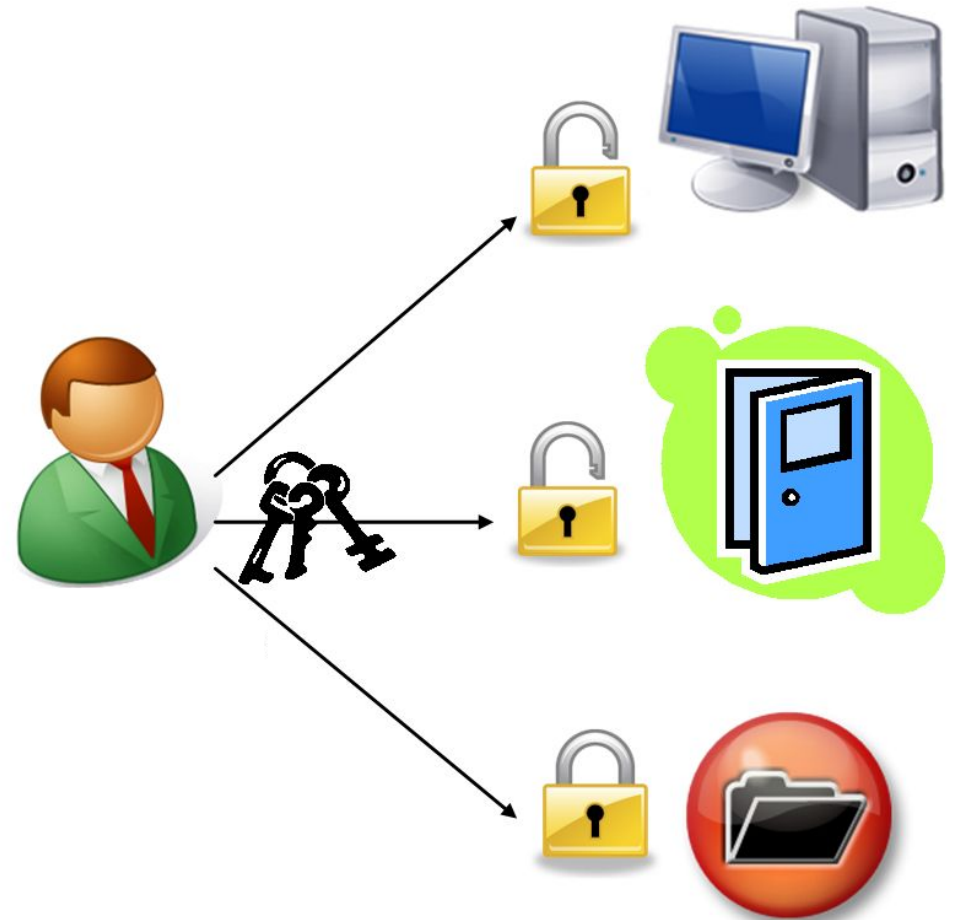
사례를 통해 본 기술적인 문제들: 데이터 ACL

- 부처간 데이터 연계

- 여전히 권한 통제는 필요

- 데이터 ACL

- 사용자 인증
 - 사용자, 역할
 - 데이터베이스/테이블 권한 관리



사례를 통해 본 기술적인 문제들: 데이터 ACL

default | Hive Privileges

Role / User	Hive Privilege								Grant Option	
	ALL	SELECT	ALTER	INSERT	UPDATE	CREATE	DROP	SHOW_DATABASE		
 UserGroup1	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	-
 User1	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	-
 User <input type="text" value="admin"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	+

access_log_ndap | Hive Privileges

Schema | Table Info | Preview | Hive Privileges

Role / User	Hive Privilege								Grant Option	
	ALL	SELECT	ALTER	INSERT	UPDATE	CREATE	DROP	SHOW_DATABASE		
 jakemoon	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	-
 ndap	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	-	Add
 Role <input type="text" value="AnHangBu"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	+

사례를 통해 본 기술적인 문제들: 클러스터 간 데이터 연계

- **DistCp**

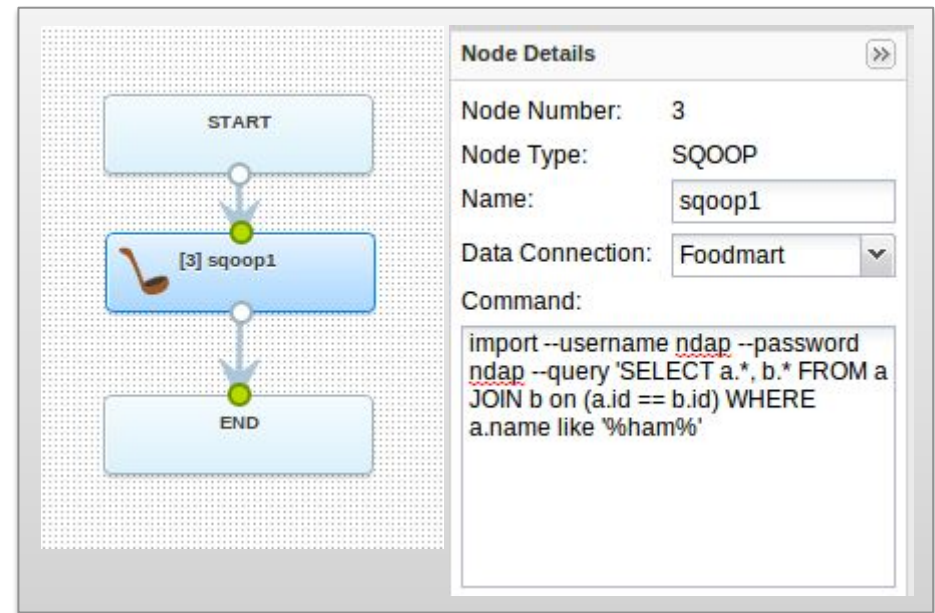
- 대규모 인터/인트라 클러스터 복사 도구
- 맵리듀스(정확히는 맵 태스크) 프레임워크 사용

- **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



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 \pm 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>



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