

# NoSQL 기반의 산업재해 빅 데이터 활용기법에 대한 연구

## A Study of Industrial Accidents' Big Data based on NoSQL

이상범, 이종혁, 김응모  
성균관대학교 정보통신대학

Sang-Bum Lee, JongHyeok Lee, Ung-Mo Kim  
College of Information & Communication Engineering,  
Sungkyunkwan University

### 요약

최근 들어, 여러 곳에서 산업재해가 발생하여 심각한 인명피해와 경제적인 손실을 불러오고 있다. 산업재해의 발생빈도를 최대한 줄이기 위해서는 산업재해의 원인이 되는 다양한 요소들을 파악하여 원인을 차단하는 것이 필수적이다. 본 논문은 여러 산업재해 간 유사성이 존재한다는 사실에 기반 하여 산업재해 빅 데이터 분석을 통해 사고를 일으킬 수 있는 원인을 파악하여 잠재적인 사고의 발생을 예방하고, 위험요인을 예측할 수 있는 효율적인 산업재해 빅 데이터 활용방법을 제시한다.

## I. 서론

2012년도에 산업재해로 사망 1,864명, 부상, 83,349명, 업무상질병 이환자 6,724명이 발생하였다. 또한 산업재해로 인한 직접손실액(산재보상금 지급액)은 3,851,287백만원으로 전년대비 6.23% 증가했고, 직접 또는 간접 손실을 포함한 경제적 손실 추정액은 19,256,435백만원으로 전년대비 6.23%가 증가하였다[1]. 이러한 피해를 줄이기 위해서는 산업의 종류, 사고 장소, 사고 형태, 사고 당시상황 등의 산업재해의 속성을 분석하여 사고원인을 도출해 내고 그것을 통해 예방하는 방법이 필요하다. 따라서 본 논문에서는 NoSQL을 기반으로 하여, 누적된 산업재해 빅 데이터의 분석을 통해 산업재해의 속성을 파악하고 원인을 분석해내어, 사전에 잠재된 사고를 방지할 수 있는 가능성을 제시한다.

## II. 관련 연구

### 1. NoSQL

No SQL은 기존의 관계형 데이터베이스 시스템(RDBMS)을 중심으로 한 데이터 저장기술이 아닌 다른 형태의 데이터 저장 기술을 의미한다. 기존 RDBMS보다 덜 제한적인 일관성 모델을 이용하는 것이 특징이다. RDBMS보다 훨씬 더 많은 데이터를 저장할 수 있다는 특징 덕에 빅 데이터를 다루는 곳에 보다 효과적으로 쓰일 수 있다[2].

### 2. MongoDB

MongoDB는 여러 개의 Collection들로 이루어진 각각의 데이터베이스들의 집합으로 구성된다. 유연한 구조를 기본적으로 가지고 있기 때문에 각각의 Collection들은 서로 다른 종류의 객체를 저장할 수 있다. 모든 객체 즉, 문서(Document)들은 Key와 Value의 쌍으로 표현된다.

### 3. MapReduce

MapReduce는 일종의 함수형 프로그래밍으로써, Map과 Reduce라는 용어가 합쳐진 용어로, 두 함수의 조합을 통해서 분산/병렬 시스템 운용을 지원한다. MapReduce를 이용한 프로그램을 만들기 위해선 Map함수와 Reduce함수를 작성해야 한다.

#### 3.1 Map

Map에는 Key가 되는 데이터와 그에 대응하는 Value 2가지가 존재한다. Map의 결과로 <key, value>쌍을 가진 결과 값들의 집합이 정의된다. Map안에서는 주어진 Key를 사용하여 새로운 Key와 Value를 원하는 만큼 생성한다. 각각의 Key는 그에 해당하는 Value를 가지고 있고 하나의 쌍으로서 처리된다. 질의에서 지정한 모든 데이터에 대해 Map과정을 실행하고, Map에 의해서 만들어진 새로운 Key값은 자동적으로 정리되어 한곳으로 모인다. 그 후 새로운 Key와 모든 Value가 Reduce로 전달된다.

#### 3.2 Reduce

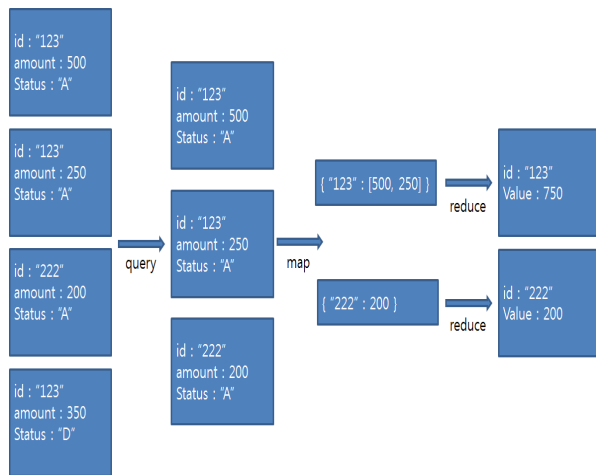
하나의 Key값에 대한 Value를 매개변수로 하여 Reduce함수를 수행한다. 같은 Key값을 가지는 결과들에 대해 통합하는 과정을 Reduce과정이라 할 수 있다. 즉, 같은 Key를 가진 결과쌍들은 Value값을 합치게 되고 그 결과 하나의 통합된 결과쌍이 도출된다. 따라서 Reduce함수는 Value List를 이용하여 각각의 Key들에 대한 최종 결과값을 생성하고, 최종 Key와 Value를 반환한다[3].

## III. 제안 기법 소개 및 구현

### 1. 제안 기법 소개

기본적인 구성은 MongoDB를 통해 작성한 MapReduce함수를 이용해 데이터를 분석하는 것이다. 그

림1을 보면, 가장먼저 질의 과정을 실행한다. 질의에서 요청한 status가 "A"인 데이터만을 따로 추출한다. 그다음 Map과정을 진행한다. id가 "123"인 데이터들은 총 두 개가 있고, 각각 500,과 250의 value를 가지고 있다. 즉 key가 123이고 value가 500, 250인 두 개의 결과쌍이 생성된다. key가 222이고 value가 200인 결과 쌍 또한 추출된다. 그 후 추출된 결과 쌍들은 Reduce함수로 보내져 통합과정을 거치게 된다. 즉 Reduce과정에 전달된 결과 값은 <123, 500>, <123, 250>, <222, 200> 이 있는데, Reduce과정을 거친 후에는 Key가"123" value가 750인 데이터와 Key가"222", value가 200인 데이터가 도출되게 된다[4].



▶▶ 그림1. MongoDB의 MapReduce과정

## 2. 구현 및 결과분석

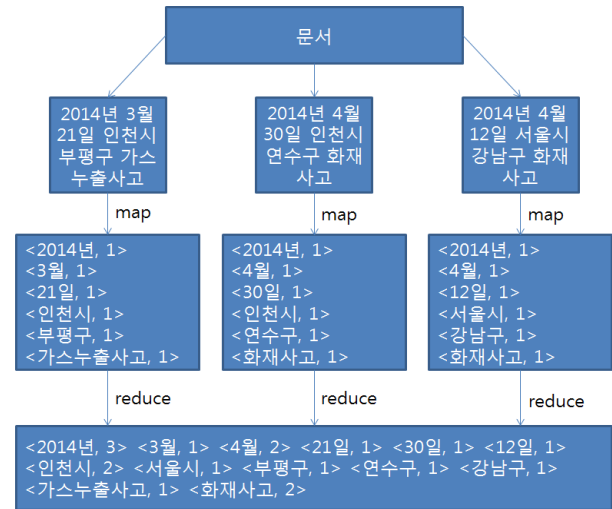
### 2.1. 초기 데이터 처리과정(Map)

Mongo DB의 문서 데이터베이스에 저장되는 산업재해관련 모든 데이터의 텍스트를 활용한다. 각 문서에 저장된 각각의 텍스트는 Map함수를 거쳐 그 자체로 Key값이 되고 그에 따른 Value값을 가지게 된다. 예를 들어 2014년 3월 21일 인천시 부평구 가스누출사고 라는 문자열이 있다면 <2014년,1>, <3월,1>, <21일,1>, <인천시,1>, <부평구,1>, <가스누출사고,1> 이라는 결과 값들을 가지게 된다. Map과정에서는 일련의 문자열에 대해 이러한 과정을 반복한다. 모든 산업재해 데이터에 대한 Map과정을 통해 생성된 결과 쌍을 Reduce함수에 전달함으로써 Map과정이 종료된다.

### 2.2 데이터 분석 및 결과 도출과정(Reduce)

Reduce과정에서는 같은 key값을 가진 결과 쌍들에 대해 통합하는 과정을 거치게 된다. 위에서 얻은 결과 값과 "2014년 4월 30일 인천시 연수구 화재사고"라는 문자열에 대한 결과 값은 <2014년,1>, <인천시,1> 이라는 동일한 key를 가진 결과 값이 있다. 같은 Key를 가진 결과쌍이 존재하기 때문에 <2014년,2>, <인천시,2>라는 새로운 결과 쌍으로 통합된다[5]. 이러한 과정을 거쳐 산업재해 데이터를 분석할 수 있고, 그 속에서 연관성을 찾아내어

사고가 빈번한 장소, 자주 일어나는 재해형태, 시간과 재해형태의 연관성 등, 다양한 각도에서 산업재해 데이터의 분석이 가능하고 그를 통해 다른 사고의 잠재적 위험을 예측하고 대비하는데 사용할 수 있다.



▶▶ 그림2. 실제 문서를 이용한 MapReduce과정

## IV. 결론

본 논문에서는 MongoDB와 MapReduce기법을 활용하여 다양한 산업재해 데이터를 텍스트로 나누어, 사고년도, 사고시간, 사고 장소, 사고형태 등으로 세분화하여 분석하는 방법에 대해 설명했다. 산업재해를 막연히 우연적인 사고로 인식하는 것이 아니라, 어떠한 원인에 의해 발생하는 막을 수 있는 결과라고 인식하고 그것을 예방하는 것이 내용의 핵심이다. 이러한 제안을 활용한다면 잠재적인 사고를 막음으로써 산업재해 발생률이 감소할 것이라 기대된다.

### ■ 사 사 표 기 ■

본 연구는 미래부가 지원한 2013년 정보통신·방송(ICT) 연구개발사업의 연구결과로 수행되었음. 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2013R1A1A2008578)

### ■ 참고 문헌 ■

- [1] 한국산업안전보건공단 "2012 산업재해현황분석" 산업재해 현황 분석 연보, 2013
- [2] Ma Zhi Zhong, "NoSQL 이용한 대용량 데이터처리 성능 향상에 관한 연구", pp.12-31, 2011
- [3] Jeffrey Dean and Sanjay Ghemawat, "MapReduce : Simplified Data Processing on Large Clusters" Google, OSDI 2004
- [4] 김용현, 허의남, "Hadoop과 MongoDB를 이용한 동적 빅 데이터 분석 시스템 구조" 경희대학교 컴퓨터공학과, 2014
- [5] 박성훈, "교통사고 예측을 위한 MapReduce 기반의 대용량 불균형 데이터 분류분석 기법", 건국대학교 대학원, 2014