

석 사 학 위 논 문

Live simulation 훈련결과 Datawarehouse
분석활용 활성화 방안 연구

-과학화훈련단 훈련분석 Datawarehouse를 중심으로-

2012년

한성대학교 국방과학대학원

국 방 M&S 학 과

국 방 M&S 학 전 공

임 회 성

석사학위논문

지도교수 김종만

Live simulation 훈련결과 Datawarehouse

분석활용 활성화 방안 연구

-과학화훈련단 훈련분석 Datawarehouse를 중심으로-

A Study of Using a Practical Application of Live Simulation

Training Datawarehouse

-Focused on the datawarehouse
of Korea Combat Training Center-

2012년 6월 일

한성대학교 국방과학대학원

국 방 M&S 학 과

국 방 M&S 학 전 공

임 회 성

석사학위논문

지도교수 김종만

Live simulation 훈련결과 Datawarehouse

분석활용 활성화 방안 연구

-과학화훈련단 훈련분석 Datawarehouse를 중심으로-

A Study of Using a Practical Application of Live Simulation
Training Datawarehouse

-Focused on the datawarehouse
of Korea Combat Training Center-

위 논문을 국방M&S학 석사학위 논문으로 제출함

2012년 6월 일

한성대학교 국방과학대학원

국 방 M&S 학 과

국 방M&S학 전공

임 회 성

임회성의 국방M&S학 석사학위논문을 인준함

2012년 6월 일

심사위원장 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

국 문 초 록

Live simulation 훈련결과 Datawarehouse 분석활용 활성화 방안 연구 -과학화훈련단 훈련분석 Datawarehouse를 중심으로-

한성대학교 국방과학대학원

국방M&S학과

국방M&S학전공

임 회 성

정보통신기술의 획기적인 발달에 따라 우리군은 미래의 효율적인 국방 경영수단으로 M&S(Modeling & Simulation)를 활용한 훈련을 확대하고 있으며, 2006년부터 실시하여 오고 있는 육군과학화훈련단(KCTC)의 실시간 동모의(Live Simulation)훈련 데이터는 훈련결과에 대한 사후 평가를 실시한 후 훈련분석 데이터웨어하우스에 종합되어 축적되어 왔으며, 이렇게 축적된 데이터를 분석하여 활용하려는 국방부 차원의 노력의 일환으로 국방기술품질원에서는 2011년에 “M&S 분석자료 활용방안 연구” 용역을 수행한 바 있다.

본 연구는 이러한 선행연구를 기반으로 KCTC의 데이터분석 활용여건에 대한 현황과 사용자 의견을 분석하고, 통계청 등 공공기관의 사례를 벤치마킹하여 데이터 활용을 활성화하기 위한 방안을 모색하였다.

본 연구를 통하여 과학화훈련단의 장기간 축적된 자료를 분석하고 군 관련 교리연구 및 M&S체계 구축에 기초자료로 활용할 수 있는 여건을 조성하고자 하며, 기존에 구축되어 있는 과학화훈련단의 훈련분석용 데이터웨어하우스를 대상으로 데이터의 신뢰성을 높이고 활용을 쉽게 하기 위한 방안을 모색하여 웹 기반 자료 검색 체계를 구축하면서 자료 제공 시

요구되는 보안관련 검토사항과 분석자를 위한 불필요한 오류정보 제거 및 선별의 필요성을 확인함으로써 축적된 자료의 적절한 가공을 통하여 분석자가 필요로 하는 자료의 추출이 용이한 서비스 환경에 대하여 연구하고 시험체계를 구현하여 가능성을 확인하였다.

KCTC의 Live simulation 데이터웨어하우스에 대한 분석자료에 일관성과 안전성을 강화하여 요구자에게 보다 원활하게 검색서비스를 제공하는 웹 기반 OLAP서비스 구축이 가능하며, KCTC의 분석자료에 대한 활용 활성화 방안으로 이를 제시한다.

오랜기간 축적된 실기동 데이터를 보다 쉽게 활용할 수 있도록 업무적 용을 위한 커스터마이징과, 자료사전의 제공 등이 추가로 요구되며, 이를 통하여 실기동훈련 데이터의 실무적인 활용 분야가 확대되기를 바란다.

【주요어】 simulation, datawarehouse, ETL



목 차

제 1 장 서론	1
제 1 절 연구의 목적	1
제 2 절 연구의 범위와 방법	2
제 2 장 이론적 배경	4
제 1 절 Live simulation 개요	4
1. M&S의 기본개념 및 효과	4
2. 육군 과학화 실기동 훈련 현황	6
3. 실기동 과학화훈련 분석 체계	8
제 2 절 데이터 웨어하우스 개요	9
1. 데이터 웨어하우스의 기본개념	9
2. OLAP 도구와 비정형 분석 활용	12
3. ETL 패키지	15
4. Micro Data Service	18
제 3 장 과학화 훈련 데이터 분석	23
제 1 절 시뮬레이션 데이터 현황	23
1. 과학화훈련 실기동 자료 축적현황	23
2. 실기동훈련 데이터의 신뢰성	23
3. 사용자 요구 분석	24

제 4 장 과학화 훈련 데이터 활용 활성화 연구	28
제 1 절 데이터 정제를 통한 활성화 방안	28
1. 데이터 공개 범위 확대	28
2. 신뢰성 있는 데이터 제공	28
제 2 절 데이터 접근성 개선을 통한 활성화 방안	29
1. 데이터 분석환경 개선	29
2. 데이터의 정제 가공	32
3. 시험체계 구현	33
제 5 장 결 론	43
【참고문헌】	44
ABSTRACT	46



【 표 목 차 】

[표 2-1] 훈련분석 데이터웨어하우스 스키마 개요	10
[표 2-2] 데이터 웨어하우스의 4가지 의미	11
[표 2-3] OLTP 시스템과 OLAP 시스템의 비교	13
[표 2-4] MOLAP 및 ROLAP 특성 비교	15
[표 2-5] OLAP 서버 제품 현황	16
[표 2-6] 데이터 마이그레이션 방법 비교	19
[표 2-7] 비밀보호를 위한 마이크로데이터 가공방법	22
[표 3-1] 훈련분석 데이터웨어하우스의 전체 데이터량	23
[표 3-2] 과학화훈련 분석결과 또는 관련자료 활용 여부	25
[표 3-3] 과학화전투훈련장의 업무활용 검토	26
[표 4-1] 훈련자 인적정보 선별 대상	28
[표 4-2] 훈련자 상태 정보 추출대상 항목	29
[표 4-3] 훈련분석 데이터웨어하우스 데이터 내용	32



HANSUNG
UNIVERSITY

【 그림 목 차 】

<그림 2-1> 국방 M&S 분류	6
<그림 2-2> KCTC 훈련 개념	8
<그림 2-3> 훈련분석 데이터웨어하우스 누적 과정	10
<그림 2-4> 스타스키마와 스노우플레이크 스키마	12
<그림 2-5> 통계청 마이크로데이터 서비스 홈페이지	20
<그림 3-1> 과학화전투훈련단 자료 활용 형태	25
<그림 3-2> 비정형분석 미활용 사유	26
<그림 3-3> 비정형분석 활용 활성화 의견	27
<그림 4-1> 외부기관용 웹 포탈 운용 구성	32
<그림 4-2> WEB-OLAP 시험체계 구성도	34
<그림 4-3> Pentaho Data Integration Database Connection	34
<그림 4-4> Pentaho Data Integration SQL select	35
<그림 4-5> Pentaho Data Integration Calculator	35
<그림 4-6> Pentaho Data Integration Filter rows	36
<그림 4-7> Pentaho Data Integration Number ranges	36
<그림 4-8> Pentaho Data Integration Table out 필드 매핑	37
<그림 4-9> Pentaho Data Integration 데이터 마이그레이션 실행	38
<그림 4-10> 사용자 그룹 관리 화면	39
<그림 4-11> 사용자 권한 설정 화면	40
<그림 4-12> 사이트 관리 화면	40
<그림 4-13> 테이블 관리 화면	41
<그림 4-14> 피격거리별 검색결과 화면	42
<그림 4-15> 거리별 피격인원	42

제 1 장 서 론

제 1 절 연구의 목적

우리 군에서 육군은 정보통신, 광학, 위치추적 등의 기술이 융합되어 실전투와 가장 유사하도록 가상의 전문 대항군과 실시간 교전을 통하여 야전부대들이 실전에 가까운 전투경험을 체득하는 과학화 전투훈련인 KCTC(Korea Combat Training Center)¹⁾ 훈련을 실시하여 많은 훈련성과를 거두고 있다.

실전과 유사한 경험을 쌓을 뿐 아니라 한국적 상황에 맞는 전투경험 및 전투제원 확보가 가능할 것으로 기대되며 실제훈련에 따른 고가의 탄약, 장비 등을 사이버효과로 대체함으로써, 연간 2,700억 원의 예산 절감 효과가 있을 뿐만 아니라, 실 사격으로 인한 전장소음, 자연훼손, 농경지 파괴 및 대민피해 요소를 최소화 할 수 있어 가장 효과적인 훈련방법으로 평가되고 있다.²⁾

이러한 과학화 전투훈련은 '06년부터 매년 00부대를 대상으로 훈련을 실시해 오고 있으며 수년 내로 연대급으로 확장하여 2년간의 의무복무기간 중에 최소 1번은 훈련을 경험할 수 있도록 확대될 예정이다.

KCTC 실기동훈련 결과 데이터는 자동화체계를 거쳐 데이터베이스에 계속 축적되고 있으며, 해당 부대 훈련 종료 후 사후검토 위주로 활용하고 있으며 '06년부터 축적된 데이터에 대한 종합적인 분석 및 활용은 최근까지 활발하지 못한 실정이다.

이러한 여건에서 '11년도에 과학화된 실기동훈련을 통해 축적된 데이터의 현황을 분석하고 활용방안을 제시하여 향후 과학화 전투훈련이 지속적으로 확대, 발전하기 위한 연구를 국방기술품질원에서 수행한 바 있으며,

1) 과학화전투훈련(KCTC) : 1997년 육군교육사 예하 과학화전투훈련 계획처로 최초 태동 후 강원 인제·홍천 지역에 118km²의 부지를 확보하고, 2005년 9월 중앙통제장비 체계를 개발하여 대대급 과학화전투훈련을 실시할 수 있는 실 기동훈련장을 구축하였다.

2) 이재완(2008), "과학화전투훈련을 통해 본 합동성 강화", 『합참 제34호』, 서울, 합동참모본부, p.62

연대급 구축시에 더욱 발전될 수 있을 것으로 기대하고 있다.

본 연구는 이제까지 축적된 KCTC의 실기동훈련분석 데이터웨어하우스로부터 국방분야의 관심있는 연구자들에게 다양한 요구에 적절한 분석 지원을 하기 위해 별도의 데이터웨어하우스와 OLAP(On-Line Analytical Processing)을 구축하여 지원할 수 있는지를 시험하기 위하여 시도된 것이다.

데이터웨어하우스는 데이터를 보유하고 있는 정보의 창고라고 한다면 OLAP은 창고안의 데이터를 가져와서 사용자가 의사결정을 하거나 필요한 정보의 보고서를 보고자 할 때 사용되는 도구라고 할 수 있다.

기존의 KCTC에는 각각의 단위 훈련주기별 데이터베이스를 통합하여 하나의 커다란 데이터 저장소인 훈련분석 데이터웨어하우스가 구축되어 있으며 이를 활용하고 분석하는 OLAP 환경이 제공되고 있으나 활용은 활발하지 못한 상태이다. 오랜기간 축적된 KCTC의 실기동 데이터를 연구 목적에 맞는 형태로 가공하고 정제하여 손쉽게 활용할 수 있는 안전한 방식이 제공되어 연구자들이 필요 적절한 시기에 원하는 정보를 장소에 구애받지 않고 원하는 형태로 KCTC의 체계운영부서 도움 없이 스스로 검색, 분석, 판단을 할 수 있는 시스템이 제공된다면 보다 연구의 활성화를 기대할 수 있을 것이다.

본 논문에서는 기존에 구축되어 있는 과학화 훈련단의 데이터웨어하우스를 대상으로 자료 제공 서비스를 위한 웹 기반 자료 검색 체계를 구축하면서 자료 제공 시 요구되는 보안관련 검토사항과 분석자를 위한 불필요한 오류정보 제거 및 선별의 필요성을 확인함으로써 축적된 자료의 적절한 가공을 통하여 분석자가 필요로 하는 자료의 추출이 용이한 서비스 환경에 대하여 연구하고 시험체계구현을 통한 확인에 목적이 있다.

제 2 절 연구의 범위와 방법

본 연구에서는 기존에 구축되어 있는 KCTC의 데이터웨어하우스를 대상으로 수록된 실기동 훈련데이터를 효과적으로 활용할 수 있도록 지원하는 자료 제공 서비스에 대하여 검토하고 시험체계를 구현한다. KCTC의

데이터웨어하우스를 기반으로 하는 OLAP 서비스를 구축하고 활용하여 데이터의 효율적인 분석활용 방안을 모색하고자 한다.

먼저 훈련분석 시스템의 근간이 되는 데이터 웨어하우스에 관련된 이론과 기술을 검토하고 시스템의 설계에 대하여 연구할 것이다. 기존에 구축되어 있는 KCTC의 훈련분석 데이터웨어하우스에 저장된 다년간의 실기동훈련 데이터를 효과적으로 활용할 수 있는 자료 제공 서비스 방법을 연구하고 구현방법을 제안하고자 하며 이를 위하여 자료 제공시 요구되는 보호조치와 수록자료의 분석을 통한 명백한 오류정보의 식별 및 제거를 통한 정제된 데이터의 서비스방법을 시험하고자 한다.

본 논문은 총 5장으로 구성되어 있다. 1장에서는 본 논문 연구의 목적 및 범위와 방법에 대하여 설명하며, 2장에서는 '국방 M&S(Modeling & Simulation)'의 기본개념을 상기하며, 우리 군의 실전에 대비한 훈련에서 중요한 부분을 차지하는 육군 과학화훈련단의 현황을 Live Simulation 측면에서 살펴보고 훈련분석용으로 구축된 데이터웨어하우스의 주요개념과 데이터의 분석활용도구에 대하여 조사한다.

3장에서는 그동안 과학화훈련을 통해서 축적된 데이터의 내용과 활용 상태 및 사용자의 의견을 분석하여, 4장에서 본 연구목적인 KCTC에 축적된 실기동훈련 데이터의 활용을 활성화하기 위한 조건과 서비스 방법을 구체화하여 제안하고 시험체계를 구현하였다. 5장에서는 결론으로 본 연구의 기대효과와 향후 연구발전분야에 대하여 설명한다.

이렇게 기존의 데이터웨어하우스에서 추출되어지는 데이터를 분석하고 활용하기 위하여 구현한 시험체계의 구성은 다음과 같다. 데이터베이스 SW로는 MySQL 5.5.21버전을 사용하여 서비스용 데이터웨어하우스를 구축하였다. 데이터 정제를 위한 ETL(Extraction, Transformation, Loading) Tool은 Pentaho사의 Data Integration 4.2버전 (kettle)을 사용하였고 Data 분석제공을 위한 OLAP Tool은 Oz EQ 5.1을 사용하여 구축하였다.

이들 SW를 탑재한 컴퓨터의 HW 사양은 Intel Quad Core 2.4GHz

CPU, 8GByte RAM, 1TB HDD 사양의 PC급이며 운영체제는 Windows server 2008 standard를 설치하였다.



제 2 장 이론적 배경

제 1 절 Live simulation 이해

첨단 정보과학기술이 발달함에 따라 컴퓨터를 이용한 시뮬레이션의 활용이 활발히 증가하고 있으며, 특히 컴퓨터게임에서 만이 아니라 실제환경과 매우 유사한 전장환경에 대한 표현기법의 발달과 조직 및 무기체계, 인간행동 등에 대한 표현기법도 정교하게 발전하고 있다.

우리 군은 1980년대 말에 을지포커스 훈련과 같은 한미 연합연습 등을 통해 부분적으로 접하던 M&S를 1990년대부터 본격적으로 광범위하게 적용하기 시작하였다.

육군의 경우 1990년대에 BCTP(Battle Combat Training Program)단을 설립하여 독자적으로 개발한 위게임 모델인 창조21을 활용하여 사·군단급 전투지휘훈련을 시작하여 한국군 고유의 모델들을 개발하여 적용해오고 있으며, 실기동훈련장인 KCTC를 구축하였다.

2000년대에 들어서면서 육군에 이어 합참, 해군, 공군, 해병대가 각각 독자 모델을 개발하는 등 교육훈련분야에의 활용뿐만 아니라 전력분석, 국방 획득 및 전투실험 분야에의 M&S 활용이 점차 증가하고 있다.

1. M&S의 기본개념 및 효과

M&S(Modeling & Simulation)란 시스템, 개체, 현상 또는 절차의 물리적, 수학적, 논리적 표현과정을 의미하는 '모델링(Modeling)'과 모델로 표현되는 모델링의 결과를 실제와 동일 또는 유사하게 시간에 따른 변화로 표현하는 방법인 '시뮬레이션(Simulation)'의 합성어를 말한다.³⁾

국방 M&S(DM&S : Defence Modeling & Simulation)는 국방분야에서 실제 시행이 제한되는 여러 현상에 대하여 분석/예측을 할 수 있는 중요한 수단으로서의 M&S로 정의가 가능하다.

3) 방위사업청 획득기반과(2010), 『무기체계 획득단계별 M&S 적용지침』, 서울, 방위사업청, p.1

모델이란 개념은 매우 광범위하고, 통상적으로는 어느 정도의 제한사항을 가지고 있다. 모델은 실제체계 즉 시스템을 그대로 복사하는 것이 아니라, 표현하고 묘사해 줌으로써 연구자와 의사결정자에게 유용한 도구로서 사용되게 된다. 모델에서 포함요소가 누락되지 않고, 실제감을 완벽하게 묘사한다는 것은 매우 복잡하고 어려운 일이다. 그런 가운데에서도 할 수만 있다면 실제체계의 특징을 잘 나타내면서 단순화해야 한다. 이렇게 모델을 만들어가는 과정을 모델링이라고 하며, 컴퓨터의 등장 및 첨단 정보과학기술의 발달과 함께 모델링 기법도 점차 발전하고 있다.

국방 M&S는 실제 전장의 특성/기능을 모형화하고, 이들을 국방분야의 분석·훈련·획득 등의 업무에 활용하는 것으로서, 그 분류체계는 <그림 2-1> 국방 M&S 분류⁴⁾와 같다.



<그림 2-1> 국방 M&S 분류

국방 M&S를 모델링의 대상에 따른 기능별 유형으로 분류하면, 과학화 전투훈련단(KCTC)과 같은 실기동모의(Live simulation)체계와 기능숙달 및 조종훈련 시뮬레이터와 같은 가상모의(Virtual simulation)체계, 창조21등의

4) 최홍필(2010), "M&S를 활용한 군 과학화 훈련체계 구축에 관한 연구", 한성대학교 경영대학원 석사학위논문, p.9

위게임 모델과 같은 구성모의(Constructive simulation)체계로 구분되어진다.

국방분야에 M&S를 활용하는 일반적인 목적과 이유는 인류역사 이래 끊임없이 반복되어온 전쟁이라는 실제체계에 대해 충분히 분석하고, 이해하여 이를 적절히 대비하고자 하는 데 있다. 적이 취할 수 있는 반응이나 행위에 대해 보다 나은 이해를 얻고, 할 수만 있다면 사전에 전쟁계획의 절차 및 과정을 분석하고 검토함으로써, 추상적이고 막연한 개념을 보다 구체화하고자 하는 것이다. 동시에 가능한 대안들에 대해서도 적용을 시도하기 전에 단순화하고, 이해할 수 있도록 준비된 상태에서 시행하고자 하는 것이다. 이런 관점에서 M&S는 충분히 과학적이고 합리적이며 시스템적인 사고의 수단이 될 수 있다. 컴퓨터를 활용함으로써, 새롭게 디자인한 체계의 잠재적 효용성을 평가하고, 동시에 다양한 체계들의 운용전략을 비교 분석하는 등 M&S의 활용 효과는 매우 크다고 할 수 있다.

2. 육군 과학화전투훈련단 현황

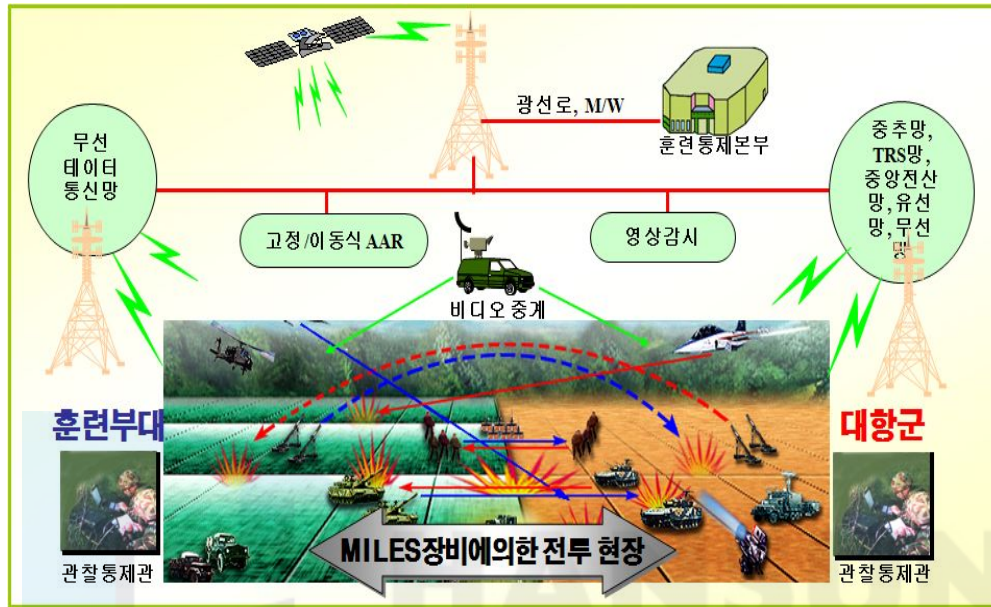
실제 전장과 유사한 훈련환경을 구축하여 "피를 흘리지 않고 전투를 체험"할 수 있도록 실제 전투현장과 유사하게 전투훈련을 하고, 실시간 자료수집 및 분석을 통해 사후검토 및 전투실험이 가능하도록 구축한 체계가 육군 KCTC(과학화전투훈련)이다.

역사적으로 전투 미경험자가 초기전투에서 큰 피해를 당한다는 것은 확인된 사실이며, 미군이 1981년에 NTC를 창설하여 걸프전 및 이라크전을 통해 그 효과성이 입증됨에 따라 과학화전투훈련장의 필요성이 부각되었다. 이에 '05년 실제 전투와 유사하게 훈련하고 무기체계 발전/교리 변화의 검증을 위해 KCTC단이 창설되었다.

대대급 과학화 전투훈련의 목표는 실전적 전투상황하에서 보병대대의 통합 전투수행능력을 향상시키고 보병연대 지휘관 및 참모의 전투지휘 능력과 전투 근무지원능력을 배양하며 전투훈련 결과 및 교훈을 야전부대에 환류시키고 전투발전 소요를 제기하는 것으로서, 실전적인 훈련묘사, 입체적인 훈련관찰, 실시간 훈련분석 및 사후검토 기능이 구비된 과학화 전투훈련장에서 적전술을 구사하는 전문 대향군과 자유기동훈련을 통해 전장

의 실상을 간접으로 체험하도록 하는 실기동훈련이다.

KCTC는 1개 보병대대와 대항군이 첨단 정보통신장비와 마일즈(MILES : Multiple Integrated Laser Engagement System)⁵⁾ 등을 이용하여 전장 실상에 부합된 쌍방 자유기동으로 훈련을 실시할 수 있는 훈련장으로, '11년까지 100여개 부대 약 12만 여명이 훈련에 참가하였다.



<그림 2-2> KCTC 훈련 개념

<그림 2-2> KCTC 훈련 개념⁶⁾에서와 같이 훈련은 부대전개/이동, 공격/방어 등으로 구성되며 주요 국면 및 종합 사후검토 등 약 2주간에 걸쳐 실시된다. 병력이동 및 교전은 자율에 의하며, 직사화기는 운용자의 격발 및 MILES에 의해 처리되고, 곡사화기/화학오염/지뢰지대 등은 모의처리된다. 훈련간 개인/장비별 위치는 GPS⁷⁾를 통해 식별되고, 상호 교전결과, 상황부여/조치 및 모의내역 등은 훈련 인원 및 장비에 부착된 PU(Personal

5) MILES : 다중 통합 레이저 교전체제로 레이저에 의한 화기발사 묘사 장비와 감지장비, 레이저 교전상황을 통합 처리하는 소프트웨어 및 운용체계를 통칭함

6) 최홍필, 전계논문, p.34

7) Global Positioning System(GPS) : 범지구위치결정시스템, 미국 국방성에서 개발, 공식 명칭은 NAVSTAR GPS임,

Unit)와 관찰통제관의 PDA에 의해 처리되어 유무선 통신을 통해 훈련통제본부의 중앙통제장비로 전송되어 저장된다. 훈련간 통신내역은 감청을 통해 분석 및 저장하고 주요 국면은 Video로 촬영하여 사후검토 자료로 활용한다.

대대급 KCTC의 훈련효과 증대와 미래전을 대비한 다양한 작전수행능력 구비를 위해 '09년부터 연대급 훈련체계를 구축 중에 있으며, '14년 시험평가를 거쳐 '15년부터 연대급 훈련을 실시할 계획이다.⁸⁾

3. 실기동 과학화훈련 분석 체계

과학화훈련장의 실기동훈련간에 수집되는 정보는 중앙통제장비의 주기훈련별 데이터베이스에 수록된다. 훈련부대 인원 및 장비로부터 수신되는 처리 내역, 관찰통제관 및 분석관에 의해 묘사/처리된 사항, 통신감청 및 영상촬영 자료가 유무선으로 전송되어 중앙통제장비로 저장된다.

1차적으로 주기훈련별 데이터베이스에 저장된 실기동 데이터는 훈련부대별 사후검토에 활용된 후, 일정기간 보관한다. 주기훈련별 데이터베이스는 1회 훈련차수마다 생성되어 훈련에 필요한 모든 데이터를 수용하고 실시간으로 발생하는 실기동훈련 결과를 기록 유지한다. 훈련이 종료되면 사후평가에 활용하기 위한 훈련분석서버의 데이터웨어하우스로 데이터 이관 및 백업을 한다.

훈련분석 데이터웨어하우스에는 각각의 훈련주기마다 수집된 주기훈련 데이터베이스로부터 사격화기, 거리, 상태 등의 개인별 정보를 축적하여 다차원으로 분석할 수 있도록 한다. 훈련분석 데이터웨어하우스는 [표 2-1]와 같은 스키마⁹⁾로 구성되어 있다.

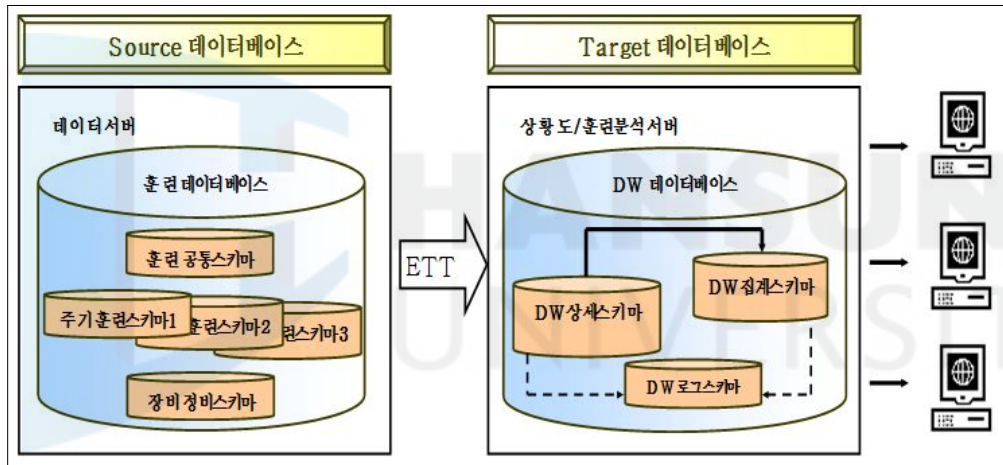
8) 김종만 외 2인(2011), “M&S 분석자료 활용방안 연구”, 국방기술품질원, p.43

9) 스키마(schema) ; 데이터시스템 언어회의(CODASYL) 데이터베이스를 기술하기 위해 사용하기 시작한 개념. 데이터베이스의 구조에 관해서 이용자가 보았을 때의 논리 구조와 컴퓨터가 보았을 때의 물리 구조에 대해 기술하고 있다. 데이터 전체의 구조를 정의하는 개념 스키마, 실제로 이용자가 취급하는 데이터 구조를 정의하는 외부 스키마 및 데이터 구조의 형식을 구체적으로 정의하는 내부 스키마가 있다. (IT용어사전)

[표 2-1] 훈련분석 데이터웨어하우스 스키마 개요

스키마 분류	내 용
DW상세스키마	훈련 종료 후, 훈련분석현황을 볼 수 있도록 교전데이터 중심으로 훈련자별 개인훈련결과정보를 누적관리
DW집계스키마	훈련결과를 제대별로 집계해서 여러차원으로 데이터를 분석할 수 있도록 집계된 자료를 축적 (다차원의 디멘전을 필요한 분석모형 위주로 구축)
DW로그스키마	ETT 프로그램 실행시 오류발생내용과 매핑정보 관리

매 주기훈련이 종료되면 <그림 2-3>과 같이 훈련데이터베이스에서 ETT ; Extract(추출)/ Transformation(변형)/ Transmission(전송) 프로그램을 실행하여 훈련분석 데이터웨어하우스에 누적한다.



<그림 2-3> 훈련분석 데이터웨어하우스 누적 과정

제 2 절 데이터 웨어하우스 개요

1. 데이터 웨어하우스의 기본개념

정보시스템을 데이터의 처리형태로 볼때 크게 두가지로 나눌 수 있다.

첫째, 물류 시스템이나 금융, 증권거래처럼 정보의 변경이 지속적으로 일어나는 운영데이터베이스(Operational Database)¹⁰⁾ 중심의 시스템으로 안정성과 정확성을 중시하는 운영계 시스템(OLTP: OnLine Transaction Processing)과 둘째, 대량의 데이터를 경영자와 같은 분석자에게 알기 쉽게 제공하기 위한 분석을 목적으로 설계된 정보계 시스템(OLAP: OnLine Analysis Processing)이다.

데이터웨어하우스에 대한 W. H. Inmon의 정의에 따르면, "데이터웨어하우스란 의사 결정 프로세스를 지원하도록 데이터를 첫째 주제지향적(Subject-Oriented)이고, 둘째 통합(Integrated)되고, 셋째 시계열적(Time-Variant)으로 시간에 따라 변화된 데이터 정보를 저장한 특정분야/주제의 데이터를 지원하며, 넷째 비휘발성(Non-Volatile)으로 모아 놓은 것(collection)"을 의미한다. 이것을 정리하면 [표 2-2]과 같다.¹¹⁾

[표 2-2] 데이터 웨어하우스의 4가지 의미

구 분	내 용
주제지향성 (Subject Oriented)	데이터웨어하우스내의 데이터는 일상적인 트랜잭션을 처리하는 프로세스 중심의 데이터와 달리 일정한 주제별로 구성된 데이터를 필요로 한다.
통합성 (Integtrated)	여러 운영계 시스템의 자료들은 표준화 및 통합화 과정을 거쳐 일관된 정보를 제공하여야 한다.
시계열성 (Time Variant)	운영계 시스템에서는 검색하는 그 순간의 데이터의 정확성이 중요시 된다. 그러나 데이터웨어하우스의 데이터는 이미 정해진 어느 한 시점의 정확성을 요구한다. 따라서 데이터 구조상 시간은 중요한 키로 작용한다.
비휘발성 (Non-Volatile)	데이터웨어하우스의 데이터에 대한 작업은 원시데이터로부터의 로딩과 적재된 자료에 대한 검색이며 자료의 변경 및 삭제와 같은 휘발성(Volatile) 작업은 없다.

10) 운영데이터베이스(Operational Database) ; 일반적인 비즈니스활동에서 지속적인 검색과 수정이 이뤄지며 매일 발생하는 트랜잭션을 처리하는 데이터베이스이다. OLTP 접근 방식을 사용하고 데이터의 저장을 위하여 최적화되도록 설계된다. (Nishith, 2005)

11) 최철민(2007), "데이터의 효율적인 분석을 위한 DataWarehouse기반 OLAP의 설계 및 구현", 한양대학교 공과대학원석사논문, p.10

가. 다차원 데이터 모델

데이터웨어하우스는 다차원(multi-dimensional) 데이터 모델을 기반으로 하는데, 조직이 원하는 여러 차원(측면, dimension)에서 데이터 모델링이 데이터 큐브(Data Cube)를 통해 이루어진다. 데이터 큐브와 다차원 데이터 모델링을 다루기 전에 먼저 차원(Dimension)과 사실(Fact)을 정의한다.

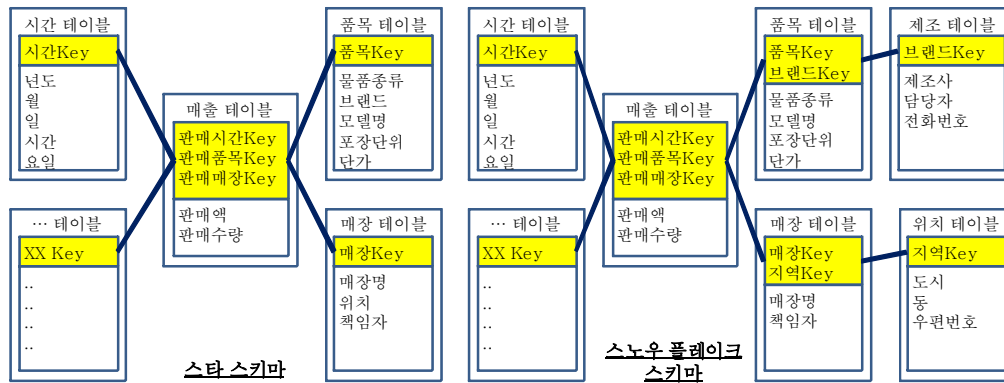
차원(Dimension)이란 조직이 데이터 레코드를 운용하는 이유의 대상이 되는 측면을 의미하며, 사실(Fact)이란 숫자적으로 표현되는 값을 의미한다. 예를 들어 마트에서 매장 관리에 따른 의사 결정을 위한 ‘판매분석 데이터 웨어하우스’를 만드는데, 연도, 품목, 지점에 따른 총 판매 금액을 분석할 필요가 있다고 하면, 이 경우 차원(Dimension)은 연도, 품목, 지점 세 가지이고, 총 판매 금액은 사실(Fact)이 된다.

정보분석자가 데이터웨어하우스를 활용하는 주된 질의 패턴에는 드릴다운과 롤업이 있다. 연도별 총 판매 금액, 다시 연도와 품목별 총 판매 금액, 그리고 연도, 품목, 지점별 총 판매 금액 정보를 계산해 내는 방식으로 점점 더 기존 차원에 또 다른 차원을 첨가해 세분화된 질의를 하는 질의 패턴을 드릴다운(Drill-Down)이라 한다. 이와 반대로 차원 수를 줄여가며 점점 요약된 형태의 정보를 얻어 나가는 질의 패턴은 롤업(Roll-Up)이라 한다. 이 밖에 슬라이싱(Slicing), 다이싱(Dicing), 피벗(Pivot) 등도 흔히 나타나는 질의 패턴이다.

나. 스타 스키마와 스노우플레이크 스키마

앞서 살펴보았듯이 데이터 웨어하우스에는 다차원 모델이 적합한데 이를 논리적 (Conceptual)으로 설계할 때 제일 자주 사용되는 스키마는 스타 스키마(Star Schema)와 스노우플레이크 스키마(Snowflake Schema)이다.

스타 스키마는 필요한 사실과 기타 속성들로 이루어진 사실 테이블과, 관심있는 차원과 그에 따른 부가적인 정보들을 각각 하나의 차원 테이블로 설계해, 사실 테이블과 차원 테이블을 외부키(Foreign Key)로 연결할 수 있도록 한다. <그림 2-4> 스타스키마와 스노우플레이크 스키마에서 보면 마치 별 모양이 되도록 구성한 스키마이다. 스타 스키마의 변형인 스노



<그림 2-4> 스타스키마와 스노우플레이크 스키마

우플레이크 스키마(Snowflake Schema)는 차원 테이블을 다시 정규화(Normalize)하여 이들 테이블들을 눈송이 모양으로 만든 스키마이다.

2. OLAP 도구와 비정형 분석 활용

일반적인 데이터베이스 시스템이 주로 조직이 필요로 하는 일상 업무(day-to-day operations)를 위한 OLTP(On-Line Transaction Processing)을 위한 시스템이라면, 데이터웨어하우스 시스템은 데이터 분석이나 의사 결정 등을 지원하는 OLAP(On-Line Analytical Processing)을 위한 시스템이다.

OLTP시스템과 OLAP시스템의 주요 차이를 비교하면 [표 2-3] OLTP 시스템과 OLAP 시스템의 비교와 같이 요약 될 수 있다.

[표 2-3] OLTP 시스템과 OLAP 시스템의 비교

특징	OLTP	OLAP
주요 용도	정보 입/출력	정보 분석
사용자	은행 창구원등의 일반 사용자	조직관리자, 분석관등의 지식 근로자
DB 설계 기법	ER 기반, 응용 프로그램 중심	스타/스노우플레이크 스키마, 주제 중심
데이터 중점	현재값 중심	시간에 따른 변화 중심

특징	OLTP	OLAP
데이터 요약	개별 데이터 값 중심	데이터 요약 중심
작업 단위	짧고 간단한 트랜잭션	대다수가 복잡한 질의
액세스 유형	읽고/쓰기 모두 필요	주로 읽기
취급 레코드 수	수십개 정도 (소량)	~ 수백만개 (대량)
사용자 수	수천명 단위 (다수)	수백명 단위 (소수)
데이터베이스 크기	100MB에서 수GB 정도	100GB에서 수 TB 정도
시스템 우선 순위	즉시성, 신뢰성	유연성, 가용성
성능 평가 척도	트랜잭션 Throughput	질의 Throughput

일반적으로 OLAP 서버는 MOLAP, ROLAP, DOLAP, HOLAP로 구분된다.

첫째로 MOLAP(Multidimensional OLAP)은 다차원 데이터베이스에 기반하여 데이터 큐브를 전용의 어레이에 기반한 다차원 저장소 엔진을 사용하여 저장 운용하는 방식이다. 비교적 네트워크 상의 데이터 이동이 최소화되며 다차원 데이터의 저장과 프로세싱에 동일한 엔진이 사용된다.

둘째로 ROLAP(Relational OLAP)은 관계형 데이터베이스나 확장된 관계형 데이터베이스를 사용해 다차원 모델링되는 데이터 큐브를 테이블 형태로 저장하고 관계형 데이터베이스와 클라이언트 사이를 연결해준다.

셋째로 DOLAP(Desktop OLAP)은 다차원 데이터의 저장 및 프로세싱이 모두 클라이언트에서 이루어지는 특징이 있으며, 타 OLAP제품에 비해 비교적 설치와 관리가 쉽고 유지 보수가 용이한 반면, 필요한 데이터가 모두 클라이언트로 이동될 필요가 있어서 대용량 데이터 처리에 한계가 있다.

넷째로 HOLAP(Hybrid OLAP)은 다차원 데이터의 저장 공간으로 다차원 데이터베이스와 관계형 데이터 베이스가 함께 사용될 수 있는 제품이며 일반적으로 요약된 데이터나 관계식에 의해 새로 계산된 데이터는 다차원 데이터베이스에 저장되며 상세데이터는 관계형 데이터베이스에 저장한다.

일반적으로 [표 2-4]와 같이 ROLAP제품은 MOLAP제품에 비해 복잡한 비즈니스 로직을 반영하기 힘들고 다차원 연산기능이 부족하지만 대용량

의 데이터를 무리없이 분석할 수 있는 장점을 가지고 있으며, KCTC에서 사용하는 오라클 디스커버리 관련제품도 여기에 속한다.

[표 2-4] MOLAP 및 ROLAP 특성 비교

구분	MOLAP	ROLAP
특성	다차원 모델링 및 질의도구	다차원 질의 도구
데이터 조작	읽기/쓰기	읽기 중심
데이터 요구량	대용량	초대용량
연산 기능	다양한 연산	제한된 연산
개발 주체	최종 사용자 주도형	전산부서 주도형

기간계에서 추출된 데이터는 변형되지 않고 그대로 임시저장소인 ODS (Operational Data Store)영역에 저장되어 진다. 그 이후에 데이터마트를 생성하기 전에 정제와 변형이 가능하다. ODS에 저장된 데이터는 특정 단일 기능의 요구를 위한 자료를 담고 있는 반면에 데이터마트는 다차원의 관점에서 전체 자료의 요약된 정보를 제공한다. ODS의 구축을 위해서는 전통적인 관계형 데이터베이스 스키마가 채택되고 데이터마트는 다차원 관점을 반영할 수 있는 보다 효율적인 데이터스키마로서 다차원 모델링 다차원 데이터베이스 스키마가 적용된다. ODS는 기간계 시스템의 부하를 줄이고 데이터마트에서 데이터를 형성하는데 문제가 생겼을 때 바로 복구가 가능하도록 하는데 그 목적이 있다.

데이터웨어하우징 시스템은 통상적으로 데이터 소스로부터 데이터를 클리닝하고 로딩하며 메타 정보 등을 관리하는 데이터웨어하우스 서버를 하위 티어(Tier)에 두고, 중간 티어에 OLAP 프로세스 서버를 두며, 상위 티어에 데이터마이닝 등의 의사 결정 응용 프로그램 도구를 두는 3계층 (3-Tier) 아키텍처를 가진다.

현재 전세계적으로 상업적으로 개발된 데이터웨어하우징 시스템 중에 대표적인 제품을 정리하면 [표 2-5]¹²⁾와 같다.

[표 2-5] OLAP 서버 제품 현황

OLAP Server	제조사	최신버전	라이선스
Essbase	Oracle	11.1.2.0	독점
icCube	MISConsulting SA	2.0.6	독점
Microsoft Analysis Services	Microsoft	2008 R2	독점
MicroStrategy Intelligence Server	MicroStrategy	9	독점
Mondrian OLAP server	Pentaho	3.2	EPL
Oracle Database OLAP Option	Oracle	11g R2	독점
Palo	Jedox	3.2 SR3	GPLv2 or EULA
SAS OLAP Server	SAS Institute	9.3	독점
SAP NetWeaver BW	SAP	7.20	독점
TM1	IBM	9.5.2	독점

국내제품으로는 LG엔시스(www.lgnsys.com)의 "이지다스(ezDAS)", 위세아이텍(www.wise.co.kr)의 "와이즈올랩(WISE OLAP)", 오브젠(www.obzen.com)의 "eCube OLAP", 포시에스(www.forcs.com)의 "OZ EQ" 등이 주로 사용되고 있다.

최근의 추세는 기존에 일반 사용자가 IT부서에 데이터를 요청하면 IT부서는 데이터 추출을 위한 쿼리를 작성하여 시스템 혹은 문서형태로 작성된 결과물을 제공하던 기존 방식에서 사용자가 직접 데이터를 추출하는 방향으로 변화하고 있다.

IT부서는 사용자/그룹 등록, 데이터 접근권한 설정 등 기본 환경을 설정하고 현업 활용도가 높은 데이터 세트 및 비정형 보고서를 개발하여 배포하면, 사용자가 직접 웹포탈에 접속하여 데이터 세트 및 보고서 파일을 검색하여 질의용 데이터 셋을 작성하고 리포트를 생성하여 보고서 생성, 편

12) 출처: http://en.wikipedia.org/wiki/Comparison_of_OLAP_Servers, 2012년 2월 접속

집, 배포까지 직접할 수 있도록 지원하는 EUC(End User Computing) 개념의 제품이 확산되고 있으며, "OZ EQ"가 그 한 예이다.

3. ETL 패키지

ETL(Extract, Transform, Load)은 데이터를 추출하고, 변환하고 적재하는 작업을 의미한다. 예를 들어, 실기동훈련의 훈련 관련 데이터가 수집된 방법에 따라서 여러 데이터베이스에 저장되어 있다면, 이 자료를 통합해 분석하려면 하나의 데이터 저장소에 모으는 작업을 해야 한다. 이럴 때 수집된 데이터를 쉽표로 구분된 CSV(Comma Separated Value) 파일 형식으로 FTP(File Transfer Protocol)를 통해 주고받는 데이터 연계 방식을 정해야 하고, 그 파일을 파싱하고 분석을 위한 데이터 저장소에 맞는 형식으로 바꿔서 저장하는 ETL 작업이 필요해진다.

많은 ETL 도구들은 데이터 저장소, 분석, 리포팅 도구들과 통합된 패키지의 일부로서 제공되기도 한다. 마이크로소프트사의 데이터베이스 패키지인 SQL Server 제품의 일부로 SQL Server Integration Services(이하 SSIS)가 제공되고 있고 Informatica Power Center에도 분석도구, 리포팅 도구들이 포함되어 있다.

데이터 연계, 변환 작업은 어느 시스템에서나 필요하며 이런 도구들을 필요시마다 개발한다는 것이 비생산적인 일이 될 수도 있다. ETL 도구들은 기본적으로 XML(eXtensible Markup Language), 엑셀 파일, CSV 등 다양한 데이터 포맷과 LDAP((Lightweight Directory Access Protocol)¹³⁾, SOAP(Simple Object Access Protocol)¹⁴⁾, FTP 등의 프로토콜을 지원하기 때문에 활용할 수 있는 용도가 많다.

ETL 솔루션들로는 마이크로소프트의 SSIS, IBM의 DataStage, Infomatica사의 제품 등 상용 제품들도 다수 나와 있다. 이런 상용제품들이 기술 지원을 받기가 용이하고, 안정된 기능들을 제공하지만, 상대적으로 도입이

13) 디렉터리 정보의 등록, 갱신, 삭제와 검색 등을 실행할 수 있으며, 통신망을 이용한 사용자 메일 주소나 이용자의 정보를 검색하는 데 주로 사용되는 통신규약.

14) 확장성 생성 언어(XML)와 하이퍼텍스트 전송 규약(HTTP) 등을 기반으로 하여 다른 컴퓨터에 있는 데이터나 서비스를 호출하기 위한 통신 규약.

쉬운 오픈소스 제품들도 개발을 주도하는 업체에서 더 기능이 강화되고 안정된 상용버전을 제공하며 교육, 컨설팅 등을 유료로 제공하고 있다.

Pentaho Data Integration 제품은 Pentaho사에서 Kettle이라는 오픈소스 프로젝트를 제품화한 것이다. Pentaho사에서 기술 지원과 안정성을 보장하는 상용 라이선스 버전도 있으며, 무료로 사용할 수 있는 부분도 비교적 기능적인 제약이 적은 편이다. 그리고 비교적 오래된 사용자 커뮤니티도 운영되고 있어 기술지원의 상당부분을 해소할 수 있는 장점이 있다. Kettle은 GUI 작업 환경을 제공하는 Spoon, 작성된 작업을 Console에서 실행할 수 있는 Kitchen과 Pan, 원격 및 분산 실행을 위한 서버인 Carte로 이루어져 있다.

이 툴은 작업에 대한 정보를 담은 메타데이터를 파일 또는 데이터베이스에 저장할 수 있고, JDBC 드라이버가 제공되는 대부분의 DB에서 제약 없이 저장소 생성이 가능하다. 작업정보가 저장될 테이블들을 자동으로 생성해주고, 저장소 간의 import, export 기능도 제공되어, DBA 등 프로그래밍 언어에 익숙하지 않은 실무자에게도 부담이 적은 등 장점이 많다.

특히 DB 테이블에 작업 로그를 기록하거나, 처리 단계마다의 성능을 그래프로 바로 GUI 도구에서 볼 수 있는 기능도 제공하여 배치 처리 과정의 가시성도 높은 편이다. 이 ETL 도구는 개별 패키지로 무료배포되며 BI 업무를 지원하는 리포팅 툴, 대시보드, OLAP, 데이터 마이닝 도구 등의 통합 패키지인 Pentaho BI Suite에도 포함되어 제공되고 있다.¹⁵⁾

이 외에도 오픈소스 ETL도구는 Talend Open Studio, CloverETL, ETL integrator, Scriptella, Jitterbit, Apatar등 여러 가지가 있으며¹⁶⁾ Java기반의 GUI를 제공하는 경우가 많아 편의성 면에서도 많이 뒤지지 않는다.

서비스를 위한 데이터 웨어하우스를 구축하기 위하여 기존의 데이터베이스에서 데이터를 적당한 형태로 가공하여 이전하는 데이터 마이그레이션 방법은 [표 2-6]¹⁷⁾와 같이 각각의 장단점이 있으며, 적용대상에 따라

15) 정상혁(2009), "오픈소스 ETL 솔루션 활용", 『월간 마이크로소프트웨어 2009년 10월 호』, 서울, 씨넷코리아 아이앤씨, pp.128~131

16) 출처: <http://www.manageability.org/blog/stuff/open-source-etl>, 2012년 4월 접속

17) 최종근(2011), "데이터 마이그레이션을 위한 오픈소스 ETL도구 평가", 숭실대학교 정보과학대학원석사논문, p.5

판단할 필요가 있다.

[표 2-6] 데이터 마이그레이션 방법 비교

수행방법	장점	단점
DBMS에 내장된 솔루션 이용	·별도의 도구를 설치하지 않아도 된다.	·한눈에 보기 어렵다. ·조작하기 불편하다
프로그래밍언어와 SQL언어를 이용	·관련언어 지식자에게는 다루기 쉽다.	·테이블수가 많으면 복잡해 진다.
ETL 도구를 이용	·비교적 단순하다. ·빠르고 안정적이다. ·실시간 동기화가 가능하다	·해당 제품 라이선스가 필요하다. ·교육이 필요하다.

4. Micro Data Service

Micro Data Service란 원자료에서 입력오류 등을 제거하여 공표하여 통계표 작성 등 데이터 가공의 기초 자료로 사용되는 통계원시자료(마이크로데이터, microdata)를 기본적인 통계표에서 얻을 수 없는 심층적인 분석을 원하는 다양한 계층의 이용자에게 제공하는 것을 말한다.

국내에서 대표적인 마이크로데이터서비스인 통계청 마이크로데이터 서비스시스템(MDSS, <http://mdss.kostat.go.kr>)¹⁸⁾은 2005년 온라인서비스로 구축에 착수하여 2006년부터 서비스를 시작하였고 2008년에 데이터웨어하우스기반으로 확장하여 2010년부터는 원격접근서비스도 제공하고 있다.

통계자료 제공에 관한 주요정책의 결정은 통계법시행령 및 통계청통계자료제공 규정에 의거하여 통계자료제공심의위원회를 구성하여 통계자료 제공과 관련된 사항을 심의하여 시행한다. 마이크로데이터의 관리는 통계청이 하며, 제공은 통계청과 한국통계진흥원이 수행한다. 심층적인 경제, 사회현상 분석을 원하는 다양한 계층의 이용자를 위해 제공대상에는 제한 없이 제공함을 원칙으로 하고 있다. 다만, 통계조사 응답자의 비밀보호를 위해 최대한 보호되는 범위 내에서 제공하며 제공분야는 사업체, 가구,

18) 출처: <http://mdss.kostat.go.kr/mdssect/>, 2012년 4월 접속

인구 등에 대해 총 38종의 통계조사자료를 제공하고 있다.

<그림 2-5>은 통계청 마이크로데이터 서비스의 홈페이지이다.



<그림 2-5> 통계청 마이크로데이터 서비스 홈페이지

일반적으로 통계자료는 사람, 사업체와 같은 개체별 기록들의 집합인 마이크로데이터와 개체별 기록들을 표 형태로 분류한 매크로 자료, 두 가지로 크게 나뉘어질 수 있다.

이러한 마이크로데이터 제공이 자료 이용자에게는 연구에 큰 도움이 되는 장점이 있지만, 개인, 사업체 등 개체에 대한 민감한 정보가 공개(노출)되는 문제가 일어날 수도 있다. 통계법 제33조 등에서는 개인 정보에 대한 비밀을 보호하도록 규정하고 있으며, 통계청에서도 마이크로데이터를 제공할 때에 개인, 사업체 등 통계조사응답자 개체의 정보 노출을 방지하기 위하여 자료공개관리(statistical disclosure control, "통계적 비밀보호기법"이라고도 함.)를 적용하고 있다.

이와 같이 마이크로데이터를 제공하려는 통계제공기관에서는 응답자의 노출위험을 줄이기 위하여 특정 개체와 연결될 수 있는 비밀정보를 주지 않고 자료를 공개할 수 있도록 통계 자료를 변경하여 새로운 마스크데이터(masking data)를 만들어 제공하는 SDC(Statistical Disclosure cntrol)¹⁹⁾를 수행한 뒤 제공하여야 한다.

통계작성기관이 가지고 있는 원래의 마이크로데이터는 개체를 직접적으로 식별할 수 있는 이름, 주민번호, 주소 등 공식적인 식별자, 성별, 나이, 학력, 직업분야, 혼인상태, 출생지 등 개체의 기본적인 사항인 식별 변수(잠재적 식별변수 또는 key변수라고도 함.)와 연 수입, 범죄기록, 건강기록 등 민감한 정보인 관심 변수로 이루어져 있다. 그리고 마이크로데이터를 외부에 공개할 때에는 개체의 비밀보호를 위하여 공식적 식별자를 제외한 익명화된 자료(원 자료)를 제공하고 있다.

그러나 외부인이 외부의 데이터베이스를 통하여 공식적 식별자와 식별 변수에 대한 정보를 갖고 있을 경우 식별 변수들의 조합을 비교하여 특정 개체의 민감한 정보인 관심 변수는 공개될 수 있을 것이다. 따라서, 익명화된 마이크로데이터도 공개위험을 평가하여 공개여부를 결정하고, 공개위험 평가결과에 따라 원 자료가 아닌 위험 축소 조치를 취한 자료로 제공되어야 할 것이다.

비밀보호방법을 마이크로데이터에 적용할 때 개체의 정보보호 외에 또한 가지 고려해야 할 점은 비밀 보호된 자료가 연구자에게 제공될 때, 연구자 입장에서 원 자료와 유사한 목적을 얻을 수 있어야 한다는 것이다. 개체의 정보보호를 위하여 원 자료를 지나치게 변경할 경우 자료 이용자는 원하는 정보를 얻지 못하거나 잘못된 정보를 얻을 수도 있을 것이다.

통계자료에 대한 비밀보호방법은 최적의 방법이 있는 것이 아니라 주어진 자료의 종류와 양, 공개할 표의 형태, 비밀보호 요구 수준 등을 고려하여 선택되어야 한다. 마이크로데이터 자료의 가공방법에 따른 비밀보호방법은 크게 외부 이용자에게 제공하는 자료의 정보량을 축소하는 방법과 자료를 변경하는 방법이 있으며 이를 분류하면 다음 [표 2-7]와 같다.²⁰⁾

19) 박영권(2007), "마이크로데이터에 대한 통계적 비밀보호기법 비교연구", 서울시립대학교 대학원석사논문, p.1

[표 2-7] 비밀보호를 위한 마이크로데이터 가공방법

자료종류 가공방법	마이크로데이터
자료 축소	·익명화(Anonymisation) ·표본(Sampling) ·모집단크기제한(Placing Restrictionon Population) ·세부 내용 축소(Reductionin Detail) ·Recoding ·감추기(Suppression)
자료 변조	·마이크로데이터에 대한 임의 변조 (Random Perturbation for Micro-Statistics) ·자료 교환(Data Swapping) ·보상에 의한 변조(Perturbation by Compensation) ·소집단화(Micro aggregation) ·Data shuffling

본 연구에서는 훈련자의 신상정보, 부대식별정보 등 기본정보에 대한 익명화와 비정상으로 추정되는 정보의 감추기 등을 우선 적용하기로 한다.

20) 권대홍(2008), "마이크로데이터의 공개 위험과 유용성을 고려한 공개방안 연구", 충남대학교 대학원석사논문, pp.1~5

제 3 장 과학화 훈련 데이터 분석

제 1 절 시뮬레이션 데이터 현황

1. 과학화훈련 실기동 자료 축적현황

KCTC 훈련분석 데이터웨어하우스에는 체계 구축시 부터 최근까지 다음의 [표 3-1]과 같은 상당량의 데이터가 누적되고 있다.

[표 3-1] 훈련분석 데이터웨어하우스의 전체 데이터량

구 분	계	상세스키마	집계스키마	로그스키마
테이블 수	169	64	93	12
데이터 건수	463,680,436	456,540,683	5,183,523	1,956,230
테이블당 평균 데이터 건수	2,743,671	7,133,448	55,737	163,019

상세스키마에는 데이터의 수집방법에 따라 위치정보와 같이 짧은 시간 간격으로 다량의 데이터(3.67억 건)가 누적되어 있는 경우도 있지만, 실기동 훈련중에 수동 입력하거나 여건상 자주 실시하지 않는 부분은 소량의 일부 데이터만 있는 경우도 있어 축적된 내용을 검토하여 제공여부를 결정하여야 하며 자동화 정보체계로 수집된 정보는 유효한 기간과 범위등을 고려하여 정제 후 활용하는 것이 필요하다. 테이블별 상세내역은 보안상 일반에 공개하기 어려울 것으로 판단하여 수록하지 않는다.

2. 실기동훈련 데이터의 신뢰성

KCTC의 데이터웨어하우스에 수록된 데이터 건수를 보면 2006년부터 2011년까지 훈련결과는 충실히 누적되어 있음을 확인할 수 있으나, 건수가 많은 데이터는 자동입력(MILES) 데이터, 단순자료 위주로 편중되어 있으며 지휘통제통신, 정보분야 등 수동으로 입력된 소량의 데이터는 정보량이

너무 적어 활용이 곤란할 것으로 판단되어 제외할 필요가 있어 보인다.

KCTC의 실기동 훈련중에 지휘·통제·통신 훈련, 기동훈련, 전투근무지원 훈련 등은 실제 상황과 동일하게 직접 실행하고 기록된 내용으로 현실세계를 그대로 반영하는 신뢰도가 높은 데이터라 할 수 있겠다.

훈련부대와 대항군의 교전상황이 발생할 때 교전훈련장비(MILES)와 PU(Personal Unit)의 센서에 의해서 수집된 모의 결과는 실제상황에 가까운 몰입도에서 얻어진 것으로 충분히 신뢰할 수 있다고 보아서 분석 서비스 대상에 포함할 수 있다.

훈련목적의 상황부여 및 모의논리에 의하여 적용된 사이버모의(포병, 화학, 항공, 장애물)에 의한 가상 피해현황 등은 일반적인 위게임 모델과 유사한 것으로 실제 기동훈련 대비 상대적으로 신뢰도가 낮은 것으로 판단되며 활용가치도 낮은 것으로 판단된다.

훈련부대의 인적사항은 일부 주기에서 부정확하거나 누락된 항목이 있어서 주기별로 정확성을 점검할 필요가 있으며 분석내용에 영향을 미칠 수 있는 요소이다.

GPS와 PU에 의하여 자동으로 수집되는 위치, 모의상황 정보는 환경적인 영향을 고려하여야 하며, 위치오차, 수신불량에 의한 손실, 장비조작에 따른 비정상값 등을 유효사거리, 이동속도 등의 논리적인 조건을 적용하여 제거할 경우 정확도를 높일 수 있을 것이다.

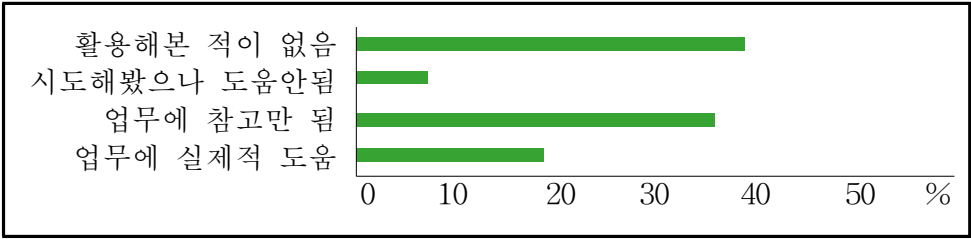
국면코드 등이 일부 훈련주기에서 코드값의 의미가 바뀌는 경우가 있어 별도의 고려가 필요한 부분이며 분석 활용을 어렵게 만드므로 일관성을 유지할 수 있도록 해야 하며, 정상적인 대항군 운용이 아닌 훈련주기에 대하여는 별도로 분리하여 분석하여야 한다.

3. 사용자 요구 분석

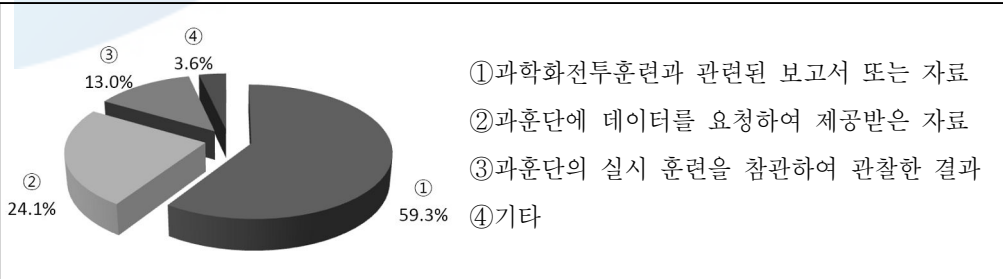
2010년 21세기 군사연구소가 수행한 정책연구보고서인 “과학화 전투훈련 성과평가”에서 육군 교육사령부 5개부서의 업무 담당자 85명을 대상으로 실시한 설문 분석결과 KCTC 훈련간 분석 데이터의 처리 및 데이터 축적에 대해 향후 발전시켜야 할 가장 큰 분야로 지적되고 있다.²¹⁾

[표 3-2] 과학화훈련 분석결과 또는 관련자료 활용 여부 설문조사 결과에서 보면 실제 활용실적이 20% 미만으로 저조하게 나타나는데, 주된 이유로 업무와 연계성이 없기 때문으로 응답하였다.

[표 3-2] 과학화훈련 분석결과 또는 관련자료 활용 여부



활용 경험자에 대해서만 추가 질문한 결과 <그림 3-2> 과학화전투훈련단 자료 활용 형태와 같이 59%이상이 과학화 전투훈련과 관련된 보고서 및 자료로 활용하고 있었다. 과학화 전투훈련단에서의 자료가 도움이 되지 않았다고 응답한 대상자들은 첫째 원하는 시기에 데이터를 얻을 수 없어서, 둘째 업무에 도움되는 데이터가 산출되지 않아서, 셋째 업무와 관련이 없어서 등의 순으로 답변하였다. 실제로 과학화전투훈련단에서는 예산, 인력, 장비 등의 제한으로 훈련결과 데이터 자료를 공식 문서화하기가 어려운 실정이다.

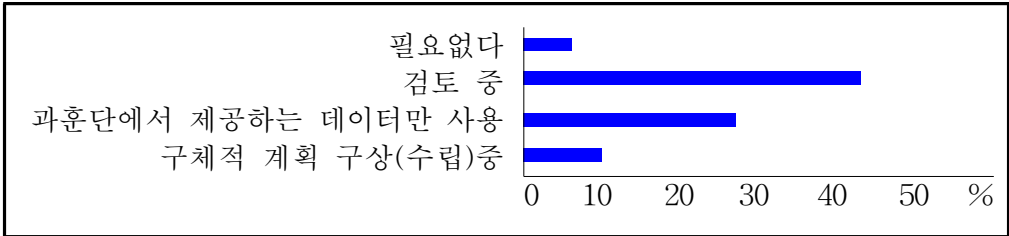


<그림 3-1> 과학화전투훈련단 자료 활용 형태

21) 김종만 외 2인, 전계서, p.133

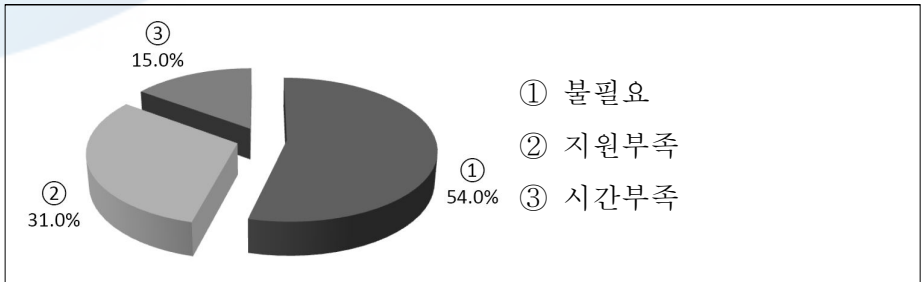
과학화전투훈련을 업무에 활용하고자 검토중인 업무담당자는 [표 3-3]과 같이 44%가 검토중이며, 구체적으로 계획 구상중인 업무담당자는 10%로 나타나, 꾸준한 관심을 보이고 있음을 알 수 있다.

[표 3-3] 과학화전투훈련장의 업무활용 검토



과학화전투훈련에서 발생하는 많은 데이터가 교육사령부 등에서 활용하기를 원하는 형태로 분석 및 가공된다면 그 활용도가 매우 향상될 것이며, 그 활용분야도 교리, 지휘관의 전투지휘능력, 소부대 전투기술 발전 등 다양해 질 것으로 분석된다. 따라서 과학화전투훈련 결과 데이터를 체계적으로 산출하고 관리·분석할 수 있는 시스템의 필요성은 높다.

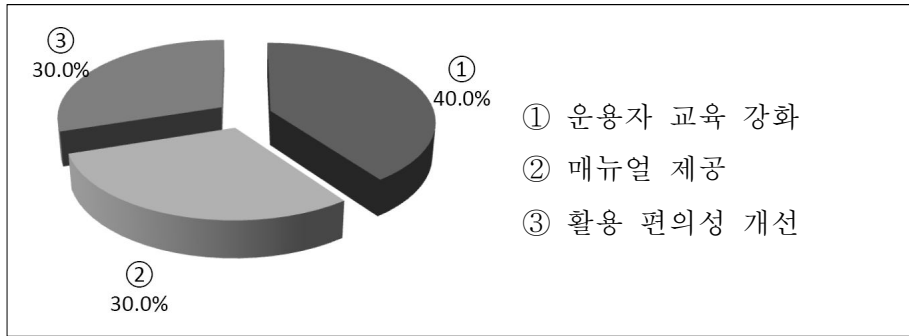
Live Simulation 결과 Data의 효율적 활용방안 모색과 KCTC 훈련간 축적된 Data의 문제점 식별 및 개선할 분야의 도출을 위하여 국방기술품질원에서 과학화전투훈련단의 훈련분석관 13명을 대상으로 실시한 의견수렴에서도 다음의 <그림 3-2> 비정형분석 미활용 사유와 같이 비정형분석의 필요성을 느끼지 못하거나(54%) 지원이 부족하다고(31%) 답하였다.



<그림 3-2> 비정형분석 미활용 사유

이에 따라 향후에 요구되는 사항은 <그림 3-3> 비정형분석 활용 활성

화 의견과 같이 운용자교육의 강화와(40%) 매뉴얼 제공 및 접속편의성 개선으로 파악되었다.



<그림 3-3> 비정형분석 활용 활성화 의견

이상의 사용자 의견을 종합해 볼때 다음의 사용자 요구를 만족한다면 데이터웨어하우스의 활용을 활성화할 수 있는 서비스체계가 될 것이다.

첫째 현재의 업무와 연관성이 있는 분야에서 활용분야를 찾을 수 있어야 한다. KCTC에서 수집이 가능한 데이터에 대한 명세서(자료사전)를 공지하며 분석사례를 공유하여 연구자로 하여금 활용의 유용성을 알릴 필요가 있다.

둘째 다양한 분석방법과 가공정보가 활발히 만들어지고 공유될 수 있어야 한다. 기존의 훈련분석에 사용되는 정형분석은 사후분석에 잘 활용되고 있다고 보인다. 비정형분석은 주로 OLAP도구(오라클 디스커버리)를 사용하는 분야인데 먼저 분석자의 컴퓨터에 클라이언트 프로그램이 설치되어야 하고, 지정된 단말에서만 사용이 가능하며, 충분한 교육지원이 되지 못하여 사용자들이 어려움을 호소하는 것으로 보인다. 사용이 용이하고 교육지원을 받기 쉬운 분석도구를 선정할 필요가 있다.

셋째 원하는 장소에서 쉽게 사용할 수 있는 편의성을 제공할 필요가 있으며 최근에는 웹기반의 분석환경으로 대부분 전환된 상태이다.

넷째 교육이 간단하거나 충분히 이뤄질 수 있어야 할 것이다.

또한, 군내 공개 시에 개인정보나 보호해야 할 특정요소가 노출되지 않도록 데이터에 대한 보호조치도 강구되어야 할 것이다.

제 4 장 과학화훈련 데이터 활용 활성화 연구

제 1 절 데이터 정제를 통한 활성화 방안

1. 데이터 공개 범위 확대

훈련분석에 영향이 없는 참여자의 신상정보를 선별하여 제거하고 공개함으로써 좀더 안전한 데이터 활용이 가능하다. 다음의 [표 4-1]과 같이 훈련자 인적사항으로부터 활용에 지장이 없다고 판단되는 개인정보항목을 식별하여 ETL 작업시에 성명, 군번, 수정일자 등을 공란으로 변조하며, 실제 전투원이 아닌 장비에 대하여는 해당 데이터를 삭제하여 OLAP전용 데이터웨어하우스에 적재한다. 이러한 조치를 하므로써 데이터를 공개할 수 있는 범위를 확대할 수 있을 것이다.

[표 4-1] 훈련자 인적정보 선별 대상

항목명	선별내용	항목명	선별내용
훈련자ID	Key값 유지	우시력	유지
훈련자명	삭제(개인정보보호)	체중	유지
훈련자유형	전투원의 장비삭제	키	유지
병력정보ID	Key값 유지	학력	유지
계급	유지	병과ID	유지
직책	유지	임관구분ID	Key값 유지
주특기	유지	훈련제대ID	Key값 유지
군번	삭제(개인식별정보)	주기훈련ID	Key값 유지
복무연차	유지	수정일자	삭제(무의미)
혈액형	유지	피아구분	유지
좌시력	유지	-	-

2. 신뢰성 있는 데이터 제공

KCTC의 실기동 훈련간 자동화체계(GPS, MILES 등)를 통하여 수집되는 데이터에는 환경조건에 따른 장애와 오차 및 훈련목적의 활동(대기, 재투입 등)에 따라 비정상적인 값이 혼입된다. 예를 들어 주요 국면간(공격, 방어)에 아군이 적에게 직사화기로 사격/명중한 거리 값 중에서 공격/방어

훈련간 발생한 17만여건의 피격데이터 중에서 일부는 피격거리 데이터가 없는 상태이며, 일부는 직사화기의 유효사거리(600m)이상에서 피격된 것으로 나타나 데이터의 신뢰성에 의문을 갖게 한다.²²⁾ 이는 GPS, MILES 등에서 여러 원인을 찾을 수 있으나, 가장 큰 원인은 GPS 오류인 것으로 과학화훈련단에서 확인하여 개선된 장비로 보급되고 있는 부분이다. 데이터가 없거나 “0” 또는 유효사거리를 넘는 등 상식적인 범위를 벗어나는 데이터에 대하여는 ETL 도구를 이용하여 서비스용 데이터웨어하우스로 이전할 때 [표 4-2]와 같이 해당항목의 값을 확인하여 제거하고 외부에 공개할 수 있도록 하였다.

[표 4-2] 훈련자 상태 정보 추출대상 항목

훈련자 상태 정보 테이블	훈련자 상태 정보
필드 설명	정제 조건
피격화기ID	피격화기ID = 2(K-1 소총) or 3(K-2 소총)
피격거리(EAST)	$0 < \sqrt{EAST^2 + NORTH^2} \leq 600$
피격거리(NORTH)	

제 2 절 데이터 접근성 개선을 통한 활성화 방안

1. 데이터 분석환경 개선

KCTC에서는 훈련이 종료됨과 동시에 빠르고 일관된 사후강평을 하기 위하여 체계구축 당시에 개발된 정형분석 모델에 의한 데이터 평가 결과를 대부분 의존하고 있으며 비정형 분석수단을 적극적으로 활용하기에는 시간이 부족한 것으로 보인다. 훈련분석관, 관찰통제관 등이 관찰자료를 입력해야 생성 가능한 수동 분석모델이 전체 분석 시스템의 절반을 차지 하며, 모든 분석모델이 사후강평에 사용되는 것은 아니기 때문에 입력이 누락되는 경우도 예상되는 상태이다.

현재 구축중인 연대급 KCTC로 확장시 기존의 어려움이 상당부분 보완

22) 군 전력에 대한 불필요한 노출의 우려가 있어 구체적인 수치는 제시하지 않음.

될 예정이지만, 훈련 규모가 더 커지는 만큼 자동화체계의 중요성은 더 커질 것이다.

KCTC가 훈련만을 위한 체계가 아니며 전투실험 등 새로운 전술교리의 개발과 신규획득장비의 야전운용 평가 등 다양한 분야로 활용되기 위해서는 여러 가지 조건에 따라 변화하며 분석할 수 있는 비정형 분석도구의 활용이 필수적이다.

2006년 구축당시에 도입하여 보유중인 비정형 분석도구(오라클 디스커버러)는 활용지원의 어려움으로 적극적인 활용사례가 적은 상태이지만, 앞으로 활성화 되어야 할 분야임에는 틀림이 없다.

보통 비정형분석 도구는 다양한 분석관의 요구에 효과적으로 대응하기 위하여 분석관이 직접 데이터웨어하우스의 데이터에 접근할 수 있는 기회를 제공하기 위한 것이다. KCTC에서 운용중인 오라클 디스커버러 패키지는 소프트웨어가 설치되어 있는 워크스테이션에서 훈련분석 프로그램의 “비정형 분석”메뉴를 호출하여 “비정형 질의”를 선택하면 미리 등록해 둔 쿼리를 선택하여 실행시키는 형태로 되어 있다. 이는 개발 당시 사용자의 편의성을 높이기 위한 방법이었는데 주 사용자인 분석관들이 업무상 반드시 사용해야 할 필요성이 많지 않아서 활성화되지 못한 것으로 판단된다.

“디스커버러”가 최종사용자를 위한 OLAP 분석도구이기는 하지만 클라이언트-서버 구조의 프로그램으로 사용자의 분석용 워크스테이션에 설치하고 동작환경을 설정하는데 불편함이 있고, 효과적으로 사용하기 위해서는 관리자의 도움과 적절한 교육 및 기술지원을 필요로 한다.

이러한 어려움을 해소하고 분석자료의 활용 활성화를 위한 개선된 환경을 제공할 수 있는 분석도구의 조건으로는 첫째, 데이터웨어하우스에 대한 다차원분석이 용이하고 둘째, 직관적이고 사용자가 직접 데이터를 분석할 수 있으며 셋째, 사용자 레벨에 따른 접근통제 및 이력관리 등 보안성을 제공하며 넷째, 표 이외의 도표, 그림 등 사용자 입맛에 맞는 효과적인 가공방법을 제공할 수 있어야 할 것이다.

2014년부터 운용할 연대급 KCTC체계에서는 EUC 기반의 쿼리 솔루션으로 “Oz EQ”가 도입될 예정이다.

“Oz EQ”는 국내업체에서 2007년 발표되었으며, 웹기반 환경의 서비스를 제공하여 사용자 PC에서 별도의 설치없이 사용이 가능하고, 데이터 분석 및 공유가 쉬워 일반 사용자 중심의 손쉬운 분석 기능을 제공 한다. 사용자는 주어진 권한을 기반으로 각 데이터웨어하우스 테이블의 데이터에 직접 접근하여 데이터 정렬, 필터링, Drill Up/Down, 그룹핑 등 다중 데이터 세트간 결합 등을 통해 데이터 비교, 분석을 할 수 있고 사용자별 접근권한관리와 국내개발에 따른 기술지원이 용이한 장점을 갖고 있다.

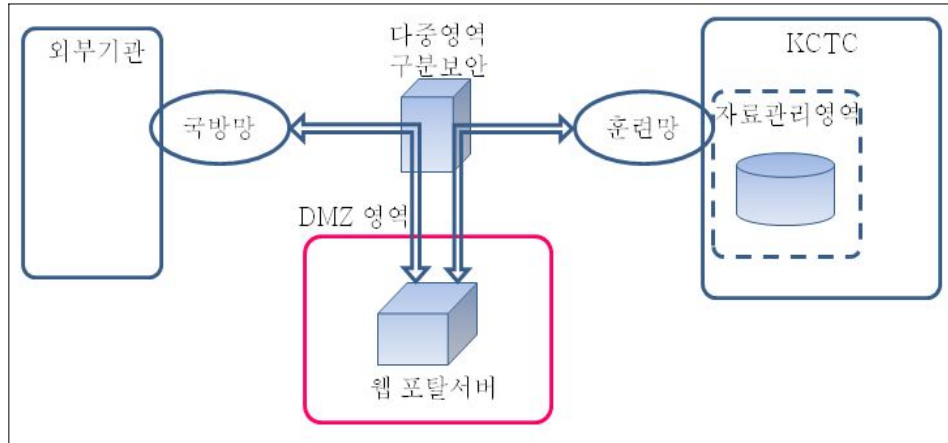
금번 연구시에는 시험용 서버PC에 “Oz EQ”를 비롯한 관련 프로그램을 설치하여 기존의 훈련분석 서버에 있는 데이터웨어하우스를 변경하지 않고 별도의 통합 데이터웨어하우스에서 비정형분석 모형을 용이하게 생성하고 운용할 수 있음을 확인하였으며 이를 통하여 분석을 위한 데이터웨어하우스를 유연하게 재구성할 수 있음을 알 수 있었다.

실기동 훈련의 결과에 대한 활용을 활성화하기 위한 방안으로 훈련결과에 대한 연구의 기회를 KCTC내로 한정하지 않고 제한적으로 공개하여 훈련참여부대나 교육기관, 연구기관에서 적극적인 실기동훈련 데이터를 분석 활용할 수 있도록 하는 것이 바람직하다.

다양한 각도에서 데이터가 분석되고 활용될 수 있도록 신뢰할 수 있는 데이터가 축적된 항목을 식별하고 비밀요소, 비정상 오류 데이터를 제거한 신뢰도 높은 마이크로데이터를 인가된 사용자에게 국방전산통신망을 통하여 온라인으로 제공한다면 활용환경이 훨씬 개선될 것이다.

연대급 KCTC 구축시 외부기관 자료제공 기능이 설계에 반영되어 있으며 <그림 4-2>과 같이 훈련망 내부에 자료교환영역을 설정하고 국방망과 격리된 상태에서 외부 웹 포탈을 통하여 외부기관(학교기관 및 훈련부대)이 KCTC의 편집된 훈련영상, 상황도, 훈련분석자료, 사후검토자료 등의 훈련자료를 활용할 수 있도록 할 예정이다. 이때 자료관리영역의 임시저장 영역에 저장하는 훈련분석자료를 생성할 때 본 연구의 적용이 가능할 것이다.

또한 통계청의 마이크로데이터시스템과 같은 통계정보제공서비스는 이를 위한 참조모델로 활용할 수 있을 것이다.



<그림 4-1> 외부기관용 웹 포탈 운용 구성

2. 데이터의 정제 가공

훈련분석을 위하여 축적된 데이터웨어하우스는 크게 상세자료와 집계자료로 구분되어 2011년 4월까지 총 4억 6,368만건이 저장되었으며 다음 [표 4-3]과 같은 내용의 데이터가 수록되어 있다.

[표 4-3] 훈련분석 데이터웨어하우스 데이터 내용

구 분	전장기능	내 용
상세자료	지휘통제통신	통신감청 이력정보 외 5종
	정보	가용자산 운용 외 1종
	기동	지뢰지대 모의정보 외 4종
	화력	곡사화력 요청/지원 정보
	방호	화생방 모의정보 외 2종
	전투근무지원	훈련자 탄약상태 정보 외 4종
집계자료	지휘통제통신	통신수단 운용시간 외 1종
	정보	기상정보 외 5종
	기동	장애물별 기동지연효과 외 2종
	화력	곡사화력 요청건수 외 3종
	전투근무지원	제대별 화기별 발사수 외 15종

각각의 데이터 중 GPS, MILES 등으로 자동화 수집된 경우에는 피격율
에서와 같이 논리적으로 타당한 범위를 판단하여 ETL도구로 걸러내며,
개인정보 등은 연구가 가능한 범위 내에서 최대한 삭제하도록 한다. 모든
데이터를 공개하는 것 보다는 요청되는 항목부터 웹-OLAP 서비스에 선
택적으로 검토하여 제공하는 것이 바람직하다.

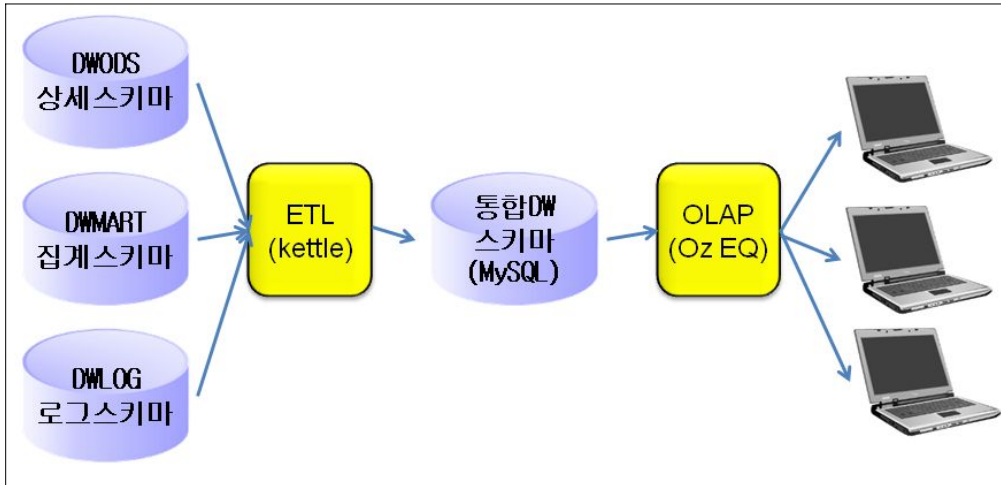
3. 시험체계 구현

본 연구에서는 기존의 KCTC 데이터웨어하우스에서 ETL 도구를 이용
하여 대외서비스를 위한 통합DW(데이터웨어하우스)를 생성하며 이를 바
탕으로 WEB방식의 OLAP 도구를 사용하여 자료요청자에게 마이크로데이
터형식 또는 집계된 데이터를 제공하는 형태로 <그림 4-2>와 같은
WEB-OLAP 시험체계를 구성하였다.

시험체계는 기존의 훈련분석 데이터웨어하우스의 제대위치정보 테이블
로부터 피격율에 대한 신뢰할 수 있는 수준의 데이터를 추출하여 통합DW
에 적재하고 외부에 제공하는 환경을 구현하고자 한다.

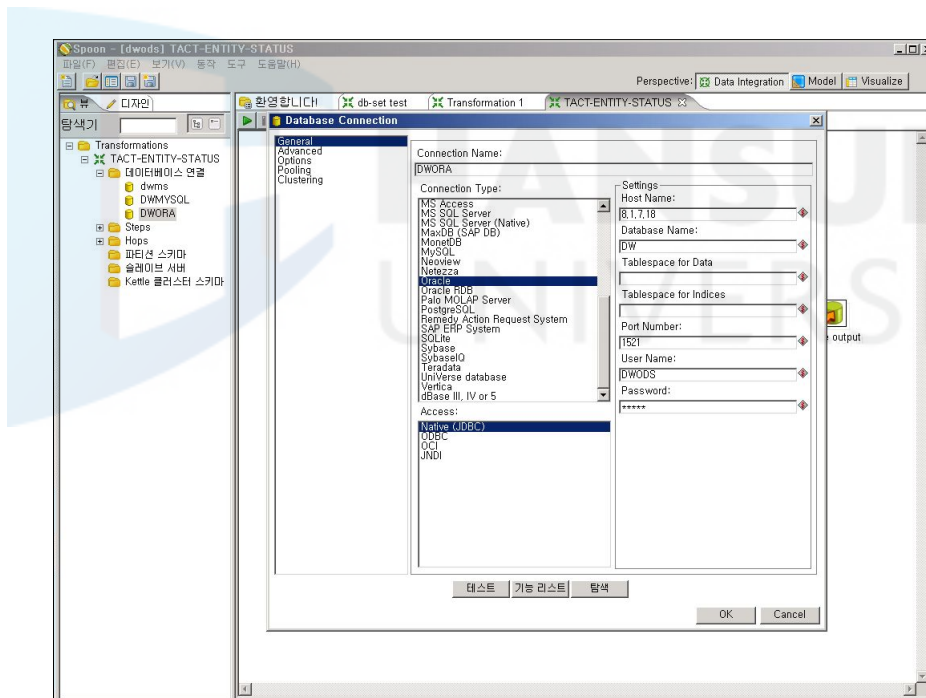
www.oracle.com에서 MySQL 5.5.21버전을 내려받아 시험체계용 PC에
데이터베이스로 설치한 다음 데이터 정제 및 변환을 위한 ETL Tool로
www.pentaho.com에서 Data Integration 4.2버전 (kettle)을 내려 받아 설
치한다. Kettle을 설치하면서 GUI 작업 환경을 제공하는 Spoon, 작성된
작업을 Console에서 실행할 수 있는 Kitchen과 Pan을 설치한다. MySQL
과 Data Integration(kettle)은 오픈소스 프로젝트로 무료로 받을 수 있다.
Data 분석제공을 위한 OLAP Tool로 Oz EQ 5.1을 시험체계를 설치하였
다. Oz EQ는 국내SW업체인 “포시에스”에서 판매하는 상용도구 SW이어서
제조사에 협조한 한시 인증을 받아 설치하였다.

KCTC 데이터웨어하우스의 데이터베이스와 서비스용 MySQL 데이터베
이스 접속정보를 kettle과 EQ솔루션에 등록한다.



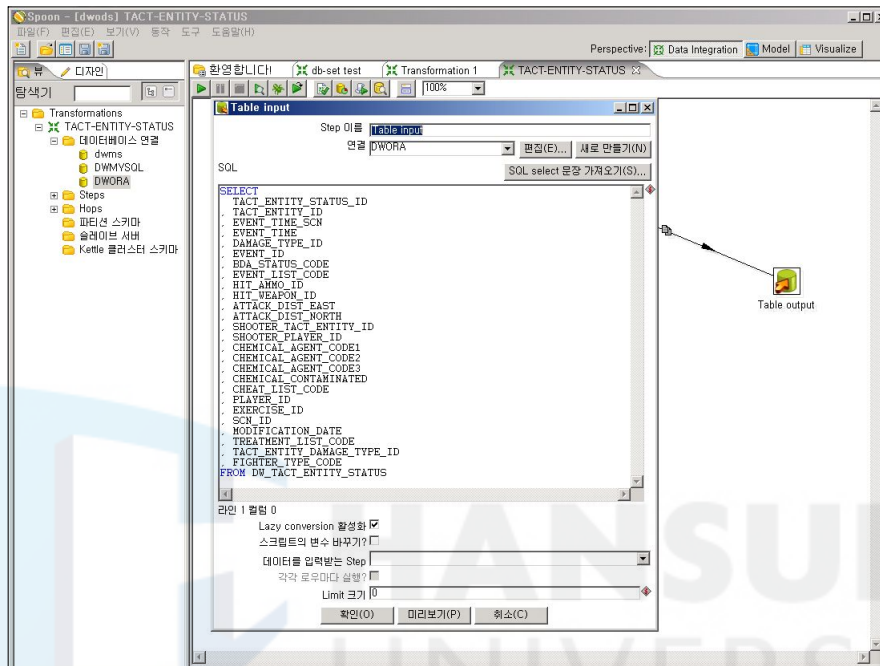
<그림 4-2> WEB-OLAP 시험체계 구성도

가. Pentaho Data Integration을 활용한 데이터 마이그레이션



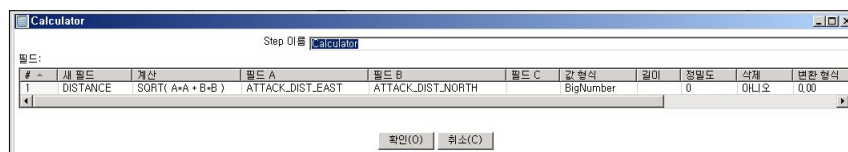
<그림 4-3 > Pentaho Data Integration Database Connection

새로운 Transformation을 생성 후 디자인 네비게이션에서 Input의 Table Input모듈을 생성한다. <그림 4-3>과 같이 Tableinput모듈에서 데이터베이스를 연결 설정을 할 수 있다. 원하는 DBMS 종류와 연결 방법별로 설정이 가능하다. 원본 데이터웨어하우스에 Oracle의 JDBC를 이용하여 데이터베이스 연결을 맺었다. 이와 같은 방법으로 디자인 네비게이션에서 Output의 Tableoutput모듈을 생성 후 데이터베이스 연결을 설정하였다.



<그림 4-4> Pentaho Data Integration SQL select

SQL select문장 가져오기를 통하여 연결된 데이터베이스 안의 테이블에서 SQL select문장을 가져 올 수 있다. <그림 4-4>과 같이 Table Input 모듈에서 연결한 데이터베이스의 As_Is테이블에서 데이터를 추출한다.



<그림 4-5> Pentaho Data Integration Calculator

Table input 모듈에서 Calculator 모듈로 hop 설정을 한다. Calculator 모듈에서 <그림 4-5>와 같이 피격거리를 계산할 새 필드와 계산식을 설정한다.

Filter rows

Step 이름: Filter rows

'true'인 경우 보낼 Step: Number range

'false'인 경우 보낼 Step: Text file output

조건:

DISTANCE > [0]

AND

DISTANCE <= [600]

확인(O) 취소(C)

<그림 4-6> Pentaho Data Integration Filter rows

Calculator 모듈에서 Filter rows 모듈로 hop 설정을 한다. 계산된 거리값이 0 보다 크고 600 이하이면 다음 단계로 진행하고 범위가 벗어나는 값은 확인을 위해서 별도로 저장한다. Filter rows 모듈에서 <그림 4-6>와 같이 피격거리의 적정범위와 처리 Flow를 설정한다.

Number ranges

Step 이름: Number range

입력 필드: DISTANCE

출력 필드: DIST_ID

기본값(범위 매칭이 없는 경우): unknown

범위 (최소값 <= x < 최대값):

#	하한	상한	값
1	0,0	50,0	0
2	50,0	100,0	1
3	100,0	150,0	2
4	150,0	200,0	3
5	200,0	250,0	4
6	250,0	601,0	5

OK Cancel

<그림 4-7> Pentaho Data Integration Number ranges

Filter rows 모듈에서 True이면 Number ranges 모듈로 hop 설정을 한다. 통계분석을 쉽게하기 위하여 <그림 4-7>과 같이 피격된 거리값을 50미터 간격으로 구분하여 그룹값을 부여하고 다음 단계로 진행한다.

Step 이름: Table output

연결: DWMYSQL

대상 스키마:

대상 테이블: dw_tact_entity_status

Commit 크기: 1000

Truncate table: ☐

Insert 오류 무시: ☐

데이터베이스 폴드 지정: ☒

메인 옵션: 데이터베이스 폴드

입력할 필드:

#	테이블 필드	스트림 필드
1	TACT_ENTIT...	TACT_ENTIT...
2	TACT_ENTIT...	TACT_ENTIT...
3	EVENT_TIME...	EVENT_TIME...
4	EVENT_TIME...	EVENT_TIME...
5	DAMAGE_TV...	DAMAGE_TV...
6	EVENT_ID	EVENT_ID
7	BDA_STATUS...	BDA_STATUS...
8	EVENT_LIST...	EVENT_LIST...
9	HIT_AMMO_ID	HIT_AMMO_ID
10	HIT_WEAPON...	HIT_WEAPON...
11	ATTACK_DIS...	ATTACK_DIS...
12	ATTACK_DIS...	ATTACK_DIS...
13	SHOOTER_T...	SHOOTER_T...
14	SHOOTER_PL...	SHOOTER_PL...
15	CHEMICAL_A...	CHEMICAL_A...
16	CHEMICAL_A...	CHEMICAL_A...
17	CHEMICAL_A...	CHEMICAL_A...
18	CHEMICAL_C...	CHEMICAL_C...
19	CHEAT_LIST...	CHEAT_LIST...
20	PLAYER_ID	PLAYER_ID
21	EXERCISE_ID	EXERCISE_ID
22	SCN_ID	SCN_ID
23	MODIFICATIO...	MODIFICATIO...
24	TREATMENT...	TREATMENT...
25	TACT_ENTIT...	TACT_ENTIT...
26	FIGHTER_TY...	FIGHTER_TY...
27	DISTANCE	DISTANCE
28	DIST_ID	DIST_ID

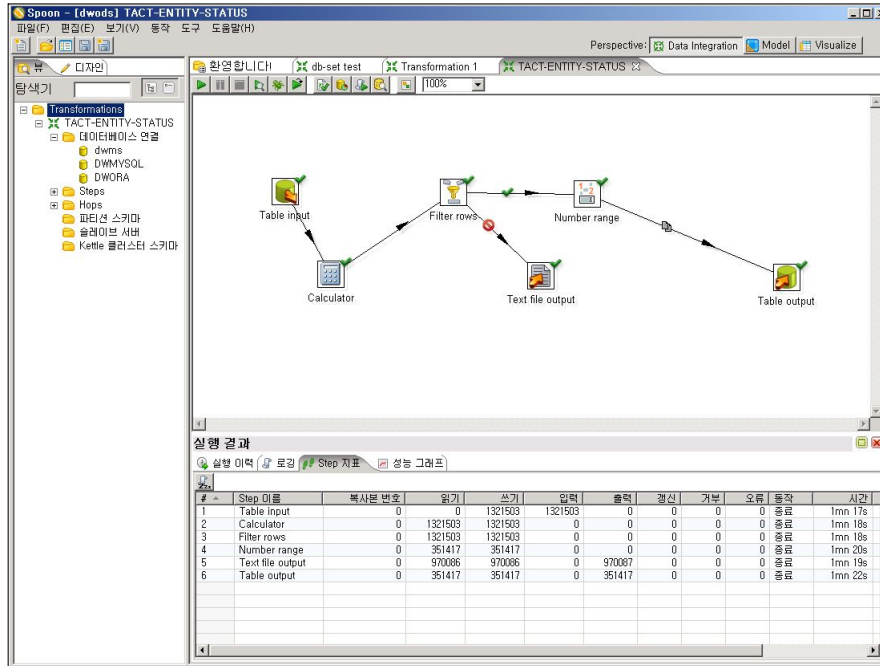
필드 가져오기(G)

필드 매핑 입력

확인(O) 취소(C) SQL

<그림 4-8> Pentaho Data Integration Table out 필드 매핑

Number ranges 모듈에서 Table output 모듈로 hop 설정을 한다. Table output 모듈에서 <그림 4-8>과 같이 스트림 필드와 데이터베이스에 기록할 테이블 필드에 1:1매핑을 설정한다.

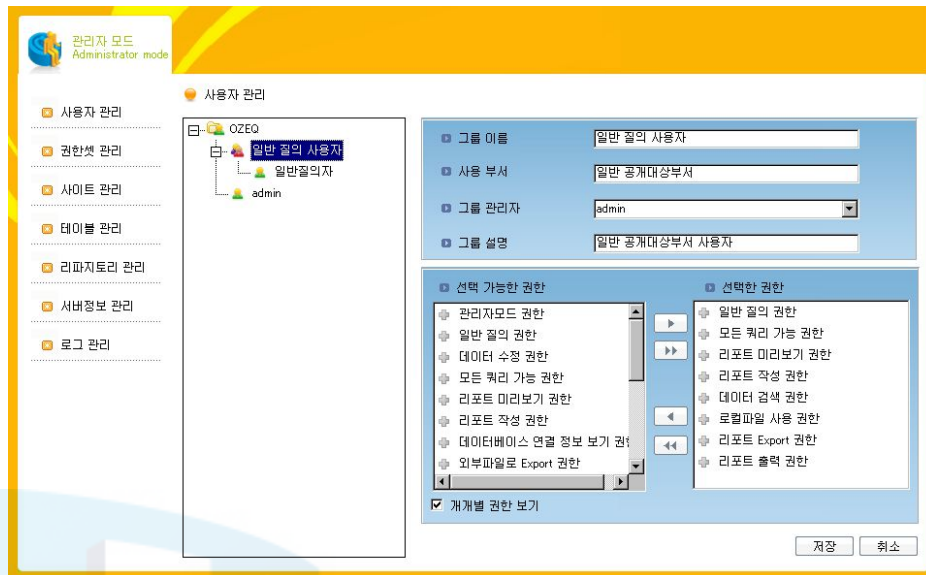


<그림 4-9> Pentaho Data Integration 데이터 마이그레이션 실행

설정과정을 맞춘 후 실행버튼을 클릭하여 데이터 마이그레이션을 수행한다. <그림 4-9>의 하단에 step지표에서 수행과정이 표시된다. 여러 개의 모듈이 거의 같은 시간에 시작되기 때문에 모듈수행 시간 중 많은 시간이 총수행시간이 된다. 원본 데이터웨어하우스의 132만건을 분류하여 유효 데이터 35만건을 추출하는데 1분 22초가 소요되어 초당 1만6천건을 처리하는 것으로 파악되었다. 유효데이터가 26%로 낮게 보이는 것은 모든 국면의 전 주기, 모든 피격정보가 자동으로 수집되어 있어서 유효한 시간과 장소, 화기, 유효한 국면을 반영하여 판단하여야 하고 실제로 GPS 장비의 오류정보도 상당부분을 차지할 것으로 예상이 되는 부분이다. KCTC의 데이터웨어하우스 전체 4억6,400만건을 모두 전환한다면 단순비례계산으로 8시간이 소요되게 되므로 시험장비가 데스크톱PC인 점을 감안하면 처리시간은 문제가 없어 보인다.

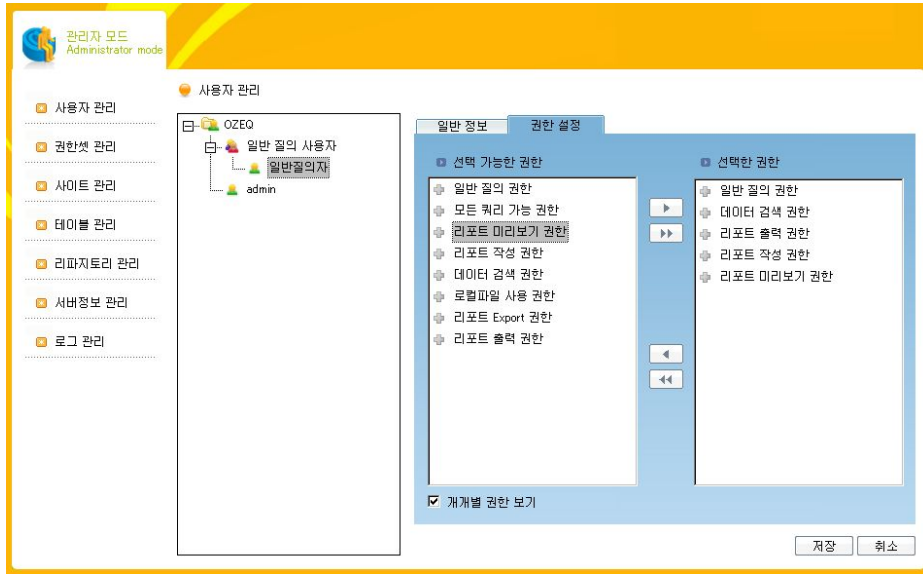
나. Oz EQ를 활용한 웹기반 OLAP 시험체계

Oz EQ 패키지를 설치하고 관리자(admin)로 접속한 후, <그림 4-10>의 화면에서와 같이 사용자 그룹을 등록하여 그룹에 공통적으로 적용할 질의 및 리포팅 권한을 부여한다.



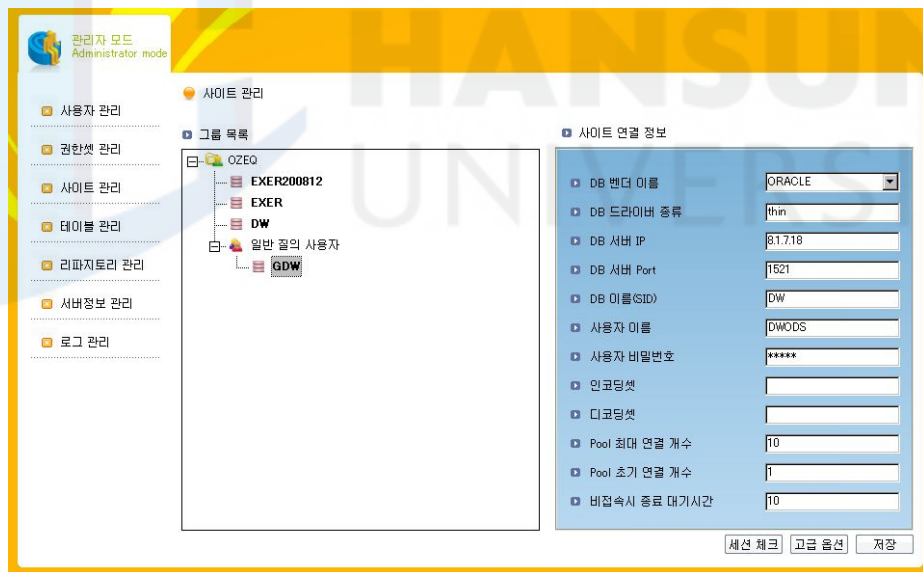
<그림 4-10> 사용자 그룹 관리 화면

등록한 사용자 그룹에 소속된 사용자를 <그림 4-11>의 화면과 같이 등록하면서 일반정보와 그룹에 부여된 권한 내에서 사용자별로 권한을 부여한다.



<그림 4-11> 사용자 권한 설정 화면

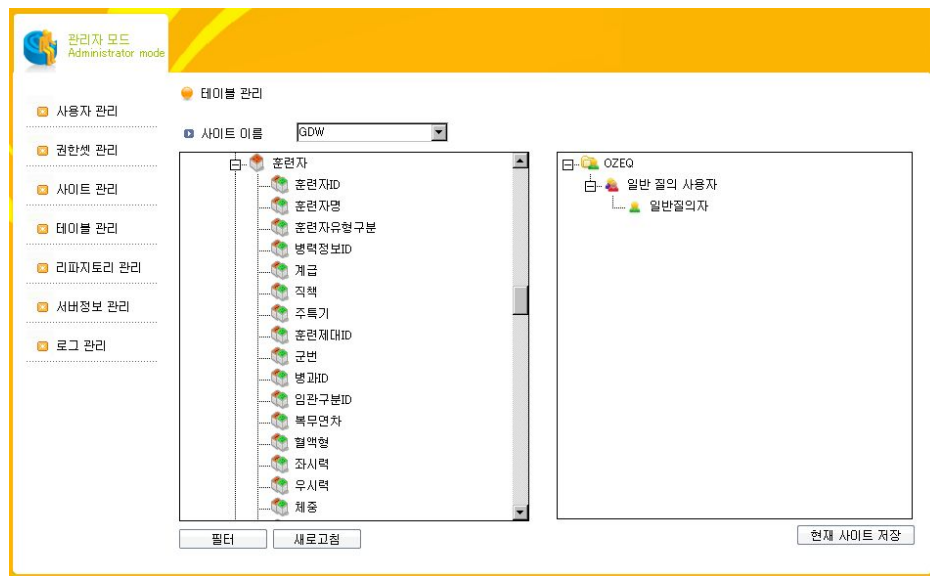
사이트 관리 메뉴를 선택하여 <그림 4-12>의 화면과 같이 해당 사용자 그룹에 접속이 허용된 데이터베이스의 사이트 연결 정보를 등록한다



<그림 4-12> 사이트 관리 화면

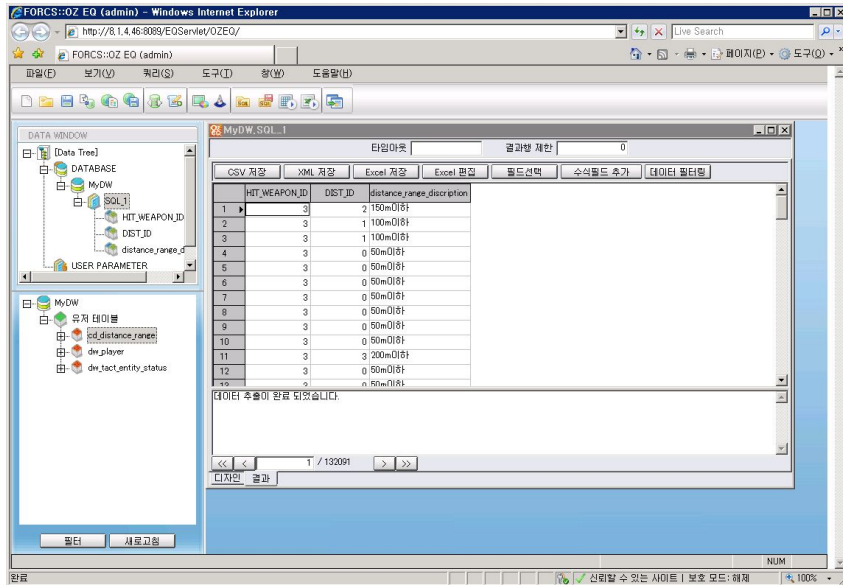
테이블 관리 메뉴를 선택하여 <그림 4-13>의 화면과 같이 해당되는 사

이트에 접속을 허용할 데이터베이스와 사용자를 선택하여 등록한다.



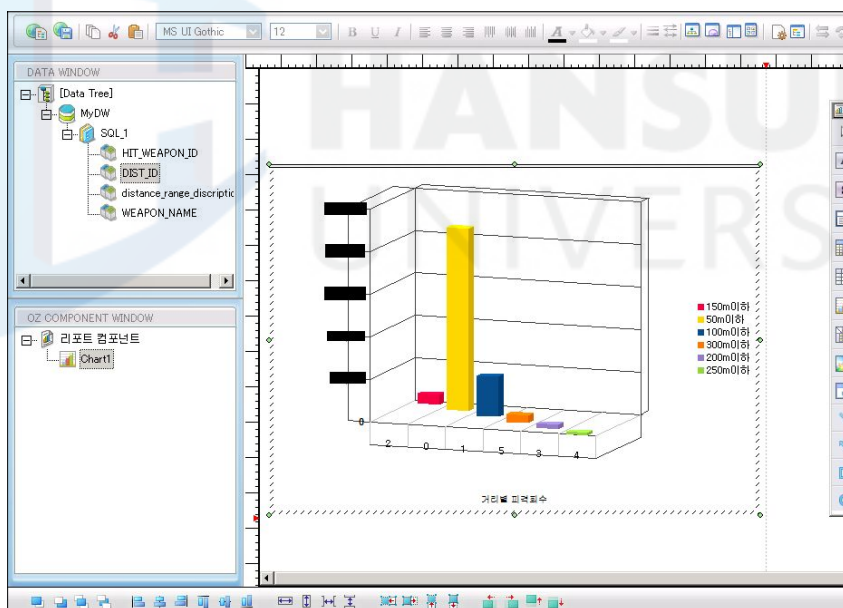
<그림 4-13> 테이블 관리 화면

EQ 시스템의 사용자로 접속하여 사용모드를 “쿼리 디자이너”로 선택한 후 검색할 데이터 창에서 해당되는 테이블과 디자인 화면에서 필요한 항목을 선택한 후 <그림 4-14>와 같이 결과를 검색한다. 이때 마법사 기능을 사용하면 손쉽게 검색결과를 얻을 수 있다.



<그림 4-14> 피격거리별 검색결과 화면

필요한 항목의 데이터를 검색한 후 <그림 4-15>와 같은 분석이 가능하다.



<그림 4-15> 거리별 피격인원

제 5 장 결 론

군사훈련 분야에서 최소의 비용으로 최대의 효과를 얻을 수 있는 새로운 대안으로서 정보통신기술 및 시뮬레이션 기법을 이용한 과학화 전투훈련은 계속 확대 적용되는 추세이다.

본 연구에서는 육군의 KCTC에서 다년간의 실기동 훈련을 통해 축적된 데이터에 대하여 분석자료를 효과적으로 활용할 수 있도록 데이터웨어하우스의 활용을 활성화할 수 있는 방안을 다음과 같이 모색하였다.

첫째, 웹기반의 비정형 분석도구를 시험 적용하여 KCTC의 실기동훈련 결과에 대한 M&S 분석자료의 일관성과 신뢰성을 확보하면서 검색 활용을 지원할 수 있는 방안을 모색하였다.

둘째, ETL도구를 시험 적용함으로써 오류로 판단되는 데이터와 공개하기 곤란한 정보를 제거하고 순수하게 분석목적으로만 활용할 수 있는 데이터를 제공할 수 있도록 하여 제공범위를 확대할 수 있도록 하였다.

본 연구를 통하여 KCTC의 Live simulation 데이터웨어하우스에 대한 데이터의 일관성과 안전성을 강화하여 요구자에게 보다 원활하게 검색서비스를 제공하는 웹기반 OLAP서비스 구축이 가능함을 제시하며, 향후 유사분야에서 다량의 데이터에 대한 선별적인 웹기반 OLAP서비스에 확대 적용도 가능할 것이다.

KCTC의 분석자료에 대한 활용 활성화는 오랜기간 축적된 실기동 데이터를 연구목적에 맞는 형태로 가공하고 정제하여 손쉽게 활용할 수 있는 안전한 방식으로 제공하는 것이 중요하며, 이를 위한 서비스체계를 정교화하고 다양한 업무적용을 위한 커스터마이징이 추가로 필요하며, 사용자에게 데이터웨어하우스의 활용을 위한 자료사전을 제공할 수 있도록 보완해야 하는 점이 향후의 과제이며, Live Simulation훈련 데이터의 실무적인 활용 분야를 확대하고, 연대급으로 확장 구축하는 KCTC 훈련의 비정형분석서비스에 대하여 모델로 활용될 수 있기를 바란다.

【참고문헌】

1. 국내문헌

- 권대홍(2008), "마이크로데이터의 공개 위험과 유용성을 고려한 공개방안 연구",
충남대학교 대학원 석사학위논문
- 김종만 외 2인(2011), "M&S 분석자료 활용방안 연구(Live Simulation data 중
심으로)", 국방기술품질원
- 김형현(2009), 『국방 M&S 개론』, 서울, 국방기술품질원
- 박영권(2007), "마이크로데이터에 대한 통계적 비밀보호기법 비교연구",
서울시립대학교 대학원 석사학위논문
- 이재완(2008), "과학화전투훈련을 통해 본 합동성 강화", 『합참 제34호』,
서울, 합동참모본부
- 정상혁(2009), "오픈소스 ETL 솔루션 활용", 『월간 마이크로소프트웨어
2009년 10월호』, 서울, 씨넷코리아 아이앤씨
- 조재희(2003), "OLAP 테크놀로지의 이해", 『정보과학회지 제21권 제10호』,
서울, 누리미디어
- 최종근(2011), "데이터 마이그레이션을 위한 오픈소스 ETL도구 평가",
승실대학교 정보과학대학원 석사학위논문
- 최철민(2007), "데이터의 효율적인 분석을 위한 Data Warehouse기반 OLAP의
설계 및 구현", 한양대학교 공과대학원 석사학위논문
- 최홍필(2010), "M&S를 활용한 군 과학화 훈련체계 구축에 관한 연구", 한
성대학교 경영대학원 석사학위논문
- 육군과학화전투훈련단 체계운영부(2010), 『디스커버러 활용법』, 인제, 육군
과학화전투훈련단
- 쌍용정보통신(2001), 『과학화전투훈련단 중앙통제장비체계 운영개념기술서』,
서울, 육군교육사령부
- 방위사업청 획득기반과(2010), 『무기체계 획득단계별 M&S 적용지침』, 서울,
방위사업청

육군교육사령부(2011), 『과학화 전투훈련 지침서』, 서울, 육군본부
포시에스 EQ개발팀(2007), 『OZ EQ 관리자 매뉴얼』, 서울, (주)포시에스
_____ (2007), 『OZ EQ 사용자 매뉴얼』, 서울, (주)포시에스

2. 국외문헌

Carlos E. Perez(2012), Open Source ETL (Extraction, Transform, Load) Written in Java, <http://www.manageability.org/blog/stuff/open-source-etl>
Matt Casters, Roland Bouman, Jos van Dongen(2010), *Pentaho Kettle solutions : building open source ETL solutions with Pentaho Data Integration*, Indianapolis, Wiley
Matteo Golfarelli(2009), Open Source BI Platforms: A Functional and Architectural Comparison, *Lecture Notes in Computer Science Vol.- No.5691*, Berlin Heidelberg, SPRINGER-VERLAG
P R Brownbridge(2002), *Oracle9i Discoverer Desktop User's Guide Version 9.0.2 for Windows*, Redwood City, Oracle Corporation
Samuel S. Conn(2005), OLTP and OLAP Data Integration: A Review of Feasible Implementation Methods and Architectures for Real Time Data Analysis, *IEEE SOUTHEASTCON*, Ft. Lauderdale FL , IEEE

ABSTRACT

A Study of Using a Practical Application of Live Simulation
Training Datawarehouse
-Focused on the datawarehouse
of Korea Combat Training Center-

Lim, Hoy Sung

Major in National Defense Modeling&Simulation

Dept. of National Defense Modeling&Simulation

Graduate School of National Defense Science

Hansung University

In an environment of rapidly changing domestic and international defense and limited training conditions by shrinking our military defense budget and concern about safety accident, our military has attempted steadily. The innovation of training methods to cultivate the combatant and the army who can fight and win. so as to meet changing battlefield in the future.

According to the development of information and communication technology, our military has delivered the training innovation by using M&S being considered as an efficient method of defense management in the future.

The data conducted by KCTC since 2006 has been accumulated on the training analysis data warehouse after estimation of the result of training.

The cumulative effects for five years of training are valuable data described by the most similar to the South Korean military. We expect to continue to rise the data by expanding domestic KCTC into the future.

This study, these previous studies based on live simulation training such exercises through the accumulated data, the utilization of employed, seeking ways to the existing KCTC of data analysis utilized conditions for the Status and feedback analysis and Statistics the case of the National Statistical Office such as utilization of the data to enable benchmarking environment was studied.

Through this study, KCTC's training only long-term accumulated data analysis and military-related research and M&S system to build the basic data can serve as the creation of conditions to, and the existing is built on KCTC of training analysis data warehouse targets to increase the reliability of the data utilized were looking for ways to ease.

Through appropriate reprocessing of accumulated data analyst needs to facilitate the extraction of data rights management service environment possible considering building a secure and provide reliable data services for Web-based test system was implemented.

【keyword】 simulation, datawarehouse, ETL