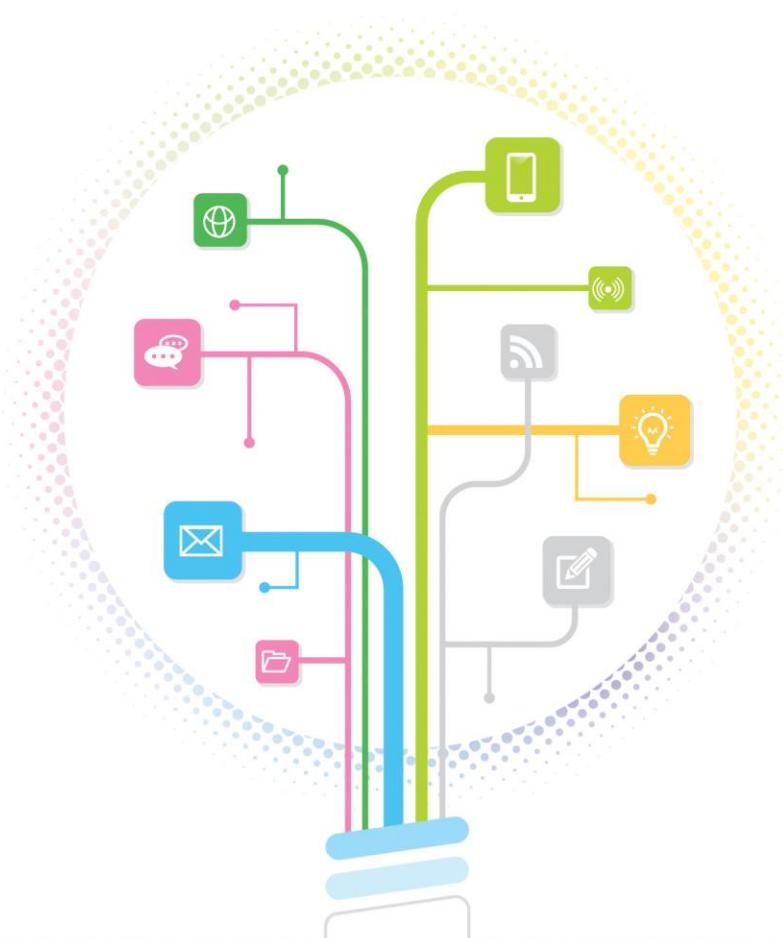


데이터 분석 콘텐츠 활용 매뉴얼

4 쇼핑



미래창조과학부



한국정보화진흥원



CONTENTS

Beginning Level 초급과정

I 개요

개요	9
----	---

II 수집

개요	13
교육용 데이터 샘플	14
데이터 수집	15
데이터 작업 영역 이동 스크립트	18

III 가공

개요	23
데이터 가공 과정	24

IV 저장

개요	29
가공 데이터 저장	30



V 분석

개요	33
데이터 분석 과정	34

VI 시각화

개요	39
시각화 과정	40
시각화 데이터 분석	42

VII 예제 문제

예제 문제1. 온라인 쇼핑몰의 연간 매출 변동을 분석하라.	45
예제 문제2. 분기별 최고 매출 아이템을 파악하라.	46

CONTENTS

Intermediate Level *중급과정*

I 개요

개요	51
----	----

II 수집

개요	55
교육용 데이터 샘플	56
데이터 수집	57
데이터 작업 영역 이동 스크립트	60

III 가공

개요	65
----	----

IV 저장

개요	69
가공 데이터 하둡 파일시스템 업로드	70
가공 데이터 하둡 파일시스템 저장	72



V 분석

개요	75
데이터 분석 스크립트	77
분석 데이터 파일 조회	84
결과 데이터 2차 분석	86

VI 시각화

개요	95
시각화 과정	96
분석 데이터 시각화	97
데이터 분석	98

VII 예제 문제

예제 문제1. 3년치 거래 단위별 판매 내역 데이터를 읽어들여 연도별로 월별 매출 추이를 지역별 비교 분석하라.	101
예제 문제2. 3년치 거래 단위별 판매 내역 데이터를 읽어들여 연도별로 최고 매출 아이템 Best5를 추출하라.	102



쇼핑 

Beginning Level

초급과정







I 개요

개요

9

8

I

개요

> 개요

국내 남성 의류 온라인 쇼핑몰 2010~2012년 남성 쇼핑몰 판매 거래 데이터를 바탕으로, 2010년 월별 거래 판매량의 변화를 시계열 분석의 패턴분석을 통해서 해본다. 오픈오피스를 활용하여 월별 판매 가격, 판매 여부에 대한 데이터를 추출 가공하고, 피벗테이블 기능을 사용하여 2010년 월별 매출 추이를 시각화(막대그래프) 하여 분석을 통해 1년간 쇼핑몰의 매출 패턴이 어떻게 나타나는지를 확인하고, 성수기와 비수기를 구분하여 마케팅에 활용하고자 한다.

> 활용 데이터

- **mall_transaction.xls :**

2010~2012년 남성 쇼핑몰 거래 데이터(2010년만 사용)

> 선행학습

- **리눅스** – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- **오픈오피스** – 피벗테이블 기능, 차트 사용 방법

> 요구사항

- 남성 쇼핑몰의 2010년 판매 내역을 기반으로 월별 매출 통계를 산출하고, 월별 쇼핑몰의 매출 변화를 시각화하여 계절에 따른 매출 추이를 분석하라.

> 분석 절차

- 2010~2012년 남성 쇼핑몰 거래 데이터를 오픈오피스를 사용하여 로드한다.
- 2010년 월 별 거래 판매량의 변화를 시계열 분석의 패턴분석에 용이한 형태로 가공하기 위해 분석에 필요 없는 항목들을 삭제한다.
- 가공되어진 2010년 남성 쇼핑몰 거래 데이터를 피벗테이블 기능을 활용하여 거래단위별 판매 데이터로부터 월별 매출 합계를 계산한다.
- 취합한 데이터를 오픈오피스 차트의 막대그래프를 활용하여 시각화해 본다.
- 시각화 되어진 그래프를 보고 2010년 1년간 쇼핑몰의 매출 패턴이 어떻게 나타나는지를 확인하고 성수기와 비수기가 명확하게 구분되는지 확인한다.



• 오픈오피스(OpenOffice)

- 마이크로소프트 오피스와 같은 오피스 스위트입니다.
- 기존 오피스 프로그램과의 뛰어난 호환성을 자랑합니다.
- 제품 개발의 전 과정이 투명한 공개 소프트웨어 프로젝트입니다.
- 라이센스 비용을 지불할 필요가 없는 무료 소프트웨어입니다.
- 윈도우 뿐만 아니라 리눅스와 솔라리스 등 다양한 운영체제를 지원합니다.



II 수집

개요	13
교육용 데이터 샘플	14
데이터 수집	15
데이터 작업 영역 이동 스크립트	18



수집

> 개요

쇼핑 데이터는 국내 남성 의류 온라인 쇼핑몰 2010~2012년 남성 쇼핑몰 판매 거래 데이터를 수집하여 분석 목적을 달성할 수 있는 한도 내에서 개인 정보 비식별화 및 특정 상품 비식별화 처리를 통해 분석에 용이하게 편집하여 제공한다.

> 수집 방법

- **데이터 제공** : 쇼핑 정보 데이터는 국내 남성 의류 온라인 쇼핑몰에서 제공하는 데이터를 OpenAPI, 자료수집기(Crawler)를 통하여 데이터를 수집하였고, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 쇼핑몰 정보 데이터 셋을 다운로드할 수 있도록 원시데이터를 제공하고 있다.



용 어 정 리

- **비식별화** : 데이터 값 삭제, 가명처리, 총계처리, 범주화, 데이터 마스킹 등을 통해 개인정보의 일부 또는 전부를 삭제하거나 대체함으로써 다른 정보와 쉽게 결합하여도 특정 개인을 식별할 수 없도록 하는 조치를 말한다.

- *출처: 방송통신위원회, “빅데이터 개인정보보호 가이드라인”, 작성일 2014.12.23

▶ 교육용 데이터 샘플

▶ 온라인 쇼핑몰 거래내역 데이터 샘플(mall_transaction.xls)

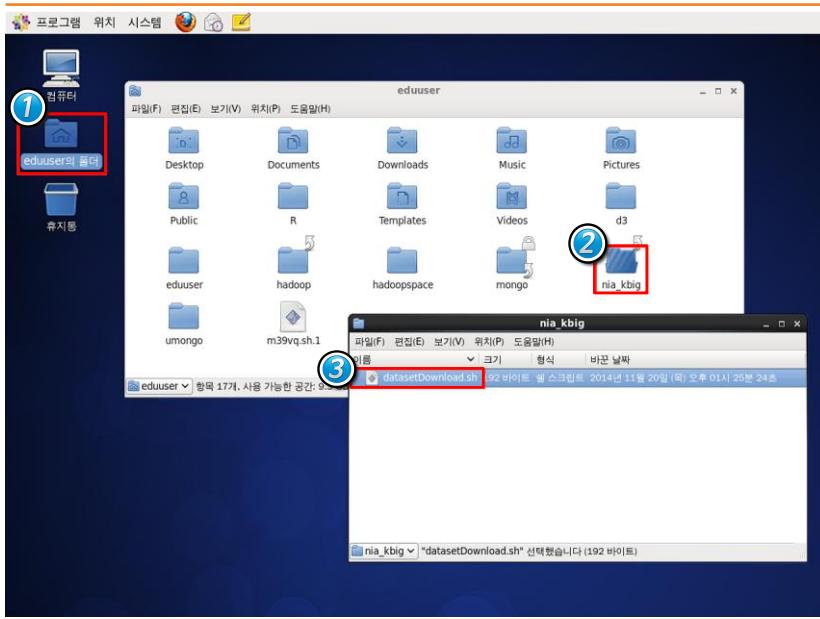
주문일자	거래상태	상품명	판매가	수량합계	합계금액
10.12.30 13:11	판매완료	ADW 짚업…	57200	4	194000
10.12.30 13:33	판매완료	스판 풀바지…	21400	1	23900
10.12.30 13:50	판매완료	무지 면 츄…	29200	1	30200
10.12.30 13:51	판매완료	스트라이프 …	15500	6	122300
10.12.30 14:07	반송(반품)	중청 스티치 …	39400	1	41900
10.12.30 13:11	판매완료	ADW 짚업 기모 후드 Tia336 – 카키(1)	57200	4	194000
10.12.30 13:33	판매완료	스판 풀바지 5310– 블랙(1)	21400	1	23900
10.12.30 13:50	판매완료	무지 면 츄리닝 바지 (추동용) 9102– 다크그레이(1)	29200	1	30200

II. 수집

> 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 쇼핑 데이터 셋을 복사해 온다.
 - `mall_transaction.xls` : 온라인 쇼핑몰 거래내역 데이터

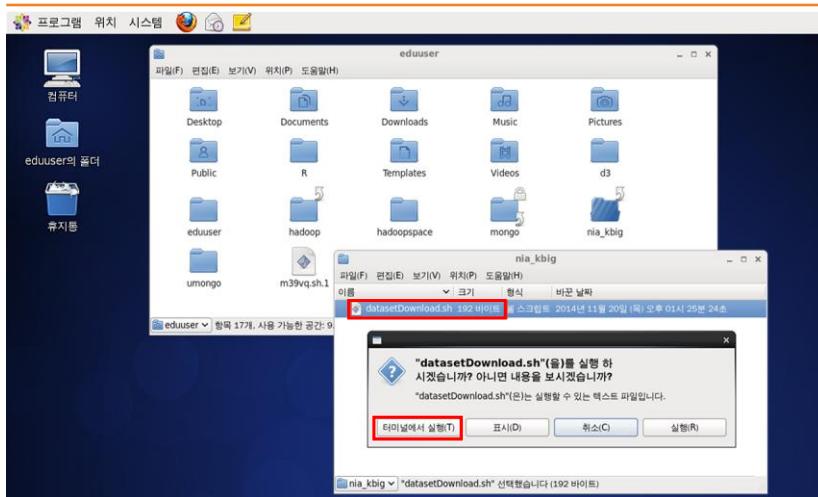
> 실습코드 디렉토리로 이동



- ① 로그인 후 바탕화면에서 `eduuser` 폴더를 오픈한다.
- ② `nia_kbig` 폴더를 오픈한다.
- ③ `datasetDownload.sh`를 더블클릭하여 실행한다.

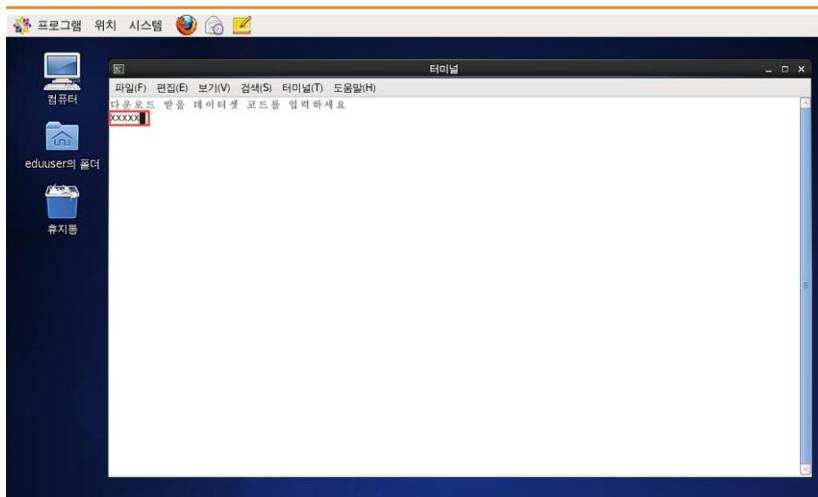
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



- '터미널에서 실행' 버튼을 클릭한다.

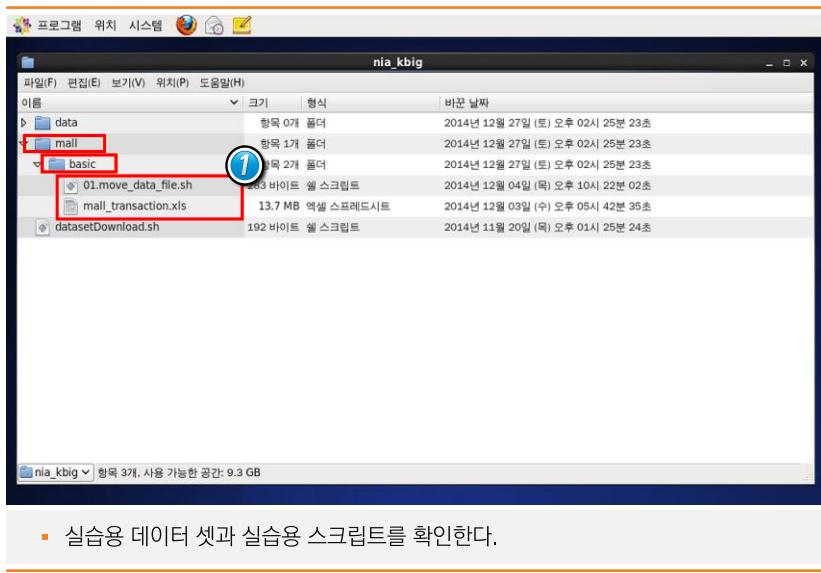
▶ 데이터셋 코드 입력



- 다운로드 받을 데이터셋 코드를 입력 후 엔터

II. 수집

▶ 데이터셋과 실습용 쉘 스크립트



- 실습용 데이터셋과 실습용 스크립트를 확인한다.

▶ ① 데이터 및 스크립트

▪ 01.move_data_file.sh :

로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

▪ mall_transaction.xls : 2010년 남성 쇼핑몰 거래 데이터

> 데이터 작업 영역 이동 스크립트(01.move_data_file.sh)

> 데이터 작업 공간으로 이동

- 로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

01.move_data_file.sh (작업영역 폴더로 원시데이터 이동)

```

01.#!/bin/bash
02. # 복사 대상 파일 정의
03. TARGET_SHOPPING_PRICE=/home/eduuser/nia_kbig/shopping/basic/mal
   ↛ l_transaction.xls
04. # 작업 디렉토리 정의
05. LOCAL_DIR=/home/eduuser/nia_kbig/data/
06. mv $TARGET_SHOPPING_PRICE $LOCAL_DIR
07. mv $TARGET_SHOPPING_PRICE $LOCAL_DIR
08.

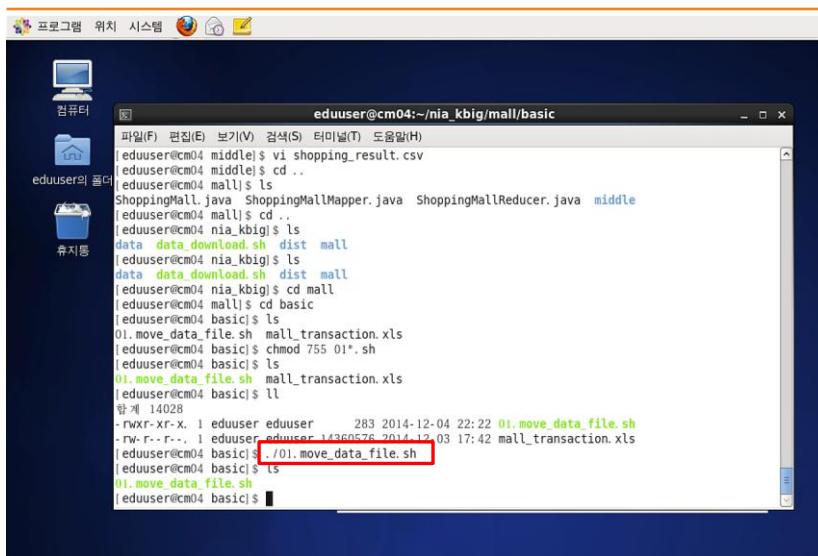
```



- 데이터 작업 영역 이동 스크립트 소스(01.move_data_file.sh)
- 라인 03 : 다운로드 받은 원시데이터 파일의 위치(path)를 변수(TARGET_SHOPPING_PRICE)로 지정하는 라인이다.
- 라인 05 : 작업영역 디렉토리의 위치(path)를 변수(LOCAL_DIR)로 지정하는 라인이다.
- 라인 06~07 : mv 명령어를 사용하여 다운로드 받은 원시데이터 파일들을 작업영역 디렉토리로 이동시키는 라인이다.

II. 수집

▶ 수집 데이터 셋 작업 영역 폴더 이동



```
eduuser@cm04:~/nia_kbig/mall/basic
[파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
eduuser@cm04 middle]$ vi shopping_result.csv
eduuser@cm04 mall]$ ls ..
eduuser@cm04 nia_kbig]$ ls
ShoppingMall.java ShoppingMallMapper.java ShoppingMallReducer.java middle
eduuser@cm04 mall]$ cd ..
eduuser@cm04 nia_kbig]$ ls
data_data_download.sh dist mall
eduuser@cm04 nia_kbig]$ cd mall
eduuser@cm04 mall]$ cd basic
eduuser@cm04 basic]$ ls
01.move_data_file.sh mall_transaction.xls
eduuser@cm04 basic]$ chmod 755 01*.sh
eduuser@cm04 basic]$ ls
01.move_data_file.sh mall_transaction.xls
eduuser@cm04 basic]$ ll
합계 14028
-rwxr-Xr-X 1 eduuser eduuser 283 2014-12-04 22:22 01.move_data_file.sh
-rw-r--r-- 1 eduuser eduuser 1436576 2014-12-03 17:42 mall_transaction.xls
[redacted] ./01.move_data_file.sh
[redacted] 01.move_data_file.sh
[redacted] ./01.move_data_file.sh
[redacted] [redacted]
```

- 로컬에 원시데이터를 작업 영역 폴더로 이동 (/home/eduuser/nia_kbig/data/) 시킨다.
- ./01.move_data_file.sh 입력 후 엔터

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화







가공

> 개요

작업 영역 폴더에 복사한 쇼핑 데이터의 가공은, 전처리 단계에서 수집된 2010~2012년 남성 의류 쇼핑몰 거래 내역 데이터를 로드하여, 소비 패턴분석에 필요한 2010년 데이터만 추출하고, 연월별 패턴분석을 위해 주문 일자 컬럼으로부터 판매월 컬럼을 생성한다.

> 가공 방법

- 쇼핑몰 판매 데이터는 2010~2012년간의 3년치 데이터가 저장되어 있다.
- 2010년 시트만 활용하며, 분석에 필요 없는 컬럼들을 삭제한다.
- 각각의 연도별 판매 데이터를 월별로 분석할 수 있도록 판매일자로부터 판매월 컬럼(열)을 생성한다.

> 데이터셋

주문일자	거래상태	상품명	판매가	수량합계	배송료	합계금액
10.12.30 13:11	판매완료	ADW 짚업…	57200	4	0	194000
10.12.30 13:33	판매완료	스판 썰바지…	21400	1	2500	23900
10.12.30 13:50	판매완료	무지 면 츄…	29200	1	2500	30200
10.12.30 13:51	판매완료	스트라이프 …	15500	6	0	122300
10.12.30 14:07	반송(반품)	중청 스티치 …	39400	1	2500	41900
10.12.30 13:11	판매완료	ADW 짚업 기모 후드티a336 - 카키(1)	57200	4	0	194000
10.12.30 13:33	판매완료	스판 썰바지 5310-블랙(1)	21400	1	2500	23900
10.12.30 13:50	판매완료	무지 면 츄리닝 바지(추동용) 9102- 다크그레이(1)	29200	1	2500	30200

> 데이터 가공 과정

> 데이터 로드

- 오픈오피스를 사용하여 연도별로 저장된 원시데이터를 읽어 들여 분석을 위한 기초 가공을 한다.

shoppingmall_data.xls - OpenOffice Calc

A	B	C	D
주문일자	거래상태	수신자우편번호	상품명(수량)
2 2010-12-31 23:41	판매확정	570-080	정시이즈 크리스마스 셀프워싱 퍼치기모 스웨터a333 - 블랙 (1)
3 2010-12-31 23:38	판매확정	712-730	BWIN 면 후드 트레이닝세트 (추정)a309 - 블랙 (1)
4 2010-12-31 20:17	판매확정	220-963	사마리비 반팔라인팔리a369 - 블랙 (1)
5 2010-12-31 18:28	판매확정	441-704	프리미엄 쿠션사 코듀로이 스웻바지 7301-풀색 (1)
6 2010-12-31 18:09	판매확정	448-549	에센셜로 기능성 밴드 카고바지 (추정) 9301-네이비 (1)
7			에센셜로 기능성 밴드 카고바지 (추정) 9301-블루 (1)
8 2010-12-31 17:47	주문취소	210-110	기모 스웨트바지a331 - 블랙 (1)
9 2010-12-31 17:21	판매확정	437-810	주머니백색 대워싱 기모 스웨트바지a327 - 친환경 (1)
10 2010-12-31 17:16	판매확정	437-810	온실업 후드 페달걸과a337 - 화이트 (1)
11 2010-12-31 16:58	판매확정	415-728	방지의 티셔츠풀라인팔리a301 - 블랙 (1)
12 2010-12-31 16:30	판매확정	402-060	포화풀리스내과 풍천기록 조끼a312 - 블랙 (1)
13			청자백색 자동 정장 뱃트9303 - 블랙-보라 3 (1)
14 2010-12-31 16:16	주문취소	462-759	주얼리 뱃드바지a302 - 미디블루 (1)

- 1. 데이터 로드 – 작업 폴더에서 mall_transaction.xls 더블클릭하여 오픈오피스 스프레드시트를 실행한다.

III. 가공

▶ 데이터 축소

A	B	C	D	E	F	G	H	I	J	K
주문일자	구매상태	주문자주소	설 서식 지정...	신분상을 주당	판매가	기본수령				
2010-12-31 23:41	판매 완료	578-088	일 내비밀...	롯데화식 쇼핑기본 스판 칠비	1	65800				
2010-12-31 23:38	판매 완료	712-730	최적 일 내비밀...	롯데화식 쇼핑기본 스판 칠비	1	112700				
2010-12-31 20:17	판매 완료	220-963	일 상업비	롯데화식 쇼핑기본 스판 칠비	1	36700				
2010-12-31 18:28	판매 완료	441-704	내용 지정...	롯데화식 쇼핑기본 스판 칠비	1	58600				
2010-12-31 18:09	판매 완료	448-549	술기기비	롯데화식 쇼핑기본 스판 칠비	1	43400				
2010-12-31 17:47	주류 완료	210-110	표시(S)	롯데화식 쇼핑기본 스판 칠비	1	43400				
2010-12-31 17:21	판매 완료	437-810	주어니 배색 폴리포장 기호	롯데화식 쇼핑기본 스판 칠비	1	41800				
2010-12-31 17:18	판매 완료	437-810	주어니 배색 폴리포장 기호	롯데화식 쇼핑기본 스판 칠비	1	53800				
2010-12-31 16:58	판매 완료	415-728	롯데화식 쇼핑기본 스판 칠비	롯데화식 쇼핑기본 스판 칠비	1	79800				
2010-12-31 16:30	판매 완료	402-060	롯데화식 쇼핑기본 스판 칠비	롯데화식 쇼핑기본 스판 칠비	1	47600				
2010-12-31 16:16	주류 완료	462-759	최적 배송 자동 정지 버튼	롯데화식 쇼핑기본 스판 칠비	1	37400				
2010-12-31 16:04	판매 완료	240-028	롯데화식 쇼핑기본 스판 칠비	롯데화식 쇼핑기본 스판 칠비	1	21400				
2010-12-31 15:57	판매 완료	520-752	주어니 배드 바지 0305 - 미리고정(1)	롯데화식 쇼핑기본 스판 칠비	1	39800				
			주어니 배드 바지 0305 - 미리고정(1)	롯데화식 쇼핑기본 스판 칠비	1	89800				
			주어니 배드 바지 0305 - 미리고정(1)	롯데화식 쇼핑기본 스판 칠비	1	19000				
			주어니 배드 바지 0305 - 미리고정(1)	롯데화식 쇼핑기본 스판 칠비	1	19000				

- 2. 데이터 축소 - 분석에 사용할 항목(주문 일자, 거래상태, 합계금액)을 제외한 모든 컬럼을 삭제한다.

▶ 신규컬럼 추가

A	B	C	D	E	F	G	H	I	J	K
주문일자	구매상태	설 서식 지정...								
2010-12-31 23:41	판매 완료	일 내비밀...								
2010-12-31 23:38	판매 완료	최적 일 내비밀...								
2010-12-31 20:17	판매 완료	최적 일 내비밀...								
2010-12-31 18:28	판매 완료	일 시장비								
2010-12-31 18:09	판매 완료	내용 지정(E...)								
2010-12-31 17:47	주류 완료	술기기비								
2010-12-31 17:21	판매 완료	표시(S)								
2010-12-31 17:16	판매 완료	최적 배송 자동 정지 버튼								
2010-12-31 16:58	판매 완료	복사(C)								
2010-12-31 16:30	판매 완료	붙여넣기(V)								
2010-12-31 16:16	주류 완료	42300								
2010-12-31 16:04	판매 완료	89800								
2010-12-31 15:57	판매 완료	54100								
2010-12-31 15:56	판매 완료	55600								
2010-12-31 15:28	판매 완료	46300								
2010-12-31 15:13	판매 완료	44000								
2010-12-31 14:15	판매 완료	287900								
2010-12-31 13:59	판매 완료	119800								
2010-12-31 13:49	판매 완료	62600								

- 3. 신규컬럼 추가 - 주문 일자 우측에 신규 컬럼을 생성하고, 컬럼명을 '판매월'로 입력한다.

> 셀 수식 계산

	A	B	C	D	E	F	G	H	I	J
1	주문일자	=주문일자	내수 대체	당일 구매						
2	2018-12-31 23:41	=TEXT(A2;"yyyymm")	판매 완료		111365					
3	2018-12-31 23:38		판매 완료		37400					
4	2018-12-31 20:17		판매 완료		55700					
5	2018-12-31 18:28		판매 완료							
6	2018-12-31 18:09		판매 완료	82500						
7	2018-12-31 17:47		주문 완료	42200						
8	2018-12-31 17:23		판매 완료	53800						
9	2018-12-31 17:16		판매 완료	79800						
10	2018-12-31 16:58		판매 완료	58300						
11	2018-12-31 16:30		판매 완료	55900						
12	2018-12-31 16:16		판매 완료	42300						
13	2018-12-31 16:04		판매 완료	89800						
14	2018-12-31 15:57		판매 완료	54100						
15	2018-12-31 15:56		판매 완료	55600						
16	2018-12-31 15:28		판매 완료	46300						
17	2018-12-31 15:13		판매 완료	44900						
18	2018-12-31 14:15		판매 완료	207900						
19	2018-12-31 13:59		판매 완료	119800						
20	2018-12-31 13:40		판매 완료	62600						
21										
22										
23										
24										
25										
26										
27										

- 4. 셀 수식 계산 – 새로 만든 컬럼의 첫번째 행에 '=TEXT(A2;"yyyymm")' 을 입력한다.

> 수식 복사

	A	B	C	D	E	F	G	H	I	J
1	주문일자	=주문일자	내수 대체	당일 구매						
2	2018-12-31 23:41	=TEXT(A2;"201812")	판매 완료	62500						
3	2018-12-31 23:38	=TEXT(A2;"201812")	판매 완료	111365						
4	2018-12-31 20:17	=TEXT(A2;"201812")	판매 완료	37400						
5	2018-12-31 18:28	=TEXT(A2;"201812")	판매 완료	55700						
6	2018-12-31 18:09	=TEXT(A2;"201812")	판매 완료	82500						
7	2018-12-31 17:47	=TEXT(A2;"201812")	주문 완료	42200						
8	2018-12-31 17:23	=TEXT(A2;"201812")	판매 완료	53800						
9	2018-12-31 17:16	=TEXT(A2;"201812")	판매 완료	79800						
10	2018-12-31 16:58	=TEXT(A2;"201812")	판매 완료	58300						
11	2018-12-31 16:30	=TEXT(A2;"201812")	판매 완료	55900						
12	2018-12-31 16:15	=TEXT(A2;"201812")	판매 완료	42300						
13	2018-12-31 16:04	=TEXT(A2;"201812")	판매 완료	89800						
14	2018-12-31 15:57	=TEXT(A2;"201812")	판매 완료	54100						
15		=TEXT(A2;"201812")								
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										

- 5. 데이터 복사 – 입력한 셀을 복사하여 컬럼 전체에 붙여넣기 한다.



IV 저 장

개요	29
가공 데이터 저장	30

IV

저장

> 개요

오픈오피스 스프레드시트를 활용하여 데이터 로드 > 가공 > 분석 > 시각화 단계를 한번에 실행할 수 있으나, 기초 데이터를 유지하기 위하여 가공 단계에서 가공한 데이터를 별도로 저장한다.



> 저장 방법

- 오픈오피스 스프레드시트 툴을 활용하여 1차 가공한 데이터를 유지하고 이후 패턴 분석 과정을 진행하기 위해 별도 파일로 저장한다.

> 가공 데이터 저장

> 오픈오피스 스프레드 시트 데이터 저장

The screenshot shows a Microsoft Windows desktop environment. In the center, there is an OpenOffice Calc window displaying a spreadsheet titled "mall_transaction.xls". The spreadsheet contains data from row 1 to 28, with columns A through M. Row 1 has headers: 주문일자, 주문번호, 거래상태, 할인증액. Rows 2 through 28 contain transaction data. A "Save as" dialog box is overlaid on the calc window. The "File name:" field is set to "mall_basic_analysis.xls", and the "File type:" dropdown is set to "Microsoft Excel 97/2000/XP (.xls)". There are several other options in the dialog, such as "Save with password", "Automatic file name extension", and "Edit filter settings".

- 오픈오피스 스프레드 시트에서 메뉴 / 파일 / 다른 이름으로 저장을 선택하여 별도 파일을 저장한다.
- 저장 위치 및 파일명은 "/home/eduuser/nia_kbig/mall_basic_analysis.xls"로 저장한다.

W





V 분석

개요

33

데이터 분석 과정

34

V 분석

> 개요

쇼핑 데이터 분석은 오픈오피스 스프레드 시트의 피봇 테이블 기능과 차트 기능을 활용한다. 가공 단계에서 가공한 2010년 남성 쇼핑몰의 거래단위별 판매 내역 데이터를 읽어 들여, 시계열 분석의 패턴 분석을 하기 위하여 피봇테이블 기능을 활용해서 2010년도 월별 매출 통계를 계산하여 매출 추이를 분석한다.

> 분석 방법

- 연도별로 1차 가공, 저장한 파일을 읽어 들인다.
- 패턴 분석을 하기 위하여 스프레드 시트의 피봇테이블 (Pivot Table) 기능을 활용, 연월(YYYYMM)별 매출(판매금액합계)을 산출한다.
- 거래 단위별 판매 내역 데이터를 읽어들여 연도별로 월별 매출 추이를 분석한다.

> 데이터 분석 과정

> 가공데이터 로드

	A	B	C	D	E	F	G	H	I	J
1	주문일자									
2	2010-12-31 23:41:201012	마이크로	62500							
3	2010-12-31 23:38:201012	마이크로	111365							
4	2010-12-31 20:17:201012	마이크로	37400							
5	2010-12-31 18:28:201012	마이크로	55700							
6	2010-12-31 18:09:201012	마이크로	82500							
7	2010-12-31 17:47:201012	주류회사	42200							
8	2010-12-31 17:23:201012	마이크로	52800							
9	2010-12-31 17:16:201012	마이크로	79800							
10	2010-12-31 16:58:201012	마이크로	50300							
11	2010-12-31 16:40:201012	마이크로	55900							
12	2010-12-31 16:30:201012	마이크로	55900							
13	2010-12-31 16:16:201012	주류회사	42300							
14	2010-12-31 16:04:201012	마이크로	89800							
15	2010-12-31 15:57:201012	마이크로	54100							
16	2010-12-31 15:57:201012	마이크로	54100							
17	2010-12-31 15:57:201012	마이크로	54100							
18	2010-12-31 15:56:201012	마이크로	55600							
19	2010-12-31 15:56:201012	마이크로	55600							
20	2010-12-31 15:56:201012	마이크로	55600							
21	2010-12-31 15:28:201012	마이크로	46300							
22	2010-12-31 15:13:201012	마이크로	44000							
23	2010-12-31 14:15:201012	마이크로	207000							
24	2010-12-31 13:59:201012	마이크로	119800							
25	2010-12-31 13:59:201012	마이크로	119800							
26	2010-12-31 13:59:201012	마이크로	119800							

- 1차 가공, 저장한 데이터 파일((mall_transaction_analysis.xls)을 오픈오피스에서 읽어들인다.

V. 분석

▶ 피봇 테이블을 활용한 매출 통계 분석

- 오픈오피스의 피봇테이블 기능을 활용하여 연월(YYYYMM)별 매출(판매금액 합계)을 산출한다.

The screenshot shows the OpenOffice Calc interface with a spreadsheet titled "mail_transaction.xls". A pivot table is being created in cells A1:D38601. The pivot table has "주체 일자" (Subject Date) in row 1, "주체 번호" (Subject Number) in row 2, and "내부 상세" (Internal Details) in row 3. The "내부 상세" row contains "판매 금액" (Sales Amount). The formula bar at the top shows the formula =SUMIF(\$A\$2:\$A\$38601, \$D1, \$C\$2:\$C\$38601). The "Data" tab of the ribbon is selected. A context menu is open over the pivot table area, with "피벗 테이블" (Pivot Table) highlighted.

- 1. 메뉴/데이터/피봇테이블/만들기 를 선택하여 피봇테이블 마법사를 실행한다.

The screenshot shows the OpenOffice Calc interface with the "mail_transaction.xls" spreadsheet. A context menu is open over the pivot table area, with "피벗 테이블" (Pivot Table) highlighted. A "원본 선택" (Source Selection) dialog box is displayed, showing the range A1:D38601. It includes options for "현재 선택" (Current Selection), "OpenOffice 드롭인 데이터 원본" (OpenOffice Drop-in Data Source), and "외부 원본 (인터넷址)" (External Source (Internet Address)). The "확인" (Confirm) button is highlighted.

- 2. 자동으로 범위가 선택되어져 있으므로 '현재선택' 상태에서 '확인'버튼을 클릭한다.

■ 3. 우측 '필드' 영역에서 아이템들(매장, 지역, 구분, 공급금액)을 각각 드래그하여 위 화면과 같이 배치한다.

■ 4. 데이터의 최하단까지 스크롤 하면 피벗테이블의 결과값이 출력된 것을 확인할 수 있다.



1

2



VI 시각화

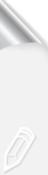
개요	39
시각화 과정	40
시각화 데이터 분석	42

VI

시각화

> 개요

쇼핑 데이터의 시각화 과정에서는, 패턴 분석을 위한 오픈오피스 피벗 테이블 기능으로 작성된 통계 데이터를 오픈 오피스 스프레드 시티의 차트 기능을 활용하여 막대그래프로 시각화한다.

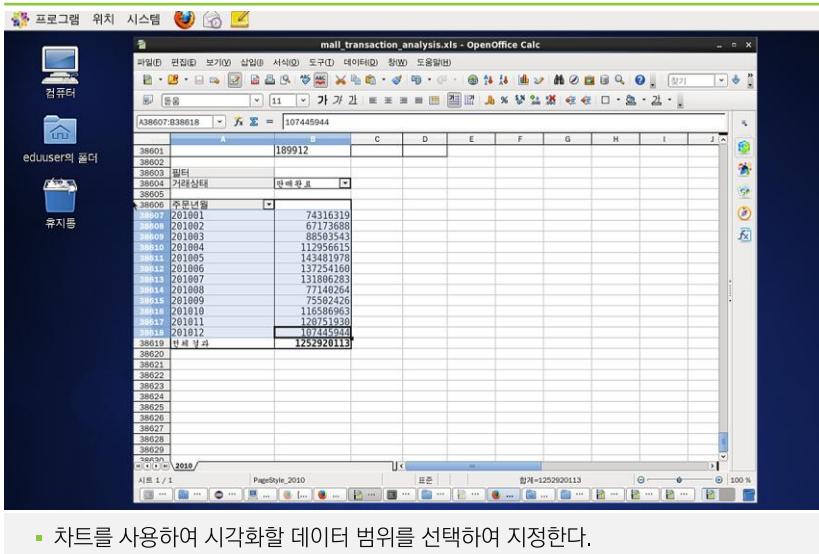


> 시각화 방법 및 활용기술

- 오픈 오피스 스프레드 시트의 차트 기능을 활용하여 멀티 컬럼 데이터를 하나의 차트로 표현한다.
- 2010년 월별 매출 패턴의 변화를 확인하기 위하여 오픈오피스 스프레드시트의 막대 그래프를 활용하여 시각화한다.

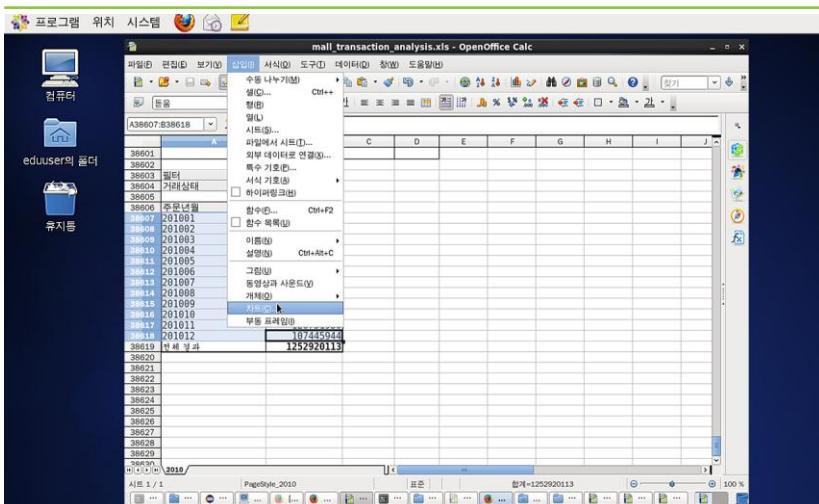
> 시각화 과정

> 데이터 범위 설정



- 차트를 사용하여 시각화할 데이터 범위를 선택하여 지정한다.

> 차트 생성



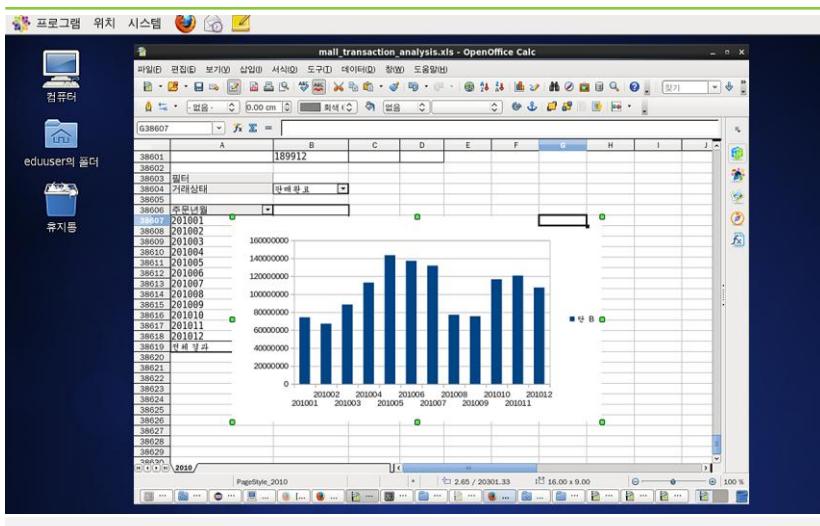
- 메뉴/삽입/차트(C)를 선택하여 차트 마법사를 실행한다.

VI. 시각화

▶ 차트 설정

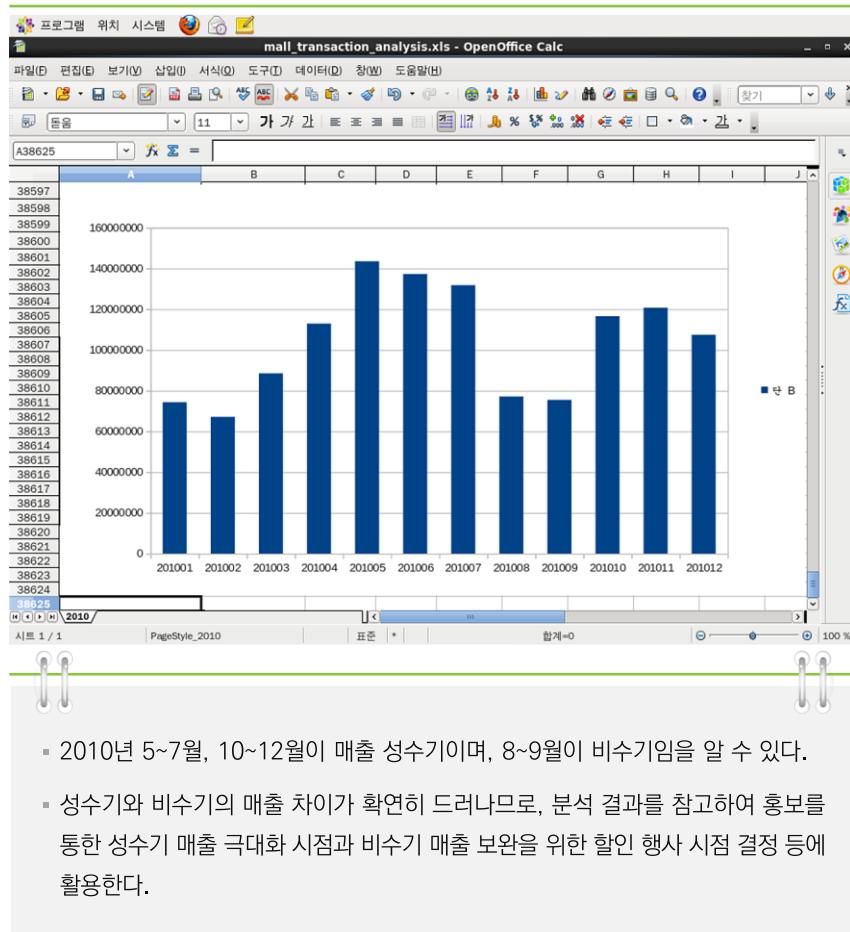
■ 차트 유형에서 막대그래프를 선택하고 마침을 누른다.

▶ 시각화 차트 확인



▶ 시각화 데이터 분석

▶ 결과 분석





VII 예제문제

예제 문제1

45

예제 문제2

46

예 / 제 / 문 / 제

예제 1

온라인 쇼핑몰의 연간 매출 변동을 분석하라.

- 2010년 월간 쇼핑몰 매출 추이를 분석하고, 막대그래프를 활용하여 시각화하라.

- 원시 데이터를 가공하여 필요한 필드만 추출한다.
- 월간 매출 추이 구한다.

예제 2

분기별 최고 매출 아이템을 파악하라.

- 분기별 최고 매출 아이템 및 해당 아이템의 매출액을 파악하고, 최고 매출 아이템이 전체 매출액에서 차지하는 비율을 분석하라.

- 쇼핑몰 매출 원시데이터로부터 분기별, 아이템별 매출액 합계를 구한다.
- 분기별로 매출액 내림차순으로 데이터를 정렬한다.
- 분기별 최고 매출 아이템과 전체 매출액 대비 비중을 계산한다.



쇼핑 

Intermediate Level

중급과정







I 개요

개요

51

I

개요

> 개요

거래 단위별로 주어진 2010년~2012년 3년간의 거래단위별 남성 쇼핑몰 판매 내역 데이터에 대하여 여러 개의 파일로 분산되어 제공된 데이터를 하둡과 자바로 한꺼번에 읽어들여 맵리듀스를 통해 연도별로 월간 거래 금액 통계를 산출하고, 오픈오피스의 꺾은선 그래프를 활용하여 하나의 차트에 연도별로 시각화하는 과정을 통해, 대용량 분산 데이터에 대한 통계 산출 방법과 시각화 방법을 학습한다. 2010년~2012년 3년간의 거래단위별 남성 쇼핑몰 판매 내역 데이터의 패턴 분석을 통하여 성수기와 비수기를 구분하여 마케팅에 활용하고자 한다.

> 활용 데이터

- **mall_transaction_2010.csv** : 2010년 남성 쇼핑몰 거래 데이터
- **mall_transaction_201101.csv** : 2011년 상반기 남성 쇼핑몰 거래 데이터
- **mall_transaction_201102.csv** : 2011년 하반기 남성 쇼핑몰 거래 데이터
- **mall_transaction_2012.csv** : 2012년 남성 쇼핑몰 거래 데이터

> 선행학습

- **리눅스** – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- **하둡 에코시스템** – 하둡 시작, 종료, 하둡 파일 시스템 명령어, 맵리듀스 실행 방법
- **자바** – 자바코딩, 자바컴파일, JDK 설치, jar 파일 만드는 방법
- **오픈오피스** – 피벗테이블 기능, 차트 사용 방법

> 요구사항

- 2010년~2012년 남성 쇼핑몰 판매 내역을 바탕으로 연도별, 월별로 매출 금액을 계산하여 3년간의 월간 매출 금액의 추이를 비교 분석한다.

> 분석 절차

- 연도별, 반기별로 제공된 판매 데이터를 하둡 파일 시스템으로 로드한다.
- 자바와 하둡으로 거래단위별로 맵리듀싱을 실행하여 연도별, 월별 판매금액을 계산한다.
- 계산 결과를 막대그래프를 활용하여 시각화해 본다.
- 시각화 한 데이터를 보고, 연도별 월별 판매금액의 추이를 분석한다.
- 연도별로 판매금액의 규모가 성장했는지, 계절별 판매금액 추이가 연도별로 일정하게 나타나는지 등을 분석하여, 계절별 마케팅 시점 결정 등에 기초 데이터로 활용한다.



1

2

II 수집

개요	55
교육용 데이터 샘플	56
데이터 수집	57
데이터 작업 영역 이동 스크립트	60



수집

▶ 개요

쇼핑 데이터는 국내 남성 의류 온라인 쇼핑몰 2010~2012년 3년간의 거래 내역 데이터를 수집하여 분석 목적을 달성할 수 있는 한도 내에서 개인정보 비식별화 및 특정 상품 비식별화 처리를 통해 분석에 용이하게 편집하여 제공한다.

▶ 수집 방법

- **데이터 제공** : 쇼핑 정보 데이터는 국내 남성 의류 온라인 쇼핑몰에서 제공하는 데이터를 OpenAPI, 자료수집기(Crawler)를 통하여 데이터를 수집하였고, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 쇼핑몰 정보 데이터 셋을 다운로드할 수 있도록 원시데이터를 제공하고 있다.



용 어 정 리

- **비식별화** : 데이터 값 삭제, 가명처리, 총계처리, 범주화, 데이터 마스킹 등을 통해 개인정보의 일부 또는 전부를 삭제하거나 대체함으로써 다른 정보와 쉽게 결합하여도 특정 개인을 식별할 수 없도록 하는 조치를 말한다.

- *출처: 방송통신위원회, “빅데이터 개인정보보호 가이드라인”, 작성일 2014.12.23

▶ 교육용 데이터 샘플

▶ 온라인 쇼핑몰 거래내역 데이터 샘플(mall_transaction_2010.csv)

주문일자	거래상태	상품명	판매가	수량합계	합계금액
10.12.30 13:11	판매완료	ADW 짚업…	57200	4	194000
10.12.30 13:33	판매완료	스판 쫄바지…	21400	1	23900
10.12.30 13:50	판매완료	무지 면 쥬…	29200	1	30200
10.12.30 13:51	판매완료	스트라이프 …	15500	6	122300
10.12.30 14:07	반송(반품)	중청 스티치 …	39400	1	41900
10.12.30 13:11	판매완료	ADW 짚업 기모 후드티 a336 – 카키(1)	57200	4	194000
10.12.30 13:33	판매완료	스판 쫄바지 5310–블랙(1)	21400	1	23900
10.12.30 13:50	판매완료	무지 면 쥬리닝 바지(추동용) 9102– 다크그레이(1)	29200	1	30200

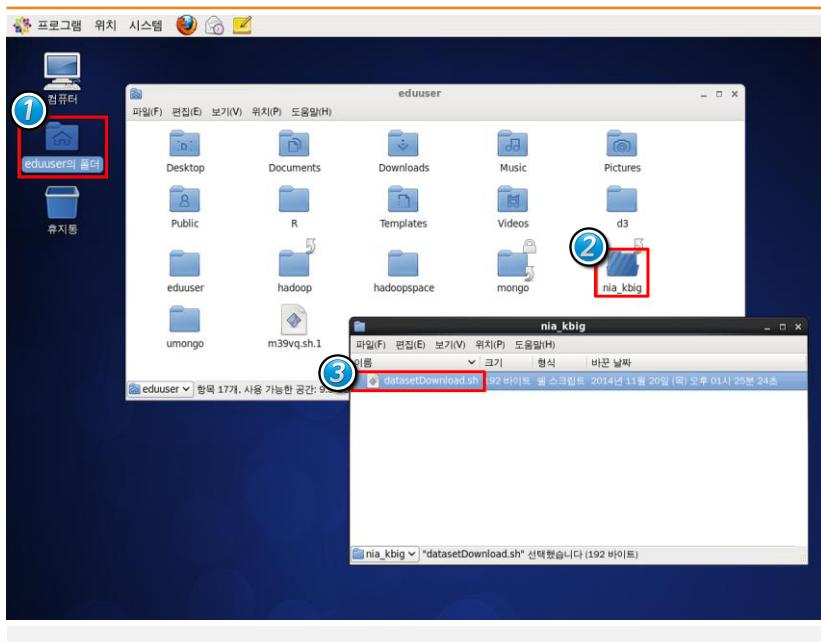
위와 같은 형태로 2011년과 2012년 데이터가 함께 제공된다.

- 2011년 상반기(mall_transaction_2011_1.csv)
- 2011년 하반기(mall_transaction_2011_2.csv)
- 2012년(mall_transaction_2012.csv)

▶ 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 쇼핑 데이터 셋을 복사해 온다.
 - `mall_transaction.xls` : 온라인 쇼핑몰 거래내역 데이터

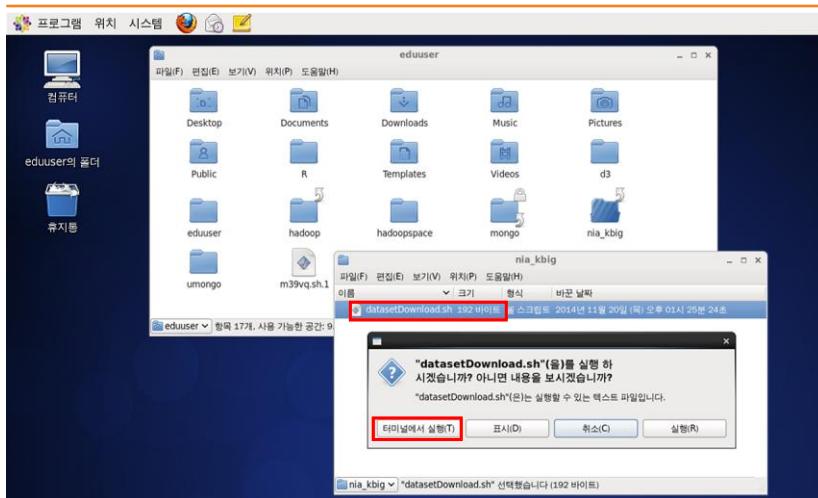
▶ 실습코드 디렉토리로 이동



- ① 로그인 후 바탕화면에서 eduuser 폴더를 오픈한다.
- ② nia_kbig 폴더를 오픈한다.
- ③ datasetDownload.sh를 더블클릭하여 실행한다.

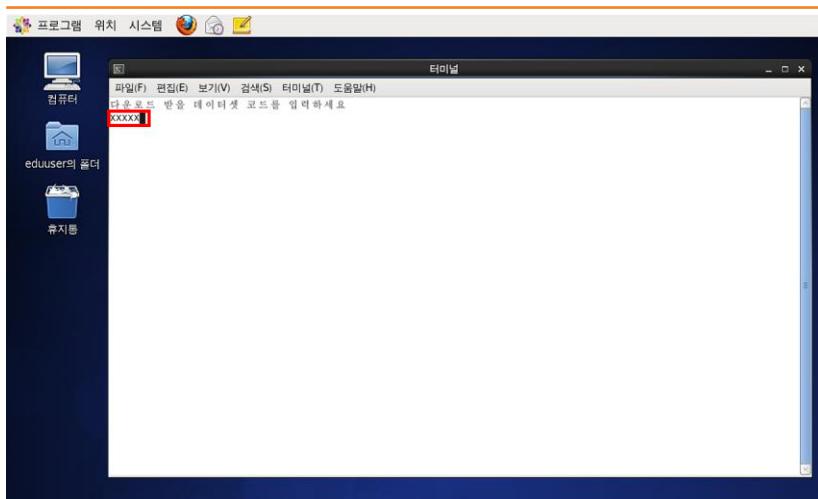
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



- '터미널에서 실행' 버튼을 클릭한다.

▶ 데이터셋 코드 입력



- 다운로드 받은 데이터셋 코드를 입력 후 엔터

▶ 데이터셋과 실습용 쉘 스크립트



- 실습용 데이터셋과 실습용 스크립트를 확인한다.

▶ ① 데이터 및 스크립트

▪ 01.move_data_file.sh :

로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

▪ 02.upload_csv.sh :

쇼핑 데이터를 하둡 파일시스템으로 업로드하는 스크립트

▪ 03.run.sh :

쇼핑 분석용 하둡 맵리듀스 프로그램 실행 스크립트

▪ 04.download_csv.sh :

하둡 맵리듀스 분석 결과 파일을 다운로드하는 스크립트

▪ mall_transaction_2010.csv : 2010년 남성 쇼핑몰 거래 데이터

▪ mall_transaction_201101.csv : 2011년 상반기 남성 쇼핑몰 거래 데이터

▪ mall_transaction_201102.csv : 2011년 하반기 남성 쇼핑몰 거래 데이터

▪ mall_transaction_2012.csv : 2012년 남성 쇼핑몰 거래 데이터

> 데이터 작업 영역 이동 스크립트(01.move_data_file.sh)

> 데이터 이동 스크립트

- 로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

01.move_data_file.sh (작업영역 폴더로 원시데이터 이동)

```

01. #!/bin/bash
02. # 복사 대상 파일 정의
03. # 쇼핑몰 거래내역 파일
04.
05. TARGET_TRANSACTION_2010=/home/eduuser/nia_kbig/mall/middle/mall_tr
   ↵ nsaction_2010.csv
06. TARGET_TRANSACTION_2011_1=/home/eduuser/nia_kbig/mall/middle/mall_
   ↵ transaction_2011_1.csv
07. TARGET_TRANSACTION_2011_2=/home/eduuser/nia_kbig/mall/middle/mall_
   ↵ transaction_2011_2.csv
08. TARGET_TRANSACTION_2012=/home/eduuser/nia_kbig/mall/middle/mall_tr
   ↵ nsaction_2012.csv
09.
10. # 작업영역 디렉토리 정의
11. LOCAL_DIR=/home/eduuser/nia_kbig/data/
12.
13. # 데이터 파일 이동
14. mv $TARGET_TRANSACTION_2010 $LOCAL_DIR
15. mv $TARGET_TRANSACTION_2011_1 $LOCAL_DIR
16. mv $TARGET_TRANSACTION_2011_2 $LOCAL_DIR
17. mv $TARGET_TRANSACTION_2012 $LOCAL_DIR

```

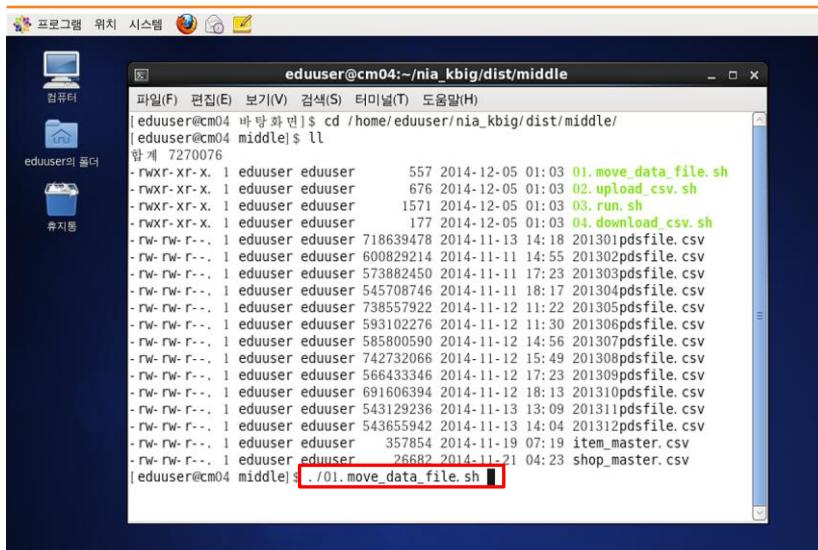


부연설명

- 데이터 작업 영역 이동 스크립트 소스(01.move_data_file.sh)
- 라인 05~08 : 다운로드 받은 원시데이터 파일들의 위치(path)를 변수 (TARGET_TRANSACTION_2010, TARGET_TRANSACTION_2011_1, TARGET_TRANSACTION_2012)로 지정하는 라인이다.
- 라인 11 : 작업영역 디렉토리의 위치(path)를 변수(LOCAL_DIR)로 지정하는 라인이다.
- 라인 14~17 : mv 명령어를 사용하여 다운로드 받은 원시데이터 파일들을 작업영역 디렉토리로 이동시키는 라인이다.

II. 수집

▶ 수집 데이터 셋 작업 영역 폴더 이동



The screenshot shows a terminal window titled 'eduuser@cm04:~/nia_kbig/dist/middle'. The window displays a file listing from the command 'll' and a command being typed at the prompt.

```
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[eduuser@cm04 바탕화면] $ cd /home/eduuser/nia_kbig/dist/middle/
[eduuser@cm04 middle]$ ll
합계 7270076
-rwxr-Xr-X 1 eduuser eduuser 557 2014-12-05 01:03 01.move_data_file.sh
-rwxr-Xr-X 1 eduuser eduuser 676 2014-12-05 01:03 02.upload_csv.sh
-rwxr-Xr-X 1 eduuser eduuser 1571 2014-12-05 01:03 03.run.sh
-rwxr-Xr-X 1 eduuser eduuser 177 2014-12-05 01:03 04.download_csv.sh
-rw-rw-r-- 1 eduuser eduuser 718639478 2014-11-13 14:18 201301pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 600829214 2014-11-11 14:55 201302pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 573882450 2014-11-11 17:23 201303pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 545708746 2014-11-11 18:17 201304pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 738557922 2014-11-12 11:22 201305pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 593102276 2014-11-12 11:30 201306pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 585800590 2014-11-12 14:56 201307pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 742732066 2014-11-12 15:49 201308pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 566433346 2014-11-12 17:23 201309pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 691606394 2014-11-12 18:13 201310pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 543129236 2014-11-13 13:09 201311pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 543655942 2014-11-13 14:04 201312pdfsfile.csv
-rw-rw-r-- 1 eduuser eduuser 357854 2014-11-19 07:19 item_master.csv
-rw-rw-r-- 1 eduuser eduuser 26682 2014-11-21 04:23 shop_master.csv
[eduuser@cm04 middle]$ ./01.move_data_file.sh
```

- 로컬에 원시데이터를 작업 영역 폴더로 이동 (/home/eduuser/nia_kbig/data/) 시킨다.
- ./01.move_data_file.sh 입력 후 엔터

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



III 가공

개요

65



64



가공

▶ 개요

작업 영역 폴더에 복사한 쇼핑 데이터의 가공은, 전처리 단계에서 수집된 2010~2012년 남성 의류 쇼핑몰 거래 내역 데이터를 로드하여, 하둡의 맵리듀스 기능을 이용한 분석 과정에서 동시에 진행되기 때문에, 별도의 가공 과정은 생략한다. 하둡에서 데이터를 읽어들여 맵리듀스 분석하는 과정에서 필요한 데이터만 추출하고 계산하는 일련의 가공 과정이 메모리상에서 동시에 이루어 진다.

▶ 가공 방법

- 쇼핑몰 판매 데이터는 2010~2012년간의 3년치 데이터가 각각의 CSV 파일로 저장되어 있다.

▶ 데이터셋

-온라인 쇼핑몰 거래내역 데이터 샘플(mall_transaction_2010.csv)

주문일자	거래상태	상품명	판매가	수량합계	합계금액
10.12.30 13:11	판매완료	ADW 짚업…	57200	4	194000
10.12.30 13:33	판매완료	스판 썰바지…	21400	1	23900
10.12.30 13:50	판매완료	무지 면 췄…	29200	1	30200
10.12.30 13:51	판매완료	스트라이프 …	15500	6	122300
10.12.30 14:07	반송(반품)	중청 스티치 …	39400	1	41900
10.12.30 13:11	판매완료	ADW 짚업 기모 후드티 a336 – 카키(1)	57200	4	194000

▶ 가공 과정

- 자바와 하둡을 이용하여 가공 및 분석 실행
- 자바를 활용하여 맵리듀스 분석을 할 경우에는 [분석]단계에서 가공 및 분석을 동시에 진행할 수 있으므로 [분석]단계에서 상세하게 설명한다.

WIT



IV 저 장

개요	69
가공 데이터 하둡 파일시스템 업로드	70
가공 데이터 하둡 파일시스템 저장	72

IV

저장

> 개요

자바와 하둡을 이용하여 맵리듀스를 실행하기 위해서는 하둡 파일 시스템에 데이터를 업로드하여야 한다. 따라서, 하둡 파일 시스템에서 제공하는 커맨드를 사용하여 2010년~2012년 쇼핑몰 거래 데이터 파일을 업로드한다.

> 저장 방법

- 2010년~2012년 쇼핑몰 거래 데이터 파일(mall_transaction*.csv)을 하둡에 업로드한다.
- 하둡 커맨드를 이용해서 업로드한다.

> 가공 데이터 하둡 파일시스템 업로드(02.upload_csv.sh)

> 하둡 파일시스템에 업로드 스크립트

- 가공 데이터를 하둡 파일시스템으로 업로드(upload_csv.sh)

02.upload_csv.sh (가공데이터를 하둡파일시스템으로 업로드)

```

01.#!/bin/bash
02.#
03.#
04.#
05.#
06.#
07.#
08.#
09.#
10.#
11.#
12.#
13.#
14.#
15.#
16.#
17.#
18.#
19.#
20.#
21.#

```

01. #!/bin/bash
02. # 2010년 쇼핑몰 거래 데이터
03. TRANS_OUTPUT_FILE_2010= '/home/eduuser/nia_kbig/data/mall_transacti
04. on_2010.csv'
05. # 2011년 상반기 쇼핑몰 거래 데이터
06. TRANS_OUTPUT_FILE_2011_1= '/home/eduuser/nia_kbig/data/mall_transa
07. ction_2011_1.csv'
08. # 2011년 하반기 쇼핑몰 거래 데이터
09. TRANS_OUTPUT_FILE_2011_2= '/home/eduuser/nia_kbig/data/mall_transa
10. ction_2011_2.csv'
11. # 하둡 파일시스템 저장 위치
12. HDFS_TRANSACTION_2010=/user/bigdata/mall_transaction_2010.csv
13. HDFS_TRANSACTION_2011_2=/user/bigdata/mall_transaction_2011_1.csv
14. HDFS_TRANSACTION_2011_1=/user/bigdata/mall_transaction_2011_2.csv
15. HDFS_TRANSACTION_2012=/user/bigdata/mall_transaction_2012.csv
16. #
17. # 파일을 하둡의 파일 시스템에 업로드
18. hadoop fs -put \$TRANS_OUTPUT_FILE_2010 \$HDFS_TRANSACTION_2010
19. hadoop fs -put \$TRANS_OUTPUT_FILE_2011_1 \$HDFS_TRANSACTION_2011_1
20. hadoop fs -put \$TRANS_OUTPUT_FILE_2011_2 \$HDFS_TRANSACTION_2011_2
21. hadoop fs -put \$TRANS_OUTPUT_FILE_2012 \$HDFS_TRANSACTION_2012

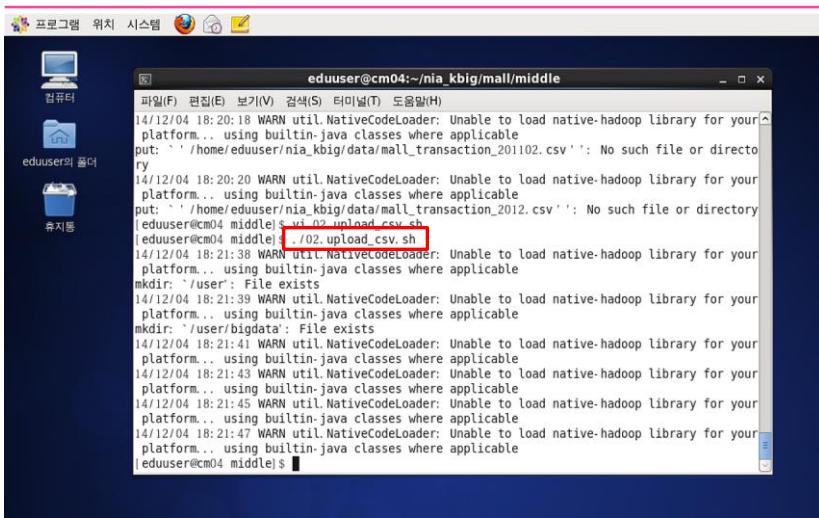
IV. 저장



- 70페이지 가공 데이터 업로드 스크립트(02.upload_csv.sh)
- 라인 02~09 : 작업영역으로 이동한 원시데이터 파일들의 위치(path)를 변수(TRANS_OUTPUT_FILE_2010, TRANS_OUTPUT_FILE_2011_1, TRANS_OUTPUT_FILE_2011_2, TRANS_OUTPUT_FILE_2012)로 지정하는 라인이다.
- 라인 11~15 : 하둡 파일시스템에 업로드할 파일들의 위치(path)를 변수(HDFS_TRANSACTION_2010, HDFS_TRANSACTION_2011_1, HDFS_TRANSACTION_2011_2, HDFS_TRANSACTION_FILE_2012)로 지정하는 라인이다.
- 라인 17~21 : hadoop fs -put 명령어를 사용하여 원시데이터 파일들을 하둡 파일시스템으로 업로드하는 라인이다.

> 가공 데이터 하둡 파일시스템 저장

> 가공 데이터 업로드 스크립트 실행(02.upload_csv.sh)



The screenshot shows a terminal window titled "eduuser@cm04:~/nia_kbig/mall/middle". The window displays the output of a command being run. The command is "/02.upload_csv.sh", which is highlighted with a red rectangle. The terminal shows several warning messages from the "NativeCodeLoader" class, indicating issues with loading native-hadoop libraries. It also shows the creation of a directory named "bigdata" and the existence of files like "bigdata" and "bigdata2". The terminal ends with a prompt "[eduuser@cm04 middle]\$".

```

파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
14/12/04 18:20:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
put: '/home/eduuser/nia_kbig/data/mall_transaction_201102.csv': No such file or directo
ry
14/12/04 18:20:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
put: '/home/eduuser/nia_kbig/data/mall_transaction_2012.csv': No such file or directory
[eduuser@cm04 middle]$ ./02.upload_csv.sh
14/12/04 18:21:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
mkdir: '/user': File exists
14/12/04 18:21:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
mkdir: '/user/bigdata': File exists
14/12/04 18:21:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
14/12/04 18:21:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
14/12/04 18:21:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
14/12/04 18:21:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
[eduuser@cm04 middle]$ 
```

- 2010년~2012년 쇼핑몰 거래 데이터 파일을 하둡 파일시스템에 업로드한다.
- ./02.upload_csv.sh* 입력 후 엔터

W





V 분석

개요	75
데이터 분석 스크립트	77
분석 데이터 파일 조회	84
결과 데이터 2차 분석	86

V 분석

> 개요

쇼핑 데이터의 분석은 자바 스크립트와 하둡 파일 시스템의 맵리듀스 기능을 활용하여 시계열 분석에 필요한 매출 패턴분석을 한다. 저장 단계에서 하둡 파일 시스템에 업로드한 2010년~2012년 남성 쇼핑몰의 거래 단위별 판매 내역 데이터를 읽어 들여, 맵리듀스를 실행하여 연월별 매출 통계를 산출하여 결과를 파일로 저장한다.

> 분석 예제

- 3년치 거래 단위별 판매 내역 데이터를 읽어들여 연도별로 월별 매출 추이를 분석한다.

> 분석 방법

- 하둡 파일 시스템에 업로드한 2010년~2012년 남성 쇼핑몰의 거래단위별 판매 내역 데이터를 로드한다.
- 패턴분석을 위한 분석 실행 스크립트를 활용하여 맵리듀스를 실행한다.
- 맵리듀스 결과로 만들어진 연월별 매출 통계 데이터를 하둡 파일 시스템으로 부터 다운로드한다.
- 계산된 데이터를 연도별로 월간 그래프로 시각화할 수 있도록 하기 위하여 오픈오피스 스프레드 시트의 피벗테이블 기능을 사용하여 연도별 월간 데이터를 산출한다.
- **통계 분석 기술** : 여러 개의 파일로 분산 저장된 데이터를 읽어들여 일괄적으로 통계를 계산하기 위해서 하둡 맵리듀스 기능을 활용한다.

> 저장 데이터

주문일자	거래상태	상품명	판매가	수량합계	합계금액
10.12.30 13:11	판매완료	ADW 짚업…	57200	4	194000
10.12.30 13:33	판매완료	스판 썰바지…	21400	1	23900
10.12.30 13:50	판매완료	무지 면 췌…	29200	1	30200
10.12.30 13:51	판매완료	스트라이프 …	15500	6	122300
10.12.30 14:07	반송(반품)	중청 스티치 …	39400	1	41900
10.12.30 13:11	판매완료	ADW 짚업 기모 후드티 a336 – 카키(1)	57200	4	194000
10.12.30 13:33	판매완료	스판 썰바지 5310-블랙(1)	21400	1	23900
10.12.30 13:50	판매완료	무지 면 췌리닝 바지(추동용) 9102- 다크그레이(1)	29200	1	30200

Tip ↗

- 가공 데이터 분석 스크립트(03.run.sh) 실행시, 맵리듀스 분석 실행 중 멈춤 현상 해결 방법
 - Ctrl+C 를 눌러 스크립트 실행 종료.
 - 하둡 종료 : 터미널 입력창에 stop-all.sh 입력 후 엔터.
 - 하둡 재실행 : 터미널 입력창에 start-all.sh 입력 후 엔터.
 - 하둡 실행 상태 확인 : 터미널 입력창에 jps 입력 후 엔터.
(목록 중에 NodeManager가 존재하는지 확인한다.)
 - 가공 데이터 분석 스크립트(03.run.sh) 재실행 : 터미널 입력창에 ./03.run.sh 입력 후 엔터

➤ 데이터 분석 스크립트(03.run.sh)

➤ 가공 데이터 분석 실행 셸스크립트

- 맵리듀스를 처리하는 프로그램은 Shopping.java에 구현되어 있다.
- 자바 프로그램을 컴파일하여 Shopping.jar 파일로 만든 후 yarn 커맨드를 이용해서 shopping.jar 파일로 맵리듀스 작업을 수행한다.
- 분석 결과는 하둡 파일시스템의 지정한 디렉토리에 저장한다.

03.run.sh (맵리듀스 실행)

```

01.#!/bin/bash
02. # 현재 위치를 지정한다.
03. CURRENT_DIR=/home/eduuser/nia_kbig/shopping/middle
04. # 컴파일하여 생성할 프로그램(jar) 경로를 지정한다.
05. TARGET_JAR=$CURRENT_DIR/shopping.jar
06. # 컴파일할 소스를 지정한다.
07. TARGET_SOURCE=$CURRENT_DIR/java_source/com/nia/hadoop/*.java
08. # jar를 생성하는데 필요한 class 파일을 지정한다.
09. TARGET_CLASSES=$CURRENT_DIR/com/nia/hadoop/*.class
10. # Hadoop상에 존재하는 농수산물 가격정보 파일을 지정한다.
11. INPUT_PRODUCT_DATA=/user/bigdata/mall_transaction*.csv
12. # 맵리듀스로 처리한 결과 데이터파일을 생성할 디렉토리를 지정한다.
13. OUTPUT_DIR=/user/bigdata/mall/out/2013
14. #컴파일에 필요한 hadoop 라이브러리 패스와 함께 source를 컴파일한다.
15. javac -classpath /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapr
   ↛ educe-client-core-2.2.0.jar:/usr/local/hadoop/share/hadoop/common/li
   b/commons-cli-1.2.jar:/usr/local/hadoop/share/hadoop/common/hadoo
   p-common-2.2.0.jar $TARGET_SOURCE
16. # 컴파일한 *.class 파일을 jar로 압축한다.
17. jar cf $TARGET_JAR $TARGET_CLASSES
18. # yarn 커맨드로 Hadoop에서 TARGET_JAR 프로그램을 돌려서 맵리듀스를 실행한다.
19. yarn jar $TARGET_JAR com.nia.hadoop.shopping $INPUT_PRODUCT_DATA
   ↛ $OUTPUT_DIR
20.

```



- 라인 03 : 현재 작업폴더 위치를 변수(CURRENT_DIR)로 지정하는 라인이다.
- 라인 05~09 : 하둡 맵리듀스 작업 수행 프로그램 파일(shopping.jar)을 컴파일하기 위한 환경을 지정하는 라인이다.
- 라인 11 : 맵리듀스 프로그램의 입력 데이터 경로를 변수(INPUT_DATA)로 저장하는 라인이다.
- 라인 13 : 맵리듀스 프로그램 실행 결과 파일을 저장할 경로를 변수(OUTPUT_DIR)로 저장하는 라인이다.
- 라인 15 : javac 명령을 사용하여 하둡 맵리듀스 작업 수행 프로그램 소스(Shopping.java)를 컴파일하여 작업 수행 프로그램(shopping.jar)을 컴파일하는 라인이다.
- 라인 17 : 하둡 맵리듀스 프로그램을 실행 가능한 작업 위치로 이동하는 라인이다.
- 라인 19 : yarn 명령을 사용하여 하둡 맵리듀스 프로그램을 실행하는 라인이다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

➤ 가공 데이터 분석 소스

Shopping.java (맵리듀스 스크립트)

```

01. package com.nia.hadoop;
02. import java.io.IOException;
03. import java.util.Iterator;
04.
05. import org.apache.hadoop.fs.Path;
06. import org.apache.hadoop.conf.*;
07. import org.apache.hadoop.io.*;
08. import org.apache.hadoop.mapred.*;
09. import org.apache.hadoop.util.*;
10.
11. public class ShoppingMall extends Configured implements Tool{
12.     public int run(String[] args) throws Exception
13.     {
14.         //하둡의 Job 이름 및 맵리듀스 클래스를 지정한다.
15.         JobConf conf = new JobConf(getConf(), ShoppingMall.class);
16.         conf.setJobName("ShoppingMallJob");
17.
18.         //맵리듀스 출력의 키와 값의 클래스 타입을 지정한다.
19.         conf.setOutputKeyClass(Text.class);
20.         conf.setOutputValueClass(IntWritable.class);
21.
22.         //매퍼 클래스와 리듀서 클래스를 지정한다.
23.         conf.setMapperClass(ShoppingMallMapper.class);
24.         conf.setReducerClass(ShoppingMallReducer.class);
25.         //맵리듀스를 실행할 입력 데이터와 출력 데이터의 파일 위치를 지정한다.
26.         Path inp = new Path(args[0]);
27.         Path out = new Path(args[1]);
28.
29.         FileInputFormat.addInputPath(conf, inp);
30.         FileOutputFormat.setOutputPath(conf, out);
31.

```

```

32.     JobClient.runJob(conf);
33.     return 0;
34. }
35.
36. public static void main(String[] args) throws Exception
37. {
38.     // 실제 실행되는 메인함수.(맵리듀스 클래스(ShoppingMall)를 생성한다.)
39.     int res = ToolRunner.run(new Configuration(), new ShoppingMall(),args);
40.     System.exit(res);
41. }
42.
43. public static class ShoppingMallMapper extends MapReduceBase
44.     ↪ implements Mapper<longWritable, text,="" intwritable="">
45. {
46.     private final static IntWritable number = new IntWritable(1);
47.     private Text word = new Text();
48.
49.     private static String linePrev = "";
50.
51.     private static String date = "";
52.     private static String status = "";
53.     private static String name = "";
54.     private static String num = "";
55.     private static String amount = "";
56.
57.     public void map(LongWritable key, Text value, OutputCollector<text,
58.         ↪ intwritable=""> output, Reporter reporter) throws IOException
59.     {
60.         //맵 함수는 한번에 입력 데이터 한 줄에 해당하는 데이터가 키(Key)/값(Value)
61.         ↪ 한쌍이 파라미터로 들어온다.
62.         String line = value.toString();
63.         String[] arr = line.split(",(?=(["""]*\"[""]*\"")*[""]*$)");
64.         String[] arrPrev = null;

```

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

V. 분석

```
65.     if (false==arr[0].equals("")){
66.     {
67.         date = arr[0];
68.         date = date.split(" ")[0];
69.
70.         String[] datesplit = date.split("/");
71.
72.         String month = datesplit[0];
73.         String year = "20" + datesplit[2];
74.         month = month.length()==1?"0"+month:month;
75.         date = year + month;
76.     }
77.
78.     if (arr.length >= 2 && false == arr[1].equals(""))
79.     {
80.         status = arr[1];
81.     }
82.     if (false == status.contains("완료"))
83.     {
84.         return;
85.     }
86.     if (true == status.trim().equals(""))
87.     {
88.         return;
89.     }
90.     if (arr.length >= 3 && false == arr[2].equals(""))
91.     {
92.         name = arr[2].trim();
93.     }
94.     if (arr.length < 6)
95.         return;
96.     if (arr.length >= 6 && false == arr[5].equals(""))
97.     {
98.         num = arr[5];
```

```

99.    }
100.   if (arr.length >= 7 && false == arr[6].equals(""))
101.   {
102.     amount = arr[7];
103.   }
104.
105.   // 구매단가
106.   int amountPerTransaction = Integer.valueOf(amount);
107.   number.set(amountPerTransaction);
108.
109.   // 년월별
110.   word.set(date);
111.
112.   // 리듀서로 넘길 키값과 데이터를 정의한다.
113.   output.collect(word, number);
114. }
115. }
116.

117. public static class ShoppingMallReducer extends MapReduceBase
118.   implements Reducer<text, IntWritable,="" text,="" intwritable="">
119. {
120.   //reduce 함수는 매퍼로부터 매핑된 키에 해당하는 값의 리스트를 파라미터로 받아
121.   //들인 후, 키에 해당하는 결과 값 한쌍을 output 콜렉션에 저장한다.
122.   public void reduce(Text key, Iterator<IntWritable> values,
123.     OutputCollector<text, intwritable=""> output, Reporter reporter)
124.     throws IOException
125.   {
126.     int sum = 0;
127.     //입력된 키에 해당하는 값들에 대하여 루프를 돌며 합계를 계산한다.
128.     while (values.hasNext())
129.     {
130.       sum += values.next().get();
131.     }
132.     output.collect(key, new IntWritable(sum));
133.   }
134. }

```

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

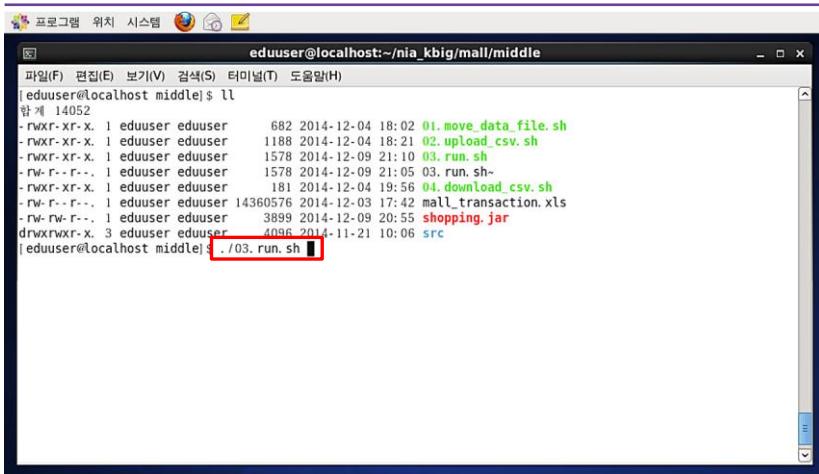
V. 분석



- 79페이지 맵리듀스 분석 프로그램 소스(Distribution.java)
- ShoppingMallMapper 클래스 (라인 43~115)
 - 맵리듀스 과정에서 매핑 기능을 정의한다.
 - map 함수(라인 57~114) : 판매년월을 기준으로 입력데이터를 매핑(묶음)한다.
 - map 함수를 수정하면 통계를 산출할 기준을 수정할 수 있다.
- ShoppingMallReducer 클래스 (라인 117~130)
 - 맵리듀스 과정에서 리듀스 기능을 정의한다.
 - reduce 함수(120~129) : 매핑을 통해 기준별로 묶인 데이터들을 합산하는 기능을 수행한다.
 - reduce 함수를 수정하면 합산 이외에 평균, 표준편차 등 원하는 통계값을 산출할 수 있다.

> 분석 데이터 파일 조회

> 분석 맵리듀스 실행



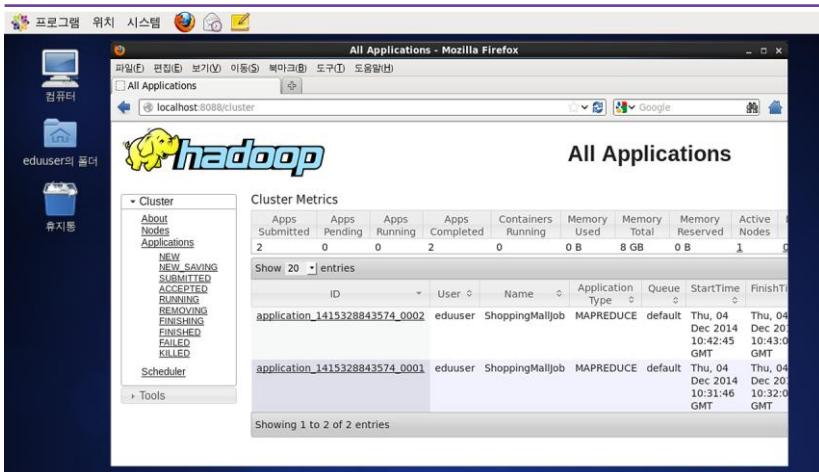
```

eduuser@localhost middle]$ ll
合계 14052
-rwxr-xr-x 1 eduuser eduuser      682 2014-12-04 18:02 01.move_data_file.sh
-rwxr-xr-x 1 eduuser eduuser    1188 2014-12-04 18:21 02.upload_csv.sh
-rwxr-xr-x 1 eduuser eduuser   1578 2014-12-09 21:10 03.run.sh
-rw-r--r-- 1 eduuser eduuser   1578 2014-12-09 21:05 03.run.sh~
-rwxr-xr-x 1 eduuser eduuser    181 2014-12-04 19:56 04.download_csv.sh
-rw-r--r-- 1 eduuser eduuser 14360576 2014-12-03 17:42 mall_transaction.xls
-rw-rw-r-- 1 eduuser eduuser   3899 2014-12-09 20:55 shopping.jar
drwxrwxr-x 3 eduuser eduuser    4096 2014-11-21 10:06 src
[eduuser@localhost middle]$ ./03.run.sh

```

- 하둡의 맵리듀스를 실행하여 데이터를 분석하여 결과 파일을 하둡 파일시스템에 생성 한다. ./03.run.sh 입력 후 엔터

> 맵리듀스 실행 현황 조회



All Applications - Mozilla Firefox

localhost:8088/cluster

hadoop

All Applications

ID	User	Name	Application Type	Queue	StartTime	FinishTime
application_1415328843574_0002	eduuser	ShoppingMalljob	MAPREDUCE	default	Thu, 04 Dec 2014 10:42:45 GMT	Thu, 04 Dec 2014 10:43:00 GMT
application_1415328843574_0001	eduuser	ShoppingMalljob	MAPREDUCE	default	Thu, 04 Dec 2014 10:31:46 GMT	Thu, 04 Dec 2014 10:32:00 GMT

- 파이어폭스 브라우저를 클릭한 후 주소 입력창에 <http://localhost:8088>을 입력 후 엔터를 치면 맵리듀스 진행과정을 볼 수 있다.

➤ 맵리듀스 분석 결과 파일 조회

The screenshot shows the Mozilla Firefox browser window with the URL `localhost:9000`. The page title is "NameNode 'localhost:9000' (active)". It displays the following information:

- Started:** Tue Nov 25 11:06:11 KST 2014
- Version:** 2.2.0, 1529768
- Compiled:** 2013-10-07T06:28Z by hortonmu from branch-2.2.0
- Cluster ID:** CID-97aafe1e3-2002-4dad-9652-1bdde0b3d5dSee
- Block Pool ID:** BP-1924143028-127.0.0.1-1413649470850

A red box highlights the "Browse the filesystem" link. Below it, the "Cluster Summary" section shows various metrics:

Configured Capacity	66.50 GB
DFS Used	3.89 MB
Non DFS Used	12.40 GB
DFS Remaining	54.10 GB
DFS Used%	0.01%
DFS Remaining%	81.35%
Block Pool Used	3.89 MB

- 맵리듀스의 분석 결과 파일을 확인하기 위해서 파일어폭스 브라우저 창에 `localhost:50070` 입력 후 엔터
- `Browse the filesystem` 링크 클릭하여 하위 폴더로 접근하면 출력 결과물 파일이 나온다,

The screenshot shows the Mozilla Firefox browser window with the URL `localhost:50070/browseDirectory.jsp?dir=%2Fuser%2Fbigdata%2Fmall%2Fout&namenodeid=0`. The page title is "HDFS:/user/bigdata/mall/out - Mozilla Firefox". It displays the following information:

Contents of directory `/user/bigdata/mall/out`

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 B	1	128 MB	2014-12-04 19:43	rw-r--r--	eduuser	supergroup
part-00000	file	477 B	1	128 MB	2014-12-04 19:43	rw-r--r--	eduuser	supergroup

Local logs

[Log directory](#)
[Hadoop](#), 2014.

- 출력 결과일 `/user/bigdata/mall/out/` 폴더에 `part-00000` (연월별 매출 통계 데이터) 파일을 조회할 수 있다.

▶ 결과 데이터 2차 분석

▶ 분석 결과 데이터 다운로드(04.download_csv.sh)

- 터미널 상에서 다운로드 스크립트(04.download_csv.sh)를 실행한다.

04.download_csv.sh (결과 파일 다운로드)

```

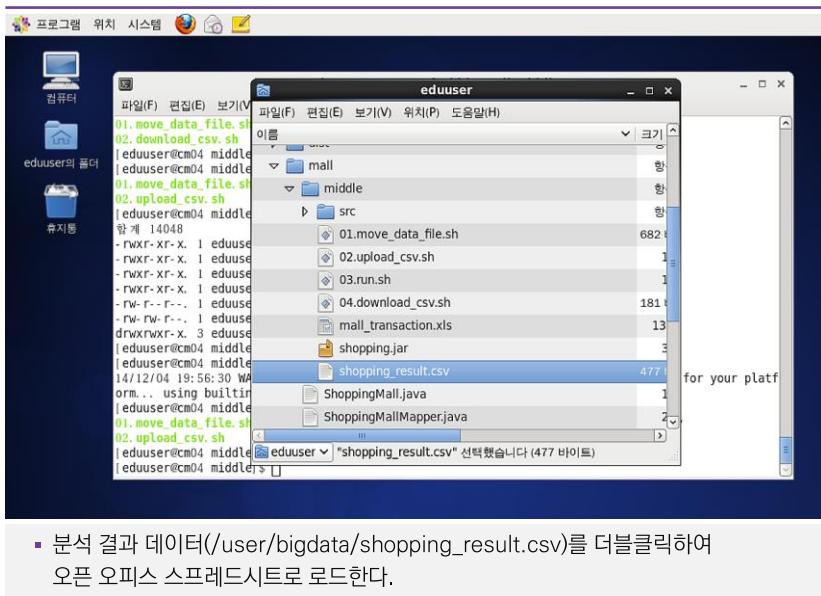
01. #!/bin/bash
02. # 파일을 하둡의 파일 시스템으로부터 다운로드
03. hadoop fs -get /user/bigdata/mall/out/part-00000 ./shopping_result.csv
04.

```

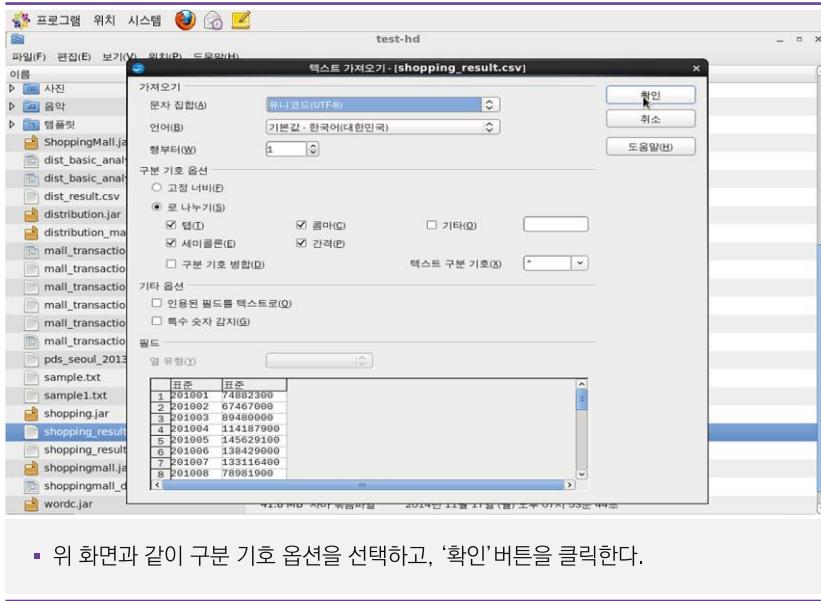


- 결과 파일 다운로드 스크립트 (04.download_csv.sh)
- 라인 3 : hadoop fs -get 명령을 이용하여 하둡 파일시스템으로부터 맵리듀스 분석 결과 파일을 다운로드하는 라인이다.

▶ 분석 결과 파일 로드



- 분석 결과 데이터(/user/bigdata/shopping_result.csv)를 더블클릭하여 오픈 오피스 스프레드시트로 로드한다.



- 위 화면과 같이 구분 기호 옵션을 선택하고, '확인'버튼을 클릭한다.

> 가공 및 분석

The screenshot shows a Microsoft Windows desktop environment. In the center is an OpenOffice Calc spreadsheet window titled "shopping_result.csv - OpenOffice Calc". The spreadsheet contains data from row 1 to 31, with columns A through K. Column C is currently selected, and a context menu is open over the data. The menu items visible are: 셀 서식 지정(B)..., 일 너비(U)..., 최적 일 너비(P)..., 일 삽입(I)..., 일 삭제(D), 내용 삭제(E)..., 습기기(B), 표시(S), 절라내기(I), 복사(C), 붙여넣기(P), 선택하여 붙여넣기(S)..., and 편집(E). The "일 삽입(I)" option is highlighted with a blue selection bar.

- 행을 삽입하여 제목(판매 연월, 판매금액)을 지정한 후 분석을 위해 두 개의 컬럼을 추가한다.
- 삽입한 행의 각 컬럼에 각각 '판매 년도', '판매월'을 입력하여 헤더를 완성한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

V. 분석

- 판매 년도를 구하기 위해 수식(LEFT(A2;4))을 입력한다.

- 판매월을 구하기 위해 수식(RIGHT(A2;2))을 입력한다.

I. 개요

shopping_result.csv - OpenOffice Calc

A	B	C
1	마트니코마트니코	판매 총액
2	2010012010	01 748823000
3	2010022010	02 67467000
4	2010032010	03 89480000
5	2010042010	04 114187900
6	2010052010	05 145629100
7	2010062010	06 138429000
8	2010072010	07 133116400
9	2010082010	08 78981900
10	2010092010	09 76223700
11	2010102010	10 118235300
12	2010112010	11 122931300
13	2010122010	12 110536200
14	2011012011	01 87977600
15	2011022011	02 63948000
16	2011032011	03 88265900
17	2011042011	04 112349300
18	2011052011	05 158127700
19	2011062011	06 125467300
20	2011072011	07 104306200
21	2011082011	08 75652200
22	2011092011	09 78635900
23	2011102011	10 121509900
24	2011112011	11 110621600
25	2011122011	12 133442400
26	2012012012	01 81255300
27	2012022012	02 72958100
28	2012032012	03 84441400
29	2012042012	04 118315600
30	2012052012	05 163054100
31	2012062012	06 125525400
32	2012072012	07 113763300

- 입력한 수식을 컬럼 전체로 복사한다.

II. 수집

shopping_result.csv - OpenOffice Calc

데이터(D) 창(W) 도움말(H)

피벗 테이블(P) > 만들기(M)...

A	B	C
1	마트니코마트니코	판매년월
2	2010012010	01 748823000
3	2010022010	02 67467000
4	2010032010	03 89480000
5	2010042010	04 114187900
6	2010052010	05 145629100
7	2010062010	06 138429000
8	2010072010	07 133116400
9	2010082010	08 78981900
10	2010092010	09 76223700
11	2010102010	10 118235300
12	2010112010	11 122931300
13	2010122010	12 110536200
14	2011012011	01 87977600
15	2011022011	02 63948000
16	2011032011	03 88265900
17	2011042011	04 112349300
18	2011052011	05 158127700
19	2011062011	06 125467300
20	2011072011	07 104306200
21	2011082011	08 75652200
22	2011092011	09 78635900
23	2011102011	10 121509900
24	2011112011	11 110621600
25	2011122011	12 133442400
26	2012012012	01 81255300
27	2012022012	02 72958100
28	2012032012	03 84441400
29	2012042012	04 118315600
30	2012052012	05 163054100
31	2012062012	06 125525400
32	2012072012	07 113763300

- 메뉴/데이터/피벗테이블/만들기 를 클릭하여 피벗테이블 마법사를 실행한다.

IV. 저장

V. 분석

VI. 시각화

V. 분석

The screenshot shows an OpenOffice Calc spreadsheet titled "shopping_result.csv". The data consists of four columns: 판매년도 (Year), 판매월 (Month), 판매일 (Date), and 판매총액 (Total Sales). A filter dialog box is open over the spreadsheet, titled "원본 선택" (Original Selection). It contains three radio button options: "현재 선택" (Current Selection) (selected), "OpenOffice에 등록된 데이터 원본" (Data source registered in OpenOffice), and "외부 원본/인터페이스" (External source/interface). Buttons for "확인" (Confirm), "취소" (Cancel), and "도움말" (Help) are at the bottom.

- 자동적으로 범위가 선택되므로 '현재 선택' 상태에서 그대로 확인을 누른다.

The screenshot shows an OpenOffice Calc spreadsheet titled "shopping_result.csv". The data includes columns for Year, Month, Date, and Total Sales. A more complex filter dialog box is open, titled "피봇 테이블" (Pivot Table). It features a "레이아웃" (Layout) section with "페이지 필드(G)" (Page Field), "판매년도" (Sales Year), "판매월" (Sales Month), and "판매일" (Sales Date). It also has a "필드" (Fields) section with "판매년도", "판매월", "판매일", and "판매총액". On the left, there's a "필드" (Fields) list containing "판매년도", "판매월", and "판매일". The right side shows a "데이터 필드" (Data Field) list with "합계 - 판매금액". Buttons for "확인" (Confirm), "취소" (Cancel), "도움말" (Help), "제거(N)" (Delete), and "옵션(O)...". A note at the bottom says "마우스를 이용하여 오른쪽에서부터 원하는 영역으로 필드를 끄십시오." (Drag the field from the right side to the desired area).

- 필드 영역의 아이템을 드래그하여 위 화면과 같이 구성한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

번호	연도	지역	판매량			판매액			평균가격		
			2010	2011	2012	2010	2011	2012	2010	2011	2012
32	2012072012	07			113763300						
33	2012082012	08			76565500						
34	2012092012	09			77432800						
35	2012102012	10			137891700						
36	2012112012	11			135160800						
37	2012122012	12			106334600						
38											
39											
40											
41	전체	전체	2010	2011	2012						
42	01	74882300	87977600	81255300	244115200						
43	02	67467000	63913300	72958100	204338400						
44	03	89480000	88265900	84441400	262187300						
45	04	114187900	112349300	118315600	344852800						
46	05	145629100	158127700	163054100	466810900						
47	06	138429000	125467300	125525400	389421700						
48	07	133116400	104306200	113763300	351185900						
49	08	78981900	75652200	76565500	231196000						
50	09	76223700	78635900	77432800	232292400						
51	10	118235300	121509900	137891700	377636900						
52	11	122931300	110621600	135160800	368713700						
53	12	110536200	113442300	106334600	330313100						
54	전체	1.27E+009	1.240269200	1.29E+009	3.603067900						
55											
56											
57											
58											
59											
60											
61											
62											
63											

- 화면의 최하단까지 스크롤하면 위와 같이 피벗테이블이 집계되어 있는 것을 볼 수 있다.(시각화를 위한 데이터 분석 완료)



1

2



VI 시각화

개요	95
시각화 과정	96
분석 데이터 시각화	97
데이터 분석	98

VI

시각화

> 개요

쇼핑 데이터의 시각화 과정에서는, 분석 과정에서 하둡 맵리듀스 분석과 오픈오피스 스프레드시트의 피봇 테이블 기능으로 작성된 통계 데이터를 오픈 오피스 스프레드 시트의 차트 기능을 활용하여 꺾은선 그래프로 시각화한다.

> 시각화 방법 및 활용기술

- 오픈 오피스 스프레드 시트의 차트 기능을 활용하여 멀티컬럼 데이터를 하나의 차트로 표현한다.
- 3개 년도의 매출 패턴을 한눈에 확인하기 위하여 오픈오피스 스프레드시트의 꺾은선 차트를 활용하여 시각화한다.



용어 정리

- **오픈오피스(OpenOffice)**

- 마이크로소프트 오피스와 같은 오피스 스위트입니다.
- 기존 오피스 프로그램과의 뛰어난 호환성을 자랑합니다.
- 제품 개발의 전 과정이 투명한 공개 소프트웨어 프로젝트입니다.
- 라이센스 비용을 지불할 필요가 없는 무료 소프트웨어입니다.
- 윈도우 뿐만 아니라 리눅스와 솔라리스 등 다양한 운영체제를 지원합니다.

> 시각화 과정

> 시각화 절차



> 시각화 과정

- Hadoop 분석 결과 파일을 로컬의 작업 폴더로 다운로드한다.
- 결과 파일을 오픈오피스 스프레드 시트에서 로드한다.
- 데이터의 헤더(제목) 부분을 입력한다.
- 시각화 할 데이터 영역을 지정한다.
- 파일/삽입/차트 를 선택하여 차트 설정을 연다.
- X 축과 Y축의 값을 지정한다.
- 완료하여 표현된 차트를 확인한다.

> 분석 데이터 시각화

> 시각화 차트 설정

The screenshot shows the OpenOffice Calc interface with a chart setup dialog box overlaid. The dialog is titled '차트 마법사' (Chart Wizard) and is set to the '선' (Line) chart type. The data series is selected as '선택 항목' (Selected items). The chart style is set to '점과 선' (Point and Line). The '선을 유연하게 조정(M)' (Adjust line flexibly) checkbox is checked. The '마침(F)' (Finish) button is highlighted.

Below the dialog, the spreadsheet contains data from rows 32 to 55, with columns A through L. The data includes dates from 2012 to 2013 and various numerical values.

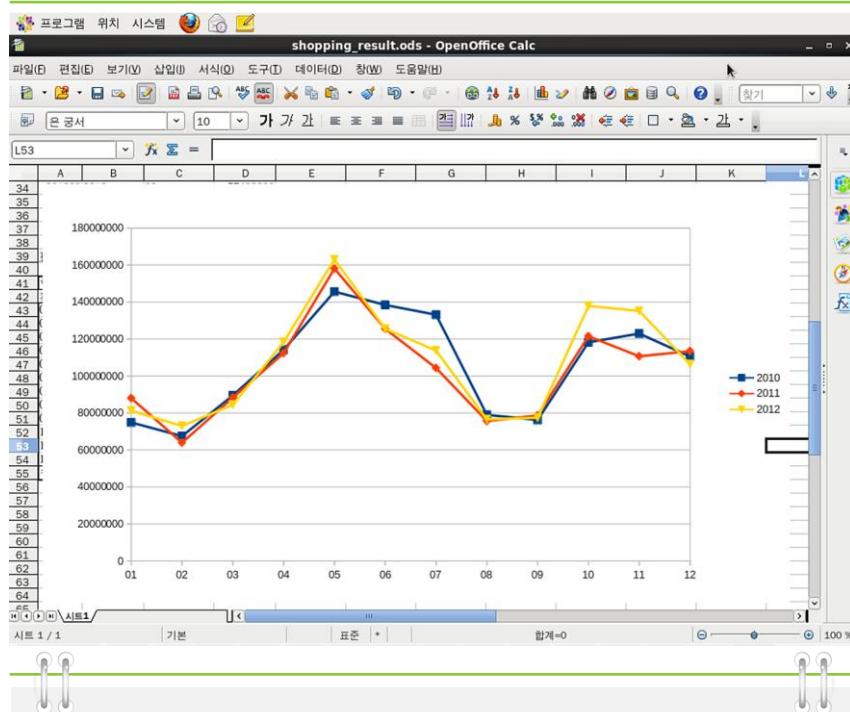
■ 시각화 할 데이터 범위를 지정한 후, 메뉴/삽입/차트를 선택하여 차트를 생성한다.

This screenshot shows the same OpenOffice Calc interface with the chart setup dialog still open. The '선' (Line) chart type is now selected in the '차트 유형' (Chart Type) section of the dialog. The other options like '열' (Column), '막대' (Bar), and '원형' (Pie) are dimmed. The '선과 선' (Point and Line) style is also selected in the preview area. The '마침(F)' (Finish) button is visible at the bottom right of the dialog.

■ 차트 유형에서 선을 선택하고, 우측에서 점과 선을 선택한 후 마침을 누른다.

> 데이터 분석

> 차트 데이터 분석



- 3개년 월별 매출 패턴을 보면, 대체로 월별 매출 추이가 비슷하게 나타남을 알 수 있다.
- 5~6월, 10~12월이 매출 성수기이며, 8~9월이 비수기임을 알 수 있다.
- 성수기와 비수기의 매출 차이가 확연히 드러나므로, 분석 결과를 참고하여 홍보를 통한 성수기 매출 극대화 시점과 비수기 매출 보완을 위한 할인 행사 시점 결정 등에 활용한다.



VII 예제문제

예제 문제1

101

예제 문제2

102

예 / 제 / 문 / 제

예제 1

3년치 거래 단위별 판매 내역 데이터를 읽어들여
연도별로 월별 매출 추이를 지역별 비교 분석하라.

- 연도별로 저장된 거래 단위별 판매 내역 데이터를 읽어들여 연도별/월별 매출을 합산하여, 월간 매출 추이를 분석하라.

- 연도별로 따로 저장된 거래 단위별 정보 데이터를 로드한다.
- 맵리듀스를 통해 연도별로 월간 매출 금액을 합산한다.
- 월간 매출 추이를 연도별로 시각화하여 분석한다.

예제 2

3년치 거래 단위별 판매 내역 데이터를 읽어들여
연도별로 최고 매출 아이템 Best5를 추출하라.

- 3년간 거래단위별 아이템명과 매출금액으로부터 연도별, 아이템별 매출 금액을 합산, 정렬하여 최고 매출 아이템을 추출하라.

- 연도별로 저장된 거래단위별 매출금액 데이터를 로드한다.
- 맵리듀스를 통해 연도별로 아이템별 매출금액을 합산한다.
- 연도별로 아이템별 매출금액을 내림차순으로 정렬한다.
- 연도별 Best5 를 추출하여 시각화한다.

데이터 분석 콘텐츠 활용 매뉴얼

2014년 12월 인쇄

2015년 1월 발행

발 행 처 한국정보화진흥원 빅데이터전략센터

집 필 신신애, 김성현, 박재원, 김현태, 김지홍, 정다운,
이승하, 신은비

주 소 서울시 중구 청계천로 14

연 락 처 (02) 2131-0114

인 쇄 HNJ Printing

〈비매품〉

[데 이 터 분 석 콘 텐 츠]

활용 매뉴얼

NIA  한국정보화진흥원

(100-775) 서울시 종구 청계천로 14 한국정보화진흥원
TEL 02-2131-0114 FAX 02-2131-0109
www.nia.or.kr

