Integrated Workflow for Large Database

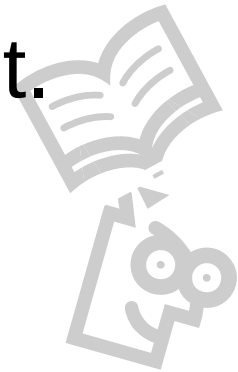
National Yang-Ming University

Health Informatics and Decision Support

Yu Chun Chen

Beautiful World

- There’s more than one way to do it.
- Which would be better ?



Size Does Matter

- Taiwan NHI database
- Manipulation of large and plain text files
  - Error Checking
    - transporting I/O error, missing word, mis-placed line terminator
  - Splitting Files
    - sampling
  - Cutting Files
    - SELECT certain fields
  - Join Files
    - Merging files

Possible Strategies-SAS

- Proven Statistical Package
- Knowledge Manager
- Various Product Line
  - Database: SAS/Datawarehouse
  - Data Mining: SAS/Enterprise Miner
  - ……

Possible Strategies-DBMS

- Data Manipulation
  - Abstract level of Data Modeling
  - Data Consistency / Correctness
  - Transaction Support
  - Security
- Available Packages
  - SQL server, Oracle, Sybase, IBM UDB2
  - mySQL

Possible Strategies- Hand-made Prog.

- Hand-made programs
  - VB, C, C++, Java, PERL … etc
- PERL
  - Jan 1988, Larry Wall
  - Practical Extract & Report Language
  - Multipurpose esp. Text Processing
    - CGI
    - Bioinformatics
    - Text parsing
    - Reporting

Possible Strategies - ?

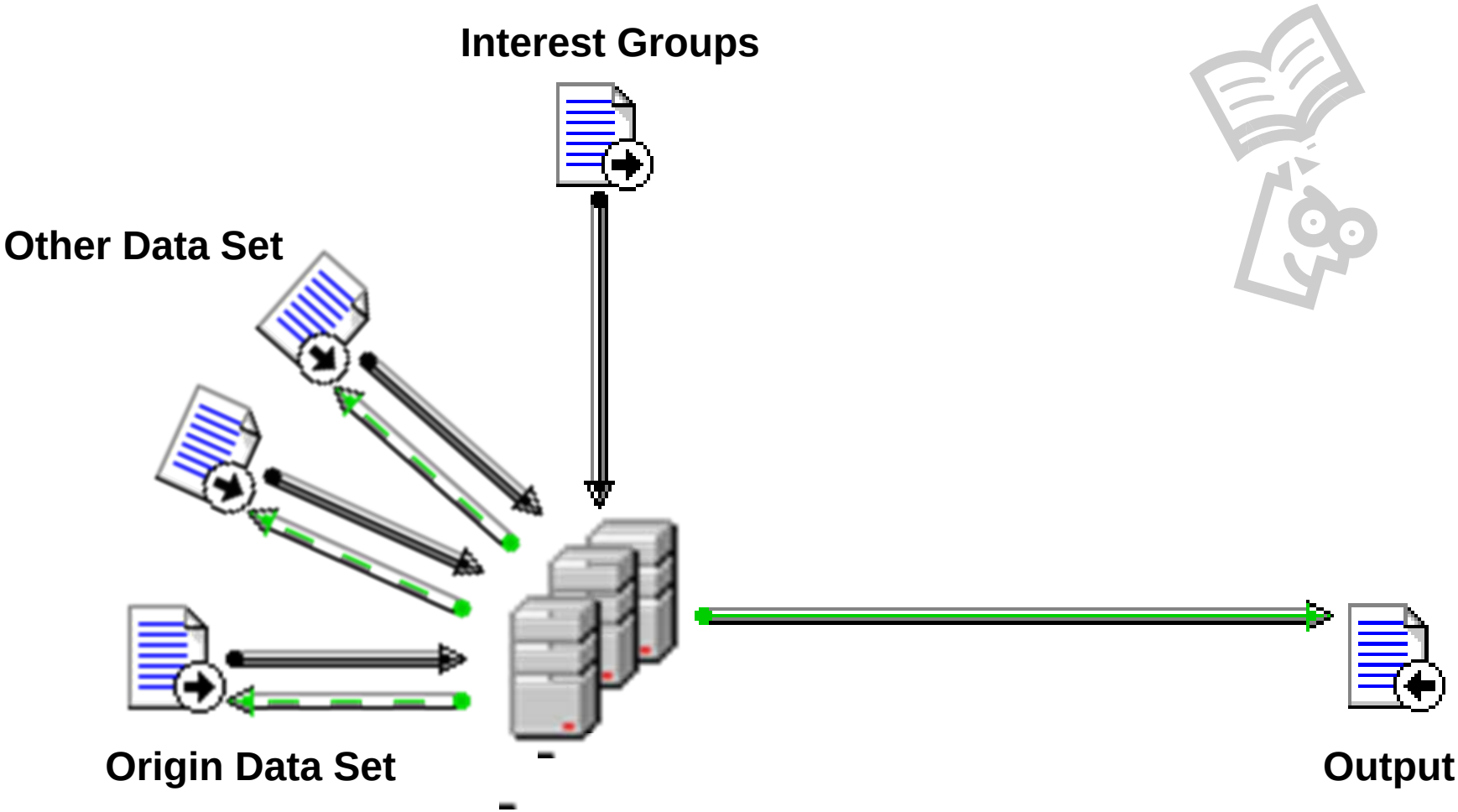
- Which one would be suitable ?
  - Efficient, price, easy
- Advantages and disadvantages

Real Work

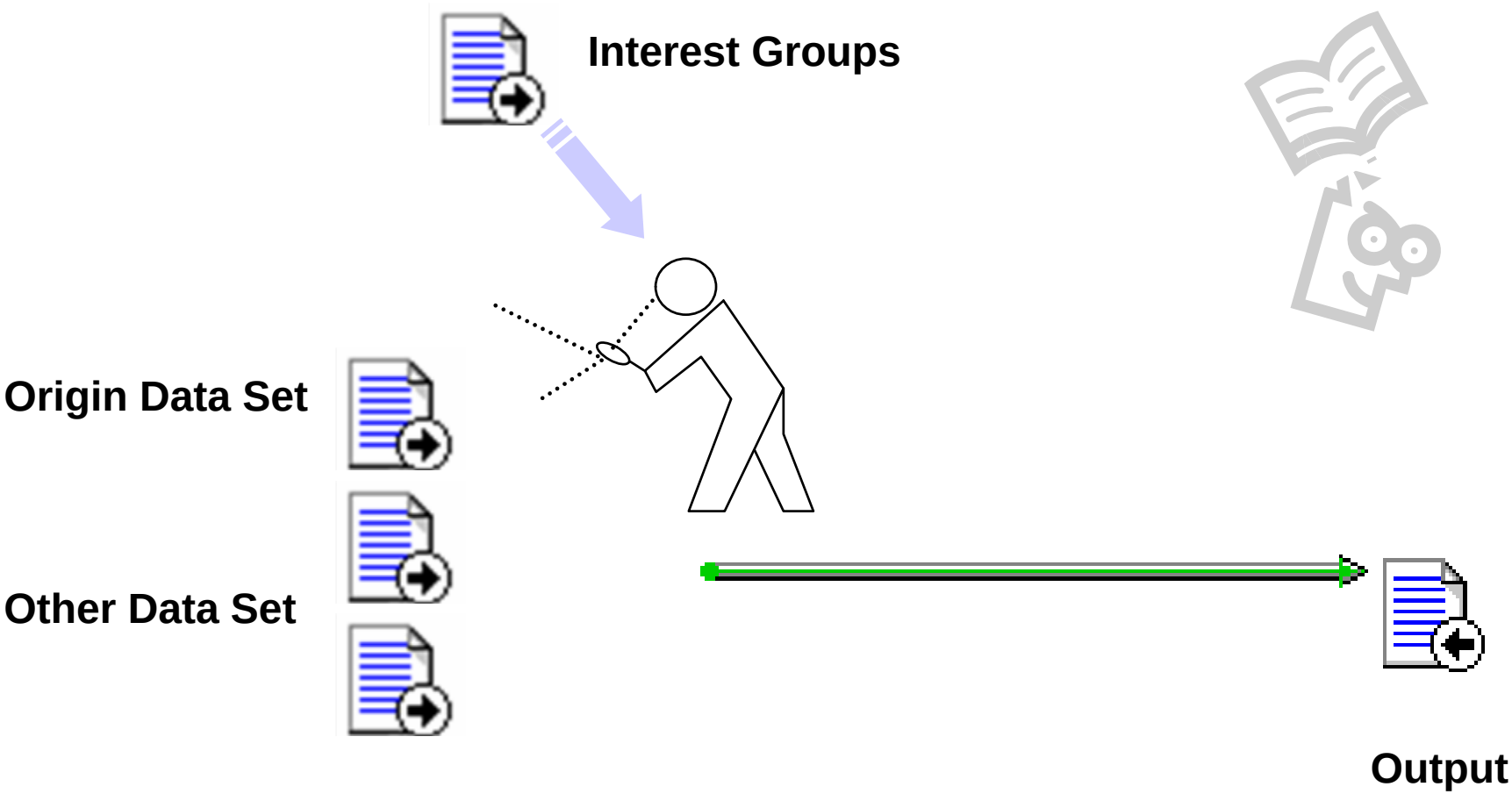
- 找出特定目標 (Interest groups)
- 從 ?? 資料檔裡面找出特定目標 的 ?? 紀錄

- 「串檔」作業：
  - 接受過某種手術病人的所有就診紀錄
  - 罹患某種病病患的所有住院紀錄
  - 某種藥物的使用情形 (什麼人在使用，什麼 狀況下使用）
  -

Generic Data Processing - I



Generic Data Processing - II



Material

- Interest Groups
  - 民眾清單 (Int5k, Int10k)
    - 只包含 ID + Birthday (18 bytes)
  - 藥物清單 (Int48M)
    - 48,000,000 records
    - Transactional ID (33 bytes)
- Sample Data Set
  - Set1, Set2, Set3
    - ≈70,000,000 records/file
    - ≈1.5 GB/file

Experiment:

- Three methods:
  - SAS
  - Database – SQL Server
  - PERL
- Run the same data files with 3 methods
- Record each elapsed time

Hardware Platform

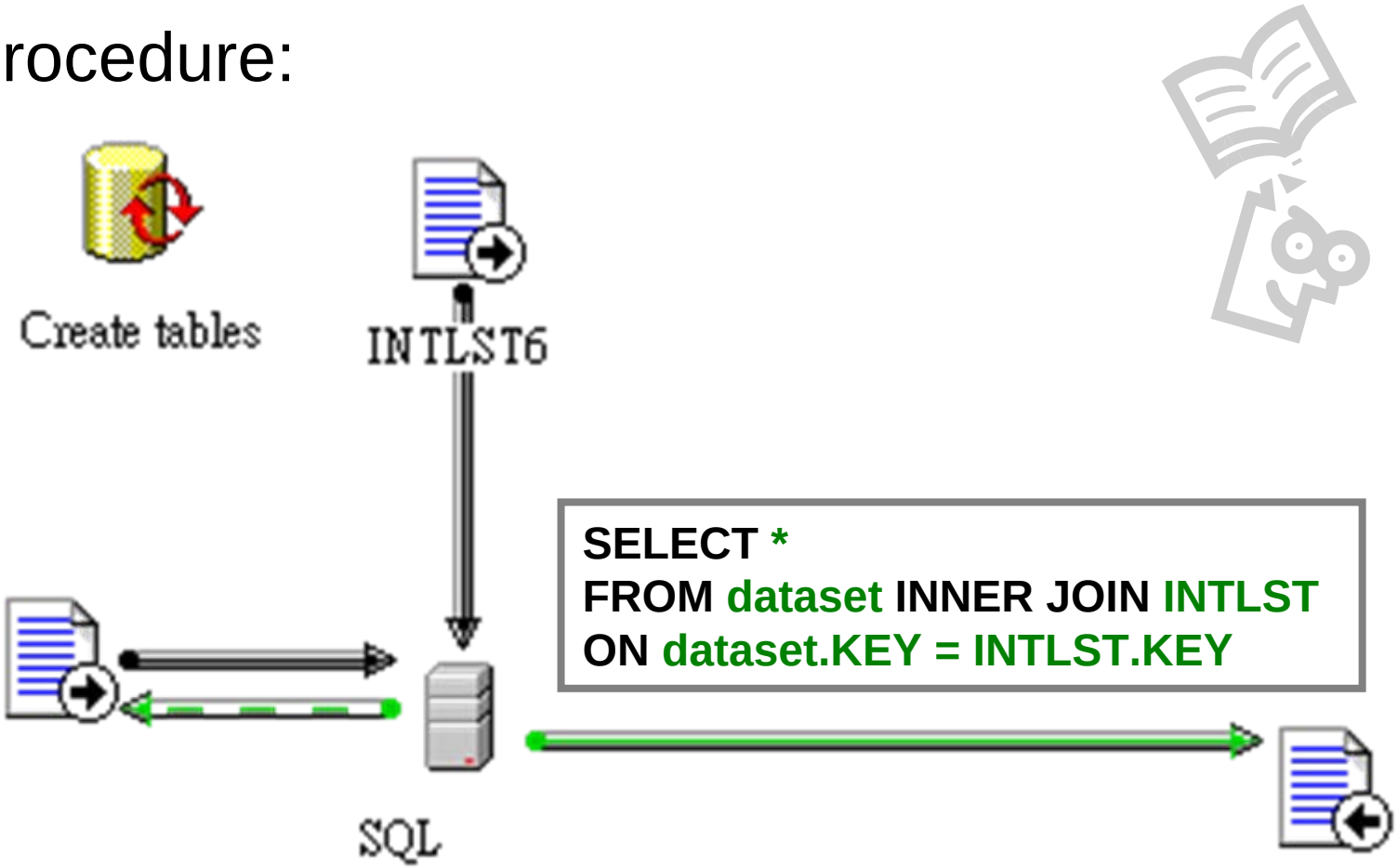
- Personal Computer
  - Athlon 1.6 GMHz
  - 512 MB
  - 20 GB Hard-Disk
- OS
  - Windows 2000 Server/Professional

SAS

- SAS 8.1
- Procedure:
  - ①.先分別匯入所有檔案
  - ②.執行 DATA MERGE / PROC SQL 敘述
  - ④.記錄時間
- 爲了加快速度，記錄檔每筆記錄只分成三部分：
  - lpart char(59) // left portion of record
  - BirthID char(18) // Birtday + ID
  - rpart char(132) // right portion of record

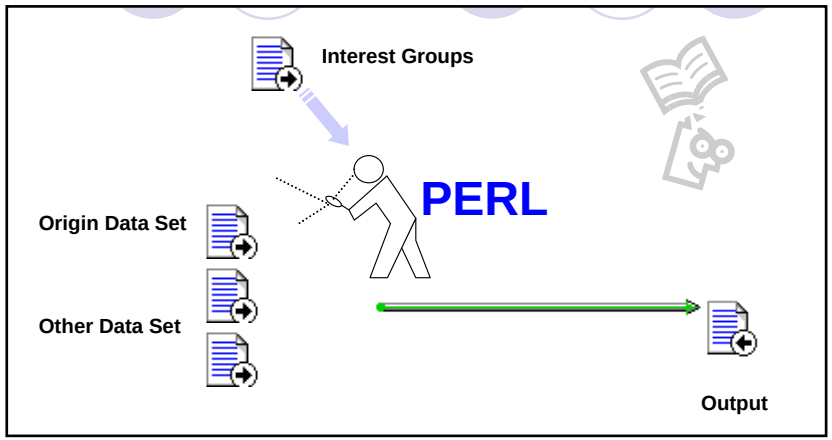
SQL Server 2000

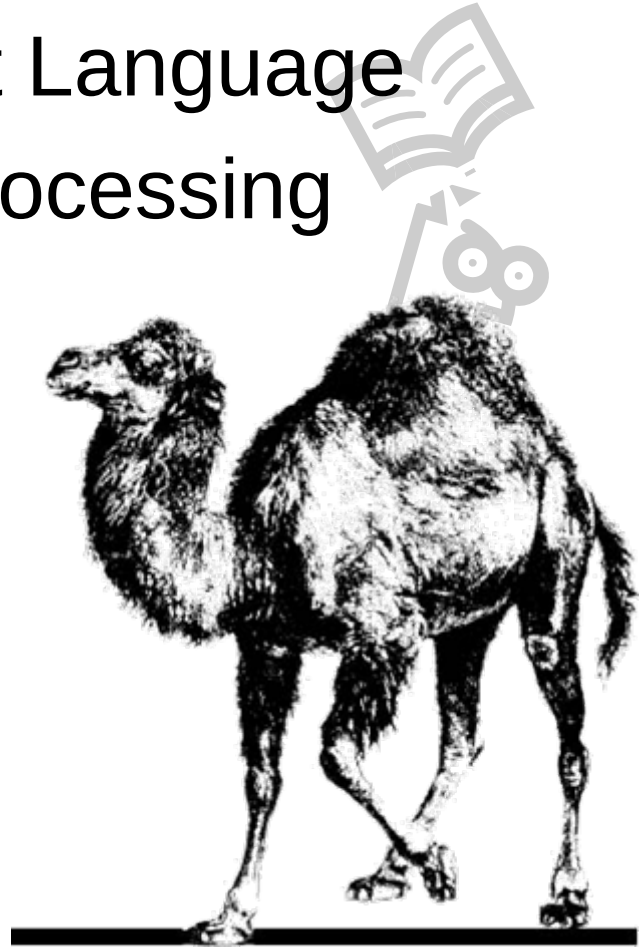
- Procedure:



PERL

- Practical Extract & Report Language
- Multipurpose esp. Text Processing
- Let Things Simple
  - Hash Join - algorithm
  - Hash Join - coding





Result

SAS

	Interest	Screen(import)	Screen(Join)	Screening Time	TOTAL
5,000	0.7	15.0	1.7	16.7	17.3
10,000	0.9	15.0	1.7	16.7	17.6
8,000,000	3.0	15.0	1.7	16.7	19.7

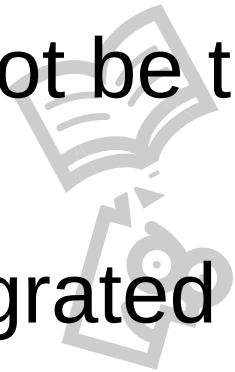
SQL

	Interest	Screen(import)	Screen(Join)	Screening Time	TOTAL
5,000	0.7	15.0	15.0	30.0	30.7
10,000	0.9	15.0	17.0	32.0	32.9
8,000,000	4.2	15.0	29.0	44.0	48.2

PERL

	Interest	Screen(import)	Screen(Join)	Screening Time	TOTAL
5,000	0.0	3.3	0.0	4.0	3.4
10,000	0.0	3.3	0.0	4.0	3.4
8,000,000	0.9	3.3	0.0	4.0	4.2

Discussion

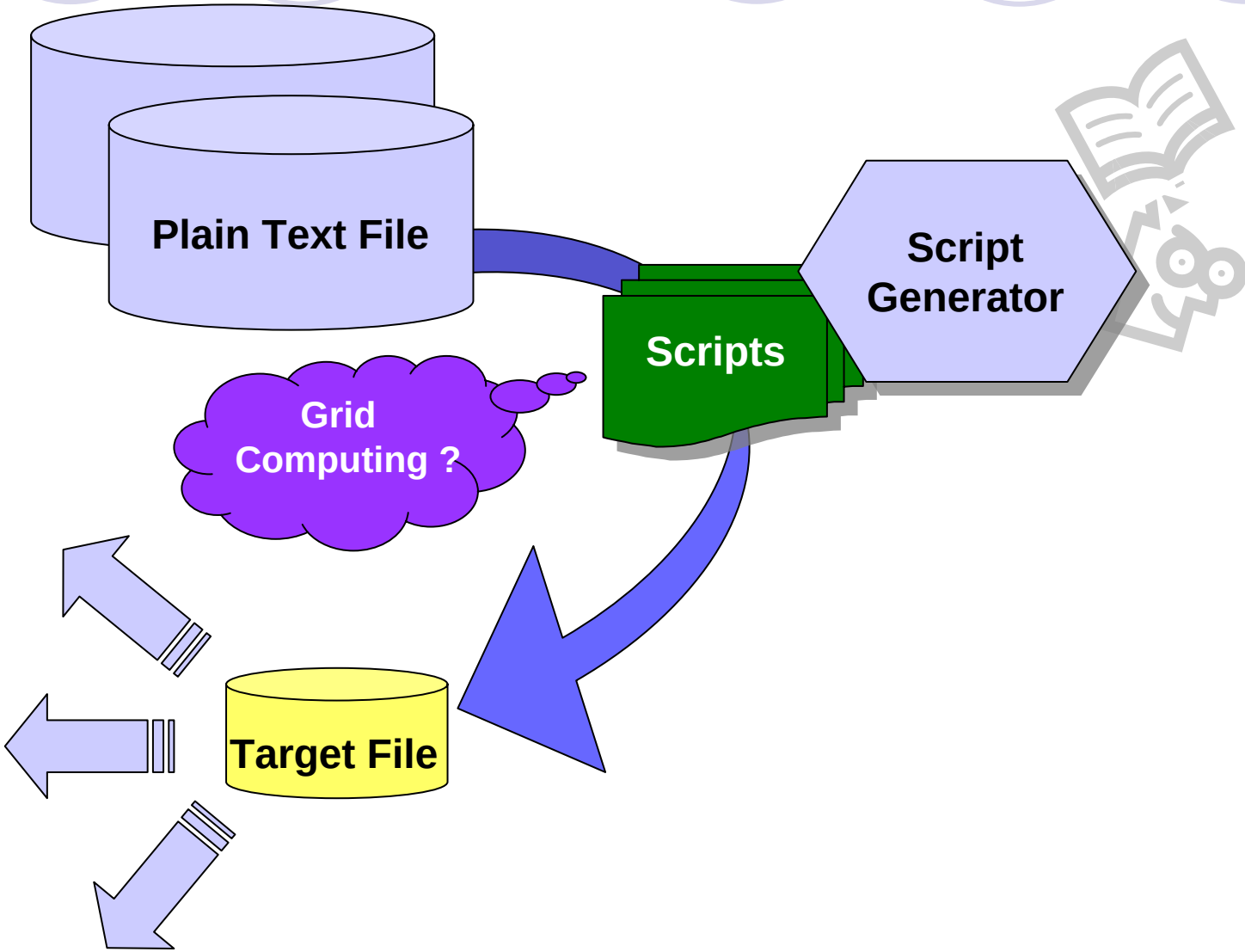


- A single, triumph software might not be the only solution
- The same work can be easily integrated into dataware-housing software
  - SAS/Dataware-house
  - SQL server, DB2 …etc
- A lightweight solution would be more cost-effective under certain scenario.

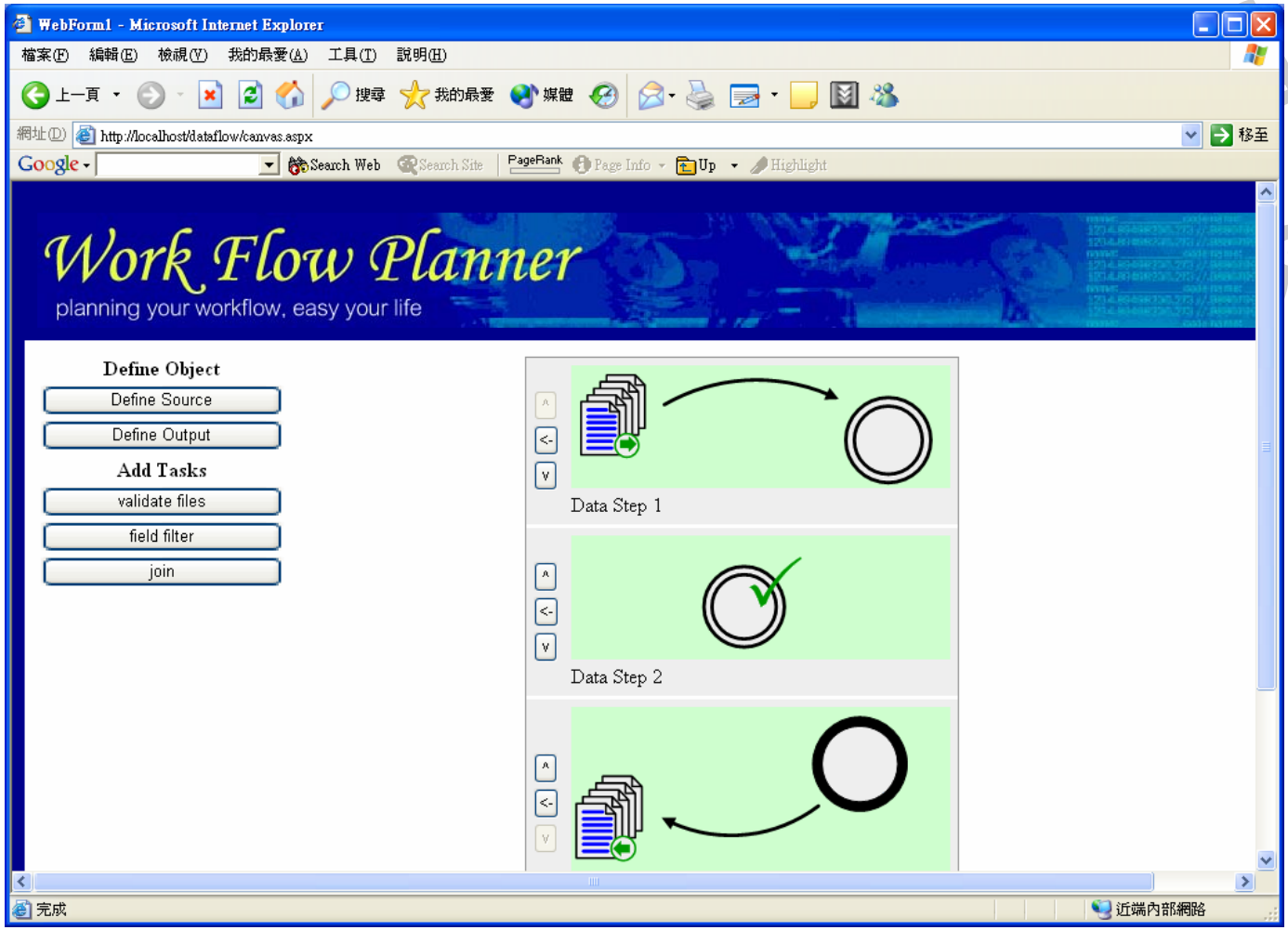
Suggestion

- Simple task
  - Data Validating
    - Check data correctness
  - Data Transformation
  - Simple join
  - Iceberg search

Current Solution



Script Generator



SAS – macro


```
/*-----------------------------------------------
              資料準備階段, 執行一次就好了
  -----------------------------------------------*/
* Int6K;
DATA cbrain.test; set cbrain.test;
    proc sort; by birth_id; run;
* Int10K;
data cbrain.one; set cbrain.one;
    proc sort; by birth_id; run;
/*-----------------------------------------------
       正式開始
-----------------------------------------------*/
* * 讀取 TEST 檔案 並排序
data cbrain.s199801;
    infile 'z:\CD199801_11.txt';
    input birth_id$ 60-77 lpart$ 1-59 rpart$ 78-209;
run;
data cbrain.s199801;
    proc sort; by birth_id; run;
* TEST1: Merge 6K;
data cbrain.one_ok;
    merge cbrain.one(in=ava1) cbrain.s199801(in=ava2); by birth_id; if(ava1)=1; run;
* TEST1: Merge 10K;
data cbrain.test_ok;
    merge cbrain.test(in=ava1) cbrain.s199801(in=ava2); by birth_id; if(ava1)=1; run;
* Rename;
```

=BACK=