

Evolutionary Design: Data Evolution - Process

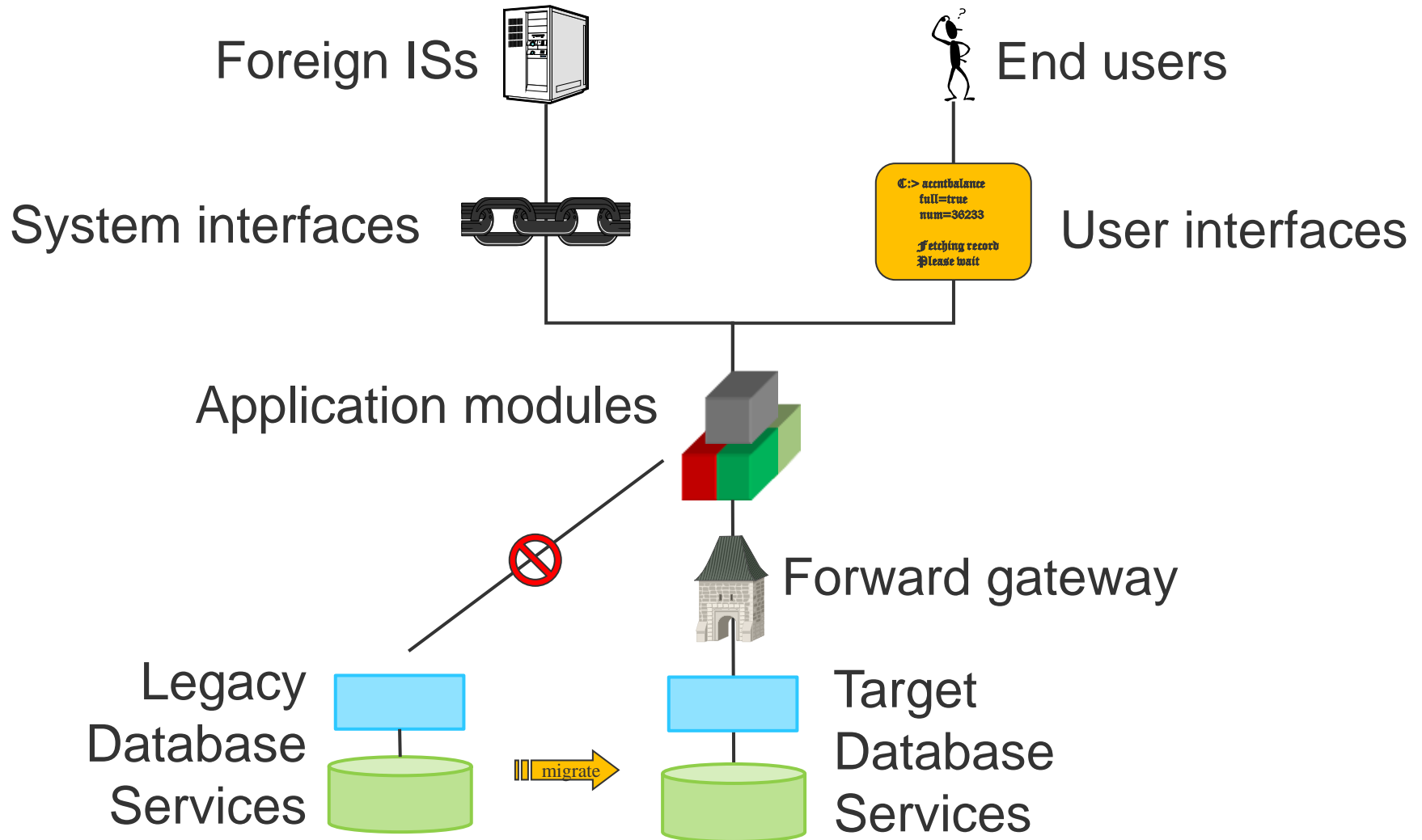
Andy Carpenter

School of Computer Science

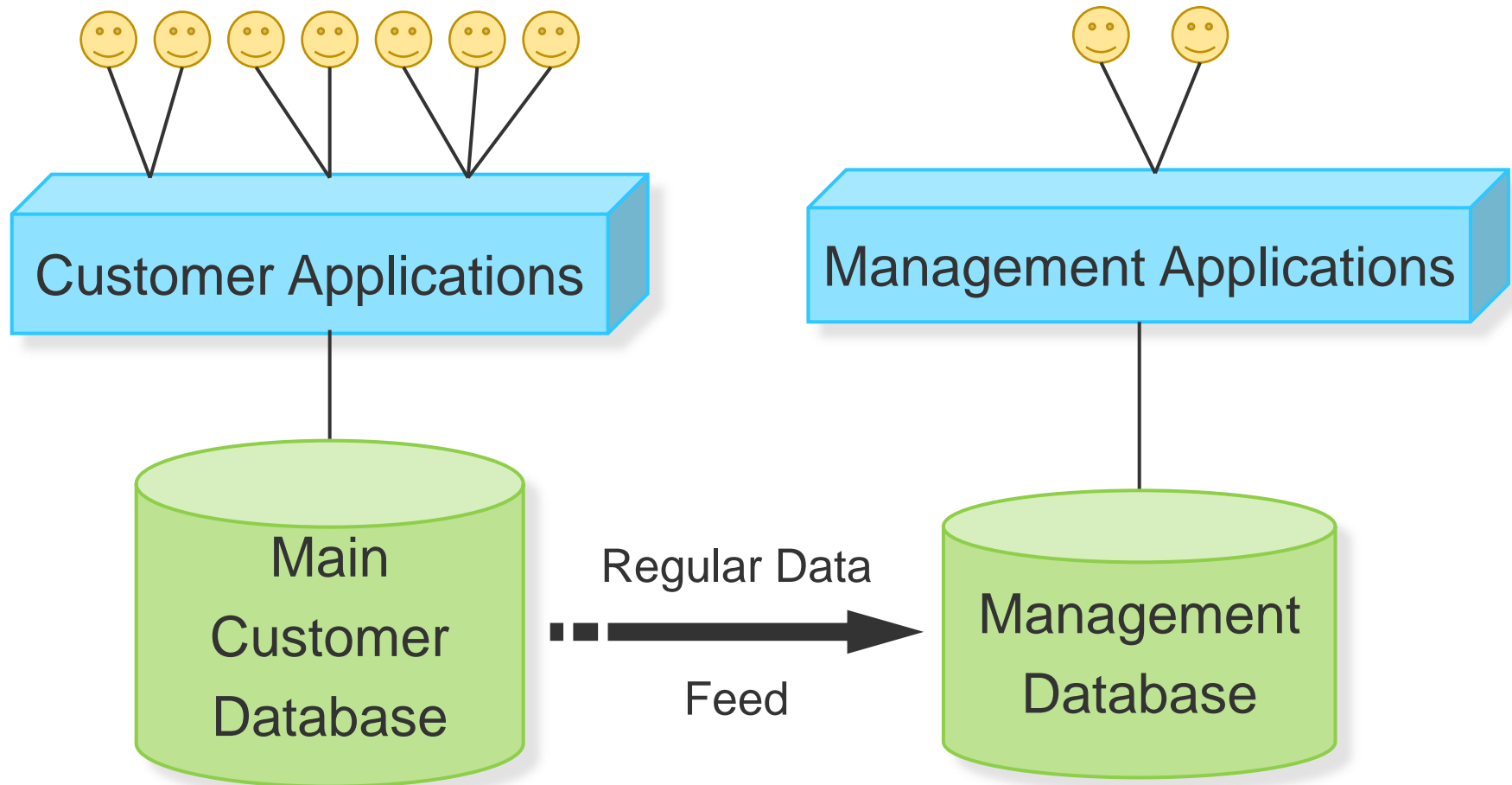
(Andy.Carpenter@manchester.ac.uk)

Elements these slides come from Sommerville, author of "Software Engineering", and are copyright Sommerville

Migration for Re-Engineering

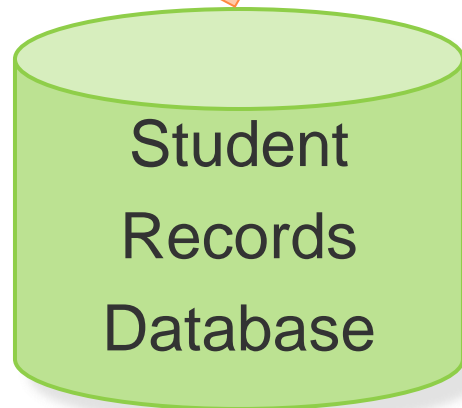


Data Extraction/Replication



Example

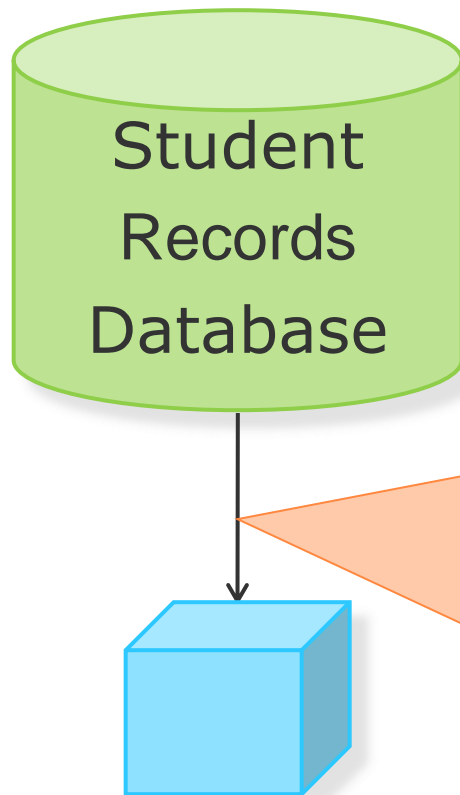
Student(RegNo, Name, Level, Gender)
Course(Code, Title, Level, Description)
Exam(RegNo, Code, Session, Mark)
Cwork(RegNo, Code, Session, Mark, I)



Mark(RegNo, Sex,
Level, Session,
Type, Mark)

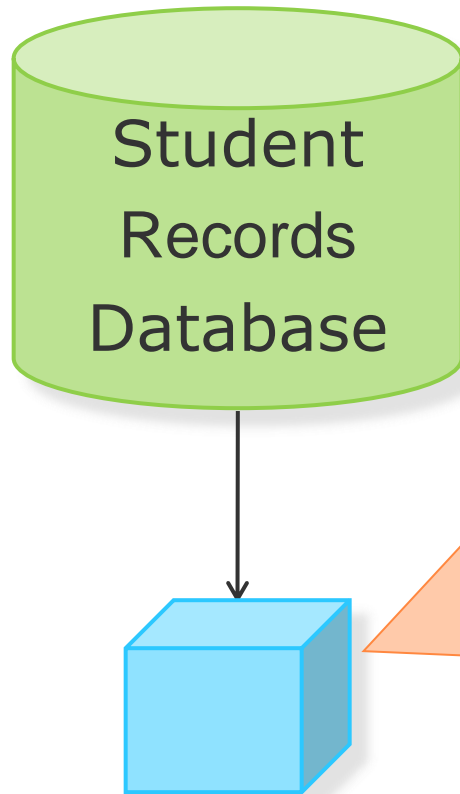


Step 1: Extract



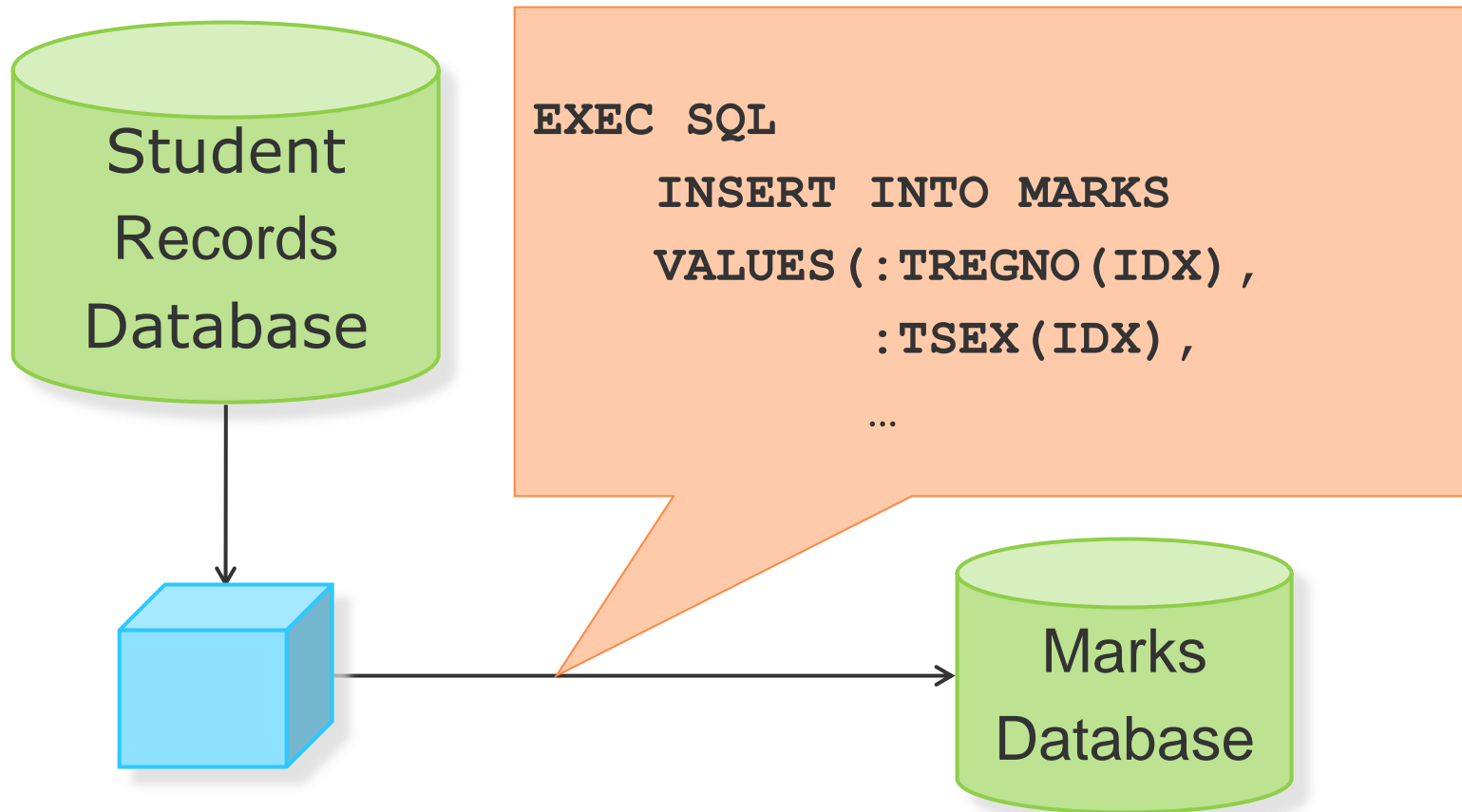
```
OBTAIN FIRST EXAM
MOVE EXAM-REGNO TO REGNO(1)
MOVE EXAM-CODE TO CODE(1)
MOVE EXAM-SESSION TO SESSION(
MOVE EXAM-MARK TO MARK(1)
MOVE 2 TO IDX
PERFORM UNTIL DB-REC-NOT-FOUND
OBTAIN NEXT EXAM
...
```

Step 2: Transform



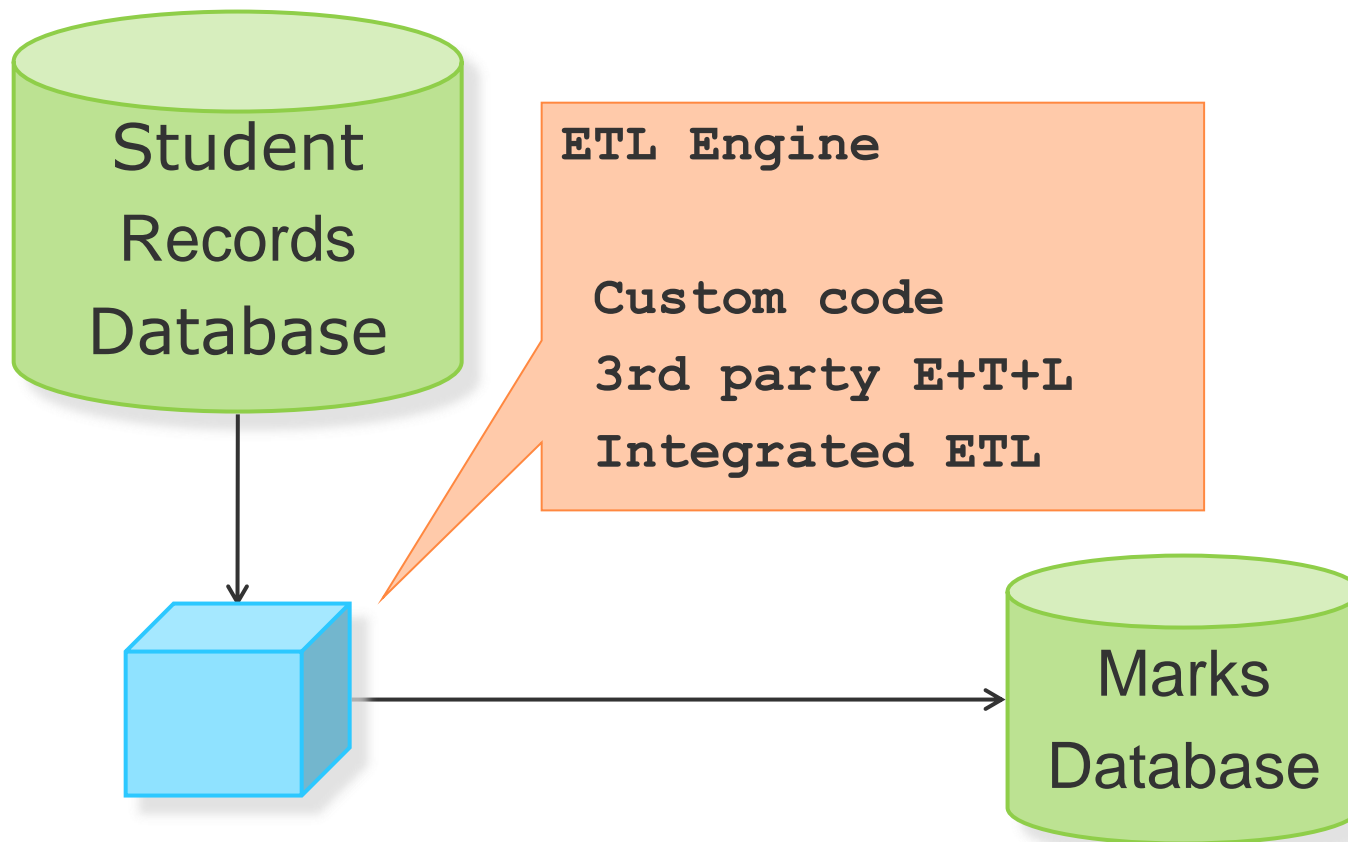
```
IF GENDER(IDX) = "M" OR "m"  
    MOVE "male" TO SEX(IDX)  
ELSE  
    IF GENDER(IDX) = "F" OR "f"  
        MOVE "female" TO SEX(IDX)  
    ELSE  
        LOG "Error: bad gender in  
        ...
```

Step 3: Load



ETL: Extract/Transform/Load

- Many commercial ETL tools now available



Migration/Replication Plan

Extract what?

Extract when?

Extract how?

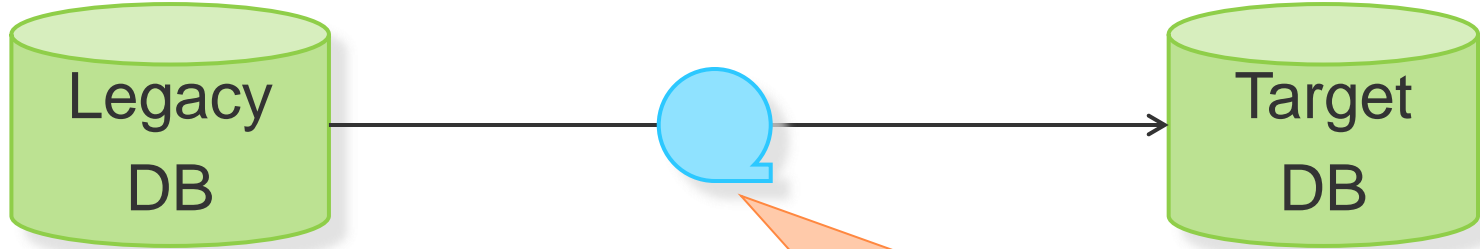
Transformations?

Transform where?

Freshness?

What to Extract?

Extract when?



Extraction Window
2.00am to 5.00am Mondays

Average of 1,000,000 records
to transfer @ 0.02sec/record
= 5 hrs 34 mins

Strategy?

One step

Snapshot +
updates

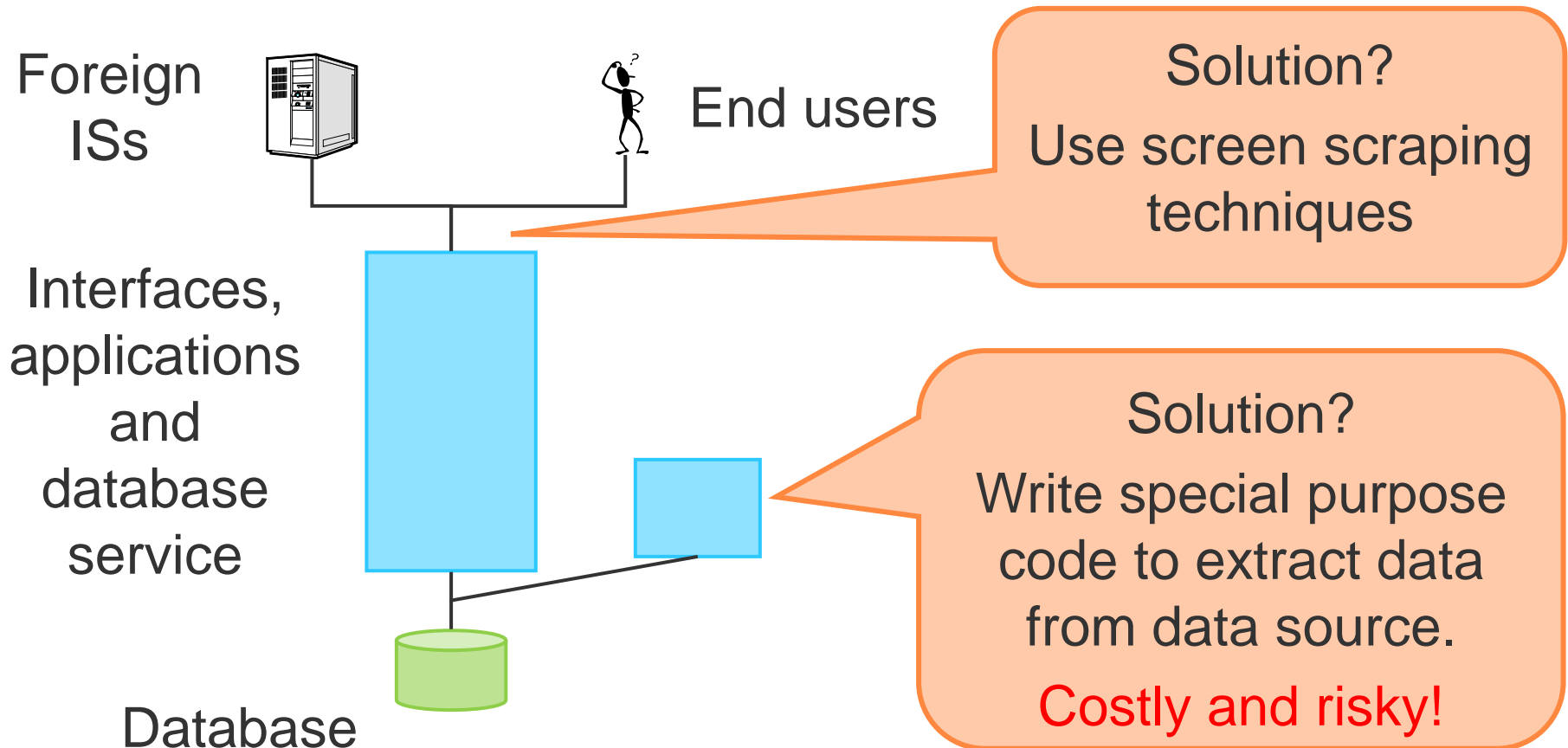
By domain
concept

Horizontal
partition

Zero downtime?

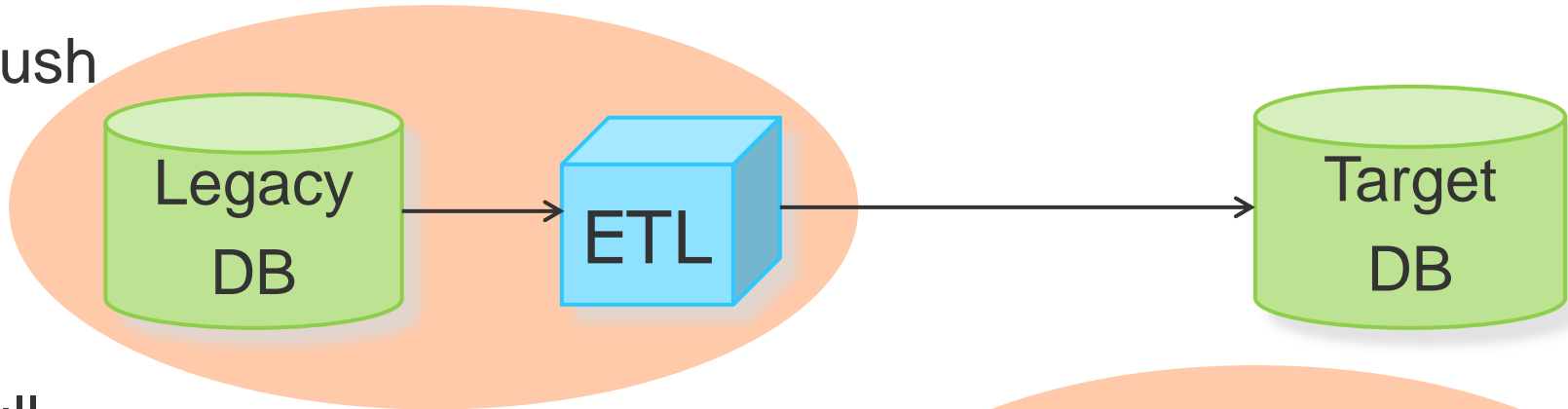
Extract how?

- Modern DBMSs have bulk import/export features
- Some legacy systems do not...

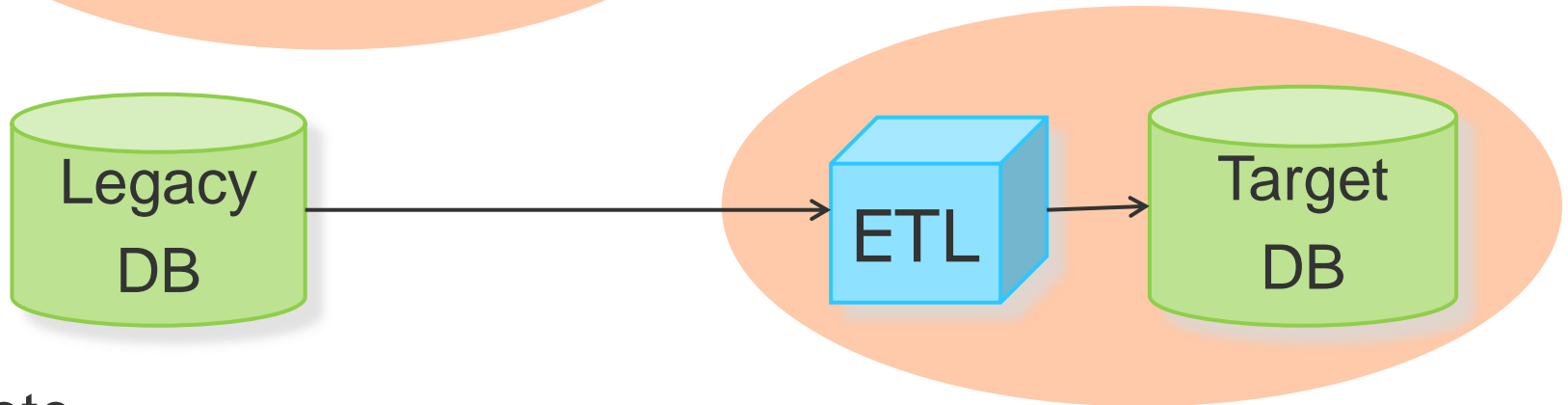


Transform where?

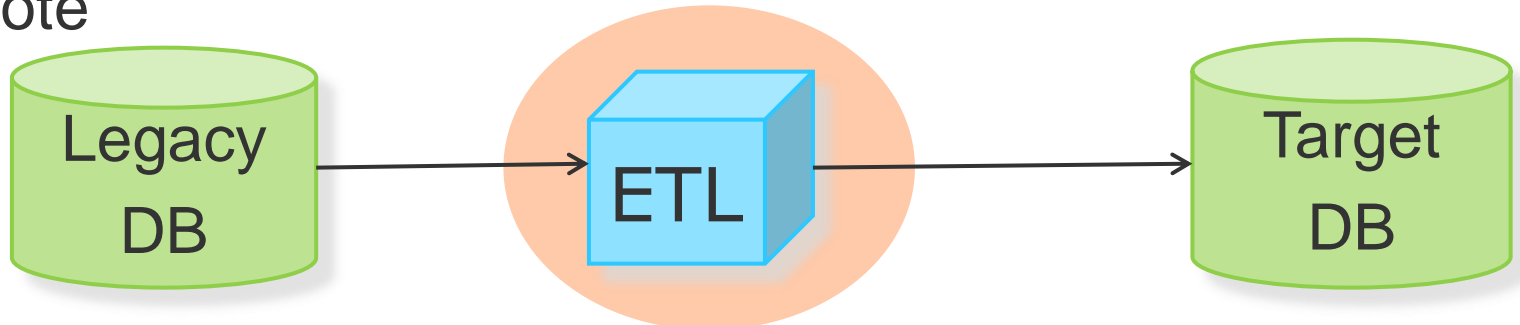
Push



Pull



Remote



Freshness

Frequency?

one-off

scheduled

triggered

on update

Depends on?

Capabilities

Change
pattern

Resources

Freshness
needs

Refresh Strategies

Snapshot

Deltas

Strategies?

Track

Timestamps

Logs

Snapshot
comparisons