Loading Type 2 SCDs in SSIS

Alex Whittles
Business Intelligence Consultant
Purple Frog

Alex@PurpleFrogSystems.com
www.PurpleFrogSystems.com
www.PurpleFrogSystems.com/blog
Twitter: @PurpleFrogSys

Accompanying Blog Post: http://www.purplefrogsystems.com/blog/2012/04/automating-t-sql-merge-to-load-dimensions-scd/

Alex Whittles

- Run SQL Server User Group in Birmingham
  www.SQLMidlands.com
- Regular speaker at SQLBits & User Groups
- Run **Purple Frog** BI Consultancy
- Studying MSc in Business Intelligence

Business Intelligence Consultancy

Dimensional Modelling	Data Quality
Data Warehousing	ETL Systems
OLAP Cubes	Reporting Systems

MSc Research Summary

"What's the best method available for loading Type 2 SCDs into a Kimball Data Warehouse using SSIS"

"Does the storage hardware effect the choice of method?"

Purple Frog®

What are SCDs?

- Historical data tracking in Data Warehouse Dimensions

Surrogate Key	Business Key (Customer ID)	Name	Address	IsRowCurrent	Valid From	ValidTo
1	123	Joe Bloggs	10 downing Street	0	10/04/2008	25/08/2010
2	123	Joe Bloggs	29 Acacia Road	0	25/08/2010	20/11/2011
3	123	Joe Bloggs	221b Baker Street	1	20/11/2011	

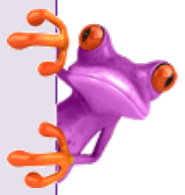
- Split incoming data into new data or changes
  - New Data: Insert
  - Change data: update & insert

Tests

- 50m rows of customer data
- 0%, 0.01%, 0.1%, 1%, 10% New records
- 0%, 0.01%, 0.1%, 1%, 10% Changed records
- Storage Hardware: Raid 10 HDD & FusionIO SSD
- 4 Different load Methods
- Each test run 3 times
- Rank each method within each test
- Statistically analyse rank & time

Purple Frog®

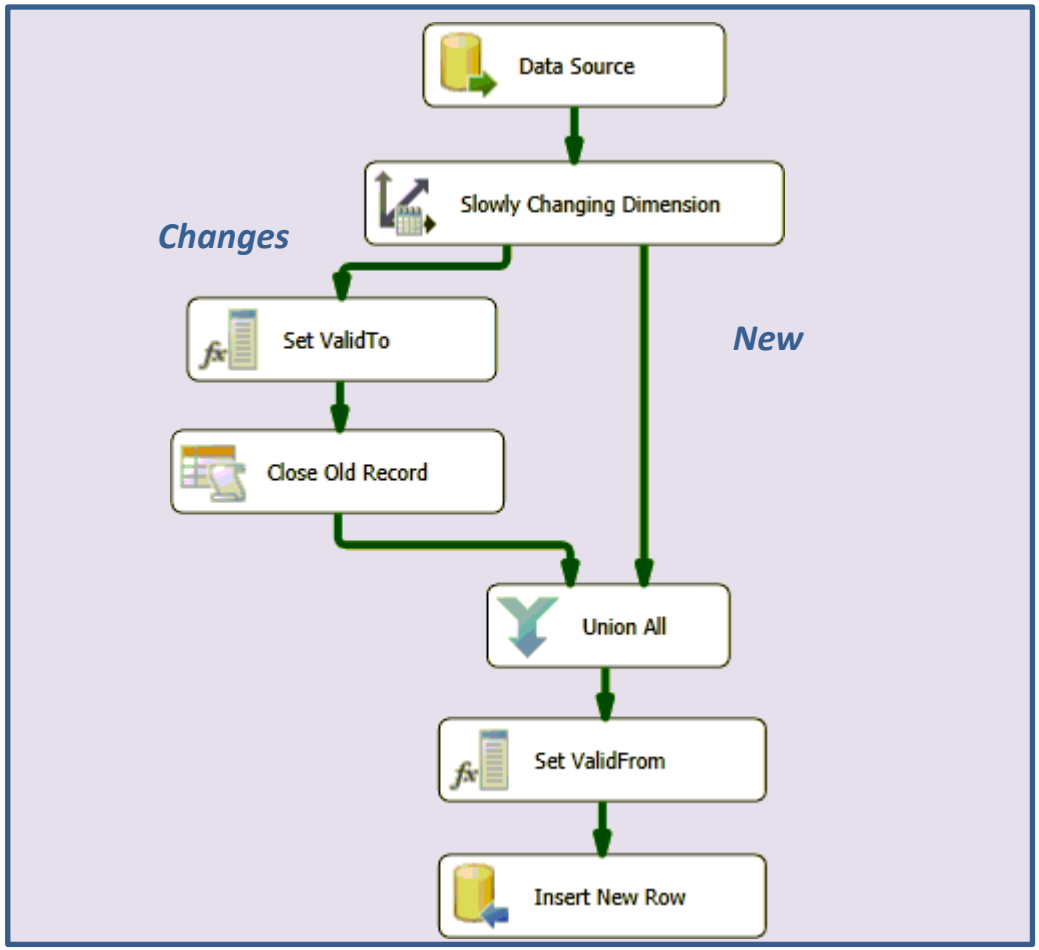
Methods



- SCD Wizard
- Lookup
- Merge Join
- T-SQL Merge

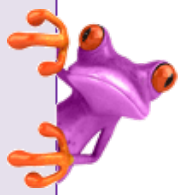
Methods – SCD Wizard

- **SCD Wizard**
- Lookup
- Merge Join
- T-SQL Merge

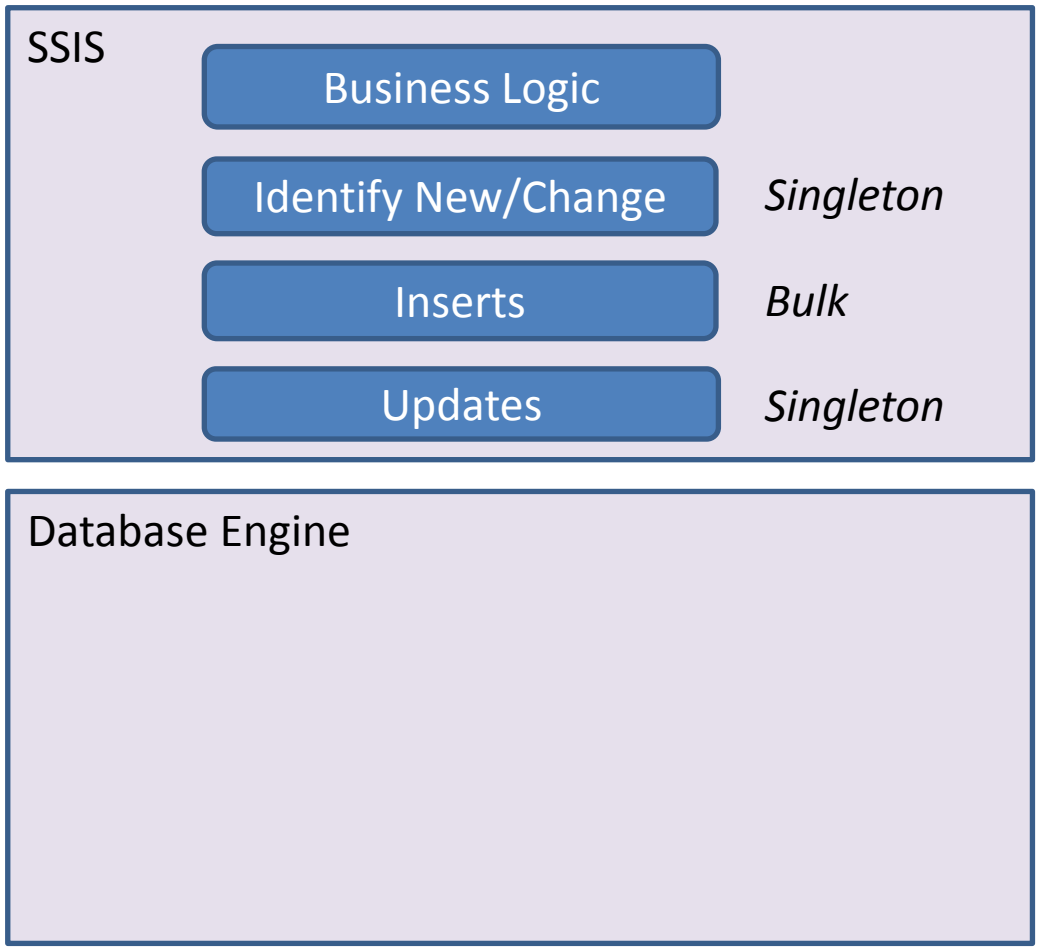


Purple Frog®

Methods – SCD Wizard



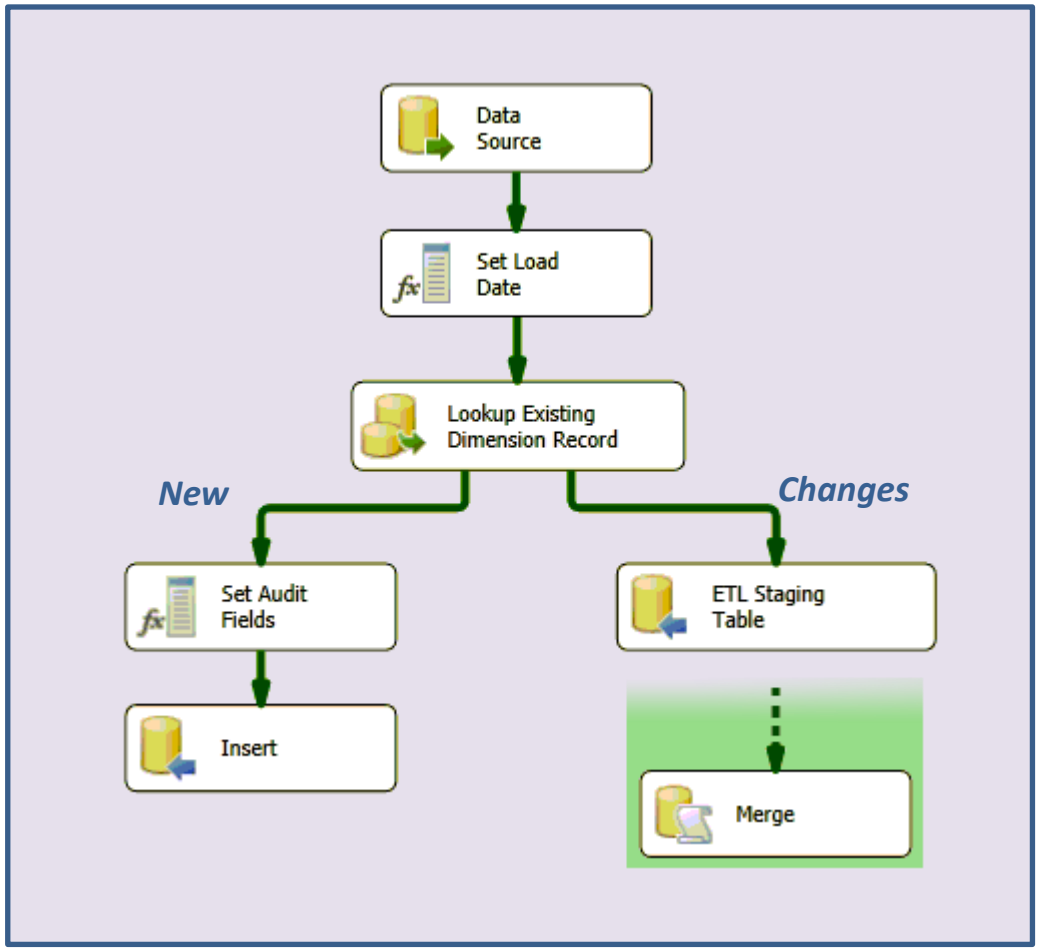
- **SCD Wizard**
- Lookup
- Merge Join
- T-SQL Merge



Purple Frog®

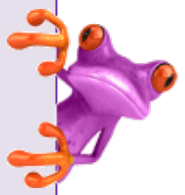
Methods – Lookup

- SCD Wizard
- **Lookup**
- Merge Join
- T-SQL Merge

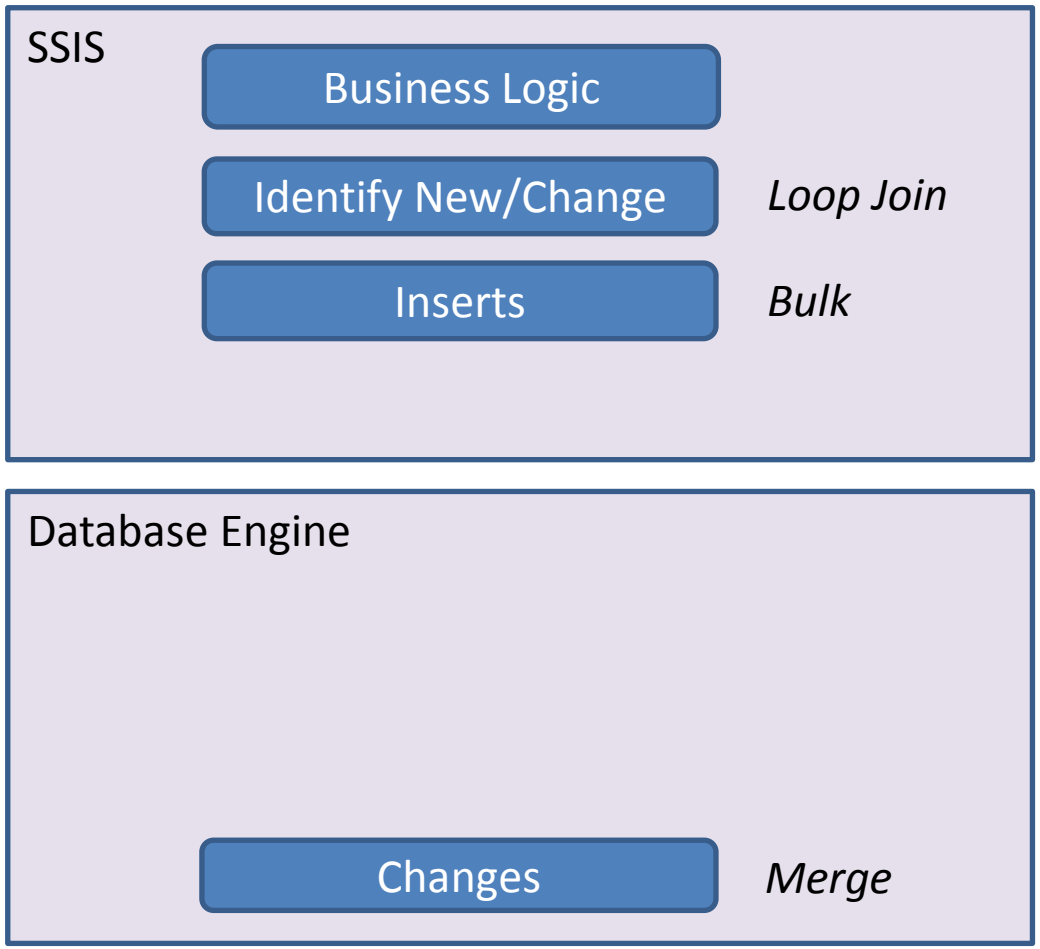


Purple Frog®

Methods – Lookup



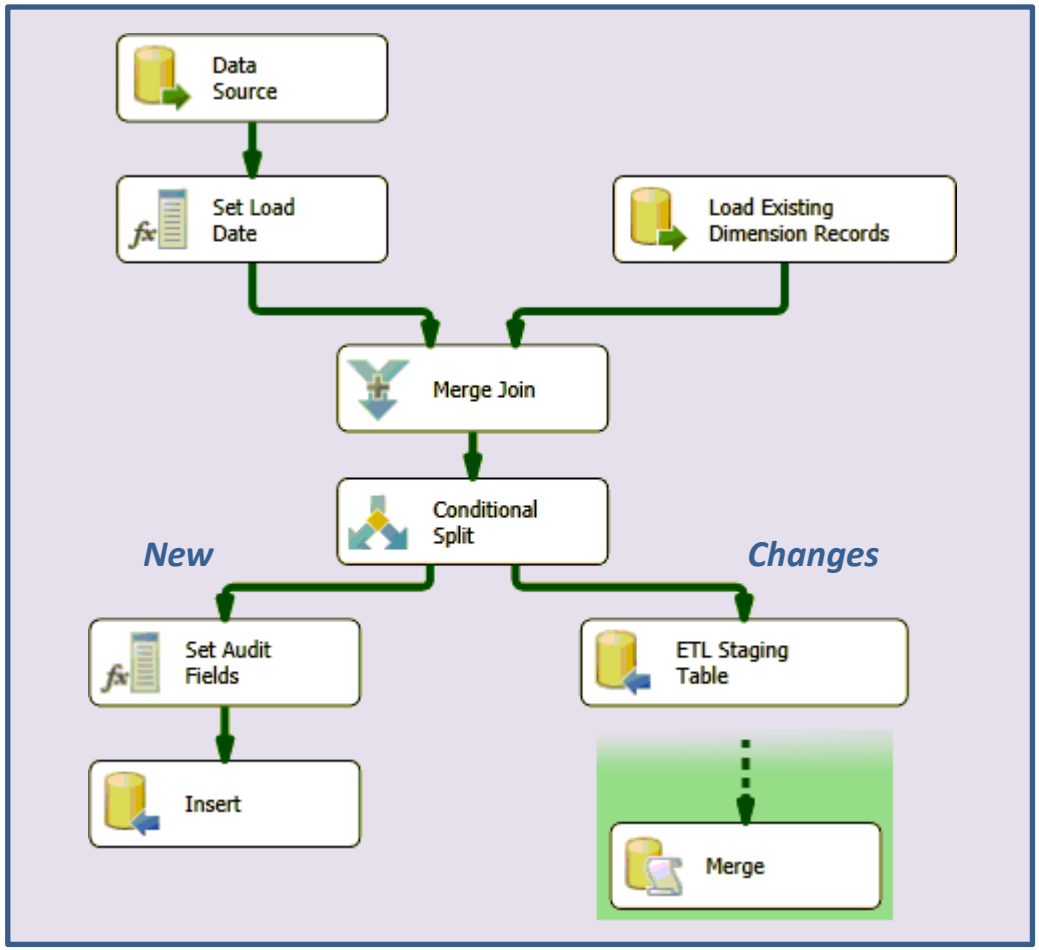
- SCD Wizard
- **Lookup**
- Merge Join
- T-SQL Merge



Purple Frog®

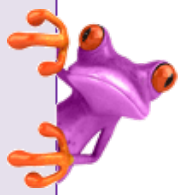
Methods – Merge Join

- SCD Wizard
- Lookup
- **Merge Join**
- T-SQL Merge

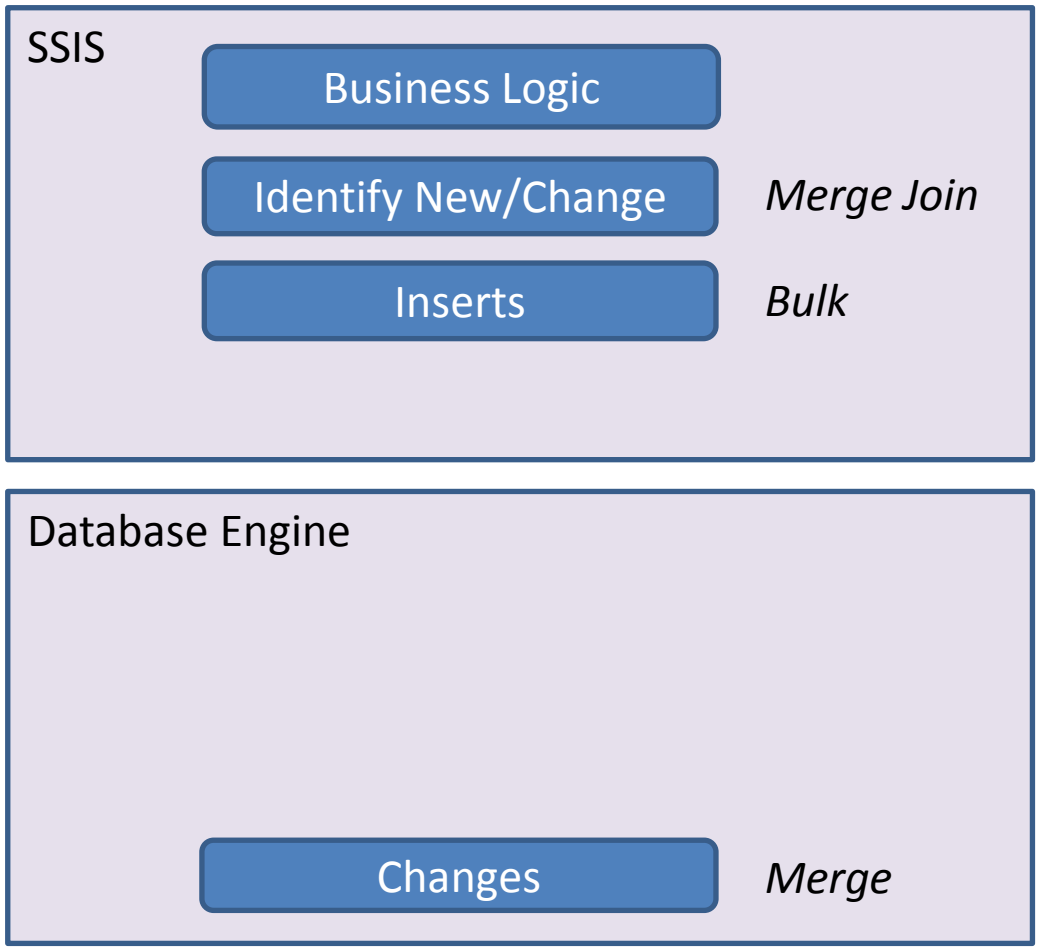


Purple Frog®

Methods – Merge Join



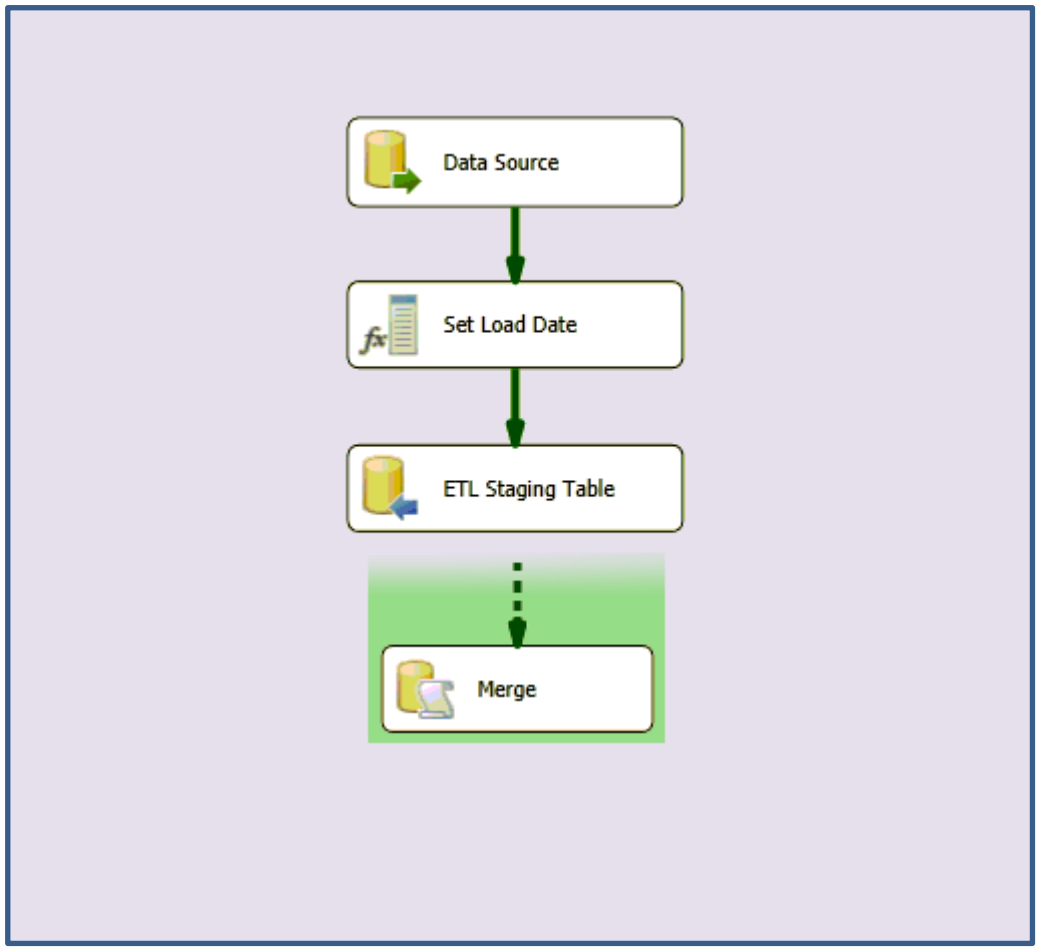
- SCD Wizard
- Lookup
- **Merge Join**
- T-SQL Merge



Purple Frog®

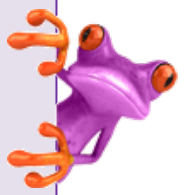
Methods – T-SQL Merge

- SCD Wizard
- Lookup
- Merge Join
- **T-SQL Merge**

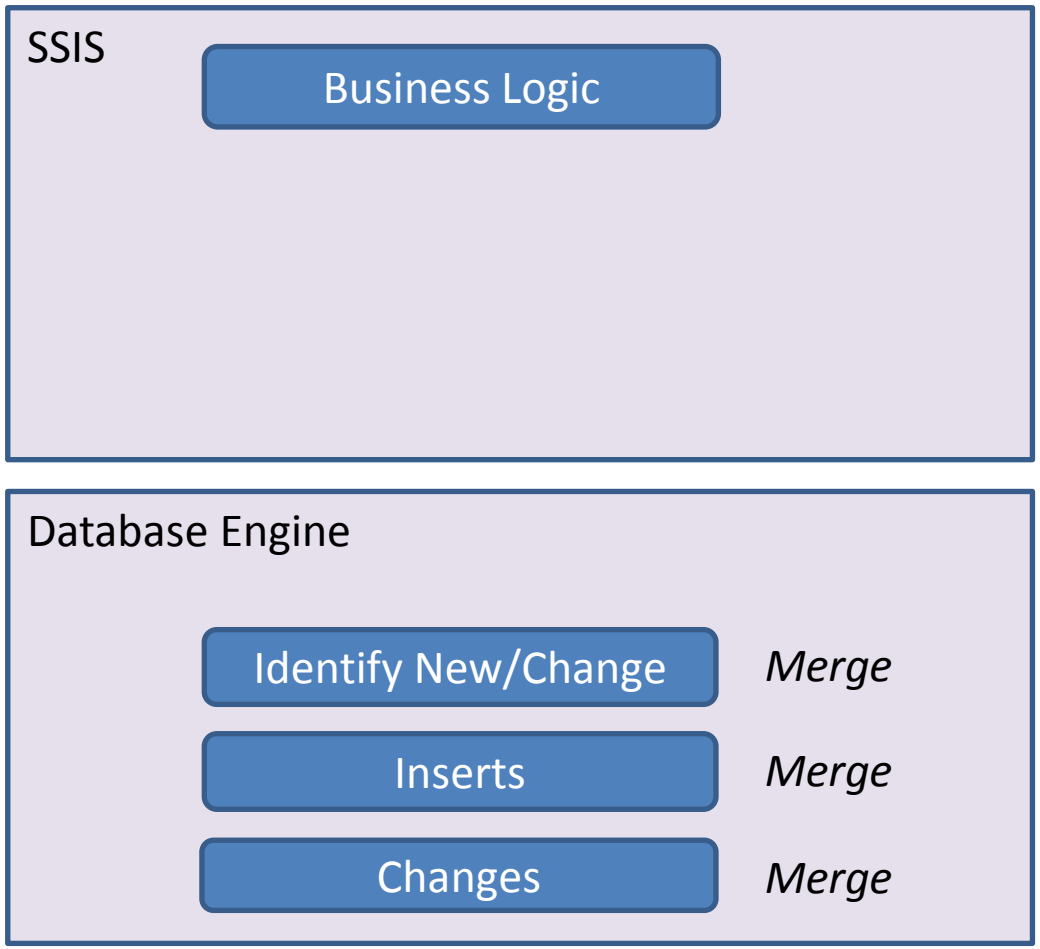


Purple Frog®

Methods – T-SQL Merge

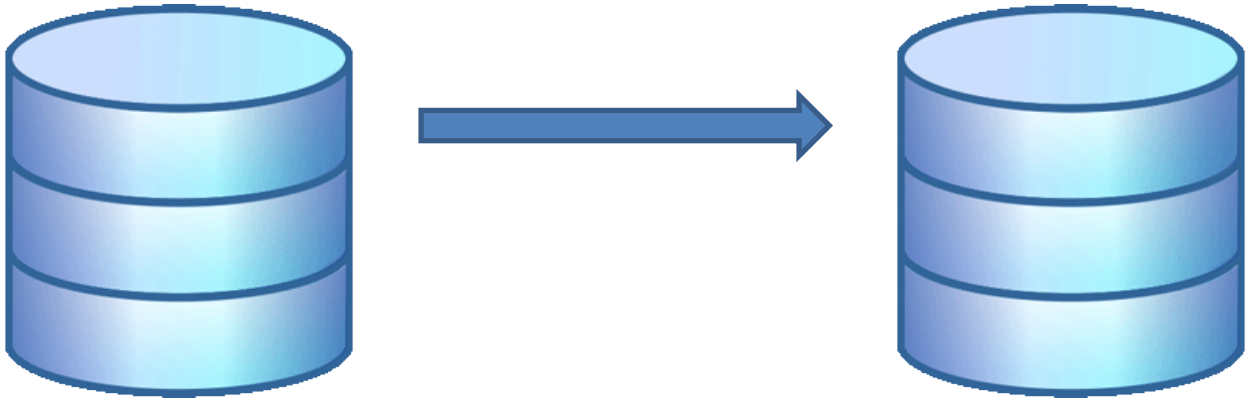


- SCD Wizard
- Lookup
- Merge Join
- **T-SQL Merge**



Purple Frog®

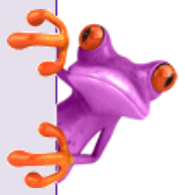
T-SQL Merge



New Data	Insert
Changed Data	Update
Both	Merge

Purple Frog®

T-SQL Merge



```
MERGE INTO [Destination]    Dest
    USING      [Source]          Src
        ON Src.[Key] = Dest.[Key]

WHEN MATCHED
   AND [Field] <> Src.[Field]
   THEN UPDATE SET
      [Field] = Src.[Field]

WHEN NOT MATCHED
   THEN INSERT
      ([Field1], [Field2])
   VALUES
      (Src.[Field1], Src.[Field2])
```

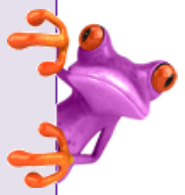
Set up Source & Destination

Perform Update

Perform Insert

Purple Frog®

T-SQL Merge



```
INSERT INTO [Destination]
   ([Field1], [Field2])
SELECT [Field1], [Field2] FROM (

  MERGE ......
    OUTPUT
        $ACTION Action_Out
        ,src.[Field1]
        ,src.[Field2]

) AS MergeOut
WHERE MergeOut.Action_Out = 'UPDATE'
```

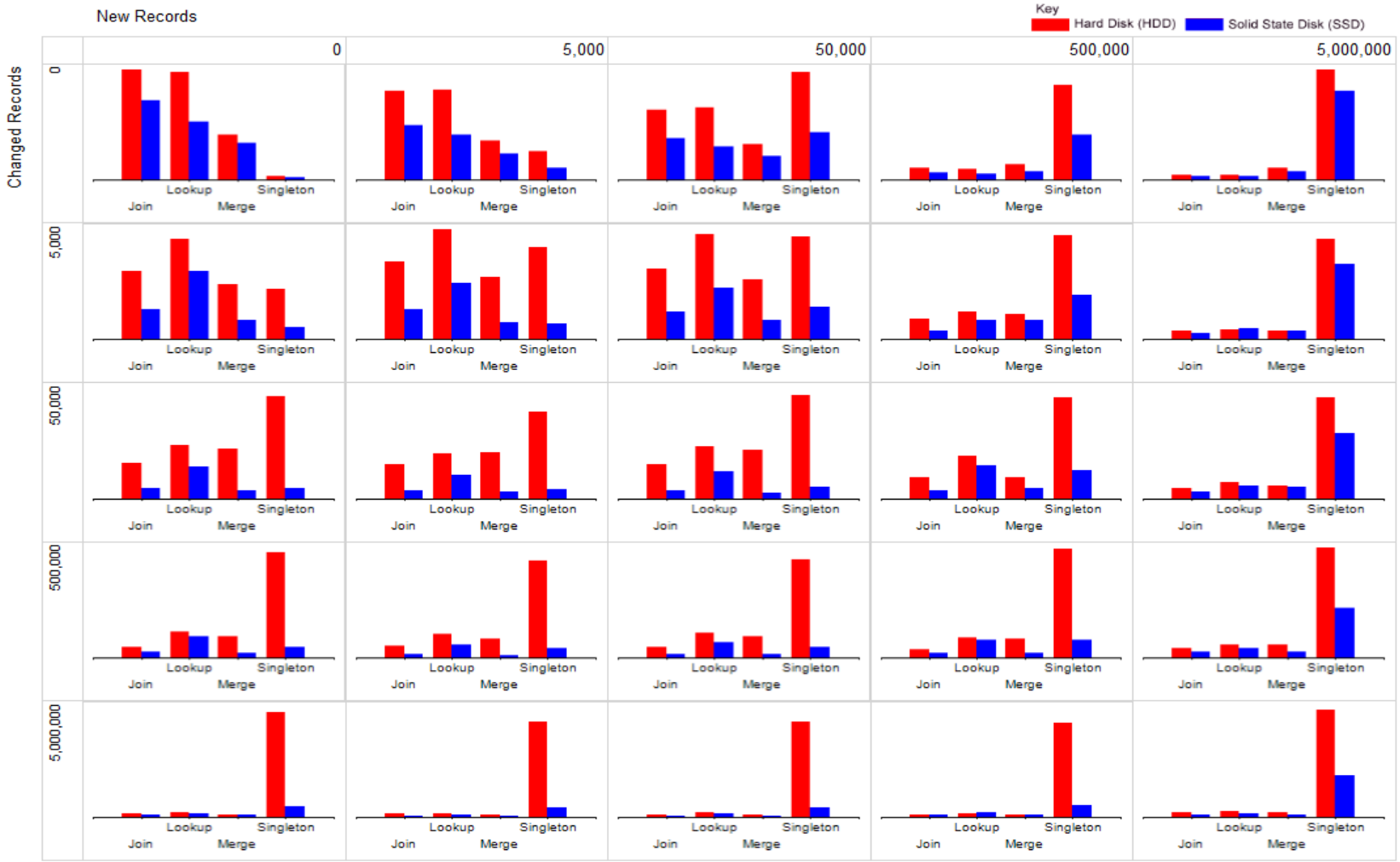
Insert New Record

Perform Merge

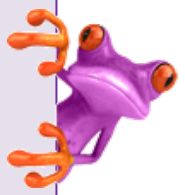
Capture Updates

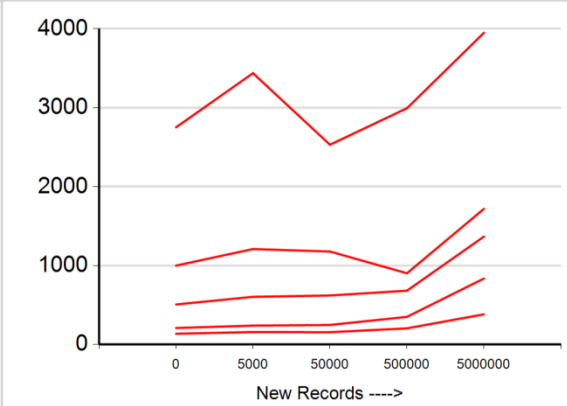
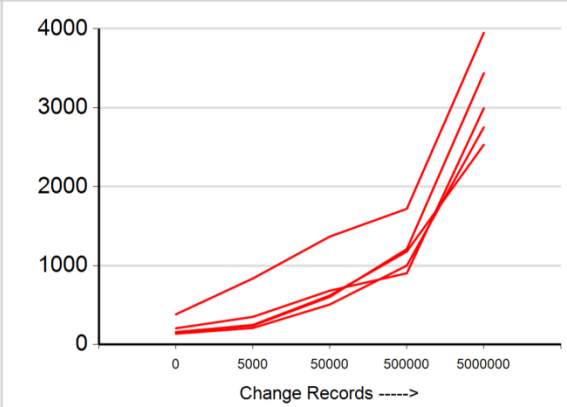
Purple Frog®

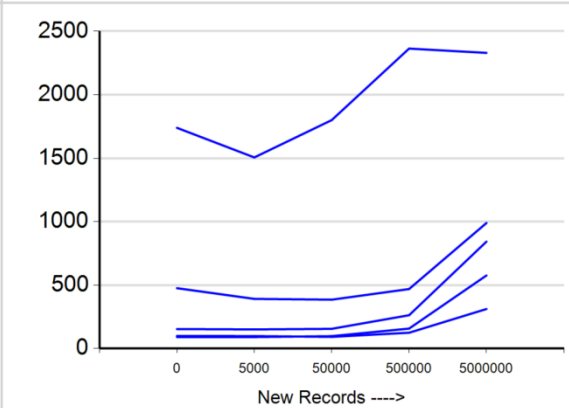
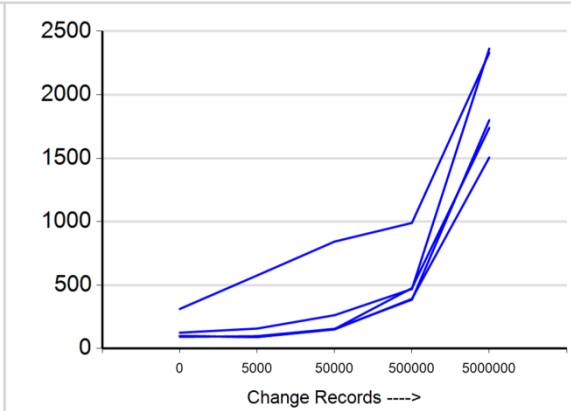
Results


New Records
Key
Hard Disk (HDD)
Solid State Disk (SSD)
Changed Records
0
5,000
50,000
500,000
5,000,000
Join
Lookup
Merge
Singleton

Results

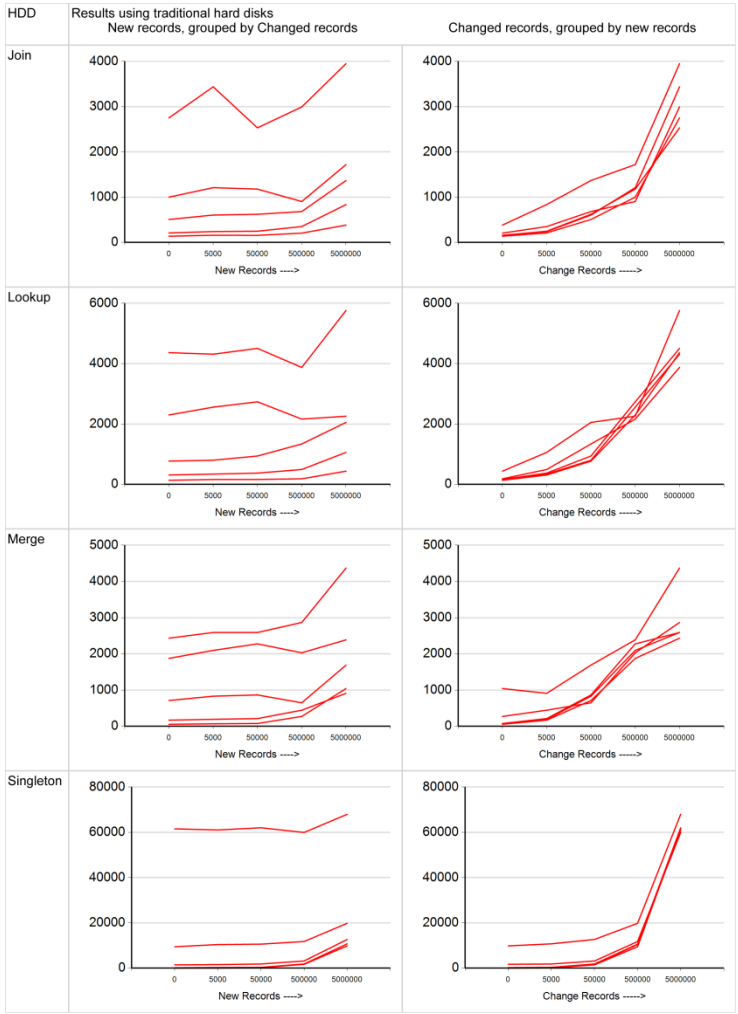


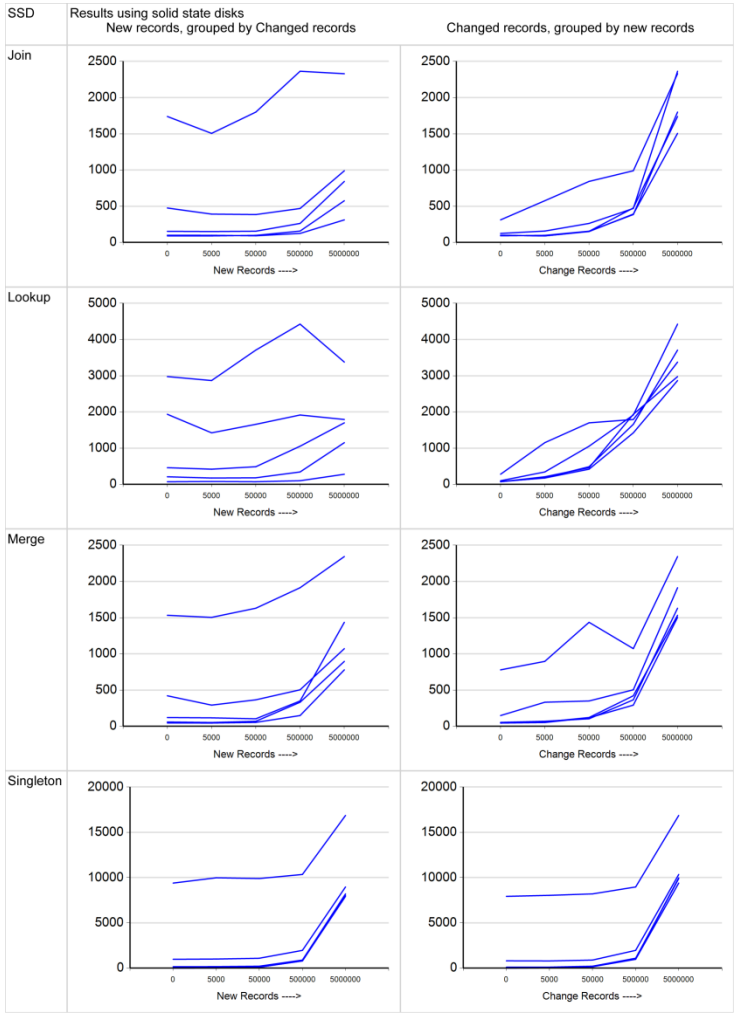
HDD	Results using traditional hard disks New records, grouped by Changed records	Changed records, grouped by new records
Join		

SSD	Results using solid state disks New records, grouped by Changed records	Changed records, grouped by new records
Join		

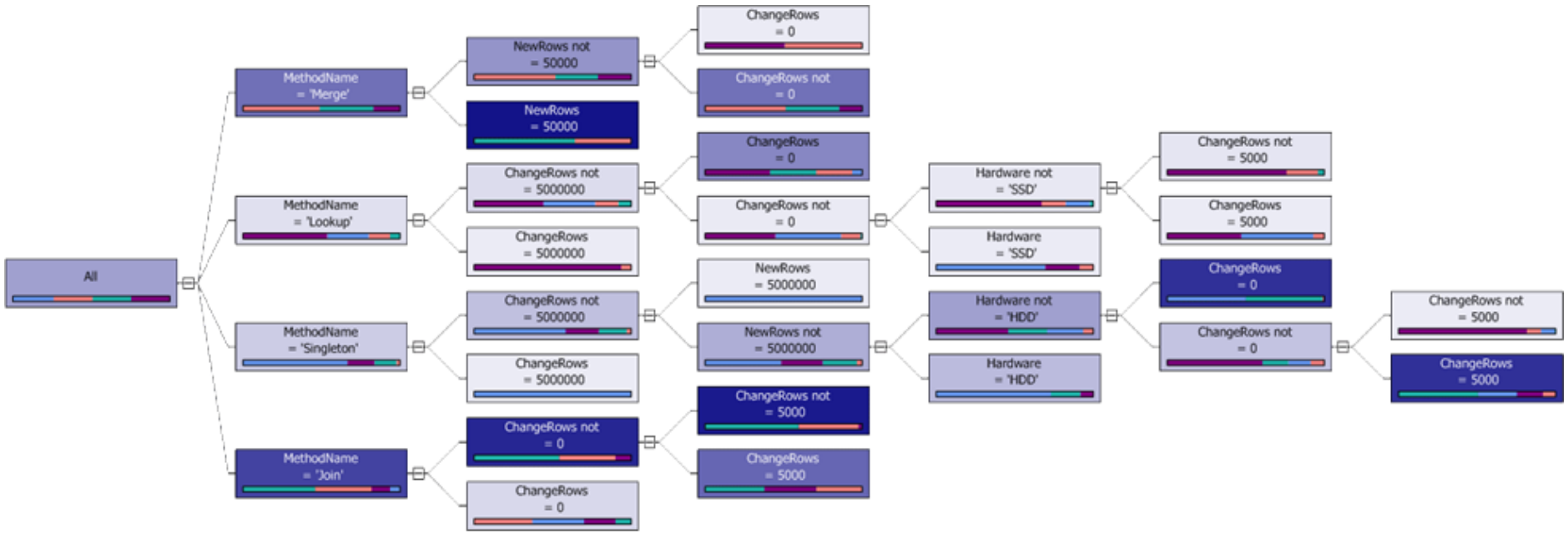
Purple Frog®

Results


HDD
Results using traditional hard disks
New records, grouped by Changed records
Changed records, grouped by new records
Join
Lookup
Merge
Singleton
New Records ---->
Change Records ---->


SSD
Results using solid state disks
New records, grouped by Changed records
Changed records, grouped by new records
Join
Lookup
Merge
Singleton
New Records ---->
Change Records ---->

Decision Tree



Key

Rank 1
Rank 2
Rank 3
Rank 4

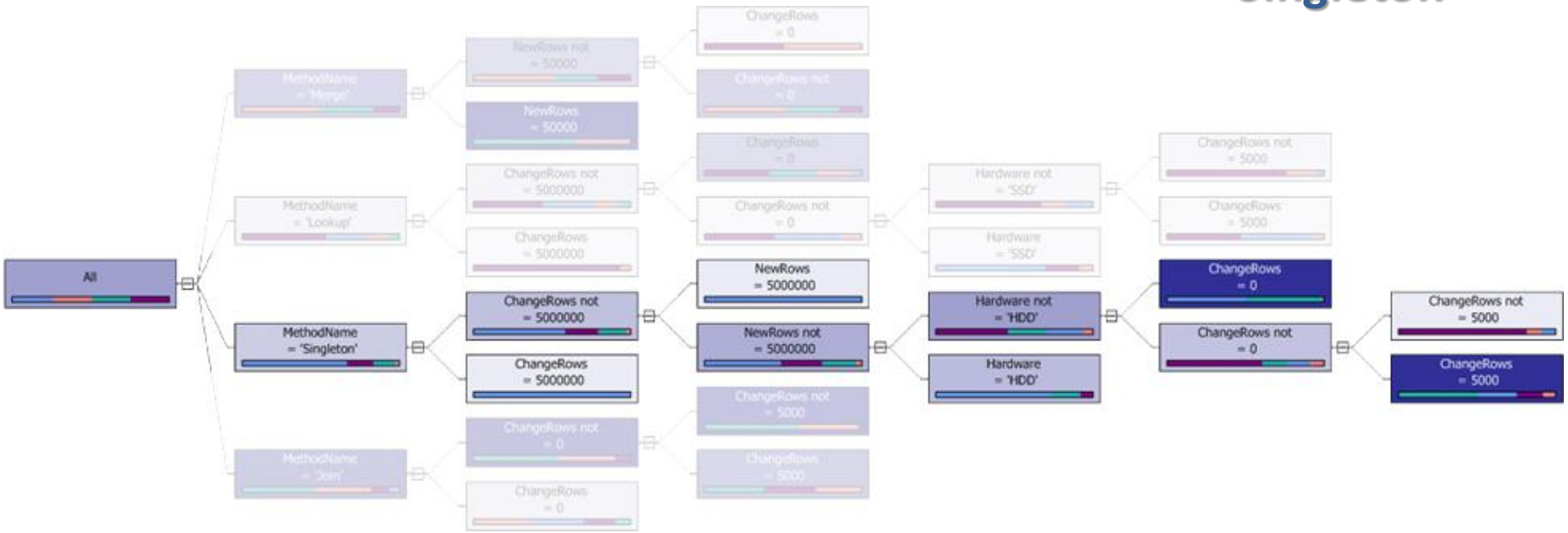
High — Low

Probability of being Rank 1

Purple Frog®

Decision Tree

Singleton



Key

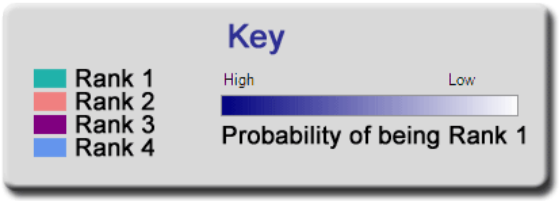
Rank 1
Rank 2
Rank 3
Rank 4

High — Low

Probability of being Rank 1

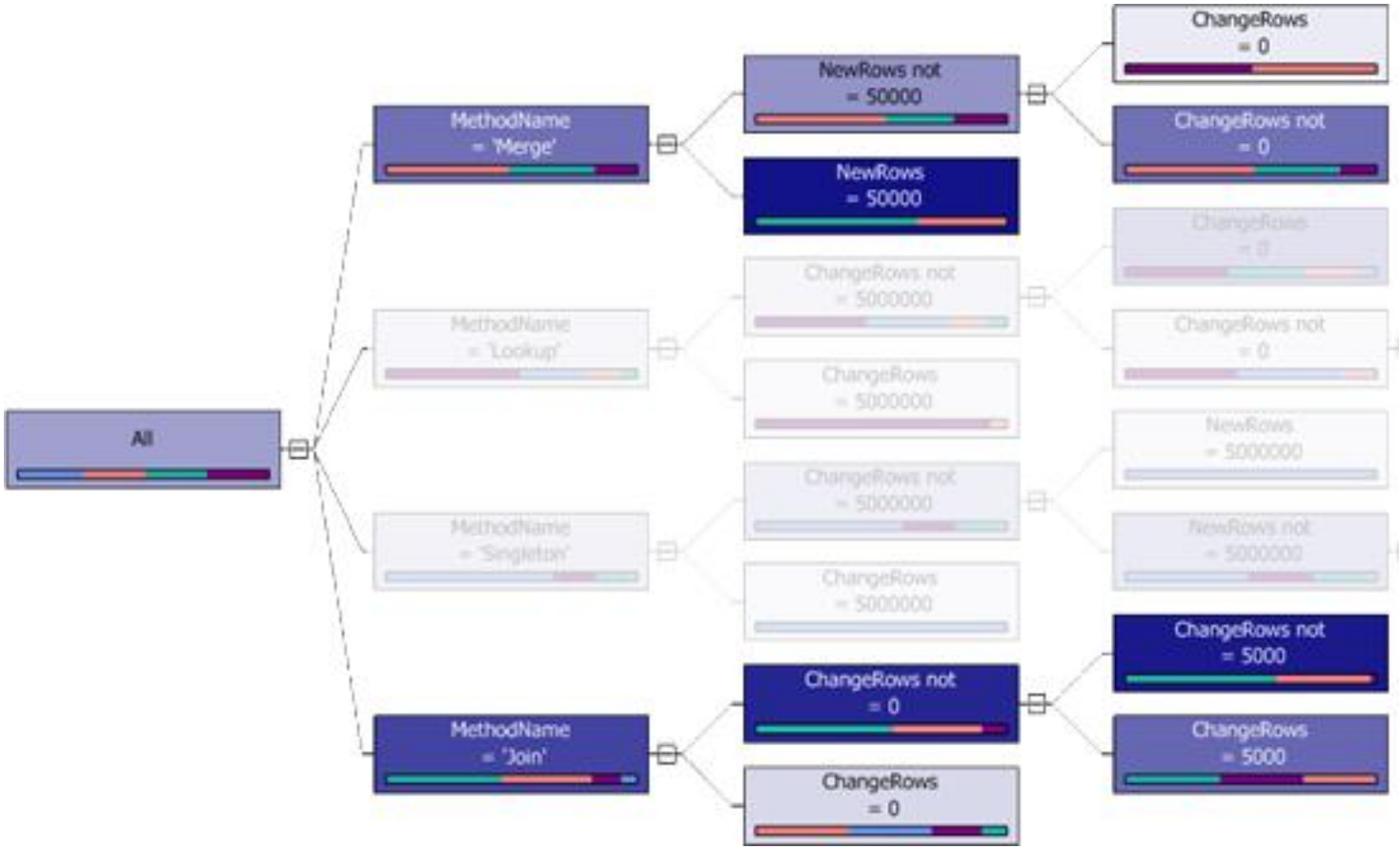
Purple Frog®

Decision Tree



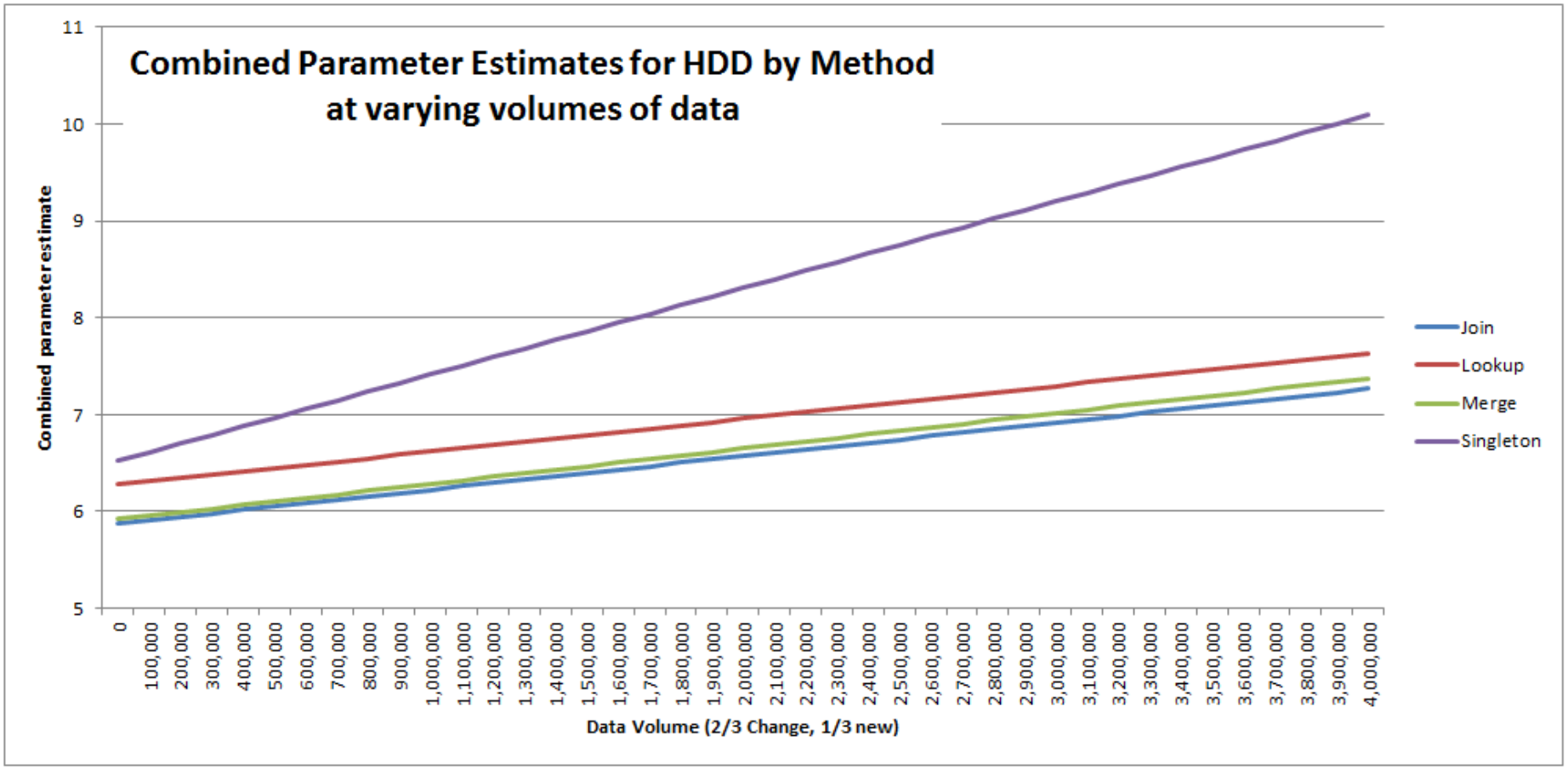
Purple Frog®

Decision Tree



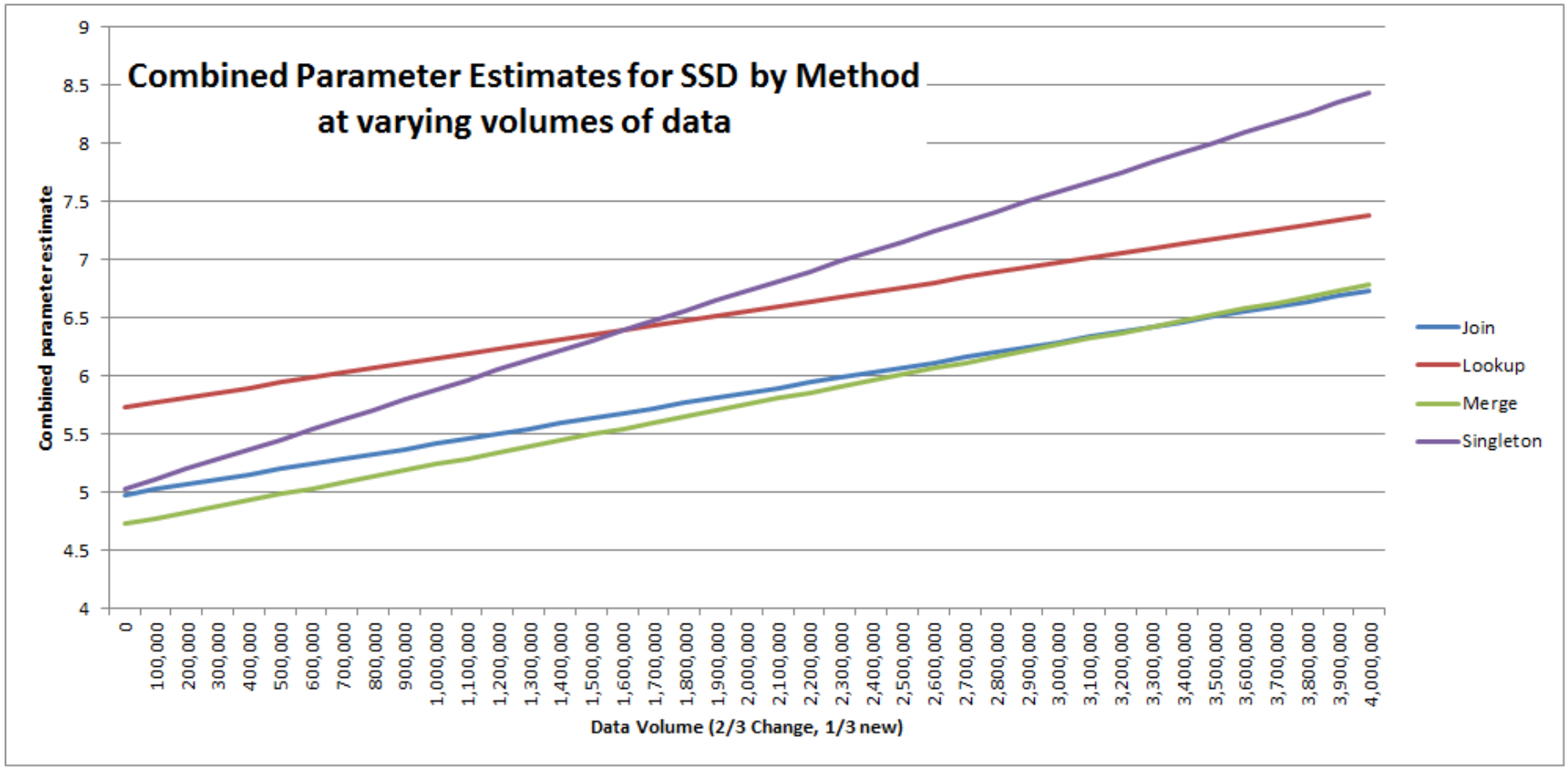
Purple Frog®

Regression Model - HDD



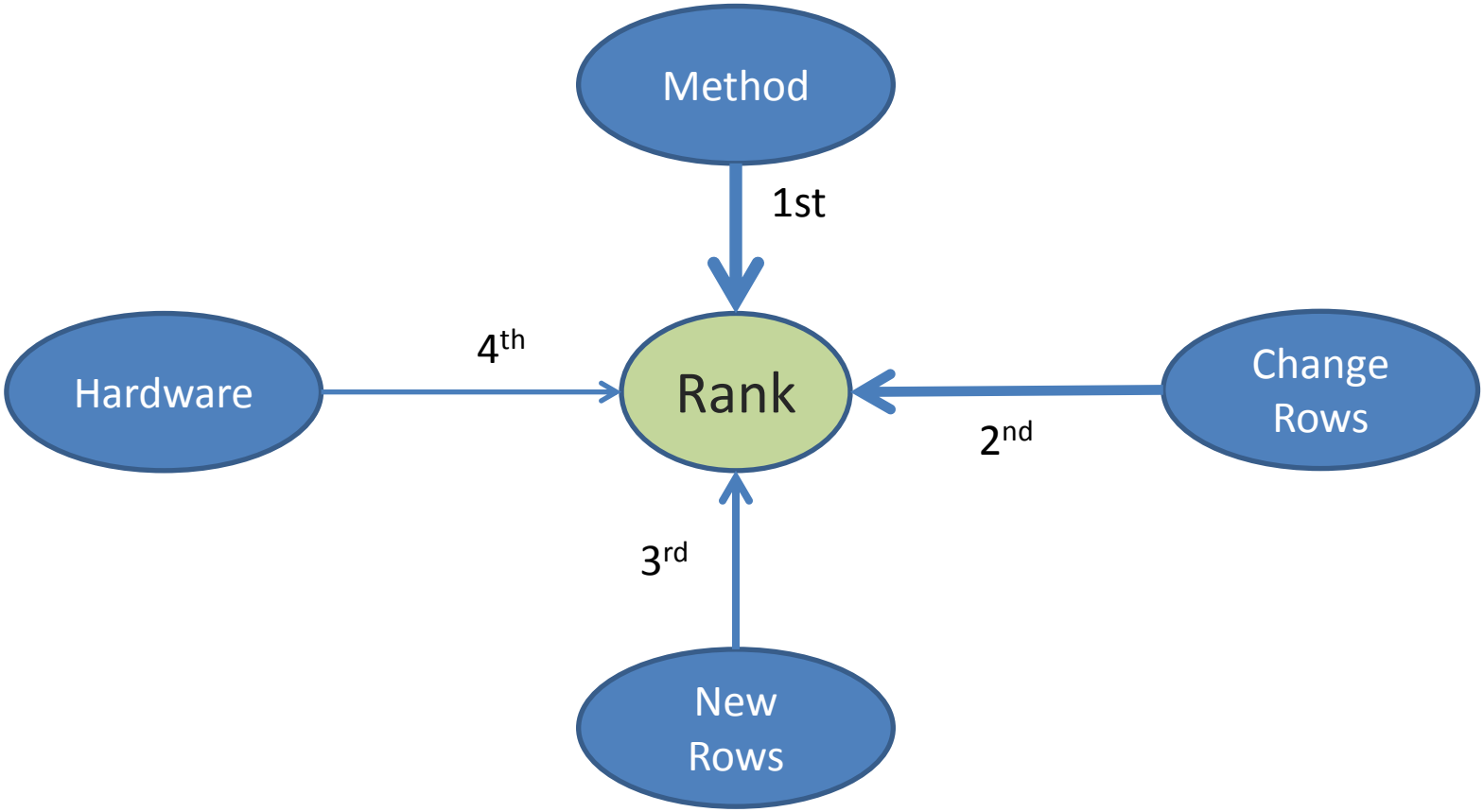
Purple Frog®

Regression Model - SSD



Purple Frog®

Key Influencers



Purple Frog®

Results - Hardware

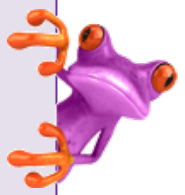
- Hardware does not influence the best choice of method
- SSD can reduce load time by up to 92% (12x)
  - Up to 92% (12x) for SCD Wizard
  - Up to 91% (11x) with T-SQL Merge
  - Up to 78% (5x) with Merge Join
  - Up to 67% (3x) with Lookup
- Most significant difference for
  - Singleton
  - Changes
- Service Orientated Architecture – real time messages

Purple Frog®

Summary

- Avoid SCD Wizard unless low volumes or prototype on SSD
- Little difference between other 3 methods
  - Lookup is the worst
  - Merge Join & T-SQL Merge statistically equivalent
  - My preference is T-SQL Merge

!! What isn't covered !!



- Type 0, 1, 3 or 6 SCDs
- Performance of data source
- CDC
- Different sizes of dimension
  - Columns & Rows
- Different methods
  - 3rd party components
  - Checksums
- Hardware
  - Memory, CPU, SAN, 15k drives, Fusion IO Octal
- SSIS & SQL Server on separate servers

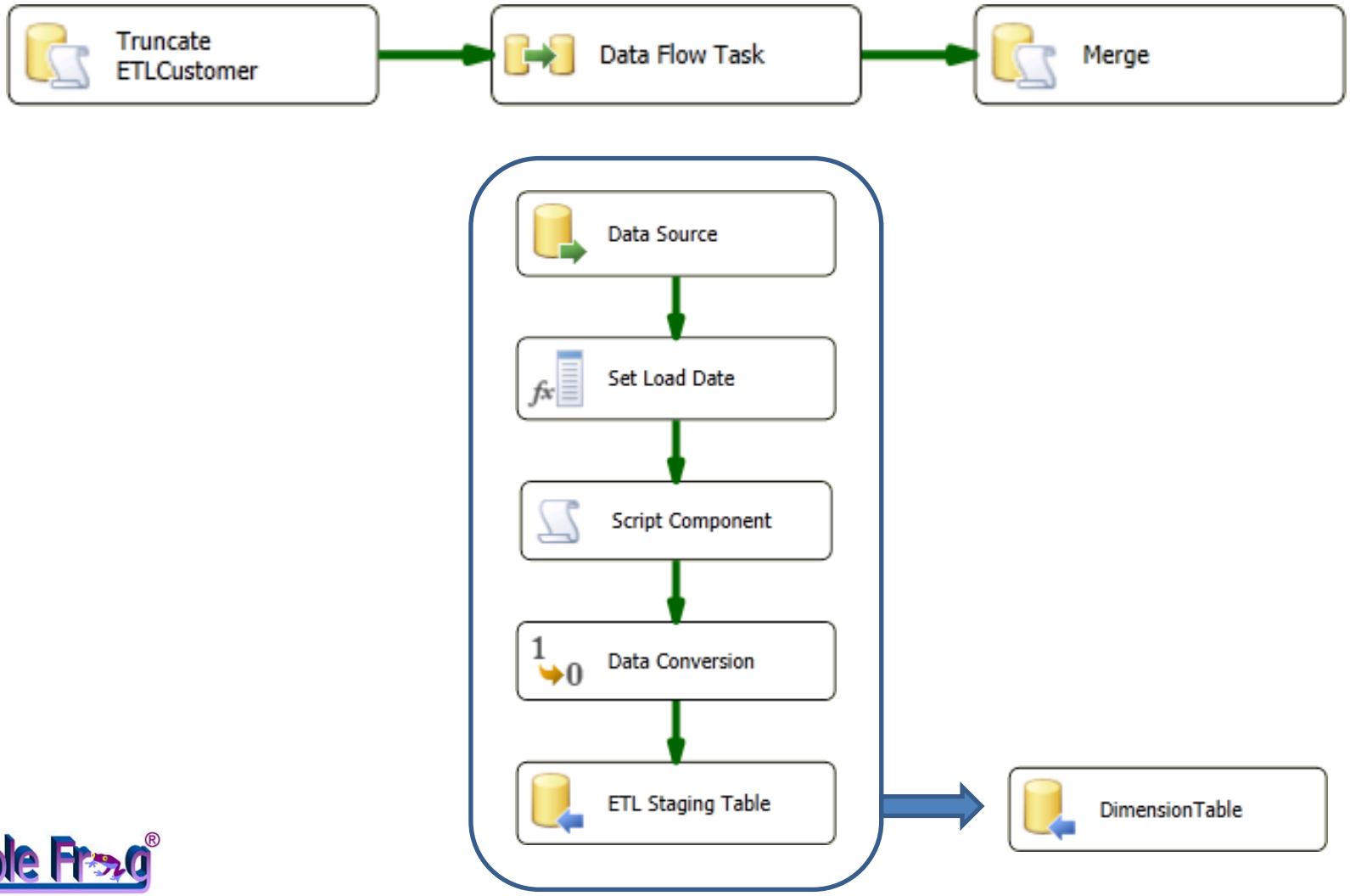
Purple Frog®

Demo

MERGE MERGE MERGE
MERGE MERGE MERGE
MERGE MERGE MERGE

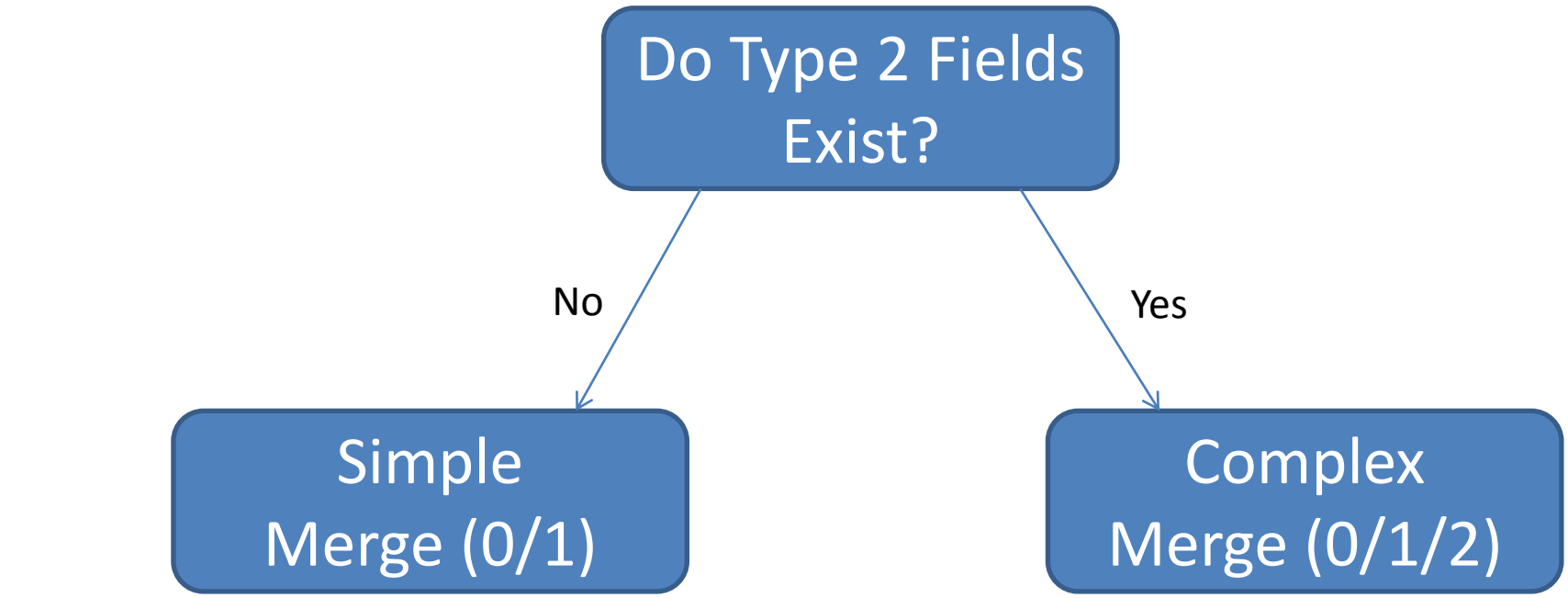
Purple Frog®

Merge



Purple Frog®

Automating Merge



Purple Frog®

Automating Merge (Type 0/1)

```
MERGE     [DIMENSION TABLE] as Target
  USING   [STAGING TABLE]     as Source
     ON   [LIST OF BUSINESS KEY FIELDS]
    AND   IsRowCurrent=1
  WHEN MATCHED AND
        Target.[LIST OF TYPE 1 FIELDS] <> Source.[LIST OF TYPE 1 FIELDS]
    THEN UPDATE SET
        [LIST OF TYPE 1 FIELDS] = Source.[LIST OF TYPE 1 FIELDS]
   WHEN NOT MATCHED THEN INSERT
        [LIST OF ALL FIELDS]
    VALUES
        Source.[LIST OF ALL FIELDS]
```

Purple Frog®

Loading Type 2 SCDs in SSIS

Alex@PurpleFrogSystems.com
Twitter: @PurpleFrogSys
www.PurpleFrogSystems.com
www.PurpleFrogSystems.com/blog