Benchmarking ETL Workflows

Alkis Simitsis¹, Panos Vassiliadis², Umeshwar Dayal¹, Anastasios Karagiannis², Vasiliki Tziovara²

¹ Intelligent Information Management Lab
   Hewlett-Packard Labs
   Palo Alto, CA, USA

² University of Ioannina
   Greece

alkis@hp.com

presented by Kevin Wilkinson¹

© 2008 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice
ETL workflows
ETL Tools

• Commercial
  – Ab Initio
  – SAP Business Objects
  – IBM WebSphere Information Integration
  – Informatica PowerCenter
  – Microsoft SSIS
  – Oracle Warehouse Builder
  – Pervasive
  – SAS Data Integration Studio

• Open Source
  – Clover
  – Pentaho Kettle
  – Talend
## ETL Tools

### Commercial

1. ActaWorks, Acta Technologies
2. Actem(TM), Actsoft
3. ASAP, ASAP Technologies
4. TextPro, TextPro Software
5. iMerge, iMerge Technologies
6. iMatrix, iMatrix Technologies
7. iLoader, iData Technologies
8. iLoader, Data Management Technologies
9. iMigrator, iMigrator Technologies
10. iJoin, iJoin Technologies
11. iRepository, iRepository Technologies
12. iSelect, iSelect Technologies
13. iSource, iSource Technologies
14. iTarget, iTarget Technologies
15. iTarget, iTarget Technologies
16. iView, iView Technologies
17. iView, iView Technologies
18. iView, iView Technologies
19. iView, iView Technologies
20. iView, iView Technologies
21. iView, iView Technologies
22. iView, iView Technologies
23. iView, iView Technologies
24. iView, iView Technologies
25. iView, iView Technologies
26. Data Extender, DogHouse Enterprises
27. Data Flow Manager, Peter's Software
28. Data Integration, Data Integration Technologies
29. Data Load, Data Load Technologies
30. Data Manager, Applied Database Technology
31. Data Migration Tools, Friedman & Associates
32. DataPower, DataPower Technologies
33. Data Propagation System, Treehouse Software
34. Data Warehouse Tools, Javacorporate
35. Data3, Inform Information Systems
36. DataBlaster 2, Bus-Tech, Inc.
37. DataExchanger SRV, CrossDataBase Technology
38. DataConvert, Metadata Information Partners
39. DataDragon, Dulcian, Inc.
40. DataExchanger SRV, CrossDataBase Technology
41. DataExchanger SRV, CrossDataBase Technology
42. DataImport, Spalding Software
43. DataLever, Data Lever
44. DataPipe, Crystal Software
45. DataManager, Joe Spanicek
46. DataPipe, Crystal Software
47. DataPipe, Crystal Software
48. DataPipe, Crystal Software
49. DataPipe, Crystal Software
50. DataPipe, Crystal Software
51. DataPipe, Crystal Software
52. DataPipe, Crystal Software
53. DataPipe, Crystal Software
54. DataPipe, Crystal Software
55. DataStage XE, Ascential Software
56. DataSuite, Pathlight Data Systems
57. DataSuite, Pathlight Data Systems
58. DataSuite, Pathlight Data Systems
59. DataStage XE, Ascential Software
60. DataStage XE, Ascential Software
61. DBMS/Copy, Conceptual Software, Inc.
62. DECISIVE Advantage, InfoSAGE, Inc.
63. DECISIVE Advantage, InfoSAGE, Inc.
64. DecisionBase, Computer Associates
65. DecisionBase, Computer Associates
66. DecisionBase, Computer Associates
67. DecisionBase, Computer Associates
68. DecisionBase, Computer Associates
69. Distribution Agent for MVS, Sybase
70. Distribution Agent for MVS, Sybase
71. DQ Now, DQ Now
72. DQtransform, Metagon Technologies
73. DT/Studio, Embarcadero Technologies
74. DTS, Microsoft
75. eWorker Portal, eWorker Legacy, entrinsic.com
76. eWorker Portal, eWorker Legacy, entrinsic.com
77. eWorker Portal, eWorker Legacy, entrinsic.com
78. e-Sense Gather, Vigil Technologies
79. ETI Extract, Evolutionary Technologies, Inc.
80. EZ-Pickin's, ExcelSystems
81. EZ-Pickin's, ExcelSystems
82. eWorker Portal, eWorker Legacy, entrinsic.com
83. eWorker Portal, eWorker Legacy, entrinsic.com
84. eWorker Portal, eWorker Legacy, entrinsic.com
85. eWorker Portal, eWorker Legacy, entrinsic.com
86. eWorker Portal, eWorker Legacy, entrinsic.com
87. eWorker Portal, eWorker Legacy, entrinsic.com
88. eWorker Portal, eWorker Legacy, entrinsic.com
89. eWorker Portal, eWorker Legacy, entrinsic.com
90. eWorker Portal, eWorker Legacy, entrinsic.com
91. Fusion FTMS, Proginet
92. Fusion FTMS, Proginet
93. Fusion FTMS, Proginet
94. Fusion FTMS, Proginet
95. Fusion FTMS, Proginet
96. Graphical Performance Series, Vanguard Solutions
97. Graphical Performance Series, Vanguard Solutions
98. HIREL, SWS Software Services
99. Hummingbird ETL, Hummingbird Ltd
100. Hummingbird ETL, Hummingbird Ltd
101. iMerge, iMerge Technologies
104. iPointOut, mSE GmbH
105. iPowerLoad, iPowerware Technologies
106. iPowerLoad, iPowerware Technologies
107. iPowerLoad, iPowerware Technologies
108. iPowerLoad, iPowerware Technologies
109. iPowerLoad, iPowerware Technologies
110. iPowerLoad, iPowerware Technologies
111. iPowerLoad, iPowerware Technologies
112. iPowerLoad, iPowerware Technologies
113. iPowerLoad, iPowerware Technologies
114. iPowerLoad, iPowerware Technologies
115. iPowerLoad, iPowerware Technologies
116. iPowerLoad, iPowerware Technologies
117. iPowerLoad, iPowerware Technologies
118. iPowerLoad, iPowerware Technologies
119. iPowerLoad, iPowerware Technologies
120. iPowerLoad, iPowerware Technologies
121. iPowerLoad, iPowerware Technologies
122. iPowerLoad, iPowerware Technologies
123. iPowerLoad, iPowerware Technologies
124. iPowerLoad, iPowerware Technologies
125. iPowerLoad, iPowerware Technologies
126. iPowerLoad, iPowerware Technologies
127. iPowerLoad, iPowerware Technologies
128. Monarch, Datawatch Corporation
129. Mozart, Magma Solutions
130. Mozart, Magma Solutions
131. Mozart, Magma Solutions
132. NatQuery, NatWorks, Inc
133. netConvert, The Workstation Group, Ltd.
134. NGS-IQ, New Generation Software
135. Oracle Warehouse Builder, Oracle Corporation
136. ODBCFace, System Tech Consulting
137. OLAP Data Migrator, Legacy to Web Solutions
138. OpenExchange, OpenData Systems
139. OpenExchange, OpenData Systems
140. OpenExchange, IST
141. OpenMigrator, PrismTech
142. OpenWizard Professional, OpenData Systems
143. OptiLoad, Leveraged Solutions, Inc.
144. Oracle Warehouse Builder, Oracle Corporation
145. Oracle Warehouse Builder, Oracle Corporation
146. Oracle Warehouse Builder, Oracle Corporation
147. Oracle Warehouse Builder, Oracle Corporation
148. Oracle Warehouse Builder, Oracle Corporation
149. pcMainframe, cfSOFTWARE
150. PinnPoint Plus, Pinnacle Decision Systems
151. PL/Loader, Hanlon Consulting
152. PointOut, mSE GmbH
153. PowerLoad, PowerLoad Technologies
154. PowerDesigner WarehouseArchitect, Powersoft
155. PowerMart, Informatica
156. PowerStage, Sybase
157. Rapid Data, Open Universal Software
158. Relational Database, Liant Software Corporation
159. Relational Tools, Princeton Software
160. ReTarGet, Tominy
161. Rodin, Rodin Mill Pty Ltd.
162. Roll-Up, Ironbridge Software
163. Sagent Solution, Sagent Technology, Inc.
164. SAS/Warehouse Administrator, SAS Institute
165. Schematic Advanced, Appligator.com
166. Scribe Integrate, Scribe Software Corporation
167. Scriptoria, Bunker Hill
168. SERdistiller, SER Solutions
169. Signiant, Signiant
170. SFINA PRO, Diagnos
171. Speedometer, Benchmark Consulting
172. SRTTransport, Schema Research Corp.
173. StarQuest Data Replicator, StarQuest Software
174. StarTools, StarQuest
175. StarTransfer, Circle Systems
176. Strategy, SPSS
177. Sunopsis, Sunopsis
178. SyncSort Unix, Syncsort
179. TableTrans, PPD Informatics
180. Text Agent, Tasc, Inc.
181. TextPipe, Crystal Software Australia
182. TextProc2000, IVRA
183. TextProc2000, IVRA
184. Tilon, Tilon
185. Transporter Fountain, Digital Fountain
186. TransportIT, Computer Associates
187. ViewShark, infoShark
188. Vignette Business Integration Studio, Vignette
189. Visual Warehouse, IBM
190. Volantia, Volantia
191. vTag Web, Cennotate Technologies
192. Waha, Beacon Information Technology
193. Warehouse, Taurus Software
194. Warehouse Executive, Ardent Software
195. Warehouse Plus, eHy Systems
196. Warehouse Workbench, Systemfabrik
197. Web Automation, webMethods
198. Web Data Kit, LOTONtech
199. Web Mining, Blossom Software
200. Web Replicator, Media Consulting
201. WebFOCUS ETL Manager, Information Builders, Inc.
202. WebSQL, Caesius Software
203. WhizBang! Extraction Library, WhizBang! Labs
204. WinPop, Tuning Point
205. Xiris, GrayMatter Software Corporation
206. XB, XB Inc.

### Open Source

- Clover
- Pentaho
- Talend

### ETL Tools, current as of 2003

Outline

- Motivation
- Goal of the benchmark
  - Effectiveness
  - Efficiency
- Benchmark parameters
  - Experimental parameters
  - Measured effects
- ETL flows
  - Micro-level: activities
  - Macro-level: workflows
- Specific scenarios
- Open issues
Outline

- **Motivation**
- Goal of the benchmark
  - Effectiveness
  - Efficiency
- Benchmark parameters
  - Experimental parameters
  - Measured effects
- ETL flows
  - Micro-level: activities
  - Macro-level: workflows
- Specific scenarios
- Open issues
Motivation

• An ETL benchmark can be used
  – as a comparison method for
    • ETL tools
    • ETL methods (algorithms)
    • ETL designs
  – for experimenting with ETL workflows
    • for optimizing ETL workflows
      – logical [ICDE05, TKDE05] and physical [DOLAP07] optimization
      – QoX-driven optimization [EDBT09, SIGMOD09]
    • what are the important problem parameters & what are the realistic values for them?
    • what test suites should we use?
Motivation

- Existing standards are insufficient
  - TPC-H
  - TPC-DS
- Practical cases are not publishable
  ... and hard to find
- We resort in devising our own ad-hoc test scenarios
  - either through a specific set of scenarios
  - or, through a scenario generator (will not touch this here)
Outline

- Motivation
- **Goal of the benchmark**
  - Effectiveness
  - Efficiency
- Benchmark parameters
  - Experimental parameters
  - Measured effects
- ETL flows
  - Micro-level: activities
  - Macro-level: workflows
- Specific scenarios
- Open issues
Goal of this work

- We are interested in understanding
  - The important parameters to be tuned in an experiment & the appropriate values for them
  - The appropriate measures to be measured during an experiment
  - The fundamental families of activities performed in an ETL scenario
  - The frequent ways with which activities and recordsets interconnect in an ETL scenario
Fundamental goals of any ETL flow

- **Effectiveness**
  - Quality objectives as
    - performance, recoverability, reliability, freshness, maintainability, scalability, availability, flexibility, robustness, affordability, consistency, traceability, auditability
  - Data should respect both database and business rules
  - Typical questions
    - Q1. Does the workflow execution reach the maximum possible level of data *freshness, completeness, and consistency in the warehouse within the necessary time* (or resource) constraints?
    - Q2. Is the workflow execution *resilient to occasional failures*?
    - Q3. Is the workflow easily *maintainable*?
Fundamental goals of any ETL flow

• **Efficiency**
  – Typically ETL processes should run within strict time windows
  – Achieving high performance enables other qualities as well
  – Typical questions
    • Q4. How *fast* is the workflow executed?
    • Q5. What degree of *parallelization* is required?
    • Q6. How much *pipelining* does the workflow use?
    • Q7. What *resource overheads* does the workflow incur at the source, *intermediate* (staging), and warehouse sites?
Outline

• Motivation
• Goal of the benchmark
  – Effectiveness
  – Efficiency
• Benchmark parameters
  – Experimental parameters
  – Measured effects
• ETL flows
  – Micro-level: activities
  – Macro-level: workflows
• Specific scenarios
• Open issues
Experimental parameters

- Parameters for the measurement of ETL workflows:
  - P1. the size of the workflow
  - P2. the structure of the workflow
  - P3. the size of input data originating from the sources,
  - P4. the workflow selectivity
  - P5. the values of probabilities of failure,
  - P6. the latency of updates at the warehouse
  - P7. the required completion time
  - P8. the system resources (e.g., memory, processing power)
  - P9. the “ETL workload” and the number of instances of the workflows that should run concurrently
Measures

- **Q1. Measures for data freshness and data consistency**
  - % data that violate business rules / are not present at the DW
- **Q2. Measures for the resilience to failures**
  - MTBF, MTTR, #rec_points, resumption type, #replicas, ETL uptime
- **Q3. Measures for maintainability (qualitative objective)**
  - Flow length, complexity, modularity, coupling
- **Q4. Measures for the speed of the overall process**
  - Throughput of workflow execution: regular, w/ failures, avg latency per tuple in regular execution
- **Q5. Measures for partitioning parallelism**
  - Partition type, number/length/data_volume of branches, #partitions,
- **Q6. Measures for pipelining parallelization**
  - CPU/mem util for flows/operators, #blocking operators, length of the largest and smaller paths containing pipelining operations
- **Q7. Measured Overheads**
  - Memory consumed at the sources/DW, elapsed time for OLTP/OLAP transactions (w/ or w/o failures)
Outline

• Motivation
• Goal of the benchmark
  – Effectiveness
  – Efficiency
• Benchmark parameters
  – Experimental parameters
  – Measured effects
• ETL flows
  – Micro-level: activities
  – Macro-level: workflows
• Specific scenarios
• Open issues
Micro-macro view of ETL flows

- **Micro-level**
  - Inside the workflow
  - A “taxonomy” for ETL activities

- **Macro-level**
  - Infinite possibilities of connecting nodes (activities and recordsets)
  - A set of “design patterns” as abstractions of how frequently encountered ETL graphs look like
Micro level

- Problem
  - derive a set of fundamental classes, where frequently encountered activities can be classified

- Why a taxonomy of ETL activities?
  - Impossible to predict any possible script / algorithm / operator
  - No algebra for ETL available right now

- Not necessary only for the benchmark, useful for other tasks (e.g., optimization, statistics, etc.)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Row-level:</strong> Function that can be applied locally to a single row</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routers: Locally decide, for each row, which of the many outputs it should be sent to</td>
<td></td>
<td>Transformer (A generic representative of a broad range of functions: date and time, logical, mathematical, null handling, number, raw, string, utility, type conversion/casting, routing)</td>
<td>- Deduplicator (distinct) - Filter - Sequence - Constant - Table function (it is applied on a set of rows for increasing the performance) - Data Cleansing Operators (Name and Address, Match-Merge) - Other SQL transformations (Character, Date, Number, XML, etc.)</td>
</tr>
<tr>
<td></td>
<td>- Conditional Split - Multicast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unary Grouping: Transform a set of rows to a single row</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Aggregate - Pivot/Unpivot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unary Holistic: Perform a transformation to the entire data set (blocking)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Sort - Percentage Sampling - Row Sampling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binary or N-ary: Combine many inputs into one output</td>
<td></td>
<td>UNION-like: - Union All - Merge Join-like: - Merge Join (MJ) - Lookup (SKJ) - Import Column (NLJ)</td>
<td></td>
</tr>
<tr>
<td>Load</td>
<td></td>
<td>- Export Column - Slowly Changing Dimension</td>
<td></td>
</tr>
</tbody>
</table>

* All ETL tools provide a set of physical operations that facilitate either the extraction or the loading phase. Such operations include: extraction from hashed/sequential files, delimited/fixed width/multi-format flat files, file set, ftp, lookup, external sort, compress/uncompress, and so on.
Macro level

- Even harder!
- How to derive a set of typical structural patterns for an ETL scenario?
  - Top down: delve to the fundamental constituents of such a scenario
  - Bottom up: explore scenarios and try to abstract common parts
- We did a little bit of both, and derived a fundamental pattern of structure
Butterflies to the rescue!

A butterfly is an ETL workflow that consists of three distinct components:

− **Body**
  - a central, detailed point of persistence (e.g., fact or dimension table) that is populated with the data produced by the left wing

− **Left wing**
  - sources, activities, intermediate results
  - performs extraction, cleaning and transformation + loads the data to the body

− **Right wing**
  - materialized views, reports, spreadsheets, as well as the activities that populate them, to support reporting and analysis
Butterflies to the rescue!
Butterflies to the rescue!
Outline

• Motivation
• Goal of the benchmark
  – Effectiveness
  – Efficiency
• Benchmark parameters
  – Experimental parameters
  – Measured effects
• ETL flows
  – Micro-level: activities
  – Macro-level: workflows
• Specific scenarios
• Open issues
Butterflies to the rescue!

- Butterflies constitute a fundamental pattern of reference
  - Line
  - Balanced butterfly
- Left-winged variants (heavy of the ETL part)
  - Primary flow
  - Wishbone
  - Tree
- Right-winged variants (heavy on the “reporting” part)
  - Fork
- Irregular variants
Line
Wishbone
Primary Flow
Fork
Balanced Butterfly
Balanced Butterfly
Slowly Changing Dimension of Type II
Outline

• Motivation
• Goal of the benchmark
  – Effectiveness
  – Efficiency
• Benchmark parameters
  – Experimental parameters
  – Measured effects
• ETL flows
  – Micro-level: activities
  – Macro-level: workflows
• Specific scenarios
• Open issues
Open issues

- **Data sizes**
  - the numbers given by TPC-H can be a valid point of reference for data warehouse contents
  - Important: fraction of source data over the warehouse contents. Values in the range 0.01 to 0.7?

- **Selectivity** of the left wing of a butterfly
  - Values between 0.5 and 1.2?

- **Failure rates**
  - Range of $10^{-4}$ and $10^{-2}$?

- **Workflow size**
  - Although we provide scenarios of small scale, medium-size and large-size scenarios are also needed
Open issues

- **Nature of data**
  - not only relational
  - also: XML, unstructured data, spatial data, multimedia, ...
- **Active vs. off-line modus operandi**
- **Auxiliary structures and processes**
  - e.g., indexes, backup & maintenance scenarios, etc.
- **Parallelism and Partitioning**
Conclusions

• We need a **commonly agreed benchmark** that realistically reflects real-world ETL scenarios

• We have provided
  – A list of **parameters and metrics**
  – A **taxonomy** for ETL activities (micro level)
  – A set of design patterns: **butterflies** (macro level)

• Future tasks
  – study more real-world scenarios for identifying
    • workflow complexity
    • workflow variants of different scale
    • frequencies of typically encountered ETL operations
Thank you!

All pictures are imported from MS Clipart
Auxiliary slides
Micro level

Physical-level characteristics

- blocking
- semi-blocking
- non-blocking

# inputs

- unary
- binary
- N-ary

row-level

- router
- grouper

Final classification
Data Warehouse:

**PART** (rkey, s_partkey, name, mfr, brand, type, size, container, comment)

**SUPPLIER** (s_suppkey, name, address, nationkey, phone, acctbal, comment, totalcost)

**PARTSUPP** (s_partkey, s_suppkey, availability, supplycost, comment)

**CUSTOMER** (s_custkey, name, address, nationkey, phone, acctbal, mktsegment, comment)

**ORDER** (s_orderkey, custkey, orderstatus, totalprice, orderdate, orderpriority, clerk, shippriority, comment)

**LINEITEM** (s_orderkey, partkey, suppkey, linenumber, quantity, extendedprice, discount, tax, returnflag, linestatus, shipdate, committdate, receiptdate, shipinstruct, shipmode, comment, profit)

Storage House:

**PART** (partkey, name, mfr, brand, type, size, container, comment)

**SUPPLIER** (suppkey, name, address, nationkey, phone, acctbal, comment)

**PARTSUPP** (partkey, suppkey, availability, supplycost, comment)

Sales Point:

**CUSTOMER** (custkey, name, address, nationkey, phone, acctbal, mktsegment, comment)

**ORDER** (orderkey, custkey, orderstatus, totalprice, orderdate, orderpriority, clerk, shippriority, comment)

**LINEITEM** (orderkey, partkey, suppkey, linenumber, quantity, extendedprice, discount, tax, returnflag, linestatus, shipdate, committdate, receiptdate, shipinstruct, shipmode, comment)
Statistics per pattern

<table>
<thead>
<tr>
<th>Filters</th>
<th>Functions</th>
<th>Routers</th>
<th>Aggr</th>
<th>Holistic f.</th>
<th>Joins</th>
<th>Diff</th>
<th>Unions</th>
<th>Load Body</th>
<th>Load Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line</td>
<td>1+1</td>
<td>2+0</td>
<td>0+1</td>
<td>0+3</td>
<td></td>
<td></td>
<td></td>
<td>INCR</td>
<td>INCR</td>
</tr>
<tr>
<td>Wishbone</td>
<td>1+0</td>
<td>4+0</td>
<td></td>
<td></td>
<td>1+0</td>
<td></td>
<td></td>
<td>INCR</td>
<td>-</td>
</tr>
<tr>
<td>Pr. Flow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I/U</td>
<td>-</td>
</tr>
<tr>
<td>Tree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3+0</td>
<td></td>
<td></td>
<td>I/U</td>
<td>I/U</td>
</tr>
<tr>
<td>Fork</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I/U</td>
<td>I/U</td>
</tr>
<tr>
<td>BB(1)</td>
<td>4+0</td>
<td>0+4</td>
<td></td>
<td>1+0</td>
<td></td>
<td></td>
<td></td>
<td>INCR</td>
<td>FULL</td>
</tr>
<tr>
<td>BB(2)</td>
<td>0+2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>-</td>
<td>I/U</td>
</tr>
<tr>
<td>2+1</td>
<td>13+2</td>
<td>0+1</td>
<td>0+12</td>
<td>1+0</td>
<td>6+0</td>
<td>1</td>
<td>1+0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:

- **N+M** (left wing + right wing)
- **INCR**: incremental maintenance
- **I/U**: insert and/or update
- **FULL**: full recomputation
Partitioning & parallelism

Original state

(a) Sequential execution

(b) Pipelining

(c) Partitioning
Measures

• Q1. Measures for **data freshness** and **data consistency**
  – The objective is to have data respect both database and business rules
  – Concrete measures are:
    • (M1.1) Percentage of data that violate business rules
    • (M1.2) Percentage of data that should be present at their appropriate warehouse targets, but they are not
Measures

- Q2. Measures for the resilience to failures
  - Test the capability of a workflow to successfully compensate within the specified time constraints
  - Concrete measures are:
    - (M2.1) Percentage of successfully resumed workflow executions
    - (M2.2) MTBF, the mean time between failures
    - (M2.3) MTTR, mean time to repair
    - (M2.4) Number of recovery points used
    - (M2.5) Resumption type: synchronous or asynchronous
    - (M2.6) Number of replicated processes (for replication)
    - (M2.7) Uptime of ETL process
Measures

• Q3. Measures for **maintainability** (qualitative objective)
  – It captures the effort needed after a change has been occurred either at the SLA’s or the underlying systems
  – Concrete measures are:
    • (M3.1) Length of the workflow (i.e., the length of its longest path)
    • (M3.2) Complexity of the workflow refers to the amount of relationships that combine its components
    • (M3.3) Modularity (or cohesion) refers to the extent to which the workflow components perform exactly one job
    • (M3.4) Coupling captures the amount of relationship among different recordsets or activities (i.e., workflow components)
Measures

• Q4. Measures for the speed of the overall process
  – The objective is to perform the ETL process as fast as possible
  – Concrete measures are:
    • (M4.1) Throughput of regular workflow execution (this may also be measured as total completion time)
    • (M4.2) Throughput of workflow execution including a specific percentage of failures and their resumption
    • (M4.3) Average latency per tuple in regular execution
Measures

• Q5. Measures for **partitioning parallelism**
  • (M5.1) Partition type (e.g., round-robin, hash-based, follow-database-partitioning, and so on)
  • (M5.2) Number and length of workflow parts that use partitioning
  • (M5.3) Number of partitions
  • (M5.4) Data volume in each partition (it is related to partition type too)

• Q6. Measures for **pipelining parallelization**
  • (M6.1) CPU and memory utilization for pipelining flows or for individual operation run in such flows
  • (M6.2) Min/Max/Avg length of the largest and smaller paths (or subgraphs) containing pipelining operations
  • (M6.3) Min/Max/Avg number of blocking operations
Q7. Measured Overheads

- The overheads at the source and DW are measured in terms of consumed memory and latency w.r.t. regular operation.

- Concrete measures are:
  
  1. (M7.1) Min/Max/Avg timeline of memory consumed at the sources
  2. (M7.2) Time needed to complete a set of OLTP transactions in the presence (vs. absence) of ETL software at the sources (normal mode)
  3. (M7.3) The same as 7.2, but with source failures (recovery mode)
  4. (M7.4) Min/Max/Avg timeline of memory consumed at the DW
  5. (M7.5) (active warehousing) Time needed to complete a set of OLAP queries in the presence (vs. absence) of ETL software at the DW (normal mode)
  6. (M7.6) The same as M7.5, but with DW failures (recovery mode)