Evaluating the Performance of ETL tool across multiple database systems

Prajesh Jhumkhawala [paj5858@rit.edu]
Advisor: Dr. Rajendra K. Raj [rkr@cs.rit.edu]

Objective
• Analyze how the ETL tools react to change in the source or destination system
• Analyze the performance of ETL tools when the size of the data increases

Introduction
• Overview of ETL process:
  - ETL processes take up to 80% of the efforts in BI projects
  - Performance of ETL depends on the source and destination database in use
  - Need to analyze the performance of ETLs with various database systems
  - Data size keeps on increasing and hence need to see how the ETL tools react when the size of data increases

Hardware/Software
• ETL Tools: Pentaho Data Integration (Kettle), Talend Open Studio, Microsoft SQL Server Integration Services (SSIS).
• Database System: MongoDB, PostgreSQL, SQL Server, MySQL
• System Configuration: Intel Core i7 and 8Gb RAM

Design
Architecture
Source database system
ETL Tools
Destination database system
Time Taken
Result and Analysis

Result
• ETL Tool v/s Database Systems

Conclusion
• The performance of an ETL is highly affected by the size of data which it operates on as well as the source and destination database in picture
• The ETL tools performance does not depend only on whether the data is SQL or NoSQL though it is one of the factor which plays an important role
• Some ETL tools have a sharp increase in the time taken when the size of data increases beyond a certain point indicating that the efficiency of that ETL tool is inversely proportional to the size of data

Future Work
• Extending the results to take into consideration other factors such as CPU time, Memory load, system specifics
• Extending the research to view the performance when the data size is further increased (5-6 times the current size)
• Use actual company data to enable real-world analysis about the ETL tool performance
• Extend the research to include more ETL tools and database systems to make the research more general

Implementation
• Citywide Payroll data published on NYC OpenData by Office of Payroll Administration
• Contains 2.19 million instances each having 16 columns
• Initial load of 1.3 million instances
• Increase the size of data by 0.1 million instance

References