
*Project Name: Implement Pilot Data Warehousing on ISI Web of Knowledge.
Company: Innovation Institute of Ontario*

Scope:

This document is designed to provide process guidelines used to Design & Implement pilot Data Warehouse structure to extract, parse and clean Scientific Information data from test ISI Web of Knowledge.

The source data is *Project Report of Scientists* submitted to Government of Canada for funding. Each report consists of 15-20 pages of word files. Since these reports have all extract of research made by these scientists throughout Canada, it is highly confidential.

Project Details:

Key Persons	:	Mr. Mark Dietrich (Project Manager) Mr. Hemal Shah (Database Administrator)
Start Date	:	October 10, 2006
End Date	:	November 12, 2006

Considerations:

- Since the source data is in word format, immense care is required to extract data into SQL Server with minimum of data loss
- Use of ETL tools is required for cleansing and parsing data.
- After cleansing and parsing, a report needs to be created to match the actual printed report.

Process:

The project comprises of two phases;

- (1) Import source data, clean and normalize it.
 - a. Extract data from around 1100 report files in word (.doc) format into SQL database on a test server 'Test' database.
 - b. Parse data from the extracted data dump
 - c. Clean the import data by removing trailing spaces, typos, misspelled words, etc.
 - d. Normalize the data by separating the entities and establishing relationship between them.
- (2) Move clean data into Data Warehouse for reporting

- a. Moving the clean-Parsed-Normalized data into Data Warehouse server 'Thomson' database.
- b. Creating desired reports as per the report sample using Crystal Reports 8.0 to view relational data
- c. Comparing report print out with the actual source printed report.

Conclusion:

The process was completed successfully with 81 % success.