CHAPTER 2:

DESCRIPTION OF THE RESEARCH WORK

2.1 Research problem

This research work emphasis on the extraction process of ETL. Extraction is the process of extracting data from a source system for further use in a data warehouse environment. This is the first step of the ETL process. After the extraction, this data can be transformed and loaded into the data warehouse.

None of the extraction processes, today address the security during the extraction process, thus there are possibilities for the data to be hacked during the process [13, 45]. If the data that is extracted content any confidential data, then just providing security after building data warehouse cannot make data secure as it would have been hacked during building process itself.

This work focuses on improving the extraction process by use of flat file and providing security to the flat files during the extraction process.

- Data Warehouse is used to store large data amounts of data.

- This data is used to take decision and for statistical analysis.

- The building Process of the Data warehouse needs to be highly effective and reliable.

- Research focus: To improve building process of the Data warehouse.
2.2 Existing Methodologies

Extraction is the operation of extracting data from a source system for further use in a data warehouse environment. This is the first step of the ETL process. After the extraction, this data can be transformed and loaded into the data warehouse. The source systems for a data warehouse are typically transaction processing applications.

For example, one of the source systems for a sales analysis data warehouse might be an order entry system that records all of the current order activities. Designing and creating the extraction process is often one of the most time-consuming tasks in the ETL process and, indeed, in the entire data warehousing process [34]. The source systems might be very complex and poorly documented, and thus determining which data needs to be extracted can be difficult. The data has to be extracted normally not only once, but several times in a periodic manner to supply all changed data to the data warehouse and keep it up-to-date. Moreover, the source system typically cannot be modified, nor can its performance or availability be adjusted, to accommodate the needs of the data warehouse extraction process [68].

These are important considerations for extraction and ETL in general. The focus is on the technical considerations of having different kinds of sources and extraction methods. It assumes that the data warehouse team has already identified the data that will be extracted, and discusses common techniques used for extracting data from source databases.

Designing this process means making decisions about the following two main aspects:

- Which extraction method can be choose which influences the source system, the transportation process, and the time needed for refreshing the warehouse.
• How can we provide the extracted data for further processing, which influences the transportation method, and the need for cleaning and transforming the data.

2.2.1 Introduction to Extraction Methods in Data Warehouses

Key References:

“Integrated data warehousing for telecommunication industries”

Outline of the paper:
• Oracle materialized views are created for saving pre-computed result to protect from the re-execution of queries
• Does not support data analysis from different perspective

“Near Real Time ETL”

Highlights of the paper:
• Use of Enterprise Integration Application(EIA) for fast extraction
• EAI is a Push System not appropriate for batch transformations
• EAI implementations are time consuming and needs lot of resources.
“Optimizing ETL processes in data warehouses”

Simitsis, A.; Vassiliadis, P.; Sellis, T.;


This paper focuses on:

- Use of heuristic search algorithm to reduce search space
- Use of signature patterns.
- Work well for small datasets.
- Very poor for medium and large datasets.
- Security is not addressed

“Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes”

Mrunalini, M.; Kumar, T.V.S.; Kanth, K.R.;


Focus of the paper:

- Addresses the security problem during extraction process of ETL
- Does not address the performance factors like time.
“Speeding ETL Processing in Data Warehouses Using High-Performance Joins for Changed Data Capture (CDC)”

Tank, D.M.; Ganatra, A.; Kosta, Y.P.; Bhensdadia, C.K.;

Advances in Recent Technologies in Communication and Computing (ARTCom), 2010 International Conference on Digital Object Identifier: 10.1109/ARTCom.2010.63 Publication Year: 2010, Page(s): 365 – 368 IEEE Conferences

Outline of the paper:

- Use of joins and aggregation for CDC.
- Execution of joins and aggregation will slow the CDC process.

“cql – A Flat File Database Query Language”


Focus of the paper:

cql is a UNIX system tool that applies C style query expressions to flat file databases

This paper describes the cql data description and query language, query optimizations, and provides comparisons with other tools.

ETL: Literature Survey

- Jian-hua Luo; Yong-ming Chen; Qing-ling Zeng; International Conference on The Design and Implementation of Data Integration System Based on the Extraction-Transformation-Loading Technology Management and Service Science (MASS), 2010 Digital Object Identifier: 10.1109/ICMSS.2010.5576560 Publication Year: 2010, Page(s): 1 – 4
Improved Extraction mechanism in ETL process for building of a Data Warehouse


- Tank, D.M.; Ganatra, A.; Kosta, Y.P.; Bhensdadia, C.K."Speeding ETL Processing in Data Warehouses Using High-Performance Joins for Changed Data Capture (CDC)” at International Conference on advances in Recent Technologies in communication and computing (ARTCom), 2010

### 2.2.2 Data Warehousing Extraction Examples

We can extract data in two ways:
- Extraction Using Data Files
- Extraction Through Distributed Operations

**Change Data Capture**

An important consideration for extraction is incremental extraction, also called Change Data Capture. If a data warehouse extracts data from an operational system on a nightly basis, then the data warehouse requires only the data that has changed since the last extraction (that is, the data that has been modified in the past 24 hours). Change Data Capture is also the key-enabling technology for providing near real-time, or on-time, data warehousing.[3]
2.2.3 Data extraction with variable datasets

Data warehouse projects involve populating databases with unique subsets of records. These record subsets are frequently extracted from much larger sets that can sometimes contain many millions of records. Data extractions is based on lookup tables, which can be implemented using several different techniques and their performance differences are examined using several test data sets and flat files [75, 76, 79]. This can help to guide the warehouse designer's decisions regarding sorting, indexing, and choice of an extraction method.[3]

Observation shows total times to extract 10,000 records from source data sets and files of 10,000, 100,000, and 1,000,000 records on the PC.

**Limitation:**

Data for the largest merge extraction is not shown because it need high specification computer to able to sort or index the source data set with one million records. This does make the point that these operations can be very resource intensive and will become impractical at some point on any computer [66].

Extractions with an indexed key take the greatest amount of time. The difference is substantial, with key times of 20 minutes to extract 1,000,000 records, versus one or two minutes for the other techniques. The SQL extraction is the fastest technique, in part because it avoids sorting or indexing overhead.
2.3 Thesis work methodology

This research work makes use of Change Data Capture mechanism along with full extraction as, every time extracting the complete data does not make sense if the copy of the unchanged data is already available at the destination. Change capture data mechanism is used to keep data updated with the existing changes at the source system. According to this mechanism trigger is raised at the destination as soon as the data base at the source is updated. The destination system can extract only the updated records to the destination in the similar fashion as that of the earlier extraction. The full extraction mechanism is used to extract the data only for the first time.

- Data warehouse is popularly used for statistical data analysis and as a decision making system.
- **More than 60%** of the data warehouse is build using ETL processes.
- **More than 70%** of the time is spent on Extraction and Transformation process.
- Improving Extraction or Transformation process will lead to improvement in ETL process and thus contributes to the improvement in building of DWH
- Motivation: Difficulty in designing the secure system and Extraction Optimization
2.4 Thesis Contributions:

- Data warehouse is popularly used for statistical data analysis and as a decision making system.

- More than 60% of the data warehouse is built using ETL processes.

- More than 70% of the time is spent on Extraction and Transformation process.

- Improving Extraction or Transformation process will lead to improvement in ETL process and thus contributes to the improvement in building of DWH.

- Motivation: Difficulty in designing the secure system and Extraction Optimization

1. This research makes use of full extraction only once and later makes use of incremental extraction, thus the data size for the later extraction will not be very bulky.

2. As the extraction process makes use of flat file during the extraction process, the size consumed by the data base in the disk is comparatively very small and thus is extracted much faster than that of database file.

3. Only factor which matters during the extraction process is the security, as this process makes use of flat files and they are more prone to violate the security features than the database files. This is over come by use of one time password over a file during the extraction.
4. File hacking during the extraction process becomes useless as the flat files which are to be extracted from source to destination are password protected and thus cannot be used by unauthorized person.

5. Earlier extraction mechanism till date did not emphasis on optimization and security both during the extraction process.

- Data Warehouse Size
- Data Access Latency Time
- **Optimized Data warehouse Building**
  - Improves **Data warehouse Architecture** Efficiency
- **Information quality information completeness**
- **System quality** includes **System scalability**
- Plat form independent Extraction
- Faster extraction
- Secure extraction and
- Reliable extraction
- Knowledge based expert system
- Social Networking Sites
• E-Commerce, M-Commerce, L-Commerce

• Shopping Cart Portals

• Management Information Systems (MIS)
  
  • Large Scale Database Applications like ATM, Payment Gateways, Online voting system etc