페트리넷 기반의 **ETL**프로세스 모델링

정성재*, 최윤호*, 황종하*,김동훈*,이화식*
*(주)엔코아 기술혁신팀
e-mail : sesame.jung@gmail.com

Petri net based ETL Process Modeling

Sung-Jae Jung*, Jongha Hwang*, Dong-Hoon Kim*, Hwasik Lee*
*Division of Technology Innovation, EN-CORE Corporation

요 약

ETL(Extraction, Transformation, Loading)작업은 데이터웨어하우스 시스템 구축 및 유지관리 뿐만 아니라 차세대 정보시스템 구축의 핵심 프로세스에 해당 한다. 특히 차세대 정보시스템 구축에 있어서 ETL 작업의 성능은 시스템오픈의 성패를 좌우하는 중요한 요소이다. 일반적으로 차세대시스템 구축의 데이터 전환을 위해 수행되는 ETL작업은 대용량데이터에 대한 다양한 형태의 데이터변형 과정을 수반하며 반드시 정해진 시간안에 완료되어야 한다. 또한, 수 많은 원천 집합을 추출하여 타겟시스템의 형태에 맞게 변형 및 적재하는 작업은 한정된 컴퓨팅 자원내에서 수행되어야 한다. 따라서 자원성능을 고려한 ETL작업 스케줄 최적화는 데이터전환 작업에 있어 필수적인 고려사항이 된다.

이 연구에서는 가용한 물리자원을 최대한 활용하 여 ETL 프로세스의 처리능(throughput)을 최대화 하는데 초점을 맞추어, ETL프로세스를 페트리넷을 이용해 모델링하는 기법을 제시한다. 이 모델에는 ETL 프로세스가 수행될 서버의 컴퓨팅자원이 토큰화(tokenize)되어 포함된다. 이 모델을 기반으로 가용자원을 최대한 활용하면서도 자원병목이 발생하지 않는 수준으로 ETL 작업이 수행될 수있도록 프로세스를 제어할 수 있는 방안을 제시한다.

1. 서론

정보기술의 발달로 정보시스템에서 발생하는 정보량이 급속히 증가함에 따라 ETL(Extraction, Transformation, Loading)작업의 성능은 매우 중요한 요소가 되었다. ETL은 데이터웨어하우스 시스템 구축 및 유지관리 뿐만 아니라 차세대 정보시스템 구축의 핵심 프로세스에 해당한다. 특히 차세대 정보시스템 구축에 있어서 ETL작업의 성능은 시스템오픈의 성패를 좌우하는 중요한 요소이다. 일반적으로 차세대시스템 구축의 데이터 전환을 위해 수행되는 ETL작업은 대용량데이터에 대한 다양한 형태의 데이터변형 과정을 수반하며 반드시 정해진 시간안에 완료되어야 한다. 또한, 수 많은 원천 집합을 추출하여 타겟시스템의 형태에 맞게 변형 및 적재하는 작업은 한정된 컴퓨팅 자원내에서 수행되어야 한다. 따라서 자원성능을 고려한 ETL작업 스케줄 최적화는 데이터전환 작업에 있어 필수적인 고려사항이 된다.

ETL작업성능을 최적화 하기 위해서는 가용한 시스템 자원을 최대한 활용 해야한다. 그렇지만 ETL프로세스가 CPU, Memory, 그리고 I/O와 같은 컴퓨팅 자원을 어느 하나라도 과다사용하면 시스템 과부하로 인해 오히려 처리성능이 저하되는 결과를 초래하게 된다. 그러므로 물리적인 시스템자원을 과다사용하지 않는 범위에서 최대한 많은 ETL 작업이 동시에 수행되도록 ETL 프로세스를 제어할 필요가 있다. 이 연구에서는 페트리넷을 이용해 ETL프로세스를 정형적으로 모델링하고, 이를 기반으로 과부하로 인한 성능저하를 피하면서도 가용자원을 최대한 활용하도록 프로세스의 수행을 제어 하는 방안을 제시하고자 한다.

Diogo 등[2012]은 ETL 프로세스 명세 및 검증을 위해 CPN (Colored Petri Net)을 ETL 프로세스 모델링에 적용하였다[1]. 이들은 CPN을 이용해 ETL 프로세스를 모델링 할 수 있다는 것을 보였다. 또한 CPN기반 ETL프로세스 모델을 이용해 SKP(Surrogate Key Pipelining) 사례에 대한 ETL 프로세스를 CPN으로 시뮬레이션 하여 데이터 흐름을 면밀히 검토하고 프로세스상의 오류를 손쉽게 인지할 수 있음을 제시하였다.

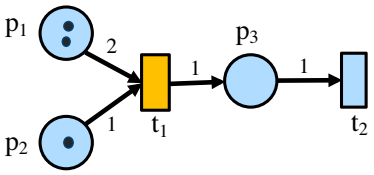
Simitis는 ETL 프로세스 명세를 위한 모델링 프레임웍을 제시하였다[2]. 그는 이 프레임웍에서 ETL 프로세스

모델링을 위해 UML(Unified Modeling Language)에 기반한 모델링 표기법을 정의하고 ETL 프로세스를 그래프형태로 나타내었다. 또한 그의 연구는 모델 기반의 프로세스 최적화 알고리즘을 제시하였다. 이 알고리즘은 논리적인 정보요소의 처리과정을 최적화 하는 ETL 프로세스의 실행계획을 수립해 준다. 그러나 Simitis의 알고리즘이 물리적인 컴퓨팅자원을 고려한 ETL 프로세스의 실행계획을 수립하는 것은 아니다.

반면, 이 연구의 초점은 가용한 물리자원을 최대한 활용하 여 ETL 프로세스의 처리능(throughput)을 최대화 하는 데 있다. ETL프로세스를 페트리넷으로 표현하고 ETL 프로세스가 수행될 서버의 컴퓨팅자원을 토큰화(tokenize)하여 페트리넷에 참여시키면 가용자원을 최대한 활용하면서도 자원병목이 발생하지 않는 수준으로 ETL 작업이 수행될 수있도록 작업을 제어할 수 있다.

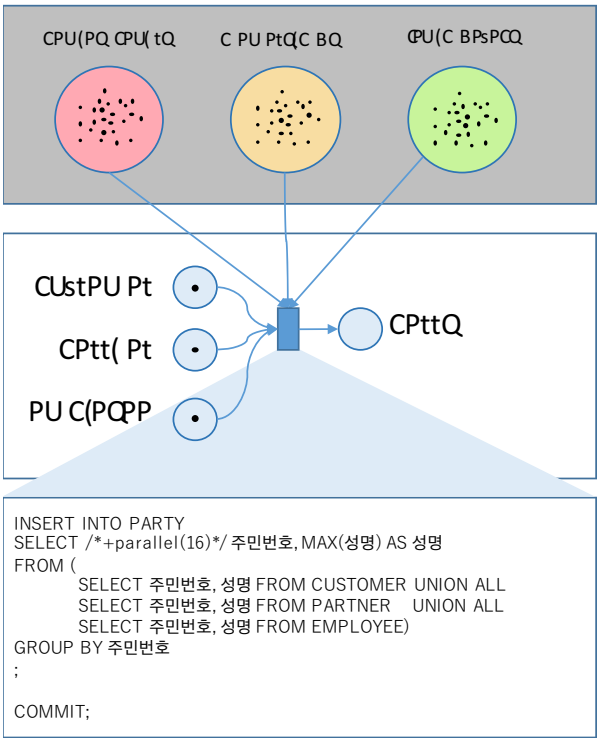
2. 페트리넷을 이용한 ETL 프로세스 모델링

페트리 넷(Petri Net)은 1962년 Carl Adam Petri가 제안한 수학적 모형으로 생화학반응 모델링 등 다양한 프로세스 모델에 널리 사용되어왔다 [3][4]. 페트리넷의 정형화된 수학적 표기는 Reddy[9]의 연구에 상세히 기술되어 있다. 페트리넷은 그래프형태의 모델링 언어로 플레이스(place) 와 트랜지션 (transition)으로 불리는 두 종류의 정점(vertex)과 방향성을 가진 연결(edge)로 구성되는 그래프이다. 페트리 넷 그래프상의 모든 연결은 플레이스에서 트랜지션으로, 트랜지션에서 플레이스로 향한다. 특정 트랜지션을 중심으로 해당 트랜지션으로 향하는 연결을 갖는 플레이스를 입력 플레이스(input place), 그리고 그 트랜지션으로부터 자신으로 향하는 연결을 갖는 플레이스를 출력 플레이스(output place)라고 부르기도 한다 [4]. 플레이스에는 토큰(token)이 위치해 플레이스가 향하는 트랜지션의 수행(fire)가능여부를 표현하며 트랜지션이 수행완료되면 트랜지션이 향하는 플레이스에 토큰이 생성된다.



(그림 1) 페트리넷 그래프

플레이스, 트랜지션, 연결, 토큰, 그리고 가중치(weight)로 구성되는 페트리넷의 일반적인 예시를 그림으로 나타내면 (그림 1)과 같다. 연결 상에 표기된 숫자는 가중치(weight)를 의미한다. 특정 트랜지션이 활성화 되어 실행(fire)되려면 해당 트랜지션으로 향하는 입력 플레이스상 토큰의 개수가 연결에 표기된 가중치의 수보다 커야만 한다. (그림 1)에서 트랜시션 t1의 입력 플레이스 p1, p2의 토큰의 개수는 2개와 1개로 가중치 2와 1이상이므로 t1은 활성화되어 실행가능한 상태가 된다.



(그림 2) 페트리넷 기반 ETL 프로세스 모델 예시

(그림 2)는 페트리넷 기반 ETL 프로세스 모델의 예시이다. 그림에서 추출원천 집합인 customer, partner, employee는 입력 플레이스로, ETL 프로세스의 출력집합인 party는 출력 플레이스로 표현되며 ETL연산을 담당하는 SQL은 트랜지션으로 표현된다. 트랜지션으로 표현된 ETL 작업은 customer, partner, 그리고 employee 세개의 테이블을 합쳐(UNION ALL)서 주민번호를 기준으로 유일(unique)한 집합을 생성하고 이를 party 테이블에 INSERT 하는 ETL 작업을 수행하고 있다. 페트리넷 기반 ETL 프로세스 모델에서의 페트리넷의 구성요소의 의미를 아래 표 1에 요약정리하였다.

<표 1> ETL 프로세스모델에서 페트리넷 구성요소의 의미

Petri net Pet r enent	Pet r e ic Per re entatien	P eaninn in ETL Prece
Place	○	sets to be extracted, transformed, and loaded comp tatbnal reso rces, be. c rrentlb aeabable CPx (Pa s), memorb, I/m
sranstbon	□	sst ob, s allbb rbtten bn dat
sdte	→	dependencbbetb een sets prereq bdte comp tatbnal reso rces for a transtbon to be enabled
s elt t	n	
soten	●	Presence of sets

3.컴퓨팅 자원의 토큰화

ETL수행 서버의 CPU, Memory, I/O등의 컴퓨팅자원을 토큰으로 표현하여 ETL프로세스 모델에 참여시키면 자원 과다 사용으로 인한 병목현상을 막을 수 있다. (그림 2)에 CPU(PQ count), Memory, I/O의 컴퓨팅자원이 페트리넷 기반 ETL프로세스 모델에 참여한 모습을 나타내었다.

가용 PQ(parallel query process)개수를 CPU자원의 토큰으로 표현할 수 있다. PQ는 병렬도를 요구하는 SQL이 수행될 때 생성되는 DBMS 서버프로세스로 동시 가용한 총 개수가 CPU 개수에 의해서 결정된다(오라클의 경우 MAX_PARALLEL_SERVERS 파라메터에 최대병렬도 값을 설정)[7]. 병렬도를 요구하는 여러개의 SQL(ETL작업)이 동시 수행되도록 요청 받았을 때 최대 PQ count를 넘기는 SQL부터는 직렬처리(serial processing) 방식으로 수행되기 때문에 전체적인 작업의 완료시간이 늦어지게 된다. 따라서 동시가용 최대 PQ개수를 토큰으로 나타내어 가용 PQ 개수가 해당 SQL의 수행에 소요되는 PQ개수보다 많을 경우에만 수행되도록 ETL프로세스를 제어할 필요가 있다.

메모리와 I/O도 ETL 작업처리에 소요되는 컴퓨팅 자원이며 토큰으로 표현할 수 있다. 예를들어 ETL 서버에서 사용자 SQL처리를 위해 할당된 메모리가 1 GB이고 I/O 대역폭이 1500 Mbps라고 가정해보자. 메모리의 경우 토큰하나를 1 MB 로 보고 1000개의 토큰을 메모리 플레이스에 배치하면 된다. I/O 대역폭의 경우 토큰하나를 1Mbps로 보고 1500개의 토큰을 I/O플레이스에 배치하면 된다. 토큰하나를 얼마만큼의 크기로 볼것인지는 ETL프로세스를 구성하는 SQL의 자원소요 특성에 따라 탄력적으로 결정하면 될것이다. PQ(CPU)자원의 경우와 마찬가지로 메모리와 I/O의 경우도 현재 남아있는 토큰의개수가 신규로 수행해야하는 SQL이 필요로하는 토큰의 개수보다 부족한 경우 기 실행중인 SQL이 자원을 반환할 때까지 기다리도록 하면 메모리나 I/O 부족으로 인한 자원 병목현상의 발생을 피할 수 있다.

4.모델기반 ETL 프로세스 제어

가용 컴퓨팅 자원이 포함된 페트리넷 기반의 ETL 프로세스 모델을 이용해 ETL프로세스의 제어가 가능하다. ETL프로세스 모델에서 트랜시션으로 표현되는 단위 SQL(ETL작업)의 수행에 소요되는 PQ개수, 메모리소요량, I/O 소요량의 측정이 필요하다. 측정된 자원소요량은 토큰의 개수로 환산하여 CPU, 메모리, I/O 플레이스에서 트랜지션으로 향하는 연결에 가중치 값으로 표기한다. 이렇게 함으로써 특정 ETL작업(SQL)의 수행요청을 받았을 때 해당 작업의 자원소요량과 해당시점에 가용한 자원량을 비교하여 ETL작업 수행 여부를 결정할 수 있다.

5.결론

이 연구에서는 페트리넷을 이용한 ETL프로세스 모델링 방안을 제시하였다. 이 모델에서는 원천집합과 목표집합 뿐만 아니라 CPU, Memory, I/O와 같은 시스템 자원이 플레이스상의 토큰으로 표현된다. 이 모델을 기반으로 ETL 시스템의 프로세스 수행을 제어하면 ETL 작업의 동시수행성을 최대화 하면서도 자원부족으로 인한 병목발생을 배제하여 ETL 프로세스의 처리능을 최대화 할 수 있을것으로 기대된다. 이 모델을 엔코아의 ETL 시스템 ETT#[6]에 적용하는 작업이 진행되고 있다. 이 작업이 완료되면, ETT#은 시스템 자원현황을 감안하여 ETL프로세스를 제어하는 최초의 ETL 툴이 될것으로 기대된다.

이 모델을 기반으로 시스템 자원부족으로 인한 병목현상이 발생하지 않는 범위내에서 동시수행성을 최대화 할 수 있는 최적화된 작업스케줄을 도출하는 ETL프로세스 전산모사기(ETL Process Simulator)의 구현이 가능할 것으로 판단되며 이를 향후 연구로 진행할 계획이다.

참고문헌

[1] Diogo Silva, Orlando Belo, and João M. Fernandes, COLORED PETRI NETS IN THE SIMULATION OF ETL STANDARD TASKS THE SURROGATE KEY PIPELINING CASE, Eurosis-ETI, 2012.

[2] Alkis Simitsis., Modeling and Managing ETL Processes, VLDB PhD Workshop, 2003.

[3] Petri, C.A., Kommunikation mit Automaten. Schriften des IIM Nr. 2, Institut für Instrumentelle Mathematik, Bonn, 1962. English translation: Technical Report RADC-TR-65-377, Griffiths Air Force Base, New York, Vol. 1, Suppl. 1, 1966.

[4] Venkatramana N. Reddy*, Michael L. Mavrovouniotis*, and Michael N. Liebmant. Petri Net Representations in Metabolic Pathways, ISMB-93 Proceedings., 1993.

[5] Oracle Corp., Oracle Database Online Documentation 11gRelease 2 (11.2), <http://docs.oracle.com/cd/E11882_01/index.htm>

[6] EN-CORE Corp., ETT# Data Integration Solution, <http://dataware.kr/solution/new_idx>