

대용량 데이터의 실시간 저장을 위한 블록 단위 삽입 트랜잭션 처리 기법과 압축-합병 저장 기법

김영기^{0 1}, 백성하¹, 이동욱¹, 정원일², 배해영¹

인하대학교 정보공학과¹, 호서대학교 정보보호학과²

{ykkim,shbaek,dwlee}@dbl.inha.ac.kr, wnchung@hoseo.edu, hybae@inha.ac.kr

Processing Method of Insert Transaction by Block and Compress-Merge to Store Real-time Very Large Data

Young-Ki Kim⁰¹, Seong-Ha Baek¹, Dong-Wook Lee¹, Warn-Il Chung², Hae-Young Bae¹

¹Department of Computer Science and Information Engineering, InhaUniversity

²Department of Information Security Engineering, HoseoUniversity

반도체나 FPD(Flat Panel Display) 산업이 발전함에 따라 생산 공정의 생산성 향상을 위하여 무인 자동화 생산 시스템을 요한다. 이러한 자동화 생산 시스템은 생산 장비에서 발생하는 데이터를 저장 관리 또는 분석 하여 생산성 향상을 위한 각종 응용에 이용한다. 응용은 생산 스케줄 관리와 에러 가능 여부를 실시간으로 체크하여 자동화 공정의 활용률을 최대한 높여서 생산 기간을 단축시키고, 생산 과정에서 발생하는 불량 을 최소화한다. 따라서 이와 같은 응용시스템은 장비에 서 발생한 데이터를 실시간으로 분석 및 적용 하기 위한 저장 관리 시스템을 사용한다. 하지만 초미세 나노 공정에 사용되는 차세대 공정제어 장비는 초당 5만 건 에서 50만 건 이상의 데이터를 발생시키기 때문에 기존의 오라클, 사이베이스, MYSQL, MSSQL과 같은 디스크 기반 DBMS로는 이러한 대용량 데이터를 실시간으로 처리하는데 한계가 있다. 따라서 본 논문에서는 초당 5만 건 이상의 실시간 대 용량 데이터를 저비용으로 디스크에 저장하기 위해 블록 단위 삽입 트랜잭션과 압축-합병 저장 기법을 제안 한다. 기존의 디스크 기반 DBMS는 튜플 단위의 트랜잭션 처리로 인한 동시성 제어 시 병목 현상과 각 튜플에 대한 로그를 남기고 검색을 위해 인덱스를 생성하는 비용으로 인해 초당 5만 건에서 50만 건 이상의 로그 데이터를 생성하는 자동화 생산 시스템의 대용량 데이터를 저장하기에는 한계가 있다. 또한 압축 저장은 압축 후 남게 되는 버퍼 공간을 활용하지 못하는 문제와 디스크 기록 시 내부 단편화 문제가 발생한다. 이와 같은 문제점을 본 논문에서는 블록 단위 삽입 트랜잭션 처리 기법과 압축-합병 저장 기법을 사용하여 해결한다. 본 논문이 제시하는 블록 단위 삽입 트랜잭션 처리 기법은 대량의 레코드를 버퍼 블록 내에 저장하여 하나의 블록이 다수의 레코드 모임인 레코드 셋의 형태로 저장된다.

클라이언트가 레코드 삽입 요청을 하면 대용량 데이터 저장 관리자는 기존의 DBMS와 달리 레코드 삽입을 위한 세션을 생성하며 각 세션은 데이터를 버퍼 블록에 지속적으로 저장하여 레코드를 모아 커다란 레코드 셋으로 만든다. 그리고 이렇게 만들어진 커다란 레코드 셋이 하나의 버퍼 블록을 모두 채우거나 삽입 대기 시간이 일정 시간을 초과하면 현재까지 버퍼 블록에 저장한 레코드 셋으로 삽입 트랜잭션을 생성하여 기록한다.

대용량 데이터 저장 관리자는 실시간 대용량 데이터를 저비용으로 저장하기 위해 데이터를 블록 단위로 압축을 한다. 하지만 버퍼 블록을 압축하게 되면 압축 후 남은 버퍼 블록 공간을 활용하지 못하고 디스크 기록 시 내부 단편화 현상이 발생한다. 따라서 압축된 블록을 합병하는 압축-합병 저장 구조를 사용한다. 압축 관리자(Compress Manager)는 삽입 요청된 버퍼를 버퍼 관리자로부터 받아서 데이터를 압축한다. 그리고 압축 관리자가 생성한 압축된 버퍼블록들을 합병하여 합병된 버퍼 블록을 생성한다. 이때 합병 대상 버퍼간의 시간 정보를 비교 후 갱신하여 합병된 버퍼 블록에 저장한다. 그 다음 합병된 블록들을 디스크에 연속적으로 기록한다. 이와 같은 압축-합병 저장 기법은 디스크 공간 활용을 최대화하여 대용량 데이터의 저장 시 디스크 공간 비용을 다소 감소시킨다.

본 논문이 제안하는 압축-합병 저장 기법의 성능 평가는 저장 속도 평가와 저장 비용 평가로 진행되었다. 저장 속도 측정 방식은 MSSQL과 Oracle에 레코드의 시간 속성을 인덱스로 생성하는 벌크 삽입 방식으로 튜플을 10,000개에서 1,000,000개까지 삽입하여 본 제안 기법의 삽입 속도와 비교하였으며 그 결과, 블록 단위 삽입 트랜잭션이 대량의 레코드들을 블록 단위로 묶어서 한 번에 삽입하기 때문에 MSSQL보다는 크게 향상되었고 오라클의 경우보다는 소폭 향상된 것을 알 수 있다. 저장 비용 평가 방식은 타 상용 DBMS와의 성능 비교를 위해 동일한 데이터를 압축 저장 하여 저장 전과 후의 크기 변화를 비교 하였다. 저장하는 데이터의 크기는 1MB ~ 100MB로 변화시켜서 비용을 측정하였다. 오라클의 경우 원본 데이터에 비해 약 75~80% 정도의 압축률을 보였고, 압축-합병 저장 기법은 약 85~90% 정도의 압축률로 다소 우수하다는 것을 알 수 있다.

본 논문은 대용량 데이터를 실시간으로 저장하기 위한 블록 단위 삽입 트랜잭션 처리 기법과 저장 공간을 감소시키기 위한 압축-합병 저장 기법을 제안하였다. 블록 단위 삽입 트랜잭션 처리 기법은 기존의 삽입 트랜잭션의 튜플 단위 수행으로 인한 일부 작업들을 생략하여 삽입 트랜잭션의 속도를 향상 시켰다. 또한 압축-합병 저장 기법은 데이터를 압축 하여 저장함으로써 저장 공간을 최소화 하였다.

향후 연구로는 블록 단위 트랜잭션을 사용하기 위한 효율적인 버퍼 교체 정책과 기존의 로깅 기법을 블록 단위로 확장한 블록 단위 로깅 기법, 그리고 본 논문에서 제시한 압축 기법을 보완하여 압축 속도를 향상 시키는 압축 기법의 연구가 필요하다.