배관 시계열 데이터를 위한 캐시 관리자의 설계 및 구현1)

김선효, 김원식, 신제용, 한욱신
경북대학교 컴퓨터공학과
e-mail:{shkim, wskim, jyshin}@www-db.knu.ac.kr,
wshan@knu.ac.kr

Design and implementation of a cache manager for pipeline time-series data

Seon-Hyo Kim, Won-Sik Kim, Je-Yong Shin, Wook-Shin Han
Dept. of Computer Engineering, Kyungpook National University

요 약

배관에 생기는 구멍이나 틈은 대형 사고의 원인이 될 수 있다. 이러한 배관의 결함을 찾기 위해서는 먼저 센서를 부착한 배관 탐사 장비를 배관에 통과시키고, 배관을 통과하는 중에 센서가 읽은 정보들을 배관 탐사 장비의 하드 디스크에 저장한다. 배관 통과가 완료된 후, 분석가는 분석 프로그램을 사용하여 탐사 장비에서 얻은 데이터에서 결함을 수동적으로 찾는다. 분석가가 데이터를 분석할 때 일반적으로 두 가지 패턴이 존재한다. 첫 번째 패턴은 일정한 구간의 센서 데이터를 순차적으로 분석하는 패턴이고, 두 번째 패턴은 현재 구간에서 이전 구간으로 되돌아가서 다시 분석하는 반복적인 패턴이다. 현재까지 만족할 만 한 수준으로 자동적으로 분석이 되지 않으므로, 분석가는 수작업으로 분석을 하는 경우가 많은데 이로 인해 최근에 읽은 부분을 전후 반복해서 액세스하는 반복적인 패턴이 많이 사용된다. 반복적 패턴의 경우 시스템의 성능을 향상시키기 위해, 이전에 읽은 배관 센서 데이터를 캐싱 할 필요가 있다. 그러나 기존의 분석 소프트웨어에는 캐싱 기능이 없으므로 반복적 패턴일 경우 데이터베이스에서 동일한 데이터를 반복적으로 읽는 문제를 가지고 있다. 본 논문에서는 배관 센서 데이터를 효율적으로 관리하는 캐쉬 관리자를 설계하고 구현하였다. 세부적으로는, 배관 센서 데이터를 시계열 데이터로 간주하고, 시계열 데이터에 대한 캐시 관리자를 제안하였다. 본 논문은 배관 탐사 장비에서 획득한 데이터들을 시계열 데이터로 간주하여 데이터베이스 측면에서 이러한 문제들을 접근하였다는 점에서 의미가 있으며, 향후 이 분야에 대한 많은 연구들이 나올 것으로 기대한다.

1. 서론

세계는 가스와 기름을 사용하기 위해 수십만 km의 배관을 사용하고 있다. 현재 이런 배관들은 안전상의 문제로 주로 지하에 매설되어 있다. 이렇게 지하에 묻혀 있는 배관의 상태를 점검하는 것은 매우 어려운 일이지만 배관의 결함으로 인해 발생하는 대형 사고를 예방하기 위해서는 필수적인 일이다. 그래서 배관의 내·외부 상태를 파악할 수 있는 지하 배관 비파괴 검사 기술[1]이 도입되었다. 그 중 인텔리전트 피그(Intelligent PIG)[4]가 현재 대부분의 배관 회사에서 사용되고 있다.

배관의 결함을 찾기 위해 피그는 배관 내의 유체의 흐름에 따라 배관을 이동하면서 피그에 탑재되어 있는 센서 장비를 사용하여 일정한 주파수 단위의 데이터를 수집한다. 분석가가 수집된 데이터를 분석할 때, 일반적으로 두 가지의 분석 패턴이 존재한다. 첫 번째 패턴은 일정한 구간의 센서 데이터를 순차적으로 분석하는 패턴이고, 두 번째 패턴은 현재 구간에서 이전 구간으로 되돌아가서 다시 분석하는 반복적인 패턴이다. 본 논문에서는 첫 번째 패턴을 순차 패턴(sequential pattern)이라 하고, 두 번째 패턴을 반복 패턴(repetitive pattern)이라 한다.

현재까지 수집된 데이터를 만족할 수준까지 자동으로 분석이 되지 않기 때문에, 분석가는 수작업으로 분석을 할 경우가 많으며, 이 때 최근에 읽은 부분을 전후 반복 패턴이 많이 사용된다. 반복 패턴의

1) 본 논문은 한국가스공사의 지원에 의하여 이루어졌으며 이에 감사 드립니다.

경우 이미 읽은 데이터들을 다시 읽어야 하기 때문에 캐시를 사용함으로써 시스템의 성능을 향상시킬 수 있다.

그러나 기존의 분석 프로그램에는 반복 패턴을 효과적으로 처리하는 캐싱 기능이 없는 것으로 추정된다. 왜냐하면 분석을 위해 화면을 앞뒤로 이동할 시 화면에 데이터를 보여주는 시간이 비슷하기 때문이다. 이것으로 매번 데이터베이스에서 데이터를 읽어 온다고 유추할 수 있다. 더군다나 분석 프로그램과 데이터베이스가 떨어져 있는 클라이언트/서버 환경이라면 분석 프로그램의 성능은 더욱 저하된다.

본 논문에서는 주파수 단위 측정된 배관 데이터를 시계열 데이터로 간주하고, 이러한 시계열 데이터를 효과적으로 관리하는 시계열 캐시 관리자(Time-series Cache: T-Cache)를 제안한다. 제안한 캐시 관리자의 자료 구조와 스마트 포인터(smart pointer)를 소개한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 기존의 데이터 분석 프로그램에 대한 관련연구에 대해 설명하고, 제 3장에서는 시계열 캐시 관리자의 구조를 설명한 뒤, 스마트 포인터의 개념에 대해 설명한다. 제 4장에서는 논문에 대한 결론을 맺는다[2].

2. 관련 연구

배관의 상태 정보를 분석하는 프로그램을 개발한 회사들에는 캐나다의 BJ Pipeline Inspection Services, CORROPRO, 영국의 PII(Pipeline Integration International), 독일의 3P Services 등의 업체가 있으며, 분석 프로그램으로는 GEODENT[2], Vectra view[3], Lina view[5] 등이 있다. 현재까지 나온 분석 프로그램에 대한 기술은 공개된 것을 찾을 수가 없었다. 배관 관련 저널과 컨퍼런스, 그리고 관련 회사 사이트를 조사해 보았으나 찾을 수가 없었다. 기술관련 자료가 없어 분석 프로그램에 캐시가 존재하는지에 대한 여부를 알 수 없었다. 다만, 분석 패턴에 맞추어 기존 분석 프로그램을 수행시켜 수행시간을 측정함으로써 캐시 기능의 사용 여부를 유추하였다.

캐시 기능의 사용 여부를 유추할 수 있는 반복 패턴에 대해 기존 분석 프로그램을 수행시켜 보았다. 그 결과 일정한 구간의 데이터를 읽기 위해 데이터를 앞으로 이동하였을 때 분석 프로그램에서 데이터를 보여주기 위해 수행된 시간과 동일한 구간을 뒤로 이동하였을 때 분석 프로그램에서 데이터를 보여주기 위해 수행된 시간이 거의 동일하였다. 이것으로 보아 기존의 분석 프로그램에서는 분석 패턴에 관계없이 데이터를 요구할 때마다 데이터베이스에서 데이터를 읽어오는 것으로 유추할 수 있다. 즉, 캐시 기능이 없거나 있더라도 그 기능이 제대로 제 역할을 하지 못하고 있다고 결론을 내릴 수 있었다.

따라서 본 논문은 자주 사용되는 분석 패턴에 대해 프로그램의 성능을 향상시키고 다른 컴퓨터에 존재하는 데이터를 주고받는 클라이언트/서버 환경 하에서도 성능을 향상시킬 수 있는 시계열 캐시 관리자를 제안한다.

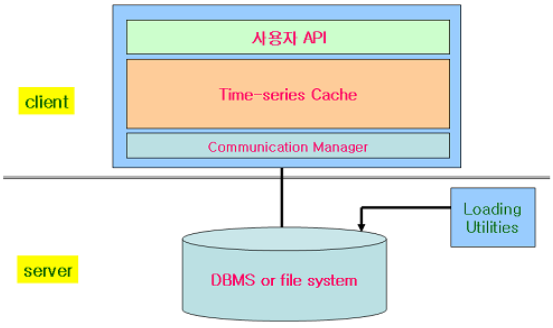
3. 시계열 캐시 관리자의 설계 및 구현

본 장에서는 논문에서 제안하는 시계열 캐시 관리자를 설명한다. 3.1에서는 배관 데이터 분석 소프트웨어의 구조를 살펴본다. 다음으로 3.2에서는 시계열 캐시 관리자의 자료구조를 설명한다.

3.1 배관 데이터 분석 소프트웨어 구조

배관 데이터 분석 소프트웨어는 센서에서 수집한 배관 데이터(raw data)를 처리, 저장, 관리하는 피그 데이터 관리자 (PIG Data Manager)와 분석가에게 요청받은 배관 데이터를 화면에 보여주는 피그 가시화 모듈(PIG visualization Module)로 구성되어 있다.

피그 데이터 관리자는 센서에서 수집한 데이터를 저장하고, 피그 가시화 모듈에서 요청받은 데이터를 전달한다. 그리고 분석자가 도출한 분석결과에 따라 데이터베이스에 저장된 데이터를 수정하는 작업을 수행한다. 피그 데이터 관리자의 자세한 구조는 그림 1과 같다.



(그림 1) PIG Data Manager

서버는 원시 배관 데이터를 처리하는 로딩 유틸리티들과 데이터를 저장하는 데이터베이스 관리 시스템 또는 파일 시스템으로 구성된다. 피그에서 수

2) 실험결과는 페이지 제약으로 인해 생략한다.

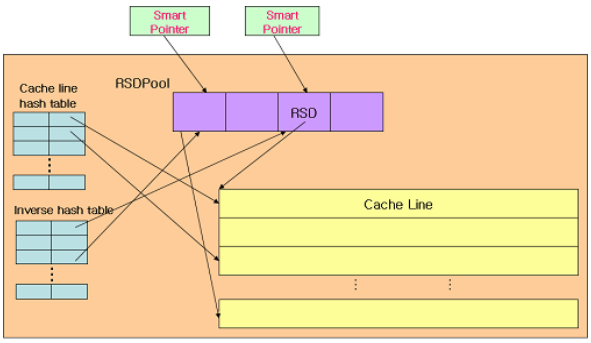
집한 원시 데이터를 로딩 유틸리티들에서 처리하여 데이터베이스에 삽입한다. 삽입된 데이터를 데이터베이스 관리 시스템에서는 저장 및 보관하고, 클라이언트의 통신 관리자로부터 데이터 요청이 있을 때 해당 데이터를 상위 계층으로 전달한다.

클라이언트는 통신 관리자, 시계열 캐시 관리자(Time-series Cache Manager) 그리고 사용자 API로 구성된다. 상위모듈이 사용자 API를 통해 데이터를 요청하면 시계열 캐시 관리자는 데이터가 캐시 내에 있는지 확인한 후, 있으면 캐시에서 데이터를 전달한다. 만일 캐시에 데이터가 없으면 통신 관리자를 통해 서버에 데이터를 요청하여 서버로부터 데이터를 받아서 상위 모듈에 전달한다.

프로그램의 성능에 가장 많은 영향을 끼치는 것은 요청받은 데이터를 상위 계층으로 전달하는 기능이다. 서버가 데이터를 요청받을 때마다 매번 데이터베이스로부터 읽어 전달하면 많은 I/O가 발생한다. 이 문제를 해결하기 위해 피그 데이터 관리자에서는 데이터베이스에 저장된 배관 센서 데이터들을 시계열 데이터로 간주하여 처리할 수 있는 시계열 캐시 관리자를 사용하여 관리하고 있다.

3.2 시계열 캐시 관리자의 자료구조

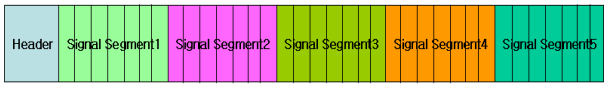
시계열 캐시 관리자의 자료구조는 그림 2와 같다. 아래에서 각각의 자료구조에 대해 자세히 설명한다.



(그림 2) 시계열 캐시 관리자의 자료구조

3.2.1 캐시라인

캐시라인(Cache Line)은 시계열 캐시 관리자의 교체 단위이다. 캐시라인은 100m단위의 배관 데이터를 보관하여 분석 소프트웨어의 의미(semantics)를 반영하였다. 이는 분석 소프트웨어가 기본적으로 100m 단위로 사용자에게 보여주기 때문이다. 그림 3과 같이 캐시 라인의 구조는 헤더(header)와 여러 개의 세그먼트(segment)들로 이루어져 있다.

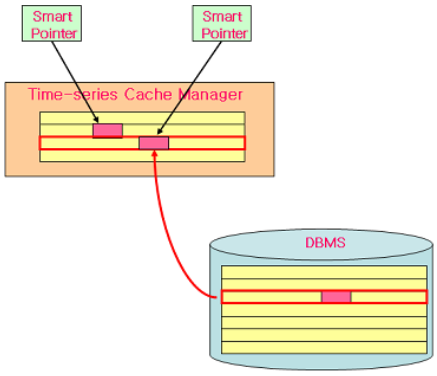


(그림 3) 캐시라인의 자료구조

헤더는 캐시 라인이 보관하고 있는 데이터들의 시작 거리와 끝 거리, 보관하고 있는 데이터의 개수에 대한 정보를 가지고 있고 캐시라인의 각 세그먼트들은 센서장비가 측정한 각 센서 데이터의 집합이다.

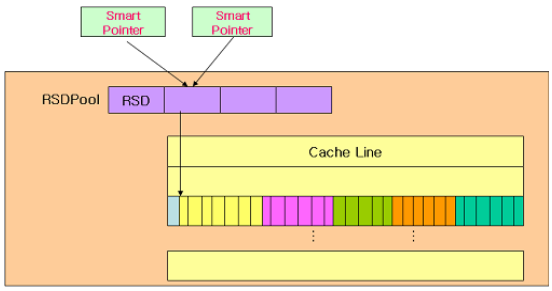
3.2.2 스마트 포인터와 RSD

스마트 포인터는 객체지향데이터베이스(Object-oriented database)에서 사용되는 개념으로써 상위 모듈에서 요청한 데이터가 캐시에 존재하는지 안 하는지에 관계없이 요청한 데이터를 동일한 방법으로 접근하는 개념이다. 즉, 요청한 객체가 시계열 캐시 관리자에 존재하는 경우, 그림 4와 같이 시계열 캐시 관리자에 바로 접근한다. 객체가 시계열 캐시 관리자에 존재하지 않는 경우에는 서버의 DB로부터 해당 객체를 시계열 캐시 관리자로 가져와서 시계열 캐시 관리자에 접근하도록 한다. 분석자는 이런 과정에 관계없이 스마트 포인터를 사용하여 동일한 방법으로 데이터에 접근할 수 있다.



(그림 4) 스마트 포인터를 통한 객체 접근

RSD(Resident Segment Descriptor)는 캐시 라인의 특정 세그먼트를 가리키는 포인터이다. RSD는 하나의 캐시라인을 가리키는 것이 아니라 캐시라인을 구성하는 세그먼트들 중 하나를 가리킨다. 캐시라인은 여러 센서 데이터의 세그먼트들을 가지고 있는데, 각 세그먼트들은 RSD가 가리키고 있다. 그래서 스마트 포인터가 접근하고자 하는 센서 데이터가 시계열 캐시 관리자에 있는지의 여부는 데이터의 세그먼트를 가리키고 있는 RSD가 존재하는지를 찾으면 된다. 그림 5는 스마트 포인터와 RSD의 관계를 나타낸 그림이다.

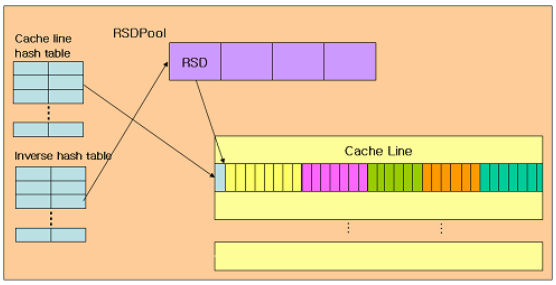


(그림 5) RSD와 스마트 포인터

RSD가 가리키고자 하는 캐시 라인이 어디에 있는지에 대한 정보를 얻기 위해서는 캐시 라인 해시 테이블(Cache line hash table)을 검색해야 한다. 다음 절에서 캐시 라인 해시 테이블과 역방향 해시 테이블에 대해 자세히 설명한다.

3.2.3 캐시 라인 해시 테이블과 역방향 해시 테이블

본 논문에서 제안하는 시계열 캐시 관리자는 그림 6에서 알 수 있듯이, 두 개의 해시 테이블로 이루어져 있다. 캐시라인을 검색하기 위한 캐시라인 해시 테이블과 캐시라인을 가리키는 RSD를 검색하기 위한 역방향 해시 테이블이다.



(그림 6) 캐시 라인 해시 테이블과 역방향 해시 테이블

캐시 라인 해시 테이블은 데이터의 구간 정보와 데이터의 캐시 내 인덱스 정보로 이루어져 있고, 역방향 해시 테이블은 캐시 라인 인덱스 정보와 그 캐시 라인을 가리키고 있는 RSD의 인덱스 정보로 이루어져 있다.

두 개의 해시 테이블을 이용하여 원하는 데이터를 접근하는 과정은 다음과 같다. 캐시 외부에서 원하는 데이터에 대한 구간 정보와 데이터의 타입 정보 그리고 구간 내에서의 위치 정보를 이용하여 데이터에 접근한다. 우선 원하는 데이터가 저장된 캐시 라인이 있는지 있다면 캐시 내에서 어디에 위치하는지 검색한다. 이 검색은 캐시 라인 해시 테이블을 이용하면 된다. 캐시 라인 해시 테이블에서 구간 정보를 가진 쌍이 존재하지 않으면 캐시에 데이터가 없으므로 데이터베이스로부터 읽어와야 한다. 새로 읽어온 캐시 라인을 해시 테이블에 저장하고 캐시내의 원하는 데이터가 있는 세그먼트를 가리키는 RSD를 생성한다. RSD를 통해 캐시 라인 내의 세그먼트에 접근하고 세그먼트 내에서의 위치 정보로 원하는 데이터에 접근한다.

만일 원하는 데이터가 속한 캐시 라인이 있다면 그 캐시 라인을 가리키는 RSD를 역방향 해시 테이블을 사용하여 검색한다. 검색하여 나온 쌍들 중에서 데이터가 속한 세그먼트를 가리키는 RSD를 검색하고 있다면 그 RSD를 통해 데이터에 접근한다. 만일 RSD가 없다면 새로운 RSD를 생성하여 데이터가 속한 세그먼트를 가리키도록 한다.

5. 결론

본 논문에서는 배관 데이터를 분석하는 프로그램의 성능 향상을 위해 일정한 주파수 단위로 측정된 배관 데이터를 시계열 데이터로 간주하고, 이러한 시계열 데이터를 효과적으로 관리하는 시계열 캐시 관리자를 제안하였다. 객체지향데이터베이스에서 사용되는 스마트 포인터를 시계열 캐시 관리자에 도입하였다. 또한 시계열 캐시 관리자는 배관 센서 데이터의 특성을 이용할 뿐 아니라 분석패턴의 특성도 활용하였다.

본 논문은 배관 탐사 장비에서 획득한 데이터들을 시계열 데이터로 간주하여 데이터베이스 측면에서 이러한 문제들을 접근하였다는 점에서 의미가 있으며, 향후 이 분야에 대해 한국가스공사에서 진행되고 있는 프로젝트의 결과로 많은 연구들이 나올 것으로 기대한다.

참고문헌

[1] K. K. Tandon, "MFL Tool Hardware for Pipeline Inspection," Materials Selection & Design, pp. 75-79, Feb. 1997.

[2] P. Michailides, et al., "NPS 8 Geopig: Inertial Measurement and Mechanical Caliper Technology," BJ Services company, Jun. 2002.

[3] S. Westwood, and D. Hektner, "Data Integration Ensures Integrity," BJ Services company, Mar. 2003.

[4] http://www.kogas.or.kr

[5] http://www.tuboscopepipeline.com/Products.htm