Source Code Analysis
RockDB 분석

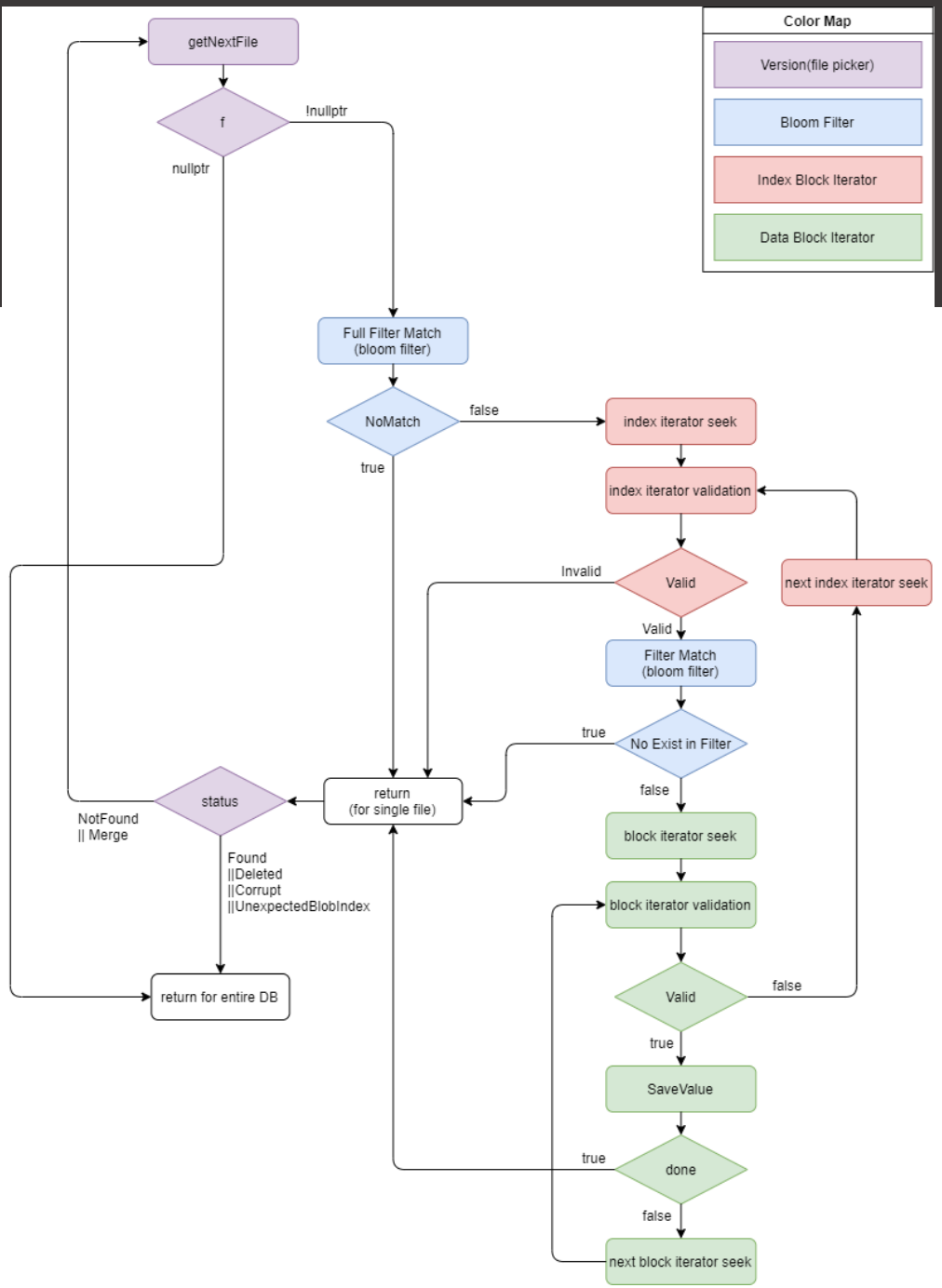
컴퓨터소프트웨어학부
김현정 박정호

CONTENTS

Flow chart

분석 결과

Flow chart



- 3개의 loop 존재
  - version_set에 하나
    (version_set.cc 1890번 Line)
  - block based table에 둘
    (block_based_table_reader.cc 2337, 2400번 Line)

각 Loop의 의미

1. Version::Get의 Loop
   SST File 단위로 이루어지는 Loop, File Picker를 사용해서 file을 순회한다.
   (앞의 flow chart에서 보라색 부분)

2. IndexIterator 의 Loop
   파일의 Index Block 을 순회하며 Key가 포함된 Data Block을 탐색하는 Loop
   (앞의 flow chart에서 빨간색 부분)

3. DataBlockIterator 의 Loop
   Index 를 통해 탐색된 Data Block 내에서 탐색하는 Loop
   (앞의 flow chart에서 초록색 부분

Sequence number와 table format, cache

1. 현재 확인한 사용중인 위치
    1. Snapshot
    2. 일부 table format(block based table, plain table)
2. default 세팅(key cache)에서 reading에 sequence number가 쓰이는 경우
    1. Key를 찾고난 이후의 validation
3. Cache option 중 row cache를 사용할때,
    1. Snapshot을 사용한 read -> snapshot이 파일의 largest_seqno보다 크면 snapshot이 모든 데이터를 포괄한다고 봄.
    2. Key를 찾고난 이후의 validation

4. Default table format(block based table)에서 sequence number
    1. table구조에서 Meta block에 저장하고 있음(key값 당 가지고 있지 않다.) range deletion에서 사용되는 것은 확인함.
5. 그 외 table format(plain, cuckoo, index table)에서 sequence number
    1. Plain table은 cache구조를 사용하지 않는 작은 파일인 대신에 key값과 함께 seq를 저장해서 사용함.
    2. 나머지(cuckoo, index) seq를 key값과 함께 저장하지 않음. (정확한 seq값을 가지고 있지 않다.)

가설

가설

질문

가설

현재 Key Value Pair의 Sequence Number로 Visibility를 확인하는 부분은 SaveValue,

즉 3중 루프 중에서 가장 안쪽에 있다. 이외의 과정에서 Sequence Number를 사용하는 부분은 발견되지 않았다.

그러나 각 파일의 메타데이터에 smallest sequence number, largest sequence number 가 유지되고 있음을 확인했고, 이 FileMetaData 형식의 변수를 사용하는 곳은 가장 바깥쪽의 루프인 Version::Get 에서 Cache에 해당 파일의 블록을 올릴 때임을 확인했다.

이 값을 이용해서 첫번째 루프에서 시간을 절약할 수 있으리라 생각된다.
(smallest sequence number가 현재 sequence number 보다 크다면 해당 파일에 대한 순회를 생략)

다만 이 방식은 sequence number에 대한 메타데이터를 관리하는 단위가 너무 크기 때문에, data block 단위로도 관리하는 것이 좀 더 좋을 것이라 생각한다.

질문

1. 지금 관리되고 있는 smallest_seqno, largest_seqno 만으로는 너무 metadata가 rough하다고 생각한다. 그렇지만 bloom filter는 대소 관계를 봐야 하는 sequence number 의 관리에 적합하지 않다.

   “어떤 값 N 이하의 값이 포함되어 있는지”에 대해 판단하는 거라면, 이진트리와 같은 구조가 적합할 듯한데, 이는 너무 공간을 많이 차지할 것이다. Data block 별로 seq의 최대 최소만 저장하는 것도 괜찮을지 궁금하다.

2.

THANK YOU!

감사합니다!