Lambda 구조를 적용한 IoT 센서 데이터 수집 환경 설계

황윤영[O], 김수현[*], 신용태[*]
[O]숭실대학교 일반대학원 컴퓨터학과,
[*]숭실대학교 컴퓨터학부
e-mail: {doublewhy, sooo0426}@soongsil.ac.kr[O*], shin@ssu.ac.kr[*]

Design on the IoT Sensor Data Collection Envionment using Lambda Architecture

Yun-Young Hwang[O], Soo-Hyun Kim[*], Yong-Tae Shin[*]
[O]Dept. of Science of Computer, Soongsil University,
[*]Dept. of Science of Computer, Soongsil University

● 요 약 ●

데이터의 양은 기술의 발전과 함께 크게 증가하였다. Hadoop은 빅데이터 분야에서 사용되는 대표적인 빅데이터 처리 플랫폼으로 IoT 분야에서도 사용된다. HDFS(Haddop Distributed File System)는 Hadoop의 코어 프로젝트로 블록 기반의 대용량 데이터 저장소다. 기존의 Hadoop 기반 IoT 센서 데이터 수집 환경은 HDFS를 사용한다. 그러나 HDFS의 Small File로 인한 네임노드의 과부하 문제와 한 번 Import된 데이터의 Update와 Delete를 지원하지 않는 Hadoop의 특징으로 인해 성능과 활용이 제한적이다. 본 논문에서는 기존 Hadoop 기반 IoT 센서 데이터 수집 환경의 단점을 극복하기 위해 Lambda 구조를 적용한 IoT 센서 데이터 수집 환경을 설계한다.

키워드: Lambda Arcitecture, Apache Kudu, Apache Impala

I. Introduction

Hadoop은 Apache 소프트웨어 재단에서 개발한 대표적인 빅데이터 처리 플랫폼으로 여러 분야에서 사용되고 있다. HDFS(Hadoop Distribute File System)는 하둡의 코어 프로젝트로 블록 기반 대용량 데이터 저장소이다. HDFS는 최소 64MB 크기의 파일로 데이터를 저장해, 데이터가 64MB가 되지 않아도 64MB로 저장되는 Small File 문제가 발생한다.

IoT 센서는 디바이스 특성상 크기가 작은 데이터를 지속적으로 생성하여 서버로 전송한다. 서버는 Batch Process를 사용해 전송받은 데이터를 IoT 센서 데이터 수집 환경의 HDFS에 저장한다. 전송 받은 데이터의 크기가 64MB가 되지 않고 Batch가 실행되면 Small File 문제가 발생한다. HDFS의 네임노드는 Small File로 인해 과부하가 발생해 전체적인 시스템의 성능이 저하된다.[1]

HDFS는 데이터를 파일 형식으로 블록에 저장하기 때문에 Import된 데이터에 대한 Update와 Delete가 불가능하다. Apache Kudu는 데이터를 컬럼 기반으로 저장하여 하둡에 Import 되는 데이터에 대한 Update와 Delete가 가능하다. Apache Impala가 쿠두의 질의 처리기로 일반적으로 함께 사용된다.

본 논문의 구성은 다음과 같다. 2장에서는 Lambda 구조의 정의와 특징, 쿠두와 임팔라에 대해 살펴본다. 3장에서는 Lambda 구조를 적용한 IoT 센서 데이터 수집 환경을 설계한다. 4장에서는 결론 및 향후 연구 방향을 제시한다.

II. Preliminaries

1. Lambda Architecture

Lambda 구조는 오래된 데이터를 보관하는 배치 테이블과 실시간 데이터를 가진 실시간 테이블을 JOIN하여 결과를 얻는 구조이다. [Fig.1]은 Lambda 구조의 구성을 나타낸다.

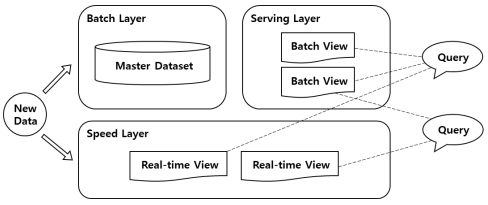

Fig. 1. Lambda Architecture' s Configuration

2. Apache Kudu

Apache Kudu는 하둡 기반의 프레임워크 대부분과 호환되며 하둡의 범용 하드웨어 사용성, 확장성, 데이터 가용성 보증의 특성을 지원한다. 쿠두는 NoSQL OLAP DB로 HDFS에서 불가능한 Update와 Delete 쿼리를 지원한다. 쿠두는 단순한 CRUD만 제공하기 때문에 복잡한 질의를 실행하기 위한 별도의 질의 처리기가 필요하다. [Fig.2]는 쿠두의 서버 구성을 나타낸다.[2]

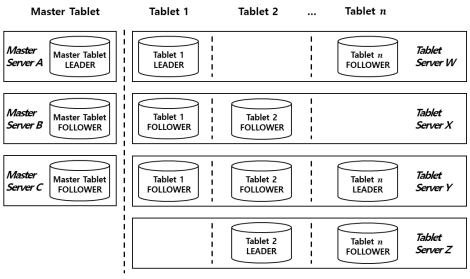

Fig. 2. Apache Kudu Server's Configuration

3. Apache Impala

Apache Impala는 분산 병렬 질의 처리엔진으로 쿠두가 제공하지 않는 연산을 실행한다. [Fig.3]은 임팔라의 구성을 나타낸다.

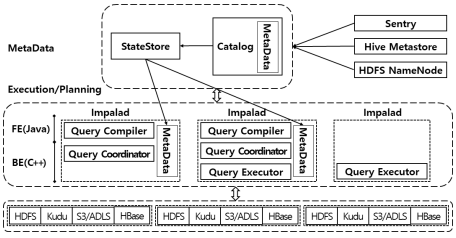

Fig. 3. Apache Impala's Configuration

III. The Proposed Scheme

제안하는 Lambda 구조를 적용한 IoT 센서 데이터 수집 환경은 데이터의 유형에 따라 저장소를 나누어 저장하고 Batch Process의 실행 빈도를 최소화한다. 이를 통해 HDFS에서 발생하는 Small File 문제를 해결하고, Kudu와 Impala를 활용해 데이터의 정확성을 향상시킨다. [Fig.4]는 제안하는 Apache Kudu와 Impala를 활용한 Lambda 구조의 구성을 나타낸다.[3]

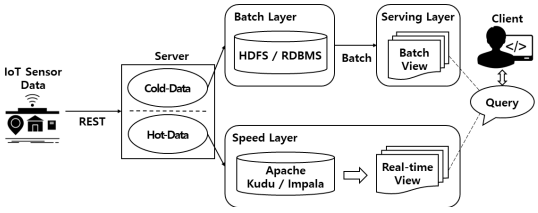

Fig. 4. Suggested Lambda Architecture's Design

IV. Conclusions

제안하는 Lamda 구조 설계는 HDFS에서 생성되는 배치 뷰와 쿠두와 임팔라를 통해 생성되는 실시간 뷰에 접근하는 시간을 단축한다. 또한, 하둡 기반의 환경에서 발생하는 Small File 문제를 해결해 네임노드의 과부하 문제가 없다. 향후 해당 설계의 구축과 구체적인 성능평가가 필요하다.

ACKNOWLEDGEMENT

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2018-0-00209)

REFERENCES

[1] Bende, Sachin, and Rajashree Shedge. "Dealing with small files problem in hadoop distributed file system." Procedia Computer Science 79 pp. 1001-1012, 2016.

[2] Lipcon, Todd, et al. "Kudu: Storage for fast analytics on fast data." Cloudera, inc 28, 2015.

[3] KORNACKER, Marcel; ERICKSON, Justin. Cloudera impala: Real time queries in apache hadoop, for real. Ht Tpblog Cloudera Comblog201210cloudera-Impala-Real-Time-Querie—-Apache-Hadoop--Real, 2012.