Sqoop

Sqoop이란 무엇인가?

- Apache Sqoop (TM)은 구조화된 관계형 데이터 베이스와 Apache Hadoop의 대용량 데이터들을 효율적으로 변환하여 주는 도구임

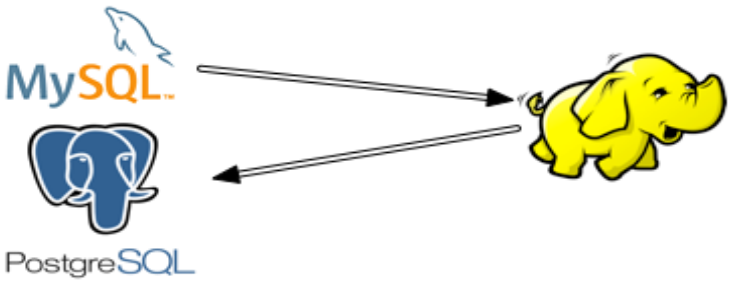
- 오라클 또는 MySQL같은 관계형 데이터 베이스에서 하둡 분산 파일 시스템으로 데이터들을 가져와서 그 데이터들을 하둡 맵리듀스로 변환을 하고, 그 변환된 데이터들을 다시 관계형 데이터 베이스로 내보낼 수 있음

- Sqoop은 데이터의 가져오기와 내보내기를 맵리듀스를 통해 처리하여 장애 허용 능력뿐만 아니라 병렬 처리가 가능함

- Sqoop은 2012년 3월 최상위 아파치 프로젝트로됨

Sqoop이란 무엇인가?

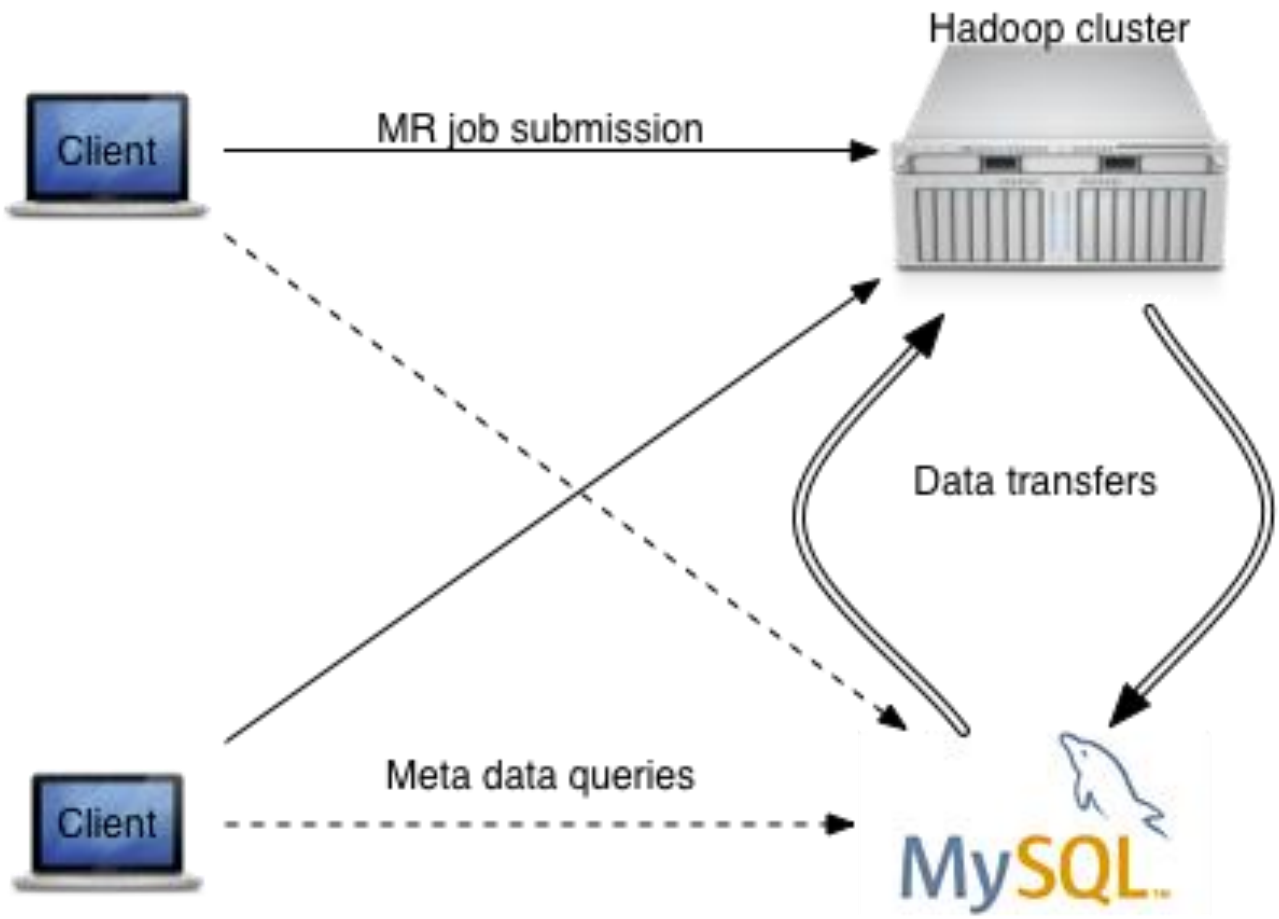
- Apache Top-Level Project
- SQl to hadOOP = SQOOP
- 관계형 데이터베이스에서 데이터를 전송하는 도구
  - Oracle, MySql, PostgreSQL, Teradata, Netezzza
- 하둡생태계
  - HDFS(text, sequence file), Hive, Hbase, Avro, ...



왜 Sqoop인가?

- 자원에 대한 처리를 효과적이고 효율적으로 제어 가능함
  - Concurrent connections, Time of operation
- 데이터 타입 mappin과 conversion
  - 자동화, 사용자 재정의
- Metadata의 적용
  - Sqoop Record
  - Hive Metastore
  - Avro
    - Apache Avro is a data serialization system

Sqoop 1

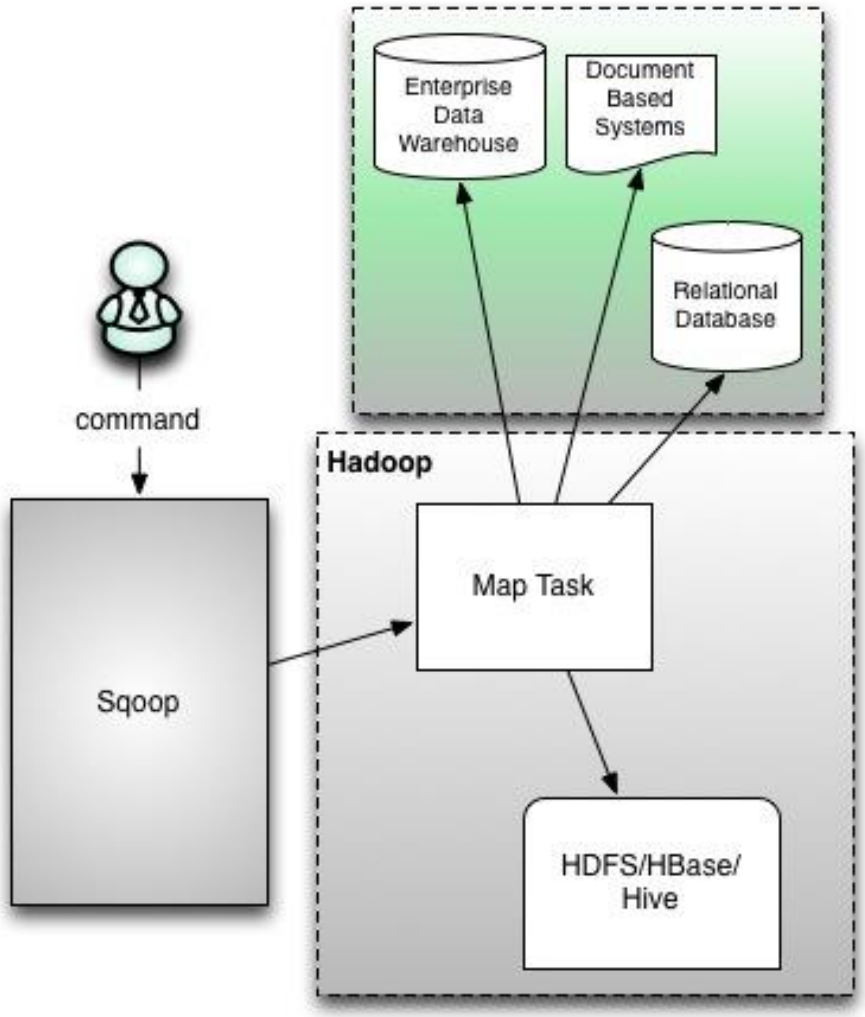

Hadoop cluster
Client
MR job submission
Data transfers
Client
Meta data queries
MySQL

Sqoop 1

- Connectors 기반
  - Metadata 조회, 데이터 전송
  - Connector는 JDBC기반으로 처리
  - Non-JDBC Connector(단지 MySql, Postgresql만 지원됨)

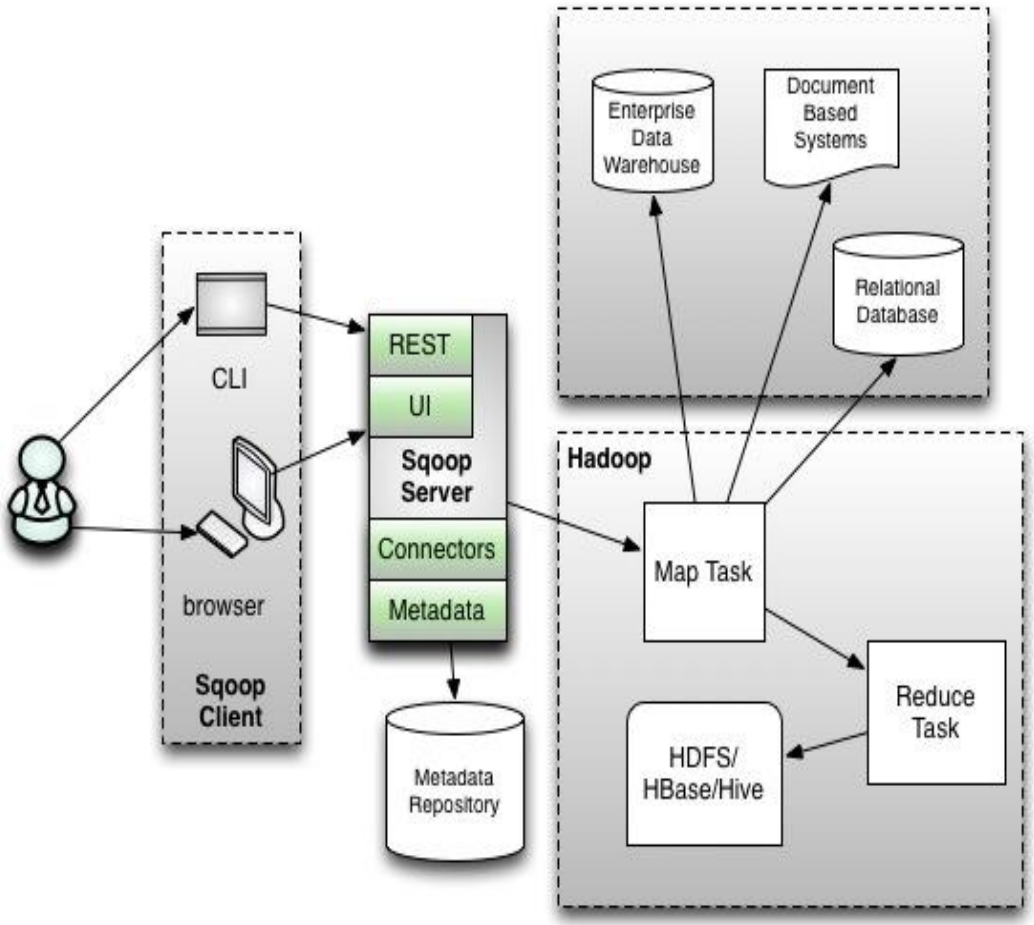
- Connectors는 모든 기능을 담당함
  - Hbase Import, Avro Support, …

Sqoop 1 문제점



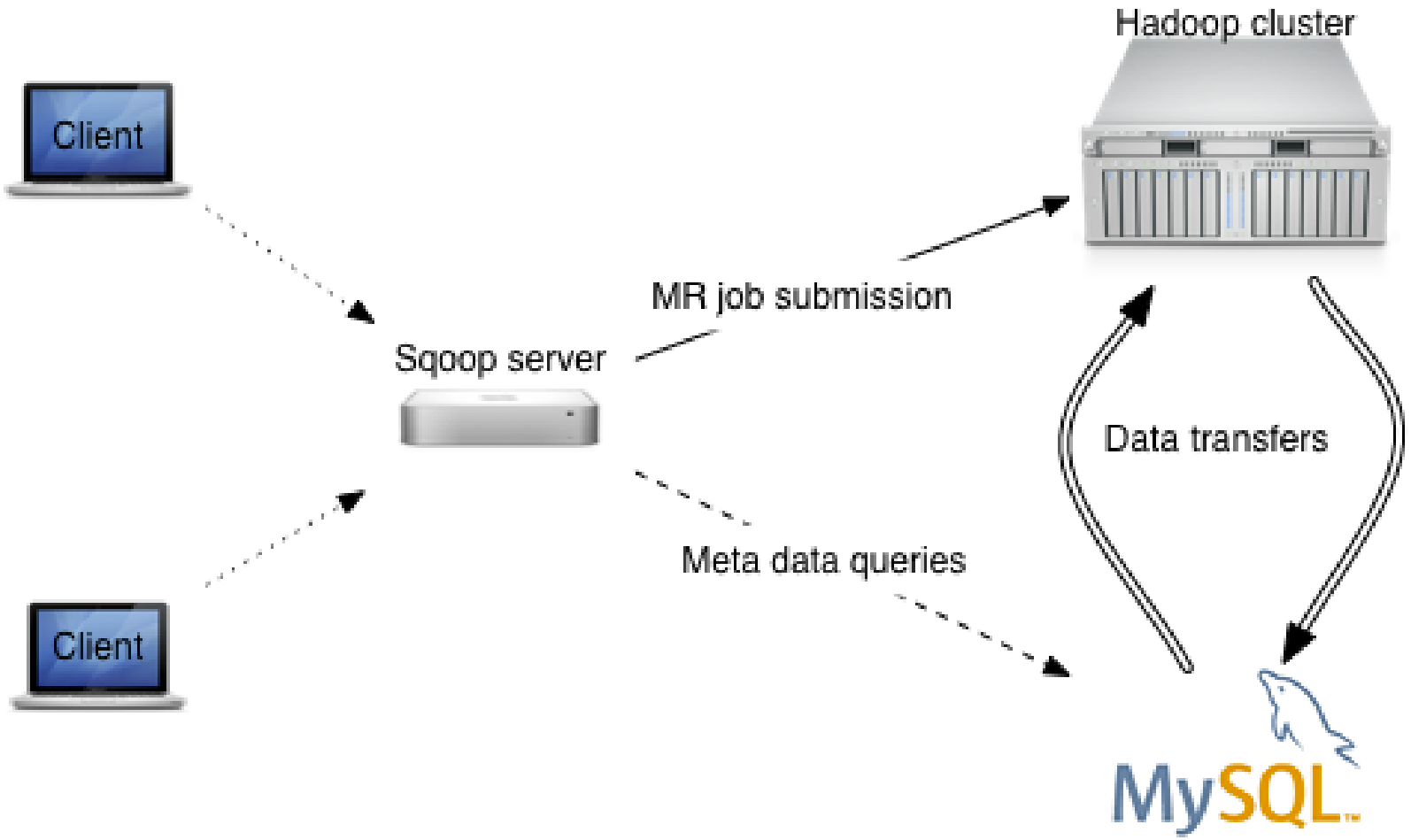
- Client-Side 설치
- Connectors가 Local에 설치됨
- JDBC Driver는 접속하는 Local마다 설치가 필요함
- CLI(Command Line Interface)제공

Sqoop 2



- Server-Side 설치
- Connetors가 필요한 서버 한 곳에만 설치하여 연결 가능함
  - JDBC Driver가 한곳에만 설치되면됨
- CLI접속 외에도 Web및 REST API를 통한 접속도 가능함
- Workflow Manager인 Apache Oozie와 통합하여 처리 가능함

Sqoop 2

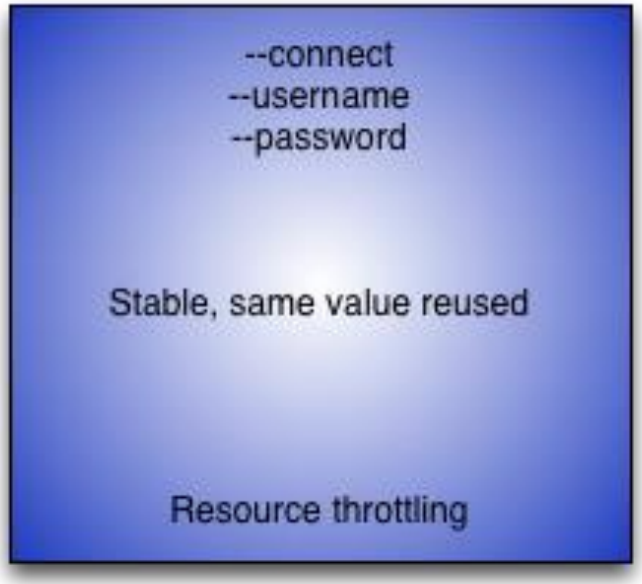

Hadoop cluster
Client
MR job submission
Sqoop server
Data transfers
Meta data queries
Client
MySQL

Sqoop 2(Design Goals)

- 보안 및 기능별 분리
  - 역할 기반 엑세스 및 사용
- 확장성이 용이함
  - 하둡에 대한 지식이 없어도 사용가능
  - Connectors간의 기능 중복성이 없음
- 사용의 용이함
  - 통합된 기능
  - 시나리오별 용이한 구성이 가능함

Sqoop 2(Connection vs Job Metadata)

- 두가지 구분된 Options
  - Connection(각 데이터베이스당)



Job(각 테이블당)

Sqoop 2(Workings)

- Metadata에 Connectors 등록함
- Metadata에 Connections와 Jobs을 생성할 수 있음
- Connections와 Jobs은 Metadata Repository에 저장됨
- 운영자는 적합한 Connections을 Jobs에서 실행가능함
- 관리자가 Connection사용에 대한 정책을 정의할 수 있음

Sqoop 2(Security)

- 역할 기반 액세스 객체를 통한 외부 시스템에 대한 보안 액세스 지원
  - 관리자가 연결 만들기 / 수정 / 삭제
  - 운영자는 Connection을 사용함

Sqoop 1 & 2 버전 상태

- http://sqoop.apache.org/
- Sqoop 1 version
  - 1.4.7
- Sqoop 2 version
  - 1.99.7
- Sqoop 1.99.7 version은 1.4.7버전과 호환되지 않음

Sqoop 1 실습

- 서버명01 : Data(MySql 또는 MariaDB설치)
- 서버명02 : Server01(Hadoop
  - Sqoop를 Server01(Hadoop MasterNode)에 설치하고 Data에 설치된 MySql의 Employees 테이블을 Server01에 Import하고 저장된 Employees데이터를 다시 MySql의 Employees2테이블(Empty테이블)에 Export해보도록 한다.