빅데이터의 데이터 파이프라인은 어디에서 데이터를 수집하여 무엇을 실현하고 싶은지에 따라 변화한다. 처음에는 간단한 구성으로도 끝나지만, 하고 싶은 일이 증가함에 따라 시스템은 점차 복잡해지고 그것을 어떻게 조합시킬지가 문제가 된다.

**데이터 수집** 벌크 형과 스트리밍 형의 데이터 전송

데이터 파이프라인은 데이터를 모으는 부분부터 시작한다. 데이터는 여러 장소에서 발생하고 각각 다른 형태를 보인다. 데이터베이스에 쓰인 거래처 데이터, 파일 서버에 축적된 로그 파일, 스마트 폰 등의 모바일 애플리케이션에서 모여진 이벤트 데이터 및 임베디드(embedded) 장비에서 보내진 센서 데이터 등 각각 서로 다른 기술로 데이터를 전송한다.

**'데이터 전송**(data transfer)'의 방법은 크게 다음의 두 가지가 있다(그림 1.4 ❶ ❷).

- 벌크(bulk) 형
- 스트리밍(streaming) 형

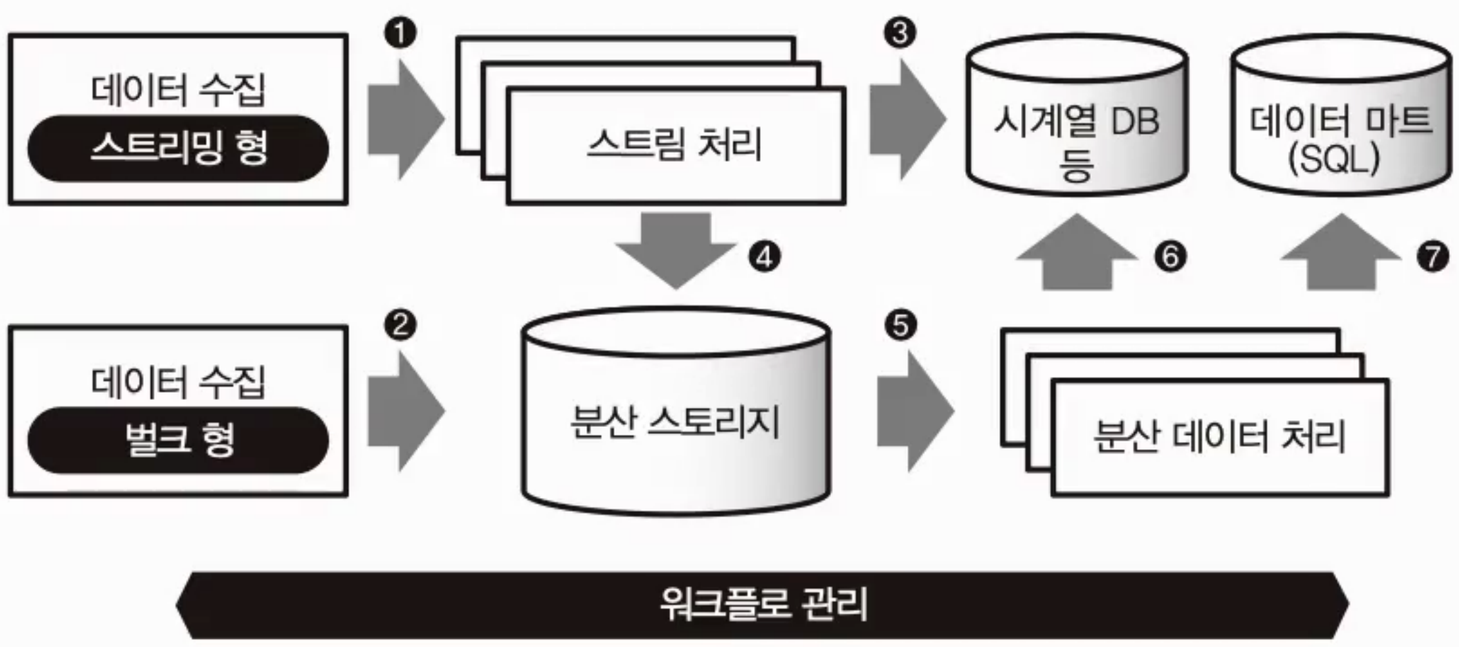


그림 1.4 빅데이터를 위한 데이터 파이프라인

벌크 형은 이미 어딘가에 존재하는 데이터를 정리해 추출하는 방법으로, 데이터베이스와 파일 서버 등에서 정기적으로 데이터를 수집하는 데에 사용한다. 한편, 스트리밍 형은 차례차례로 생성되는 데이터를 끊임없이 계속해서 보내는 방법으로 모바일 애플리케이션과 임베디드 장비 등에서 널리 데이터를 수집하는 데 사용된다.