Linear Regression, Normal Equation, QR Factorization, Numerical Stability, RHadoop

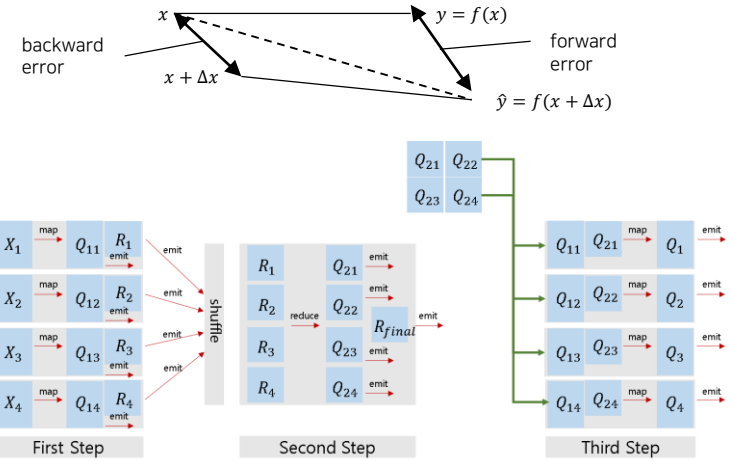
Numerically Stable Algorithms estimating Linear Regression Coefficients in RHadoop

Purpose

- RHadoop에서의 **QR분해**를 이용한 선형회귀계수 추정 알고리즘 개발
- 기존 **병렬 Normal Equation** 알고리즘과 비교

Methodology

- **Direct TSQR** 알고리즘을 수정, 보완하여 알고리즘 작성
- **수치적 안정성(numerical stability)**과 시스템 처리시간을 기준으로 비교



Environment

노드	master	1
	slave	4
서버 사양	CPU	Intel i7-8700K
	RAM(master/slave)	64GB / 32GB
	HDD	1TB
소프트웨어 버전	OS	Ubuntu 16.04 LTS
	Java	1.8.0
	Hadoop	2.6.0
	R	3.4.4
	rhdfs	1.0.8
	rmr2	3.3.1

Issues

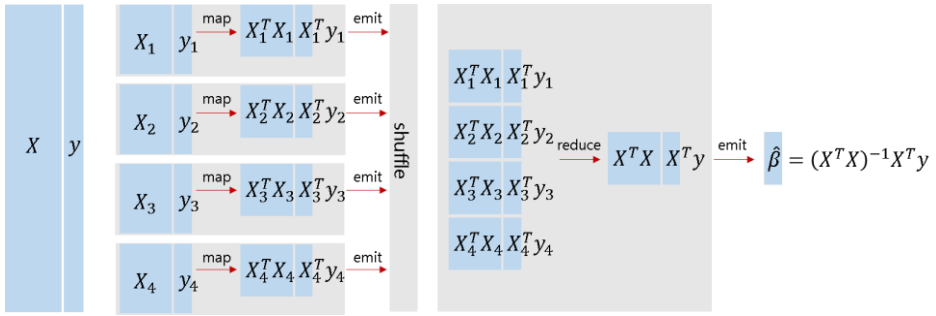
기존의 Direct TSQR 알고리즘은 **총 3단계의 맵리듀스 과정**을 거치기 때문에 비효율적

Solutions

두 단계에 걸쳐 계산되는 Q_1, Q_2를 마지막에 결합하지 않고, **계산될 때마다 블록 별로 필요한 연산을 수행** (다음 페이지 참고)

Numerically Stable Algorithms estimating Linear Regression Coefficients in RHadoop

1) Algorithms

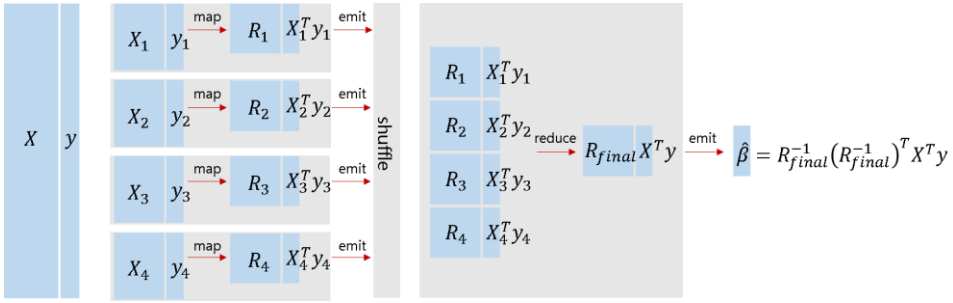


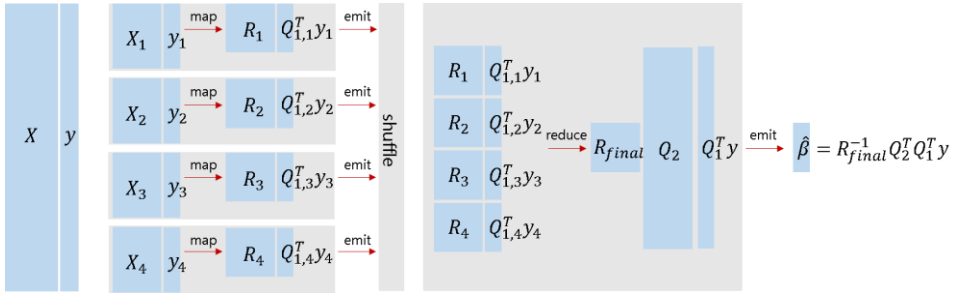
기존 병렬 Normal Equation 알고리즘

시스템 처리시간 빠르지만, 단일 컴퓨팅에서의 연산과 동일한 연산을 수행하기 때문에 수치적으로 불안정할 것으로 예상

Indirect TSQR 수정 보완한 선형회귀계수 추정 알고리즘

기존 Indirect TSQR 알고리즘이 QR의 Q를 구하는데 있어 수치적으로 불안정하기 때문에 수치적으로 불안정할 것으로 예상



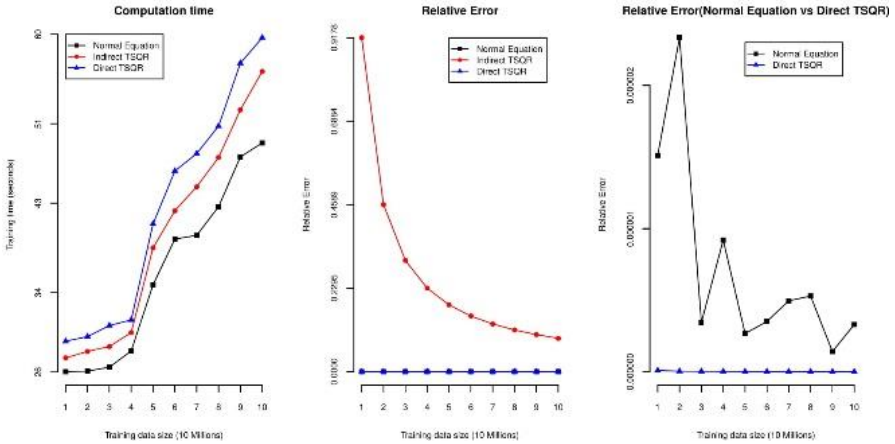


Direct TSQR 수정 보완한 선형회귀계수 추정 알고리즘

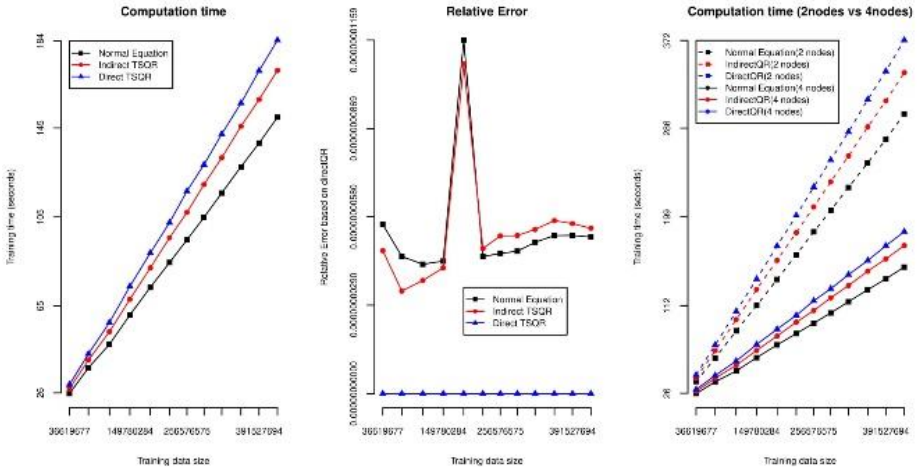
수치적으로 안정적이고 시스템 처리시간은 다른 알고리즘보다 조금 느릴 것으로 예상

Numerically Stable Algorithms estimating Linear Regression Coefficients in RHadoop

2) Results



sparse한 자료에 대한 각 알고리즘의 시스템 처리시간(좌),
상대 오차(가운데, 우)



뉴욕 옐로우 택시 자료에 대한 각 알고리즘의 시스템 처리시간(좌), 상대 오차(가운데)
노드 수 2개일 때와 4개일 때의 시스템 처리시간 비교(우)

결론

수치적 안정성 : Normal Equation≈ Indirect TSQR ≪Direct TSQR
시스템 처리시간 : Normal Equation ≤ Indirect TSQR ≤ Direct TSQR