SNS data를 이용한 Hadoop & DBMS의 성능비교

Comparison between Hadoop and DBMS for SNS data

pidl Petabyte-scale In-memory Database Lab. 이창형 한겨레

INTRODUCTION

What is Hadoop?

Apache Hadoop is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of nodes and petabytes of data. Hadoop was inspired by Google's MapReduce and Google File System (GFS) papers.

Map & Reduce

MapReduce is a software framework introduced by Google in 2004 to support distributed computing on large data sets on clusters of computers. Parts of the framework are patented in some countries.

What is DBMS?

- Database : a very large, integrated collection of data
  - Models real-world enterprise
    - Entities(e.g., customers, supllies, products)
    - Relationships(e.g., John orders two wifi i-pads)
- A Database Management System (DBMS) is a software package designed to store and manage databases

What is PageRank?

PageRank is a link analysis algorithm, named after Larry Page and used by the Google Internet search engine, that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references.

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + .. + \frac{PR(Tn)}{C(Tn)}\right)$$

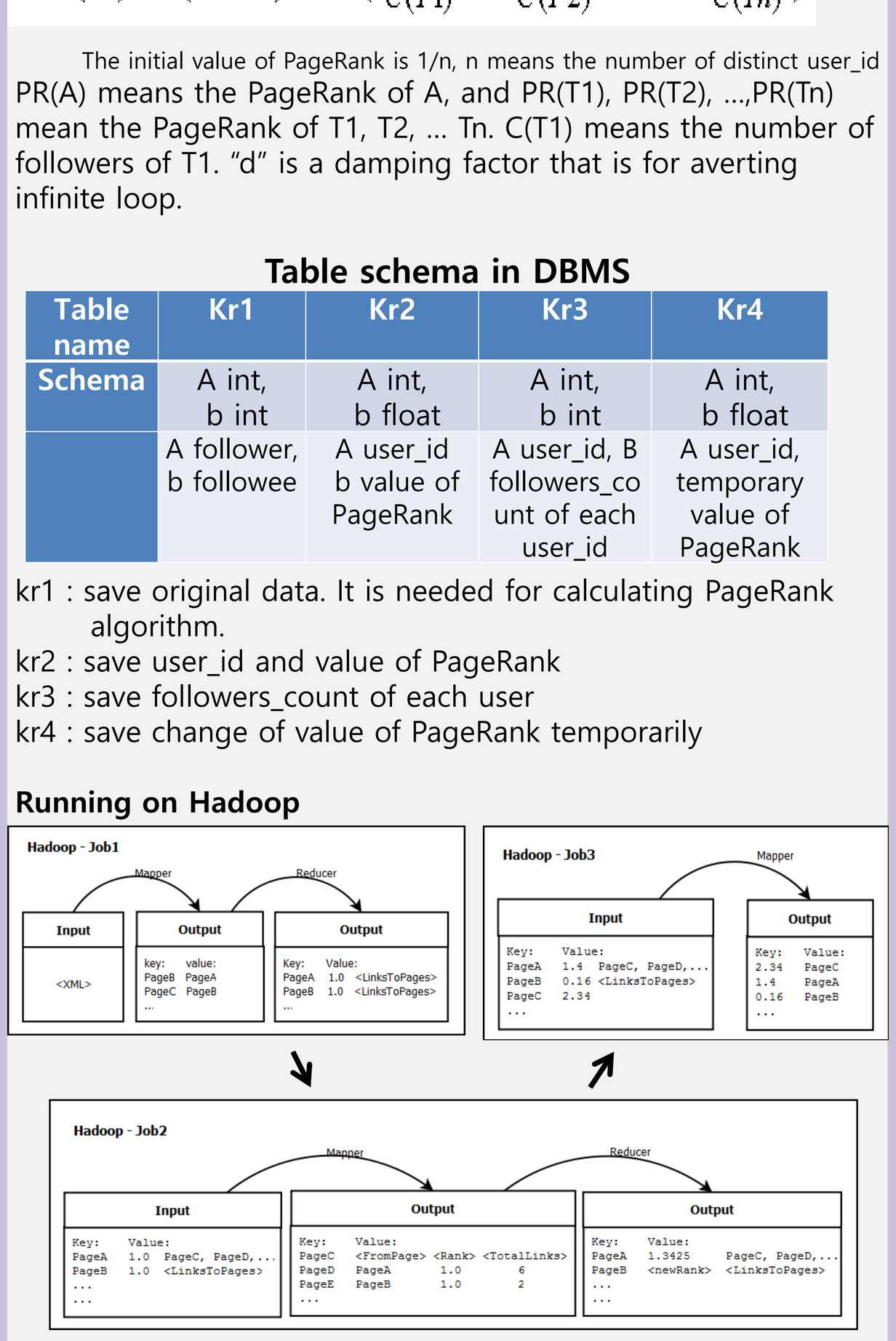
The initial value of PageRank is 1/n, n means the number of distinct user_id PR(A) means the PageRank of A, and PR(T1), PR(T2), ...,PR(Tn) mean the PageRank of T1, T2, ... Tn. C(T1) means the number of followers of T1. "d" is a damping factor that is for averting infinite loop.

Table schema in DBMS

Table name	Kr1	Kr2	Kr3	Kr4
Schema	A int, b int	A int, b float	A int, b int	A int, b float
	A follower, b followee	A user_id b value of PageRank	A user_id, B followers_count of each user_id	A user_id, temporary value of PageRank

kr1 : save original data. It is needed for calculating PageRank algorithm.
kr2 : save user_id and value of PageRank
kr3 : save followers_count of each user
kr4 : save change of value of PageRank temporarily

Running on Hadoop

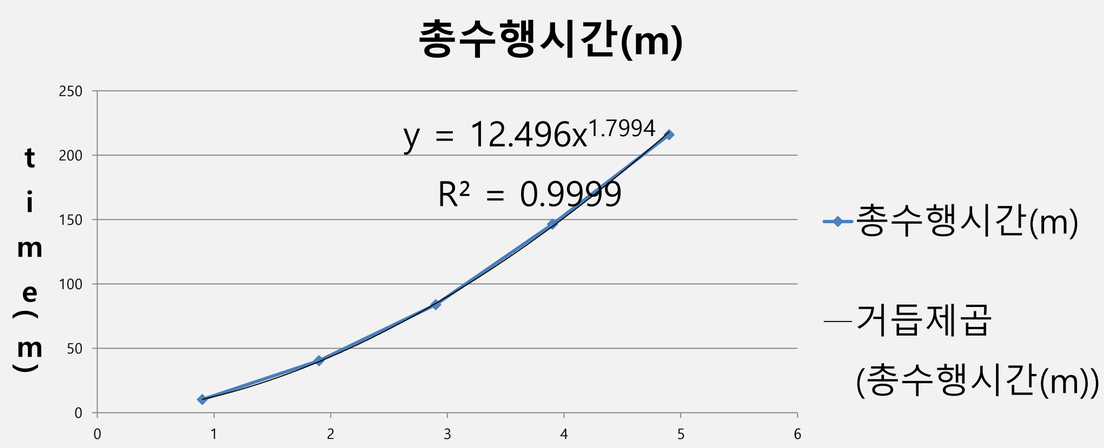


Result of DBMS

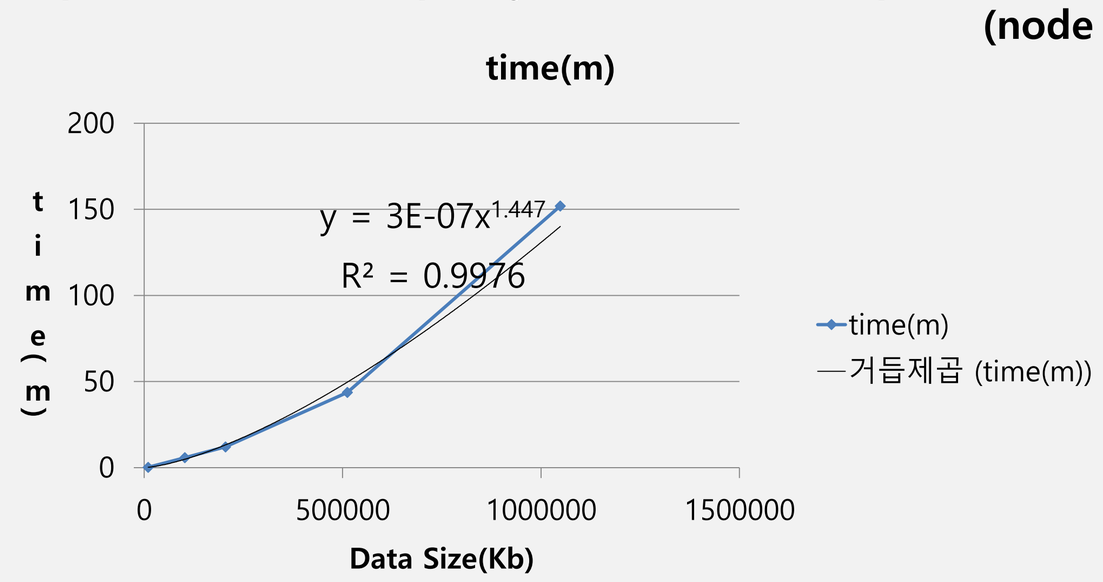
- **Comparison of performance between row store and column store in DBMS**

4.9G		총수행시간(m)	loop 1회 평균 수행시간(s)	row->column
Row table	insert	546	1,091.5	시간 단축률
	update	25	50.0	
	sum	581	1,141.5	
Column table	insert	162	323.1	70.40%
	update	54	108.7	-117.37%
	sum	216	431.9	62.17%

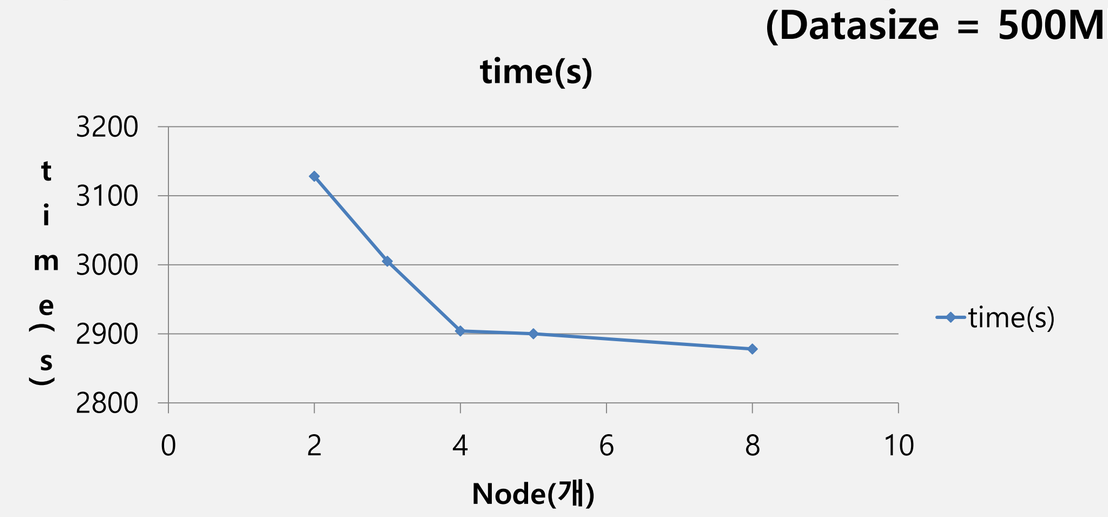
- **performance as capacity of data in DBMS**



- **performance as capacity of data in Hadoop (node 8)**



- **performance as number of node in Hadoop (Datasize = 500Mb)**



- **performance between DBMS and Hadoop (DataSize = 1Gb)**