MapReduce

Tomáš Vlk (vlktoma5@fit.cvut.cz)

May 26, 2019

Úvod

MapReduce je programovací model používaný pro distribuované paralelní výpočty. Původní návrch pochází z Googlu.

Základní myšlenka

Základní myšlenkou MapReduce přístupu je distrubuce dat na navzájem nezávislé části. Důsledkem toho lze dosáhnout snadného zpracování v jednotlivých uzlech.

Další důležitou myšlenkou MapReduce je zpracovávání dat v místě jejich uložení. Tudíž není nutné zbytečně data přesouvat a je tak dosaženo lepší škálovatelnosti.

Samotný přístup se pak dělí na dvě části:

1. Mapování
2. Redukce

Popis částí

Nyní poskytneneme popis a příklady obou částí MapReduce přístupu.

Mapovací funkce

Vstupem mapovací funkce je seznam prvků, najehož prvky je následně aplikovaná transformace. Mapovací funkce je bezstavová a příkladem takovéto funkce může například být převod stringů na velká písmena.

Redukční funkce

Redukční funkce pouze agreguje seznam hodnot, který jí je poskytnut, do obvykle menšího množství hodnot. Příkladem takové funkce může být i jednoduchý součet.

MapReduce přístup

MapReduce kombinuje oba přístupy zmíněné výše. První je mapovací funkce, po které následuje shuffle a nakonec reduční funkce. U každé části je uvedený příklad na součtu počtu slov stejné délky.

Mapování

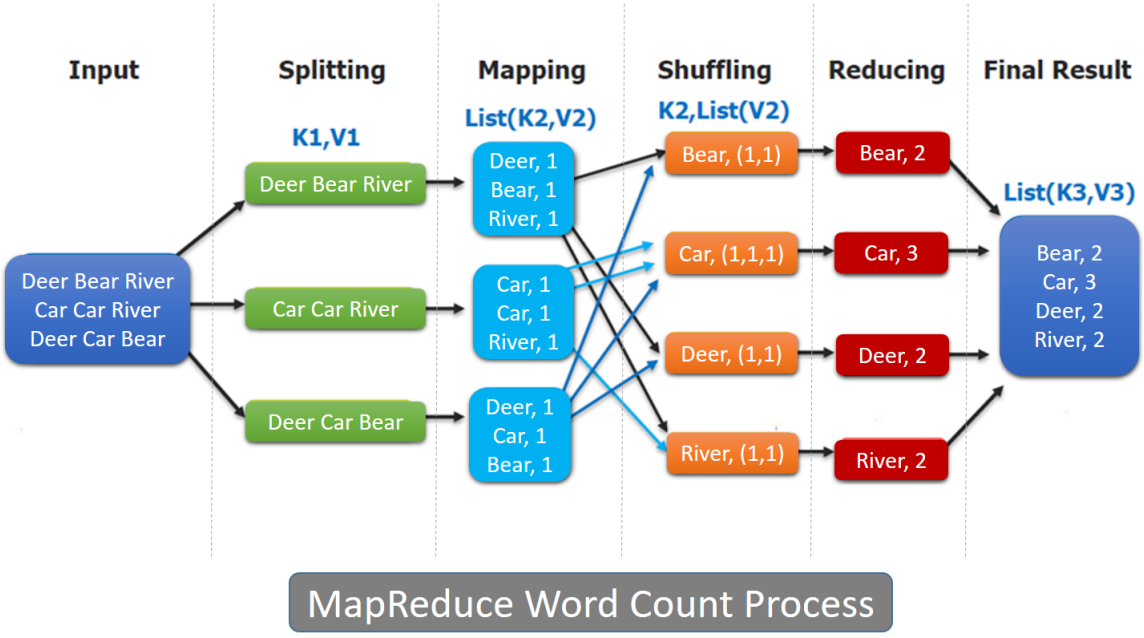
Vstupem do mapovací funkce je seznam hodnot. Mapovací funkce za něj vytvoří pár ¡klíč, hodnota¿. V případě počítání slov vytvoří pár ¡délka slova, slovo¿.

Shuffle

Vstupem do shuffle funkce je list párů vytvořených pomocí mapovací funkce. Shuffle je promíchá a předá redukční funkci. Každé redukční funkce dostane páry se stejným klíčem.

Redukce

Redukční funkce provede agregaci párů se stejným klíčem. V případě počítání slov, redukční funkce pouze sečte výskyty prvků se stejným klíčem.



MapReduce Word Count Process

Využití

MapReduce je používaný v Googlu pro generování Google indexu, nebo v Hadoopu ve spojeni s HDFS.