

[SUBSCRIBE](#)[SIGN IN](#)

BIZ & IT —

Google's MapReduce patent: what does it mean for Hadoop?

Google has obtained a patent on MapReduce, a technique for efficient ...

RYAN PAUL - 1/20/2010, 10:10 AM

The USPTO awarded search giant Google a software method patent that covers the principle of distributed MapReduce, a strategy for parallel processing that is used by the search giant. If Google chooses to aggressively enforce the patent, it could have significant implications for some open source software projects that use the technique, including the Apache Foundation's popular Hadoop software framework.

"Map" and "reduce" are functional programming primitives that have been used in software development for decades. A "map" operation allows you to apply a function to every item in a sequence, returning a sequence of equal size with the processed values. A "reduce" operation, also called "fold," accumulates the contents of a sequence into a single return value by performing a function that combines each item in the sequence with the return value of the previous iteration.



Join Ars Technica and

Get Our Best Tech Stories

DELIVERED STRAIGHT TO YOUR INBOX.

SIGN ME UP

By signing up, you agree to our user agreement (including the class action waiver and arbitration provisions), our privacy policy and cookie statement, and to receive marketing and account-related emails from Ars Technica. You can unsubscribe at any time.

Google's MapReduce framework is roughly based on those concepts. A series of data elements is processed in a map operation, then combined at the end with a reduce operation to produce the finished output. The advantage of partitioning a workload this way is that it's extremely conducive to parallelization. Each discrete unit of data in the series can be processed individually and combined at

the end, making it possible to spread the workload across multiple processors or computers. It's a fairly elegant approach to scalable concurrency, one that offers efficiency regardless of whether your environment is a single multicore processor or a massive grid in a data center.

Google **published a paper** in 2004 that described how it uses MapReduce. The paper attracted considerable interest and paved the way for the MapReduce pattern to become a common technique for parallelization. One of the most well-known third-party implementations of MapReduce for distributed computing is Hadoop, an open source Apache project now used by Yahoo, Amazon, IBM, Facebook, Rackspace, Hulu, the New York Times, and a growing number of other companies.

Google's patent on MapReduce could potentially pose a problem for those using third-party open source implementations. Patent #7,650,331, which was granted to Google on Tuesday, defines a **system and method for efficient large-scale data processing**:

A large-scale data processing system and method includes one or more application-independent map modules configured to read input data and to apply at least one application-specific map operation to the input data to produce intermediate data values, wherein the map operation is automatically parallelized across multiple processors in the parallel processing environment. A plurality of intermediate data structures are used to store the intermediate data values. One or more application-independent reduce modules are configured to retrieve the intermediate data values and to apply at least one application-specific reduce operation to the intermediate data values to provide output data.

Hadoop isn't the only open source project that uses MapReduce technology. As some readers may know, I've recently **been experimenting** with CouchDB, an open source database system that allows developers to perform queries with map and reduce functions. Another place where I've seen MapReduce is Nokia's **QtConcurrent framework**, an extremely elegant parallel programming library for Qt desktop applications.

It's unclear what Google's patent will mean for all of these MapReduce adopters. Fortunately, Google does not have a history of aggressive patent enforcement. It's certainly possible that the company obtained the patent for "defensive" purposes. Like virtually all major software companies, Google is **frequently the target of patent lawsuits**. Many companies in technical fields attempt to collect as many broad patents as they can so that they will have ammunition with which to retaliate when they are faced with patent infringement lawsuits.

Google's MapReduce patent raises some troubling questions for software like Hadoop, but it looks unlikely that Google will assert the patent in the near future; Google itself **uses Hadoop** for its Code University program.

Even if Google takes the unlikely course of action and does decide to target Hadoop users with patent litigation, the company would face significant resistance from the open source project's deep-pocketed backers—including IBM, which holds the industry's largest patent arsenal.

Another dimension of this issue is the patent's validity. On one hand, it's unclear if taking age-old principles of functional software development and applying them to a cluster constitutes a patentable innovation. On the other hand, Google's MapReduce paper indisputably popularized the concept and is freely characterized by Hadoop's developers as the inspiration behind their project. This suggests that Google is owed some credit by the industry for advancing distributed computing with its MapReduce paper, a factor that could strengthen the patent.

Listing image by [Han Soete](#)

READER COMMENTS 21

RYAN PAUL

Ryan is an Ars editor emeritus in the field of open source, and still contributes regularly. He manages developer relations at Montage Studio.



Unsolved Mysteries Of Quantum Leap With Donald P. Bellisario

Today "Quantum Leap" series creator Donald P. Bellisario joins Ars Technica to answer once and for all the lingering questions we have about his enduringly popular show. Was Dr. Sam Beckett really leaping between all those time periods and people or did he simply imagine it all? What do people in the waiting room do while Sam is in their bodies? What happens to Sam's loyal ally Al? 30 years following the series finale, answers to these mysteries and more await.



Unsolved Mysteries Of Quantum Leap With Donald P. Bellisario



Unsolved Mysteries Of Warhammer 40K With Author Dan Abnett



SITREP: F-16 replacement search a signal of F-35 fail?



Sitrep: Boeing 707

[+ More videos](#)

[← PREVIOUS STORY](#)[NEXT STORY →](#)

Related Stories

Hadoop becomes more business-friendly with HortonWorks and VMware partnership

Teradata goes after "big data" with Hadoop-SQL hybrid

The hot new technology in Big Data is decades old: SQL

Go: new open source programming language from Google

Today on Ars

CDC is slashing funding for states' childhood vaccination data systems

Android phone hits 24GB of RAM, as much as a 13-inch MacBook Pro

State Dept. cancels election meetings with Facebook after "free speech" ruling

Review: Forcite's MK1S smart helmet stays on your head, not in your face

Volkswagen will start testing its driverless ID. Buzz in Austin, Texas

June extremes suggest parts of climate system are reaching tipping points

India, a growing space power, is forging closer ties with NASA

US maternal deaths more than doubled over two decades, study estimates

[STORE](#)
[SUBSCRIBE](#)

[CONTACT US](#)
[STAFF](#)

[NEWSLETTER SIGNUP](#)

[ABOUT US](#)
[RSS FEEDS](#)
[VIEW MOBILE SITE](#)

[ADVERTISE WITH US](#)
[REPRINTS](#)

Join the Ars Orbital Transmission mailing list to get weekly updates delivered to your inbox. [Sign me up →](#)



CNMN Collection
WIRED Media Group
© 2023 Condé Nast. All rights reserved. Use of and/or registration on any portion of this site constitutes acceptance of our User Agreement (updated 1/1/20) and Privacy Policy and Cookie Statement (updated 1/1/20) and Ars Technica Addendum (effective 8/21/2018). Ars may earn compensation on sales from links on this site. Read our affiliate link policy.
Your California Privacy Rights | [Cookies Settings](#)
The material on this site may not be reproduced, distributed, transmitted, cached or otherwise used, except with the prior written permission of Condé Nast.
[Ad Choices](#)