imaginea

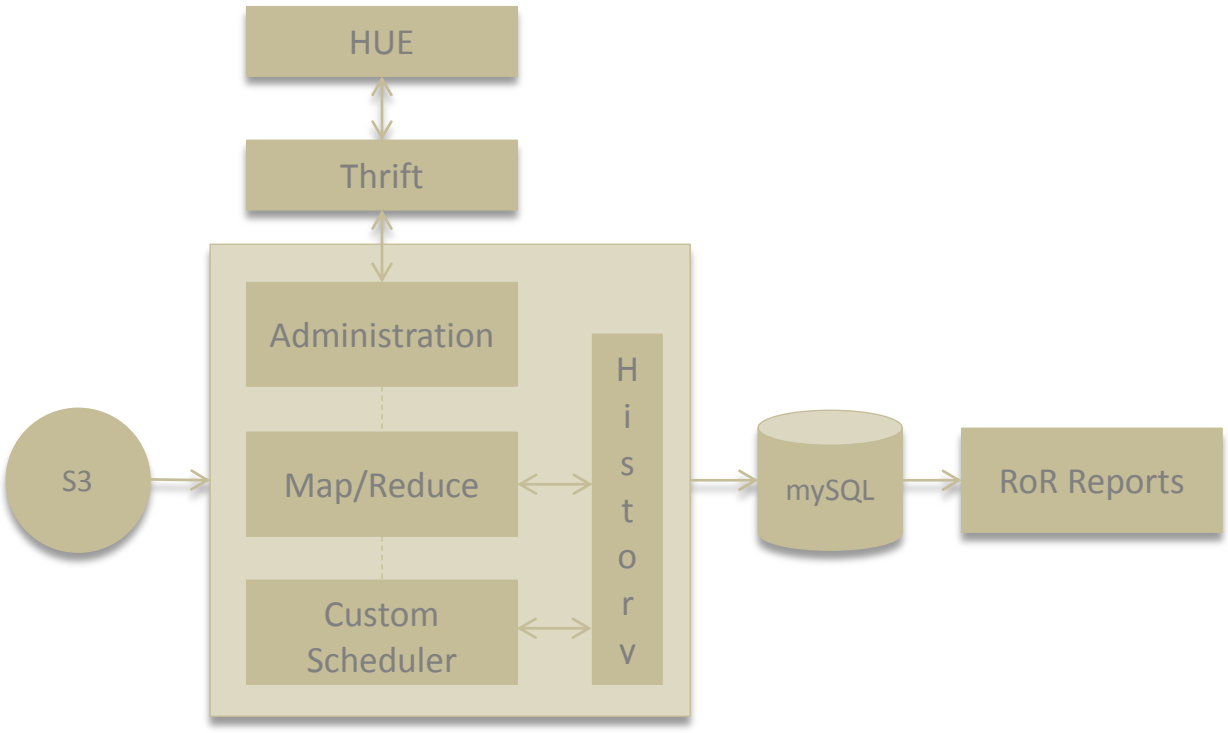
# Using Hadoop MapReduce to analyze terrabytes of web activity logs

## Social Marketing Platform – Activity Processing

Millions of people go about their day in the maze of internet, navigating hundreds of pages - liking, clicking, sharing and tweeting. In this process, they leave behind valuable crumbs of information everywhere. As an observer, how do you make sense of it all? More importantly, *how fast can you react?*

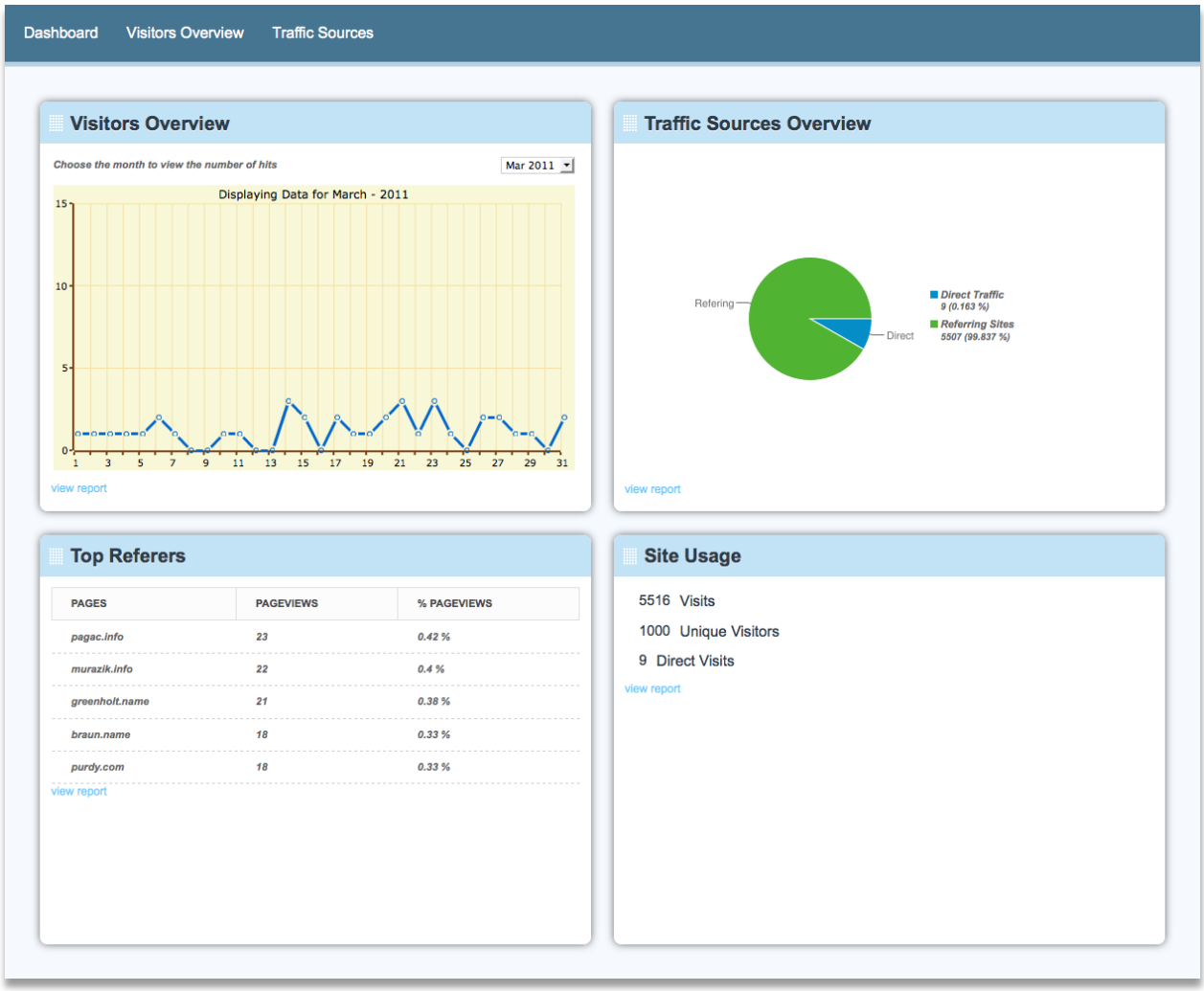
SocialTwist helps you do just that by using Hadoop Map/Reduce, a software framework for writing applications which process multi-terabytes of data in parallel on large clusters of commodity hardware, all this in a reliable, fault-tolerant way.

What's SocialTwist? It's a social marketing platform that promotes highly viral, word-of-mouth marketing campaigns online, through its embeddable widget, Tell-A-Friend. This viral leads to continuous exponentially rising rate of sharing that creates tons of information that needs to be processed.



# Actionable Information – Analysis & Reporting

SocialTwist's homegrown solution is fast on a single machine but cannot be parallelized to multiple machines. So it takes days to analyze terabytes of weblogs, looking for key pieces of information. They needed a faster solution that is easy to develop and implement over their existing infrastructure. Therefore, we applied Hadoop Map/Reduce and were able to analyze terabytes of web activity logs at just one-tenth of the processing time. SocialTwist can now relay key customer behavior and campaign efficacy information that much faster to their clients.




imaginea

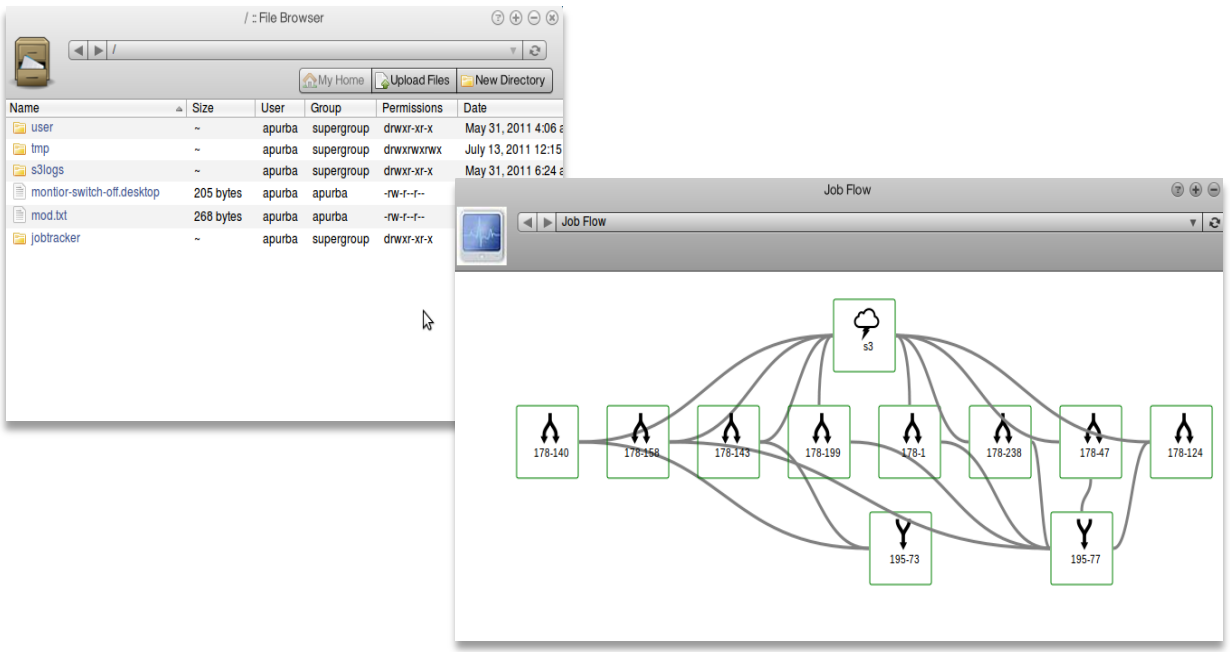
Crunching and analyzing data is one thing but presenting it in an actionable and usable format is quite another. To convey the bigger picture out of these crumbs of information, we built a front-end using Ruby on Rails. Not only does it give detailed information on the underlying tasks and data in the database, it also serves as a dashboard of vital metrics like Top Sites and Top Users.

Together with the speed of processing data, the reporting mechanism helps relay campaign efficacy data in a timely and actionable manner.

## The new HUE – Administration & Management

Moreover, we made it extremely simple to administer Hadoop Map/Reduce tasks by modifying HUE (Hadoop User Experience) - a web UI for Hadoop, started by Cloudera. It essentially is a suite of web applications like User Admin, File Browser, Job Designer, Job Browser and Beeswax.

HUE is also a platform for building custom applications with a nice UI library. The backend of HUE uses Django, a popular MVC Python web framework that understands the application namespaces and encourages rapid development and clean, pragmatic design. The frontend uses MooTools, and allows the developer to reuse common interactive UI elements in a declarative way (via CSS), as opposed to writing lots of JavaScript.




imaginea

Applying our expertise and knowledge of Django, Python & MooTools, we forked out a modified HUE that lets you do more with Hadoop. It fills some of the gaps left by HUE and goes further in managing and administering Map/Reduce tasks.

- You can plan job execution better by deciding how many nodes you would need. There is also a wizard guides on history of job execution.
- You get the best of both worlds with connectivity to Apache Hadoop.
- Also, there is better upload support with uploads to S3.

With richer information of job execution and easy to use UI, the modified HUE takes administration of Hadoop to the next level – making it easier for developers to get more out of Hadoop in a short period of time.

**Technologies**: Hadoop, Ruby on Rails, Python, dJango, Amazon S3, DSS

Copyright © 2010, Pramati Technologies Private Limited. Imaginea is a Pramati business.
All Trade Names and Marks belong to their respective owners.

Imaginea provides product engineering services to independent software vendors, enterprises and online SaaS businesses looking for reliable technology partner. Services stretch end to end, from interaction design to development, testing and managing clouds. For more information on Imaginea, visit www.imaginea.com. Imaginea is a business unit of Pramati Technologies (www.pramati.com)

sales@imaginea.com | www.imaginea.com

imaginea