

## White Paper



### Real Time Recommendations for the New Age Customer

---

Dr. Sujatha R Upadhyaya

#### Abstract

Value of a recommendation grows multifold if it reaches the customer at real time, just when the information is most needed. However, there are practical problems with running real time analytics. The ability to view and process the real time data, and make the recommendations at real time is the most crucial of them. Machine learning algorithms that are used to learn preferences from data and to make personalized recommendations take a long processing time owing to their computation complexity and also because of the need to process large amount data. Recent studies suggest that parallel architecture of graphics processors (GPU) can be exploited to enhance the performance of machine learning algorithms. A GPU/ CPU cluster based methodology to address real time analytics needs is presented in this paper.

## Now or Never - Instant Recommendation is the key

Situations demand that the right recommendations are made quickly while the opportunity lasts. Most real time recommendation opportunities are of this nature. Such situations require that the incoming stream of data is analyzed at real time, results of which are combined with the analysis of past and current behavior of the particular customer, his demographic information, and general user behavior trends to arrive at appropriate personalized recommendations. However, dissemination of such recommendation at a later time / date may not make much sense.

A good example would be an online retail context; where thousands of users logged on to the website. The retailer wants to make personalized recommendations to the most prospective buyers. In this scenario, the situation is dynamic; customers keep entering and leaving site. The selection of personalized recommendation has to be carried out online and sent out as a display of a personalized offer / a personalized advertisement or it may be even a live chat invitation to prospective buyers. Display of personalized advertisement is found to have very high eCPM (effective cost per Mille impressions) value for the publisher and is one of the applications in highest demand today. The most difficult part of such tasks is that they must be accomplished before the user leaves the site.

Long processing requirements, server delays, large amount of real time data (heavy online traffic) and limited number of resources (in case of chat invitations) are the other problems adding to the vow. There are a few applications that handle similar scenarios; however the real time decision (generative and discriminative model)

## Parallel Machine Learning on GPUs

A common grouse about the machine learning / data mining / text mining application is that they are computationally intensive and as a result, tend to be slow with increasing volume of data. So, typical data mining applications depend on offline processing (typically, model building is considered as an offline task) which has obvious limitations in the real time scenarios. Amount of data that has to be handled can still make things slow. Data and task parallelism paradigms are often applied to machine learning / data mining contexts to reduce computation time. The parallel architecture of GPUs makes them an attractive candidate for running machine learning tasks and in fact, now machine learning has been recognized as one of the most popular general purpose applications of GPGPUs

Map reduce framework is viewed as yet another boon for handling large amount data effortlessly to give performance lifts. In a distributed environment, the map reduce framework provides two functionalities; 'map' and 'reduce'. While the map functionality is helpful in distributing the work across processors / nodes, the reduce functionality combines the individual results and process them together to produce the final output. In the recent years the map-reduce technique has been one of the most sought after technology trends that has been instrumental in building scalable, effective applications in a distributed environment.

While research on parallel approximation of machine learning algorithm is much older than a decade, the attempts to improve performance by employing graphics processors are still in infancy. However, initial attempts have been successful attaining as much as 80% increase in speed in particular cases. GPU architectures provide an excellent opportunity to employ data parallelism in order to reduce the time required to process voluminous data. Using the map reduce framework to reduce the computation time in distributed environment consisting of GPU nodes is one of the most recent offshoot of research. The Hadoop framework can be utilized to realize a cluster of GPU nodes. This architecture gives the double benefit of processing large data across a network of distributed nodes and within each node the computation/ processing speed is further enhanced by effective use of multi-core processors of GPUs. In a real time environment, where instant response is the key, the capability of a GPU cluster can be effectively leveraged to improve performance.

## Typical Concerns and Solution approach based on GPU / CPU cluster on Hadoop Framework

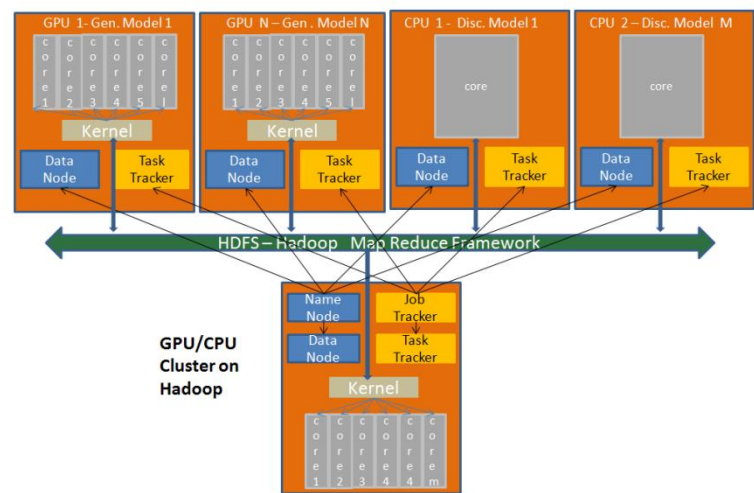
In a true online decision making scenario machine learning models are built offline and classification is done in real time, which saves a lot of time by avoiding high computation time required for model building. It is possible that the amount of real time data coming in is so high that one has to save the models in multiple servers and run the classification

independently on different servers. It is observed that in a typical practice, one machine learning model is used for making decision. Usually the chosen model is a discriminative model that makes the real time computation for classification easy and time required for the same is not really high. However, using one model across global population is not the right practice as user behavior differs across products, verticals and demographic variations. This calls for building and using multiple machine learning models to make better predictions. Besides, the current practice almost disallows the use of generative models in classification as they use a lot of computation time. However the generative models might work much better in certain contexts and it is good to for them for better prediction accuracy.

A map reduce framework based on Hadoop that uses a GPU / CPU cluster has been proposed to facilitate multiple machine learning models with necessary support for running generative machine learning models.

## Architecture

Matrix manipulation, computation complexity, and need for processing of huge data makes analytics applications that employ typical machine learning algorithms an ideal case for introducing parallelism into the process. The proposed architecture uses a combination of GPU and CPU nodes in a Hadoop cluster, each of them devoted for running one machine learning (ML) model. While GPUs are found to be helpful in introducing parallelism in generative machine learning procedures and thus reducing time required for classification, CPU nodes can be employed for using discriminative models. Against the typical procedure where copy of the same model is saved on many servers and the models use data on the respective servers, the proposed architecture will have one server is reserved for a model which makes uses data across the servers leveraging the Hadoop's map reduce framework. A 1000 node cluster will be able to support 1000 models. In this scheme it is possible to include generative models too; as it generative algorithms can be run on GPU nodes making the best use its of parallel processing capabilities.



In this set up, NVIDIA's GPGPU nodes can be employed along with regular / multicore Intel machines. From algorithm point of view, it requires that the parts of the procedure that can proceed in parallel and those that will have to be run in a sequential manner are identified. This in turn helps in making suitable approximation and modifications to algorithms. Real time data that needs to be processed is partitioned across different nodes.

Periodic off line analysis of data must reveal which model is suitable for what kind of data and the result of analysis may be crafted as rules to identify the right model for the specific data. A software module will be responsible for identifying the portion of the data in the node suitable for classification by a specific ML model and making it available for it. Generative models that require longer time run on GPUs make use of the parallel architecture of GPUs and discriminative models that requires relatively less time run on CPUs in a sequential manner. The master node architecture should ideally be GPU or even better a most advanced CUDA architecture, NVIDIA Fermi architecture where, multiple processes can proceed at a time.

This architecture is built to support real time analytics applications and is a classic example of high performance computing being put to use.

## Application Contexts

Although much has been spoken about real time analytics, accomplishing such tasks at real time remains a challenge. The current developments in internet and mobile technology have enhanced the possibilities of accomplishing such tasks to as they provide access to real time data. The need for processing large amount of real time data to run some analytics or to make

recommendations has triggered this research. Some example from the real time retail and mobile analytics scenarios are presented below and of course, there could be applications of mobile analytics in retail context.

#### Real Time Mobile Analytics:

Real time mobile data can be used to provide quick and useful services and personalized recommendations to customers. Analysis of mobile tracking data collected at various mobile towers can reveal the routes of regular use, most preferred routes, preferred locations of hangout during the weekend etc. This information can be used for building personal profiles of users, which in turn is helpful in making personalized recommendations based on the current location of the phone.

Sending traffic jam reports while the customer is still on road well before entering the area of traffic jam, news about the events happening around local hangouts when the customer is much around the area, offering assistance on finding taxis, hotels while away from home or on frequent trips etc., are some of the useful services that can be realized with real time mobile data analytics.

#### Real Time Retail Analytics:

In retail sector customer interaction points are the data pools on which real time analytics is run for providing services such as personalized recommendations. Be it online retail or the brick and mortar, retail industry is much customer focused and much aligned to customer analytics needs. Growing popularity of online retailing has been instrumental in driving online support for many supporting businesses such as delivery, packing, servicing industries etc. In fact, all these businesses also got channelized to work via the internet. Along with the retail industry, these supporting businesses also do get to benefit from real time retail analytics.

Typical examples of real time analytics solutions for the retail industry:

Live chat support for online retail sales and service: While the history of footprints of customers is helpful in customer behavior analytics, funnel analysis helps in understanding the customer drop off. With thousands of customer logged on to the retail web site; many of them entering and leaving the website simultaneously, it is important that the right customers who is most likely to be a buyer is identified and offered an invitation to chat. Given that it is important to establish connection before the customer leaves the site, a real time analytics and instant recommendation capability has a role to play here.

Real time recommendations: Online identification of customers and sending instant, personalized recommendations to them based on their past and current status is yet another application that uses the real time recommendation capability. The recommendations may be in the form of advertisements, personalized offer / discounts that act as motivators for customer retention. Even in a brick and mortar scenario, real time recommendations can play a great role. Depending on the position of the cart, the details of the customer and his past behavior, recommendations on the best offers available may be sent to his mobile. Although the number of customers to whom the recommendations are being sent may not be huge as in case of online retails, it is equally crucial to send the recommendations in real time. The processing capability depends on the size of the stores and degree of personalization proposed.

## Summary

Real time analytics is a subject that has gathered a lot of attention in recent times. Raised as an offshoot of real time BI, it has played an important role in near real time applications such as capacity utilization, demand forecast etc. Multi core platforms that process large amount of data and run complex computations such as matrix manipulations quickly can be effectively used to run true real time applications. A cluster of CPUs nodes in a hadoop framework with map reduce capability is being proposed to counter performance issues. Machine learning algorithms are modified or approximated to reap the benefits of data and task parallelisms. Voluminous data processing and real time recommendation situations as the retail and mobile analytics examples illustrated here would stand to gain from the processing capabilities of the GPUs and hadoop's map reduce framework



## Infosys among the world's top 50 most respected companies

Reputation Institute's Global Reputation Pulse 2009 ranked Infosys among the world's top 50 most respected companies.

---



### About Infosys

Many of the world's most successful organizations rely on Infosys to deliver measurable business value. Infosys provides business consulting, technology, engineering and outsourcing services to help clients in over 30 countries build tomorrow's enterprise.

For more information about Infosys (NASDAQ:INFY), visit [www.infosys.com](http://www.infosys.com).

### Global presence

#### Americas

**Brazil:** Nova Lima **Canada:** Calgary, Toronto **Mexico:** Monterrey **United States:** Atlanta, Bellevue, Bentonville, Bridgewater, Charlotte, Fremont, Hartford, Houston, Lakeforest, Lisle, Minnesota, New York, Phoenix, Plano, Quincy, Reston, Southfield

#### Asia Pacific

**Australia:** Brisbane, Melbourne, Perth, Sydney **China:** Beijing, Dalian, Hangzhou, Shanghai **Hong Kong:** Central **India:** Bangalore, Bhubaneshwar, Chandigarh, Chennai, New Delhi, Gurgaon, Hyderabad, Jaipur, Mangalore, Mumbai, Mysore, Pune, Thiruvananthapuram **Japan:** Tokyo **Malaysia:** Kuala Lumpur **New Zealand:** Auckland, Christchurch, Wellington **Philippines:** Metro Manila **Singapore:** Singapore

#### Europe

**Belgium:** Brussels **Czech Republic:** Brno, Prague **Denmark:** Copenhagen **Finland:** Helsinki **France:** Paris, Toulouse **Germany:** Eschborn, Frankfurt, Stuttgart, Waldorf **Greece:** Maroussi **Ireland:** Dublin **Netherlands:** Amsterdam **Norway:** Oslo **Poland:** Lodz **Russia:** Moscow **Spain:** Madrid **Sweden:** Stockholm **Switzerland:** Basel, Geneva, Zurich **United Kingdom (UK):** London, Swindon

#### Middle East and Africa

**Mauritius:** Reunion **UAE:** Dubai, Sharjah

---

For more information, contact [askus@infosys.com](mailto:askus@infosys.com)

[www.infosys.com](http://www.infosys.com)

© 2011 Infosys Limited, Bangalore, India. Infosys believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of the trademarks and product names of other companies mentioned in this document.