

## **White Paper**

### **Defending Against Modern Web Threats**

#### ***Introducing Data Cloud URL Filtering***

CTM90-100-128-002-R1



## Table of Contents

Executive Summary.....	3
Introduction .....	3
The Evolution of URL Filtering.....	4
Early URL Filtering .....	5
Traditional URL Filtering.....	6
Data Cloud URL Filtering .....	7
Data Cloud vs. Traditional URL Filtering .....	8
Storage Size Impacts Accuracy and Coverage.....	8
Access to Security Expertise .....	9
Updates to Local Premises.....	9
Comparison Summary .....	10
Data Cloud Essential Elements .....	11
Broad Network of Data Sources .....	11
Data Mining Capabilities.....	11
Security Expertise.....	12
GlobalView™ URL Filtering .....	12
How It Works.....	13
About Commtouch .....	14

## Table of Figures

<i>Figure 1 — Database URL Filtering Architecture.....</i>	6
<i>Figure 2 — Data Cloud URL Filtering Architecture.....</i>	8
<i>Figure 3 — URL Filtering Technology Comparison .....</i>	10



## Executive Summary

The explosion in Internet use has brought with it new threats for both home and business. URL Filtering, which has previously been used primarily for parental control, compliance and productivity, is now starting to be implemented as an additional layer to mitigate these web-based threats. However, earlier generations of URL Filtering solutions – based on black/white lists, local databases and on-session heuristics, and still in widespread use – cannot keep up with today's threats due to their technical limitations. This white paper introduces a new type of URL Filtering technology known as "Data Cloud URL Filtering" technology that overcomes the limitations of older solutions and provides the necessary protection against the growing dangers on the Web.

## Introduction

The Internet has become an indispensable part of everyday life and work, yet the massive growth of data coupled with a rapid increase in the number of individuals with Web access has introduced a variety of productivity, compliance and security challenges to today's Internet users.

In private homes, the Internet can easily turn into a dangerous tool if not used and monitored properly. Pornography, nudity, hate sites and violence are just a few examples of unwanted content to which children can easily be exposed if not properly protected.

In the business arena, organizations must defend themselves against legal liabilities that arise when employees access inappropriate or illegal Web sites or content; implementing self-enforced use policies has not been enough to prevent these activities. Productivity challenges for businesses include crippling bandwidth consumption, network latency and even network downtime due to the rise in corporate users and their bandwidth-intensive applications such as streaming media. Security risks are also an important and growing issue for both home and business users, as the enormous popularity and reach of the Internet has made it the preferred attack vector for hackers.

Attackers are finding increasingly sophisticated ways to utilize the Web for their activities, such as infecting websites with malware, both in legitimate web sites as well as less reputable sites such as those hosting pornography. Industry analyst firm Gartner pointed out that in the first quarter of 2008, more than 50 percent of infected sites were in fact legitimate ones that had been silently infected by attackers<sup>1</sup> – an alarming statistic that shows how important it is to have highly accurate solutions to identify and block access to malicious web sites.

URL Filtering (URLF) technology was originally intended to protect minors surfing the web and prevent employee access to inappropriate content or sites which had a negative impact on employee productivity; however, URLF is now being added to the security arsenal as an additional layer of protection. URL Filtering is becoming a key addition to the multi-layered approach which includes gateway firewalls, anti-virus solutions and intrusion prevention.

---

<sup>1</sup>Peter Firstbrook, "Why Malware Filtering is Necessary in the Web Gateway," Gartner Publication G00158459, 26 August 2008, 2.



According to Gartner, the Web threat environment has evolved dramatically during the past three years; however, most organizations continue to rely on the same defensive technology introduced more than ten years ago<sup>2</sup> — a practice that leaves them exposed to business, legal and security risks.

Not all URLF solutions are equal in the way they analyze a site. Today's Web sites are increasingly diversified, and user-generated content in Web 2.0 and social networking sites has led to the grouping of a variety of media and content types on a single site. In this scenario, a malware site can easily hide in a subdomain or page within a legitimate domain such as Microsoft's live.spaces.com or Facebook.

Web use and Web threats happen in real-time and organizations require next generation URLF technology advanced enough to keep up with the complex and quickly shifting threatscape in a user-generated content world.

## The Evolution of URL Filtering

In order to understand the evolution of URLF implementations, it is important first to understand the main processes of most URLF systems. The chart below defines these processes of categorization, storage and filtering:

Process	Mission
<b>Categorization</b>	Analyzes Web data to associate URLs with one or more pre-defined categories
<b>Storage</b>	Stores the categorized sites – typically in a database – for access as users browse
<b>Filtering</b>	Takes action (block/allow) based on how sites are categorized

The different approaches and implementations of each of these URLF elements can have a tremendous effect on systems' infrastructure, performance and accuracy.

---

<sup>2</sup>Firstbrook, 2.



## Early URL Filtering

In the mid 1990s, URLF solutions relied on black lists (prohibited) and white lists (approved) to control access to Web sites. While the lists of good/bad URLs were managed locally by each organization's IT department, in most cases they were generated by a central service and then stored at each customer's site, where filtering also took place. There were, however, a number of drawbacks to this approach.

Inaccurate site categorizations caused many approved sites to be blocked while allowing some prohibited sites through the filters. It was a resource-intensive process requiring IT departments to handle frequent updates and edits to the lists. List-based URLF solutions were limited to filtering "known" sites, and suffered a notable lack of granularity that left them unable to adequately process new Web requests. As the Web grew and the number and complexity of Web threats increased, a black/white list approach was no longer a viable approach for organizations seeking a reliable filtering solution.

*Most organizations continue to rely on the same defensive technology invented ten years ago.*

*- Gartner*

In the late 1990s, URLF technology began leveraging category engines coupled with a local database. The category engines were typically located in a remote location (i.e. "in the cloud") rather than a local machine within a corporate network. This remote categorization approach is still the preferred architecture for today's standard URLF solutions. Using this technology, URLs and their contents were analyzed and classified under a predefined category (e.g., pornography, employment searches, sports) stored in a centralized master database and then transferred in batches to local customer databases which were exact replicas of the master database. Unlike with black and white lists, where policies were across-the-board, the centralized category engine approach introduced the capability to enforce granular policies at each organization.

While providing more depth than the previous incarnation of URLF, the fact that these solutions were restricted by the size of the local database impeded their ability to provide extensive coverage and in-depth site classification. This limitation had an even greater effect on system accuracy as sites have become more dynamic and complex.

Figure 1 demonstrates a typical database URLF implementation, where although the categorization process is done in a remote location (in the cloud), the storage is kept at the device level. In that type of architecture, every customer has the exact same database, regardless of the fact that each different customer has different needs.

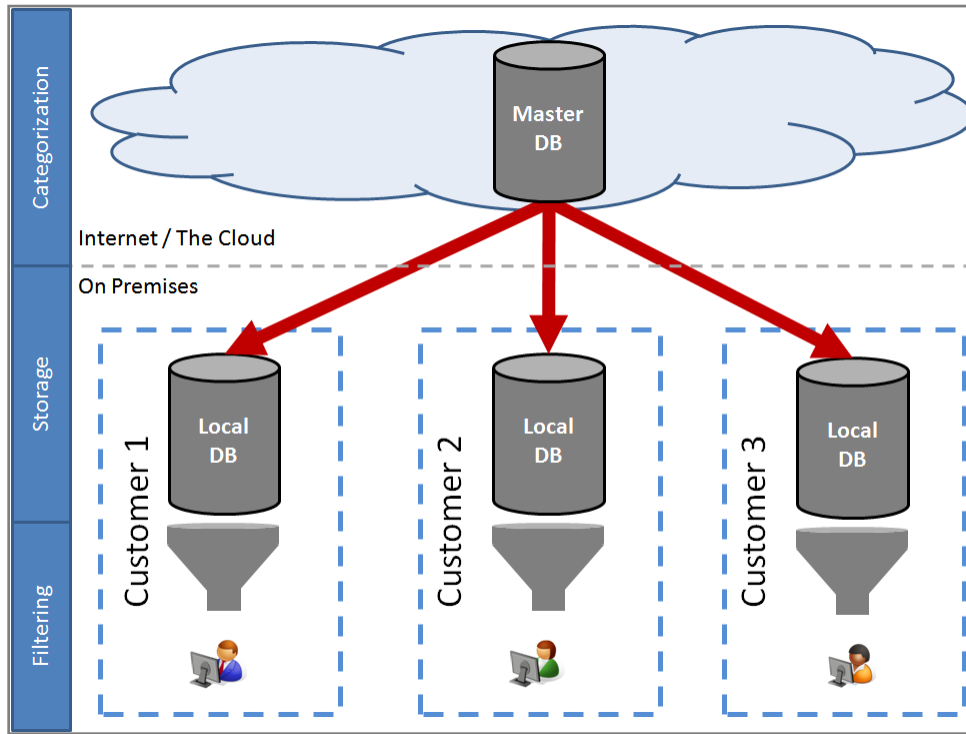


Figure 1 — Database URL Filtering Architecture

## Traditional URL Filtering

In the early 2000s, URLF solutions began to employ *dynamic engines* as a means of analyzing the flood of dynamic content and Web sites on the fly. These on-session heuristic engines examine probability — i.e., how likely it is that the text in a particular site signifies the type of site it actually is — in order to determine whether or not a site should be blocked. Local heuristic engines attempted to solve the main problems inherent in database solutions: their “one size fits all” approach and limited capacity.

Since local heuristic engines must analyze a site without interrupting user experience, they are supposed to work quickly, but unlike categorization in the cloud, they do so with limited resources because their categorization engine typically runs on a desktop or appliance. Impaired by time and resource limitations, these engines still are not capable of providing an in-depth analysis of both content *and* context – two vital points when dealing with the complexities of a Web 2.0 environment. Furthermore, these engines have been prone to harm organizations’ productivity or even break security policies by blocking legitimate sites or allowing malicious content to pass through.

Since this type of heuristic technology has not yet stabilized, these engines are typically used in tandem with older, more established local database solutions. While there are several database solutions that employ heuristic engines in their offering, their use is limited to specific categorization tasks, while leaving the database-building process to the more in-depth and mature technology.

Although deploying them together provides a slight improvement over each one alone, blending database URLF together with on-session heuristic URLF still keeps both



technologies' shortcomings. Overall, these database/heuristic URLF solutions are limited enough that they tend to disappoint when it comes to accuracy and performance.

## Data Cloud URL Filtering

The Web is an almost *infinite* collection of data that is growing at a speed of a billion pages per day, according to Google.<sup>3</sup> The Internet's vast size coupled with the unique and specific needs of individual customers has created the need to move beyond physical and technical limitations of local databases and on-session heuristics; the most innovative and effective approach has been to transfer the local database into the cloud. This provides a more flexible, comprehensive and accurate form of URLF technology which can be called "Data Cloud URL Filtering."

Data Cloud URLF does not need to rely upon the limited resources of an on-premise database for analysis and detection, nor is it dependent on database updates for the latest available information. Instead, since the database itself is located in the cloud, Data Cloud URLF is able to leverage the enormous depth and breadth of data available in the cloud for threat detection, analysis and prevention. Working in the cloud, multiple engines can work massively on classification, growing the Data Cloud's contents based on actual usage and popularity patterns.

The Data Cloud itself is an extensive, nearly infinite database existing outside of a fixed location. Comprised of relevant, up-to-date content needed to facilitate quick and accurate detection, its massive computing and storage capabilities are leveraged by a local cache that unlike the traditional database, stores only the data required to fulfill each customer's needs. Thus the data cloud overcomes many of the problems inherent in the earlier generations of URLF.

Figure 2 demonstrates a typical Data Cloud URLF implementation, where the categorization and database itself is stored in a remote location (in the cloud). In this architecture, the local cache stores the necessary information locally, and can expand and contract based on customer needs.

*Since the database itself is located in the cloud, Data Cloud URLF is able to leverage the enormous depth and breadth of data available in the cloud for threat detection, analysis and prevention*

---

<sup>3</sup>Jesse Alpert & Nissan Hajaj, "We knew the web was big..." The official Google Blog, July 25, 2008 <<http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>>.

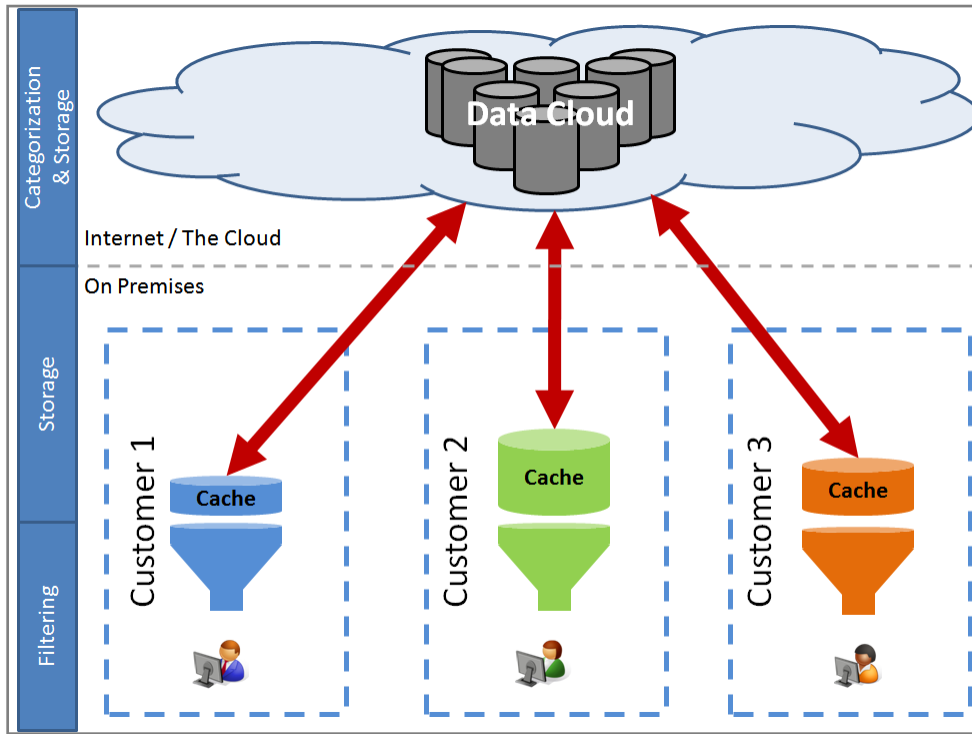


Figure 2 — Data Cloud URL Filtering Architecture

## Data Cloud vs. Traditional URL Filtering

### Storage Size Impacts Accuracy and Coverage

The innovation of Data Cloud technology is in its storage process, where it eliminates the limitation of local, on-premise storage capacity, allowing URLF solutions to focus on accuracy and coverage.

Note that even in traditional database URLF, the categorization is done in the cloud, but this differs significantly from Data Cloud URLF. A local database can contain millions of categorized sites and tens of millions of Web pages, but this represents a tiny fraction of sites/pages on the Internet overall, so the chances of missing something important are substantial. The size of the database dictates how accurate or complete it can be; every new site that needs to be added to the local database comes at the expense of data that is already there and will need to be pushed out.

For example, a popular blog hosting site such as blogspot.com may contain more than 20 million Web pages, each fitting into an entirely different category — one blog may be about science, another about sex, even another about politics — totaling 20 million unique entries in the database. As the local database URLF solution cannot afford to have 20 million different entries for a single Web site, it will typically generalize its findings. By trying to fit each page into the *most similar* categories it can find — such as “blogs” or “computers” — the potential exists for it to block legitimate pages while allowing some prohibited content through.



Furthermore, database URLF technology provides the same sites and their global classifications to all customers, regardless of its relevance to each individual user or organization. Taking into consideration the behavioral, cultural, geographical and professional differences that can vary between organizations, this dictates a necessary compromise in terms of coverage.

Even when deployed with together a heuristic engine, the limitation for database URLF technology is still a major factor. Trying to assess a URL's categories with on-session local heuristic engines under the restrictions of limited resources, performance issues and latency, impedes the ability to provide accurate classification. This means that in mixed solutions, local heuristic engines cannot function as a complementary layer to categorize every request that the database lacks; instead, they are used only a small percentage of the time and only for well-defined categorization tasks.

Data Cloud URLF, on the other hand, has no local storage limitations, so this technology can store all the Web pages that its customer-base needs *in the cloud*, while each customer's local cache stores only the sites that are relevant to that customer. Returning to the previous example, while all 20 million pages in a particular domain can be stored in the Data Cloud, each customer can store the exact required subset based on its unique browsing behavior and needs.

*The size of the database dictates how accurate or complete it can be; every new site that needs to be added to the local database comes at the expense of data that is already there.*

## Access to Security Expertise

Considering URLF as an integral part of the overall security infrastructure is an important step toward detecting and preventing Web threats. However, the design limitations of earlier generation URLF solutions prevent them from prioritizing these threats quickly and accurately. Furthermore, they are limited in their access to security sources from which to extract updated information. With the speed at which Web threats appear and evolve, this drawback has a serious impact upon the effectiveness of traditional URLF solutions.

Data Cloud URLF solutions, on the other hand, can leverage the extensive security expertise available in the cloud such as zero-hour malware patterns, zombie data and phishing feeds. The Cloud provides continual access to updates on the latest threats, which allows Data Cloud URLF solutions to quickly and efficiently place potential threats into core security categories such as compromised sites, malware and phishing. Retaining these categorizations in the Data Cloud enables all users, regardless of geographic or physical location, to benefit from the continuous analysis and identification of threats.

## Updates to Local Premises

Updates are another differentiator between Data Cloud and traditional URLF solutions. In order for database URLF to obtain the most recent information, frequent high-volume and often time-consuming updates must be performed. With the growth of dynamic and short-lived threats, such as phishing sites that are online for less than 24 hours, database URLF solutions relying on periodic updates cannot stay up-to-date. Additionally, each update may inflate the local database unnecessarily with large amounts of irrelevant information, affecting both system and network resources.



Data Cloud URLF solutions are not burdened with the need for scheduled updates, as they only make real-time requests for information not previously known in the local cache to receive highly specific and relevant information from the Data Cloud as needed. Thus, these solutions avoid the data dump inherent in traditional URLF updates; storing this specific information in the local cache creates a streamlined, efficient resource for users — storing nothing more and nothing less than they need at any given time.

## Comparison Summary

While most URL filtering technologies were adequate solutions at the time they were introduced, they failed to adapt to the changing environment. The following figure summarizes each technology’s advantages and disadvantages when examined in light of the demands of today’s dynamic Internet environment.

	Mid 90s Local black white lists	Late 90s Local Databases	Early 2000 Heuristic engines	2008 Data Cloud
Accuracy	Poor	Basic	Poor	Excellent
Coverage	Poor	Poor	Excellent	Excellent
Bandwidth	Excellent	Poor	Excellent	Excellent
Up-to-date	Poor	Poor	Basic	Excellent
Summary	Not scalable for today’s dynamic Internet demands	Local DB can’t cope with exponential Internet growth and diversity	Not enough local resources to analyze data	Unlimited central power, advanced auto-learning local cache

Figure 3 — URL Filtering Technology Comparison in Today’s Internet Environment



## Data Cloud Essential Elements

Implementing a Data Cloud URLF solution involves more than just storing a database in a remote location — different infrastructure, analysis methodology and expertise are required to leverage its considerable advantages. Following are descriptions of three areas that are a crucial part of an effective Data Cloud URLF implementation: broad data source network, data mining capabilities and security expertise.

### Broad Network of Data Sources

The sheer size and complexity of the Internet make it impossible for any one vendor to possess enough expertise to address every possible Web threat. Therefore, it is critical for a Data Cloud URLF solution to gather data from multiple specialized sources and protocols to obtain truly accurate information.

Messaging and Web protocols are now frequently used as a channel for malicious activities. Whether from phishing sites that are distributed through spam or malicious Web sites hosted on infected machines, a Data Cloud solution must be able to gather and analyze a substantial amount of complementary sources in order to better detect new and emerging threats.

Browsing behavior from both individual users and backbone traffic is a valuable source of information for URLF solutions. With no database size limitations, each request can be analyzed and processed to identify key trends from around the world. For example, if a small site receives a sudden spike in traffic, it should be analyzed for possible threats since malware writers frequently increase traffic to an infected site to achieve results.

Zero hour malware is also a concern. While identifying malware variants lies in the hands of expert anti-virus vendors, Data Cloud URLF can utilize this expertise in a mutually beneficial manner – the URLF acts as a source of information about new variants, informing anti-virus vendors; while the anti-virus vendors act as an authority for blacklisting malicious files on Web sites.

This sort of sharing and analysis requires a dedicated infrastructure for trading information not only across different protocols, but across companies as well.

### Data Mining Capabilities

Since the amount of information available for analysis and processing is nearly infinite, accurate and comprehensive data mining capability is a key function of Data Cloud URLF solutions. The ability to extract patterns and threats from an enormous and diversified group of data sources — in mere seconds — is one of the most important differentiators among Data Cloud URLF solutions.

Cross referencing different data sources in order to identify new patterns and threats is an integral part of URLF data mining. The ability to place all the pieces together in order to see the big picture requires a technology infrastructure that feeds one data source engine output with another, in order to extract more value from existing sources. For example,

*It is critical for a Data Cloud URLF solution to gather data from multiple specialized sources and protocols to obtain truly accurate information.*



most web-based phishing and malware outbreaks are delivered via email, so feeding these attacks into the URL engine can allow a URLF engine to block attacks at the zero hour. By utilizing this expertise, Data Cloud URLF enables the identification and extraction of new threats in real-time. When data is gathered from many different protocols including external sources, Data Cloud URLF can provide a 360-degree view of the entire threatscape.

## Security Expertise

Web threats are increasing at an alarming rate. According to Gartner, the threat environment will continue to evolve rapidly due to attackers' specializations and the proliferation of adaptable malware tools.<sup>4</sup> Thus, they stress the need for organizations to change their defensive strategies or accept increased business risk and potentially costly disruptions.

The original URLF technologies were designed to handle productivity and inappropriate content, rather than security. Today's Web threats, however, require URLF solutions to include proven security expertise to detect current and even future Internet threats. There are millions of data points constantly being analyzed with Data Cloud URLF and without proven security expertise, crucial elements can be overlooked.

## GlobalView™ URL Filtering

Commtouch GlobalView™ URL Filtering leverages Commtouch's Data Cloud technology to provide the most accurate, up-to-the-minute analysis and categorization available in a URLF solution today. It is provided as a software development kit for security vendors and service providers to integrate into their offerings.

By taking full advantage of the depth and breadth of resources available in the cloud, the Commtouch GlobalView URLF engine is able to extract patterns and threats from an enormous and diversified group of data sources in mere seconds. Built on proven, award-winning security technology, Commtouch and its expansive global network of security partners — the Commtouch Security Alliance — are able to provide timely, detailed information on coverage areas ranging from anti-spyware, anti-phishing and anti-child pornography, to site rank technologies and search engine pattern analysis, among others. Once accessed, this information is stored in the Data Cloud for ongoing reference. Updated in real-time, these same sources inform the rapid analysis and categorization process of GlobalView URLF, allowing users to extract highly specific and detailed results wherever and whenever they require it.

*Sharing resources amongst vast numbers of users in the Data Cloud enables organizations to access more comprehensive, granular and specific filtering data than is available with a database-dependent URLF solution.*

---

<sup>4</sup>Firstbrook, 2



## How It Works

Commtouch GlobalView URL Filtering is a software development kit that has a small local client that is integrated within a licensee's web security device (e.g. secure web gateway, unified threat management). As a user browses, the web security device sends the requested URL to the Commtouch URLF solution and within microseconds, its smart client checks the local cache for information. If it finds the relevant classification in the cache, it extracts the necessary information, enabling integrating partners to determine if that URL should be allowed or blocked (fully or partially). If the information does not reside in the local cache, it quickly sends out a request to the Commtouch Data Cloud and in return receives the latest, most relevant and accurate available classification. It then stores the new information in the local cache for future reference while leaving a copy in the Data Cloud for other users to access when necessary. In a sense, Data Cloud users contribute valuable knowledge to the centralized data center, where sites growing in popularity can be queried once by an end-user and already categorized for the later users looking for that same site. This incorporation of real user-browsing data ensures that the Data Cloud is continually updated and expanded.

For similar queries in the future, the results are retrieved directly from the local cache. Since users and organizations tend to use many of the same sites repeatedly, the cache learns the specific requirements of that organization's users and is intelligently built to suit their specific needs. Unless a new, previously uncategorized request is made, there is no need to access the Data Cloud.

The Commtouch Data Cloud can be simultaneously accessed by hundreds of millions of users at one time, each extracting and contributing relevant data to grow the knowledgebase. Sharing resources amongst vast numbers of users in the Data Cloud enables organizations to access more comprehensive, granular and specific filtering data than is available with a database-dependent URLF solution.

The Commtouch Data Cloud is modeled on highly successful data mining and detection engines the Company developed for messaging security. The Data Cloud provides GlobalView URLF with information on millions of categorized Web sites in 64 individual categories, enabling the licensing partner's offering to accurately block or allow browsing to specific URLs. Of the wealth of categories GlobalView URLF employs, ten are dedicated to security issues. This deep level of granularity allows GlobalView URLF to provide comprehensive data on areas of concern such as compromised Web sites, known spam sites, malware sites and other emerging Web 2.0 threats.

GlobalView URLF also enables the more conventional content-filtering applications such as parental control, corporate productivity and compliance. By blocking access to prohibited sites (e.g., pornography, social networking or job search), GlobalView URLF helps to provide a safe browsing environment, increase employee productivity, and ensure corporate compliance, while at the same time securing against Web-based malware and fraud. Bandwidth consumption can also be reduced by eliminating improper downloads of streaming audio and video content.

**commtouch®**  
Security Alliance

*The Commtouch Security Alliance is a framework for the sharing of information between Commtouch and leading information security organizations.*



Combining years of security expertise with best-in-class technology, GlobalView URLF supports improved network safety and management by quickly and accurately processing massive amounts of Web traffic with minimal latency. Commtouch GlobalView URLF is the solution of choice to protect against today's and tomorrow's growing Web threats.

## About Commtouch

Commtouch® (NASDAQ: CTCH) is the source of proven messaging and Web security technology for scores of security companies and service providers, founded on a unique cloud-based datacenter approach. Commtouch's expertise in building efficient, massive-scale security services has resulted in its patented technology mitigating Internet threats for thousands of organizations and hundreds of millions of users in more than 100 countries. Commtouch technology automatically analyzes billions of Internet transactions in real-time to identify new threats as they are initiated, protecting email infrastructures and enabling safe, compliant browsing. The unmatched suite of Commtouch security offerings is based on patented Recurrent Pattern Detection (RPD™) and GlobalView™ technologies, which work together in a comprehensive feedback loop and offer equally effective protection for all languages and formats. Commtouch was founded in 1991, is headquartered in Netanya, Israel, and has a subsidiary in Sunnyvale, Calif.

Stay abreast of the latest at the Commtouch Café: <http://blog.commtouch.com>. For more information about enhancing security offerings with Commtouch technology, see [www.commtouch.com](http://www.commtouch.com) or write [info@commtouch.com](mailto:info@commtouch.com).