Enabling Real-Time Big Data Movement in the Constantly Connected World

Huge volumes of data are on the move. The average person uploads 15 times more data today than they did just three years ago, and estimates are that the total amount of information in the world is doubling every 18 months. Enterprises are storing an exponentially increasing amount of data about the customer activity and conditions that surround their business in the form of clickstreams, digitized call center interactions, sensor readings, transaction records and more.

Apache Hadoop has emerged as the de facto standard way of storing all of this “big data,” mostly in the form of commercial implementations from HortonWorks, Cloudera and MAPR. Associated technologies such as Flume, HBAse, Hive, Kafka, MapReduce, Spark and Storm offer different ways to get information into and out of Hadoop Distributed File Systems (HDFS) so it can be shared with analytics engines, enterprise applications and user interfaces.

As the practice of collecting big data in batches has given way to more frequent and even real-time updates, the need for a new kind of data collection and distribution infrastructure has emerged – systems with the capacity, performance and scalability to enable *real-time* big data.

Solace message routers excel at efficiently collecting, filtering and distributing large volumes of information. By handling all aspects of routing and delivery in a pure hardware datapath, Solace can move more information than software-based data distribution technologies with much higher performance, all in a compact hardware devices that’s easy and less expensive to deploy, scale and operate.

This whitepaper explores the various forces driving big data applications to become more real-time and examines what that trend means to the business owners, application designers, and the people who build the infrastructure to move and manage all of that data.

Solace Systems®

http://www.solacesystems.com

The Internet of Things and Tidal Wave of Data

Remote sensors are being deployed in just about every place imaginable: police cruisers, soda machines, smart meters, railway locomotives, home thermostats, solar panels, etc. Private and public spaces, both commercial and residential, are teeming with “always on” producers of information, so the Internet is evolving from a network of computers into an “Internet of Things.”

Even the phone in your pocket is a sensor of sorts, generating streams of location and status data including:

- Location via GPS, WiFi triangulation, Bluetooth and near-field communication
- Movement utilizing location sensors plus compasses, accelerometers, and gyroscopes for tilt and vibration.
- Activity including interest and purchase tracked via barcodes, QR codes, RFID tags and near-field communication.
- Audio and images via microphones and cameras capable of HD pictures and video.

Half a billion people accessed the Internet from a mobile device in 2009, and usage is expected to double within five years as mobile overtakes the PC as the most popular way to get on the Web. The biggest source of edge generated data is the apps running on these increasingly powerful mobile platforms. The average smart phone user has installed over two dozen applications on their phone.

The nature of those applications is changing too, from centralized top-down broadcast and consumption, (e.g. stock market prices, news, weather, and other syndicated content) to user generated content. Social networks, location based services, activity streams, and telemetry information are combining to drive a torrent of upstream data that needs to be captured and analysed so it can be monetized and used to create innovative services.

Social networks, location based services, activity streams, and telemetry information are driving a flood of upstream data that needs to be captured,

Big Data in the Cloud and Enterprise

At the other end of the equation are the datacenters full of servers that make the Internet of Things possible: in-memory data grids, compute clusters full of virtual machines, Software as a Service systems and cloud computing environments. Major players run dozens of globally distributed datacenters, each with hundreds or thousands of servers.

Since these millions of virtual machines produce more data than commercial monitoring software can keep up with, most operations teams have built custom systems to capture and make sense of all the data generated and stored by their server farms. Inevitably, the rest of the company wants to tap in to this treasure trove of data:

- Marketing wants to use it to better upsell, cross-sell, and target advertise.
- Sales wants to increase average order size and reduce customer churn.
- Security wants to reduce fraud.

Datacenters are full of systems that collectively produce more data than conventional monitoring and management software can

- The user experience team wants to perform multivariate testing and optimize human-computer interaction.
- Support wants to monitor compliance to SLAs, enable user self-support, and reduce call center volume.
- Legal wants to ensure regulatory compliance and avoid fines and penalties.
- Management wants to support the six sigma initiative by measuring performance and meeting performance goals.

In summary, the data needs to be made available everywhere, be sorted and stored in different formats, and used in a wide variety of ways. Capturing a single copy of these data sets in a centralized data warehouse for after the fact analysis represents a huge missed opportunity to realize the full value of "big data". Realizing the full value of these data sets requires a new scale of real-time data movement.

Big Data Applications by Industry

Here are just a few applications that demand big data to be accurate in real-time:

- Transportation: track & trace of cargo and equipment, real-time safety monitoring, logistics and scheduling, optimization of locomotives, trucks, crew, cargo, and passengers
- Finance: market data distribution, risk analytics, fraud prevention, SLA compliance, smart order routing, algorithmic trading, high frequency trading, arbitrage transaction processing
- Government: homeland security, CBRN sensors, video surveillance, weather monitoring, situational awareness, emergency management, cyber security
- Smart Grid: wide area situational awareness, disturbance correction, fraud detection, outage detection, plug-In electric vehicle infrastructure
- Web 2.0: real-time auctioning, micropayment processing, geospatial social networking, online gaming, status update analytics
- Telco: distributed data grids, call detail record (CDR) capture, fraud detection, network monitoring, customer analytics, location based advertising

Capturing a single copy of these data sets in a centralized data warehouse for after the fact analysis misses the opportunity of big data.

Big Data Inside and Outside the Datacenter

Inside a datacenter means that the information is coming from applications and systems connected directly via high-speed Ethernet or InfiniBand switching fabrics. Examples include: ○ Click streams ○ Application events ○ Transaction records ○ Syslog/Monitoring data ○ Cloud Computing VM Infrastructure monitoring	Outside the datacenter, data sources include phones, distributed sensors, and partners. The massive number of endpoints and connections becomes the main challenge. Examples include: ○ Remote Telemetry ○ SCADA & Sensors Nets ○ Smart Phone & Tablet Apps ○ RFID and barcode scanners ○ GPS Position Tracking ○ Point-of-Sale Terminals

Moving Big Data in Real-Time

Wikipedia defines big data as "datasets that grow so large that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analytics, and visualizing." Applications that generate huge continuous streams of data can make the *capture* and *storage* functions most challenging. We call these types of systems "Real-Time Big Data" to differentiate them from use cases where *search*, and *analytics* are the most difficult functions.

When big data sets are relatively static, and updated incrementally or infrequently, Extract Transform and Load (ETL) and Data Warehousing tools are often sufficient to enable batch movement of records. However, processing real-time streams of updating data elements is more complicated and requires some form of publish/subscribe data movement. For example, a full stock market data feed with all North American instruments updates at rates of about 2 million updates per second.

Just collecting basic activity stream data from a datacenter with thousands of servers can easily exceed the capability of most systems to capture reliably. Storing information generated from millions of concurrently connected users can require enormous numbers of servers when done in software.

The most effective way to handle the distribution of big data is with a data movement layer that lets systems send data into the cloud without worrying about where it needs to go

The most effective way to handle this volume of data is with a data movement layer that takes care of message delivery so publishers/sensors can send data without worrying about where it needs to go or how it needs to get there. This entails the establishment and management of queues and topics, application of subsequent routing rules, intelligent handling of fault conditions such as applications or network links being down or slow, etc.

The Solace Solution

Solace message routers are built with high-speed ASICs, FPGAs, Network Processors and non-volatile memory-based storage in order to scale to a level that would otherwise require 30-100 general purpose computers running software-based solutions. They support all kinds of data movement including reliable, guaranteed and JMS messaging, along with the ability to optimize for the WAN and stream data via the Web, with a single unified API and administration framework.

A single Solace message router can deliver up to 80 Gbps of information, and easily networked to handle millions of concurrent users or devices.

With such high capacity, Solace enables you to capture and process massive amounts of data in a small footprint that costs less in terms to operate in terms of rack space, connectivity, power, cooling, and administrative manpower.

Rich Internet Applications written in JavaScript, Flash, Silverlight, or other common mobile platform development frameworks are all supported with semantics developers are used to inside the firewall. Data communication can be provided over HTTP(s), COMET, HTML5 Websockets, or dedicated TCP socket connections with hardware acceleration, compression, and security.

Solace's performance advantage and intelligent routing protocols enable a number of capabilities that make the Solace solution uniquely well-suited to meeting the needs of big data capture, processing and distribution:

- Sharding: Solace message routers deliver route data using metadata tags such as topic and queue names, or key on content contained in the data stream itself. Routing rules allow for guaranteed 1 of N delivery (for load balancing) or more complicated multi-mode or high fan-out delivery of data.
- Sequencing: Some use cases like fraud detection are sensitive to out-of-order data capture and require global data sequence to be preserved. When data updates rates become extremely high, timestamps cannot reliably order messages and absolute global ordering must be preserved by the data capture infrastructure. Solace can guarantee the sequential delivery of data elements across many downstream storage nodes and applications.
- De-duping: Wireless sensors and network probes will frequently generate the same information multiple times. Finding and removing these redundant data elements cleanses the incoming data stream reducing the need to store and process extraneous information. Solace message routers use various de-duplication algorithms to handle some of the most challenging use cases, such as cleaning multi-path effects from wireless video streams.
- Affinity Based Routing: Captured data elements need to be routed to the right storage location, and subsequent updates need to have affinity to that location. It can be challenging to efficiently get the right data to the right place(s) in real-time

Solace message routers act as a "shock absorber" to smooth out the peaks without losing information. Incoming data can be buffered in non-volatile memory, and spooled out at an optimal rate for the downstream storage systems so no data is lost when the data sources and storage location move around the globe. Solace uses dynamic address space and routing protocols to make sure every data element is routed appropriately.

- Data Buffering: Incoming data can spike, resulting in peak rates that far exceed the average inbound rate. Solace message routers can act as a big data “shock absorber” to smooth out the peaks without losing information. Incoming data can be buffered in non-volatile memory, and spooled out at an optimal rate for the downstream storage systems so no data is lost during overflow conditions. It also eliminates the costly and wasteful need to size for peak rate.
- Replication & Disaster Recovery: Capturing data in a single location can expose your business to continuity problems if the primary location becomes unavailable. Solace routers can efficiently replicate data streams across the WAN to multiple sites and provide failsafe delivery to remote DR and active/active datacenters. Copying and delivering data in the Solace hardware does not add load to producing or consuming apps or storage nodes.

Summary

Moving high volume data streams within and between datacenters can be complicated and expensive using traditional software-based messaging middleware on general purpose servers. Even with open source solutions, the cost of horizontally scaling out the hardware, and the added cost of support, administration, network ports, bandwidth, etc. all add up to a significant investment. Solace message routers can meet the daunting demands of real-time big data movement with better performance, lower TCO and less complexity than any other solution.

To learn more visit solacesystems.com or call +1 613-271-1010.