SOLUTION BRIEF

ACT INSTANTLY ON YOUR DATA

Plug into the most powerful, real-time stream processing platform for Big Data

DataTorrent

TRACE3

BIG DATA. DONE NOW.

Your ability to recognize and react to events instantaneously isn't just a business advantage. In today's world—it's a necessity. Data is happening NOW, streaming in from various sources—in real-time, all the time.

Many organizations struggle with processing, analyzing and acting on this never ending and ever growing stream of information. Sometimes, by the time data is stored to disk, analyzed, and responded to, it's already too late to gain maximum value from that data.

Big Data Real-Time Stream Processing enables you to process and monitor **data in motion**, so you can respond to events in a matter of milliseconds. This capability is crucial for many industries including finance, fraud detection, machine-generated data, utilities, geo-location services, dynamic pricing applications, and telecom.

ENTERPRISE-SCALE THAT'S FUTURE-READY

DataTorrent is the most powerful, real-time stream processing platform for Big Data. Unlike traditional batch processing that can literally take hours, DataTorrent automatically scales to process 1 billion data items per second, enabling you to analyze and act instantly based on your data as it comes in—not in "near real time," but now.

DataTorrent supports today's most demanding, mission-critical, Big Data applications. It is designed to enable highly scalable, massively distributed real-time computations—all managed automatically by the platform itself. The platform's unparalleled performance and unique features enable organizations to focus on their business logic, rather than on the operations of managing the infrastructure. Furthermore, it ensures your real-time streaming applications are future ready—employing a solution designed to automatically sustain any future changes to load, distribution, or business logic, as your needs evolve.

EXTREME SCALABILITY - 1 billion events per second? No problem!

- DataTorrent automatically scales out or in to accommodate any data size and processing need you may have to support—now and in the future
- Linear scalability with sub-second latency guaranteed—even while processing 1 billion events per second
- Automatic scaling is handled by the platform, so you don't need to worry about ensuring scalability or capacity planning for future loads

HIGH AVAILABILITY GUARANTEED

- Built-in fault tolerance ensures your application is running smoothly at all times
- Fully-automated resource management
- Update your application while it's running! You can enhance your code and expand your business logic with no need for application downtime, all the while ensuring business continuity.

HADOOP-NATIVE FROM THE GROUND UP

- DataTorrent runs on your existing Hadoop 2.x cluster—enabling enterprises to leverage their investment in Hadoop for real-time computations
- DataTorrent's open source framework and application templates enable you to develop faster and support any business logic
- Seamlessly integrate with your existing data flow

REAL-TIME STREAM PROCESSING USE CASES

A real-time stream processing strategy can pay dividends for your organization if you need to: 1) Process data in real time; 2) Monitor and alert for any system event or business metric in real time; 3) Make decisions and take action in real time based on incoming data.

Some of the use cases identified for real-time stream processing are:

SECURITY AND INTELLIGENCE

- Continuously track millions of events and process massive amounts of sensor data
- Real-time event correlation from diverse data sources to detect anomalies and generate automatic alerts

FINANCE

- Risk analysis and fraud detection: analyze complex patterns across multiple sources and trigger actions automatically within seconds of detected events
- Deliver personalized products and improve financial services strategies and offerings
- Make decisions in real time for trading and transactional platforms

MACHINE-GENERATED DATA (INDUSTRIAL INTERNET)

- Analyze and act on massive amounts of machine-generated data from sensors and actuators in real time
- Predict system failures and perform smart maintenance
- Correlate usage and performance data across multiple sources

TELECOM & NETWORKING

- Real-time network monitoring and protection; automatic resource allocation and load balancing
- Fraud detection & revenue assurance
- Easily support tiered service levels and billing
- Take action based on user location

SMART GRID & UTILITIES

- Monitor and optimize your infrastructure and resource allocation
- React in real time to abnormal usage fluctuations
- Predict spikes in demand or underutilization and trigger action

SITE OPERATIONS

- Optimize infrastructure resources across a global, distributed network
- Monitor and react in real time to outages or variations in demand
- Detect security attacks
- Predict potential bottlenecks and failures

WEB & MOBILE APPLICATIONS

- Ingest and react in real time to massive amounts of data in motion
- Analyze and take action based on user-generated content and social media interactions
- Actionable analytics for gaming, location services, travel, and more

MEDIA & ONLINE ADVERTISING

- Dynamic bidding
- Real time targeting & personalization
- Maximize click-through and conversion rates

RETAIL & E-COMMERCE

- Real time targeting & personalization
- Streamline supply chain processes
- Improve customer satisfaction and retention

PLATFORM ARCHITECTURE

Built exclusively on Hadoop 2.0, DataTorrent enables you to utilize your existing infrastructure and commodity hardware for stream processing on a massive scale. DataTorrent's processing is done in-memory, parallel to your existing Hadoop batch jobs. It can read and write to external persistent storage, to both SQL and NoSQL databases.

Application development is as simple as defining business logic—DataTorrent takes care of the rest. The platform delivers superior adaptability with scaling, resource optimization, and dynamic application modification—all done on the fly, with no downtime or management hassles.

REAL-TIME ENGINE

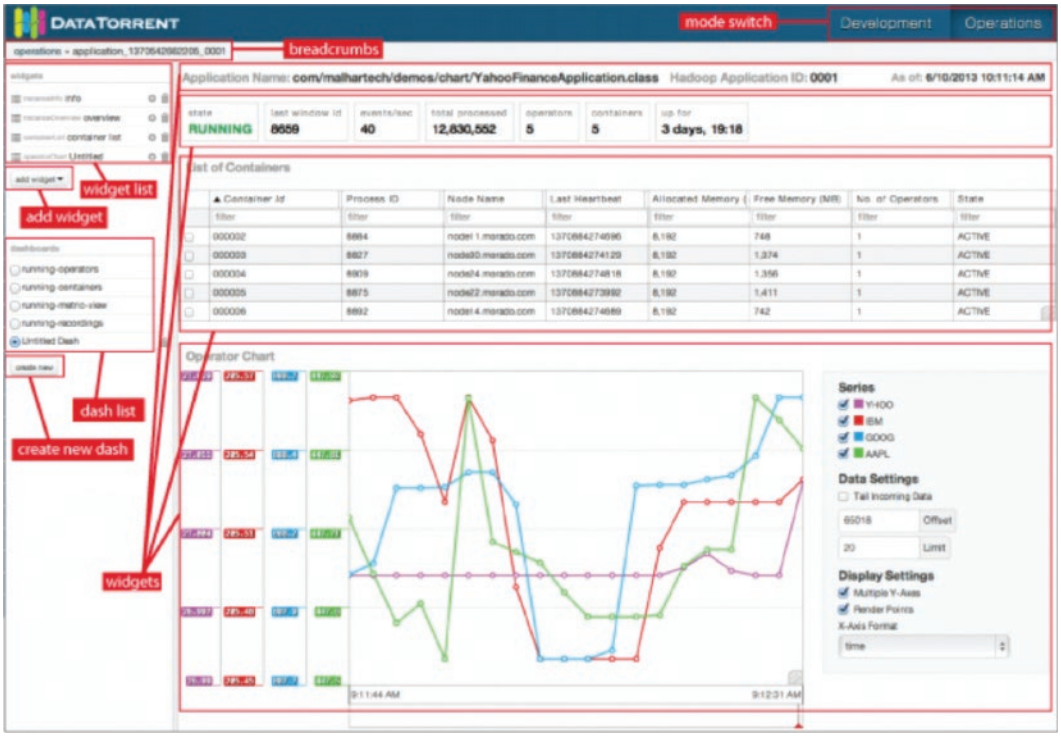
- A native Hadoop 2.x platform designed to enable distributed, asynchronous, real-time Big Data computations in as unblocked a way as possible, with minimal overhead and all computations done in memory to ensure low latency.
- This is enabled by tracking computations on a streaming window basis, and treating each window as an atomic "micro batch" of a finite time period.
- Application windows are supported natively as integral multiples of atomic micro batches.
- Resource utilization is distributed across a Hadoop cluster to support linear scaling to 1 billion events per second with sub-second latency.
- Fully stateful and with operability as a core building block, the DataTorrent platform is self-healing, adjusts to load via runtime automatic dynamic partitioning, and enables runtime functional changes including bucket testing.

DEVELOPMENT ENVIRONMENT AND OPEN SOURCE FRAMEWORK

- Use your favorite IDE and develop in Java—there's no need to learn domain-specific language to code applications to use DataTorrent.
- A set of standard library operator and application templates are provided that include a host of commonly used functions like adapters to various message busses, databases, and other key interfaces.
- The command line interface has the ability to access all data across the application including metrics and DAG.
- Specify applications via configuration files or the command line for easy integration with other tools and technologies in the enterprise data flow.

DATATORRENT CONSOLE

An interactive user interface is provided for monitoring, debugging, and charting real-time stream processing applications (used by developers and IT administrators).



DEPLOYMENT OPTIONS

DATATORRENT CAN BE DEPLOYED ON ANY HADOOP 2.X CLUSTER.

Firewall Considerations:	If there is a firewall in the configuration, DataTorrent servers should be placed on the Hadoop side of the firewall.
Data Ingestion & Sanitation:	DataTorrent offers an open source application for data ingestion, which can be used parallel to other tools such as Trifacta.
Storage Requirements:	DataTorrent processes in-memory and uses persistent storage for resiliency. HDFS is used by default, but any DFS-compliant storage can be used.
Compute Server Selection:	In general one container is about 2GB-4GB, and servers are about 64GB-128GB nowadays. At an average size of 96GB, 24 to 48 containers recommended per server. The performance depends on the type of computations.

SYSTEM REQUIREMENTS:

MINIMUM SERVER CONFIGURATION:

- i7, dual quad core, hyperthreaded
- Recommended containers are 2GB to 4GB. The server can have between 64GB to 256GB
- NIC should be at least 1GB. A 10GB NIC improves performance and may pay back by needing fewer servers.

OPTIONAL: Hard disk is not directly part of the spec. But given that this is Hadoop, if needing to run MapReduce, recommended hard disk is 18TB-36TB.

VIRTUAL SERVER: Using virtual machines for lab or test environments is recommended due to the high memory requirements. For virtual environments, a configuration similar to the physical server recommendations should be followed, keeping performance implications in mind.

GUI REQUIREMENTS: Supports Chrome 24 or higher (preferred browser); Firefox 18 or higher; Safari 6.0.2 or higher; Internet Explorer 9 or higher

COMPETITIVE ANALYSIS

Real-time streaming processing requirements have increased dramatically in the last couple of years due to the popularity and widespread adoption of Big Data analysis. There are just a few players in the market today that offer competitive solutions to DataTorrent. The following matrix summarizes some of our findings.

Solution Matrix	DataTorrent	Apache Storm	IBM InfoSphere Streams	TIBCO StreamBase	Apache S4	Amazon Kinesis	Apache Flume NG
Proprietary / Open Source	O	O	P	P	O	P	O
Support for Hadoop 1.x	●	●	●	●	●	●	●
Support for Hadoop 2.x	●	●	●	●	●	●	●
Native YARN	●	●	●	●	●	●	●
Dashboard	●	●	●	●	●	●	●
Extensible via Modules	●	●	●	●	●	●	●
Technical Support	●	●	●	●	●	●	●
Atomic Micro-batch	●	●	●	●	●	●	●
Events per Second	1 Billion	Thousands	Millions	Thousands	Thousands	Thousands	Thousands
Automated Parallelism	●	●	●	●	●	●	●
Dynamic Runtime Changes	●	●	●	●	●	●	●
High Availability	●	●	●	●	●	●	●
Prog. Languages Supported	Java, Python, etc.	Java, Python, etc.	SPL	Java, Python, etc.	Java, Python, etc.	●	●
Log Analysis	●	●	●	●	●	●	●
Site Operations	●	●	●	●	●	●	●
MapReduce Diagnostics	●	●	●	●	●	●	●
Open Source Operators Library	●	●	●	●	●	●	●
Open Source Application Templates	●	●	●	●	●	●	●
Complex Computations (DAG)	●	●	●	●	●	●	●
Linear Scalability	●	●	●	●	●	●	●
Security	●	●	●	●	●	●	●
CLI and Macros	●	●	●	●	●	●	●
Configuration Based Specification	●	●	●	●	●	●	●
State Checkpointing	●	●	●	●	●	●	●

ABOUT TRACE3'S BIG DATA INTELLIGENCE TECHNICAL BRIEFS

The Trace3 Big Data Intelligence (BDI) team finds and vets innovative solutions that apply to common, real-world use cases. Once a solution has been identified and gone through our technical vetting, it is installed in our BDI Innovation Labs for extensive integration testing with other complementary technologies. A Technical Brief document is developed with the participation of the solution providers involved.

Trace3 enables business transformation through a continuum of IT expertise, services and solutions designed to give organizations the ability to remain competitive in today's ever evolving marketplace.

FOR MORE DETAILED TECHNICAL INFORMATION OR TO RECEIVE A DEMO OF THIS SOLUTION, CONTACT THE TRACE3 BIG DATA INTELLIGENCE TEAM.

BDI@TRACE3.COM | WWW.TRACE3.COM/BIGDATA

TRACE3