SOLUTION BRIEF

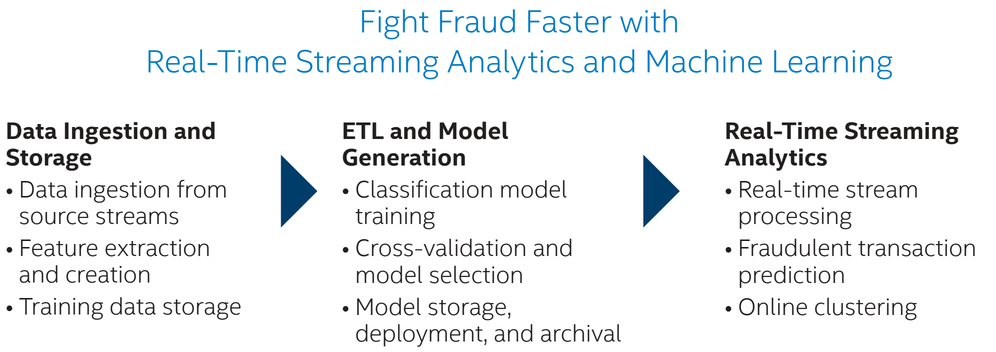
Maximize Data Value with a Real-Time Streaming Analytics Solution

Increase competitiveness and enable data-driven, split-second decisions by deploying a machine learning-enabled solution stack based on a collaborative solution from Lenovo* and Intel on the Cloudera Distribution of Hadoop*

Executive Summary

In today's business world, data never stops flowing. Traditionally decisions were made based on reports of monthly data roll-ups, and this approach still serves well for historical and archived data where time is not a crucial factor. But for time-sensitive data, answers must be delivered within seconds to be of great value for organizations. To stay competitive, enterprises must deploy real-time streaming machine learning applications in their data centers—91 percent of CIOs say streaming data analytics can positively impact their company's bottom line.[1]

Intel and Lenovo have collaborated to offer a unique high-performance, reliable reference architecture for a real-time streaming analytics solution that includes machine learning. This reference architecture is focused on fraud detection for the financial services industry, but it can be easily adapted to other use cases. The solution stack includes trusted software technology from other major industry players such as Cloudera* and Elastic*. At the foundation of this solution stack are Lenovo ThinkSystem* servers based on the powerful Intel® Xeon® processor Scalable family. Figure 1 shows how the real-time streaming reference architecture helps enterprises gain business insights and maintain a competitive edge.



Fight Fraud Faster with
Real-Time Streaming Analytics and Machine Learning

Data Ingestion and Storage

- Data ingestion from source streams
- Feature extraction and creation
- Training data storage

ETL and Model Generation

- Classification model training
- Cross-validation and model selection
- Model storage, deployment, and archival

Real-Time Streaming Analytics

- Real-time stream processing
- Fraudulent transaction prediction
- Online clustering

Figure 1. Intel and Lenovo* have collaborated to create a real-time streaming architecture tailored to the financial services industry that uses machine learning to turn raw data into deep business insights, expediting fraud detection.

Solution Benefits for the 24/7 Business

A real-time streaming architecture can offer the following benefits to enterprises in a wide variety of industries, including financial services:

- Fast insights and decision making
- Accelerated fraud detection
- Enhanced regulatory compliance
- Increased competitiveness

Business Challenge: Mountains of Data Bury Insights, Hide Fraud

Enterprises are being buried in a mountain of data. 2.7 zetabytes of data exist in the digital universe today.[2] In just a few years, business transactions on the Internet, including business-to-business and business-to-consumer, will reach 450 billion per day.[3] But all data is not created equal—according to one estimate, in 2016 bad data cost the United States 3.1 trillion dollars[4]—and quickly distilling raw data into valuable business insights can be challenging. In the financial industry, consider these facts:

- In 2015, 11 out of the 15 small credit union closures were caused by fraud.[5]
- According to the 2015 Identity Fraud Study, USD 16 billion was stolen from 12.7 million US consumers in 2014, resulting in a new identity fraud victim every two seconds.[6]
- Fraud costs the property and casualty insurance industry USD 30 billion each year in the United States alone.[7]

In the last decade, massive amounts of data have poured into enterprise data lakes, requiring equally massive extract, transform, and load pipelines to aggregate this data and generate monthly, quarterly, and yearly reports and projections. However, businesses today want to move faster. Enterprises need to adopt new business models and innovative technology to thrive in a data-first world that never sleeps.

Reference Architecture Overview: Streaming Analytics with Machine Learning

Lenovo* and Intel are taking advantage of innovations in silicon and software by creating a real-time streaming analytics reference architecture tailored for financial transaction fraud detection. As shown in Figure 2, the reference architecture uses several components to accomplish the workflow. The following sections describe each step, as well as the foundational hardware and software components (see Figure 3 on the next page), in more detail.

- **Data ingestion.** Historical transactions are stored in the Hadoop Distributed File System* (HDFS*) for batch analytics while real-time data from the Web, smartphones, and card readers are streamed to Kafka*. Records are represented in CSV strings and may use Apache Avro* for serialization and Snappy* compression. For secure implementations, encryption can be added at every stage of the pipeline along with Kerberos* for authentication. Logging and monitoring are accomplished with tools such as Elasticsearch*, Kibana*, and Logstash*.
- **Batch analytics.** The reference architecture uses Spark* and Yarn* along with Oozie* and HDFS to perform feature engineering. This task includes dropping duplicates or filtering unformatted records, and feature imputation and binning, as well as attribute generation using aggregations.

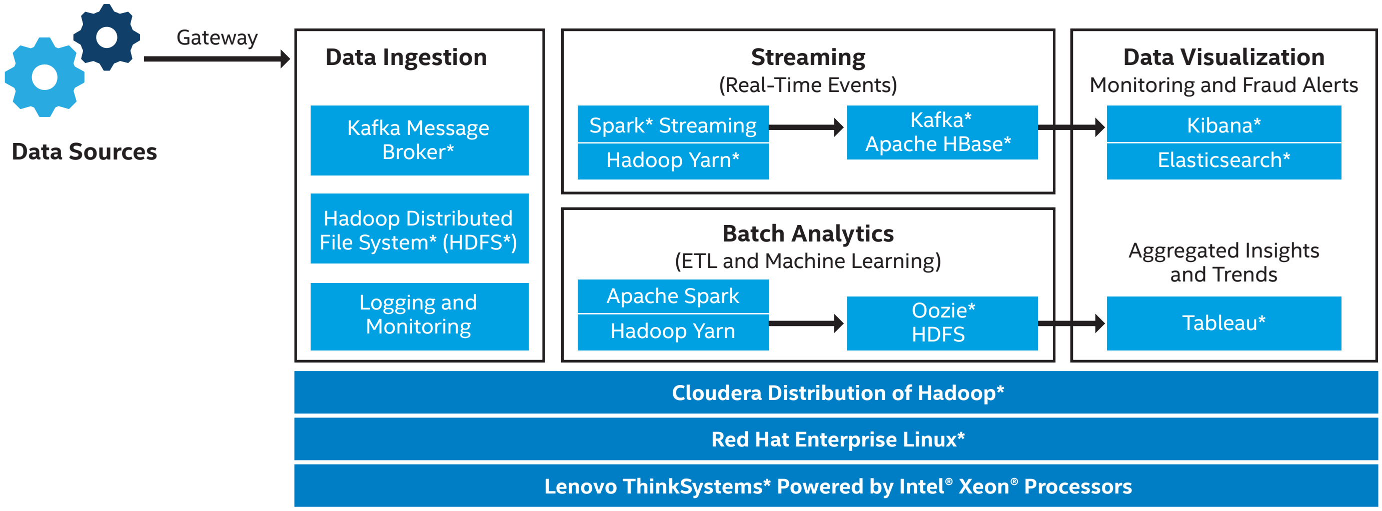
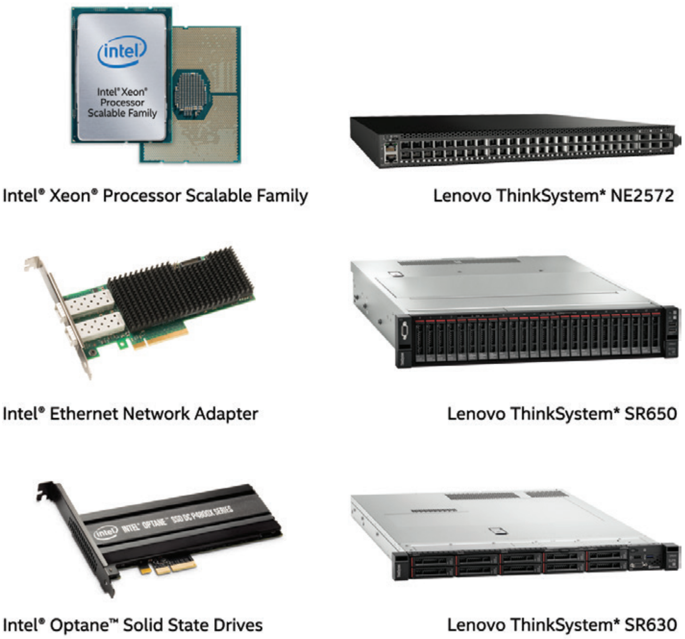


Figure 2. Lenovo ThinkSystem* servers running Intel® Xeon® processors power a Cloudera Distribution of Hadoop*-based real-time streaming analytics reference architecture suited for financial transaction fraud detection workflows.

Seaborn* or matplotlib* can be used for visualizing feature correlations. Once all features have been created, a features vector is assembled to train the machine learning algorithms. The whole machine learning pipeline is packaged and saved for auditing purposes and for reuse during the streaming pipeline.

- **Streaming analytics.** New transactions are pulled directly from Kafka brokers at various frequencies. Sizing and mini-batch intervals depend on throughput and latency service-level agreements (SLAs). For example, 20 seconds might be appropriate for certain workloads, while others might need to process the pipeline every second, or several times per second. The reference architecture uses Spark Streaming* for stream processing. Spark Streaming, apart from its maturity, provides seamless integration with Kafka and Yarn and with available libraries that support CSV parsing, Spark ML*, and elasticsearch-Hadoop connectors. Spark Streaming also enjoys widespread community involvement. The streaming workflow involves processing data through online clustering (K-Means) and fraud detection (Logistic Regression and Random Forest Classification) algorithm pipelines. Transactions are initially processed using feature engineering pipelines and then fitted into the models to obtain predictions. Fraudulent records are pushed out to Elasticsearch for rapid indexing and search as well as saved to HBase* for long-term storage (Cassandra* can act as an alternative).



Intel® Xeon® Processor Scalable Family

Lenovo ThinkSystem* NE2572

Intel® Ethernet Network Adapter

Lenovo ThinkSystem* SR650

Intel® Optane™ Solid State Drives

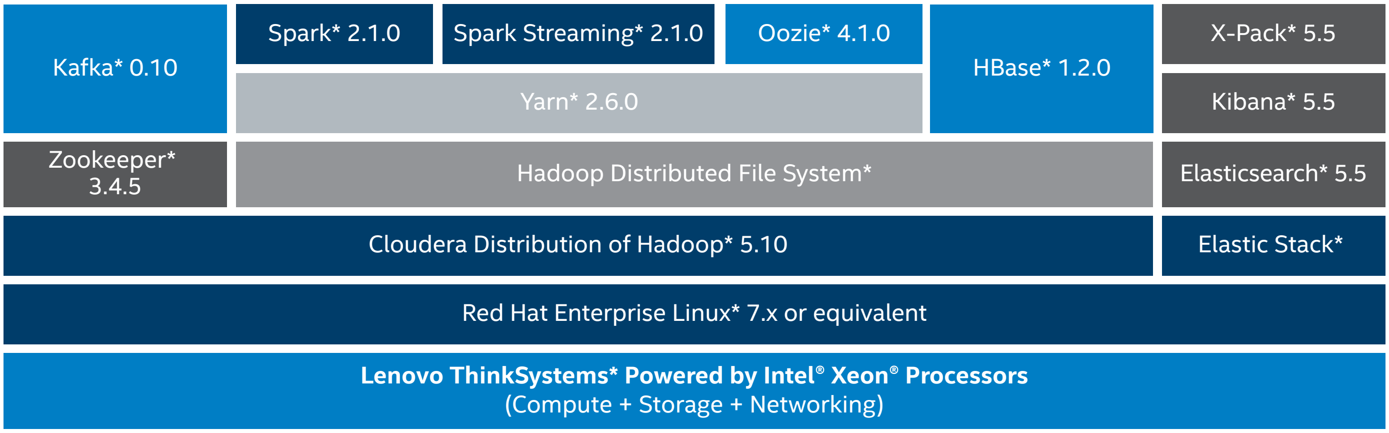
Lenovo ThinkSystem* SR630

Advanced technology from Intel and Lenovo* combine to create a unique high-performance real-time streaming analytics platform that can help financial organizations detect and prevent fraud.

- **Data visualization.** The reference architecture uses Kibana as the analytics and visualization platform. Kibana is open source and is designed to work with Elasticsearch (as part of the Elastic Stack*). Additionally, Elastic offers X-Pack* extensions for the enterprise that bundle security, alerting, monitoring, reporting, and graph capabilities.

Reference Architecture Details

- **Software.** The real-time streaming analytics platform runs on Red Hat Enterprise Linux*. Cloudera Distribution of Hadoop* provides seamless deployment and management of various Hadoop components.
- **Compute.** The server cluster consists of Lenovo ThinkSystem* servers powered by the Intel® Xeon® processor Scalable family. Data nodes use the Lenovo ThinkSystem SR650 Server, while management nodes use the Lenovo



Kafka* 0.10

Spark* 2.1.0

Spark Streaming* 2.1.0

Oozie* 4.1.0

HBase* 1.2.0

X-Pack* 5.5

Yarn* 2.6.0

Kibana* 5.5

Zookeeper* 3.4.5

Hadoop Distributed File System*

Elasticsearch* 5.5

Cloudera Distribution of Hadoop* 5.10

Elastic Stack*

Red Hat Enterprise Linux* 7.x or equivalent

Lenovo ThinkSystems* Powered by Intel® Xeon® Processors
(Compute + Storage + Networking)

Figure 3. The reference architecture is based on Intel® technology and Lenovo* technology and uses Cloudera Distribution of Hadoop* and many components of the Apache ecosystem.

ThinkSystem SR630 Server. ThinkSystem servers, combined with ThinkSystem network switches, are designed to provide the high-performance compute power, security, and agility necessary for real-time streaming analytics.

- **Networking and memory.** The reference architecture uses the Lenovo ThinkSystem NE2572 Network Switch for the data nodes (two per rack) and the Lenovo ThinkSystem G8052 Network Switch for the management nodes (one per rack).
- **Storage.** Storage for Kafka was deployed on 4-TB Intel® Solid State Drive Data Center P4500 Series (featuring NVMe*) for high throughput and bandwidth, while Elasticsearch stored its data on low-latency Intel® Optane™ SSD DC P4800X Series.

Test Configuration

The following list describes a representative deployment that can be scaled up or down based on throughput needs and capacity planning:

- 18-node deployment of servers based on the latest generation Intel® Xeon® Scalable processor.
- Six nodes dedicated to Spark, Spark Streaming, and Yarn (processing).
- Four nodes shared between Kafka and Elasticsearch (data).
- Four nodes for long-term storage (HBase).
- Three nodes required for Cloudera management nodes (Primary, Secondary, and Journal node).
- One node acts as the bastion and gateway for incoming transactions and as the Elasticsearch Load Balancer. It also hosts Kibana visualization dashboards.

In our tests, we used a 10 Gbps network switch; however, network utilization was fairly high on all nodes, effectively saturating the network. Therefore, for even better performance, we highly recommend using high-throughput networking and memory, such as a 25 Gbps Ethernet switch.

Solution Performance: Fight Fraud Faster Using the Intel® Xeon® Processor Scalable Family

The Intel® Xeon® processor Scalable family is a new micro-architecture with many additional features compared to the previous-generation Intel® Xeon® processor E5-2600 v4 product family. These features include increased processor cores, increased memory bandwidth, non-inclusive cache, Intel® Advanced Vector Extensions 512, Intel® Memory Protection Extensions, Intel® Ultra Path Interconnect, and sub-NUMA clusters. Highlights for enterprise innovation include support for Intel® Optane™ SSDs and Intel® 3D NAND SSDs as well as Intel® Run Sure Technology.

The real-time streaming analytics solution combined with machine learning enables firms to detect suspicious transaction or network traffic patterns in seconds more accurately than traditional rule based analytical engines. The solution also ingests external data sources that can augment the streaming data. Stream-processing transaction data helps detect anomalies that signal fraud in real time, and then stops fraudulent transactions before they are completed.

The setup is able to process a steady-state throughput of 1 million transactions per second and peak throughput (to accommodate data spikes) of **1.5 million transactions per second** at a capacity of **24 to 36 GB of data ingestion per minute**. The deployed model shows **up to 95 percent accuracy** against the test data. The streaming pipeline combines the best components in the Hadoop ecosystem along with the raw computation power of the Intel Xeon processor Scalable family to deliver a comprehensive feature engineering and machine learning inference pipeline. This sort of accurate, high-speed performance means that financial organizations can significantly reduce the billions of dollars lost to fraud each year (see Figure 4).

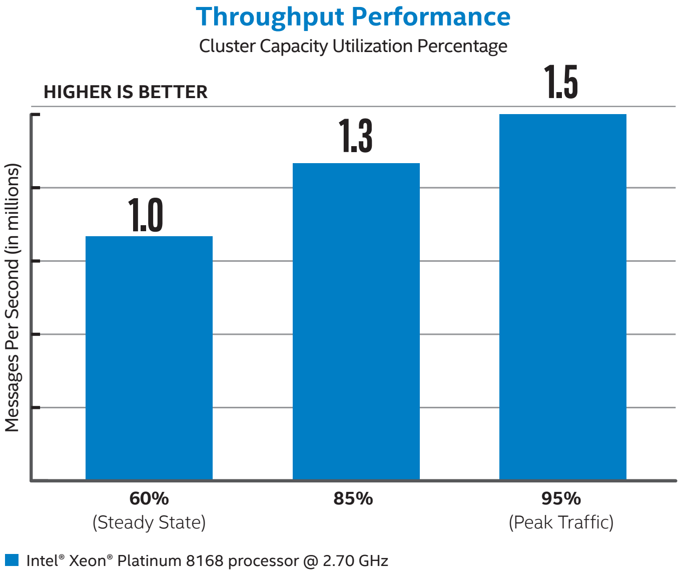


Figure 4. Complex stream processing Throughput with Machine Learning on Intel® Xeon® Scalable processors compute cluster.

Pushing Storage Latency with Intel® 3D NAND SSDs and Intel® Optane™ SSDs

Intel 3D NAND SSDs P4500 Series and Intel Optane SSDs P4800X Series provide high throughput and low latency storage for real-time frameworks such as Kafka and Elasticsearch (see Figure 5).

We executed an ensemble of queries resembling real-world web traffic against the Elasticsearch cluster, such as Exact Search, Range Query, Term Match, Geo Bounding, and Random Score. Table 1 shows the results.

These workloads were run primarily with the out-of-the-box settings for most open source components. While the workflow utilizes the compute, memory, and networking to its capacity and efficiently uses fast NVMe and solid-state drives for I/O, the pipeline can be further optimized for different SLAs targeted for high-throughput or low-latency scenarios, as business needs dictate.

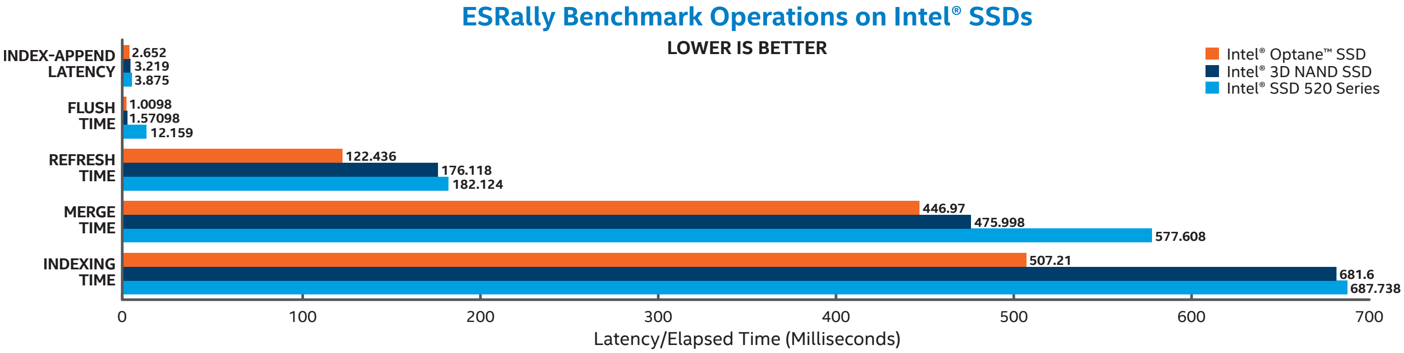


Figure 5. Deploying Kafka* on Intel® 3D NAND SSDs provides exceptionally high throughput for incoming raw messages while Intel® Optane™ SSDs provide low indexing and search latencies boosting overall performance of Elasticsearch*.

Table 1. Test results from queries

METRIC	EXACT SEARCH	RANGE QUERY	TERM MATCH	GEO BOUNDING	RANDOM SCORE
Load Split	80 percent	6.75 percent	6.75 percent	6 percent	0.5 percent
Throughput	1608 requests/sec	136 requests/sec	136 requests/sec	120 requests/sec	10 requests/sec
Average Latency	8 ms	79 ms	68 ms	99 ms	561 ms

Real-Time Streaming Analysis for a Broad Set of Functionalities

- **Clinical healthcare.** Analytics in healthcare is a rapidly growing area for implementations of big data and machine learning. Real-time analytics provides the ability to monitor patient safety, personalize patient results, and assess clinical risk as well as reduce patient readmission—all of which improve organizational efficiency and patient experience.
- **Transportation and fleet operations.** The transportation industry faces rapid changes. Autonomous vehicles, advanced driver assistance systems, and the Internet of Things are increasing transportation data set volumes exponentially. Plus, more and more of the population is moving to urban areas, increasing traffic density. Real-time streaming of data coming from edge devices can help predict speed and travel times, best routes, and manage traffic density.
- **Retail.** Retailers, who often have stores worldwide as well as an online presence, struggle with real-time inventory tracking. Real-time streaming analytics based on a central scalable repository can improve operational efficiency, as well as drive higher sale volumes, more trend insights, and enhanced customer satisfaction.
- **Customer churn and call center analytics.** Call center logs can act as a central input to predict customer churn. These logs can include data from sensors, online interactions, interactive voice response systems, and IT support systems. By streaming this data into a data lake, enterprises can create models for churn prediction as well as derive insights into customer metrics and incidents.
- **Network traffic monitoring and fraud detection.** Cybersecurity is among the most critical threats enterprises face today. Increasingly sophisticated hackers continuously seek vulnerabilities they can use to steal data. Tools such as Apache Spot* can help expedite threat detection with machine learning models that perform streaming analysis of network flows, DNS data, proxies, and so on.
- **Financial fraud detection and credit transaction monitoring.** Financial transactions represent a continuous stream of data from several sources. Real-time streaming analytics can aggregate this data to improve business health and reveal insights. In addition, previous transactional data can be used to train machine learning models to predict fraudulent transactions.

Conclusion

The modern data explosion, combined with an increasing need for decisions based on real-time data, makes real-time streaming analytics bolstered by machine learning an absolute necessity for firms—especially financial services organizations—who want to remain competitive. Organizations adopting this reference architecture for their infrastructure platform enables real-time decision making and helps prevent fraudulent transactions.

By using the reference architecture described here, customers can be confident they will get the performance and reliability they need.

Intel Xeon Scalable processors enable new, differentiated services that will allow customers to innovate and grow into new markets; but to do so, performance efficiency must be as disruptive as the services being created. The Intel Xeon Scalable processor platform along with Intel 3D NAND and Intel Optane SSD storage delivers disruptive performance efficiency across workloads, and early adopters of the technology are experiencing performance improvements.

To learn more about this solution and real-time streaming analytics, contact your Lenovo representative or visit **lenovo.com/us/en/data-center/solutions/big-data**.

Learn More

You may also find the following resources helpful:

- Lenovo ThinkSystem portfolio
- Intel Xeon Scalable Processors
- Intel Data Center Information
- Cloudera Enterprise Data Hub
- Intel Financial Services Solutions

IMPORTANT NOTE

Note that while this solution was tailored for the financial services industry, it can easily be modified to suit the needs of a wide variety of sectors, including transportation, retail, healthcare, and many others.

A Closer Look

Lenovo is a leading provider of x86 servers for the data center. Featuring rack, tower, blade, dense, and converged systems, the Lenovo server portfolio provides excellent performance, reliability, and security. Lenovo also offers a full range of networking, storage, software, solutions, and comprehensive services supporting business needs throughout the IT life cycle. With options for planning, deployment, and support, Lenovo offers the expertise and services needed to deliver better service-level agreements and generate greater end-user satisfaction.

Intel is a world leader in computing innovation. The company designs and builds the essential technologies that serve as the foundation for the world's computing devices. Additional information about Intel is available at newsroom.intel.com and blogs.intel.com.

Solution Provided By:

Lenovo intel

[1] How Real-Time Data Boosts the Bottom Line. cioinsight.com/it-management/innovation/slideshows/how-real-time-data-boosts-the-bottom-line.html

[2] Big Data Brings Marketing Big Numbers. marketingtechblog.com/ibm-big-data-marketing

[3] Revisited: The Rapid Growth in Unstructured Data. wikibon.org/blog/unstructured-data

[4] The Cost of Bad Data. information-age.com/cost-bad-data-123465972

[5] Fraud a Growing Risk for Credit Unions. casewareanalytics.com/blog/fraud-growing-risk-credit-unions

[6] 5 Top Fraud Risks for Financial Institutions in 2016. securitytoday.com/articles/2016/01/27/5-top-fraud-risks-for-financial-institutions-in-2016

[7] Deloitte: Workers' Compensation, Auto Lead in P&C Fraud Claims. wci360.com/news/article/deloitte-workers-compensation-auto-lead-in-pc-fraud-claims

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to intel.com/performance.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Test configuration (actual deployment configuration may vary): a) Stream Processing: 2S Intel® Xeon® Platinum 8168 processor @ 2.70 GHz, 384 GB memory, Red Hat Enterprise Linux* 7.3; b) Kafka*/Elasticsearch*: 2S Intel® Xeon® Gold 6138 processor @ 2.00 GHz, 384 GB memory, Red Hat Enterprise Linux 7.3; c) Storage: ST2000NX0403 8 SATA Seagate* drives (HDFS), Intel SSD DC S3710 Series (OS), Intel® SSD DC P4500 Series 4-TB (Yarn*, Kafka*), Intel® Optane™ SSD P4800X 375 GB (Elastic Search; d) Ethernet controller: Intel® Ethernet Connection X722 for 10GBASE-T (rev 03), Intel® Ethernet Controller 10 Gigabit X540-AT2 (rev 01). For more complete information visit intel.com/performance/datacenter.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Learn About Intel® Processor Numbers.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Copyright © 2017 Intel Corporation. All rights reserved. Intel, the Intel logo, Intel Optane, and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

0917/BGOW/KC/PDF ♻ Please Recycle 336460-002US