Streaming Data: Why It Makes Sense and How to Work With It

White paper

September 2015

This white paper is sponsored by Zoomdata.

Introduction

To build and maintain competitive advantage, organizations are squeezing more and more value out of the data they collect and the analytics they perform on that data. The faster and more timely the analysis the more value it holds in today's rapidly shifting competitive business world—time is money. Insights and opportunities may exist only for a brief period of time. By relying on outdated data storage and analytics methods, organizations put themselves at risk, failing to uncover and act on critical insights before their competitors. Batch-oriented data warehouses are too cumbersome and unwieldy to provide the required timely analytics. Guided by a new generation of technologies, organizations are beginning to reap the rewards of streaming-data analysis of both real-time and historical data.

This white paper examines the evolution from batch to streaming-data analysis and provides key insight into building streaming data capacity. The paper then moves to an evaluation of Zoomdata's strength as a visual streaming-data analytics solution.

Data Naturally Exists in Streams

All commerce, whether conducted online or in person, takes place as a stream of events and transactions; only the limitations of technology forced it into batches. In the beginning, the stream was recorded in a book—an actual book that held inventories and sales, with each transaction penned in on its own line on the page. Over time, the practice evolved. Books yielded to computers and databases, but practical limitations still constrained data processing to local operations. Later on, data was packaged, written to disk, and shipped between locations for further processing and analysis. Concatenating the data stream into batches made it easier to store and transport.

Technology marches on, and it now has evolved to the point that, in many cases, batching is no longer necessary. Systems are faster, networks are faster and more reliable, and programming languages and databases have evolved to accommodate a more distributed streaming architecture. Physical retail stores used to close for a day each quarter to conduct inventory. Then they evolved and did batch analysis of various locations on a weekly basis, and then a daily basis. Now they keep a running inventory that is accurate through the most recent transaction. Not only does consuming data in streams more closely reflect the reality of the data, but there are numerous technological drawbacks to consuming data in batches. Batches themselves can get lost or stolen, especially if there are multiple copies of data, in multiple locations, each of which must be tracked and secured. Version control is also of great concern when working with batches.

Traditional, batch-oriented data warehouses pull data from multiple sources at regular periods, bringing it to a central location and assembling it for analysis. The practice causes data management and security headaches that grow larger over time as the number of data sources and the size of each batch grows. It takes a lot of time to export batches from the data source and import them into the data warehouse. In very large organizations, for which time is of the essence, batching can cause conflicts with backup operations. And the process of batching, transporting, and analysis often takes so much time that it becomes impossible for a complex business to know what happened yesterday or even last week.

By contrast, with streaming-data analysis, organizations know they are working with the most recent—and timely—version of data because they stream the data on demand. By tapping into data sources only when they need the data, organizations eliminate the problems that storing and managing multiple versions of data present. Data governance and security are simplified; working with streaming data means not having to track and secure multiple batches.

We live in an on-demand world. It's time to leave behind the model of the monolithic, complex, batch-oriented data warehouse and move toward a flexible architecture built for streaming-data analysis. New technologies such as Zoomdata that are able to process and visualize data streams ensure the timely and accurate analysis that enables enterprises to harness the value of the data they work so hard to collect, and tap into it to build competitive advantage.

Moving Toward Streaming-Data Analysis

Previously, building streaming-data analysis environments was complex and costly. Traditional ETL and data management solutions took months or years to deploy. They required expensive, dedicated infrastructure, suffered from a lack of interoperability, required specialized developers and data architects, and failed to adapt to rapid changes in the database world (such as the rise of unstructured data).

In the past few years we have witnessed a flurry of activity in the streaming-data analysis space, both in terms of the development of new software and in the evolution of hardware and networking technology. Always-on, low-latency, high-bandwidth networks are less expensive and more reliable than ever before. Inexpensive and fast memory and storage allow for more efficient data analysis.

This has been accompanied by explosive growth in the number of streaming data sources and the volume of streaming data. It's no longer enough to look to historical data for business insight. Organizations require timely analysis of streaming data from such sources as Internet of Things (IoT), social media, location, market feeds, news feeds, weather feeds, website clickstream analysis, and live transactional data. Examples of streaming-data analytics include telecommunications companies optimizing mobile networks on the fly using network device log and subscriber location data, hospitals decreasing the risk of nosocomial (hospital-acquired) infections by capturing and analyzing real-time data from monitors on newborn babies, and office equipment vendors alerting service technicians to respond to impending equipment failures.

Microservices

The rise of microservices is changing the landscape of enterprise database and application architecture. Software applications are designed around business capabilities as suites of independently deployable services that communicate using lightweight mechanisms.

Microservices can be coupled with bite-size compute instances using an automated deployment framework to scale up and down transparently. This architecture is extremely valuable in designing streaming-data analytics.

Coupled with the increase of streaming-data sources is a rapid increase in the number of easy-to-use, inexpensive, and open-source streaming-data-platform components. Apache Storm, a Hadoop-compatible add-on (developed by Twitter) for rapid data transformation, has been

implemented by The Weather Channel, Spotify, WebMD, and Alibaba.com. Apache Spark, a fast and general engine for large-scale data processing, supports SQL, machine learning, and streaming-data analysis. Apache Kafka, an open-source message broker, is widely used for consumption of streaming data. And Amazon Kinesis, a fully managed, cloud-based service for real-time data processing over large, distributed data streams, can continuously capture large volumes of data from streaming sources. New databases such as MongoDB, Cassandra, HBase, and DynamoDB are popular data stores for sinking and persisting streaming data.

However, a streaming data infrastructure and myriad sources of streaming data are useful only when subjected to the proper analytics. The burgeoning field of analytics uses techniques such as machine learning and predictive modeling to uncover the valuable business insights that would otherwise be trapped within virtually unreadable columns of numbers. A key component of a successful analytics program is a visual analytics technology, like Zoomdata, that provides an intuitive GUI through which a user can manipulate and report on data to test hypotheses. Visual analytics technologies serve as a gateway to unlock the value contained within historical and streaming data sources.

A Practical Approach to Making the Transition

Many large organizations have already made a considerable investment in building batch-oriented data warehouses. They work with batch-based analytics and gain valuable insight from them. For practical purposes, where companies have already realized value in batch-based analysis, they should continue to work with batches for those existing projects.

Beginning to work with streaming data does not have to mean discarding those batch-oriented architectures wholesale. As IT organizations test the waters of streaming-data analytics, they can and should build hybrid batch- and streaming-data analytics programs. Leave existing batch processing in place for now, build new applications that use streaming data as needed, and run a hybrid system that uses both streaming- and batch-data sources.

New projects, however, should work with streaming data where applicable. As new data sources become available, they should be analyzed as data streams. These data sources include IoT, logs from devices, systems, servers, and transactions, and streams offered for consumption such as market feeds, news feeds, and weather. Location-based data is also well-suited for streaming analytics. New data sources, formats, and structures provide greater flexibility when analyzed as streams.

Several architectural solutions are available for organizations seeking to analyze a combination of streaming- and batched-data sources. For example, the Lambda Architecture includes a batch layer for historical data, a speed layer for analyzing real-time data streams, and a serving layer that indexes the batches so they can be queried side-by-side with the data streams. However, although the Lambda Architecture allows organizations to preserve their existing investment in batch-processing, it is a complex, multitier architecture that requires the development and maintenance of multiple code bases. Another option is the Kappa Architecture, which treats all data, including batches, as a stream. The primary advantage of a Kappa Architecture is that it requires a single code base, avoiding the complexity of a Lambda Architecture.

Although architecture is a fundamental consideration for infrastructure teams, it is nevertheless critical to choose a visual analytics solution that can provide real value for an organization regardless of architecture because the visual analytics solution serves as the business user's interface to the data. Therefore, a flexible and extensible visual analytics solution such as Zoomdata that can leverage any architecture is a requirement for organizations for access to critical insight and business intelligence.

Zoomdata Is Built for Streaming-Data Analysis

Zoomdata excels as a visual analytics solution for streaming-data analysis. The interactive visual interface is designed for business users to explore data sets without having to learn complex coding. Users can consume, filter, pause, rewind, replay and otherwise interact with streams of data. As users interact with data, Zoomdata pushes the questions into the stream so that they can be executed in the big data source.

At the core of Zoomdata is a purpose-built stream-processing engine that relies on massively parallel processing technologies to quickly work with huge volumes of streaming data. This streaming architecture provides the fastest visual analytics experience for both historical and real-time data. Data streams from the source, through the stream-processing engine, and to the user’s browser via a WebSockets connection. Using sophisticated microquery and data-sharpening techniques, Zoomdata shows complex results instantly and refines them as further analysis takes place. As new data is entered into the original data source, push-based updates allow for real-time data analysis. Because the system interacts seamlessly with streaming data from live and historical sources, end users experience no delay.

Behind the scenes, Zoomdata was designed and architected for big data analysis, and it connects to a multitude of big data platforms using native APIs. In addition to working with streaming technologies like those mentioned above, Zoomdata also connects to any of the modern big data sources such as Hadoop, search (ElasticSearch, Cloudera Search, Apache Solr), NoSQL databases (Cassandra, MongoDB), cloud apps such as Salesforce, and SQL databases (MySQL, Oracle).

Unlike other visual data analysis tools, Zoomdata does not build cubes or lenses, or move and import data into an additional data source, before a user can query it. A BI tool that relies on cubes or lenses cannot work with streaming data efficiently and quickly because the stream of data is simply moving too fast to pump it all into a cube. Zoomdata scales efficiently for streaming by pushing the query to the source and then continuing to treat the data as a stream all the way through to the visualization.

While visual analytics over an individual source is powerful, Zoomdata Fusion enables visual analytics across multiple sources. Zoomdata Fusion makes multiple data sources appear as one, combining historical and real-time data to enrich both and provide deep insights. Many enterprises have discovered that relying on multiple disparate data sources means not being able to store data a single framework. Zoomdata overcomes that obstacle by processing data in place, transparently combining query results from new and traditional, structured and unstructured sources for both real-time and historical systems. This approach greatly simplifies the process

compared with batch-based data warehousing, most notably removing the need for a data architect to predefine destination tables and then conduct analyses.

Conclusion

Streaming-data analysis is rapidly overtaking batch-processed data analysis in today's business world. This trend makes the ability to rapidly visualize streaming data in a fast, agile, and responsive manner a critical capability for businesses seeking to gain maximum benefit from big data initiatives. Data is a valuable commodity, and working with the right data in a timely manner to gain business insight confers a competitive advantage.

Zoomdata is optimally built to provide visual analytics for any streaming-data platform or architecture. Zoomdata is purpose-built for big data visualization and analysis. The platform uses native APIs to connect to data sources and execute queries directly against source data. This process allows Zoomdata to perform complex concurrent analyses of multiple data sources without the need for time-consuming batch imports.

About Sarrel Group

Sarrel Group is a technology product assessment, editorial services, and IT consulting firm with offices in New York City and San Francisco. Sarrel Group helps technology companies develop sales and marketing programs based on lab-tested validation of their products' competitive advantages. For more information, please visit www.sarrelgroup.com or call 866-MSARREL.

About Zoomdata

Zoomdata develops the world's fastest visual analytics solution for big data. Using patented Data Sharpening™ and Micro-query technologies, Zoomdata empowers business users to visually consume data in seconds, even across billions of rows of data. Zoomdata Fusion enables interactive analytics across disparate data sources, bridging modern and legacy data architectures, blending real-time streams and historical data, and unifying enterprise data with data in the cloud. Delivered in a micro-services architecture for elastic scalability, Zoomdata runs on premises, in the cloud or embedded in an application. With offices in Washington, D.C. and Silicon Valley, Zoomdata is venture-backed by Accel Partners, Columbus Nova Technology Partners and NEA.