the
Incubator
research

DOMAIN PERSPECTIVE

Observability Pipeline

Key Benefits

For large enterprises, delivering an observability pipeline is often the first place to start on an observability journey. By doing so, an organization can achieve the following benefits:

- Provides data autonomy which prevents vendor lock-in and supports a data independence strategy
- Supports a managed self-service model, aiding an org. shift to *Product Mode* while improving the *Developer Experience* (DX)
- Allows capture of arbitrarily wide events which are critical to high cardinality services
- Consolidates data collection and instrumentation across Teams, the estate, and/or the enterprise
- Decouples data sources from data sinks for experimentation flexibility
- Supports input-to-output schema normalization
- Provides a mechanism to encode routing, filtering, and transformation logic

A TAXONOMY

Observability Pipeline

Due to concepts of continuous delivery & deployment, assurance strategies such as periodic service repaving, architectural patterns such as EMA, EDA, SEDA, and vendors in the observability space converging on features – have led to a desire to capture, process, and route data in a manner that promotes more comprehensive observability.

This facility is an Observability Pipeline, and like any conceptual architecture it has a ***taxonomy*** of requisite elements as well as both commercial and open-source options. These include:

- **Data Specification/Protocol Support** – structuring logging aides in debugging. If you cannot structure logging across Product Teams, then you will need to provide support across a broad range of protocols, both proprietary and open.
- **Data Collector** – responsible for collecting data from sources and writing it to the data pipeline (Logstash, FluentD, Beats, daemon sets)
- **Data Pipeline** – highly scalable and available data stream which can handle and manipulate data being generated (Apache Kafka, Apache NiFi, Google Cloud Pub/Sub, Amazon Kinesis Data Streams, Liftbridge)
- **Data Router** – consumes data from the pipeline, performs filtering, and writes it to the appropriate destinations (using client libraries in stateless delivery)
- **Management and Auditability** – configuration management requires unit testing frameworks for configurations and intense code and configuration review processes, and monitoring must be implemented to ensure introspection, data capture/recapture, and troubleshooting is possible.
- **Performance at Scale** – the system must be able to accommodate potentially multi-petabyte daily volumes and commercial products should have benchmark data available.

Source Site: https://ink8r.com

For more information, please contact us:

Satbir Sran
satbir@ink8r.com

Darren Boyd
darren@ink8r.com

Event Streaming (ES)

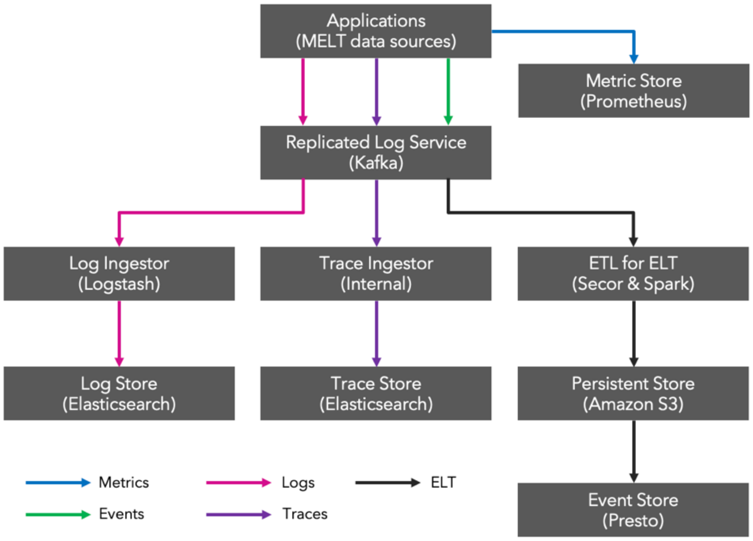
Event streaming is a core component of an observability pipeline, so it is therefore tempting to leverage existing stream processing platforms in use in the enterprise. There are well documented architectures (both open and closed) that leverage event streaming technologies as receiver of *Metrics, Events, Logs, and Traces* (MELT) data which then subsequently leverages various transform technologies to eventually write to persistent stores.

When contemplating an observability pipeline solution, consider its overall intention of routing MELT data from any source to any destination, while enriching, transforming, and optimizing the data via compliant standards. For most organizations this means considering 10s of thousands of existing agents (closed and open) already deployed, along with various wire protocols that must be considered. Replacing or adding agents contributes to agent fatigue, as is introducing potentially additive connectors to handle wire protocols (e.g. Syslog, Kafka, Amazon Kinesis, Splunk's HTTP Event Collector).

Leveraging a stateful receiver like Kafka by itself may be an interim measure, but for enterprises that are looking to perform transforms such as populating data in index-time fields or as time-series metrics, and processing of the data to include data enrichment, data aggregations, and/or removal of superfluous data, would require multiple processing steps (e.g. Fluentd and Sanitizer as a filter plugin could pull data off of Kafka topics onto a secondary topic, while Logstash could enrich the information and store on a final topic for consumption).

ES Products & Solutions

Slack uses Apache Kafka to collect MELT data, and Pinterest Secor/Apache Spark to write events, logs, and traces to Amazon S3 in a columnar Parquet file that they then query using Presto.



Applications (MELT data sources)
Metric Store (Prometheus)
Replicated Log Service (Kafka)
Log Ingestor (Logstash)
Trace Ingestor (Internal)
ETL for ELT (Secor & Spark)
Log Store (Elasticsearch)
Trace Store (Elasticsearch)
Persistent Store (Amazon S3)
Event Store (Presto)
Metrics
Logs
ELT
Events
Traces

Humio is a commercial product acquired by Crowdstrike and is an example of a platform that leverages Kafka internally for queuing incoming messages and for storing shared state when running in a cluster. If Humio manages Kafka (default, but you can add a prefix to topics in an existing Kafka deployment), then their front end will accept ingest requests, parse them, and put them on the queue where their backend processes the events and stores them into their datastore.

Splunk Data Stream Processor (DSP), through their acquisition of Streamlio, leverages both Kafka as well as Apache Pulsar's decoupled architecture (advantageous under conditions such as latency problems).

Cribl, an enterprise solution in the observability pipeline space, leverages proprietary technology to handle streaming while promoting transforms and enrichment.

StreamWeaver, a BMC acquisition also leverages the Apache Kafka architecture in their solution.

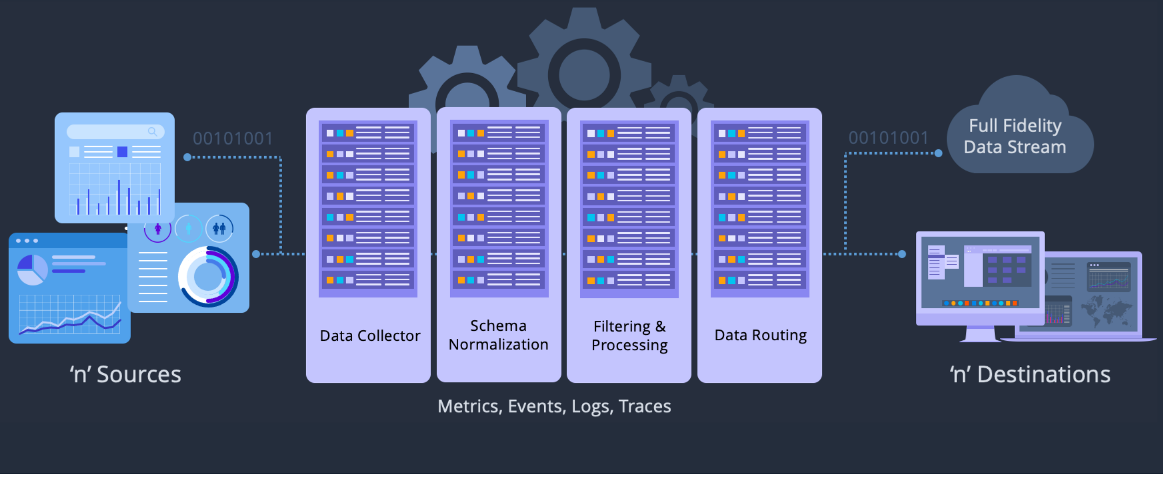
Source Site: https://ink8r.com

For more information, please contact us:

Satbir Sran
satbir@ink8r.com

Darren Boyd
darren@ink8r.com

Observability Pipeline Diagram



Commercial Products

Observability Pipelines at enterprise scale represent a relatively new technology focus, and therefore does not have its own category in any research publications. Due to data stream processing being a core requirement, often this technology focus gets conflated with Event-Driven Integration (EDI) solutions, which aims to solve a different set of challenges. Herein we discuss some entrants in this space.

Market Leader

Cribl

Cribl came to market with its flagship product, LogStream, around the same time that Splunk began talking about its own stream-processing engine. Cribl LogStream provides a flexible approach to managing logs such that large organizations can gain much better control over their logging operations in a way that could deliver several important outcomes, including increased security and governance, lower costs (processing and storage), and the ability to collect the volume of data that internal users demand. Cribl have established themselves as the leader in this space. Ink8r has interviewed several prominent customers leveraging Cribl to reduce log management costs (25%-50%), achieve vendor independence (e.g. seamless SaaS migrations), data independence, and provide data enrichment for security use cases.

Datadog

Datadog acquired Timber Technologies to boost its capabilities in data stream processing. Timber was originally developing a log management platform but pivoted to data stream processing with their flagship product Vector, after seeing the swell of log management services available on the market and that user challenges often revolved around data movement itself. Vector enables data to be transformed as it is routed to multiple destinations and therefore has potential to broaden their appeal to different personas from site reliability engineers (SREs), to developers, to security and IT ops teams. At the time of this paper Vector is not ready for GA with enterprise features.

Source Site: https://ink8r.com

For more information, please contact us:

Satbir Sran
satbir@ink8r.com

Darren Boyd
darren@ink8r.com

Commercial Products (cont'd)

Splunk

Splunk's Data Stream Processor (DSP) can collect data from several sources, including Splunk, Apache Kafka, Amazon Kinesis and Azure Event Hubs. Once processed, the data can be sent to Splunk Enterprise, Apache Kafka, and Amazon Kinesis. At this time of this publication DSP was unable to collect data from tools such as Elastic Beats or write to destinations like Elastic and Honeycomb, though we expect to see the product evolve. Splunk's Data Stream Processor appeals to organizations that primarily use Splunk but can realize benefits from data processing.

Humio

Humio was founded in 2016 and acquired by Crowdstrike in 2021. Humio offers cloud log management and observability capability. Its index-free design ingests structured and semi-structured log, application, and feed data, which it efficiently manages through data compression and flexible ingest plans (including an unlimited option). Users can view their data in real time through visualizations that can enable DevOps, SecOps, and ITOps teams alike. Users can also query this data to gain context about specific use cases. A broad range of sources are supported with parsers/log shippers for syslog, fluentd, beats, and others.

Blue Medora

Blue Medora's BindPlane collects data once and ships to various destinations. Their approach to architecture reduces agent overhead by enabling end users to deploy BindPlane collectors to collect operations data and logs from a growing list of infrastructure and application components, and then send data to the monitoring tools of their choice. BindPlane doesn't offer the same kinds of fine-grained controls that other platforms supply to enrich and transform the data at this point.

StreamWeaver (a BMC Company)

StreamWeaver helps its customers move data between monitoring tools, breaking down silos but not addressing headaches related to running many agents at the edge. While capable of filtering and masking transforms the product is currently limited in its integrations.

Logzilla

LogZilla's Network Event Orchestration (NEO) is a technology platform that can significantly lower the volume of data sent to software like Splunk and Elastic, thus reducing licensing, hardware, and maintenance costs associated with deploying those tools. NEO ingests events and logs, de-dupes the information, and forwards the resultant data set to destinations of choice. Logzilla claims that NEO can ingest 855,000 events per second per server, or about 40TB a day and has unlimited consumption models. The solution can perform as a Centralized Log Management platform (CLM), or as a forwarder to other products/services such as Splunk.

Edge Delta

Tech to Watch

Having raised $15M in June 2021, Edge Delta offers a decentralized agent-based model that allows metrics, events, logs, traces (MELT) data to begin analysis at source. This allows local event analysis which can then be combined with other data sets at a later phase. The vendor publishes the performance improvements this decentralized model promotes, relative to solutions such as SplunkUF/HF, FluentD, Vector, Logstash, etc. As their product is nascent, time will validate their claims of approximately 90% improvement in TCO compared to traditional centralized services. This is a company to watch for sure.

Source Site: https://ink8r.com

For more information, please contact us:

Satbir Sran
satbir@ink8r.com

Darren Boyd
darren@ink8r.com