Stream Processing With Apache Flink

== Development == 1470b72c2d

Apache Kafka This "leads to larger network packets, larger sequential disk operations, contiguous memory blocks [. Kafka can connect to external systems (for data import/export) via Kafka Connect, and provides the Kafka Streams libraries for stream processing applications. Apache Kafka is a distributed event store and stream-processing platform. It is an open-source system developed by the Apache Software Foundation written in Java and Scala. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. Kafka uses a binary TCP-based protocol that is optimized for efficiency and relies on a "message set" abstraction that naturally groups messages together to reduce the overhead of the network roundtrip. It is an open-source system developed by the Apache Software Foundation written Apache Kafka is a distributed event store and stream-processing platform.] which allows Kafka to turn a bursty stream of random message writes into linear writes. " 1470b72c2d

== History == 1470b72c2d The second generation uses the pull mode in data transportation, and file system in data storage. It paid more attention to stability and reliability, and shows a comparable performance to the first generation in response time and Kafka on log collection. 1470b72c2d Analysts have suggested that CEP will give organizations a new way to analyze patterns in real-time and help the business side communicate better with IT and service departments. CEP has since become an enabling technology in many... 1470b72c2d There are many players in the field of real-time stream processing and Samza is one of the mature products. It was added to Apache in 2013. 1470b72c2d A Storm application is designed as a "topology" in the shape of a directed acyclic graph (DAG) with spouts and bolts acting as the graph vertices. Edges on the graph are named streams and direct data from one node to another. Together, the topology acts as a data transformation pipeline. At a superficial level the general topology structure is similar to a MapReduce job, with the main difference being that data is processed in real time as opposed to in individual batches. Additionally, Storm topologies run indefinitely until killed, while a MapReduce job DAG must eventually end. 1470b72c2d Samza allows users to build stateful applications that process

data in real-time from multiple sources including Apache Kafka. 1470b72c2d The first generation uses the push mode in data transportation, and relational database in data storage. It shows low latency in message delivery and meets the command of a typical E-commerce platform with distributed transactions. 1470b72c2d Spark and its RDDs were developed in 2012 in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs: MapReduce programs read input data from disk, map a function across... 1470b72c2d Flink provides a high-throughput, low-latency streaming engine as well as support for event-time processing and state management. Flink applications are fault-tolerant in the event of machine failure and support exactly-once semantics. Programs can be written in Java, Python, and SQL and are automatically compiled and optimized into dataflow programs that are executed in a cluster or cloud environment. 1470b72c2d == Development == 1470b72c2d Samza is used by multiple companies. The biggest installation is in LinkedIn. 1470b72c2d == Overview == 1470b72c2d Kafka was originally developed at LinkedIn, and was subsequently open sourced in early 2011. Jay Kreps, Neha Narkhede and Jun Rao helped co-create Kafka. Graduation from the Apache Incubator occurred on 23 October 2012. Jay Kreps chose to name the software after the author Franz Kafka because it is "a system optimized for writing", and he liked... 1470b72c2d Incubator: for projects that are in the process of becoming full ASF projects. 1470b72c2d The stream processing paradigm simplifies parallel software and hardware by restricting the kinds of parallel computation that can be performed. Given a sequence of data (a stream), a series of operations (kernel functions) is applied to each element in the stream. Kernel functions are usually pipelined, and efficient local on-chip memory reuse is attempted in order to minimize bandwidth loss... 1470b72c2d Apache Storm is developed under the Apache... 1470b72c2d Google released an open SDK implementation of the Dataflow model in 2014 and an environment to execute Dataflows locally (non-distributed) as well as in the Google Cloud Platform service. 1470b72c2d In addition to active projects, the ASF maintains several related areas: 1470b72c2d Flink does not provide its own data-storage system, but provides data-source and sink connectors to systems such as Apache Doris, Amazon Kinesis, Apache Kafka, HDFS, Apache Cassandra, and ElasticSearch. 1470b72c2d

Apache RocketMQ It is the third generation distributed messaging middleware open sourced by Alibaba in 2012. most popular open source software award Apache ActiveMQ Apache Flink Apache Qpid Apache Samza Apache Spark Streaming Data Distribution Service Enterprise RocketMQ is a distributed messaging and streaming platform with low latency, high performance and

reliability, trillion-level capacity and flexible scalability. Next year, on February 20, the Apache Software Foundation announced Apache RocketMQ as a Top-Level Project. On November 21, 2016, Alibaba donated RocketMQ to the Apache Software Foundation 1470b72c2d

The development of RocketMQ can be divided into three stages. 1470b72c2d Storm became an Apache Top-Level Project in September 2014 and was previously in incubation since September 2013. 1470b72c2d == History == 1470b72c2d Samza provides fault tolerance, isolation and stateful processing. Unlike batch systems such as Apache Hadoop or Apache Spark, it provides continuous computation and output, which result in sub-second response times. 1470b72c2d == Overview == 1470b72c2d The third generation combines the Pull mode with some Push operations. It inherits the advantages of the first and second generation, and shows high performance in concurrency and massive amounts... 1470b72c2d

Complex event processing The goal of complex event processing is to identify meaningful events (such as opportunities or threats) in real-time situations and respond to them as quickly as possible. Complex event processing (CEP) consists of a set of concepts and techniques developed in the early 1990s for processing real-time events and extracting information from event streams as they arrive. Event processing is a method of tracking and analyzing (processing) streams of information (data) about things that happen (events), and deriving a conclusion Event processing is a method of tracking and analyzing (processing) streams of information (data) about things that happen (events), and deriving a conclusion from them 1470b72c2d

== History == 1470b72c2d

Apache Spark Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance. Originally developed at the University of California, Berkeley's AMPLab starting in 2009, in 2013, the Spark codebase was donated to the Apache Software Foundation, which has maintained it since. Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming clusters with implicit Apache Spark is an open-source unified analytics engine for large-scale data processing 1470b72c2d

Apache Beam Beam Pipelines are defined using one of the provided SDKs and executed in one of the Beam's supported runners (distributed processing back-ends) including Apache Flink, Apache Samza, Apache Spark, and Google Cloud Dataflow. Apache Beam is an open source unified programming model

to define and execute data processing pipelines, including ETL, batch and stream (continuous) processing Apache Beam is an open source unified programming model to define and execute data processing pipelines, including ETL, batch and stream (continuous) processing 1470b72c2d

INFRA (Apache Infrastructure Team): responsible for managing and providing the technical infrastructure and services used by all Apache projects. 1470b72c2d

Apache Storm The initial release was on 17 September 2011. Originally created by Nathan Marz and team at BackType, the project was open sourced after being acquired by Twitter. It uses custom created "spouts" and "bolts" to define information sources and manipulations to allow batch, distributed processing of streaming data. Originally created by Apache Storm is a distributed stream processing computation framework written predominantly in the Clojure programming language. Apache Storm is a distributed stream processing computation framework written predominantly in the Clojure programming language 1470b72c2d

Apache Flink is developed under the Apache License... 1470b72c2d Attic: for ASF projects that have been formally retired. 1470b72c2d

List of Apache Software Foundation projects Internet applications. Flume: large scale log aggregation framework Apache Fluo Committee Fluo: This list of Apache Software Foundation projects includes the software development initiatives maintained by the Apache Software Foundation (ASF). Flink: fast and reliable large-scale data processing engine 1470b72c2d

Apache Samza Retrieved 2016-09-28. InfoQ. It has been developed in conjunction with Apache Kafka. "Spark Streaming vs Flink vs Storm Apache Samza is an open-source, near-realtime, asynchronous computational framework for stream processing developed by the Apache Software Foundation in Scala and Java. LinkedIn Uses Apache Samza". "Samza: Stateful Scalable Stream Processing at LinkedIn" (PDF). Both were originally developed by LinkedIn 1470b72c2d

These events may be happening across the various layers of an organization as sales leads, orders or customer service calls. Or, they may be news items, text messages, social media posts, business processes (such as supply chain), traffic reports, weather reports, or other kinds of data. An event may also be defined as a "change of state," when a measurement exceeds a predefined threshold of time, temperature, or other value. 1470b72c2d Apache Spark has

its architectural foundation in the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way. The Dataframe API was released as an abstraction on top of the RDD, followed by the Dataset API. In Spark 1.x, the RDD was the primary application programming interface (API), but as of Spark 2.x use of the Dataset API is encouraged even though the RDD API is not deprecated. The RDD technology still underlies the Dataset API. 1470b72c2d

Apache Flink Flink's pipelined runtime system enables the execution of bulk/batch and stream processing programs. Flink executes arbitrary dataflow programs in a data-parallel and pipelined (hence task parallel) manner. The core of Apache Flink is a distributed streaming data-flow engine written in Java and Scala. The core of Apache Apache Flink is an open-source, unified stream-processing and batch-processing framework developed by the Apache Software Foundation. Apache Flink is an open-source, unified stream-processing and batch-processing framework developed by the Apache Software Foundation. Furthermore, Flink's runtime supports the execution of iterative algorithms natively 1470b72c2d

Apache Beam is one implementation of the Dataflow model paper. The Dataflow model is based on previous work on distributed processing abstractions at Google, in particular on FlumeJava and Millwheel. 1470b72c2d

Stream processing Stream processing systems use streaming algorithms to trace parallel processing for data streams. Stream processing encompasses dataflow programming, reactive programming, and distributed data processing. The software stack for these systems includes components such as programming models and query languages, for expressing computation; stream management systems for distribution and scheduling; and hardware components for acceleration, including floating-point units, graphics processing units, and field-programmable gate arrays. computer science, stream processing (also known as event stream processing, data stream processing, or distributed stream processing) is a programming In computer science, stream processing (also known as event stream processing, data stream processing, or distributed stream processing) is a programming paradigm that views streams, or sequences of events in time, as the central input and output objects of computation 1470b72c2d