**Định nghĩa**

Apache Storm là một hệ thống xử lý dữ liệu luồng theo thời gian thực.

Storm được thiết kế để xử lý khối lượng lớn dữ liệu với khả năng chịu lỗi và có thể mở rộng theo chiều ngang.

Storm quản lý môi trường phân tán và trạng thái cluster thông qua Apache Zookeeper. Nó khá đơn giản và chúng ta có thể thực hiện song song nhiều thao tác trên các dữ liệu thời gian thực.

**Apache Storm vs Spark.**

**Giống nhau:**

Được viết trên ngôn ngữ dựa trên nền tảng JAVA: Scala và Clojure.

|  |  |  |
| --- | --- | --- |
|  | Spark | Storm |
| Mô hình xử lí | Spark Streaming là bao bọc lại mô hình xử lí theo lô của Spark | Xử lí trực tiếp thông qua các lõi storm layer. |
| Các hàm | Gồm 2 loại phương thức phổ biến : Stream transformation operators và ouput operators. | Cung cấp rất nhiều các hàm để xử lí ở mức tuple , trong các khoảng khác nhau trong 1 luồng.  Hỗ trợ kết bảng trái, phải |
| Cơ chế đảm bảo an toàn cho dữ liệu | Tự định nghĩa mức độ sai lệch (fault tolerance). Mức độ đảm bảo phục thuộc vào công thức và toán tử output (output operator).  Khi gặp lỗi , thì các executors sẽ mất hết tất cả những mà nó đã nhận trong bộ nhớ => Spark cung cấp data checkpointing để khắc phục. | Cung cấp 3 tiến trình đảm bảo: at least once, at-most-once, exactly once.  Khi gặp lỗi, các storm deamons được tạo ra để fail-fast( sẽ có phương thức tự huỷ khi gặp sự cố) và phi trạnh thái. |
| Auto Scaling. | Hiện đang được phát triển dynamic scaling để streaming dữ liệu. | Hỗ trợ việc thiết lập cấu hình song song ở các mức topology.  Hỗ trợ cân bằng động, cho phép thêm hoặc giảm số lượng worker process mà không cần khởi động lại cluster. |
| Các ngôn ngữ tuỳ chọn | Java, Scala , Python, R. | Java, Clojure, Scala |
|  |  |  |

**Apache Storm vs Hadoop**

|  |  |
| --- | --- |
| Storm | Hadoop |
| Xử lý phân luồng thời gian thực | Xử lý hàng loạt |
| Phi trạng thái | Có trạng thái |
| Kiến trúc Master-Slave dựa trên sự kết hợp với Zookeeper. Thành phần master node gọi là nimbus và các slave node gọi là supervisor | Kiến trúc Master-Slave có thể/không dựa trên sự kết hợp với Zookeeper. Thành phần master node gọi là job tracker và các slave node gọi là task tracker |
| Xử lý phân luồng của Storm có thể truy cập đến 10000 thông điệp/giây trong các cluster | Hadoop Distributed File System (HDFS) sử dụng MapReduce để xữ lý khối lượng lớn dữ liệu mất vài phút hay vài giờ |
| Cấu trúc liên kết của Storm chỉ dừng khi bị tắt bởi người dùng hay gặp sự cố bất ngờ không thể khắc phục được. | Các công việc Mapreduce được thực hiện một cách tuần tự |
| Cả 2 đều phân luồng và có khả năng chịu lỗi | |
| Nếu nimbus/supervisor chết, sau khi khởi động lại sẽ tiếp tục từ chỗ dừng trước đó, do đó không có gì bị ảnh hưởng | Nếu job tracker chết thì các công việc đang thực hiện đều bị mất |

**Các thuật ngữ cần lưu ý trong Storm**

**Distributed**: phân luồng.

**Tuple**: kiến trúc dữ liệu chính của Storm, là một danh sách các yếu tố được xếp thứ tự. Mặc định thì tuple hỗ trợ tất cả kiểu dữ liệu. Thông thường, nó được mô hình hình hóa dưới dạng một tập các giá trị được phân cách bởi dấu phẩy và được truyền cho cluster của Storm.

**Stream**: Một chuỗi các Tuple không có thứ tự.

**Spout**: Nguồn của Stream. Thông thường Storm sẽ lấy dữ liệu input từ các nguồn dữ liệu thô như Twitter Streaming API, Apache Kafka queue, Kestrel queue, …

**Bolt**: là những đơn vị xử lý luận lý. Spout truyền dữ liệu cho bolt, bolt tiến hành xử lý và xuất ra output stream mới. Bolt có thể thực hiện các hoạt động như: lọc, kết hợp, tương tác với các nguồn dữ liệu và CSDL. Mỗi bolt có thể nhận nhiều dữ liệu từ các bolt hoặc là đầu vào dữ liệu cho các bolt khác.

**Topology**: được hình thành khi spout và bolt liên kết lại với nhau.

**Nimbus**: master node của Storm cluster, tất cả các node khác gọi là worker node. Chịu trách nhiệm phân luồng dữ liệu giữa các worker node, phân công task cho worker node và giám sát những task thất bại.

**Supervisor**: những node thực hiện những chỉ thị từ Nimbus. Các supervisor có nhiều worker process và quản lý các worker process này hoàn thành các task giao bởi nimbus.

**Worker process**: thực thi các task liên quan đến 1 topology cụ thể. Worker process sẽ ko chạy các task mà sẽ tạo ra các executor để thực thi chúng.

**Executor**: là một thread (luồng) tạo ra bởi worker process. Một executor thực hiện 1 hoặc nhiều task nhưng chỉ cho spout hoặc bolt cụ thể.

**Task**: task thực hiện xử lý dữ liệu thực, vì vậy nó là spout hoặc bolt.

**Các trường hợp sử dụng Apache Storm**

**Twitter**

Storm được sử dụng trong nhiều hệ thống của Twitter như: phân tích thời gian thực, cá nhân hóa, tìm kiếm, tối ưu hóa tổng thu nhập, … Storm đồng hóa với các cơ sở hạ tầng khác của Twitter bao gồm hệ thống CSDL như Cassandra, Memcached, hệ thống giám sát và cảnh báo.

Lịch cô lập của Storm cho phép nó tận dụng chung một cluster cho các ứng dụng thuộc dạng sản xuất hoặc đang phát triển.

**Yahoo!**

Yahoo! hiện đang làm việc trên platform cho phép hợp nhất giữa Big Data và xử lý độ trễ thấp. Mặc dù Hadoop là công nghệ chính dùng để xử lý hàng loạt, Storm cho phép xử lý theo luồng các sự kiện của người dùng, nhật ký.

**Infochimps**

Infochimps sử dụng Apache Storm như là nguồn cho một trong ba dịch vụ dữ liệu đám mây của mình - DDS (Data Delivery Services). Tập dữ liệu thu được bằng cách dùng Storm có khả năng chịu lỗi và mở rộng cao. Giống với Hadoop trong việc thực hiện quá trình ETL và xử lý phân tích hàng loạt, DDS còn hỗ trợ thực hiện chúng theo thời gian thực.

**Flipboard**

FiIpboard là một ứng dụng tổng hợp những bài báo, blog hay những tin tức. Nó sử dụng Storm trong các dịch vụ như tìm kiếm, phân tích theo thời gian thực, tùy chỉnh nguồn các tạp chí, … Storm được tích hợp với cơ sở hạ tầng bao gồm các hệ thống như ElasticSearch, Hadoop, HBase và HDFS để tạo ra các nền tảng dữ liệu mà có khả năng mở rộng cao.

**Ooyala**

Ooyala là một công ty tư nhân với sự hỗ trợ của nhiều doanh ngiệp, nó chủ yếu cung cấp các sản phẩm về công nghệ video trực tuyến và các dịch vụ liên quan cho các nhà mạng, thương hiệu và các công ty truyền thông lớn trên thế giới. Ooyala có công cụ phân tích có thể xử lý hơn 2 tỉ sự kiện mỗi ngày được tạo ra từ 200 triệu người xem trên Ooyala-power player trên toàn thế giới. Ooyala sử dụng Storm để thực hiện phân tích trực tuyến hành vi của người xem. Storm cho phép khai thác tập dữ liệu video trực tuyến một cách nhanh chóng để cung cấp các phân tích kinh doanh như các mẫu view theo thời gian thực, cá nhân hóa những nội dung đề nghị, …

**Taobao**

Dùng Storm đễ tạo ra các số liệu thống kê về log và trích xuất các thông tin hữu ích từ các số liệu theo thời gian thực. Các log được đọc từ các thông điệp sẽ được đưa vào hàng đợi trong các spout, sau đó được xử lý và truyền thông qua các cấu trúc liên kết để tính toán các kết quả cần thiết. Số lượng input log của Taobao nằm từ 2 triệu đến 1.5 tỷ mỗi ngày.

**Klout**

Klout là một ứng dụng sử dụng phân tích truyền thông cộng đồng để xếp hạng người dùng của nó. Xếp hạng sẽ dựa trên mức ảnh hưởng xã hội trực tuyến thông qua “Klout Score” là con số từ 1 đến 100. Storm được sử dụng để tạo ra các cấu trúc liên kết phức tạp để truyền dữ liệu từ các mạng thông qua Kafka, sau đó được xử lý và ghi xuống HDFS.

**Wego**

Là một công cụ tìm kiếm về du lịch và được sử dụng bởi nhiều người. Weigo so sánh và hiển thị lịch các chuyến bay theo thời gian thực, các khách sạn hiện có, giá cả và các địa điểm du lịch trên thế giới. Storm truyền các dữ liệu tìm kiếm thời gian thực từ các nguồn trên khắp thế giới đến các người dùng cuối. Khái niệm về cấu trúc liên kết trong Storm giải quyết nhiều vấn đề và cùng lúc có thể thực hiện các công việc như tích hợp, phân tích và làm sạch dữ liệu.

**NaviSite**

Sử dụng Storm để giám sát các bản ghi sự kiện (Event logs) của hệ thống. Mỗi bản ghi khi được tạo đều phải thông qua Storm. Storm sẽ kiểm tra các thông điệp với các biểu thức đã được cấu hình, nếu có sự trùng khớp thì thông điệp sẽ được lưu xuống CSDL.

Tham khảo

<https://www.edureka.co/blog/apache-storm-use-cases/>

<https://www.tutorialspoint.com/apache_storm/>

<https://zdatainc.com/2014/09/apache-storm-apache-spark/>