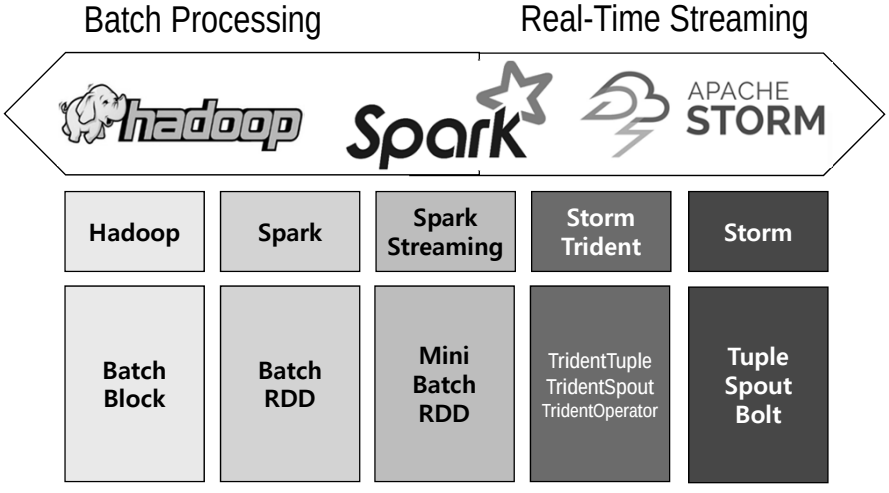
Big Data

Apache Storm

Apache Storm



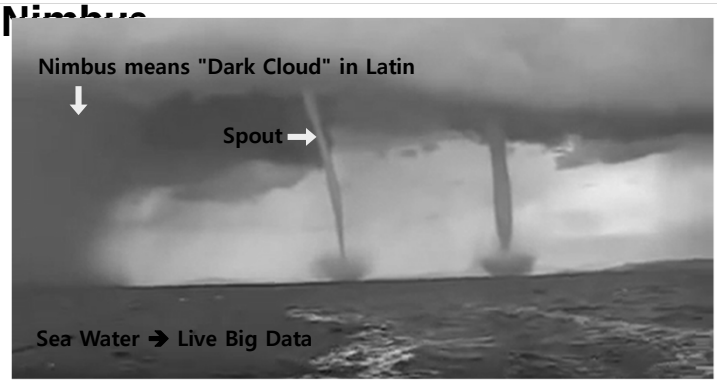
Apache Storm

❖ Apache Storm

- Introduction to Storm
  - Distributed and fault-tolerant real-time data stream processing big data framework
  - Original developers are **Backtype** & **Twitter**

Apache Storm

❖ STORM's Terminology: Nimbus, Spout



Picture Reference : Vidaluz , Waterspout project 2012, "What exactly is a Waterspout?," YouTube, released March 26, 2012, https://www.youtube.com/watch?v=yqXl8QDZ5Ew

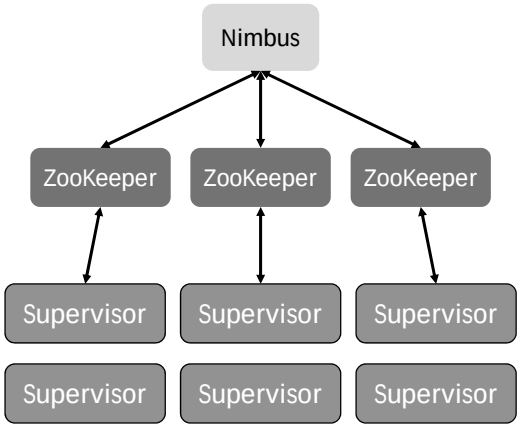
Apache Storm

❖ STORM's Terminology: Bolt

Picture Reference : Pecos Hank, "TOP 10 BEST LIGHTNING STRIKES," YouTube, released Oct. 28, 2014, https://www.youtube.com/watch?v=TaPgSWdcYtY

Apache Storm

❖ Storm Components



Apache Storm

❖ Storm Components

- Nimbus
  - Distributes program codes throughout the cluster
  - Schedules Topology processes
  - Assigns tasks to Worker nodes
  - Monitors Worker nodes
  - Monitors for node failures

Apache Storm

❖ Storm Components

- ZooKeeper
  - Apache ZooKeeper provides highly reliable distributed coordination for clusters
  - HA (High Availability) fault-tolerance provided through replicated master/agents and non-disruptive upgrades

Apache Storm

❖ Storm Components

- Supervisor
  - Waits for task assignments from the Master node
  - Each worker process executes a subset of a Topology process
  - Topology process involves many Worker processes distributed across multiple Worker nodes of the cluster

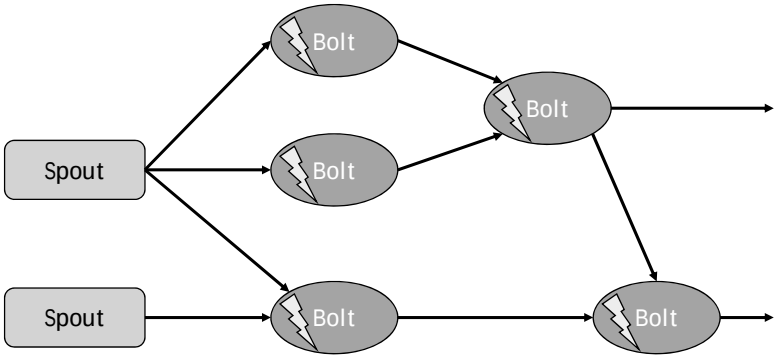
Apache Storm

❖ Storm Components

- Topology
  - Overall computation process represented in a graph form based on Spouts and Bolts connected together to form a DAG (Directed Acyclic Graph) of operations

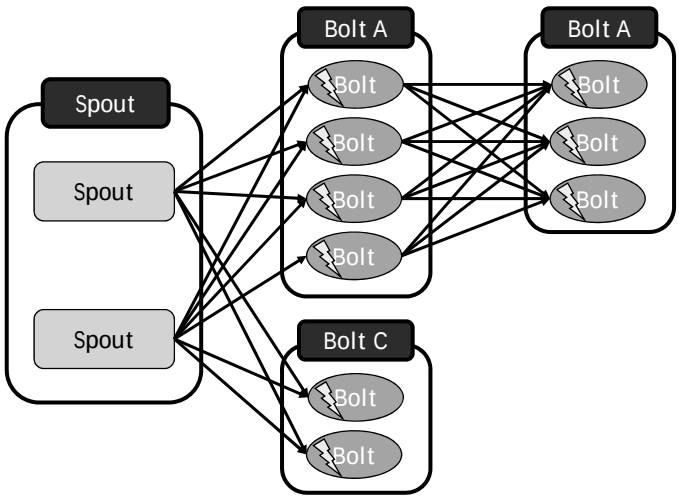
Apache Storm

❖ **Storm Topology DAG with Single Components**



Apache Storm

❖ **Storm Topology DAG with Multiple Components**



Apache Storm

❖ Storm Components

- Spout
  - Source of the Tuple stream
  - Input to the Topology
    - Example: Twitter API

Apache Storm

❖ Storm Components

- Tuple
  - Sequence of (finite number) data values
  - N-tuple is a sequence of N elements
    - 5-tuple example (5, 3, 9, 2, 7)

- Tuple Stream
  - Continuous stream of tuples

Apache Storm

❖ Storm Components

- Bolt
  - Processing unit of the tuple stream
  - Receives input tuple streams
  - Bolts will process and emit a tuple stream
  - Bolts at the end of a topology may or may not emit tuples
  - Process types include functions, filter, aggregate, join, talk to database, etc.

Apache Storm

❖ Apache Storm

- Introduction to Storm

Storm programs are written in Java and Clojure

- Initially released on September 17, 2011
- 2011 September 19 the first major released version was 0.5.0
  - Originally developed by Nathan Marz and team at BackType

Apache Storm

❖ Apache Storm

- Introduction to Storm
  - Twitter acquired Storm and made it an open source Apache Storm project
    - 2014 February 10, Apache Storm's first major release was version 0.9.1
  - 2014 September, Apache Storm became an Apache Top-Level Project

Apache Storm

❖ Apache Storm

- Features of Storm
  - Can process millions of tuples per second
  - Highly fault tolerant with fast recoveries
  - Efficient and fast Cluster control
  - Simple to program and supports multiple programming languages
    - Supports various cross-platform OSs (Operating Systems)

Apache Storm

❖ Apache Storm

- Features of Storm
  - Topology based on Spouts and Bolts form a DAG (Directed Acyclic Graph)
    - Topology's DAG Edges are streams that direct data from one node to another
    - Topology forms data stream transformation pipeline(s)

Big Data

Reference

References

- Apache Storm, http://storm.apache.org
- Nathan Marz, "ETE 2012 - Nathan Marz on Storm," https://www.youtube.com/watch?v=bdps8tE0gYo&t=542s, Feb. 15, 2012.
- Wikipedia, https://en.wikipedia.org
- edureka!, "Understanding Spout In Apache Storm | Edureka," https://www.youtube.com/watch?v=5kiZs1a8UPM, Oct. 10, 2014
- https://www.webopedia.com/TERM/E/ETL.html
- Sean T. Allen, Matthew Jankowski, "Storm Applied: Strategies for real-time event processing", Apr. 12, 2015.
- Jonathan Leibiusky, Gabriel Eisbruch, Dario Simonassi, "Getting Started with Storm: Continuous Streaming Computation with Twitter's Cluster Technology",Sep. 17, 2012.
- Flavio Junqueira, Benjamin Reed, "ZooKeeper: Distributed Process Coordinationt", Dec. 5, 2013.

References

- **Image source**
  - https://upload.wikimedia.org/wikipedia/commons/c/cc/Stock_Tips.jpg
  - https://upload.wikimedia.org/wikipedia/commons/8/80/New_York_Stock_Exchange_trading_floor_on_Wall_Street%2C_New_York%2C_New_York_LCCN2011634218.tif
  - https://upload.wikimedia.org/wikipedia/commons/b/bd/USB-thumb-drive-16-GB.jpg