Big Data

Storm Spouts & Streams

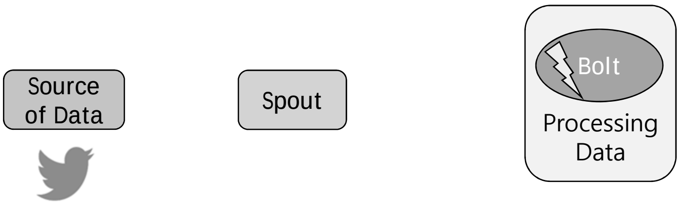
Storm Spouts & Streams

❖ Storm Spouts Pull

- Spout Data Stream Pull Process
  - Spout contacts that Data Source and requests for the data stream to be sent to itself
  - Pull mechanism of the Spout enables Spout to be the active controller of the source stream of data to the topology
    - Therefore, no Queueing node is needed

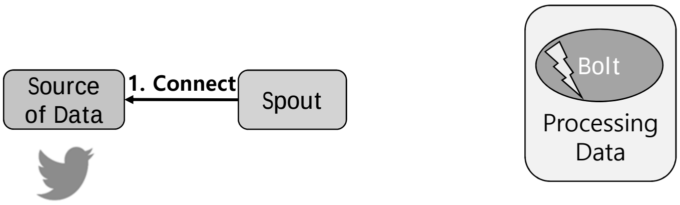
Storm Spouts & Streams

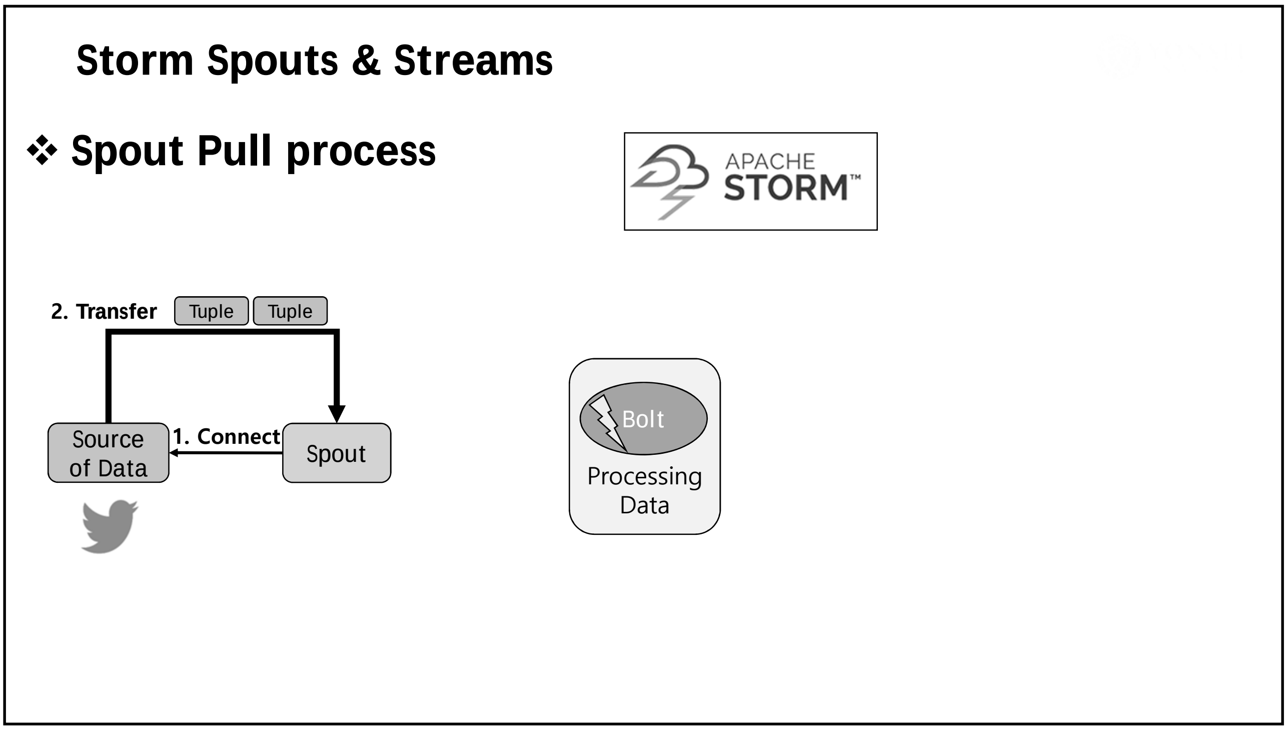
❖ **Spout Pull process**



Storm Spouts & Streams

❖ **Spout Pull process**



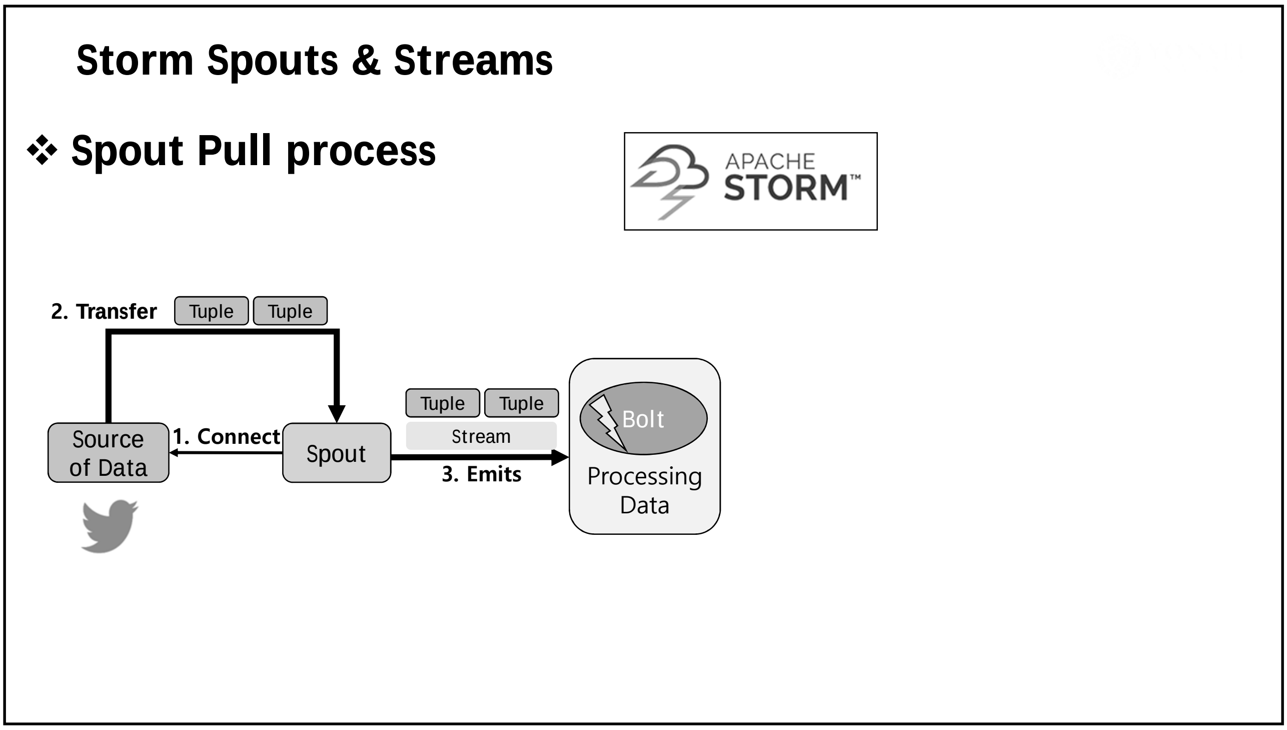


Storm Spouts & Streams

❖ **Spout Pull process**

APACHE STORM™

2. Transfer
Tuple
Tuple
Source of Data
1. Connect
Spout
Bolt
Processing Data



Storm Spouts & Streams

❖ **Spout Pull process**

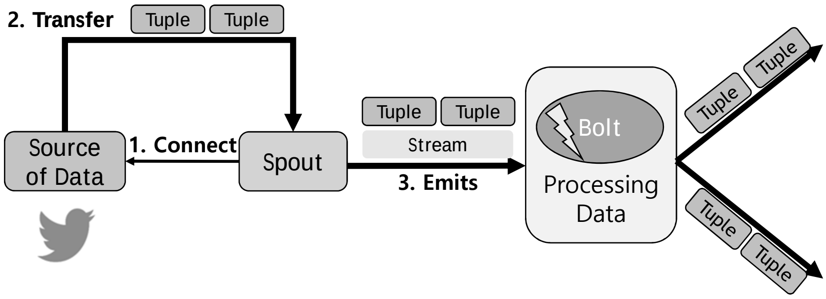
APACHE STORM™

2. Transfer
Tuple
Tuple
Source of Data
1. Connect
Spout
Tuple
Tuple
Stream
3. Emits
Bolt
Processing Data

Storm Spouts & Streams

❖ **Spout Pull process**

APACHE STORM™



Storm Spouts & Streams

❖ **Spout Pull process**

- Spout Data Stream Pull Process
  - Spout Pull operation enables the Spout to control the start, stream termination, flow speed, manage the Topology process, and control the changes that need to be made to the topology of the steam data processing

Storm Spouts & Streams

❖ **Spout Pull process**

- Spout Data Stream Pull Process
  - **`nextTuple( )`** is the message that a Spout uses to receive tuples from the data source, which is based on a form of “one time” or an “on user request” method

Storm Spouts & Streams

❖ **Reliable & Unreliable Spouts & Bolts**

- Reliable Spouts & Bolts
  - Reliable Spouts will use acknowledgment (using “**`ack(object tupId)`**”) and failure (using “**`fail(object tupId)`**”) to monitor the tuples processing status of the Bolts
  - When using Reliable Spouts, Reliable Bolts should also be used

Storm Spouts & Streams

❖ Reliable & Unreliable Spouts & Bolts

- Unreliable Spouts & Bolts
  - Unreliable Spouts (and Unreliable Bolts) will not conduct Bolt acknowledgment and failure monitoring
  - The ack and fail monitoring and messaging of Reliable Spouts & Bolts requires more overhead in messaging & processing

Storm Spouts & Streams

❖ Reliable & Unreliable Spouts & Bolts

- Reliable & Unreliable Application Examples
  - Reliable Spouts & Bolts should be used for Stock Market analysis since each transaction must be processed, monitored, and protected

Author:FreeStockTips
https://commons.wikimedia.org/wiki/File%3AStock_Tips.j

Author: Carol M. Highsmitht
https://commons.wikimedia.org/wiki/File%3ANew_York_Stock_Exchange_trading_floor_on_Wall_Street%2C_New_York%2C_New_York_LCCN2011634218.tif.jpg

Storm Spouts & Streams

❖ **Reliable & Unreliable Spouts & Bolts**

- Reliable & Unreliable Application Examples
  - Unreliable Spouts & Bolts can be used for monitoring and analyzing SNS (Social Network Services) streams, since missing a message would be less critical and Unreliable Spouts & Bolts requires a much smaller load in processing on the topology and cluster

Storm Spouts & Streams

❖ **Storm Data Stream Queue Management**

- Can process over 1 million messages per second
- Storm can control the maximum number of tuples that can be pending and not processed yet
  - Using **`topologyMaxSpoutPending( )`**

Storm Spouts & Streams

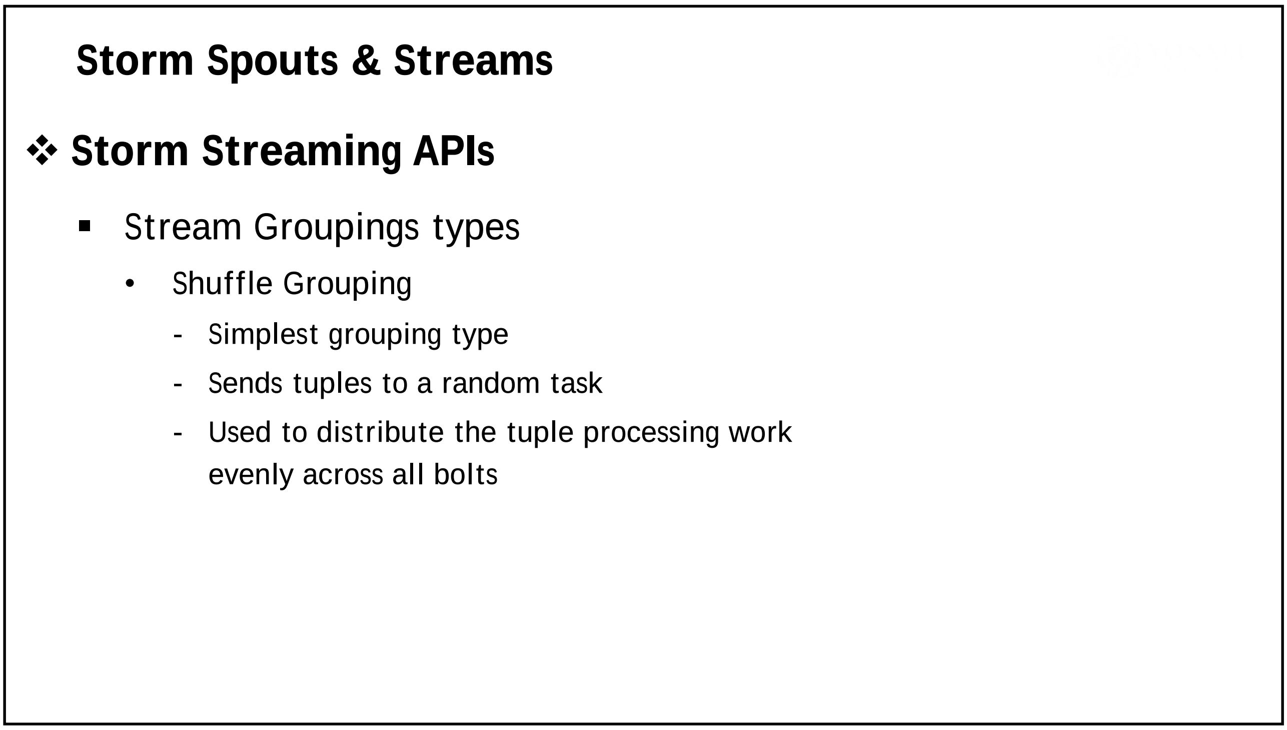
❖ Storm Streaming APIs

- Tuple
  - Set of a finite number of data values
  - N-tuple is a sequence of N elements
- Trident
  - Processes the incoming stream of Micro-Batches one-at-a-time
    - Micro-Batch is a collection of Tuples
  - More efficient → Higher throughput performance

Storm Spouts & Streams

❖ Storm Streaming APIs

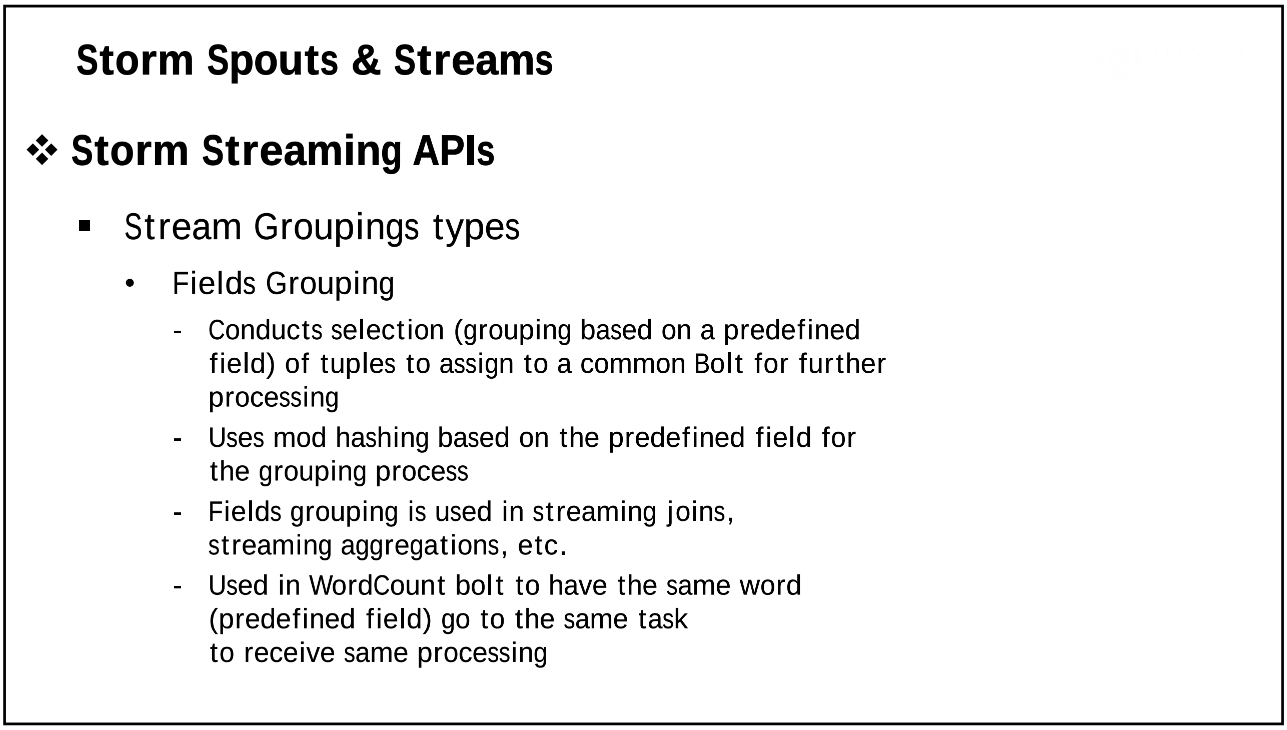
- Stream Grouping
  - For a given Topology, Stream Grouping informs how to send tuples in parallel across the interface of the Spout-Bolt and Bolt-Bolt
  - Spout-Bolt and Bolt-Bolt connections are used to process multiple tuples in a parallel way over a distributed topology in the Cluster



Storm Spouts & Streams

❖ Storm Streaming APIs

- Stream Groupings types
  - Shuffle Grouping
    - Simplest grouping type
    - Sends tuples to a random task
    - Used to distribute the tuple processing work evenly across all bolts



Storm Spouts & Streams

❖ Storm Streaming APIs

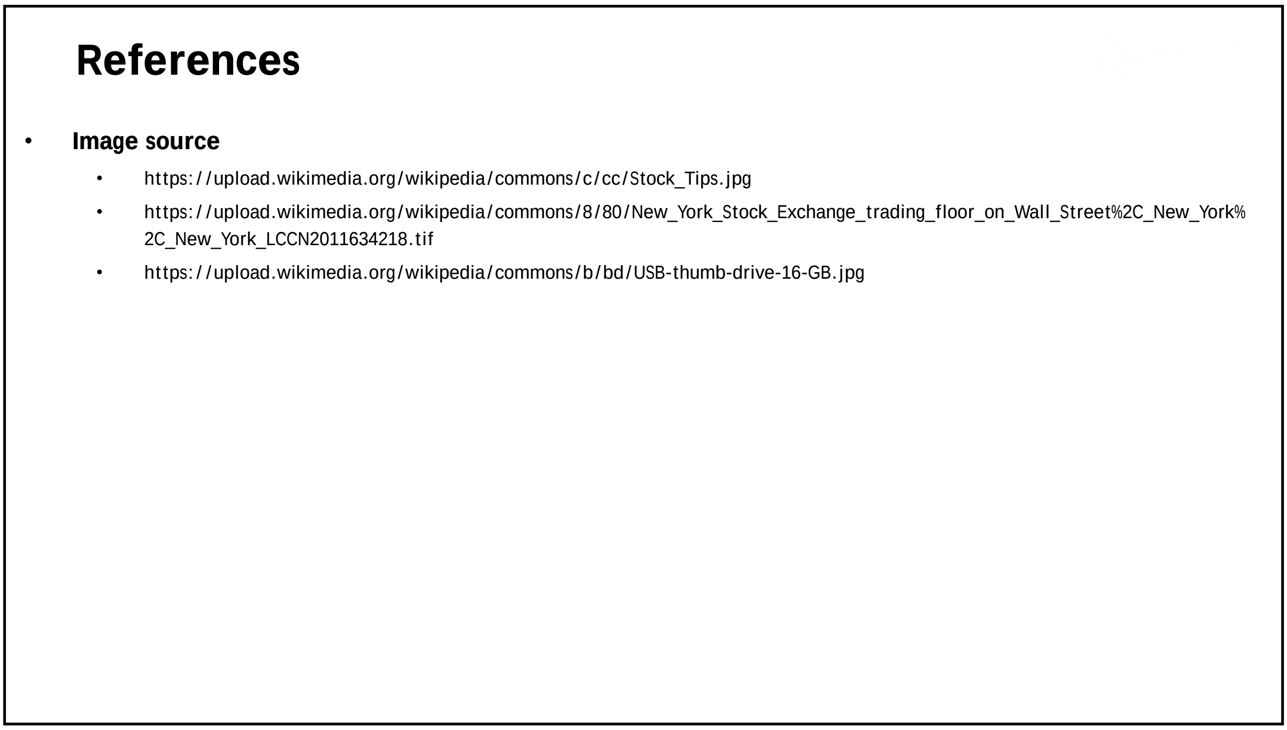
- Stream Groupings types
  - Fields Grouping
    - Conducts selection (grouping based on a predefined field) of tuples to assign to a common Bolt for further processing
    - Uses mod hashing based on the predefined field for the grouping process
    - Fields grouping is used in streaming joins, streaming aggregations, etc.
    - Used in WordCount bolt to have the same word (predefined field) go to the same task to receive same processing

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