Big Data

Storm Applications

Storm Applications

❖ Apache Storm

- Storm Use Cases
  - Stream Processing
    - Direct stream processing without any intermediate queues
  - DRPC (Distributed Remote Procedure Call)
    - Computation load is efficiently distributed over parallel processing CPUs for real-time reliable results

Storm Applications

❖ Apache Storm

- Storm Use Cases
  - Continuous Computation
    - Continuous real-time updates to (complex) computations as the input data streams are continuously analyzed

Storm Applications

❖ Storm Applications

- Real-time analytics
- Online ML (Machine Learning)
- Continuous computation
- DRPC (Distributed Remote Procedure Call)
- ETL (Extract, Transform, Load)
- etc.

Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- DRPC (Distributed Remote Procedure Call)
  - RPC (Remote Procedure Call) is a program that can execute its process on a remote computer with a different IP address
  - Distributed RPC is a RPC executed in a distributed parallel form over a cluster of nodes

Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- ETL (Extract, Transform, Load)
  - Three essential database functions Extract + Transform + Load combined into one function
  - ETL function extracts data from a database and transforms it and loads it into a data warehouse

Storm Applications

❖ Storm Applications

- Financial Services
  - Security fraud detection
  - Policy violation detection
  - Stock Monitoring & Alert Notification
  - Pricing control & adaptation

Storm Applications

❖ Storm Applications

- Retail Services
  - Logistic scheduling
  - Sales monitoring
  - Discount rate control
  - Coupon distribution
  - Shrinkage & Stock outs
  - Pricing & Offers

Storm Applications

❖ Storm Applications

- Web Services
  - Web server & system support
  - Security breach monitoring
  - Application failure recovery
  - User problem solving
  - Personalized content protection

Storm Applications

❖ Storm Applications

- ICT & Telecom Services
  - SNS (Social Network Services) services
  - Mobile Apps & Cloud services
  - Customer services
  - Security breach monitoring
  - Network (gateways/routers, switches) outage
  - Bandwidth allocation
  - Mobility support
  - RAN (Radio Access Network) connectivity

Storm Applications

❖ Understanding Storm: Storm vs. Hadoop

- Cluster & Operation
  - Hadoop Cluster
    - Hadoop runs MapReduce jobs on HDFS
  - Storm Cluster
    - Storm runs Topology processes on the DAG

Storm Applications

❖ Understanding Storm: Storm vs. Hadoop

- Process Ending
  - Hadoop
    - MapReduce jobs end when the assigned dataset is completely processed
  - Storm
    - Topology processes streaming messages continuously until the stream is intentionally terminated by the user

Storm Applications

❖ Understanding Storm: Storm vs. Hadoop

- Node Types
  - Hadoop
    - NameNode (Master) and DataNode (Worker)
      - NameNode runs the JobTracker daemon
      - DataNode runs the TaskTracker daemon
  - Storm
    - Master node and Worker node
      - Master node runs the Nimbus daemon
      - Worker node runs the Supervisor daemon

Storm Applications

❖ Example of Monitoring Accessed Web URLs in Twitter Retweets

- Hadoop URL Retweet Counting Example

Twitter Firehose provides streams of Twitter messages to Queueing (intermediate Message Broker) nodes that serve as intermediate queues

- Workers collect and store data in HDFS
- MapReduce reads from HDFS to find web URLs

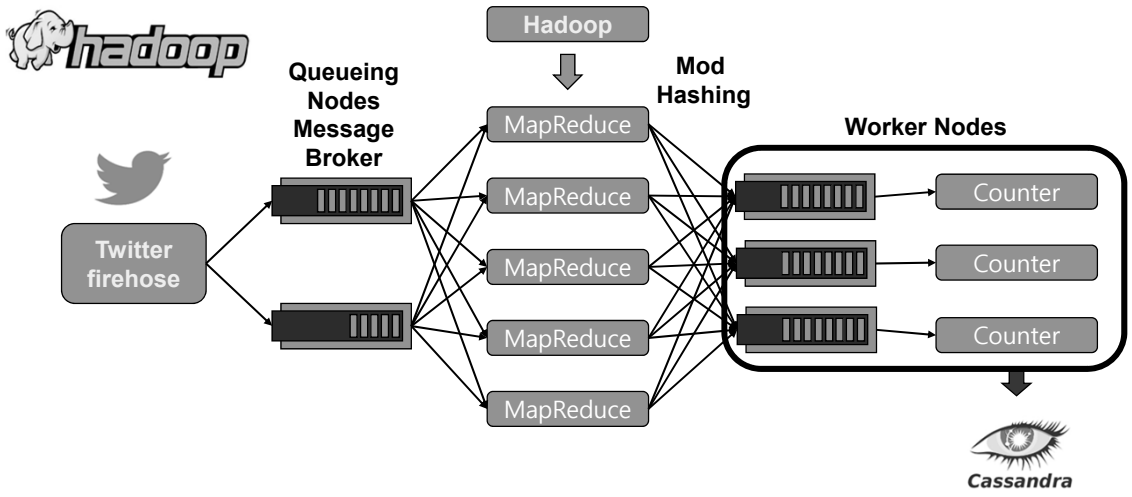
Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- Hadoop URL Retweet Counting Example
  - Mod Hashing operation is used to send same URL MapReduce processed <Key, Value> results to the same Worker node (for counting & queueing)
  - Updated URL counts and statistics are collected and sent to Cassandra

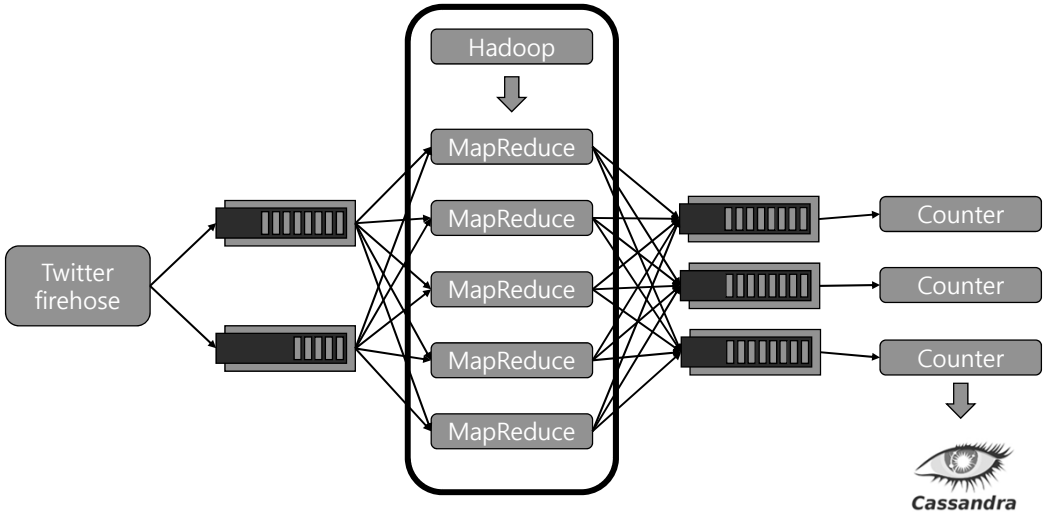
Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**



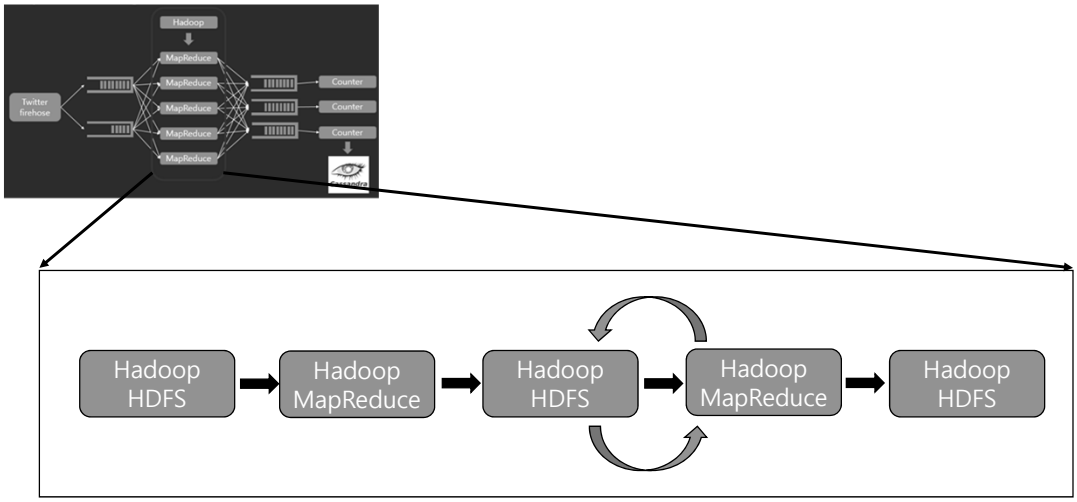
Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**



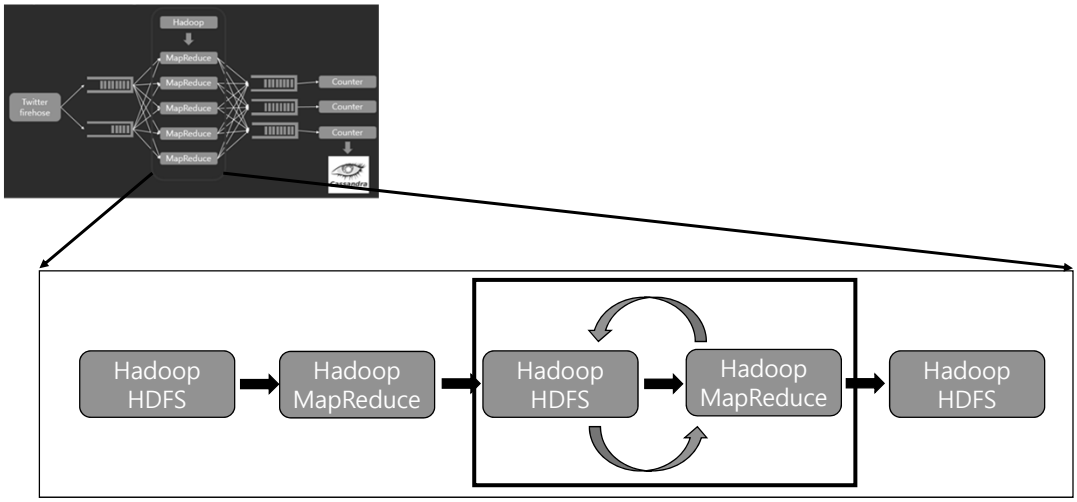
Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**



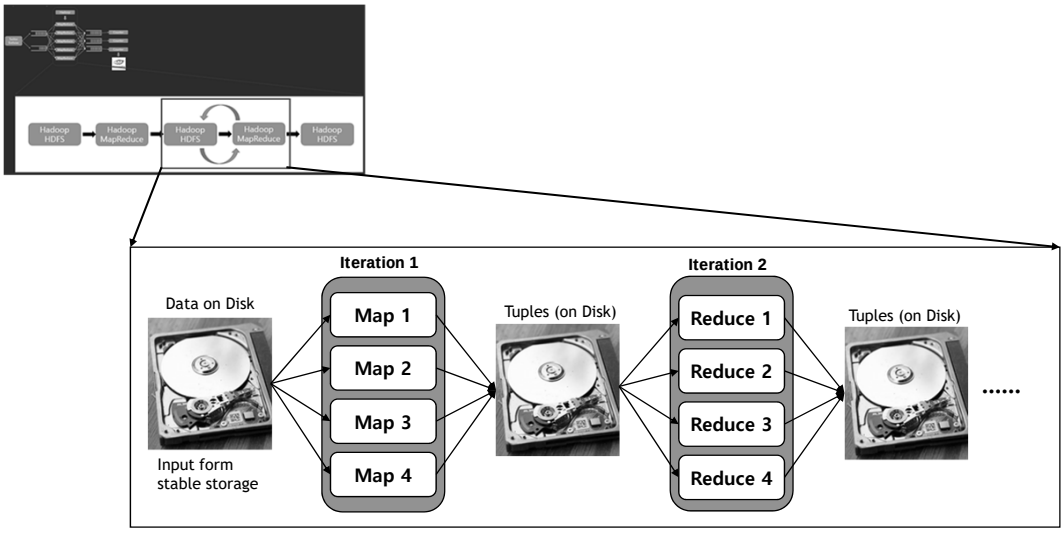
Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**



Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**



Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- Issues when using Hadoop
  - Intermediate message broker functioning Queueing nodes are needed
  - Write and Read with HDFS is slow so multiple HDFS Writes and Reads will slow down the analysis significantly
  - Needing to compute the Mod Hashing operation and send the same URL message to the same counting Worker node is inconvenient

Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- Issues when using Hadoop
  - Adding an additional URL search Worker node (DataNode) in parallel is complex and requires changes in all interacting Nodes and Worker programs (JVMs) in the cluster
    - Significant time is required to conduct program (JVM) replacements throughout the cluster
    - Hadoop program for this Job is complicated and difficult to program, and also difficult to debug and upgrade

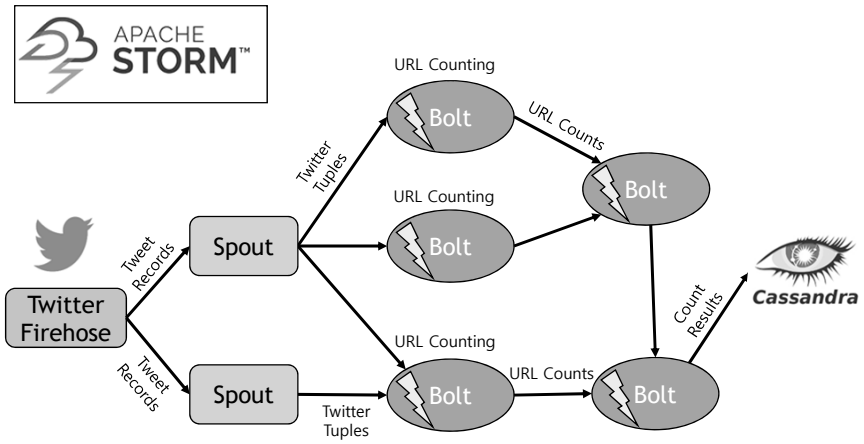
Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- Storm URL Retweet Counting Example
  - Spouts receive Tweeter messages in Tuple stream form
  - Multiple Bolts in parallel are assigned the task of finding URLs and make local counts
  - Following stages of Bolts aggregate the URL counts to find summed numbers
  - Final stage Bolts save the Twitter URL statistics to Cassandra

Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**



Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- Advantages when using Storm
  - Spout controls the message input stream from Twitter
    - Pull scheme
  - Guaranteed data processing of all Tweet messages
    - Reliable vs. Unreliable Spouts/Bolts

Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- Advantages when using Storm
  - Built in self-healing fault tolerance capability
  - No intermediate Queueing nodes needed
  - No Mod Hashing operations needed
  - Great horizontal scalability of Spouts and multistage Bolts

Storm Applications

❖ **Example of Monitoring Accessed Web URLs in Twitter Retweets**

- Advantages when using Storm
  - Fast RAM based in-memory Write and Read speeds in Bolts
  - Simple and easy to program, debug, and upgrade
  - Instantaneous processing topology updates
    - Rebalance feature

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
  - By FreeStockTips (Own work) [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons
  - Carol M. Highsmith [Public domain], via Wikimedia Commons
  - Teravolt at English Wikipedia [CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons