Apache Storm

|  |  |
| --- | --- |
| Soundharya Khanapur | Pratima Bhat |
| *Computer Science Engineering* | *Computer Science Engineering* |
| *KLS’s Gogte Institute Of Technology*  *Belagavi, India*  [*soundarya.m.khanapur@gmail.com*](mailto:soundarya.m.khanapur@gmail.com) | *KLS’s Gogte Institute Of Technology*  *Belagavi, India*  prathimabhat98@gmail.com |

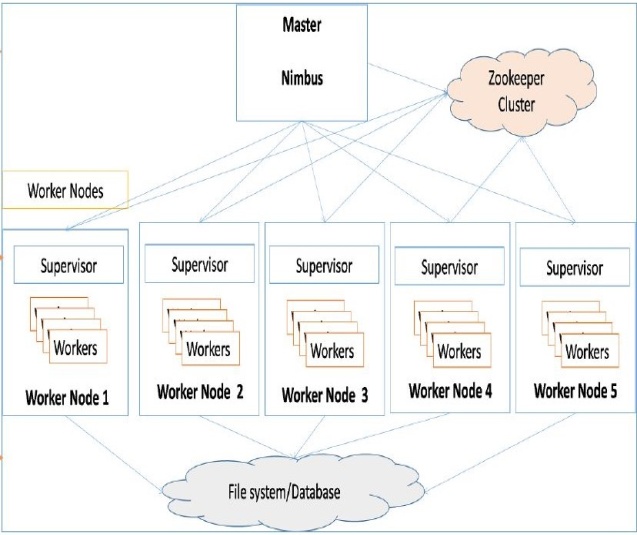
***Abstract*** *—* **The advancement and development in the field of Technology and Communication calls in for a need of Real time data processing that is fast and fault tolerant. Apache Storm provides an epoch platform to develop applications that can process multitude of data in real time. Being distributed, Storm is predominantly fast and maintains high accuracy with its topological analysis and task completion checks.**

Keywords—Nimbus, Zookeeper, spouts, bolts, stream processing

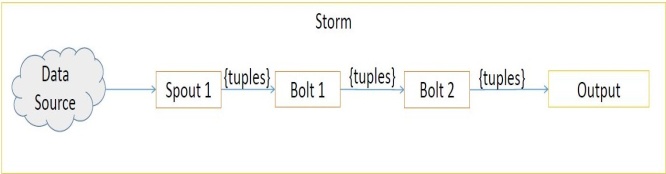
# Introduction

Apache Storm is a distributed framework for real time processing of Big Data. Parallelism is the principle of storm. Here the same code is executed on multiple nodes with different input data. Apache Storm utilizes ZooKeeper for coordination. This enables Storm to start right from where it left even after the restart. Moreover, it is free, reliable, scalable and fault tolerant.

# Design And Working



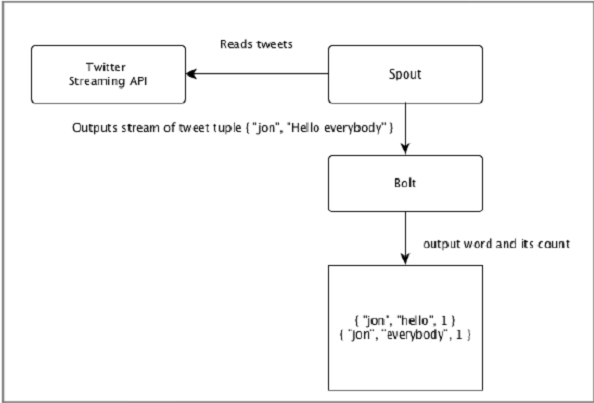
Apache Storm has master-slave architecture, where Nimbus is the master daemon and Supervisor is the worker daemon of Storm clusters. Nimbus is responsible for distributing the data among various worker nodes. There is no need to learn a new language to work with Apache Storm as Nimbus is an Apache Thrift service which allows us to submit code in any programming language. Storm has Zookeeper for coordination between the nodes. Zookeeper monitors the message processing tasks as all the worker nodes update their tasks status in it. Worker nodes have slave services called Supervisors, which supervises the workers working in the Worker node. Workers get the input from file system or database and store the output also to a file system or database. Nimbus and Supervisor are stateless, which helps Storm process real-time data rapidly. Storm is not entirely stateless as it stores its state in Zookeeper, which helps a failed Nimbus to restart from where it stopped.



Storm data model consists of tuples and streams. Tuple is an ordered list of named values having dynamic data type. Data type can be a string, integer, float, double, boolean or even user-defined data type. Stream is an unbounded sequence of tuples. A Storm component consists of spouts and bolts to process the input stream. Spouts ingest the data and process it into streams of tuples and send them to bolts. Bolts process this tuples and produce some output tuples. Input to bolts is either from a spout or from another bolt. Bolt works as a filter, join, and aggregation. Bolts can process any number of input streams. Input data is deserialized and the output data is serialized. Serialization is the process to convert data structures or objects into a platform independent stream of bytes. Serialization makes the data readable by other programs. To read the data, programs have to reverse the serialization process, this is called deserialization. Spouts and bolts run in parallel and can distribute across machines in the Storm cluster.

# Use Cases

Twitter – Twitter is a social networking site where users post and interact with messages known as tweets. The Storm spout gets the tweets submitted by the users with the help of Twitter Streaming API. The API can be accessed in any programming language. The spout needs OAuth authentication details and at least a keyword. The spout will emit real-time tweets based on keywords. The spout’s output is fed to HashtagReaderBolt, which processes the tweets and emits all the available hashtags. The emitted hashtag is forwarded to HashtagCounterBolt to process all hashtags and save each and every hashtag and its count in memory.



Yahoo! Finance – Yahoo Finance is a part of Yahoo network. It is a website to provide leading business news and financial data. The Storm spout gets the details of the company and emits the prices to bolts. The bolts process the given company’s prices when the price falls below cutoff limit value.

# Advantages

* Storm is straightforward, easy and works with almost all programming language.
* Storm can ingest high volume and high velocity data and is highly parallelizable and fault tolerant.
* It is reliable, versatile, flexible and can transfer billions tuples within seconds.
* Storm automatically does the data refresh, so no need to wait for loading the page when it is manipulated or changed.
* Storm is used to create log statistics and helps to fetch the information from the statistics in real time.
* Based on some real time data, storm helps to take smarter decision in business operations.
* Storm ensures the processing of data, even though some of the nodes die in the cluster or when messages are exhausted.
* Storm can be used to create complex topologies that may use in social media analytics.
* Storm helps metasearch engine to search all real time data and solves some concurrency related issues.
* Performance of storm is very high; under increasing loading condition, it keeps the performance up by adding resources in continuous manner.

# Conclusion

Data has always been huge but, real time processing is what we are looking for to provide applications which give us analysis in real time and help us with communication problems. This is not possible with Hadoop and Spark, because these are also clustered services but they are not real time. Apache is real time and it can help us with all such issues. We have Twitter, WhatsApp and IOT devices where we need to communicate. Consider a coffee machine which is connected to wifi and user can just tap it on the mobile application describing the type of coffee. By the time the user walks towards the machine, the coffee has to be ready. The user cannot wait in such cases for data to be moved from different clusters, process it and come back. This is what happens in Hadoop but, Apache Storm is real time and the job has to be completed within given time.

# Acknowledgement

We would like to express our special thanks of gratitude to Dr.A.S.Deshpande, principal of KLS’s Gogte Institute Of Technology, Belagavi and Dr.Veena Desai, Head of Computer Science and Engineering and to entire Avalanche team for giving this wonderful opportunity to do paper presentation on the topic “Apache Storm”. It helped us to explore to a new technology and also helped us in doing a lot of research in the same field. Secondly, we would like to thank teaching and non teaching staff of CSE department who helped us with college resources.

# References

1. [www.whatis.techtarget.com/definition/Apache-Storm](http://www.whatis.techtarget.com/definition/Apache-Storm).

www.simplilearn.com/introduction-to-storm-tutorial-video

1. [www.simplilearn.com/introduction-to-storm-tutorial-video](http://www.simplilearn.com/introduction-to-storm-tutorial-video).
2. [www.tutorialspoint.com/apache\_storm/apache\_storm\_in\_yahoo\_finance.htm](http://www.tutorialspoint.com/apache_storm/apache_storm_in_yahoo_finance.htm).
3. [www.scribd.com/document/238317376/Benefits-of-Apache-Storm](http://www.scribd.com/document/238317376/Benefits-of-Apache-Storm).