60_Subrato-D16AD

BDA-Experiment 8

- **Aim:** To study Twitter data analysis using Flume

- **Theary:**
  - Apache Flume is a distributed, reliable and scalable tool for efficiently collecting, aggregating and transporting large volumes of log data, events or streaming data from various sources to centralised data storage, HDFS or HBase.
  - Flume is designed to collect data from various sources, such as web server logs, social media feeds, etc.
  - It uses a distributed architecture, and data collection is performed by agents.
  - Once the data is collected, it is placed in a channel where the data is temporarily stored before it is forwarded to a sink.
  - Sinks are responsible for delivering data from the channel from final destination, which is often HDFS or HBase.
  - Data collected by flume can be processed and analysed using tools like MapReduce or Apache Spark.

- **Conclusion:** Apache Flume was used to gather and study Twitter Data.