Design and Development of BigData Retail Analytics using Hadoop Ecosystem

Student's Name	**J. Uma Shankar**	**CSN (FT-2012)**
Academic Supervisor(s)	V. S. Yerragudi and N. D. Gangadhar	
Industrial Supervisor(s)		

J. Uma Shankar

umashankar.jayaram@gmail.com
Ph. No: 0 88615 38222

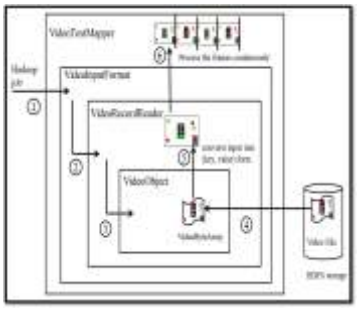
Keywords: Hadoop, Retail Video Analytics

Abstract:

Organisations in domains such as the retail sector generate huge amount of data consisting of billing, customer information, sales history, inventory and video data. The amount of video data generated here is huge in volume. Retail industries can improve their business value using this video data. Storing huge amounts of data is easy since large scale storage devices are available at low costs. However, retrieval and processing of the large scale data requires an appropriate solution platform. Hadoop addresses this problem by providing fault tolerant storage system with parallel processing framework enabling high throughput. However, since video data is not a standard input format for Hadoop Distributed File System (HDFS) it is necessary to evolve the Hadoop environment to accept video data and process it.

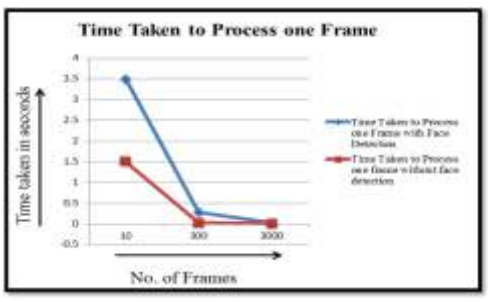
This project describes the design and development of Hadoop environment which accepts video data and processes it as per the requirement. The design is based on the use cases identified by Tyco Fire and Security. A Proof of Concept (POC) is developed to make Hadoop accept video data and do the processing as per the use cases. The development effort includes implementation of a new Hadoop InputFormat class for video input along with VideoObject class and related classes that make Hadoop's MapReduce framework accept video data. Two image processing algorithms in OpenCV libraries were adopted to suit the use case requirement: Identifying the red/green colours in a traffic signal video and identifying number of people standing in a billing queue.

The system was tested and validated successfully for both the use cases. The performance of video processing improved with increasing number of data nodes in the Hadoop cluster along with video data partitioning. However, the processing time of user defined video input format is more when compared with Hadoop's standard file input format. In the future, the performance can be improved further by creating a HDFS sequential file from the video input.

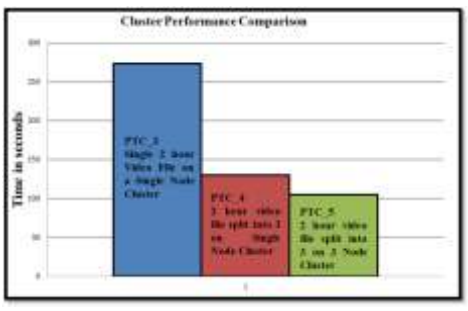


Video data flow in Hadoop environment

Output of people count on Hadoop cluster



Average time taken to process one frame



Cluster comparison