Abstract

Big data is the word used to describe structured and unstructured data. The term big data is originated from the web search companies who had to query loosely structured very large
Various Data-Mining Techniques for Big Data

distributed data. Big Data is a new term used to identify the datasets that due to their large size and complexity. Big data mining is the capabilities of extracting useful information from these large datasets or streams data that due to its volume, variability and velocity. This data is going to be more diverse larger and faster. Mapreduce provides to the application programmer the abstraction of the map and reduce. Mapreduce is a framework used to write applications that process large amounts of data in parallel on clusters. Mapreduce framework for processing large amount of data. The main aim of this system is to improve performance through parallelization of various operations such as loading the data. This paper explores the efficient implementation of bisecting clustering algorithm with mapreduce in the context of grouping along with a new fully distributed architecture to implement the mapreduce programming model. The architecture also uses queries to shuffle results from map to reduce the cluster results also indicate that queues to overlap the map and shuffling stage seems to be a promising approach to improve mapreduce performance.

References

Various Data-Mining Techniques for Big Data

Data Problem Using Hadoop and Map Reduce:
- Dong, X. L.; Srivastava, D. Data Engineering (ICDE), "Big data integration"; IEEE International Conference on, 29(2013)1245-1248.

Index Terms

Computer Science
Information Sciences

Keywords
Big Data  Clustering  Classification  Clustering Algorithms  Data Mining  Map-reduce.