**RANDOM SAMPLING IN APACHE HIVE**

Data is growing at a rapid pace. Traditional databases has been storing huge amounts of data for the past few decades. Analyzing large amounts data helps to find patterns, trends and data insight that helps organizational decisions. In traditional databases, handling and analyzing massive amounts of data needs lots of resources and time. Apache Hadoop is a collection of open source big data software’s that can eﬃciently handle storing huge amounts of data into small blocks. Data is analyzed based on the concept of parallel computation. Apache Hive is a data warehousing software that works on top of Hadoop ﬁle system. For aggregate queries and for analyzing trends in data, small population gives good approximation about overall population. The process of selecting limited number of elements from whole population is called Sampling. We have investigated three techniques to perform random sampling on Hive: simple random sampling using sorting, Bernoulli’s sampling, and our algorithm random sampling using bucketing. Simple random sampling using sorting , and Bernoulli’s sampling goes through the whole data to perform sampling in Hive. This slows down the performance when the data is huge. To avoid whole table scan while performing simple random sampling, our algorithm uses bucketing in hive architecture to manage the data stored on Hadoop Distributed File System. Bucketing divides the whole data into specified number of small blocks. Data is divided into buckets based on a specified column in a table. Bucketing allows to select any bucket of required size without scanning the whole table. Limiting data scan and sorting fewer elements decreases the time taken to perform simple random sampling using bucketing. Our experiments shows random sampling using bucketing performs much faster than random sampling using sorting and Bernoulli sampling when the data sizes or sample sizes are large.