**Project Title:** Benchmark analysis of Cassandra NoSQL database.

**Project Description:** This work studies the popular Cassandra database in a distributed environment through an exhaustive performance analysis. Two types of scalability analysis are conducted namely scale-out and scale-up leveraging the datasets of Yahoo's Cloud Serving Benchmark (YCSB).
The six core workloads provided by YCSB are used in this performance evaluation. Each workload included one or more of the following atomic operations, read, insert, update, read-modify-write, and scan.

**Framework used:** Yahoo under YCSB project developed a framework and a common set of workloads to evaluate the performance of data stores. The project has two key components. First, the client that is an extensible workload generator. Second, the core workloads that are a set of workload scenarios to be executed by the generator.

**Workloads Data**: Each of the six workloads provided by YCSB is composed of one or more following atomic query operations: read, insert, update, read-modify-write, and scan (read up to 100 consecutive rows in the database from a random start point).
Each workload goes through two phases, the "LOAD" phase in which random data is generated and written in the data-store, and the "RUN" phase in which atomic operations are performed on the written data. Workload composition parameters are as follows.

| Workload | Composition |
|---|---|
| Workload A | 50% Read + 50% Update |
| Workload B | 95% Read + 5% Update |
| Workload C | 100% Read |
| Workload D | 95% Read + 5% Insert |
| Workload E | 95% Scan + 5% Insert |
| Workload F | 50% Read + 50% Read-Modify-Write |

**Metrics for Measurement:** The evaluation results are based on latency and throughput based metrics. Latency metric represents the time interval between the function call and the response to the call. Throughput metric represents the rate of successful execution of a functional call. Two different approaches are followed for "LOAD" and "RUN" phases. In the "LOAD" phase, multiple YCSB instances run in parallel. Thus, throughput is measured as the addition of throughput measures from all the instances and average of latency measures from all the instances. In the "RUN" phase, a single instance is used thus, reported output measures are used.

**Scalability Assessment:** The cluster nodes are varied from 1 to 8 to measure scale-up performance of the Cassandra data store. The increase in number of nodes is accompanied by proportional increase in data. A single node is associated with 5GB of raw data.