Performance Testing of Big Data Applications
Key Big Data Technologies

• Map Reduce
  • Apache Hadoop, Cloudera, Hortonworks, MapR etc.

• NO SQL
  • Cassandra, Mongo DB, Oracle NoSQL, Neo4j etc.

• Messaging queues
  • Kafka, ActiveMQ, RabbitMQ, ZeroMQ etc.

• Search
  • Lucene, Elastic Search, Solr
Agenda

- Big Data Performance Testing Focus Areas
- Challenges
- Performance Testing Approach
- Solutions
Big Data Performance Test Focus Areas

Data Ingestion → Queues → mapReduce/ETL → Data Store/NoSQL → Reporting & Analytics

1. Sensor Data
2. Log
3. Facebook
4. Twitter

Data Processing

Data Persistence

Reporting & Analytics
Performance Testing Challenges

- Diverse technologies
- Unavailability of tools
- Test scripting
- Test environment
- Limited monitoring solutions
- Lack of diagnostic solutions
Performance Testing Approach

1. Setup Test Environment
2. Identify Workload Characteristics
3. Prepare Test Clients
4. Performance & Benchmark Tests
5. Optimum Configuration
6. Analyze Results & Tune Components
7. Failover & Reliability Tests
Performance Testing Solutions

- **Performance Test Tools**
  - YCSB (Yahoo Cloud Serving Benchmark), SandStorm, JMeter

- **Monitoring Tools**
  - Nagios, Zabbix, Ganglia, JMX utilities

- **Diagnostic Tools (APM)**
  - visualVM, AppDynamics, Compuware
Critical Performance Parameters

- Data Storage
- Commit Logs
- Concurrency
- Caching
- JVM parameters
- mapReduce configurations
- Message queue configurations
Real World Experience

About the application

• Online Social networking website with almost 100 million registered users across the globe
• Hundreds and thousands of users are online at any given time

The challenges

• Develop a near real time analytics solution to analyze the user feedback and interactions
• Solution uses Kafka, HBase and Hive as major technologies
• SLA to support 50K messages per minute
• Optimize and tune the Kafka clusters for maximum throughput
• Real time monitoring of test environment for bottleneck identification
Real World Experience

Impetus contributions

• Proposed and implemented performance test strategy for analytics solution
• Identified key performance components namely Kafka and Hbase for focus testing
• Proposed SandStorm as performance testing tool based on project requirements
• Prepared Kafka clients to simulate expected data ingestion volumes
• Optimized and tune single Kafka cluster in EC2 on medium instance
• Executed tests with varying message rate and reached to the max throughput of 50k messages per minute using 3 server cluster
• Real time monitoring using SandStorm to identify performance bottlenecks

Benefits Realized

• Optimum hardware utilization for complete solution
• Zero performance issues on Go live
• Maximum throughput of Kafka servers
Summary

• Test each component of the eco-system
  • Data Ingestion Rate, Throughput, System Resources

• Automate test environment in Cloud

• Appropriate tools and plug-ins
  • Cloud v/s On-premise solutions
  • Custom test harness
  • Monitoring and Diagnostic plug-ins

• Failover and Reliability testing
Q&A
Thank You

For more info mbatterywala@impetus.co.in