HBase Cluster Implementation and WebTable Use Case Project

Project Overview

This project involves the implementation of a Highly Available (HA) HBase cluster integrated with an existing Hadoop HA cluster. The implementation will include setting up the cluster infrastructure, testing failover mechanisms, creating custom Docker files, deploying the cluster, and implementing a WebTable use case to demonstrate HBase functionality.

Project Timeline

Start Date: May 1, 2025
End Date: May 15, 2025
Duration: 2 weeks

Technical Requirements

1. HBase Cluster Architecture and Integration

- **Cluster Configuration:**
  - 2 Master Nodes (HMaster nodes) (Active/Standby configuration)
  - 2-3 RegionServers
  - 3 ZooKeeper Quorum servers
  - Hadoop Cluster HA is up and running.
- **Integration Requirements:**
  - Integrate the HBase cluster with the existing Hadoop HA cluster.
  - Configure proper networking between all components
  - Ensure HBase can access the HDFS storage layer properly
  - Set up appropriate security and authentication mechanisms.

2. High Availability and Failover Configuration

- **Master Node Failover:**
  - Implement automatic failover for HBase Master nodes
  - Configure ZooKeeper for master election
  - Document the failover process.

- **RegionServer Failover:**
    - Configure automatic recovery of regions when a RegionServer fails
    - Ensure region failover and assignment mechanisms
    - Document the failover process
- **Testing Requirements:**
    - Create comprehensive test scenarios for both Master and RegionServer failures.
    - Simulate different failure conditions (process crash, host failure).
    - Measure and document recovery times.
    - Verify data integrity after failover events.

3. Containerization and Deployment

- **Docker Requirements:**
    - Create custom Docker files for all HBase components (not using public registry images)
    - Include all necessary dependencies and configuration.
- **Deployment Script Requirements:**
    - Develop automated deployment scripts to provision the entire cluster
    - Include configuration management for different environments
    - Implement proper sequencing for component startup
    - Add validation steps to ensure correct deployment

4. WebTable Use Case Implementation

- **Application Requirements:**
    - Implement the WebTable schema as specified in the requirements document.
    - Create necessary tables, column families, and performance tuning methods.
    - Develop data ingestion processes.

5. Documentation Requirements

- **Setup Documentation:**
    - Detailed step-by-step installation guide.
    - Configuration files with explanations.
    - Network architecture diagrams.
    - Security considerations and implementations.
- **Use Case Documentation:**
    - WebTable schema design explanation.
    - Data modeling decisions.
    - Performance optimization techniques.
    - Sample queries and expected results.

Deliverables

1. Infrastructure Code

- Complete Docker files for all HBase components
- Deployment scripts with documentation
- Configuration files for all components
- Testing scripts for validating the setup

2. WebTable Use Case Implementation

- Schema creation scripts
- Design documentation
- Data loading utilities
- Queries commands and results
- Performance tuning documentation

3. Testing Documentation

- Test plans for HA and failover scenarios
- Failure recovery documentation

4. Technical Documentation

- Architecture overview document
- Detailed setup guide
- Operational procedures

Evaluation Criteria

The project will be evaluated based on the following criteria:

HBase Cluster Setup Evaluation:

1. **Functionality (50%)**
   a. Successful HA configuration and failover
   b. Correct integration with Hadoop
   c. Working WebTable use case

2. **Code Quality (30%)**
    a. Clean, maintainable Docker files
    b. Well-structured deployment scripts
    c. Proper error handling
3. **Documentation (20%)**
    a. Completeness of documentation
    b. Clarity of instructions
    c. Quality of diagrams and explanations

WebTable Use-case Evaluation:

1. **Functionality (50%)**
    a. Schem and Key Design of Table, and Data Generation Script.
    b. Business Access Patterns.
    c. System Functionalities Implementation
2. **Code Quality (30%)**
    a. Clean, organized, and Well-structured commands/scripts files.
    b. Configuration optimization strategies and techniques.
3. **Documentation (20%)**
    a. Completeness of documentation
    b. Clarity of instructions
    c. Quality of diagrams and explanations

Bouns Section

- **WebTable Case Study:**
    - Backup and Recovery: build a backup and recovery mechnism in case of failure tools (CopyTable, or Import and Export).
    - Try another Client API (Java).
- **HBase HA Cluster Setup:**
    - Adding more regions servers and hbase master.
    - Troubleshooting and Log Analysis: implement and log analyzer for hbase log to detect a failures and documnet the failure solutions.
    - Monitoring: add a monitor tool like JMX into your cluster and integrate it with hbase for performance analysis.

Project Milestones

1. **Environment Setup (Days 1-3)**

    - Prepare Docker files and base configurations
    - Set up networking and security

2. **Cluster Deployment (Days 4-6)**

    - Deploy ZooKeeper ensemble
    - Deploy HBase Master and RegionServers
    - Integrate with Hadoop

3. **HA Configuration (Days 7-9)**

    - Configure and test the Master failover
    - Configure and test RegionServer failover
    - Document recovery procedures

4. **WebTable Implementation (Days 10-15)**

    - Implement schema and data loading
    - Develop and test queries
    - Optimize performance

5. **Final Testing and Documentation (Days 13-15)**

    - Comprehensive testing
    - Complete all documentation
    - Prepare final deliverables

References:

https://blog.newnius.com/setup-apache-hbase-in-docker.html

https://github.com/khalidmammadov/hbase_docker

https://github.com/mjaglan/docker-hbase-distributed-mode/tree/master

https://hub.docker.com/r/krejcmat/hadoop-hbase-master

https://hbase.apache.org/book.html

- Production image that we used during our course.