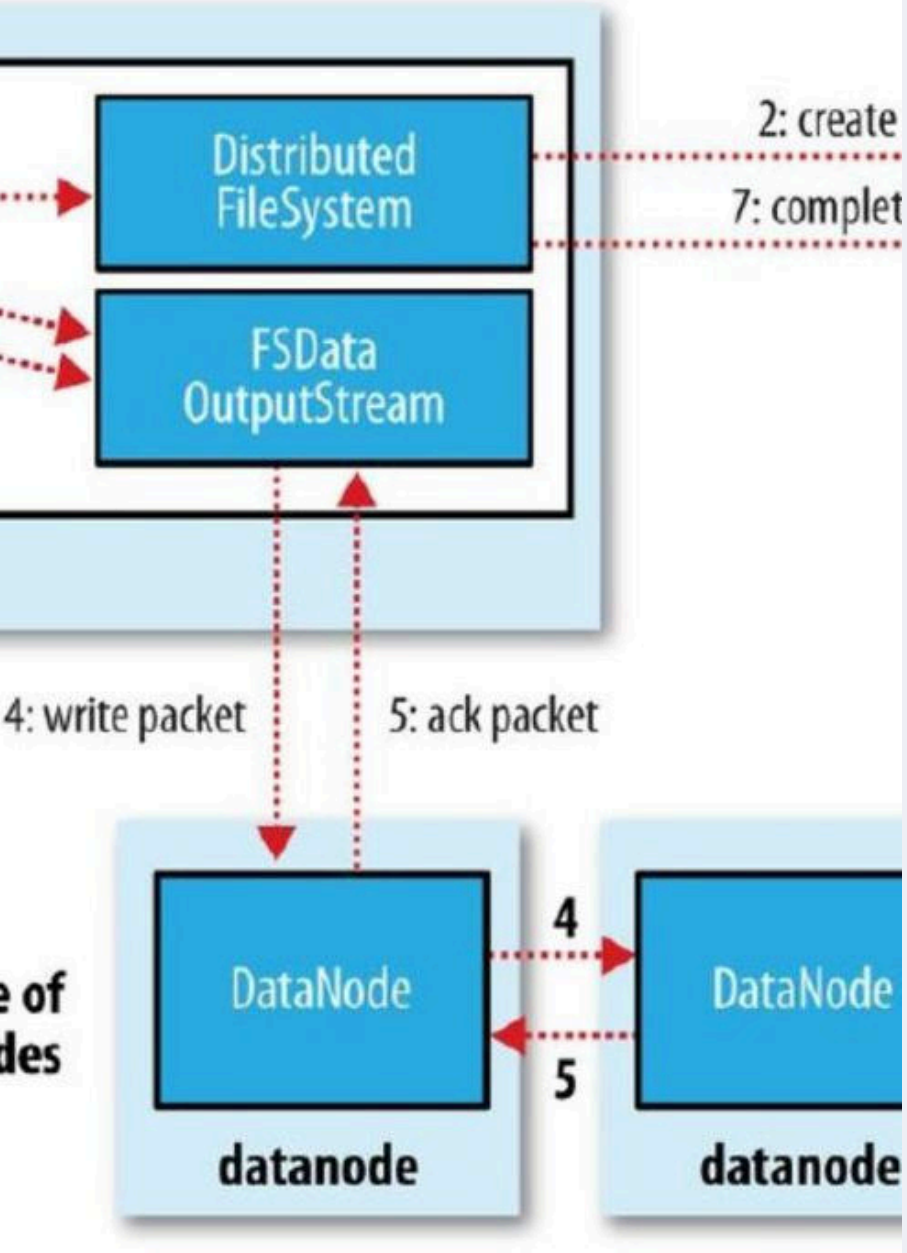


Creating Directory Structure in HDFS

Creating the directory structure in HDFS is the first step to organizing data in Hadoop. It involves planning the hierarchy of directories for efficient data storage and access. The structure should align with the organization's data requirements and access patterns.

KA **by Kakumanu Kalyani**

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Overview 'hadoop-master:9000' (active)

Setting Up Hadoop Cluster

Started:	Tue Apr 03 14:20:51 -0400 2018
Version:	3.0.1, r496dc57cc2e4f4da117f7a8e3840aaeac0c1d2d0
Compiled:	Fri Mar 16 19:00:00 -0400 2018 by lei from branch-3.0.1
Cluster ID:	CID
Block Pool ID:	BP

1 Resource Allocation

Allocate and configure hardware resources for the Hadoop cluster.

2 Cluster Configuration

Install and set up Hadoop components across the cluster nodes.

3 Network Configuration

Optimize network settings for seamless communication within the cluster.

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks = 1 total filesystem object(s).

Heap Memory used 22.77 MB of 29.29 MB Heap Memory. Max Heap Memory is 241.69 MB.

Non Heap Memory used 48.81 MB of 49.88 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configuring Hadoop Environment

Hadoop Configuration Files

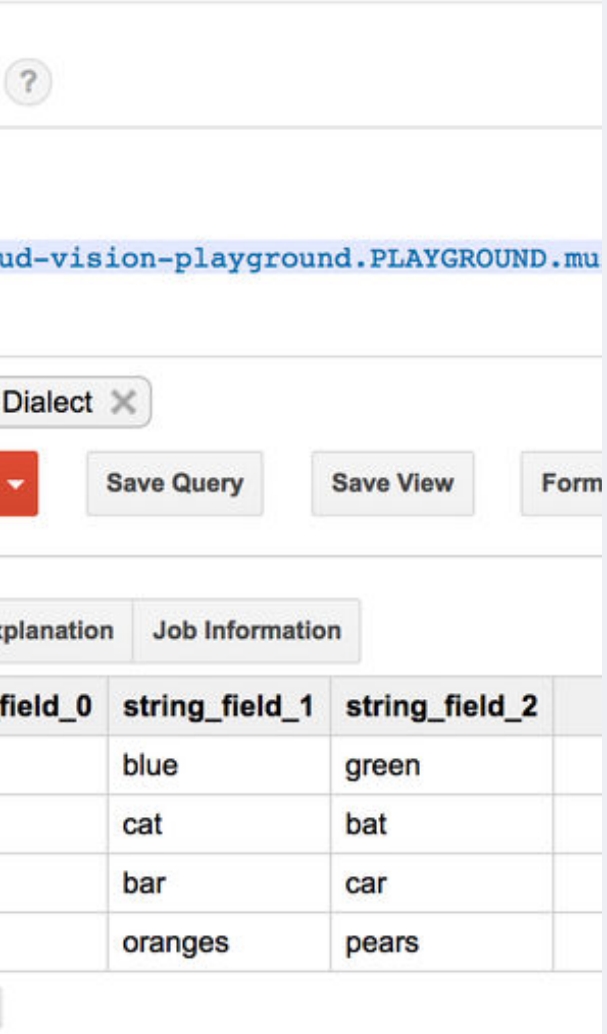
Modify core-site.xml and hdfs-site.xml to define Hadoop cluster configurations.

Memory Management

Adjust memory settings for efficient execution of Hadoop components.

Security Setup

Implement security measures, including user authentication and authorization.



Loading Data into HDFS

1 Data Ingestion Methods

Choose between direct data load, file upload, or distributed copy methodologies.

2 Data Replication Strategy

Determine the replication factor for fault tolerance and data redundancy.

3 Integration with ETL Tools

Integrate with ETL tools for seamless data extraction, transformation, and loading.

Running MapReduce Jobs

Map Phase

Divide and conquer data processing using the map phase.

Reduce Phase

Aggregate and summarize map outputs during the reduce phase.

Job Execution

Monitor and manage MapReduce jobs for optimal execution.

Monitoring Hadoop Cluster

Cluster Health

Monitor node status, resource utilization, and overall cluster health.

Alerting & Notifications

Set up alerts and notifications for critical cluster events and anomalies.

Performance Optimization

Analyze and optimize cluster performance based on resource usage and job execution.

Troubleshooting Hadoop Issues

1 2 3

Issue Identification

Thoroughly diagnose and identify the root cause of Hadoop issues.

Troubleshooting Steps

Follow systematic troubleshooting steps to resolve identified issues.

Log Analysis

Review logs and error messages to trace and resolve problems.

Best Practices for Hadoop Deployment

1

Performance Tuning

Fine-tune configurations for optimal performance and resource utilization.

2

Security Optimization

Implement encryption, authentication, and role-based access controls for data security.

3

Backup & Recovery

Establish robust backup and recovery mechanisms for data protection and disaster recovery.