Date: / /

Demonstrate Ambari's process to start and stop Hadoop services.

Apache Ambari is a web based management tool for provisioning, managing, and monitoring Apache Hadoop clusters.

Here a step-by-step guide on how to start and stop Hadoop services using Ambari

→ Access Ambari web interface
→ Navigate to Services Page
→ Start or stop a service
→ Confirm the Action
→ Monitor Service Status

Show how to monitor health using Ambari metrics.

Apache Ambari provides a comprehensive set of metrics to monitor the health and performance of your Hadoop cluster. Step-by-step guide on how to monitor cluster health using Ambari metrics.

→ Access Ambari web interface
→ Navigate to Dashboard
→ View Cluster Metrics
→ Drill Down into Service Metrics
→ View Host Metrics
→ Set up Alerts
→ Analyze Historical Data

Date: / /

Write and explain a Pig Script for Student marks Analysis.

→ Loading the Data
→ Grouping by student
→ Calculating Total marks
→ Calculating Average Marks
→ Filtering High Scores
→ output

Describe a real-world use case of IoT Contributing to Big Data.

Smart Cities

Traffic management in smart cities, IoT sensors are embedded in traffic lights, roadways, and vehicles to collect real-time data on traffic flow, vehicle speed, congestion levels, and accidents.

The data is transmitted continously to centralized Big Data systems, where it is processed and analyzed

City planners and traffic control systems use this Big Data to:
optimize traffic signal timings.
Predict and reduce congestion.
Provide real-time updates to commuters via

Date: / /

apps.

Improve emergency response times.

Q Big Data ~~Pipeline~~ Pipeline using Hive and Pig

A Big Data pipeline using Hive and Pig enables efficient processing, querying and transformation of large datasets stored in Hadoop Distributed File system.

Here a pipeline:- Data Ingestion
Data cleaning & Transformation
Data warehousing & Querying
output / visualization

Qy Show how to monitor cluster health using Ambari metrics.

To monitor cluster health using Apache Ambari follow using its metrics and dashboard tools.
Ambari provides real-time metrics and visualization for hadoop services, helping administrators detect and resolve performance issues.

Steps to monitor cluster Health using Ambari metrics

① Login to Ambari Web UI
② Dashboard Overview

Date: / /

Service Health monitoring
Host Metrics view
Ambari Metrics system (AMS)
Set and manage Alert
Use Heatmaps
Check Logs for Deep Analysis

Common metrics to watch

HDFS :- Live / Dead DataNodes, Disk usage, Block status

YARN :- Running Applications, containaiers, Resource usage.

Hive :- Query Latency, connections

HBase :- Region server status, Read/Write Latency

3° General :- CPU, Memory, Network, Disk Usage

Q2 Describe how Spark handles lazy evaluation.

Lazy evaluation means Spark does not execute transformations (like map, Filter, etc) immediately when they are called.

Instead, it builds a logical execution plan and waits until an action (like Co

Date: / /

count(), collect(), saveAsTextFile() is called.

How it works

Transformation
Actions trigger execution

Benefits of Lazy Evaluation

Performance optimization
Failure recovery
Efficient memory use

Q. Show a use case of real-time processing using spark streaming.

Real time Fraud Detection in Bank using Spark Streaming :-

A bank wants to monitor credit card transactions in real time to detect fraudulent activity. They need to Analyze transaction data as it arrives to flag suspicious patterns like:

→ Unusually high-value transactions
→ multiple transactions from different locations within a short time
→ Transactions from blacklisted IPs or merchants

Date: / /

How Spark streaming Helps

1. Data ingestion
2. Real Time Processing
3. Action

Example Technologies used :-

- Apache Kafka
  For real time transaction data ingestion
- Apache Spark Streaming
  For processing and Analysis.
- HBase / Cassandra
  To store flagged transactions.
- Dashboard
  For visualization and alert monitoring.

Benefits

- Detect fraud within seconds instead of hours
- Prevent financial losses and improve customer trust.
- Scalable and fault tolerant real-time analytics.

(A) Write A command to list all files in an HDFS directory.

Ans: To list all files in an HDFS directory use the following command

hdfs dfs -ls <directory_path>

(B) Which SQL command is used to create a table in RDBMS?

M The CREAT TABLE command is used to create a new table in RDBMS.

The CREAT command is in DDL Command.

(C) Which is the purpose of the hdfs-rm command?

M The rm command is used to remove objects such as files, directories, Symbolic links and so on from the file system like Unix

d) Name any one DCL (Data Control Language) command.

M The GRANT and Revoke command is used to assign specific privileges to a user.

(c) Which Command is used to insert data into an RDBMS table?

M The INSERT command is used to insert data into a table in a RDBMS

(a) What is cloudera?

M cloudera is a hybrid data platform designed for unmatched freedom to choose - any cloud, any analytics, any data. cloudera delivers faster and easier data management and data analytics for data anywhere, with optimal performance, scalability, and security.

(b) What is the Role of Apache Ambari in Hadoop?

M Apache Ambari, an open-source software is a management platform that provides web-based user interfaces and APIs for monitoring, managing and provisioning apache Hadoop clusters.

(c) Define mapreduce.

M mapreduce is a programming model that runs

Date: / /

on hadoop - a data analytics engine ~~that~~ widely used for Big data - and writes applications that run in parallel to process large volumes of data stored on clusters.

d) What is partitioning in Hive?

A partitioning strategy that is used to split a table into multiple files based on partition keys.

e) Define a Resilient Distributed Dataset (RDD)?

A Resilient Distributed Dataset (RDD) in Apache Spark is a fault tolerant distributed collection of data elements that can be processed in parallel across a cluster of machines.

Q. Discuss the role and usage of Apache Spark in the Hadoop ecosystem.

Apache Spark is a powerful open source distributed processing system designed for big data workloads. It enhances the Hadoop ecosystem by providing faster and more versatile data processing capabilities than Hadoop's traditional

Date: / /

Roles in the Hadoop Ecosystem

Processing Engine
Resource management
Enhanced Capabilities

Uses of Apache Spark

Batch Processing
Real-time Stream Processing
Interactive Queries
Machine Learning
Graph Processing

Key Advantages
Speed, Flexibility, Unified Engine,
Ease of use

Date: / /

5. Explain the evolution from traditional data processing to Big Data Processing

M The evolution from traditional data processing to Big data processing has been driven by the rapid growth of data volumes, varieties, and velocities.

Traditional data processing involved:-
Structured data
Batch processing
Centralized Architecture

Big Data Processing —

Distributed Architecture
Scalability
Flexibility
Real time processing

Some key technologies that have enabled the evolution to Big Data processing include:
Hadoop
Spark
NO SQL databases
cloud computing

The evolution to Big Data processing has provided several benefits:- Improved insights
Increased efficiency
Competitive advantage

Date: / /

46. Explain Ambari alerts and their role in cluster management.

Ambari alerts are automated health checks that monitor the status and performance of various services and components in a Hadoop cluster.

Types and Customization

Ambari support several types of alerts, including:

→ web, port, metric, aggregate, script, server, and recovery alerts.

→ Alerts can be customized for check intervals, thresholds and notification methods.

Role in cluster management

→ Continuous monitoring of cluster components and services for health and availability.

→ Immediate notification when thresholds are breached.

→ Quick identification and troubleshooting of issues

→ Centralized management.

Date: / /

10. Create a step-by-step guide to installing and configuring Apache Spark.

Step-by-step Guide:

Installing and Configuring Apache Spark Using Ambari

1. Download Apache Spark
2. Install Apache Spark
3. Configure Environment Variables
4. Configure Spark
5. Start Spark
6. Configure Spark Cluster
7. Verify Spark Cluster

List and describe Apache Spark libraries.

Apache Spark offers several core and extension libraries designed to handle different data processing and analytics tasks. Here are main ones:

Core Spark Libraries

- Spark Core

The foundation of Spark providing in memory computing and distributed task scheduling.

Date: / /

Handles basic I/O task scheduling and fault recovery.

Spark SQL

→ Enables querying structured data using SQL or Data Frame API.

→ Integrates with Hive, HDFS and other data sources for relational processing.

Spark Streaming

→ Process live data streams in real time.
→ Integrates with sources like Kafka, Flume, and Kinesis.

MLib

→ Scalable ML library.
→ Includes algorithms for classification, regression, clustering and collaborative filtering.
→ Compatible with Java, Scala, Python, and R.

GraphX

→ Graph processing library.
→ Provides APIs for graph-parallel computation and analytics.