HADOOP

INFORMACION

FORMATO
Presencial
En Sitio
A partir de 3 participantes

DURACIÓN:
60 Horas
10 días
Lunes a Viernes

DIRIGIDO A
Analistas y programadores que requieren procesar y analisar grandes volumenes de datos.

REQUISITOS:
Conocimientos básicos de Linux.

MATERIAL:
Manual Oficial
DVD de la distribución Linux más reciente

DOCUMENTO
Diploma expedido por PLCT S.A. DE C.V.

DESCRIPCION GENERAL

Curso orientado a usuarios y desarrolladores de aplicaciones que requieren accesar a a grandes volúmenes de datos para llevar acabo el procesamiento y análisis de información.

OBJETIVOS

Que el estudiante aprenda y aplique las técnicas y metodologías de procesamiento y análisis de grandes volúmenes de datos en un sistema Hadoop.

1. Meet Hadoop

Data!
Data Storage and Analysis
Comparison with Other Systems
Rational Database Management System
Grid Computing
Volunteer Computing
A Brief History of Hadoop
Apache Hadoop and the Hadoop Ecosystem
Hadoop Releases
What's Covered in This Book
Compatibility

2. MapReduce

A Weather Dataset
Data Format
Analyzing the Data with Unix Tools
Analyzing the Data with Hadoop
Map and Reduce
Java MapReduce
Scaling Out
Data Flow
Combiner Functions
Running a Distributed MapReduce Job
Hadoop Streaming
Ruby
Python
Hadoop Pipes
Compiling and Running

3. The Hadoop Distributed Filesystem

The Design of HDFS
HDFS Concepts
Blocks
Namenodes and Datanodes
HDFS Federation
HDFS High-Availability
The Command-Line Interface
Basic Filesystem Operations
Hadoop Filesystems
Interfaces
The Java Interface
Reading Data from a Hadoop URL
Reading Data Using the FileSystem API
Writing Data
Directories
Querying the Filesystem
Deleting Data
Data Flow
Anatomy of a File Read
Anatomy of a File Write
Coherency Model
Data Ingest with Flume and Sqoop
Parallel Copying with distcp
Keeping an HDFS Cluster Balanced
Hadoop Archives
Using Hadoop Archives
Limitations

4. Hadoop I/O

Data Integrity
Data Integrity in HDFS
LocalFileSystem
ChecksumFileSystem
Compression
Codecs
Compression and Input Splits
Using Compression in MapReduce
Serialization
The Writable Interface
Writable Classes
Implementing a Custom Writable
Serialization Frameworks
Avro
Avro Data Types and Schemas
In-Memory Serialization and Deserialization
Avro Datafiles
Interoperability
Schema Resolution
Sort Order
Avro MapReduce
Sorting Using Avro MapReduce
Avro MapReduce in Other Languages
File-Based Data Structures
SequenceFile
MapFile

5. Developing a MapReduce Application

The Configuration API
Combining Resources
Variable Expansion
Setting Up the Development Environment
Managing Configuration
GenericOptionsParser, Tool, and ToolRunner
Writing a Unit Test with MRUnit
Mapper
Reducer
Running Locally on Test Data
Running a Job in a Local Job Runner
Testing the Driver
Running on a Cluster
Packaging a Job
Launching a Job
The MapReduce Web UI
Retrieving the Results
Debugging a Job
Hadoop Logs
Remote Debugging
Tuning a Job
Profiling Tasks
MapReduce Workflows
Decomposing a Problem into MapReduce Jobs
JobControl
Apache Oozie

6. How MapReduce Works

Anatomy of a MapReduce Job Run
Classic MapReduce (MapReduce 1)
YARN (MapReduce 2)
Failures
Failures in Classic MapReduce
Failures in YARN
Job Scheduling
The Fair Scheduler
The Capacity Scheduler
Shuffle and Sort
The Map Side
The Reduce Side
Configuration Tuning
Task Execution
The Task Execution Environment
Speculative Execution
Output Committers
Task JVM Reuse
Skipping Bad Records

7. MapReduce Types and Formats

MapReduce Types
The Default MapReduce Job
Input Formats
Input Splits and Records
Text Input
Binary Input
Multiple Inputs
Database Input (and Output)
Output Formats
Text Output
Binary Output
Multiple Outputs
Lazy Output
Database Output

8. MapReduce Features

Counters
Built-in Counters
User-Defined Java Counters
User-Defined Streaming Counters
Sorting
Preparation
Partial Sort
Total Sort
Secondary Sort
Joins
Map-Side Joins
Reduce-Side Joins
Side Data Distribution
Using the Job Configuration
Distributed Cache
MapReduce Library Classes

9. Setting Up a Hadoop Cluster

Cluster Specification
Network Topology
Cluster Setup and Installation
Installing Java
Creating a Hadoop User
Installing Hadoop
Testing the Installation
SSH Configuration
Hadoop Configuration
Configuration Management
Environment Settings
Important Hadoop Daemon Properties
Hadoop Daemon Addresses and Ports
Other Hadoop Properties
User Account Creation
YARN Configuration
Important YARN Daemon Properties
YARN Daemon Addresses and Ports
Security
Kerberos and Hadoop
Delegation Tokens
Other Security Enhancements
Benchmarking a Hadoop Cluster
Hadoop Benchmarks
User Jobs
Hadoop in the Cloud
Apache Whirr

10. Administering Hadoop

HDFS
Persistent Data Structures
Safe Mode
Audit Logging
Tools
Monitoring
Logging
Metrics
Java Management Extensions
Maintenance
Routine Administration Procedures
Commissioning and Decommissioning Nodes
Upgrades

11. Pig

Installing and Running Pig
Execution Types
Running Pig Programs
Grunt
Pig Latin Editors
An Example
Generating Examples
Comparison with Databases
Pig Latin
Structure
Statements
Expressions
Types

Schemas
Functions
Macros
User-Defined Functions
A Filter UDF
An Eval UDF
A Load UDF
Data Processing Operators
Loading and Storing Data
Filtering Data
Grouping and Joining Data
Sorting Data
Combining and Splitting Data
Pig in Practice
Parallelism
Parameter Substitution

12. Hive

Installing Hive
The Hive Shell
An Example
Running Hive
Configuring Hive
Hive Services
The Metastore
Comparison with Traditional Databases
Schema on Read Versus Schema on Write
Updates, Transactions, and Indexes
HiveQL
Data Types
Operators and Functions
Tables
Managed Tables and External Tables
Partitions and Buckets
Storage Formats
Importing Data
Altering Tables
Dropping Tables
Querying Data
Sorting and Aggregating
MapReduce Scripts
Joins
Subqueries
Views
User-Defined Functions
Writing a UDF
Writing a UDAF

13. HBase

HBasics
Backdrop
Concepts
Whirlwind Tour of the Data Model
Implementation
Installation
Test Drive
Clients
Java
Avro, REST, and Thrift
Example
Schemas
Loading Data
Web Queries
HBase Versus RDBMS
Successful Service
HBase
Use Case: HBase at Streamy.com
Praxis
Versions
HDFS
UI
Metrics
Schema Design
Counters
Bulk Load

14. ZooKeeper

Installing and Running ZooKeeper
An Example
Group Membership in ZooKeeper
Creating the Group
Joining a Group
Listing Members in a Group
Deleting a Group
The ZooKeeper Service
Data Model
Operations
Implementation
Consistency
Sessions
States
Building Applications with ZooKeeper
A Configuration Service
The Resilient ZooKeeper Application
A Lock Service
More Distributed Data Structures and Protocols
ZooKeeper in Production
Resilience and Performance
Configuration

15. Sqoop

Getting Sqoop
Sqoop Connectors
A Sample Import
Text and Binary File Formats
Generated Code
Additional Serialization Systems
Imports: A Deeper Look
Controlling the Import
Imports and Consistency
Direct-mode Imports
Working with Imported Data
Imported Data and Hive
Importing Large Objects
Performing an Export
Exports: A Deeper Look
Exports and Transactionality
Exports and SequenceFiles

16. Case Studies

Hadoop Usage at Last.fm
Last.fm: The Social Music Revolution
Hadoop at Last.fm
Generating Charts with Hadoop
The Track Statistics Program
Summary
Hadoop and Hive at Facebook
Hadoop at Facebook
Hypothetical Use Case Studies
Hive
Problems and Future Work
Nutch Search Engine
Data Structures
Selected Examples of Hadoop Data Processing in Nutch
Summary
Log Processing at Rackspace
Requirements/The Problem
Brief History
Choosing Hadoop
Collection and Storage
MapReduce for Logs
Cascading
Fields, Tuples, and Pipes
Operations
Taps, Schemes, and Flows
Cascading in Practice
Flexibility
Hadoop and Cascading at ShareThis
Summary
TeraByte Sort on Apache Hadoop
Using Pig and Wukong to Explore Billion-edge Network Graphs
Measuring Community
Everybody's Talkin' at Me: The Twitter Reply Graph
Symmetric Links
Community Extraction