About Us

About Dilişim

Dilişim was founded in 2009 by Dr. Özgür Yılmazel who has a PhD in natural language processing and information extraction. Dilişim has expertise in Big Data Systems, Natural Language Processing and Search. Dilişim's vision and goal is to support its clients and create measurable value to its customers by utilizing data at their hand. Dilişim is Cloudera's first and only training partner in Turkey and also the only silver-level integrator partner in Turkey since 2012. Dilişim deployed first commercial Hadoop Cluster in Turkey, and it now runs the largest Hadoop Cluster in Turkey.

About Cloudera

Founded in 2008, Cloudera was the first, and is currently, the leading provider and supporter of Apache Hadoop for the enterprise. Cloudera also offers software for business critical data challenges including storage, access, management, analysis, security, and search. Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data: The Enterprise Data Hub.

What are Cloudera trainings?

Dilişim offers the following Cloudera trainings:

- Cloudera Developer Training for Spark and Hadoop (4 days)
- Cloudera Developer Training for Apache Spark (3 days)
- Cloudera Administrator Training for Apache Hadoop (4 days)
- Cloudera Data Analyst Training: Using Pig, Hive and Impala with Hadoop (4 days)
- Cloudera Training for Apache HBase (3 days)

The trainings in Turkey are delivered by Dilişim by being the only training partner of Cloudera in Turkey.

Why Cloudera Training?

Through instructor-led discussion and interactive, hands-on exercises, participants will navigatethe Hadoop ecosystem and experience the following:

- Most comprehensive suite of courses to address the Hadoop objectives of every data professional: developers, administrators, and data analysts.
- The industry's only truly dynamic and up-to-date Hadoop training curriculum
- Delivered by full-time technical and Cloudera certified instructors
- Industry leader in Hadoop with over 100.000 participants
- Video tutorials and e-learning services

dilişim™

cloudera®
TRAINING PARTNER

Cloudera Developer Training for Spark and Hadoop

Learn how to import data into your Apache Hadoop cluster and process it with Spark, Hive, Flume, Sqoop, Impala, and other Hadoop ecosystem tools

This four-day hands-on training course delivers the key concepts and expertise developers need to develop high-performance parallel applications with Apache Spark 2. Participants will learn how to use Spark SQL to query structure data and Spark Streaming to perform real-time processing on streaming data from a variety of sources. Developers will also practice writing applications that use core Spark to perform ETL processing and iterative algorithms. The course covers how to work with large datasets stored in distributed file system, and execute Spark applications on a Hadoop cluster. After taking this course, participants will be prepared to face real-world challenges and build applications to execute faster decisions, better decisions and interactive analysis, applied to a wide variety of use cases, architectures and industries. With this course update, we streamlined the agenda to help you quickly become productive with the most important technologies including Spark 2.

Hands-On Hadoop

Hands-on exercises take place on a live cluster, running in the cloud. A private cluster will be built for each student to use during the class. Through instructor-led discussion and interactive, hands-on exercises, participants will learn Apache Spark and how it integrates with the entire Hadoop ecosystem, learning:

- Distribute, store and process data in a Hadoop cluster
- Write, configure, and deploy Apache Spark applications on a Hadoop cluster
- Use the Spark shell for interactive data analysis
- Process and query structured data using Spark SQL
- Use Spark Streaming to process a live data stream

Audience and Prerequisites

This course is designed for developers and engineers who have programming experience. Apache Spark examples and hands-on exercises are presented in Scala and Python, so the ability to program in one of those languages is required. Basic familiarity with the Linux command line is assumed. Basic knowledge of SQL is helpful. Prior knowledge of Hadoop is not required.

CCA Spark & Hadoop Developer

This course is an excellent place to start for people working towards the CCA Spark & Hadoop Developer certification. It covers many of the subjects tested in the certification exam.

TRAINING SHEET

Cloudera Developer Training for Spark and Hadoop

Introduction to Hadoop and the Hadoop Ecosystem

- Apache Hadoop Overview
- Data Ingestion and Storage
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to the Hands-on Exercises

Apache Hadoop File Storage

- Apache Hadoop Cluster Components
- HDFS Architecture
- Using HDFS

Distributed Processing on an Apache Hadoop Cluster

- YARN Architecture
- Working with YARN

Apache Spark Basics

- What is Apache Spark?
- Starting the Spark Shell
- Using the Spark Shell
- Getting Started with Datasets and DataFrames
- DataFrame Operations

Working with DataFrames and Schemas

- Creating DataFrames from DataSources
- Saving DataFrames to Data Sources
- DataFrames Schemas
- Eager and Lazy Execution

Analyzing Data with DataFrame Queries

- Querying DataFrames Using Column Expressions
- Grouping and Aggregation Queries
- Joining DataFrames

RDD Overview

- RDD Overview
- RDD Data Sources
- Creating and Saving RDDs
- RDD Operations

Transforming Data with RDDs

- Writing and Passing Transformation Functions
- Tranforming Execution
- Converting Between RDDs and DataFrames

Aggregating Data with Pair RDDs

- Key-Value Pair RDDs
- Map-Reduce
- Other Pair RDD Operations

Querying Tables and Views with Apache Spark SQL

- Querying Tables in Spark Using SQL
- Querying Files and Views
- The Catalog API
- Comparing Spark SQL, Apache Impala and Apache Hive-on-Spark

Working with Datasets in Scala

- Datasets and DataFrames
- Creating Datasets
- Loading and Saving Datasets
- Dataset Operations

Writing Configuring and Running Apache Spark Applications

- Writing a Spark Application
- Building and Running an Application
- Application Deployment Mode
- The Spark and Application Web UI
- Configuring Application Properties

Distributed Processing

- Review: Apache Spark on a Cluster
- RDD Partitions
- Example: Partitioning in Queries
- Stages and Tasks
- Job Execution Planning
- Example: Catalyst Execution Plan
- Example: RDD Execution Plan

Distributed Data Persistence

- DataFrame and Dataset Persistence
- Persistence Storage Levels
- Viewing Persisted RDDs

Common Patterns in Apache Spark Data Processing

- Common Apache Spark Use Cases
- Iterative Algorithms in Apache Spark
- Machine Learning
- Example: k-means

Apache Spark Streaming: Introduction to DStreams

- Apache Spark Streaming Overview
- Example: Streaming Request Count
- DStreams
- Developing Streaming Applications

Apache Spark Streaming: Processing Multiple Batches

- Multi-Batch Operations
- Time Slicing
- State Operations
- Sliding Window Operations
- Preview: Structured Streaming

Apache Spark Streaming: Data Sources

- Streaming Data Sources Overview
- Apache Flume and Apache Kafka Data Sources
- Example: Using a Kafka Direct Data Source

dilişim™

E: info@dilisim.com
W: www.dilisim.com

dilişim™

cloudera®
TRAINING PARTNER

Cloudera Administrator Training for Apache Hadoop

Take your knowledge to the next level with Cloudera's Apache Hadoop

This four-day administrator training course for Apache Hadoop provides participants with a comprehensive understanding of all the steps necessary to operate and maintain a Hadoop cluster using Cloudera Manager. From installation and configuration through load balancing and tuning, Cloudera's training course is the best preparation for the real-world challenges faced by Hadoop administrators.

Hands-On Hadoop

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, learning topics such as:

- Cloudera Manager features that make managing your clusters easier, such as aggregated logging, configuration management, resource management, reports, alerts, and service management.
- The internals of YARN, MapReduce, Spark, and HDFS
- Determining the correct hardware and infrastructure for your cluster
- Proper cluster configuration and deployment to integrate with the data center
- How to load data into the cluster from dynamically-generated files using Flume and from RDBMS using Sqoop
- Configuring the FairScheduler to provide service-level agreements for multiple users of a cluster
- Best practices for preparing and maintaining Apache Hadoop in production
- Troubleshooting, diagnosing, tuning, and solving Hadoop issues

Audience & Prerequisites

This course is best suited to systems administrators and IT managers who have basic Linux experience. Prior knowledge of Apache Hadoop is not required.

Administrator Certification

Upon completion of the course, attendees are encouraged to continue their study and register for the Cloudera Certified Administrator for Apache Hadoop (CCAH) exam. Certification is a great differentiator. It helps establish you as a leader in the field, providing employers and customers with tangible evidence of your skills and expertise.

TRAINING SHEET

Cloudera Administrator Training for Apache Hadoop

Introduction

The Case for Apache Hadoop

- Why Hadoop?
- Fundamental Concepts
- Core Hadoop Components

Hadoop Cluster Installation

- Rationale for a Cluster Management Solution
- Cloudera Manager Features
- Cloudera Manager Installation
- Hadoop (CDH) Installation

The Hadoop Distributed File System (HDFS)

- HDFS Features
- Writing and Reading Files
- NameNode Memory Considerations
- Overview of HDFS Security
- Web UIs for HDFS
- Using the Hadoop File Shell

MapReduce and Spark on YARN

- The Role of Computational Frameworks
- YARN: The Cluster Resource Manager
- MapReduce Concepts
- Apache Spark Concepts
- Running Computational Frameworks on YARN
- Exploring YARN Applications Through the Web UIs, and the Shell
- YARN Application Logs

Hadoop Configuration and Daemon Logs

- Cloudera Manager Constructs for Managing Configurations
- Locating Configurations and Applying Configuration Changes
- Managing Role Instances and Adding Services
- Configuring the HDFS Service
- Configuring Hadoop Daemon Logs
- Configuring the YARN Service

Getting Data Into HDFS

- Ingesting Data From External Sources With Flume
- Ingesting Data From Relational Databases With Sqoop
- REST Interfaces
- Best Practices for Importing Data

Planning Your Hadoop Cluster

- General Planning Considerations
- Choosing the Right Hardware
- Virtualization Options
- Network Considerations
- Configuring Nodes

Installing and Configuring Hive, Impala, and Pig

- Hive
- Impala
- Pig

Hadoop Clients Including Hue

- What Are Hadoop Clients?
- Installing and Configuring Hadoop Clients
- Installing and Configuring Hue
- Hue Authentication and Authorization

Advanced Cluster Configuration

- Advanced Configuration Parameters
- Configuring Hadoop Ports
- Configuring HDFS for Rack Awareness
- Configuring HDFS High Availability

Hadoop Security

- Why Hadoop Security Is Important
- Hadoop's Security System Concepts
- What Kerberos Is and how it Works
- Securing a Hadoop Cluster With Kerberos
- Other Security Concepts

Managing Resources

- Configuring cgroups with Static Service Pools
- The Fair Scheduler
- Configuring Dynamic Resource Pools
- YARN Memory and CPU Settings
- Impala Query Scheduling

Cluster Maintenance

- Checking HDFS Status
- Copying Data Between Clusters
- Adding and Removing Cluster Nodes
- Rebalancing the Cluster
- Directory Snapshots
- Cluster Upgrading

Cluster Monitoring and Troubleshooting

- Cloudera Manager Monitoring Features
- Monitoring Hadoop Clusters
- Troubleshooting Hadoop Clusters
- Common Misconfigurations

Conclusion

dilişim™

E: info@dilisim.com
W: www.dilisim.com

dilişim™

cloudera®
TRAINING PARTNER

Cloudera Data Analyst Training: Using Pig, Hive and Impala with Hadoop

Take your knowledge to the next level with Cloudera's Apache Hadoop Training

This four-day data analyst training course focusing on Apache Pig and Hive and Cloudera Impala will teach you to apply traditional data analytics and business intelligence skills to big data. Cloudera presents the tools data professionals need to access, manipulate, transform, and analyze complex data sets using SQL and familiar scripting languages.

Advance Your Ecosystem Expertise

Apache Hive makes multi-structured data accessible to analysts, database administrators, and others without Java programming expertise. Apache Pig applies the fundamentals of familiar scripting languages to the Hadoop cluster. Cloudera Impala enables real-time interactive analysis of the data stored in Hadoop via a native SQL environment.

Hands-On Hadoop

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, learning topics such as:

- The features that Pig, Hive, and Impala offer for data acquisition, storage, and analysis
- The fundamentals of Apache Hadoop and data ETL (extract, transform, load), ingestion, and processing with Hadoop tools
- How Pig, Hive, and Impala improve productivity for typical analysis tasks
- Joining diverse datasets to gain valuable business insight
- Performing real-time, complex queries on datasets

Audience & Prerequisites

This course is designed for data analysts, business intelligence specialists, developers, system architects, and database administrators. Knowledge of SQL is assumed, as is basic Linux command-line familiarity. Knowledge of at least one scripting language (e.g., Bash scripting, Perl, Python, Ruby) would be helpful but is not essential. Prior knowledge of Apache Hadoop is not required.

TRAINING SHEET

Cloudera Data Analyst Training: Using Pig, Hive, and Impala with Hadoop

Introduction

Hadoop Fundamentals

- The Motivation for Hadoop
- Hadoop Overview
- Data Storage: HDFS
- Distributed Data Processing: YARN, MapReduce, and Spark
- Data Processing and Analysis: Pig, Hive, and Impala
- Data Integration: Sqoop
- Other Hadoop Data Tools
- Exercise Scenarios Explanation

Introduction to Pig

- What Is Pig?
- Pig's Features
- Pig Use Cases
- Interacting with Pig

Basic Data Analysis with Pig

- Pig Latin Syntax
- Loading Data
- Simple Data Types
- Field Definitions
- Data Output
- Viewing the Schema
- Filtering and Sorting Data
- Commonly-Used Functions

Processing Complex Data with Pig

- Storage Formats
- Complex/Nested Data Types
- Grouping
- Built-In Functions for Complex Data
- Iterating Grouped Data

Multi-Dataset Operations with Pig

- Techniques for Combining Data Sets
- Joining Data Sets in Pig
- Set Operations
- Splitting Data Sets

Pig Troubleshooting and Optimization

- Troubleshooting Pig
- Logging
- Using Hadoop's Web UI
- Data Sampling and Debugging
- Performance Overview
- Understanding the Execution Plan
- Tips for Improving the Performance of Your Pig Jobs

Introduction to Hive and Impala

- What Is Hive?
- What Is Impala?
- Schema and Data Storage
- Comparing Hive to Traditional Databases
- Hive Use Cases

Querying with Hive and Impala

- Databases and Tables
- Basic Hive and Impala Query Language Syntax
- Data Types
- Differences Between Hive and Impala Query Syntax
- Using Hue to Execute Queries
- Using the Impala Shell

Data Management

- Data Storage
- Creating Databases and Tables
- Loading Data
- Altering Databases and Tables
- Simplifying Queries with Views
- Storing Query Results

Data Storage and Performance

- Partitioning Tables
- Choosing a File Format
- Managing Metadata
- Controlling Access to Data

Relational Data Analysis with Hive and Impala

- Joining Datasets
- Common Built-In Functions
- Aggregation and Windowing

Working with Impala

- How Impala Executes Queries
- Extending Impala with User-Defined Functions
- Improving Impala Performance

Analyzing Text and Complex Data with Hive

- Complex Values in Hive
- Using Regular Expressions in Hive
- Sentiment Analysis and N-Grams
- Conclusion

Hive Optimization

- Understanding Query Performance
- Controlling Job Execution Plan
- Bucketing
- Indexing Data

Extending Hive

- SerDes
- Data Transformation with Custom Scripts
- User-Defined Functions
- Parameterized Queries

Choosing the Best Tool for the Job

- Comparing MapReduce, Pig, Hive, Impala, and Relational Databases
- Which to Choose?

Conclusion

dilişim™

E: info@dilisim.com
W: www.dilisim.com

dilişim™

cloudera®
TRAINING PARTNER

Cloudera Training for Apache HBase

Take your knowledge to the next level with Cloudera's Apache Hadoop Training and Certification

This three-day training course for Apache HBase enables participants to store and access massive quantities of multi-structured data and perform hundreds of thousands of operations per second.

Advance Your Ecosystem Expertise

Apache HBase is a distributed, scalable, NoSQL database built on Apache Hadoop. HBase can store data in massive tables consisting of billions of rows and millions of columns, serve data to many users and applications in real time, and provide fast, random read/write access to users and applications.

Hands-On Hadoop

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, learning topics such as:

- The use cases and usage occasions for HBase, Hadoop, and RDBMS
- Using the HBase shell to directly manipulate HBase tables
- Designing optimal HBase schemas for efficient data storage and recovery
- How to connect to HBase using the Java API to insert and retrieve data in real time
- Best practices for identifying and resolving performance bottlenecks

Audience & Prerequisites

This course is appropriate for developers and administrators who intend to use HBase. Prior experience with databases and data modeling is helpful, but not required. Knowledge of Java is assumed. Prior knowledge of Hadoop is not required, but Cloudera Developer Training for Apache Hadoop provides an excellent foundation for this course.

HBase Certification

Upon completion of the course, attendees are encouraged to continue their study and register for the Cloudera Certified Specialist in Apache HBase (CCSHB) exam. Certification is a great differentiator; it helps establish you as a leader in the field, providing employers and customers with tangible evidence of your expertise.

TRAINING SHEET

Cloudera Training for Apache HBase

Introduction

Introduction to Hadoop and HBase

- What Is Big Data?
- Introducing Hadoop
- Hadoop Components
- What Is HBase?
- Why Use HBase?
- Strengths of HBase
- HBase in Production
- Weaknesses of HBase

HBase Tables

- HBase Concepts
- HBase Table Fundamentals
- Thinking About Table Design

The HBase Shell

- Creating Tables with the HBase Shell
- Working with Tables
- Working with Table Data

HBase Architecture Fundamentals

- HBase Regions
- HBase Cluster Architecture
- HBase and HDFS Data Locality

HBase Schema Design

- General Design Considerations
- Application-Centric Design
- Designing HBase Row Keys
- Other HBase Table Features

Basic Data Access with the HBase API

- Options to Access HBase Data
- Creating and Deleting HBase Tables
- Retrieving Data with Get
- Retrieving Data with Scan
- Inserting and Updating Data
- Deleting Data

More Advanced HBase API Features

- Filtering Scans
- Best Practices
- HBase Coprocessors

HBase on the Cluster

- How HBase Uses HDFS
- Compactions and Splits

HBase Reads and Writes

- How HBase Writes Data
- How HBase Reads Data
- Block Caches for Reading

HBase Performance Tuning

- Column Family Considerations
- Schema Design Considerations
- Configuring for Caching
- Dealing with Time Series and Sequential Data
- Pre-Splitting Regions

HBase Administration and Cluster Management

- HBase Daemons
- ZooKeeper Considerations
- HBase High Availability
- Using the HBase Balancer
- Fixing Tables with hbck
- HBase Security

HBase Replication and Backup

- HBase Replication
- HBase Backup
- MapReduce and HBase Clusters

Using Hive and Impala with HBase

- Using Hive and Impala with HBase

Conclusion

Appendix A: Accessing Data with Python and Thrift

- Thrift Usage
- Working with Tables
- Getting and Putting Data
- Scanning Data
- Deleting Data
- Counters
- Filters

Appendix B: OpenTSDB

dilişim™

E: info@dilisim.com
W: www.dilisim.com

dilişim™

cloudera®
TRAINING PARTNER

Cloudera Essentials for Apache Hadoop

Summary

This one-day course gives decision-makers an overview of Apache Hadoop and how it can help them meet business goals.

You Will Learn

- When is Hadoop appropriate?
- What are people using Hadoop for?
- How does Hadoop fit into our existing environment?
- What do I need to know about choosing Hadoop?
- What resources will I need to deploy Hadoop?

Audience & Prerequisites

Architects, Technical Managers, CTOs, Engineering Managers, etc. No prior Hadoop experience is required.

Outline

- Introduction
- The Motivation for Hadoop
- Hadoop: Basic Concepts
- Hadoop Solutions
- The Hadoop Ecosystem
- Hadoop in the Data Center
- Managing the Elephant in the Room

dilişim™

cloudera®
TRAINING PARTNER

Just Enough Scala

Summary

This one-day Scala training course will teach you the key language concepts and programming techniques you need so that you can concentrate on the subjects covered in Cloudera's developer courses without also having to learn a complex programming language and a new programming paradigm on the fly.

Prerequisites

Prior knowledge of Hadoop is not required. Since this course is intended for developers who do not yet have the prerequisite skills writing code in Scala, basic programming experience in at least one commonly-used programming language (ideally Java, but Python, Ruby, Perl, C, C++, PHP, or Javascript will suffice) is assumed. NOTE: This course does not teach Big Data concepts, nor does it cover how to use Cloudera software. Instead, it is meant as a precursor for one of our developer-focused training courses that provide those skills.

Outline

Introduction

Scala Basics

- Scala Background Information
- Key Scala Concepts
- Programming in Scala

Variables

- Scala Variables
- Numerical
- Boolean
- String

Collections

- Tuples
- The Collections Hierarchy
- Sets
- Lists
- Arrays
- Maps
- Common Conversions

Flow Control

- Looping
- Iterators
- Functions
- Passing Functions
- Collection Iteration Methods
- Pattern Matching

Libraries

- Classes and Objects
- Packages
- Import

Conclusion

dilişim™

cloudera®
TRAINING PARTNER

Just Enough Python

Summary

This one-day Python training course will teach you the key language concepts and programming techniques you need so that you can concentrate on the subjects covered in Cloudera's developer courses without also having to learn a complex programming language and a new programming paradigm on the fly.

Prerequisites

Prior knowledge of Hadoop is not required. Since this course is intended for developers who do not yet have the prerequisite skills writing code in Scala, basic programming experience in at least one commonly-used programming language (ideally Java, but Ruby, Perl, Scala, C, C++, PHP, or Javascript will suffice) is assumed. NOTE: This course does not teach Big Data concepts, nor does it cover how to use Cloudera software. Instead, it is meant as a precursor for one of our developer-focused training courses that provide those skills.

Outline

Introduction

Introduction to Python

- Python Background Information
- Scope
- Exercises

Variables

- Python Variables
- Numerical
- Boolean
- String

Collections

- Lists
- Tuples
- Sets
- Dictionaries

Flow Control

- Code Blocks
- Repetitive Execution
- Iterative Execution
- Conditional Execution
- Tentative Execution (Exception Handling)

Program Structure

- Named Functions
- Anonymous Functions (Lambda)
- Generator Functions

Working with Libraries

- Storing and Retrieving Functions
- Module Control
- Common Standard Libraries

Conclusion

dilişim™

Bigdata References