HACKERU

Hadoop and Spark

DM103

32
Academic Hours

Hadoop and Spark

Outline

The course delivers the key concepts and expertise participants need to ingest and process data on a Hadoop cluster using the most up-to-date tools and techniques. How to employ Hadoop ecosystem projects such as Spark, Hive, Flume, Sqoop, and Impala. Learning about the challenges faced by Hadoop developers. Participants learn to identify which tool is the right one to use in a given situation, and will gain hands-on experience in developing using those tools.

Target Audience

- Software developers
- Software architects
- Professionals willing to start working with Big Data

Prerequisites

Basic knowledge of database concepts and development environments

Content

Module 01

Introduction to Hadoop and the Hadoop Ecosystem

- Problems with Traditional Large-scale Systems
- The Hadoop EcoSystem

Module 02

Hadoop Architecture and HDFS

- Distributed Processing on a Cluster
- Storage: HDFS Architecture
- Storage: Using HDFS
- Resource Management: YARN Architecture
- Resource Management: Working with YARN
- Exercise

Module 03

Importing Relational Data with Apache Sqoop

- Sqoop Overview
- Basic Imports and Exports
- Limiting Results
- Improving Sqoop's Performance
- Sqoop 2
- Exercise

Module 04

Introduction to Impala and Hive

- Introduction to Impala and Hive
- Why Use Impala and Hive?
- Comparing Hive to Traditional Databases
- Hive Use Cases

Module 05

Modeling and Managing Data with Impala and Hive

- Data Storage Overview
- Creating Databases and Tables
- Loading Data into Tables
- HCatalog
- Impala Metadata Caching
- Exercise

> "Hadoop and Spark- **Learning about the challenges** faced by Hadoop developers"

Module 06

Data Formats

- Selecting a File Format
- Hadoop Tool Support for File Formats
- Avro Schemas
- Using Avro with Hive and Sqoop
- Avro Schema Evolution
- Compression
- Exercise

Module 07

Data Partitioning

- Partitioning Overview
- Partitioning in Impala and Hive
- Exercise

Module 08

Capturing Data with Apache Flume

- What is Apache Flume?
- Basic Flume Architecture
- Flume Sources
- Flume Sinks
- Flume Channels
- Flume Configuration
- Exercise

Module 09

Spark Basics

- What is Apache Spark?
- Using the Spark Shell
- RDDs (Resilient Distributed Datasets)
- Functional Programming in Spark
- Exercise

Module 10
Working with RDDs in Spark

- A Closer Look at RDDs
- Key-Value Pair RDDs
- MapReduce
- Other Pair RDD Operations
- Exercise

Module 11
Writing and Deploying Spark Applications

- Spark Applications vs. Spark Shell
- Creating the SparkContext
- Building a Spark Application (Scala & Java)
- Running a Spark Application
- The Spark Application Web UI
- Configuring Spark Properties
- Logging
- Exercise

Module 12
Parallel Programming with Spark

- Review: Spark on a Cluster
- RDD Partitions
- Partitioning of File-based RDDs
- HDFS and Data Locality
- Executing Parallel Operations
- Stages and Tasks
- Exercise

Module 13
Spark Caching and Persistence

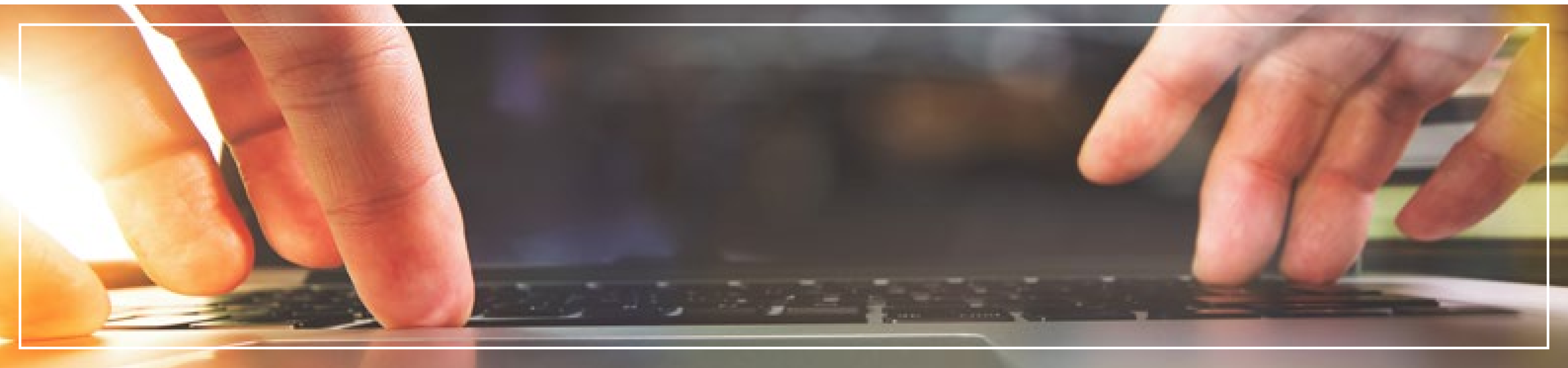
- RDD Lineage
- Caching Overview
- Distributed Persistence
- Exercise

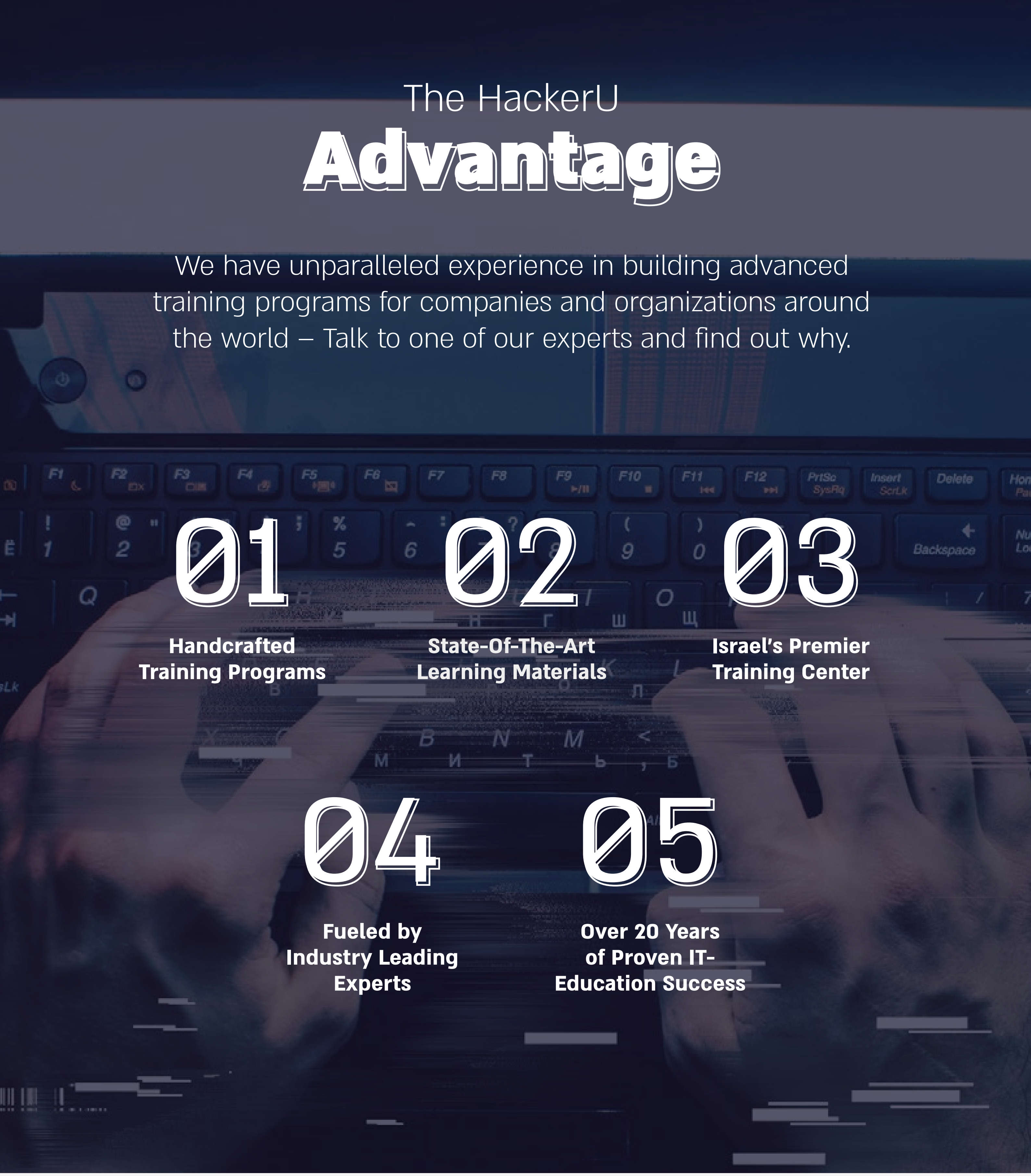
Module 14
Common Patterns in Spark Data Processing

- Common Spark Use Cases
- Iterative Algorithms in Spark
- Graph Processing and Analysis
- Machine Learning
- Example: k-means
- Exercise

Module 15
Spark SQL

- Spark SQL and the SQL Context
- Creating DataFrames
- Transforming and Querying DataFrames
- Saving DataFrames
- Comparing Spark SQL with Impala
- Exercise





The HackerU Advantage

We have unparalleled experience in building advanced training programs for companies and organizations around the world – Talk to one of our experts and find out why.

01

Handcrafted Training Programs

02

State-Of-The-Art Learning Materials

03

Israel's Premier Training Center

04

Fueled by Industry Leading Experts

05

Over 20 Years of Proven IT-Education Success

info@hackerupro.com

www.hackerupro.com