Hewlett Packard Enterprise

HPE Digital Learner Big Data/Data Sciences Fundamentals Content Pack

HPE Content Pack number	CP009
Content Pack length	21 Hours
Content Pack category	Category 2
Learn more	**View now**

Why HPE Education Services?

- IDC MarketScape leader 5 years running for IT education and training*
- Recognized by IDC for leading with global coverage, unmatched technical expertise, and targeted education consulting services*
- Key partnerships with industry leaders OpenStack®, VMware®, Linux®, Microsoft®, ITIL, PMI, CSA, and SUSE
- Complete continuum of training delivery options—self-paced eLearning, custom education consulting, traditional classroom, video on-demand instruction, live virtual instructor-led with hands-on lab, dedicated onsite training
- Simplified purchase option with HPE Training Credits

*Realize Technology Value with Training, IDC Infographic 2037, Sponsored by HPE, October 2017

This 21-hour collection of self-paced eLearning content provides technical and non-technical users with insight into the benefits of implementing big data projects. In addition, users gain practical insight on how to transform their business into a data driven enterprise. Topics include: why businesses should implement big data initiatives; a discussion of key data analytics concepts; process/governance suggestions to ensure repeatable and accurate data gathering; a review of key data analysis tools such as Hadoop; and the role data scientists can play in driving enterprise success.

Audience

- Anyone interested in a practical primer in big data
- Individuals looking for practical and useable lessons learned" related to the challenges, risks, opportunities and benefits associated with big data implementations

Content Pack objectives

This Content Pack provides the information necessary to:

- Understand the fundamentals of, and the tools available for, working with big data
- See the challenges, risks, opportunities and benefits related to becoming a data driven enterprise
- Explain, in detail, the features, benefits, and uses of Apache Hadoop
- Grasp how teams work in big data organizations and explore key big data use cases
- Understand the value of a data governance strategy and learn how to create such a strategy
- Recognize the importance of securing your enterprise data and the consequences of not securing data
- Learn the difference between "big data" and "data science"
- Successfully navigate the privacy and legal concerns surrounding big data when transforming into a data driven enterprise
- Absorb the foundational elements of data science while exploring practical tools for data exploration, gathering and filtering
- Explore the conceptual elements of key machine learning techniques

Detailed Content Pack outline

Big Data Overview

This session provides an overview of the opportunities and challenges that customers face when they evaluate and implement big data solutions. During this webinar, we will discuss HPE's big data strategies and how HPE helps customers solve their most difficult big data problems.

Outline

- Module 1 : Welcome to the Course
  - Introduction
  - Business intelligence platform types
  - HPE strategy and services
- Module 2 : Challenges and Opportunities
  - Data landscape
  - Data processing
  - Barriers to success
  - Big data lessons learned
  - Hadoop and the data lake
  - Challenges in maximizing value with Hadoop
  - Challenges with legacy systems
  - Evolution of Hadoop
  - Typical use case: EDW modernization with Hadoop
  - Deep learning
  - Evolution of leveraging analytics
  - Machine learning and common techniques
  - Trends in deep learning hardware
- Module 3 : HPE Big Data Strategy
  - Big data trends
  - Data solutions
  - Data lakes
  - Hadoop ecosystem
  - SMACK stack
  - Hadoop add-on solutions
  - HPE Enterprise Hadoop
- Module 4 : HPE Pointnext
  - HPE Pointnext and the big data ecosystem portfolio
  - HPE Pointnext lifecycle
  - Hadoop services across the lifecycle

The Big Data Technology Wave

A number of tools are available for working with big data. Many of the tools are open source and Linux distribution based. This course covers the fundamentals of big data, positioning big data in a historical IT context, the tools available for working with big data, the big data stack, and an in-depth look at Apache Hadoop.

Outline

- Put big data into the perspective of supercomputing
- Describe big data in context of technology waves and put it into perspective by comparing to previous technology waves
- List the six emerging technologies and relate them to big data
- Define big data and describe Gartner's vectors
- Define structured and unstructured data in terms of Gartner's model
- List the standard sizes used in big data to determine sizes of data sets
- List the three primary key contributors to the origins of big data
- List the primary big data distro companies
- Describe the Apache Software Foundation
- List projects attributable to the Apache Software Foundation
- Describe cascading and MongoDB
- List the layers of the big data stack
- List the common big data components
- Describe columnar databases and Hbase
- Describe solutions for scaling computing
- Describe the design principles of Hadoop
- Map out the functional view of Hadoop
- Describe the architecture of HDFS
- Describe the architecture of YARN
- Describe the attributes and processes of MapReduce
- Describe the architecture of Spark
- Describe big data in a historical context and the tools available for working with big data

Big Data Opportunities and Challenges

Big Data requires a holistic approach and a change to regular working practices. This course covers the way teams work in big data organizations, projects and use cases for big data, and challenges and opportunities related to big data.

Outline

- List the team members in a big data team and describe their interrelationship
- List the team members in a big data business team and describe their interrelationship
- List the team members in a big data analytics team and describe their interrelationship
- List the team members in a big data architects team and describe their interrelationship
- List the team members in a big data Hadoop operators team and describe their interrelationship
- Describe the factors impacting big data teams
- Describe DIY supercomputing in big data projects
- Describe the use of Hadoop in cloud computing
- Describe data warehousing for big data
- Identify the key factors impacting the cost of big data
- Assess relational database management systems in the context of big data
- Assess big data infrastructure projects
- Describe data mining
- Describe big data recommendation engines
- Describe common use cases for big data analytics
- Describe the global impact of big data
- Describe the global increase in produced data volume
- List the main companies involved in big data
- Identify big data business opportunities
- Describe the big data physical challenges
- Describe the privacy and security challenges of big data
- Describe the required planning activities for big data
- Describe big data industry trends
- Describe and identify the challenges and opportunities presented by big data

Big Data Corporate Leadership Perspective

Big data leaders must have skill sets that differ from the skills sets of leaders in the past. They must be able to show how big data generates value; how investments in big data initiatives should be targeted; and how fast the organization should move to implement them. In this course, you will learn how to create a governance strategy, examine security concerns, and explore how big data impacts human resources.

Outline

- Decide where to start
- Establish economic value
- Learn how big data products are economic engines
- Compare building vs. buying approaches to big data skill
- Design a governance strategy
- Identify security concerns
- Identify stewardship issues
- Describe the impact to the human resources department
- Describe the big data culture clash
- Learn what incubators are and how they can improve companies
- Become more familiar with big data as it relates to corporate leadership

Big Data Engineering Perspectives

Big data is a term for data sets so large that traditional data processing applications cannot be used to perform analysis. It is often semi-structured or unstructured in form. There are a number of unique challenges that arise when companies begin to use big data, the least of which are engineering concerns. This course introduces engineering challenges and describes the solutions created by various companies.

Outline

- Describe big data and the three v's
- Identify important factors when considering a big data infrastructure
- Examine options when building a cloud-based big data infrastructure
- Compare the pros and cons of building an in-house big data infrastructure
- Understand the skill sets necessary for individuals on the big data team
- Describe the different software options that are available for big data analytics
- Identify the types of data analytics
- Compare the different types of data visualization that can be done with the analytics
- Review the effective use of big data analytics among large companies
- Describe the risks involved when considering big data as a solution
- Understand big data engineering concerns, such as choosing storage and software

Big Data Marketing Perspective

Big data dramatically impacts all aspects of business culture. Companies need to evolve from traditional methods and practices as they learn to use big data to improve their organization. This course examines the impacts of big data from the marketing perspective. We look at how the mobile effect has changed marketing, how purchasing habits have changed and how datafying has impacted consumer behavior.

Outline

- Describe the value of repeat shoppers
- Describe how marking strategies need to change with technology
- Describe how the mobile effect has changed marketing
- Describe how technology has changed the way that product discovery and product research is done
- Describe how big data is used in marketing
- Describe the new market research methods
- Describe different ways that companies can misuse the new technology
- Describe how big data has datafied consumer behavior
- Describe technique changes in split testing and cross-channel marketing
- Describe how using user profiling can assist in advertising
- Understand how big data affects marketing and various marketing techniques

Big Data Strategic Planning

In order to adopt big data, senior leadership must be able to establish investment priorities, balance speed and cost, and ensure acceptance by the front line. They must also build a plan based on data, analytic models, and tools. In this course, you will compare scaling up to scaling out, identify different analytical models, and learn how to secure funding for data initiatives.

Outline

- Describe what big data is and is not
- Identify the value of big data
- Describe the four main challenges of big data
- Understand why big data is a leadership problem
- Integrate from multiple data sources
- Compare scaling up to scaling out
- Manage risk and governance issues
- Identify examples of analytical models
- Secure funding for big data initiatives
- Review some of the top 10 companies that use big data solutions
- Become more familiar with big data as it relates to strategic planning

Big Data – The Sales Perspective

Big data allows salespeople to adopt data driven methodologies to target high value prospects rather than rely on relationships and other soft factors to target and close business deals. In this course, you will learn the difference between big data and data science. You will take a look at different algorithms and technology accelerators

Outline

- Describe what big data is
- Compare data science to big data
- Describe how big data entered into the public consciousness
- Find leads using big data
- Describe the different algorithms behind the systems we know
- Explore different software implementations for big sales data
- Describe the term The Internet of Things
- Identify different technologies that are accelerating the sales world
- Describe the most common barrier to technology adaptation in the work place
- Learn how different companies embraced the big data movement
- Become more familiar with big data from the sales perspective

Big Data - The Legal Perspective

By proactively dealing with privacy issues, organizations can safely leverage big data while retaining customers, and avoiding reputational harm, litigation, and regulatory scrutiny. In this course, you will examine privacy concerns, how data can be used ethically, and what to do about social media.

Outline

- Describe the privacy concerns over big data
- Learn some of the results that can occur when data is not properly protected
- Describe why the governance of the tools used to collect and analyze the data is important
- Learn how constantly changing laws and regulations are a challenge
- Describe the dilemma of predictive capabilities
- Learn how technology assisted reviews can reduce workload
- Describe the goals and the hazards of data monetization
- Describe the areas of concern to address before data collection
- Learn how transparency helps companies improve their customer relationships
- Explore the legal concerns related to social media sites used to recruit potential employees
- Understand how big data relates to legal concerns

Managing Big Data Operations

This course covers the challenges faced when rethinking data systems from the ground up. This course also covers in-depth topics such as monitoring tools, orchestrations, and performance modeling.

Outline

- Describe big data and where it's heading
- Explore the trends and the various industries that are exposed to big data operations
- Identify the technologies and advancements in big data
- Describe the process of monitoring big data repositories and predictive modeling
- Identify the various big data KPIs and how each can be used
- Describe the various performance issues and how to solve them using data monitoring
- Explore big data network monitoring operations and its importance
- List the various software and applications that can be used to provide big data orchestrations
- Learn the operations that integrate big data
- Explore the various processes of automating ETL jobs
- Describe big data trends and characteristics

Quality and Security of Big Data Operations

In this course, you will learn about big data testing and challenges that organizations face with big data operations.

Outline

- Describe key similarities and differences between data warehousing testing vs. big data testing
- Identify the various testing methods and strategies
- Learn the testing methods in ETL processes
- Describe the various methods in performance testing
- Demonstrate key strategies in big data testing
- Identify important tools in the Hadoop ecosystem that are available for testing big data
- Describe the main challenges businesses currently face with big data
- Describe data leakage and its prevention
- Identify the various ways to keep data clean, maintained, and secure
- Describe important characteristics of big data testing

Data Science Overview

Data science differentiates itself from academic statistics and application programming by drawing from a variety of disciplines. In this course, you will explore what it is to be a data scientist and learn what sets data science apart from other disciplines. The course prepares learners to navigate the foundational elements of data science.

Outline

- Define data science and what it is to be a data scientist
- Describe the data wrangling aspect of data science
- Describe the big data aspect of data science
- Describe the machine learning aspect of data science
- Use common data science terminology
- Identify ways to communicate results of data science
- Learn the steps in data science analysis
- Compare various tools and software libraries used for data science
- Exercise: Explore Your Data Science Needs

Data Gathering

To carry out data science, you need to gather data. Extracting, parsing, and scraping data from various sources, both internal and external, is a critical first part in the data science pipeline. In this course, you will explore examples of practical tools for data gathering

Outline

- Describe problems and software tools associated with data gathering
- Use cURL to gather data from the Web
- Use in2csv to convert spreadsheet data to CSV format
- Use agate to extract data from spreadsheets
- Use agate to extract tabular data from dbf files
- Extract data from particular tags in an HTML document
- Distinguish between metadata and data
- Work with metadata in HTTP headers
- Work with Linux log files
- Work with metadata in email headers
- Perform a secure shell connection to a remote server
- Copy remote data using a secure copy
- Synchronize data from a remote server
- Download an HTML file and explore table data

Data Filtering

Once data is gathered for data science, it is often in an unstructured or raw format. Data must be filtered for content and validity. In this course, you will explore examples of practical tools and techniques for data filtering.

Outline

- Identify common filtering techniques and tools
- Extract date elements from common date formats
- Parse content types in HTTP headers
- Use csvcut to filter CSV data
- Use sed to replace values in a text data stream
- Drop duplicate records from data
- Extract headers from a jpeg image
- Use pdfgrep to extract data from searchable pdf files
- Detect invalid or impossible data combinations
- Parse robots.txt from a web site to decide what should and should not be crawled orindexed
- Drop records from a CSV file based on date range

Data Transformation

Once data is filtered, the next step is to transform it into a usable format. In this course, you will explore examples of practical tools and techniques for data transformation.

Outline

- Convert CSV data to JSON format
- Convert XML data to JSON format
- Create SQL inserts from CSV data
- Extract CSV data from SQL
- Change delimiters in a CSV file from commas to tabs
- Convert basic date formats to standard ISO 8601 format
- Convert numeric formats within a CSV document
- Round floating point decimals to two places within a CSV document
- Use optical character recognition (OCR) to extract text from a jpeg image
- Use optical character recognition (OCR) to extract text from a pdf document
- Read various date formats and convert to standard compliant ISO 8601 format

Data Exploration

Once data is transformed into a useable format, the next step is to carry out preliminary data exploration. In this course, you will explore examples of practical tools and techniques for data exploration

Outline

- Use csvgrep to explore data in CSV data
- Use csvstat to explore values in CSV data
- Use csvsql to query CSV data like an SQL database
- Use gnuplot to quickly plot data on the command line
- Use wc to count words, characters, and lines within a text file
- Explore a subdirectory tree from the command line
- Use natural language processing to count word frequencies in a text document
- Take random samples from a list of records
- Find the top rows by value and percent in a data set
- Find repeated records in a data set
- Identify outliers using standard deviation
- Perform a word frequency count on a classic book from Project Gutenberg

Data Integration

Data integration, the last step in the data wrangling process, is where data is put into a useable and structured format for analysis. In this course, you will explore examples of practical tools and techniques for data integration.

Outline

- Use csvjoin to concatenate CSV data
- Use the cat function to concatenate separate logs into a single file
- Sort lines in a text file
- Merge separate xml files into a single schema
- Aggregate data from a CSV file into a table of summarized values
- Normalize data from unstructured sources
- Denormalize data from a structured source
- Use pivot tables to cross tabulate data
- Insert missing values in a data set
- Use csvjoin to merge two compatible CSV documents into one

Data Analysis Concepts

There are many software and programming tools available to data scientists. In order to apply those tools effectively, you must understand the underlying concepts. In this course, you will explore the underlying data analysis concepts needed to employ software and programming tools effectively.

Outline

- Perform basic math operations required by data scientists
- Perform basic vector math operations required by data scientists
- Perform basic matrix math operations required by data scientists
- Perform a matrix decomposition
- Identify different forms of data
- Describe probability in terms of events and sample space size
- Describe basic properties of outcomes
- Apply probability rules in calculation
- Identify common continuous probability distributions
- Identify common discrete probability distributions
- Apply Baye's theorem and describe how it is used in email spam algorithms
- Apply random sampling to A/B tests
- Identify and describe various statistical measures
- Describe the difference between an unbiased and biased estimator
- Describe sampling distributions and recognize the central limit theorem
- Define confidence intervals and work with margins of error
- Carry out hypothesis tests and work with p-values
- Apply the chi-square test for categorical values
- Identify the given data set descriptions by their types

Data Classification and Machine Learning

Machine learning is the area of data science that uses techniques to create models from data without being explicitly programmed. In this course, you will explore the conceptual elements of various machine learning techniques.

Outline

- Identify problems in which supervised learning techniques apply
- Identify problems in which unsupervised learning techniques apply
- Apply linear regression to machine learning problems
- Identify predictors in machine learning
- Apply logistic regression to machine learning problems
- Describe the use of dummy variables
- Use naive Baye's classification techniques
- Work with decision trees
- Describe k-means clustering
- Define cluster validation
- Define principal component analysis
- Describe machine learning errors
- Describe underfitting
- Describe overfitting
- Apply k-folds cross validation
- Describe fall-forward and back-propagation in neural networks
- Describe SVMs and their use
- Choose the appropriate machine learning method for the given example problems

Data Communication and Visualization

The final step in the data science pipeline is to communicate the results or findings. In this course, you will explore communication and visualization concepts needed by data scientists.

Outline

- Choose appropriate visualization techniques
- Describe the difference between correlation and causation
- Define Simpson's paradox
- Communicate data science results informally
- Communicate data science results formally
- Implement strategies for effective data communication
- Use scatter plots
- Use line graphs
- Use bar charts
- Use histograms
- Use box plots
- Create a network visualization
- Create a bubble plot
- Create an interactive plot
- Find an appropriate data set to visually represent using a scatter plot and plot it

Learn more at
www.hpe.com/ww/digitallearner

Interested in purchase of this Content Pack as a stand-alone WBT? Contact Us for information on purchasing this Content Pack for individual use.

www.hpe.com/ww/digitallearner-contentpack

Follow us:

CP009 A.00 , January 2019

Hewlett Packard Enterprise