

Oracle Cloud Infrastructure Data Science Professional: Hands-on Workshop

Student Guide
D1111082GC10

Learn more from Oracle University at education.oracle.com



Copyright © 2025, Oracle and/or its affiliates.

Disclaimer

This document contains proprietary information and is protected by copyright and other intellectual property laws. The document may not be modified or altered in any way. Except where your use constitutes "fair use" under copyright law, you may not use, share, download, upload, copy, print, display, perform, reproduce, publish, license, post, transmit, or distribute this document in whole or in part without the express authorization of Oracle.

The information contained in this document is subject to change without notice and is not warranted to be error-free. If you find any errors, please report them to us in writing.

Restricted Rights Notice

If this documentation is delivered to the United States Government or anyone using the documentation on behalf of the United States Government, the following notice is applicable:

U.S. GOVERNMENT END USERS: Oracle programs (including any operating system, integrated software, any programs embedded, installed or activated on delivered hardware, and modifications of such programs) and Oracle computer documentation or other Oracle data delivered to or accessed by U.S. Government end users are "commercial computer software" or "commercial computer software documentation" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, reproduction, duplication, release, display, disclosure, modification, preparation of derivative works, and/or adaptation of i) Oracle programs (including any operating system, integrated software, any programs embedded, installed or activated on delivered hardware, and modifications of such programs), ii) Oracle computer documentation and/or iii) other Oracle data, is subject to the rights and limitations specified in the license contained in the applicable contract. The terms governing the U.S. Government's use of Oracle cloud services are defined by the applicable contract for such services. No other rights are granted to the U.S. Government.

Trademark Notice

Oracle®, Java, MySQL, and NetSuite are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Inside are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Epyc, and the AMD logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group.

Third-Party Content, Products, and Services Disclaimer

This documentation may provide access to or information about content, products, and services from third parties. Oracle Corporation and its affiliates are not responsible for and expressly disclaim all warranties of any kind with respect to third-party content, products, and services unless otherwise set forth in an applicable agreement between you and Oracle. Oracle Corporation and its affiliates will not be responsible for any loss, costs, or damages incurred due to your access to or use of third-party content, products, or services, except as set forth in an applicable agreement between you and Oracle.

1006032025

Table of Contents

Module 1: Welcome to Data Science	19
Lesson 1: Data Science Professional Course Overview	19
Course Speakers	20
Course Contributors	21
For whom is this course intended?	22
Prerequisites	23
What does the Data Science exam validate?	24
Course Outline	25
Data Science Demos and Labs	26
Get the Most Out of This Course	28
Ratings and Feedback	29
Module 2: Introduction and Configuration	30
Lesson 1: Data Science: Introduction	31
Data Science and Machine Learning (ML) in History	32
Data Science and Machine Learning Today	33
Importance of Data Science and AI	34
What is Oracle AI?	35
OCI Services That Support AI and ML	36
What is Oracle Cloud Infrastructure Data Science?	37
Core Principles of Oracle Cloud Infrastructure Data Science	38
OCI Data Science Details	39
Data Science Features and Terminology	40
Ways to Access Oracle Cloud Infrastructure Data Science	41
Where to Find Data Science	42
Lesson 2: ADS SDK Overview	43

Accelerated Data Science (ADS) Software Development Kit (SDK)	44
Ways to Access ADS SDK	45
ADS SDK: Features	46
Data Visualization	48
Feature Engineering	49
Model Training	50
Model Evaluations	51
Model Interpretation and Explainability	52
Model Deployment	53
Lesson 3: Tenancy Configuration Basics	54
Tenancy Configuration Concepts	55
How Data Science Components Work Together	56
Compartments	57
Creating a Compartment	58
User Groups	59
Dynamic Groups	60
Dynamic Groups: Matching Rules	61
Policies	62
Policy Syntax	63
Policy Basics: Verb	64
Policy Basics: Resource Type	65
Required Data Science Policies	66
More Required Data Science Policies	67
Optional Data Science Policies	68
Optional Policies for Data Science–Related Services	69
Demo	70
Lesson 4: Configure a Tenancy with OCI Resource Manager	71
Automatic Configuration and the Data Science Service Template	72
What does the Data Science Service template create?	73

Running the Oracle Resource Manager (ORM) Stack	74
Accessing the Terraform Script	75
Demo	76
Lesson 5: Networking for Data Science	77
Cloud Networking Components: Overview	79
Data Science and Networking Connectivity	80
Default Networking	81
Custom Networking	82
Demo	83
Lesson 6: Authenticate to OCI APIs	84
Importance of Authentication in Data Science	85
Authenticating Different Interfaces	86
What are resource principals?	87
Resource Principals and Data Science Service Authentication	88
Resource Principals via Different Interfaces	89
OCI Configuration File	90
OCI Config File Format and Example	91
Demo	92
Module 3: Workspace Design and Setup	93
Lesson 1: Projects	94
What are projects?	95
Create Projects	96
Viewing, Editing, and Deleting Projects	97
Demo	98
Lesson 2: Notebook Sessions	99
What are notebook sessions?	100
Create Notebook Sessions from the Console	101
Viewing, Editing, and Deleting Notebook Sessions	102

Activating and Deactivating Notebook Sessions	103
Notebook Session Metrics	104
Demo	105
Lesson 3: How to Work with JupyterLab	106
Jupyter Lab: Overview	107
JupyterLab Interfaces	108
Features of JupyterLab	109
Demo	110
Lesson 4: Conda Environments: Overview	111
What is a conda environment?	112
Benefits of Conda Environments	113
Environment Explorer	114
Data Science Conda Environments	115
Published Conda Environments	116
Installed Environments	117
Demo	118
Lesson 5: Data Science Conda Environments	119
Types of Conda Environments	121
Conda Environment Families	122
Conda Environment Naming Conventions	123
Computer Vision	124
Data Exploration and Manipulation	125
General Machine Learning	126
Natural Language Processing	127
ONNX	128
Oracle Database	129
PyTorch	130
PySpark	131
TensorFlow	132

Lesson 6: Manage Conda Environments	133
Conda Environments: Recap	134
Conda Environment Functionality	135
Browse	136
Search	137
Install	138
Clone	139
Modify	140
Publish	141
Delete	142
Create from YAML	143
Demo	144
Lesson 7: OCI Vault: Introduction	145
Why Is the OCI Vault Important in Data Science?	146
OCI Vault	147
Vaults	148
Keys	149
Master and Data Encryption Keys	150
Rotating Keys	151
What Are Secrets?	152
Secrets	153
Lesson 8: Using OCI Vault in OCI Data Science	154
Encryption Using Oracle-Managed Keys	155
Encryption Using Customer-Managed Keys	156
Encrypt Data in the Vault	157
Encode the Secret	158
Store Secret in the Vault	159
Retrieve the Secret from the Vault	160
Using OCI Vault with ADS	161

MySQLDBSecretKeeper	162
Lesson 9: Code Repositories (Git)	163
What is a version control system?	164
Code Repositories	165
Centralized Versus Distributed Version Control	166
Using Git in Data Science Workflows	167
OCI Git Extension	168
Git Terminology	169
Git: Flow Diagram	170
OCI Code Repository	171
Working with a Remote Repo: GitHub	172
Commands in Git	173
Demo	174

Module 4: Machine Learning Lifecycle **175**

Lesson 1: ML Lifecycle: Overview	176
Machine Learning Lifecycle	177
Data Access and Collection	178
Data Exploration and Preparation	179
Feature Exploration	180
Feature Engineering	181
Modeling	182
Validation	183
Model Deployment	184
Model Monitoring	185
Lesson 2: Access Data	186
Machine Learning Life Cycle: Data Access	187
Bring Data to OCI	188
Why Need Data?	189

Data Collection	190
Access Data from Some Common Sources	191
Access Data from Object Storage	192
Access Data from Local Storage	193
Access Data from Oracle Autonomous Database	194
Access Data from MySQL	196
Access Data from Amazon S3	197
Access Data Using HTTP and HTTPs Endpoints	198
Access Data Using DatasetBrowser	199
Access Data Using PyArrow	200
Data Types	201
Supported Sources/Formats by Oracle ADS	202
Demo	203
Lesson 3: Data Preprocessing	204
Machine Learning Life Cycle:Data Transformations	205
Preprocessing Data	206
Combine and Clean Data	207
Data Imputation	208
Dummy Variables	209
Outlier Detection	210
Feature Scaling	211
Dimensionality	212
Text Data	213
ADS Data Transformations	214
Apply Automated Transformations	215
Optimization Features	216
suggest_recommendations(): Employee Attrition Data Set	217
auto_transform(): Employee Attrition Data Set	218
visualize_transforms(): Employee Attrition Data Set	219

Split Data Set into Train, Validation, and Test Data	220
Demo	221
Lesson 4: Introduction to Feature Types	222
What are feature types?	223
Feature Type: Example	224
Types of Feature Types	225
Exploratory Data Analysis (EDA)	226
Multiple Inheritance	227
Feature Type Selection	228
Feature Type Count	229
Correlation Tables	230
Correlation Plots	231
Lesson 5: Custom Feature Types	232
Feature Statistics	233
Customized Feature Statistics	234
Feature Plots	235
Custom Feature Plot	236
Feature Type Warnings	237
Creating Warnings	238
Defining Warning Handler and Registering Warnings	239
Feature Type Validators	240
Create Validators	241
Types of Feature Type Validators	242
Creating a Custom Feature Type	243
Lesson 6: Data Visualization	244
Machine Learning Life Cycle: Data Visualization	245
Data Visualization Tool	246
ADS Smart Visualization Tool	248
Automatic Visualization	249

Automatic Visualization: Methods	250
corr() Method	251
show_in_notebook() Method	252
plot() Method	253
plot(): Examples	254
Feature Type	255
feature_plot() Method	256
Customized Visualization	257
Seaborn	258
Matplotlib	259
GIS	260
Machine Learning Life Cycle: Data Profiling	261
Data Profiling	262
Simple Statements to Generate Data Set Diagnostics	263
Simple Python Code for Data and Feature Profiling	264
Lesson 7: Model Training	265
Machine Learning Life Cycle: Model Training	266
Model Training	267
Training Process	268
Model Training Libraries	269
Ways to Train	270
Lesson 8: Oracle AutoML: Introduction	271
Machine Learning Life Cycle: AutoML	272
AutoML: What and Why	273
AutoML Approaches	274
Oracle Automated Machine Learning (Oracle AutoML)	275
Benefits	276
Oracle AutoML Workflow	277
AutoML Pipeline	278

Algorithm Selection	279
How Algorithms Are Selected	280
Adaptive Sampling	281
How Is Adaptive Sampling Done?	282
Feature Selection	283
How Is Feature Selection Done?	284
Hyperparameter Tuning	285
How Is Hyperparameter Tuning Done?	286
Building with Oracle AutoML	287
Oracle AutoML: Time Budget	289
Oracle AutoML: Minimum Feature List	290
Demo	291
Lesson 9: Hyperparameter Tuning: ADSTuner	292
ADSTuner Search Spaces	293
Tuning Process	297
Custom Search Space	298
Lesson 10: Model Evaluation	299
The Machine Learning Lifecycle: Model Evaluation	300
Model Evaluation	301
Benefits	302
Types of ADS Evaluators	303
Binary Classification Metrics	304
Binary Classification Charts and Plots	308
Multiclass Classification Metrics	309
Multiclass Classification Charts and Plots	313
Regression Metrics	314
Regression Charts and Plots	316
Lesson 11: Model Explanations: Global Explainer	317
The Machine Learning Lifecycle: Model Explanations	318

Model Explanations	319
Explanation Types	320
Global Explainers	321
Global Explainers: Feature Permutation Importance Explanations	322
ADS Model Explanation: Visualization	323
Global Explainers: Feature Permutation Importance Explanations	324
Global Explainers: Feature Dependence Explanations	327
Feature Dependence Explanations	328
Global Explainers: Feature Dependence Explanations	329
Global Explainers: Accumulated Local Effects (ALE)	331
Lesson 12: Model Explanations: Local and WhatIf	334
Local Explainers	335
Compute a Local Explanation	336
Local Explainers	337
Model Section	338
Explainer Section	339
Explanations Section	340
WhatIf Explainer	341
Explore Sample	342
Explore Predictions	343
Lesson 13: Model Catalog: Overview	344
Machine Learning Life Cycle: Model Catalog	345
Model Catalog: Overview	346
Entry Point to the Model Catalog	347
Model Artifacts	348
Preparing Artifacts	349
Components of a Model Artifact	350
Model Artifact Directory	351
Custom Logic: Score.py	352

Deployment Configuration File	353
Additional Artifacts	354
Preparing Model Metadata	355
Model Catalog Documentation	356
Model Input and Output Schema	357
Input Schema: Example	358
Output Schema: Example	359
Model Provenance	360
Model Introspection Tests	361
Model Taxonomy	362
Types of Model Taxonomy	363
Lesson 14: Model Serialization	364
Serialization and Deserialization	365
Saving Models by Using ADS SDK: Model Serialization and Model Save	366
Saving Models by Using ADS SDK: Generic Model Approach	367
Saving Models to the Model Catalog	368
Managing Models in the Model Catalog	369
Viewing Models	370
Editing Models	371
Other Operations	372
Lesson 15: Model Deployment	373
Machine Learning Life Cycle: Model Deployment	374
Model Deployment	375
Model Deployment Architecture	376
Key Components	377
Create and Invoke Model Deployment	378
Model Deployment Components	379
Creating a Model Deployment from the UI Console	380
Creating a Model Deployment from ADS	381

Creating a Model Deployment from the CLI	382
Generating Predictions with the Deployed Model	383
Invoking Your Model	384
Managing a Model Deployment	385
Deactivating or Reactivating	386
Monitoring a Model Deployment Logs	387
Monitoring a Model Deployment Metrics	388
Demo	389
Lesson 16: LLM Training & LangChain Integration	390
LLM Training in OCI Data Science	391
Fine-Tuning a Pre-Trained Model	392
LangChain Integration	393
Demo	394
Lesson 17: OCI Data Science Operators	395
Data Science Operators	396
Types of Data Science Operators	397
Demo - Operators	398
Lesson 18: OCI Data Science AI Quick Actions	399
AI Quick Actions	400
Demo – AI Quick Actions	401
Module 5: MLOps Practices	402
Lesson 1: MLOps Architecture	403
What is MLOps?	406
Continuous Practices in MLOps	407
Why is MLOps important?	408
Automation in MLOps	410
MLOps Architecture in OCI	411
Lesson 2: Oracle Cloud Infrastructure Data Science Jobs	412

Jobs Service	413
Jobs in MLOps Life Cycle	414
OCI Jobs: Benefits	415
Jobs Versus Job Runs	416
Data Science Jobs	417
Data Science Job Runs	419
Jobs Life Cycle	420
Ways to Run Jobs	421
Oracle Cloud Access for Jobs	422
External Access for Jobs	423
Batch Inference	424
Mini Batch	425
Distributed Batch	426
Batch Comparison Table	427
Scaling	428
Lesson 3: Jobs Monitoring and Logging	429
Jobs Life Cycle	430
Monitoring Service	431
Job Run Metrics	432
Job Run Logs	433
OCI Job Events and Services	434
Events, Rules, and Actions	435
Lesson 4: Data Science Pipeline	436
OCI Data Science Pipeline	437
Pipeline	438
Pipeline Step	439
Pipeline Life Cycle and Runs	440
Pipeline: Demo Scenario	441
Lesson 5: Model Deployment: Autoscaling	442

Data Science Model Scaling Issues	443
Autoscaling - Data Science Model Deployment	444
Metric-Based Autoscaling	445
Metrics	447
Module 6: Related OCI Services	448
Lesson 1: Spark Applications, Data Flow,and Data Science	449
Introduction to Oracle Cloud Infrastructure Data Flow	450
Data Flow: Overview	451
Data Flow: Components	452
Data Flow: Capabilities	453
Data Flow: Security	454
Using Data Flow and Spark with AI/ML Workloads	455
Spark Application Configuration	456
Integration with Data Science	457
Create and Manage Spark Applications by Using Data Flow and Data Science	458
Prerequisites for Data Science	459
Prerequisites for Data Flow	460
Build and Train ML Models with Data Flow	461
Best Practices for Developing Spark Applications from a Notebook Session Environment	462
Learn More About ADS	463
Lesson 2: Oracle Open Data	464
Oracle Open Data: Overview	465
Oracle Open Data: Benefits	467
Accessing Oracle Open Data	468
Lesson 3: Oracle Cloud Infrastructure Data Labeling	469
What Is Data Labeling?	470
Who Uses Data Labeling?	472
How Industries Use Data Labeling	473

Data Labeling Within AI/ML Life Cycle474

Data Labeling Use Cases475

Scenario: Data Labeling for AI476

Data Label Types477

Data Labeling Integrations478