Anaconda Training
Data Science Foundations

At the conclusion of this 4-day course you will have a solid understanding of how Anaconda Enterprise and the Python ecosystem work together to help you perform quantitative and qualitative analyses. This course covers the core libraries for data processing and analysis, statistical computation, and an overview of machine learning. You'll learn how to access tabular data stored in various file formats along with data stored in relational databases and distributed storage systems.

Preparation

Students will connect to an Anaconda Enterprise instance maintained by the Anaconda Training department.

This course is meant for all levels of Python and Data Science backgrounds. However, these DataCamp courses are very useful to provide background preparation.

- Intro to Python for Data Science
- Intermediate Python for Data Science

Curriculum

See following pages for a detailed outline of each section.

- Getting Started with Anaconda Enterprise (1 day)
- Essential Pandas (1 day)
- Access Big Data with Anaconda Enterprise (1/2 day)
- Statistical Modeling and Analysis (1/2 day)
- Machine Learning with Scikit-Learn (1 day)

About Anaconda, Inc.

With more than 13 million users, Anaconda is the world's most popular data science platform and the foundation of modern machine learning. Anaconda Enterprise delivers data science and machine learning at speed and scale, unleashing the full potential of our customers' data science and machine learning initiatives.

ANACONDA

Day 1

Getting Started with Anaconda Enterprise

Duration: 1 day

Anaconda Enterprise Platform

- **Log into Anaconda Enterprise**
  - Overview of *Projects, Deployments, and Channels*
- **Working with projects**
  - Install packages and environments
  - Commit and share projects
- **Working in Sessions**
  - Jupyter Notebook
  - JupyterLab
  - Zeppelin

Review Core Python Syntax

- **Python concepts & constructs**
  - Python object model & modules
  - Flow control
  - Functions
- **Data structures**
  - Methods/attributes for common data structures
  - Idioms for slicing, indexing, iteration, and comprehension
  - Files & file I/O: common methods & idioms, context managers

ANACONDA

Day 2

Essential Pandas

Duration: 1 day

- **Data Exploration**
  reading data sources
  selections & summary statistics
  data analysis methods
  filtering data using logical conditions
  plotting in Jupyter notebooks
- **Data Formats**
  Flat text files: CSV, TSV
  Binary Formats: HDF5, SAS, Excel, databases
  Structured files: JSON
- **Data Processing**
  using vectorized operations
  transforming strings & datetimes
- **Time series**
  creating & using datetime indexes
  resampling time series
  using rolling windows
- **Grouping**
  Grouping data by column values
  resampling time series
- **Merging & Joining DataFrames**
  appending & concatenating DataFrames
  joins/merges on Index
  merges on multiple columns
  merging with missing values

ANACONDA

Day 3

Access Big Data with Anaconda Enterprise

Duration: 1/2 day

- **Accessing Databases**
  Connect with SQLAlchemy
  Execute queries
  Retrieve results
- **PySpark**
  HIVE tables and storage
  DataFrame objects
  Processing Hadoop files

Statistical Modeling and Analysis

Duration: 1/2 day

- **Overview of scipy.stats**
- **Sampling empirical distributions**
- **Construct PDF and CDF**
- **Hypothesis testing**
- **Statsmodels**
  linear regression
  regression analysis
  logistic regression
  building design matrices with R-like equations

ANACONDA

Day 4

Machine Learning with Scikit-Learn

Duration: 1 day

Supervised Learning

- **Model Training and validation**
- **Regression problems**
  - Linear models
  - Support vector machines
  - Decision trees
- **Classification problems**
  - K-nearest neighbors classification
  - Naive Bayes classification
  - Support vector machines
  - Decision trees and ensemble strategies
- **Model building and scoring**
  - Scoring functions & cross-validation
  - Feature selection
  - Feature extraction
- **Pipelines**
- **Grid search parameter optimization**

Unsupervised Learning

- **Feature extraction**
- **Clustering problems**
  - K-means & hierarchical clustering
  - DBScan
- **Dimensionality reduction**
  - PCA, LDA, NMF
- **Detection & treatment of outliers**