National Czech Programme | eosc

Základy Pythonu pro data stewardy

Tomáš Martinovič

03.06.2025

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

National Czech Programme | eosc

Python Environment Management - Rye

- Modern Python package management tool
- Focus on reproducible environments for data projects
- A modern Python package management tool
- Focuses on dependency management and virtual environments
- Similar to tools like Poetry, but with additional features
- Helps maintain consistent and reproducible Python environments

National Czech Programme | eosc

Installing Rye

On macOS/Linux:

```
curl -sSf https://rye-up.com/get | bash
```

On Windows:

- Official guide - https://rye.astral.sh/guide/installation/
- Video tutorial - https://youtu.be/xJdIKQ84s8E

After installation, restart your terminal or run:

```
source "$HOME/.rye/env"
```

National Czech Programme | eosc

Rye Basic Commands

Create a new project:

```
rye init my-project
cd my-project
```

Add dependencies:

```
rye add pandas numpy
```

Install dependencies:

```
rye sync
```

Execute commands:

```
rye run python script.py
```

Activate virtual environment (direct interaction with environment):

```
source .venv/bin/activate  # Unix
.venv\Scripts\activate  # Windows
```

National Czech Programme | eosc

Rye for Data Projects

Benefits for data stewards:

- Reproducible environments across team members
- Lockfiles for exact dependency versions
- Easy switching between Python versions
- Simplified package management workflow
- Compatible with requirements.txt and pyproject.toml

National Czech Programme | eosc

Essential Python Libraries
A toolkit for modern data stewardship

- Python has become the go-to language for data stewards and engineers
- Rich ecosystem of libraries for data processing, analysis, and management
- This presentation covers essential libraries for data stewards

National Czech Programme | eosc

1. Pandas

The backbone of data manipulation and analysis

- Data structures like DataFrames for structured data
- Key features:
  - Data wrangling and cleaning
  - Data aggregation and grouping
  - Merging and joining datasets
  - Handling missing data

2. NumPy

Foundation for numerical computations

- Support for arrays, matrices, and mathematical functions
- Key features:
  - Efficient array computations
  - Linear algebra and statistical operations
  - Integration with other libraries
- Essential foundation for scientific computing in Python

3. SQLAlchemy

Powerful ORM for database interactions

- Python interface to SQL databases
- Key features:
  - Database-agnostic SQL querying
  - ORM capabilities
  - Database schema generation

4. Boto3

AWS SDK for Python

- Interact with AWS services programmatically
- Key features:
  - Access to AWS services
  - S3 file management
  - EC2 and RDS control

5. Requests

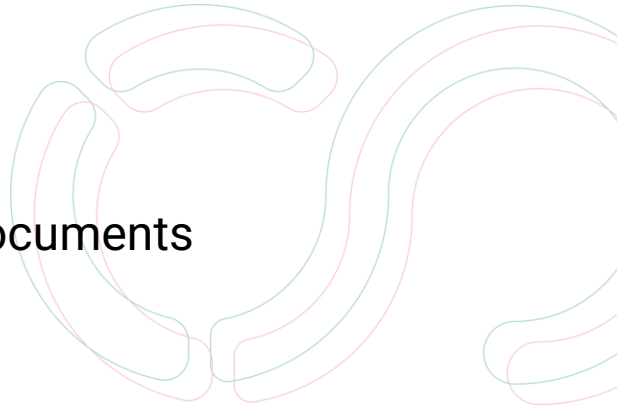
HTTP library for API interactions

- Simple yet powerful for web data retrieval
- Key features:
  - Sending HTTP requests (GET, POST, etc.)
  - Handling JSON responses
  - Supports authentication and sessions

6. Beautiful Soup

Web scraping library for data acquisition

- Extract data from HTML and XML files
- Key features:
  - Parses HTML and XML documents
  - Navigates parse trees to extract data
  - Supports different parsers
- Essential for acquiring data from websites and HTML documents



National Czech Programme | eosc

7. Great Expectations

Data validation framework for data quality

- Ensure data quality and integrity throughout the data lifecycle
- Key features:
  - Data validation and profiling
  - Data documentation
  - Integration with data pipelines
- Essential for data stewards responsible for data quality

National Czech Programme | eosc

8. Dask

Scale Python for larger-than-memory data

- Extends Python’s parallelism
- Key features:
  - Parallel computing
  - Scales Python code across multiple cores
  - Handles large datasets

9. PySpark

Python API for Apache Spark

- Essential for big data processing
- Key features:
  - Distributed data processing
  - Integration with Hadoop and HDFS
  - Machine learning capabilities with MLlib

National Czech Programme | eosc

10. Airflow

Workflow orchestration tool

- Define, schedule, and monitor workflows
- Key features:
  - Workflow automation
  - Task scheduling and monitoring
  - Dynamic pipeline generation

National Czech Programme | eosc

Conclusion

Building a comprehensive data toolkit

- These libraries form a comprehensive toolkit for data stewards
  - Acquisition (Requests, Beautiful Soup, Boto3)
  - Processing (Pandas, NumPy, PySpark, Dask)
  - Quality Check (Great Expectations)
  - Storage (SQLAlchemy, Boto3)
  - Orchestration (Airflow)

National Czech Programme | eosc

Let’s get our hands dirty!

Presentation and code on opencode.it4i

https://opencode.it4i.eu/eosc-trainings/python_data_stewards_june_2025

National Czech Programme | eosc

Thank you for attention

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY