

ETL mit Python

Daten aus unterschiedlichen Quellen zuverlässig extrahieren, transformieren und laden

01 INTRO

Viele Datenanalysen beginnen nicht im Notebook, sondern mit der Frage, wie Daten zuverlässig aus unterschiedlichen Systemen zusammengeführt und für die weitere Verarbeitung bereitgestellt werden. ETL-Prozesse bilden dafür eine zentrale technische Grundlage.

Der Kurs behandelt die Entwicklung reproduzierbarer ETL-Prozesse mit Python. Neben Datenimport und Transformation stehen robuste Verarbeitung, Fehlerbehandlung, Validierung, Logging, Wiederanlauf und Automatisierung im Mittelpunkt. Damit schlägt der Kurs bewusst die Brücke zwischen Data Analytics und Software Engineering.

DAUER

3 Tage

FORMAT

Präsenz · Live Online · Inhouse

VORKENNTNISSE

Praktische Python-Grundkenntnisse; SQL-Grundkenntnisse sind hilfreich

02 TRAININGSPROGRAMM

Inhalte

ETL und Datenpipelines

- Extract, Transform, Load
- ETL versus ELT
- Batch-Verarbeitung
- typische Datenarchitekturen
- Quellen und Ziele
- Anforderungen an produktive Pipelines

ETL-Projekte mit Python strukturieren

- Projektaufbau
- Konfiguration
- Trennung der Verarbeitungsschritte
- wiederverwendbare Komponenten
- Abhängigkeiten
- Entwicklungs- und Laufzeitumgebung

Dateien als Datenquellen

- CSV
- Excel
- JSON
- strukturierte Textdateien
- große Dateien
- Encoding
- fehlerhafte Eingabedaten
- Dateiverarbeitung automatisieren

Datenbanken als Quelle und Ziel

- Verbindungen
- SQL-Abfragen
- Lesen und Schreiben
- Transaktionen
- Bulk Operations
- Datentypen
- sichere Parametrisierung

REST APIs als Datenquelle

- HTTP-Grundlagen
- Requests
- JSON
- Authentifizierung
- Pagination
- Rate Limits
- Fehlerbehandlung
- Retries

Daten transformieren

- Mapping
- Normalisierung
- Datentypen
- Berechnungen
- Filter
- Aggregationen
- Joins
- Referenzdaten
- Transformation mit Pandas und Python

Datenqualität und Validierung

- Eingangsdaten prüfen
- Schemas
- Pflichtfelder
- Wertebereiche
- fachliche Regeln
- fehlerhafte Datensätze
- Qualitätsmetriken

Ladeprozesse entwickeln

- Insert
- Update
- Upsert
- vollständige und inkrementelle Loads
- Schlüssel
- Duplikate
- Ladezustände
- atomare Verarbeitung

Fehlerbehandlung und Wiederanlauf

- technische und fachliche Fehler
- Exceptions
- Retry-Strategien
- partielle Verarbeitung
- Checkpoints
- Idempotenz
- Wiederaufnahme abgebrochener Läufe

Logging und Observability

- strukturiertes Logging
- Laufzeiten
- Mengen
- Fehlerraten
- Datenqualitätsinformationen
- technische Diagnose
- Auditierbarkeit

Automatisierung und Betrieb

- Kommandozeilenprogramme
- Konfiguration
- Scheduling
- Container
- Tests
- Deployment
- Secrets
- Monitoring
- CI/CD-Grundlagen für ETL

Praktisches ETL-Projekt

- mehrere Datenquellen anbinden
- Daten extrahieren
- transformieren und validieren
- Zielsystem beladen
- Fehlerfälle behandeln
- Logging integrieren
- Pipeline wiederholbar ausführen

03 DURCHFÜHRUNG

Rahmen und Organisation

Dauer	3 Tage
Durchführung	Präsenz oder Live Online, vorzugsweise als Inhouse-Schulung
Schwerpunkt	Inhouse-Schulungen für Unternehmen und Teams
Sprache	Deutsch, englische Fachbegriffe und Dokumentation
Unterlagen	Kursunterlagen, Beispiele und Übungsaufgaben
Technik	Eigener Rechner mit Python, Entwicklungsumgebung und Zugang zu einer Datenbank

Inhouse-Schulungen

Vorhandene Datenquellen, Zielsysteme und bestehende Verarbeitungsprozesse können in die Übungen einbezogen werden.

04 ZIELGRUPPE

Für wen ist der Kurs gedacht?

Python-Entwickler, Data Analysts, Data Engineers und technische Mitarbeiter, die Datenpipelines mit Python entwickeln oder bestehende Datenverarbeitungsprozesse automatisieren möchten.

Voraussetzungen

Praktische Python-Grundkenntnisse. Grundkenntnisse in SQL und relationalen Datenbanken sind hilfreich.

05 LERNZIELE

Was Sie nach dem Kurs können

Nach dem Kurs können die Teilnehmer ETL-Pipelines mit Python strukturieren, unterschiedliche Datenquellen anbinden, Transformationen und Qualitätsprüfungen implementieren sowie robuste, automatisierbare Ladeprozesse entwickeln.

06 FAQ

Häufige Fragen

Was unterscheidet den Kurs von Datenaufbereitung und Data Cleaning?

Der Cleaning-Kurs konzentriert sich auf Qualität und Bereinigung der Daten selbst. Hier steht der gesamte technische Datenfluss im Mittelpunkt: von der Quelle über Transformation und Validierung bis zum Zielsystem.

Wird ein bestimmtes ETL-Framework verwendet?

Der Kurs vermittelt zunächst die zugrunde liegenden Engineering-Prinzipien mit Python. Frameworks können ergänzend betrachtet werden, ohne den Kurs an ein bestimmtes Produkt zu binden.

Ist der Kurs auch für Data Analysts geeignet?

Ja, wenn grundlegende Python-Kenntnisse vorhanden sind und Datenverarbeitungsprozesse automatisiert oder produktionsnäher aufgebaut werden sollen.

UC IT Service · Ulrich Cuber · kontakt@uc-it.de
<https://www.uc-it.de/coaching/etl-mit-python/>