Python ETL Projects

A community-driven, ceramics haven

By Dhandapani Yedappalli Krishnamurthi Nov 4, 2025

ETL projects using Spark, PySpark, and Python, covering real-life data workflows and pipeline best practices.

ETL Project Repositories

- ETL-PySpark (rvilla87/ETL-PySpark): Real-world ETL project using PySpark, Spark SQL, and Hadoop Distributed File System. Demonstrates CSV to Parquet transformations, performance testing, and integration with HDFS.
- pyspark-example-project (AlexIoannides/pyspark-example-project): Focuses on best practices for structuring ETL jobs with PySpark. Includes testable, modular code, configuration management, dependency handling, and meaningful tests for ETL jobs.
- pysetl (JhossePaul/pysetl): A Python Spark ETL framework designed to improve readability, maintainability, and type safety for large PySpark ETL pipelines. Useful for modeling complex data workflows.
- Advanced ETL with Databricks and PySpark (JANHMS/Advanced-ETL-Azure-Databricks-Pyspark): Shows large-scale ETL workflows on Azure Databricks using PySpark and Data Lake Storage.

- Metorikku (YotpoLtd/metorikku): A lightweight ETL framework for Spark, supporting YAML-based configuration files for defining ETL workflows and a variety of input/output sources (CSV, JSON, Parquet, JDBC, Kafka, Cassandra, Elasticsearch, etc.).

Sample Repository Table

Repository Name	**Technologies**	**Features**
rvilla87/ETL-PySpark	PySpark, Spark	HDFS, Spark SQL, Parquet format, real CSV demo
AlexIoannides/pyspark-example-project	PySpark, Python	Project structure, config files, unit tests, modular code
JhossePaul/pysetl	PySpark, Python	Framework for large, type-safe ETL pipelines
JANHMS/Advanced-ETL-Azure-Databricks-Pyspark	PySpark, Databricks	Azure Data Lake, complex transformations
YotpoLtd/metorikku	Spark, Scala, YAML	Config-driven ETL for Spark, multiple formats, easy pipeline

Each repository has concrete code and documentation for building real ETL pipelines—from data extraction to transformation and loading—using Spark, PySpark, and Python.