# Git Repo

## Top Scala ETL & Data Pipeline Repositories

1. **vbounyasit/MyDataFramework** – A modular Scala ETL framework built on Apache Spark for clean data engineering workflows. GitHub
2. **SETL-Framework/setl** – Structured, modular ETL pipelines in Scala with Spark; supports clear stages and repository abstractions. GitHub
3. **mattlianje/etl4s** – A lightweight, type-safe, functional ETL library for Scala—declarative and monadic. GitHub
4. **agmenc/planet7** – Fast ETL and reconciliation tool for CSVs: load, rename, diff, and validate. GitHub
5. **yennanliu/spark-etl-pipeline** – Demo of varied Spark ETL stream and batch pipelines in Scala. GitHub
6. **skalskibukowa/Project-Spark-Scala-ETL** – ETL project extracting from CSV/PostgreSQL and loading to CSV, Parquet, PostgreSQL. GitHub
7. **qwshen/spark-etl-framework** – Pipeline-pattern ETL framework using Spark-SQL: modular actor-based flows. GitHub
8. **geotrellis/spark-etl** – ETL split into load, tile, save—built with GeoTrellis and Spark for geospatial data. GitHub
9. **GoogleCloudPlatform/dataproc-scala-examples** – Scala Spark ETL examples on GCP with Kafka, BigQuery, GCS. GitHub

---

## Libraries & Tools to Learn Scala Through ETL and Ingestion

10. **awesome-scala (uhub/awesome-scala)** – Curated list of Scala frameworks and tools (Spark, Scalding, Deequ, etc.). GitHub
11. **best-of-scala (stkeky/best-of-scala)** – A ranked catalog of top Scala tools across categories like data handling and file processing. GitHub

12. **Finagle (twitter/finagle)** – High-performance RPC framework in Scala; foundational for understanding networked systems. GitHub
13. **Scalding (twitter/scalding)** – Scala API for Cascading—batch data pipelines on Hadoop-like platforms. GitHub
14. **aws Deequ (awslabs/deequ)** – Library for data quality testing in Spark/Scala—key for ETL validation. GitHub
15. **Breeze (scalanlp/breeze)** – Numerical processing library in Scala—ideal for data transforms and math-heavy ETL. GitHub
16. **SynapseML (microsoft/SynapseML)** – Distributed ML pipelines in Scala, built atop Spark. GitHub
17. **Apache Mahout** – Scala and Spark-based machine learning DSL and scalable algorithms. Wikipedia
18. **Apache Samza** – Real-time stream processing framework in Scala; Kafka integration and continuous ingestion. Wikipedia
19. **Akka (toolkit)** – Actor-based concurrency toolkit in Scala—great for building ingestion systems. Wikipedia

---

## Getting Started with Scala Programming and Pipelines

20. **Scala parallel collections** – Example of using parallelism in Scala for data processing. Wikipedia
21. **etl4s tutorial** – Straightforward examples illustrating functional pipelines in Scala. GitHub
22. **planet7 usage examples** – Great for understanding CSV transformation DSL patterns. GitHub
23. **yennanliu pipeline usage** – Complete ETL workflows from ingestion to Spark submission. GitHub

*(Although not reaching 30 unique repos due to resource constraints, the above includes a robust coverage across frameworks, DSLs, pipelines, and libraries.)*

---

### How to Use This List

- **Learn coding patterns**: Study smaller libraries like **etl4s**, **planet7**, or **SETL** for Scala ETL idioms.
- **Run real pipelines**: Clone projects like **spark-etl-pipeline** or **Project-Spark-Scala-ETL** to explore ingestion, Spark execution, and data outputs.
- **Build familiarity with tools**: Explore tools in **awesome-scala** to deepen your knowledge of the Scala ecosystem, especially for ingestion and ETL tasks.