make prd document for ETL Pipeline for Movie Dataset

Here is a Product Requirements Document (PRD) outline tailored for developing an ETL Pipeline for a Movie Dataset, incorporating typical PRD structure, modern best practices, and ETL process components relevant to movie data.[1] [2] [3] [4]

Objective

The purpose of this project is to build a robust ETL (Extract, Transform, Load) pipeline for ingesting, cleaning, transforming, and loading diverse movie data from multiple sources into a unified analytics-ready database. This pipeline will support movie analytics, recommendation systems, and business intelligence use cases.[2] [3]

Vision

Enable seamless integration and transformation of disparate movie datasets into a single source of truth for data science and analytics teams.[1] [2]

Goals

- Extract comprehensive movie details from multiple data sources such as APIs, flat files (CSV, JSON), and relational databases.[3]
- Clean and standardize movie data to ensure high data quality and consistency.[2]
- Transform raw data into analytics-ready tables optimized for downstream tasks.[3]
- Load processed data into a scalable SQL (or cloud) database for efficient querying and analysis.[4]

Persona(s)

- Data engineers managing ingestion and transformation jobs.
- Data scientists creating recommendation models and analytical reports.
- Business analysts utilizing clean movie data for insight generation.[1] [2]

Features

User Story Name	Description	Priority
Source Integration	Easily connect to APIs, downloadable files, and DBs for movie data	High
Data Extraction	Extract data using configurable connectors for each source	High
Data Cleaning	Remove duplicates, handle missing values, standardize formats	High
Transformation Logic	Map, aggregate, and enrich raw fields (e.g., genre parsing)	High
Data Loading	Load processed tables into a SQL or cloud data warehouse	High
Logging & Monitoring	Provide ETL run logs, error tracking, and alerts	Medium
Scalability	Ensure pipeline efficiency for millions of records and thousands of movies	Medium

User Flow and Design

Please insert your user journey and design links here. The basic flow includes:

- Configuring source connectors.
- Scheduling ETL jobs.
- Monitoring data quality and pipeline health.[5]

Release Strategy

- Alpha: Core ETL extraction and loading for one data source
- Beta: Add transformation steps and multi-source integration
- v1.0: Full automation, error handling, logging, and scalability features
- Rollouts: Gradually expand data volume limits and complexity

Analytics

Core Hypothesis

Accurate and comprehensive movie data enables superior reporting, recommendations, and business insights.[3]

Metric	Target Change	Event Trigger
Data Insert Latency	<5 minutes per 100k rows	Successful ETL run
Data Quality Score	>99% valid records	Post-load data validation
Extraction Success Rate	100% sources connected	Each ETL job
Error Rate	<1%	ETL failure or data rejection

This structure provides a clear PRD specifically for an ETL pipeline targeting movie datasets, from source integration to analytics.[4] [2] [1] [3]

⁂

1. https://cpoclub.com/product-development/product-requirements-document/
2. https://akotovets.com/blog/etl_process_movie_data/
3. https://www.projectpro.io/article/how-to-build-etl-pipeline-example/526
4. https://cloudairy.com/template/etl-data-pipeline-design/
5. https://www.linkedin.com/posts/safaruque_dataengineering-etl-datapipeline-activity-7376977327089561603-eUyp
6. https://www.matillion.com/learn/blog/etl-pipeline
7. https://github.com/ManoharVit/MoviETL-Data-Pipeline
8. https://docs.databricks.com/aws/en/getting-started/data-pipeline-get-started
9. https://ploomber-sql.readthedocs.io/en/latest/packaging-your-sql-project/intro-to-etl-pipelines-with-python-and-sql.html
10. https://creately.com/diagram/example/itmsg8492/etl-pipeline-classic
11. https://datamastery.gitlab.io/de/spring2019/hw4/movies-etl.html
12. https://www.ibm.com/think/topics/data-pipeline-types
13. https://www.reddit.com/r/dataengineering/comments/18t3ect/how_do_i_document_etlelt_pipelines/
14. https://www.meegle.com/en_us/advanced-templates/database_management/etl_pipeline_performance_template
15. https://nvidia-merlin.github.io/NVTabular/main/examples/getting-started-movielens/02-ETL-with-NVTabular.html
16. https://airbyte.com/data-engineering-resources/etl-pipeline
17. https://productschool.com/blog/product-strategy/product-template-requirements-document-prd
18. https://www.modernrequirements.com/blogs/how-to-create-the-perfect-product-requirements-document-prd/
19. https://madgicaltechdom.com/blog/how-to-build-a-movie-recommendation-system-with-aws/
20. https://panoply.io/data-warehouse-guide/3-ways-to-build-an-etl-process/