Design Documentation

None

Real Time Data Ingestion Platform

AMOS Group 1

Table of contents

1. Design Documentation 3

1.1 Ingestion 3

1.2 Transform 3

1.3 Monitoring 3

1.4 Destination 3

1. Design Documentation

The RTDIP pipeline SDK allows you to ingest data in batches from various sources, transform it, and send the data to your desired destination. The SDK uses PySpark internally to process data, and provides a modular interface that allows you to build your pipeline, by using the provided components in any desired sequence.

1.1 Ingestion

The SDK provides ingestion components that facilitate the integration of data from various sources, such as industrial IoT systems, databases, and streaming platforms. These components abstract complexities, standardizing data into PySpark DataFrames, which act as the primary data structure for subsequent stages.

1.2 Transform

Transformation components allow users to modify, clean, and enrich ingested data. They are designed to be composable and configurable, enabling tasks such as data validation, format normalization, or feature engineering.

1.3 Monitoring

Monitoring components analyze DataFrames to provide insights into the data quality. You have to provide them with a Logger, from which you can extract the results of the monitoring.

1.4 Destination

The RTDIP SDK provides pipeline components to write your processed data to sink/destination systems.