The simulator is designed to provide flexibility and control in generating large data sets with minimal effort or coding. Users just need to provide the specifications in a template or through a RESTful web service which the simulator picks up. It then uses the specifications to generate random data as per the specifications. Since much of the data generation process is an independent task, multiple Simulator Instances can run independently on different machines. Which together produce large data sets and push the data to a common data storage or stream it. This simulated data sets could be used for various tasks like perform load analysis of analytics platforms which require large data input of specific format, extrapolate small data sets with certain amount of randomness so as to simulate real-world data sets, fill in the missing data in incomplete data sets and many other data simulation requirements.

Data Specifications

Multi-document Storage

RESTful Web Services

Simulator Instance 1

Database Storage

Simulator Instance N

<xml> templates

Data Stream

Workflow of the Simulation Process

**Simulator Instance**

sss

Connection Manager

Data Generation Specifications

Data Push Process

Setup Connection

Random Data Generation Process

Generated Data

Seed Data

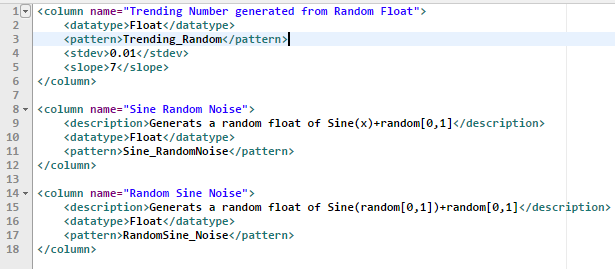
In-built data types specifications

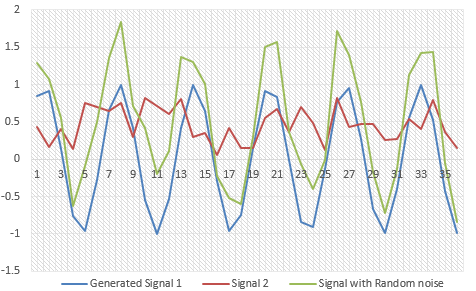
# Features

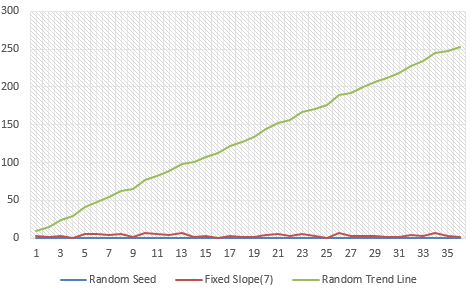
# Extensibility and Adaptability

The instance is designed in such a way that all sub modules are easily replaceable. For example the connection process to data storage in connection manager are independent and as per requirement they can be installed. A user could can easily push data to MSSQL and by just adding Cassandra Connection manager start pushing to Cassandra.

The simulator can easily be changed to generate custom patterns extending predefined patterns as required. Below snippet shows how with minimal pattern definition new types can be created and used for simulating various real-world trending data.



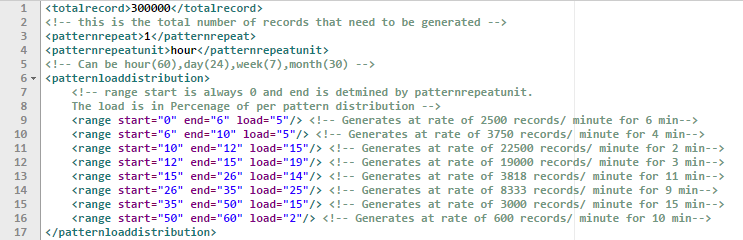




# Fine Grain Control

One of the important features of any data analysis platform is should be able to handle variation in data load and sudden surge in input data. The Simulator Instance can create this load variation in data generation and stream it real time or generate records with timestamp and push to storage that can be consumed by different analytic process.

This below load distribution template can create records for an hour with varying load during each of the minutes in an hour. Thus enabling simulation of high data load and spikes in a random fashion.



# Support data in motion and data at rest

Using different connectors the simulation engine can create data in different formats and push the data to file storage or directly into SQL or NOSQL databases. It can also stream the generated data in real-time or at required intervals that can be consumed by different services.

The below snippet lists the configuration that needs to be provided to let the simulation engine know what in what format and storage the output needs to be pushed or streamed. With this minimal change the data can be pushed once the connection manager is installed.

