**Project Description**

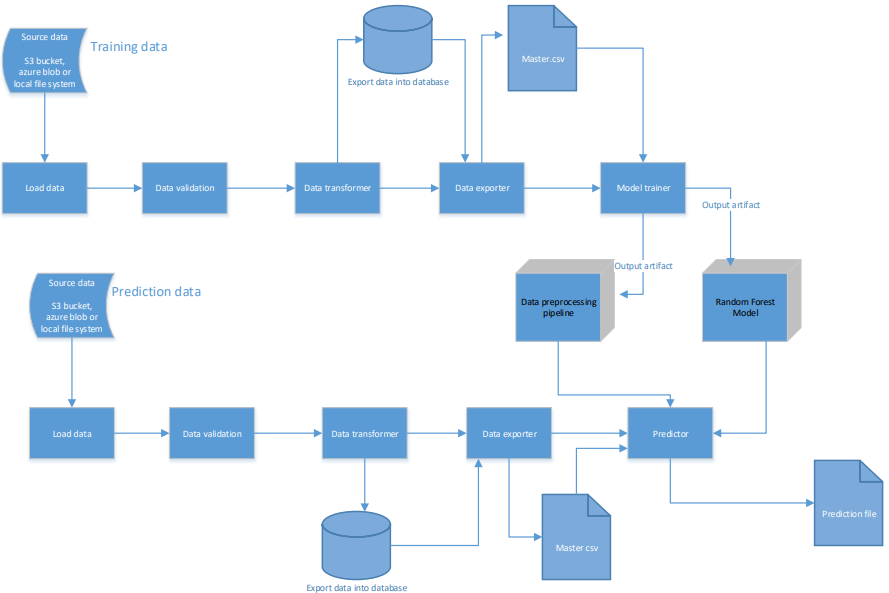
Project ‘Insurance Prediction’ is from Finance sector wherby we are trying to predict insurance expenses based on given inputs.

This is a regression problem and I’ve used Random Forest algorithm to predict Target data. The main objective of this project is to perform real time prediction on a simulated Streaming data and perform insertion of predicted data in MongoDB database in real time.

I’ve used both python and pySpark code in this project. Pipeline and regression model is created using batch data as first part. At second part, data from a local CSV file is simulated as a data stream and published to Kafka Topics. This streaming data is consumed by Spark consumer from Kafka Topics and previously saved pipeline & regression model is used for Streaming data pre-processing, transformation and prediction.

1. **Architecture for Batch Data Prediction**

Architecture for Batch data prediction is as below.



**Data Description**

The training data file contains insurance expenses as Target and six other given columns as input. Apart from training file, we will also have a "schema" file that contains all relevant information about training file such as: Name of files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of Columns and their datatype etc.

**Data Validation**

Various validation activities has performed on the training file such as File Name Validation, Number of Columns, Name of Columns, Null values in columns etc. If validation is passed, then file will proceed for data transformation and Insertion in database.

**Data Transformation**

Transformation activites include removal of unwanted columns and concatenation of files if have multiple files.

**Data Insertion in Database**

MongoDB is used as database for this project. Apart from training & prediction data, logging information has also inserted into database.

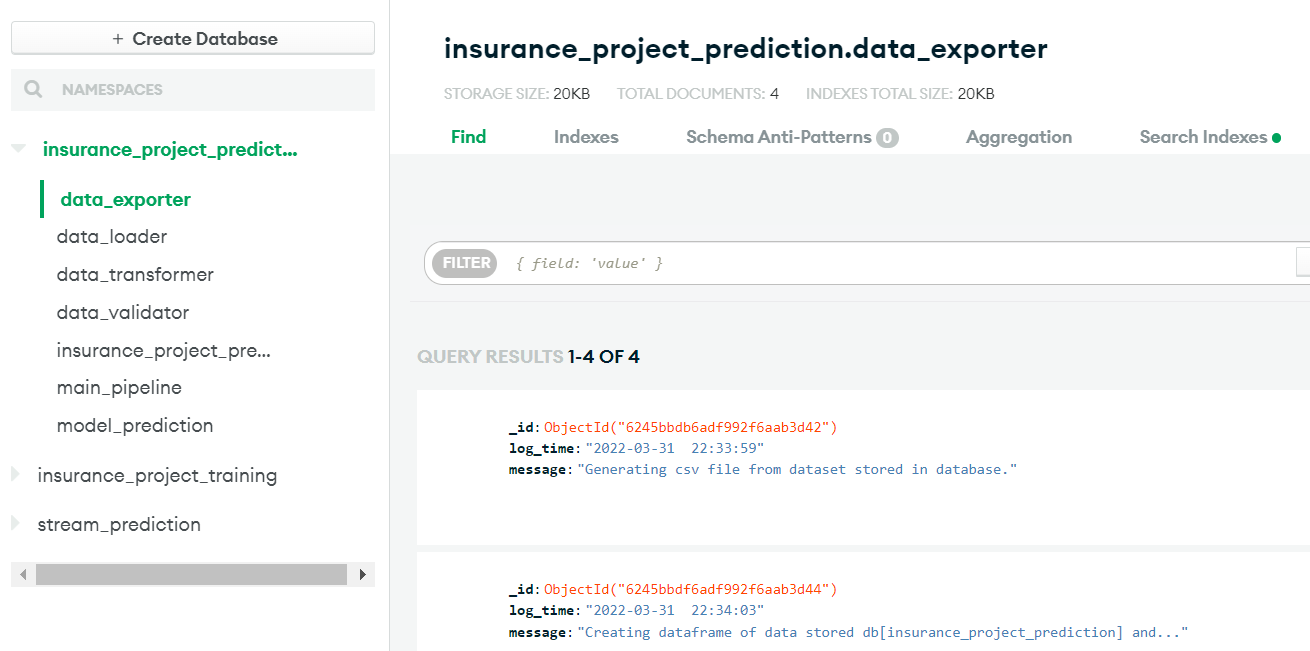


Fig : Execution logs & dataset saved in MongoDB

**Model Training**

Below activities has done as part of model training:

1. Data Export from Db : The data in training database is exported as a CSV file to be used for model training.
2. Data Preprocessing : StringIndexer, OneHotEncoder, VectorAssembler used in Pipeline
3. Model : RandomForestRegressor is used for model creation and training.

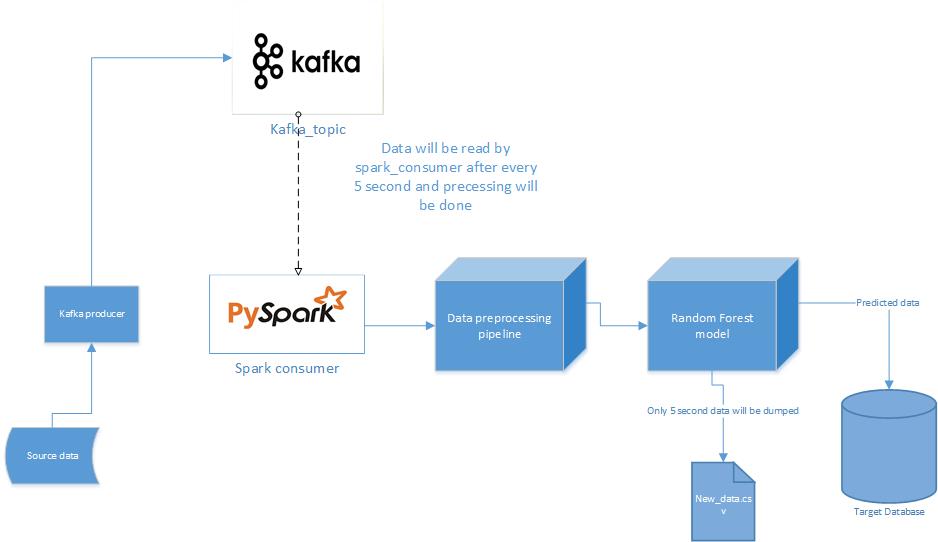
Finally Pipeline and model saved for prediction on batch data as well as Streaming data

**Batch Data Prediction**

Prediction file will also go through all activites as mentioned above. Saved pipeline & model will be loaded for Prediction.

1. **Architecture for Streaming Data Prediction**

Architecture for Streaming data prediction is as below.

****

**Producer**

Data from a ‘.csv’ file is simulated as streaming data and publish to Kafka Topics. This is done by reading ‘.csv’ file as Spark DataFrame and converting it to a collection of Rows (rdd.toLocalIterator()). Then iteratively send Rows to Kafka topics with a time delay of 1 sec each between rows. Prior to code execution, we need to ensure zookeeper and kafka server is running in the system.

**Consumer**

Streaming data from Kafka topics was read by Spark consumer. Saved pipeline and model will be loaded and perform prediction on Streaming data. Parameter processing\_interval was set as ‘5 seconds’. Hence spark consumer will wait for 5 seconds and whichever data that comes in 5 seconds will be taken for processing.

**Predicton and Data Insertion into Database**

Prediction on streaming data is inserted into MongoDB database. Database gets updated with new data at every 5seconds.

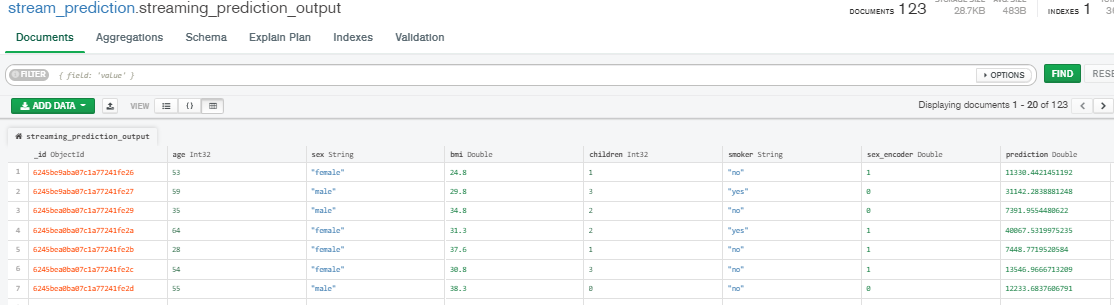
****

Fig : Streaming data Prediction inserted into Database