Stock Price Prediction

COURSE: BIG DATA SYSTEM ENGINEERING USING SCALA (CSYE 7200)
PROFESSOR: ROBIN C. HILLYARD
TEAM: ASWATHNARAYAN KIRUBAKARAN, MEENAKSHI MUTHIAH

Goals

- To determine the future value of a company stock or other financial instrument traded on an exchange
- Perform Time series analysis on Stock data using Scala and Spark

Data Sources

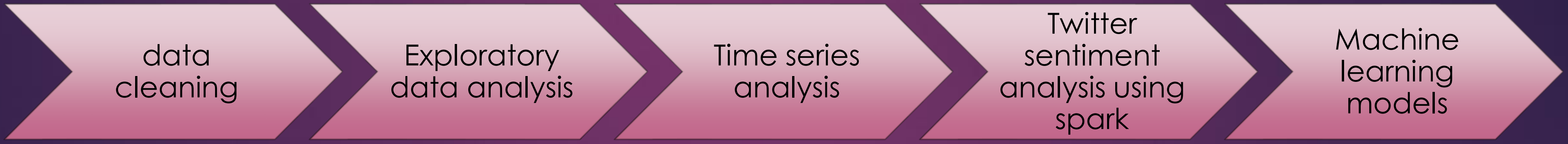
- Huge stock market Dataset: https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs/version/2/home
- S&P 500 Stock Data: https://www.kaggle.com/camnugent/sandp500
- Yahoo Finance
- The dataset is taken is from Kaggle and has data for around 500 companies.
- Each data file has 8 columns and 2000 rows per file.

Use cases

- Provide the user with future stock price for each company

- Provide the sentiment analysis for individual stock which the user wants to invest

Methodology



data cleaning

Exploratory data analysis

Time series analysis

Twitter sentiment analysis using spark

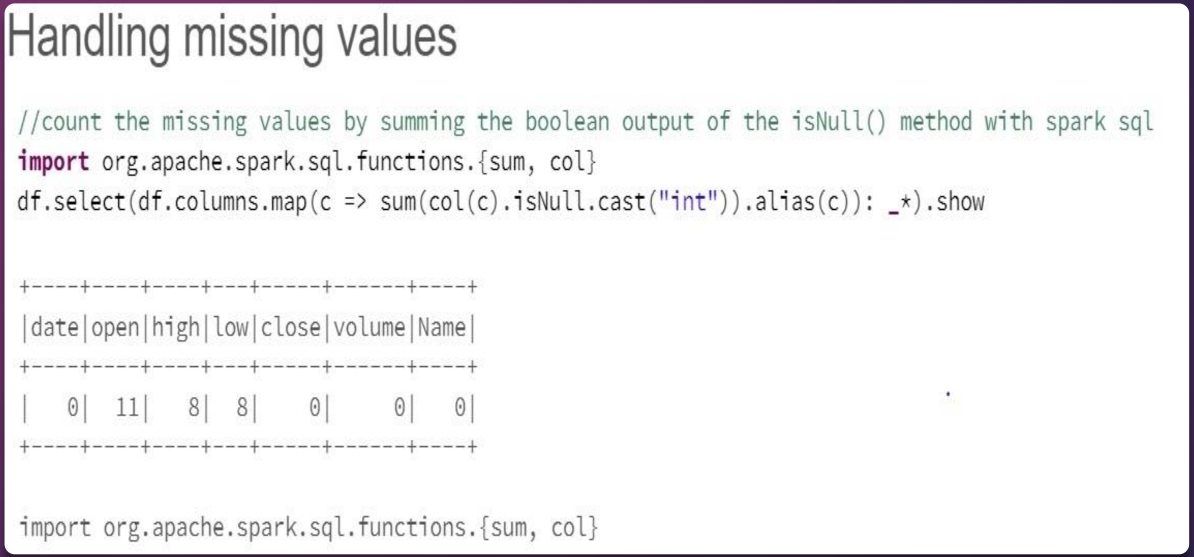
Machine learning models

Data Cleaning-Databricks

- Handle missing values

  removing nulls

  substituting nulls with the median



Handling missing values

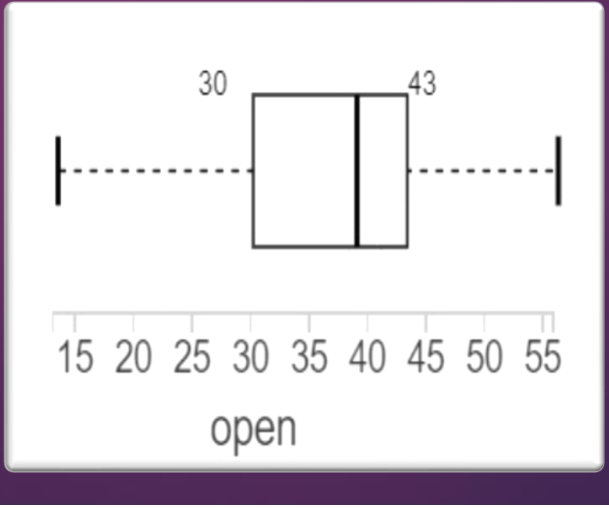
```
//count the missing values by summing the boolean output of the isNull() method with spark sql
import org.apache.spark.sql.functions.{sum, col}
df.select(df.columns.map(c => sum(col(c).isNull.cast("int")).alias(c)): _*).show
```

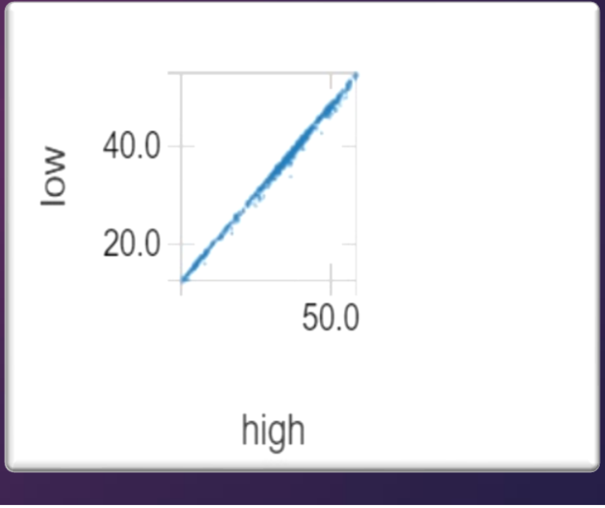
```
+----+----+----+---+-----+------+----+
|date|open|high|low|close|volume|Name|
+----+----+----+---+-----+------+----+
|   0|  11|   8|  8|    0|     0|   0|
+----+----+----+---+-----+------+----+

import org.apache.spark.sql.functions.{sum, col}
```

EDA with Databricks

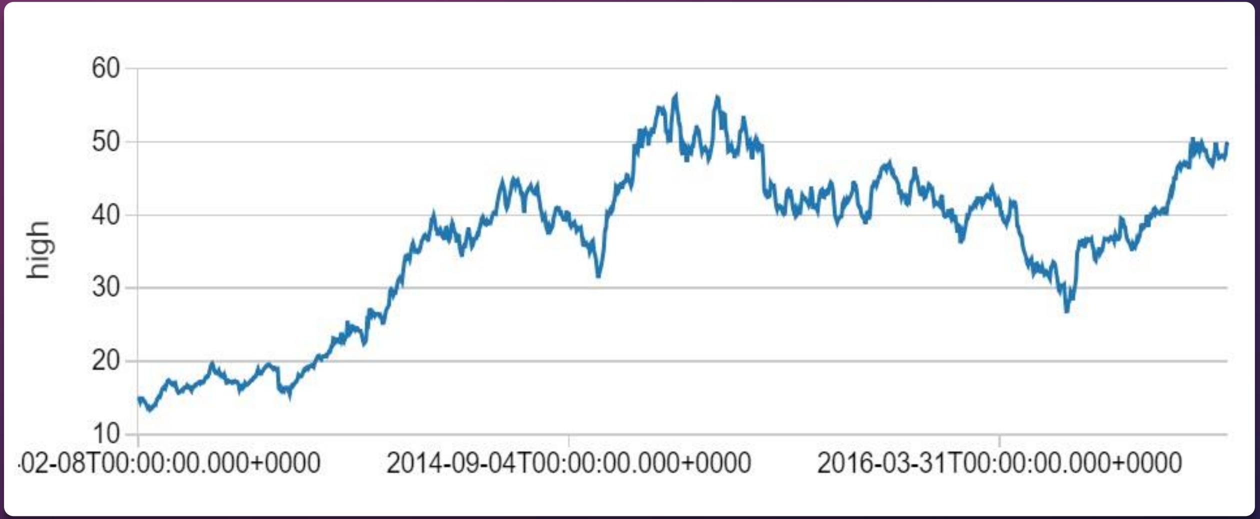
- Summary statistics
- Correlation
- Trend analysis
- Outlier detection





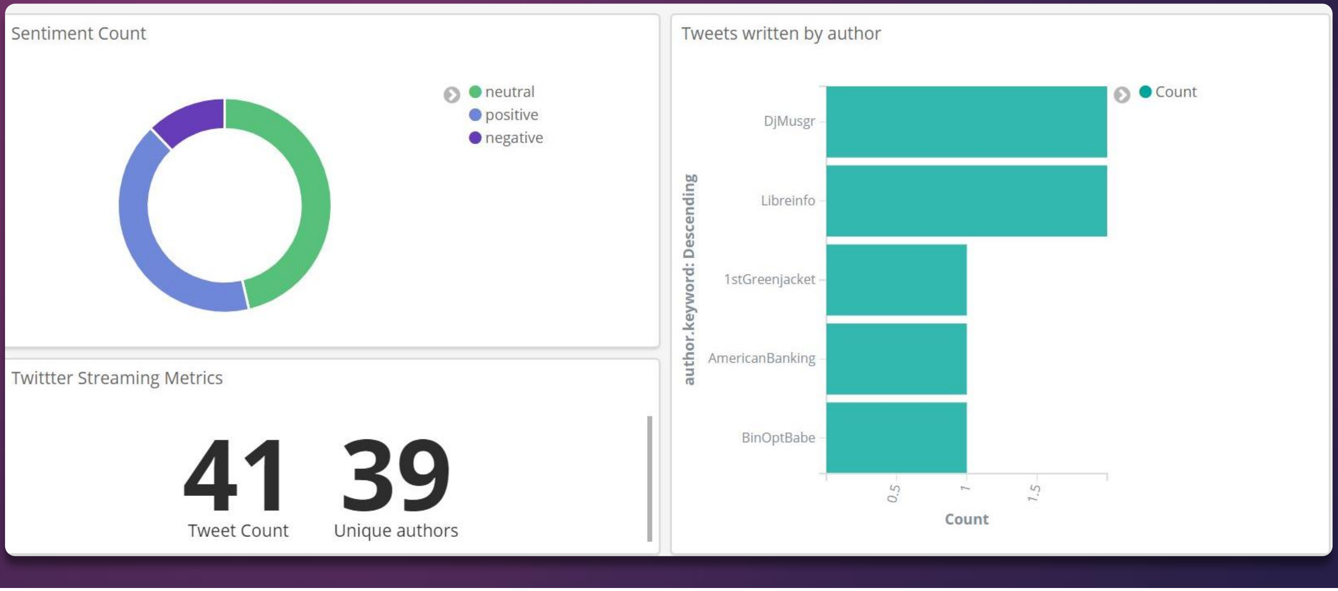
Times Series Analysis

- Stationary check
- Durbin-Watson Test for Auto Correlation
- Smoothing

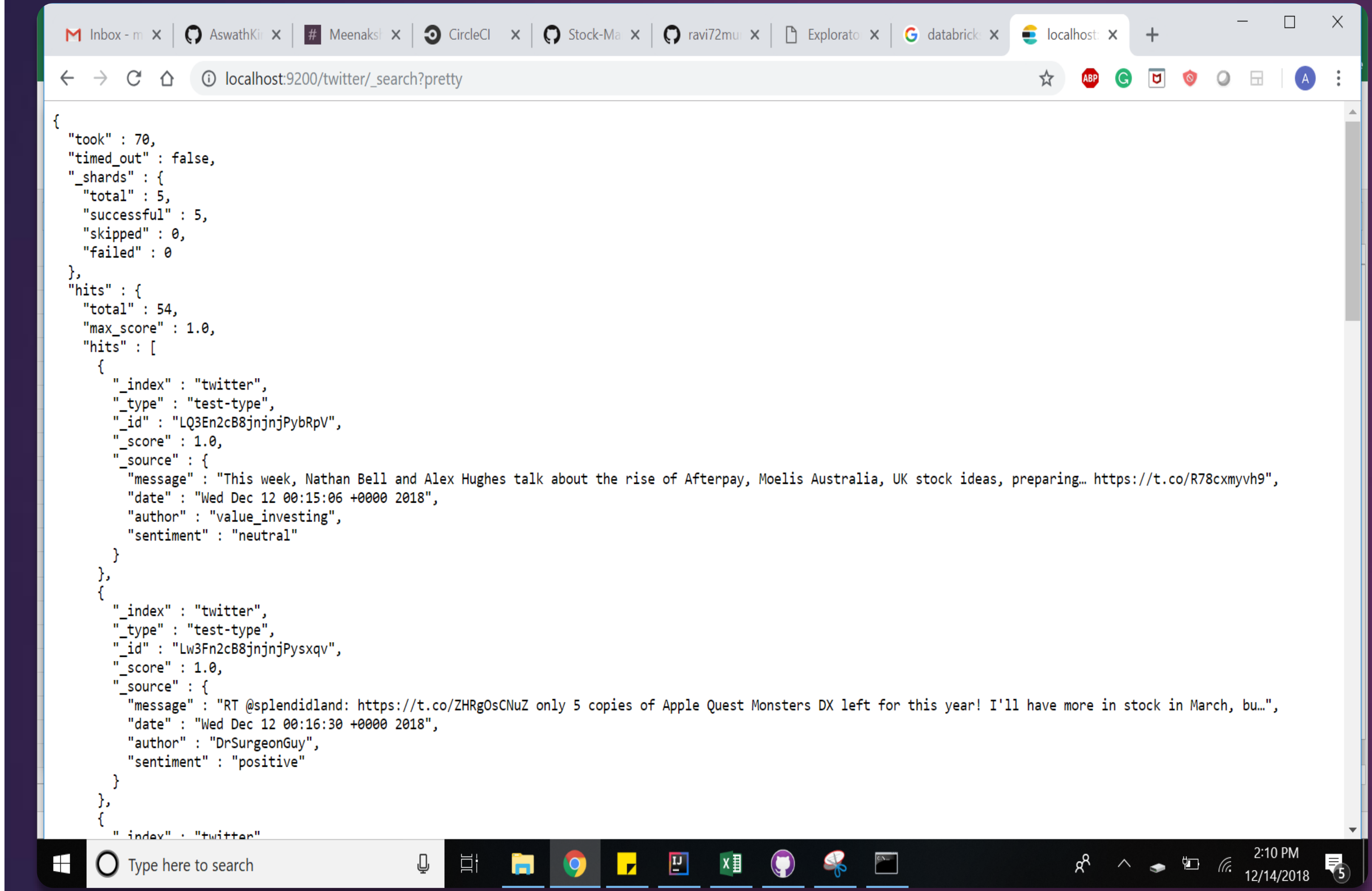


Twitter Sentiment Analysis

- Spark Twitter Streaming
- Sentiment Analysis using Stanford NLP
- Stored the results in ElasticSearch
- Visualization using Kibana



Elastic Search



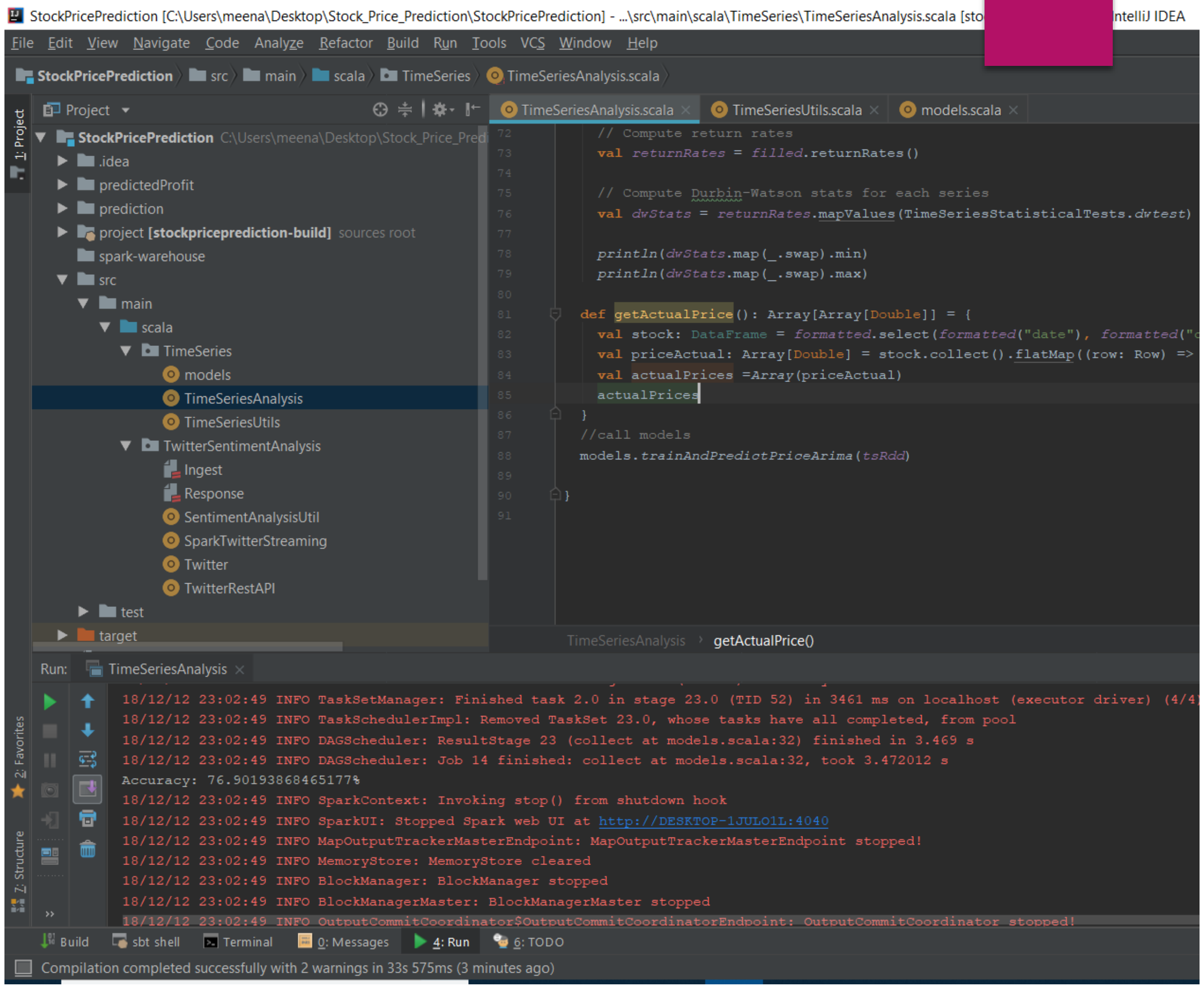
```
{
  "took" : 70,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 54,
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "twitter",
        "_type" : "test-type",
        "_id" : "LQ3En2cB8jnjnjPybRpV",
        "_score" : 1.0,
        "_source" : {
          "message" : "This week, Nathan Bell and Alex Hughes talk about the rise of Afterpay, Moelis Australia, UK stock ideas, preparing… https://t.co/R78cxmyvh9",
          "date" : "Wed Dec 12 00:15:06 +0000 2018",
          "author" : "value_investing",
          "sentiment" : "neutral"
        }
      },
      {
        "_index" : "twitter",
        "_type" : "test-type",
        "_id" : "Lw3Fn2cB8jnjnjPysxqv",
        "_score" : 1.0,
        "_source" : {
          "message" : "RT @splendidland: https://t.co/ZHRgOsCNuZ only 5 copies of Apple Quest Monsters DX left for this year! I'll have more in stock in March, bu…",
          "date" : "Wed Dec 12 00:16:30 +0000 2018",
          "author" : "DrSurgeonGuy",
          "sentiment" : "positive"
        }
      },
      {
        "_index" : "twitter",
```

Machine Learning Models-Time Series Forecasting

- *Feature Engineering* : Converted the stock prices to vectors and applied smoothing
- *Models*: Utilized ARIMA model for stock price forecasting
- *Evaluation Metrics*: Accuracy and RMSE

Acceptance Criteria

- Stock price prediction accuracy should provide accuracy above 72%


StockPricePrediction [C:\Users\meena\Desktop\Stock_Price_Prediction\StockPricePrediction] - ...\src\main\scala\TimeSeries\TimeSeriesAnalysis.scala [sto
ntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
StockPricePrediction › src › main › scala › TimeSeries › TimeSeriesAnalysis.scala
Project
StockPricePrediction C:\Users\meena\Desktop\Stock_Price_Pred
.idea
predictedProfit
prediction
project [stockpriceprediction-build] sources root
spark-warehouse
src
main
scala
TimeSeries
models
TimeSeriesAnalysis
TimeSeriesUtils
TwitterSentimentAnalysis
Ingest
Response
SentimentAnalysisUtil
SparkTwitterStreaming
Twitter
TwitterRestAPI
test
target
TimeSeriesAnalysis.scala
TimeSeriesUtils.scala
models.scala
TimeSeriesAnalysis › getActualPrice()
Run: TimeSeriesAnalysis
Build
sbt shell
Terminal
0: Messages
4: Run
6: TODO
1: Project
2: Favorites
Z: Structure

```scala
      // Compute return rates
      val returnRates = filled.returnRates()

      // Compute Durbin-Watson stats for each series
      val dwStats = returnRates.mapValues(TimeSeriesStatisticalTests.dwtest)

      println(dwStats.map(_.swap).min)
      println(dwStats.map(_.swap).max)

    def getActualPrice(): Array[Array[Double]] = {
      val stock: DataFrame = formatted.select(formatted("date"), formatted("c
      val priceActual: Array[Double] = stock.collect().flatMap((row: Row) =>
      val actualPrices =Array(priceActual)
      actualPrices
    }
    //call models
    models.trainAndPredictPriceArima(tsRdd)

}
```

```
18/12/12 23:02:49 INFO TaskSetManager: Finished task 2.0 in stage 23.0 (TID 52) in 3461 ms on localhost (executor driver) (4/4
18/12/12 23:02:49 INFO TaskSchedulerImpl: Removed TaskSet 23.0, whose tasks have all completed, from pool
18/12/12 23:02:49 INFO DAGScheduler: ResultStage 23 (collect at models.scala:32) finished in 3.469 s
18/12/12 23:02:49 INFO DAGScheduler: Job 14 finished: collect at models.scala:32, took 3.472012 s
Accuracy: 76.90193868465177%
18/12/12 23:02:49 INFO SparkContext: Invoking stop() from shutdown hook
18/12/12 23:02:49 INFO SparkUI: Stopped Spark web UI at http://DESKTOP-1JUL01L:4040
18/12/12 23:02:49 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/12/12 23:02:49 INFO MemoryStore: MemoryStore cleared
18/12/12 23:02:49 INFO BlockManager: BlockManager stopped
18/12/12 23:02:49 INFO BlockManagerMaster: BlockManagerMaster stopped
18/12/12 23:02:49 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
```

Compilation completed successfully with 2 warnings in 33s 575ms (3 minutes ago)

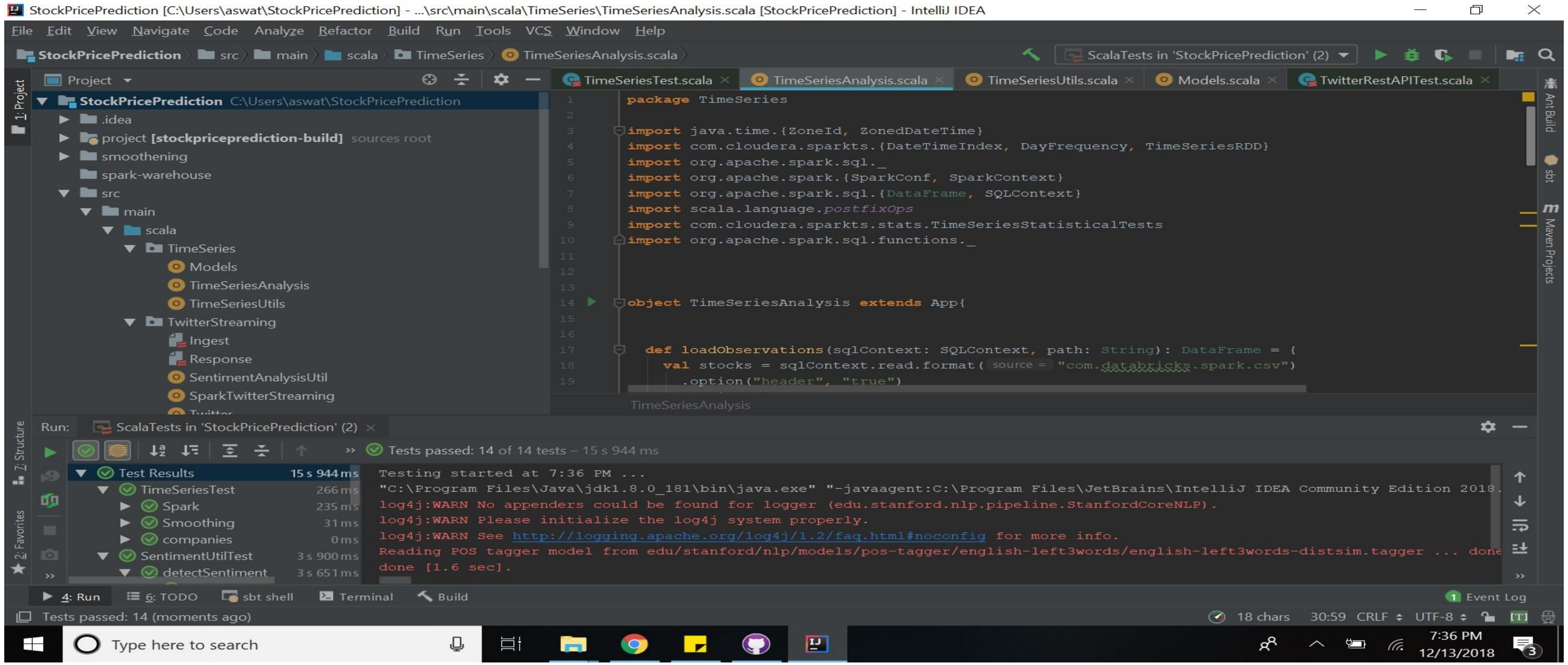
MileStones/Sprints

- Sprint 1: Perform data wrangling, integration and exploratory data analysis
- Sprint 2: Build machine learning models and perform training
- Sprint 3: Perform twitter streaming. Combine the work
- Sprint 4: Testing and Documentation

Code

- Everything in Scala
  - Spark
  - Spark Mllib
  - Spark Streaming

Unit Test



StockPricePrediction [C:\Users\aswat\StockPricePrediction] - ...\src\main\scala\TimeSeries\TimeSeriesAnalysis.scala [StockPricePrediction] - IntelliJ IDEA

```scala
package TimeSeries

import java.time.{ZoneId, ZonedDateTime}
import com.cloudera.sparkts.{DateTimeIndex, DayFrequency, TimeSeriesRDD}
import org.apache.spark.sql._
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.sql.{DataFrame, SQLContext}
import scala.language.postfixOps
import com.cloudera.sparkts.stats.TimeSeriesStatisticalTests
import org.apache.spark.sql.functions._



object TimeSeriesAnalysis extends App{


  def loadObservations(sqlContext: SQLContext, path: String): DataFrame = {
    val stocks = sqlContext.read.format( source = "com.databricks.spark.csv")
      .option("header", "true")
```

TimeSeriesAnalysis

Run: ScalaTests in 'StockPricePrediction' (2)

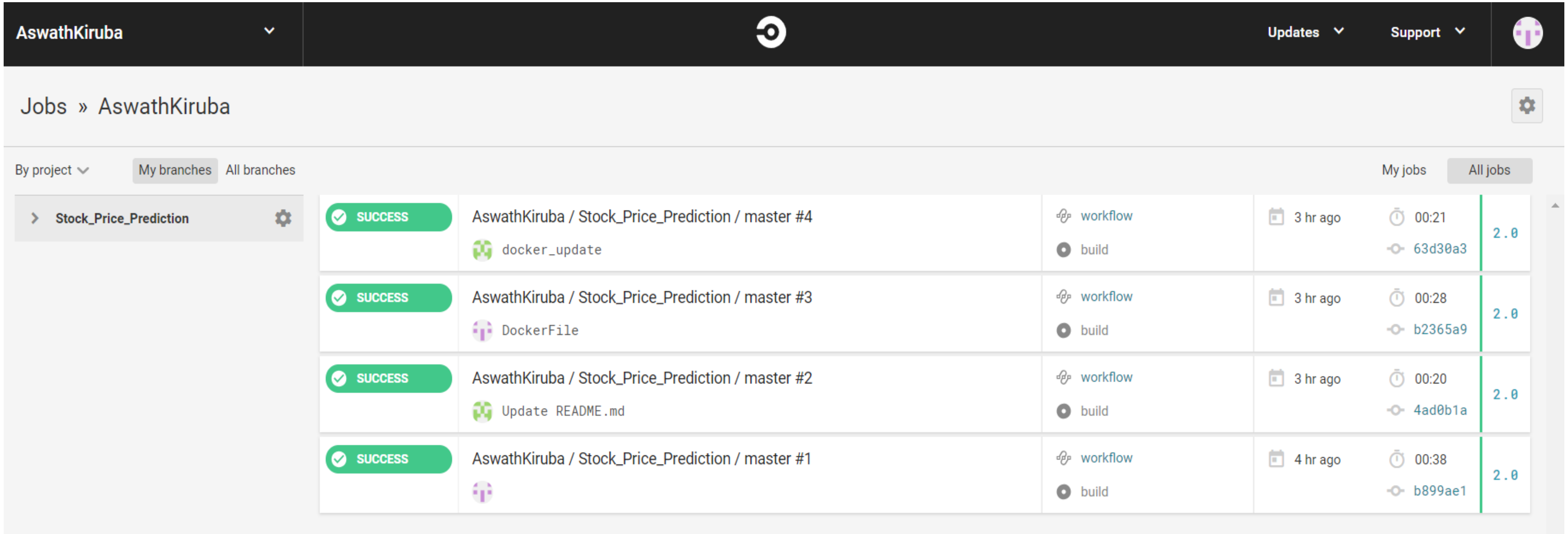
Tests passed: 14 of 14 tests – 15 s 944 ms

```
Test Results              15 s 944 ms
  TimeSeriesTest             266 ms
    Spark                    235 ms
    Smoothing                 31 ms
    companies                  0 ms
  SentimentUtilTest        3 s 900 ms
    detectSentiment        3 s 651 ms
```

```
Testing started at 7:36 PM ...
"C:\Program Files\Java\jdk1.8.0_181\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA Community Edition 2018.
log4j:WARN No appenders could be found for logger (edu.stanford.nlp.pipeline.StanfordCoreNLP).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done
done [1.6 sec].
```

Tests passed: 14 (moments ago)

Continuous Integration and Docker



Jobs » AswathKiruba

SUCCESS	AswathKiruba / Stock_Price_Prediction / master #4 docker_update	workflow build	3 hr ago	00:21 63d30a3	2.0
SUCCESS	AswathKiruba / Stock_Price_Prediction / master #3 DockerFile	workflow build	3 hr ago	00:28 b2365a9	2.0
SUCCESS	AswathKiruba / Stock_Price_Prediction / master #2 Update README.md	workflow build	3 hr ago	00:20 4ad0b1a	2.0
SUCCESS	AswathKiruba / Stock_Price_Prediction / master #1	workflow build	4 hr ago	00:38 b899ae1	2.0

Thank you