Coding tasks for data engineers

Hello and welcome to the coding tasks for data engineers!

Test 1:

The exercise covers [Apache Spark](http://spark.apache.org/docs/2.1.0/programming-guide.html), [Dataframes](http://spark.apache.org/docs/2.1.0/sql-programming-guide.html), Scala/Java and is all about ETL.

Tasks

The goals of the tasks are to pass all tests and implement it in a way that it performs very well for big data. The steps in detail:

1. Fork the project and create a new branch for your implementation
2. Implement the function, according to documentation and tests
3. Run and pass all tests
4. Update the chapter "Describe your solution" in the README files
5. Create a pull request

Comment

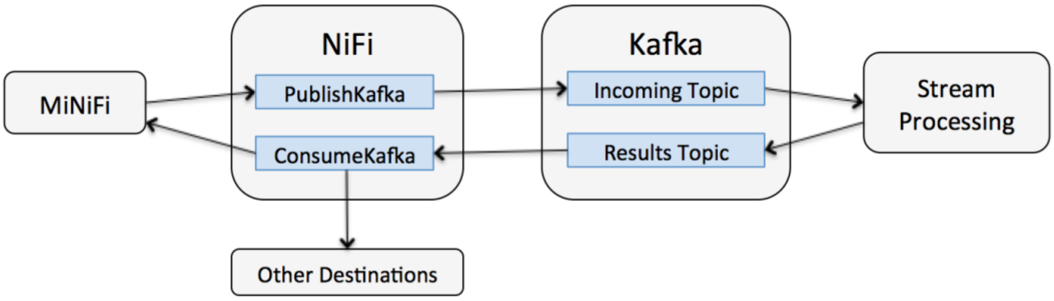
Feedback/Questions

Please contact `bigdata-team@evobanco` for feedback and questions.

Test 2:
Integrating Apache NiFi and Apache Kafka

A common scenario is for NiFi to act as a Kafka producer. With the advent of the Apache MiNiFi sub-project, MiNiFi can bring data from sources directly to a central NiFi instance, which can then deliver data to the appropriate Kafka topic. The major benefit here is being able to bring data to Kafka without writing any code, by simply dragging and dropping a series of processors in NiFi, and being able to visually monitor and control this pipeline.

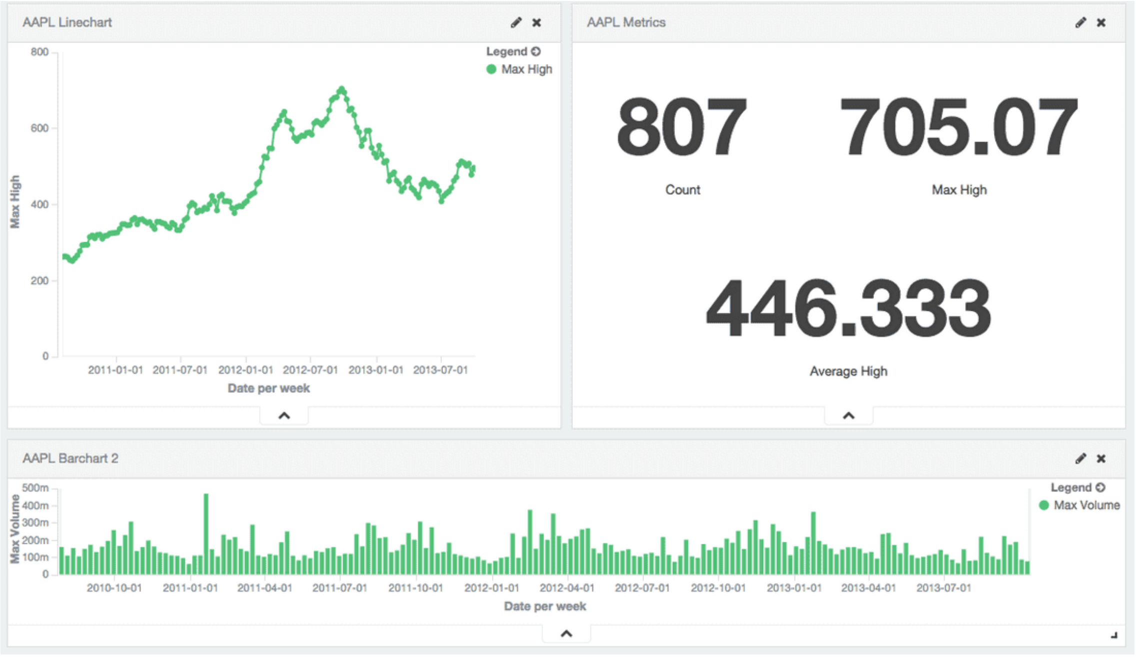
Bi-Directional Data Flows
A more complex scenario could involve combining the power of NiFi, Kafka, and a stream processing platform (Kafka Streams or Spark Streaming) to create a dynamic self-adjusting data flow. In this case, MiNiFi and NiFi bring data to Kafka which makes it available to a stream processing platform, or other analytic platforms, with the results being written back to a different Kafka topic where NiFi is consuming from, and the results being pushed back to MiNiFi to adjust collection.



Building a 311 Data Pipeline

- Building the data pipeline: NIFI, KAFKA, Stream Processing (Kafka Streams / Spark Streaming)
- Building a ELK/Kibana dashboard
  - data-cleaning
  - streaming the data into Elasticsearch
  - creating visualizations using Kibana
- Persist the data into Cocuhbase / MongoDB

Yahoo stock Price data:
https://finance.yahoo.com/quote/AAPL/history?ltr=1


AAPL Linechart
Legend
Max High
800
600
400
200
0
Max High
2011-01-01
2011-07-01
2012-01-01
2012-07-01
2013-01-01
2013-07-01
Date per week
AAPL Metrics
807
Count
705.07
Max High
446.333
Average High
AAPL Barchart 2
Legend
Max Volume
500m
400m
300m
200m
100m
0
Max Volume
2010-10-01
2011-01-01
2011-04-01
2011-07-01
2011-10-01
2012-01-01
2012-04-01
2012-07-01
2012-10-01
2013-01-01
2013-04-01
2013-07-01
Date per week