JOINT INSTITUTE FOR NUCLEAR RESEARCH
Laboratory of Information Technologies

FINAL REPORT ON THE SUMMER STUDENT PROGRAM

Big Data Analysis

Supervisor:
Vladimir Korenkov
Sergey Belov

Student:
Nijat Mursali, Azerbaijan
ADA University

Participation period:
July 15 – August 31

Dubna, 2019

Table of Contents

List of Figures

List of Abbreviations

Abbreviation	Explanation
HTTP	HyperText Transfer Protocol
API	Application Program Interface
CSV	Comma Separate Value(s)
VM	Virtual Machines
UI	User Interface
SQL	Standardized Query Language

Abstract

The report describes the fundamental components of Big Data Analysis and the development process for my project. During the lifetime of my project the fundamental objective was to make deep analysis degrading to Docker, NIFI, Kafka and Spark which are the main components of Apache. The main component of my project was to parse the data from different websites about vacancies in Azerbaijan. That was the first challenge that I have faced in the beginning of this project because in Azerbaijani websites there's no API that I could get the data. Instead of that, I have created my own Python script in order to parse the required information (name of vacancy and company and description about the vacancy) from the main website for vacancies. For this project, I have parsed more than five hundred vacancies with different types of names and cleaned it with my own written Python script in order to get rid of duplicates if it had. After acquiring the data, I have imported it into NIFI and setup the processors to generate the flow file and using other processors update the data and put it into the Kafka. After putting the data into Kafka I have managed to create the project for PySpark and send the information from Kafka to Spark. The outcome of this project is a fully working Spark Streaming application that reads the data within Spark master and workers.

Introduction

As seen from news, technologies and articles the concept of Big Data has been evolving during the last decades. Big Data is a huge field that analyzes and extracts the large sets of data that is not exactly easy to do with traditional data-processing application software. Usage of Big Data helps to reduce the complexity and usage of the large datasets in a sense to help scientists to analyze the huge volumes of data where the data can be either structured or unstructured. One important example for big data analytics is Hadoop which is the distributed processing framework that is an open source project by Apache. In this article, we will not be talking too much about the Big Data itself, but talk about the concepts like NIFI, Kafka and Spark inside the Docker images. In the following paragraphs, I will be giving more information about those concepts and the way we have used them in our project.

How it All Started

As I have mentioned in introduction part, there was no any API for getting the data in JSON format, thus I had to create my own script to parse the data from website. I have used https://boss.az/ which is the most famous website for vacancies with detailed information which makes it more reliable than others. I have parsed the information in CSV format with columns/headers of name of vacancy, company and description for them.

	A	B	
1	Title	Company	Summary
2	SATICI	Smarton MMC	- Uşaq oyuncaqlarının satışı - Müştəriyə səbr və təmkinlə xidmət göstərməli - Qoyulan satış hədəflərinə çatmalı - İki növbə: 09:00-dan 18:00-dək, 13:00-dan 22:00-dək (uşaq oyuncaqları) - İş aparıcı ticarət mərkəzlərində yerləşir - Ayda altı gün istirahət
3	DAYƏ (XƏSTƏYƏ QULLUQ)	Novco Group of Companies	- Əmək şəraiti: Həftəiçi 3 gün, saat 10:00-dan 19:00-dək - Xəstəyə nəzarət olunması - Dərman vasitələrin vaxtı-vaxtında verilməsi - Xəstənin nitqinin açılması üçün onunla danışıqların aparılması, şifahi çalışr - Xəstənin gəzintiyə çıxarılması - Gündəlik ev işlərində yardımın göstərilməsi (yemək bişirilməsi daxil deyil) - Mənzilin yığışdırılması (2 otaq)
4	MEXANİK	Konfirom MMC	- İş yeri: Sabirabad
5	MEXANİK	Konfirom MMC	- İş yeri: Badamdar, Qurd qapısı restoranı yaxınlığında
6	XADİMƏ	Retro Holding MMC	- İş yeri: Buzovna qəsəbəsi , Retro Avtoservisi - İş vaxtı: Həftə içi 6 gün, saat 09:00-dan 18:00-dək - Nahar şirkət tərəfindən verilir - Avtoservis sahəsində təmizliyə riayət etmək

Fig 1. Parsed CSV File

Big Data Concepts

Docker

The first time the idea of using containers instead of virtual machines has been developed in 2014 and it was the new era of technologies because that was the time numerous companies moved their applications from VM (virtual machines) into containers which are the part of Docker. Docker is in a sense virtual machine itself, but unlike virtual machines it is more efficient to create containers because it allows applications to use only ports in the server which creates and gives significant performance boost and beneficial to reduce the size of application. Another beneficial side of Docker is that it is open source software which other people are able to contribute to the flow of it. From those reasons, we have downloaded the images and containers for NiFi, Kafka and Spark with Docker in order to run them in the same server that could be easy to connect them together. It's also crucial to mention that to use Docker images, it was required to install both Docker and docker-compose. In the following paragraphs, I will be introducing the topics that we have learned and practiced using Docker containers.

NIFI

NiFi is one of the best open source software of Apache that manages and automates the flow of data between the systems which we call processors. There is several other software like NiFi but having great web-based user interface makes it much easier and reliable in order to create and control the flows of data between processors. Some of the advantages of using NiFi are that it is good and easy to do data ingestion, easy and usable UI, doing everything in real-time where it is connected to the server and exporting the data into other software like Kafka.

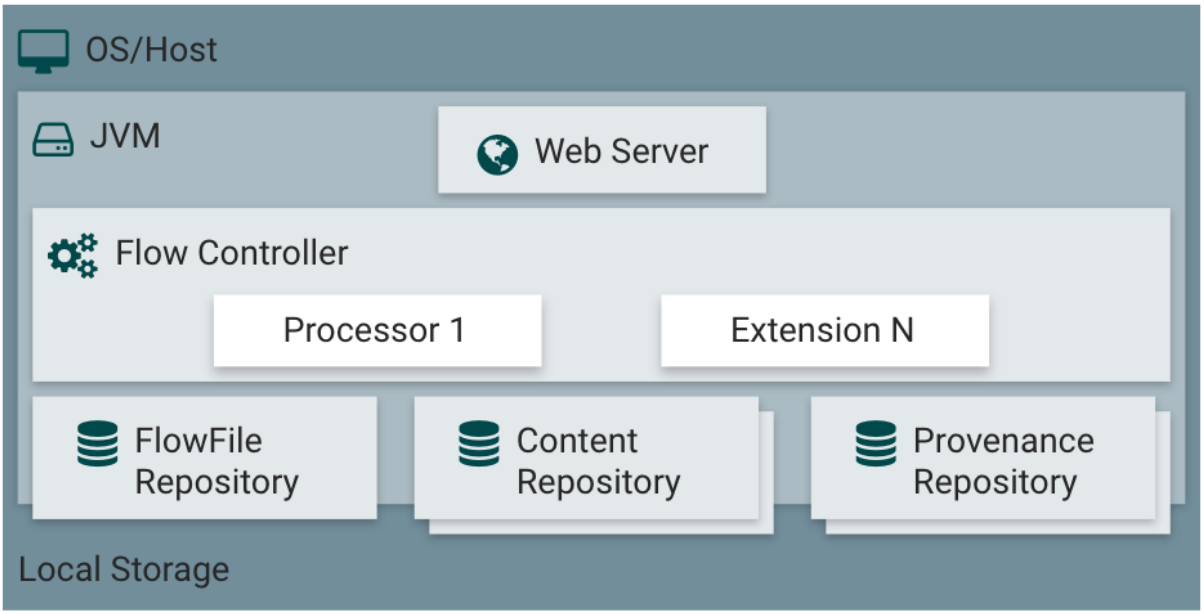

Fig 2. NiFi Architecture

For our project we have created several NiFi processors to accomplish the task which was to send the CSV file that contained more than five hundred vacancies into Kafka topics. To accomplish this, we had to create processors that could get the CSV data and put it into Kafka, so we have used GenerateFlowFile, UpdateAttribute and PublishKafka/PutKafka. Inside the GenerateFlowFile we have added all the vacancies and from GenerateFlowFile we have made connections into UpdateAttribute in order to send the data without storing it in UpdateAttribute and sending it into PublishKafka or PutKafka processors. Additionally, we have added PublishKafkaRecord in order to send the CSV file to Kafka topics in Avro format which is the famous format for Kafka topics.

After putting our data into GenerateFlowFile processor we connected processor with UpdateAttribute processor and put the termination in *success* state like in the *Figure 3*. We update the data there, but we don't store it and then send it to another and last processor to put into kafka topics.

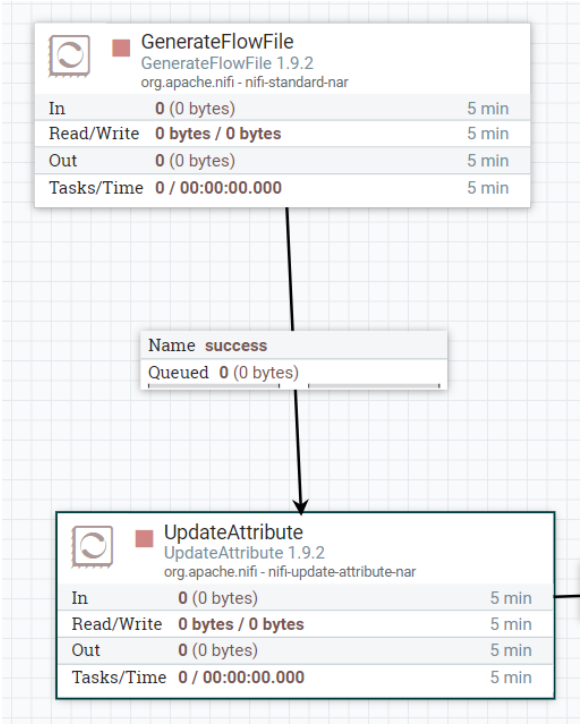

Fig 3. Making Connections Between Two Processors

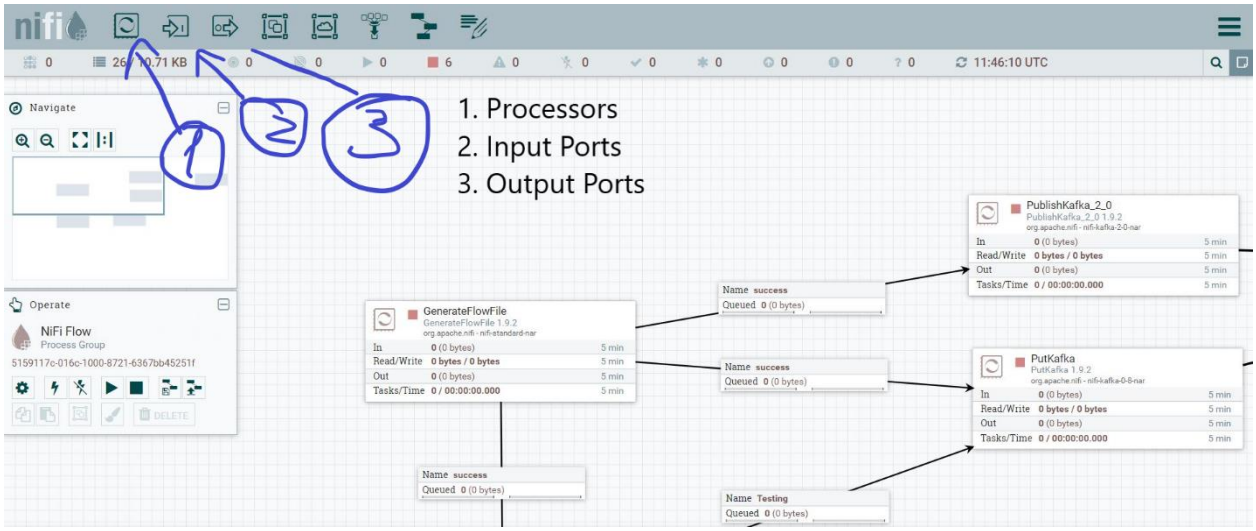

Fig 4. NiFi Interface with Workflow

Kafka

Kafka is one of the free to use open source software that provides framework for creating, storing and analyzing the data. As mentioned above, it is one of the free to use applications which means everyone can contribute towards updates and supports for the development of application

and from this reason, it is developing in today's world. Same as NiFi, we have installed Kafka using Docker images to make the life easier and faster. Additionally, Kafka gives the possibility of running several servers which makes it distributed system. In order to send the data from NiFi into Kafka, we had to create the topic inside a cluster and add it into the NiFi PublishKafka/PutKafka processors.

```
kafka-topics --zookeeper zookeeper:2181 --create --topic
vacancy_topic --partitions 1 --replication-factor 1
```

Fig 5. Creating Kafka topic

As we have created the topics inside Kafka cluster, we were able to add it inside the NiFi processors to put the information about vacancies inside Kafka topics.

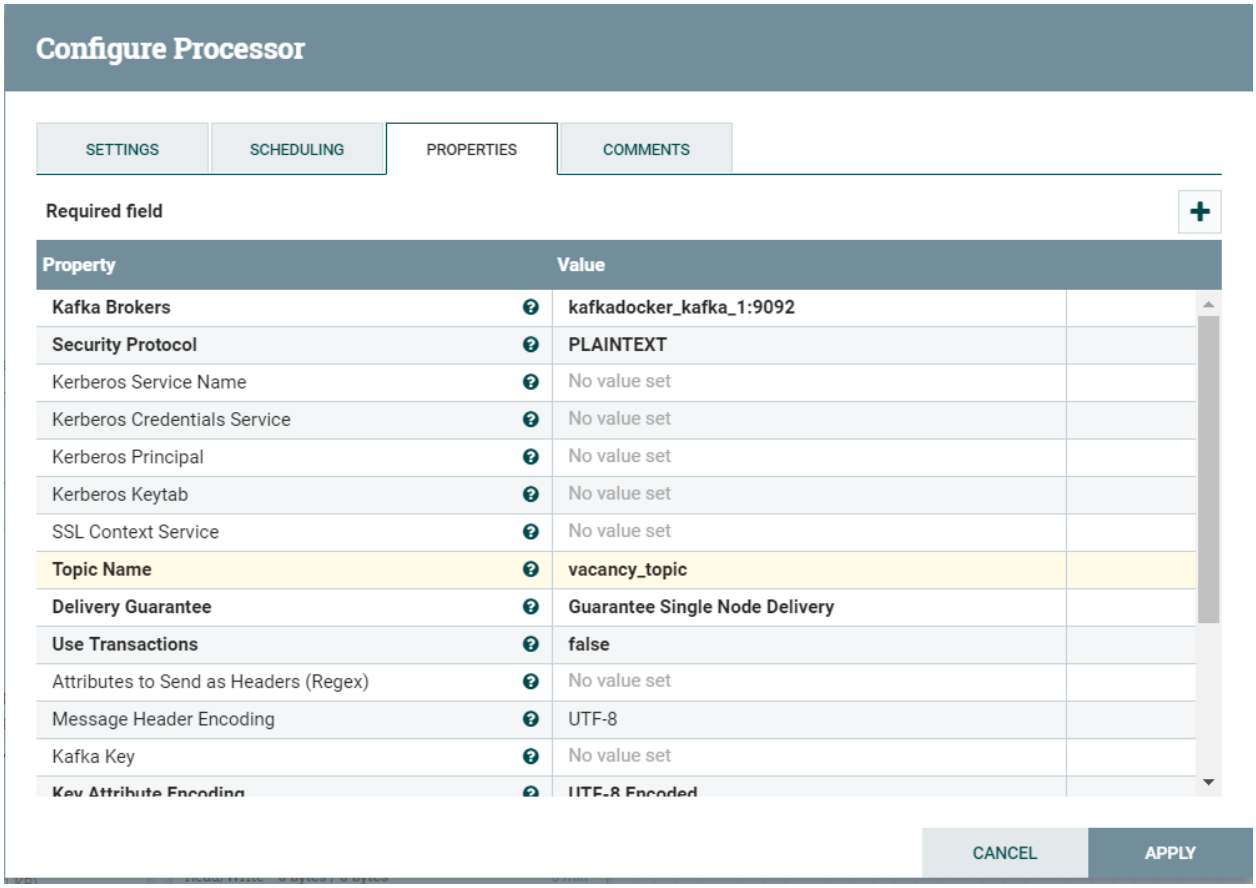

Fig 6. Kafka Topics inside NiFi

As we have accomplished adding required information about Kafka brokers and topics inside NiFi processor, we could make connection between (as in the Fig. 2) NiFi processors and start the data flow between them in order to send the data from NiFi to Kafka topic.

Spark Streaming

Spark is one of the important distributed data processing engines that is suitable in a wide range of circumstances. The workflow of Spark is huge that it can work with several libraries such as ML (Machine Learning), stream processing and SQL which is mainly used by several companies in today's world. Additionally, Spark can work with different programming languages such as Python (PySpark is the library for it), Java, Scala and R which are the most popular languages. Distributed data means that data can work with several clusters where we mainly divide it into data blocks that can work separately. We have used *SparkSession* in order to run Spark application as independent processes. For this article we have used Spark Streaming processing which takes logs files and puts into sensor data. PySpark was the best option to use Spark Streaming for our project where we have initialized *SparkSession* and *KafkaUtils* where we set Kafka brokers and name of the topics in order to read the data from Kafka topics into Spark.

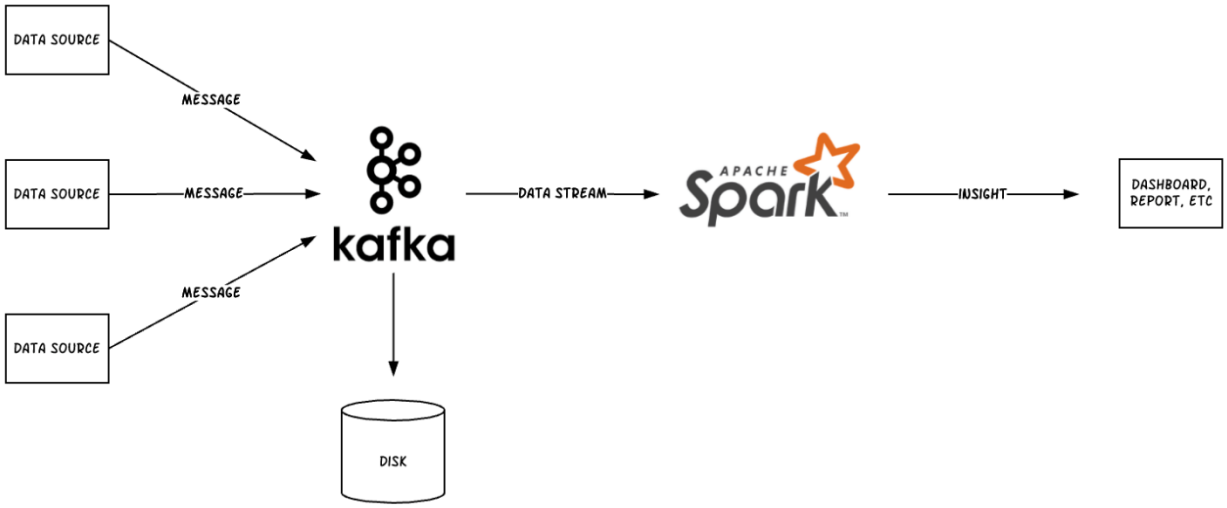

Fig 7. Spark Streaming

Additionally, creating pipelines and combining different techniques and processes into single whole was the fundamental idea of this project, so we could be working in one single cluster. As we have been mentioning beforehand, to do this we have used Docker containers to combine into one whole.

Conclusion

To conclude all the ideas and concepts I have been mentioning, Big Data is very important concept to understand and implement in order to. In this final report, I have explained the fundamental concepts of Big Data such as NIFI, Kafka, Spark by using Docker and Spark Streaming was the main outcome of this project. Additionally, I have managed to give enough information about NIFI, Kafka, Spark and the implementation of them in our project. With the help of my supervisor I have managed to understand the ways to implement these concepts, but usage of Docker images was much beneficial and better in a sense of time and compatibility. By installing Docker images for NIFI, Kafka and Spark we have managed to create our own topics for Kafka and get the data from NIFI and put in into Kafka which we eventually used for Spark Streaming.

References

1. NiFi Documentation
    - https://nifi.apache.org/
2. Docker Documentation
    - https://docs.docker.com/
3. Kafka Documentation
    - https://kafka.apache.org/documentation/
4. Spark Documentation
    - https://spark.apache.org/docs/latest/

I would like to share my own articles that I wrote during this internship on Medium.

1. Installation and Introduction to Spark
    - https://medium.com/@nijatmursali/installation-and-introduction-to-spark-d38130b97ad7
2. Installation for Docker and More
    - https://medium.com/@nijatmursali/installation-for-docker-and-more-2c0a6f39777a
3. Configuration of NiFi and Kafka Docker
    - https://medium.com/@nijatmursali/configuration-of-nifi-and-kafka-docker-abb43e023d23

Acknowledgements

During the lifetime of one month and fifteen days, I have gained a lot of new information regarding to the concepts that I have mentioned above. From this reason, I would like to express my deepest thanks and gratitude to Mr. Korenkov and Mr. Belov who were my supervisors during this internship. From another perspective, I would like to give my thanks to Ivan Kadochnikov who helped me by getting more information regarding to concepts about NIFI and Spark. I would like to say that having the opportunity to come and get knowledge in Joint Institute for Nuclear Research was one of the greatest moments of my life and I'm really thankful for that.

Additionally, I would like to give my thanks to Elena Karpova who was the responsible person for international students and during the arrival and departure she helped me a lot and Elizabeth Budennaya who helped us to visit laboratories of JINR during this internship.

Once more, I would like to thank you all for creating such an opportunity for students who are interested in learning new technologies and knowledge regarding to their field.