Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science


# A Framework for Real-Time Big Data Analytics

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Computer Science**

**By**

**Rana Abdulghafoor Mohammed Taher Algalwi**

**Supervised by**

**Prof.Dr.Dhuha Basheer Abdullah Albazaz**

---

**2022 A.D.** **1443 A.H.**

# Abstract

Big data analytics enables organizations to obtain new insights and the potential to make better and faster decisions to achieve maximum benefits. Performing real-time data analytics involves arranging, preparing, and processing streaming data to gain comprehensive insights to aid appropriate decision-making. To support real-time analytics, the system effectively processes data immediately as it comes, rather than storing and retrieving it later, and it makes use of available resources to gather, analyze, and produce findings in real-time.

A framework based on a big data environment using the Apache Spark engine to perform real-time data analytics has been proposed. In addition to integration with Artificial Intelligence to analyze data in real-time and support decision-making. Modern techniques have been used to perform big data analytics in real-time, like in-memory processing and structured streaming. In addition to implementing parallel processing in which the computer's CPU was utilized across multiple cores, allowing 4 multithreaded spark instances to run effectively in the same JVM, this implies high computational power utilization and finishing several tasks within the same period.

Finance and stock markets were the case study for the proposed framework. In this field, the data represented by stock prices is characterized by being of a volatile nature, which leads to the difficulty of accurately predicting the future of these prices and helping the investors in making the decision. However, stock prices are affected by a variety of political and economic factors, investor sentiments, and people's opinions about a particular product or company, and this makes stock price predictions based on analysis of historical data insufficient. That is why, in addition to historical stock prices, the analytics of social media data was used in this work to assist the investor in making a decision. The framework includes two modes in terms of dealing with data: offline and real-time (online).

In the offline mode, a binary text classifier was built for classifying tweets according to their polarity, either positive or negative. Several machine learning algorithms were tested using SparkMLlib, a machine learning library from Apache Spark, and through the PySpark application interface. The model trained by the Support Vector Machine algorithm gained the best accuracy among the algorithms used (81.6%) and was adopted for real-time data classification. Regarding stock price data, the (LSTM) model was used to

forecast the future price of the stock. The stock of Google was used as an example, and its historical data was obtained from the Yahoo Finance API. The model was evaluated with a Mean Square Error of 0.058.

The second mode (Real-time) works with data in real-time. The Twitter platform was chosen as the source of the streaming data that will be processed using the automated model adopted in the offline stage. Processing real-time data (tweets) was performed by Spark Structured Streaming, a data flow processing engine based on Spark SQL. Initially, tweets related to the GOOGL stock were captured using Tweepy, a Python language library. Then a pipeline has been initialized to stream tweets from the Twitter API to the local system on the computer. Stream data was processed as it arrived with Spark Structured Streaming as an unlimited table in memory and within the time constraints of each batch. To aid the investor in his decision-making, the results of the analyses were shown alongside the share price display straight from the trading platforms.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب


# اطارعمل لتحليلات البيانات الضخمة في الزمن الحقيقي

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
علوم الحاسوب
من قبل

رنا عبدالغفور محمد طاهر الكَلاوي

بإشراف
الاستاذ الدكتور ضحى بشير عبدالله البزاز

1443هـ 2022م

## الملخص

تمكن تحليلات البيانات الضخمة المؤسسات من الحصول على رؤى جديدة وإمكانية اتخاذ قرارات أفضل وأسرع للحصول على أقصى قدر من الفوائد. ان إجراء تحليلات البيانات في الزمن الحقيقي يتضمن ترتيب البيانات المتدفقة وإعدادها ومعالجتها للحصول على رؤى شاملة للمساعدة في اتخاذ القرار المناسب. ولتمكين النظام من إجراء هذا التحليل، فعليه معالجة البيانات بشكل فعال بمجرد وصولها دون تخزينها واسترجاعها في وقت لاحق وإعداد النتائج في الزمن الحقيقي باستخدام الموارد المتاحة.

في هذا العمل، تم اقتراح إطار عمل لتحليلات البيانات الضخمة في الزمن الحقيقي بالاعتماد على Apache Spark بالتكامل مع تقنيات الذكاء الاصطناعي لدعم اتخاذ القرار .كان للتقنيات الحديثة دور فاعل في اجراء تحليلات البيانات في الزمن الحقيقي مثل in-memory processing , structured streaming .اضافة الى تطبيق مبدأ المعالجة المتوازية ضمن وحدة المعالجة المركزية الخاصة بالحاسوب المستخدم عبر نوى متعددة حيث تم تشغيل 4 multithread spark instance في نفس JVM بشكل فعال لتحقيق استفادة عالية في القدرة الحسابية وانهاء عدد من المهام ضمن نفس الفترة الزمنية.

تم تطبيق الاطار المقترح في مجال التمويل والاسواق المالية حيث تتميز اسعار الاسهم بكونها بيانات ذات طبيعة متقلبة، مما يؤدي إلى صعوبة التنبؤ الدقيق بالمستقبل لهذه الأسعار. اضافة الى ان أسعار الأسهم تتأثر بمجموعة متنوعة من العوامل السياسية والاقتصادية ، ومشاعر المستثمرين وكذلك آراء الناس حول منتج أو شركة معينة ، مما يجعل توقعات أسعار الأسهم بناءً على تحليل البيانات التاريخية غير كافية. لذلك تم اعتماد تحليل بيانات مواقع التواصل الاجتماعي بالإضافة إلى أسعار الأسهم التاريخية في هذا العمل لدعم المستثمر في اتخاذ القرار.ينقسم العمل مع البيانات في الاطار المقترح الى مرحلتين : ( Real-time , Offline).

كان الهدف من المرحلة الاولى (Offline) هو بناء نموذج تعلم الي ثنائي لتصنيف التغريدات وفقًا لقطبيتها الى موجبة وسالبة. حيث تم اختبار العديد من خوارزميات التعلم الآلي باستخدام SparkMLlib ,وهي مكتبة التعلم الالي من Apache Spark ومن خلال واجهة التطبيقات PySpark . اكتسب النموذج الذي تم تدريبه بواسطة خوارزمية الة المتجه الداعم Support Vector Machine أفضل نسبة دقة من بين الخوارزميات المستخدمة بنسبة(81.6) وتم اعتماده لتصنيف البيانات في الزمن الحقيقي. اما البيانات الخاصة بأسعار الأسهم ، فانه تم استخدام نموذج LSTM (الذاكرة طويلة المدى) للتنبؤ بالسعر المستقبلي للسهم ,اعتمادًا على البيانات التاريخية المكتسبة من Yahoo Finance API. وقد كانت الاسهم لشركة Google هي المثال المعتمد في هذا العمل.و قد كانت قيمة متوسط مربع الخطأ للنموذج (0.058) .

في المرحلة الثانية (Real-time) التي تعمل مع البيانات في الزمن الحقيقي،تم اختيار منصة Twitter كمصدر للبيانات المتدفقة التي سيتم معالجتها باستخدام النموذج الالي المعتمد في المرحلة الاولى.اعتمدت معالجة (التغريدات) في الزمن الحقيقي على Spark Structured Streaming ، وهو محرك معالجة لتدفق البيانات يعتمد في عمله على .Spark SQL . في البداية تم التقاط التغريدات المتعلقة بالسهم GOOGL باستخدام مكتبة Tweepy ،احدى مكتبات لغة python ليتم من بعدها انشاء خط أنابيب لتدفق التغريدات من Twitter API إلى النظام المحلي ضمن جهاز الحاسوب. ومن ثم معالجة البيانات المتدفقة بمجرد وصولها بالاعتماد على Spark Structured Streaming كجدول غير محدود ضمن الذاكرة وضمن المحددات الزمنية لكل حزمة بيانات , ومن ثم اظهار نتائج التحليلات للمستثمر بالاضافة الى عرض سعر السهم بشكل مباشر من اسواق المال لمساعدته في اتخاذ القرار.