Real-Time Big Data Analytics Pipeline

Ahmed DJEBAL[a]

[a] *Big Data Scientist at Ooredoo, Nigeria*

Keywords:

1. Introduction

In the era of the IoT, social media, e-commerce, huge volumes of data becoming available at an incredibly fast velocity, the need for an analytics system could not be more relevant. Also, the variety of data is coming from various sources in various formats, such as sensors, logs, structured data from an RDBMS, APIs, files etc. The need of the hour is having an efficient, automated, centralized and real-time big data analytics pipeline which can derive insights from data and help businesses. This paper explores building one.

2. Aim

Considering the huge volume and the incredible rate at which data is being collected,having an efficient, automated, centralized and real-time big data analytics pipeline that should have the capability to collect and extract data from multiple internal and external data sources, retain the data, process the data,deriveinsights,provideinformation,determinehiddenweaknes and/or opportunities in an acceptable time frame (so that it is not too late for the business to respond), and be flexible enough to process a variety of use cases. Also findings must be in one place that is accessible by all departments (stakeholders, managers, market research, customer experience, sales, finance, risk etc).

3. Material and methods

The major constituents of the real-time big data analytics pipeline is as follows:

• Data collection and extraction • Distribution of data to various nodes for further processing. • Analytic processing, to derive inferences fromdata,includingtheapplicationofmachine learning. • Data storage system for storing results and related information. • Interfaces or consumption of results data, e.g. visualization tools, alerts, trends etc.

And that should:

• Handle huge volumes and variety of data • Show Low Latency • Be scalable • Be diverse (serve a variety of use cases, including new and unknown ones) • Be flexible • Be economical

January 24, 2018

So in order to respond to business needs efficiently the recommended features are: • Handling high volume of data – Using a big data framework like Hadoop to retain data. • Real-time data processing – A streaming solution like Kafka coupled with Spark Streaming would be a good option. • Predictive learning– Various machine learning algorithms can be supported by Spark's MLLib library or Hadoop Mahout Library. • Storing the results and data. A NoSQL system like MongoDB could be good choice because it provides the flexibility of storing JSON data in schema less fashion. The pipeline which we are trying to build will consist of machine generated data, hence Mongo DB could be a useful candidate. • Reporting the results – For a user interface, a Tableau-like tool could be useful. Other choices may include Qlikview. Open source tools could be Jasper or Birt. Having a mature user interface will cover the aspects of historical reporting, drill down information, etc. • Alerts - e.g. Twilio, can be used to deliver Textmessages. Sending alerts through emails could also be an option

4. Results

A suitable real-time big data analytics pipeline that can be utilized by many businesses (so different scenarios). When deployed, can add several benefitsbecauseofitsgenericnature.This pipeline has the capability to apply itself in a number of domains, from healthcare to travel. It can filter anomalous data points in medical measurements, and it can filter the most travelled destinations. This pipeline provides one very important dimension to businesses: optimization.

5. Conclusions

Going forward, having a real-time big data analytics pipeline will be a pressing need for several organizations who want to derive value from data. Building such a system is acomplex task, as it requires flexibility on an unprecedented scale, not only to handle the volume of data, but a variety of data at high velocity as well. Thankfully, with the availability of technology, it is no longer an alien concept, but a reality. The flexibility can be further extended to integrate with other cloud-based machine learning systems, such as Azure ML, Yhat, etc. Integrating these services with the pipeline will make the system more usable and more versatile

6. Keywords

Big data; analytics; data science; internet of things; data scaling; framework scaling; competitive intelligence; analysis; machine learning; hadoop, spark; sql; nosql

Author biography

Ahmed DJEBAL Ahmed DJEBALI has completed his bachelor degree of computer science from the Higher National School of Computer Science, Algeria. Since then, he dedicated most of his time studying internet of things, big data and building efficient analytics systems. He is big data scientist at Ooredoo and currently managing and working on projects such as tracking data within mobile and web applications, building data lakes, behavioral analysis against telecom data, predictions, real time recommendations, sentiment analysis against social media data, etc.