Integration of Spark framework in Supply Chain Management

Mr. Harjeet Singh Jaggi, Mr. Sunny S. Kadam*

University of Mumbai, Kalina, Mumbai-400098, India

Abstract

There has been a big boom in the field of supply chain management. With almost all the processes related to supply chain are being carried out digitally the data has expanded to a limit where in it has to handled meticulously. Handling and extracting such an intricate widespread information big data analytics is the only rescue to resort to. Big data analysis runs on the Hadoop framework so far which allows an organization to store large quantity of data on disc and also process it. Due to increase in demand for data analytics, specifically speaking faster data analytics framework Hadoop is considered to be not fast enough. This is where Apache Spark came into picture 18 months ago. Apache Spark is termed to be 100 times faster than Hadoop and this allows an organization to process the same big data in a shorter span of time thus, addressing the increasing demand effectively. In this research paper we put forward the details about the Apache Spark framework while comparing it with the existing Hadoop framework alongside, how using Apache Spark framework in Supply chain management for data analytics would benefit the organization and businesses.

Keywords: Supply chain management, Data, analytics, Hadoop, Spark

1. Introduction

This research paper talks about the recent findings about apache spark on a global scale. It specifically points out the acute advantages that apache spark framework would bring along when practiced on a large scale and that how different it is from the existing Hadoop framework on various lines. This paper also portrays information about big data analytics, supply chain management and how the former two are currently integrated into one another. Finally, through this research brings about a suggestion or an experimentation for how the latest Apache Spark framework would help Supply Chain Management once successfully integrated.
1.1. Big data analytics

The term big always puts someone into jeopardy or into a series of emotions leading to meltdown. Of course in cases like ‘a big box of candy’ the above mentioned repercussions never happen but, when it’s the case where in you are given the task of managing something big and complex sure does puts in some serious repercussions. Such series of untoward expressions related to handling something big in size exists both in human as well as machine world.

When we say BIG in machine language, the only thing that comes to mind is the ‘data’. Data simply means information; this information could be anything, think about something bizarre, something so unusual that never comes your mind. Well, whatever you thought about is generally attributed as ‘DATA’ because it carries some piece of information, no matter what it is.

Now, what is ‘BIG-DATA’? when you combine several data sets consisting of a variety of data types, correlations, trends and patterns etc. it usually forms a complex and a huge cluster like formation which we call it as ‘big data’. Good thing to know that how all the intricate piece of information gets clubbed into one big cluster but, data as a cluster does not give any kind of output whatsoever, BIG-DATA is like unrefined gold, it has no value unless it is properly processed and put into right terms so as to know about the information hidden inside in detail. The process of examining the big data or data sets consists of vital business information like, market trends, customer reviews, order tracking etc. is called Big data analytic.

Big data analytics helps in figuring out the key components of a business or an organization, it’s like breaking up the cluster of data and extracting the important or relevant information from it so as to make an informed decision taking into consideration the present scenario. Big data analytics if precise in nature catapults a particular organization way ahead of its rivals in the market. This is possible because the analytical findings procured through big data analytics can lead to more accurate operational planning, better consumer services, and faster logistical operations. Big data analytics also helps in getting ready for the future as one can easily plot out the current market trends and with a little foresight the same could be converted into future trends.

Fig. 1: Visualization of daily Wikipedia edits created by IBM. At multiple terabytes in size, the text and images of Wikipedia are an example of big data.
1.2. Big data analytics in supply chain management

Supply chain management is simply the management of transport or flow of goods and services, it also includes storage, shelf life, analysis of goods procured and goods sold, logistics etc. Supply chain management helps in planning and executing various supply chain activities of a particular organization so as to build up a net value of the organization, determining the current market trend related to the demand and supply of any goods or services and also synchronizing the same for measuring the performance of the organization.

Fig. 2: Supply chain management managing complex and dynamic supply and demand networks.

Now, what exactly is the role of Big data analytics in supply chain management?

As defined earlier, big data is a huge complex cluster of data consisting useful information. The field of supply chain since its origin has expanded widely in almost all the sectors of industry right from online shopping to chemical industries. Taking this into consideration and considering the virtual global market, there are over a billions of orders from about millions of clients each day globally and hence, big data comes into existence for such a huge entity and where there’s big data, there’s always a question of processing it. This is where big data analytics comes into picture for supply chain management.
Big data helps the supplier in collecting accurate information which has a level of clarity and insight never achieved before. Analytics help in extracting such valuable resources which leads to sharing of an information across the supply chains in a contextualized manner.

Because of the widespread use of digital technologies in today’s world, companies are collecting huge amounts of data related to supply chain and so to make proper sense of such a collection data analytics comes into picture. What’s more? Big data analytics incorporates the dimensions of big data which gives the analysis a way to process the data. One of the dimensions helps the analysis to keep in mind the qualities required to process the big data. These qualities are nothing but the 3 V’s of big data namely,

- Velocity
- Variety
- Volume

The velocity quality depicts that the data should be in real time or close to real time, variety depicts that there is no fixed data model and that the data varies in time and context. Lastly, volume depicts that for every significant volume there should be a different approach. Let’s see how these qualities help in solving the big data analysis for supply chain management.

Many manufacturers believe that big data is a catalyst for greater contributions towards increasing the profit margins as majority of supply chain data is generated outside the organization, this can be proved by plotting the data sources by variety, volume and velocity by the relative level of structured or unstructured data. This is shown in the following plot:

![Fig. 3: SCM Data volume and velocity vs variety](image-url)
The amount of data generated by the supply chains today is moving on a superfast pace thus, providing ample of data to compliment contextual intelligence. From the plot it’s clear how on the basis of the qualities of big data dimension does analytics segregate so as to easily extract the most relevant or important information.

2. The Apache Spark framework for Big Data Analytics

2.1. The Origin

With the existing techniques and growing number of demand in this case ‘data’ data analytics consumes the most of the time in a given time frame. To cut time and speed things ‘APACHE SPARK’ was the framework that was released initially 18 months ago. The technology as born out of a project at University of California, Berkeley and is used for machine learning where, algorithms continuously make predictions from the same set of data. It’s basically an open source cluster computing framework which allows its user programs to load data into a cluster’s memory ad query it repeatedly.

Spark was initially started by Matei Zaharia at UC Berkeley, AMP Lab in 2009, and open sourced in 2010 under a BSD license. In 2013 the project was donated to the Apache software foundation and switched its license to Apache 2.0 in 2014, Spark became a top level Apache project. In November 2014, an engineering team at ‘Databricks’ used Spark and set a new world record in large scale sorting.

2.2. Features

Before Spark was released all the big data analytics was done with the help of another open source known as Hadoop which processed vast amounts of data using cheap off the shelf hardware. Much like Hadoop, Spark processes vast amount of data but it does it with a speed 100 times faster than that of Hadoop. This key difference is because Hadoop is supported on the backend by MapReduce which makes processing data relatively cumbersome. According to an ad network InMobi, it would take around seven months to develop a machine learning model using Hadoop whereas, currently using Spark they complete building about 4 models a day. Thus, the demand for Spark has recently surged.
Some of the notable features of Apache Spark are as follows:

- **Speed**
  
  As mentioned earlier Spark runs 100 times faster than Hadoop. It has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.

- **Ease of use**
  
  Spark offers over 80 high level operators that makes it easy to build parallel apps which makes one use it interactively from the Scala, Python and R shells.

- **Runs everywhere**
  
  Spark runs on Hadoop, Mesos, Standalone or in the cloud. It can also access diverse data sources including HDFS, Cassandra, HBase and S3

### 2.3. Apache Spark benefits for Supply Chain Management

Supply chain management churns out multiple terabytes of data every second on a global scale. This data is not only important but also sensitive and should be processed as fast as possible because, it actually determines how efficient and how well versed the organization is with its business. Faster big data analytics leads to faster decision making scenarios which indeed lead to faster growth of the organization and when we consider the current market: fast is what sells, slow is no longer an option. As mentioned earlier Big data related to supply chain management consists of information like: orders, demand and supply trends, customer services etc. addressing such an information so as to come up with an efficient solution should be a fast process so as to beat the competition at every level.

When the demand is for faster big data analytics ‘Apache Spark’ would suffice all the needs. Spark has an infrastructure which allows it to be the data processing engine which operates in-memory, on RAM; this means there’s no more reading and writing the data back to the discs or nodes like it happens in Hadoop. Currently, using the Hadoop framework underpinned by MapReduce, between every map and reduce task data has to be read to the disc and written to disc which makes it a slower process taking into consideration the huge amount of data available for processing. If in the same scenario Spark is put into use, the information would largely stay in-memory and so would the processing thus, making it iterate the data faster. How fast? Well, tests prove that Apache Spark is 100 times faster than Hadoop in memory and 10 times faster on disc. For example: Last year, Spark Pureplay Databricks used Spark to sort 100 Terabytes of records within 23 minutes thus, beating Hadoop MapReduce’s 72 minutes.

In short, we can term Spark as the standard data processing framework. Apart from its aggressive speed Spark also incorporates a few of the Big Data bonuses:

- Spark SQL and SQL for structured data processing
- MLlib for machine learning
- GraphX for graph processing
- Spark streaming, for streaming data
Completely replacing the existing technology with something new is a kind of an upgrade nobody wants. It consumes time, work force, infrastructure utilities etc. Apache Spark seamlessly runs on the existing data analytics framework Hadoop thus, making it a complement to Hadoop thus, saving all the time and efforts which would have been required for an overall upgrade. Thus, we can say, Spark is here to enhance the existing system to give out faster and more effective data analytics. When we look at this from a supply chain management point of view, an organization would benefit in many ways like, upgrading the framework without spending a lot of money as they can improve their processing capabilities by combining Spark with Hadoop MapReduce, HBase and other big data frameworks. Apache Spark being highly compatible in nature is easy for every Hadoop user to take advantage of Spark’s capabilities and it doesn’t matter if we use Hadoop 1.0 or YARN, Spark runs on all of it.

![Fig. 5: Ways to deploy Spark in Hadoop Cluster](image)

3. Conclusion:

Like any other new technology, Spark tends to be immature for any organization to adopt it on a large scale. On paper it seems really easy just by comparing it with the existing framework i.e. the Hadoop but, it seems complicated to imply in the manner stated before. Yes, there is scope but, instead of completely replacing Hadoop, Spark could just complement it as in a reinforcement so as to make the analytics easier and faster than usual. Like, MapReduce Spark runs on Hadoop but, unlike MapReduce it is not limited to Hadoop. All that Spark needs is a set of best practices and some tools so as to figure out its actual potential and implementation.

**Acknowledgment**

We hereby extend our respectful gratitude towards ICCCV 2016 for giving us the opportunity to carry out an extensive research on ‘Integration of Spark framework in Supply Chain Management’

We further extend our gratitude towards the University of Mumbai who has been our alma mater throughout our course of academic study which has made us capable of carrying out such projects.
References

6. Figure 1 attribution: "Viegas-UserActivityonWikipedia" by Fernanda B. Viégas - User activity on Wikipedia. Licensed under CC BY 2.0 via Commons - https://commons.wikimedia.org/wiki/File:Viegas-UserActivityonWikipedia.gif#/media/File:Viegas-UserActivityonWikipedia.gif
8. Figure 3 source: Big Data Analytics in Supply Chain Management: Trends and Related Research. Presented at 6th International Conference on Operations and Supply Chain Management, Bali, 2014
9. Figure 4 source: https://sparkhub.databricks.com/wp-content/uploads/2015/06/sparklogo_homepage.png