ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Integration of Spark into Splunk for High-Scale Datasets to augment Performance, Scalability, Liability and Strength

Rohini More [1], Smita Konda [2]

Assistant Professor, Department of Computer Engineering, A.G.Patil Institute of Technology, Solapur, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, A.G.Patil Institute of Technology, Solapur, Maharashtra, India[2]

ABSTRACT: The past few years have seen huge awareness in large-scale data analysis. As we are aware about the fact that data volumes in both industry and research are keep on growing. So we need high processing individual machines. Hence to handle such large clusters of information Google's MapReduce model and it's open source implementation, Hadoop is used. We used various Hadoop parallel data analysis tools such as Apache's Hive & Pig engines for SQL processing & to handle large clusters. Though, these tools have been optimized for one pass batch processing of on-disk data, which makes them time-consuming for interactive data discovery and for more complex multipass analytics algorithms that are becoming ordinary. In this article, we are introducing Spark & Splunk. Spark- a new cluster computing framework that can execute applications up to 40× faster as compared to hadoop. It keeps data in memory and can be used interactively to query huge datasets with sub-second latency. Splunk- a platform for machine data. It collects, indexes, and harnesses machine data generated by any IT system and infrastructure—whether it's virtual, physical or in the cloud. Splunk laid its foundation helping IT for finding and fixing problems faster. So in this article we will discuss how to make use of Splunk with spark for performing analysis and machine learning on data.

KEYWORDS: Apache Spark, Splunk, RDD, machine data.

I. INTRODUCTION

Apache Spark is a high-speed and general-purpose cluster computing system. It abstracts data as a distributed collection of Resilient Distributed Data set (RDD) that can be operated in parallel with each other. Resilient Distributed Datasets (RDD) is a primary data structure of Spark. Every dataset in RDD is divided into logical partitions. It may be computed on different nodes of the cluster. RDDs can have any type of scala, java or python objects, as well as user-defined classes. Apache Spark has emerged as a widely used open-source engine. Spark is a fault-tolerant and general-purpose cluster computing system providing APIs in Java, Scala, Python, and R,along with an optimized engine that supports general execution graphs. Moreover, Spark is efficient at iterative computations and is thus well-suited for the development of large-scale machine learning applications. Spark enables the development of efficient implementations of large-scale machine learning algorithms since they are typically iterative in nature.

An emerging class of data is being generated by web servers, applications, machines, on-premises applications, and SaaS systems of all types. It includes click streams, GPS readings, call logs, RSS feeds, social media comments, weather data, fleet locators, and more. This data is one of the fastest growing categories of big data. Today's systems generate massive amounts of information that can be ignored, used for finding and fixing problems. Until now, analysis of this data focused on finding out about machines and their operations. As exponential growth in machine data over decades, due to growing number of machines in the IT infrastructure, the machine data has a lot of valuable information that can drive efficiency, productivity and visibility for the business. Machine data is complex to

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

understand, not suitable for making analysis or visualization, in an unstructured format. Splunk is a tool used to store and process large amounts of data. Splunk is a software platform to analyze, visualize and search the machine-generated data gathered from the websites, applications, sensors, devices etc. which make up your IT infrastructure and business.

This is how splunk collects data from multiple systems. Splunk's mission is to make machine data accessible across an organization by identifying data patterns. Splunk captures, indexes, and correlates real-time data in a searchable repository from which it can generate graphs, reports, alerts, dashboards, and visualizations.

II. RELATED WORK

In paper [4], Apache Spark is a general-purpose cluster computing engine with APIs in Scala, Java and Python and libraries for streaming, graph processing and machine learning. Spark was handling important feature to manipulate distributed collections of data called Resilient Distributed Datasets (RDDs). Each RDD is a collection of languages like Java, Python objects partitioned across a cluster. RDDs can be manipulated through operations like map, filter, and reduce, which take functions in the programming language and distribute them to nodes on the cluster. Spark enables difficult workflows where any SQL queries can be used to pre-process data and the results can then be analyzed using advanced machine learning algorithms. In link [2], the authors have discussed about a new programming language with a number of extensions that support data processing and machine learning tasks is the R language. An R package provides a frontend to Apache Spark . It also allows users to run large scale data analysis using Spark's distributed computation engine.

For complex and huge data processing and analysis was not only possible with Map reduce, Hadoop. So authors have found this technique called Splunk. Splunk handles complex data, huge data and in an unstructured format

III. HOW SPARK WORKS?

In MapReduce, data/information sharing is slow due to serialization replication, and disk IO. To overcome this drawback, researchers developed a specialized framework called Apache Spark. The key idea of spark is Resilient Distributed Datasets (RDD). Abstract data as a distributed collecton of RDD that can be operated in parallel. Spark uses the concept of RDD to accomplish efficient and more rapid MapReduce operations. Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark[5]. RDD supports in-memory processing computation. It means, it stores the state of memory as an object across the jobs and the object is sharable between those jobs. RDD performs the following operations

–Transformations: create a new dataset from an existing one

– Actions: return a value after running a computation on the dataset. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes. RDDs can be created through deterministic operations on either data on stable storage. An RDD is a read-only, partitioned collection of records. RDD is a fault-tolerant collection of elements that can be operated on in parallel. Parallelizing and referencing a dataset are two ways to create RDDs where parallelizing an existing collection in your driver program, or referencing a dataset in an external storage system, such as HDFS, HBase, a shared file system[6].

Iterative Operations on Spark RDD

Iterations operations 1,2..n perform read and write operations. Data is taken from disk and iterations are performed and the results are stored in the distributed memory instead of disk, to make the system faster [5]. If the Distributed memory (RAM) is not enough to store intermediate results, then it will store those results on the disk.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

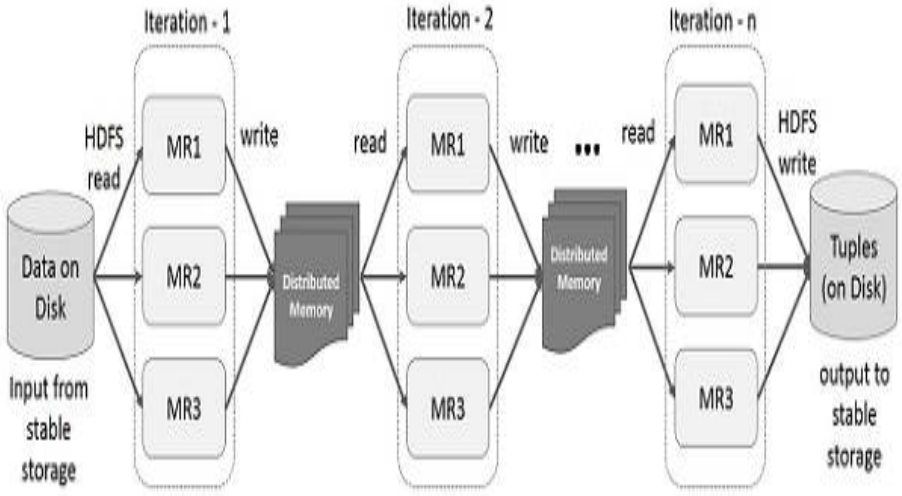

Fig 1: Iterative Operations on Spark RDD [5]

Interactive Operations on Spark RDD

If different queries are run on the same set of data repeatedly, this particular data can be kept in memory for better execution times[5].

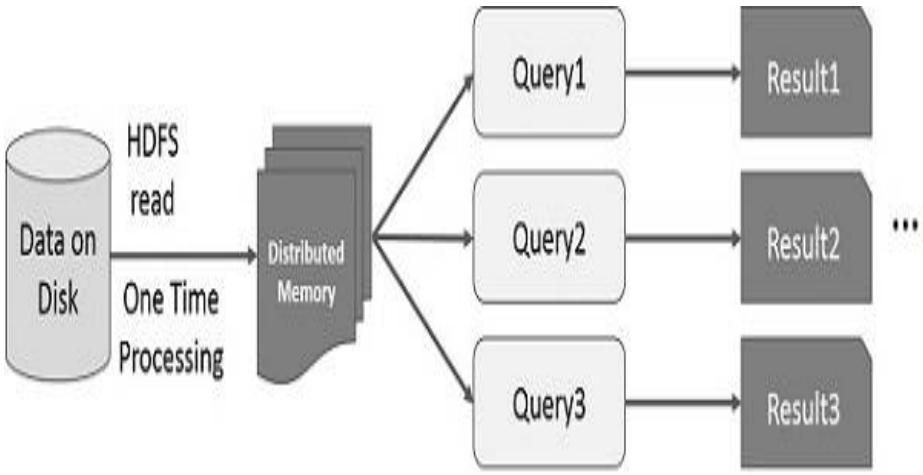

Fig 2: Iterative Operations on Spark RDD [5]

IV. BENEFITS OF SPLUNK

If we have a machine which is generating data constantly and if we want to analyze the machine state in real time, then how will we achieve it? Can we do it with the help of Splunk? Yes! we can. The image given below will help us to see how Splunk collects data from multiple sources.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

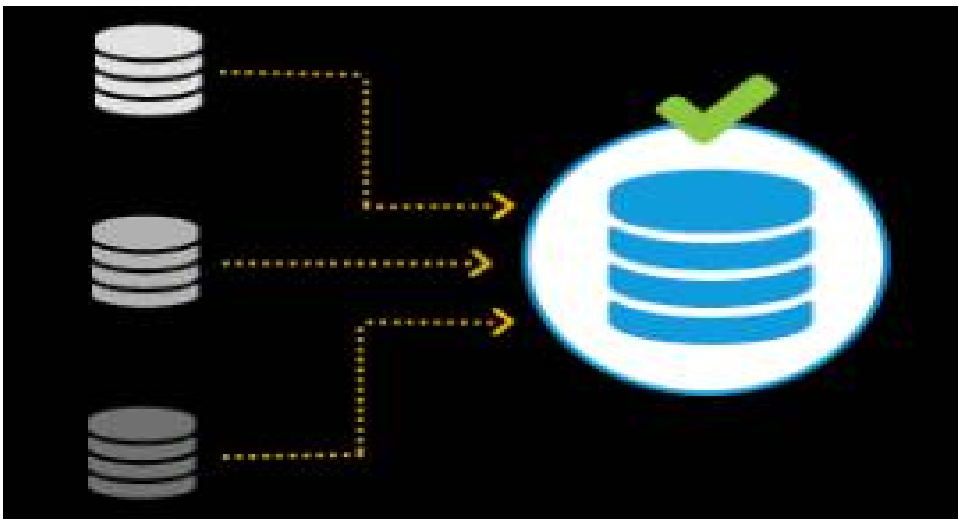

Fig.3: Splunk: Collects data from multiple systems

The benefits with implementing Splunk are[6]:

- Your input data can be in any format for e.g. .csv, or json or other formats
- You can configure Splunk to give Alerts / Events notification at the onset of a machine state
- You can accurately predict the resources needed for scaling up the infrastructure
- You can create knowledge objects for Operational Intelligence

V. SPLUNK WITH SPARK

How Splunk with Spark works?

When splunk works with spark, beyond the data processing, they have ability to do analysis and machine learning on data. On the other hand, when we extend splunk with spark, both can perform distributed computing for complex operations. Impressive cache by Spark stack is readily available. Users can write their own spark programs modified to their dataset. We can also connect other data sources, through RDD/DStream.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

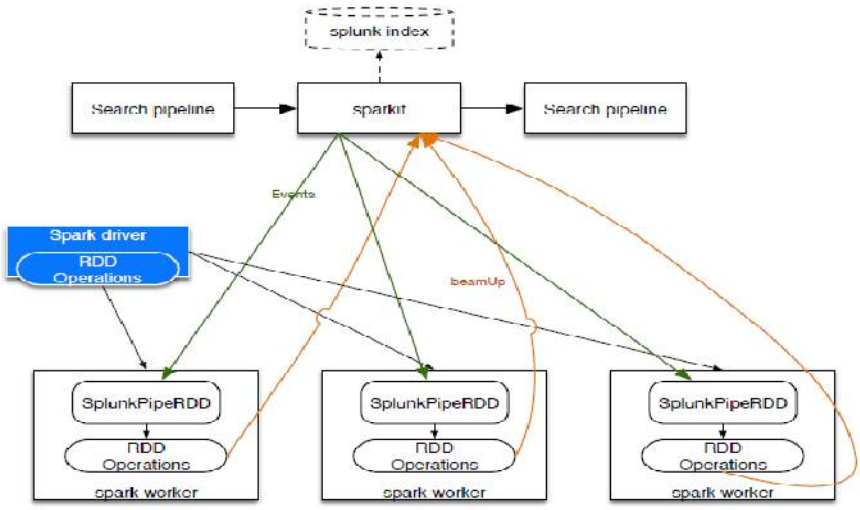

Fig 4: Architecture [7]

As shown in above figure, We use search command sparkit, which is starting point to distribute search pipeline results to spark. SplunkPipeRDD is a RDD which pulls data from search pipeline (sparkit). "beamUp" is a Custom RDD operation used to push computation results to search pipeline

VI. CONCLUSION

In paper we discussed the role of Spark using RDD to achieve efficient and faster MapReduce operations. Data sharing and analysis of such huge and complex data is processed by RDD. It supports in-memory processing computation it stores the state of memory as an object across the jobs and the object is sharable between those jobs. Data sharing in memory is 10 to 100 times faster than network and Disk.

REFERENCES

[1] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, "Fast and Interactive Analytics over Hadoop Data with Spark", https://www.usenix.org, pp. 45-51.
[2] Shivaram Venkataraman, Zongheng Yang, Davies Liu "SparkR: Scaling R Programs with Spark.", SIGMOD Conference, 2016.
[3] Xiangrui Meng, Joseph Bradley Burak Yavuz, Evan Sparks, "MLlib: Machine Learning in Apache Spark", Journal of Machine Learning Research 17 (2016) 1-7.
[4] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, et al. Spark SQL:Relational data processing in Spark. In SIGMOD, pages 1383–1394, 2015.
[5] Franklin Jr., Curtis (16 October 2015). "SAP, Splunk Dashboards Aim To Satisfy Data Hunger". InformationWeek. Retrieved 24 March 2016.
[6] M. Armbrust, T. Das, A. Davidson, A. Ghodsi, A. Or, J. Rosen, I. Stoica, P. Wendell, R. Xin, and M. Zaharia. Scaling spark in the real world: performance and usability. Proceedings of the VLDB Endowment, 8(12):1840–1843, 2015.
[7] Evan R Sparks, Ameet Talwalkar, Daniel Haas, Michael J. Franklin, Michael I. Jordan, and Tim Kraska. Automating model search for large scale machine learning. Symposium on Cloud Computing, 2015.
K. Zeng et al. G-OLA: Generalized online aggregation for interactive analysis on big data. In SIGMOD, 2015.