Techniques Management Big data Apache Hadoop and Apache Spark and which is better In structuring and processing data Ahmed M Altriki Department of Information System IT College, University of Benghazi Benghazi, Libya Ahmed.Altriki@uob.edu.ly Osama Alarafee Department of Information System IT College, University of Benghazi Benghazi, Libya Osama.Alarafee@uob.edu.ly Abstract: Newly considered both Big data Apache Hadoop and Apache Spark It is an important techniques in managing huge databases and thus this article provides an overview of Hadoop MapReduce.Big data analytics in the cloud: Spark on hadoop vs mpi/openmp on beowulf.Standard Hadoop Apache Spark Map reduce Data Processing Slow as MapReduce operates in various sequential steps Its real-time data processing capability makes Spark a top choice for big data analytics Real-Time MapReduce Analysis fails when it It can process real- time data comes to real- time data processing, as it was designed to perform batch processing on voluminous amounts of data Ease of Use easy to use Spark easier to use than Hadoop Graph Processing Slow in graphic processing Fast in graphic processing Fault Tolerance Hadoop achieves fault tolerance through replication (have good fault tolerance ability more tolerant than Spark) (have good fault tolerance ability )Spark uses RDD and various data storage models for fault tolerance by minimizing network I/O Security Hadoop MapReduce has better security features than Spark Spark's security is currently in its infancy, offering only authentication support through shared secret (password authentication) Cost Low cost High cost Compatibility yes yes Terms of Fast performance Performance is 100 times speedier than Hadoop in performance Resilience is naturally resilient to system faults or failures as data are written to disk after every operation has built-in resiliency by virtue of the fact that it arranges data in Resilient Distributed Datasets yes yes fault tolerance ability yes yes 6. References [1] Chavan, V . and Phursule, R.N., 2014. Survey paper on big data. Int. J. Comput. Sci. Inf. Big data management processing with Hadoop MapReduce and spark technology: A comparison. 1- 4).In this research clarification of previous studies of a number of research and then was studied all points of difference and similarity in the comparison of Apache Spark vs Hadoop Map reduce Clarify which is best in handling and storing data in terms of several criteria (security, performance, flexibility, cost, processing speed, etc.) As a future work, we seek to expand our study of more papers in this field and apply one of the two methods to a case study and to obtain results that contribute to explain the differences in more details.The working and architecture of Spark when it is used along with Hadoop YARN is shown by the authors Wei Huang, Lingkui Meng, Dongying Zhang, and Wen Zhang [3].On a similar note, Spark can also be integrated into any cloud- based data platform, besides HDFS, whose data can be used for its analytics function.[6] We have also tried to establish criteria for a clearer and more comprehensive comparison and to shed some more light on the Spark vs Hadoop debate, let's take a look at each of them separately.Also spoken by Priya Dahiya 1, Chaitra.B 2, Usha Kumari [4] regarding their search which focus on the architecture and working of Apache Hadoop and Apache Spark and the challenges faced by MapReduce.The two main key concepts used in Apache Spark are Resilient Distributed Datasets (RDD) and Directed Acyclic Graph (DAG)[3].Neither are they exclusive of each other.[7] Also, several big data projects require installing Spark on top of Hadoop so that the advanced analytics applications of Spark can work on the data stored by Hadoop Distributed File System (HDFS).IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(1), pp.3-19 Open

Source [4] Marone, R.M., Camara, F. and Ndiaye, S., 2017, December.Discussion After studying considering the previous studies in this research remains There is an increasing curiosity among big data professionals to choose the best framework between Apache Spark and Hadoop, often mistaking them to be the same.IEEE [6] Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J. and Ghodsi, A., 2016.MapReduce is one of the key approaches to meet the demands of computing massive datasets as well as it processes information in the nodes by executing parallelly in the system.Apache Spark is a cluster computing framework which is open-source and used to program clusters with implicit parallelism and fault- tolerance.Processes used for applications are assigned uniquely i.e. they all have their processes and due to these tasks run in multiple threads, and they must have connectivity to worker nodes.Spark uses (DAG) to divide operators into many stages of tasks and (RDD) for fault tolerance.Finally, Katarina Grolinger, Michael Hayes [5] proposed a paper which describes briefly about the challenges faced by MapReduce while processing this huge amount of data.Task Tracker of slave node gets the tasks to be completed from the Master node Job Tracker.[5] 3.2 Spark Architecture It is an open-source big data framework.Spark supports machine learning, SQL queries, graph data processing and streaming data for analysis of big data [4].Architecture of Spark 3.2.1 Resilient Distributed Dataset (RDD): it works as a collection of elements which are operated in parallel.In the distributed file system Spark runs on Hadoop cluster, and RDD is created from files in the format of text or sequence files.Apache Spark is used in big data increase the number of users of social networking sites of various types or industries.Introduction: Recently, Big data is gaining popularity with the increase in the number of users of social networking sites of various types.N. Pursue [1] describes what is Big data and what are five V's, i.e., velocity, volume, veracity, variability, variety.Hadoop consists of HDFS and MR. It consists of two layers (HDFS Layer and MapReduce layer) Fig 1.Name node has information about all Data Nodes i.e., data stored by them, which nodes are working actively and which are not i.e. passive nodes, about the free space and job tracker is working efficiently or not.As shown in fig.1 MapReduce layer has a Job Tracker which is used for assigning tasks to the Task Tracker 3.Spark consists of cluster manager, driver program (spark context), executor or worker and HDFS as shown in fig 2.In-memory parallel processing of massive remotely sensed data using an apache spark on hadoop yarn model.Traditional data processing techniques are unable to store and process this huge amount of data, and due to this, they face many challenges in processing Big Data and demands of the end user.Traditional data processing techniques are unable to store and process this huge amount of data, and due to this, they face many challenges in processing Big Data and demands of the end user.Hadoop and Spark Architecture 3.1 Hadoop Architecture Hadoop Map Reduce: It is an open-source framework for writing applications.1.2.3.4.5.