Apache Spark in Mainframe Systems

James Ash

Wentworth Institute of Technology

Author Note

First paragraph: Introduction to Spark

Second paragraph: Importance in the Industry

Third paragraph: Application in Enterprise Computing

Fourth paragraph: Data movement in Mainframe Systems

Fifth paragraph: Compliance and Security

Sixth paragraph: Conclusion

Apache Spark is an open-source framework used for high performance, scalable, data processing. Spark lets you connect multiple computers or nodes to work on different parts of your data sets at the same time. This method is called parallel computing. Spark is designed to be fast. Other traditional parallel computing systems read data directly from disk which can be slow. Spark stores everything in memory (RAM) which speeds up the data reading and writing process.

Importance in the Industry

Today in the world all organizations are looking for ways to process and manipulate data at a large scale efficiently. Spark, due to its speed, reliability, and versatility has become a staple in the world of big data with over 1000 organizations running on Apache Spark. Sparks' technique of managing data in memory as opposed to the traditional disk memory management of traditional systems allow engineers to achieve noticeable performance increases. Spark's infrastructure is essentially the same as previous frameworks with this minor but significant change, allowing existing mainframe infrastructures to completely switch or work alongside Spark to achieve its performance gains.

Application in Enterprise Computing

Enterprise computing is usually run on legacy systems such as IBM z/OS. These legacy systems are used to manage mission-critical applications and vast amounts of transactional data. These systems are used for their reliability and efficiency as those are the two most important aspects in enterprise computing. While these systems are reliable and efficient, they were not designed to handle the data driven world that we have developed into. Significant improvements

were needed to enhance these legacy systems and bring them up to speed in the modern world. To bridge this gap, organizations have increasingly turned to modern technologies like Apache Spark, which can be integrated with legacy systems without requiring a complete overhaul. By adding the power of distributed computing and in-memory processing, Spark enables these traditional systems to handle larger, more complex datasets, and run advanced analytics in real time. With Sparks real time analytics capabilities it allows industries to develop more dependable features such as fraud detection, or machine learning.

Data movement in Mainframe Systems

The only issue with using Spark is having to move data from the mainframe to Spark to allow it to use its speed to compute data. Mainframes store an absurd amount of data. First you will need to move the data from the Mainframe to a different storage system. Once the data is ready Spark processes it quickly by splitting up the task across different nodes. Once the processing is complete the user can decide where the data goes next.

Compliance and Security

The industries that use mainframes are very highly regulated industries. Following rules and regulations are very important. Spark supports encryption and uses authorization tools to make sure only users who are supposed to see the data are able to do so.

Conclusion

Spark on mainframes is a way to bring old systems into the modern world. It lets users analyze and manipulate data faster. For industries that rely on mainframes, this allows them to update their aging infrastructure for the modern day. As technology and user demand keeps

expanding, technologies like Spark will allow mainframes to stay the powerhouse of computing infrastructure.

References

Apache Spark. (2021). *Apache Spark Overview*

Retrieved from https://spark.apache.org/

IBM. (2020). *Mainframe and Apache Spark: Modernizing Legacy Systems*

Retrieved from https://www.ibm.com/blogs/mainframe/apache-spark/

TechTarget. (2022). *Mainframe Modernization with Spark*

Retrieved from https://www.techtarget.com

Databricks. (2023). *Real-time Data Processing with Spark*

Retrieved from https://databricks.com

Gartner. (2022). *The Role of Spark in Enterprise Compliance*

Retrieved from https://gartner.com