A Short Glossary of High Performance Computing and Big Data Terms

Rajani Kumar Pradhan & Pedro Hernandez Gelado

July 2021

1 Introduction

Glossary

Apache Hadoop

is a data storage and processing platform.

Apache Spark

is a framework for large-scale data processing

Dataframe

From Databricks: "A Spark DataFrame is a distributed collection of data organized into named columns. It is conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer optimizations under the hood." [9]

domain knowledge

valid knowledge used to refer to an area of human endeavour, an autonomous computer activity, or other specialized discipline

ETL

Extract, transform, load

MR MapReduce

flatMap is an extension of the vanilla map function, that is used when you need to return an interable, expandable input, e.g. multiple rows from a single row.

HPC

High Performance Computing

HDFS

Hadoop Distributed File System

MR MapReduce

is a distributed programming paradigm for parallel data processing.

MPI Message Passing Interface

is a library standard for distributed memory parallelization allowing writing portable parallel programs for all kinds of parallel systems.

RDD Resilient Distributed Dataset

An RDD is a collection of records that is partitioned and can be acted on in parallel [5]. RDDs can be arbitrary Java, Scala or Python objects, and they are the basis of data storage in Apache™ Hadoop ® and Apache ™ Spark ®

UDF

User Defined Functions

YARN Yet Another Resource Negotiator

is a resource and node manager and job scheduler for Hadoop.

References

[1] White T. *Hadoop: The Definitive Guide*. O'Reilly, 1988.

[2] Apache Software Foundation. *Hadoop*. URL: https://hadoop.apache.org.

[3] Apache Software Foundation. *Hadoop*. URL: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html.

[4] Little Big Data cluster. *Hadoop*. URL: https://lbd.zserv.tuwien.ac.at/.

[5] Bill Chambers and Matei Zaharia. *Spark: The definitive guide: Big data processing made simple*. " O'Reilly Media, Inc.", 2018.

[6] Warren J. and Marz N. *Big Data*. Manning publications, 1988.

[7] M. Zaharia et al. "Spark: Cluster Computing with Working Sets". In: *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*. HotCloud'10. Boston, MA: USENIX Association, 2010.

[8] F. Yin and F. Shi. "A Comparative Survey of Big Data Computing and HPC: From a Parallel Programming Model to a Cluster Architecture". In: *Int J Parallel Prog* (2021). DOI: https://doi.org/10.1007/s10766-021-00717-y.

[9] *Dataframe Description*. https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html#:~:text=In%20Spark%2C%20a%20DataFrame%20is,richer%20optimizations%20under%20the%20hood.. Accessed: 2021-09-01.

[10] European associations for HPC (ETP4HPC.eu) and Big Data Value (BDVA.eu). *THE TECHNOLOGY STACKS OF HIGH PERFORMANCE COMPUTING AND BIG DATA COMPUTING:What they can learn from each other*. Tech. rep. 2018. URL: https://www.etp4hpc.eu/pujades/files/bigdata_and_hpc_FINAL_20Nov18.pdf.

[11] Intel® Corporation. *Big Data Meets High Performance Computing*. Tech. rep. 2014. URL: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-meets-high-performance-computing-white-paper.pdf.

[12] A.R. Pathak, M. Pandey, and S.S. Rautaray. "Approaches of enhancing interoperations among high performance computing and big data analytics via augmentation". In: *Cluster Comput* 23 (2019). DOI: https://doi.org/10.1007/s10586-019-02960-y.