Data Engineering Notes

Created: 2024-10-30

Updated: 2024-10-30

References:

- https://www.coursera.org/learn/spark-hadoop-snowflake-data-engineering
- https://learn.microsoft.com/en-us/training/modules/use-apache-spark-azure-databricks/
- ChatGPT 4o mini
- Google search

Apache Hadoop: open source ecosystem of software enabling parallel processing of big data.

- Hadoop Distributed File System (HDFS): storage system.
- MapReduce: framework within Hadoop that:
    - maps (distributes) tasks across a cluster of computers that store intermediate results (key-value pair) on disk (slower than in-memory).
    - Intermediate results are grouped by keys and sent to the cluster of computers (parallel processing) responsible for reducing (performing user reduction function) and results are combined.

Apache Spark: built on top of Hadoop but stores intermediate results in-memory instead of on disk. Leverages parallelism for task completion via horizontal scaling (more nodes).

- Each worker/executor node runs a JVM (multi-threaded). Each has multiple slots based on #cores #cpus of node.
- Parallelized jobs are broken down into stages to be performed in order.

Azure Databricks: parallelized data processing on Apache Spark clusters.

- A notebook instance (*SparkSession* object) controls the driver node which distributes work across worker nodes.