What is Apache Hadoop in Azure HDInsight?

Article • 04/24/2023

Apache Hadoop was the original open-source framework for distributed processing and analysis of big data sets on clusters. The Hadoop ecosystem includes related software and utilities, including Apache Hive, Apache HBase, Spark, Kafka, and many others.

Azure HDInsight is a fully managed, full-spectrum, open-source analytics service in the cloud for enterprises. The Apache Hadoop cluster type in Azure HDInsight allows you to use the Apache Hadoop Distributed File System (HDFS), Apache Hadoop YARN resource management, and a simple MapReduce programming model to process and analyze batch data in parallel. Hadoop clusters in HDInsight are compatible with Azure Blob storage, Azure Data Lake Storage Gen1, or Azure Data Lake Storage Gen2.

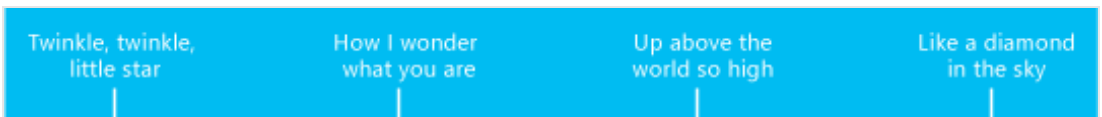
To see available Hadoop technology stack components on HDInsight, see Components and versions available with HDInsight. To read more about Hadoop in HDInsight, see the Azure features page for HDInsight.

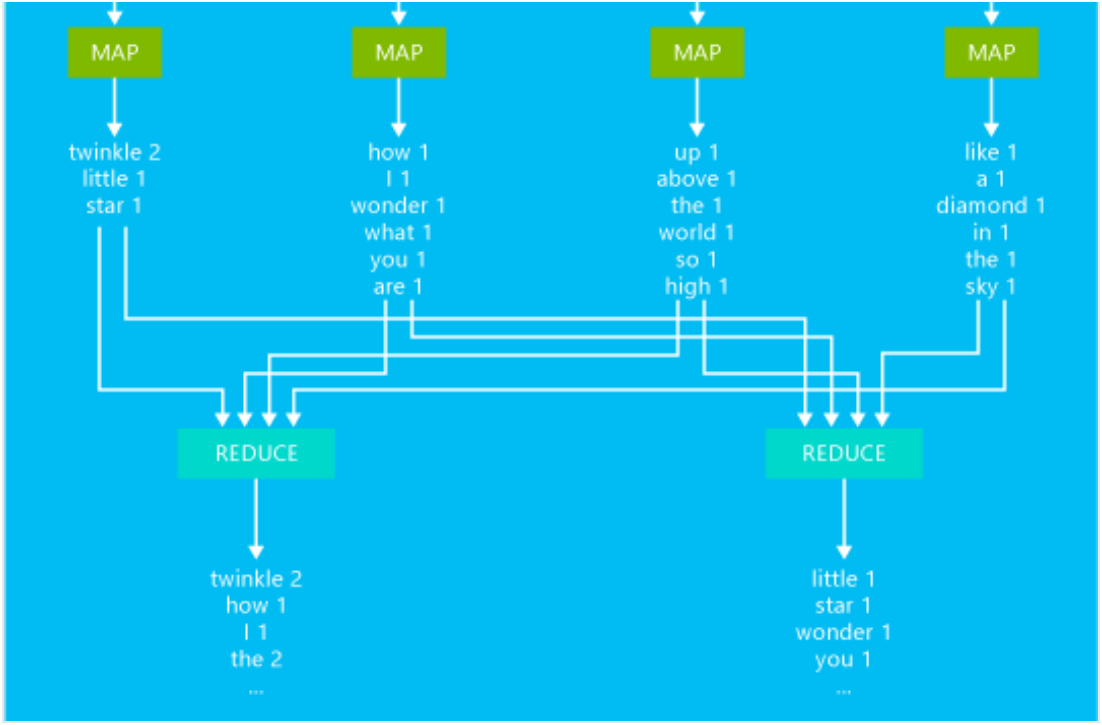
What is MapReduce

Apache Hadoop MapReduce is a software framework for writing jobs that process vast amounts of data. Input data is split into independent chunks. Each chunk is processed in parallel across the nodes in your cluster. A MapReduce job consists of two functions:

- **Mapper**: Consumes input data, analyzes it (usually with filter and sorting operations), and emits tuples (key-value pairs)
- **Reducer**: Consumes tuples emitted by the Mapper and performs a summary operation that creates a smaller, combined result from the Mapper data

A basic word count MapReduce job example is illustrated in the following diagram:





The output of this job is a count of how many times each word occurred in the text.

- The mapper takes each line from the input text as an input and breaks it into words. It emits a key/value pair each time a word occurs of the word is followed by a 1. The output is sorted before sending it to reducer.
- The reducer sums these individual counts for each word and emits a single key/value pair that contains the word followed by the sum of its occurrences.

MapReduce can be implemented in various languages. Java is the most common implementation, and is used for demonstration purposes in this document.

Development languages

Languages or frameworks that are based on Java and the Java Virtual Machine can be ran directly as a MapReduce job. The example used in this document is a Java MapReduce application. Non-Java languages, such as C#, Python, or standalone executables, must use **Hadoop streaming**.

Hadoop streaming communicates with the mapper and reducer over STDIN and STDOUT. The mapper and reducer read data a line at a time from STDIN, and write the output to STDOUT. Each line read or emitted by the mapper and reducer must be in the format of a key/value pair, delimited by a tab character:

```
[key]\t[value]
```

For more information, see Hadoop Streaming .

For examples of using Hadoop streaming with HDInsight, see the following document:

- Develop C# MapReduce jobs

Where do I start

- Quickstart: Create Apache Hadoop cluster in Azure HDInsight using Azure portal
- Tutorial: Submit Apache Hadoop jobs in HDInsight
- Develop Java MapReduce programs for Apache Hadoop on HDInsight
- Use Apache Hive as an Extract, Transform, and Load (ETL) tool
- Extract, transform, and load (ETL) at scale
- Operationalize a data analytics pipeline

Next steps

- Create Apache Hadoop cluster in HDInsight using the portal
- Create Apache Hadoop cluster in HDInsight using ARM template