

Data Engineering 101 Hadoop Q&As

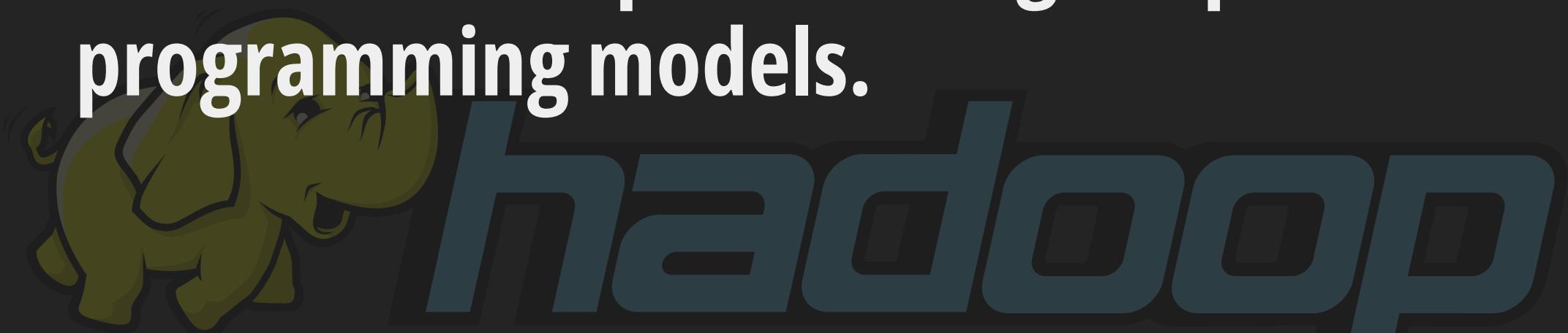


Shwetank Singh
GritSetGrow - GSGLearn.com



What is Hadoop?

Hadoop is an open-source framework designed for distributed storage and processing of large datasets across clusters of computers using simple programming models.



What is a SequenceFile in Hadoop?

A SequenceFile is a flat file consisting of binary key/value pairs. It's used to combine many small files into a larger file for more efficient storage and processing.



Shwetank Singh
GritSetGrow - GSGLearn.com



Explain Hadoop's architecture

Hadoop's architecture consists of two main components:

- 1) HDFS (Hadoop Distributed File System) for storage,
- 2) YARN (Yet Another Resource Negotiator) for resource management and job scheduling.

MapReduce is the processing paradigm.

Explain the concept of Partitioner in Hadoop

A Partitioner determines which reducer an output key-value pair from a mapper should be sent to. It helps in load balancing by ensuring even distribution of data across reducers.



Shwetank Singh
GritSetGrow - GSGLearn.com



What are the core components of Hadoop?

The core components are:

HDFS,

YARN,

MapReduce, and

Hadoop Common (libraries and utilities).

What is the purpose of InputFormat in Hadoop?

InputFormat defines how input data is split and read. It validates input specs and splits input files into logical InputSplits, then creates RecordReader to read them.



Shwetank Singh

GritSetGrow - GSGLearn.com



Describe HDFS architecture

HDFS has a master/slave architecture with a NameNode (master) managing the file system metadata and DataNodes (slaves) storing the actual data.



What is the role of OutputFormat in Hadoop?

OutputFormat specifies how the output files of a MapReduce job should be written. It provides a RecordWriter to write the output key-value pairs to files.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is a NameNode in Hadoop?

The NameNode is the master node in HDFS that manages the file system namespace and regulates access to files by clients.



Describe the function of RecordReader in Hadoop

RecordReader reads data from input splits and converts them into key-value pairs that can be processed by the Mapper. It's responsible for parsing input data.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is a DataNode in Hadoop?

DataNodes are the slave nodes in HDFS that store and retrieve blocks when told to by clients or the NameNode. They report back to the NameNode periodically with lists of blocks they store.

What is a JobConf in Hadoop?

JobConf represents a MapReduce job configuration. It specifies the MapReduce job parameters like input/output paths, input/output formats, mapper and reducer classes, etc.



Shwetank Singh
GritSetGrow - GSGLearn.com



Explain the concept of blocks in HDFS

In HDFS, files are divided into large blocks (typically 128MB or 256MB) for efficient storage and processing. Each block is replicated across multiple DataNodes for fault tolerance.



Explain the concept of Distributed Cache in Hadoop

Distributed Cache is a facility provided by MapReduce framework to cache files needed by applications. Files are copied to all the nodes before job execution starts.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is YARN and its role in Hadoop?

YARN (Yet Another Resource Negotiator) is Hadoop's cluster resource management system. It separates the resource management and job scheduling/monitoring functions.



What is the purpose of `fs.defaultFS` property in Hadoop?

`fs.defaultFS` specifies the default file system to be used. It's typically set to the NameNode's address and port, like `hdfs://namenode:8020/`.

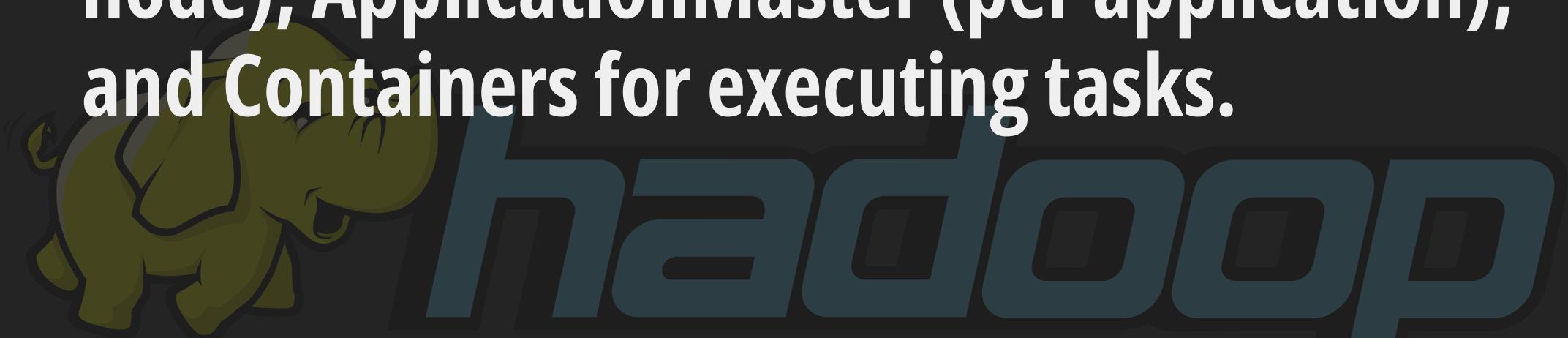


Shwetank Singh
GritSetGrow - GSGLearn.com



Describe the components of YARN

YARN consists of a ResourceManager (master), NodeManagers (on each slave node), ApplicationMaster (per application), and Containers for executing tasks.



How does data replication work in HDFS?

HDFS replicates each block of a file multiple times (default is 3) and distributes these replicas across the cluster for fault tolerance and availability.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is MapReduce?

MapReduce is a programming model and processing technique for distributed computing. It consists of a `Map()` function that processes input data and a `Reduce()` function that aggregates the results.



What is the role of the edit log in HDFS?

The edit log in HDFS records all changes to file system metadata. It's maintained by the NameNode and is crucial for maintaining consistency of the file system.



Shwetank Singh
GritSetGrow - GSGLearn.com



Explain the MapReduce workflow

MapReduce workflow:

Input data → Splitting → Mapping → Shuffling → Reducing → Final output.

Each step distributes the workload across the cluster.

Explain the concept of heartbeats in Hadoop

Heartbeats are periodic signals sent from DataNodes to the NameNode and from NodeManagers to the ResourceManager to indicate they are alive and functioning.

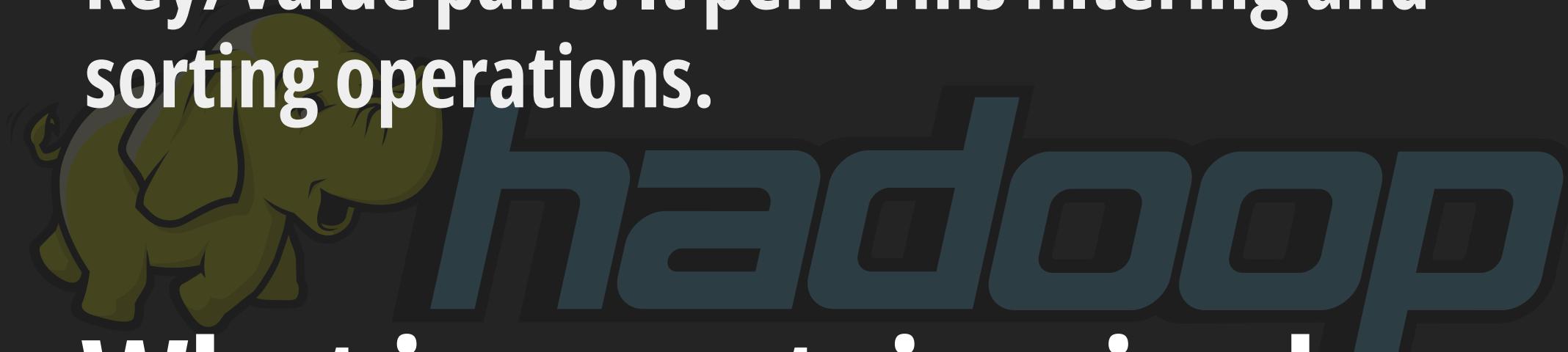


Shwetank Singh
GritSetGrow - GSGLearn.com



What is the purpose of the 'Map' function?

The Map function processes input key/value pairs to generate a set of intermediate key/value pairs. It performs filtering and sorting operations.



What is a containerized application in YARN?

In YARN, a containerized application is an application that runs within a Container, which is a logical bundle of resources (CPU, memory, etc.) allocated by the NodeManager.

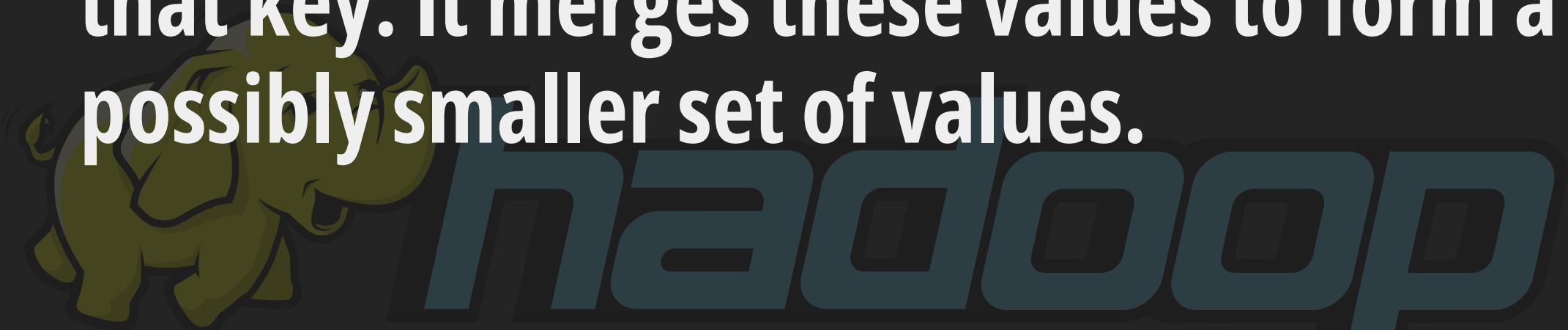


Shwetank Singh
GritSetGrow - GSGLearn.com



What is the purpose of the 'Reduce' function?

The Reduce function accepts an intermediate key and a set of values for that key. It merges these values to form a possibly smaller set of values.



How does Hadoop ensure data integrity?

Hadoop ensures data integrity through checksums. When a client writes data to HDFS, checksums are computed and stored. These are verified when data is retrieved.



Shwetank Singh
GritSetGrow - GSGLearn.com



How does Hadoop achieve fault tolerance?

Hadoop achieves fault tolerance through data replication in HDFS, speculative execution of tasks, and the ability to restart failed tasks.



What is the significance of the 'dfs.replication' property?

`dfs.replication` sets the default number of replicas of each block in HDFS. The default value is 3, which provides a good balance between fault tolerance and storage efficiency.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is rack awareness in Hadoop?

Rack awareness is Hadoop's ability to identify the rack in which a node is located. This helps in optimizing data placement and network usage by keeping data on the same rack when possible.



Explain the concept of 'Junk Record' in Hadoop

A junk record is corrupt or improperly formatted data that can't be processed correctly. Hadoop provides mechanisms to handle these, such as skipping bad records.



Shwetank Singh
GritSetGrow - GSGLearn.com



Explain the concept of data locality in Hadoop

Data locality refers to moving computation close to where the data resides, rather than moving large data to computation. This reduces network congestion and increases the overall throughput of the system.

What is a TextInputFormat in Hadoop?

TextInputFormat is the default **InputFormat** in Hadoop. It reads lines of text files. The key is the byte offset of the line, and the value is the content of the line.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is a JobTracker in Hadoop?

In Hadoop 1.x, JobTracker was responsible for resource management and job scheduling/monitoring. In Hadoop 2.x and later, these responsibilities are split between the ResourceManager and ApplicationMaster.

Describe the role of Configuration class in Hadoop

The Configuration class in Hadoop represents a set of configuration parameters. It reads configuration from resources like core-site.xml, hdfs-site.xml, etc.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is a TaskTracker in Hadoop?

In Hadoop 1.x, TaskTracker was a daemon that accepted tasks (Map, Reduce, and Shuffle) from the JobTracker. In Hadoop 2.x and later, its functionality is covered by the NodeManager.

What is the purpose of the 'hadoop-env.sh' file?

`hadoop-env.sh` is a shell script that sets up the environment for running Hadoop. It allows you to customize Hadoop daemon settings like heap size, log directories, etc.



Shwetank Singh
GritSetGrow - GSGLearn.com



How does Hadoop handle node failures?

Hadoop handles node failures by re-executing failed tasks on other healthy nodes, replicating data across multiple nodes, and using heartbeat messages to detect node failures.

Explain the concept of 'rack awareness' in HDFS

Rack awareness is HDFS's ability to identify the physical rack location of each DataNode. This information is used to optimize data placement and network usage.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is the purpose of the Secondary NameNode?

The Secondary NameNode periodically merges the namespace image with the edit log to prevent the edit log from becoming too large. It's not a real-time backup of the Primary NameNode.

What is the role of the ApplicationMaster in YARN?

The ApplicationMaster is responsible for negotiating resources from the ResourceManager and working with the NodeManager(s) to execute and monitor the tasks.

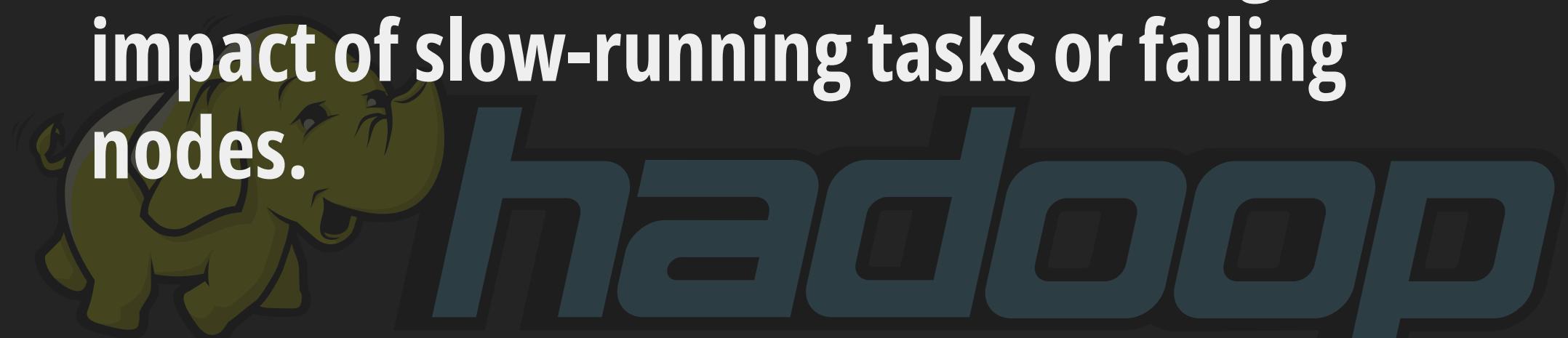


Shwetank Singh
GritSetGrow - GSGLearn.com



Explain the concept of speculative execution in Hadoop

Speculative execution is a process where multiple instances of the same task are executed on different nodes to mitigate the impact of slow-running tasks or failing nodes.



How does Hadoop handle data skew?

Data skew occurs when data is unevenly distributed across partitions. Hadoop can handle this through techniques like custom partitioning, sampling, or using combiners.

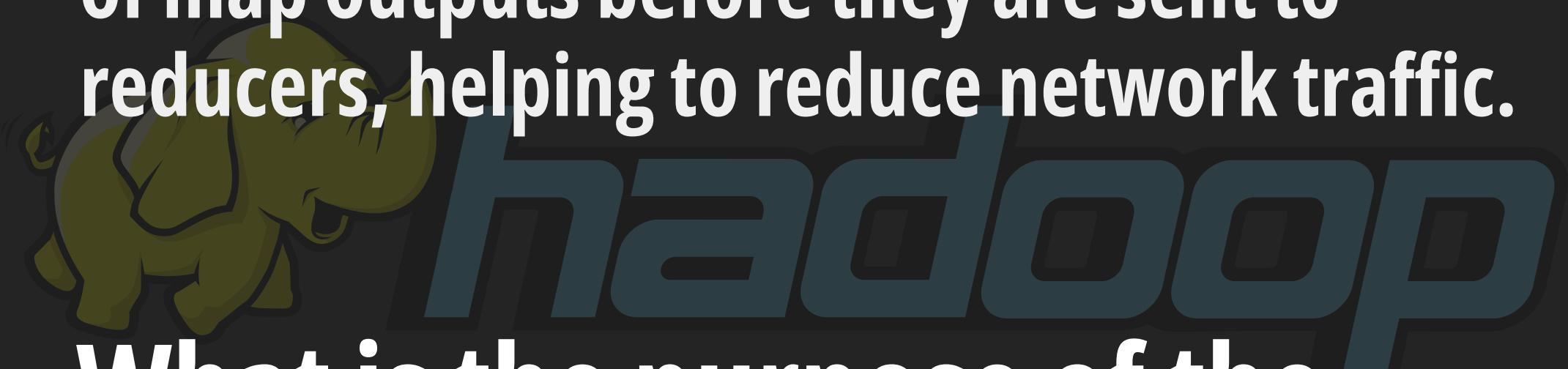


Shwetank Singh
GritSetGrow - GSGLearn.com



What is a Combiner in Hadoop MapReduce?

A Combiner is an optional component in MapReduce that performs local aggregation of map outputs before they are sent to reducers, helping to reduce network traffic.



What is the purpose of the DistCp (Distributed Copy) tool?

DistCp is a tool for large inter/intra-cluster copying. It uses MapReduce to effect its distribution, error handling, recovery, and reporting.



Shwetank Singh
GritSetGrow - GSGLearn.com



What is the significance of 'shuffle and sort' phase in MapReduce?

The shuffle and sort phase transfers the mapped output to the reducers and sorts them by key. This ensures that all values for a given key are grouped together before reduction.



Explain the concept of 'split' in Hadoop MapReduce

A split in MapReduce is a chunk of the input that is processed by a single map task. The number of splits determines the number of mappers.



Shwetank Singh
GritSetGrow - GSGLearn.com

How does Hadoop handle small files problem?

Hadoop is not efficient with small files as they create many map tasks and overwhelm the NameNode. Solutions include using HAR (Hadoop Archive) files, sequence files, or CombineFileInputFormat.

What is the significance of the 'fair scheduler' in Hadoop?

The fair scheduler is a method of assigning resources to jobs such that all jobs get, on average, an equal share of resources over time. It helps prevent small jobs from getting starved in the presence of larger jobs.

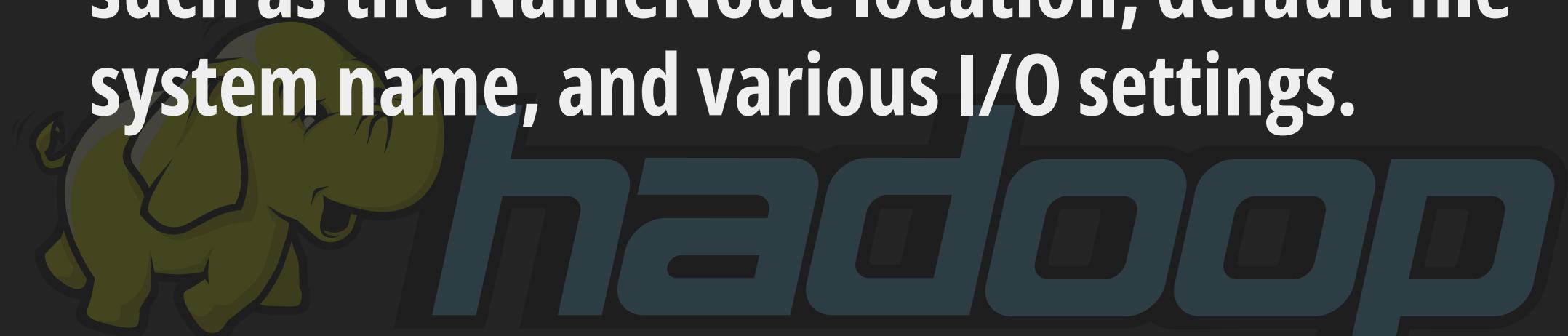


Shwetank Singh
GritSetGrow - GSGLearn.com



What is the role of the 'core-site.xml' file in Hadoop?

core-site.xml is a configuration file that contains core properties for Hadoop clusters, such as the **NameNode** location, default file system name, and various I/O settings.



How does Hadoop handle hot spotting?

Hot spotting occurs when a particular node receives a disproportionate amount of work.

Hadoop mitigates this through techniques like data replication, speculative execution, and proper partitioning.



Shwetank Singh
GritSetGrow - GSGLearn.com



THANK
you,