Cauvery College for Women (Autonomous)

Nationally Accredited (III Cycle) with 'A' Grade by NAAC

Annamalai Nagar, Tiruchirappalli-18.

Name of the Faculty	: Dr Sinthu Janita
Designation	: Professor & Head
Department	: Computer Science
Contact Number	: 9894484436
Programme	: MSc Computer Science
Batch	: 2016-2017 Onwards
Semester	: IV
Course	: Big Data Analytics
Course Code	:P16CSE5A
Unit	: V
Topics Covered	: Introduction to MapReduce, Processing data with Hadoop

using MapReduce, MapReduce Applications

Hadoop MapReduce and YARN Framework

- Introduction to MapReduce
- Processing data with Hadoop using MapReduce5
- Introduction to YARN
- Components
- Need and Challenges of YARN
- Dissecting YARN
- MapReduce applications
- Data serialization and Working with common serialization formats
- Big data serialization formats

Introduction to MapReduce

MapReduce Daemons

Job Tracker

- It provides connectivity between HADOOP and application.
- When a code is submitted to cluster, JobTracker creates the execution plan by deciding which task to assign to which node.
- It also monitors all the running tasks. When a task fails, it automatically reschedules the task to a different node after a predefined number of retries.
- JobTracker is a master daemon responsible for executing over all map reducing jobs
- There is a single JobTracker per HADOOP cluster.

Introduction to MapReduce ctd...

MapReduce Daemons

- **TaskTracker** :
- This daemon is responsible for executing individual tasks that is assigned by the JobTracker
- There is a single TaskTracker per slave and spawns multiple Java Virtual Machines (JVM) to handle multiple Map or Reduce tasks in parallel
- TaskTracker continuously sends heartbeat message to JobTracker. When the JobTracker fails to receive a heartbeat from a TaskTracker, the JobTracker assumes that the TaskTracker has failed and resubmits the task to another available node in the cluster.
- Once the client submitted the job to the JobTracker, it partitions and assigns diverse MapReduce tasks for each TaskTracker in the cluster.

How does MapReducer work ?

MapReducer divides the data analysis task into two parts namely Map and Reduce

Each Mapper works on the partial data set that is stored on the node . And the Reducer combines the output from the Mappers to produce the result set .

First the input data set is split into multiple pieces of data.

The framework creates a master and several workers processes and execute the worker processes remotely

Several map tasks works simultaneously and read pieces of data that were assigned to each map tasks .

How does MapReducer work ?

Map worker uses partitioner function to divide the data into regions.

Partitioner decides which reducer should get the output of the specified mapper .

When the map workers complete their work the master instructs the reduced workers to begin their work .

The reduce workers in turn contact the map workers to get the key value data for their partition .

The data thus received is shuffled and sorted as per keys.

Then it calls reduced function for every unique key . This function writes the output to the file .

When all the reduced workers complete their work the master transfers the control to the user programs.

Processing data with Hadoop using MapReduce

In map reducing programming , jobs are split into a set of map tasks and reduce tasks . Then these tasks are executed in a distributed fashion on HADOOP cluster. Each task processes small subset of data that has been assigned to it. This way HADOOP distributes the load across the cluster. Map reducing job takes a set of files that is stored in HDFS as input .

Map task takes care of loading , parsing , transforming and filtering . The responsibility of reduced tasks is grouping and aggregating data that is produced by map task to generate final output. Each map tasks is broken into the following phases :

RecordReader

Mapper

Combiner

Partitioner

Processing data with Hadoop using MapReduce ctd...

The output produced by map tasks is known as intermediate keys and values . These intermediate keys and values are sent to the reducer . The reducer tasks are broken in the following phases :

Shuffle

Sort

Reducer

Output format

HADOOP assigns map tasks through a DateNode where actual data to be processed resides . This way HADOOP ensures data locality. Data locality means the data is not moved over network . Only computational code is moved to process data which saves network bandwidth.

Processing data with Hadoop using MapReduce ctd...

MAPPER

The mapper maps the input key value pairs into a set of intermediate key value pairs . Maps are individual tasks that have the responsibility of transforming input records into intermediate key value pair.

RecordReader: RecordReader converts a byte oriented view of the input into a record oriented view and presents it to the mapper tasks .It presents the tasks with keys and values . Generally the key is the positional information and value is the chunk of data that constitutes the record .

Map: Map function works on the key value pair produce by RecordReader and generates zero or more intermediate key value pair . The Map Reduce decides the key value pair based on the content .

Combiner : It is a optional function but provides high performance in terms of network bandwidth and disk space . It takes intermediate key value pair provided by the mapper and applies user specific aggregate function to only that mapper. It is also known as local reducer.

Partitioner: Partitioner takes the intermediate key value pair roduced by the mapper , splits them into sharred and send the sharred to the particular reducer as per the user specific code. The key with the same value goes to the same reducer.The partitioned data of each map tasks is written to the local disks of that machine and pulled by the respective reducer.

Processing data with Hadoop using MapReduce ctd...

REDUCER

The primary chore of the reducer is to reduce a set of intermediate values to a smaller set of values . The reducer has three primary phases :Shuffle and Sort , Reduce , Output Format.

Shuffle and Sort : This phase takes the output of all the partitioners and downloads them into the local machine where the reducer is running . Then these individual data pipes are sorted by keys which produce large data lists . The main purpose of the sort is grouping similar words so that their values can be easily iterated over by the reduce tasks.

Reduce: The reducer takes the grouped data produced by the shuffle and sort phase , applies reduce function and process one group at a time. The reduce function iterates all the values assosciated with the key . Reducer function provides various operations such as aggregation , filtering and combining data .Once it is done , the output of the reducer is sent to the output format.

Output Format: The output format seperates key value pairs with tab and writes it out to a file using record writer .

Processing data with Hadoop using MapReduce ctd...

COMBINER:

It is an optimisation technique for map reducing job .Generally the reducer class is a set to be the combiner class . The difference between the combiner class and reducer class is as follows :

The output generated by the combiner is intermediate data and it is passed to the reducer.

Output of the reducer is passed to the output file on the desk.

PARTITIONER:

The partitioning phase happens after map phase and before reduce phase. Usually the number of partitions are equal to the number of reducers . The default partitioner is harsh partitioner.

COMPRESSION:

In MapReduce Programming, you can compress the map reduce output file. Compression provides two benefits as follows.

Reduces the space to store files.

Speeds up data transfer across the network.

MapReduce Application

Analysis of logs, data analysis, recommendation mechanisms, fraud detection, user behavior analysis, genetic algorithms, scheduling problems, resource planning among others, is applications that use MapReduce.

- **1 Social Networks**

  Social networking users like Facebook, Twitter, and LinkedIn to connect with friends and community.Many of the features, such as who visited your LinkedIn profile, who read the post on Facebook or Twitter, can be evaluated using the MapReduce, programming model.

- **2 — Entertainment**

  Netflix uses Hadoop and MapReduce to solve problems such as discovering the most popular movies, based on what you watched, what do you like? Providing suggestions to registered users taken into account their interests.MapReduce can determine how users are watching movies, analyzing their logs and clicks.

MapReduce Application ctd...

3 — Electronic Commerce

Many e-commerce providers, such as the Amazon, Walmart, and eBay, use the MapReduce programming model to identify favorite products based on users' interests or buying behavior.

It includes creating product ecommendation mechanisms for e-commerce catalogs, analyzing site records, purchase history, user interaction logs, and so on.It's used to establish a user's sentimental profile for a particular product by reviewing comments or reviews or analyzing search logs by identifying which items are most popular based on the search and which products are missing.Many Internet service providers use MapReduce to analyze site records and understand site visits, engagement, locations, mobile devices, and browsers.

MapReduce Application ctd...

4 — Fraud Detection

- Hadoop and MapReduce are used in the financial industries, including companies such as banks, insurance providers, payment locations for fraud detection, trend identification or business metrics through transaction analysis.Banks analyze the data of the credit card and the related expenses, for categorization of these expenses and make recommendations for different offers, analyzing anonymous purchasing behavior.

5 — Search and Advertisement Mechanisms

- Can utilize it to analyze and understand search behavior, trends, and missing results for specific keywords.
- Google and Yahoo use MapReduce to understand users' behavior, such as popular searches over a period of an event such as presidential elections.Google AdWords uses MapReduce to understand the impressions of ads served, click-through rates, and engagement behavior of users.

6— Data Warehouse

- MapReduce can be utilised to analyze large data volumes in data warehouses while implementing specific business logic for data insights.