MapReduce

Date: 5/8/2019 Page:

- MR is a Programming model for data Processing
- MR Programs are inherently parallel.
- Enables massive Scalability across 100/1000 of m/c in Hadoop Cluster.
- MR - heart of Apache Hadoop.
- MR refers to 2 Seperate and distinct task that Hadoop beform.

1) MapJob - take set of data and convert into another set of data.
   - Individual elements are broken down into tuples (Key/value) pair.
2) Reduce Job - take o/p from Map as input and Combines those data tuples into Smaller Set of tuples

Benefits

1) Scalability - Can process PB of data Stored into HDFS
2) Flexibility - Easier access to m/t Source of data m/t type of data
3) Speed - with parallel Processing, minimal data movement Hadoop offer fast processing of massive amount of data
4) Simple - Developer Can write Code in choice of language - Java/C++/Python.

Date: Page:

Block – min amount of data that can be read or write

Compare with RDBMS

- Seeking is the process of moving ~~object~~ disk's head to a particular place on the disk to read and write data
  - Characterise the latency of disk op.
  - Transfer rate corresponding to disk's bandwidth.
  - Data access dominated by seek – take longer to read or write
  - Updating small proportion of record in DB – BTree works well.
  - Updating majority of DB – BTree is less efficient than MR.
  - MR uses Sort/Merge to rebuild DB
  -

→ MR good fit for Problems
  - Analyze whole data set in batch fashion
  - Particularly adhoc Analysis

→ RDBMS good fit
  - good point queries or updates where DataSet indexed to deliver low latency retrieval

MR

→ * MR scales linearly with
  ↳ Size of data
  * Data is partitioned
  * function primitives can work in parallel on seperate partitions.
  * Double the size of input data – job will run

Date: Page:

twice as slowly.
- But if double size of Cluster a job will run as fast as the original one.
- Not true with SQL queries

Features	RDBMS	MR
Data Size	GB	PB
Access	Interactive and Batch	Batch
Updates	Read & Write many time	Write once Read times
Transactions	ACID	None
Structure	Schema on write	Schema on Read
Integrity	High	Low
Scaling	Non linear	Linear

Schema on Read

Interpret data at processing time.

MapReduce

* Map Reduce is a Programing model for data Processing
* Map Reduce programs are inherently parallel.
* Distributing task across multiple nodes.
* Each node process data stored in that nodes
* Consist of 2 phases - map - Reduce

Date:_____ Page:_____

Features of map Reduce

- Automatic parallelization and distribution
- Fault tolerance
- Clean abstraction for programmer
- MR Programs usually written in Java
- Can be written in any language using Hadoop Streaming
- MR developed in Java
- Abstracts all housekeeping away from Developer.
  - ↳ Concentrate on writing MR functions.
- Data locality
- In built redundancy.
- abstract from Complexity of DIST Programing lay.
- manages all inter process Communication
- Shared nothing than architectural model

Cluster - Group of m/c working together to store and process data.

Worker Node

- HDFS to store data - data node
- map Reduce to process data - task tracker

Master Node

- Name node - manages HDFS
- Job tracker - manages MapReduce