SE ZG522: Big Data Systems

02 Analytics on Large Data

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Feb 06, 2021 ONLINE (WILP @ BITS-Pilani Jan-May 2022)

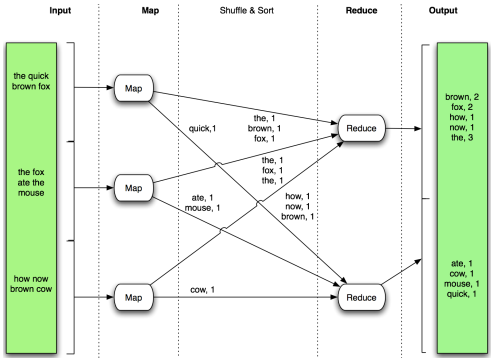
http://ktiwari.in/bds

Big Data

- **Grid** (heterogeneous) or **cluster** (homogeneous) computing
- Distributed computing **have to deal with** synchronization, deadlocks, data dependency, mutual exclusion, replication, reliability, platform scalability and provisioning
- Ready-made solution is **MapReduce**/**Hadoop**[1] or **spark** that provides a high level of abstraction for data parallel tasks
- Hadoop/PIG combo is very effective
- Need is there for a scalable distributed computing framework that provides both abstraction and performance (by exploiting all kinds of parallelisms that exist in an algorithm)

[1] MapReduce: Simplified Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemawat, OSDI, 2004

Big Data



map is a function that executes on each partition.

PK-Means

- **K-Means:** steps involved are
  1. Select Seed
  2. Assignment //(most compute intensive)
  3. Compute Centroid
- **PK-Means:** [2]
  - Input dataset is stored on GFS/HDFS. A sequence file of $< key, value >$ pairs. (Key: offset, Value: string of whole record)
  - Dataset is split and globally, and broadcast to all mappers

  Map
  - Distance calculations are parallel executed
  - Each mapper has array of centers
  - Computes closest center for each sample
  - Intermediate values/output: $< key', value' >$ (Key': index of the closed centre, value': sample)

[2] Du, Zhihua and Wang, Yiwei and Ji, Zhen, "PK-means: A new algorithm for gene clustering" in Computational Biology and Chemistry, pages=243–247, vol 32(4), Elsevier 2008

PK-Means

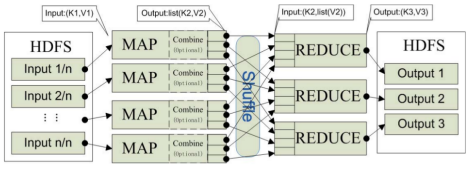
- **PK-Means:**
  Combiner
  - Combines intermediate data of each map task and stores locally
  - Partial sum the values assigned to the same cluster
    - Record number of samples in each cluster and
    - Sum of values at each dimension
  - Key': key' & Value': string of num and sums

  Reducer
  - Input: output of combiner
  - Compute all the samples assigned to a center
  - Calculate new centers

Scalability is high

MR-DBSCAN

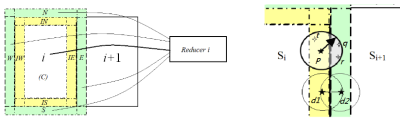
- MR-DBSCAN [3] involves following steps
  1. Preprocessing
  2. Local DBSCAN
  3. Find Merging Mapping
- Uses quadtree, a spacial data structure
- Extended regions (ϵ-extended) is taken in partition



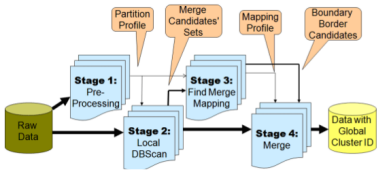
[3] MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm using MapReduce He, Yaobin et.al. Parallel and Distributed Systems (ICPADS) 473–480, IEEE 2011

MR-DBSCAN

- Extended regions (ϵ-extended) is taken in partition



- Cross connection files are processed during reduce
- This makes the algorithm data parallel



Thank You!

Thank you very much for your attention!

Queries ?