Big Data Lab Activity - Week 04

Basic Tasks

1. File Replication

Advantages:

- Faster access
- High availability
- Parallel access

Drawbacks:

- Sync complexity
- More storage
- Network load

2. Remote Access and Data Locality

Drawback:

- High latency

Better solution:

- Use data locality

Data Locality:

- Move computation to data

3. Sequential/Parallel Processing

Terms:

- Sequential processing: Step by step
- Parallel processing: Multiple tasks

Matrix multiplication:

- Divide into submatrices
- Distribute to cores

4. Formula Calculation

Parallel parts:

- $(a + b) \times (c/d)$
- $(e \times f) + (g/h)$

Limits:

- Sync cost

Cross-node limits:

- Network delay

5. Finding Max Number

Method:

- Split file
- Calculate max per node
- Merge results

6. Different Parallel Methods

- Task 4: Independent parts
- Task 5: Split data

Medium Tasks

7. Hadoop HDFS

NameNode and DataNode:

- Manage metadata
- Store data

File Splits:

- 3 splits: 64MB, 64MB, 52MB
- 3 replicas

Example:

- Split S1: Node 3, Node 5, Node 7

Node Failure:

- Node 5 crash
- NameNode re-replicates

Advanced Tasks

8. Hadoop MapReduce

JobTracker and TaskTracker:

- JobTracker: Manage tasks
- TaskTracker: Execute tasks

Map Function:

- Create key-value pairs

Reduce Function:

- Merge values

9. MapReduce for Orders

Keys:

- Input: order_id
- Output: equip_name

Map:

- Create (equip_name, qty)

Reduce:

- Sum qty for each equip_name

10. MapReduce for Streaming Service

Map:

- (film_id, user_id)

Reduce:

- Count per film_id