# Graph Partitioning Techniques for Large Graphs

## Graph Partitioning

Arik Pamnani
Shivdutt Sharma
Ravi Shrimal

Supervised by: Prof. Anirban Dasgupta and Rachit Chhaya

IIT Gandhinagar

# Motivation and Problem

- Graph datasets are huge!
- Crawls by a search engine currently amount to 1 trillion links
- Therefore, performing computations and algorithms is difficult

# Motivation and Problem

- An approach could be to distribute the graph on a cluster of nodes
- Distributing the graph can be expensive (in terms of inter-partition communication)

## Goal

- Minimize the number of inter-partition edges
- Number of nodes (of the graph) must be almost the same on all partitions

# Overview

We have -

- Implemented some graph partitioning techniques (Hashing, Chunking, Balanced, Weighted Greedy, etc.)
- Then, run algorithms like PageRank on the graph and checked for the inter-partition communication and the computing time

# Tools/Datsets

- **Tools:** Apache Spark (Map - Reduce operation performed on the dataset)
- **Dataset:** Amazon Web Data (There is an edge between $i$ and $j$, if product $i$ is bought frequently with product $j$)
    - Nodes: 262111
    - Edges: 1234877

# Apache Spark

- Spark runs MapReduce jobs in stages
- Stages are built up by DAG Scheduler
- RDD (Resilient Distributed Datasets) is the fundamental data structure of Spark
- RDDs are immutable and all MapReduce operations are performed on an RDD
- Each RDD is divided into partitions, and can be computations can be done on different nodes of the cluster

# Symbols

- Each individual partition at time $t$ is referred to by its index $P^t(i)$.
- $\Gamma(v)$ refers to the set neighbours of $v$.

# Partitioning Algorithms - Balanced

- We assign $v$ (the current vertex in the stream) to a partition of minimal size (ties are broken randomly)

$$ind = arg \min_{i \in [k]} |P^t(i)|$$

- *ind* is the index of the partition to which the vertex $v$ is assigned

# Partitioning Algorithms - Chunking

- We divide the stream into chunks of size, $C$, and fill the partitions in order

$$ind = \lceil t/C \rceil$$

- $t$ is the time at which $v$ is encountered in the stream, $ind$ is the index of the partition to which $v$ is assigned

# Partitioning Algorithms - Hashing

- We take a hash function, $H : V \to \{1...k\}$ and assign $v$ to

$$ind = H(v)$$
$$H(v) = (v \ mod \ k) + 1$$

# Partitioning Algorithms - Deterministic Greedy

- We assign $v$ to a partition where it has the most edges in common
- Also, weight this by a penalty function which imposes a penalty on larger partitions (ensures that number of nodes in each partitions are almost even)
- Break ties randomly

# Partitioning Algorithms - Deterministic Greedy

- Symbolically,

$$ind = arg \max_{i \in [k]} (|P^t(i) \cap \Gamma(v)| \times w(t, i))$$

- $w(t, i)$ is the penalty function of $P^t(i)$
- $w(t, i)$ can be any one of the following -
    - $w(t, i) = 1$ (unweighted greedy)
    - $w(t, i) = 1 - \frac{|P^t(i)|}{C}$ (linear weighted)
    - $w(t, i) = 1 - exp\{|P^t(i)| - C\}$ (exponentially weighted)

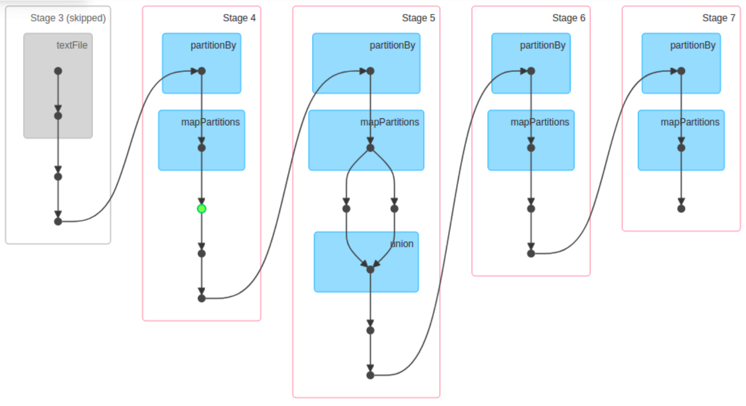
# Ordering

- The paper (which we have followed) discussed the use of different orderings of the stream
- Based on BFS, DFS, Random, etc
- We have used the same (Random) ordering as given by the Database

# Results

- Single machine implementation of the PageRank on the graph dataset
- Deployed it on a 3-node cluster
- Spark has UI Metrics (such as, Shuffle Read/Write) which gives us an idea about the inter-partition communication taking place

# DAG - Balanced



# Statistics - Balanced

### Summary

| | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Active(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 32 | 32 | 32 s (0.3 s) | 7.9 MB | 22.7 MB | 37.1 MB |
| **Dead(0)** | 0 | 0.0 B / 0.0 B | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B |
| **Total(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 32 | 32 | 32 s (0.3 s) | 7.9 MB | 22.7 MB | 37.1 MB |

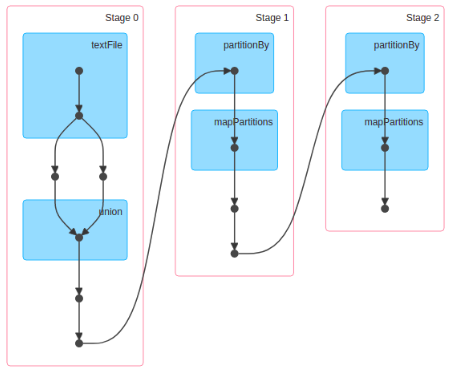
### Executors

Show 20 entries

Search:

| Executor ID | Address | Status | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write | Logs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| driver | 192.168.0.3:38623 | Active | 0 | 0.0 B / 6.7 GB | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B | |
| 0 | 192.168.0.1:34160 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B | stdout stderr |
| 1 | 192.168.0.3:43500 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 15 | 15 | 13 s (74 ms) | 3.9 MB | 11.3 MB | 18.6 MB | stdout stderr |
| 2 | 192.168.0.2:46019 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 17 | 17 | 19 s (0.2 s) | 4 MB | 11.4 MB | 18.5 MB | stdout stderr |

# DAG - Chunking



# Statistics - Chunking

## Executors

### Summary

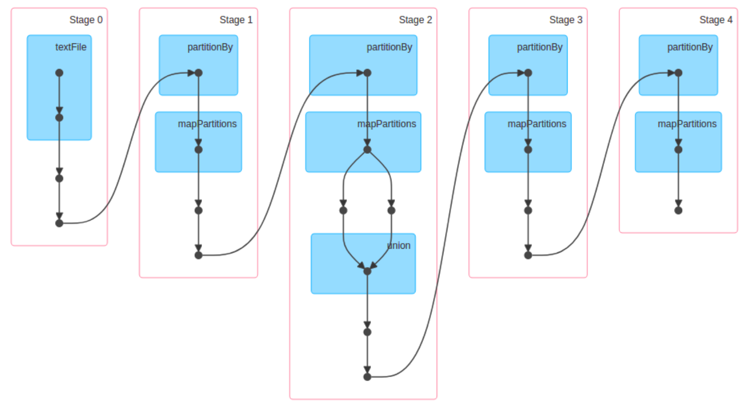
| | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Active(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 24 | 24 | 51 s (0.8 s) | 391.8 KB | 22.6 MB | 33.6 MB |
| **Dead(0)** | 0 | 0.0 B / 0.0 B | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B |
| **Total(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 24 | 24 | 51 s (0.8 s) | 391.8 KB | 22.6 MB | 33.6 MB |

### Executors

Show 20 entries Search:

| Executor ID | Address | Status | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write | Logs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| driver | 192.168.0.3:45146 | Active | 0 | 0.0 B / 6.7 GB | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B | |
| 0 | 192.168.0.1:41132 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 7 | 7 | 21 s (0.4 s) | 130.6 KB | 5.8 MB | 10.5 MB | stdout stderr |
| 1 | 192.168.0.3:42514 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 9 | 9 | 18 s (0.2 s) | 130.7 KB | 7.5 MB | 14 MB | stdout stderr |
| 2 | 192.168.0.2:40551 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 8 | 8 | 13 s (0.2 s) | 130.5 KB | 9.3 MB | 9.1 MB | stdout stderr |

# DAG - Hashing


Stage 0
textFile
Stage 1
partitionBy
mapPartitions
Stage 2
partitionBy
mapPartitions
union
Stage 3
partitionBy
mapPartitions
Stage 4
partitionBy
mapPartitions

# Statistics - Hashing

### Summary

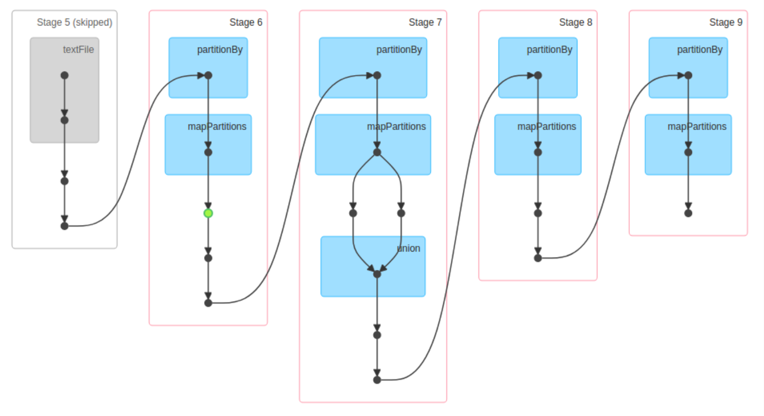
| | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Active(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 32 | 32 | **29 s (0.4 s)** | 195.9 KB | 10.6 MB | 37.9 MB |
| **Dead(0)** | 0 | 0.0 B / 0.0 B | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B |
| **Total(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 32 | 32 | **29 s (0.4 s)** | 195.9 KB | 10.6 MB | 37.9 MB |

### Executors

Show 20 entries Search:

| Executor ID | Address | Status | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write | Logs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| driver | 192.168.0.3:46019 | Active | 0 | 0.0 B / 6.7 GB | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B | |
| 0 | 192.168.0.1:40993 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 9 | 9 | **9 s (0.1 s)** | 65.3 KB | 3.6 MB | 9.5 MB | stdout stderr |
| 1 | 192.168.0.3:40837 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 13 | 13 | **11 s (77 ms)** | 65.2 KB | 3.8 MB | 17 MB | stdout stderr |
| 2 | 192.168.0.2:42474 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 10 | 10 | **9 s (0.1 s)** | 65.4 KB | 3.1 MB | 11.4 MB | stdout stderr |

# DAG - Deterministic Greedy



# Statistics - Deterministic Greedy

**Summary**

| | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Active(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 34 | 34 | 9 s (0.3 s) | 476.6 KB | 252.7 KB | 765.4 KB |
| **Dead(0)** | 0 | 0.0 B / 0.0 B | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B |
| **Total(4)** | 0 | 0.0 B / 23.3 GB | 0.0 B | 18 | 0 | 0 | 34 | 34 | 9 s (0.3 s) | 476.6 KB | 252.7 KB | 765.4 KB |

**Executors**

Show 20 entries Search:

| Executor ID | Address | Status | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write | Logs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| driver | 192.168.0.3:40552 | Active | 0 | 0.0 B / 6.7 GB | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B | |
| 0 | 192.168.0.1:42906 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 8 | 8 | 2 s (80 ms) | 173.2 KB | 40.3 KB | 169.6 KB | stdout stderr |
| 1 | 192.168.0.3:42276 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 24 | 24 | 4 s (0.1 s) | 238.3 KB | 212.4 KB | 595.9 KB | stdout stderr |
| 2 | 192.168.0.2:38216 | Active | 0 | 0.0 B / 5.5 GB | 0.0 B | 6 | 0 | 0 | 2 | 2 | 2 s (0.1 s) | 65.1 KB | 0.0 B | 0.0 B | stdout stderr |

# References


- Isabelle Stanton, Gabriel Kliot: Streaming Graph Partitioning for Large Distributed Graphs
- `https://github.com/apache/spark/blob/master/examples/src/main/python/pagerank.py`

**Thank You!**

# Questions?