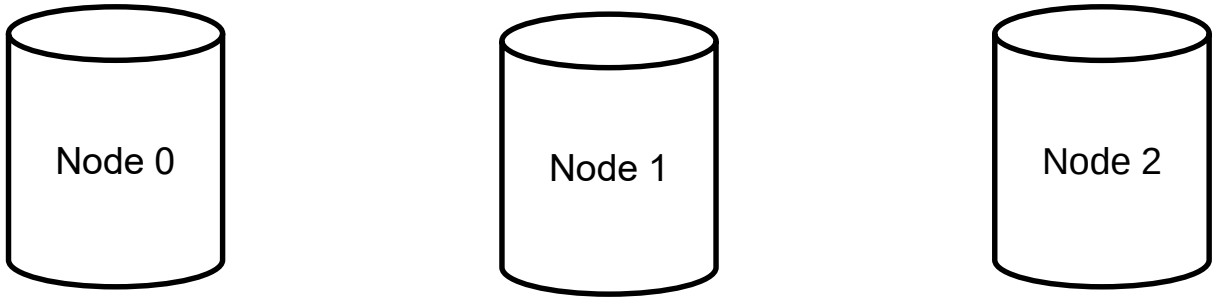
Improving Data Ingestion Performance in Apache AsterixDB

Qiyang He
Southern University of Science and Technology

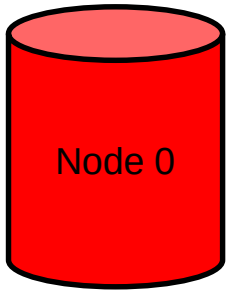
Outline

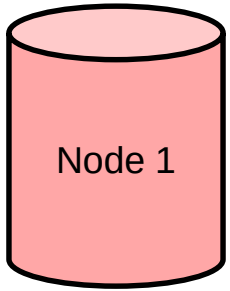
- Introduction
- Benchmark

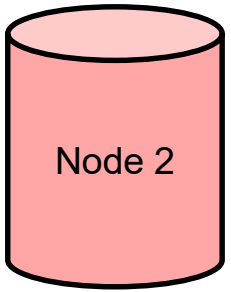
Use Case



Use Case





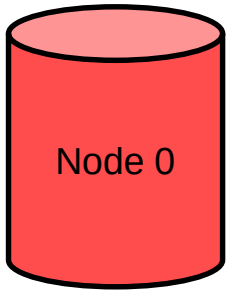


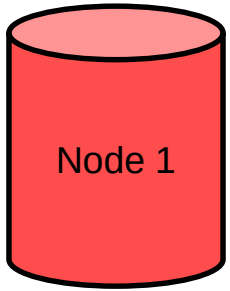
Parse: 100%
Store: 33.3%

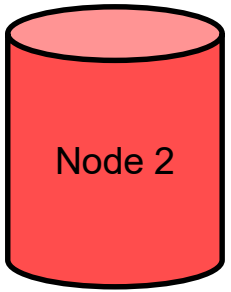
Parse: 100%
Store: 33.3%

Parse: 100%
Store: 33.3%

Use Case





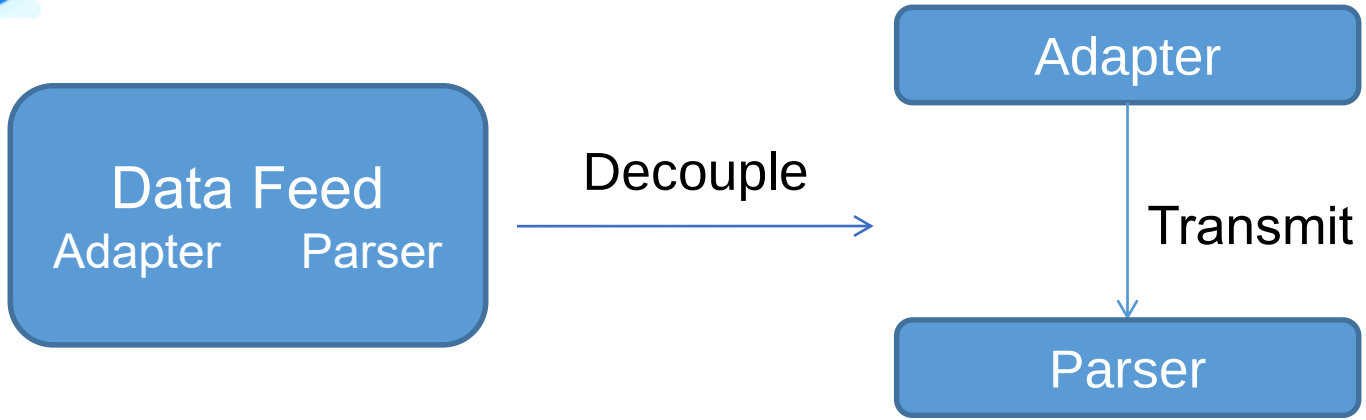


Parse: 33.3%
Store: 33.3%

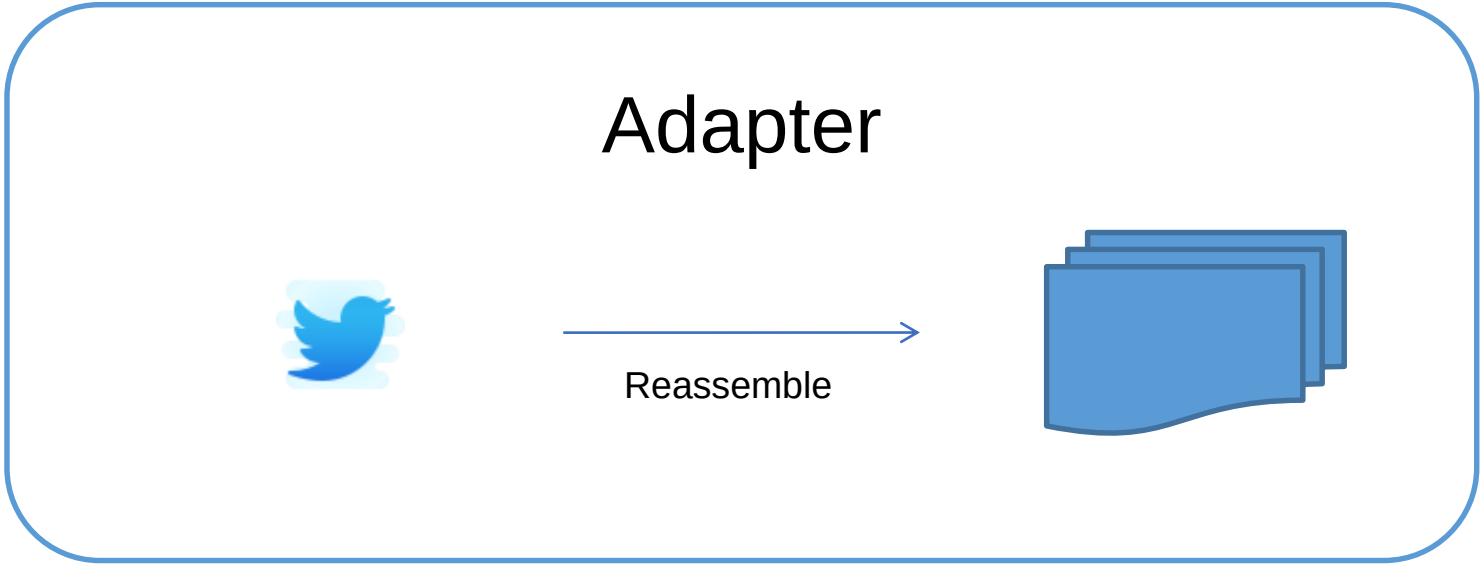
Parse: 33.3%
Store: 33.3%

Parse: 33.3%
Store: 33.3%

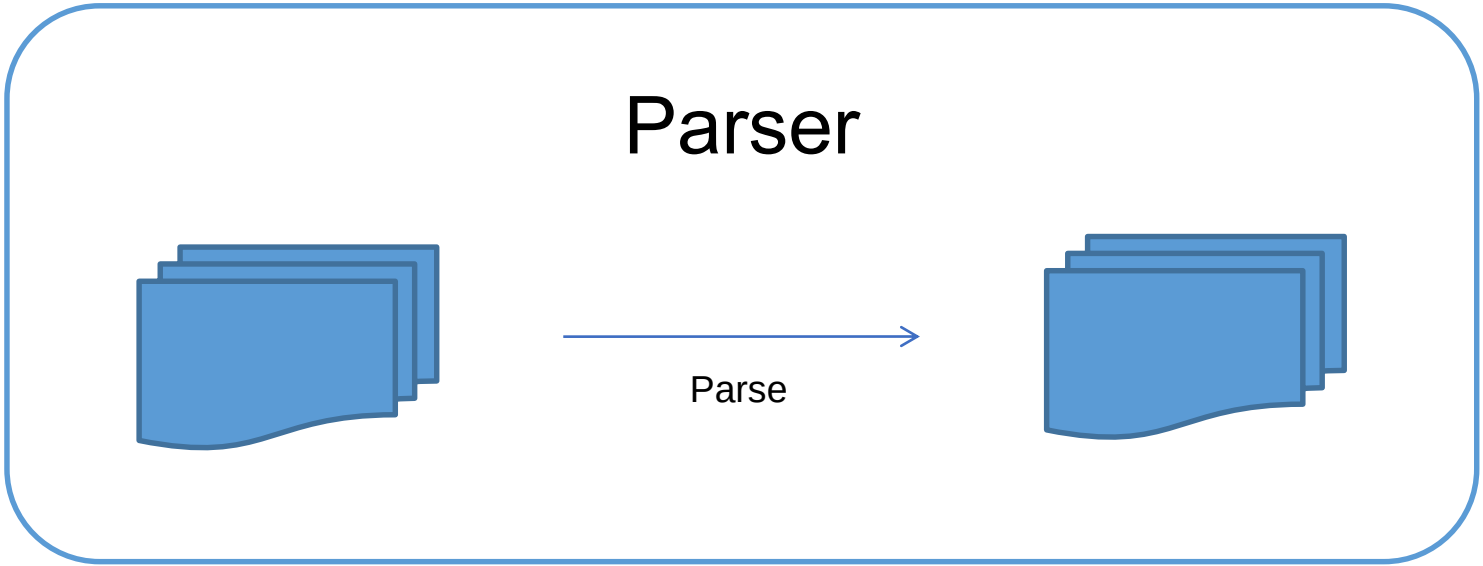
My work


Data Feed
Adapter
Parser
Decouple
Adapter
Transmit
Parser

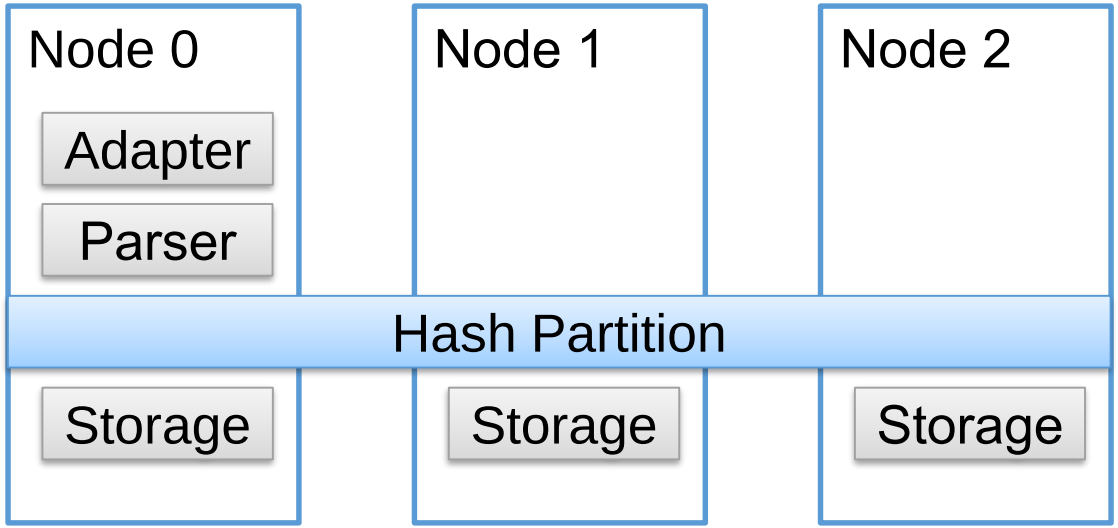
My work



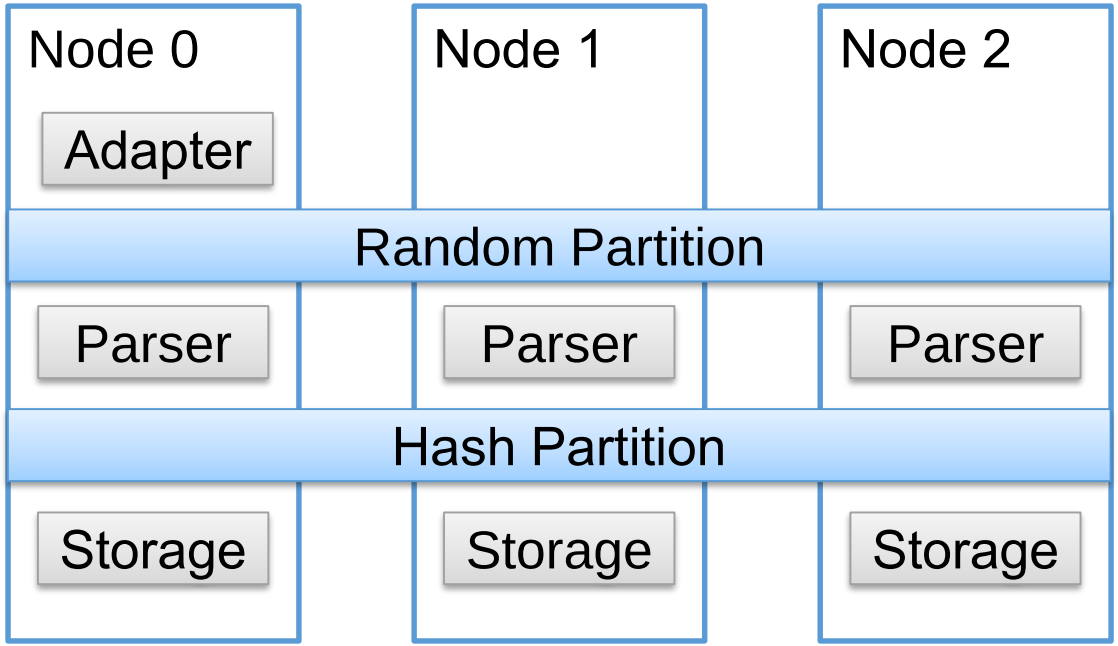
My work



Existing Framework



Proposed Framework



Restriction

```
create type TwitterUser as open{
    name: string,
    friends_count: int32,
    followers_count: int32
};

create dataset TwitterUsers(TwitterUser) primary key name;
```

Restriction

{“name”:”Alice”,”friends_count”:18,”followers_count”:49416}
{“name”:”Bob”,”friends_count”:445,”followers_count”:22649}
{“name”:”Alice”,”friends_count”:18,”followers_count”:4}
{“name”:”Bob”,”friends_count”:455,”followers_count”:22649}

Restriction

Original:
{“name”:”Alice”,”friends_count”:18,”followers_count”:49416}
{“name”:”Bob”,”friends_count”:445,”followers_count”:22649}

Parallel Ingestion:

Experiment Settings

- Processor: i7-5575 CPU @ 3.30GHz (2 cores)
- Memory: 16GB
- Dataset: 10M tweets (each about 377Bytes)

Benchmark

2 Nodes:

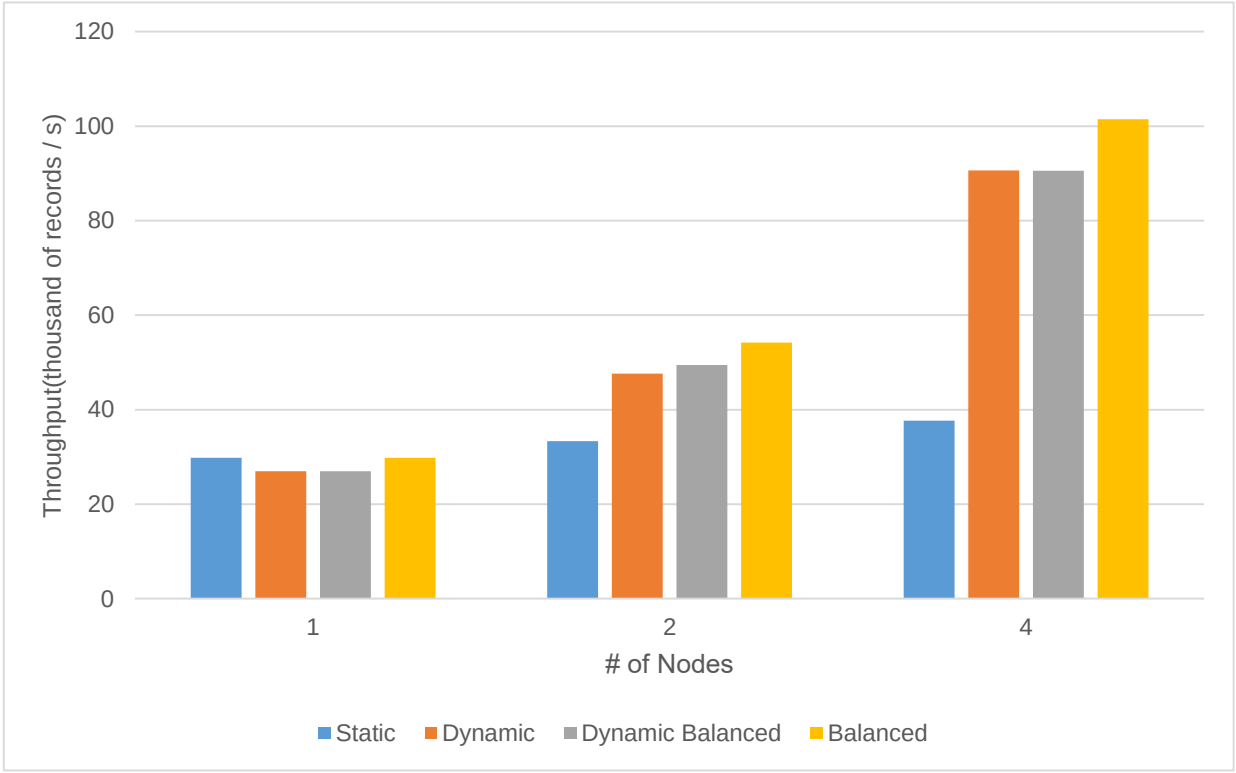
+42.8% Current Framework

-12.2% Ideal Framework

4 Nodes:

+140.9% Current Framework

-10.7% Ideal Framework



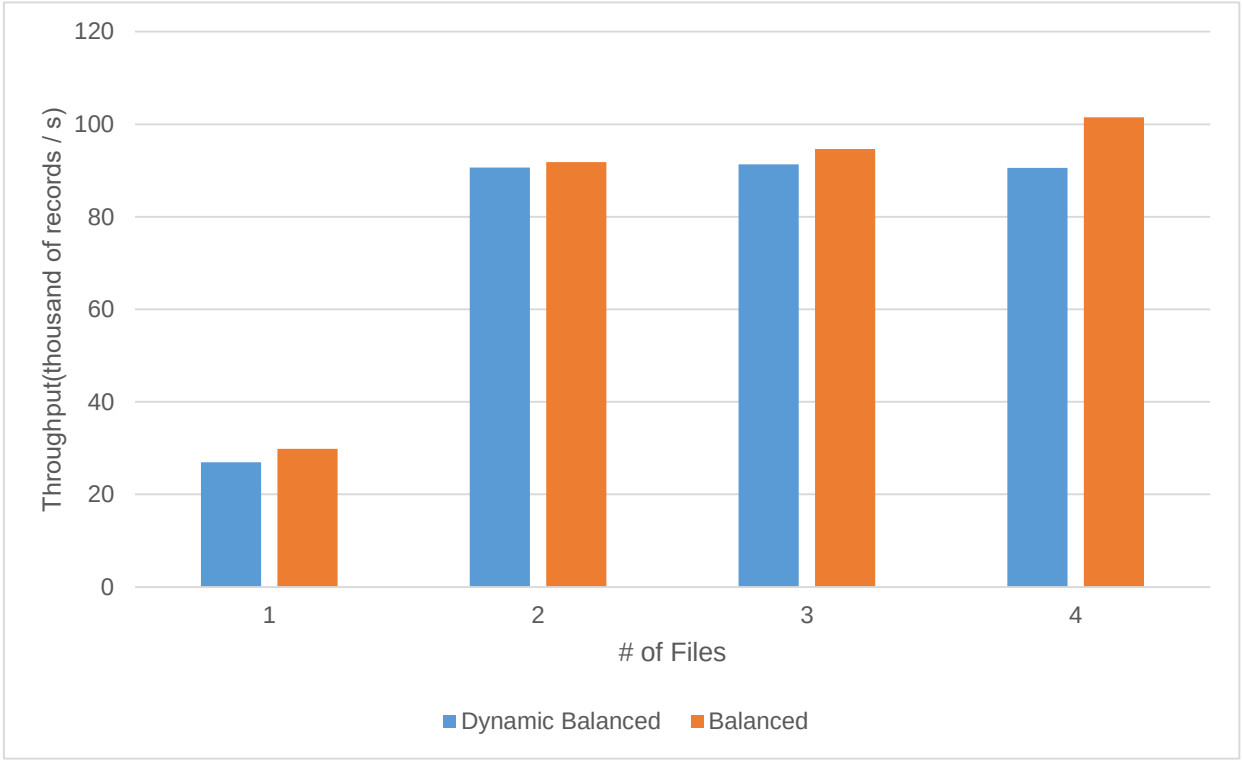
Benchmark

Dynamic:

No difference after 2 files

Balanced:

Tiny improvement after 2 files



UCI

Thank you