Databend

Databend hash join spill

王旭东

Databend

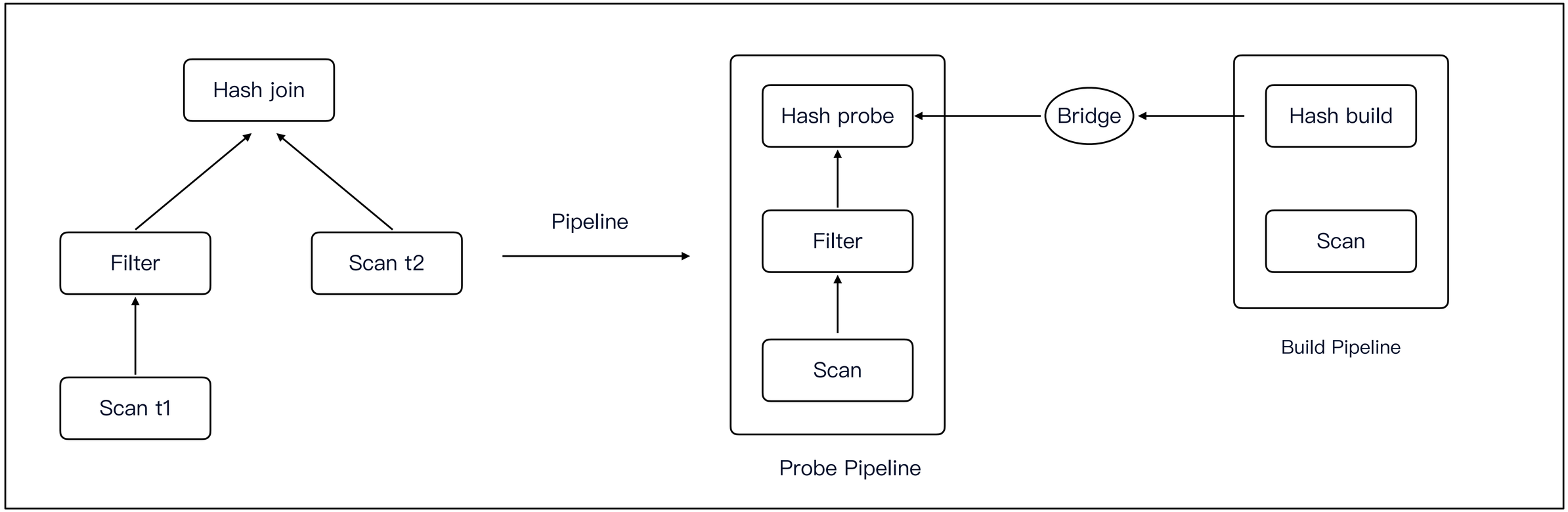
Content

Databend

Part 1

Hash join 架构

Hash join 实现


Hash join
Filter
Scan t2
Scan t1
Pipeline
Hash probe
Filter
Scan
Probe Pipeline
Bridge
Hash build
Scan
Build Pipeline

Hash join 实现

```rust
#[derive(Clone, Debug)]
enum HashJoinBuildStep {
    // The running step of the build phase.
    Running,
    // The finalize step is waiting all build threads to finish and build the hash table.
    Finalize,
    // The fast return step indicates there is no data in build side,
    // so we can directly finish the following steps for hash join and return empty result.
    FastReturn,
    // Wait to spill
    WaitSpill,
    // Start the first spill
    FirstSpill,
    // Following spill after the first spill
    FollowSpill,
    // Wait probe
    WaitProbe,
    // The whole build phase is finished.
    Finished,
}
```

Hash join 实现

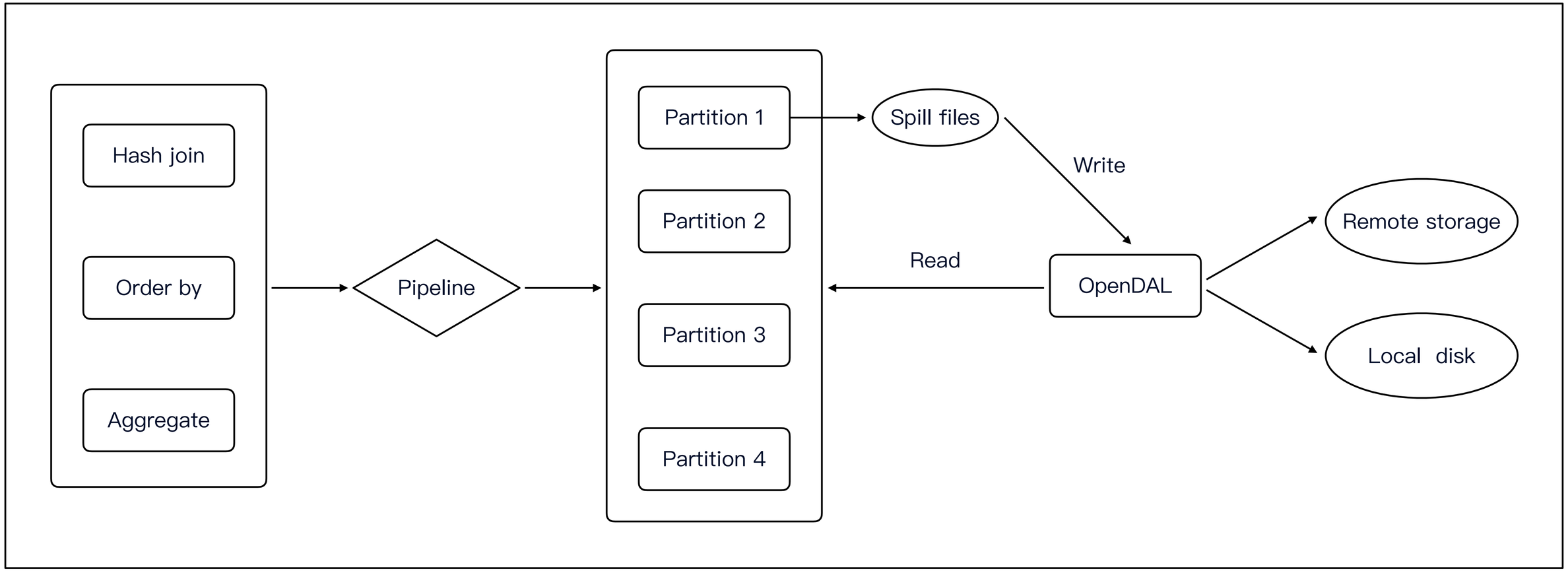
```rust
#[derive(Debug)]
enum HashJoinProbeStep {
    // The step is to wait build phase finished.
    WaitBuild,
    // The running step of the probe phase.
    Running,
    // The final scan step is used to fill missing rows for non-inner join.
    FinalScan,
    // The fast return step indicates we can directly finish the probe phase.
    FastReturn,
    // Spill step is used to spill the probe side data.
    Spill,
    // Async running will read the spilled data, then go to probe
    AsyncRunning,
}
```

Databend

Part 2

Spiller 模块设计

Spiller 模块


Hash join
Order by
Aggregate
Pipeline
Partition 1
Partition 2
Partition 3
Partition 4
Spill files
Write
Read
OpenDAL
Remote storage
Local disk

Spiller 模块

Spiller 提供的接口

```
pub async fn spill(&mut self, partitions: &[(u8, DataBlock)], worker_id: usize) -> Result<()>
```

```
pub(crate) async fn spill_input(
    &mut self,
    data_block: DataBlock,
    hashes: &[u64],
    spilled_partition_set: &HashSet<u8>,
    worker_id: usize,
) -> Result<DataBlock>
```

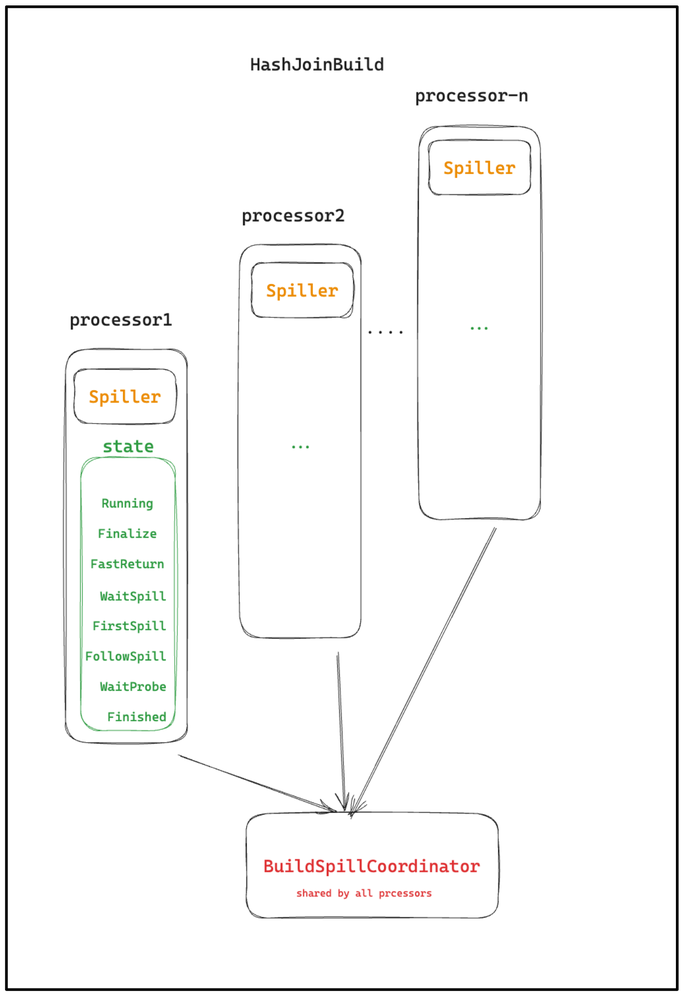
```
pub async fn read_spilled_data(&self, p_id: &u8) -> Result<Vec<DataBlock>>
```

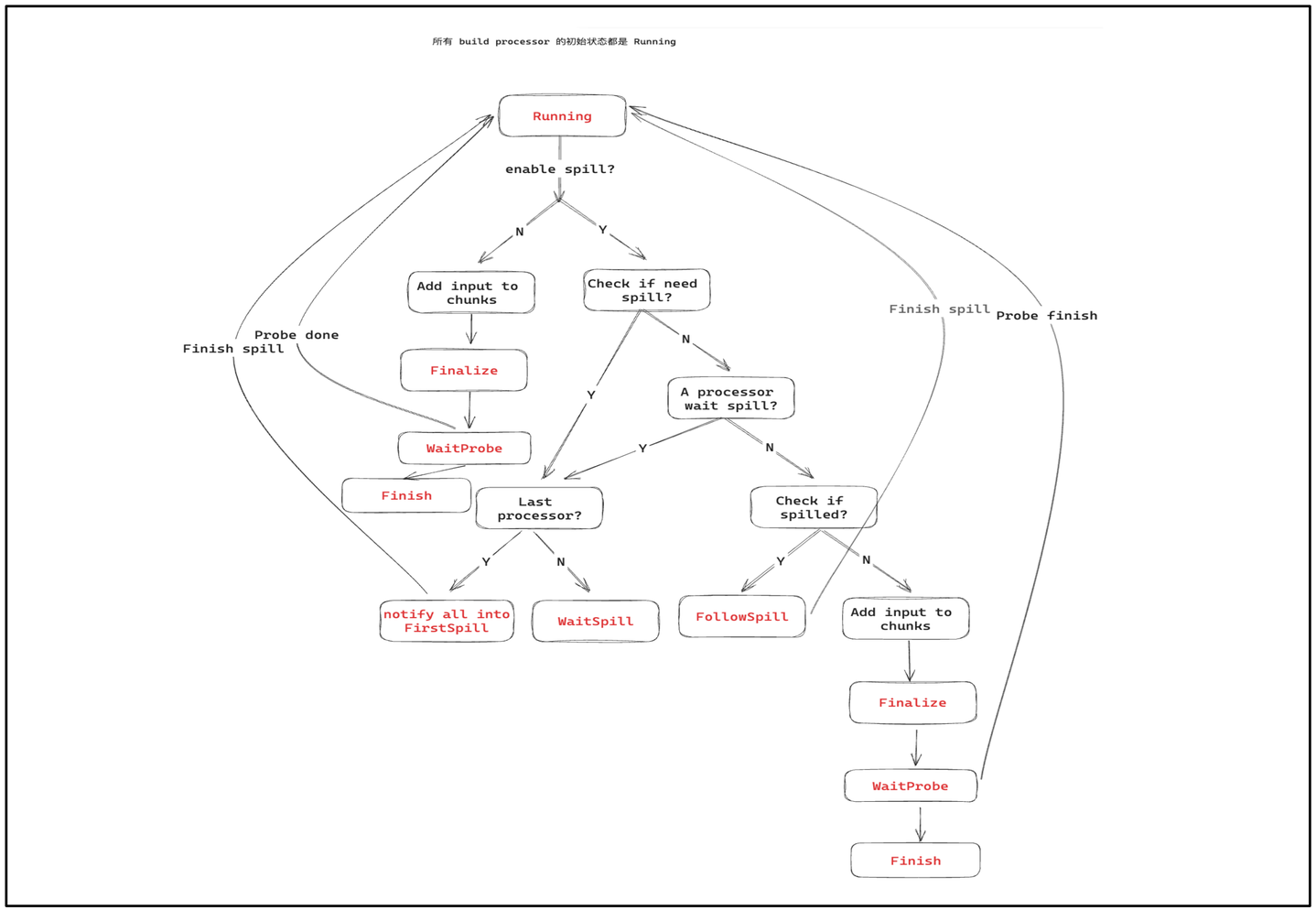
Databend

Part 3

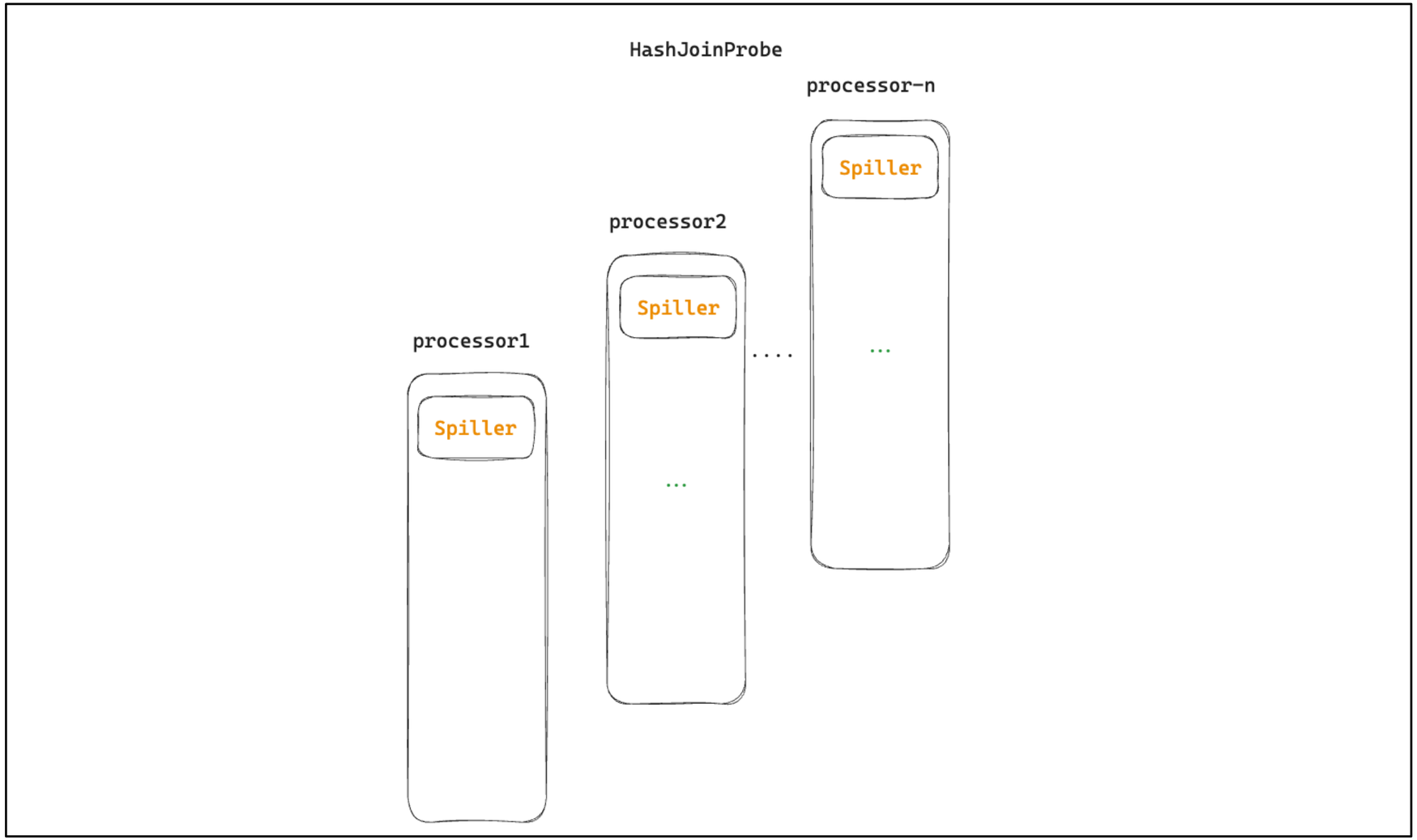
Hash join spill 设计与实现

Hash join spill 实现





Hash join spill 实现



Databend

Thank you!