Serverless Distributed Data Processing

Event-driven Architecture with Azure Functions

Daniel Bin Schmid, Ermias, Kris

Chair of Database Systems,
Technical University of Munich

February 9, 2023

Table of Contents

Pipeline as Black Box

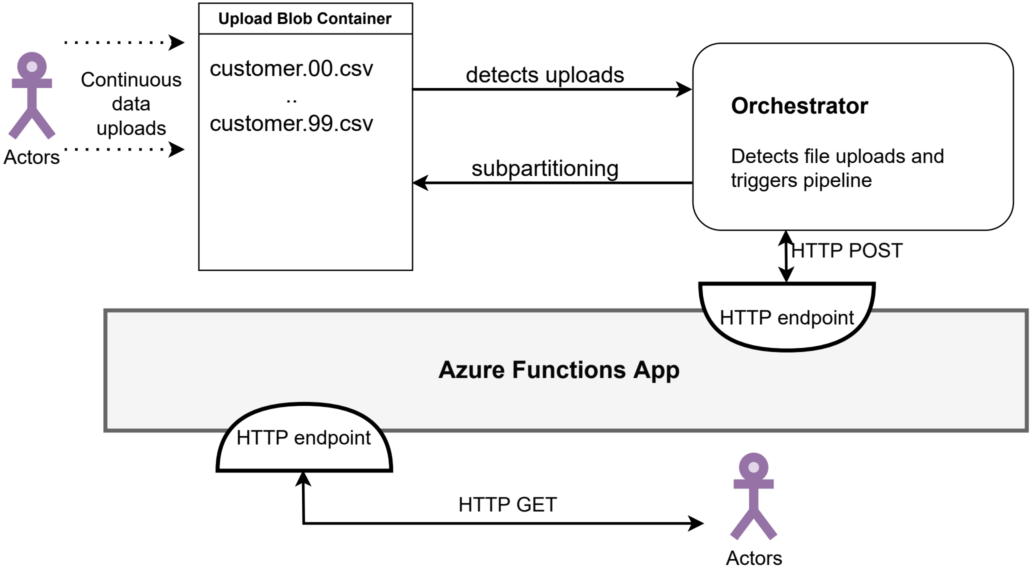

Figure. Functions app is defined by an HTTP GET/ POST API.

Pipeline as White Box

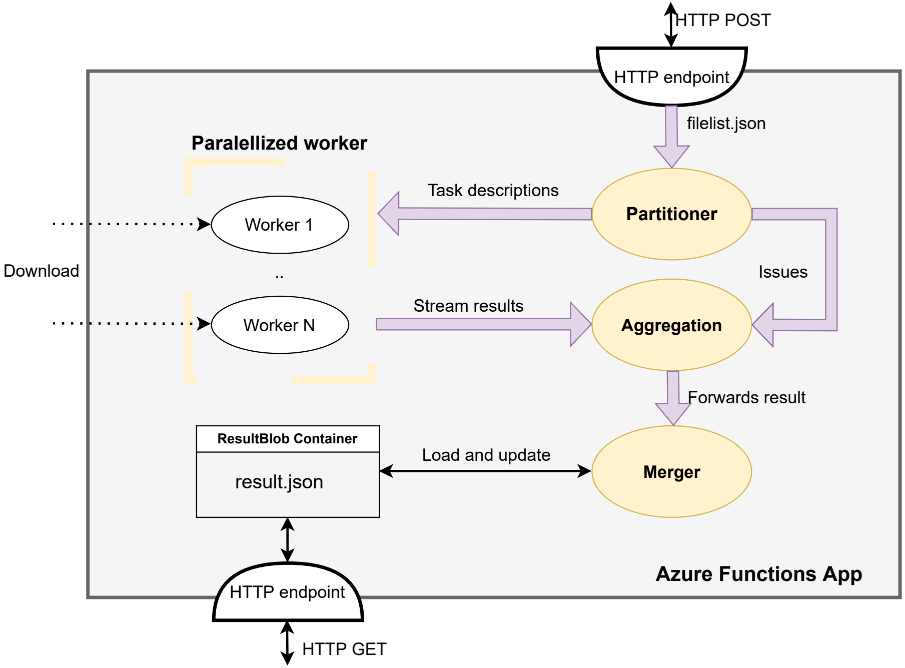

Figure. High level system diagram of Function App.

Scalability: Batch Size and Number of Batches

Scaling the number of batches: $n_{batches} \to \infty$

Assumption: Uniform batches

- Merger must coordinate race conditions
- Merger becomes bottleneck

Scaling the batch size: $s_{batch} \to \infty$

Assumption: Non-uniform tasks within batch

- Aggregation waits for slowest worker
- Aggregation becomes bottleneck

$\implies$ Good scalability if s_{batch} large $\wedge$ tasks within batch uniform

Queue versus Blob Implementation

Azure Queue Storage Implementation

Context

- Information exchange between stages
- Blob-only implementation and Queue-only implementation

General - Azure Queue Storage

- Designed for large amounts of small messages
- Trigger: Function instance for every message
- Fault tolerance: Queue trigger timeouts

Usage - Azure Queue Storage

- Fault tolerant instantiation of functions
- Result collection in aggregation

Queue versus Blob Implementation

Azure Blob Storage Implementation

General - Blob Storage

- Blob of bytes to download and upload via HTTP
- Function trigger for new blob uploads
- Blob directly downloaded for trigger
- Fault tolerance: Poison blobs

Usage - Blob Storage

- Fault tolerant instantiation of functions
- Result collection

BENCHMARKS

Deployment type	$n_{batches}$	Queue pipeline runtime	Blob pipeline runtime
Azure	1	184s	157s
Azure	5	137s	153s
Azure	10	102s	149s
Azure	20	132s	147s
Azure	50	145s	139s
Azure	100	159s	128s
Azure	250	167s	125s
Local	Average	366.33s	780.5s

Figure. Time taken to process 5 GB with different number of batches.

- No significant difference in pricing.

Conclusion

Blob-based pipeline

- Non-uniform completion time of tasks

Queue-based pipeline

- Trade-off between batch size and number of batches scales well