Cloud Computing

Winter Term 2020/2021
Tutorial Session 5

Ilja Behnke, Alexander Acker
Distributed and Operating Systems
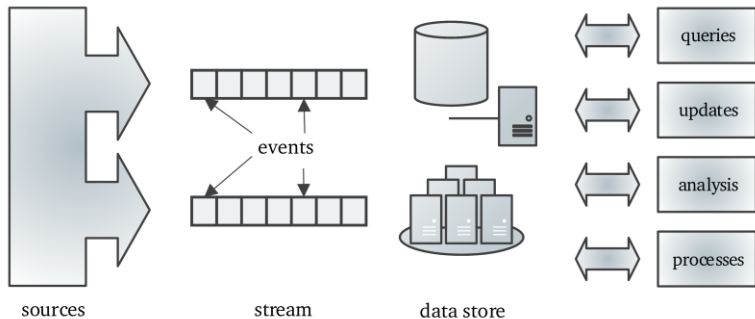
i.behnke@tu-berlin.de

Assignment 5

- Distributed stream processing
- Goal:
  - Get to know Apache Flink and run a distributed application on a small virtual cluster with it.
- Tasks:
  1. Set up infrastructure using virtual machines
  2. Set up Flink on the cluster
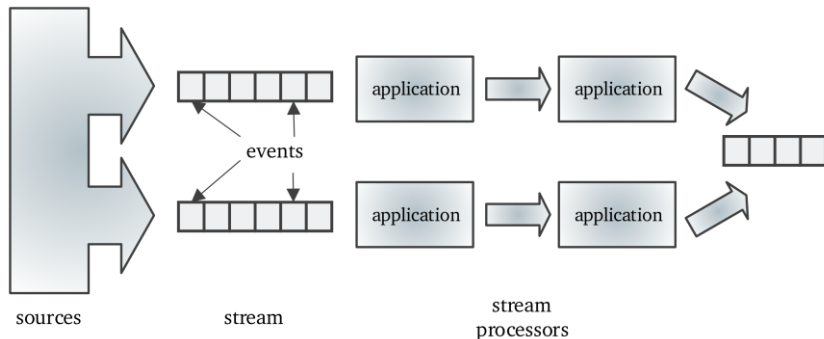 3. Run example application

Batch vs Stream Processing

- Many cloud scenarios work with streams as data sources as events happen continuously (Twitter, customer transactions, website visits…)
- Processing of continuous streams in batches oftentimes not feasible
- Batch processing requires accesses to storage for every action
- Solution: Work on the data stream before storing it (if at all)
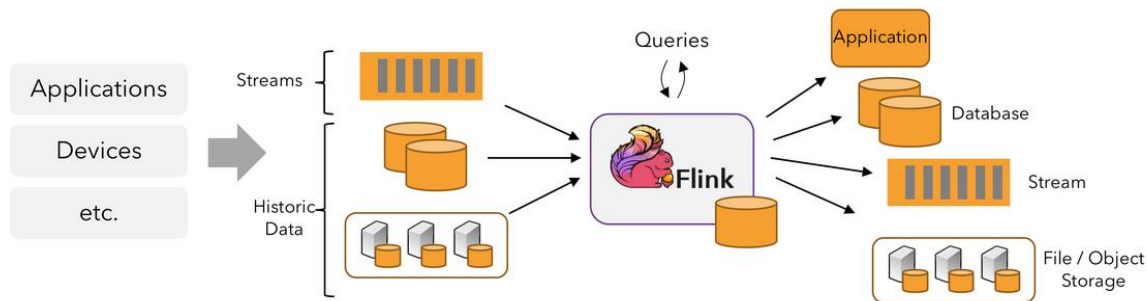
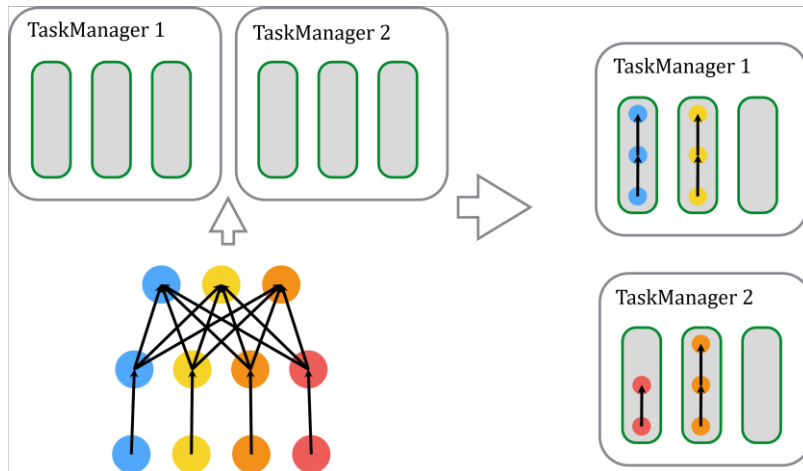Batch Processing

Stream Processing

Apache Flink

- 2nd Gen. distributed stream processing framework started at TU, HU and HPI
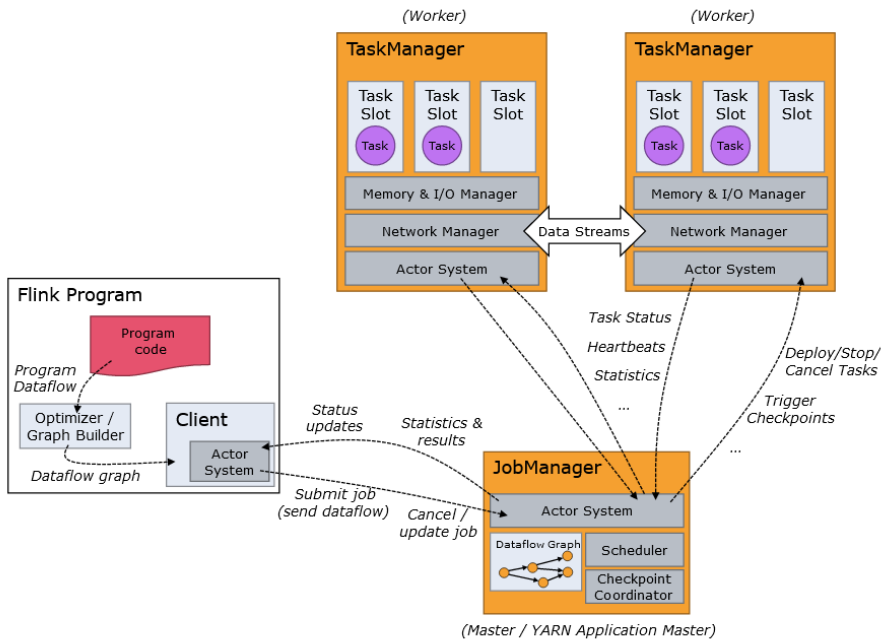- Now in use all over the industry (ebay, pinterest, aws…)

Flink Architecture I

- Flink is designed to work in a distributed manner
  - JobManagers as master nodes
  - TaskManagers as workers
  - Each task manager can run multiple parallel threads
- Jobs (programs) are directed acyclic graphs
- JobManager responsible to schedule a job across infrastructure

Flink Architecture II

Full docs: *https://ci.apache.org/projects/flink/flink-docs-release-1.12/*

Example: Word Count

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Word	Occurrences
lorem	1
ipsum	2
dolor	4
sit	1
amet	5

Practical Assignment 5

- Due: 31.01.2021
- Summary:
  - Get to know Apache Flink in a small experiment on your localhost
  - Prepare 3-10 VMs (in public cloud or locally)
  - Set up Apache Flink to work distributed
  - Count words in a book