Tracing and debugging of parallel computing frameworks for streaming data

Daniel Capelo Borges
May 15, 2020

Pr. Michel Dagenais

Polytechnique Montréal
Laboratoire **DORSAL**

Agenda

- What Streaming Data is...
- Streaming Processing Frameworks
- Research Challenges
- Early results
- Future work
- References

Agenda

- What Streaming Data is...
- Streaming Processing Frameworks
- Research Challenges
- Early results
- Future work
- References

What Streaming Data is...

- Definition from Amazon AWS [1]

  - " Streaming Data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously, and in small sizes (order of Kilobytes). Streaming data includes a wide variety of data such as log files generated by customers using your mobile or web applications, ecommerce purchases, in-game player activity, information from social networks, financial trading floors, or geospatial services, and telemetry from connected devices or instrumentation in data centers. "

What Streaming Data is...

- Definition from Neumeyer [2]
  - “ We define a stream as a sequence of elements (”events”) of the form (K, A) where K and A are the tuple-valued keys and attributes respectively ”

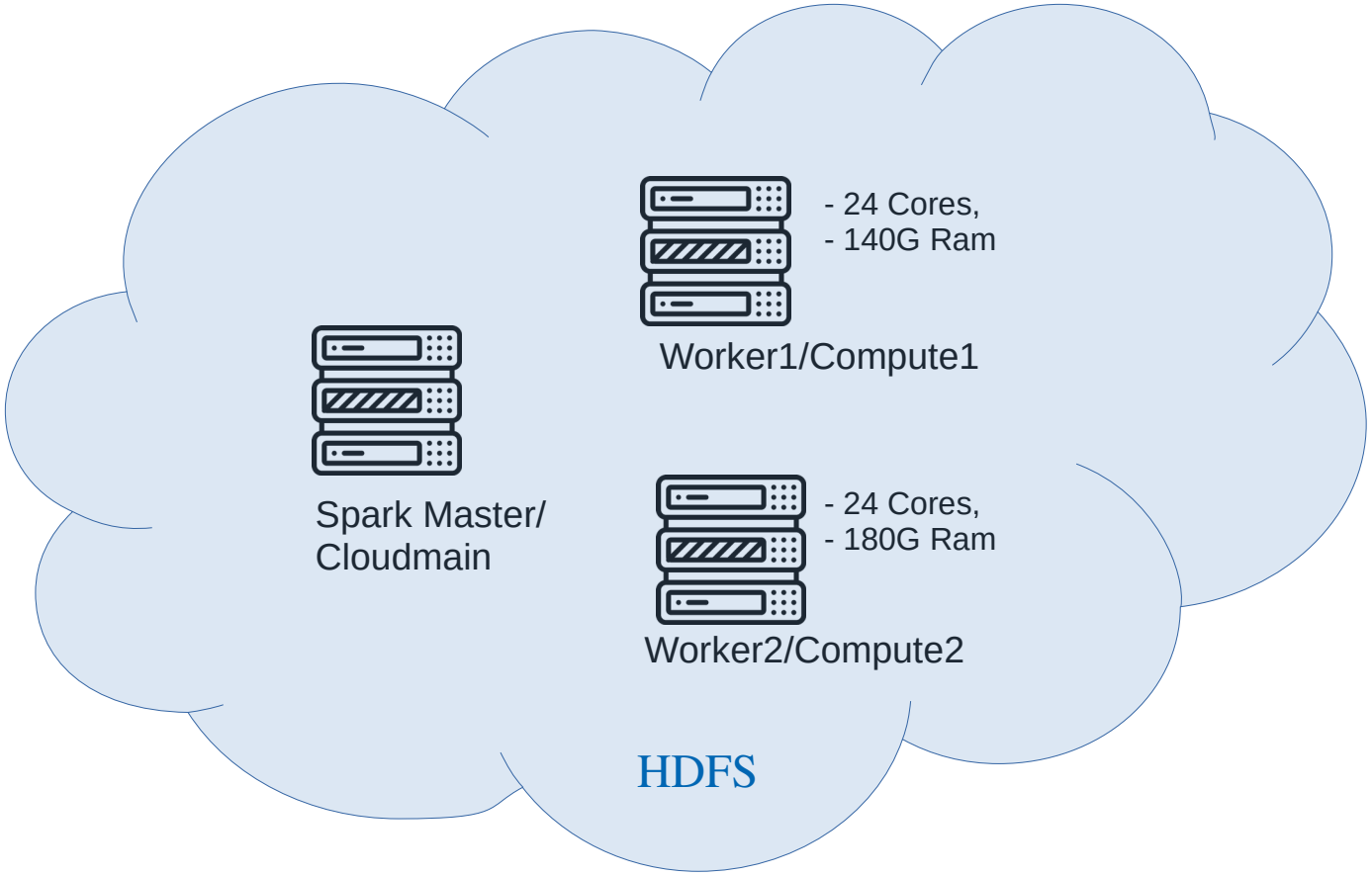
Agenda

- What Streaming Data is...
- Streaming Processing Frameworks
- Research Challenges
- Early results
- Future work
- References

Streaming Processing Frameworks

- We can define Stream Processing as the processing (computing) of data received through data streams as they are generated or received.
- Some use cases:
  - Anomaly Detection,
  - Computer systems and network monitoring [3],
  - Fraud Detection [4],
  - IOT/Smart Devices (Smart Car, Smart Home, etc.) [5],
  - Monitoring (as a production line, supply chain optimizations, etc.),
  - Predictive Maintenance, (e.g. Machine Learning Techniques for Predictive Maintenance),
  - Smart Patient Care,
  - Stock Market Surveillance,
  - Trace Analysis,
  - Traffic Monitoring.
- Some commonly used OS SPF: *Flink, Kafka, Samza, Spark, Storm*, etc.

Streaming Processing Frameworks (Distributed at Dorsal)



Our Little cluster at Dorsal

Agenda

- What Streaming Data is...
- Streaming Processing Frameworks
- Research Challenges
- Early results
- Future work
- References

Research Challenges

- How can we speed up trace analysis with a Streaming Processing Framework?
  - How can we feed data (specifically LTTng traces) into a SPF?
  - How to filter and structure the data before further processing?
- How to efficiently trace execution in a Streaming Processing Framework?
  - What are the most useful metrics, analysis and views to develop to study SPF?
  - How can we use LTTng to trace SPF?
- How can we analyse traces in real-time/streaming with Streaming Processing Framework?
  - How can we use the Spark environment to analyse large/"infinite" traces being streamed out of a large system?

Agenda

- What Apache Spark is ...
- Cloud Platform
- Research Challenges
- Early results
- Future work
- References

Early results…

- Importing/feeding LTTng traces into Apache Spark,
- Streaming Processing Frameworks literature review (in progress).

Agenda

- What Apache Spark is ...
- Cloud Platform
- Research Challenges
- Early results
- Future work
- References

Future work

- Evaluate some Streaming Processing Framework in order to :
  - Speed up trace analysis with a Streaming Processing Framework,
    - feeding data (specifically LTTng traces),
    - filtering and structuring the data before further processing.
  - Efficiently trace execution in a Streaming Processing Framework,
    - choosing the most useful metrics, analysis and views to develop to study SPF.
    - using LTTng to trace SPF.
  - Analyse traces in real-time/streaming with Streaming Processing Framework,
    - using a SPF to analyse large/"infinite" traces being streamed out of a large system.

Agenda

- What Apache Spark is ...
- Cloud Platform
- Research Challenges
- Early results
- Future work
- References

References

- [1] What is Streaming Data?, https://aws.amazon.com/streaming-data/
- [2] Neumeyer, L., Robbins, B., Nair, A., & Kesari, A. (2010, December). S4: Distributed stream computing platform. In 2010 IEEE International Conference on Data Mining Workshops (pp. 170-177). IEEE.
- [3] Gupta, A., Birkner, R., Canini, M., Feamster, N., Mac-Stoker, C., & Willinger, W. (2016, November). Network monitoring as a streaming analytics problem. In Proceedings of the 15th ACM Workshop on Hot Topics in Networks (pp. 106-112).
- [4] Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004, March). Survey of fraud detection techniques. In IEEE International Conference on Networking, Sensing and Control, 2004 (Vol. 2, pp. 749-754). IEEE.
- [5] D'silva, G. M., Khan, A., & Bari, S. (2017, May). Real-time processing of IoT events with historic data using Apache Kafka and Apache Spark with dashing framework. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1804-1809). IEEE.

Questions?

Thanks for your attention !

Daniel Capelo Borges

daniel.capelo@polymtl.ca