Parallel streaming trace processing in Spark

Daniel Capelo Borges
December 9, 2019

Pr. Michel Dagenais

Polytechnique Montréal
Laboratoire **DORSAL**

Agenda

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

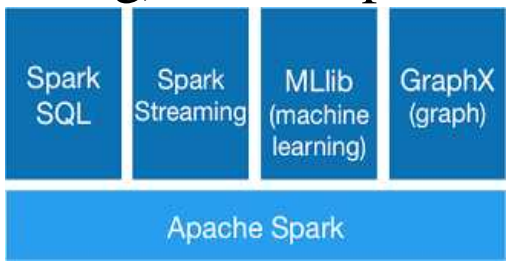
Agenda

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

What Apache Spark is… 1/4

Introduction

- Apache Spark is an open source cluster computing framework and it uses memory operations divided into several processing phases. Like Apache Hadoop, Apache Spark makes uses of the MapReduce[1] programming model published by Google to perform distributed computing across several nodes.
- Apache Spark works with batch and stream processing.
- Apache Spark includes built-in libraries like Spark SQL (for working with structured data), Spark Streaming (for building scalable fault-tolerant streaming applications), MLib (Machine Learning) and GraphX (for graphs and graph-parallel computation).



Spark libraries – from https://spark.apache.org/

What Apache Spark is… 2/4

- **Some use cases...**

Analyzing Log Data With Apache Spark

- https://databricks.com/session/analyzing-log-data-with-apache-spark

Netflix's Recommendation ML Pipeline Using Apache Spark

- https://databricks.com/session/netflixs-recommendation-ml-pipeline-using-apache-spark

Analyzing 2TB of Raw Trace Data from a Manufacturing Process: A First Use Case of Apache Spark for Semiconductor Wafers from Real Industry

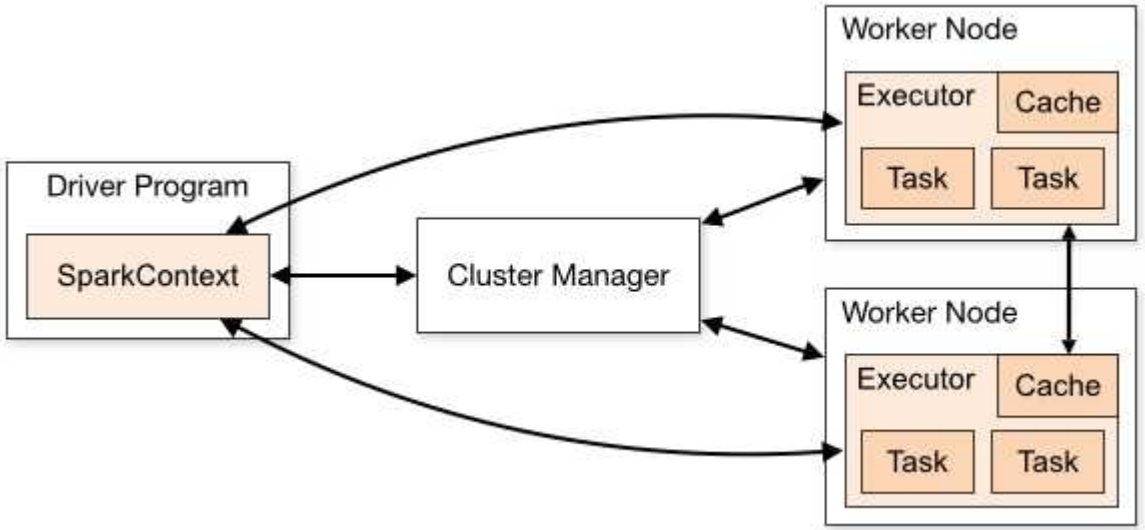
- https://databricks.com/session/analyzing-2tb-of-raw-trace-data-from-a-manufacturing-process-a-first-use-case-of-apache-spark-for-semiconductor-wafers-from-real-industry

Automating Predictive Modeling at Zynga with PySpark and Pandas UDFs

- https://databricks.com/session/automating-predictive-modeling-at-zynga-with-pyspark-and-pandas-udfs

What Apache Spark is… 3/4

- Architecture

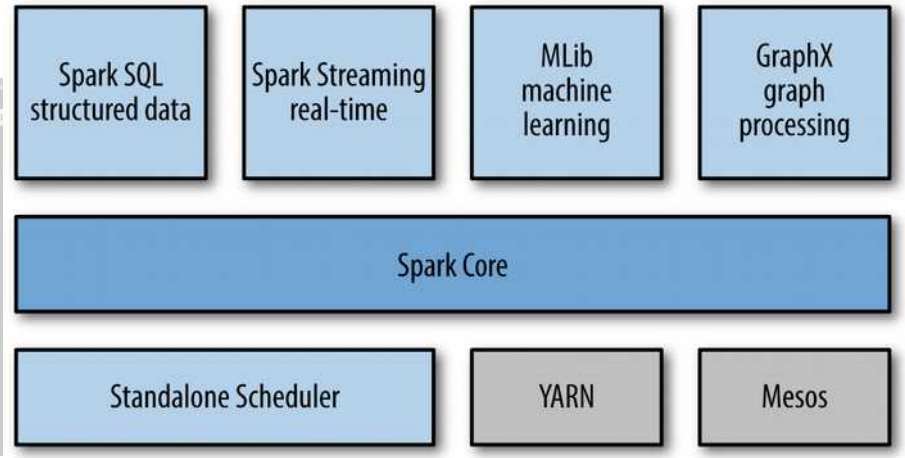


Spark architecture – from https://spark.apache.org/docs/latest/cluster-overview.html

What Apache Spark is… 4/4

Deployment

- Apache Spark can run locally,
- as a Standalone Cluster or on Hadoop YARN Cluster !

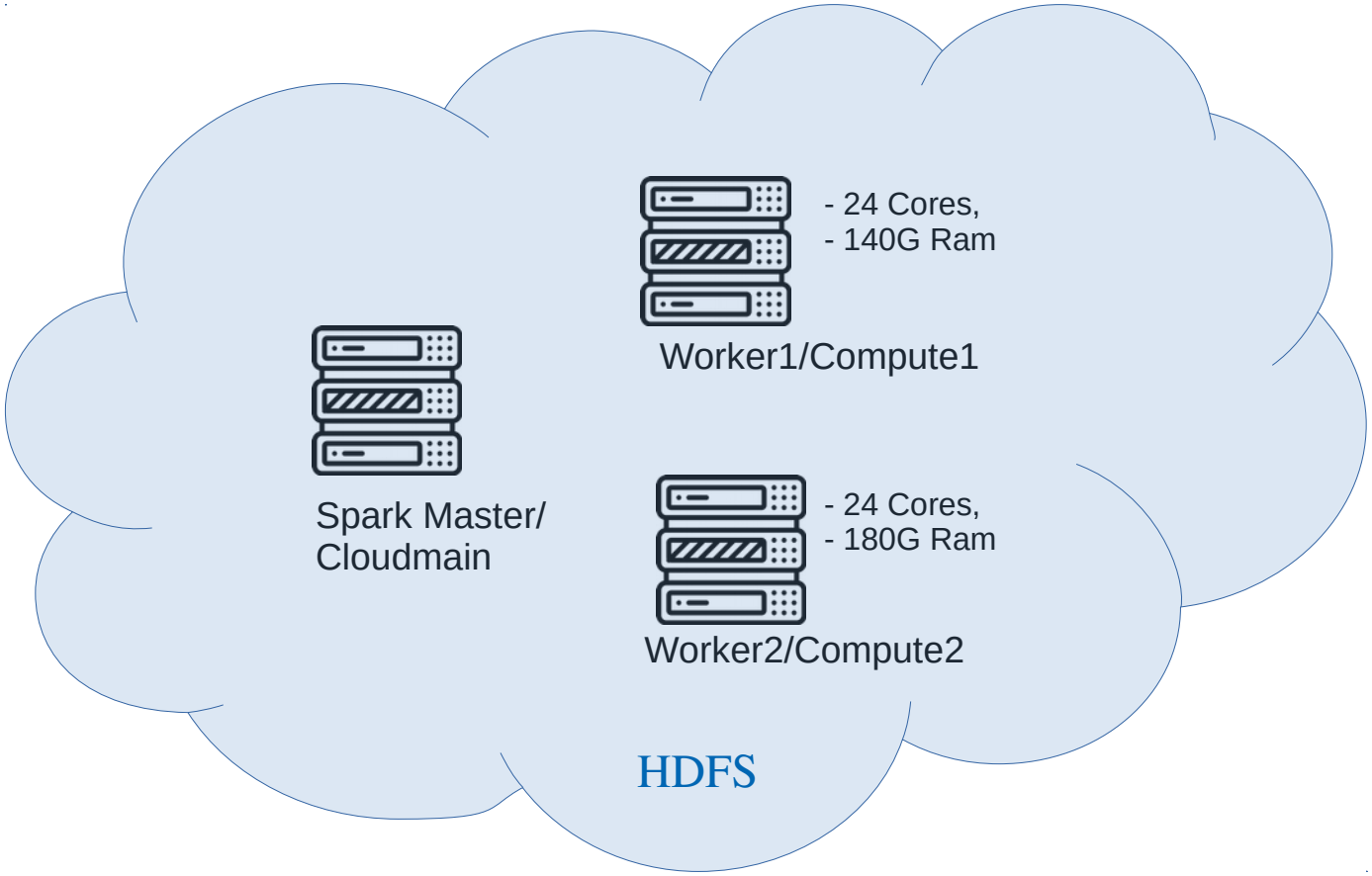


The Apache Spark stack from [2]

Agenda

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

Cloud Platform 1/1



Agenda

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

Research Challenges

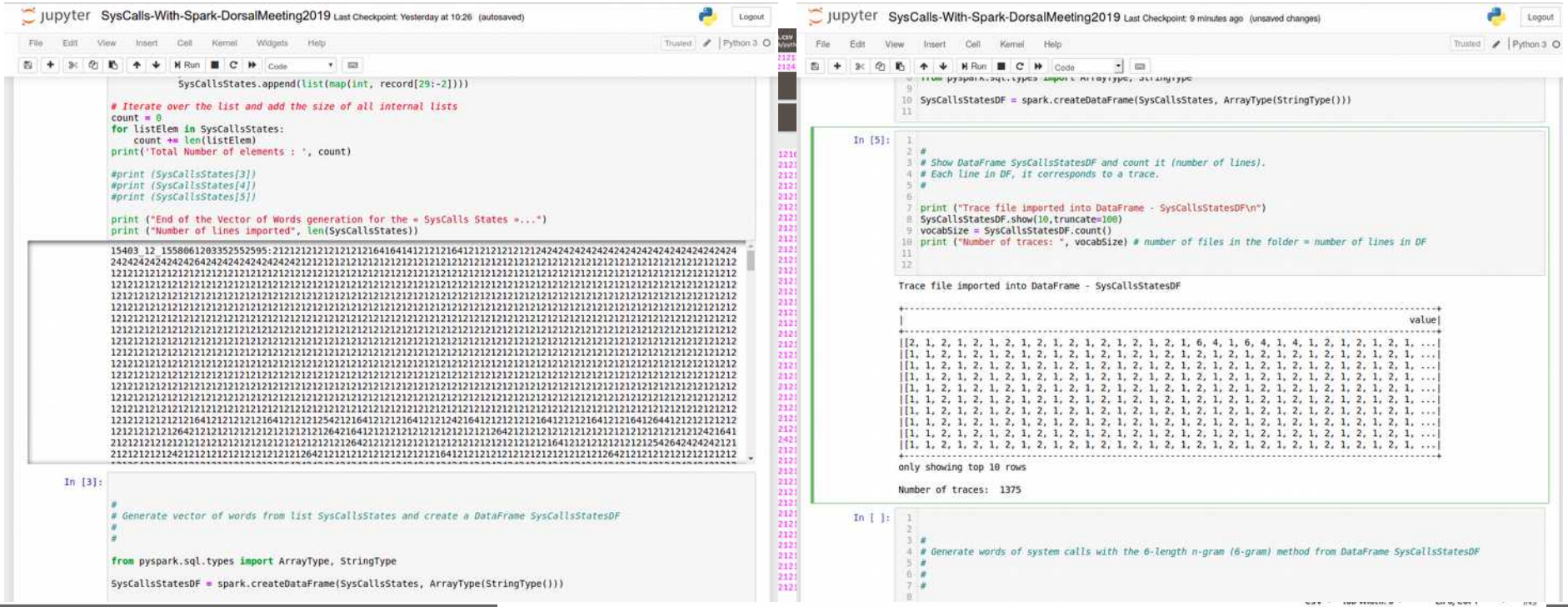
- How can we speed up trace analysis with Spark?
  - How can we import data (specifically LTTng traces) into Spark?
  - How to filter and structure the data before further processing?
- How to efficiently trace execution in a Spark environment?
  - What are the most useful metrics, analysis and views to develop to study Spark systems?
  - How can we use LTTng to trace Spark?
- How can we analyse traces in real-time/streaming with Spark?
  - 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

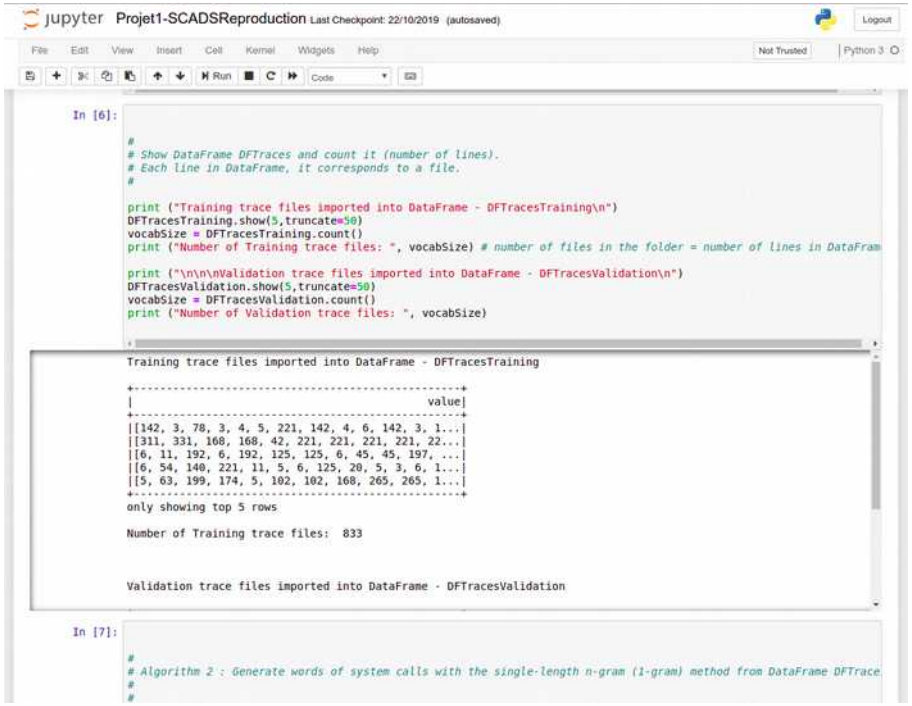
Early results… 1/2

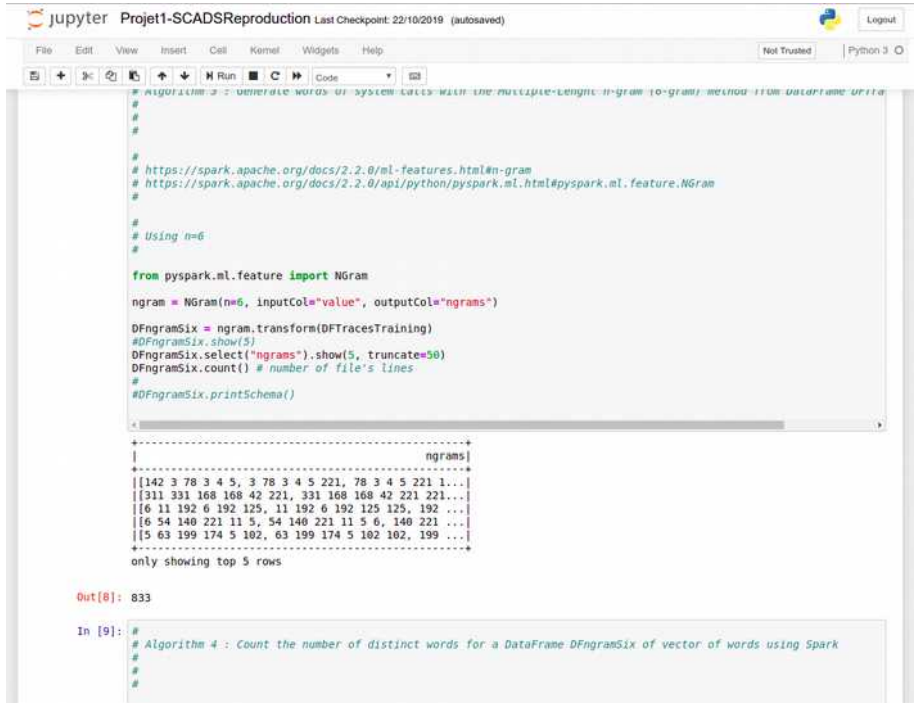
- Many tests/experiments with Apache Spark locally, Standalone Cluster and Hadoop YARN cluster,
- Importing/feeding LTTng traces into Apache Spark,
- Some experiments with Apache Spark + ML libraries (clustering) and tracing data.



Early results… 2/2

- Reproducing some papers (trace, logs and sentiment analysis; anomaly detection, etc).





Agenda

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

Future work

- ✗ How can we speed up trace analysis with Spark?
  - ✔ How can we import data (specifically LTTng traces) into Spark?
  - ✗ How to filter and structure the data before further processing?
- ✗ How to efficiently trace execution in a Spark environment?
  - ✗ What are the most useful metrics, analysis and views to develop to study Spark systems?
  - ✗ How can we use LTTng to trace Spark?
- ✗ How can we analyse traces in real-time/streaming with Spark?
  - ✗ 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

References

- [1] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
- [2] Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). Learning spark: lightning-fast big data analysis. " O'Reilly Media, Inc.".
- [3] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. HotCloud, 10(10-10), 95.
- [4] Veiga, J., Expósito, R. R., Pardo, X. C., Taboada, G. L., & Tourifio, J. (2016, December). Performance evaluation of big data frameworks for large-scale data analytics. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 424-431). IEEE.

Questions?

Thanks for your attention !

Daniel Capelo Borges

daniel.capelo@polymtl.ca