Mississauga Big Data Meetup 21

-Kumar V

October 8, 2016

www.meetup.com/Mississauga-Big-Data-Analytics-Meetup/

About This Meetup

- Nothing net new in the presentations done in this meetup
- A number of videos, software, books, etc are available on the internet, for free.
- This meetup will provide you a venue to present the stuff you learn, while you also learn from others’ presentations. This is a sure way to accelerate your learning.
- So, this is an invitation for all of you to learn some aspect of the very big ecosystem of Big Data, and present to the group.

About this meetup

- The plan is to meet at least once a month
- We can also plan virtual meetings, also once a month
- The next meet is planned on Nov 13, at the same place as this.

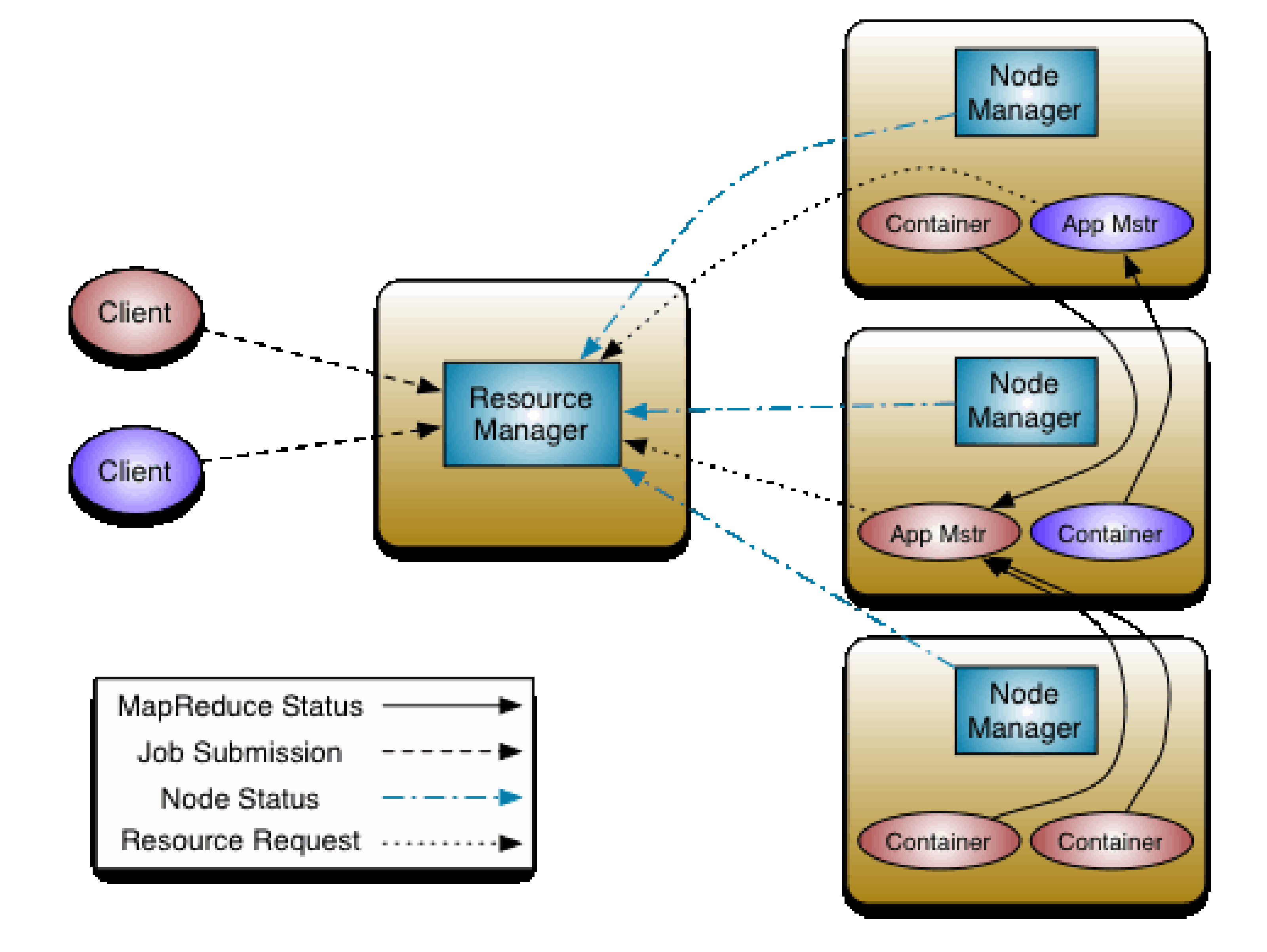
Agenda for today

- A quick introduction to YARN
- Spark Computing Model (We will not focus on installation, configuration, etc)
- SparkSQL
- Examples run on Dataproc (Hadoop on Google Cloud Platform)

- Introduction to Machine Learning
  - Cure Fitting. The idea of over fitting.
  - Intro to Python ecosystem
  - Using Scikit-learn

- Plan for next sessions

YARN

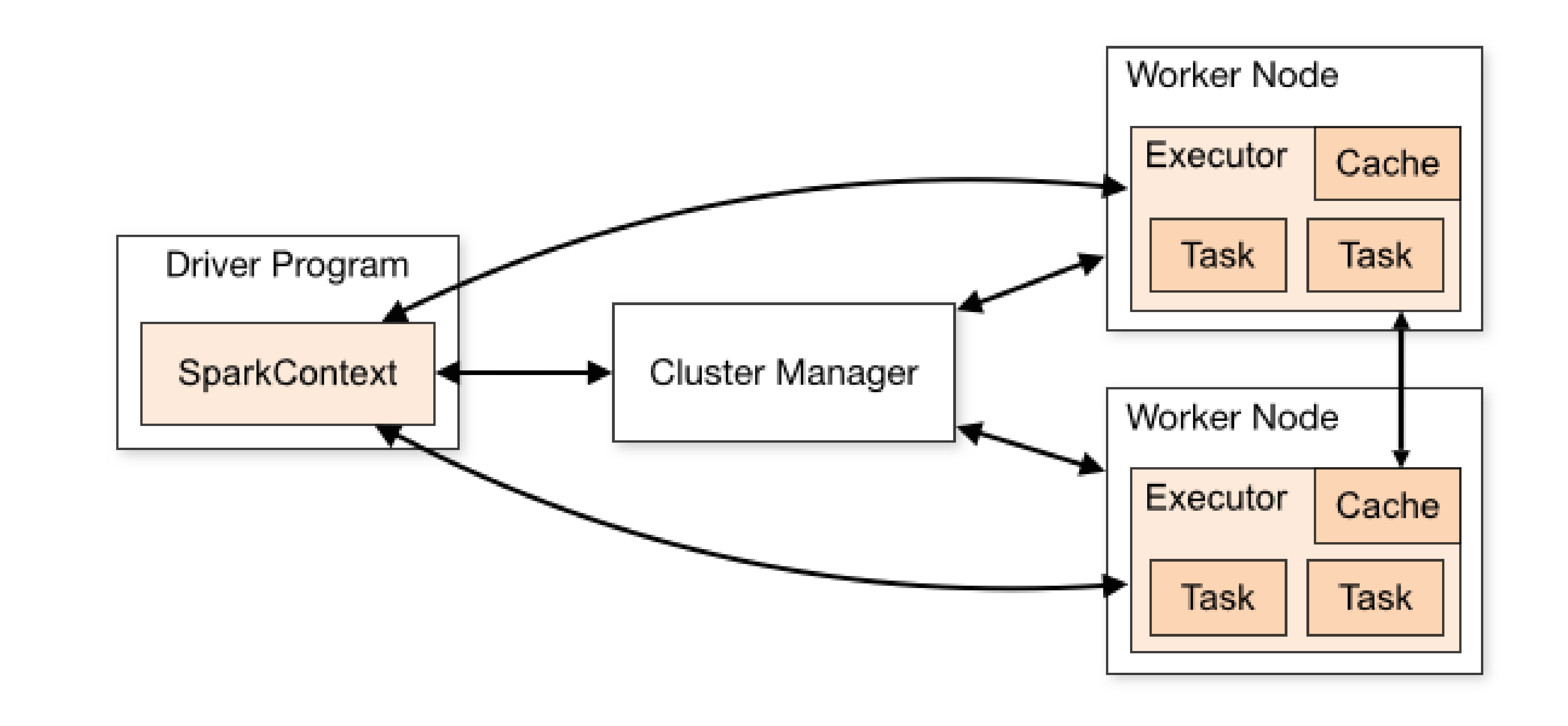

Node Manager
Container
App Mstr
Client
Resource Manager
Client
Node Manager
App Mstr
Container
Node Manager
Container
Container
MapReduce Status
Job Submission
Node Status
Resource Request

Resource Manager

- Scheduler
  - Fair Scheduler (Scheduling based on tenants/apps getting equal share of resources. By default it is only memory that is considered. CPU can be added.
  - Capacity Scheduler (Capacity guarantees)
- Applications Manager
  - Accepts job submissions, negotiates the first container for job specific app master.

Process Model

Spark jobs on a cluster are co-ordinated by SparkContext object that runs on the driver program. SparkContext represents the application session.



Spark Environments

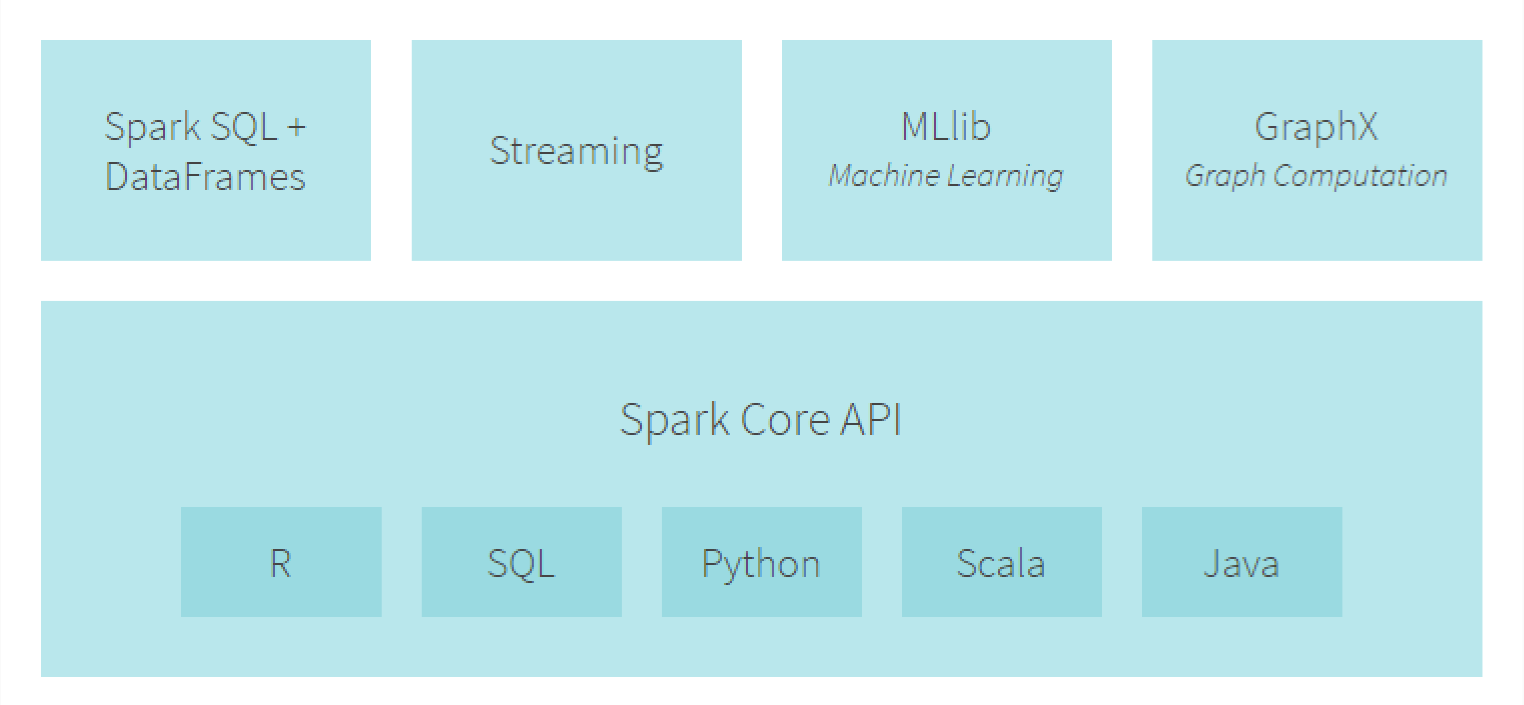
Spark supports three kinds of Cluster Managers

- Standalone
- Apache Mesos
- YARN

You can program Spark in the following languages:

- Scala, Python, Java and R

Apache Spark Ecosystem



Spark SQL + DataFrames

Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark Core API

R

SQL

Python

Scala

Java

Spark Core API

- Spark Computing Model
  - RDDs / Pair RDDs
  - RDD Operations
    - Transformations
    - Actions
  - Shared Variables
    - Broadcast Variables
    - Accumulators
  - Shuffle
  - Partitions and Repartitioning
  - RDD persistence
  - Stages in a Spark Job

Spark SQL

- Using Dataframes

Python Ecosystem

- Python Scripting Language
- Numpy, Matplotlib, Pandas, Scikit-learn
- REPL (Read-eval-print Loop)
- Jupyter

Next Steps

- Spark Internals (Look under the hood)
- More Spark Programs
- MLLib and GraphX
- Using UDFs in SparkSQL
- Internal workings of Spark SQL Joins
- Use Scikit-learn to run Machine Learning Algorithms on data (on Kaggle, for instance)
- Pick a Machine Learning Algorithm for study