Practical Introduction to Apache Spark

About Me

About Me

I'm not...

About Me

I'm not...

- A guru in Apache Spark

About Me

I'm not...

- A guru in Apache Spark
- A writer, I didn't write 📕 about Spark

About Me

I'm not...

- A guru in Apache Spark
- A writer, I didn't write 📕 about Spark
- 10x Engineer

About Me

I am…

- @marcraminv in Twitter
- 0.8976x Data Engineer last 4 years
- ScalaBcn & SparkBcn co-organizer
- Everis, Billy Mobile & LIDL … Now IntentHQ

About Me

I am…

- @marcraminv in Twitter
- 0.8976x Data Engineer last 4 years
- ScalaBcn & SparkBcn co-organizer
- Everis, Billy Mobile & LIDL … Now IntentHQ

And THIS IS MY FIRST MEETUP 🤦‍♂️🎉

About you

1. Go to https://www.mentimeter.com
2. Put the code
3. Vote 🖖🙂

Today we are going to talk...

- What is Spark
- How it works
- SparkSession
- Operations
- Your first Spark Application

Time ago when people were using
Map and Reduce...

What is Apache Spark

Spark: Cluster Computing with Working Sets

Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica
University of California, Berkeley

Abstract

MapReduce and its variants have been highly successful in implementing large-scale data-intensive applications on commodity clusters. However, most of these systems are built around an acyclic data flow model that is not suitable for other popular applications. This paper focuses on one such class of applications: those that reuse a working set of data across multiple parallel operations. This includes many iterative machine learning algorithms, as well as interactive data analysis tools. We propose a new framework called Spark that supports these applications while retaining the scalability and fault tolerance of MapReduce. To achieve these goals, Spark introduces an abstraction called resilient distributed datasets (RDDs). An RDD is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost. Spark can outperform Hadoop by 10x in iterative machine learning jobs, and can be used to interactively query a 39 GB dataset with sub-second response time.

1 Introduction

A new model of cluster computing has become widely

MapReduce/Dryad job, each job must reload the data from disk, incurring a significant performance penalty.

- **Interactive analytics**: Hadoop is often used to run ad-hoc exploratory queries on large datasets, through SQL interfaces such as Pig [21] and Hive [1]. Ideally, a user would be able to load a dataset of interest into memory across a number of machines and query it repeatedly. However, with Hadoop, each query incurs significant latency (tens of seconds) because it runs as a separate MapReduce job and reads data from disk.

This paper presents a new cluster computing framework called Spark, which supports applications with working sets while providing similar scalability and fault tolerance properties to MapReduce.

The main abstraction in Spark is that of a *resilient distributed dataset* (RDD), which represents a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost. Users can explicitly cache an RDD in memory across machines and reuse it in multiple MapReduce-like *parallel operations*. RDDs achieve fault tolerance through a notion of *lineage*: if a partition of an RDD is lost, the RDD has enough infor-

What is Apache Spark

What is Apache Spark

Web definitions

Spark is a unified analytics engine for large-scale data processing

What is Apache Spark

Web definitions

Spark is a unified analytics engine for large-scale data processing

Fast and general-purpose cluster computing system

What is Apache Spark

Web definitions

Spark is a unified analytics engine for large-scale data processing

Fast and general-purpose cluster computing system

Books

Unified computing engine for parallel processing

What is Apache Spark

Web definitions

Spark is a unified analytics engine for large-scale data processing

Fast and general-purpose cluster computing system

Books

Unified computing engine for parallel processing

Multiple language support: Scala, Python, SQL, R, Java, .Net

What is Apache Spark

Web definitions

Spark is a unified analytics engine for large-scale data processing

Fast and general-purpose cluster computing system

Books

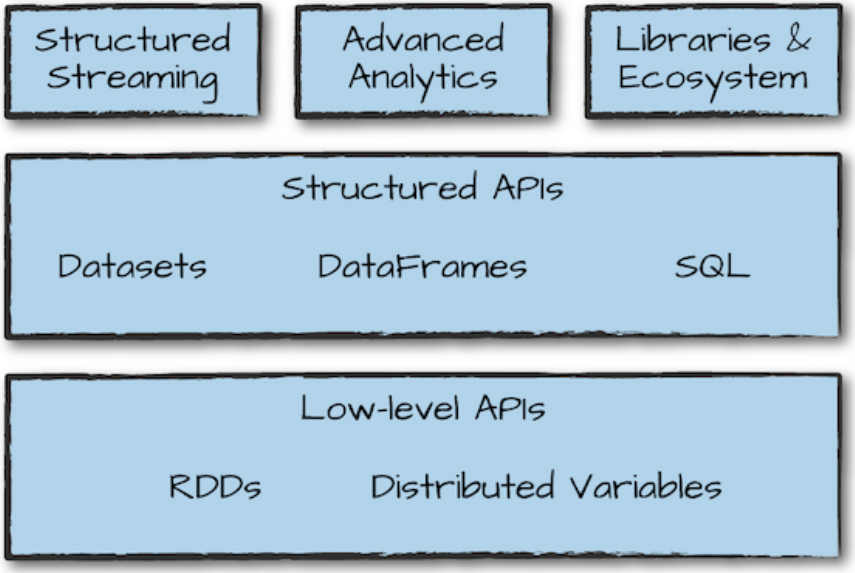
Unified computing engine for parallel processing

Multiple language support: Scala, Python, SQL, R, Java, .Net

Then…

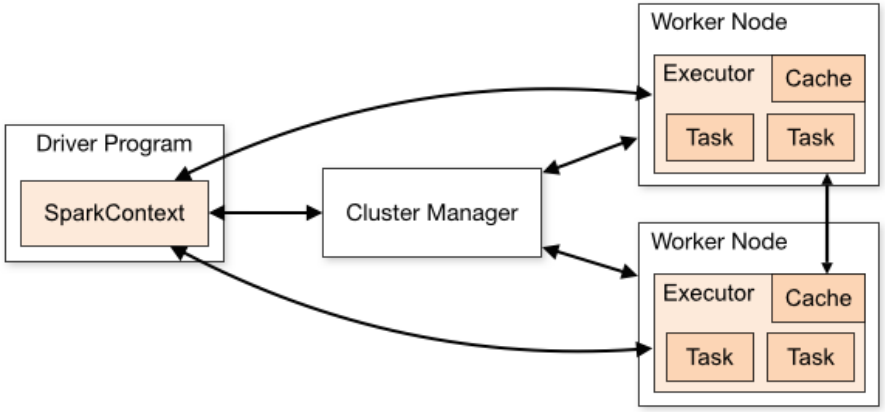
- Unified engine
- Parallel/cluster computing system
- Multiple platform

What is Apache Spark



[Oreilly Spark Definitive Guide Preview]

How it Works



Spark Cluster Overview

SparkSession => Unit

EntryPoint: SparkSession

class `pyspark.sql.`**`SparkSession`**(*sparkContext*, *jsparkSession=None*) [source]

The entry point to programming Spark with the Dataset and DataFrame API.

A SparkSession can be used create **`DataFrame`**, register **`DataFrame`** as tables, execute SQL over tables, cache tables, and read parquet files. To create a SparkSession, use the following builder pattern:

```
>>> spark = SparkSession.builder \
...     .master("local") \
...     .appName("Word Count") \
...     .config("spark.some.config.option", "some-value") \
...     .getOrCreate()
```

`builder`

A class attribute having a **`Builder`** to construct **`SparkSession`** instances

class **`Builder`** [source]

Builder for **`SparkSession`**.

`appName`(*name*) [source]

Sets a name for the application, which will be shown in the Spark web UI.

If no application name is set, a randomly generated name will be used.

Parameters: **name** – an application name

New in version 2.0.

Python API

EntryPoint: SparkSession

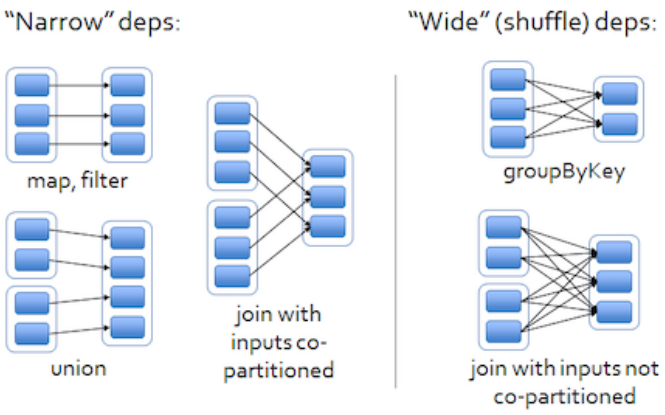
- This is the door to code your applications
- Access to the Spark world...

```
>>> spark = SparkSession.builder \
...     .master("local") \
...     .appName("Word Count") \
...     .config("spark.some.config.option", "some-value") \
...     .getOrCreate()

// generate a 1 column with N rows
spark.range(100)

// create a custom dataframe frm DataTypes
spark.createDataFrame(...)

// build a sql expression
spark.sql("SELECT 'HOLA IRONHACK 👋'")

// read from csv file
spark.read.csv(...).load()
```

Operations

Transformations

- Narrow (isolated operation)
- Wide (network communication between workers)

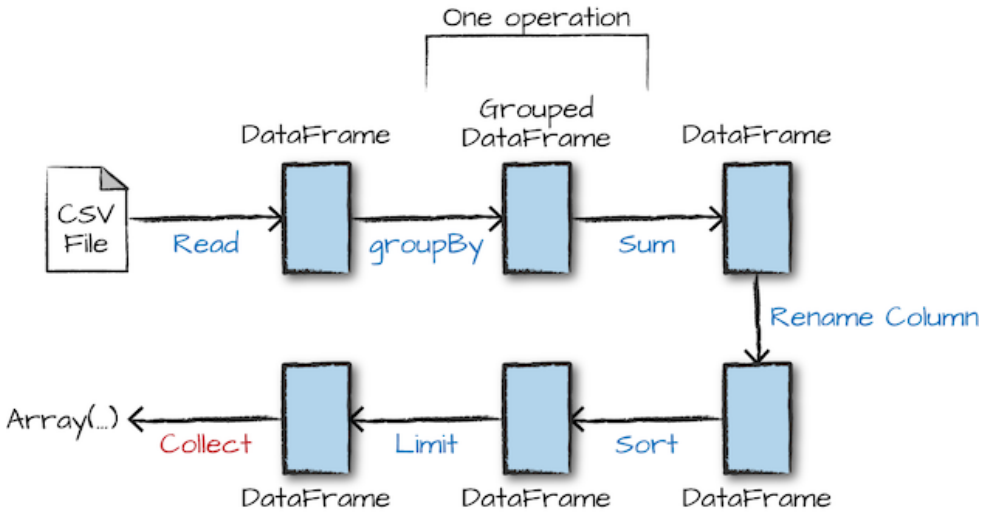


*All transformations in Spark are lazy

Operations

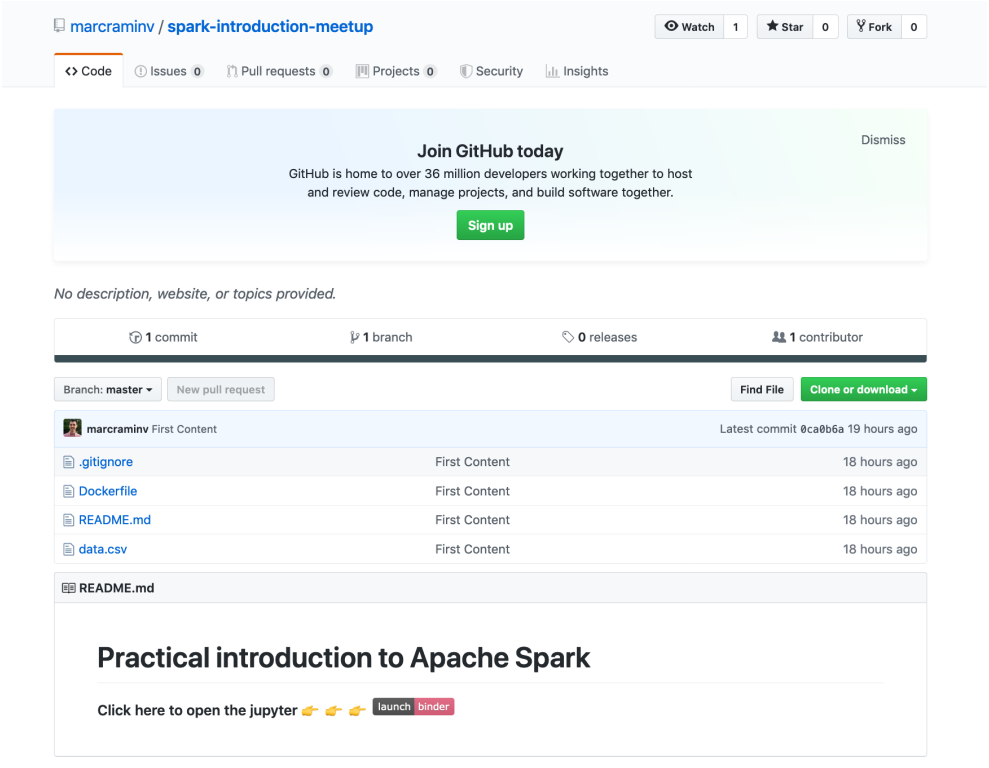
Actions

Return a value to the driver program after running a computation on the dataset



Source: Spark Definitive Guide

Time to code


marcraminv / spark-introduction-meetup
Watch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 0
Security
Insights
Dismiss
Join GitHub today
GitHub is home to over 36 million developers working together to host and review code, manage projects, and build software together.
Sign up
No description, website, or topics provided.
1 commit
1 branch
0 releases
1 contributor
Branch: master
New pull request
Find File
Clone or download
marcraminv First Content
Latest commit 0ca0b6a 19 hours ago
.gitignore First Content 18 hours ago
Dockerfile First Content 18 hours ago
README.md First Content 18 hours ago
data.csv First Content 18 hours ago
README.md
Practical introduction to Apache Spark
Click here to open the jupyter 👉 👉 👉 launch binder

Start with...

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("SimpleApp").getOrCreate()
```

Questions to answers

- Top 5 of best rated players
- Top 5 of pokemon with best attack
- Top 5 of pokemon group by `pokemon_type`
- Top 5 of players group by teams