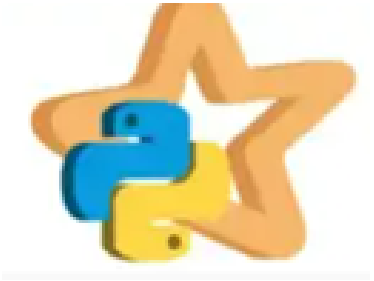


PYSPARK

Introduction to PySpark

What is PySpark?

- PySpark is a Spark library written in Python to run Python applications using Apache Spark capabilities, using PySpark we can run applications parallelly on the distributed cluster (multiple nodes).
- PySpark is a powerful tool for distributed data processing and analysis, and provides a convenient and efficient programming interface for working with Spark in Python.
- PySpark is a Python API for Apache Spark. Apache Spark is an analytical processing engine for large scale powerful distributed data processing and machine learning applications.

Introduction to PySpark

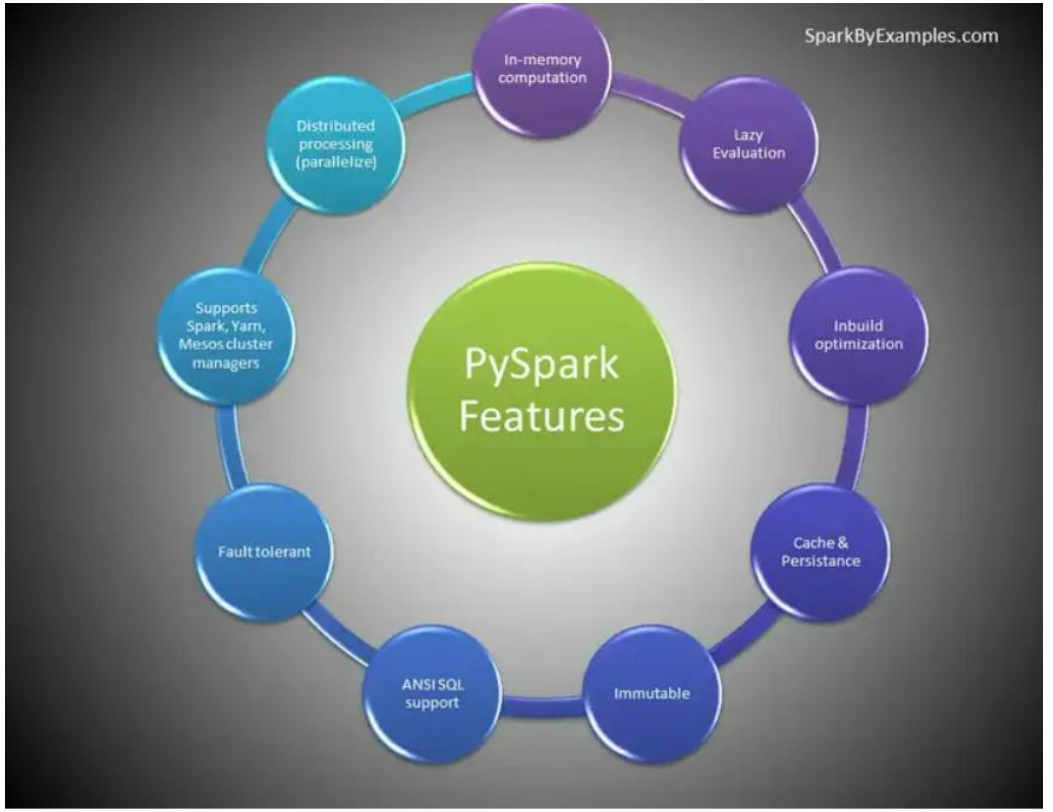
- PySpark provides a simple and easy-to-use programming interface for working with Spark, allowing developers to quickly and easily develop data processing and analysis applications.
- It provides access to a wide range of Spark functionality, including distributed data processing, machine learning, graph processing, and more.

One of the main benefits

- PySpark is its ability to handle large-scale datasets efficiently.
- PySpark distributes data across multiple nodes in a cluster, allowing for parallel processing and significantly reducing processing times for large datasets.
- PySpark also provides a variety of built-in functions and libraries for data processing and analysis, making it a powerful tool for data scientists and engineers.

Features

- In-memory computation
- Distributed processing using parallelize
- Can be used with many cluster managers (Spark, Yarn, Mesos e.t.c)
- Fault-tolerant
- Immutable
- Lazy evaluation
- Cache & persistence
- Inbuild-optimization when using DataFrames
- Supports ANSI SQL

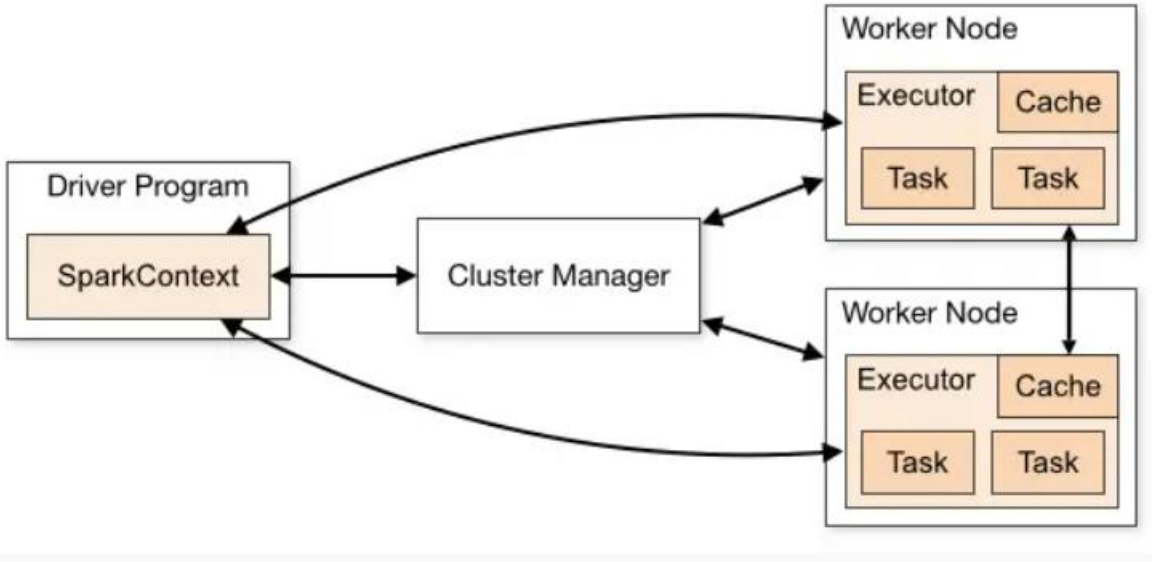


Advantages of PySpark

- PySpark is a general-purpose, in-memory, distributed processing engine that allows you to process data efficiently in a distributed fashion.
- Applications running on PySpark are 100x faster than traditional systems.
- You will get great benefits using PySpark for data ingestion pipelines.
- Using PySpark we can process data from Hadoop HDFS, AWS S3, and many file systems.
- PySpark also is used to process real-time data using Streaming and Kafka.
- Using PySpark streaming you can also stream files from the file system and also stream from the socket.
- PySpark natively has machine learning and graph libraries.

PySpark Architecture

- Apache Spark works in a master-slave architecture where the master is called “Driver” and slaves are called “Workers”. When you run a Spark application, Spark Driver creates a context that is an entry point to your application, and all operations (transformations and actions) are executed on worker nodes, and the resources are managed by Cluster Manager.



Cluster Manager Types

Spark supports below cluster managers:

- Standalone – a simple cluster manager included with Spark that makes it easy to set up a cluster.
- Apache Mesos – Mesons is a Cluster manager that can also run Hadoop MapReduce and PySpark applications.
- Hadoop YARN – the resource manager in Hadoop 2. This is mostly used, cluster manager.
- Kubernetes – an open-source system for automating deployment, scaling, and management of containerized applications.
- local – which is not really a cluster manager but still I wanted to mention as we use “local” for master() in order to run Spark on your laptop/computer.

PySpark Modules & Packages