University of Novi Sad
Faculty of Technical Sciences
November 2022

Introduction to Apache Spark with Python

Miroslav Tomić, Teaching Assistant

DATA SCIENCE
/CONFERENCE/

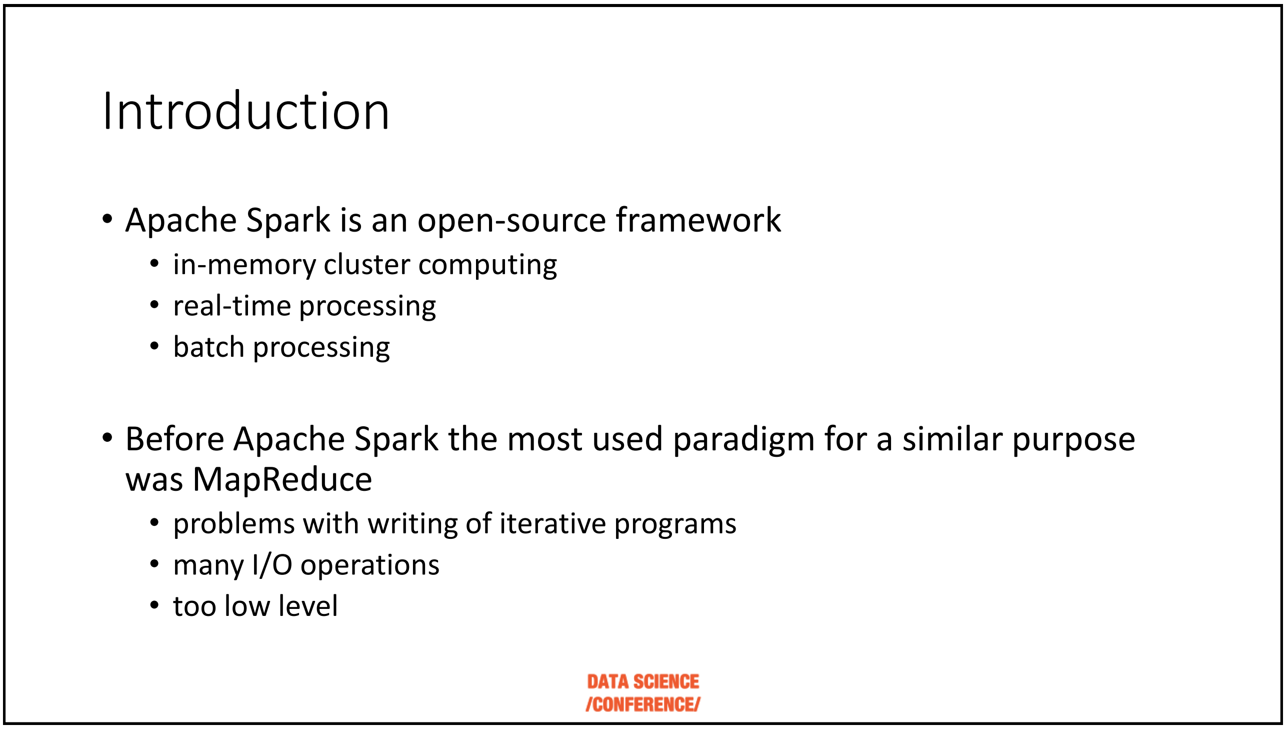
1

Agenda

- Introduction
- Apache Spark
- PySpark
- Tutorial environment
- Resilient Distributed Dataset
- DataFrame
- Dataset
- Spark MLlib

DATA SCIENCE
/CONFERENCE/

2

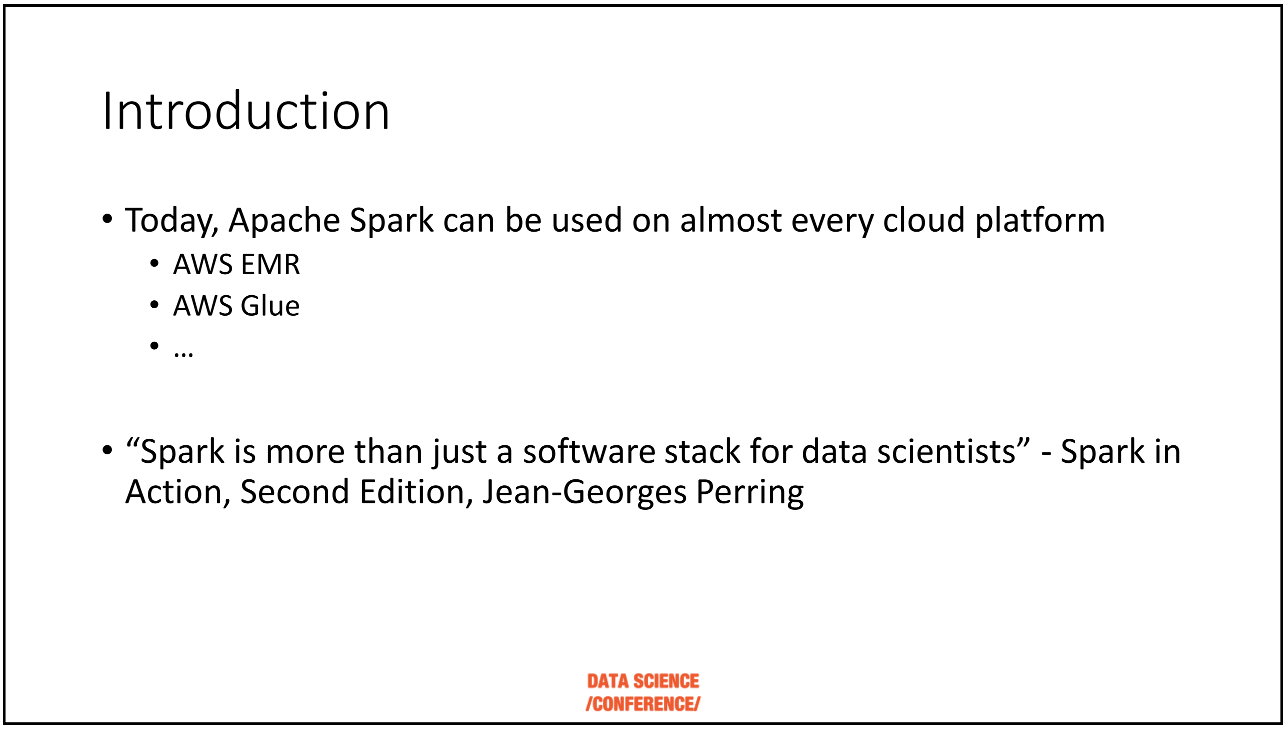


Introduction

- Apache Spark is an open-source framework
  - in-memory cluster computing
  - real-time processing
  - batch processing
- Before Apache Spark the most used paradigm for a similar purpose was MapReduce
  - problems with writing of iterative programs
  - many I/O operations
  - too low level

DATA SCIENCE
/CONFERENCE/

3



Introduction

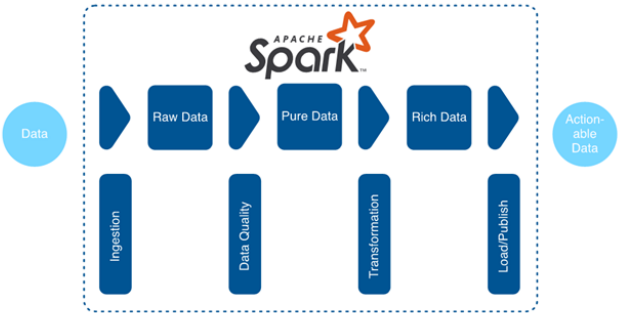
- Today, Apache Spark can be used on almost every cloud platform
  - AWS EMR
  - AWS Glue
  - …
- “Spark is more than just a software stack for data scientists” - Spark in Action, Second Edition, Jean-Georges Perring

DATA SCIENCE
/CONFERENCE/

4

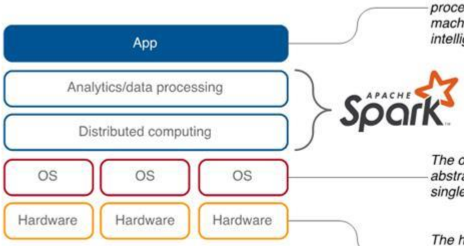
Introduction

- Apache Spark is used for
  - data ingestion
    - from multiple sources
  - data cleansing
    - data quality of processed data
  - data transformation
  - data load/publish
    - loading data in data warehouse, a BI, saving in file…



DATA SCIENCE /CONFERENCE/

Apache Spark - architecture



Your applications leverage a powerful set of API for data processing, analytics, machine learning, artificial intelligence…

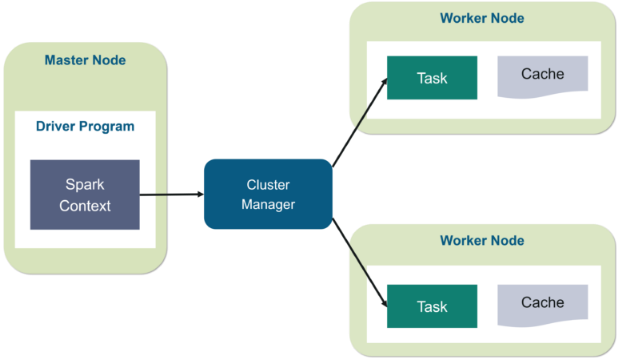
The operating system abstracts the hardware of a single node

The hardware is completely abstracted by the operating system and Spark

DATA SCIENCE /CONFERENCE/

Apache Spark - architecture

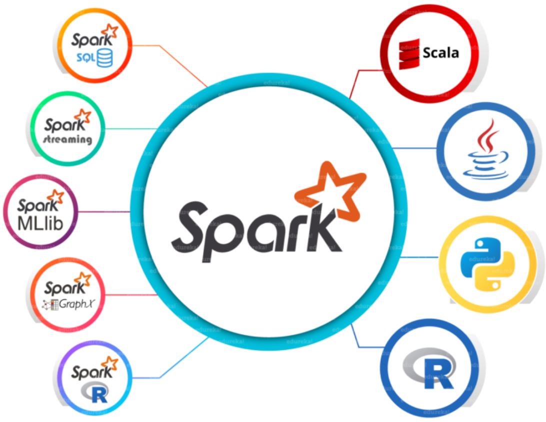
- Master/worker architecture



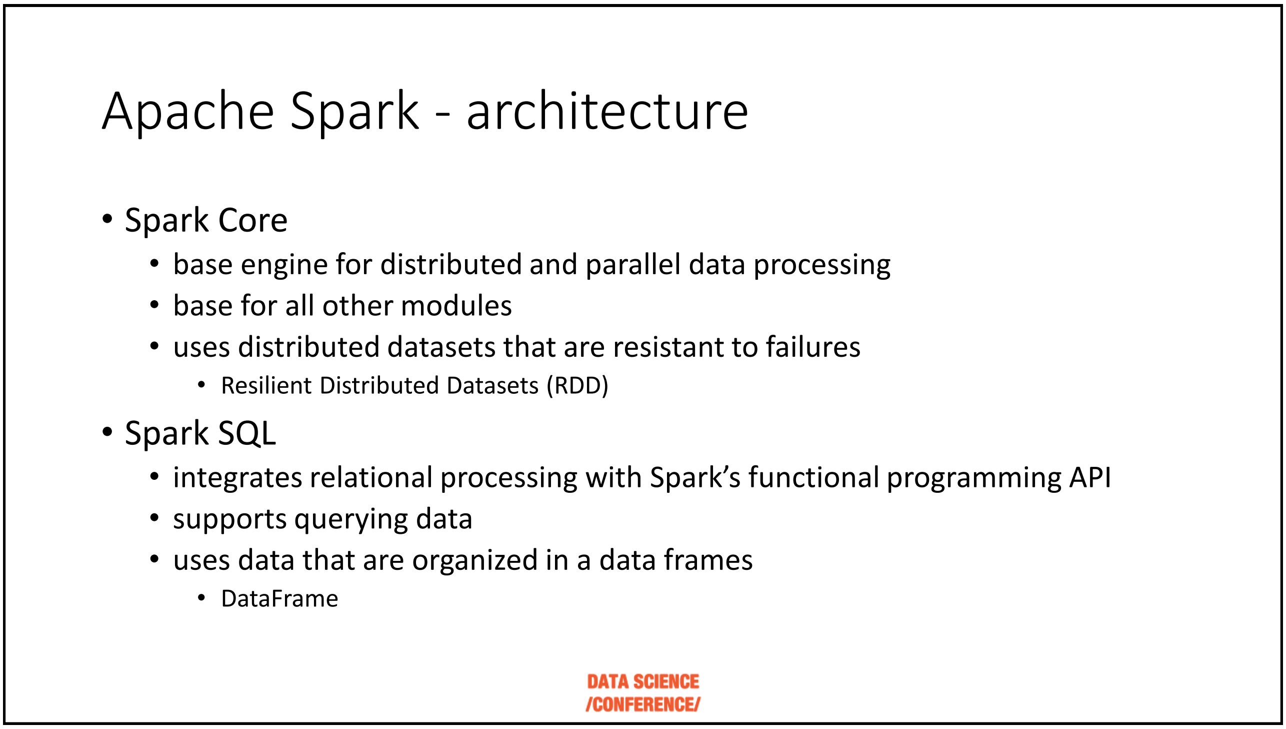
7

Apache Spark - architecture

- Spark components and layers are loosely coupled
- Architecture base components
  - Spark Core
  - Spark SQL
  - Spark MLlib
  - Spark Streaming
  - GraphX
  - SparkR



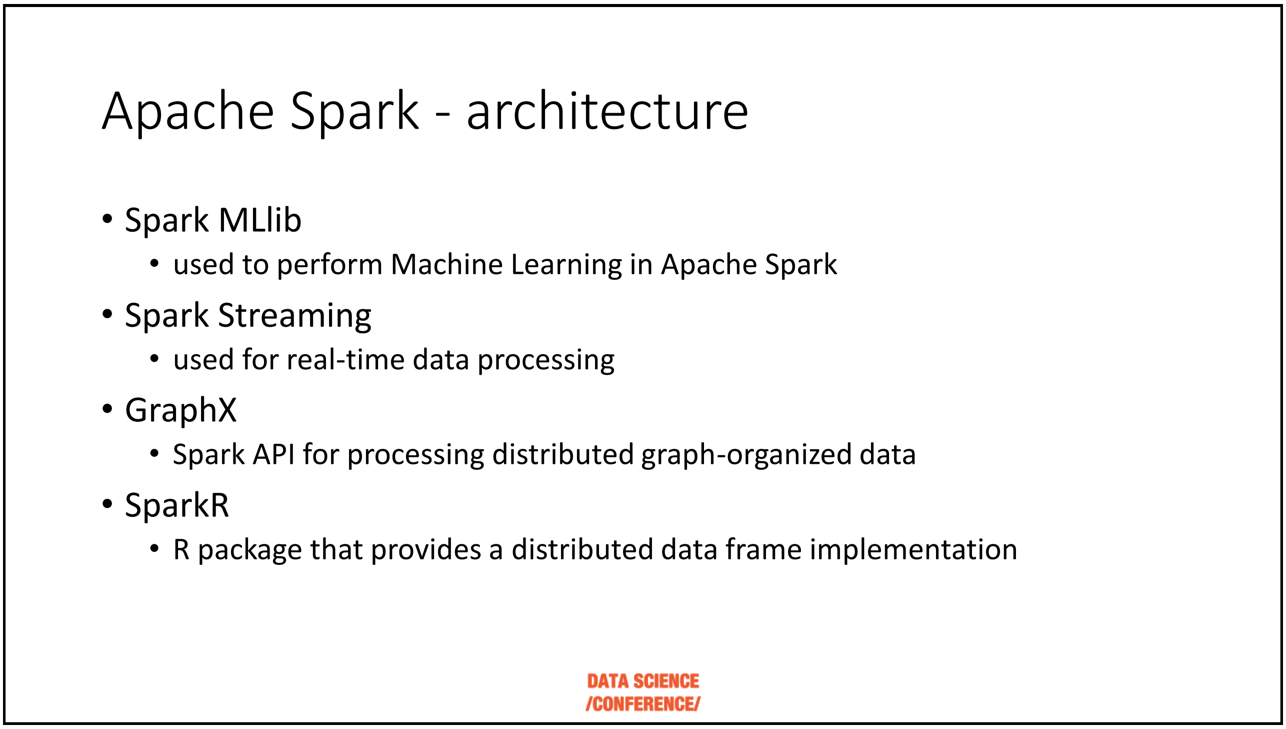
8



Apache Spark - architecture

- Spark Core
  - base engine for distributed and parallel data processing
  - base for all other modules
  - uses distributed datasets that are resistant to failures
    - Resilient Distributed Datasets (RDD)
- Spark SQL
  - integrates relational processing with Spark’s functional programming API
  - supports querying data
  - uses data that are organized in a data frames
    - DataFrame

9



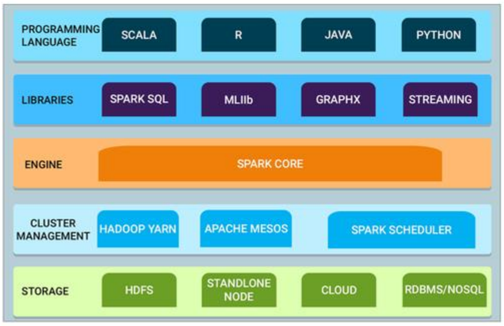
Apache Spark - architecture

- Spark MLlib
  - used to perform Machine Learning in Apache Spark
- Spark Streaming
  - used for real-time data processing
- GraphX
  - Spark API for processing distributed graph-organized data
- SparkR
  - R package that provides a distributed data frame implementation

10

Apache Spark - architecture

- Spark code can be written and provides high-level API in
  - Java
  - Scala
  - Python
  - R
- Provides shell in Scala and Python



DATA SCIENCE
/CONFERENCE/

11

PySpark

- PySpark is an interface for Apache Spark in Python
- Current version 3.3.1
- Official documentation
  - https://spark.apache.org/docs/latest/api/python/index.html

DATA SCIENCE
/CONFERENCE/

12

PySpark

- PySpark supports most of Spark’s features such as Spark SQL, DataFrame, Streaming, MLlib and Spark Core

DATA SCIENCE
/CONFERENCE/

13

Tutorial environment

- Docker is used for simulating distributed computing
  - docker-spark/docker-compose.yml
  - spark containers – distributed executors
    - spark-master
    - spark-worker1
    - spark-worker2
  - hdfs containers – distributed file system
    - namenode
    - datanode1
    - datanode2
  - visualizing data
    - hue

DATA SCIENCE
/CONFERENCE/

14

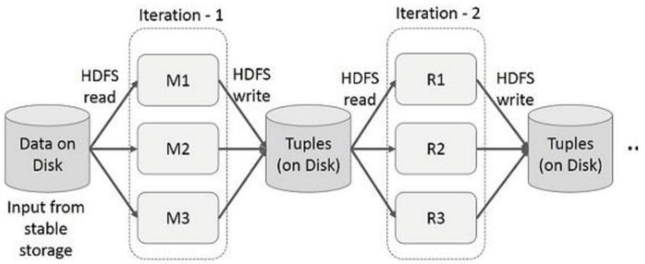
Tutorial environment

- For the purpose of python coding
  - anaconda environment with python 3.10
    - jupyter lab
    - PySpark

DATA SCIENCE
/CONFERENCE/

Resilient Distributed Dataset

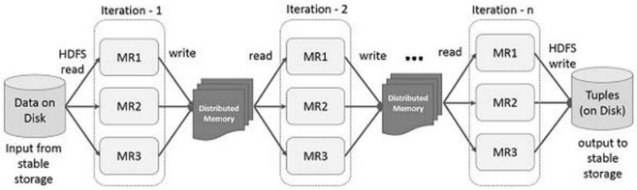
- Iterative map-reduce programs had trouble with
  - slow memory sharing
  - writing to disk after every map-reduce step
  - many I/O calls are required for the desired result



DATA SCIENCE
/CONFERENCE/

Resilient Distributed Dataset

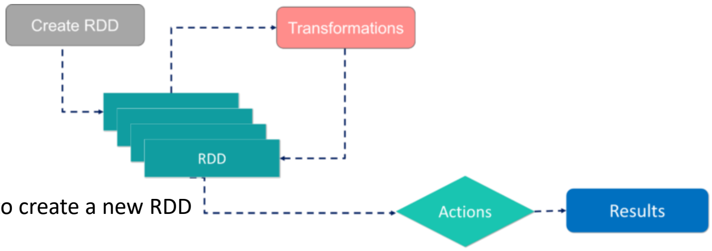
- RDD
  - data is distributed across nodes that belong to the cluster
  - stored in-memory
  - immutable dataset
  - resistant to failures due to partitioning and data replication
  - eliminates many I/O calls



DATA SCIENCE
/CONFERENCE/

17

Resilient Distributed Dataset



- RDD
  - supports
    - transformations
      - operations that are applied to create a new RDD
      - on one partition – pipelining
      - multiple partitions – shuffle
    - actions
      - applied on an RDD to instruct Apache Spark to apply computation and pass the result back to the driver

DATA SCIENCE
/CONFERENCE/

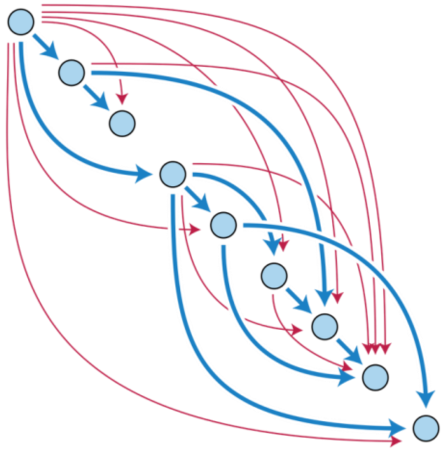
18

Resilient Distributed Dataset

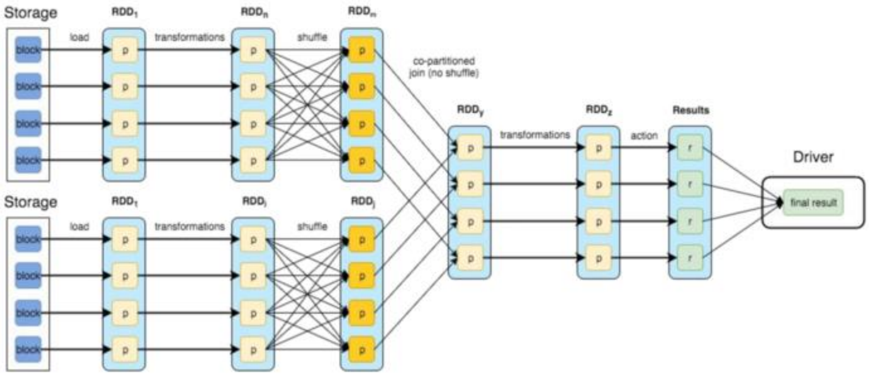
- Transformations
  - pipelining
    - Map
    - FlatMap
    - MapPartition
    - Filter
    - Sample
    - Union
  - shuffle
    - Intersection
    - Distinct
    - ReduceByKey
    - AggregateByKey
    - SortByKey
    - Join
    - Cartesian
    - Repartition
    - Coalesce
- Actions
  - Count
  - CounyByKey
  - Collect
  - First
  - Take
  - Top
  - CountByValue
  - Reduce
  - Fold
  - Aggregate
  - Foreach
  - SaveAsText
  - SaveAsSequenceFile
  - SaveAsObjectFile

Resilient Distributed Dataset - Directed Acyclic Graph

- Directed Acyclic Graph (DAG)
  - contain a series of actions connected to each other in a workflow
  - internal representation of programs for data processing
    - a base for distributing RDDs and tasks in the cluster
  - unlike the MapReduce, it supports the existence of more than two processing phases

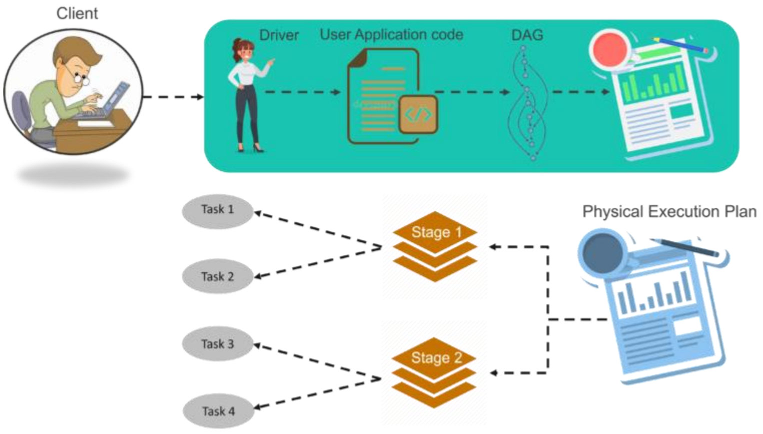


Resilient Distributed Dataset - Directed Acyclic Graph

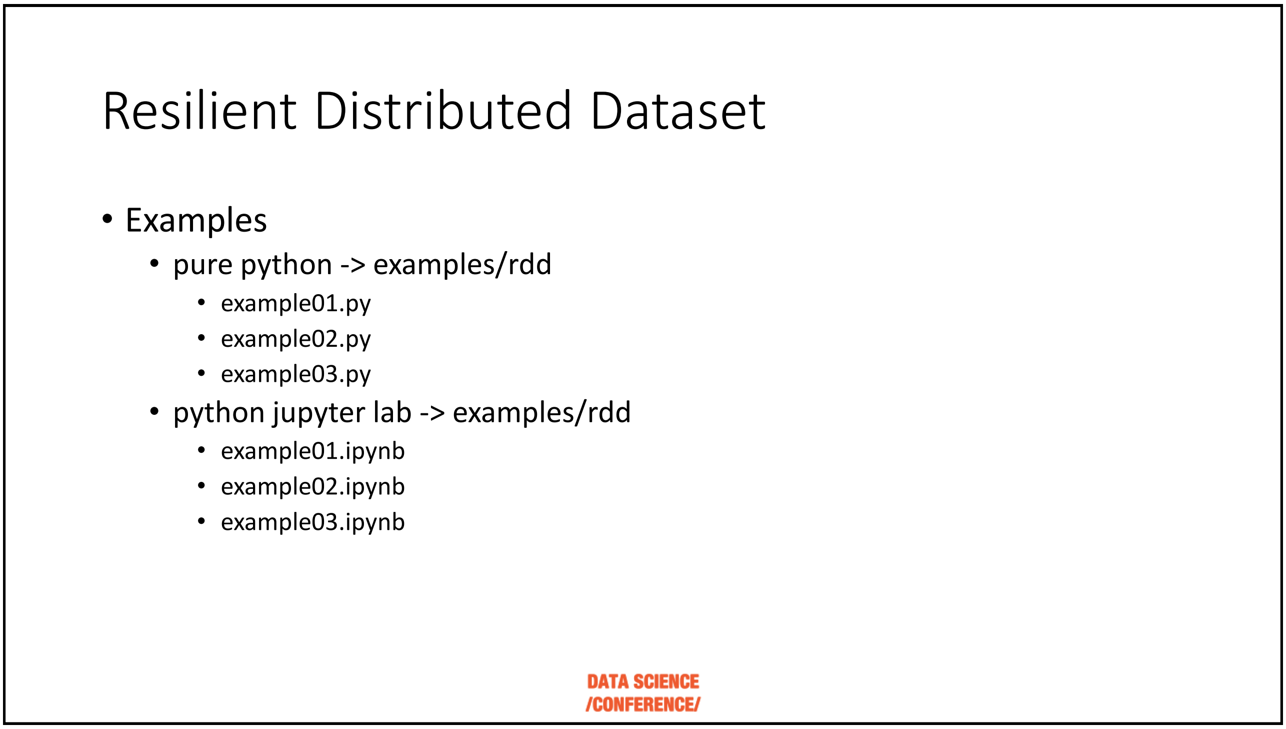


21

Resilient Distributed Dataset



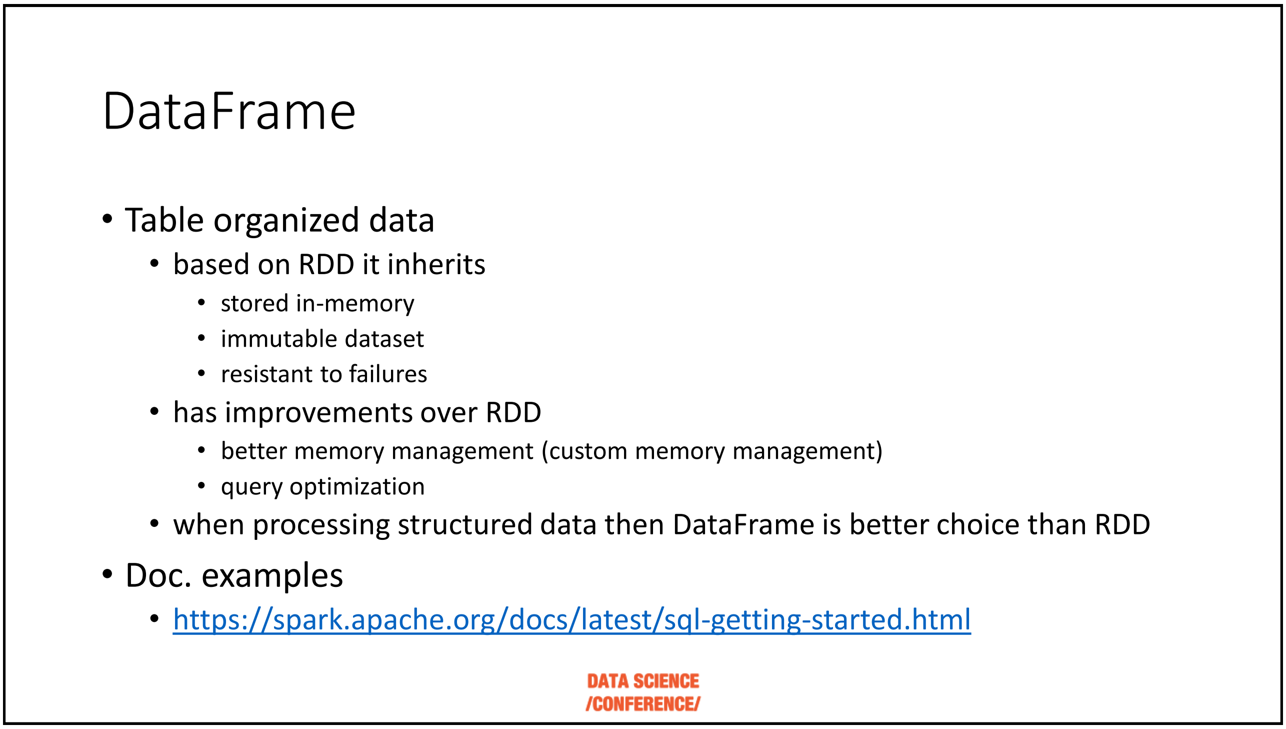
22



Resilient Distributed Dataset

- Examples
  - pure python -> examples/rdd
    - example01.py
    - example02.py
    - example03.py
  - python jupyter lab -> examples/rdd
    - example01.ipynb
    - example02.ipynb
    - example03.ipynb

23



DataFrame

- Table organized data
  - based on RDD it inherits
    - stored in-memory
    - immutable dataset
    - resistant to failures
  - has improvements over RDD
    - better memory management (custom memory management)
    - query optimization
  - when processing structured data then DataFrame is better choice than RDD
- Doc. examples
  - https://spark.apache.org/docs/latest/sql-getting-started.html

24

DataFrame

- Examples
  - pure python -> examples/df
    - example01.py
    - example02.py
    - example03.py
  - python jupyter lab -> examples/df
    - example01.ipynb
    - example02.ipynb
    - example03.ipynb

DATA SCIENCE
/CONFERENCE/

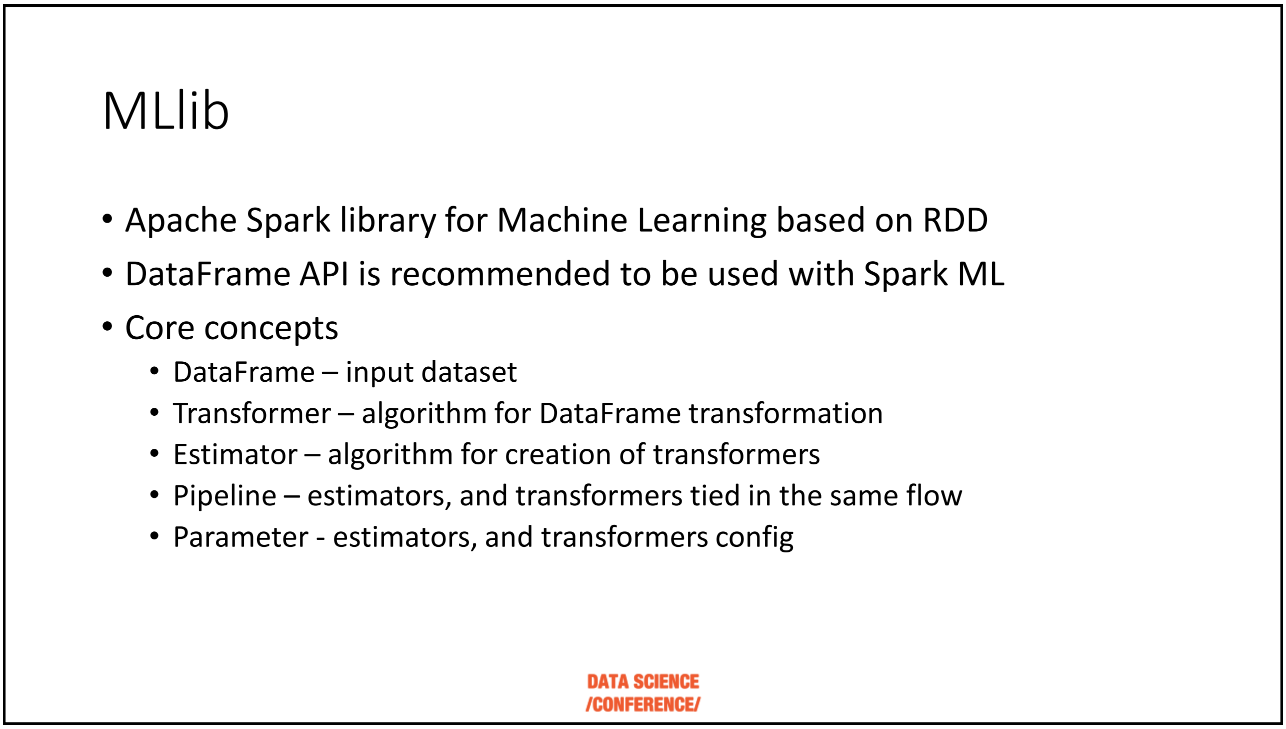
25

Dataset

- Spark dataset
  - represents DataFrame and RDD extension
  - provides an object-oriented interface
    - working with classes and objects
    - collection of JVM objects
- Doc. examples
  - https://spark.apache.org/docs/latest/sql-getting-started.html#creating-datasets

DATA SCIENCE
/CONFERENCE/

26

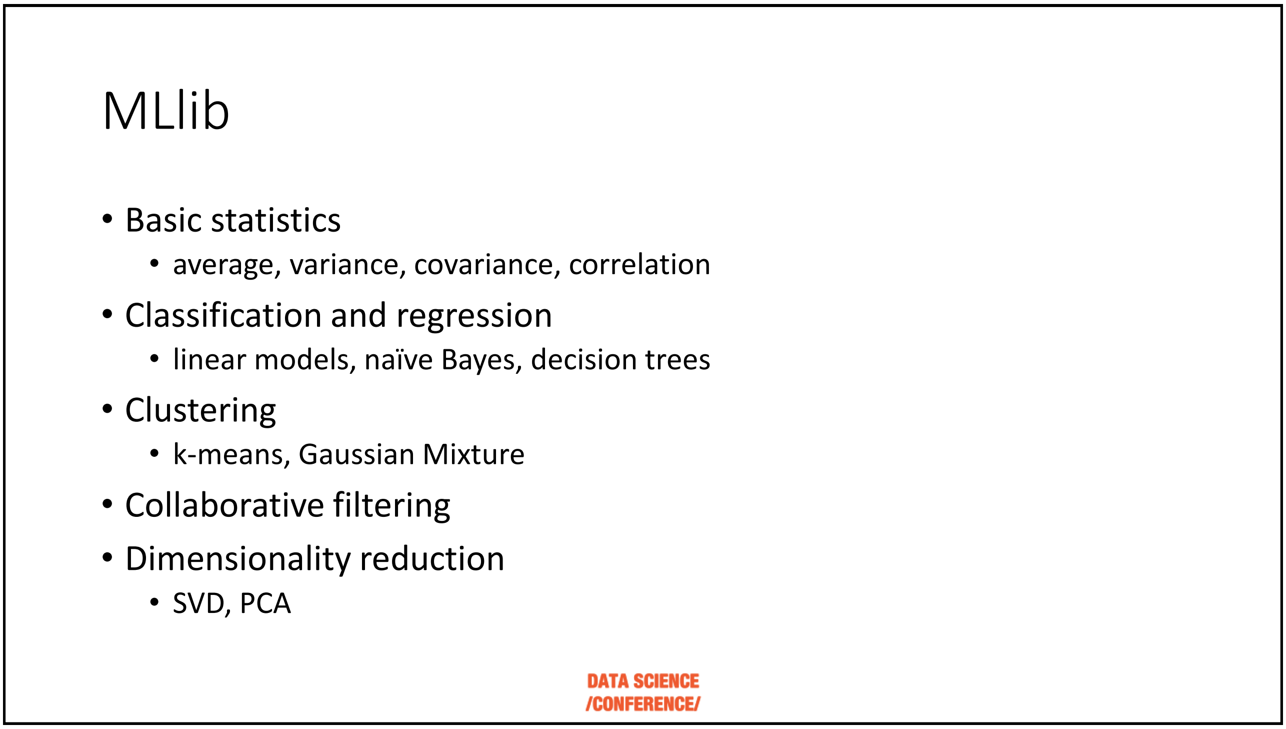


MLlib

- Apache Spark library for Machine Learning based on RDD
- DataFrame API is recommended to be used with Spark ML
- Core concepts
  - DataFrame – input dataset
  - Transformer – algorithm for DataFrame transformation
  - Estimator – algorithm for creation of transformers
  - Pipeline – estimators, and transformers tied in the same flow
  - Parameter - estimators, and transformers config

DATA SCIENCE
/CONFERENCE/

27



MLlib

- Basic statistics
  - average, variance, covariance, correlation
- Classification and regression
  - linear models, naïve Bayes, decision trees
- Clustering
  - k-means, Gaussian Mixture
- Collaborative filtering
- Dimensionality reduction
  - SVD, PCA

DATA SCIENCE
/CONFERENCE/

28

MLlib

- Examples
  - pure python -> examples/ml
    - 01-logistic-regression.py
    - 02-random-forest.py
  - python jupyter lab -> examples/ml
    - 01-logistic-regression.ipynb
    - 02-random-forest.ipynb

DATA SCIENCE
/CONFERENCE/

29

References

- Spark in Action 2nd Edition - Jean-Georges Perring
- Spark: The Definitive Guide: Big Data Processing Made Simple 1st Edition – Bill Chambers
- https://spark.apache.org/docs/latest/api/python/
- https://intellipaat.com/blog/tutorial/spark-tutorial/programming-with-rdds/
- https://www.edureka.co/blog/spark-architecture/
- https://blog.k2datascience.com/batch-processing-apache-spark-a67016008167

DATA SCIENCE
/CONFERENCE/

30

Thank you for your attention!

DATA SCIENCE
/CONFERENCE/