ADALTAS

Data Analysis with

4. App Packaging and Submitting

Spark functionalities recap

ADALTAS

- Batch processing with :
  - Resilient Distributed Datasets
  - DataFrames
- Structured Streaming processing

Spark functionalities recap

ADALTAS

But also :

- Spark MLlib:
  - Distributed model training
  - Apply models to data
- GraphX: distributed graph processing

Spark application components

ADALTAS

- Java and Scala: “Uber” **jar** file
  - Spark code
  - Dependencies (libraries)
- Python: .py, .zip or .egg files

Spark application configuration

ADALTAS

- Deploy mode: client or cluster
- Driver and executors:
  - Memory
  - Cores
  - Number of executors

Spark application monitoring

ADALTAS

- One web UI per application
- Track:
  - Jobs
  - Stages
  - Executors

Spark application performance tuning

ADALTAS

- Limit number of shuffles
- 3 partitions / CPU
- Start big, then reduce memory
- Lot of testing, monitoring, benchmarking (optimize properties one by one)