Getting to grips with Databricks

Gianluca Campanella

Contents

Databricks and Spark

What is Databricks?

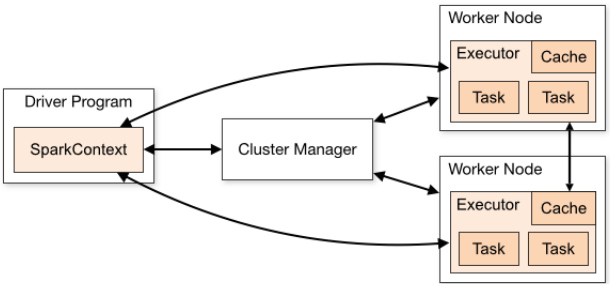
- Cluster computing system
- Java, Python, R, Scala and SQL
- Apache License 2.0

- Spark as-a-service
- Notebook-oriented
- Available on Azure and AWS

What is Spark?

Spark is…

- Scalable
- Fast (*ish*)
- Simple (*ish*)



From the Spark documentation

Spark use cases

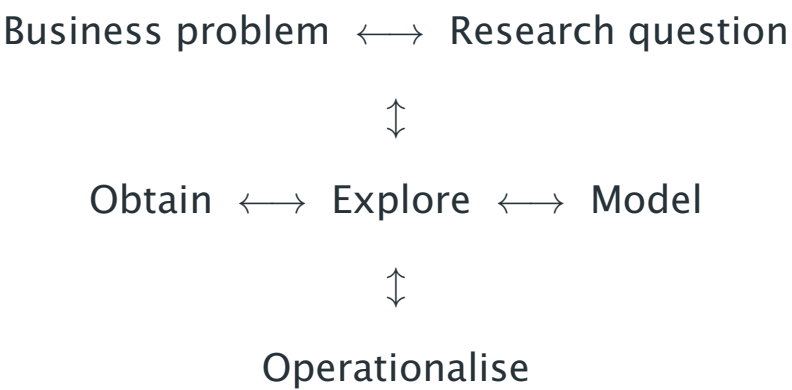
Library	Use case
Spark SQL	Read and process *huge* data sets (ETL)
Spark Streaming	Process streaming data
MLlib	Train and (batch) score ML models
GraphX	Analyse large graphs

Spark DataFrames

- Functionally similar to pandas and R DataFrames
- Backed by Resilient Distributed Datasets (RDDs)
- Typed ($\rightarrow$ querying can be optimised)

Building E2E solutions

Data Science workflow



Running example

movielens

'Latest' dataset (9/2018)

- 5.8×10^4 movies
- 2.8×10^5 users
- 2.7×10^7 ratings

Steps

1. Download data
2. ETL
3. EDA
4. Modelling