# Spark extra slides

# Spark SQL, DataFrames and Datasets

- Build on top of the **RDD API**
- Provide more information about the structure of the data and the computation being performed
    - Leads to **more optimized processing** than RDDs

# Dataset API

- A dataset represent a distributed collection of data
- Has a **strongly-typed** API (Scala/Java)
- Has an **untyped** API also called “DataFrame” (Scala/Java/python/R)

# DataFrame API

- Is a Dataset organized into named columns (a Dataset of Row objects) (like a table in a Database)
- Available in Scala, Java, Python & R
- Lazy evaluation (like in RDDs)

# DataFrame API

- **Getting Started (Spark SQL DataFrame API)**
    - https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_df.html
- **Getting Started (Pandas API on Spark)**
    - https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_ps.html

- Note: to run the pandas example in the VM you will need to install it first: “pip install pandas pyarrow”

# User Guides and Useful links

- Data Sources
  - https://spark.apache.org/docs/latest/sql-data-sources.html
- Pandas on Spark User Guide
  - https://spark.apache.org/docs/3.2.0/api/python/user_guide/pandas_on_spark/index.html
- Spark MLlib User Guide
  - https://spark.apache.org/docs/latest/ml-guide.html
- Spark Python API reference
  - https://spark.apache.org/docs/latest/api/python/reference/index.html
- Spark RDD User Guide
  - https://spark.apache.org/docs/latest/rdd-programming-guide.html
- Spark Structured Streaming User Guide
  - https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html