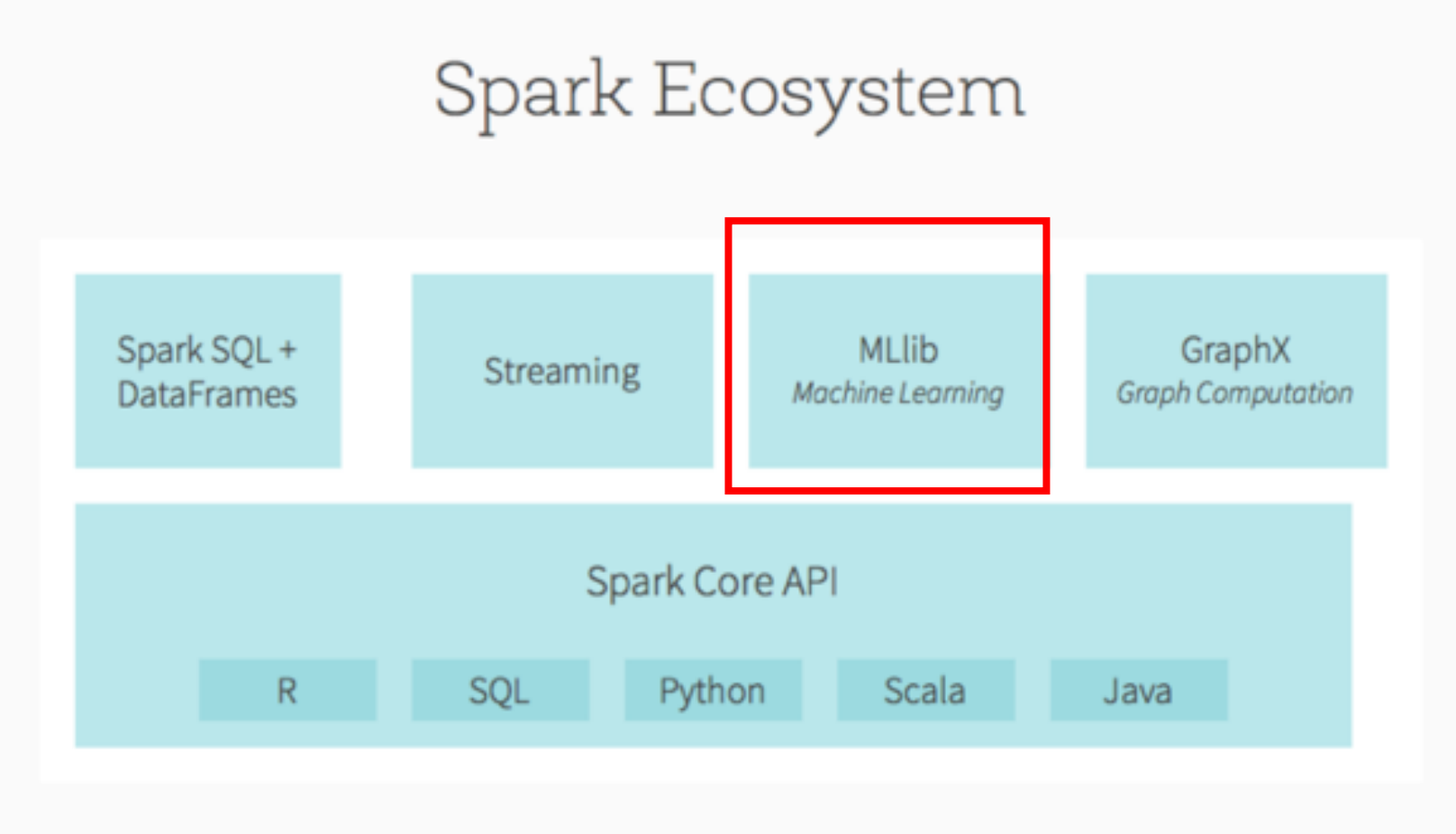
10.05 Classification & Regression in Spark

Recap: Spark Ecosystem



Spark’s design for machine learning

- From the inception of the Apache Spark project, MLlib was considered foundational for Spark’s success.
- The key benefit of MLlib is that it allows data scientists to focus on their data problems and models instead of solving the complexities surrounding distributed data (such as infrastructure, configurations, and so on).
- The data engineers can focus on distributed systems engineering using Spark’s easy-to-use APIs, while the data scientists can leverage the scale and speed of Spark core.
- Just as important, Spark MLlib is a general-purpose library, providing algorithms for most use cases while at the same time allowing the community to build upon and extend it for specialised use cases.

Spark’s design for machine learning

- The advantages of MLlib’s design include:
  - **Simplicity:** Simple APIs familiar to data scientists coming from tools like Python. Novices are able to run algorithms out of the box while experts can easily tune the system by adjusting important knobs and switches (parameters)
  - **Scalability:** Ability to run the same ML code on your laptop and on a big cluster seamlessly without breaking down. This lets businesses use the same workflows as their user base and data sets grow

Spark’s design for machine learning

- The advantages of MLlib’s design include:
  - **Streamlined end-to-end:** Developing machine learning models is a multistep journey from data ingest through trial and error to production. Building MLlib on top of Spark makes it possible to tackle these distinct needs with a single tool instead of many disjointed ones. The advantages are lower learning curves, less complex development and production environments, and ultimately shorter times to deliver high-performing models.
  - **Compatibility:** Data scientists often have workflows built up in common data science tools, such as R, Python pandas, and scikit-learn. Spark DataFrames and MLlib provide tooling that makes it easier to integrate these existing workflows with Spark.
    - For example, Databricks is writing Spark packages in Python to allow users to distribute parts of scikit-learn workflows

Spark’s design for machine learning

- The advantages of MLlib’s design include:
    - **Role Focus:** Spark MLlib allows novice data practitioners to easily work with their algorithms out of the box while experts can tune as desired. Data engineers can focus on distributed systems, and data scientists can focus on their machine learning algorithms and models. Spark enhances machine learning because data scientists can focus on the data problems they really care about while transparently leveraging the speed, ease, and integration of Spark’s unified platform.

Spark’s design for machine learning

- At the same time, Spark allows data scientists to solve multiple data problems in addition to their machine learning problems.
- The Spark ecosystem can also solve graph computations (via GraphX), streaming (real-time calculations), and real-time interactive query processing with Spark SQL and DataFrames.
- The ability to employ the same framework to solve many different problems and use cases allows data professionals to focus on solving their data problems instead of learning and maintaining a different tool for each scenario.

Spark MLlib use cases

- Many compelling business scenarios and technical solutions are being solved today with Spark MLlib, including:
  - Huawei on Frequent Pattern Mining
  - OpenTable's Dining Recommendations
  - Verizon's Spark MLlib's ALS-based Matrix Factorization (Collab Filtering)

Spark MLlib use cases

- **NBC Universal** stores hundreds of terabytes of media for international cable TV. To save on costs, it takes the media offline when it is unlikely to be used soon. The company uses Spark MLlib Support Vector Machines to predict which files will not be used.

- The **Toyota** Customer 360 Insights Platform and Social Media Intelligence Center is powered by Spark MLlib. Toyota uses MLlib to categorise and prioritize social media interactions in real-time.

- **Radius Intelligence** uses Spark MLlib to process billions of data points from customers and external data sources, including 25 million businesses and hundreds of millions of business listings from various sources.

- **ING** uses Spark in its data analytics pipeline for anomaly detection. The company's machine learning pipeline uses Spark decision tree ensembles and k-means clustering.