Spark MLlib

Dr. Villanes

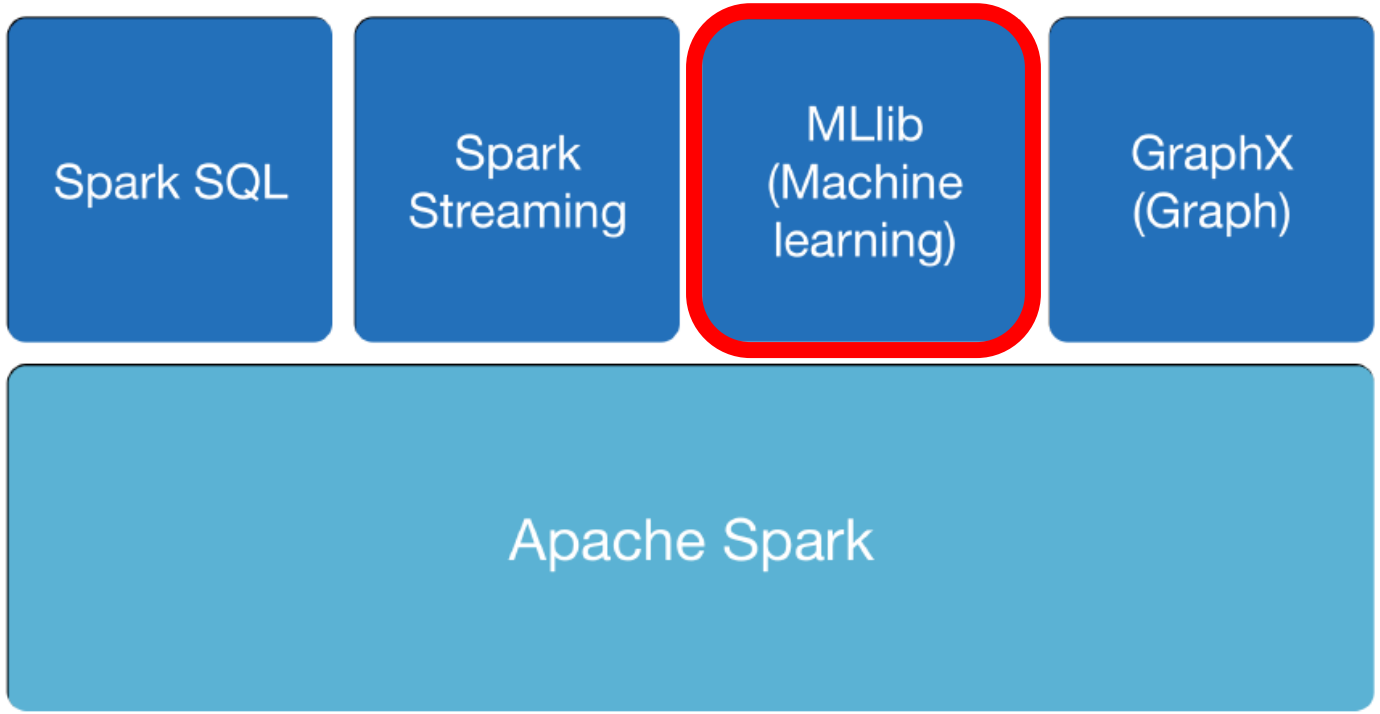
- Login to AWS Academy…

What are we going to do today?

- Spark Mllib
- Why Mllib?
- Documentation page
- Example: Predicting Flights delays
- Optional Homework (due today at midnight)

Apache Spark Components

- Combine SQL, streaming, and complex analytics.



Why would we want to use Spark for Machine Learning?

Spark comes into play when we need to **scale**.

Spark solves the **big data problem** by using a **cluster of machines/nodes/instances, rather than one single machine**.

Spark MLlib

Apache Spark MLlib

- MLlib is developed as part of the Apache Spark project. It thus gets tested and updated with each Spark release.
- MLlib is Spark's scalable machine learning library consisting of common learning algorithms and utilities, including **classification, regression, clustering, collaborative filtering, dimensionality reduction**, and more.
- List of algorithms implemented in Mllib: http://spark.apache.org/docs/latest/ml-guide.html

Does Spark MLlib supports Dataframes?

- Starting in Spark 2.0, the RDD-based APIs in the spark.mllib package have entered maintenance mode. The primary Machine Learning API for Spark is now the DataFrame-based API in the spark.ml package.
- Why is MLlib switching to the DataFrame-based API?
  - DataFrames provide a more user-friendly API than RDDs. The many benefits of DataFrames include Spark Datasources, SQL/DataFrame queries.
  - The DataFrame-based API for MLlib provides a uniform API across ML algorithms and across multiple languages.
  - DataFrames facilitate practical ML Pipelines, particularly feature transformations. See the Pipelines guide for details.
- *What is "Spark ML"?:* not an official name but occasionally used to refer to the MLlib DataFrame-based API.

Why Mllib?

- Spark is a general-purpose big data platform.
  - Reads from HDFS, S3, HBase, and any Hadoop data source.
  - Runs in standalone mode, on Hadoop, EC2, etc.
- MLlib is a standard component of Spark providing machine learning primitives on top of Spark.
- Provides scalability, and integration with Spark and its other components

Spark Documentation: http://spark.apache.org/docs/latest/ml-guide.html

Classification and regression

This page covers algorithms for Classification and Regression. It also includes sections discussing specific classes of algorithms, such as linear methods, trees, and ensembles.

Table of Contents

Spark Documentation: http://spark.apache.org/docs/latest/ml-guide.html

Binomial logistic regression

For more background and more details about the implementation of binomial logistic regression, refer to the documentation of logistic regression in `spark.mllib`.

Examples

The following example shows how to train binomial and multinomial logistic regression models for binary classification with elastic net regularization. `elasticNetParam` corresponds to α and `regParam` corresponds to λ.

Python | Scala | Java | R

More details on parameters can be found in the Python API documentation.

```python
from pyspark.ml.classification import LogisticRegression

# Load training data
training = spark.read.format("libsvm").load("data/mllib/sample_libsvm_data.txt")

lr = LogisticRegression(maxIter=10, regParam=0.3, elasticNetParam=0.8)

# Fit the model
lrModel = lr.fit(training)

# Print the coefficients and intercept for logistic regression
print("Coefficients: " + str(lrModel.coefficients))
print("Intercept: " + str(lrModel.intercept))

# We can also use the multinomial family for binary classification
mlr = LogisticRegression(maxIter=10, regParam=0.3, elasticNetParam=0.8, family="multinomial")

# Fit the model
mlrModel = mlr.fit(training)

# Print the coefficients and intercepts for logistic regression with multinomial family
print("Multinomial coefficients: " + str(mlrModel.coefficientMatrix))
print("Multinomial intercepts: " + str(mlrModel.interceptVector))
```

Other Resources

- Databricks notebook: Binary Classification Example
  - Logistic Regression, Decision Trees, Random Forest
  - https://docs.databricks.com/applications/machine-learning/train-model/mllib/index.html#binary-classification-example
  - Decision Trees Examples:
  - https://docs.microsoft.com/en-us/azure/databricks/applications/machine-learning/train-model/mllib/#decision-trees-examples

“Can I use the remaining of my AWS Academy credits?”

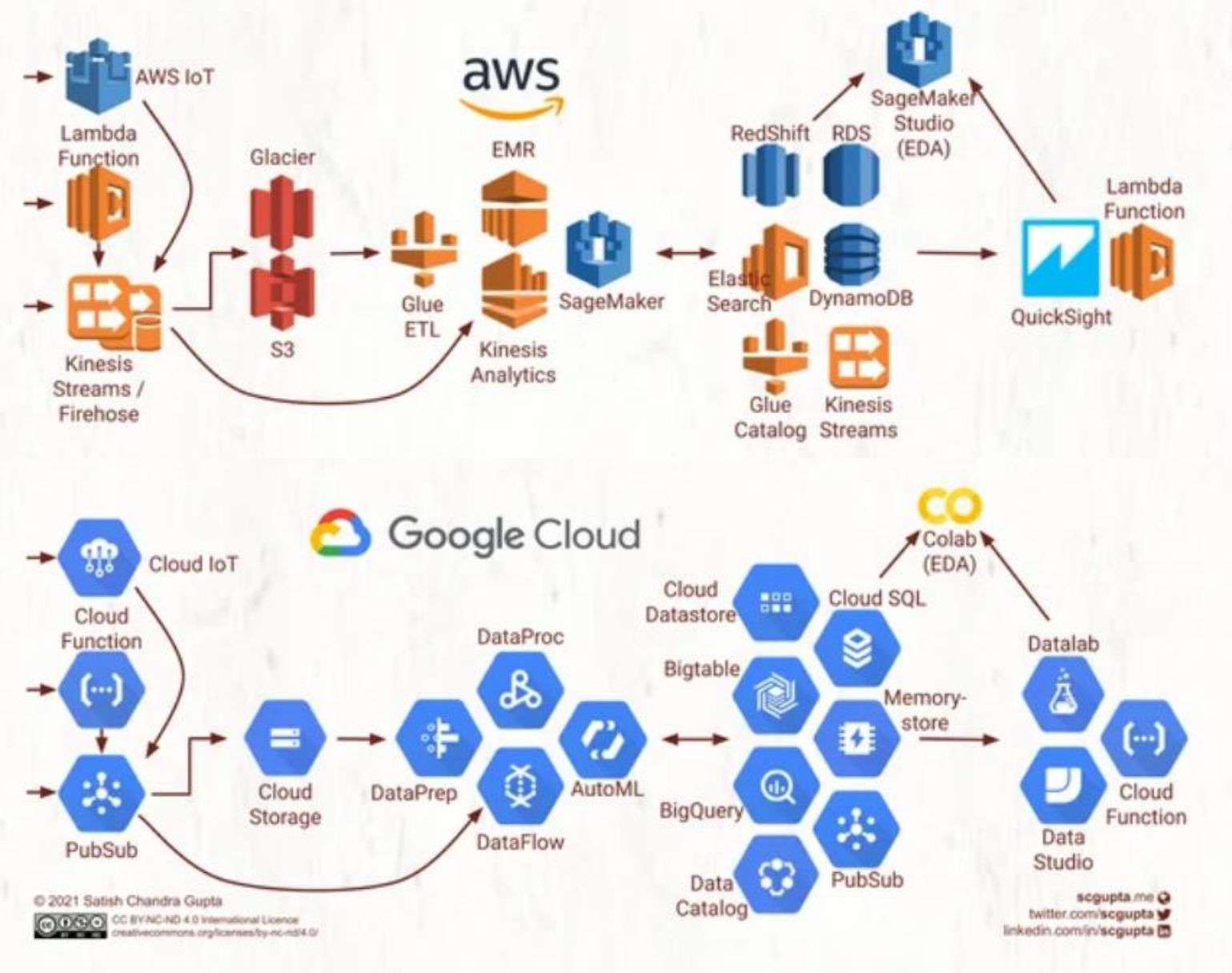
YES!

With your account, you have many services available…

Cloud Comparison Cheat Sheet

blog.bytebytego.com

aws	Azure	Google Cloud	ORACLE CLOUD	Alibaba Cloud
Elastic Compute Cloud (EC2)	Virtual Machine	Compute Engine	Virtual Machine Instance	Elastic Compute Service
Elastic Kubernetes Service (EKS)	Azure Kubernetes Service (AKS)	Google Kubernetes Engine (GKE)	Oracle Container Engine	Alibaba Cloud Kubernetes Service
Lambda	Azure Functions	Cloud Functions	OCI Functions	Function Compute
Simple Storage Service (S3)	Blob Storage	Cloud Storage	Object Storage	Object Storage Service
Elastic Block Store	Managed Disk	Persistent Disk	Persistent Volume	Block Storage
Elastic File System	File Storage	File Store	File Storage	Network Attached Storage
Virtual Private Cloud	Virtual Network	Virtual Private Cloud	Virtual Cloud Network	Virtual Private Cloud
Route 53	DNS	Cloud DNS	DNS	DNS
Elastic Load Balancing	Load Balancer	Cloud Load Balancing	Load Balancer	Server Load Balancer
Web Application Firewall	Web Application Firewall	Cloud Armor	Web Application Firewall	Web Application Firewall
RDS	SQL Database	Cloud SQL	ATP	ApsaraDB RDS
DynamoDB	Cosmos DB	Firebase Realtime Database	NoSQL Database	Table Store
Redshift	Synapse Analytics	BigQuery	Autonomous Data Warehouse	AnalyticDB
Elastic MapReduce	HDInsight	Dataproc	Big Data	Elastic MapReduce
Kinesis	Streaming Analytics	Dataflow	Streaming	DataHub
SageMaker	Machine Learning	Vertex AI	Data Science	Platform for AI
Glue	Data Factory	Data Fusion	Data Integration	DataWorks
EventBridge	Event Grid	Eventarc	Events	Eventbridge
Simple Queuing Service	Storage Queues	Pub/Sub	Streaming	Message Queue
Simple Notification Service	Service Bus	Firebase Cloud Messaging	Notifications	Message Service
CloudWatch	Monitor	Cloud Monitoring	Monitoring	CloudMonitor
CloudFormation	Resource Manager	Deployment Manager	Resource Manager	Resource Orchestration
IAM	Active Directory	Cloud Identity	IAM	Resource Access Management
KMS	Key Vault	Cloud KMS	Vault	KMS


aws
AWS IoT
Lambda Function
Kinesis Streams / Firehose
Glacier
S3
Glue ETL
EMR
Kinesis Analytics
SageMaker
RedShift
RDS
SageMaker Studio (EDA)
Elastic Search
DynamoDB
Glue Catalog
Kinesis Streams
Lambda Function
QuickSight
Google Cloud
Cloud IoT
Cloud Function
PubSub
Cloud Storage
DataPrep
DataProc
DataFlow
AutoML
Cloud Datastore
Cloud SQL
Colab (EDA)
Bigtable
Memory-store
BigQuery
Data Catalog
PubSub
Datalab
Data Studio
Cloud Function
© 2021 Satish Chandra Gupta
CC BY-NC-ND 4.0 International Licence
creativecommons.org/licenses/by-nc-nd/4.0/
scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta