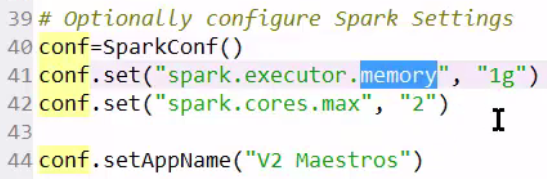
1. What are components of SPARK?
2. What are functions of Spark core?
3. What is Spark Scheduler?
4. Why Spark is faster than Hadoop MapR?
5. What is Driver Program in Spark? Give Example?
6. What is Pair RDD?
7. How to create broadcast variable in Spark?
8. Why Lazy evaluation by Spark?
9. **How to start pyspark interactive shell?**

Type pyspark at the CLI

1. How to Start Spark Scala interactive shell?
2. Which version of Spark have you worked on?
3. Which version of python is supported by Spark latest?
4. How to set memory and cores for my pyspark script?



1. **How to create local sparkContext?**

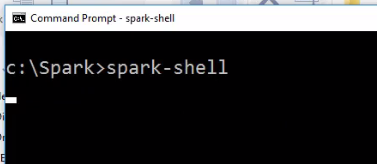


1. Cache vs Persists in Saprk?
2. Map vs Flat Map function in Spark?
3. Can Spark be run without Hadoop? Yes
4. Spark performance optimization?
5. DataSets Vs DataFrames in Spark?
6. How to write SQL in pySpark?
7. What is SparkSQL?
8. What is Serialization and Deserialization? When do we use it in Spark?
9. How to run code on SPARK 1 or 2

Export SPARK\_MAJOR\_VERSON=2

1. What is difference between hiveCtx.sql vs pySpark.sql
2. What transformations and Actions have you done on data frame?
3. What do we mean by schema in database?
4. **How to build a sparkSession in Spark 2.0 using pyspark?**
5. **Multiline comment in python?**
6. **How to create pipeline in Spark?**
7. How to Start Spark?

Type spark-shell from spark directory, or from anywhere.



1. What is Scala REPL?
2. How to load scala programming file in Spark?

: load my\_first\_program

1. How to control messages displaying on spark console? How to set logging level in Spark?

Edit your conf/log4j.properties file and change the following line:

log4j.rootCategory=INFO, console

to  
log4j.rootCategory=ERROR, console

Another approach would be to:

Start spark-shell and type in the following:

import org.apache.log4j.Logger

import org.apache.log4j.Level

Logger.getLogger("org").setLevel(Level.OFF)

Logger.getLogger("akka").setLevel(Level.OFF)

1. How to unzip file in Ubuntu?

tar xvf file\_name

1. What is difference between Spark Submit and Load method to run codes in Spark?
2. What is **libsvm** format file?
3. What are some sectors using SparkML? List SparkML projects.

• Fraud detection.

• Web search results.

• Real-time ads on web pages

• Credit scoring and next-best offers.

• Prediction of equipment failures.

• New pricing models.

• Network intrusion detection.

• Recommendation Engines

• Customer Segmentation

• Text Sentiment Analysis

• Predicting Customer Churn

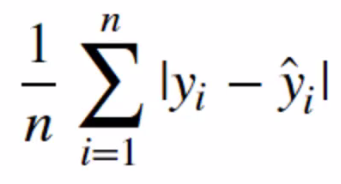
• Pattern and image recognition.

• Email spam filtering.

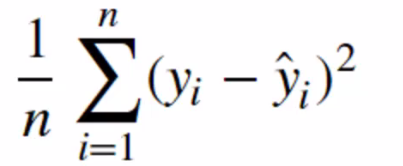
• Financial Modeling

1. What are different types of errors and their significance in Data Science?

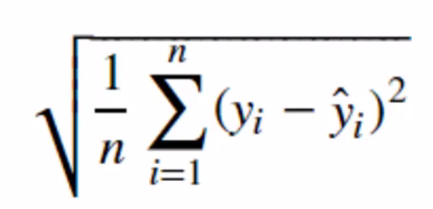
* Mean Absolute Error



* Mean Square Error: More popular because it punishes larger error by squaring



* Root Mean Squared Error: More popular compared to MSE, because units are interpretable.



1. What is difference between list and tuple in Scala?

tup has type (Int, String, Double), so you can get data back with its correct type: tup.\_1 has type Int.

list has type List[Any], so you've lost all type information: list(0)'s type is Any.

Don't use Any (or List[Any], etc.) unless you have to; certainly don't use it when a tuple will do.

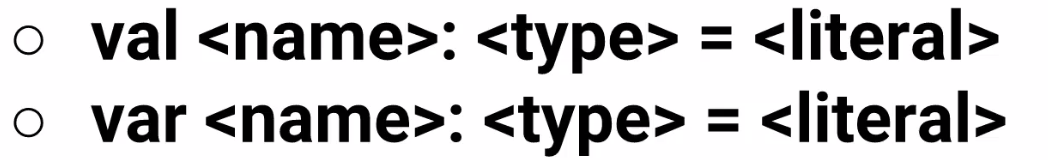
Lists are immutable.

1. What is difference between range and Range?
2. **Difference between Method and Function?**
3. Difference between val and var in Scala?

Values (val) and Variables (var):

* Values (val) are immutable, once they are assigned they cannot be changed. Err: reassignment to Val
* Variables (var) can be reassigned.
* Note, when reassigning you must use the same Data Type! Err: type mismatch

General Format when creating an object:

* 
* Later on we will see that Scala can also infer the data type, so we will skip defining it.
* Let's get started with some examples!

1. What is registerTempTable in Spark? What is createOrReplaceTempView in Spark?
2. Sliding window and batch in Spark? What is **Window length** and **Sliding interval** in Spark?
3. Why we need to import pyspark.SQL to work with DataFrames?

Because DataFrame is based on SparkSQL.

1. What is use of tilde operator in pySpark dataframes?

df.filter( (df["Close"] < 200) & ~(df['Open'] < 200) ).show()

1. How to Start a simple Spark Session?

**from** **pyspark.sql** **import** SparkSession

spark = SparkSession.builder.appName("walmart").getOrCreate()

1. How to Load the Walmart Stock CSV File, have Spark infer the data types?

df = spark.read.csv('walmart\_stock.csv',header=**True**,inferSchema=**True**)

1. What does the Schema look like?

df.printSchema()

1. How to Print out the first 5 columns?

*# Didn't strictly need a for loop, could have just then head()*

**for** row **in** df.head(5):

print(row)

print('**\n**')

1. What is significance of R2 in regression model?
2. What is significance of Serialization and Deserialization in data processing?
3. How to start Spark thrift serve from pyspark?

.config('spark.sql.hive.thriftServer.singleSession', True)