<https://mahout.apache.org/users/sparkbindings/play-with-shell.html>

sh /home/deploy/spark-applications/merge/SparkMergeDataFiles\_yarn.sh M '|' UCM\_1\_\_28-04-2018-16:32 yarn adl://developmentexcaliberdls.azuredatalakestore.net/RAW/STAGING/UCM/S\_CONTACT/\*.csv adl://developmentexcaliberdls.azuredatalakestore.net/RAW/LANDING/UCM/S\_CONTACT/2018/04/26/\*.csv '' adl://developmentexcaliberdls.azuredatalakestore.net/RAW/LANDING/UCM/S\_CONTACT/TEMP

**What is big data and data analytics?**

**Data analytics** is the science of drawing insights from raw information sources. Many of the techniques and processes of **data analytics** have been automated into mechanical processes and algorithms that work over raw **data** for human consumption.

**What is the difference between data science and analytics?**

In that sense, it's similar in nature to business **analytics**, another umbrella term for approaches to analyzing **data** -- with the **difference** that the latter is oriented to business uses, while **data analytics** has a broader focus. ... In addition, it can come from a mix of internal systems and external **data** sources.

you control your Spark Application through a

driver process called the SparkSession

os.system('clear')

sqlDF = spark.sql("**select** accountid **from** vv\_db.hvtb\_nbx\_landing\_crm\_case\_cdc **limit** 1")

sqlDF.show()

polyglot – multiple languauges scala, python, java, R

spark > Tez > Map Reduce

spark 1 – RDDs resilient distributed datasets

spark 2 – API

driver or context

cluster Manager – Hadoop Yarn, Apache Mesos

storage –HDFS, Map R, s3, kudu

executors

Data frames

Spark SQL , MLib, GraphX, Spark Streaming

Read data from HDFS,S3

Insert data into Hive

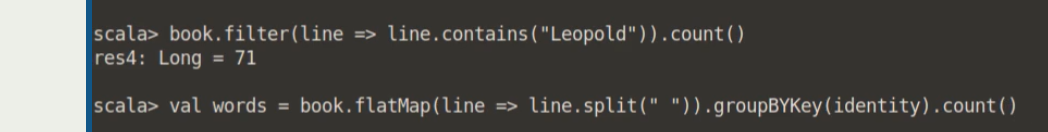
Filter operations

Head

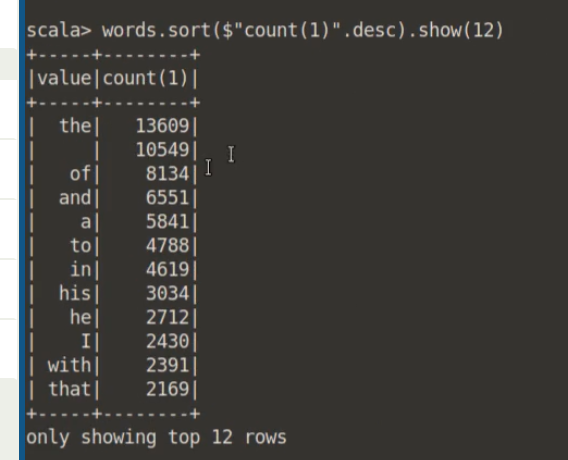
Df.show(100, False)

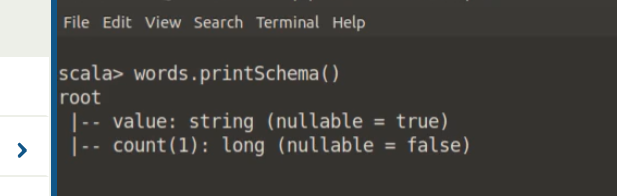
Df.count()

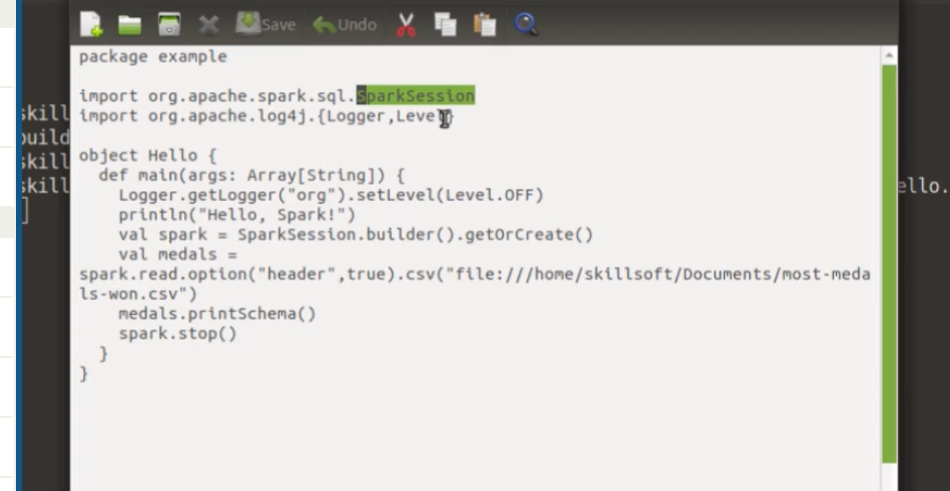
Df.distinct().count()

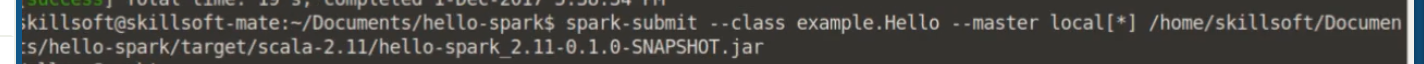


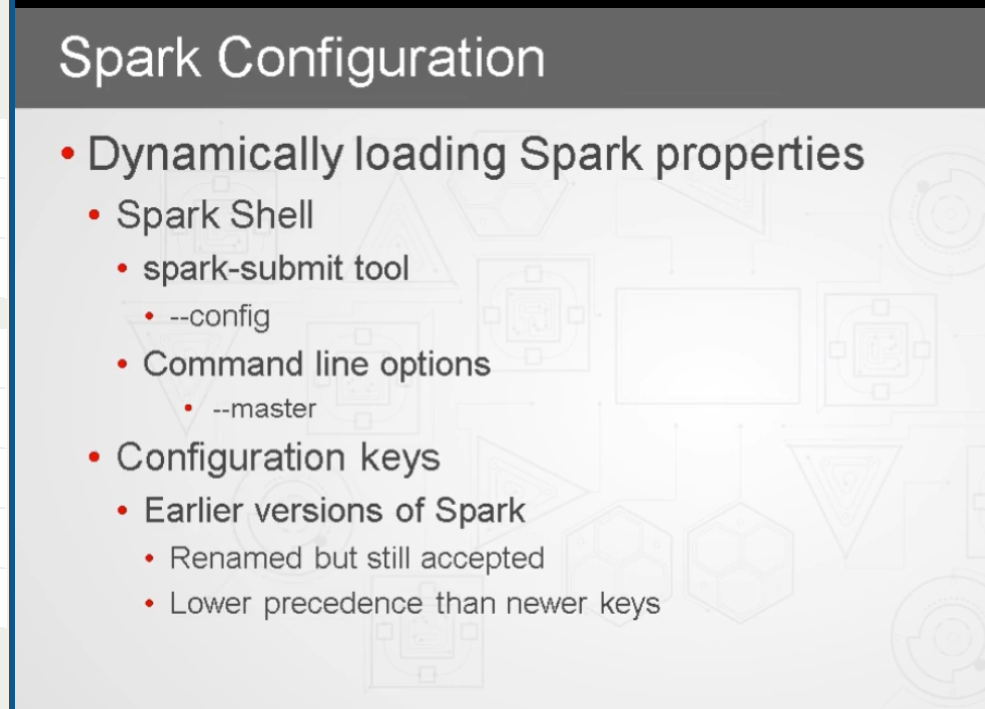
Data set





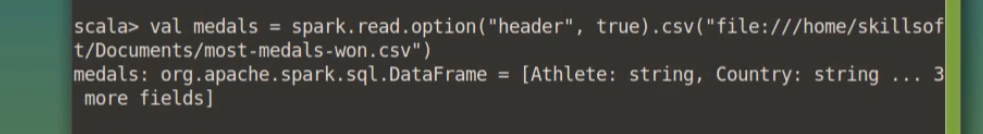


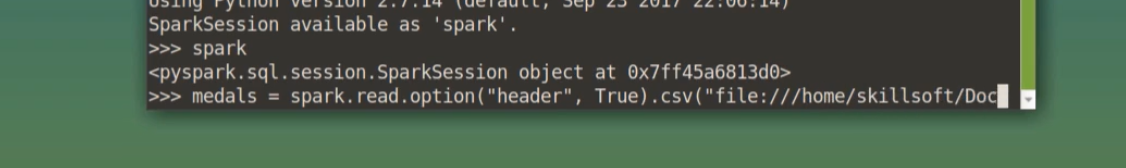


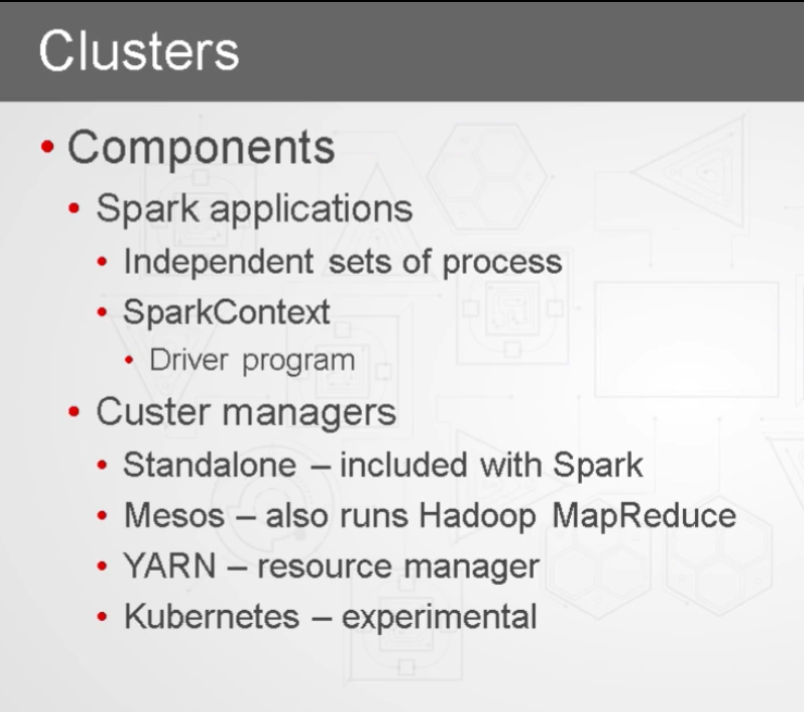




Spark-shell

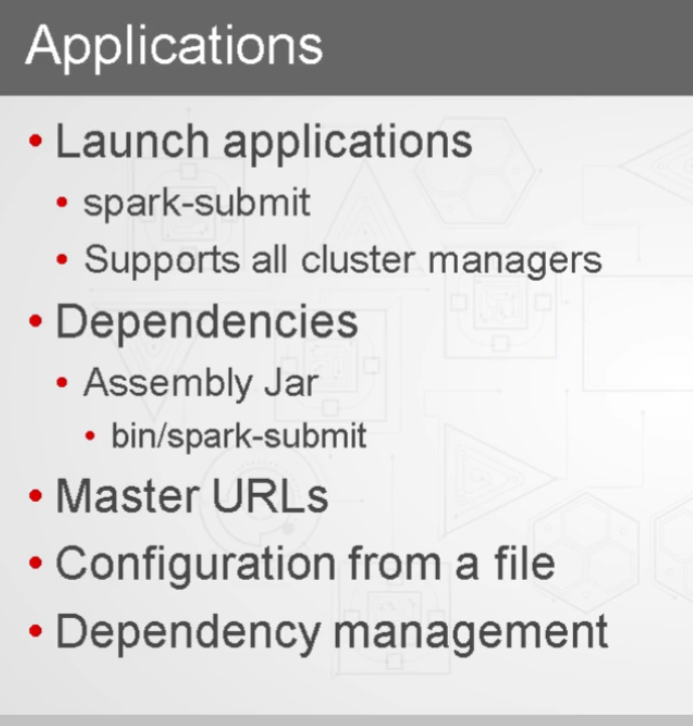


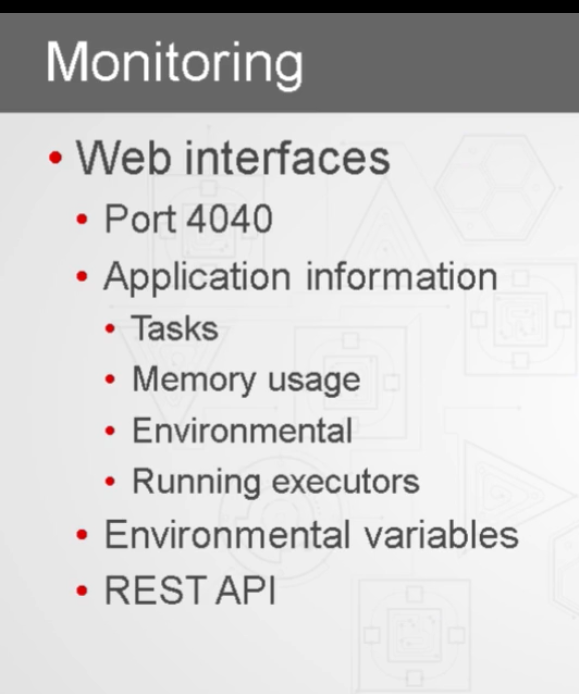




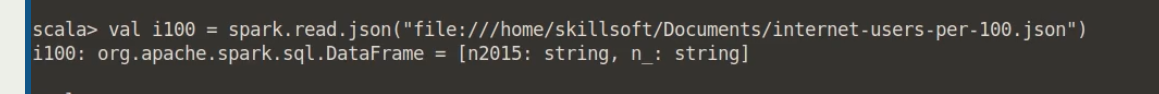
Spark Session on 2.0

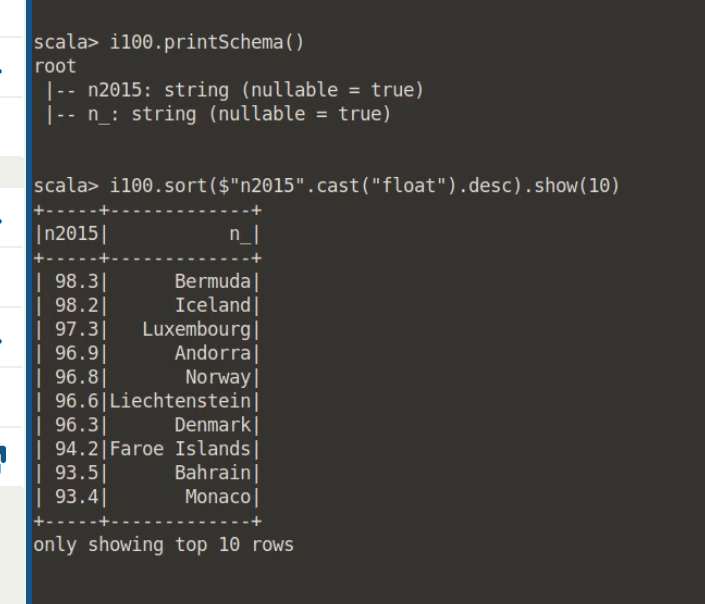
Spark Context



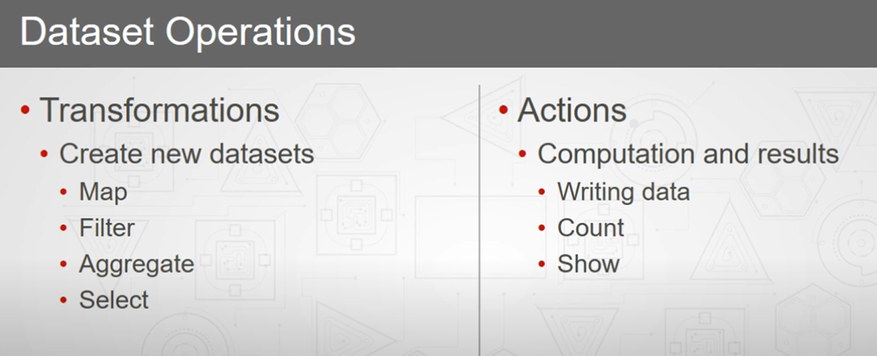


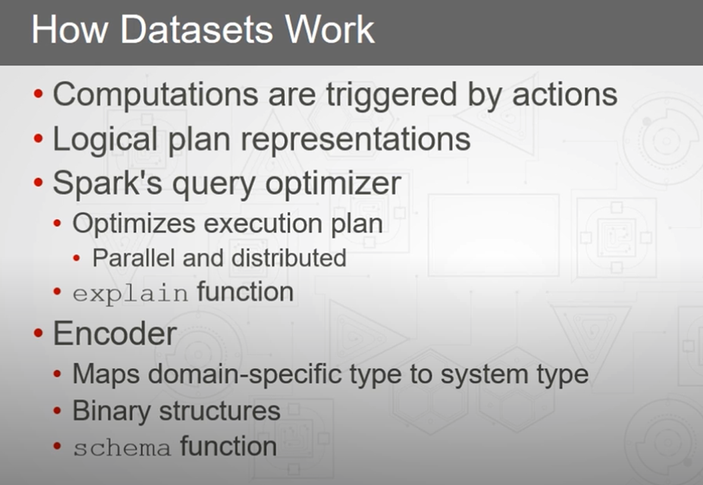


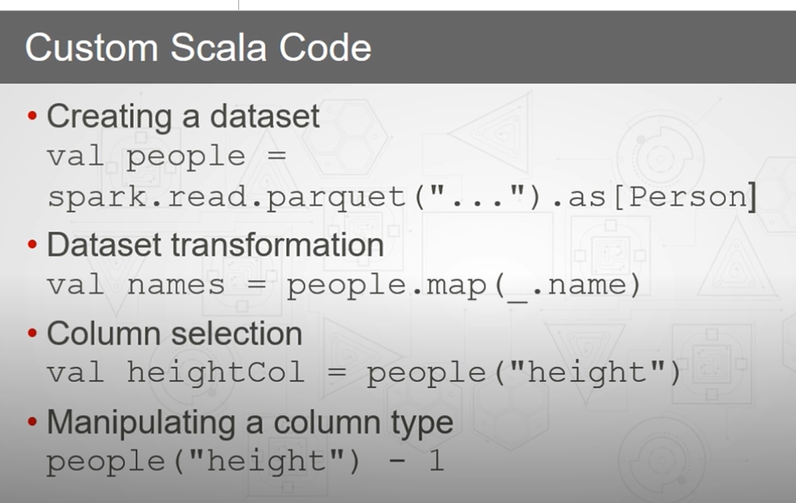












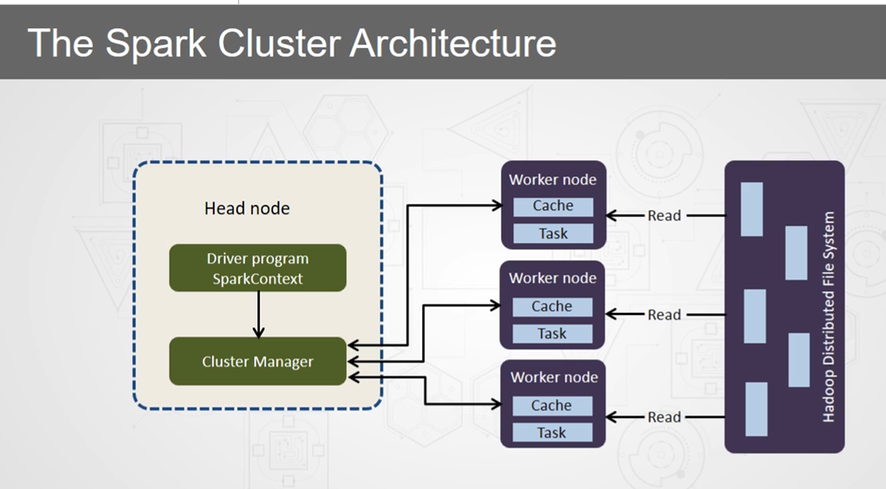
Data sets

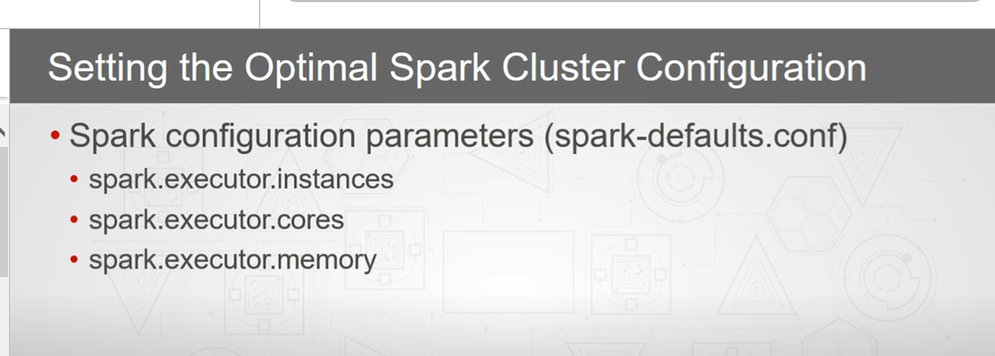
RDD

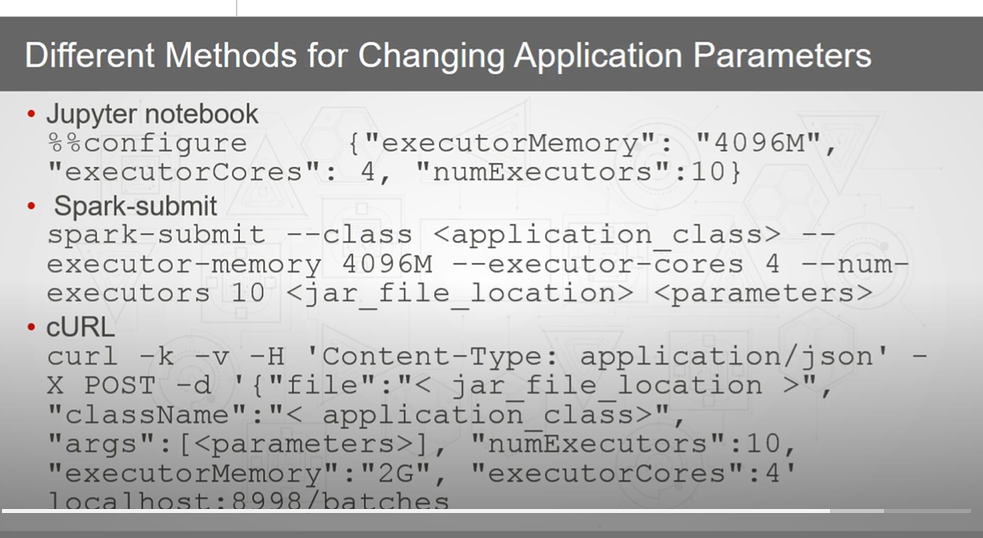
Azure Databricks

Spark sql graph

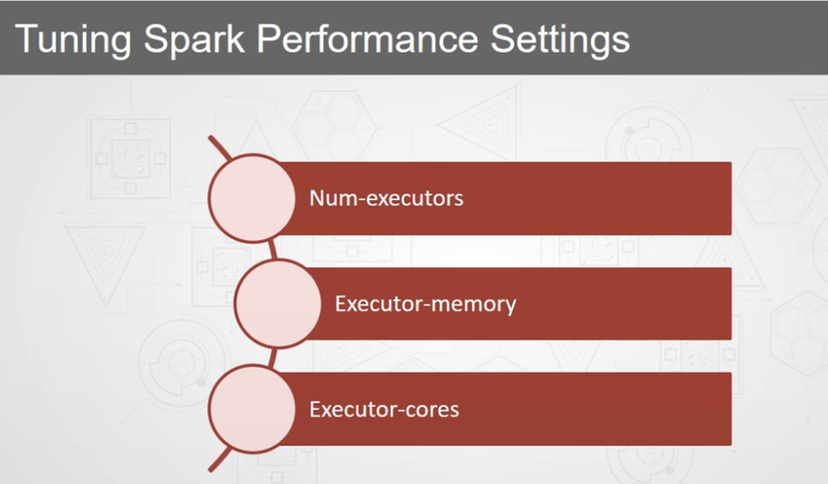
Flat map

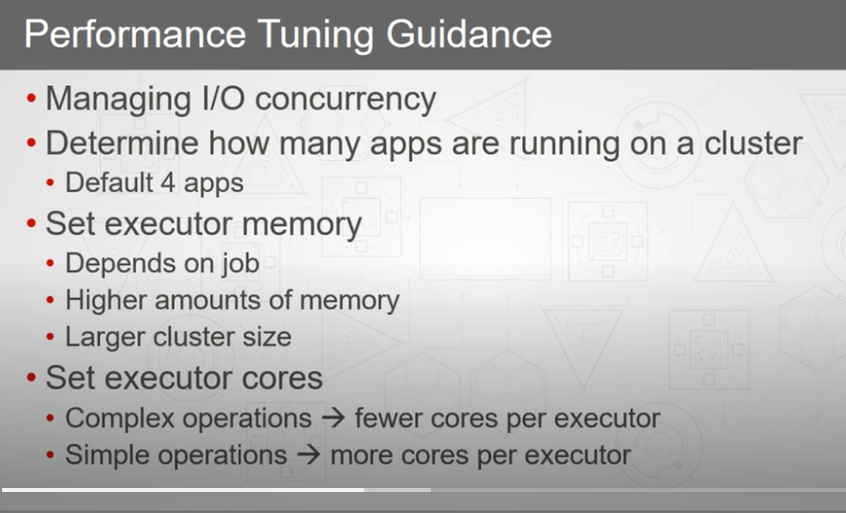


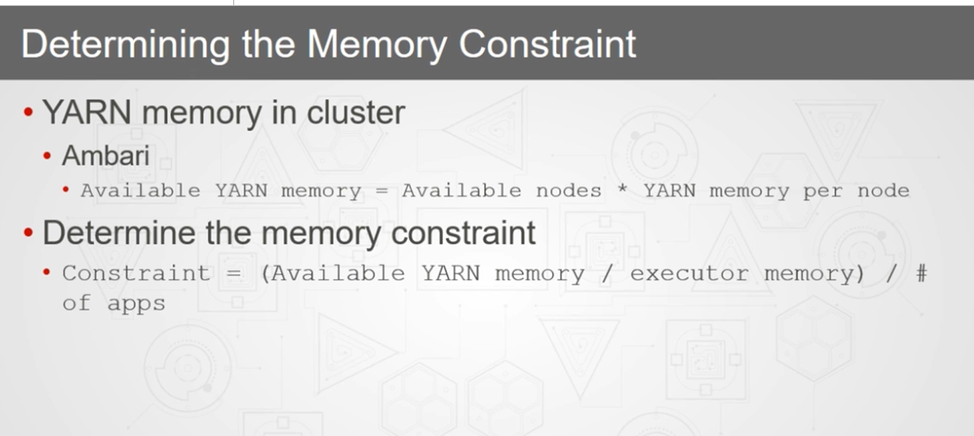








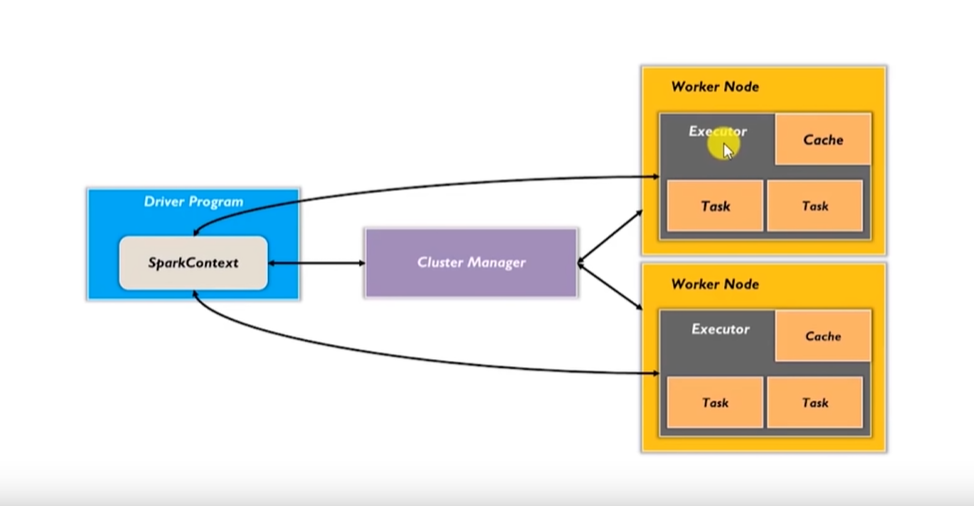


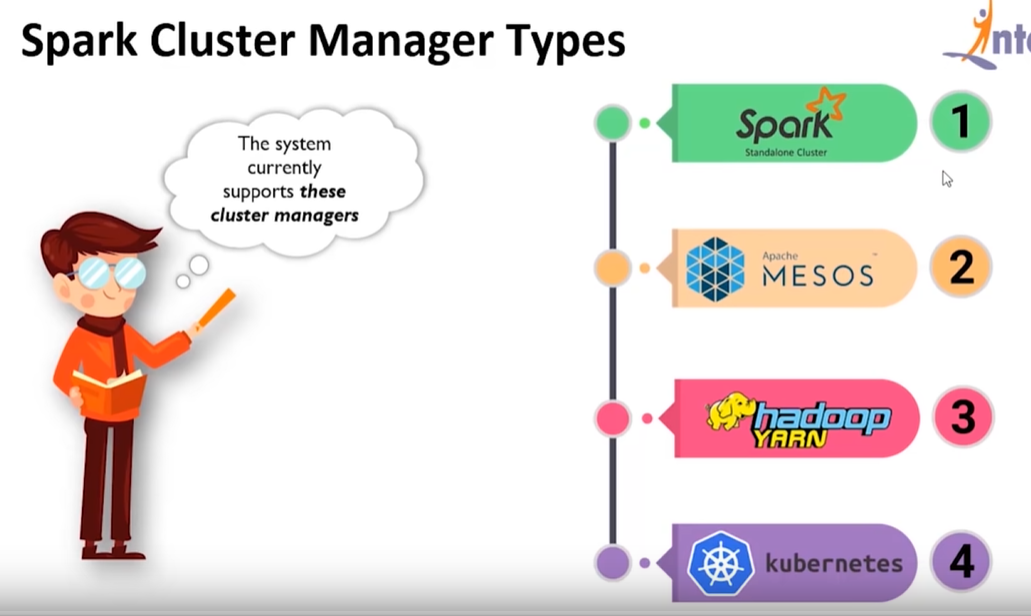




Veracity = quality of data

Mesos, yarn , spark cluster manager





RDDs

words = sc.parallelize (["scala","java","hadoop","spark","akka","spark vs hadoop","pyspark","pyspark and spark"])

>>> words.count()

8

>>>

words\_spark = words.filter(lambda x: 'spark' in x)

words\_spark.collect() – prints the RDD

words.take(1) – takes the first word in RDD

words\_subtract = words.subtract(words\_spark) – subtracts RDD

words\_map = words.map(lambda x:(x,1)) – adding a key value

[('scala', 1), ('java', 1), ('hadoop', 1), ('spark', 1), ('akka', 1), ('spark vs hadoop', 1), ('pyspark', 1), ('pyspark and spark', 1)]

words\_flatmap = words.flatMap(lambda x:(x,1))

['scala', 1, 'java', 1, 'hadoop', 1, 'spark', 1, 'akka', 1, 'spark vs hadoop', 1, 'pyspark', 1, 'pyspark and spark', 1]

numbers = sc.parallelize([1,2,3,4,5,6])

numbers\_filter = numbers.filter(lambda x: x%2==0)

intersection

numbers2 = sc.parallelize([4,5,6])

numbers.intersection(numbers2).collect()

spark-submit count.py

coll = words.collect()

spark-submit collect.py

normal\_raw\_data = raw\_data.filter(**lambda** x: 'normal.' **in** x)

filter

initialize spark session

<https://spark.apache.org/docs/latest/rdd-programming-guide.html>

df = spark.sql("REFRESH TABLE vv\_db.hvtb\_nbx\_core\_sw\_guest\_dim")

df.show(100, False)

df.printSchema()

Actions are Spark RDD operations that give non RDD values

Reduce

a = sc.parallelize(1 to 10)

[Swaroop's answer to Which book is good to learn Spark and Scala for beginners?](https://www.quora.com/Which-book-is-good-to-learn-Spark-and-Scala-for-beginners/answer/DurgaSwaroop-Perla)

[Swaroop's answer to Is it necessary to learn Scala in order to learn Spark?](https://www.quora.com/Is-it-necessary-to-learn-Scala-in-order-to-learn-Spark/answer/DurgaSwaroop-Perla)

[Swaroop's answer to What are good books or websites for learning Apache Spark and Scala?](https://www.quora.com/What-are-good-books-or-websites-for-learning-Apache-Spark-and-Scala/answer/DurgaSwaroop-Perla)

df.filter("res\_id = 55892434").show()

df.select('column1').distinct().show()

df.filter("src\_res\_id = 55892434").show()

dupPkg.createOrReplaceTempView("dupPkg")