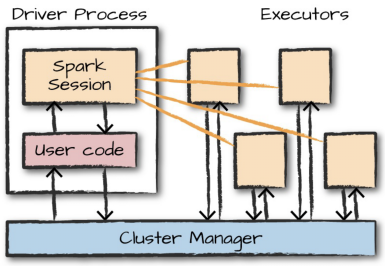
Spark浅析

- Spark的基本架构
  - Spark是什么?
    - 管理和协调跨多台计算机的计算任务的软件框架
    - SPARK的集群管理器YARN或Mesos用来分发计算任务到不同的机器
  - Spark的应用程序
    - 应用程序架构图



- Spark's Language APIs
  - Scala
  - Java
  - Python
  - SQL
  - R
- Spark's APIs
  - Low Level API
    - RDD
  - Structured APIs
    - Dataframe
    - Dataset
- The SparkSession
  - Spark executes user-defined manipulations across the cluster via spark session
- Dataframe
  - Partitioned across sercers in a data center
- Transformations

- Wide Transformations

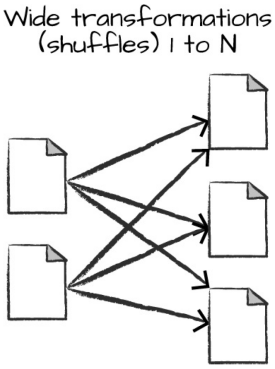


Figure 2-5. A wide dependency

  - spark.sql.shuffle.partitions控制了shuffle后的output

```
spark.conf.set("spark.sql.shuffle.partitions", "5")

flightData2015.sort("count").take(2)
```

  - Narrow Transformations

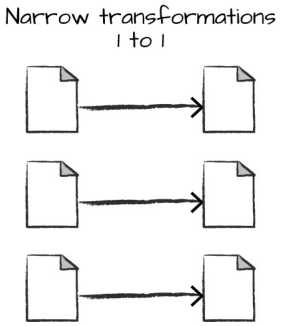


Figure 2-4. A narrow dependency

—个input partition 对应 —个output partitions

- Actions
  - df.count()
- Spark UI
  - Available on port 4040 of the driver node http://localhost:4040
- An End to End Example
  - Dataframe

```
// in Scala
flightData2015
  .groupBy("DEST_COUNTRY_NAME")
  .sum("count")
  .withColumnRenamed("sum(count)", "destination_total")
  .sort(desc("destination_total"))
  .limit(5)
  .explain()
```

  - Physical Plan

```
== Physical Plan ==
TakeOrderedAndProject(limit=5, orderBy=[destination_total#16194L DESC], outpu...
+- *HashAggregate(keys=[DEST_COUNTRY_NAME#7323], functions=[sum(count#7325L)])
   +- Exchange hashpartitioning(DEST_COUNTRY_NAME#7323, 5)
      +- *HashAggregate(keys=[DEST_COUNTRY_NAME#7323], functions=[partial_sum...
         +- InMemoryTableScan [DEST_COUNTRY_NAME#7323, count#7325L]
               +- InMemoryRelation [DEST_COUNTRY_NAME#7323, ORIGIN_COUNTRY_NA...
                  +- *Scan csv [DEST_COUNTRY_NAME#7578,ORIGIN_COUNTRY_NAME...
```

- Spark SQL

```
// in Scala
val maxSql = spark.sql("""
SELECT DEST_COUNTRY_NAME, sum(count) as destination_total
FROM flight_data_2015
GROUP BY DEST_COUNTRY_NAME
ORDER BY sum(count) DESC
LIMIT 5
""")

maxSql.show()
```

- A_Gentle_Introduction_to_Spark-Chapter_2_A_Gentle_Introduction_to_Spark.scala
  SCALA 文件