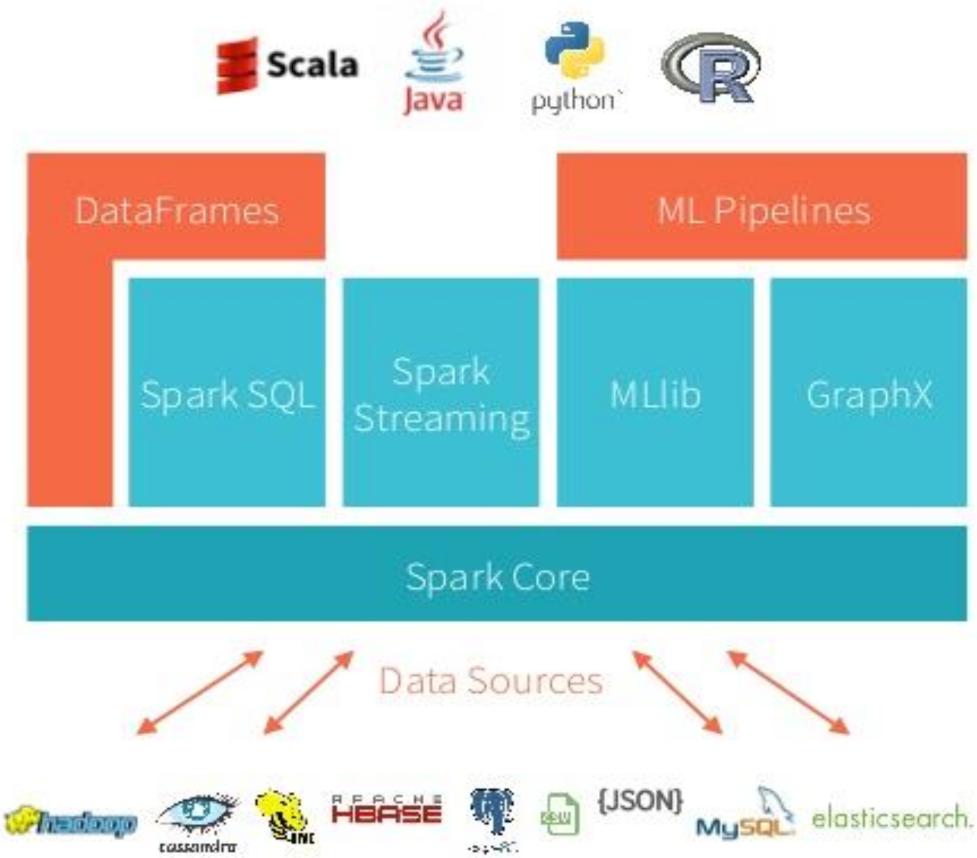
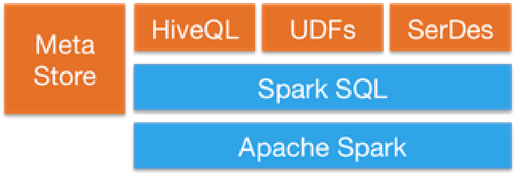
Apache Spark
SQL

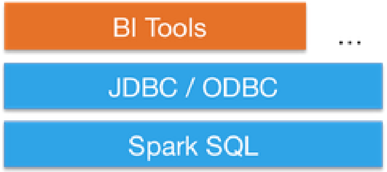
Spark Unified Stack



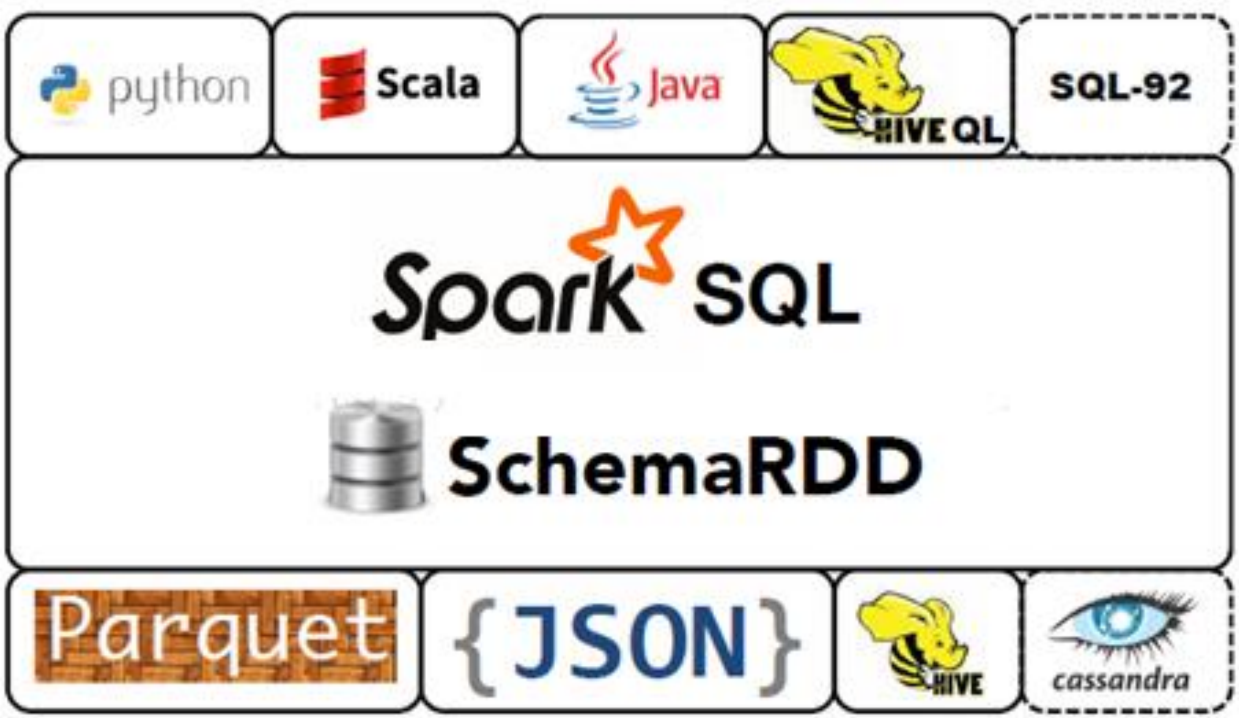
Spark SQL



Spark SQL can use existing Hive metastores, SerDes, and UDFs.



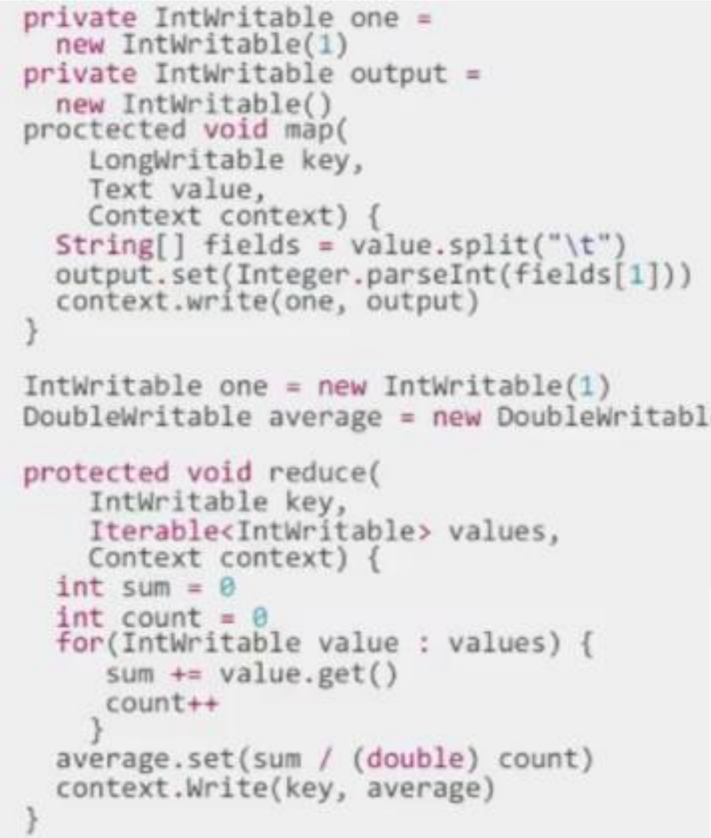
Use your existing BI tools to query big data.



https://spark.apache.org/docs/latest/sql-programming-guide.html

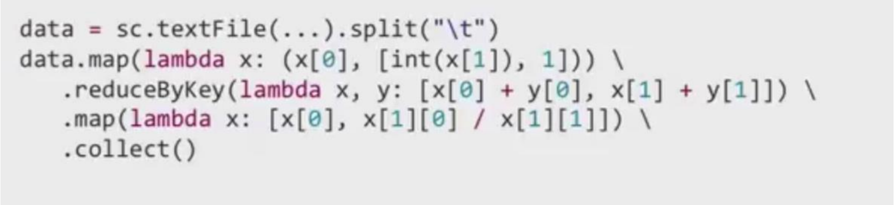
Spark SQL, DataFrames and Datasets

MapReduce代码



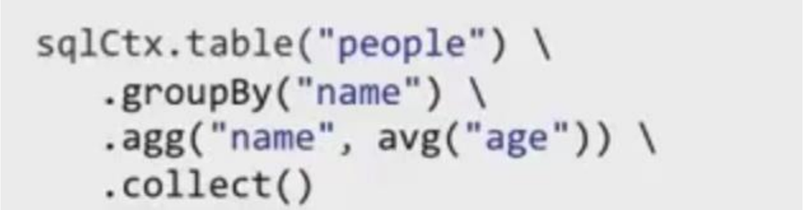
```
private IntWritable one =
  new IntWritable(1)
private IntWritable output =
  new IntWritable()
proctected void map(
    LongWritable key,
    Text value,
    Context context) {
  String[] fields = value.split("\t")
  output.set(Integer.parseInt(fields[1]))
  context.write(one, output)
}

IntWritable one = new IntWritable(1)
DoubleWritable average = new DoubleWritable()

protected void reduce(
    IntWritable key,
    Iterable<IntWritable> values,
    Context context) {
  int sum = 0
  int count = 0
  for(IntWritable value : values) {
     sum += value.get()
     count++
    }
  average.set(sum / (double) count)
  context.Write(key, average)
}
```

Spark RDD代码

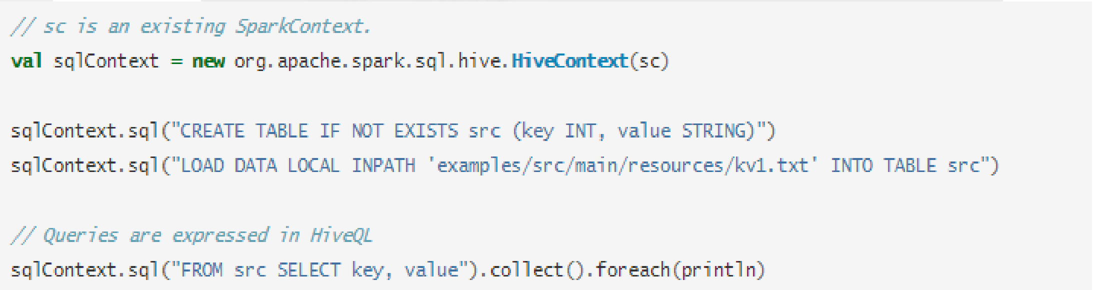


```
data = sc.textFile(...).split("\t")
data.map(lambda x: (x[0], [int(x[1]), 1])) \
   .reduceByKey(lambda x, y: [x[0] + y[0], x[1] + y[1]]) \
   .map(lambda x: [x[0], x[1][0] / x[1][1]]) \
   .collect()
```

Spark DataFrame代码



```
sqlCtx.table("people") \
    .groupBy("name") \
    .agg("name", avg("age")) \
    .collect()
```

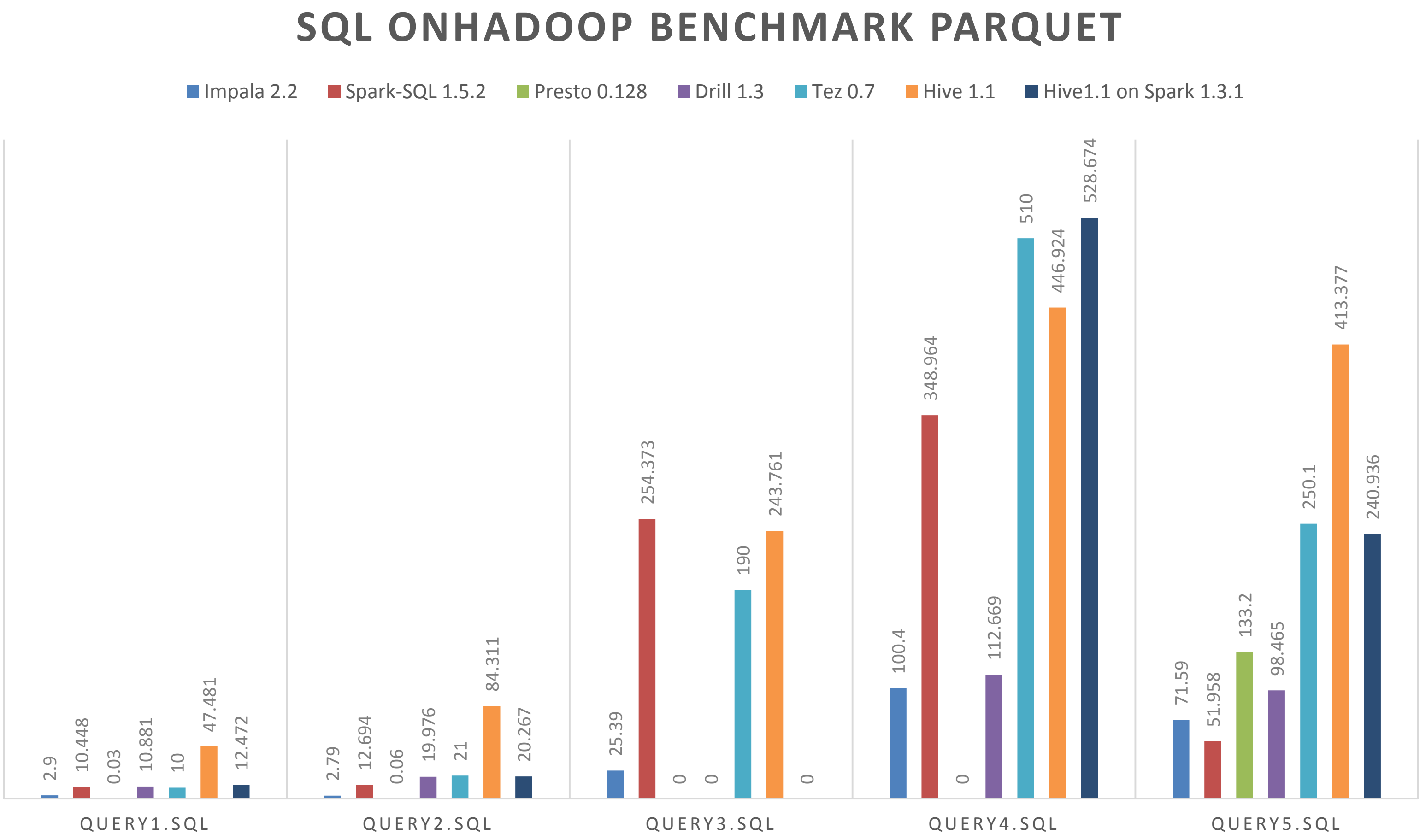


```
// sc is an existing SparkContext.
val sqlContext = new org.apache.spark.sql.hive.HiveContext(sc)

sqlContext.sql("CREATE TABLE IF NOT EXISTS src (key INT, value STRING)")
sqlContext.sql("LOAD DATA LOCAL INPATH 'examples/src/main/resources/kv1.txt' INTO TABLE src")

// Queries are expressed in HiveQL
sqlContext.sql("FROM src SELECT key, value").collect().foreach(println)
```

- 简化编程，让不熟悉函数式编程的人也能高效使用。

https://spark.apache.org/docs/latest/sql-programming-guide.html

性能测试



执行时间为0s，0是语法不支持,0.1执行失败

https://spark.apache.org/sql/

www.itweet.cn

性能测试

SQL	Phoenix	Hbase	Imapla on Hbase	Drill 1.3	Hive on Hbase
query1.sql	54.255	45.1	1861.45	0.1	510.406
query2.sql	18.416	20.2	835.26	285.483	444.557
query3.sql	0	0	0.1	0	1128.97
query4.sql	0	0	0.1	331.574	418.449
query5.sql	0	0	0.1	333.974	1092.463

执行时间为0s，0是语法不支持,0.1执行失败

SQL and Datasource： http://www.itweet.cn/2016/03/20/Impala-Hive-performance-tuning/

在线演示

Thank you

提问时间?

Blog： http://www.itweet.cn

PPT：https://github.com/itweet/course

Video：http://www.tudou.com/home/sparkjvm/