Spark Walk-through

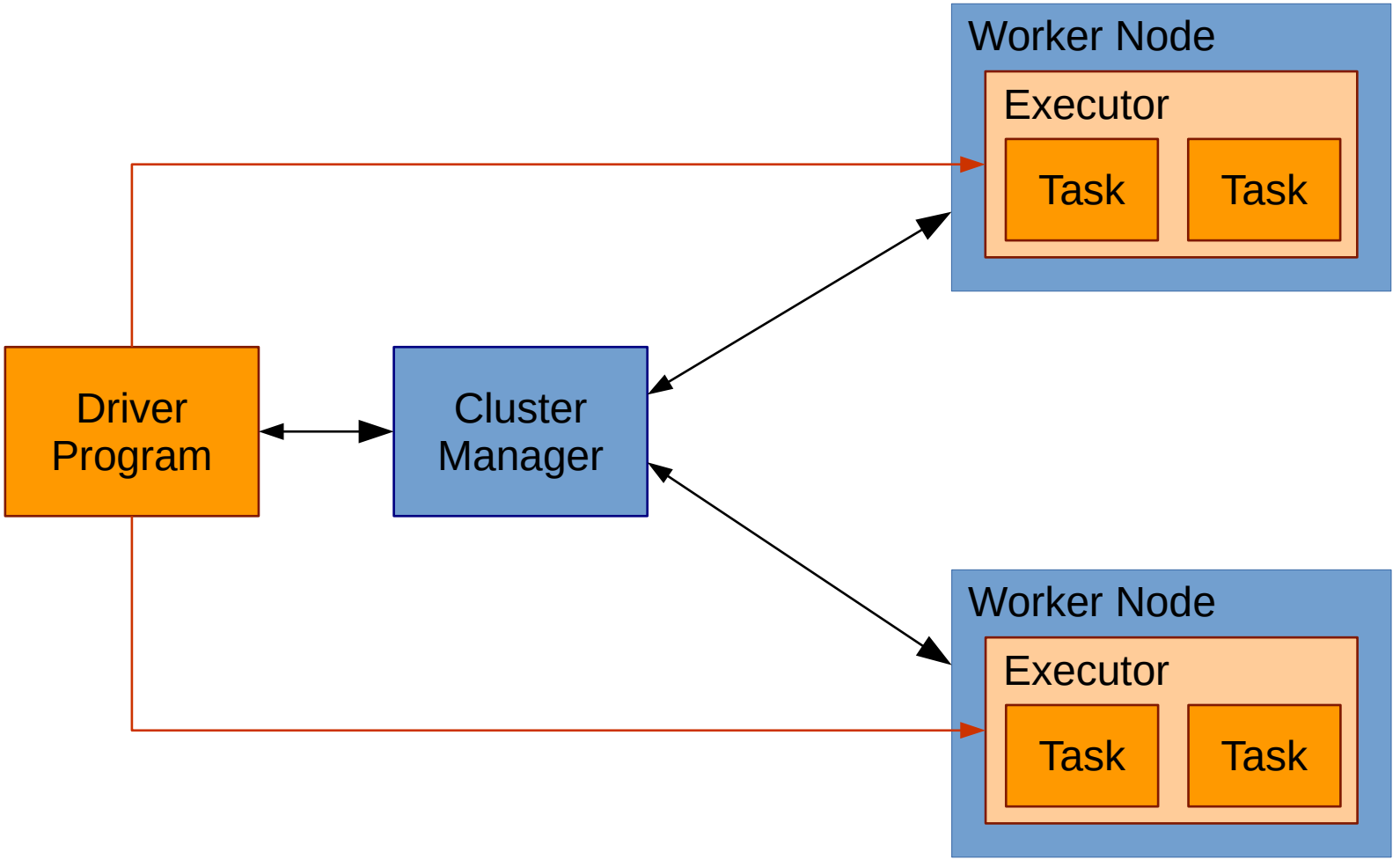
Ivan Georgiev

Timm Rüger IT Consulting
http://tr-itconsulting.com

Spark Key Features

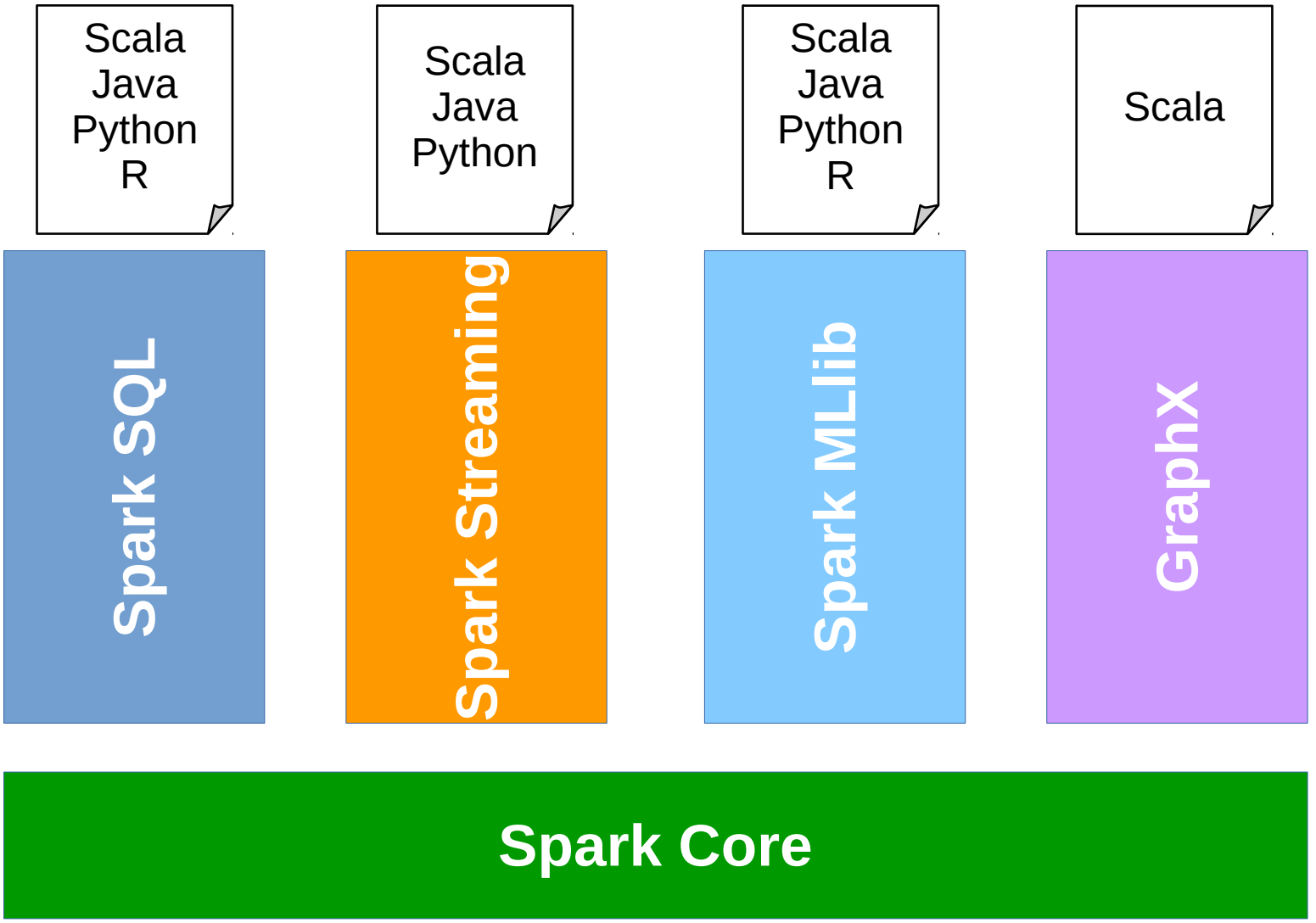
- Easy to use
  - Rich API, abstractions, programming languages
  - Cluster managers
- Fast
  - In-memory cluster computing, caching, DAG
- General Purpose
  - Batch, interactive, stream, ML, graph
- Scalable
- Fault tolerant
  - Automatic node failure handling

High Level Architecture



- Shuffle
- Job
- Stage

Spark Libraries



Spark Core API

- SparkContext
  - Main entry point
- Resilient Distributed Datasets (RDD)
  - immutable, partitioned, fault-tolerant, abstract programming interface, typed, in-memory
- Programming Model
  - Functional programming
  - Create RDD – textFile(), parallelize(), wholeTextFile(), sequenceFile(), JDBC, Cassandra, HBase, Hive, JSON, compressed formats etc.
  - Transformations – map, filter, flatMap, union, join, groupByKey, reduceByKey, etc.
  - Actions – count, countByValue, collect, first, max, min, take, fold, reduce, sum, etc.
  - Save RDD – saveAsTextFile, saveAsSequenceFile.
  - Lazy operations – RDD creation and transformation

Let's do Some Practice

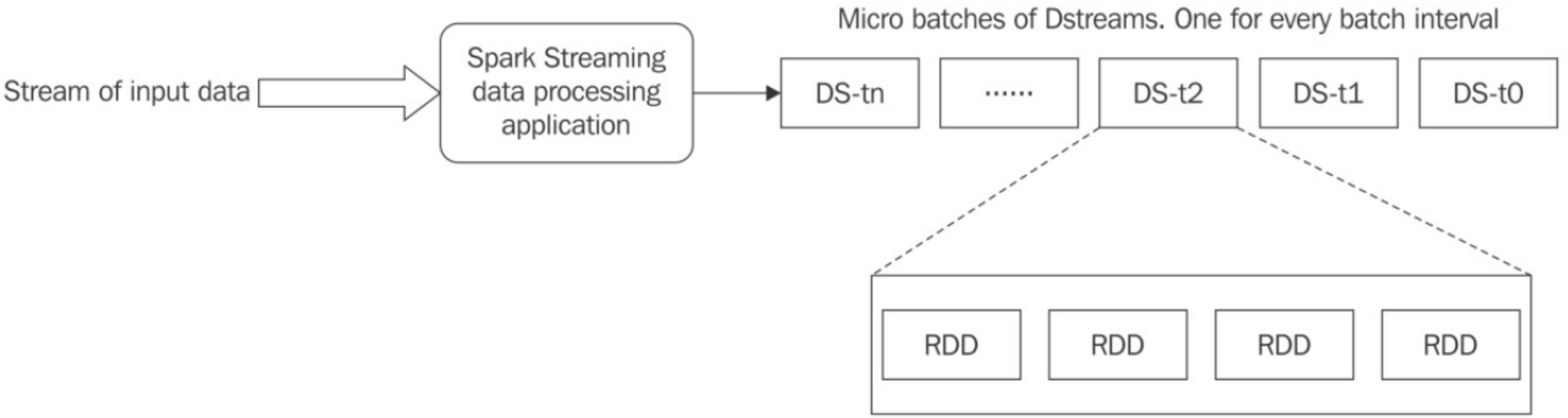
- Environment
  - Cloudera CDH 5.7.0 Quick Start VM, running on Virtual Box in Windows
  - Python 3
  - Jupyter Notebook (Could also use Apache Zeppelin – still in baby stage)
  - Apache Spark 2.0.1 for Hadoop 2.7
-

Summary I

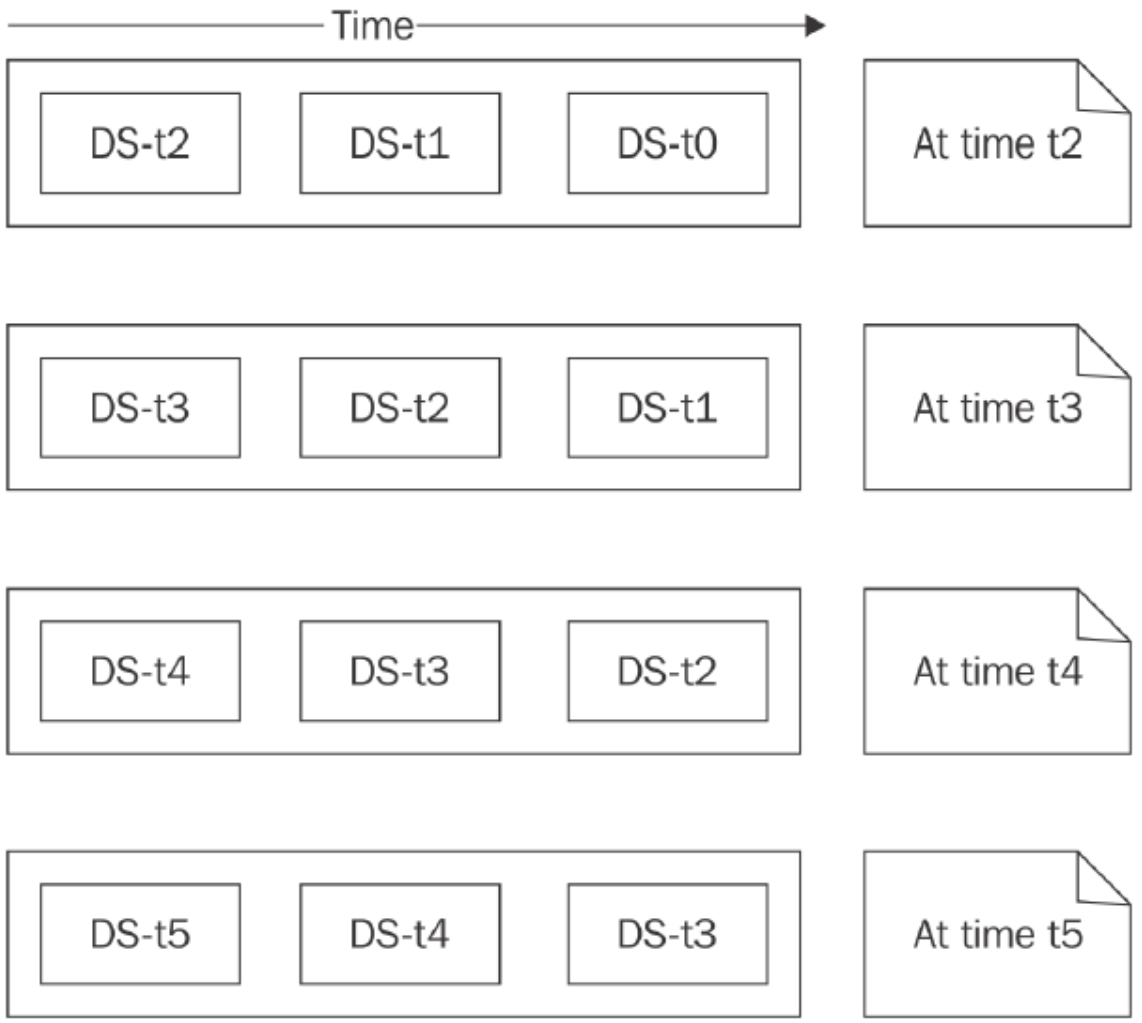
- Spark Core Features
- High Level Architecture
- Spark Libraries
- Spark Core API and Programming Model
- Practice
  - Core API with RDDs
  - Read JSON and Parquet files with DataFrame API
  - DataFrame operations – project and aggregate
-

Spark Streaming

DStream


Micro batches of Dstreams. One for every batch interval
Stream of input data
Spark Streaming data processing application
DS-tn
.....
DS-t2
DS-t1
DS-t0
RDD
RDD
RDD
RDD

Window Processing



Slide-Interval = m * Batch-Interval

Window-Length = n * Slide-Interval

Summary II

- All data – Parquet files on HDFS
- Implemented movie recommend engine using SparkSQL
- Created DataFrame views
- Materialized view into a physical Hive table
- Explored catalog
- Real time word popularity with Spark Streaming

What Else

- There is much more hidden in the core RDD and DataFrame APIs
- Sources and targets – just scratched the surface
- Machine Learning
- Graph processing
- Unit testing
- Monitoring

Time for Discussion