SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

Spark Ecosystem & Internals

陈 超 @CrazyJvm

outline

- ***basis & internals***
- ecosystem

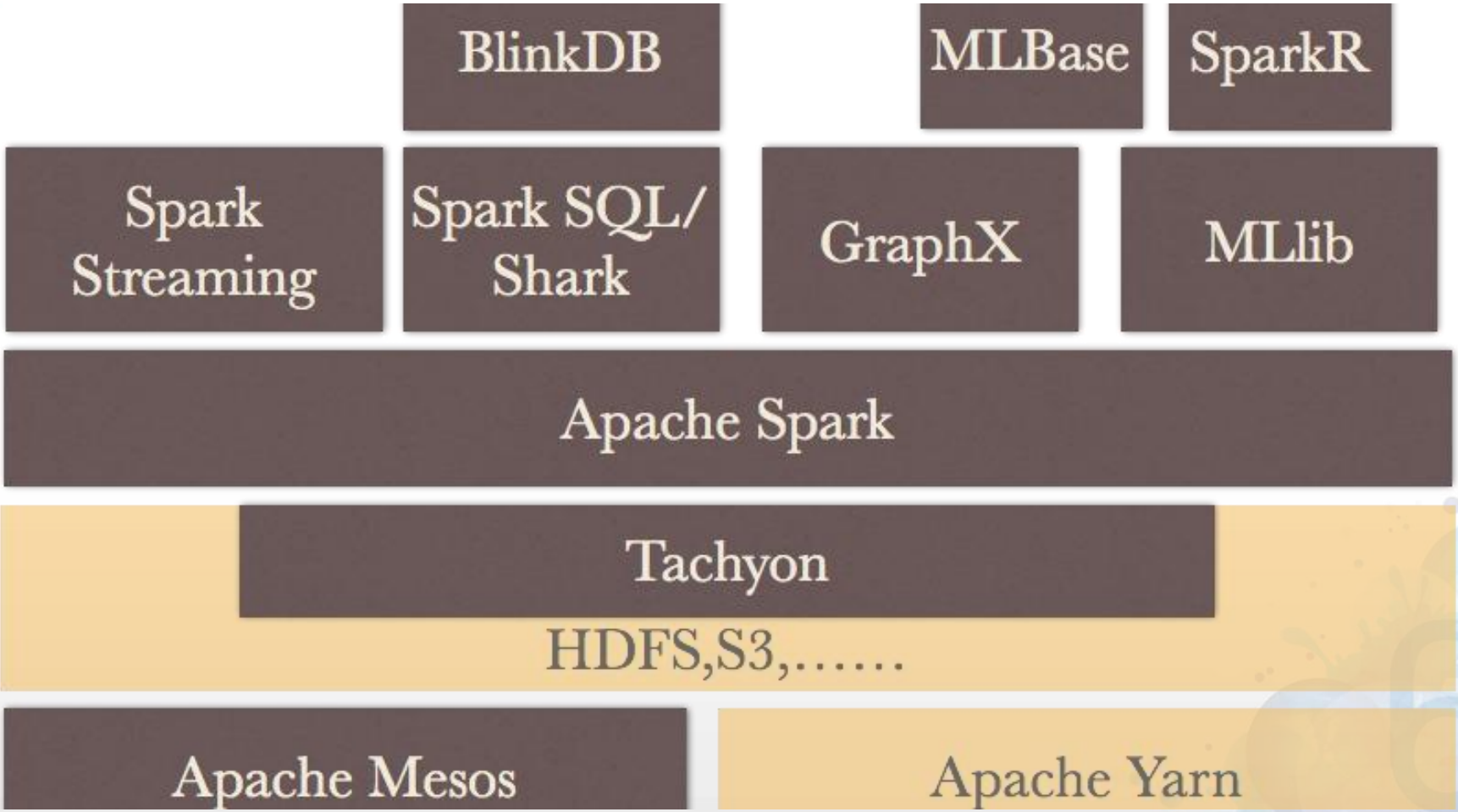
Current Major Release

- **Just released Spark 1.1**

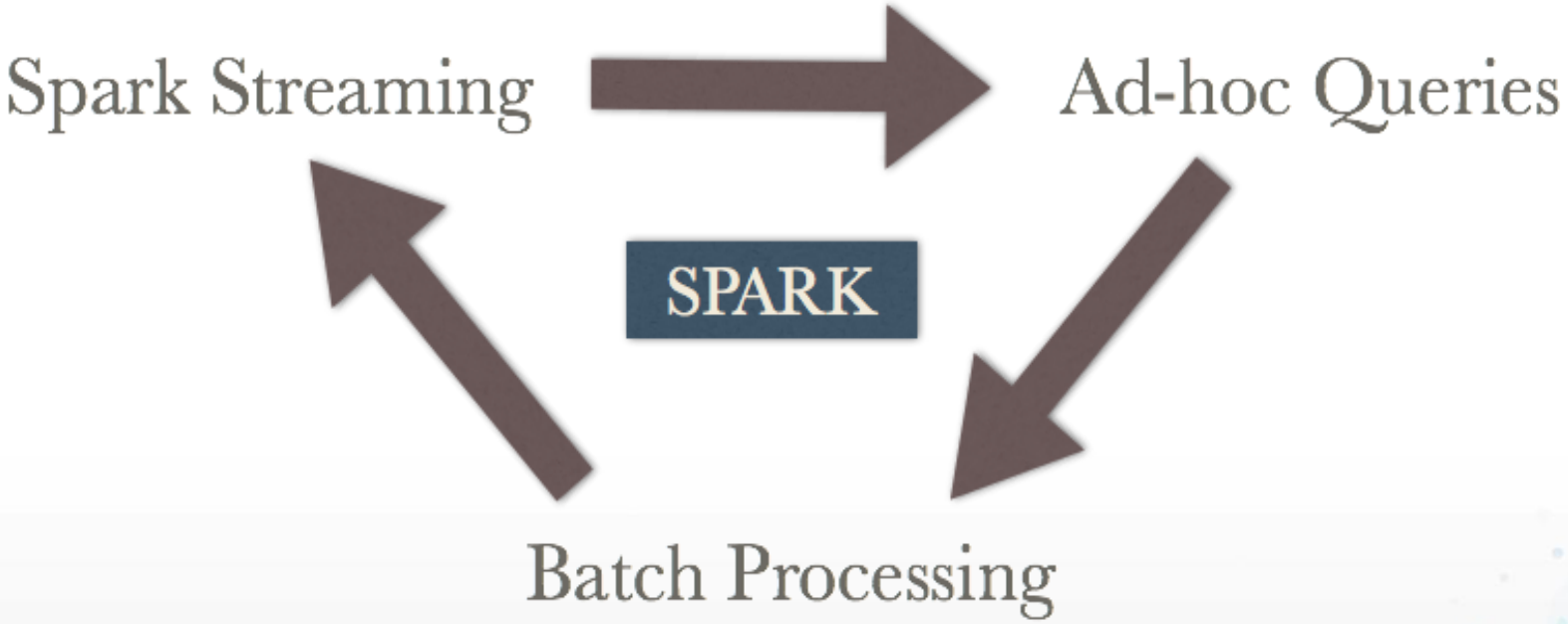
Spark : What & Why

- Apache Spark is a fast and general engine for large-scale data processing.
- Speed
- Ease of Use
- Generality
- Integrated with Hadoop

BDAS


BlinkDB
MLBase
SparkR
Spark Streaming
Spark SQL/ Shark
GraphX
MLlib
Apache Spark
Tachyon
HDFS,S3,……
Apache Mesos
Apache Yarn

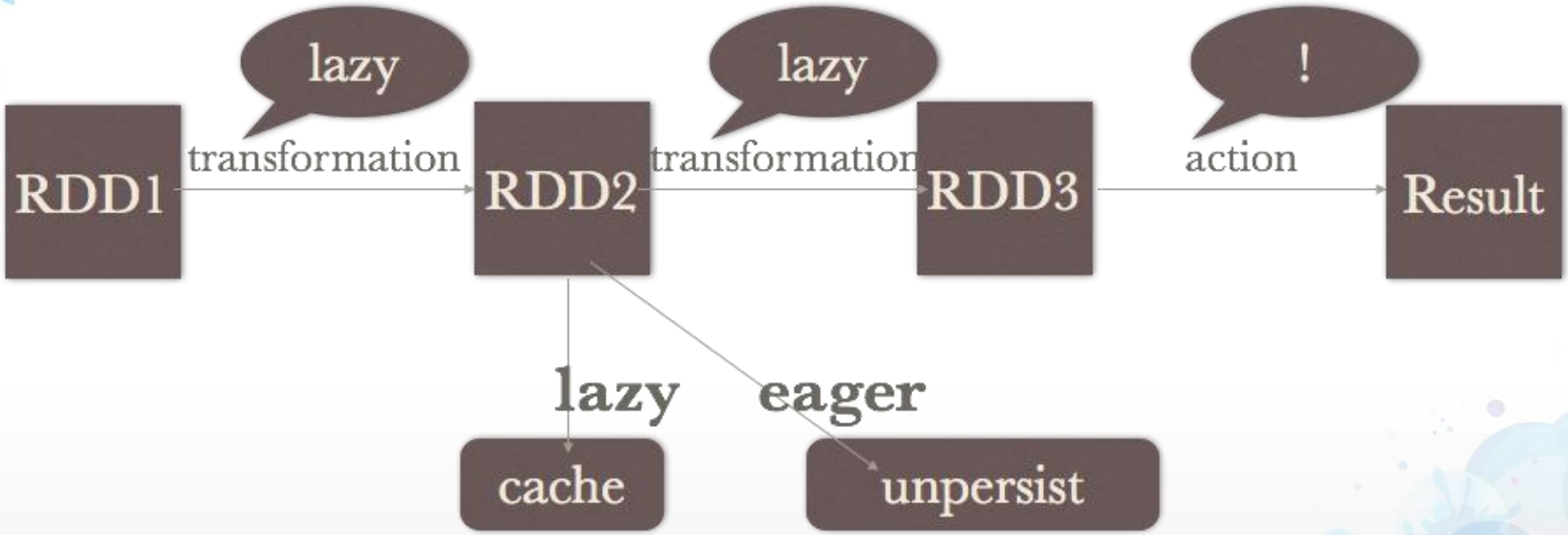
one stack to rule them all



Key Concept-RDD

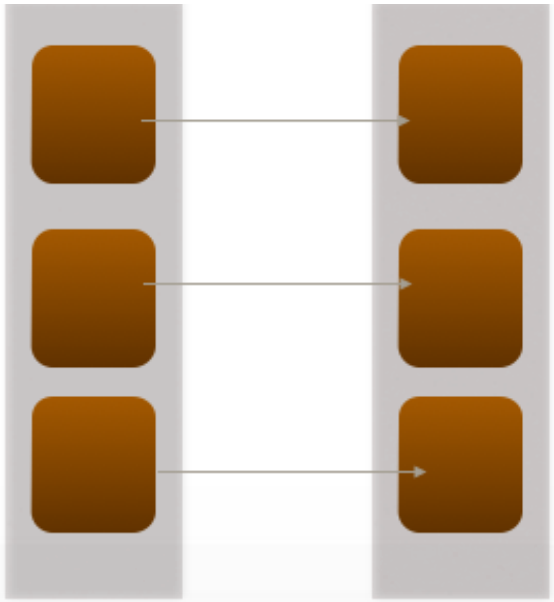
- A list of partitions
- A function for computing each split
- A list of dependencies on other RDDs
- Optionally, a Partitioner for key-value RDDs
- Optionally, a list of preferred locations to compute each split on

Key Concept-Lineage

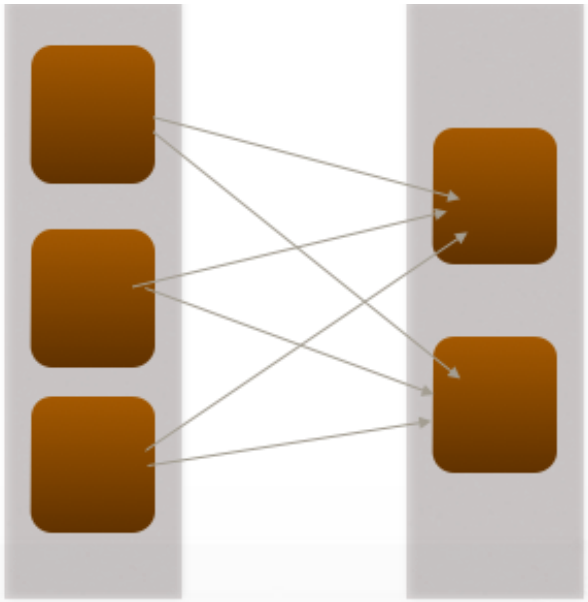


unroll partition safely when caching

Key Concept-Dependency



Narrow Dependency

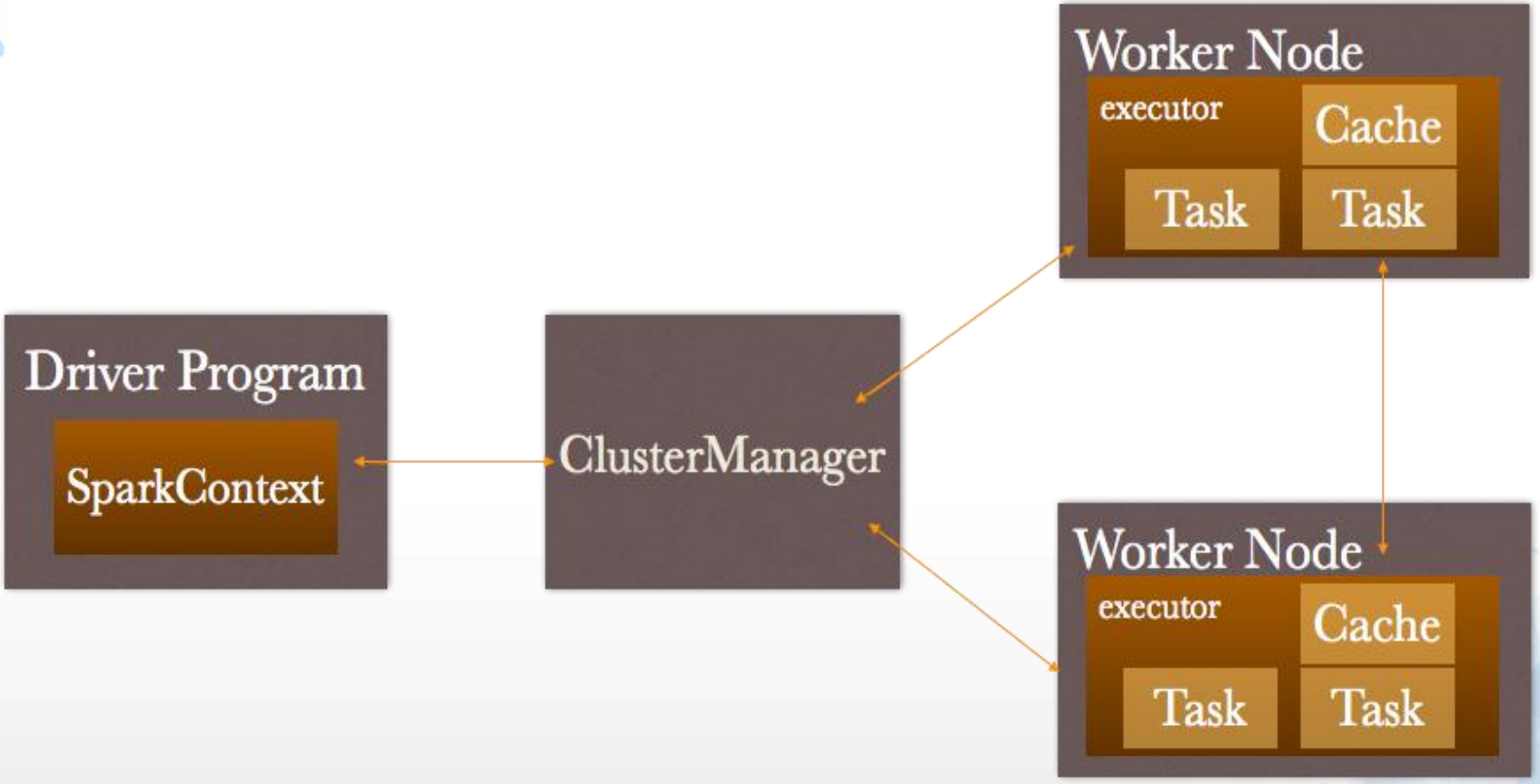


Wide Dependency

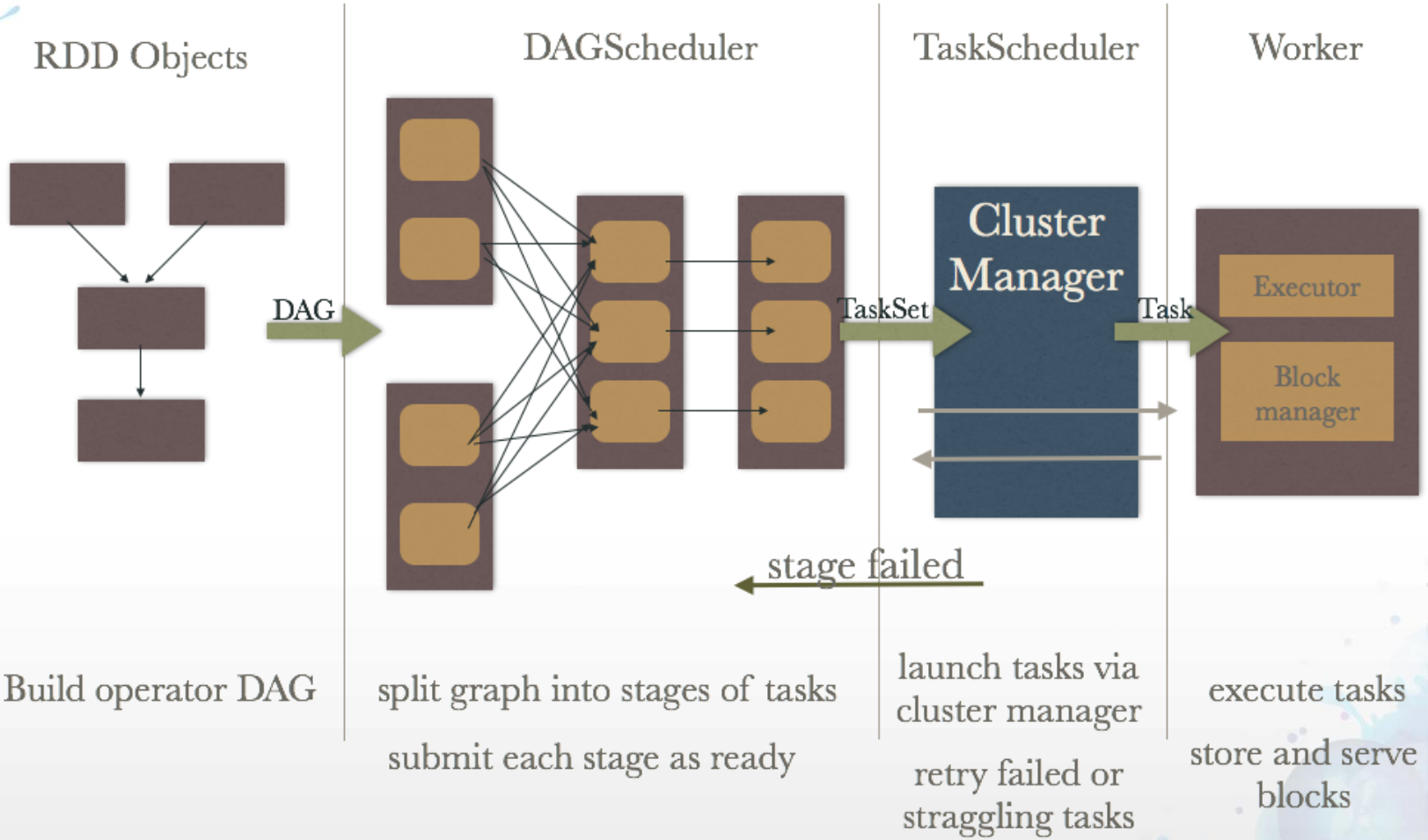
Key Concept-ClusterManager

- Local
- Standalone
- Yarn
- Mesos

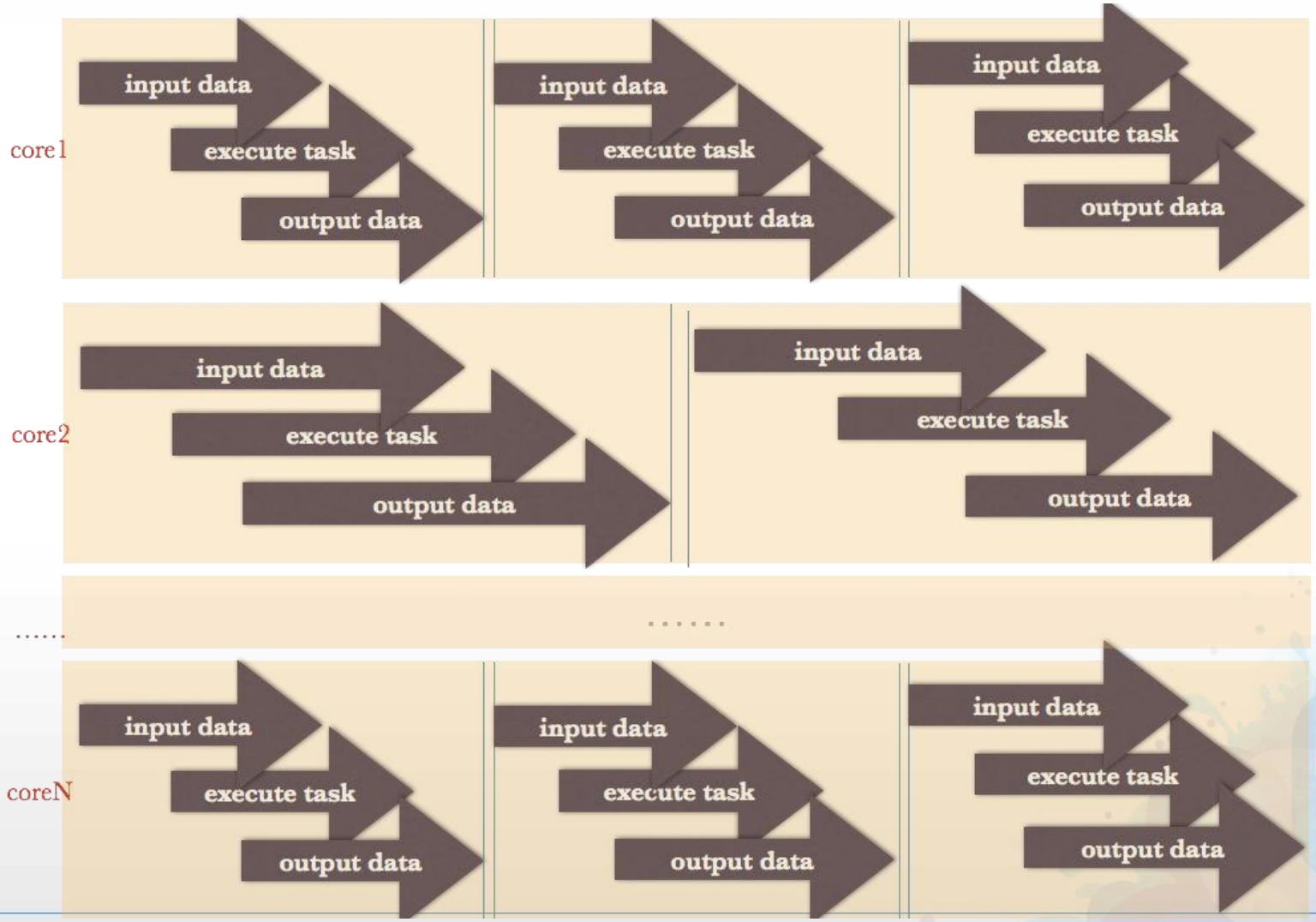
Cluster Overview



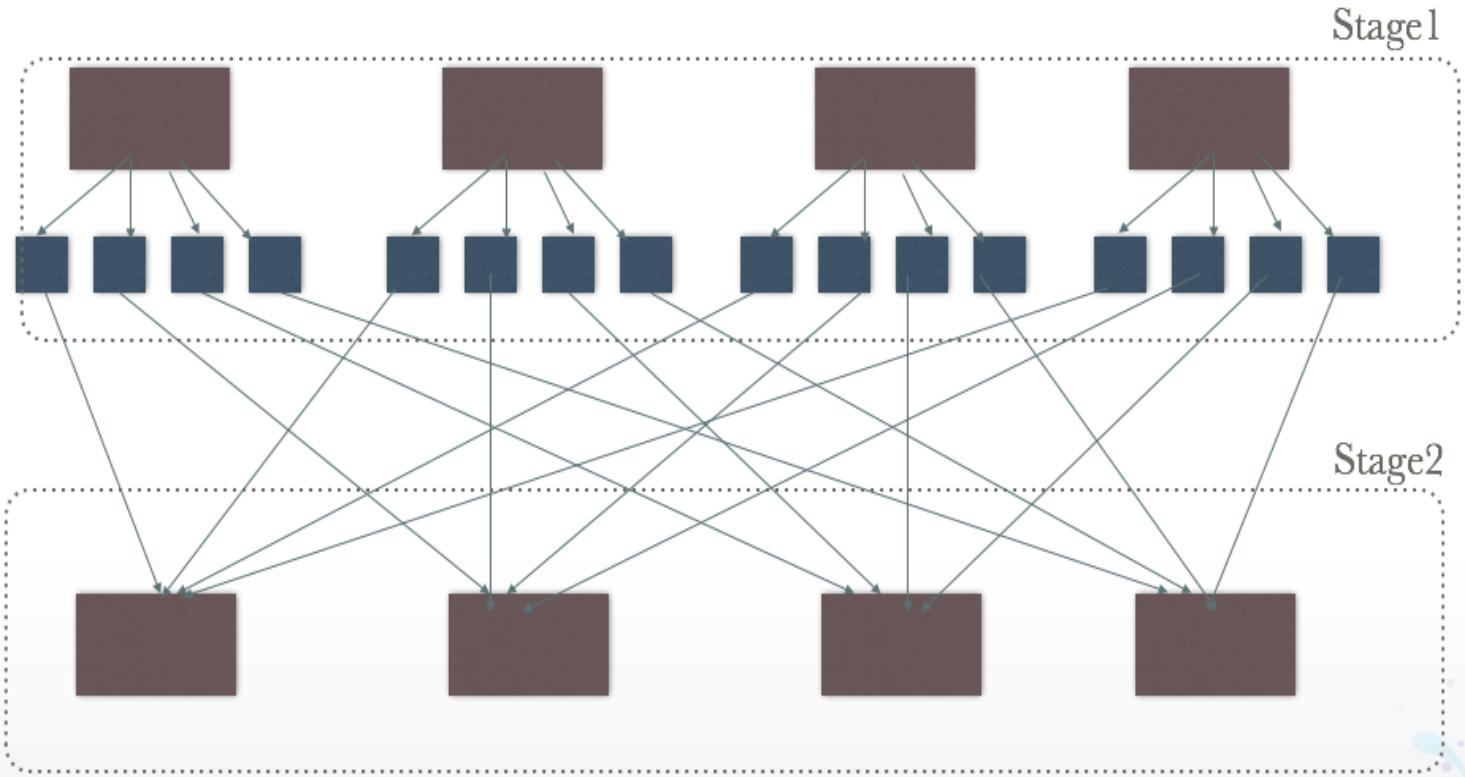
schedule


RDD Objects
DAGScheduler
TaskScheduler
Worker
DAG
TaskSet
Task
Cluster Manager
Executor
Block manager
stage failed
Build operator DAG
split graph into stages of tasks
submit each stage as ready
launch tasks via cluster manager
retry failed or straggling tasks
execute tasks
store and serve blocks

Executor


core1
input data
execute task
output data
input data
execute task
output data
input data
execute task
output data
core2
input data
execute task
output data
input data
execute task
output data
.....
.....
coreN
input data
execute task
output data
input data
execute task
output data
input data
execute task
output data

Shuffle



Sort-based shuffle supported

Shuffle

- Pull-based (not push-based)
- Write intermediate files to disk
- Build hash map within each partition
- Can spill across keys
- A single key-value pair must fit in memory

Better Metrics System

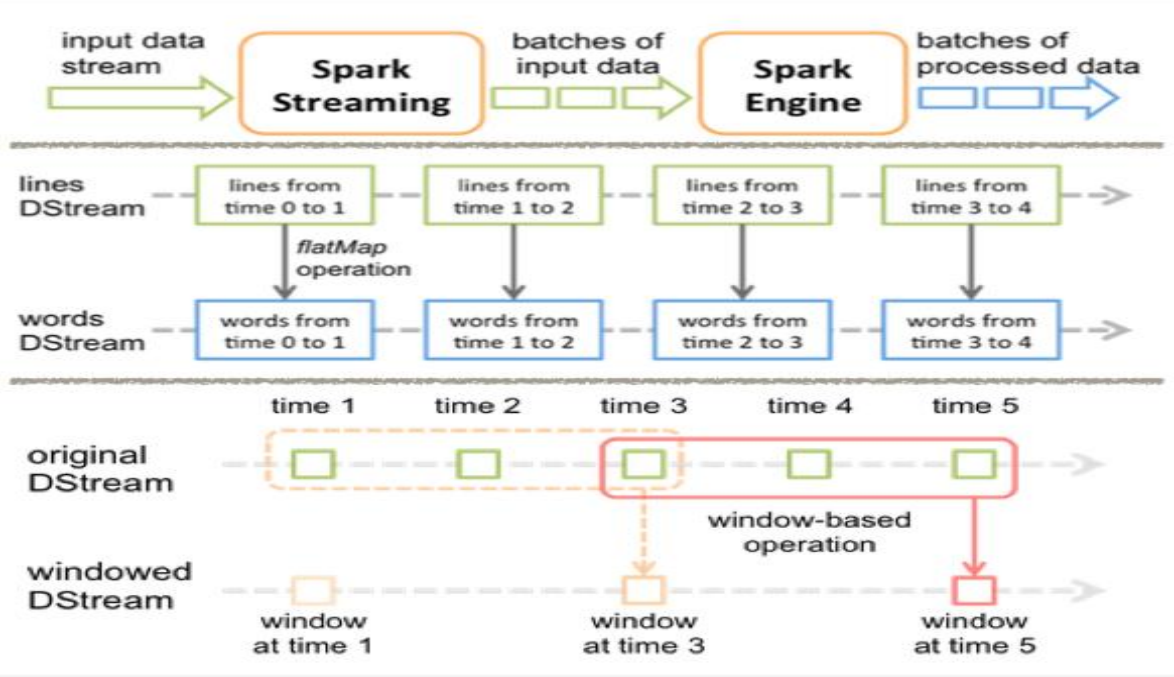
- Previously: only collect after task completed
- Now : report when task is still running

outline

- basis & internals
- ***ecosystem***

Spark Streaming

- Mini-batch

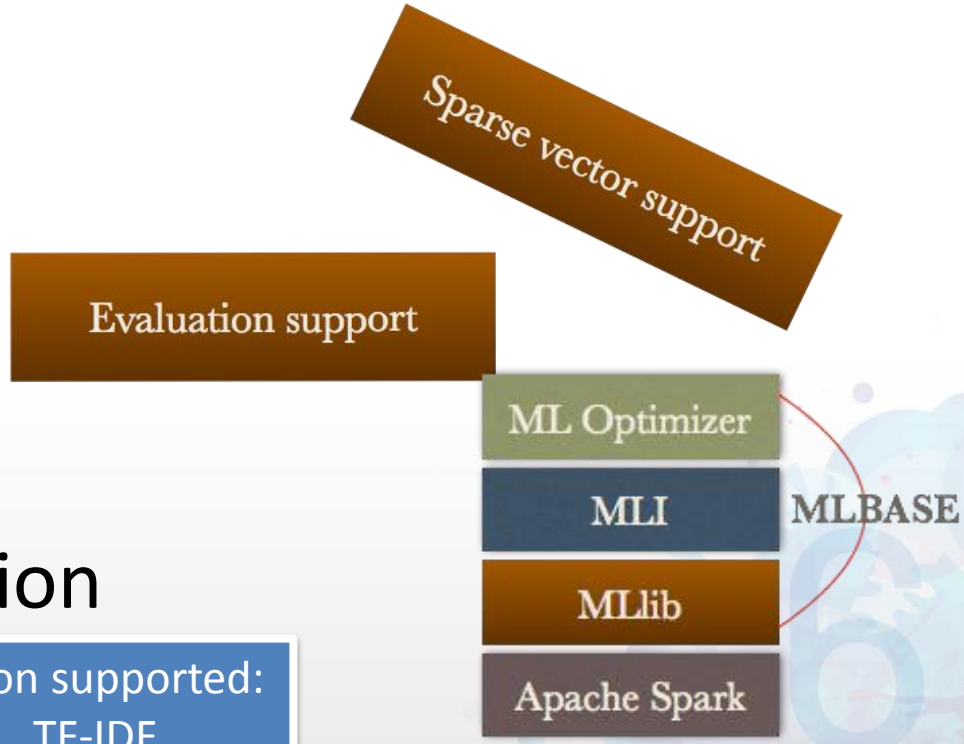


rate limiting supported!

Streaming + MLlib

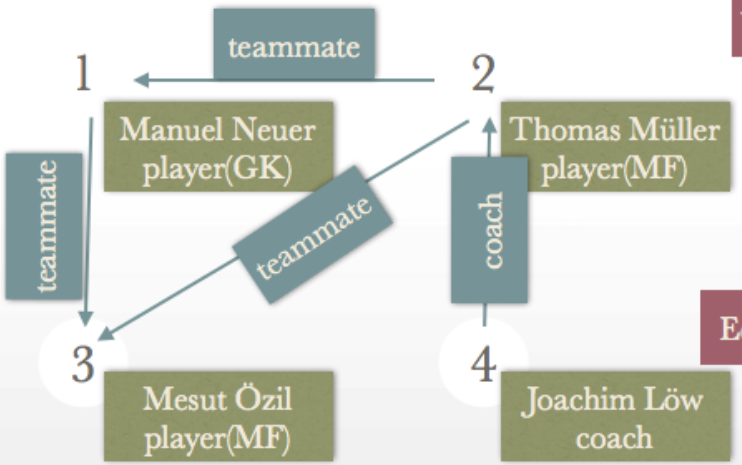
MLlib

- Spark implementation of some common machine learning algorithms and utilities
- classification
- regression
- clustering
- collaborative filtering
- dimensionality reduction



feature extraction supported:
Word2Vec , TF-IDF

GraphX



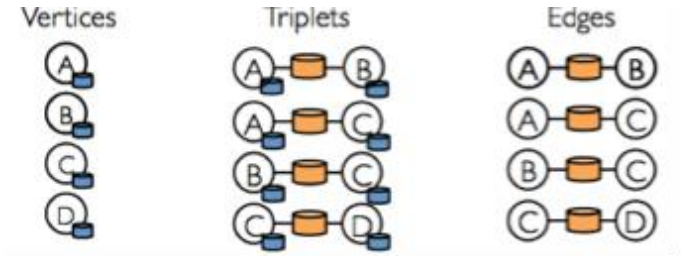
Vertex Table

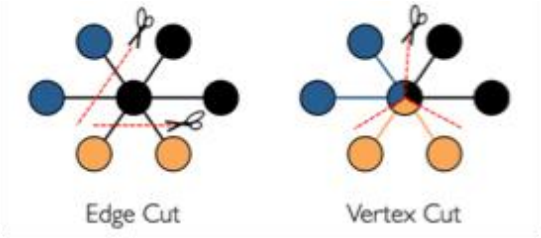
id	
1	(Manuel Neuer,player)
2	(Thomas Müller,player)
3	(Mesut Özil,player)
4	(Joachim Löw,coach)

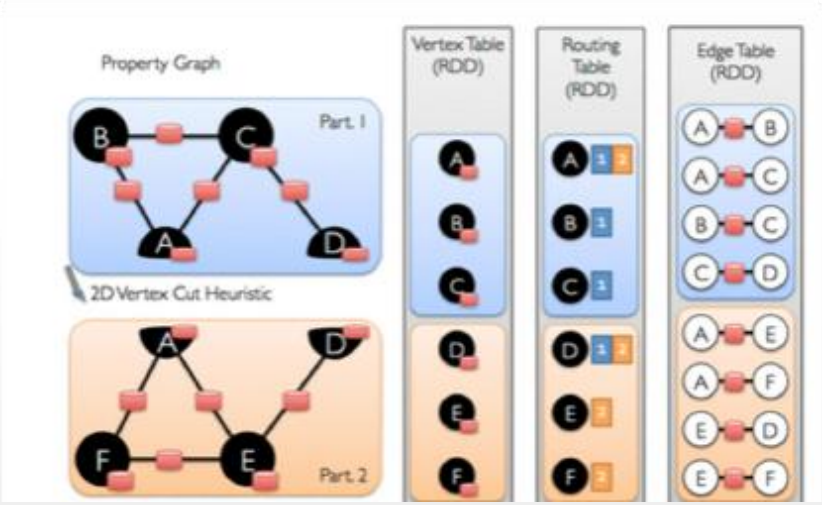
Edge Table

SrcId	DstId	Property(E)
2	1	teammate
2	3	teammate
1	3	teammate
4	2	coach

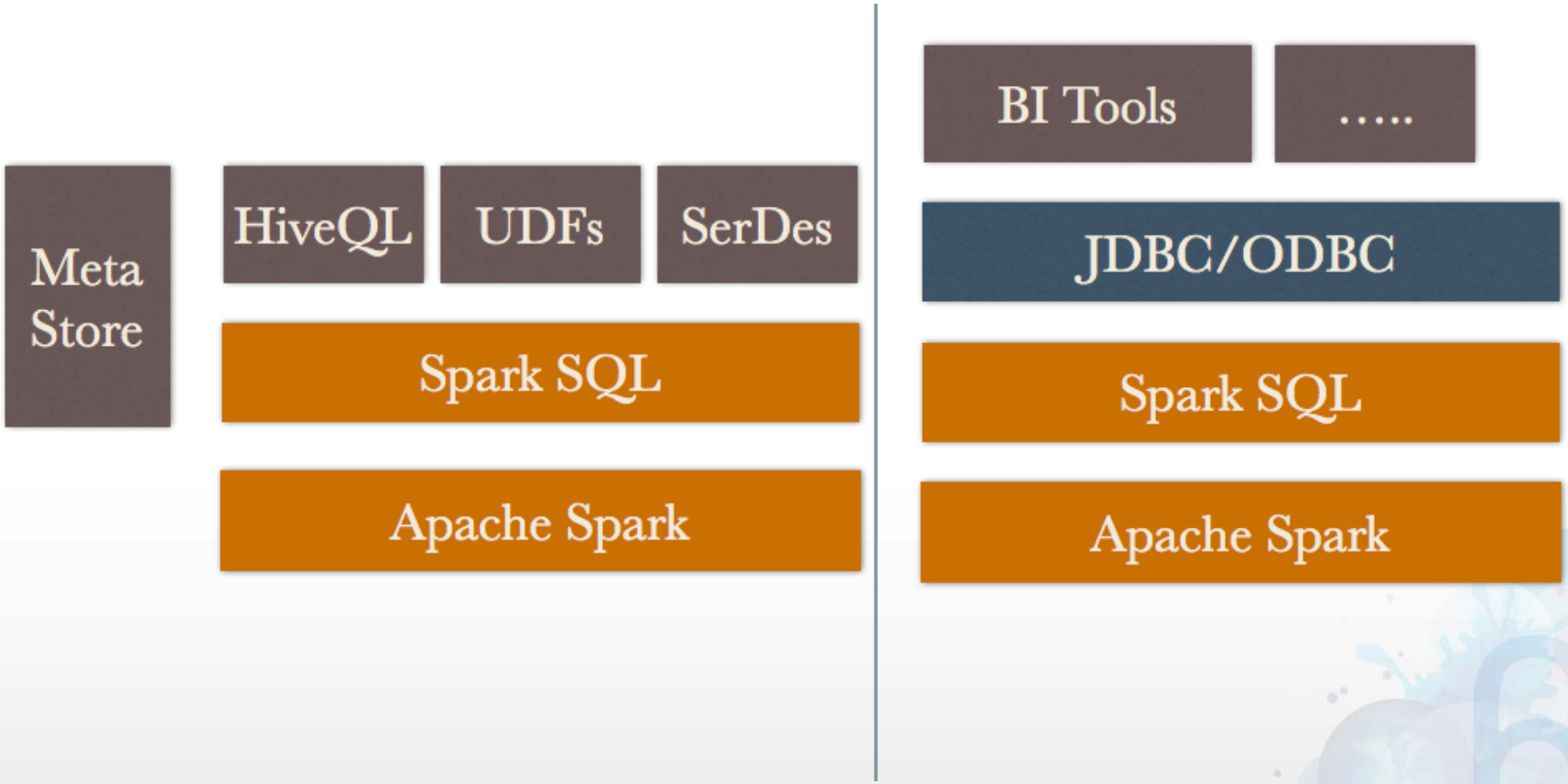
GraphX


Vertices
Triplets
Edges


Edge Cut
Vertex Cut


Property Graph
Part. 1
2D Vertex Cut Heuristic
Part. 2
Vertex Table (RDD)
Routing Table (RDD)
Edge Table (RDD)

Spark SQL


Meta Store
HiveQL
UDFs
SerDes
Spark SQL
Apache Spark
BI Tools
.....
JDBC/ODBC
Spark SQL
Apache Spark

Spark SQL

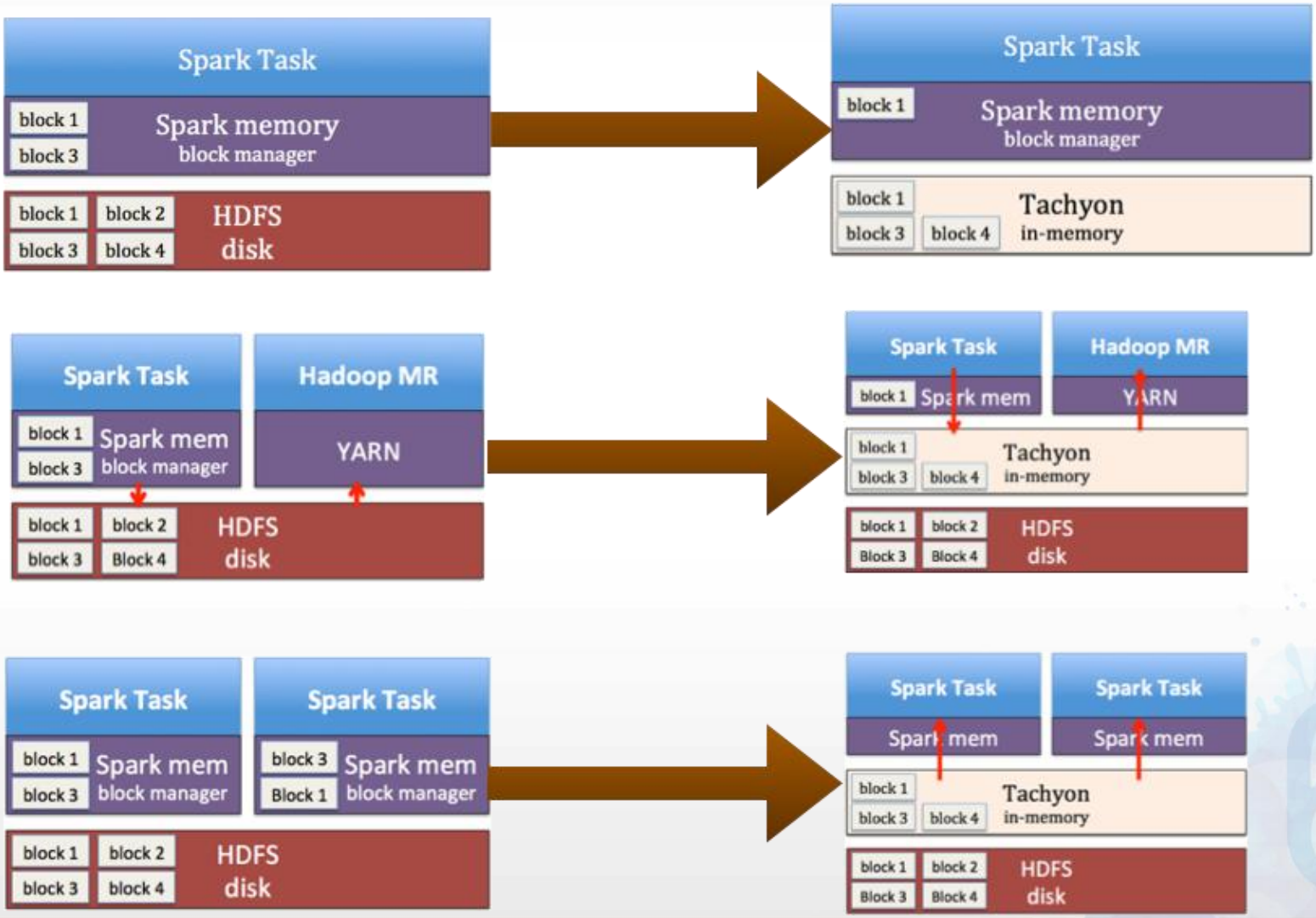
- Data Sources
- RDDs/Parquet Files/JSON Datasets/Hive Table
- DSL
- JDBC Server

Programmatically Specifying the Schema

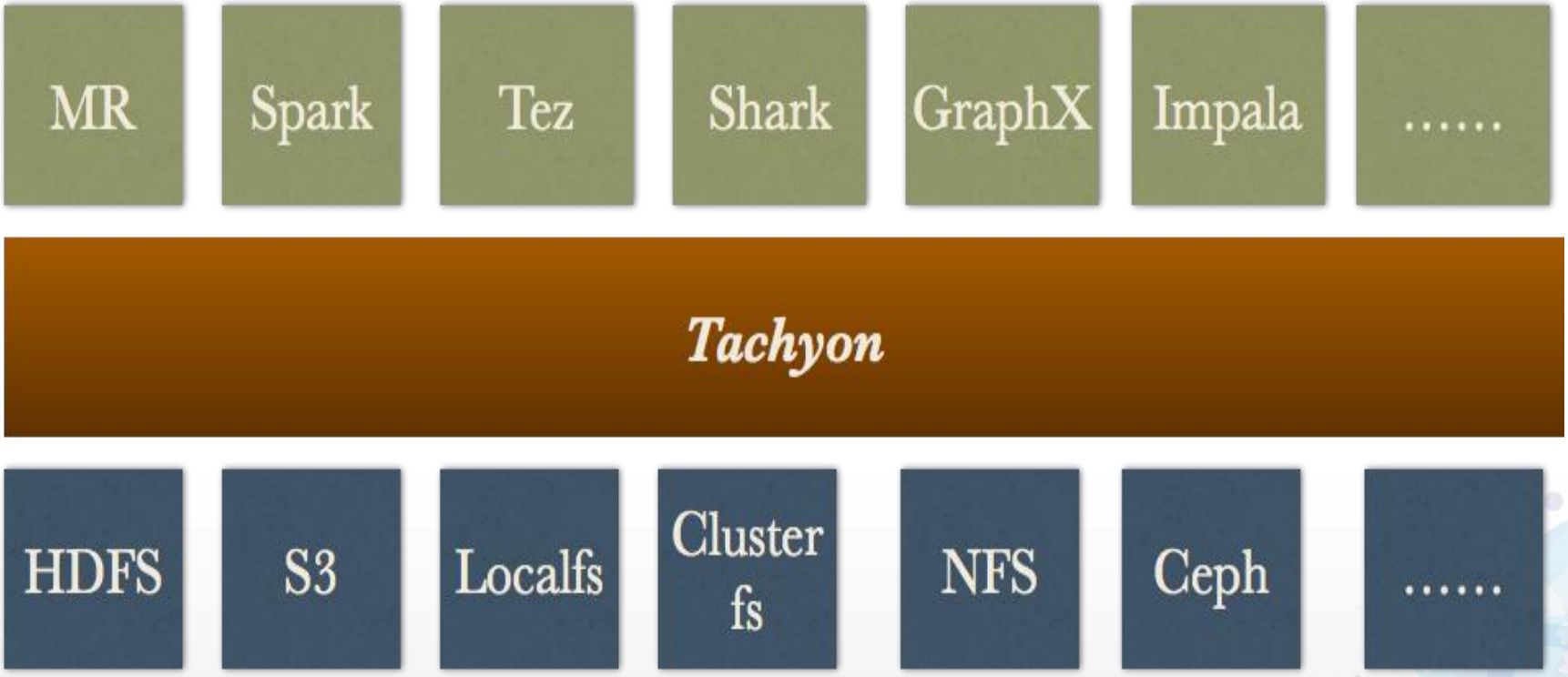
Shark

- Development in Shark has been ended and subsumed by Spark SQL
- Mission completed !!!

Tachyon


Spark Task
block 1
block 3
Spark memory
block manager
block 1
block 2
block 3
block 4
HDFS
disk
Spark Task
block 1
Spark memory
block manager
block 1
block 3
block 4
Tachyon
in-memory
Spark Task
Hadoop MR
block 1
block 3
Spark mem
block manager
YARN
block 1
block 2
block 3
Block 4
HDFS
disk
Spark Task
Hadoop MR
block 1
Spark mem
YARN
block 1
block 3
block 4
Tachyon
in-memory
block 1
block 2
Block 3
Block 4
HDFS
disk
Spark Task
Spark Task
block 1
block 3
Spark mem
block manager
block 3
Block 1
Spark mem
block manager
block 1
block 2
block 3
block 4
HDFS
disk
Spark Task
Spark Task
Spark mem
Spark mem
block 1
block 3
block 4
Tachyon
in-memory
block 1
block 2
Block 3
Block 4
HDFS
disk

Tachyon


MR
Spark
Tez
Shark
GraphX
Impala
……
Tachyon
HDFS
S3
Localfs
Cluster fs
NFS
Ceph
……

SparkR

RDDs as Distributed Lists

```
sc <- sparkR.init("local")
lines <- textFile(sc, "hdfs://data.txt")
wordsPerLine <- lapply(lines, function(line) { length(unlist(strsplit(line, " "))) })
```

BlinkDB

- Queries with Bounded Errors and Bounded Response Times on Very Large Data

```
SELECT avg(sessionTime)
FROM Table
WHERE city='San Francisco'
WITHIN 2 SECONDS
```

Queries with Time Bounds

```
SELECT avg(sessionTime)
FROM Table
WHERE city='San Francisco'
ERROR 0.1 CONFIDENCE 95.0%
```

Queries with Error Bounds

contact me

weibo:@CrazyJvm

wechat public account : ChinaScala

Q&A

THANKS

SequeMedia 盛拓传媒 | IT168 www.it168.com | ChinaUnix | ITPUB