

Spark Ecosystem

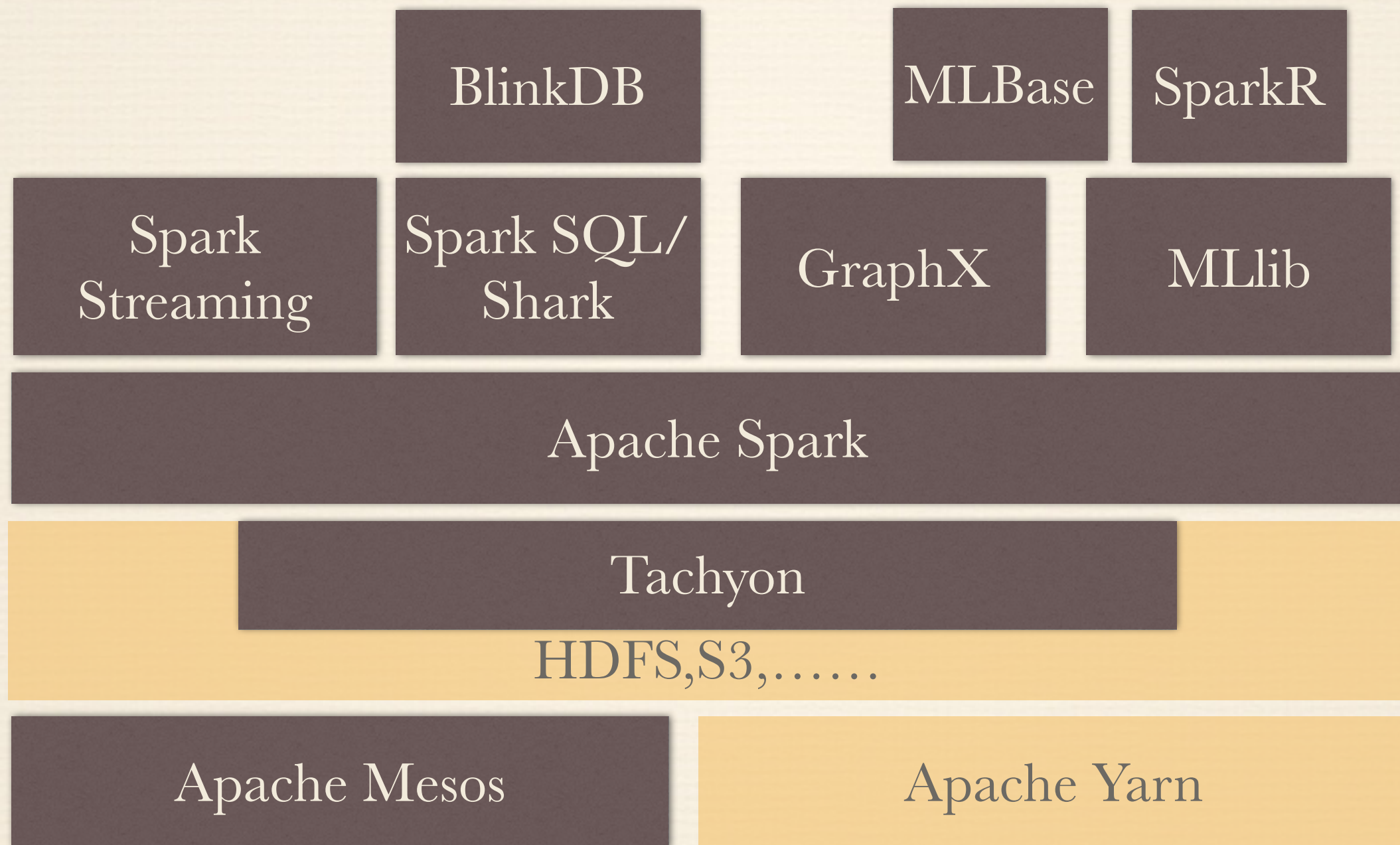


陈超 @CrazyFm
Spark Meetup @Hangzhou
2014.08.31

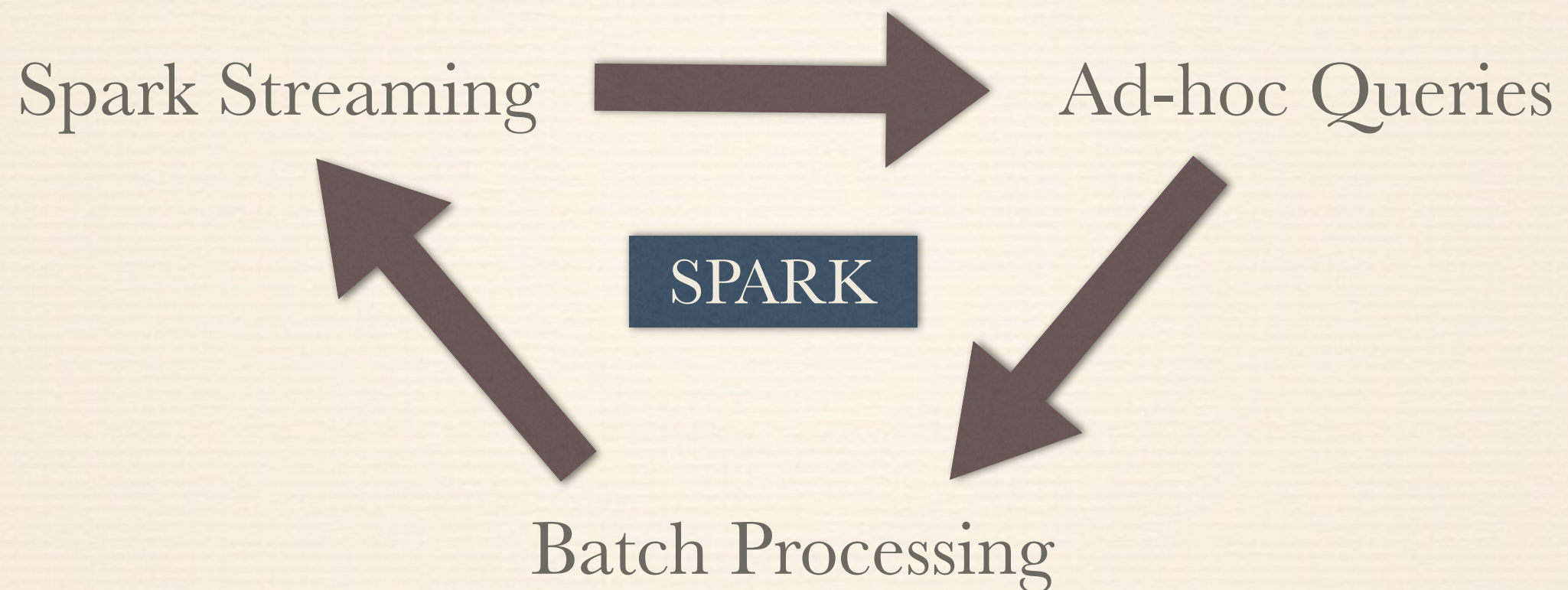
What is Spark

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

BDAS



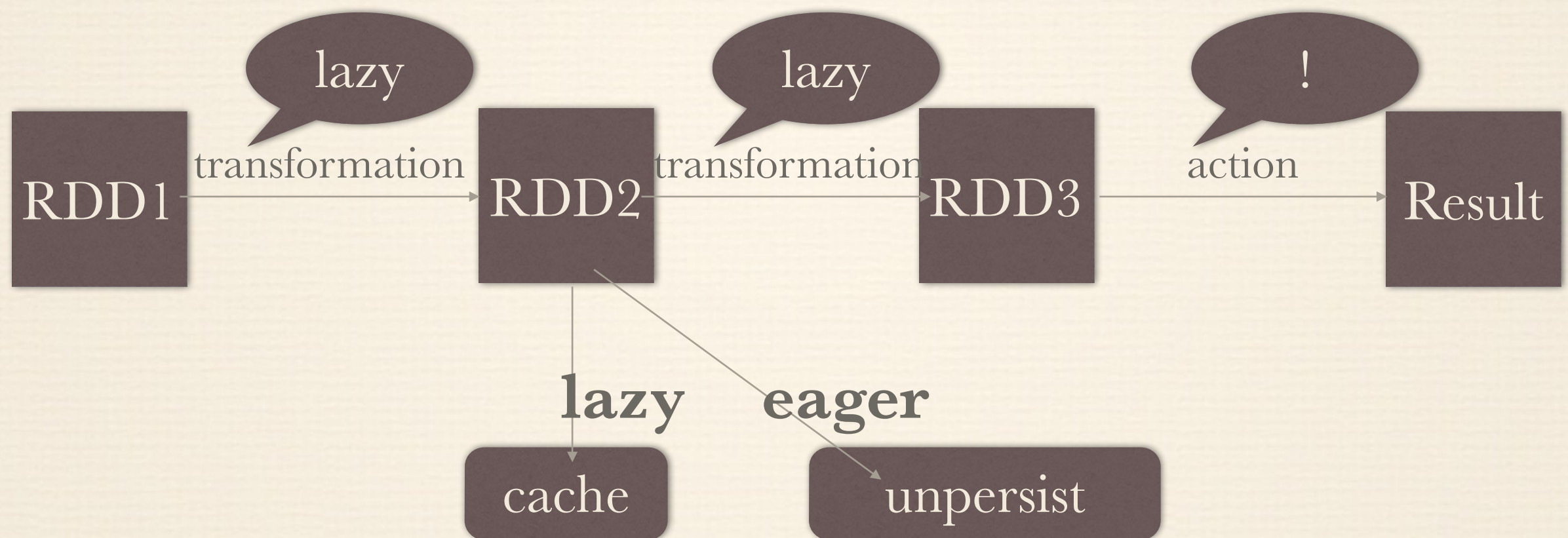
one stack to rule them all



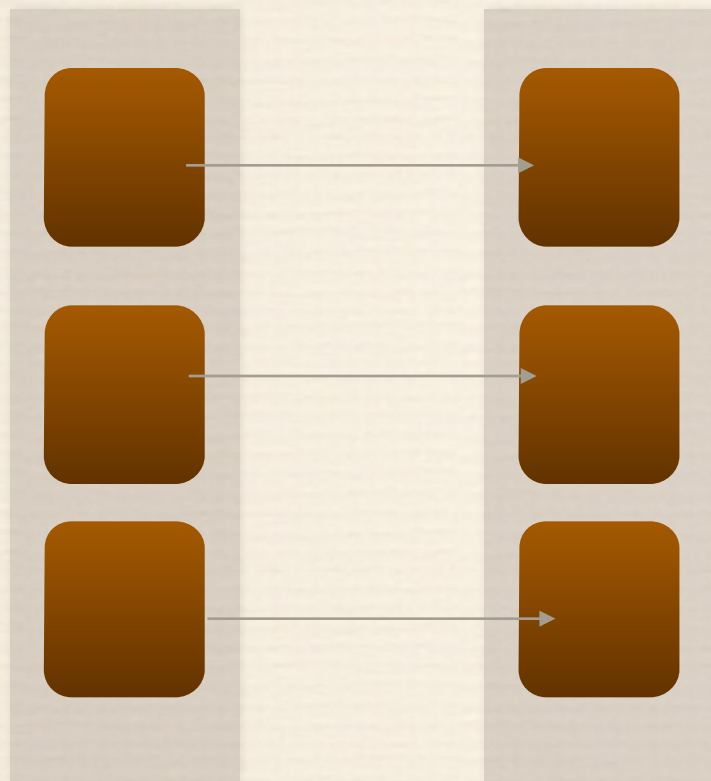
Key Concept-RDD

- ❖ 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

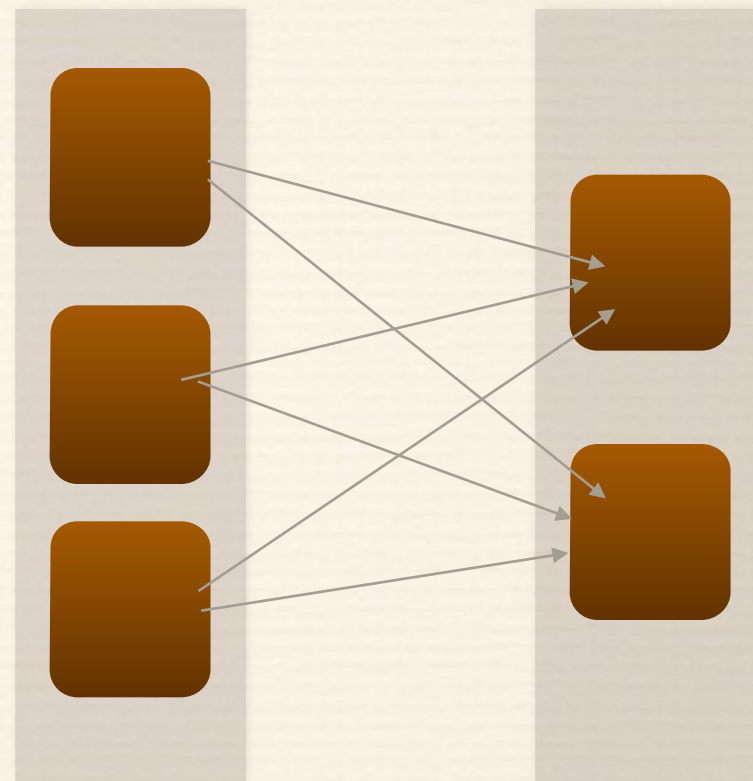
Key Concept-Lineage



Key Concept-Dependency



Narrow
Dependency

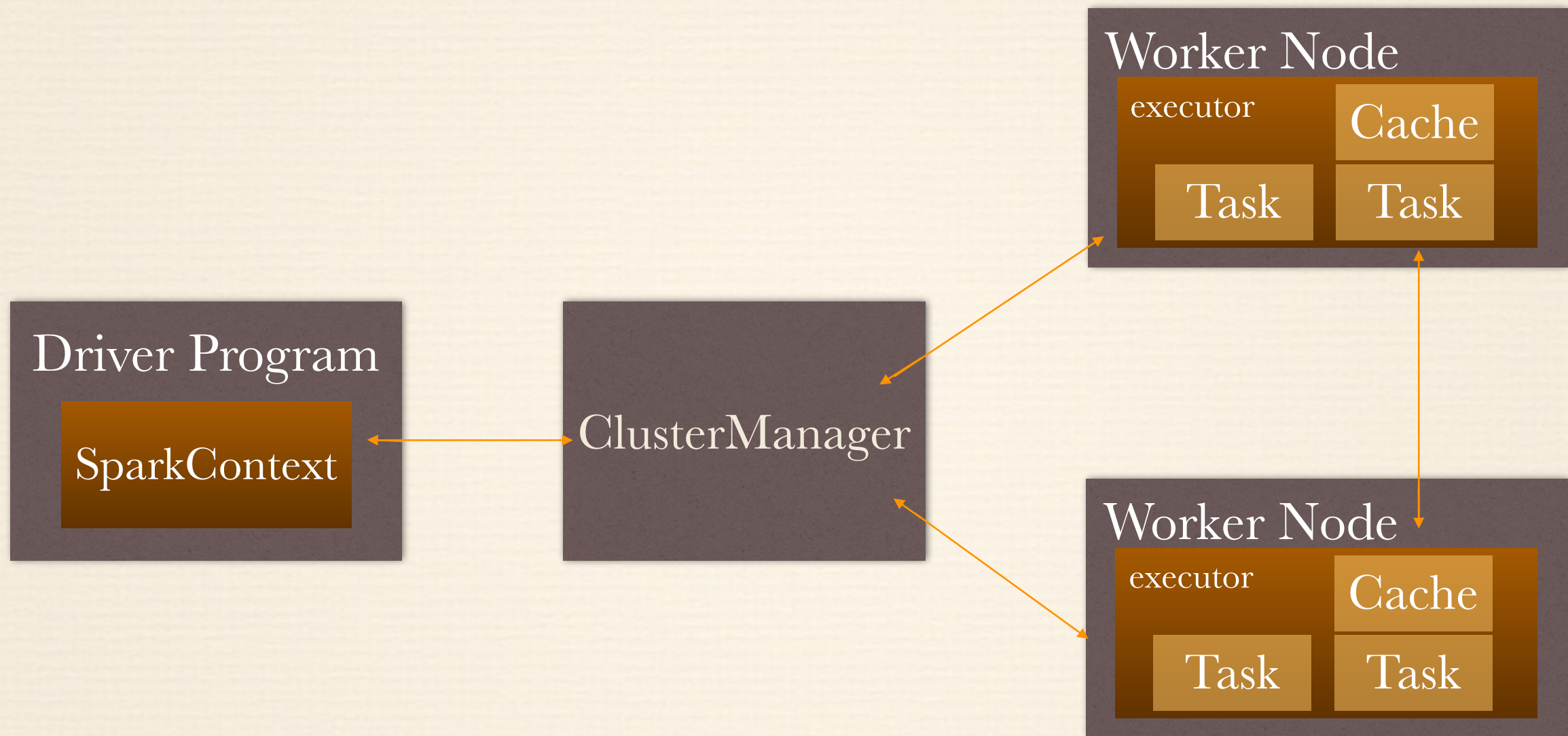


Wide
Dependency

Key Concept-ClusterManager

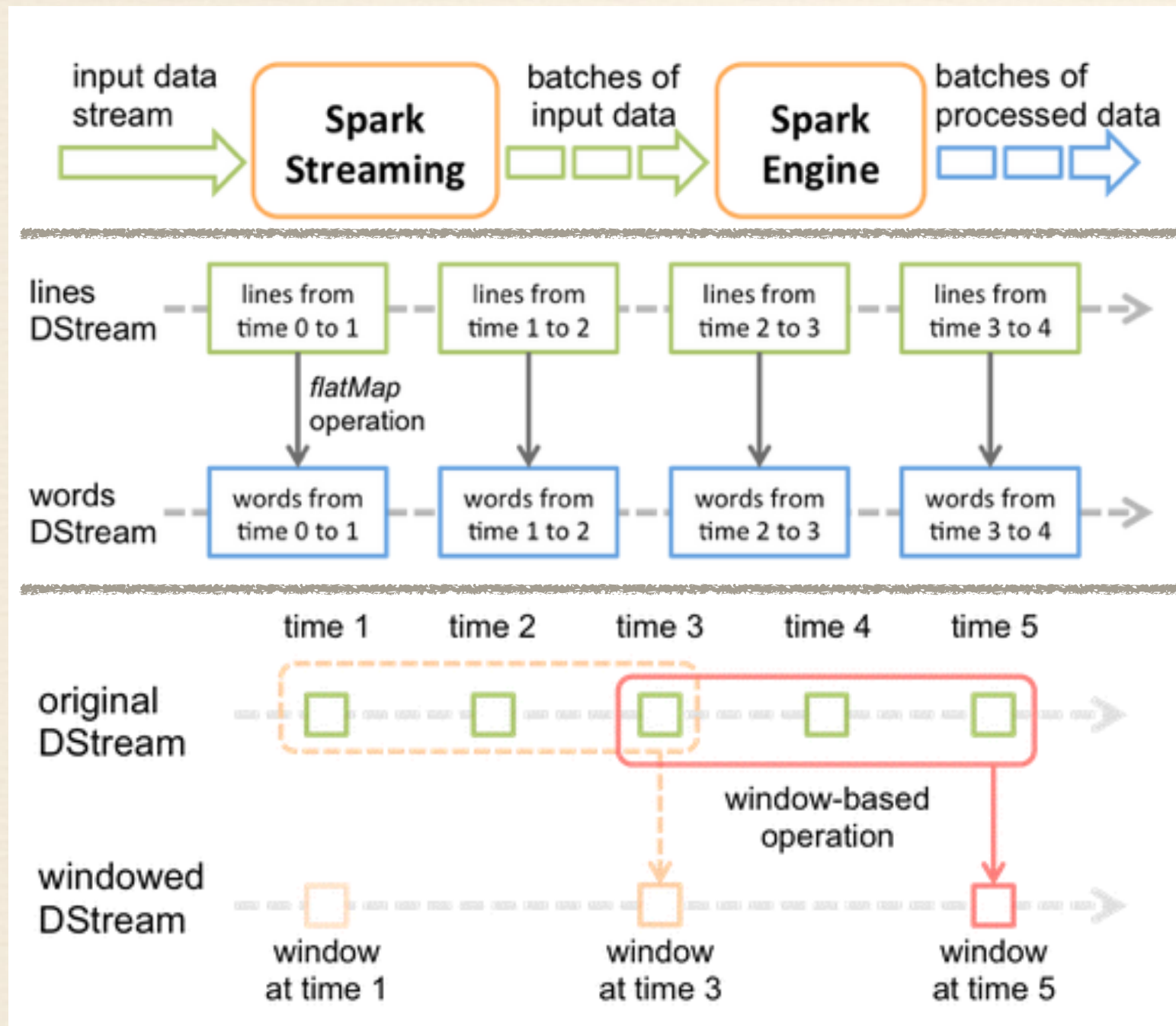
- ❖ Local
- ❖ Standalone
- ❖ Yarn
- ❖ Mesos

Cluster Overview



Spark Streaming

❖ mini-batch



MLlib

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

Sparse vector support

Evaluation support

ML Optimizer

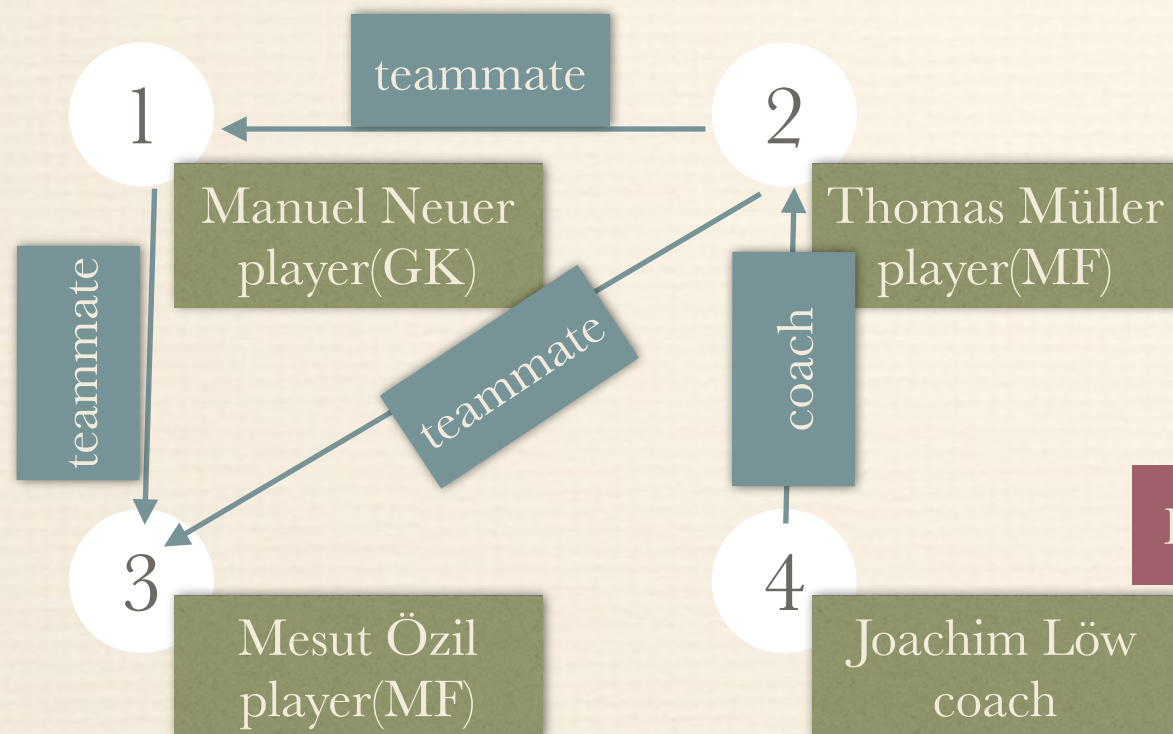
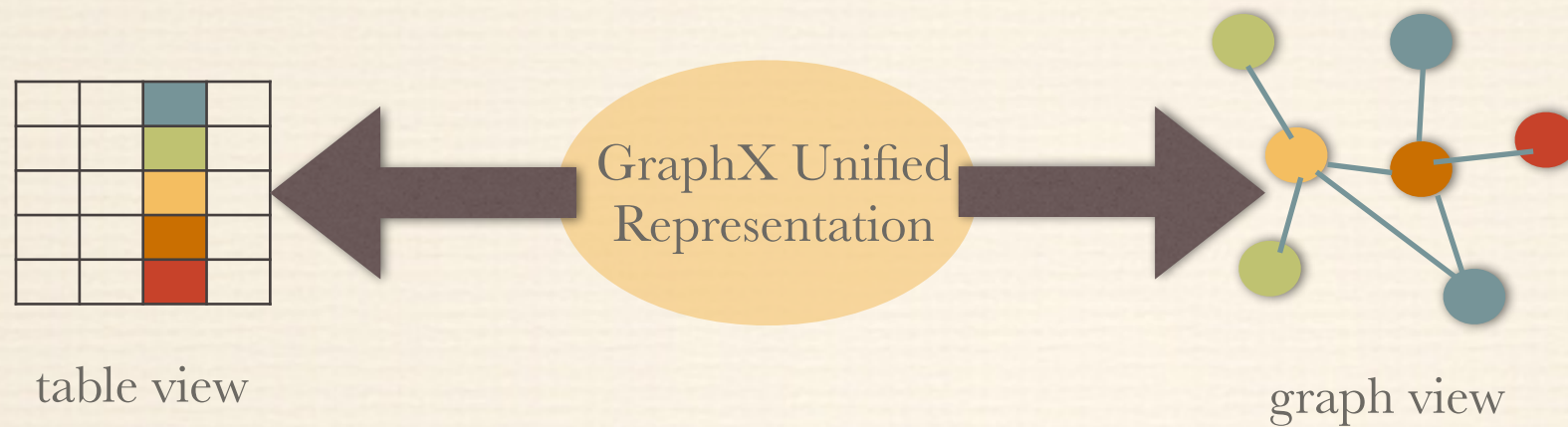
MLI

MLlib

Apache Spark

MLBASE

GraphX



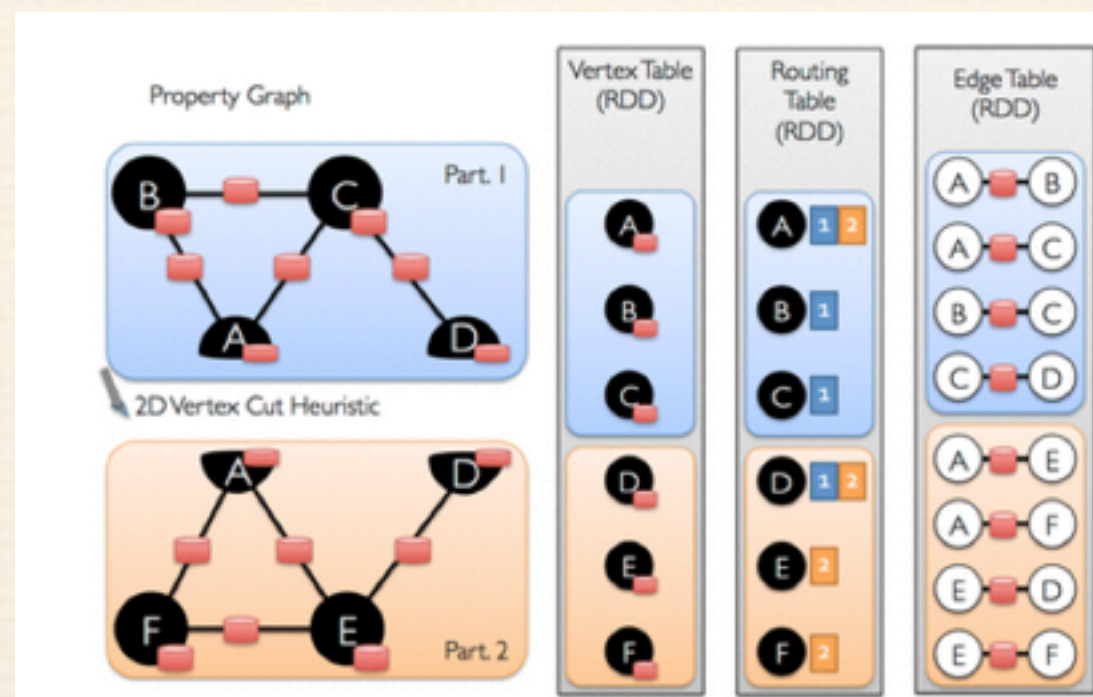
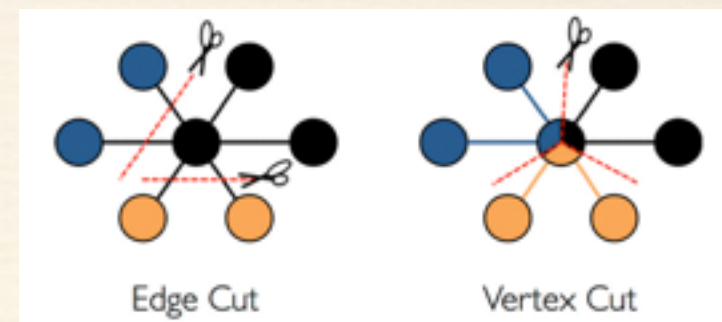
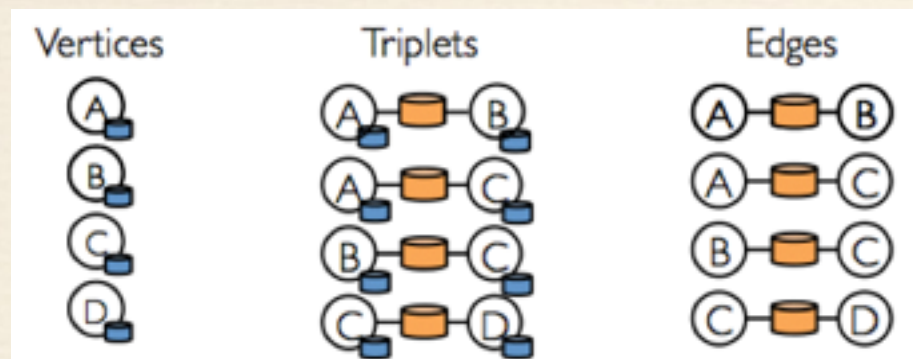
Vertex Table

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

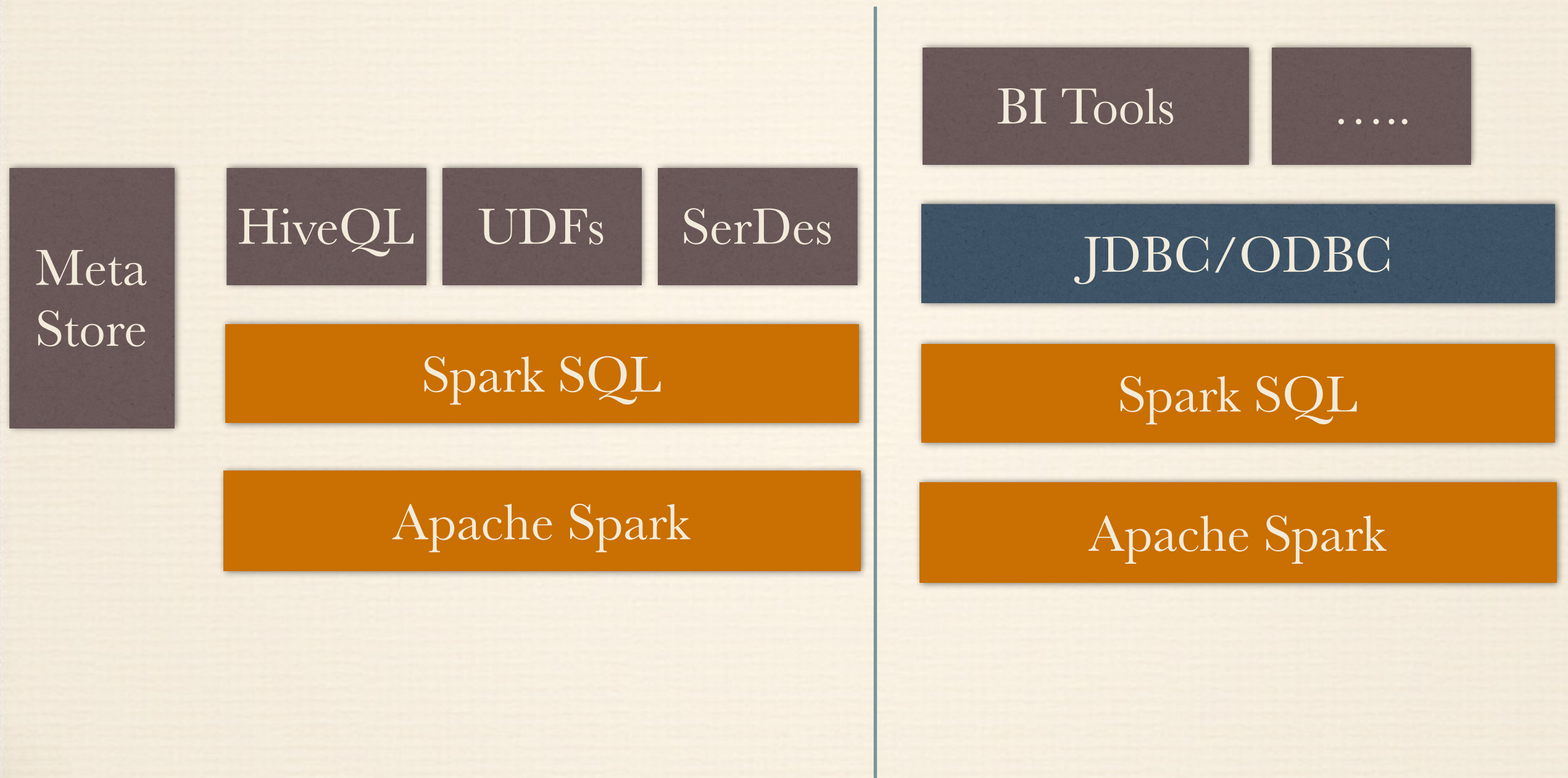
Edge Table

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

GraphX



Spark SQL



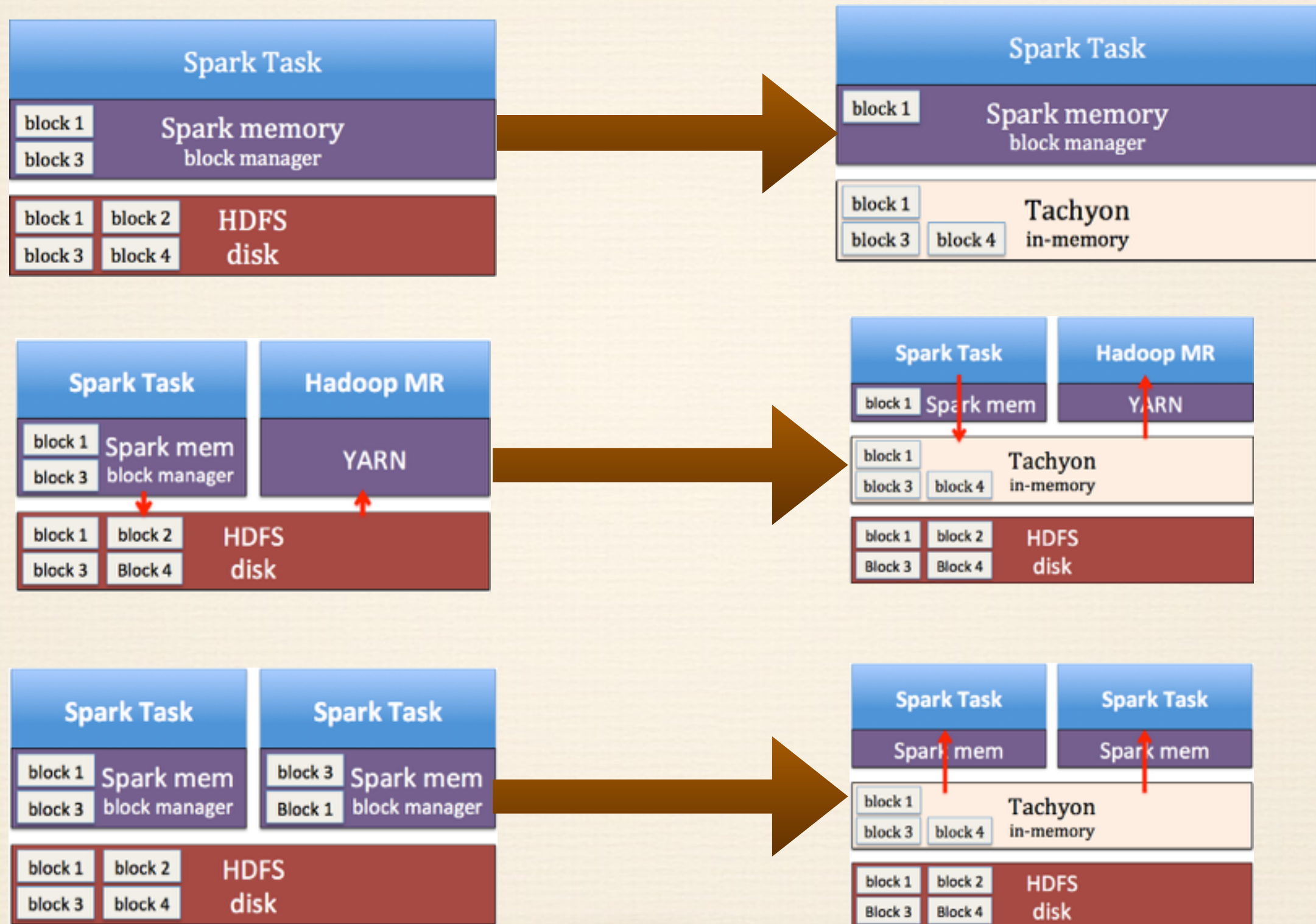
Spark SQL

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

Shark

❖ Mission Completed!!!

Tachyon



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

QA & Thanks

weibo:@CrazyJvm

wechat public account : ChinaScala