中国新一代金融服务平台的开拓者

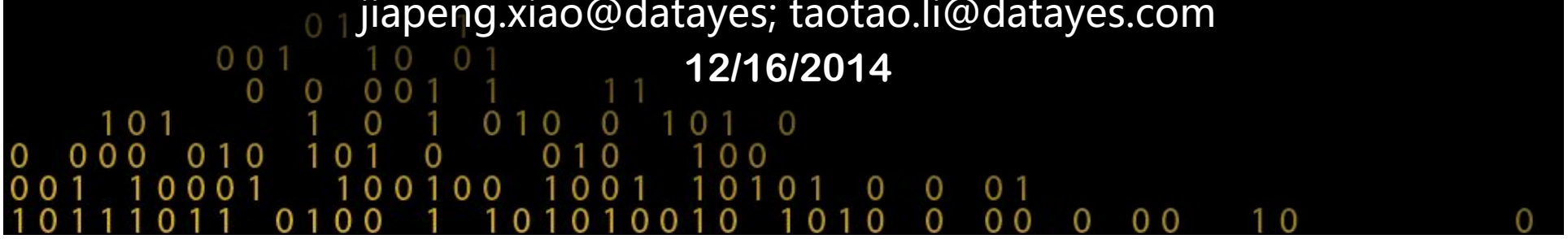
通联数据

Spark Fire

Where there is spark, there is fire

jiapeng.xiao@datayes; taotao.li@datayes.com

12/16/2014



Content

通联数据

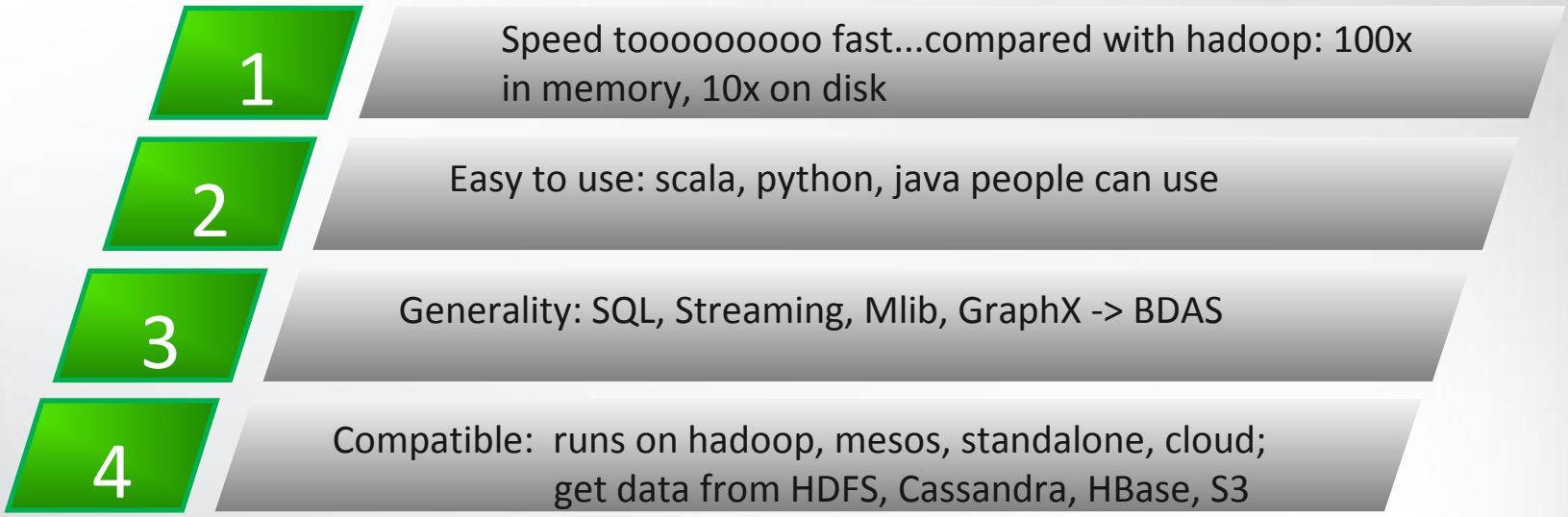
Just the beginning of BDAS

2014

2013

2009~2010

The Birth and Power of Spark：想不爱，太难！



通联数据

The Birth and Power of Spark：天下武功，唯快不破！

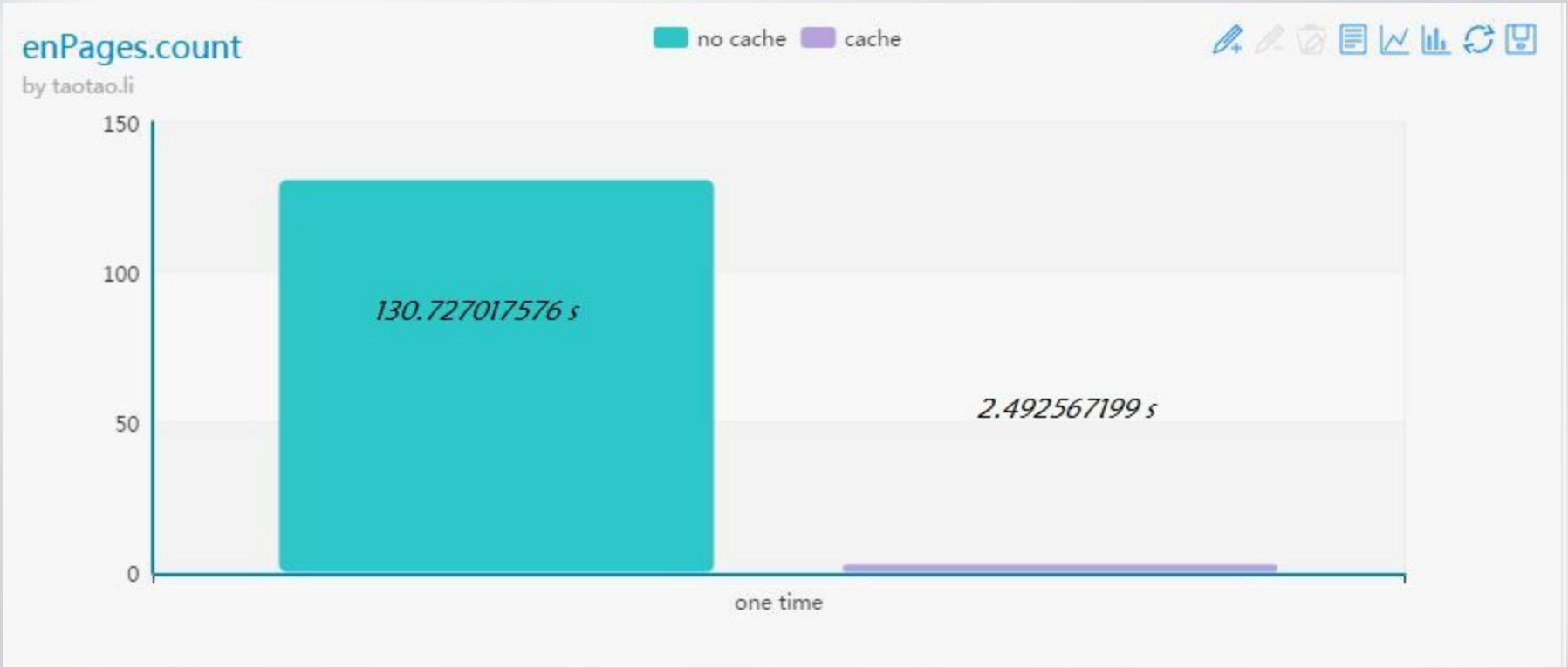
配置：EC2

master node: *1
slave node: *3 [each 2 cpus, 15.7GB mem]

任务：
20GB wikipedia 流量数据，计算英文条目数量。
本例中所有条目数量： 329,641,466
其中有英文条目数量： 122,352,588

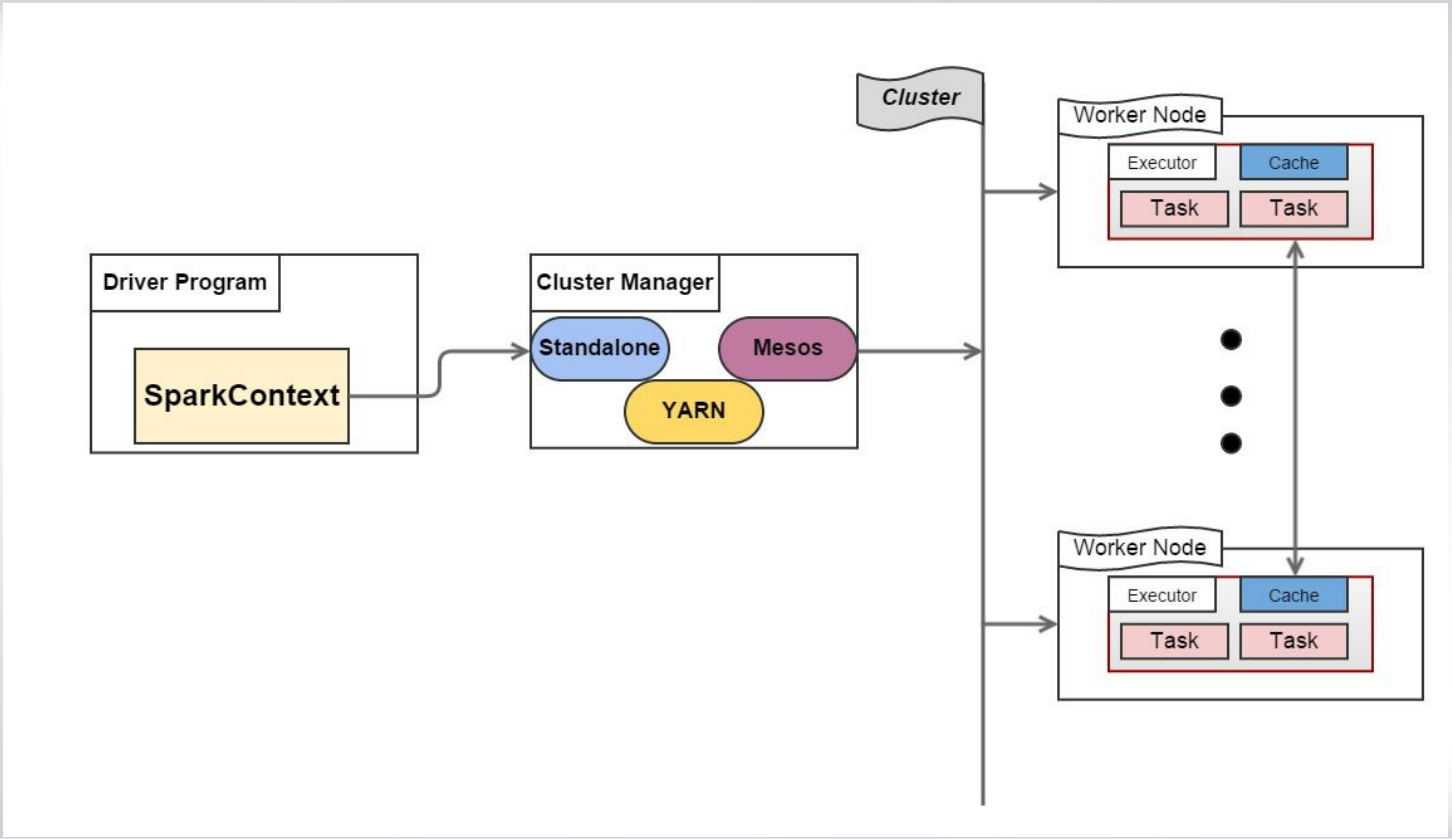
比较：
On disk: 90-150s;
In mem: 2-3s;

The Birth and Power of Spark：天下武功，唯快不破！



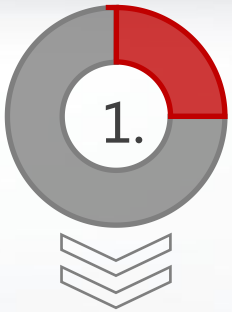
通联数据

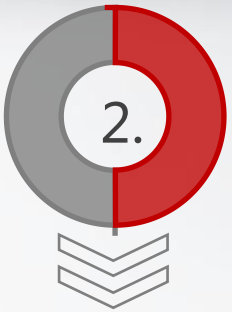
Spark Cluster Configure



通联数据

Spark Cluster Configure

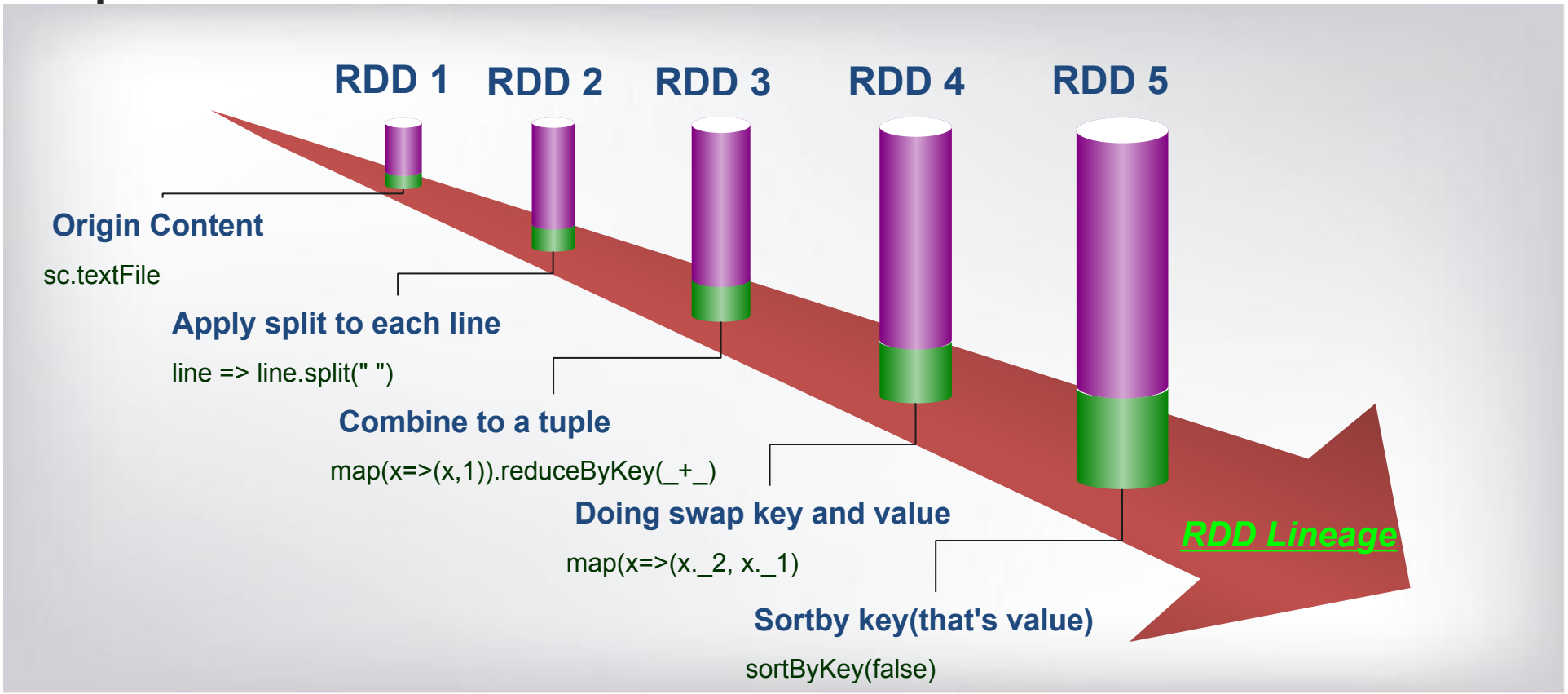




Driver Program	Cluster Manager	Nodes
应用程序逻辑 .	给应用分配、管理计算资源。一旦你的Driver Node连接上Cluster Manager，spark将会处理下面三件事: 1. 连接计算节点，这些节点是用来运行你的应用程序和存储应用数据的； 2. 把你在Driver Program里定义的应用逻辑发送到计算节点上； 3. 在每个节点上分配计算任务；	你的所有计算资源 .

通联数据

Spark Task Demo : word count



通联数据

Spark Task Demo : my resource

Executors (10)

Memory: 0.0 B Used (2.6 GB Total)
Disk: 0.0 B Used

Executor ID	Address ▲	RDD Blocks	Memory Used	Disk Used	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time	Input	Shuffle Read	Shuffle Write
4	sh-demo-hadoop-10:40603	0	0.0 B / 265.0 MB	0.0 B	0	0	3	3	4.4 s	1536.2 KB	157.7 KB	344.2 KB
1	sh-demo-hadoop-09:60154	0	0.0 B / 265.0 MB	0.0 B	0	0	2	2	1.9 s	1536.2 KB	0.0 B	312.5 KB
3	sh-demo-hadoop-08:49379	0	0.0 B / 265.0 MB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B
5	sh-demo-hadoop-07:60848	0	0.0 B / 265.0 MB	0.0 B	0	0	2	2	3.1 s	0.0 B	516.9 KB	0.0 B
6	sh-demo-hadoop-06:34556	0	0.0 B / 265.0 MB	0.0 B	0	0	1	1	1.7 s	0.0 B	329.1 KB	0.0 B
8	sh-demo-hadoop-05:53285	0	0.0 B / 265.0 MB	0.0 B	0	0	3	3	3.3 s	0.0 B	985.8 KB	263.4 KB
0	sh-demo-hadoop-04:49493	0	0.0 B / 265.0 MB	0.0 B	0	0	1	1	1.9 s	0.0 B	327.5 KB	262.1 KB
7	sh-demo-hadoop-03:57154	0	0.0 B / 265.0 MB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B
2	sh-demo-hadoop-02:58348	0	0.0 B / 265.0 MB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B
<driver>	10.20.70.80:56084	0	0.0 B / 265.0 MB	0.0 B	0	0	0	0	0 ms	0.0 B	0.0 B	0.0 B

通联数据

Spark Task Demo : cluster resource

Workers

Id	Address	State	Cores	Memory
worker-20141210113956-sh-demo-hadoop-02-33140	sh-demo-hadoop-02:33140	ALIVE	32 (2 Used)	61.9 GB (512.0 MB Used)
worker-20141210113956-sh-demo-hadoop-09-42053	sh-demo-hadoop-09:42053	ALIVE	32 (2 Used)	61.9 GB (512.0 MB Used)
worker-20141210113958-sh-demo-hadoop-03-45309	sh-demo-hadoop-03:45309	ALIVE	32 (1 Used)	61.9 GB (512.0 MB Used)
worker-20141210113958-sh-demo-hadoop-04-33111	sh-demo-hadoop-04:33111	ALIVE	32 (2 Used)	61.9 GB (512.0 MB Used)
worker-20141210113958-sh-demo-hadoop-05-58058	sh-demo-hadoop-05:58058	ALIVE	32 (1 Used)	61.9 GB (512.0 MB Used)
worker-20141210113958-sh-demo-hadoop-06-46985	sh-demo-hadoop-06:46985	ALIVE	32 (2 Used)	61.9 GB (512.0 MB Used)
worker-20141210113958-sh-demo-hadoop-07-53244	sh-demo-hadoop-07:53244	ALIVE	32 (2 Used)	61.9 GB (512.0 MB Used)
worker-20141210113958-sh-demo-hadoop-08-48448	sh-demo-hadoop-08:48448	ALIVE	32 (2 Used)	61.9 GB (512.0 MB Used)
worker-20141210113958-sh-demo-hadoop-10-39476	sh-demo-hadoop-10:39476	ALIVE	32 (2 Used)	61.9 GB (512.0 MB Used)

每个应用的资源可以申请，现在假设每个应用都只需要2.6GB，现在的集群规模可以容纳61.9*9/2.6 = 214个应用。按每个用户只运行一个app的话，可以供214个用户。

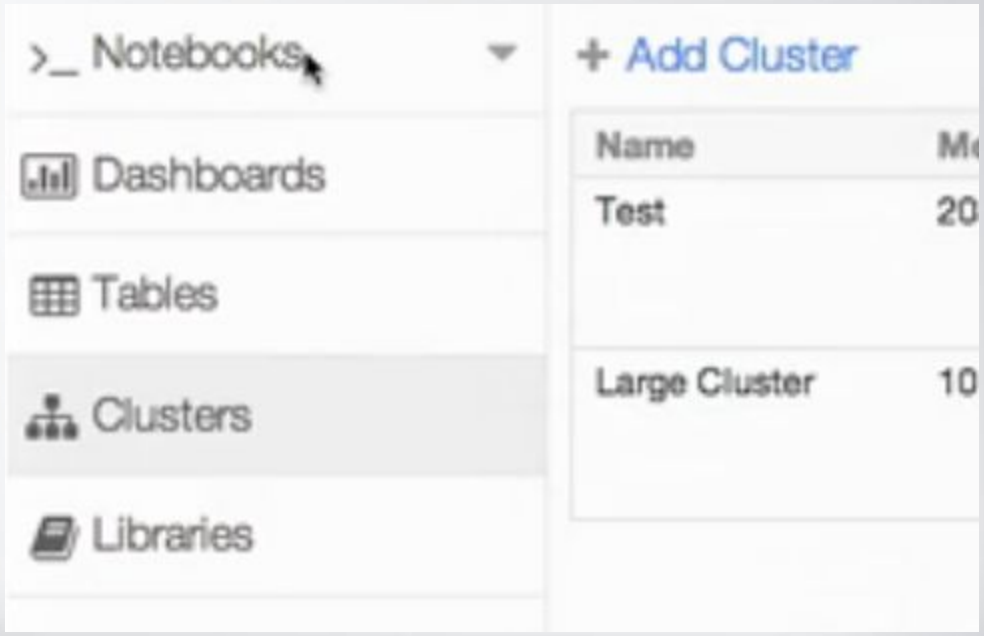
参考：EC2免费版配置：6 CPUs，45 GB MEM；

通联数据

Deep Into Databricks : components

components

01. **Notebooks**
02. **Dashboards**
03. **Tables**
04. **Clusters**
05. **Libraries**



通联数据

Deep Into Databricks : Feature

Notebooks

- 选语言
- 选集群
- 分类解析
  - %md
  - %sql
  - default sql

Dashboards

- 协作/分享
- 实时更新

Tables

- 上传数据
- 每一个数据文件视为一个完整的数据库/数据表，可在notebooks里面用sql来操作

Deep Into Databricks : Feature

Clusters

- 创建/修改集群配置
- S3/EC2

Libraries

- 上传个人代码库
- 可在notebooks里导入

通联数据

通联数据

谢　　谢