DTCC

2017第八届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2017

Spark SQL优化与硬件选型

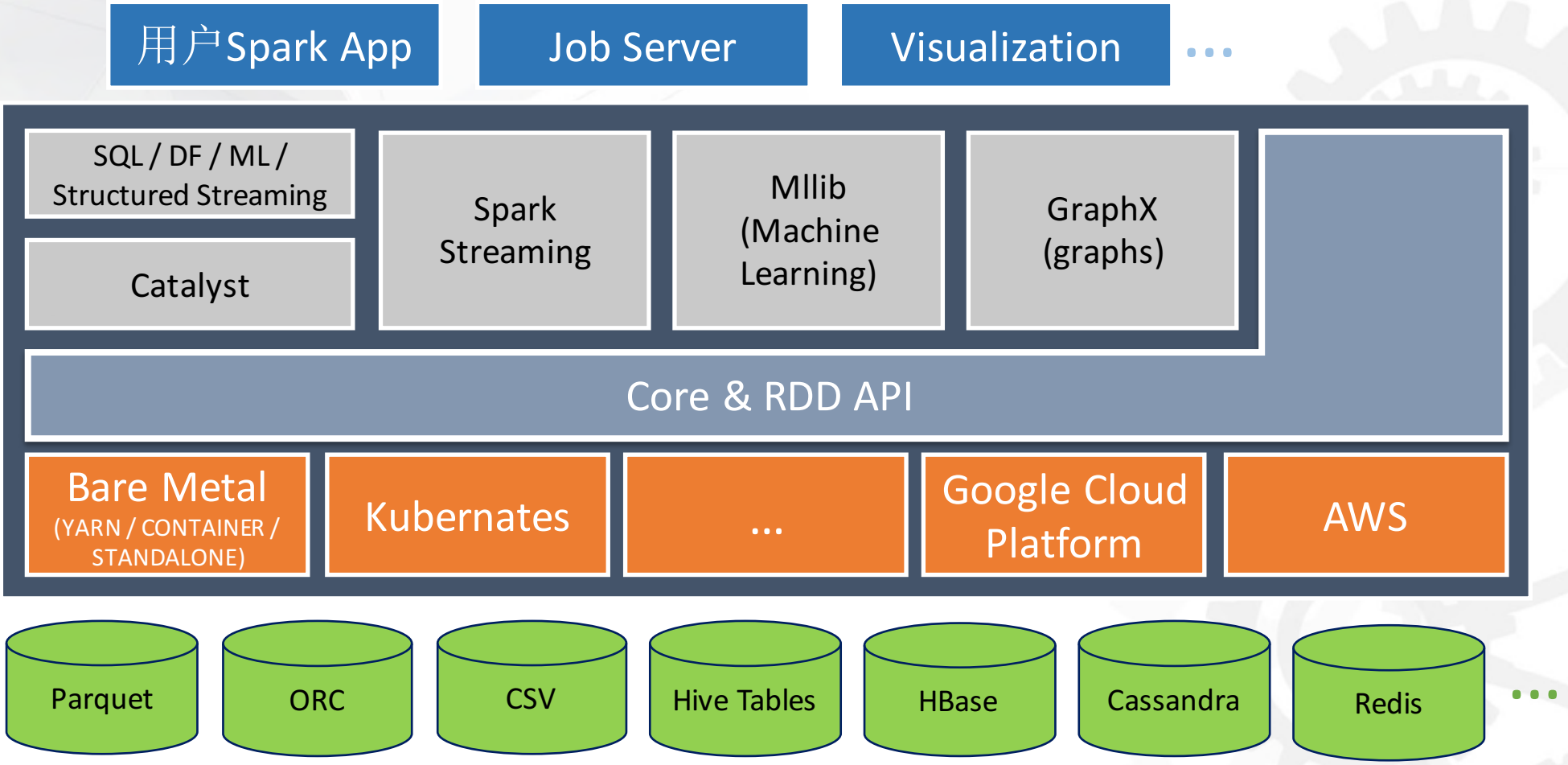
程浩

主要内容

- Spark概要简介
- Spark SQL基准测试
- 性能比较分析
- 推荐硬件选型
- 下一步？

Spark概要简介

Spark软件栈

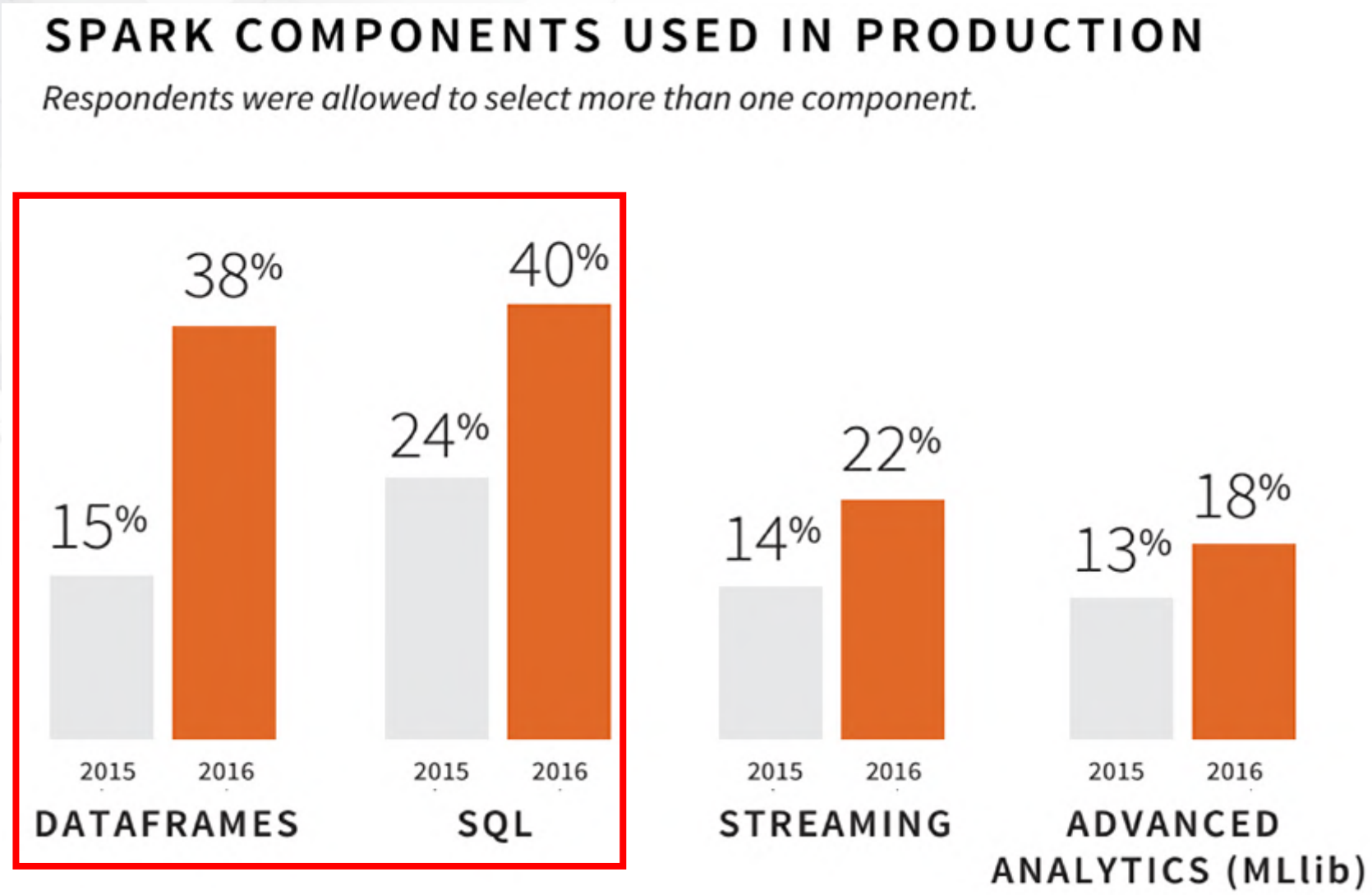

用户Spark App
Job Server
Visualization
...
SQL / DF / ML / Structured Streaming
Catalyst
Spark Streaming
Mllib (Machine Learning)
GraphX (graphs)
Core & RDD API
Bare Metal (YARN / CONTAINER / STANDALONE)
Kubernates
...
Google Cloud Platform
AWS
Parquet
ORC
CSV
Hive Tables
HBase
Cassandra
Redis
...

为什么选Spark？而不是MR？

- 简单易用
  - 同一个软件栈搞定一切(Streaming, SQL, GraphX, Machine Learning, BigDL)
  - 多种语言支持(SQL, Java, Scala, Python, R)
  - Declarative (DataSet / DataFrames / RDD) API VS. Imperative API
  - 活跃的数据源连接器开源组件(Hbase, Cassandra, Redis, ElasticSearch, MongoDB ...)

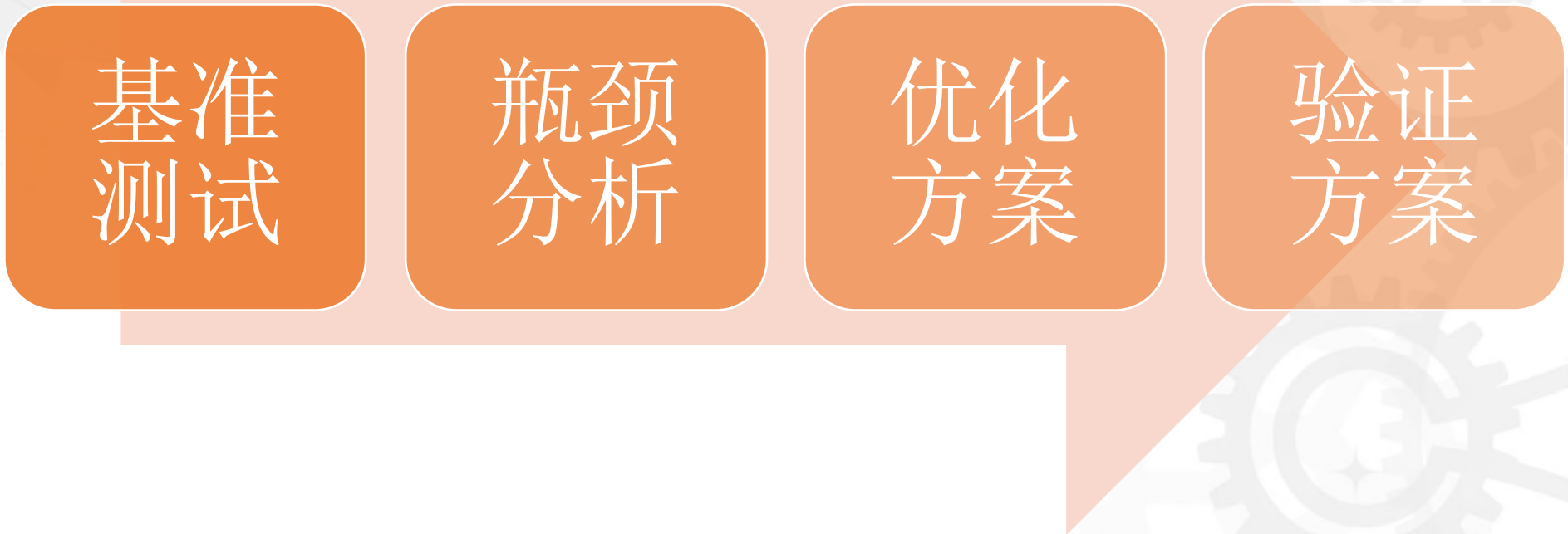
- 更快的处理引擎
  - DAG Based任务调度机制
  - 缓存API与内存计算
  - 开放式的Catalyst执行计划优化器&Tungsten系列优化执行加速

Spark生态圈组件使用比例



https://databricks.com/blog/2016/09/27/spark-survey-2016-released.html

性能优化一般步骤



Spark SQL性能基准测试

实验环境和测试集

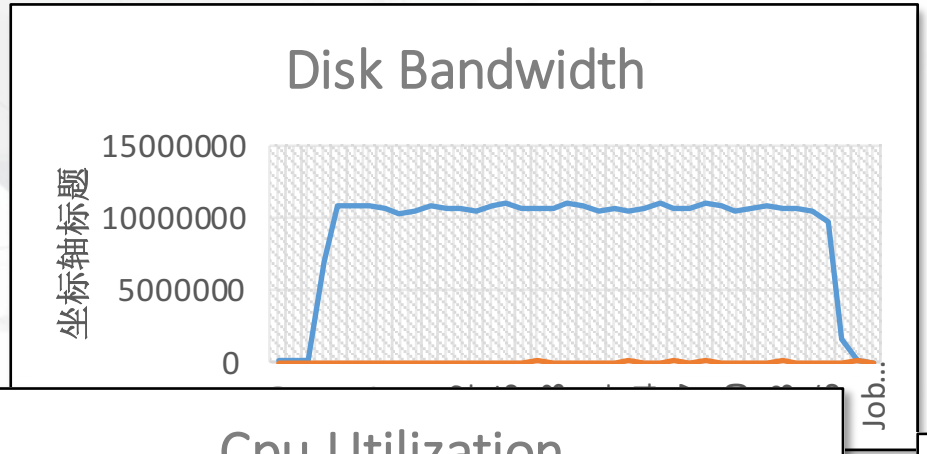
Nodes	Master	Slave
Roles	Hadoop Name Node, Spark Master	Hadoop Data Node, Spark Slaves
Services	Name Node, Resource Manager	Data Node, Node Manager
Numbers	1	7
Processer	Intel Xeon E5-2650 v3 (HSW) / Intel Xeon E5-2680 v4 (BDW) (Dual Socket / node)	
Memory	256GB	256GB
Storage	OS Disk: 480GB SSD	OS Disk: 480GB SSD Data Disk: 1TB SATA HDD x 8 / Data Disk: Intel S3520 SSD x 8 / Data Disk: Intel P3600 SSD x 3
Network	10Gb	10Gb

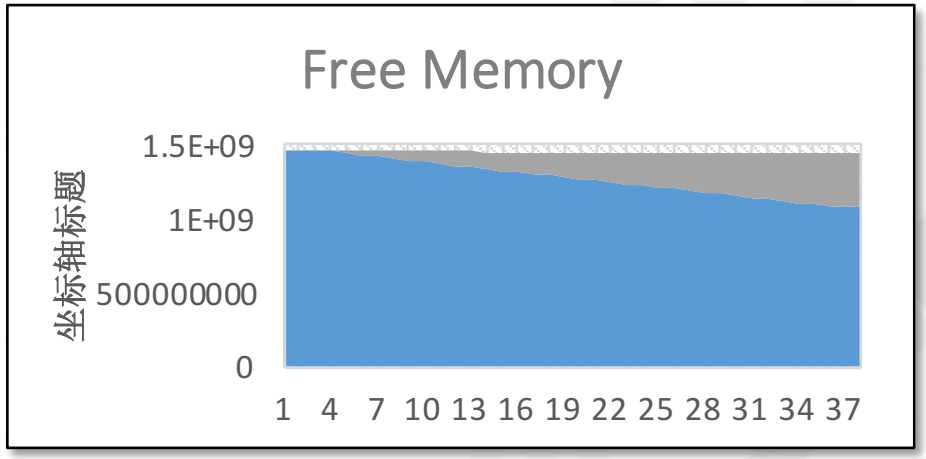
Workload (TPC-DS)	
Queries	19,42,43,52,55,63,68,72,98
Data Scale (Raw Data)	10 TB
Data Format	Parquet
Compression Codec	Snappy
Data Size	~3TB

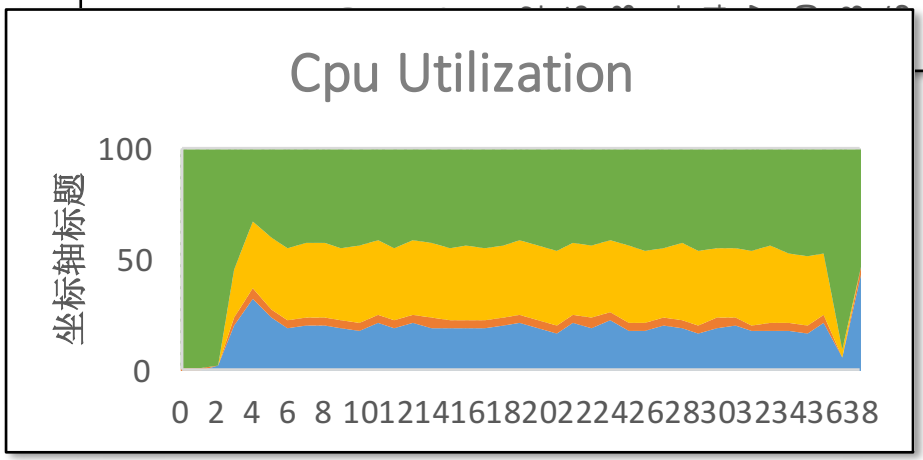
Hadoop/Spark Configuration	
Hadoop version	2.7.3
Spark version	2.1.0
Executor memory	25~40 GB
Executor Cores	8 – 10 / executors
Executor Number	5 / nodes
Spark Mode	yarn-client
JDK Version	1.8.0_112
memory.Overhead	10% Executor Memory
Shuffle Partition #	200
Broadcast threshold	30MB
broadcastTimeout	3600 sec
GC	Parallel GC

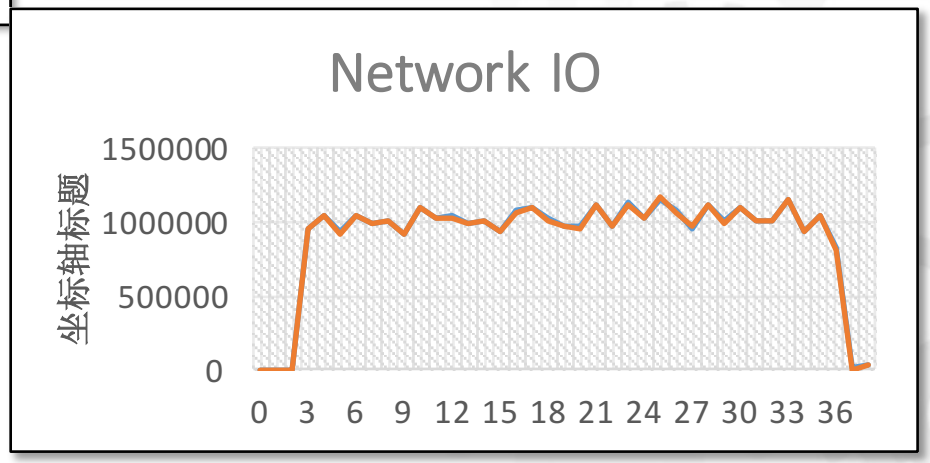
Intel Performance Analysis Tool

Performance **A**nalysis **T**ool(PAT) 适用于与在分布式环境下收集系统资源信息，包括CPU、磁盘、网络、内存等，并以图形化的形式展现出来。







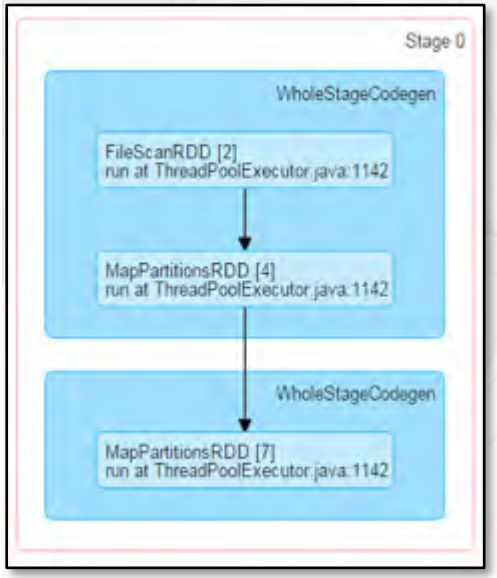


PAT: https://github.com/intel-hadoop/PAT

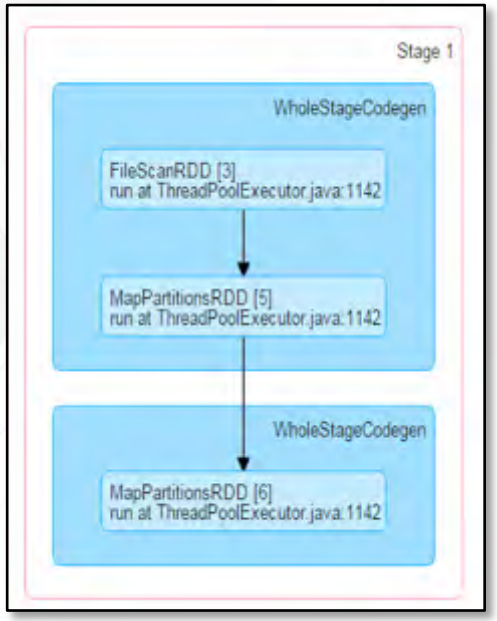
TPC-DS Q42 举例

```
SELECT dt.d_year, item.i_category_id, item.i_category, sum(ss_ext_sales_price)
FROM date_dim dt, store_sales, item
WHERE dt.d_date_sk = store_sales.ss_sold_date_sk
      AND store_sales.ss_item_sk  = item.i_item_sk
      AND item.i_manager_id = 1
      AND dt.d_moy=11
      AND dt.d_year=2000
GROUP BY      dt.d_year
            ,item.i_category_id
            ,item.i_category
ORDER BY   sum(ss_ext_sales_price)  DESC , dt.d_year
            ,item.i_category_id
            ,item.i_category
```

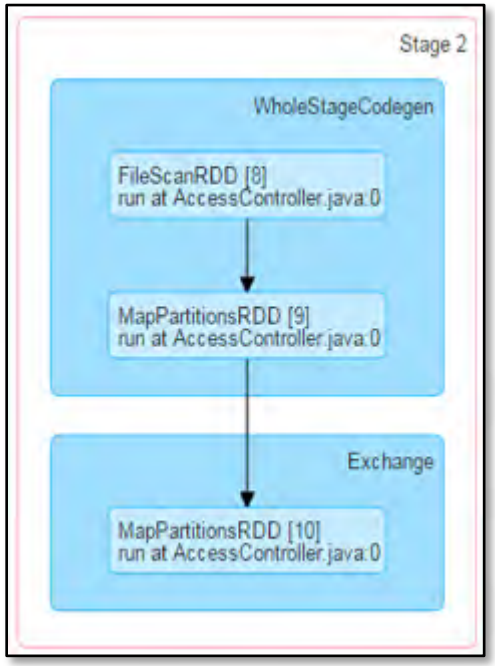
Spark SQL DAG for Q42



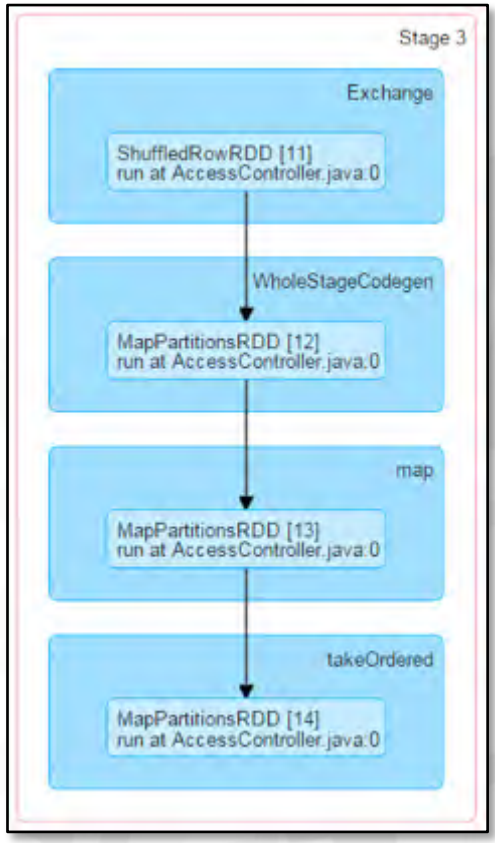
Input Data Size: 291.5KB
Duration: 0.2sec



Input Data Size: 2.3MB
Duration: 0.2sec



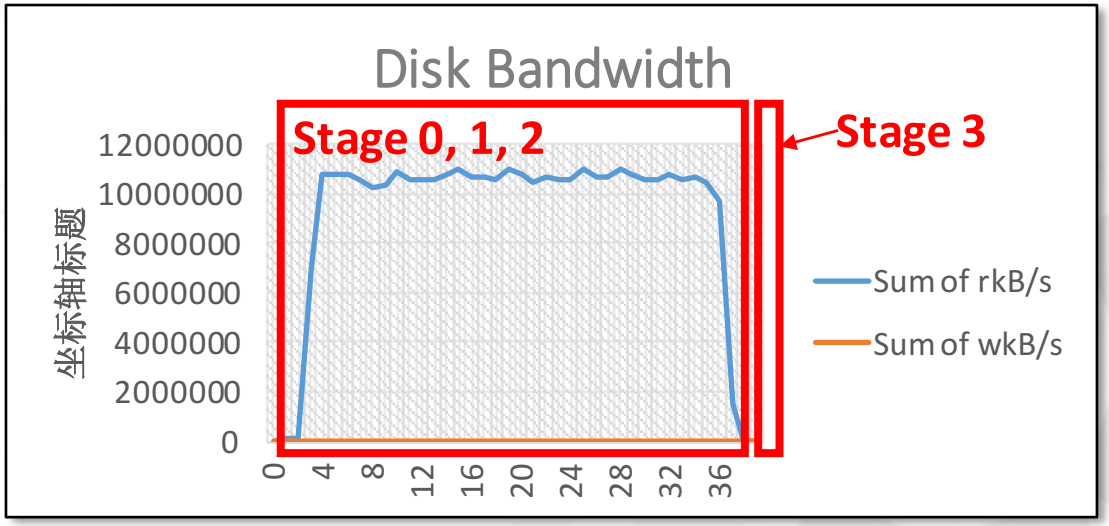
Input Data Size: 221.4GB
Duration: 35s
Shuffle Write Size: 8.8MB

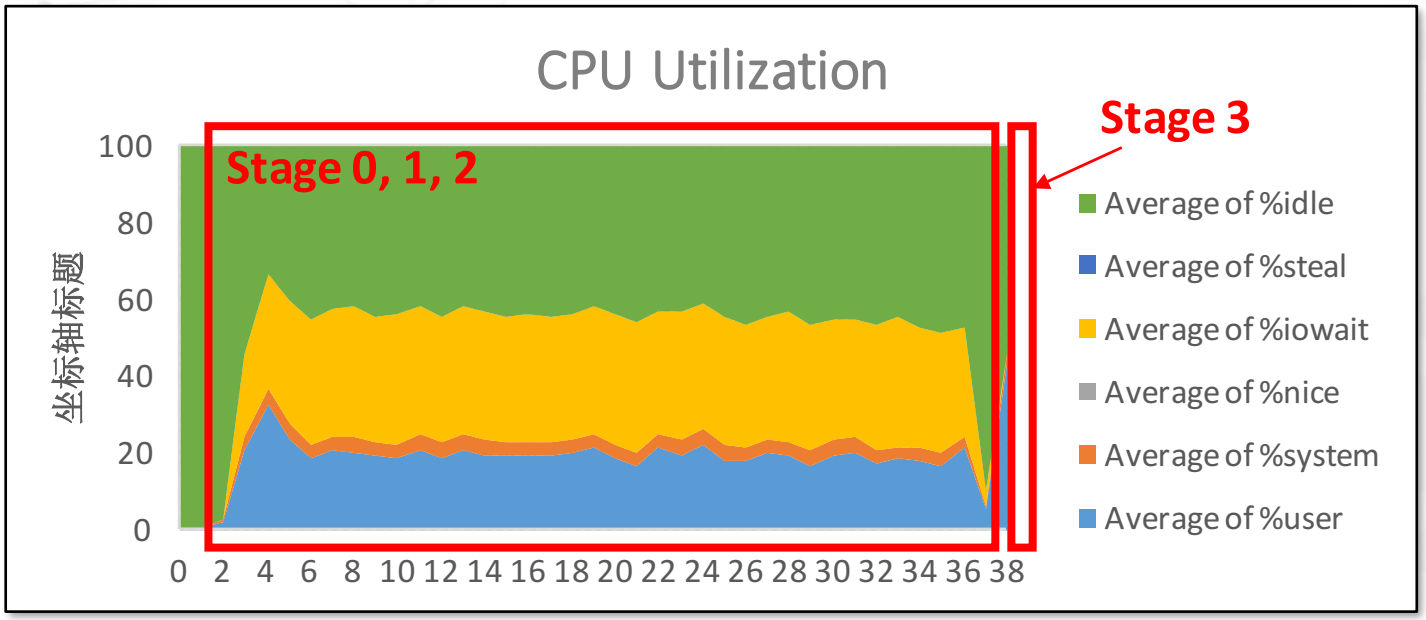


Input Data Size: n/a
Duration: 1s
Shuffle Read Size: 8.8MB

Q42的CPU & Disk使用情况

- 1 Master + 7 Workers
- Spark 2.1 on YARN
- Total Data Size = 10TB
- Use Intel S3520 1.6 TB SATA SSD * 8
- CPU: Intel HSW E5 2650 v3 (20 vcore)
- Spark Cores: 40 per node (dual sockets)
- Execution Time: 38sec

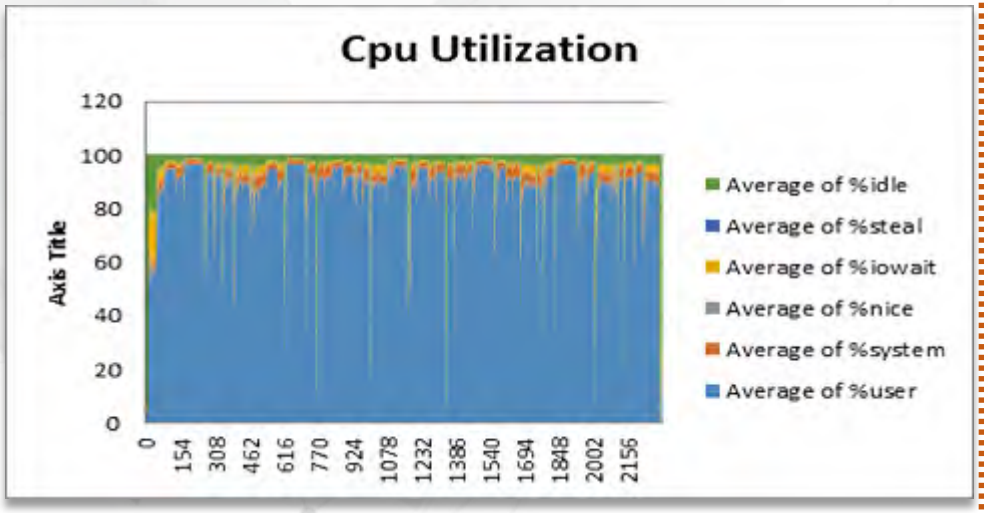


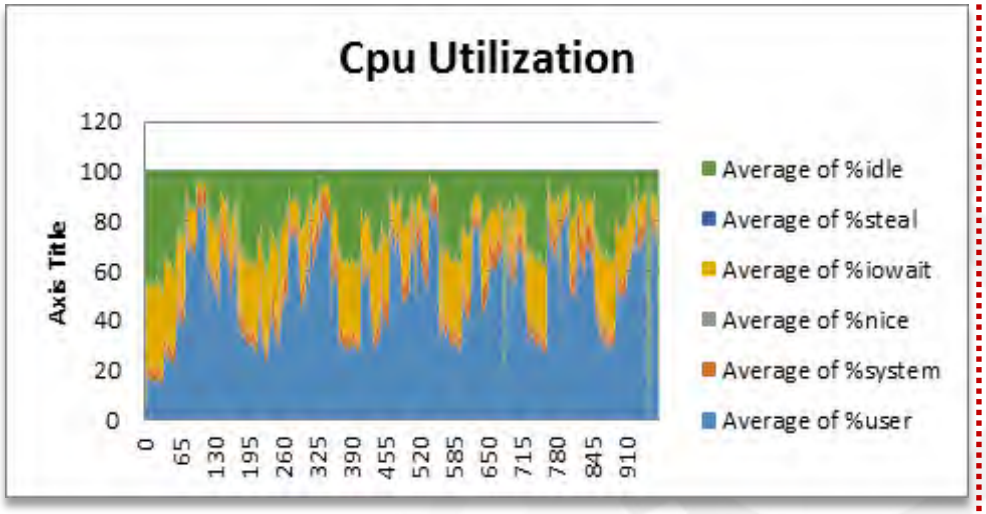


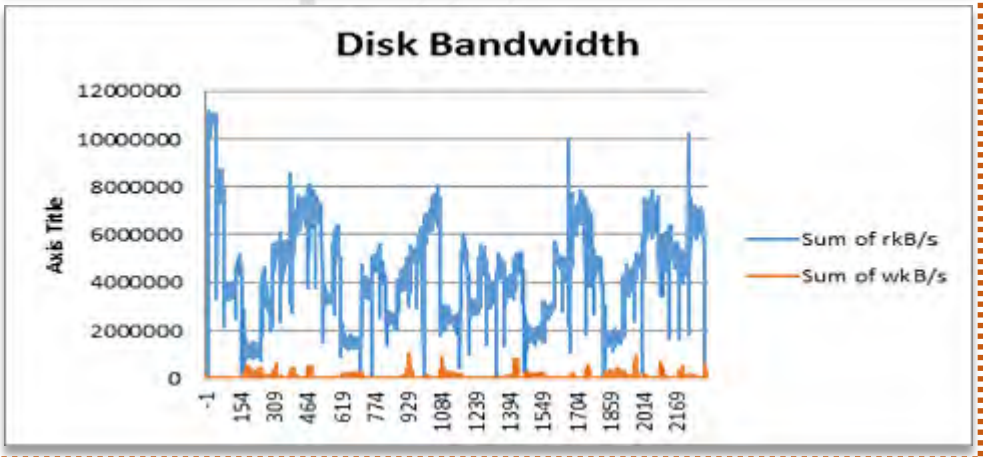
Is an IO intensive workload w/ lots of Disk Read I/O in Stage 0, 1, 2.

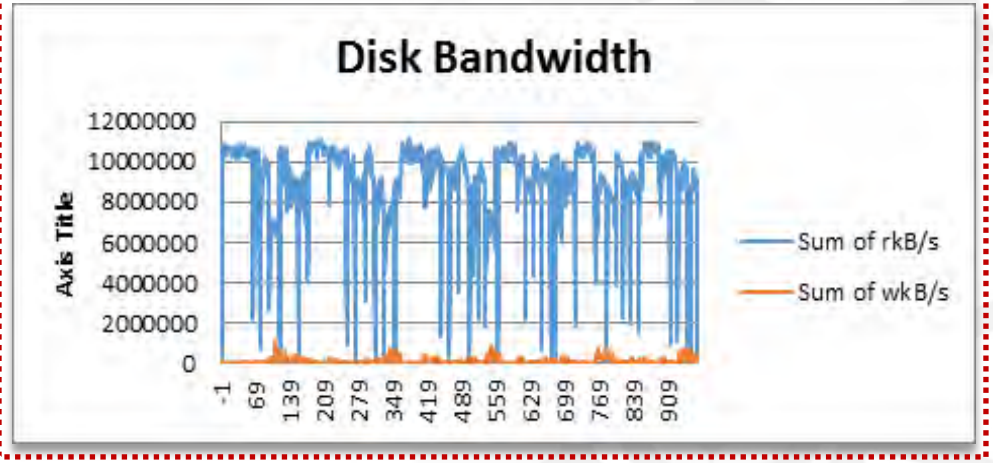
性能比较与分析

#1 SPARK 1.6 VS. SPARK 2.1 (吞吐量测试)





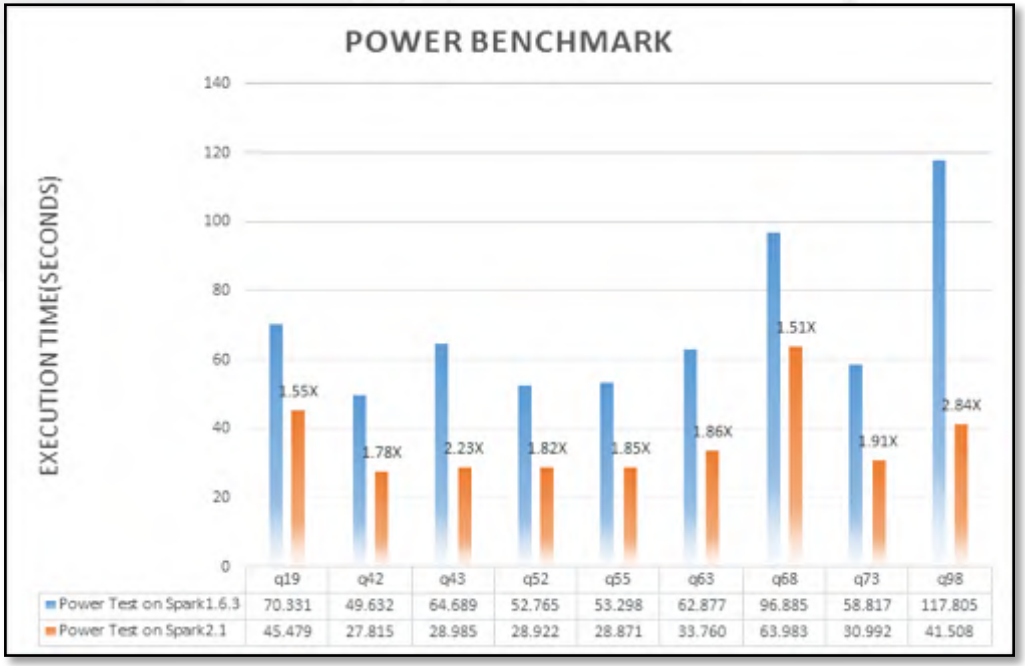


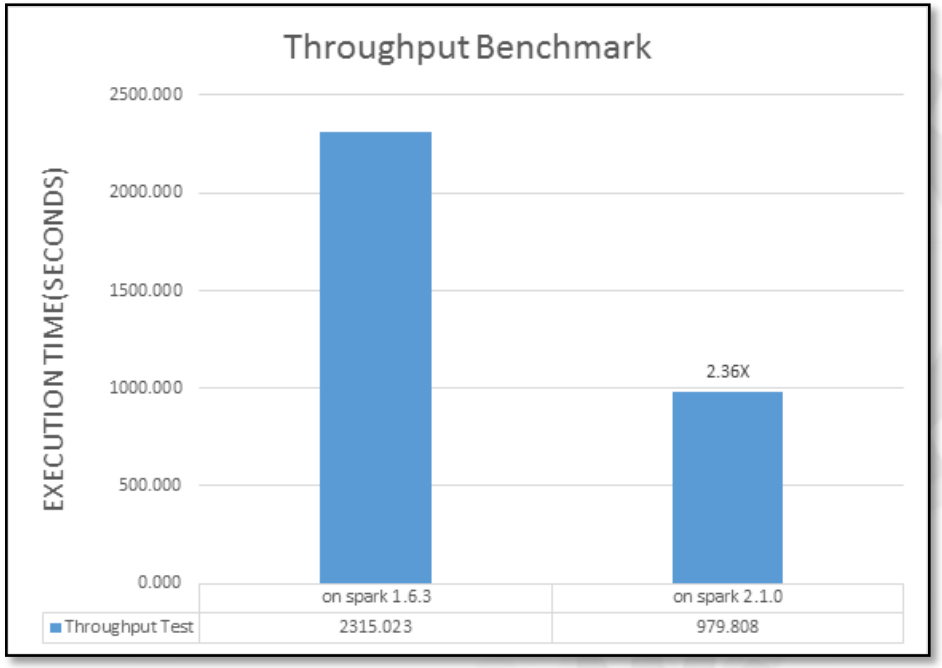


SPARK 1.6

SPARK 2.1

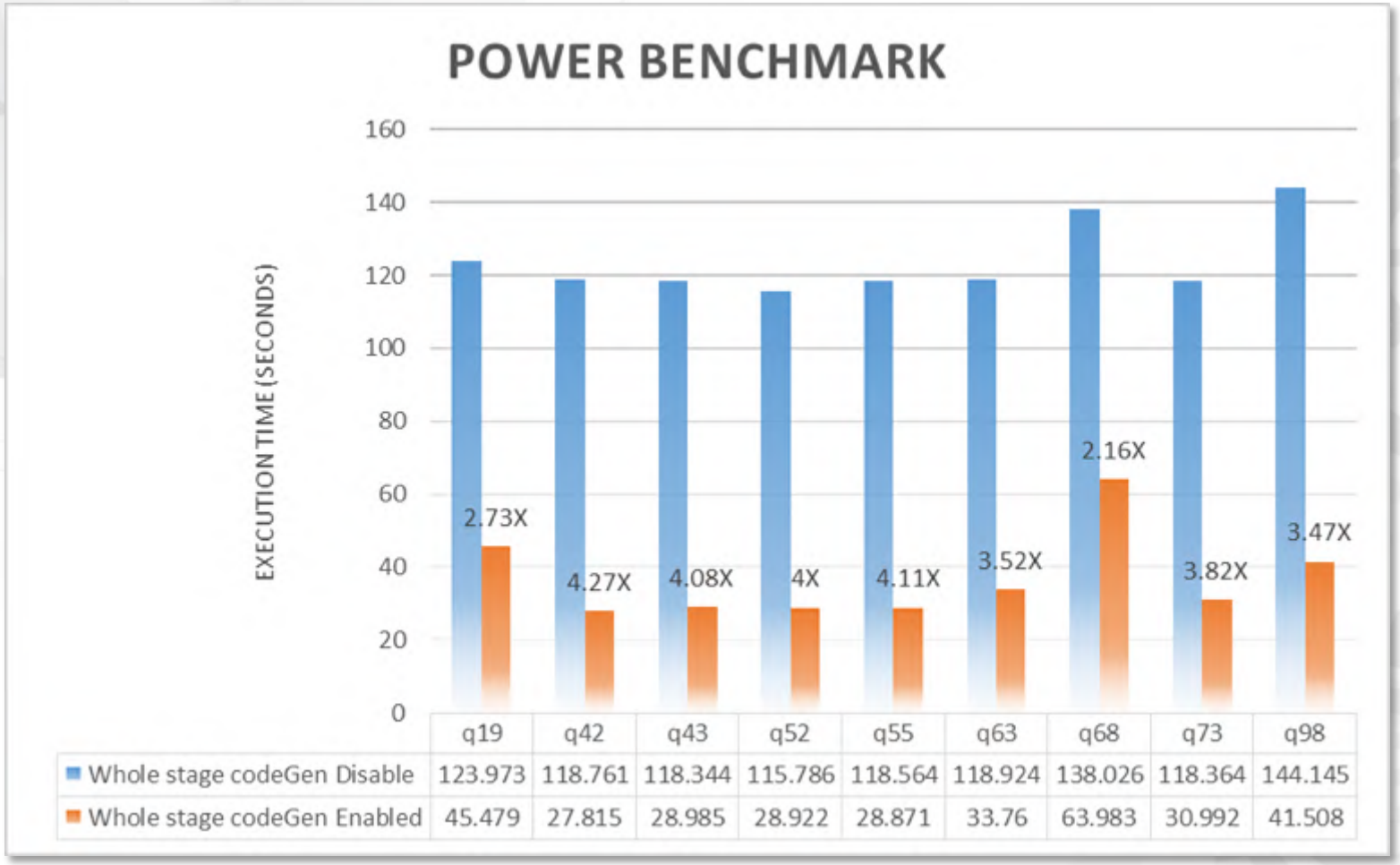
#1 SPARK 1.6 vs SPARK 2.1(吞吐量测试)





Spark 2.1 boost 1.5X~2.8X performance!

#2 Tungsten Phase 2 in Spark 2.x

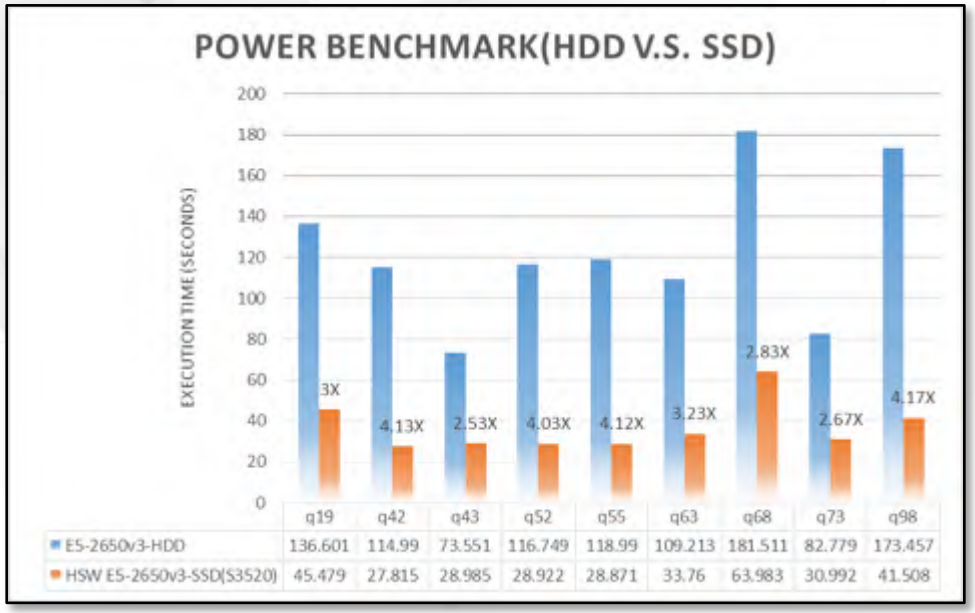


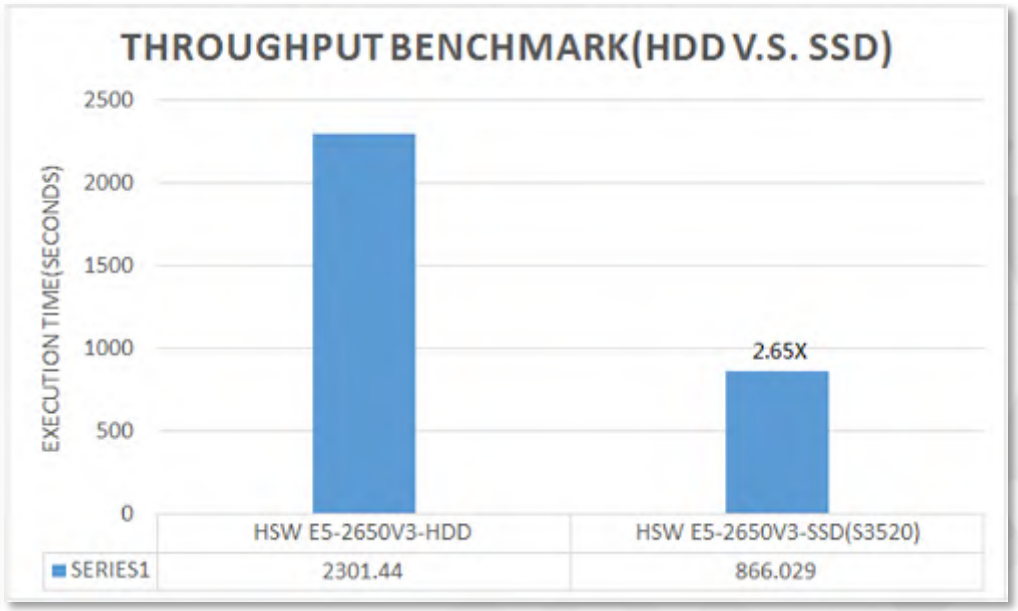
	q19	q42	q43	q52	q55	q63	q68	q73	q98
Whole stage codeGen Disable	123.973	118.761	118.344	115.786	118.564	118.924	138.026	118.364	144.145
Whole stage codeGen Enabled	45.479	27.815	28.985	28.922	28.871	33.76	63.983	30.992	41.508

"Whole Stage Codegen" provides 3X performance boost!

https://databricks.com/blog/2016/05/23/apache-spark-as-a-compiler-joining-a-billion-rows-per-second-on-a-laptop.html

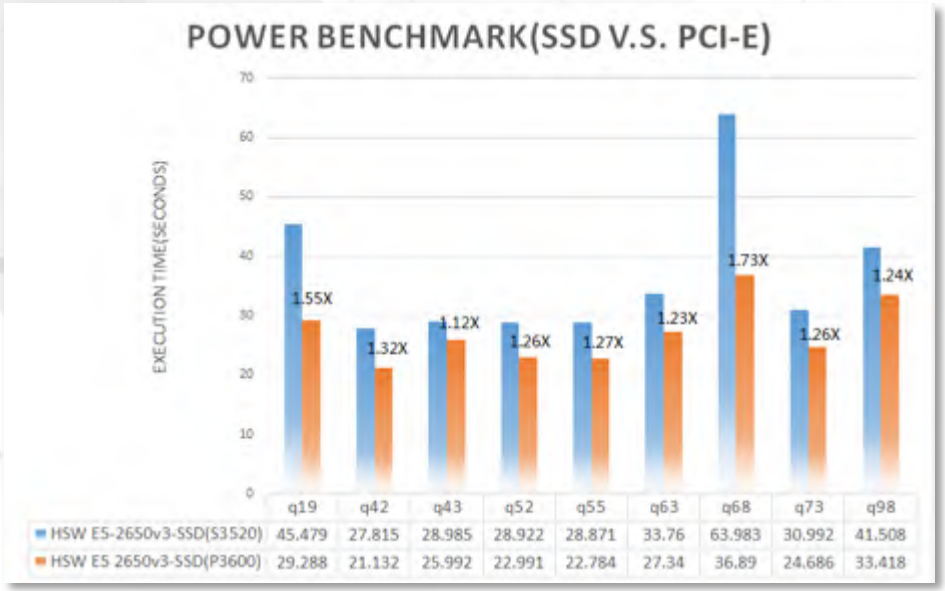
#3 HDD VS. Intel SATA SSD S3520 (2650v3)

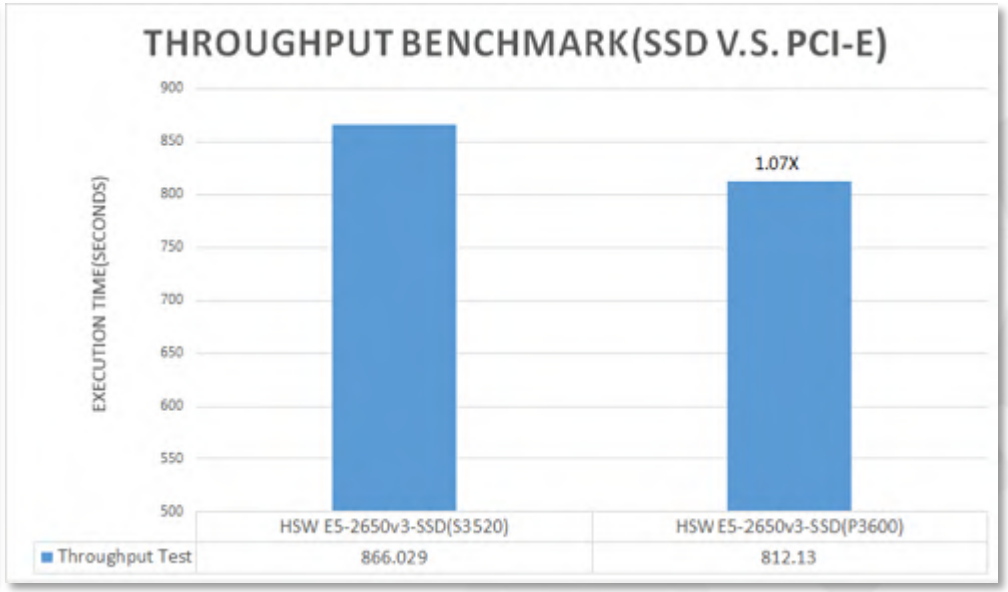




Use SSD brings avg. 2.82X performance improvement!

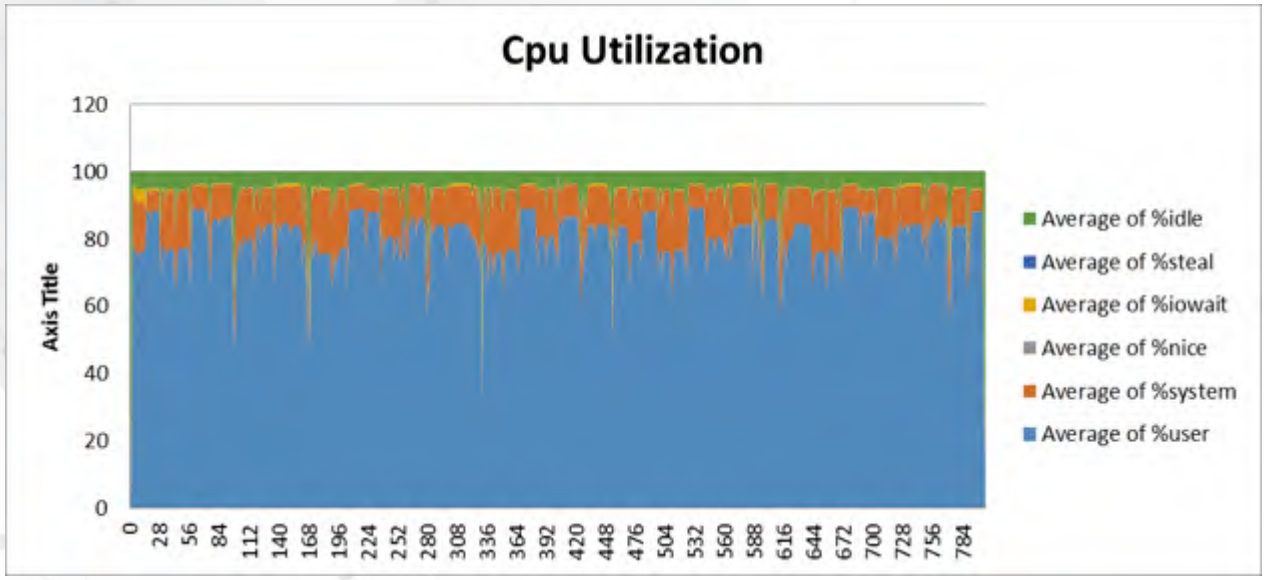
#4 升级磁盘到Intel PCI-E SSD P3600 (2650v3)

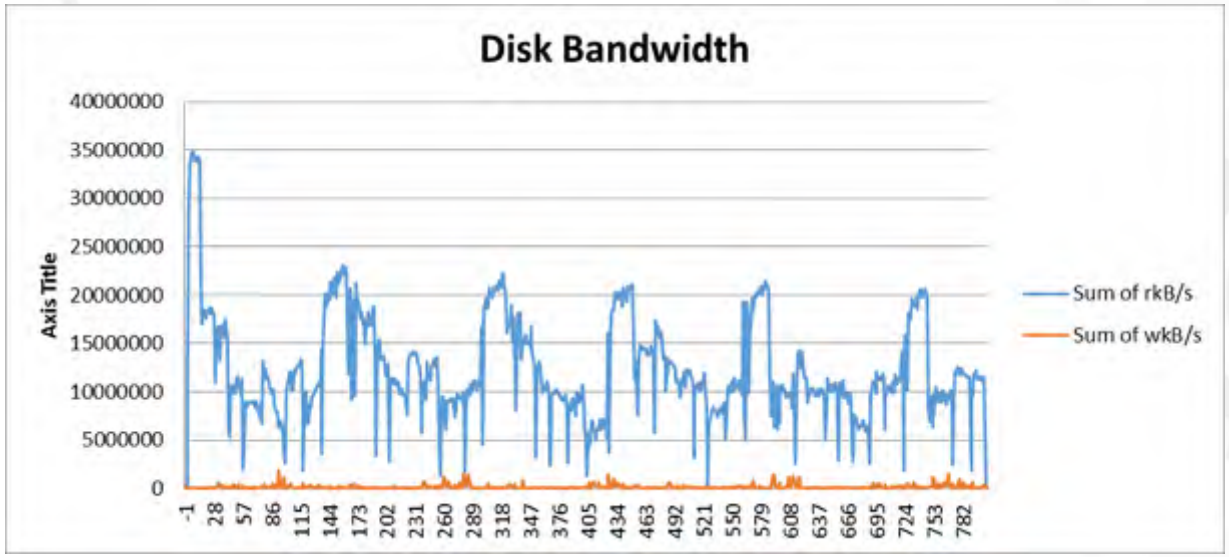




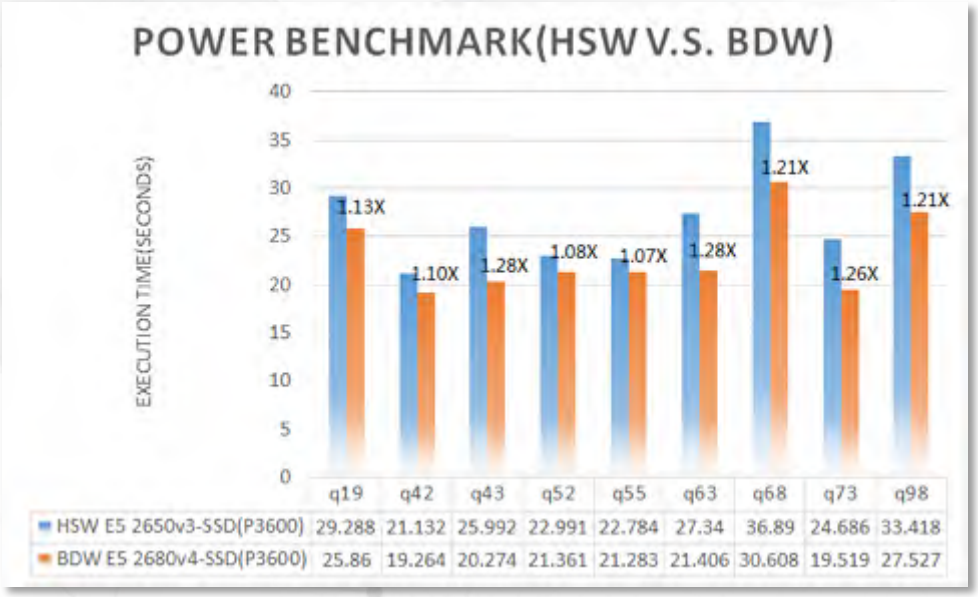
Use PCI-E SSD brings avg. 1.11X performance improvement!

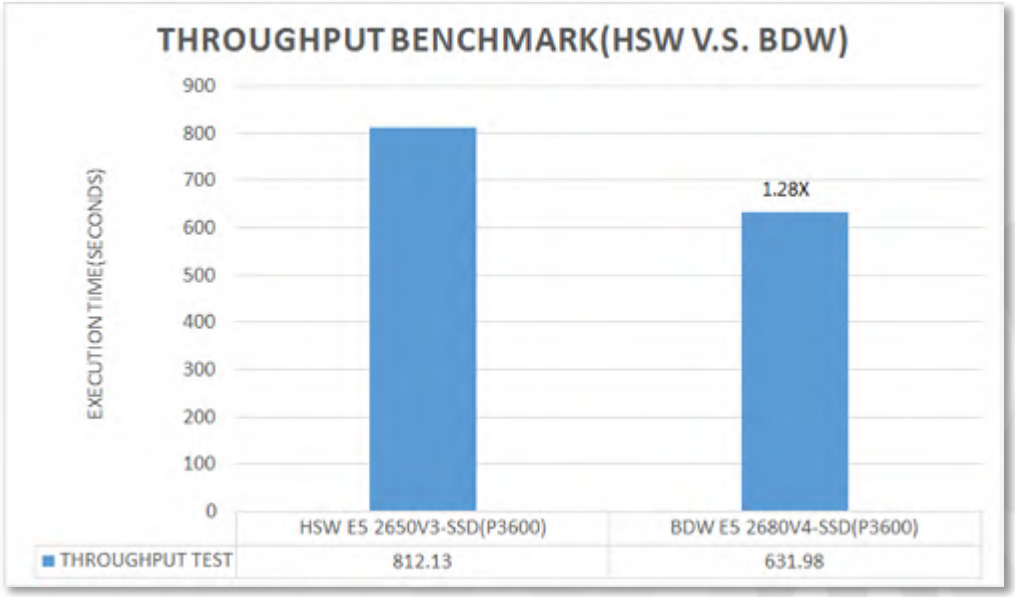
#4 升级磁盘到Intel PCI-E SSD P3600 (2650v3)



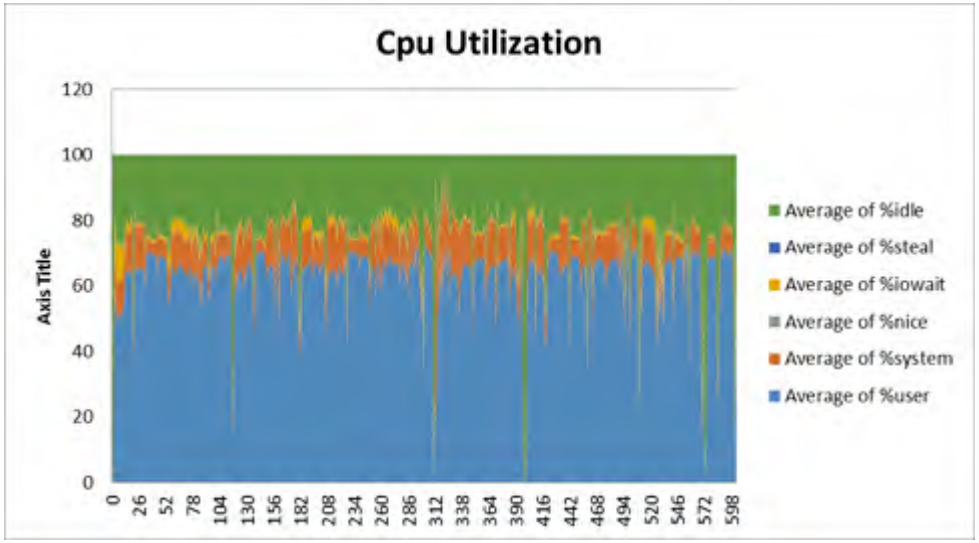


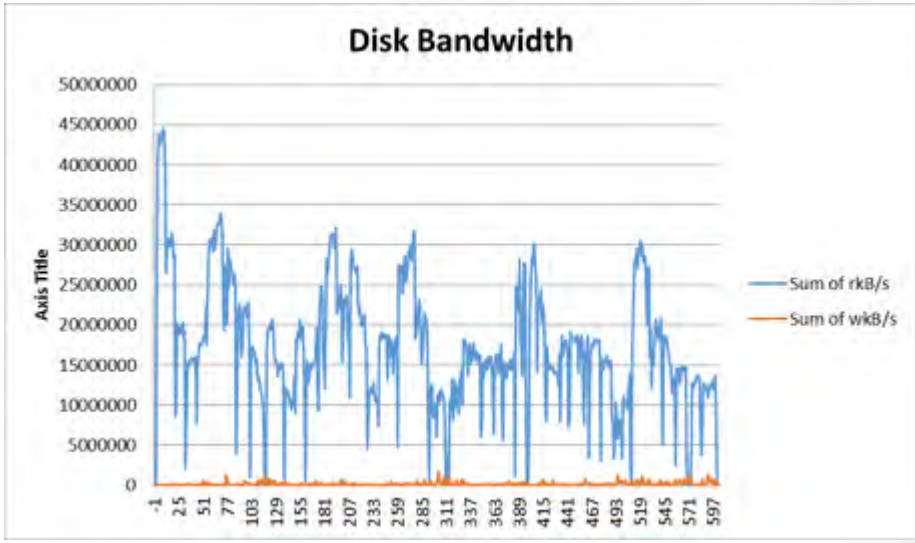
#5 升级CPU至BDW 2680 (P3600 SSD)



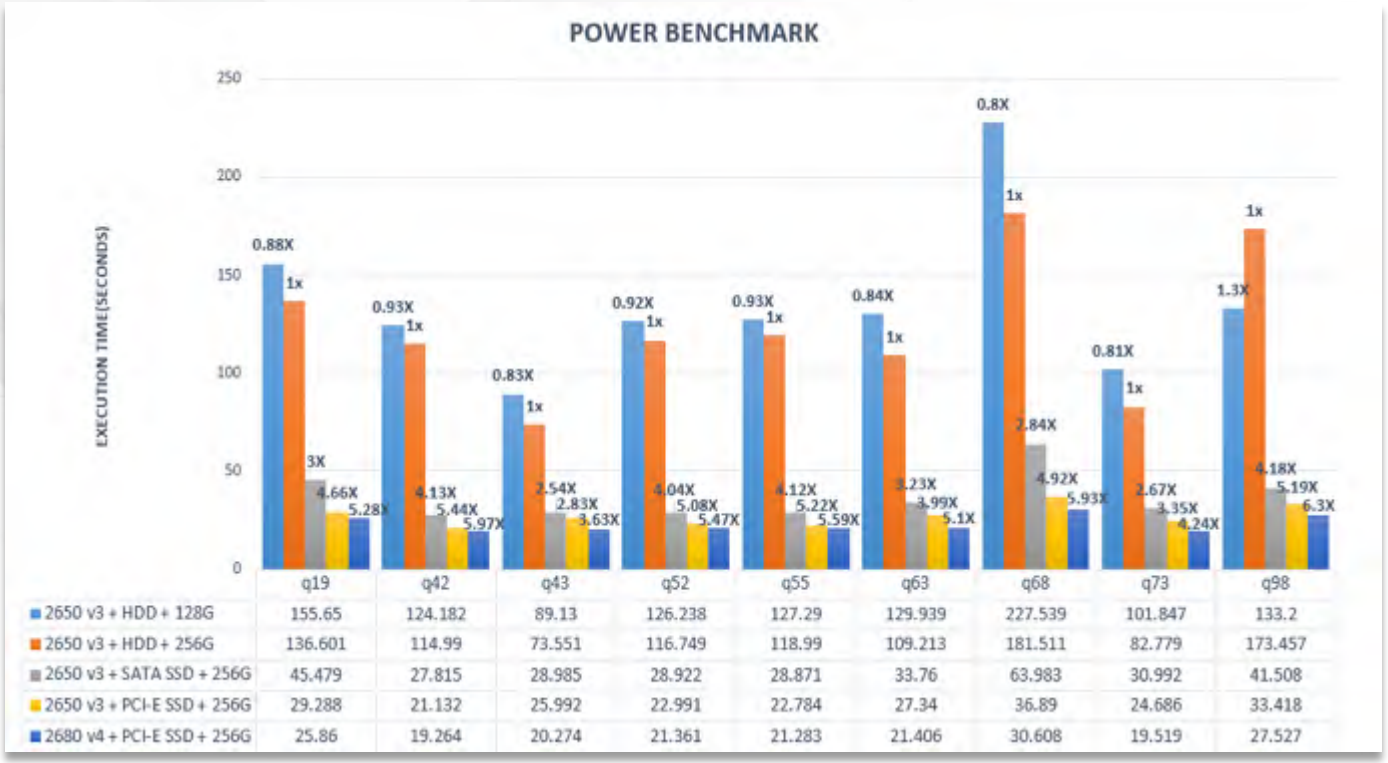


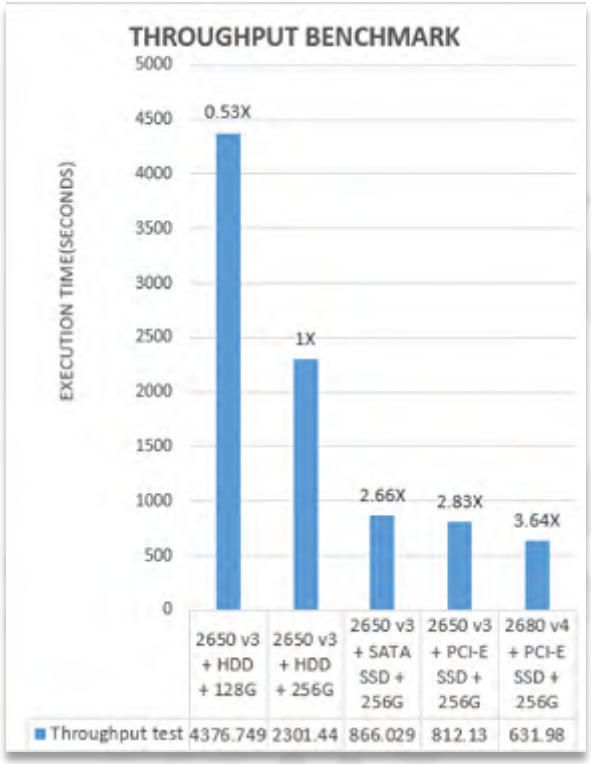
Use BDW(2680 v4) brings avg. 1.23X performance improvement!





完整性能分析对比





- 128 GB 内存可能会极大制约性能
- SATA SSD (Intel S3520) 相对HDD性能提升非常明显
- PCI-E SSD (Intel P3600) 相对HDD性能也会有明显提升，但是只有配合更高端的CPU(E5 2680v4) 才能充分发挥其优势

磁盘TCO 模型(顺序读写)

	Intel HSW(2650v3)	
Disk Types	HDD(1TB SATA)	SATA SSD(s3520)
Numbers of Drives	8	8
Total Capacity	1TB x 8	1.2TB x 8
Performance Gain	1x	~2.65x
Drive Cost	$800	$4224
Power Cost	$421	$68
Cooling Cost	$505	$82
Enclosure Cost	$3943	$3943
Reliability	$1008	$339
Total Cost	$6677	$8656
Cost (per GB)	1x	1.08x
Perf (per Dollar)	1.0x	~2.4x

	Intel HSW(2650v3)	Intel BDW(2680v4)
Disk Types	PCIe SSD(p3600)	PCIe SSD(p3600)
Numbers of Drives	3	3
Total Capacity	1.6TB x 3	1.6TB x 3
Performance Gain	~2.83x	~3.64x
Drive Cost	$4782	$4782
Power Cost	$35	$35
Cooling Cost	$42	$42
Enclosure Cost	$0	$0
Reliability	$287	$287
Total Cost	$5146	$5146
Cost (per GB)	-	-
Perf (per Dollar)	1.0x	~1.28x

http://estimator.intel.com/ssddc/

推荐配置

SPARK SQL 推荐硬件配置

- **Good:** 更出色的性价比
  - 用更多的磁盘(HDD)或者SSD来提升磁盘并发吞吐.

Spark/Hadoop Cluster - Good(HASWELL)	
CPU	Intel® Xeon® CPU E5-2650v3**
Memory	256GB
NIC	10GbE x1
Disks	SATA HDD (4TB) x 16+/ Intel® s3520 1.2TB SSD x8+

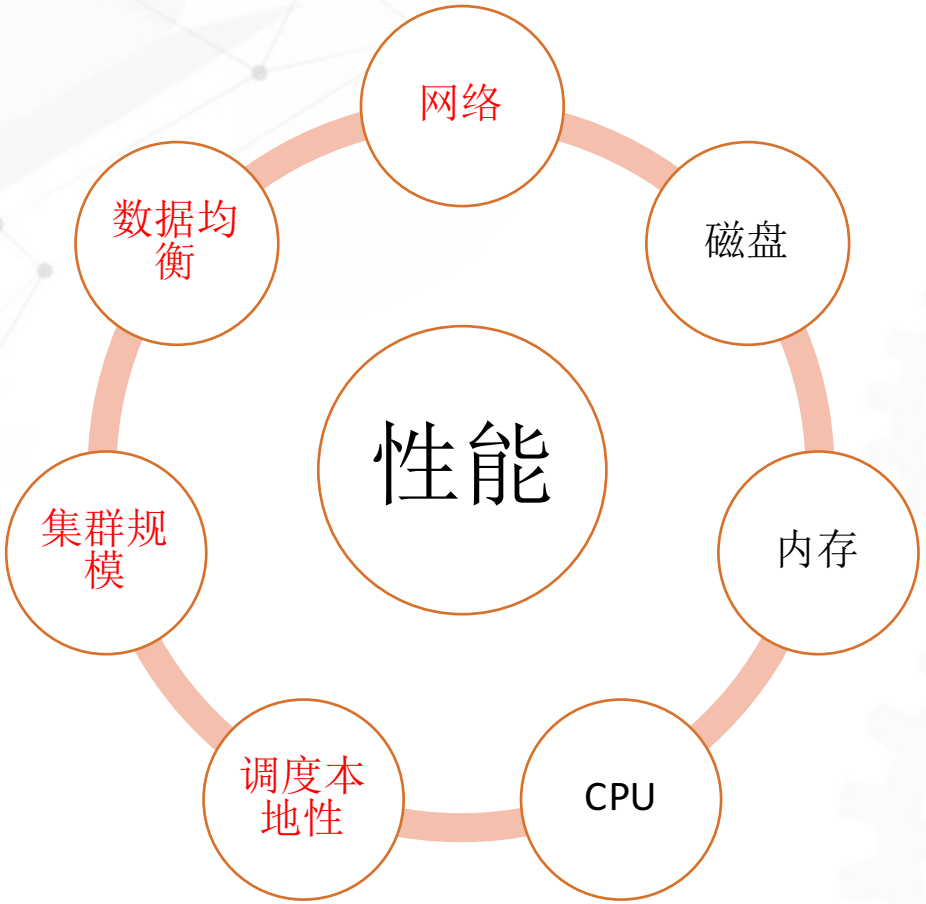
- **Better:** 更出色的性能
  - 至少是多块盘的Intel S3520 SSD甚至是P3600系列，更高端的CPU

Spark/Hadoop Cluster - Better(BROADWELL)	
CPU	Intel® Xeon® CPU E5-2680v4**
Memory	256 GB
NIC	10GbE x1 or 25GbE
Disks	Intel® s3520 1.2TB SSD x8+ / Intel® p3600 SSD x2+

- **Best: ? ? ?**

下一步？

影响Spark(SQL)性能的主要因素



可以优化的方向思考

- 更高的数据压缩比
- HDFS缓存或者层次化数据存储
- 更加合理的任务调度（利用数据本地性）
- 避免数据倾斜
- 自适应的任务数调节


THANKS
SequeMedia
盛拓传媒
IT168.com
ITPUB
ChinaUnix