How Spark Fits Baidu's Scale

Dr. James Peng

Principal Architect, Baidu U.S.A.

Baidu is Leading Chinese Search Engine

Search market share in China

No. 1

Website in China
By average daily visitors and page views

No. 1

Chinese search engine
By number of search question conducted

No. 1

Most valuable brand
Among private sector companies in China

$73 bn.

Market capitalization
Premium global Internet company

$2 bn.

Q1 2015 total revenues
By number of search question conducted

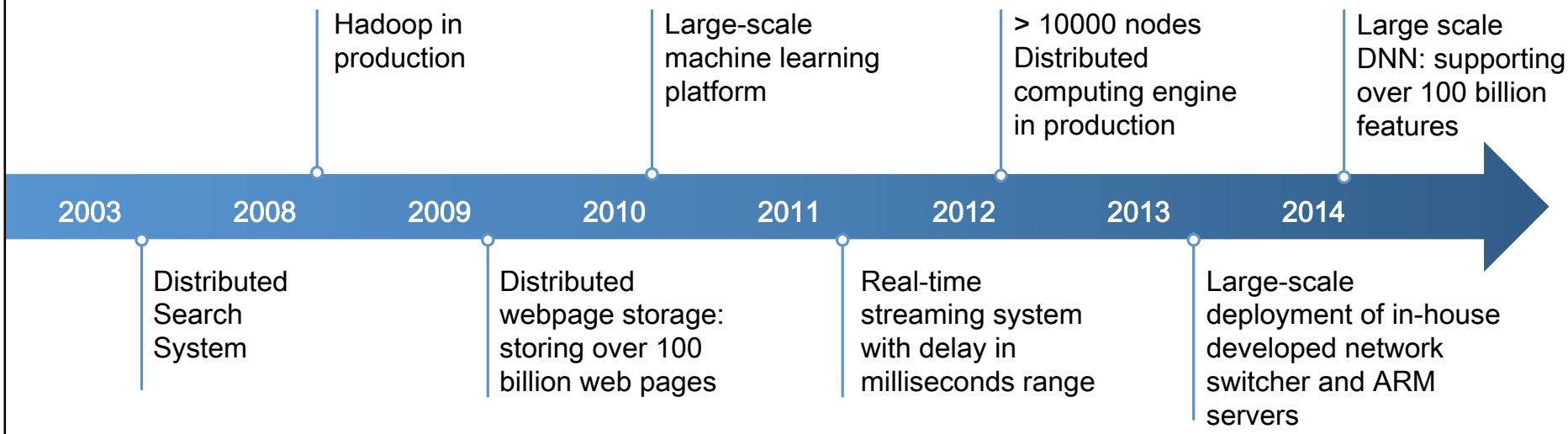
47,000

Employees
Top 3 China's best employer

Baidu's Big Data Infrastructure

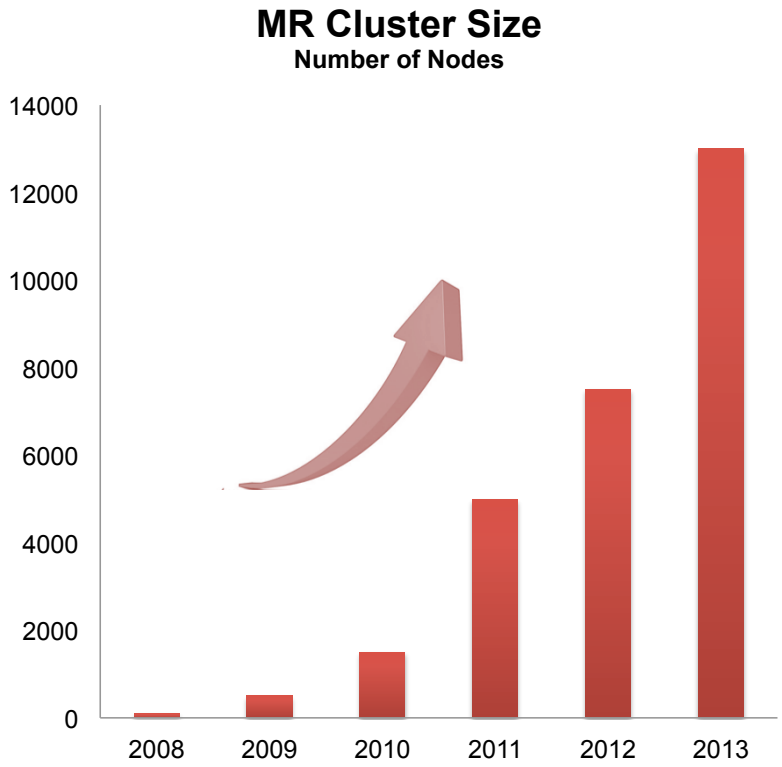
Baidu 百度

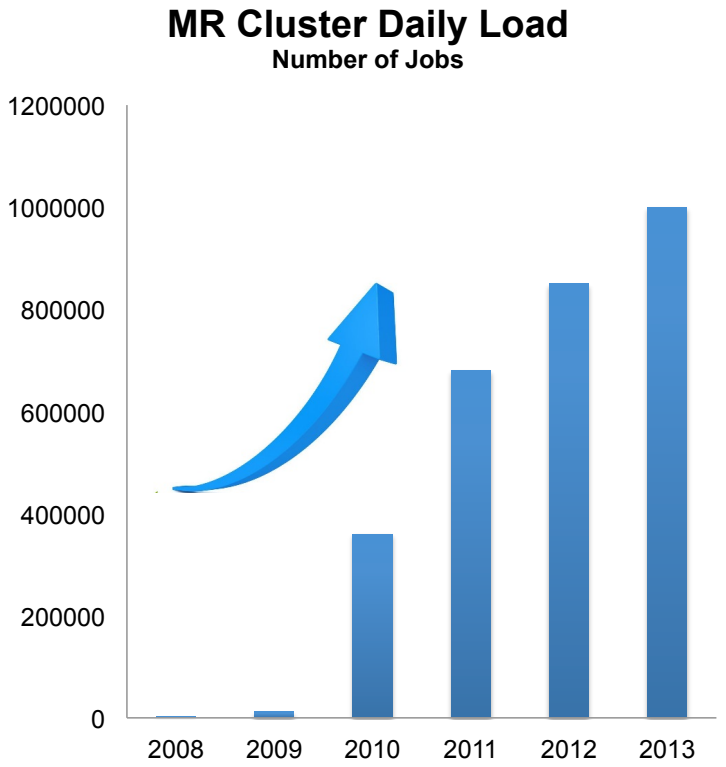
Baidu's History in Big Data Infrastructure



Baidu 百度

Largest MR Cluster





Baidu 百度

Baidu's Distributed Shuffle Engine

Champion of 2014 Sort Benchmark

Indy

2014, 8.38 TB/min

BaiduSort

100 TB in 716 seconds

982 nodes x

(2 2.10Ghz Intel Xeon E5-2450, 192 GB memory, 8x3TB 7200 RPM SATA)

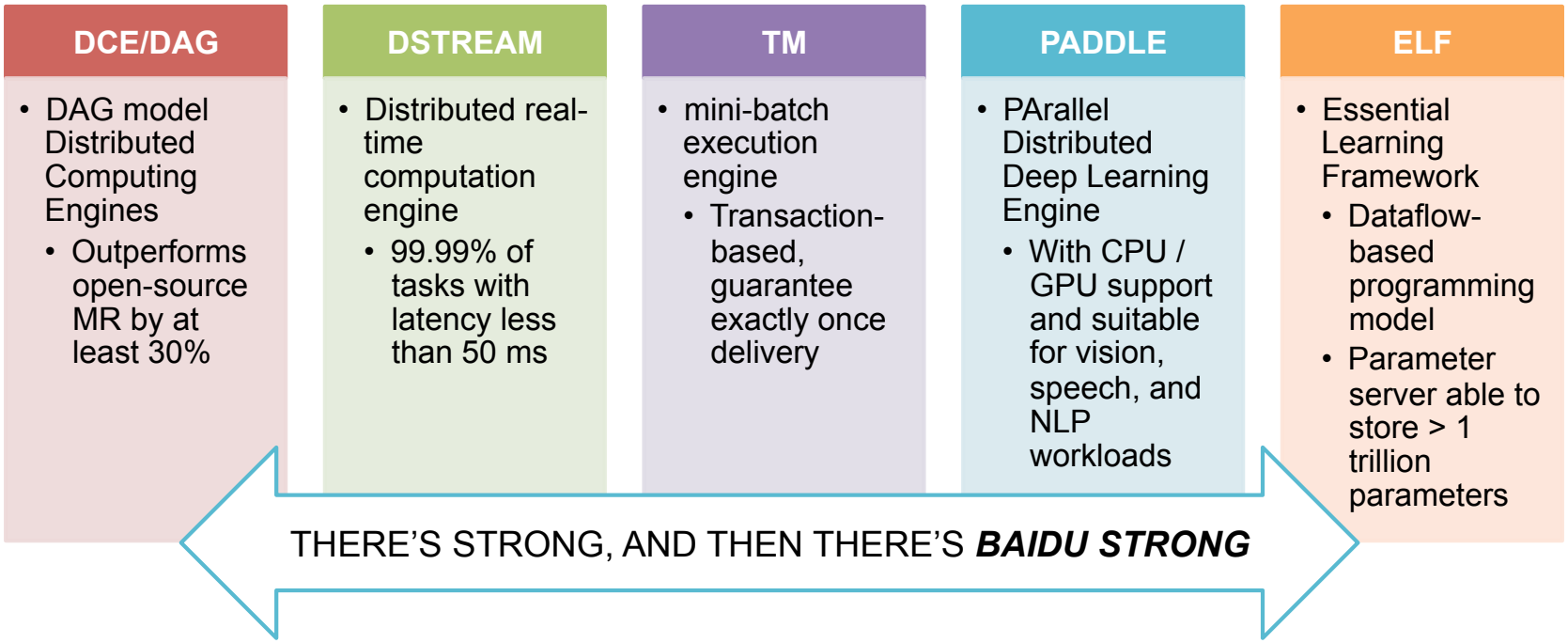
Dasheng Jiang

Baidu Inc. and Peking University

http://sortbenchmark.org/BaiduSort2014.pdf

Baidu 百度

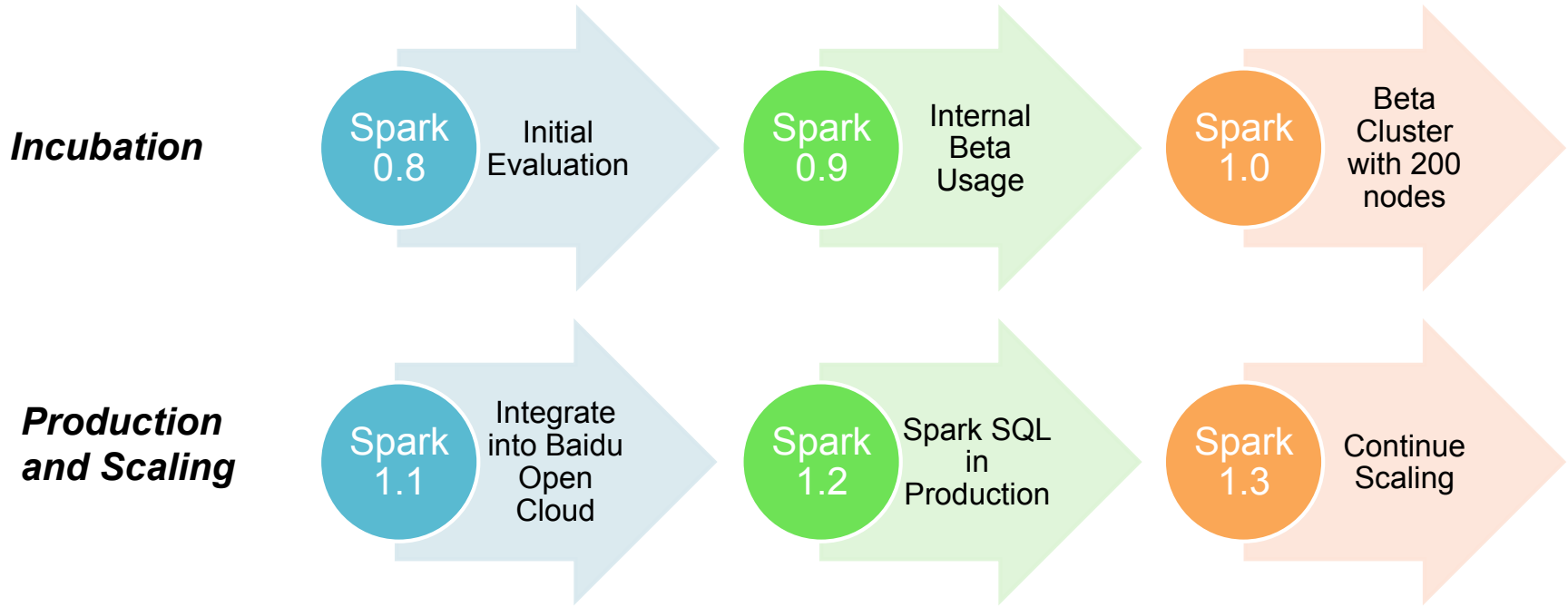
Baidu Proprietary Platforms



A Spark in Baidu

Bai du 百度

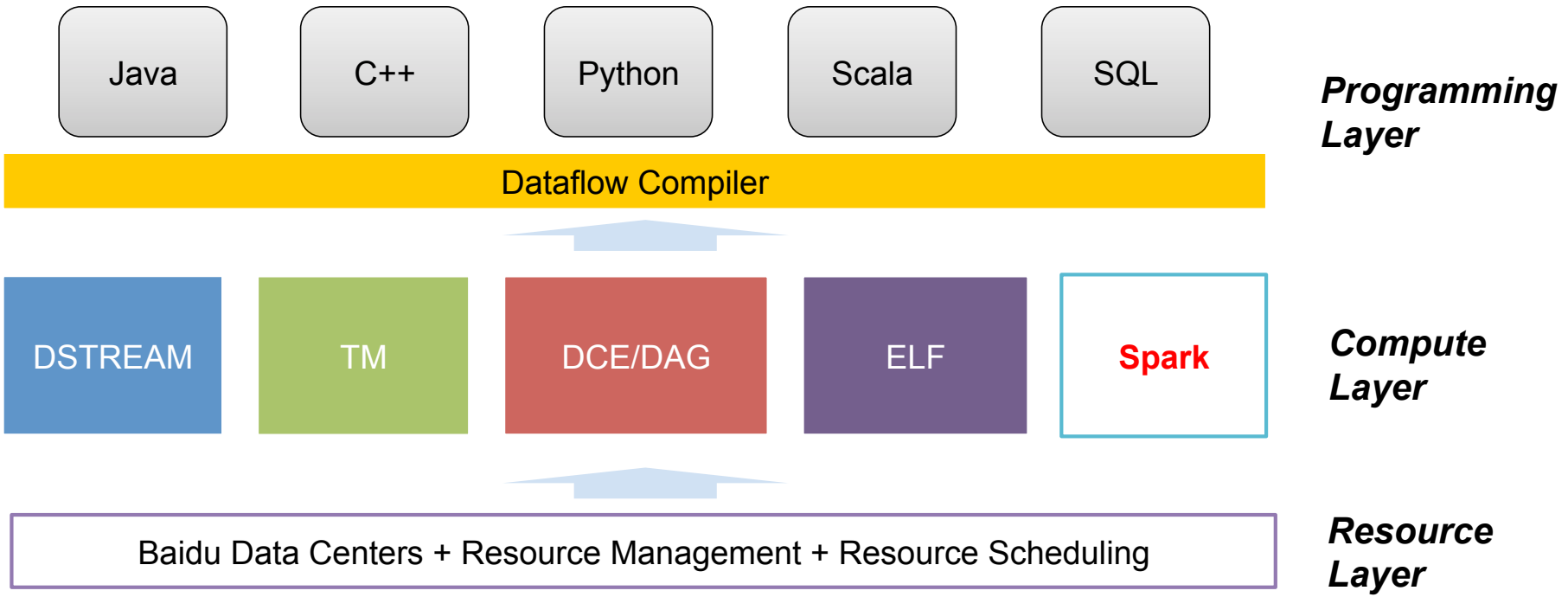
Spark's History in Baidu



- Scale: > 1000 nodes (20000 Cores, 100 TB Memory)
- Daily Jobs：2000~3000
- Supported Business Units: Ads, Search, Map, Commerce, *etc*

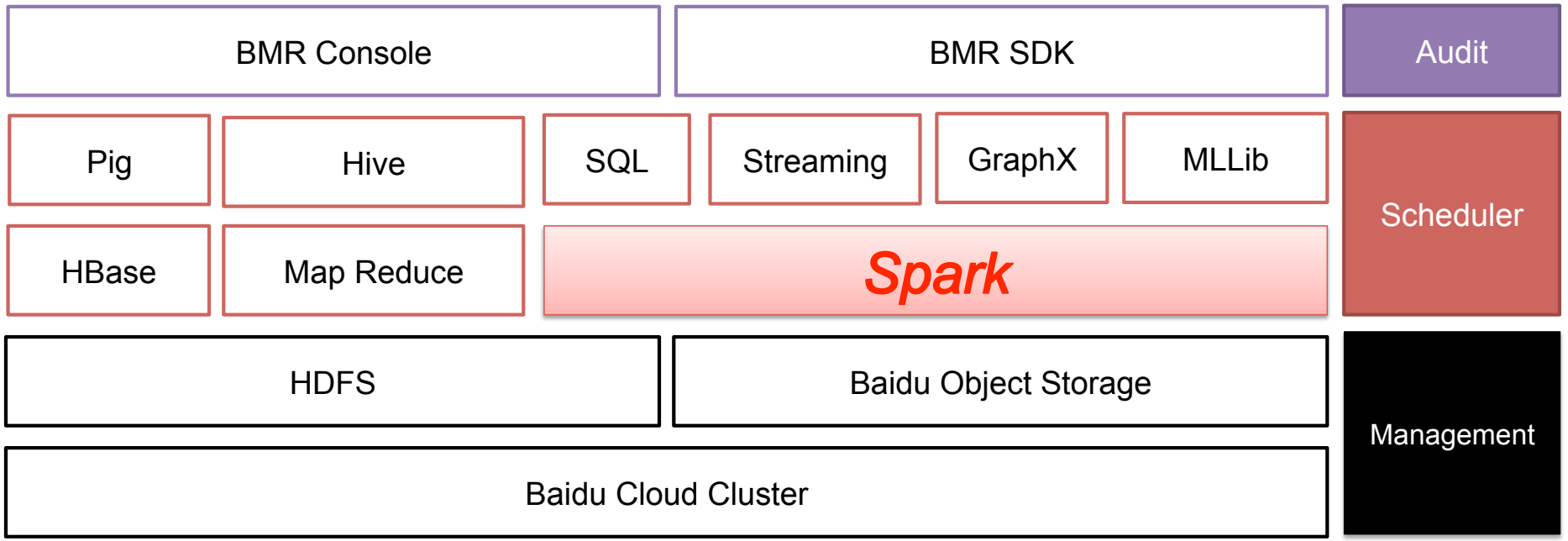
Baidu 百度

Spark in Baidu Internal Infrastructure



Spark in Baidu Open Cloud

http://bce.baidu.com/

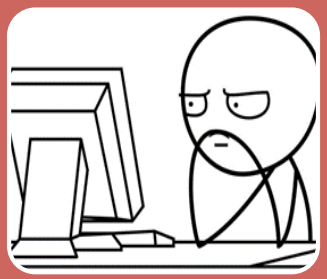


Baidu 百度

Enabling Interactive Queries with Spark and Tachyon

Baidu 百度

Need for Interaction (Speed)



Problem

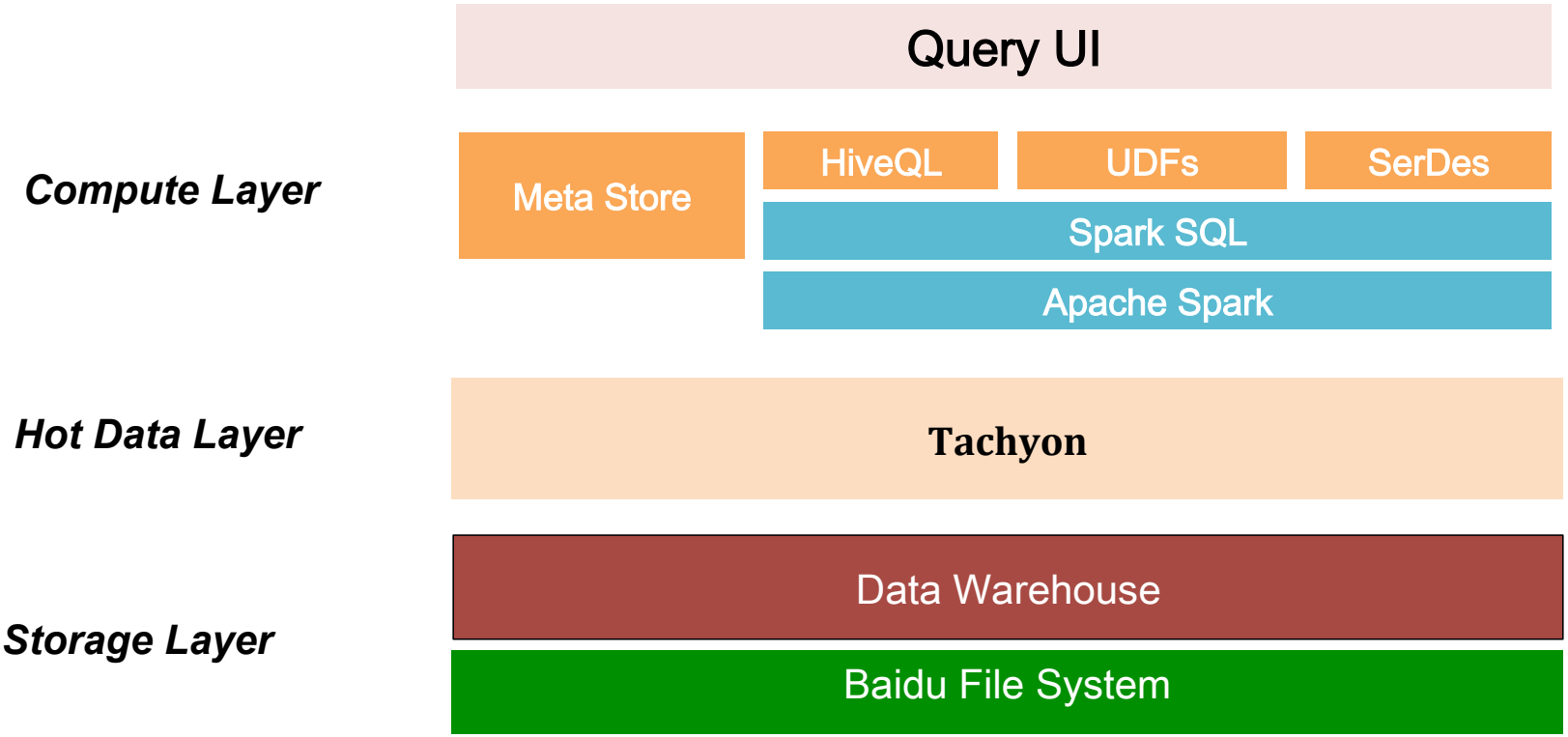
- John is a PM and he needs to keep track of the top queries submitted to Baidu everyday
- Based on the top queries of the day, he will perform additional analysis
- But John is very frustrated that each query takes ***tens of minutes***

Requirements for Baidu Interactive Query Engine

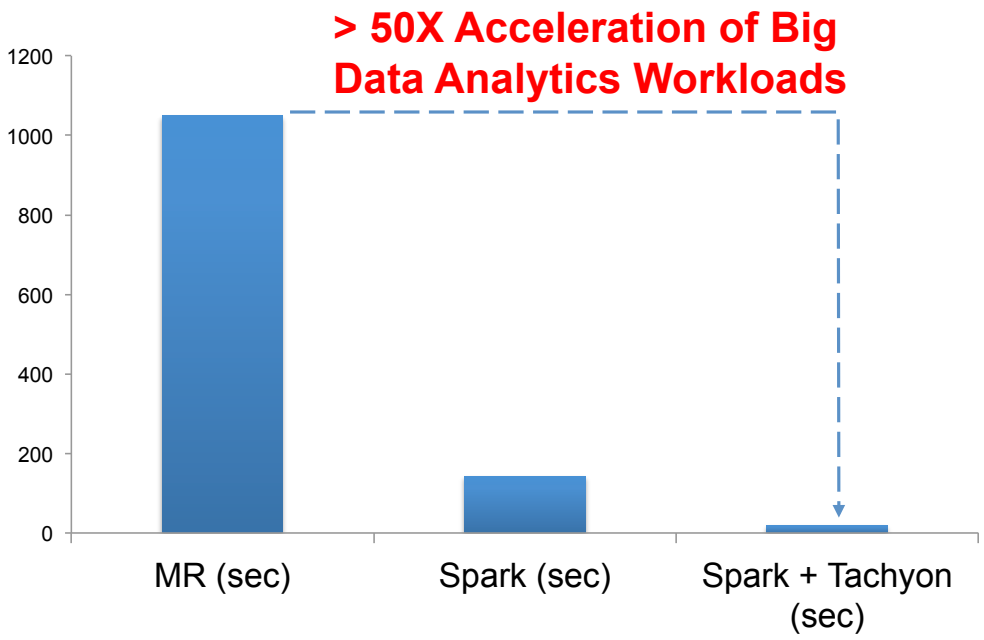
- Manages Petabytes of data
- Able to finish 95% of queries within ***30 seconds***

Baidu 百度

Baidu Interactive Query Architecture



Baidu Interactive Query Performance



Setup:

1. Use MR to query 6 TB of data
2. Use Spark to query 6 TB of data
3. Use Spark + Tachyon to query 6 TB of data

Baidu 百度

Summary

Spark and Tachyon provides significant performance advantage

- Suitable for ***latency sensitive*** workloads, such as interactive ad-hoc query

Takes time for Spark to reach MR scale

- We still use MR for ***high throughput*** batch workloads
- Deployment Scale: Spark ***1000*** nodes vs. MR ***13000*** nodes
- Memory Management is one of the main blockers of scalability

Look forward to Spark Improvements

- Project ***Tungsten***: better memory management: main blocker of scalability
- Continuous improvement of ***Catalyst***
- Support for multiple meta stores

Baidu 百度

Looking Into the Future