Why Spark on Hadoop Matters

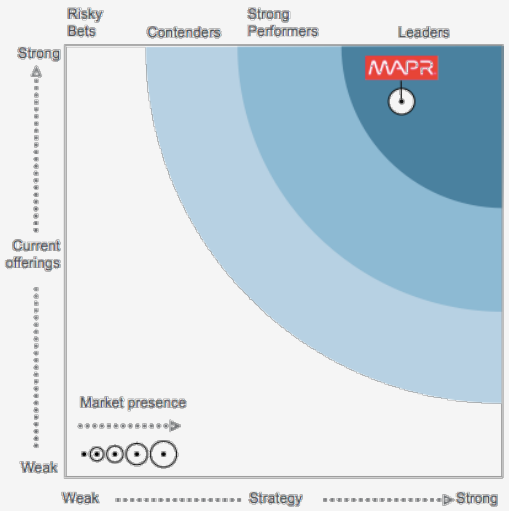
MC Srivas, CTO and Founder, MapR Technologies

Apache Spark Summit - July 1, 2014

MapR Overview

Top Ranked

FORRESTER®



Exponential Growth

- **3X** bookings Q1 '13 – Q1 '14
- **90%** software licenses
- **80%** of accounts expand 3X
- **<1%** lifetime churn
- **>$1B** in incremental revenue generated by 1 customer

500+ Customers

LIVE NATION ENTERTAINMENT

hp

TransUnion

SAMSUNG

AADHAAR

CISCO

rubicon PROJECT

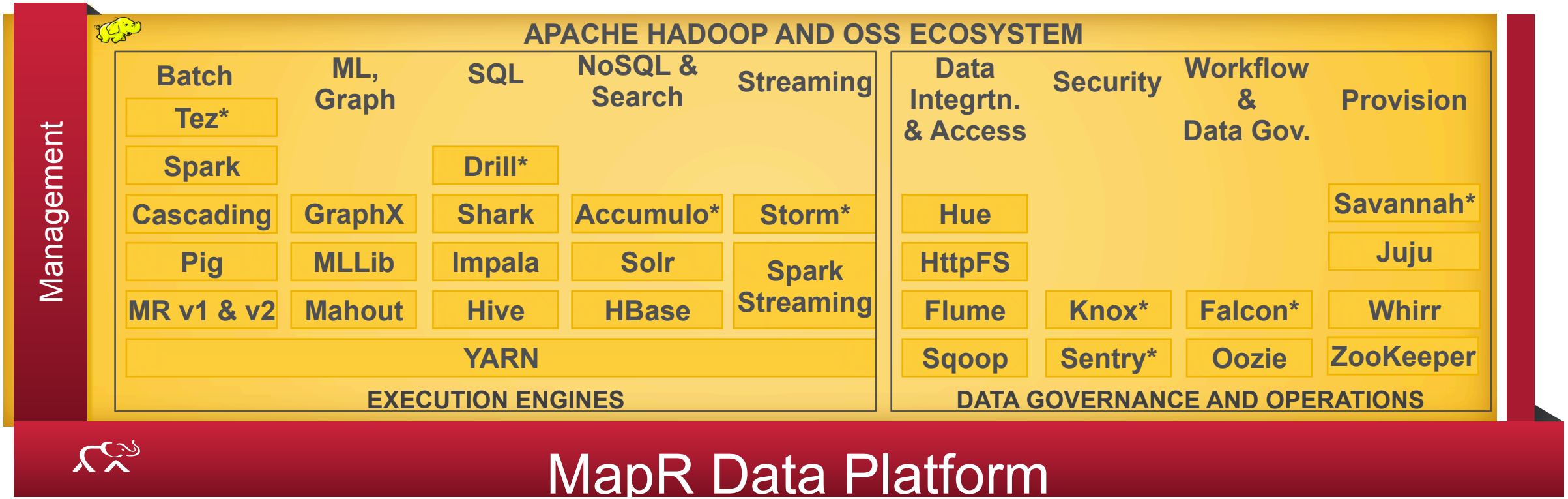
comScore

Cloud Leaders

Google

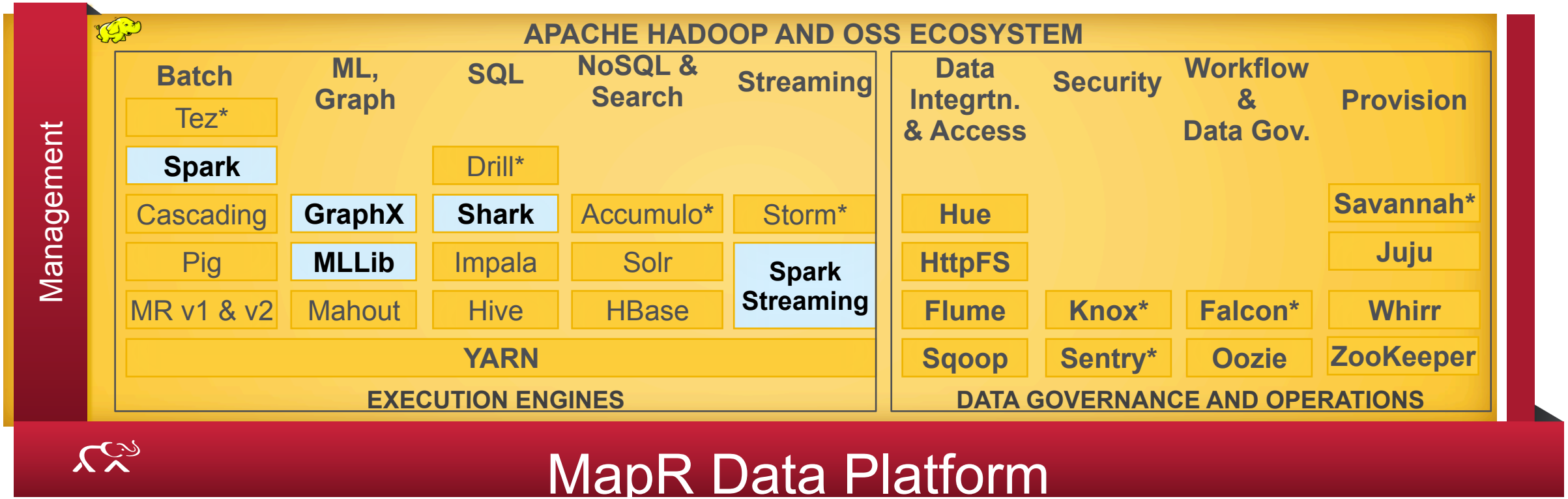
amazon web services™

Rapidly Evolving Landscape



* 2014 TIMELINE

The Complete Spark Stack on Hadoop



* 2014 TIMELINE


hadoop
+
Spark
=
A Winning Combination

Spark Advantages:

- Easier APIs
- Python, Scala, Java

EASE OF DEVELOPMENT

IN-MEMORY PERFORMANCE

- RDDs
- DAGs Unify Processing

- Shark, ML, Streaming, GraphX

COMBINE WORKFLOWS

Hadoop Advantages:

UNLIMITED SCALE

- Multiple data sources
- Multiple applications
- Multiple users

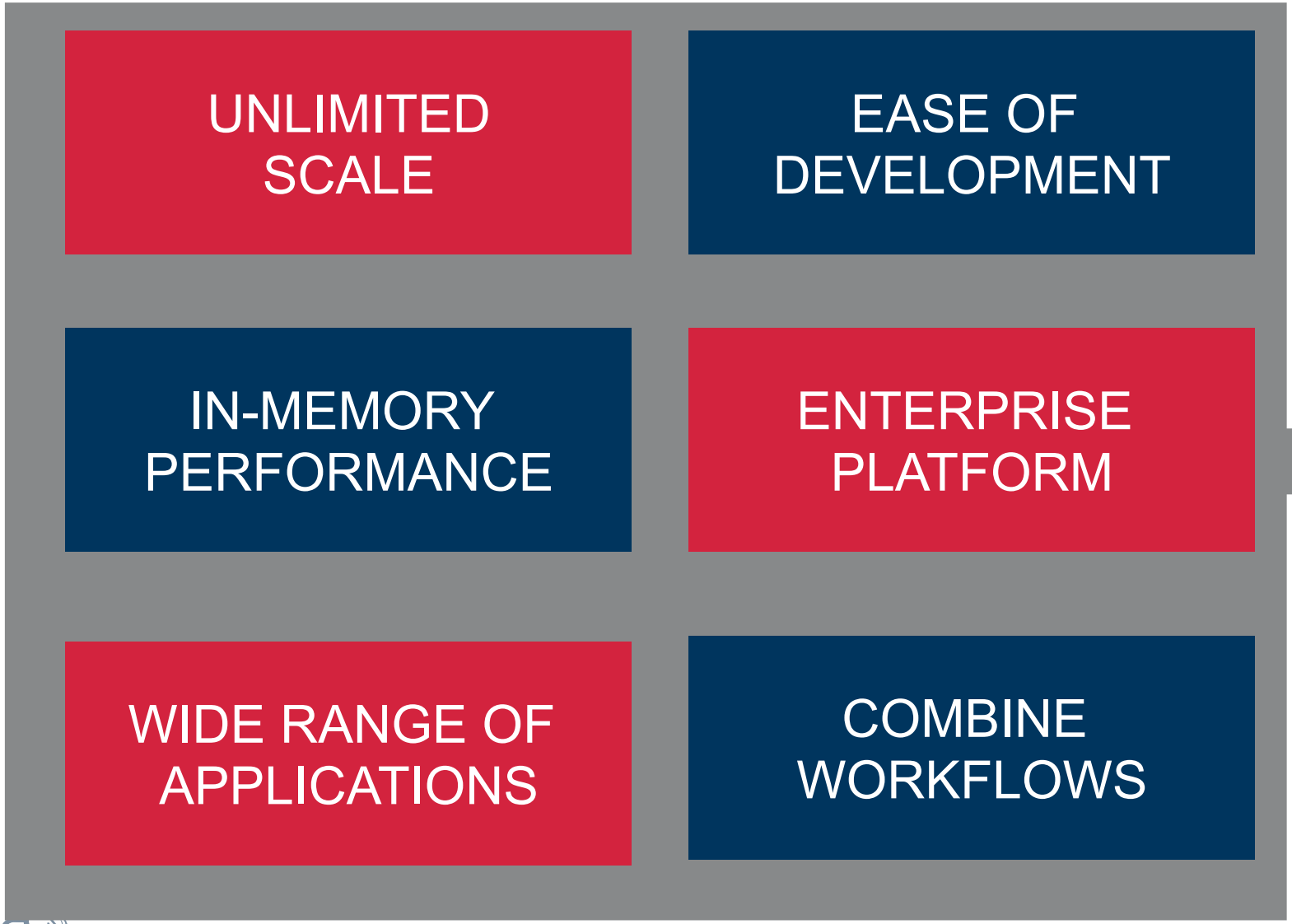
ENTERPRISE PLATFORM

- Reliability
- Multi-tenancy
- Security

WIDE RANGE OF APPLICATIONS

- Files
- Databases
- Semi-structured

The Combination of Spark on Hadoop



Operational Applications Augmented by In-Memory Performance

Case Studies

Industry Leading Ad-Targeting Platform

- High performance analytics over MapR M7 NoSQL
- Load from M7 table into RDD to augment scoring in real-time
- Results fed back to M7 for other applications

Leading Pharma Company: NextGen Genomics

Existing process takes **several weeks** to align chemical compounds with genes

ADAM on Spark allows realignment in a **few hours**

Geneticists can **minimize** engineering **dependency**

Cisco: Security Intelligence Operations

Sensor data lands in M7

Spark Streaming on M7 for first check on known threats

Data next processed on GraphX and Mahout

Results queried using SQL via Shark and Impala

Insurance Giant: Addressing Health Care Regulations

Patient information in M7 combined with clinical records to compute re-admittance probability

Process uses Spark with transactional data in M7

Insurance options decided in real-time on online portals

In Summary

Spark on **Hadoop** gains traction for **Real-time** applications

Pick the **Right Tool** for the **Job**

MapR is Unbiased Open Source (a la Linux)

- Open source distribution is about providing choice
  - Linux includes MySQL, PostgreSQL **and** SQLite
  - Linux includes Apache httpd, nginx **and** Lighttpd

	MapR Distribution for Hadoop	Distribution C	Distribution H
Spark	Spark (*all* of it) and Shark	Spark only	No
Interactive SQL	Shark, Impala, Drill, Hive/Tez	One option (Impala)	One option (Hive/Tez)
Versions	Hive 0.10, 0.11, 0.12, 0.13 Pig 0.11, 012 HBase 0.94, 0.98	One version	One version

Thank you

Engage with us!

@mapr   maprtech

mapr-technologies   MapR

srivas@mapr.com   maprtech