MAPR®

Real Experiences from a Hadoop Veteran

Jim Scott, Director, Enterprise Strategy & Architecture

Strata New York - 2014

Agenda

- How much does it cost to play this game?
  - Soft costs, hidden costs and cluster sizing
  - Shaping a team
- Limitations and cause for concern
  - Everything isn’t rainbows and unicorns
- Technologies to consider
  - NoSQL
- Architectures

Free BEER!

FREE BEER
tomorrow

Total Cost of Ownership and Operations

- More than just upfront costs
- Multiple environments for testing
- Total servers in a cluster
- Your time – It has a cost!

Skillsets

- Who is going to administer your platform?
- Monitoring
- Data engineering and enabling others

“60% of the time, it works every time”

Hands on Experience!

Arming You, with the Facts!

- YARN does not give you multi-tenancy
- Apache Hadoop does not have true NFS
- Apache Hadoop does not have consistent Snapshots
- Apache Hadoop does not have disaster recovery
- MapR supports the entire Hadoop API

Pick the Right Tool for the Job

Friends don’t let friends run NameNodes.

Technologies to Consider

Languages and Frameworks

- Languages
  - Java, Scala, Clojure
  - Python, Ruby
- Higher Level Languages
  - Hive
  - Pig
- Frameworks
  - Cascading, Crunch
- DSLs
  - Scalding, Scrunch, Scoobi, Cascalog

Once Upon a Time…

- Dissipating usage
  - Pig

- DSLs that opened doors
  - Scalding

- Hive
  - As an engine

Databases

- MongoDB
- Cassandra
- MapR-DB and HBase
  - Key design
  - Usage of column families

Data Movement and Time Series

- Flume

- Kafka

- OpenTSDB
  - Grafana

SQL on Hadoop

- Generates MapReduce jobs
  - Hive
- Do NOT generate MapReduce jobs
  - Tez
  - Impala
  - SparkSQL
    - Runs on Spark Engine
  - Apache Drill

Drill Supports *Schema Discovery On-The-Fly*

Schema Declared In Advance

- Fixed schema
- Leverage schema in centralized repository (Hive Metastore)

Schema Discovered On-The-Fly

- Fixed schema, evolving schema or schema-less
- Leverage schema in centralized repository or self-describing data

SCHEMA ON WRITE

ORACLE

Microsoft SQL Server

IBM DB2

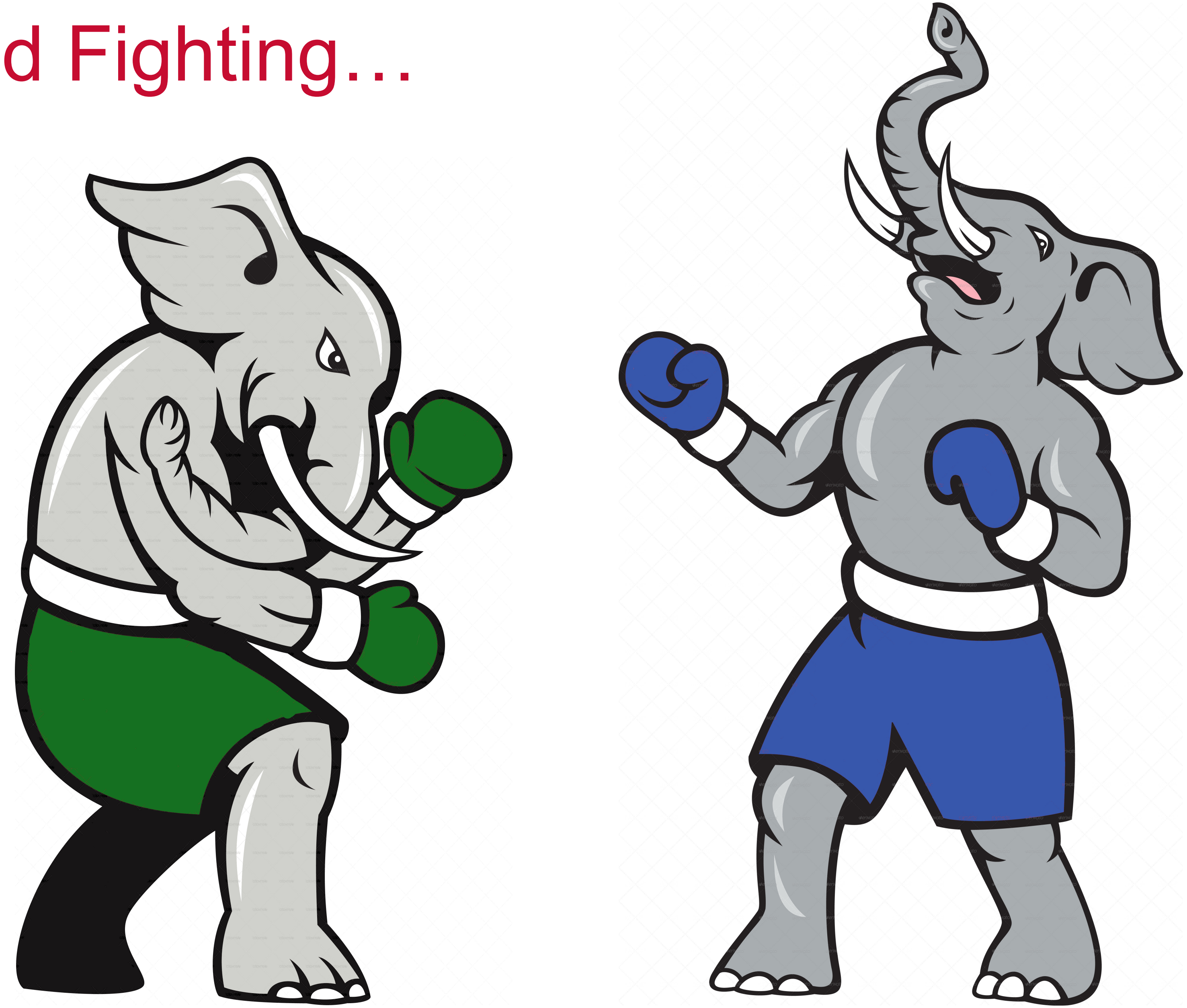
SCHEMA BEFORE READ

HIVE

SCHEMA ON THE FLY

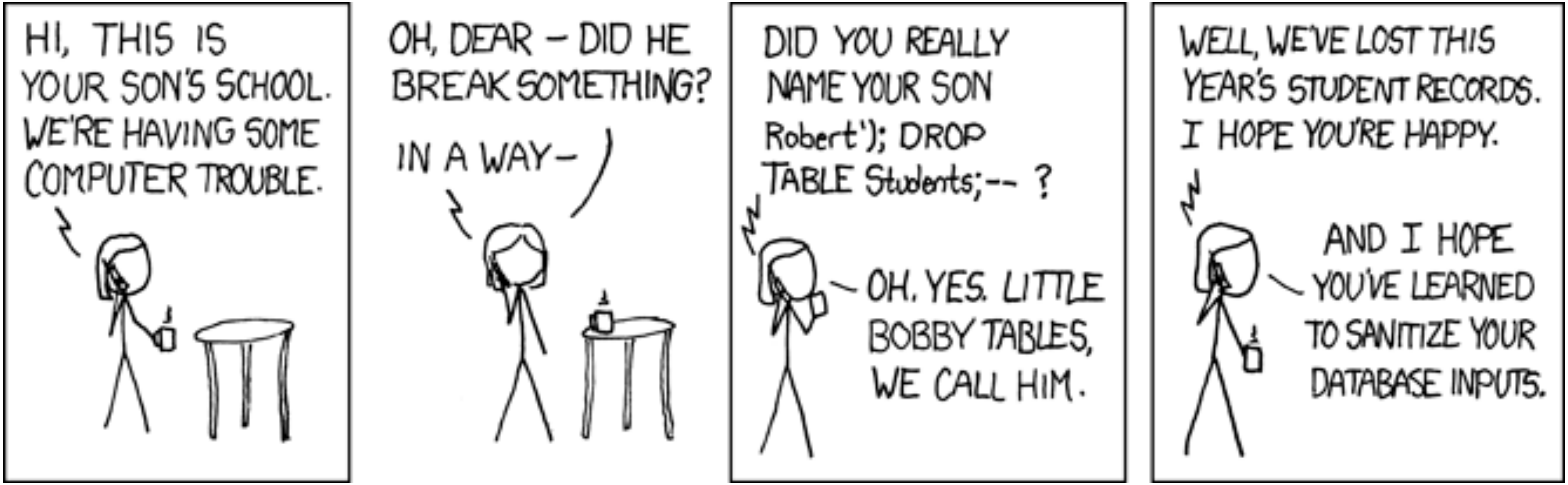
APACHE DRILL

Politics and Fighting…



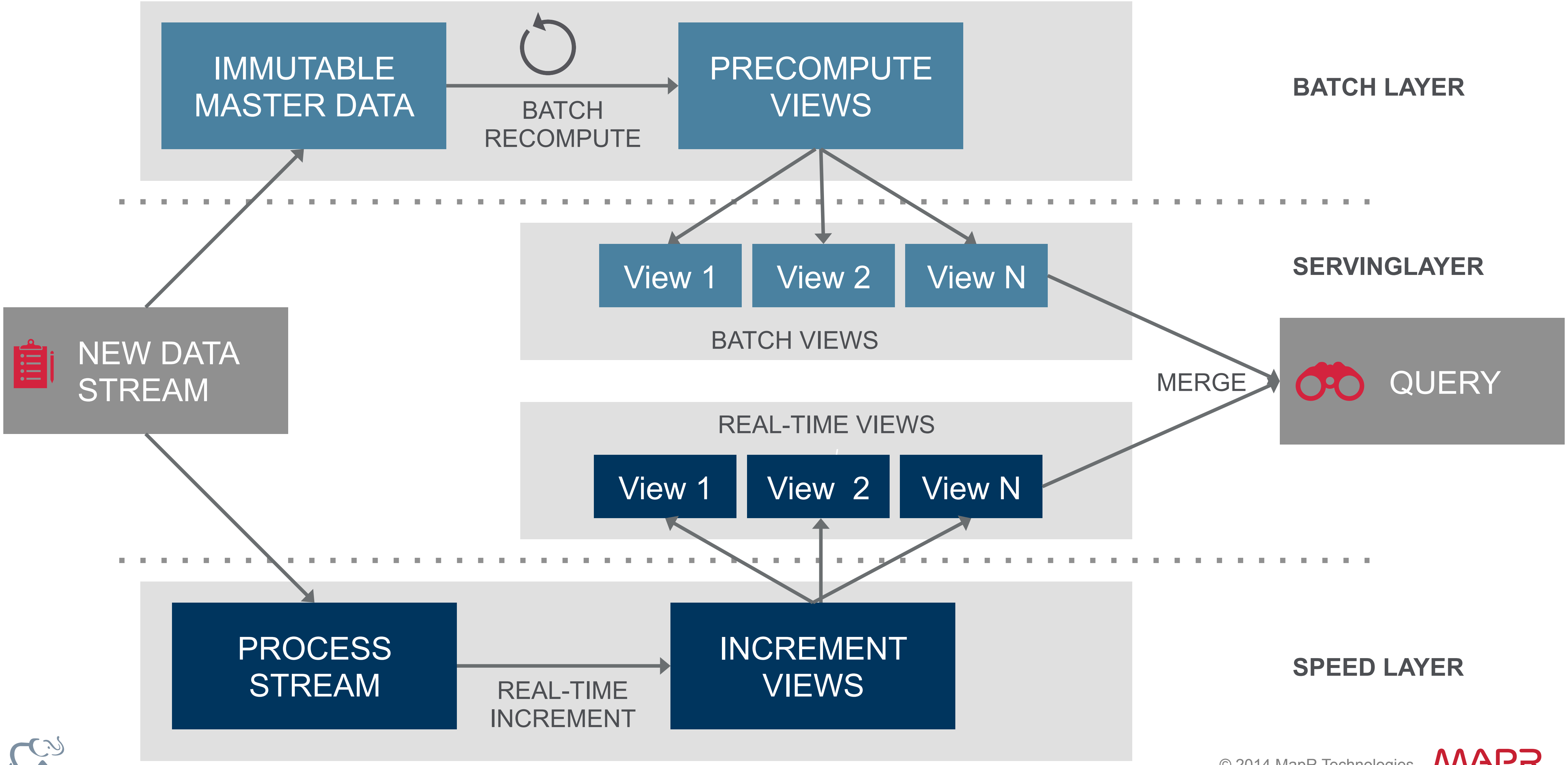
Design and Architecture

WORKED FINE IN DEV
OPS PROBLEM NOW
memegenerator.net

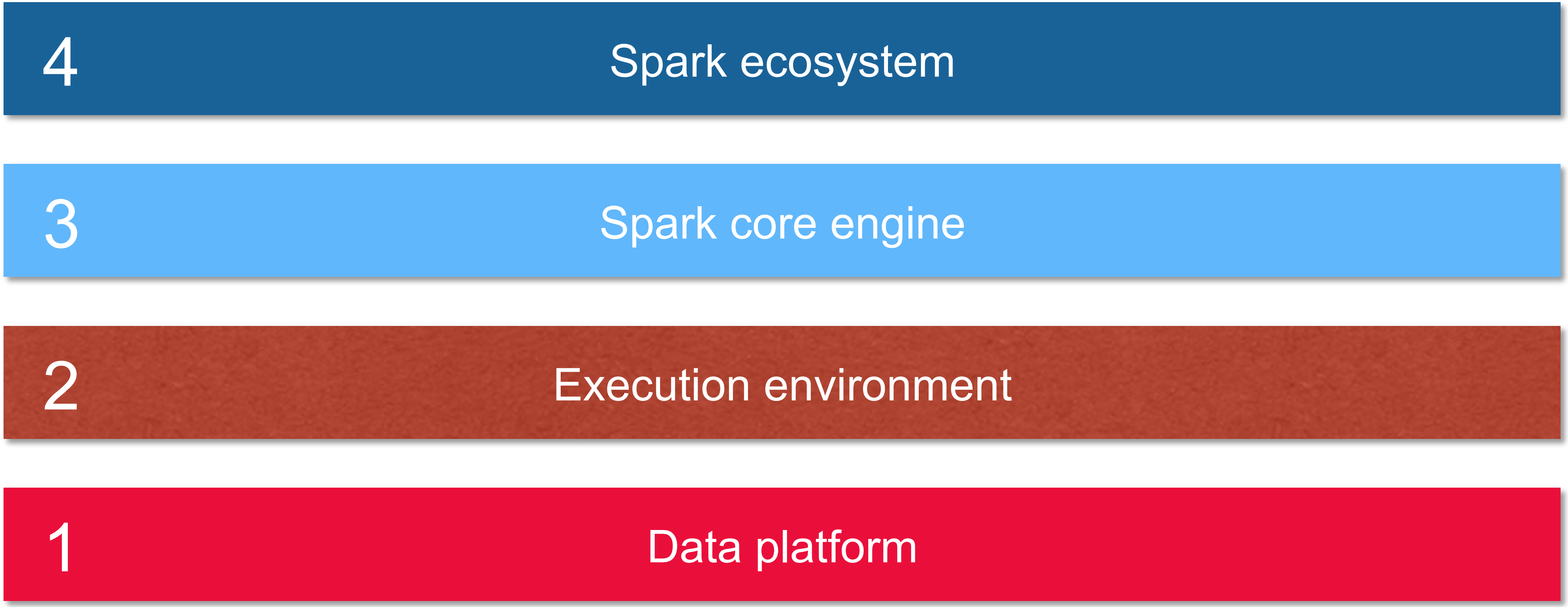

HI, THIS IS YOUR SON'S SCHOOL. WE'RE HAVING SOME COMPUTER TROUBLE.
OH, DEAR – DID HE BREAK SOMETHING?
IN A WAY–
DID YOU REALLY NAME YOUR SON Robert'); DROP TABLE Students;-- ?
OH. YES. LITTLE BOBBY TABLES, WE CALL HIM.
WELL, WE'VE LOST THIS YEAR'S STUDENT RECORDS. I HOPE YOU'RE HAPPY.
AND I HOPE YOU'VE LEARNED TO SANITIZE YOUR DATABASE INPUTS.

http://xkcd.com/327/

Lambda Architecture



The Spark Stack from 100,000 ft



4 Spark ecosystem

3 Spark core engine

2 Execution environment

1 Data platform

Q&A

Engage with us!

@kingmesal	t	f	maprtech
mapr-technologies	in	g+	MapR
jsccot@mapr.com	✉	You Tube	maprtech