HBase/PHOENIX @ Scale

A study of Salesforce's use of HBase and Phoenix

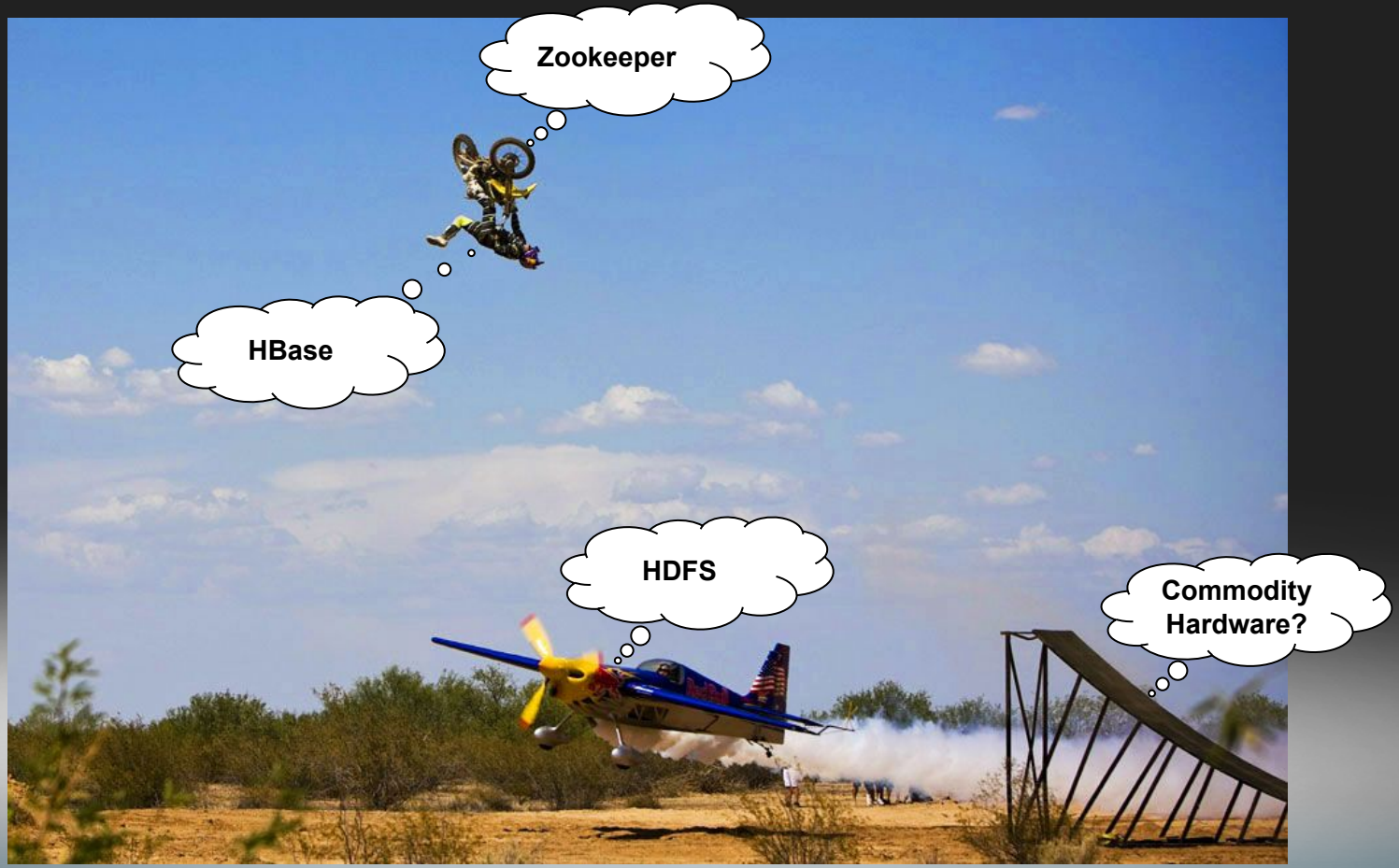
Lars Hofhansl
Vice President and Principal Architect at Salesforce
Apache HBase, Apache Phoenix committer and PMC member

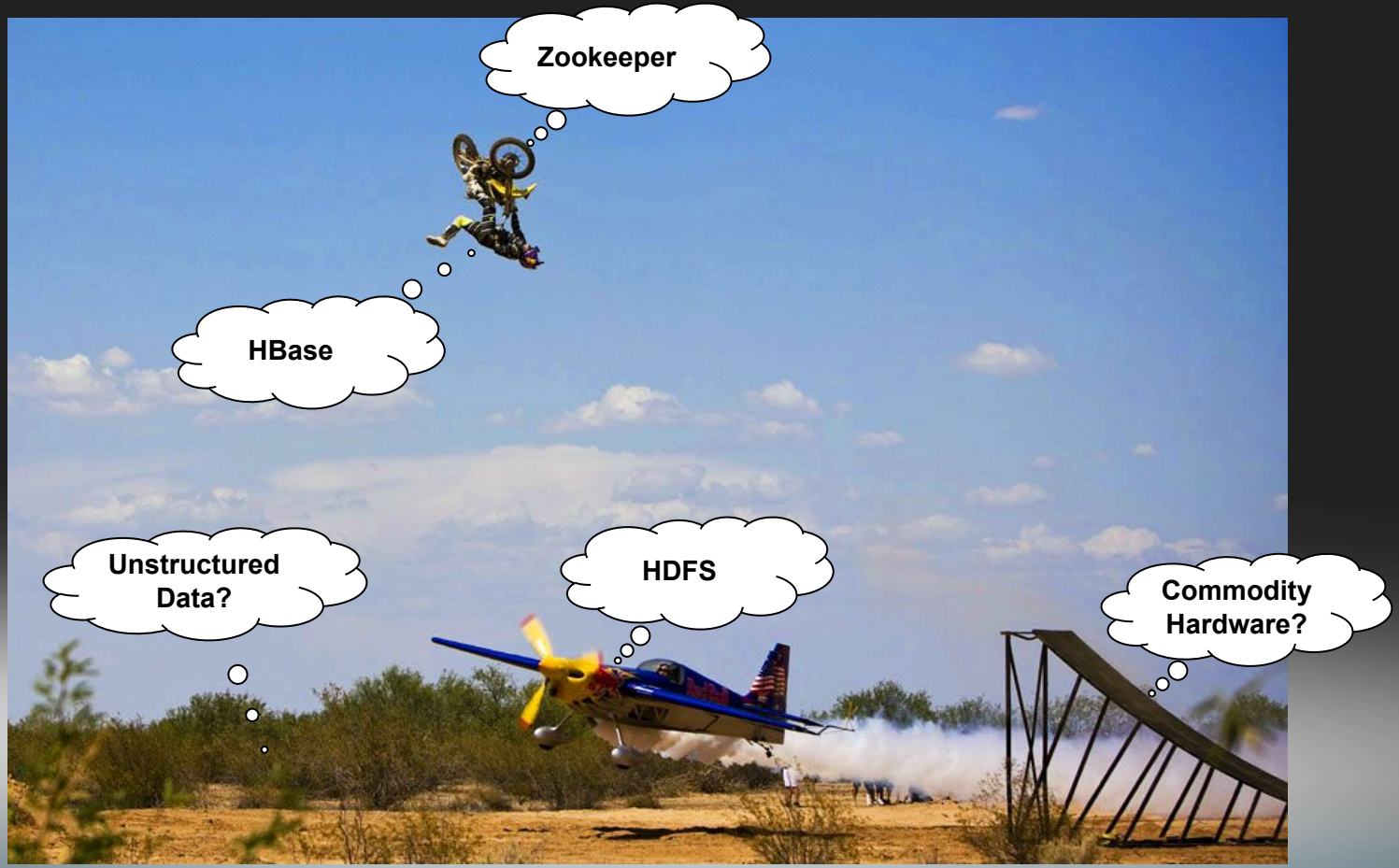
Two Years Ago, I showed you this

Zookeeper

Zookeeper
HBase

Zookeeper
HBase
HDFS


Zookeeper
HBase
HDFS
Commodity Hardware?


Zookeeper
HBase
Unstructured Data?
HDFS
Commodity Hardware?

HBase/Phoenix are BIG* at Salesforce

* Numbers are from some time in the past and do not reflect the current scale

Heavy users of relational and semi-structure data

Mix of customer* and internal data

* Through the Salesforce Platform Offering

Typical Use Cases

- Samples of Customer Data:
  - Login data to track anomalies in real-time
  - Archiving, historical data moved from operational, relational storage to HBase
  - Denormalized feed views for Salesforce Chatter
  - Chatter @mention low-latency relevancy queries
  - Storage of user activity on marketing campaigns for reporting and AI/ML
- Samples of Internal usage:
  - Periodic thread dumps from all AppServers
  - Machine metrics from all machines

Customer Facing

> 100 clusters of varying size

~4bn write requests / day

~80TB written / day

That’s about ~8 gbit/s, sustained

~600m read requests / day

~500GB read / day

Central Metrics Database

Central Metrics Database

Collecting data from > 80.000 machines

11.4 trillion metrics stored and growing

2.8 tn metrics in 6 months and growing

210 bn reads in 6 months