

Opportunities and Challenges of Data Processing in the Internet of Things (IoT)

Michael Hausenblas, Chief Data Engineer
Strata Barcelona, 2014-11-20

IoT—a superset of the Internet



IoT—a superset of the Internet

What is the IoT?

“The idea of an all-encompassing and ubiquitous **network of devices** to facilitate **co-ordination** and **communication** between the devices themselves as well as between the devices and **human** end-users. The involved devices are typically **constrained devices** such as RFID sensors, but may also more **sophisticated** ones like smartphones.”



IoT—a superset of the Internet

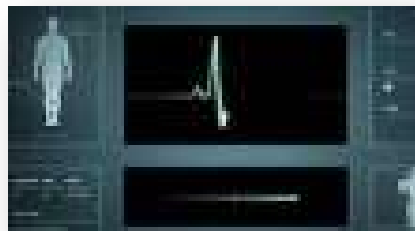


devices and their deployment



The IoT landscape

apps



data

Apache Kafka
A high-throughput distributed messaging system.

Spark



APACHE
HBASE



infrastructure



Google



the thing system

MAPR



QUALCOMM



IoT application areas

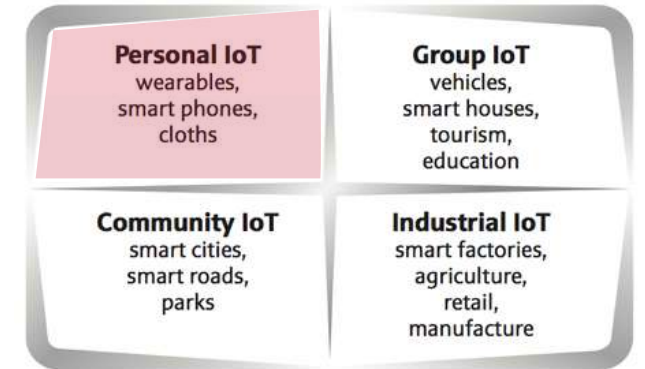


Use cases



Categorization & use cases: Personal IoT

Scope is on a *single person*, for example a smartphone equipped with GPS sensor or a fitness device that measures the heart and sharing this data with her GP. One of the fastest growing, rather consumer-oriented areas of IoT.



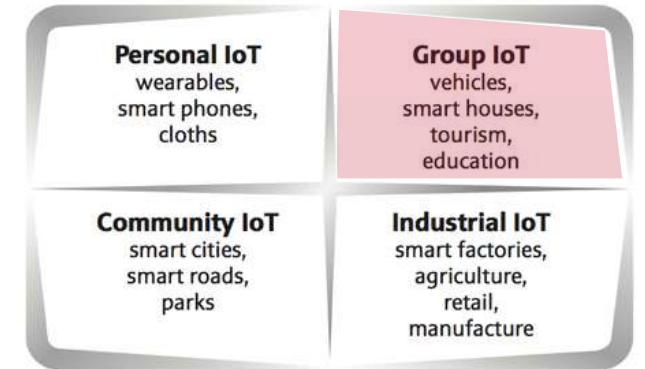
Use cases and apps

- Quantified self
- Smart jackets
- Personal digital assistant



Categorization & use cases: Group IoT

Focuses on a *small group of people*, for example a family in the context of a smart home where the deployed sensors capture temperature and lighting conditions for optimal comfort. One of the most challenging areas and yet early days.



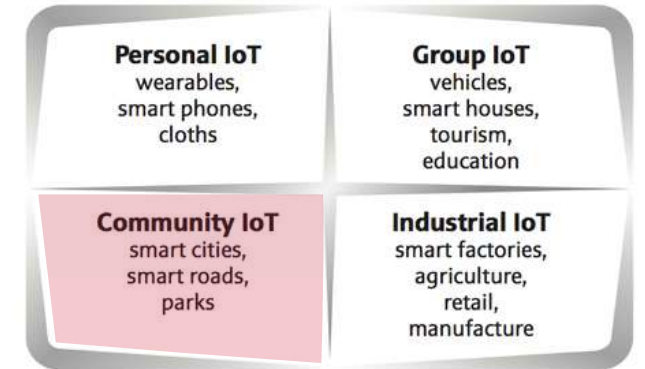
Use cases and apps

- Smart homes
- Proactive/predictive car maintenance
- Interactive tourism



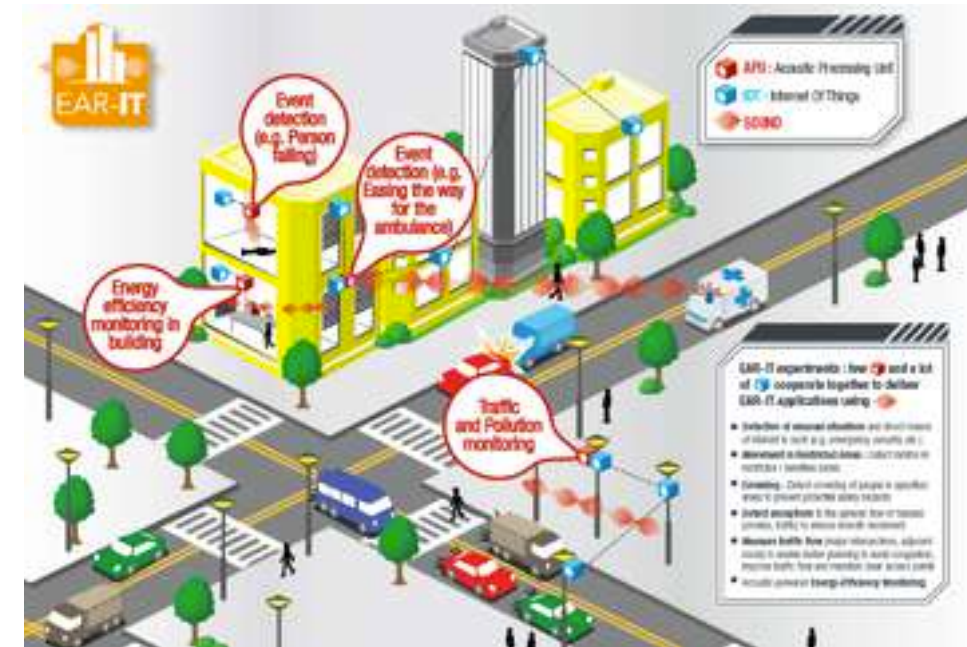
Categorization & use cases: Community IoT

Considers a *large group of people*, potentially tens of thousands, usually in the context of public infrastructure, such as smart cities. Some immature from a commercial POV but potentially promising IoT area.



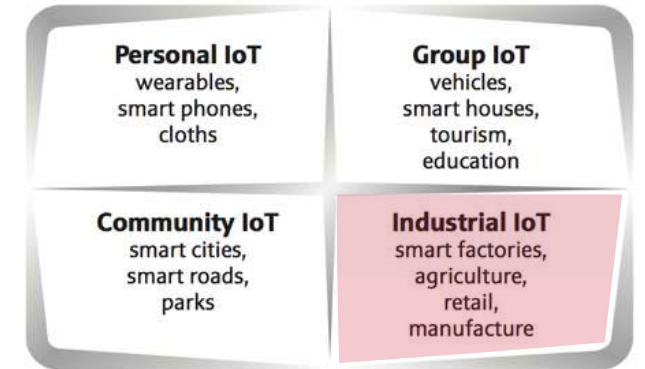
Use cases and apps

- Smart cities
- Health care (monitoring, trackers)



Categorization & use cases: Industrial IoT

Scope can be either *within* an organization or *between* organizations and/or individuals. This is arguably the most established and mature part of IoT, see also [M2M](#).



Use cases and apps

- [Smart factory](#)
- [Retailer](#) supply chain
- [Agriculture](#)
- [Waste management](#)



IoT use case examples





Largest biometric database in the world



1.2B

PEOPLE

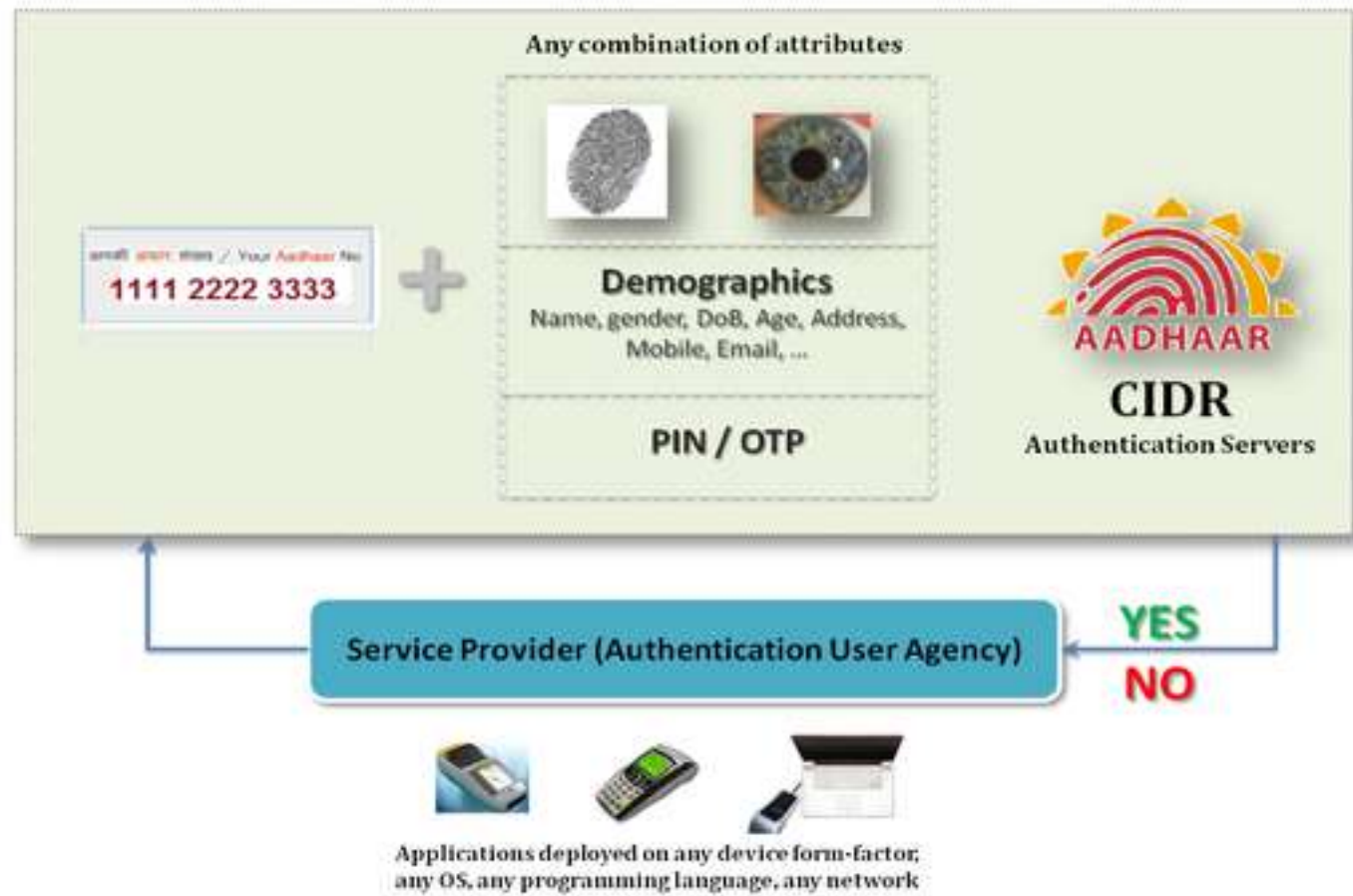


Largest biometric database in the world

- Goals:
 - Enable residents to participate in daily commercial business
 - Decrease embezzlement of government subsidies \$1.3+ billion
- Introduced in 2010 now over 500 million residents are registered
- Performs > 4.73 million authentications per minute with a latency SLA of 200 milliseconds

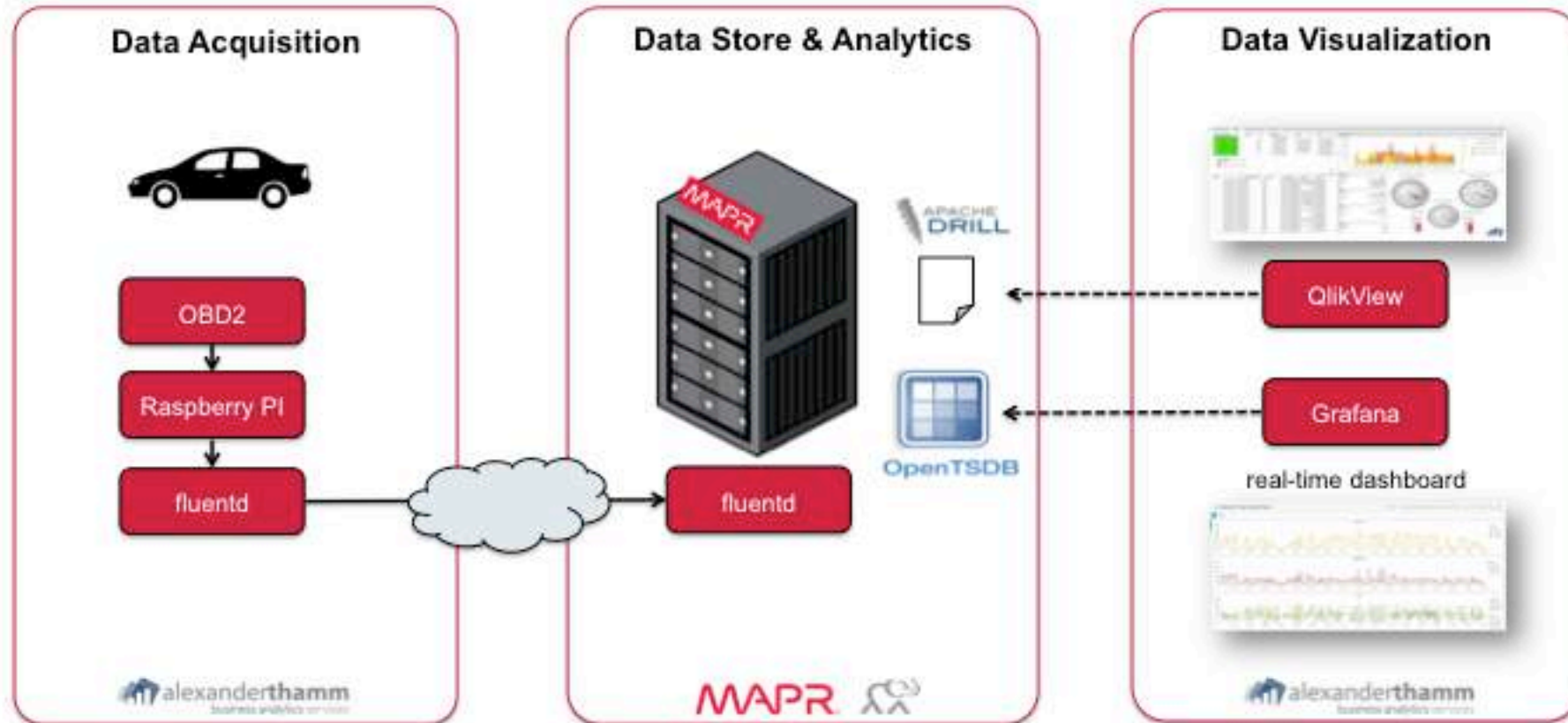


Largest biometric database in the world



<http://uidai.gov.in/publication-and-reports.html>

A proof of concept from the automotive sector



A proof of concept from the automotive sector



Waste & Recycling Leader

DATA

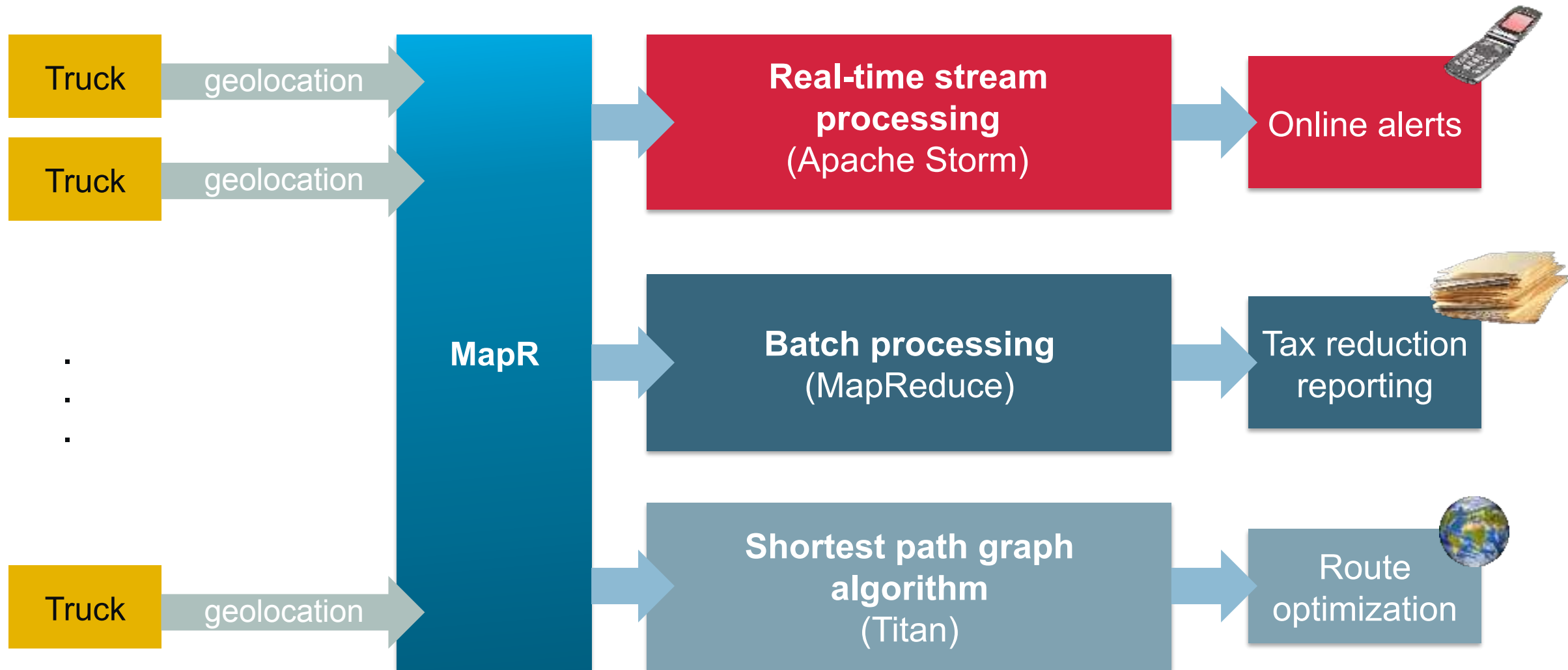
- Geolocation of 20,000 trucks
- Arriving every 5 seconds
- Geographic boundaries of landfills

GOALS

- Online alerts
- Tax reduction reporting
- Route optimization



Waste & Recycling Leader—Architecture



Business Intelligence and the IoT



What is Business Intelligence (BI)?

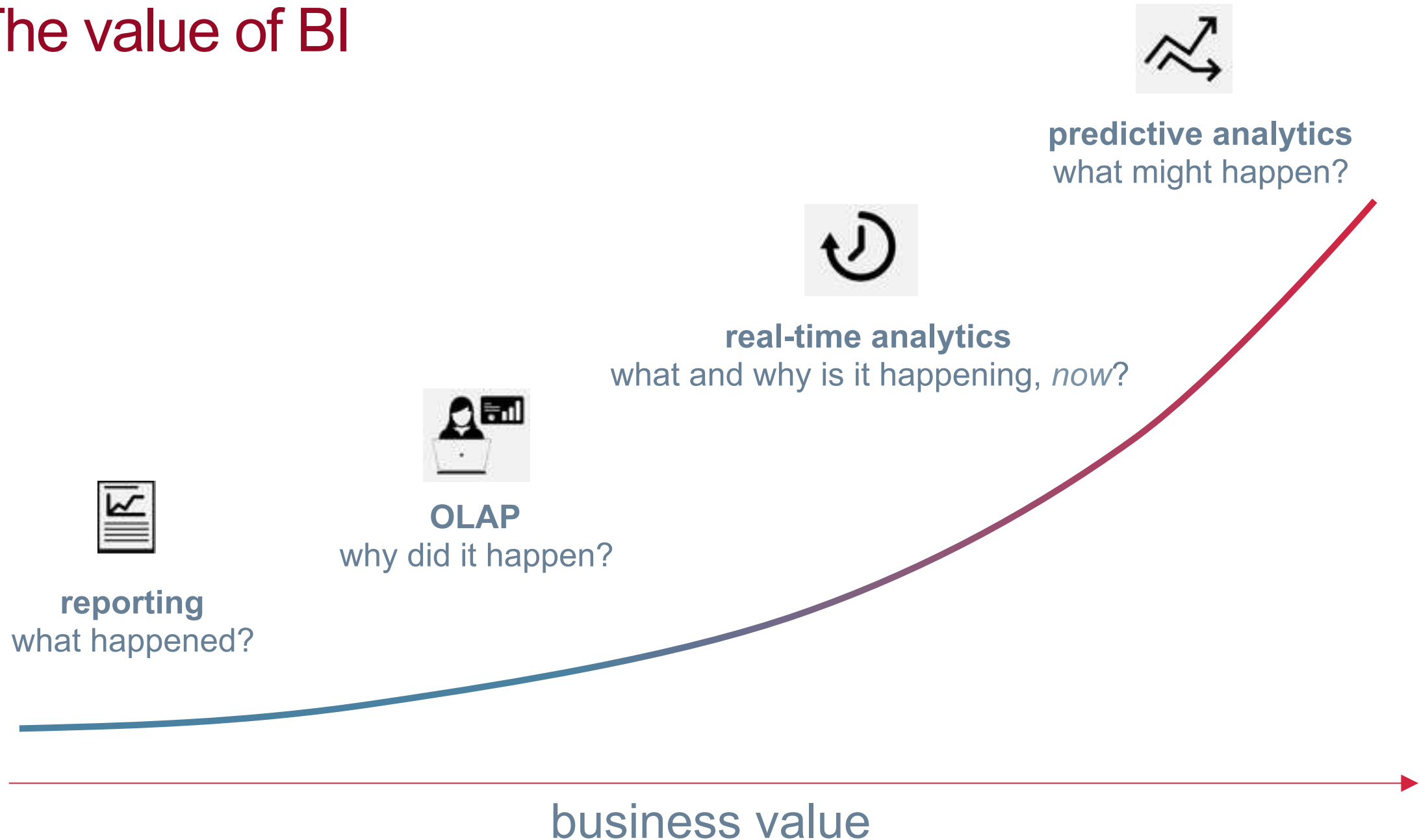
Umbrella term for methodologies, architectures, and technologies that **transform raw data into meaningful business insights**.

This includes related activities such as:





- **reporting**
- **online analytical processing (OLAP)**
- **real-time analytics**
- **predictive analytics**



The value of BI



Technologies typically used to realise ...

		Back-end	Front-end
	reporting	Hive, Pig, Spark SQL (Shark)	Excel, Tableau, Datameer
	OLAP	Hive/HBase/M7 tables + Drill/Impala/Vertica	SAS, QlikView, Tableau, Datameer
	real-time analytics	Kafka + Storm/Spark Streaming + Hive	DataTorrent, bespoke dashboards (web applications)
	predictive analytics	Potentially all above + Elasticsearch/Solr Mahout/Spark MLlib	Often bespoke apps



The Internet of Things architecture: iot-a



IoT lends itself to Big Data approach

“Using scale-out techniques on commodity hardware in a schema-on-read fashion along with community-defined interfaces”

Volume: store all incoming sensor data for historical references

Variety: dozens of data formats in use in the IoT world, none is relational

Velocity: many devices generate data at a high rate; usually data streams

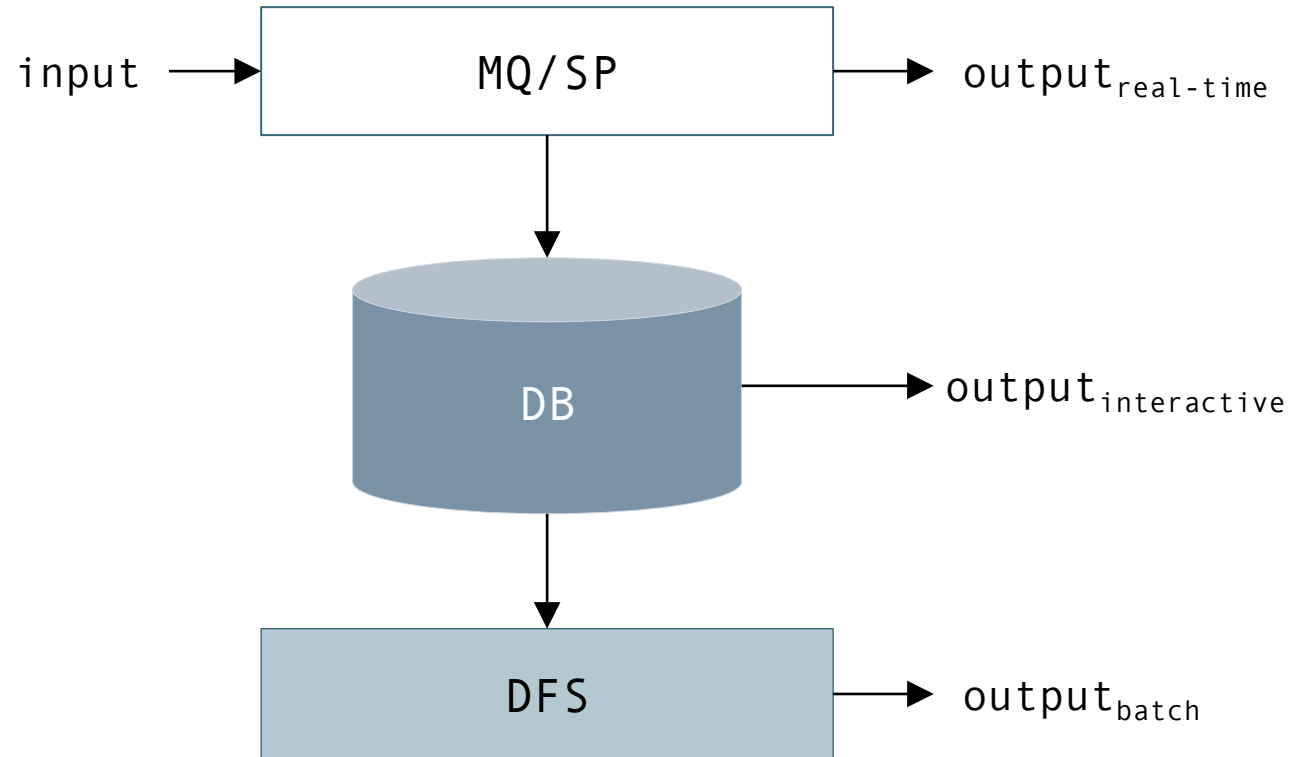


Requirements

- Able to natively deal with the raw data from devices, typically many (trillions) of small files in non-relation formats
- Support a range of workloads, especially streaming as first-class citizen
- Ensure business continuity to meet SLAs
- Provide for a secure, safe and privacy-aware end-to-end operation



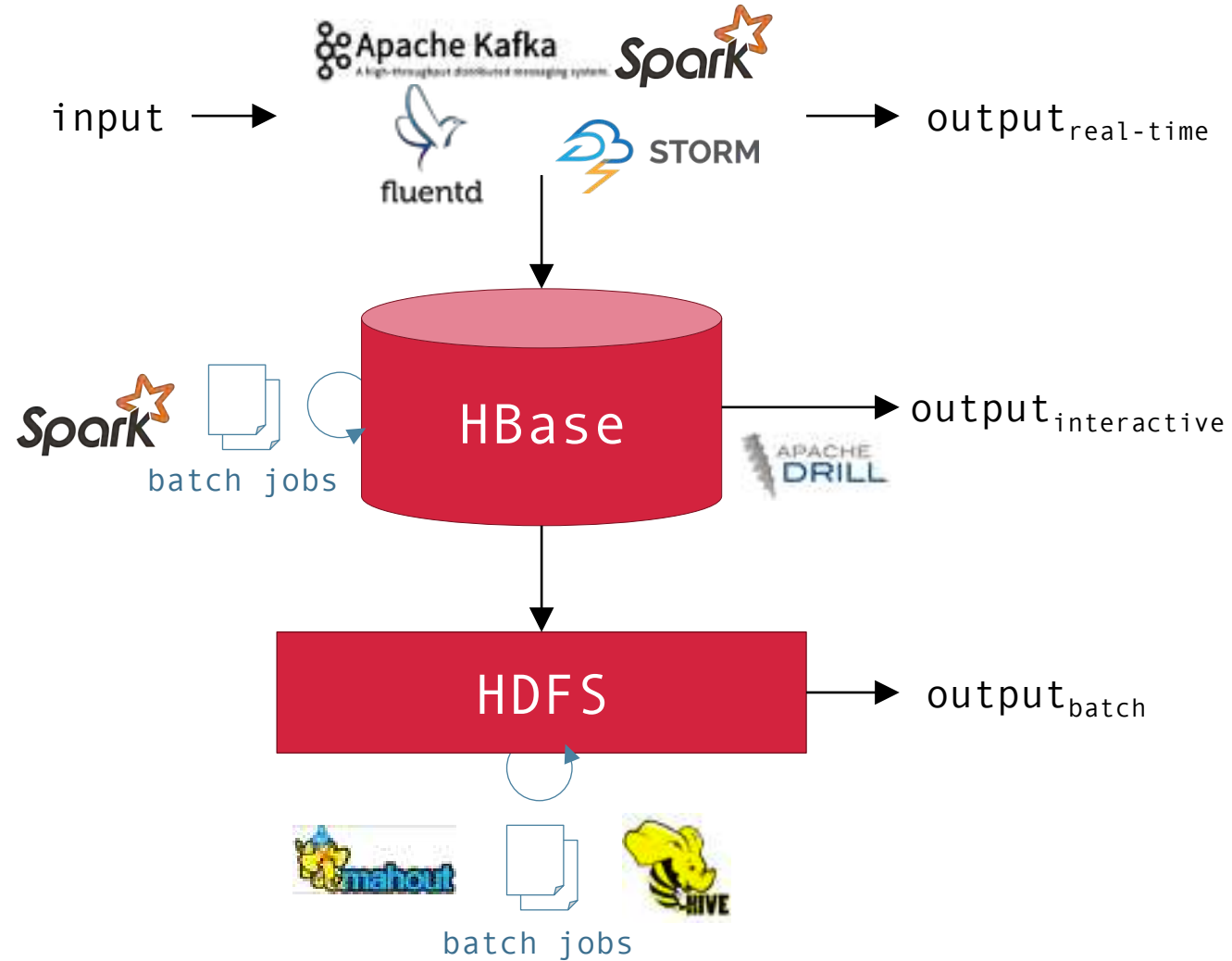
The IoT architecture (iot-a)



<http://iot-a.info/>



Example iot-a



Apache Kafka

- High-throughput, distributed, persistent publish-subscribe messaging system
- Typically used together with Storm/Spark for online stream processing



<http://kafka.apache.org/>



Apache Storm

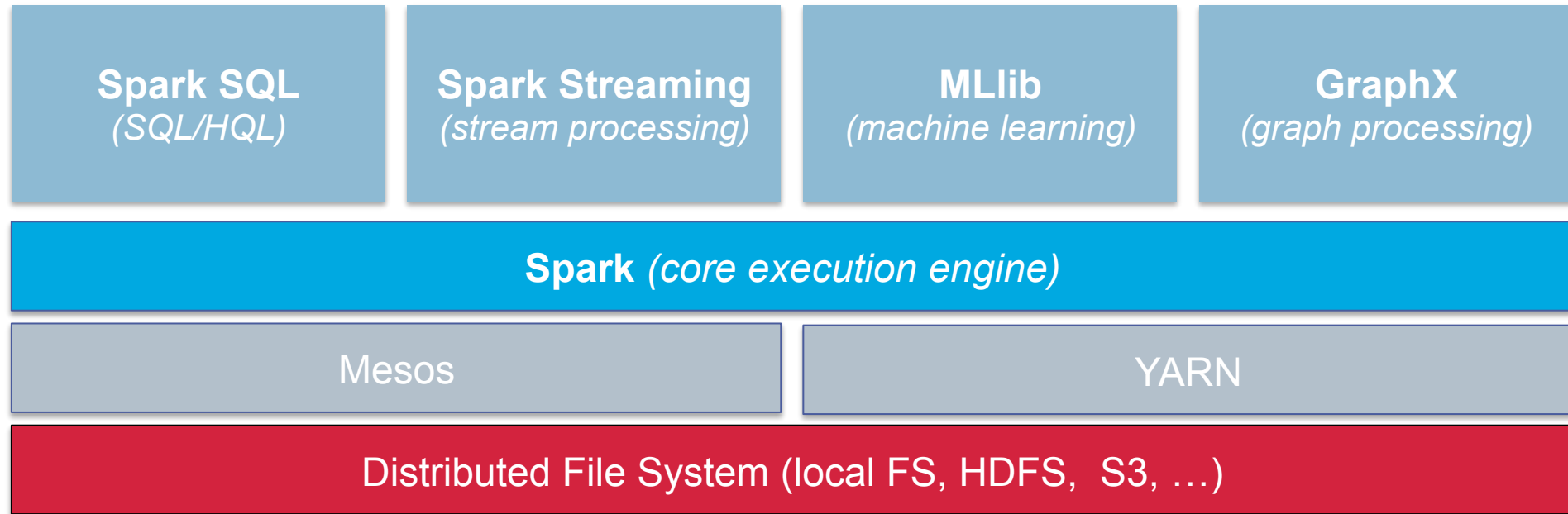
- Distributed, fault-tolerant stream-processing platform
- Guaranteed message processing; takes care of replaying messages on failure
- Concepts: tuples, streams, spouts, bolts, topologies



<http://storm.apache.org/>



Apache Spark



<https://spark.apache.org/>

Apache HBase

- Distributed, column-oriented database built on top of HDFS
- Based on Google's BigTable technology
- Able to scale horizontally to 1,000s of commodity servers, petabytes of data with low-latency get/put ops



<http://hbase.apache.org/>



Managing Time Series at Scale



Stream data sources

- physical sources such as IoT devices
- social media streams such as Twitter firehose



Stream data sources

What about development and testing?

- synthetic sources
 - <https://github.com/tdunning/log-synth>
 - <https://github.com/mapr-demos/gess>
 - <https://github.com/mapr-demos/direhose>



OpenTSDB

OpenTSDB is a distributed Time Series Database on top of HBase, enabling you ...

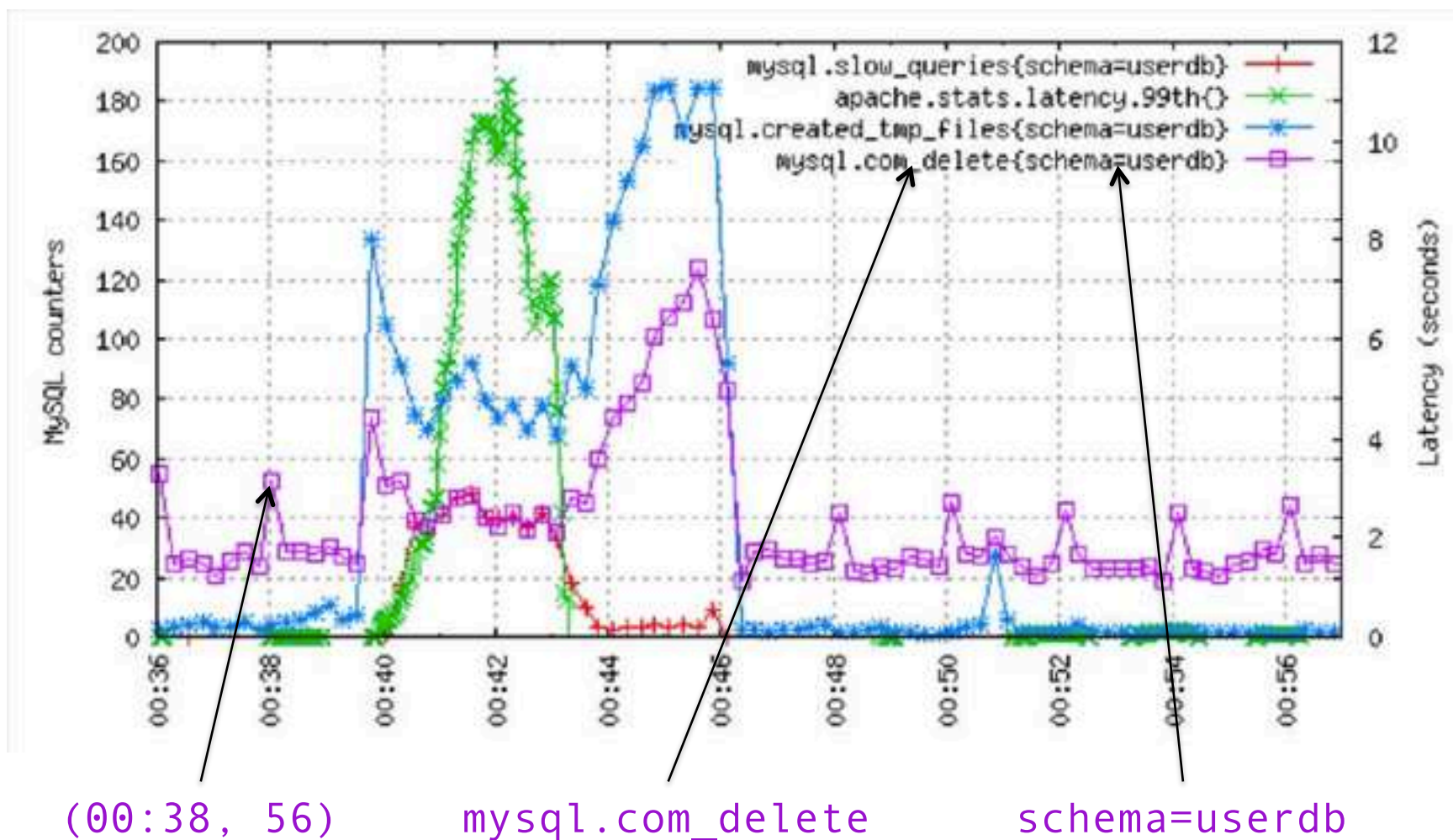
- to store & index, as well as
- to query & plot

... metrics at scale.

<http://opentsdb.net/>



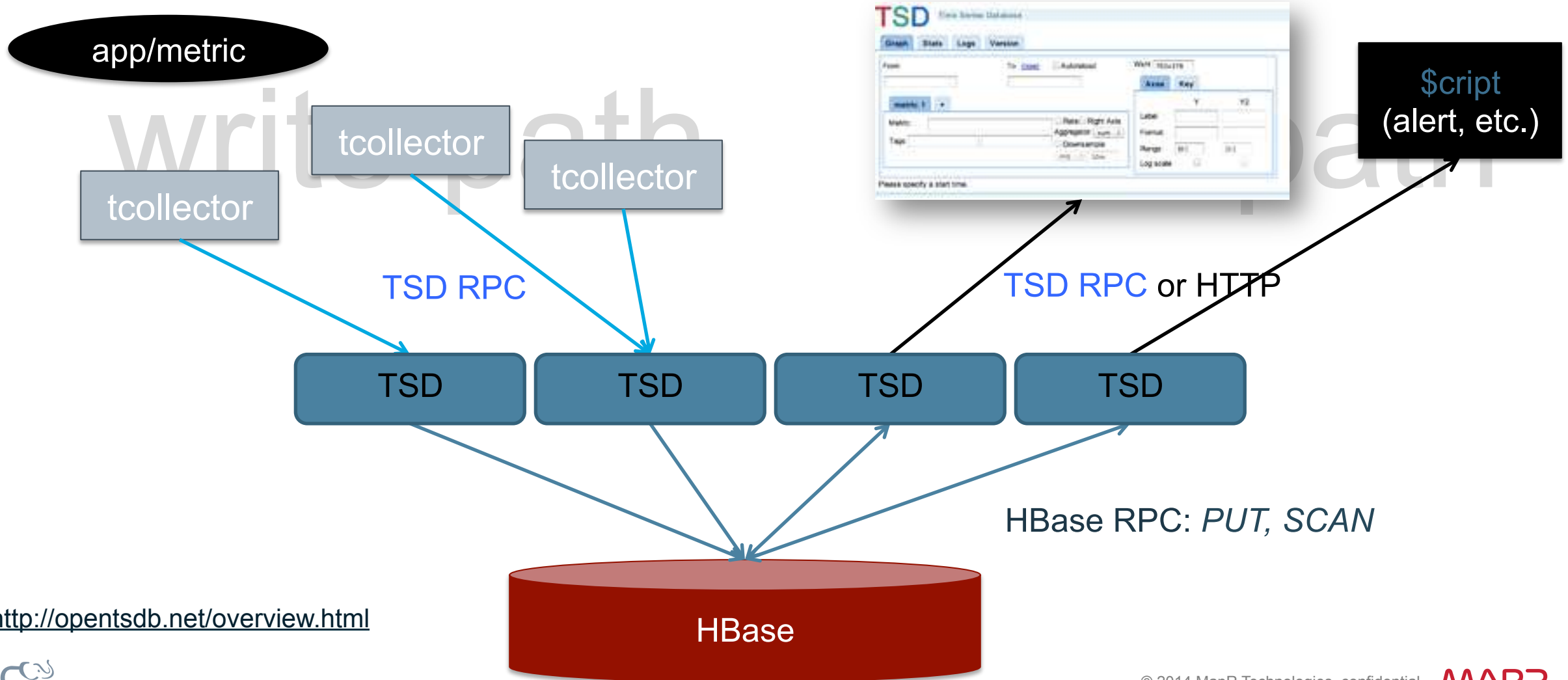
OpenTSDB: key concepts



data point: (timestamp, value) + metric + tag: key=value → time series



OpenTSDB: high-level architecture



<http://opentsdb.net/overview.html>



OpenTSDB: collecting metrics

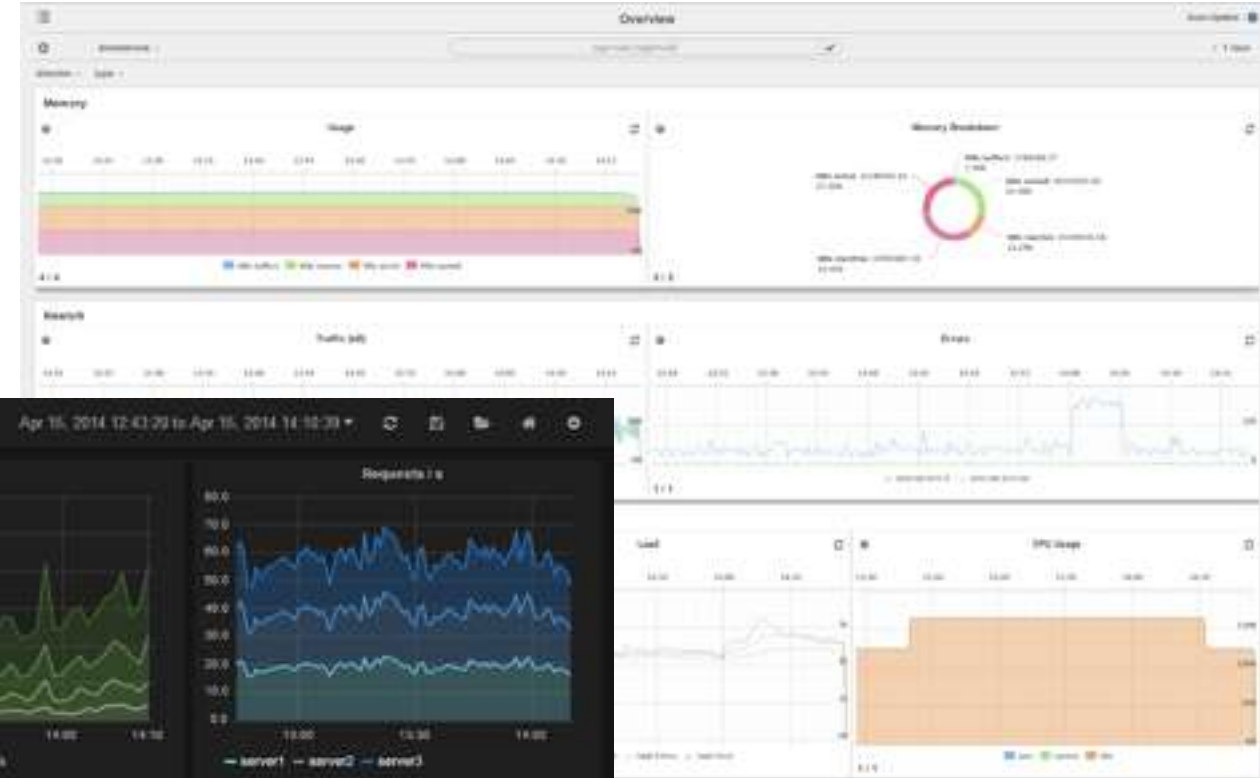
- tcollector: gathers data from local collectors, pushes to TSDs and providing deduplication
- lots bundled
 - General: iostat, netstat, etc.
 - Others: MySQL, HBase, etc.
- ... or roll your own

couchbase.py	Removed err function from individual collectors
dflstat.py	dflstat.py: python2.4 compatibility support
elasticsearch.py	Removed err function from individual collectors
graphite_bridge.py	Add a collector bridge for the basic graphite protocol.
hadoop_datanode.py	Permissions change of hadoop_datanode.py and hadoop_namenode.py
hadoop_namenode.py	Permissions change of hadoop_datanode.py and hadoop_namenode.py
haproxy.py	Move hbase and hadoop to http /jmx collection
hbase_master.py	Move hbase and hadoop to http /jmx collection
hbase_regionserver.py	hbase_regionserver.py: fix region metrics splitting
ifstat.py	ifstat.py: code cleanup
iostat.py	iostat.py: code cleanup
mongo.py	Remove code duplication by introducing a utils library.
mysql.py	Move hbase and hadoop to http /jmx collection
netstat.py	netstat.py: handle multiple sets of data for the same data type, e.g.,...
nfsstat.py	Make nfs stats per version, and support v3.
opentsdb.sh	Move copyright info to AUTHORS, and add a THANKS file.
postgresql.py	Move hbase and hadoop to http /jmx collection
procnettcp.py	Remove code duplication by introducing a utils library.
procstats.py	NUMA stat files have to be opened/closed each time to get fresh stats
redis-stats.py	Redis: add per database metrics

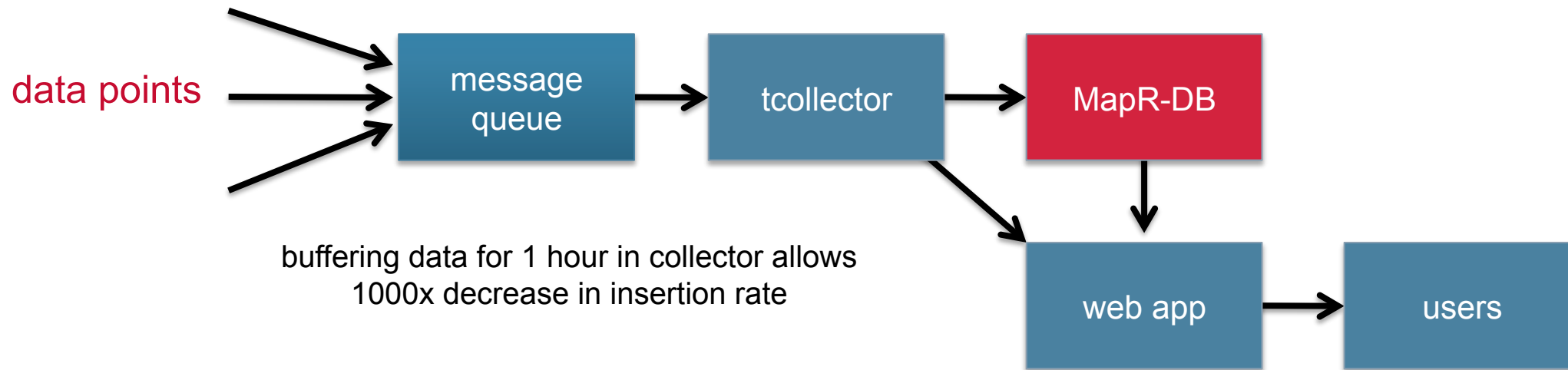


OpenTSDB: interfacing

- HTTP API
- CLI (tsd, query, mkmetric, etc.)
- Java lib: asynchbase
- Dashboards (Grafana, etc.)



OpenTSDB with MapR (HBase \leftrightarrow MapR-DB)



<https://github.com/mapr-demos/opentsdb>

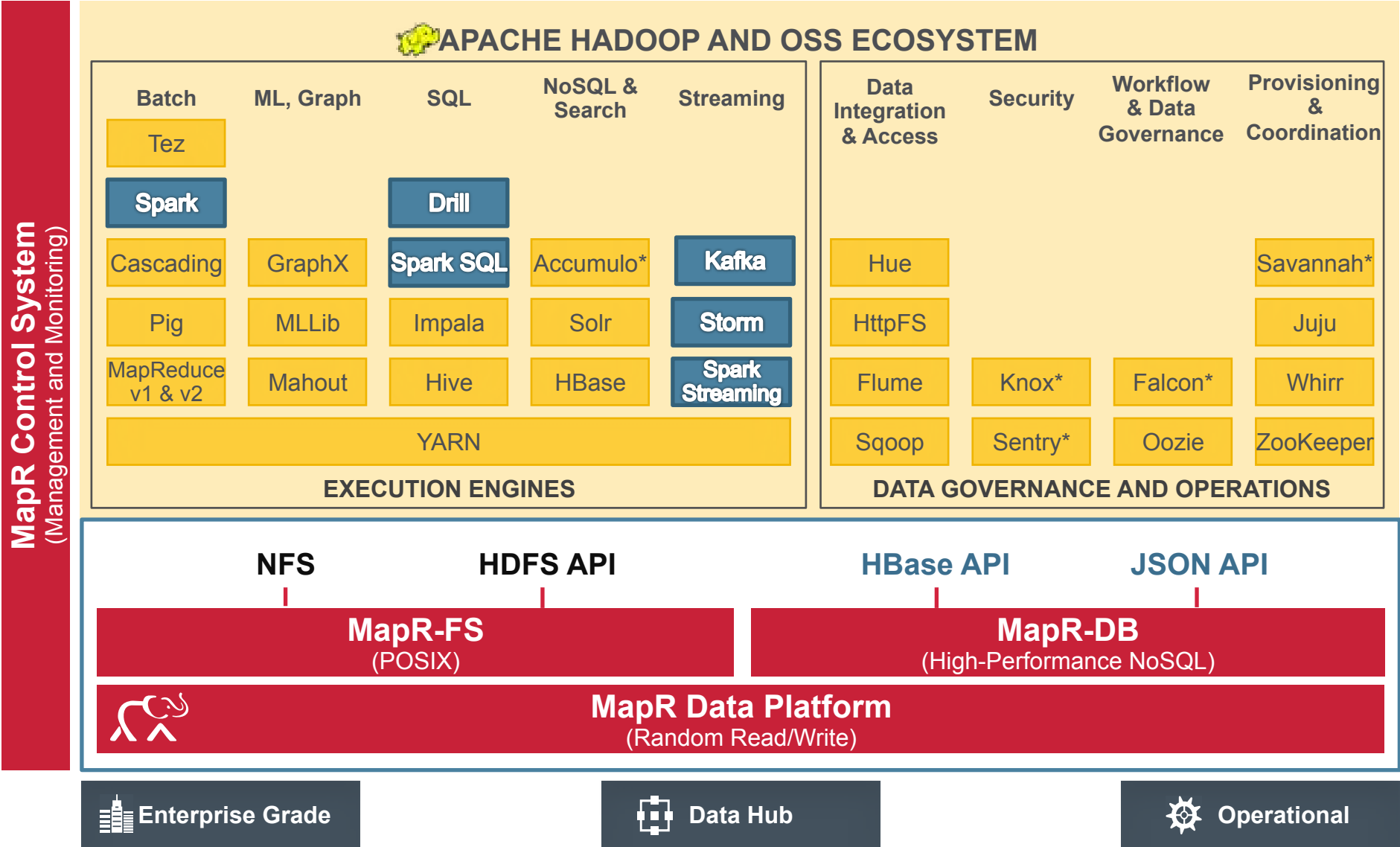


Alternative TSDB for smaller scales

- InfluxDB (written in Go)
- lots of client libs
- (cluster support via Raft)
- powerful query language

```
select mean(value), percentile(90, value) as percentile_90
from /^stats.*/
group by time(10m)
into 10m.:series_name
```

MapR's IoT offering



idential

MAPR

Q&A

Engage with us!

@mhausenblas



maprttech

mapr-technologies



MapR

mhausenblas@mapr.com



maprttech

