MAPR

Time-Series Databases and Machine Learning

Jimmy Bates

November 2017

MAPR®

Top-Ranked **Hadoop** Distribution

1. Read Write File System
2. Ease of Data Ingestion
3. World Record Performance
4. Complete Data Protection
5. High Availability
6. Real Multi-tenancy
7. Enterprise-grade Security
8. Unbiased Open Source

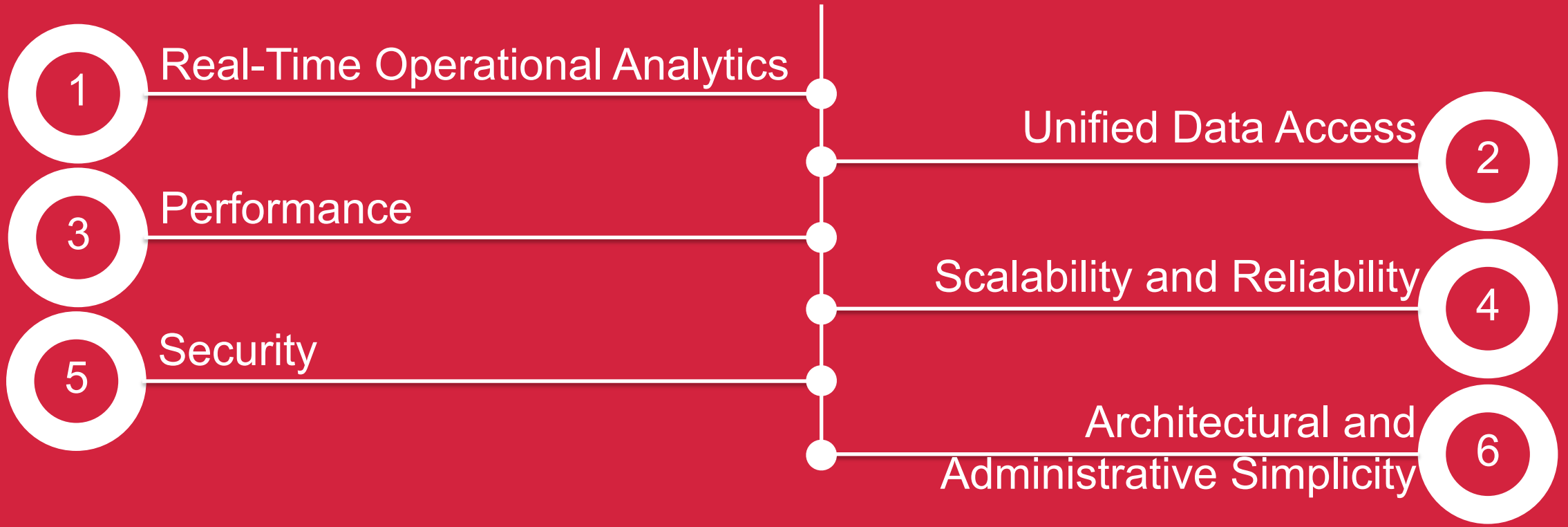
MAPR®

Top-Ranked **Hadoop** Distribution

Top-Ranked **NoSQL**

MAPR®

Top-Ranked **NoSQL**



MAPR®

Top-Ranked **Hadoop** Distribution

Top-Ranked **NoSQL**

Top-Ranked **SQL-on-Hadoop** Solution

MAPR®

Top-Ranked **SQL-on-Hadoop** Solution

1. The Most SQL Options on Hadoop
2. Including our work with Apache Drill

MAPR®

Agenda

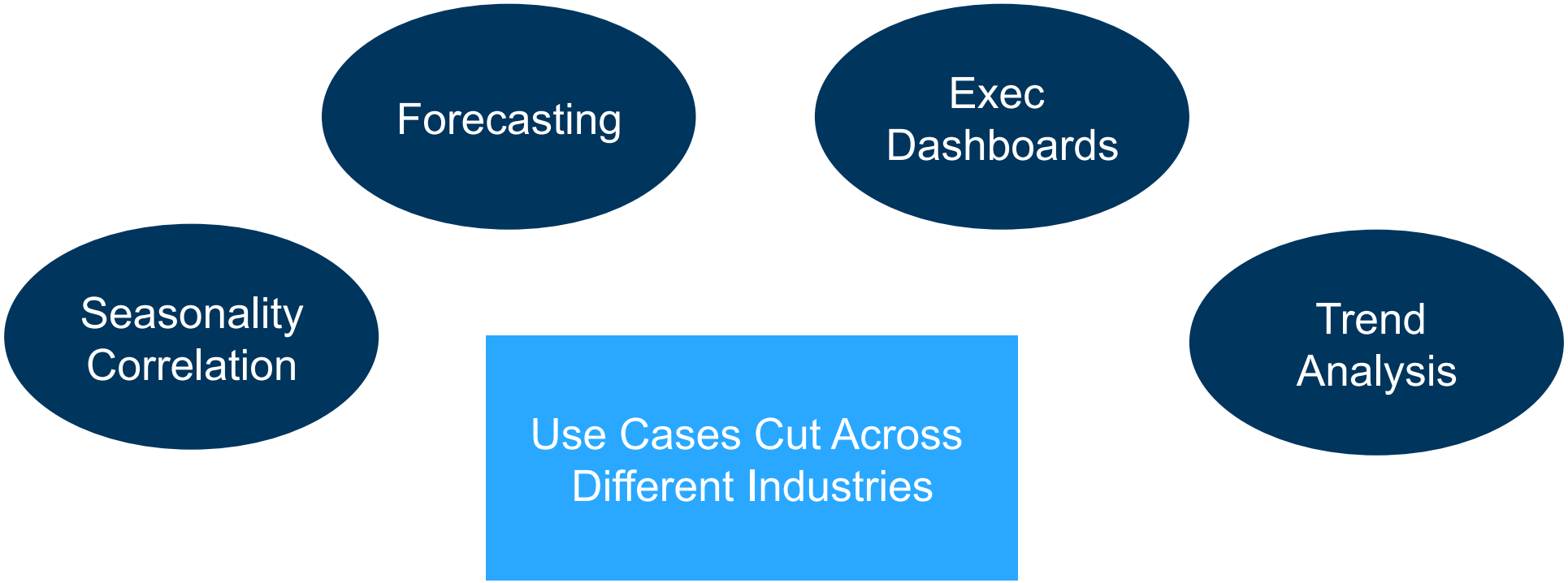
1. What is Time Series
2. A Basic Workflow
3. The Data Science User
4. A Quick Look at OpenTSDB
5. Resources for Getting Started

Time Series Data

What is Time Series Data?

- Any set of data-points that have time associated with it.
- Examples of time series data include
  - weather,
  - web clickstream,
  - product sale,
  - stock trade,
  - machine logs,
  - fleets,
  - sensors,
  - devices,
  - network traffic,
  - system login

Value of Time Series Analytics



IOT Sensor Data

Operational Behavioral Analysis

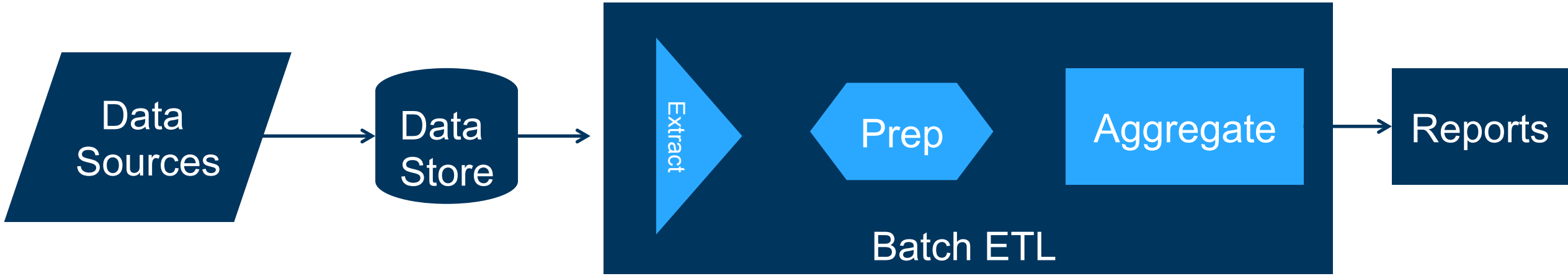
Preventative Maintenance

Environmental Monitoring (Climate)

Cell Tower Data Analysis

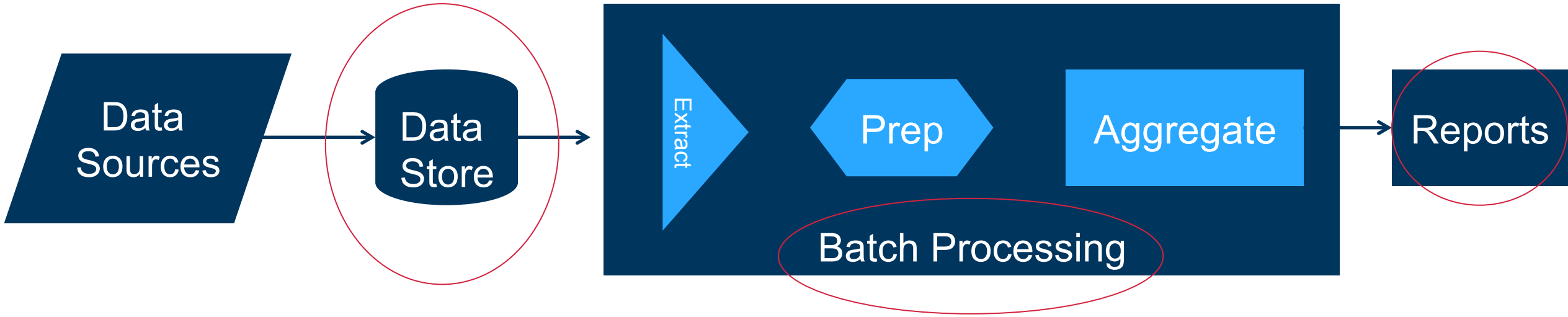
Stock Trade Data Analysis

Typical Time Series Analytics Flow



Environment: IT teams/Application Development groups/Analytics groups
Persona: Enterprise Architect, NoSQL Developer, Database developer, Analyst
Task: Conduct analyses or provide reports to management

Typical Time Series Analytics Flow - Challenges



- Larger Volumes of data from newer sources
- Expensive to store

- Takes a lot of time to process information and do simple aggregations

- Not Real-time

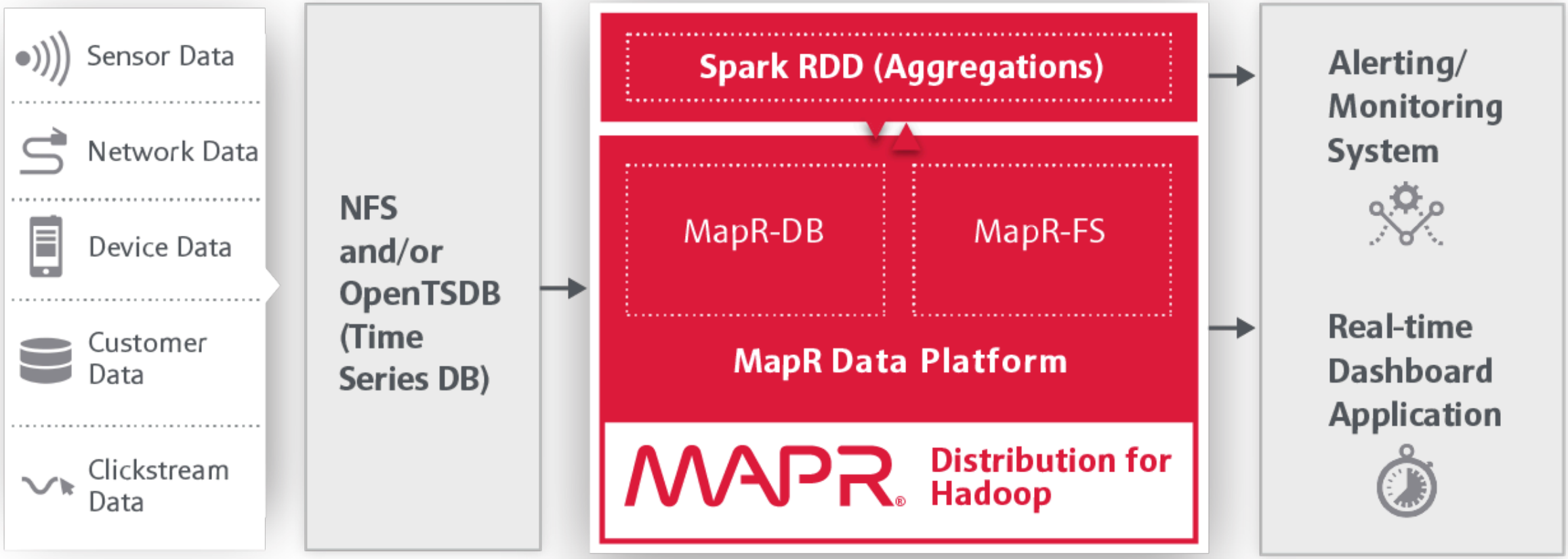
Challenges with Status Quo

- Existing systems aren't well suited to store high volumes of time-series data
- Building aggregations and statistical computations from time-series data is currently done in batch

Need:

- Cost effective and reliable way to store and analyze large amounts of time series data from various sources.
- Ability to deploy real-time dash-boarding and monitoring capabilities on aggregated data

Time Series Solution Example



MapR Data Exploration Advantages for IoT

Self-Service Data Exploration

Single SQL Interface for Structured and Semi-Structured Data

Data Agility with Less IT Required

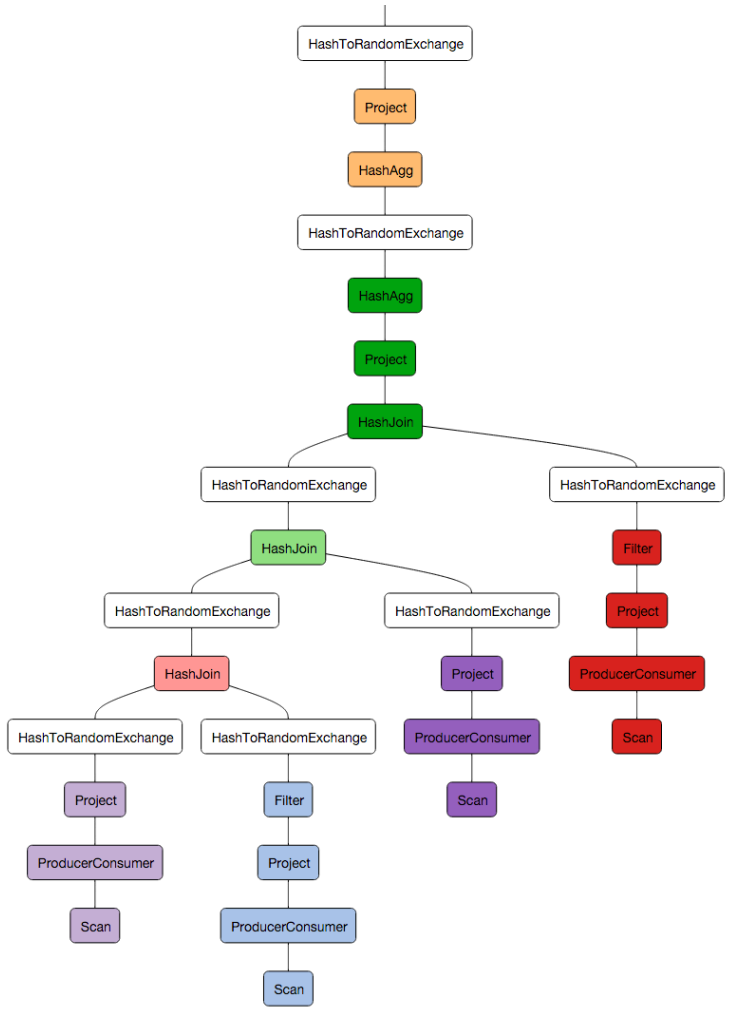
Squaring the Circle

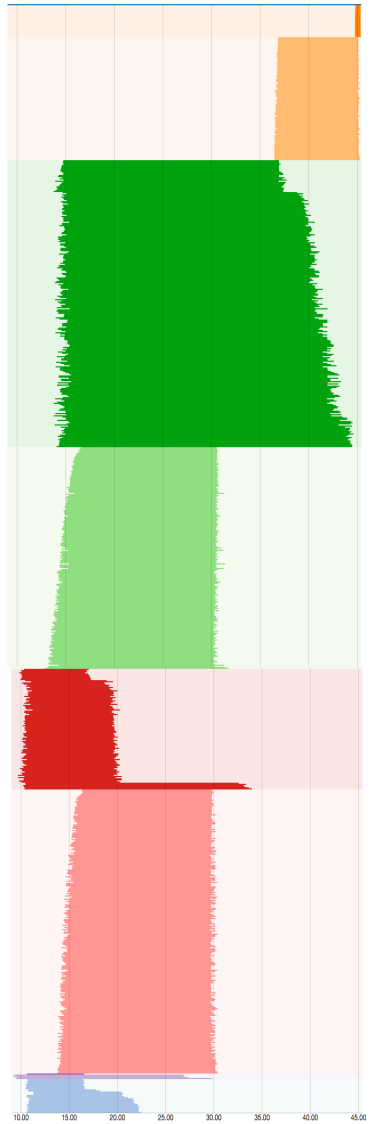
- Enter Apache Drill

- Drill is SQL *compliant*
  - Uses standard syntax and semantics

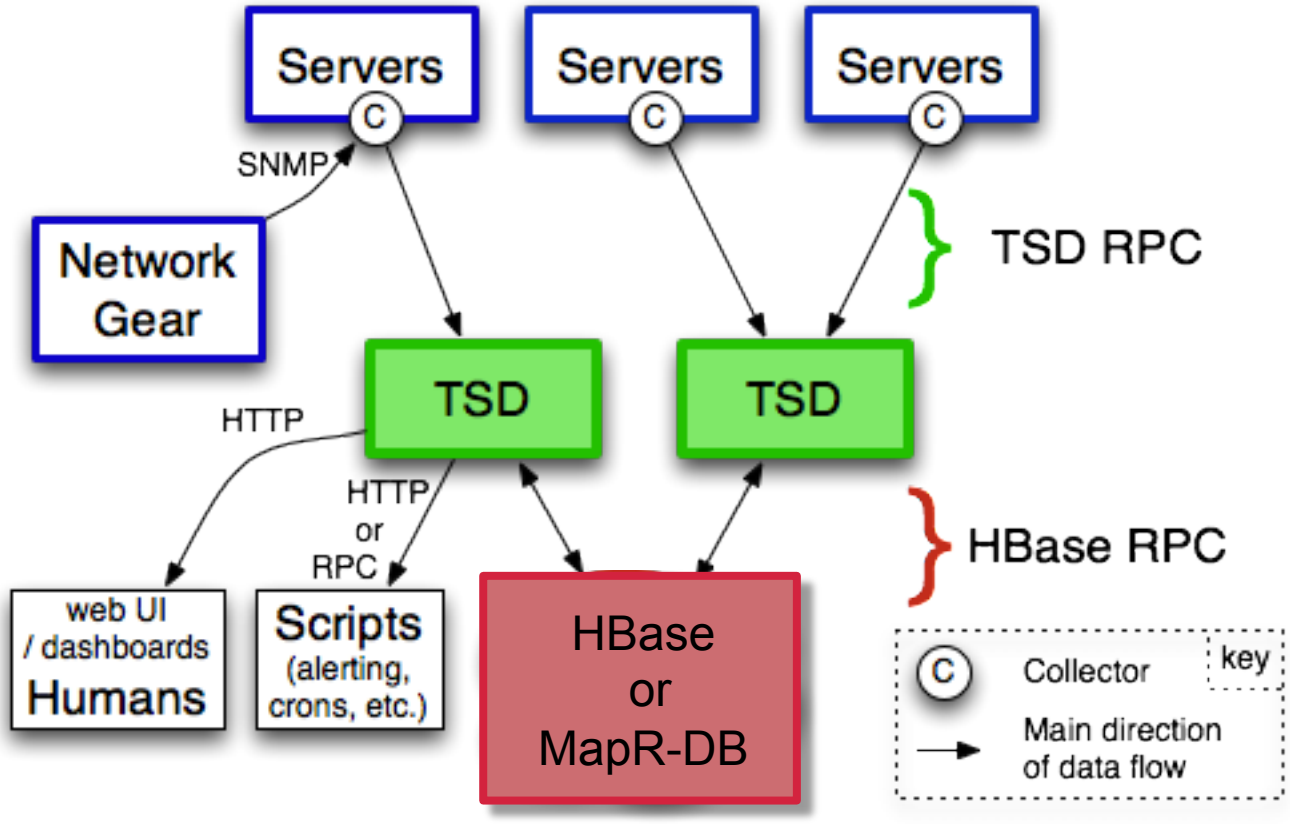
- Drill extends SQL
  - First class treatment of objects, lists
  - Full support for destructuring, flattening
  - Full power of relational model can be applied to complex data

Drill Provides Scalable and Extended SQL



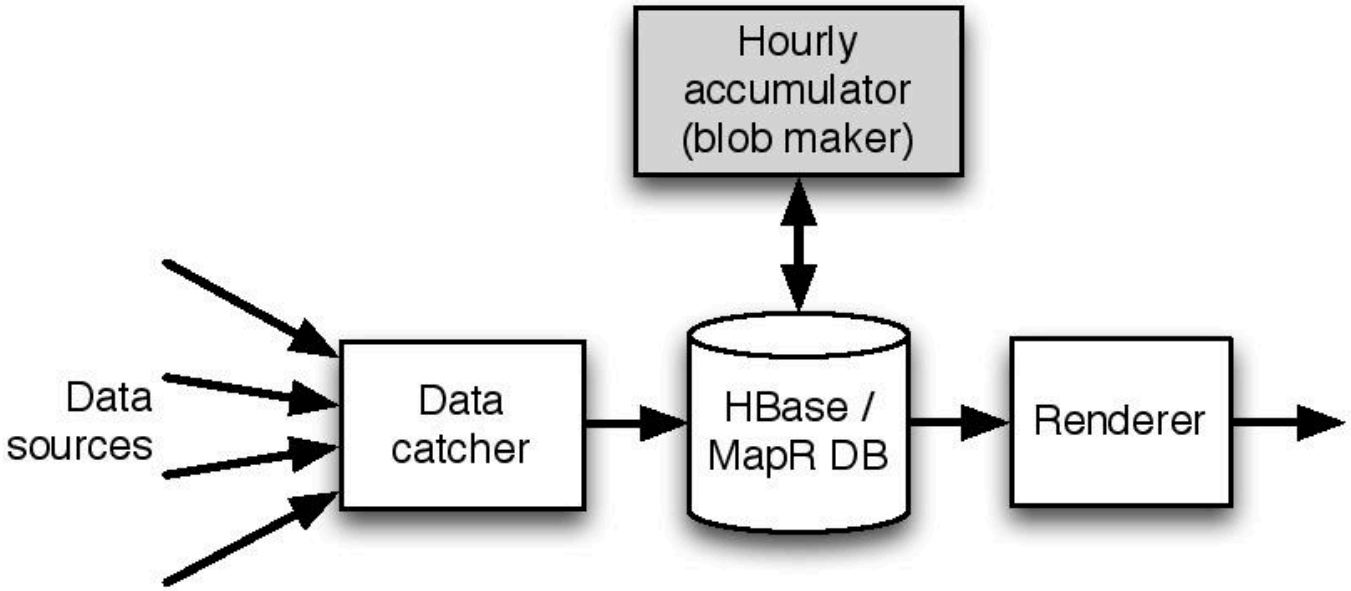


Introduction to Open TSDB



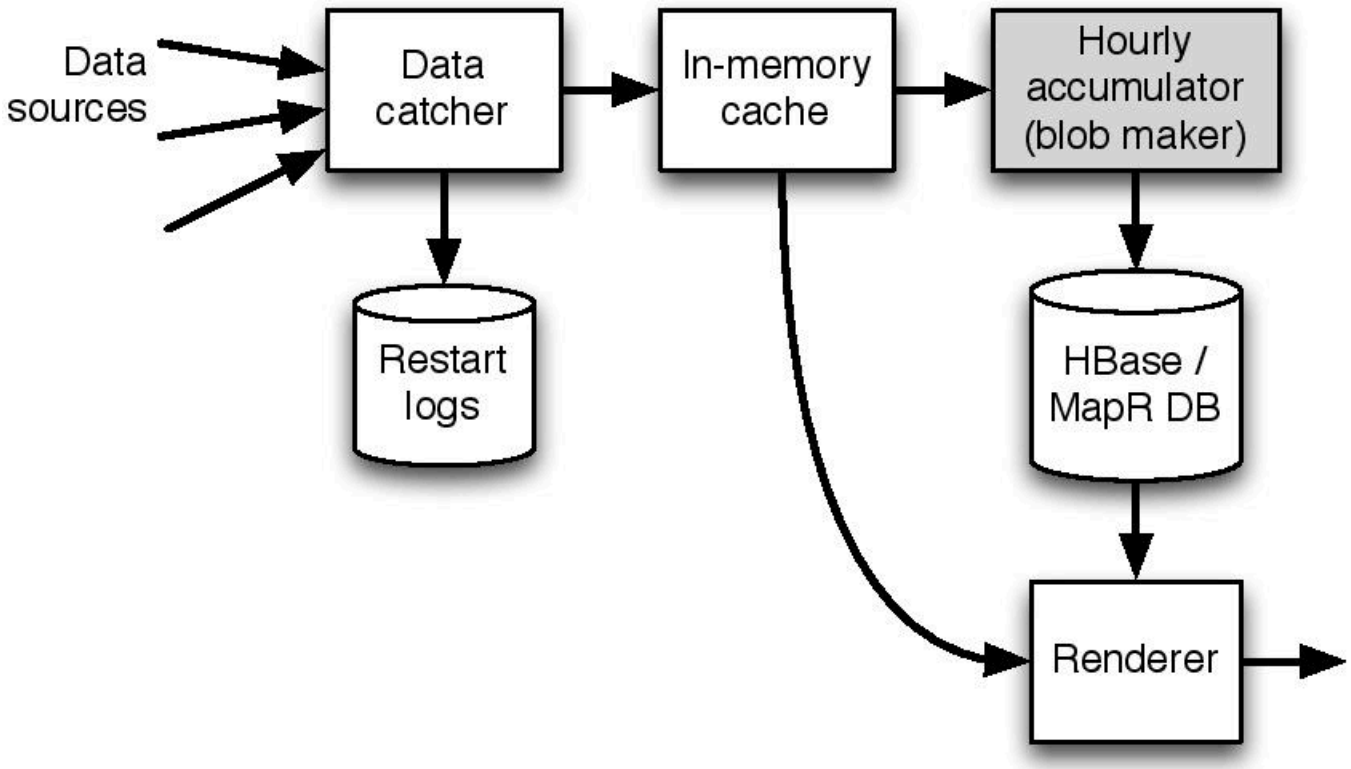
Speeding up OpenTSDB

Why can’t it be faster ?



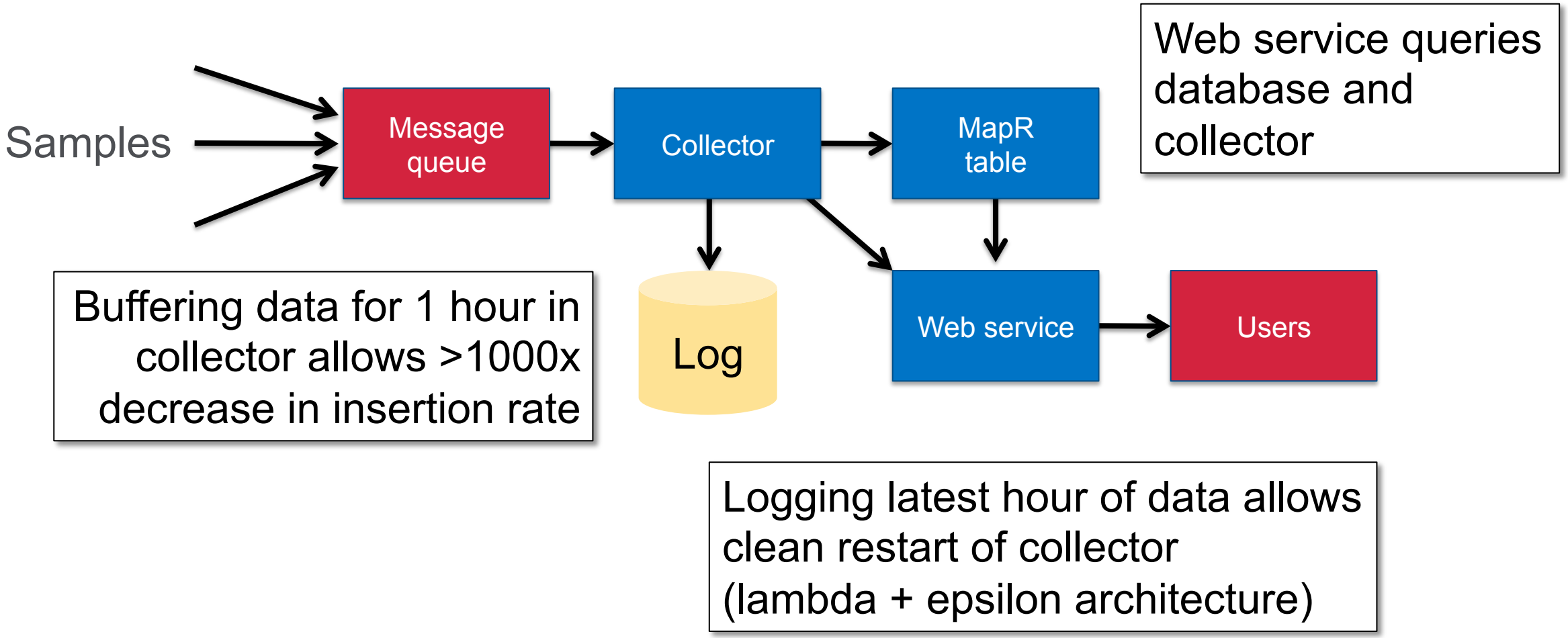
20,000 data points per second per node in the cluster

Speeding up OpenTSDB: open source MapR extensions

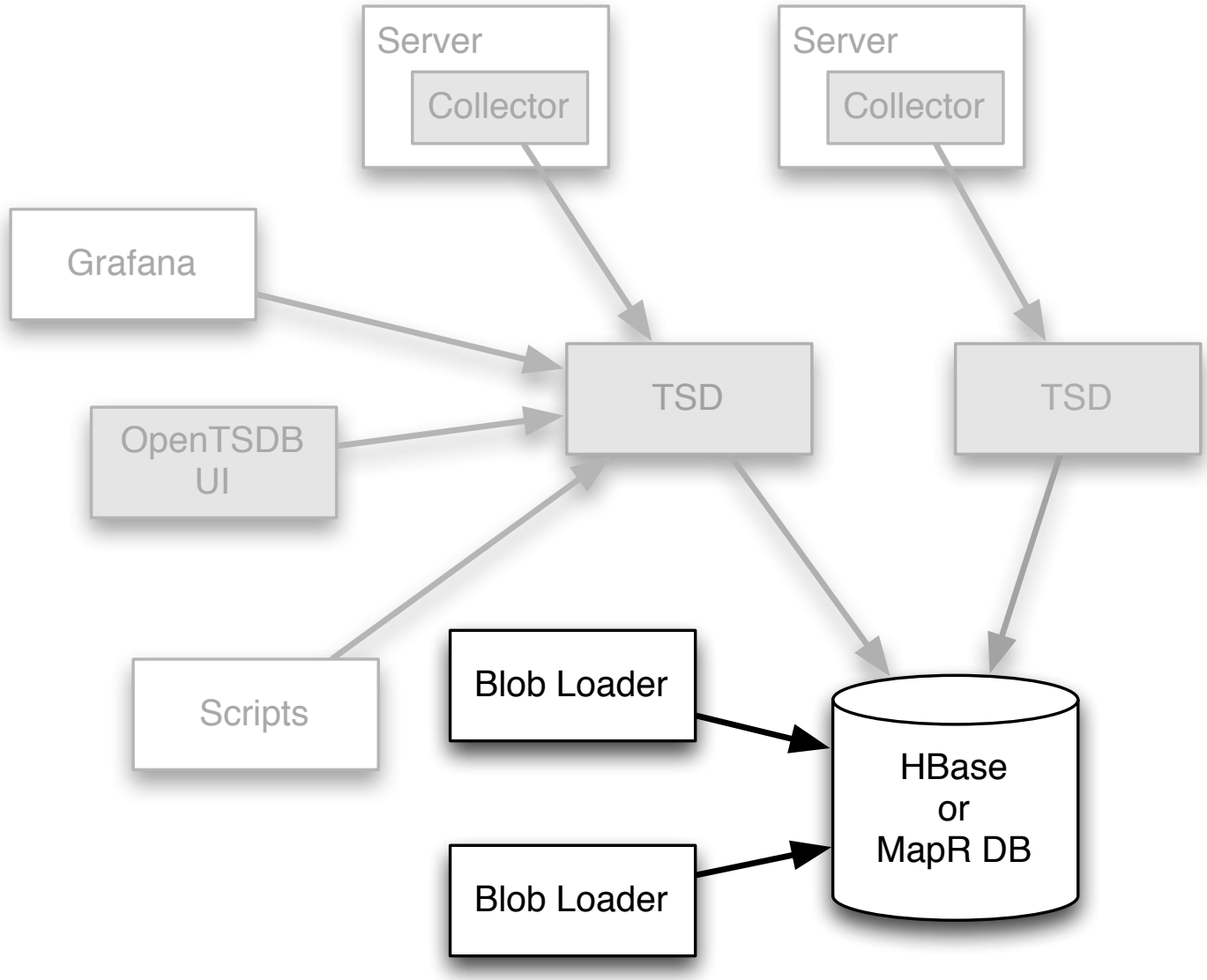


Available on Github: https://github.com/mapr-demos/opentsdb

Speeding up OpenTSDB: open source MapR extensions



Direct Blob Loading for Testing



A first example: Time-series data

Column names as data

- When column names are not pre-defined, they can convey information

- Examples
  - Time offsets within a window for time series
  - Top-level domains for web crawlers
  - Vendor id's for customer purchase profiles

- Predefined schema is impossible for this idiom

Relational Model for Time-series

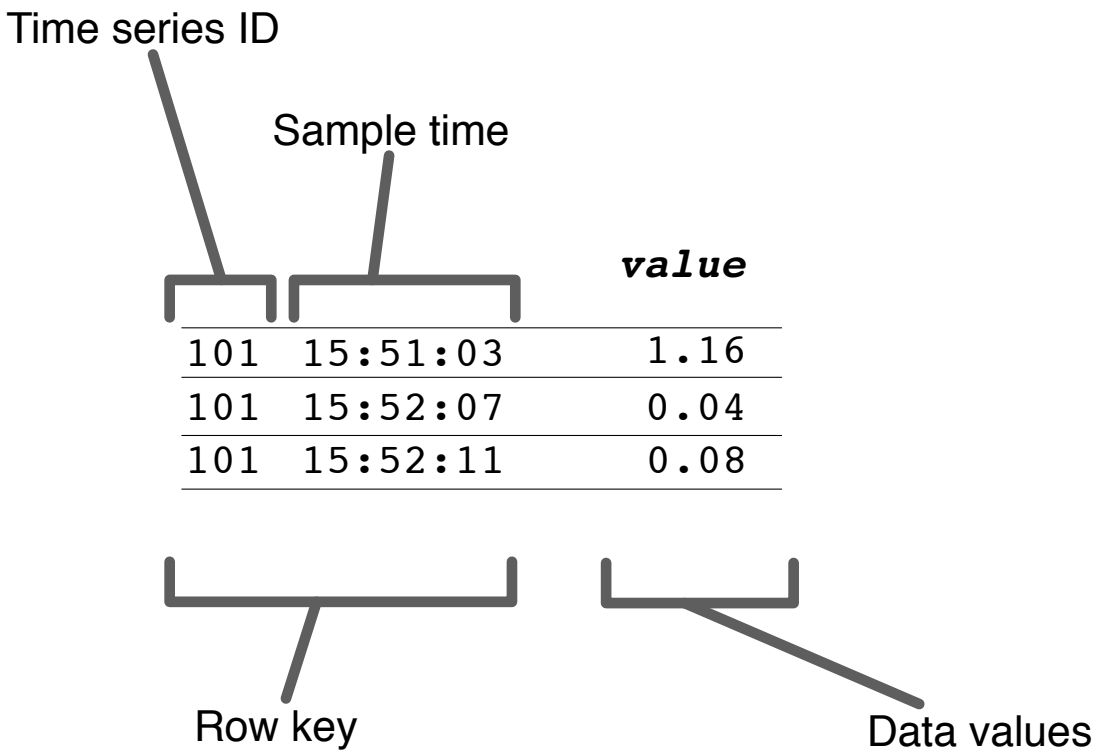


Table Design: Point-by-Point

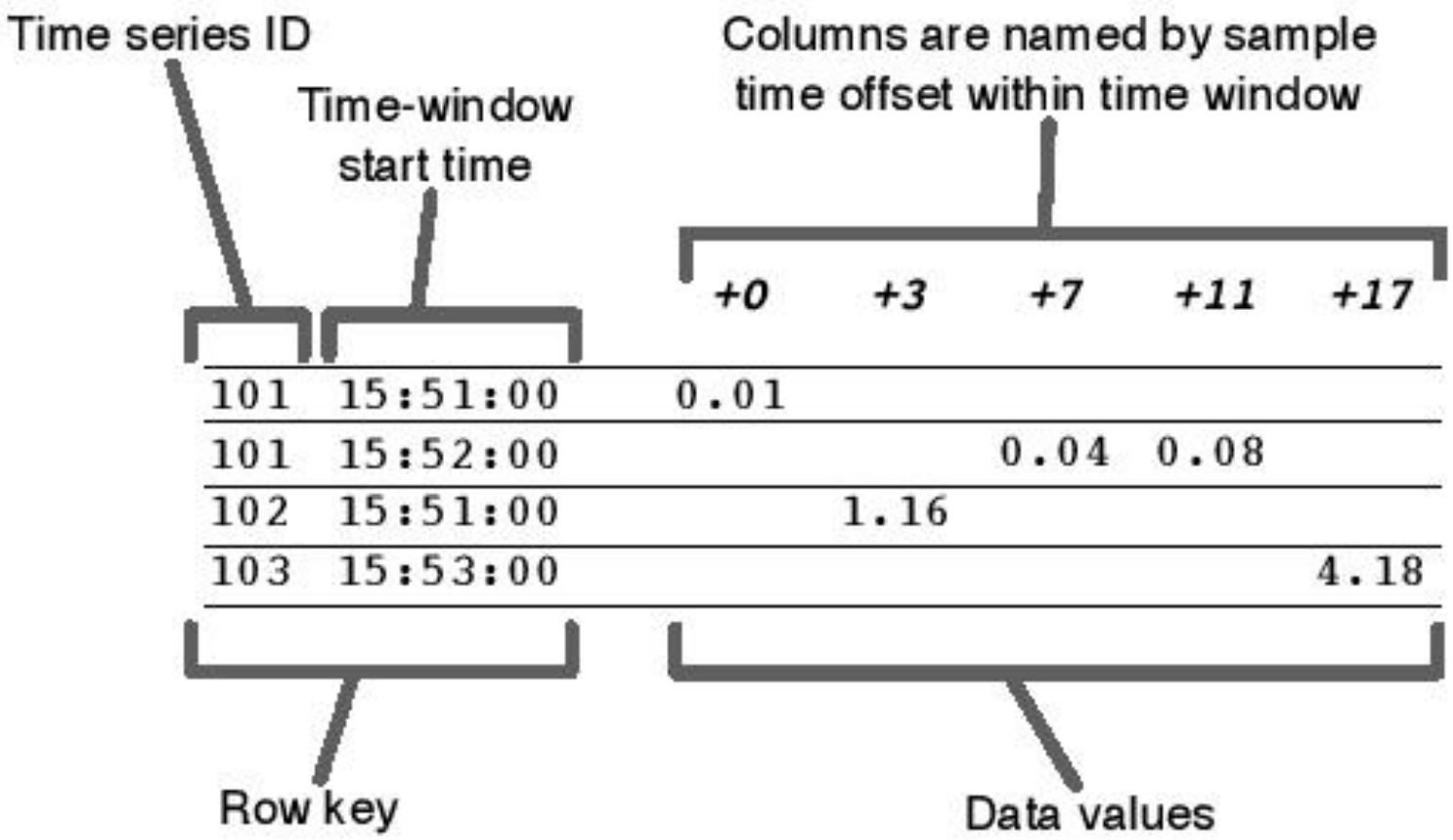
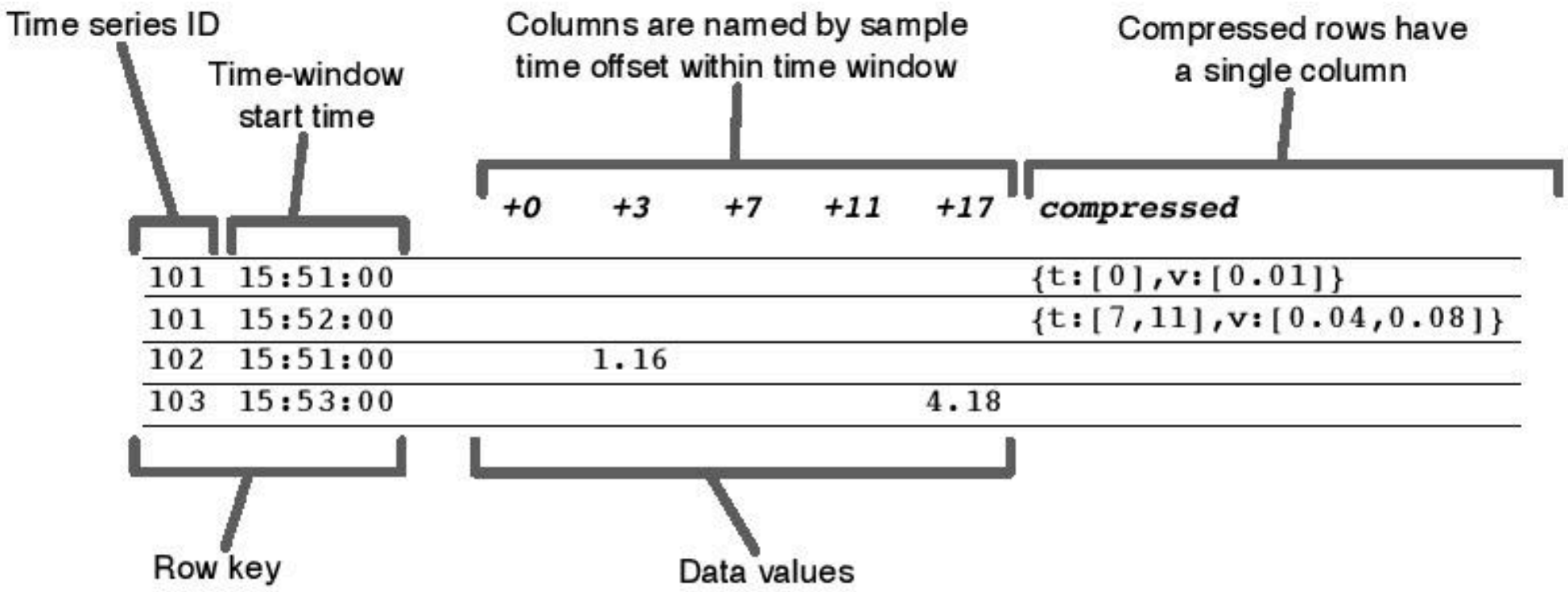


Table Design: Hybrid Point-by-Point + Sub-table



After close of window, data in row is restated as column-oriented tabular value in different column family.

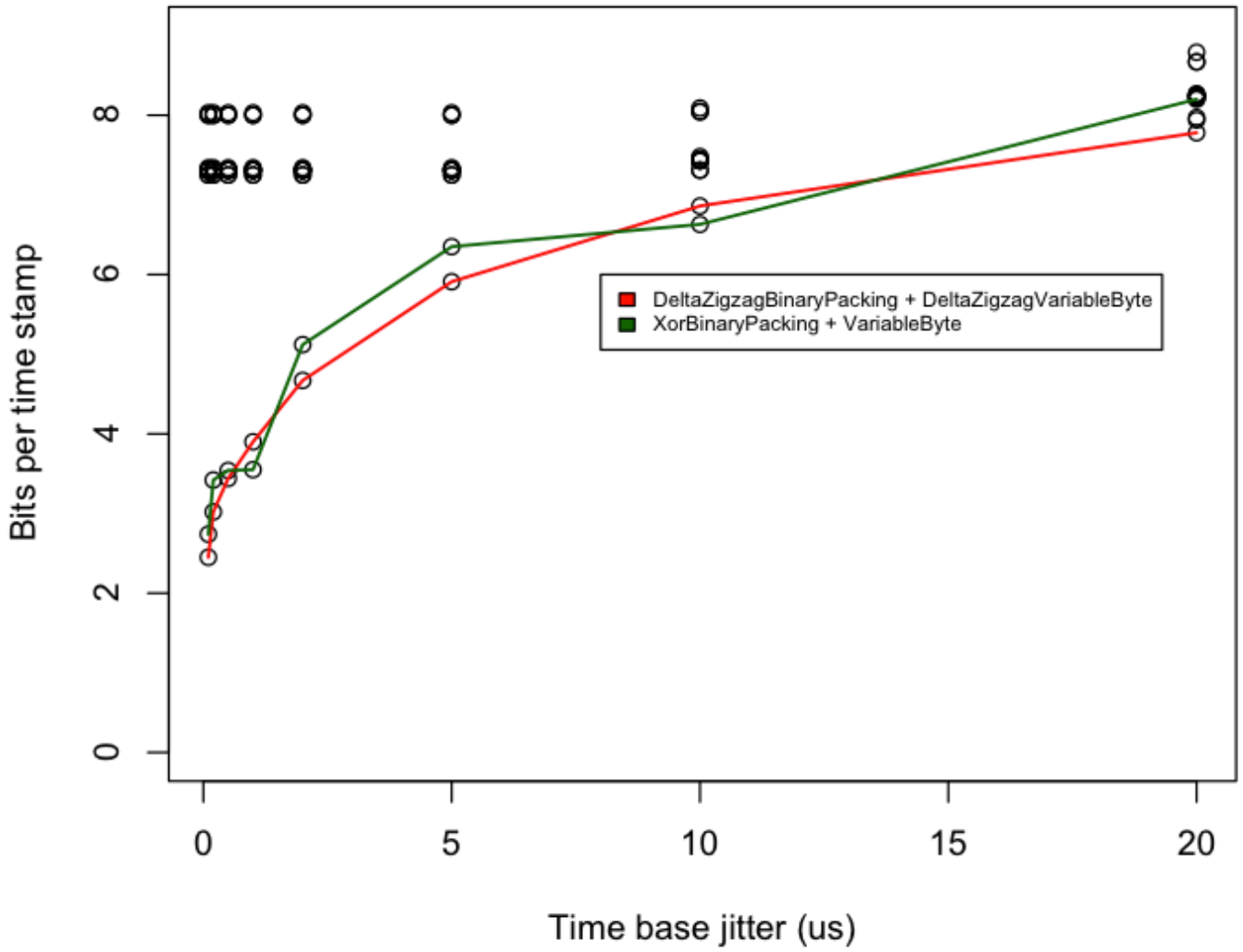
Compression Results

Samples are
64b time, 16 bit sample

Sample time at 10kHz

Sample time jitter makes it important to keep original time-stamp

How much overhead to retain time-stamp?



Insertion Speeds

- Inserting pre-bundled data allows humongous data rates

- >100 M s/s on 4 nodes
- >200 M s/s on 8 nodes

- Linear scaling if enough data sources

How fast is OpenTSDB ?

- OpenTSDB can scale to writing millions of data points per 'second' on commodity servers with regular spinning hard drives
- What if we needed 100-1000x faster writing speed ?

Problem: How Do Load Test Data at Large Scale?

- Testing at large scale is a more realistic measure of performance than on a small sample
- But with high velocity data, how do you set up test?
  - For a sample equivalent to long term data, you need ingest rates of 100 to 1000x faster than normal production

    - OR-

  - You must wait years to load up test data.

What's the solution?

The need for rapid data loading

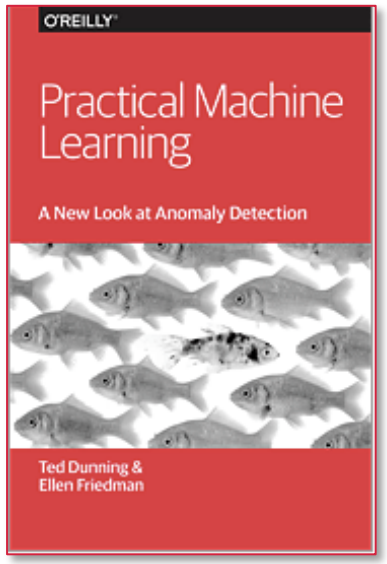
- Let's say you have 1M samples / second
- Suppose you want to *test* your system
- Perhaps with a year of data
- And you want to load that data in << 1 year

- 100x real-time = 100M samples / second

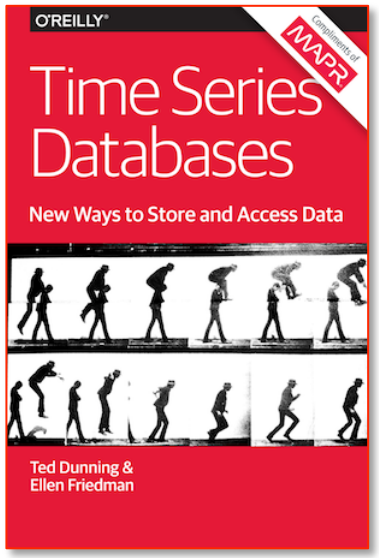
Short Books by Ted Dunning & Ellen Friedman

- Published by O'Reilly in 2014 and 2015
- For sale from Amazon or O'Reilly
- Free e-books currently available courtesy of MapR

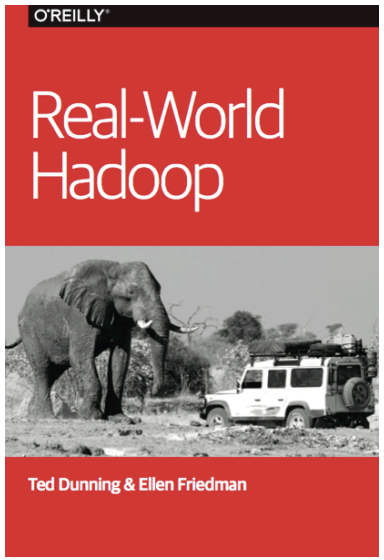
http://bit.ly/recommendation-ebook



http://bit.ly/ebook-anomaly



http://bit.ly/mapr-tsdb-ebook



http://bit.ly/ebook-real-world-hadoop

Thank You

@mapr   maprtech

mapr-technologies   MapRTechnologies

jbates@mapr.com   maprtech