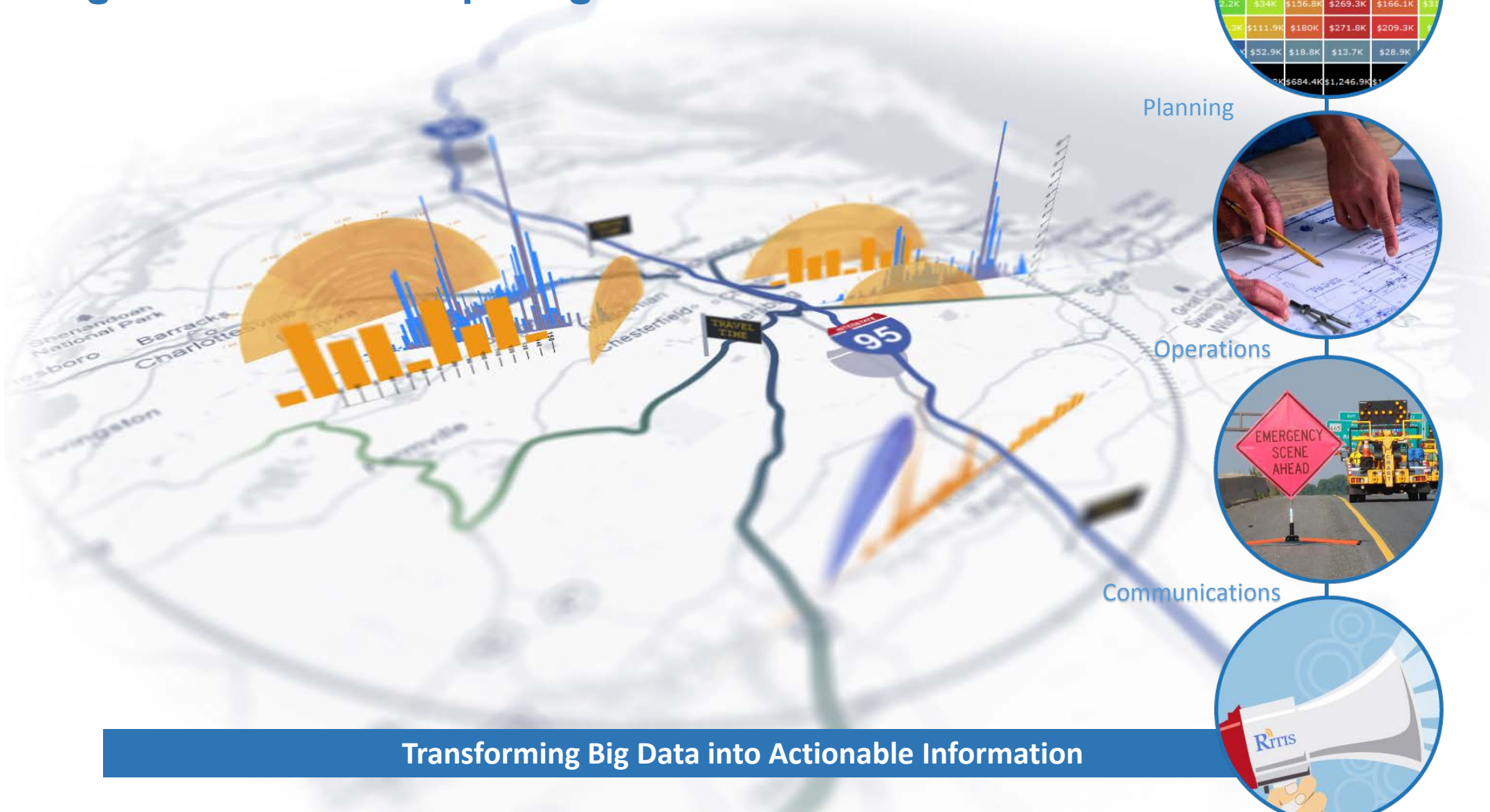
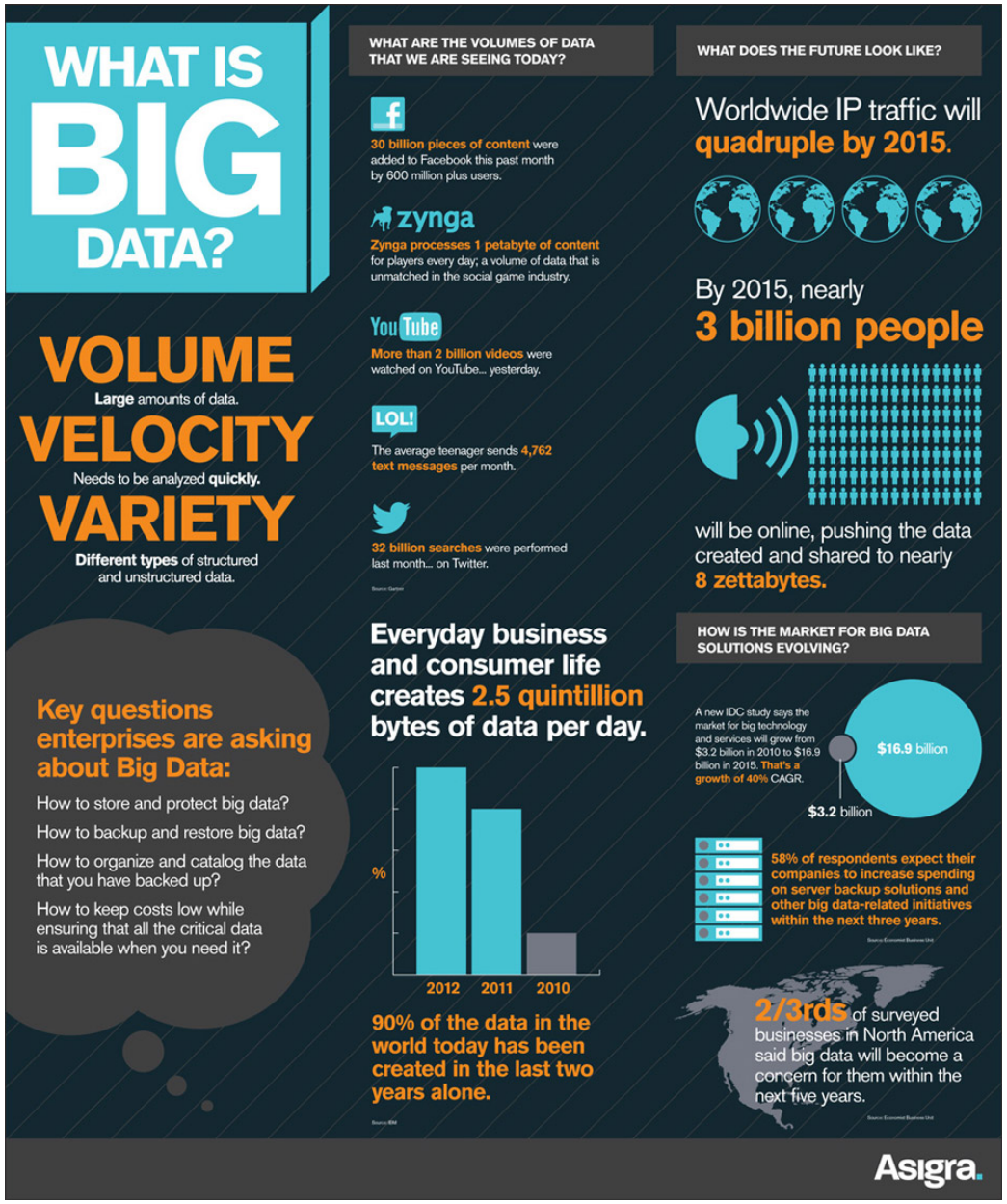
Storing and Analyzing Transportation Big Data using Distributed Computing



Transforming Big Data into Actionable Information

What is big data?



In 1992 global internet traffic amounted to 100 GB per day.

In 2019 global internet traffic will be ~52,000 GB per second!

Photo by: BBVATech

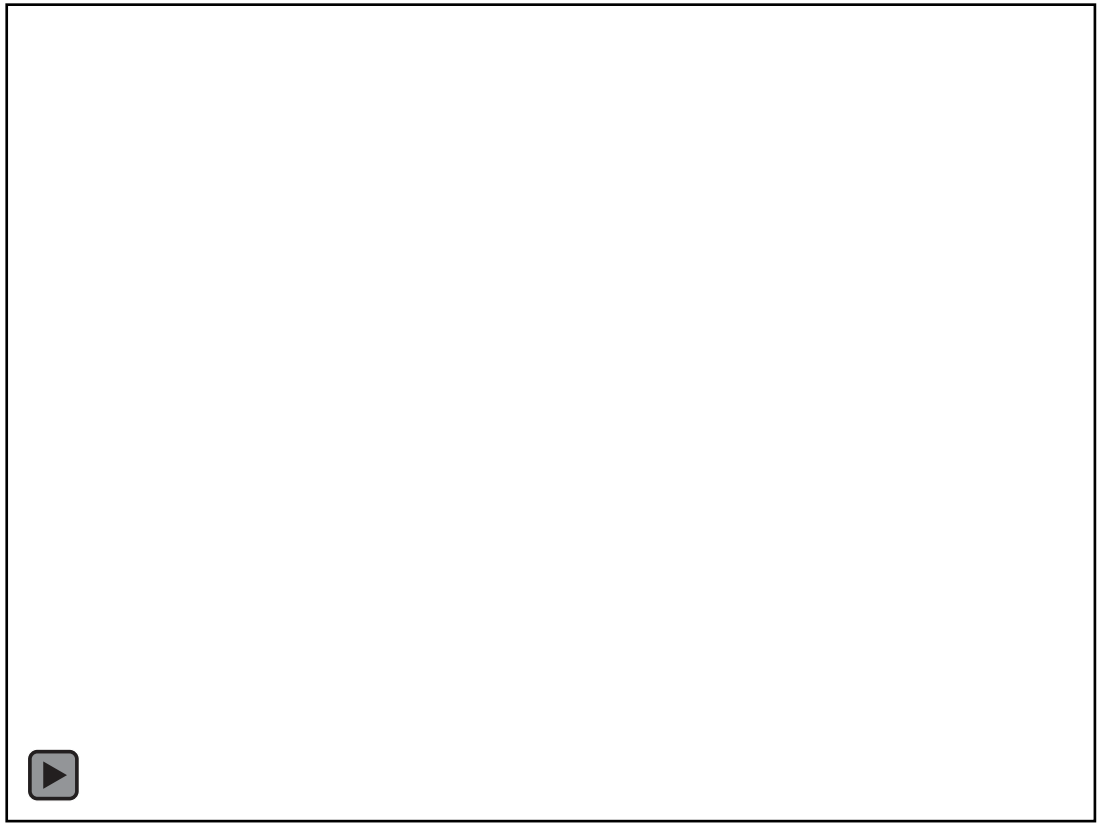
Source: Cisco VNI, 2015

What does transportation big data look like?

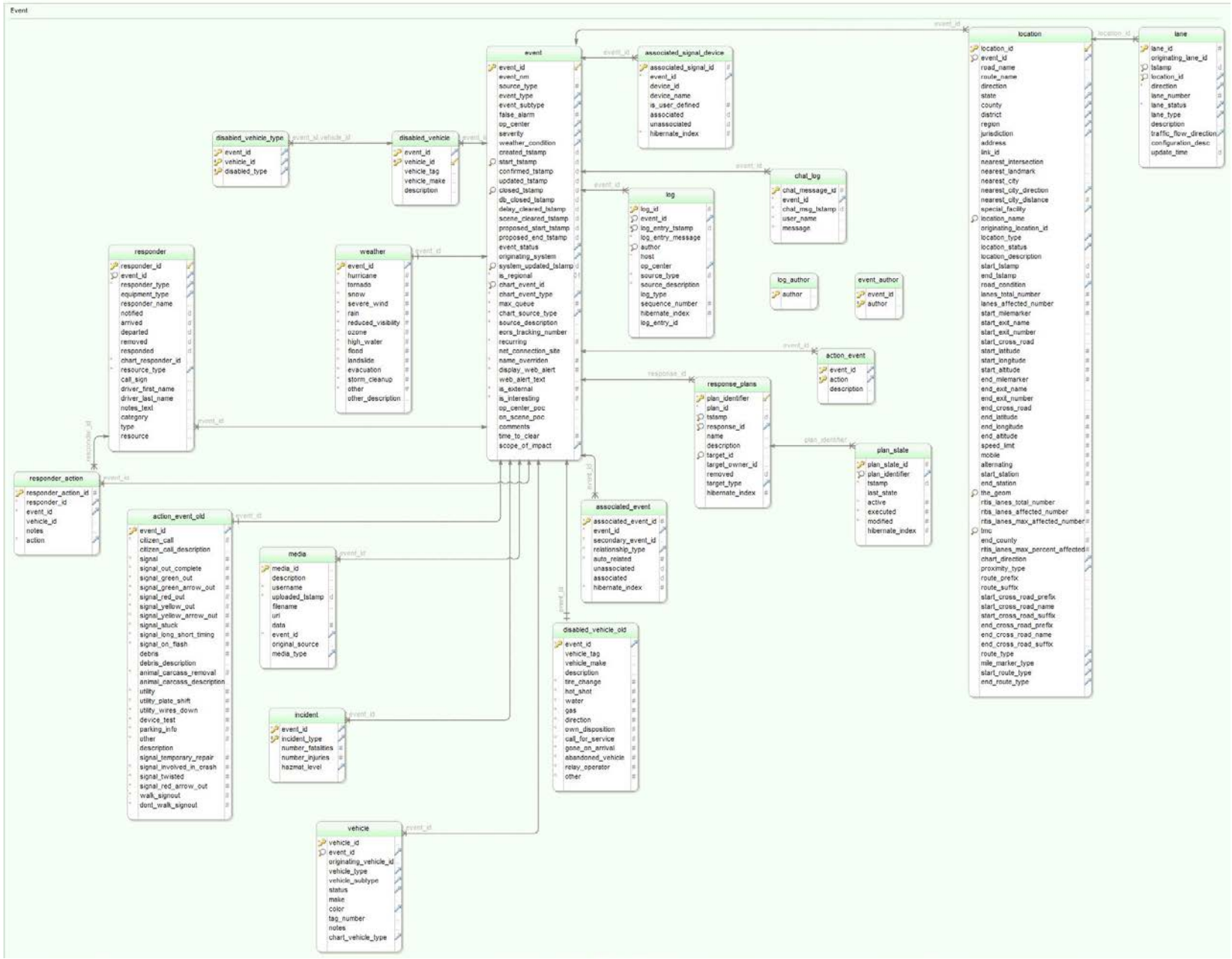
Traffic events: 40,000 records per day (0.001 GB/day)
Traffic detectors: 35,000,000 records per day (5 GB/day)
Probe vehicles: 4,200,000,000 records per day (550 GB/day)
V2X: X,XXX,XXX,XXX,XXX records per day (??? ?B/day)

Connected vehicle will generate
25 GB of data per hour.

- Hitachi, Ltd.



Traditional storage methods and challenges



Relational Database

- Normalization
- Indexing
- Distributed access
- On-demand calculations
- Rolling calculations

Distributed computing approach

Core Modules:

- Hadoop-common
- HDFS
- MapReduce
- YARN

Other Modules:

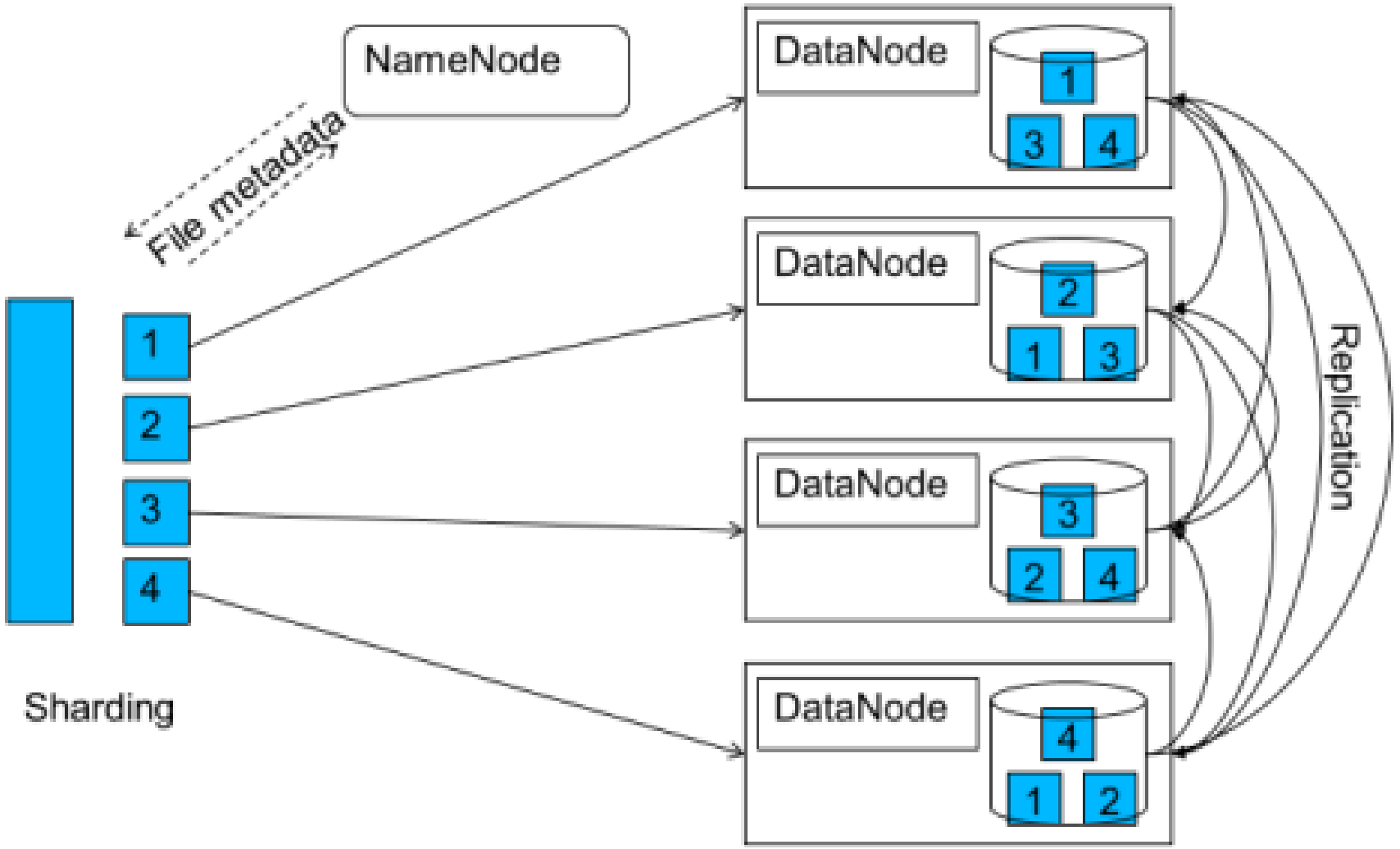
- Hbase
- ZooKeeper
- Impala
- Storm

Source: Intel Free Press

Hadoop Distributed File System (HDFS)

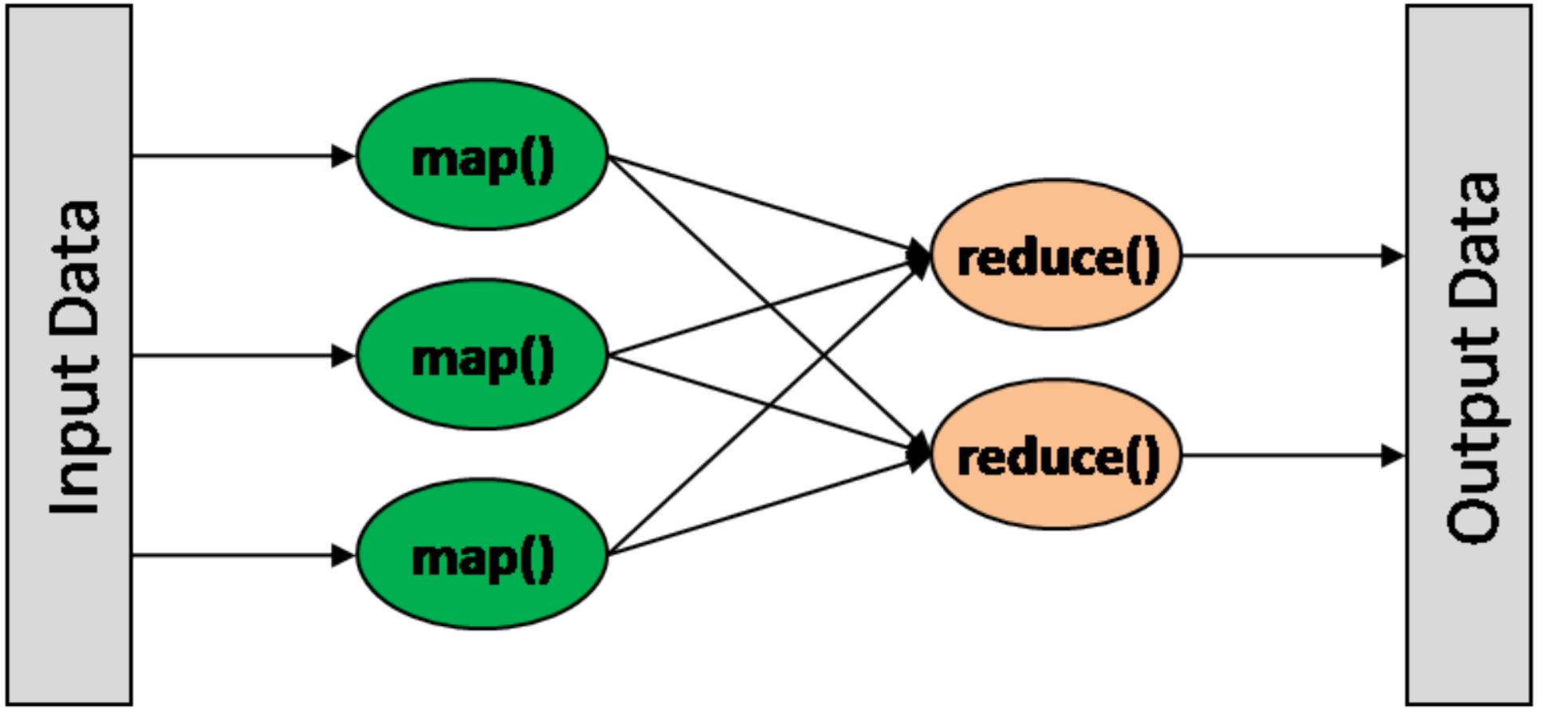
HDFS

- Distributed
- Redundant
- Block based



Source: https://cvw.cac.cornell.edu/mapreduce/dfs

MapReduce



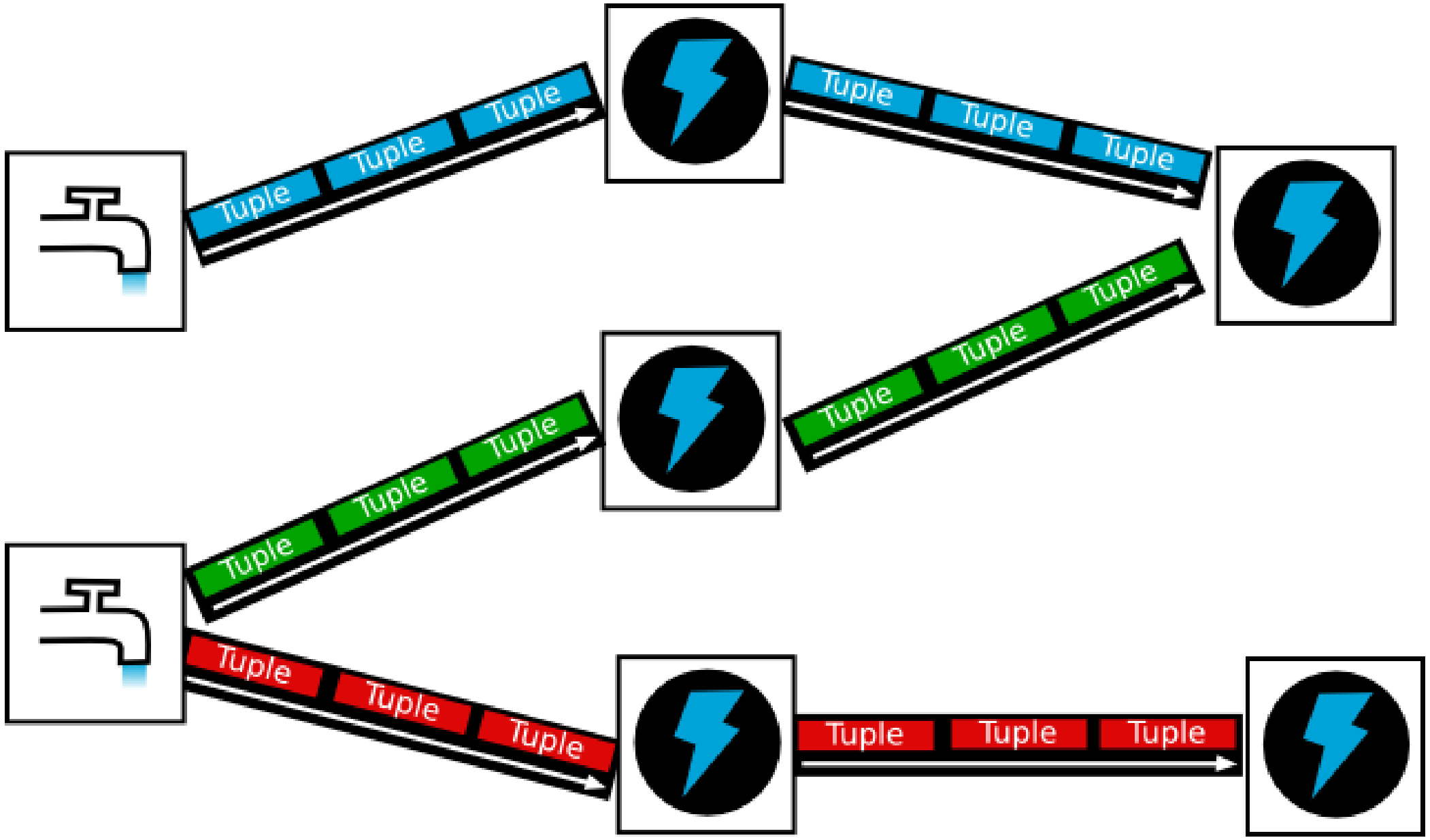
Split
[k1, v1]

Sort
by k1

Merge
[k1, [v1, v2, v3 ...]]

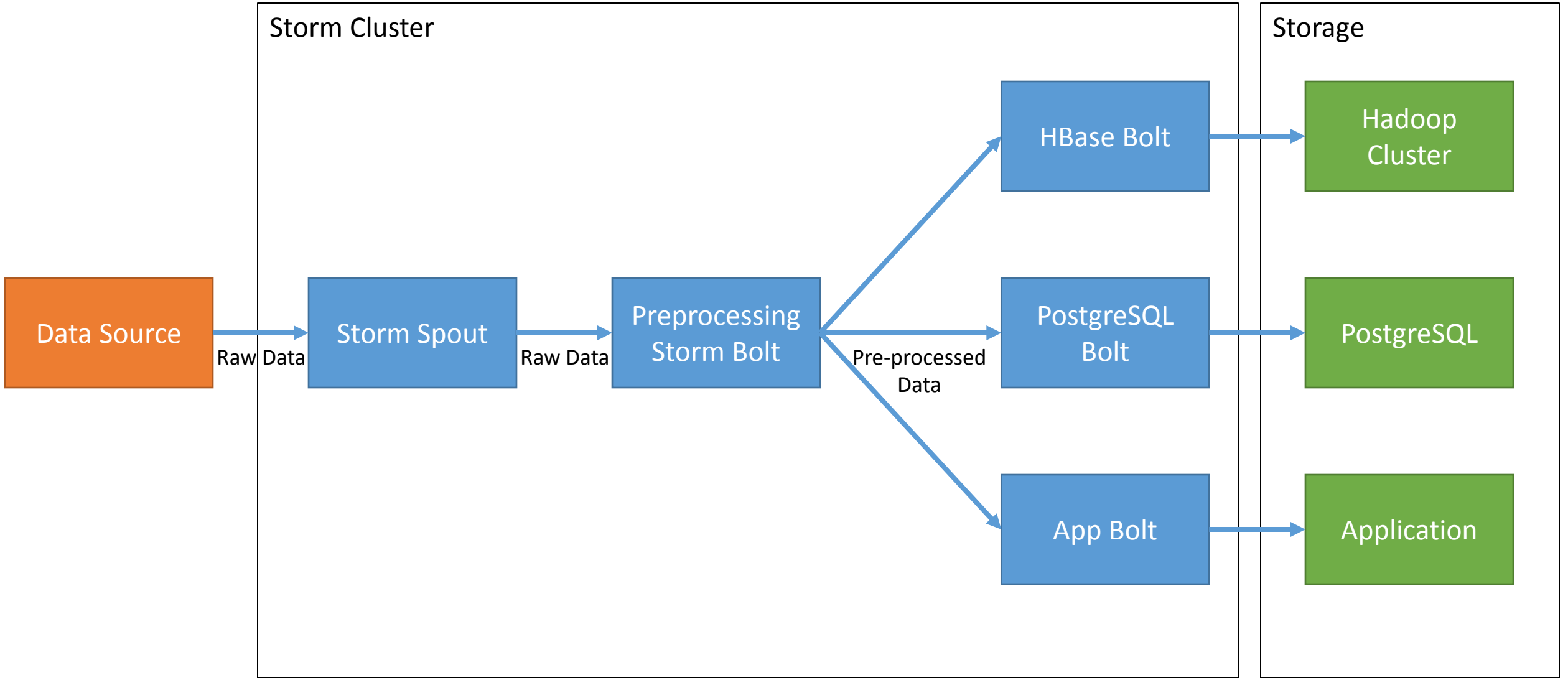
Source: https://dzone.com/articles/how-hadoop-mapreduce-works

Apache Storm



Source: https://cwiki.apache.org/confluence/display/PIG/Pig+on+Storm+Proposal

Data ingress



Data storage structure

Typical speed data record

Element	Type	Size (bytes)
TMC Code	String – 9 characters	9
Timestamp	Unsigned integer	4
Speed	Unsigned byte	1
Average Speed	Unsigned byte	1
Quality	Unsigned byte	1
Value 1	Unsigned byte	1
Value 2	Unsigned byte	1
Total per record		18

Rolling calculations challenge

Good:

- Runs on Hadoop
- Allows multiple MapReduce cycles
- Reduces number of (long) startups
- Keep intermediate results in memory

Bad:

- Difficult to configure
- Performance starts high, degrades badly
- Frequent hangs and crashes
- Startup time still significant

Source: Spark project team

Example overnight run

Configuration

- 1 million records
- 4 Hadoop nodes
- 128 GB of memory per node

Results

- 1,280 iterations
- Average run: 45 seconds
- 20+ minute hangs at the end

Rolling calculations challenge

- Functional language built on top of JVM
- Lisp dialect
- Specifically designed for concurrency

Example overnight run (again)

Configuration

- 1 million records
- 1 developer workstation
- 16 GB of memory

Results

- ~50,000 iterations (vs 1,280)
- Average run: < 1 second (vs 45 seconds)

Source: clojure.org

Next steps

- Continue to tune the system for performance and storage
- Scale
- Consider moving to cloud

Thank You!

Nikola Ivanov

ivanovn@umd.edu

301-405-3626