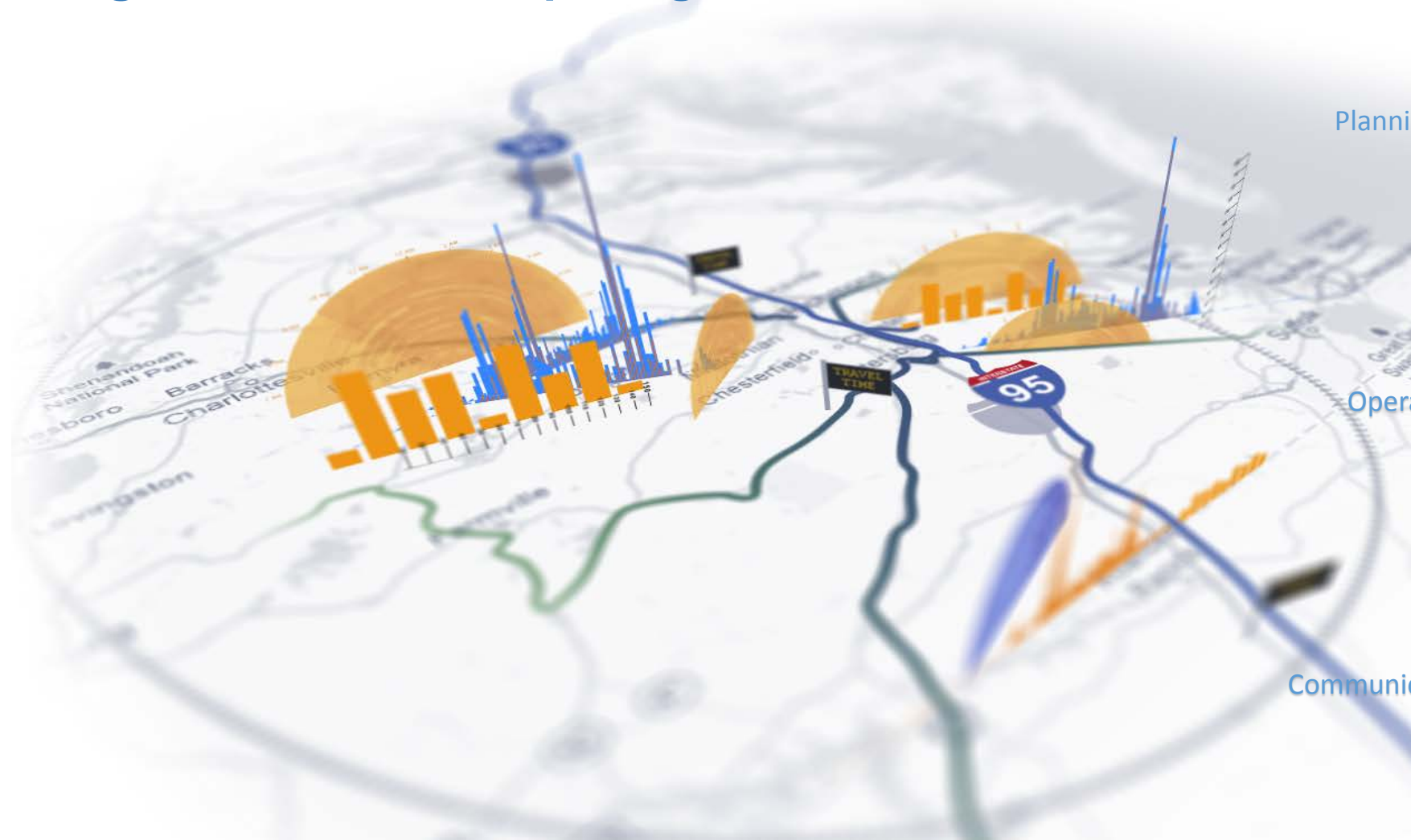


Storing and Analyzing Transportation Big Data using Distributed Computing



Performance Measures

	4 PM	5 PM	
\$31.5K	\$23.8K	\$120.3K	\$132.7K
\$8K	\$16.4K	\$42.1K	\$110.4K
\$10K	\$10.5K	\$131.8K	\$246.8K
\$4.7K	\$27K	\$131.2K	\$214.7K
\$2.2K	\$34K	\$156.8K	\$269.3K
\$3K	\$111.9K	\$180K	\$271.8K
\$52.9K	\$18.8K	\$13.7K	\$28.9K
\$2K	\$684.4K	\$1,246.9K	\$1

Planning



Operations



Communications



Transforming Big Data into Actionable Information

What is big data?

WHAT IS BIG DATA?

VOLUME
Large amounts of data.

VELOCITY
Needs to be analyzed quickly.

VARIETY
Different types of structured and unstructured data.

Key questions enterprises are asking about Big Data:

- How to store and protect big data?
- How to backup and restore big data?
- How to organize and catalog the data that you have backed up?
- How to keep costs low while ensuring that all the critical data is available when you need it?

WHAT ARE THE VOLUMES OF DATA THAT WE ARE SEEING TODAY?

- f**
30 billion pieces of content were added to Facebook this past month by 800 million plus users.
- zynga**
Zynga processes 1 petabyte of content for players every day; a volume of data that is unmatched in the social game industry.
- YouTube**
More than 2 billion videos were watched on YouTube... yesterday.
- LOL!**
The average teenager sends 4,762 text messages per month.
- Twitter**
32 billion searches were performed last month... on Twitter.

WHAT DOES THE FUTURE LOOK LIKE?

Worldwide IP traffic will **quadruple by 2015.**

By 2015, nearly **3 billion people** will be online, pushing the data created and shared to nearly **8 zettabytes.**

HOW IS THE MARKET FOR BIG DATA SOLUTIONS EVOLVING?

A new IDC study says the market for big technology and services will grow from \$3.2 billion in 2010 to \$16.9 billion in 2015. That's a growth of 40% CAGR.

58% of respondents expect their companies to increase spending on server backup solutions and other big data-related initiatives within the next three years.

2/3rds of surveyed businesses in North America said big data will become a concern for them within the next five years.

Everyday business and consumer life creates 2.5 quintillion bytes of data per day.

90% of the data in the world today has been created in the last two years alone.

Asigra.

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

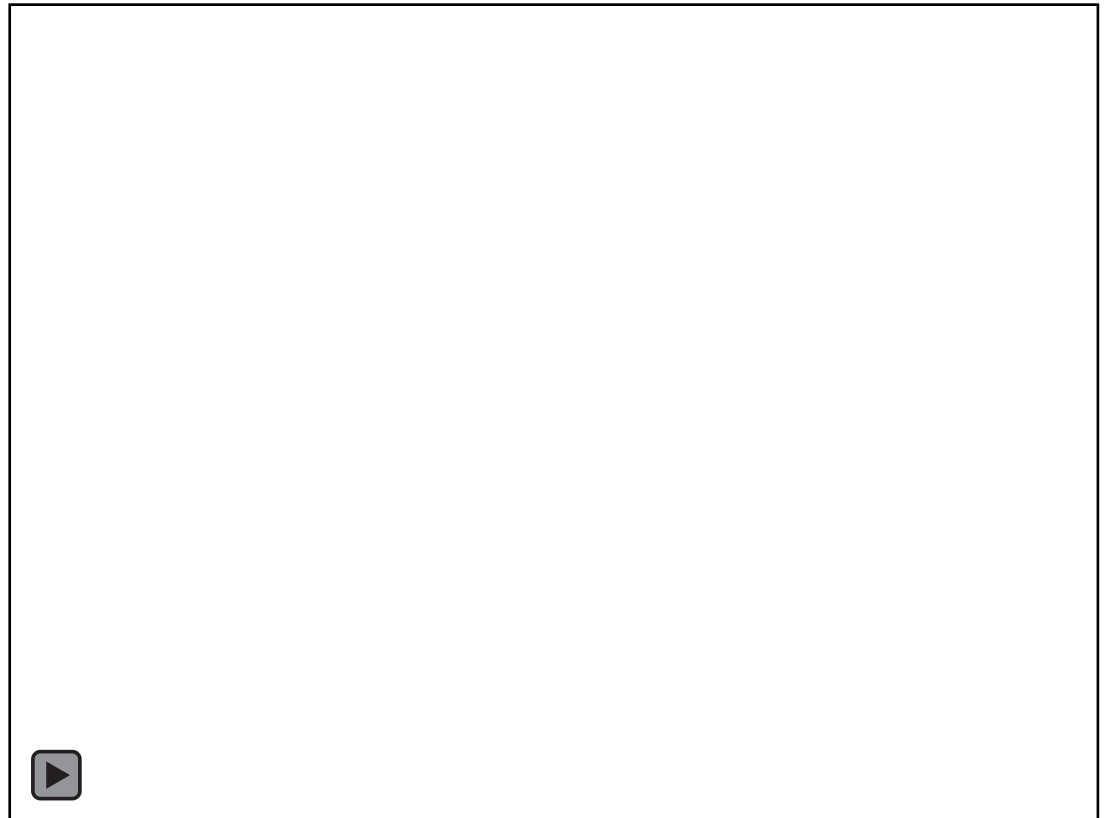
In 2019 global internet traffic will be ~52,000 GB per second!

What does transportation big data look like?

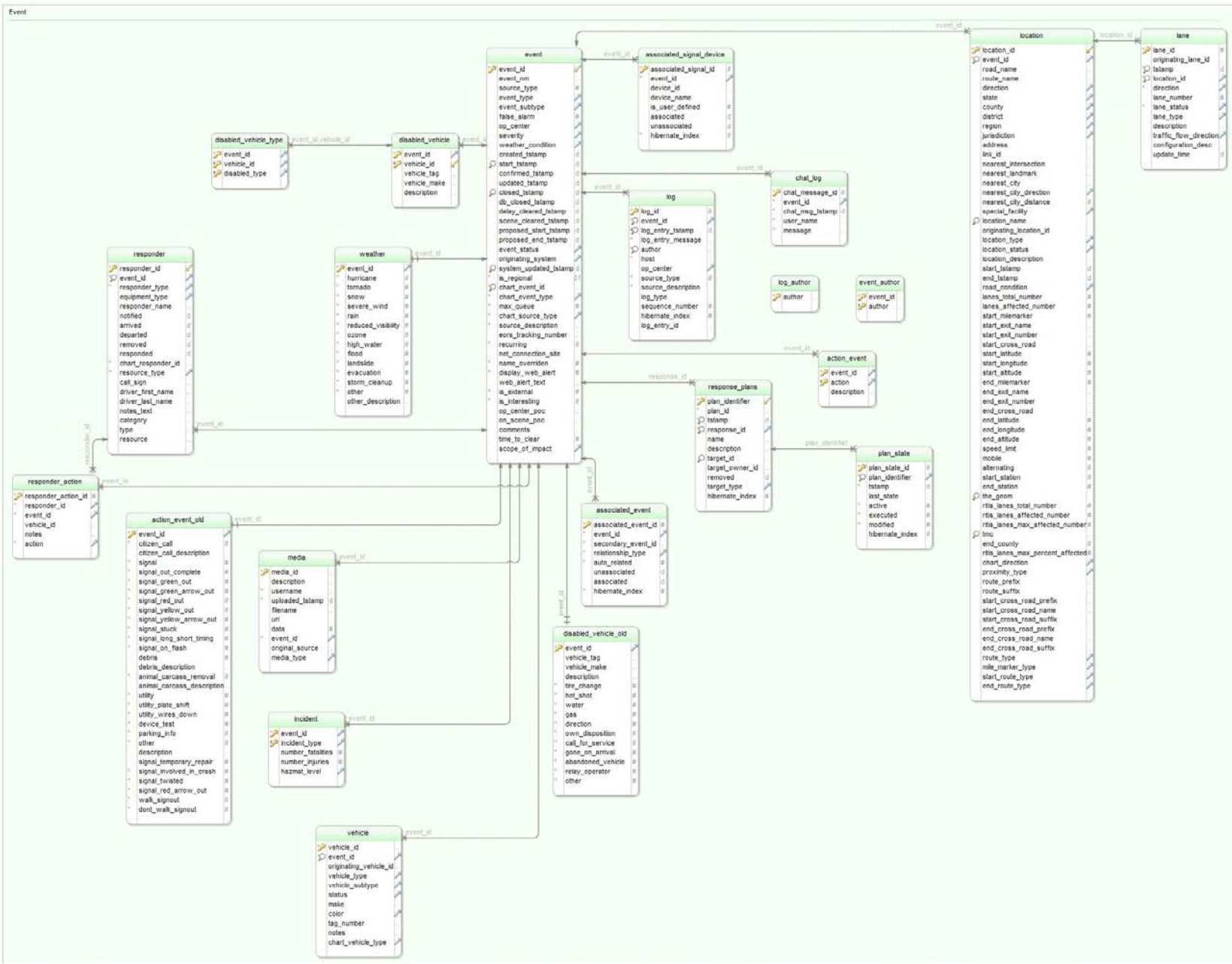
Traffic events:	40,000 records per day (0.001 GB/day)
Traffic detectors:	35,000,000 records per day (5 GB/day)
Probe vehicles:	4,200,000,000 records per day (550 GB/day)
V2X:	X,XXX,XXX,XXX,XXX records per day (??? ?B/day)

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:

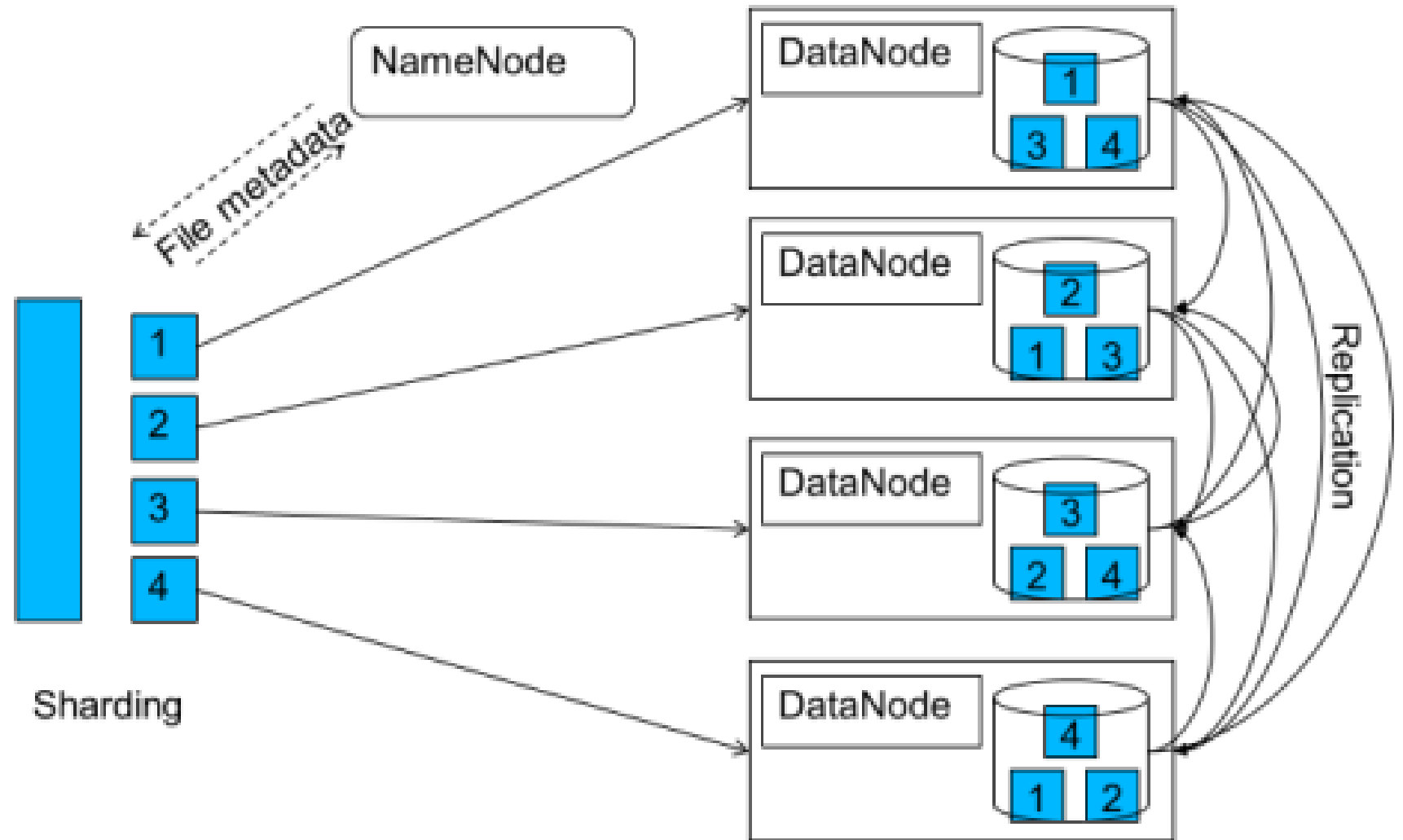
- Hbase
- ZooKeeper
- Impala
- Storm



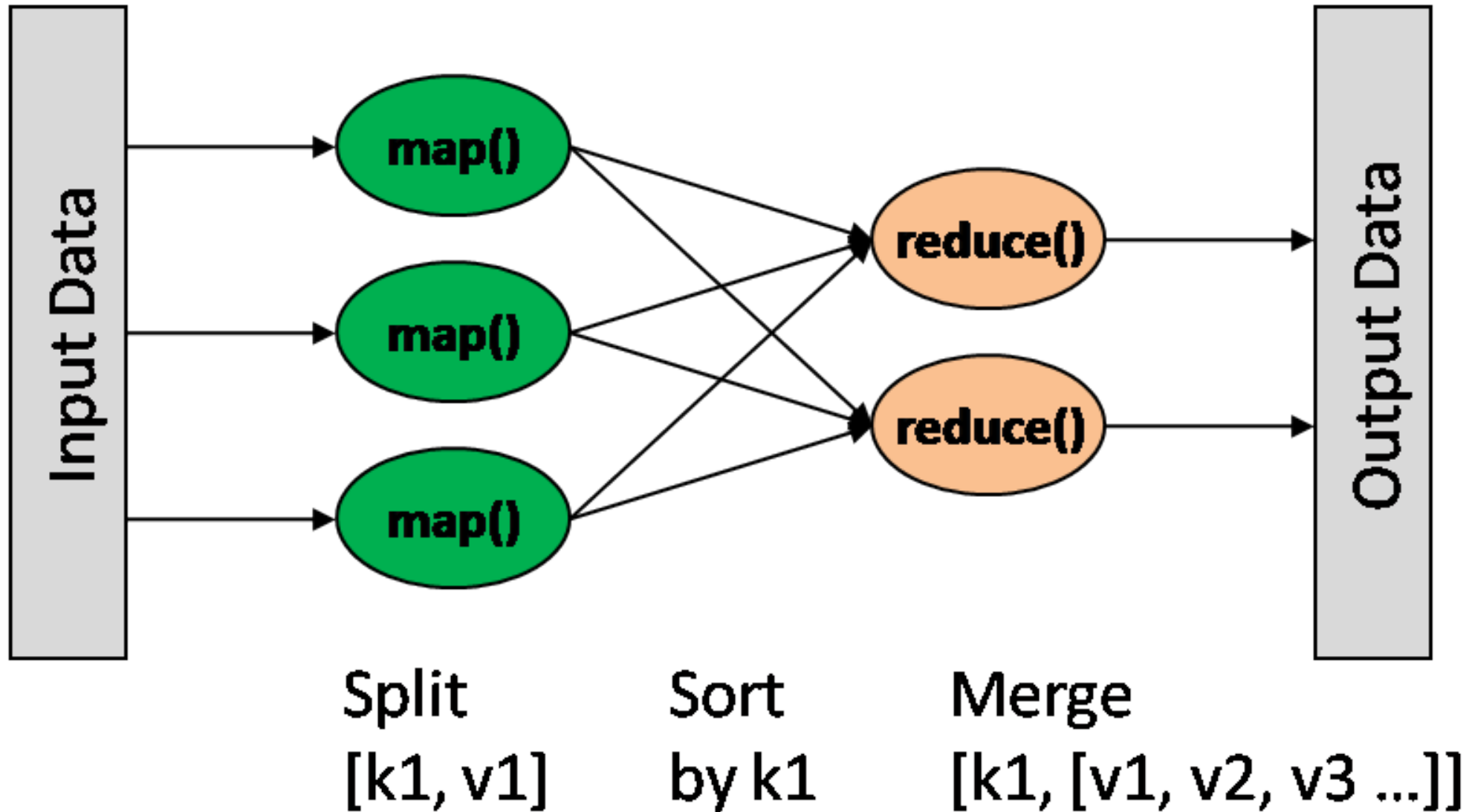
Hadoop Distributed File System (HDFS)

HDFS

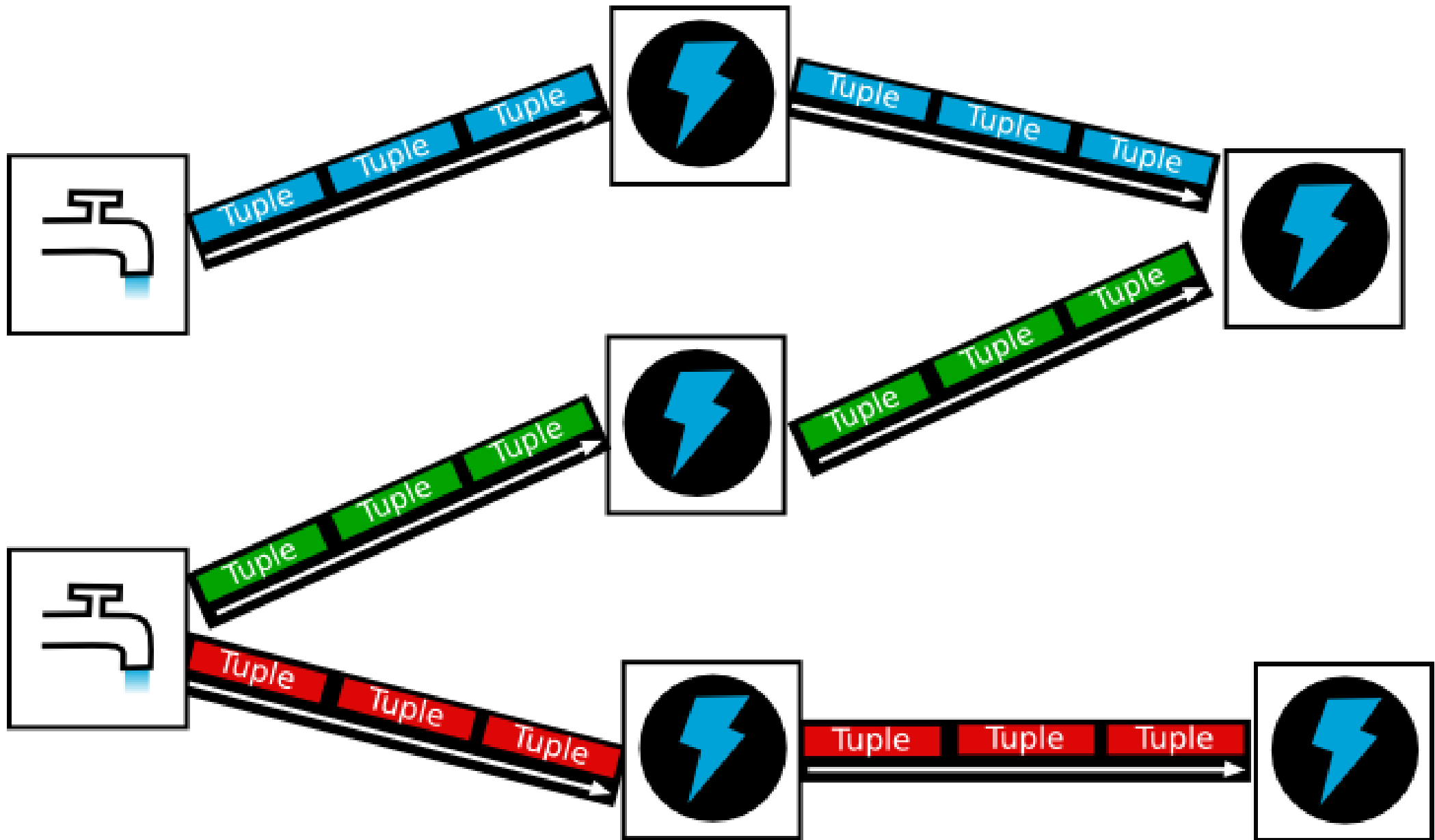
- Distributed
- Redundant
- Block based



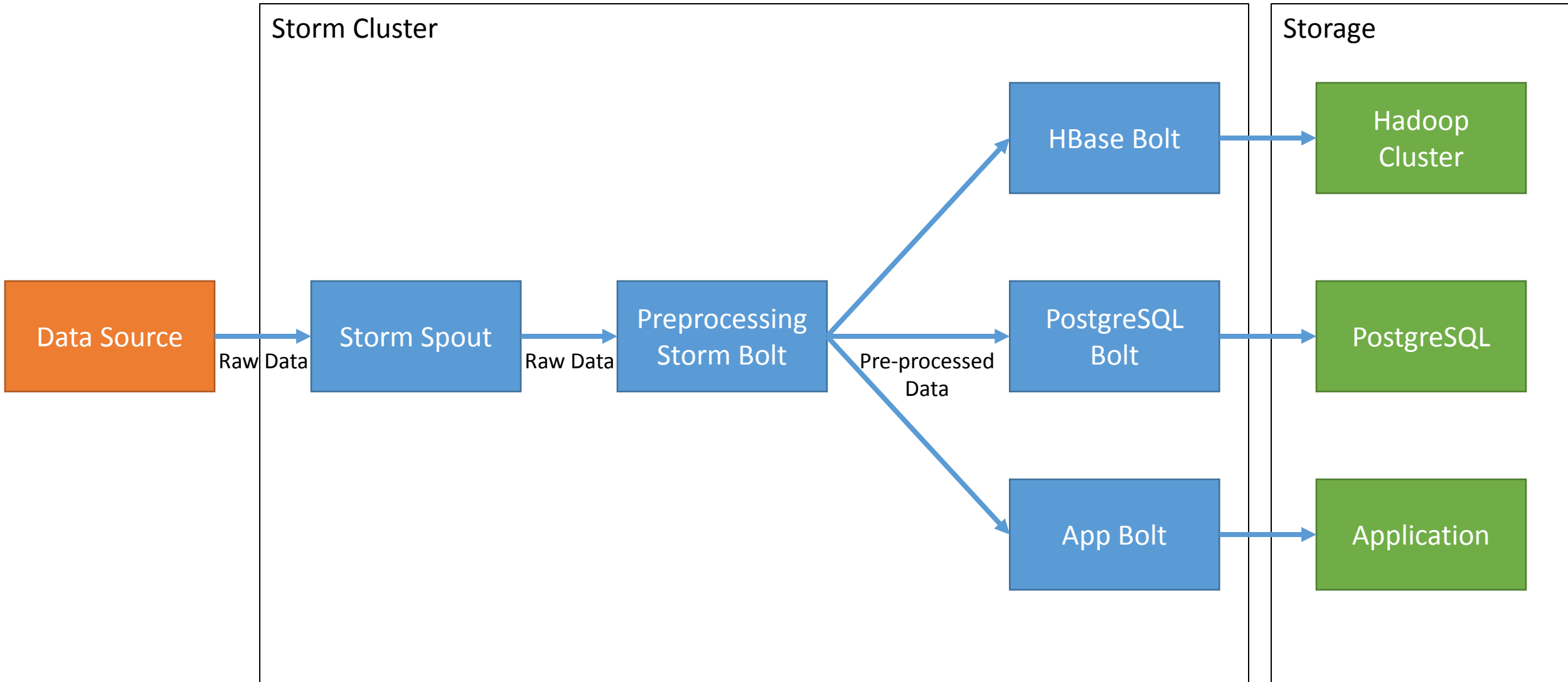
MapReduce



Apache Storm



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



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

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



- 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)

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

CATT
LABORATORY