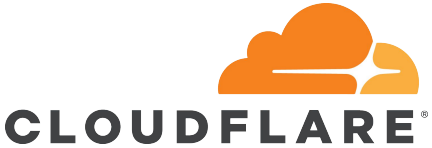


RIPE75 - Network monitoring at scale

Louis Poinsignon

Why monitoring and what to monitor?

Why do we monitor?

- Billing
  - Reducing costs
- Traffic engineering
  - Where should we peer?
  - Where should we set-up a new PoP?
  - Optimizing our network
- Anomaly detection
  - Troubleshooting
  - Proactive monitoring and predictions

Sources of information

- SNMP
- Flow data
- BGP/routing table

Sources of information

- SNMP
- Flow data
- BGP/routing table

Flow sampling protocols

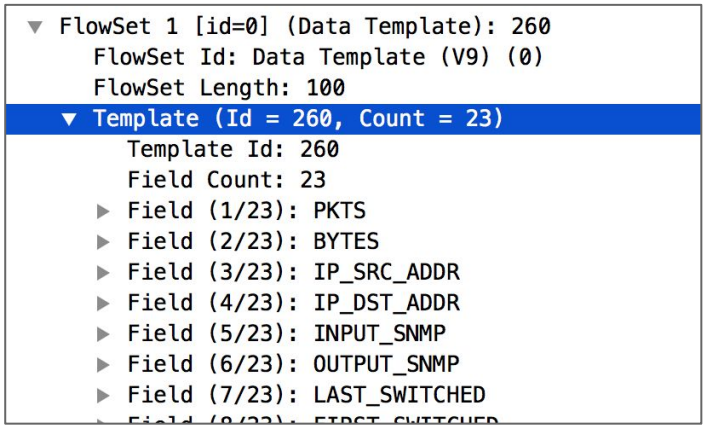
NetFlow: protocol from Cisco. IPFIX: the open standard.

Template based.

Takes 11 minutes to gather all the templates.

Between sampling and collection:

delay of **23 seconds** for **NetFlow v9 (Cisco)**
and **65 seconds** for **IPFIX (Juniper)**.



CLOUDFLARE

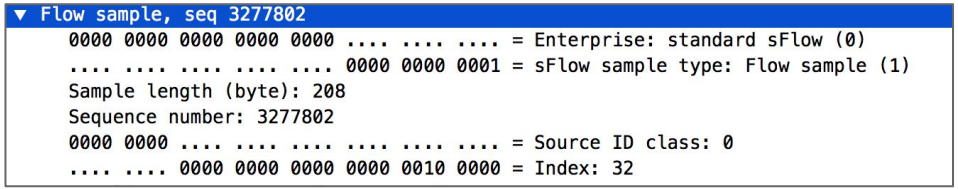
Flow sampling protocols

sFlow:

Each structure is specified (HTTP, network, Wi-Fi...)

Counters and packet sampling (headers)

Instantaneous



CLOUDFLARE

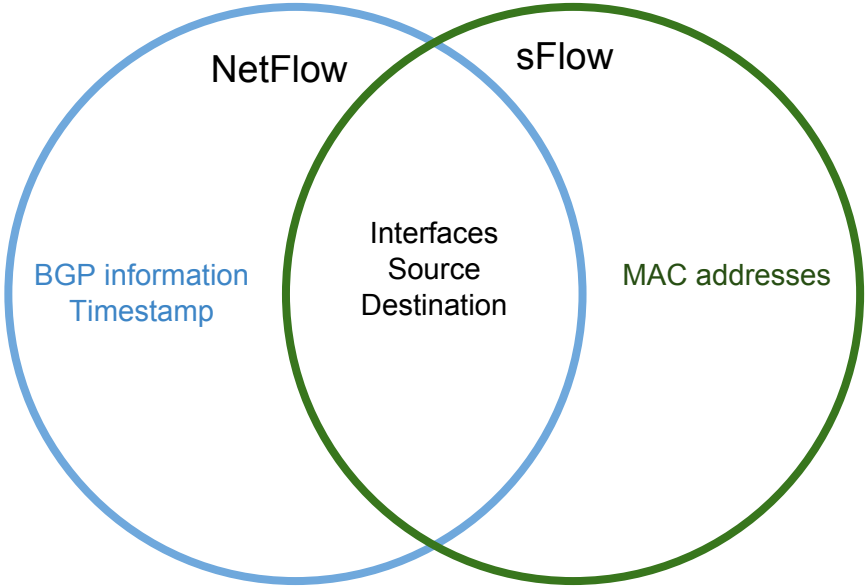
What we want

Sampling information:

- Rate
- Source router
- Timestamp

Network information

- AS number / next-hop
- Mac-addresses
- Interfaces
- Source/destination
  - IP
  - Ports
  - Protocol
  - TCP flags



CLOUDFLARE

Cloudflare today

100+ edge routers

- various vendors
- all around the globe

Different environments

Terabits of traffic

It’s already too late if a user notifies us about an issue.

CLOUDFLARE

What we used before

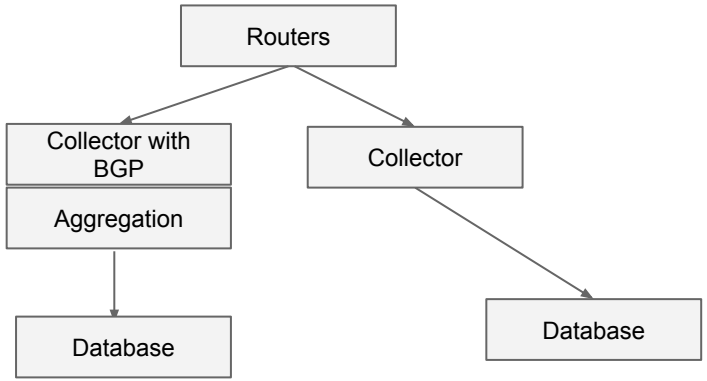
nfdump : collection / local storage

nfacctd : aggregation

Two separates path.

nfacctd was able to correct BGP information.

No sFlow.



CLOUDFLARE

Why we stopped using them

They are great tools but they became unfit to our situation.

Limitations:

- Vendor bug: corrupting ASN information
- Too many packets a single collector could not process them

Adding sFlow visualization:

- Limited ASN information
- Two aggregations in parallel

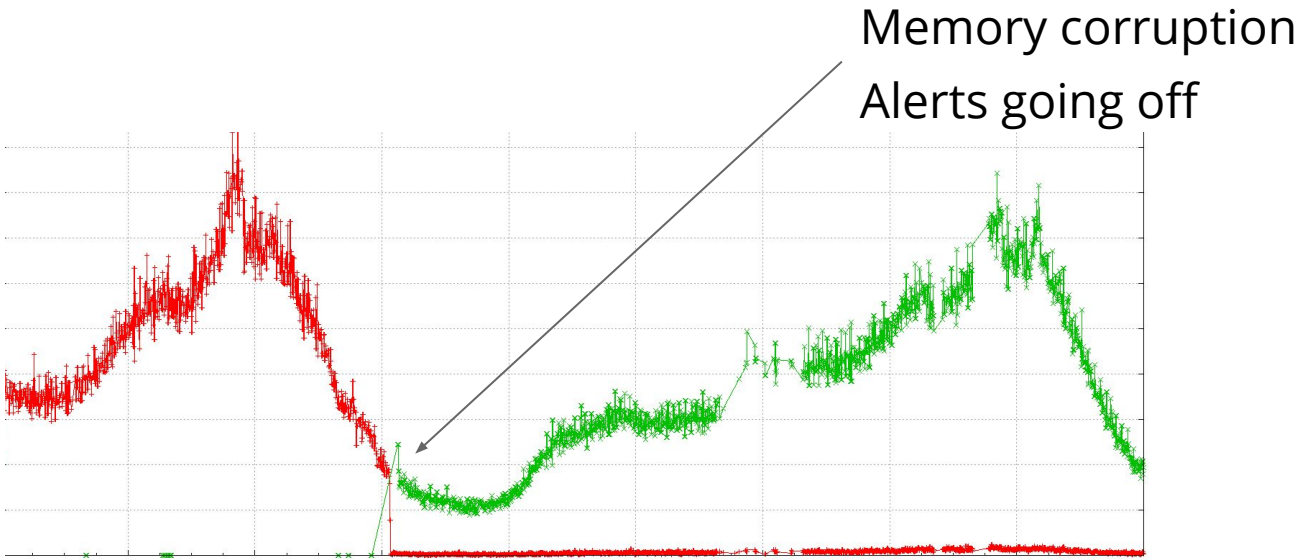
Need to monitor the collection

Anybody should be able to build tools from this data

Create aggregations *for* Cloudflare (type of plan, region, etc.)

Vendor bug

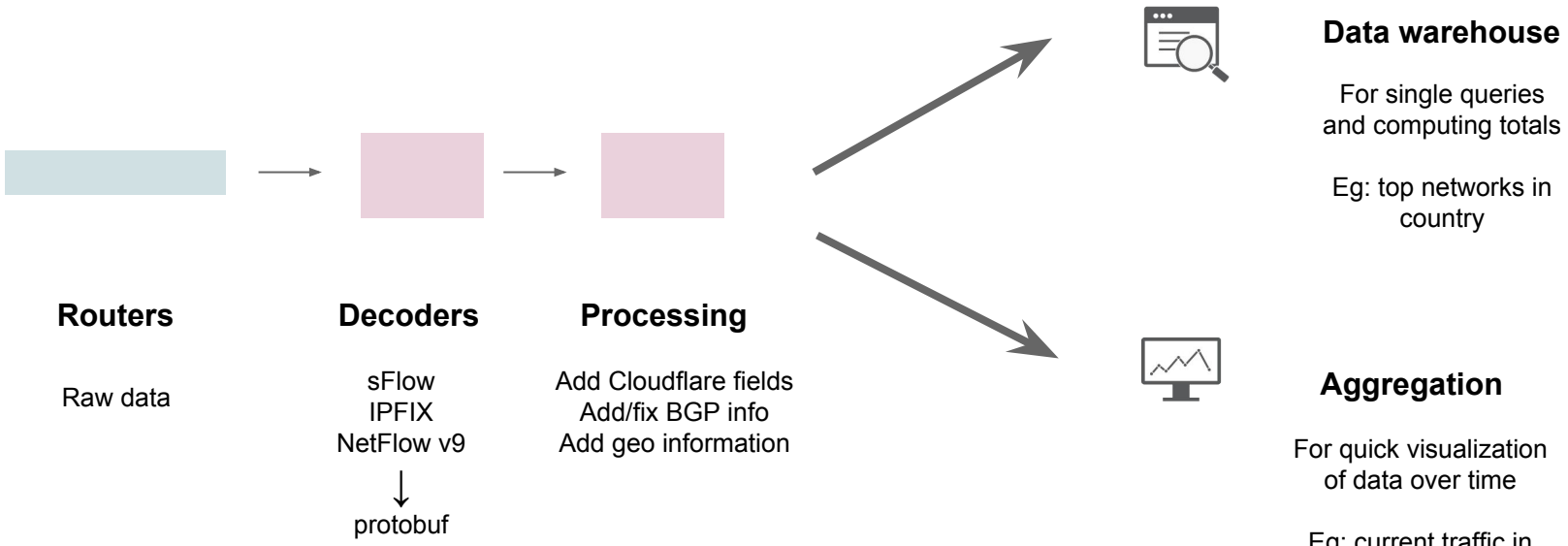
Losing a major ISP in Europe. Replaced by a small ISP from Brazil.



CLOUDFLARE

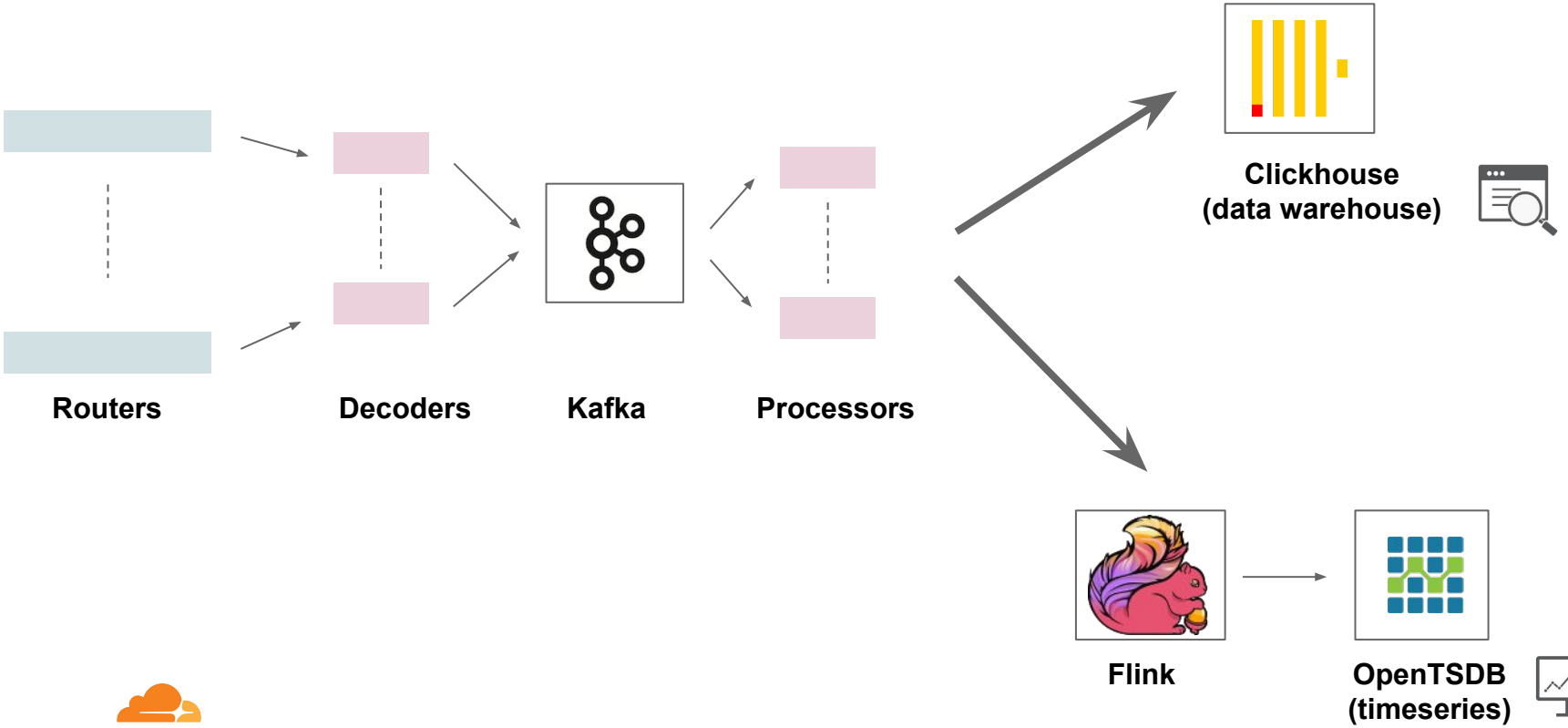
What we built

What we built



CLOUDFLARE

What we built



CLOUDFLARE

What we built

Own NetFlow+IPFIX+sFlow collector **GoFlow**:

- In **Go**
- Easily extensible for new protocols
- Outputs to protobuf format
- Can be parallelized
- Benchmarked to **30 000** messages a second
- Running in production at Cloudflare
- Living in containers

Parallel processing **units** using BGP data, geolocation databases, Cloudflare APIs to:

- Correct/add fields
- Add Cloudflare specific information

Inserters to populate databases.

Message broker to connect the pieces: **Kafka**.

Aggregation done by **Flink**.

Stored in **OpenTSDB** and **Clickhouse**.

CLOUDFLARE

Aggregations

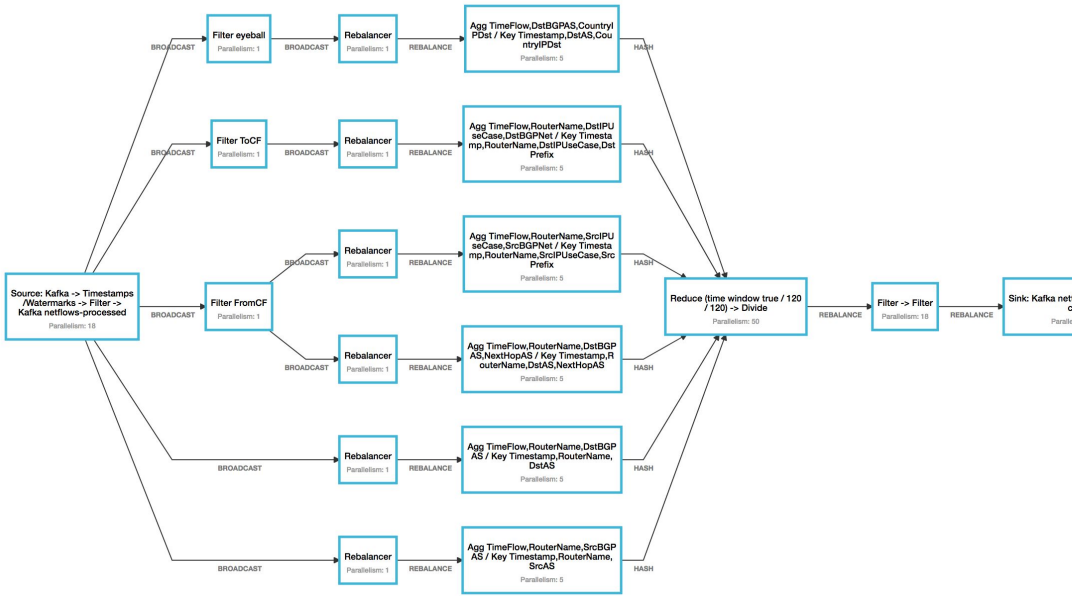
Flink is a Java **framework** for building stream-processing apps (jobs).

Jobs are split into **tasks** and sent to a cluster.

Easy to scale, balance and reorder tasks.

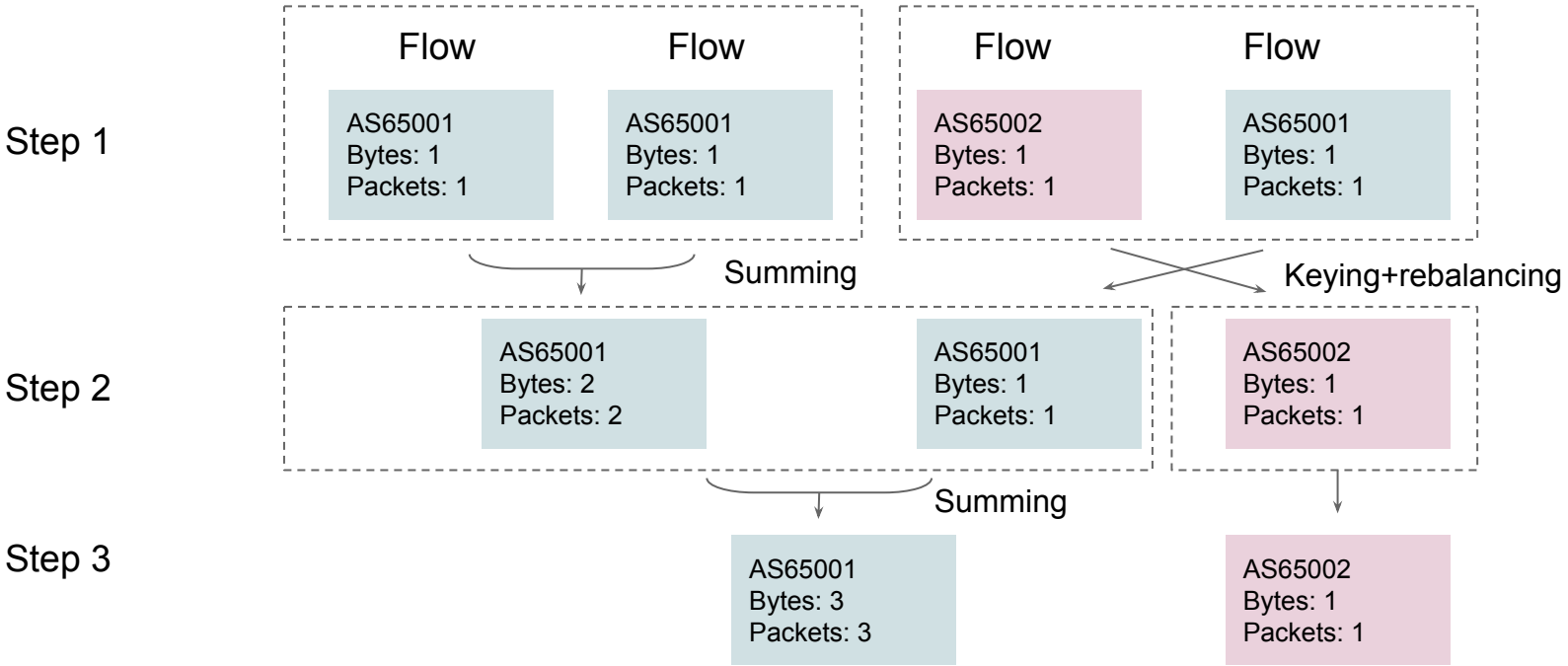
Schematic view of the app.

Accurate time-aggregation.



CLOUDFLARE

Flink - MapReduce



CLOUDFLARE

Flink - Sample program

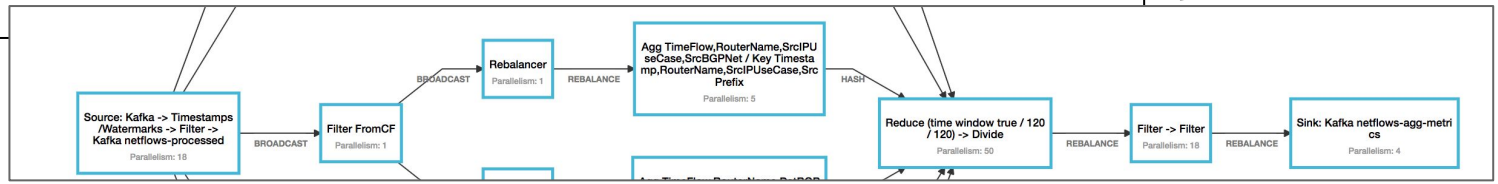
```
DataStream<FlowMessage> inData =
new FlinkKafkaConsumer09<FlowMessage>(
                "netflows-processed",
                new FlowMessageDeserializer(),
                propertiesConsumer);

DataStream<FlowMessage> inDataEyeball =
                inData.filter(new FlowFilter.EyeballFilter()).
                setParallelism(1).broadcast();

DataStream<FlowAggMessage> inDataAgg =
                inDataEyeball.map(new FlowUtils.Mapper("DstAS,colo"));

inDataAgg.reduce(new FlowTransformations.FlowAggReduceKey());
```

Source (Kafka)

Filter

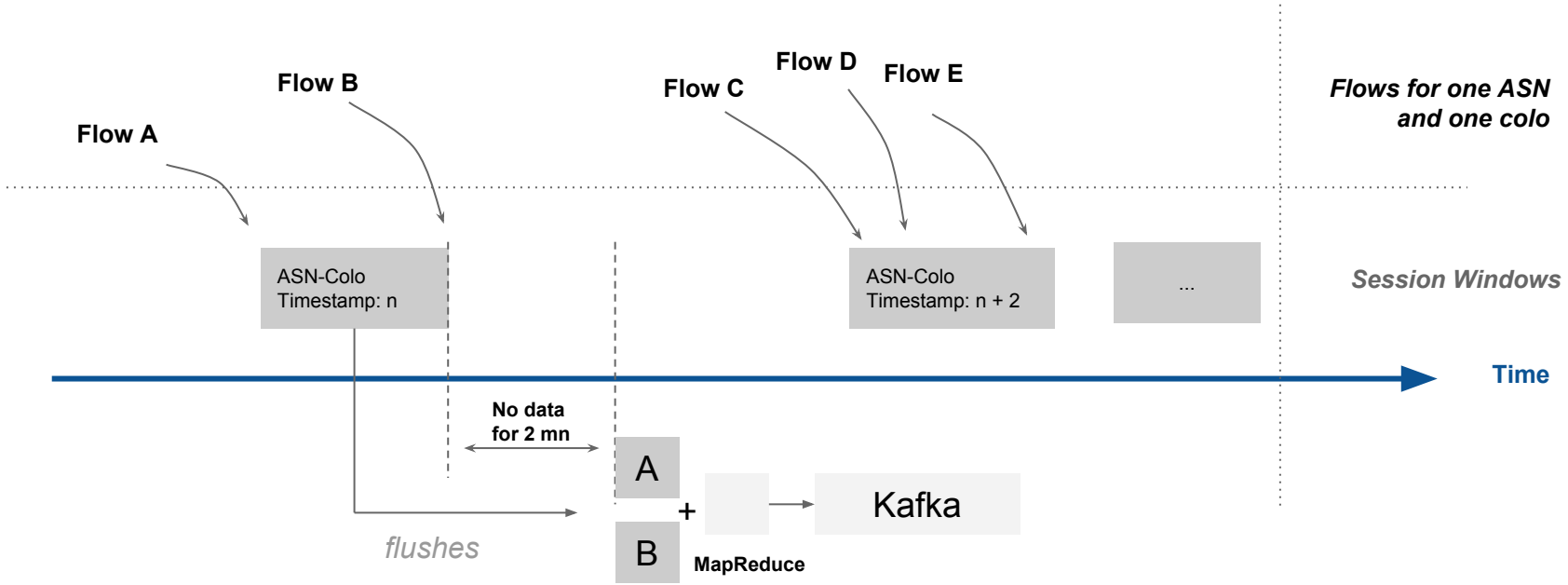
Mapping

Reduce



CLOUDFLARE

Flink - Windowing



CLOUDFLARE

Results - Flows

Business intelligence:

Simple as a SQL query

Or an API call

Top networks per country, datacenters, plan, transit providers...

IPv6 share for biggest networks

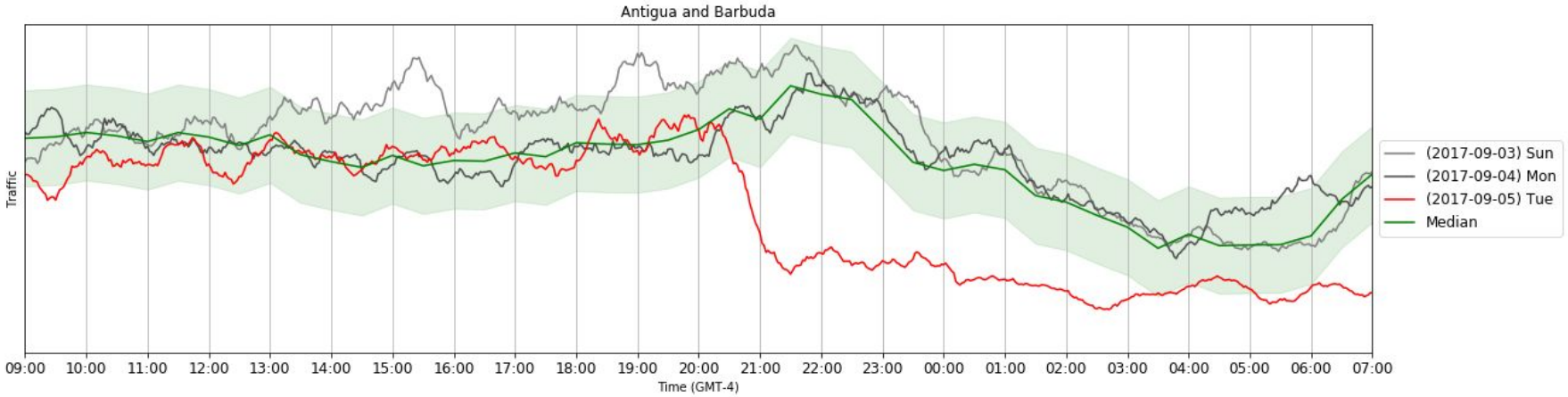
	AS	Ratio IPv6
- Comcast Cable Communications, LLC	7922	46.11%
- AT&T Services, Inc.	7018	57.21%
le Inc.	15169	0.37%
cebook, Inc.	32934	88.49%
mmunications Services, Inc. d/b/a Verizon Business	701	0%

CLOUDFLARE

Results - Aggregations

Traffic of every ASN. By data centers. By country. By interface. By type of traffic. By transit/peer...

Other teams started using the data to troubleshoot non networks problems.

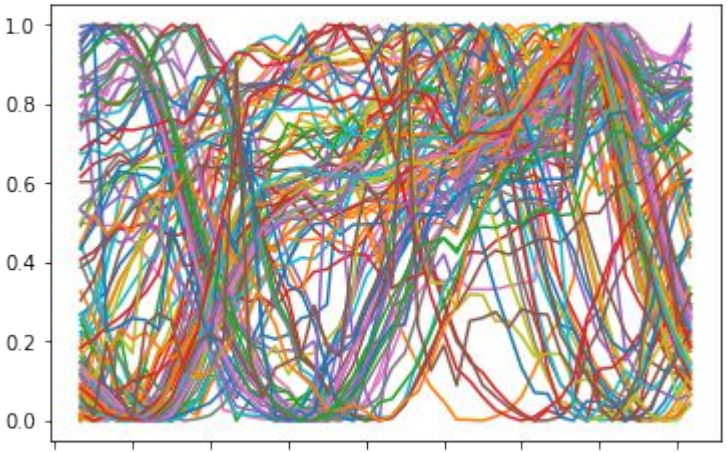


CLOUDFLARE

https://blog.cloudflare.com/the-story-of-two-outages/

Results - Example: maintenance

Automatically build a list of best hours for a maintenance.



Normalized traffic of our PoP over a day

CLOUDFLARE

Anomaly detection

Traffic variations are visible:

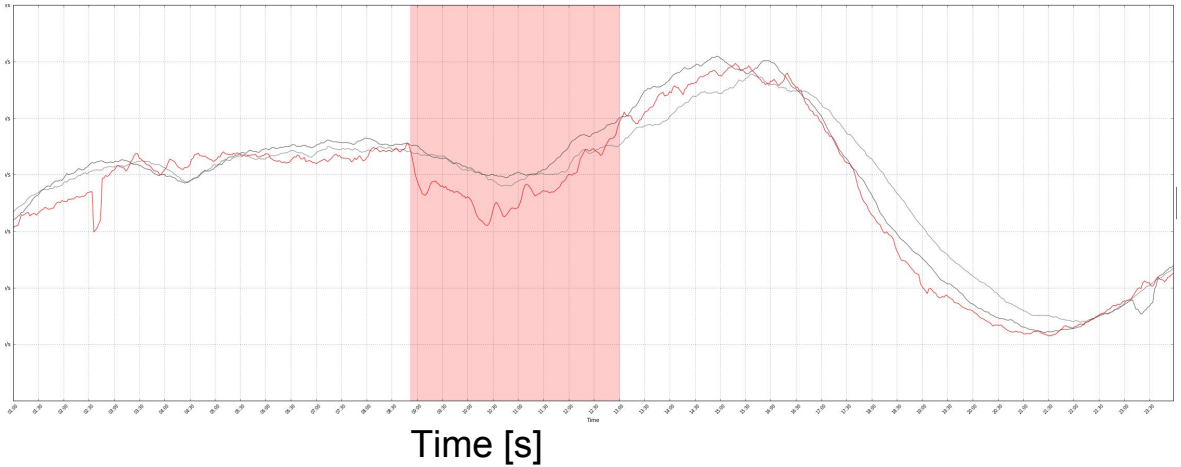
- Turkey rate-limiting (15/07/2016)
- Iraq shutting down Internet during exams
- Country wide power failure
- ...

Machine-learning to **classify**.

Automatic detection.

Mostly **Python**.

Example: Taiwan power cut



Time [s]

https://blog.cloudflare.com/power-outage-taiwan/

CLOUDFLARE

Algorithms

Derivation

Correlation

Pearson coefficient

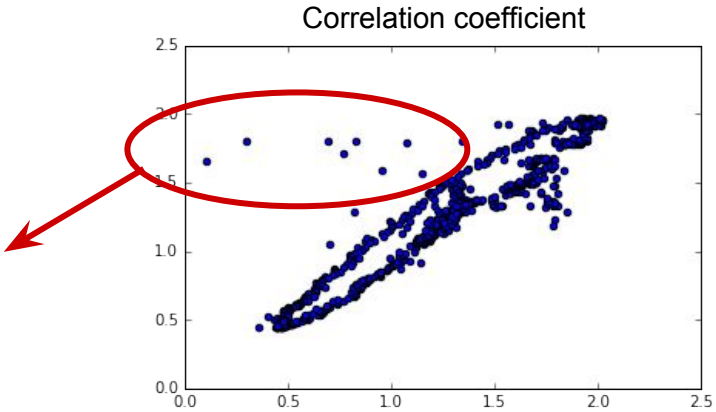
How different is it from usual

Median

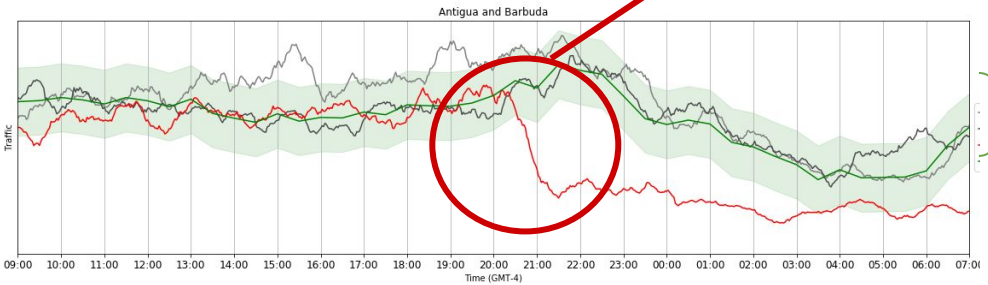
Remove small artifacts

Variance

Intensity of variation



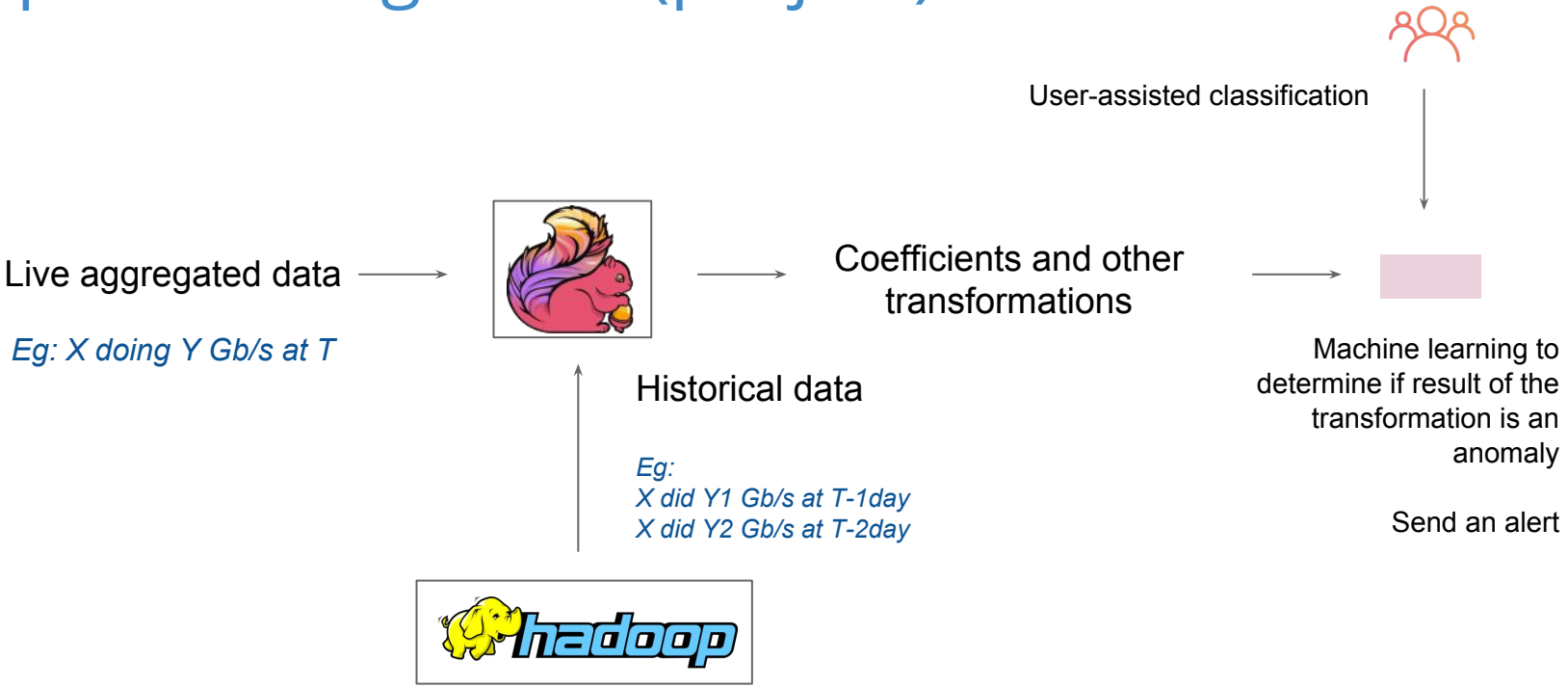
Outliers



Local variance following median

CLOUDFLARE

Pipeline integration (project)



CLOUDFLARE

Sources of information

- SNMP
- Flow data
- BGP/routing table

BGP collection

- 100’s of routers, 100’s of full tables, millions of routes
  - RIPE RIS has 15 peers (rrc00)
  - Route-views has 47 peers (route-views2.oregon-ix.net)
- View of route-leaks
- Similar pipeline and tools

CLOUDFLARE

What we built

BGP information

Updates containing routes

Collectors

Full table and forwarding updates

Processing

Add Cloudflare fields
Add/fix BGP info
Add geo information

Data warehouse

Updates history

Eg: route leak

Aggregation

Full table

Eg: routing intelligence

CLOUDFLARE

Full table?

- Stream processing versus Batch processing
- Spark (or Flink)
- Examples of what we did:
  - Find out the longest AS-Path
  - Peered prefixes
  - Mapping IP → ASN

Open-source

Flow collector

Flow collector will be open-sourced soon.

What it does:

- Decode NetFlow/IPFIX/sFlow network fields
- Encode them into a generic “network sample” format (interface, ASN, src/dst IP...)
- Provide metrics
- Filters corrupted data (garbage value)
- Provides framework for parallel processing/decoding
- 23µS for NetFlow decoding / 80µS for sFlow decoding

What it does not do:

- Decode any field (eg: Wi-Fi, GSM specific fields, etc.)
  - But, you can extend it with a new protobuf format and decoder
- Aggregation

CLOUDFLARE

https://github.com/cloudflare/

Costs

Do you want to run it on a Cloud?

Product	Amazon	Google	Azure
Collection	Compute/Docker	Compute/Docker	Compute/Docker
Stream processing	Kinesis (Firehose+Stream)	DataFlow/DataProc	Stream Analytics
Storage	Redshift	BigQuery	SQL Data Warehouse

CLOUDFLARE

Costs

UDP: about 70 bytes per flow

Message: around 100 bytes

Aggregation:
based on cardinality and time-windows

Message aggregated: 100 bytes

	Case 1	Case 2	Case 3
Traffic	10Gb/s | 1Mpps	100Gb/s | 20Mpps	1Tb/s | 100Mpps
Sampling	8192	16384	16384
Number of samples	120/sec	1200/sec	6100/sec
Aggregation window	120 s	300 s	120 s
Cardinality	120	12000	120000
Processor units	0.5	1	2
Throughput	< 1 Mb/s	1 Mb/s	5 Mb/s
Aggregation throughput	< 1 Mb/s	< 1 Mb/s	1 Mb/s
Monthly data raw	32 GB	320 GB	1.6 TB
Monthly data agg	270 MB	100 GB	270 GB

CLOUDFLARE

Costs

Case 1	Amazon	Google
Compute units	T2.micro $15	micro $10
Storage	RDS (db.T2.medium) or Dynamo $50	BigQuery $10
Aggregation	Firehose $2 Analytics $79	Dataflow $50
Total	$200/month	$100/month

Case 2	Amazon	Google
Compute units	T2.medium $80	standard $80
Storage	RDS $300	BigQuery $100
Aggregation	Firehose $20 Analytics $160	Dataflow $300
Total	$500/month	$500/month

Case 3	Amazon	Google
Compute units	T2.medium $100	standard $80
Storage	(+Redshift) $650	BigQuery $200
Aggregation	Analytics $200	Dataflow $300
Total	$1000/month	$800/month

CLOUDFLARE

BGP library

The BGP library will also be released

What it does:

- Decode BGP packets
- Can maintain session and a RIB with peers
- Encode/decode MRT
- Includes RFC and extensions

What you can do:

- Implement the behavior you want (route-reflector)
- Event-based API

CLOUDFLARE

https://github.com/cloudflare/

Thank you

louis@cloudflare.com
@lpoinsig