Image source: goo.gl/utySEK

Challenges in building learning models when traffic is encrypted

Vijay K. Gurbani, Ph.D. |
Network Data Science |
vijay.gurbani@nokia.com |
March 16, 2018 | London

mami

measurement and architecture for a middleboxed internet

NOKIA

About Nokia

Mobile Networks

Higher quality and more reliable mobile broadband experiences

Fixed Networks

More bandwidth in more places giving communities more access to the world

IP/Optical Networks

Massively scalable networks securely connecting everyone and everything to the Cloud

Applications & Analytics

Intelligent software platforms optimizing and automating network performance

Nokia Technologies

Connected health devices; professional Virtual Reality capture and broadcast; and highly valuable brand, intellectual property and technologies

Countries of operation

100+

Number of employees at the end of 2016

~101 000

R&D investment in 2016

EUR 4.9bn

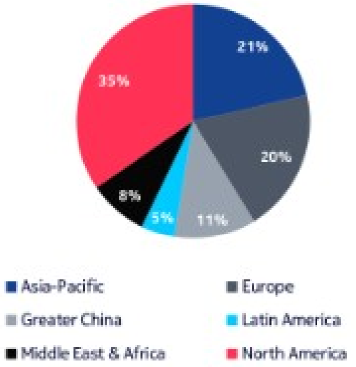
Net sales 2016

EUR 23.6bn

Nokia's Networks business

Net sales by geographic area

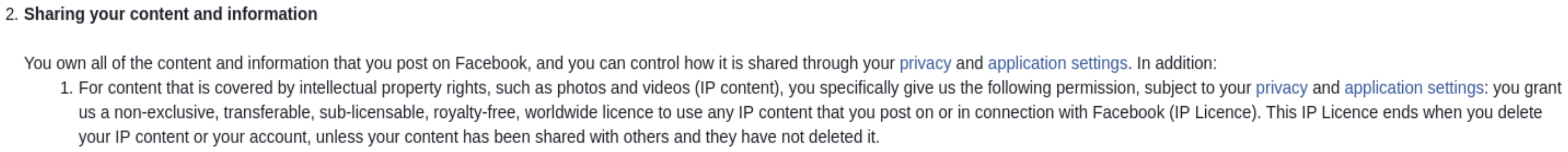
Q1 2017



We are never definitely right;
we can only be sure when we are wrong.
- Richard P. Feynman
(Lectures on *The Character of Physical Law*, 1964)

Analytic architectures: The players

- Application providers (Google, Facebook, Instagram, …)
  - Own the data, have a business relationship with the account holders and can mine information the account holder makes available.
  - Not as much interested in DPI and URI analysis (after all, the URI is destined to Google or Facebook).
  - Mostly interested in *content*-based analytics.



2. **Sharing your content and information**

You own all of the content and information that you post on Facebook, and you can control how it is shared through your privacy and application settings. In addition:

1. For content that is covered by intellectual property rights, such as photos and videos (IP content), you specifically give us the following permission, subject to your privacy and application settings: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide licence to use any IP content that you post on or in connection with Facebook (IP Licence). This IP Licence ends when you delete your IP content or your account, unless your content has been shared with others and they have not deleted it.

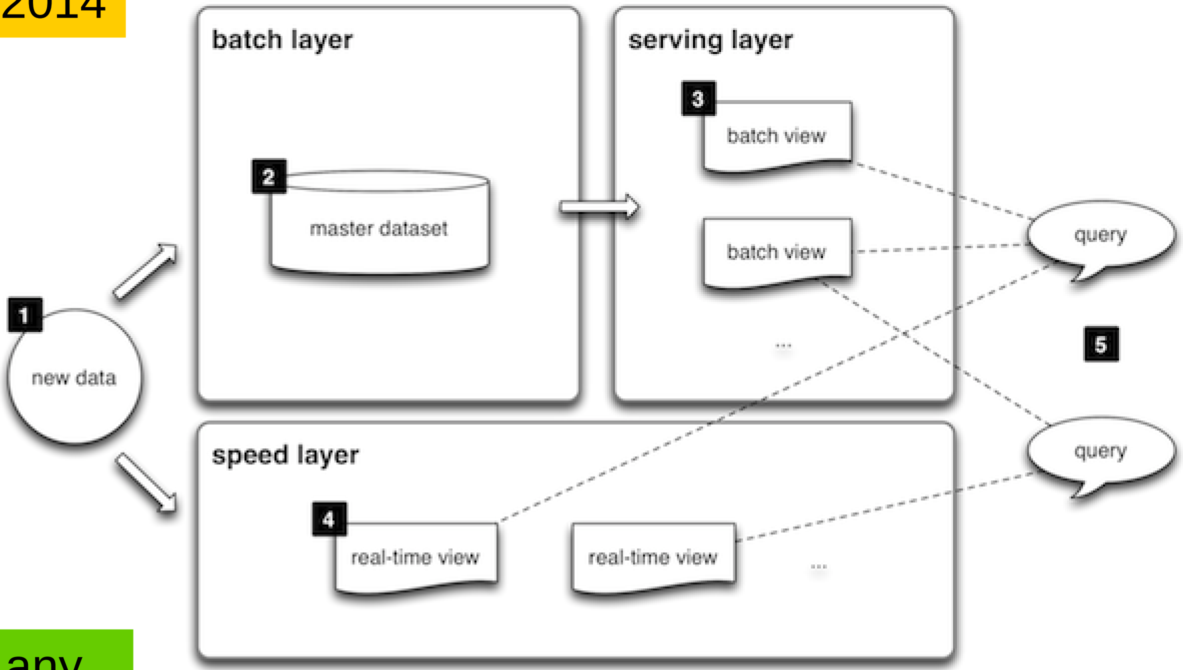
https://www.facebook.com/terms.php

Analytic architectures: The players

- Service providers
  - Moving beyond flow analysis.
  - Do not own the data, have a business relationship with the account holder but cannot mine other aspects besides the bytes traversing their networks.
  - Is interested in DPI (traffic engineering) and URI analysis, but doing so is exceedingly hard given pervasively ubiquitous encryption.
    - Some aspects are in cleartext on certain packets when a flow starts, but limited information gathered from them.
- Our focus.

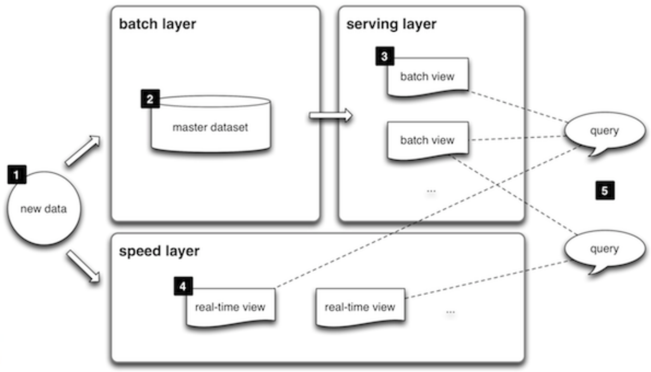
So, what is an big data analytic architecture?

The Lambda Architecture
State-of-the art, circa 2014

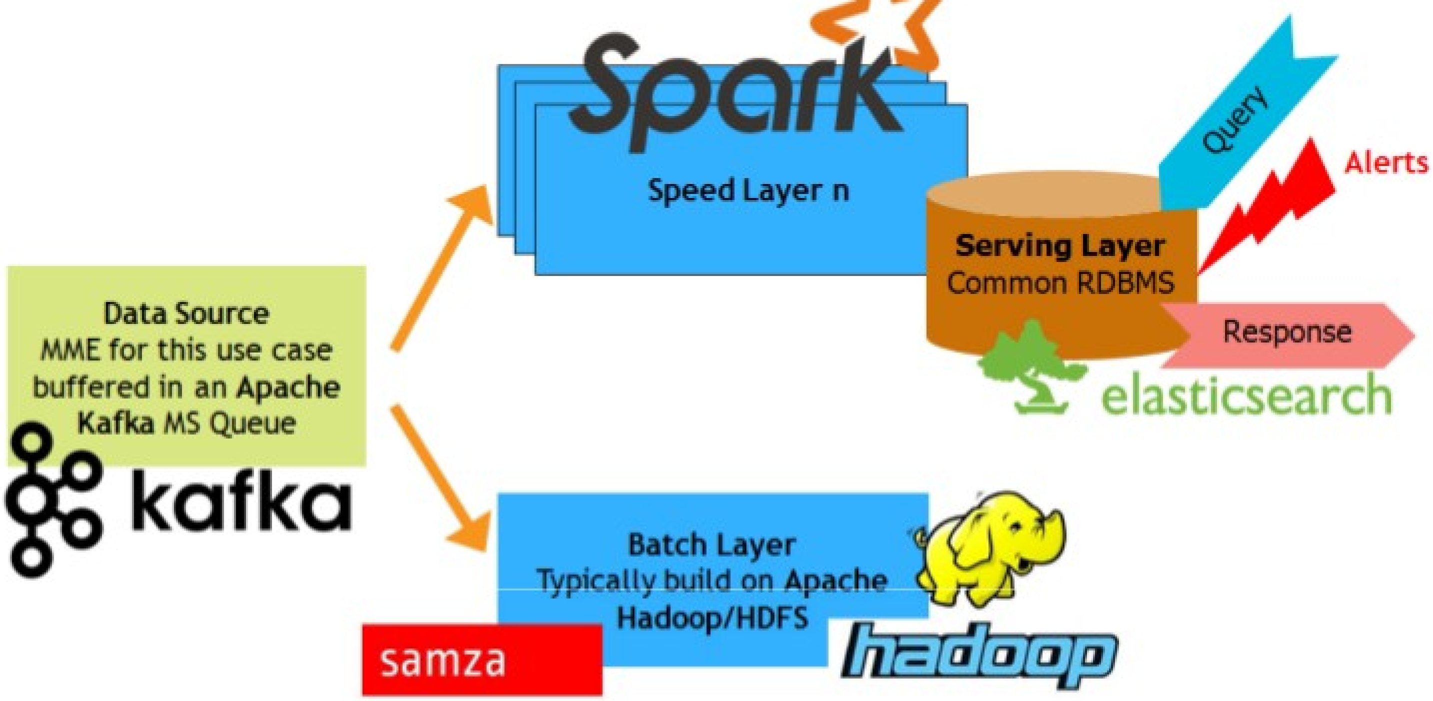


Not much attention, if any, paid to security during data movement.

So, what is an big data analytic architecture?



The logical architecture mapped to physical tools.



Speed Layer n

Serving Layer
Common RDBMS

Query

Alerts

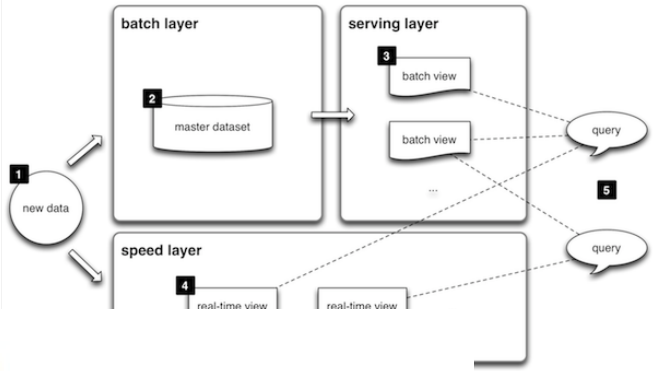
Response

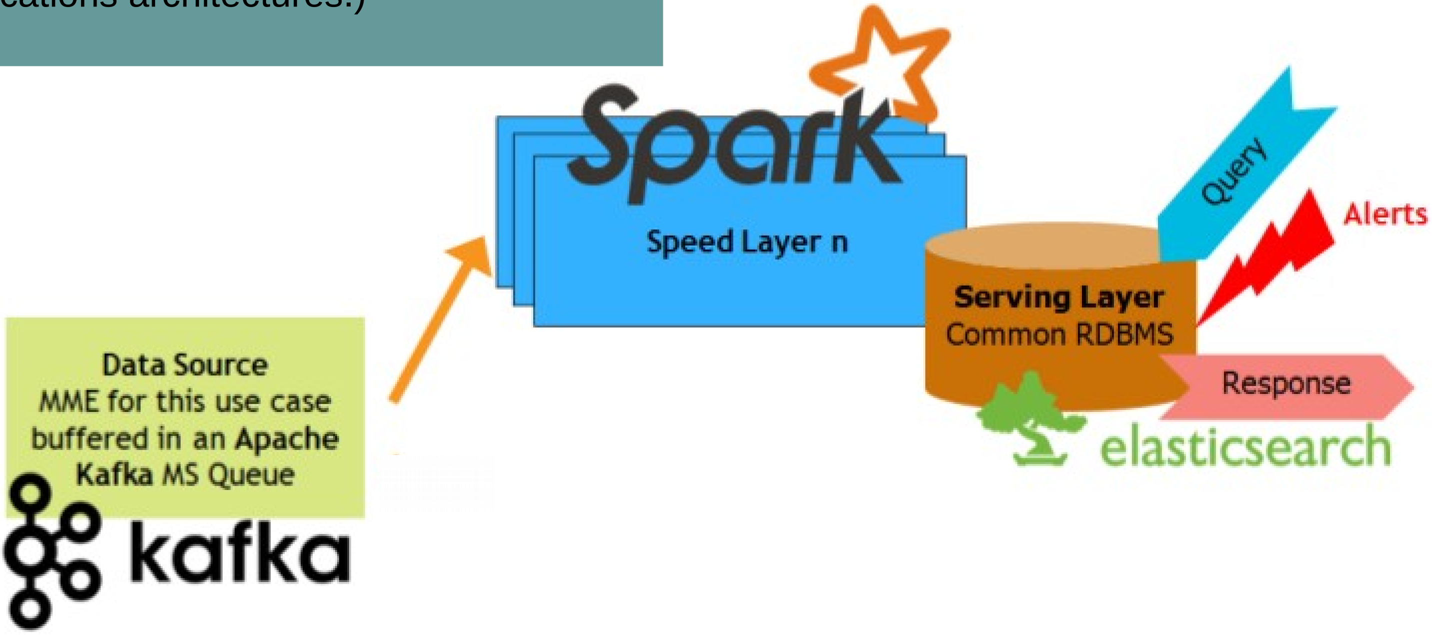
Data Source
MME for this use case
buffered in an Apache
Kafka MS Queue

Batch Layer
Typically build on Apache
Hadoop/HDFS

So, what is an big data analytic architecture?

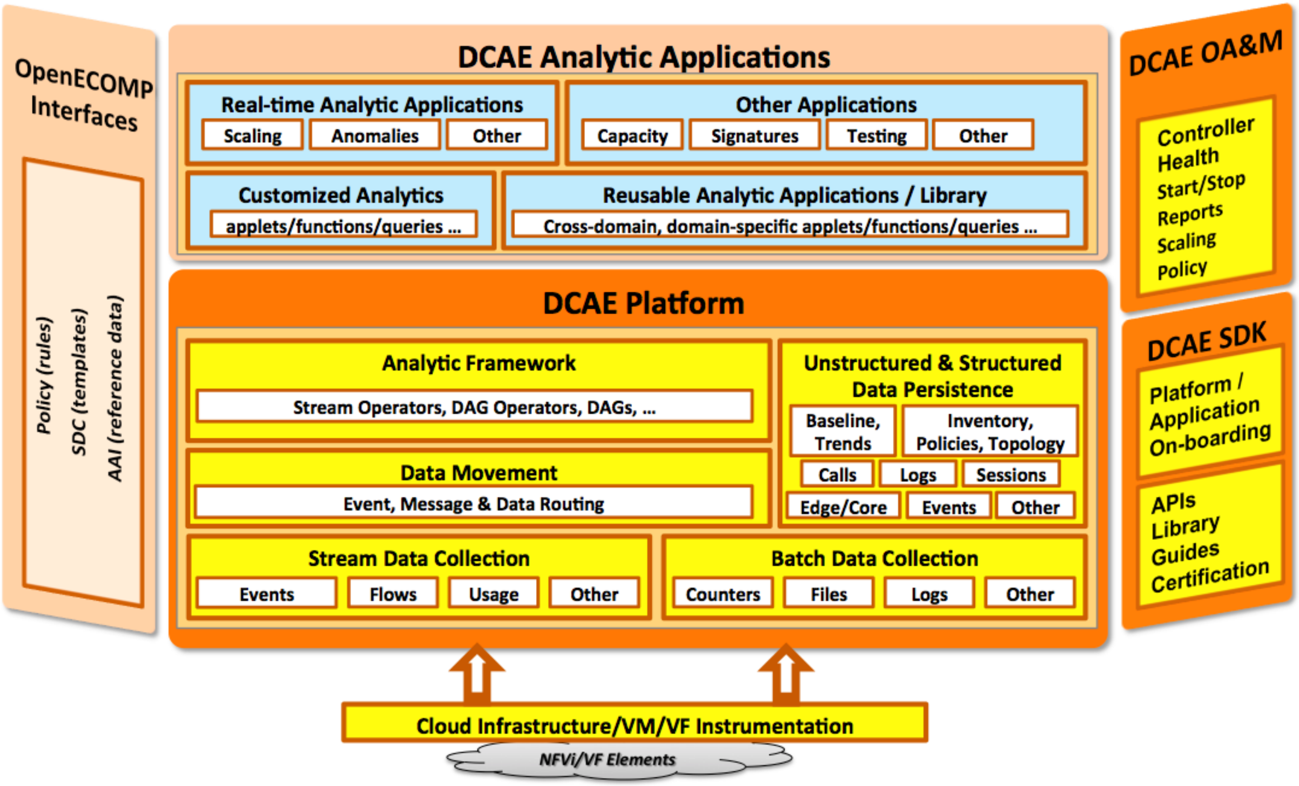
Around 2016, based on our work [Falk ‘17], we discovered that for we can simplify Lambda by removing the batch layer for telecommunication architectures.
=> Kappa Architecture (Jay Kreps, July 2014)
(But this was eliminate code redundancy, not because of performance issues related to telecommunications architectures.)





… And all this is important because …?

ONAP
OPEN NETWORK AUTOMATION PLATFORM



OpenECOMP Interfaces

Policy (rules)

SDC (templates)

AAI (reference data)

DCAE Analytic Applications

Real-time Analytic Applications: Scaling, Anomalies, Other

Other Applications: Capacity, Signatures, Testing, Other

Customized Analytics: applets/functions/queries …

Reusable Analytic Applications / Library: Cross-domain, domain-specific applets/functions/queries …

DCAE Platform

Analytic Framework: Stream Operators, DAG Operators, DAGs, …

Unstructured & Structured Data Persistence: Baseline, Trends; Inventory, Policies, Topology; Calls; Logs; Sessions; Edge/Core; Events; Other

Data Movement: Event, Message & Data Routing

Stream Data Collection: Events, Flows, Usage, Other

Batch Data Collection: Counters, Files, Logs, Other

Cloud Infrastructure/VM/VF Instrumentation

NFVi/VF Elements

DCAE OA&M

Controller
Health
Start/Stop
Reports
Scaling
Policy

DCAE SDK

Platform / Application On-boarding

APIs
Library
Guides
Certification

Challenges

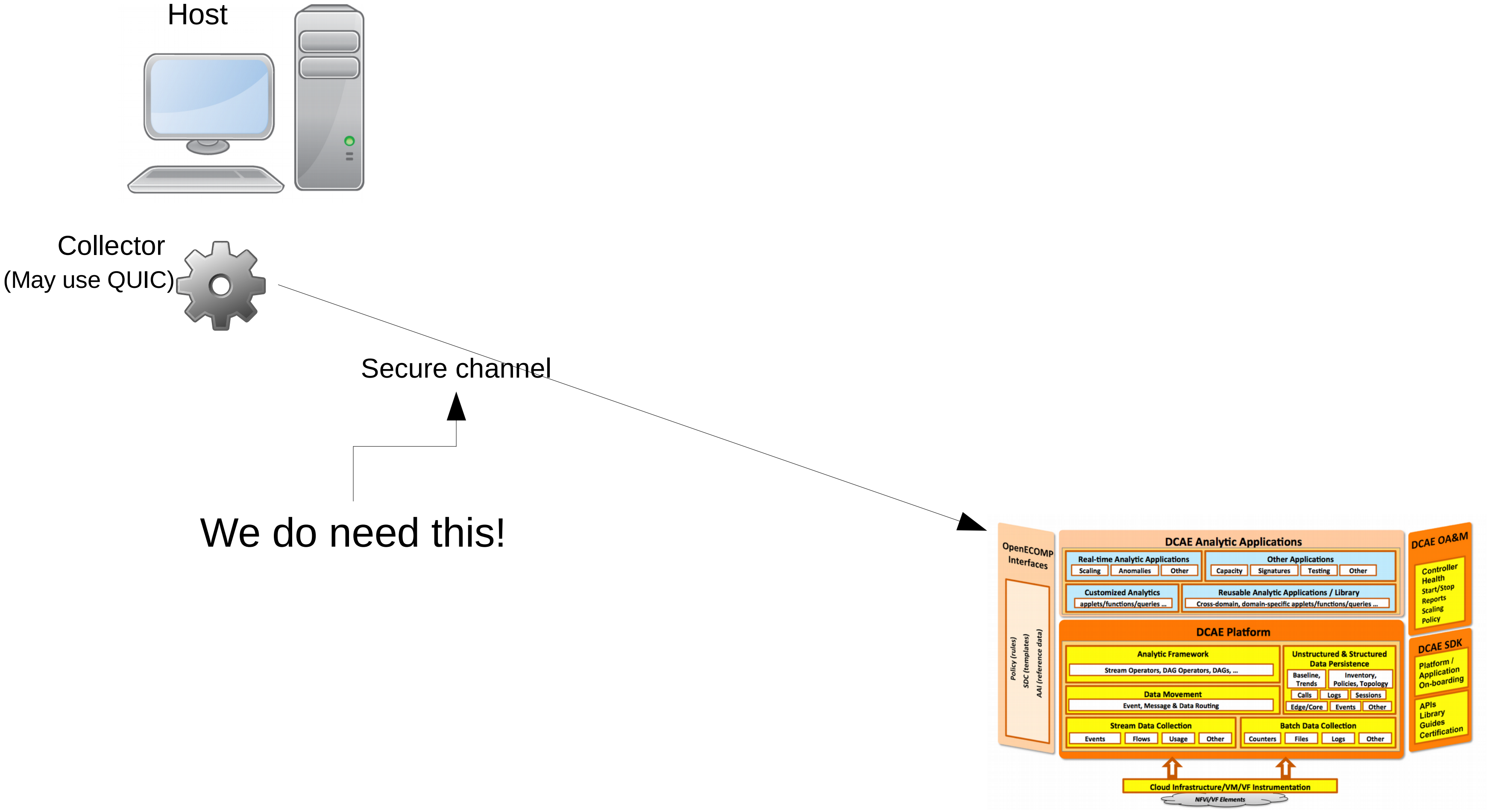
- For telecommunication service providers, the analytic architecture is composed of:
    - Distributed components
    - Collectors: run on remote hosts and collect data (performance counters, KPIs).
    - Aggregation points: Aggregate data collected by the collectors.
    - Machine learning models that fit the collected data to a model.
    - Visualization components to visualize the results of the fitted model and provide insight.

Challenges

- For telecommunication service providers, the analytic architecture is composed of:
  - Distributed components
  - Collectors: run on remote hosts and collect da ✗ (performance counters, KPIs).
  - Aggregation points: Aggregate data collected ✗ the collectors.
  - Machine learning models that fit the collecte ✗ data to a model.
  - Visualization components to visualize the resu ✗ of the fitted model and provide insight.

✗ = Will not tolerate encrypted data

What are the requirements?



Host

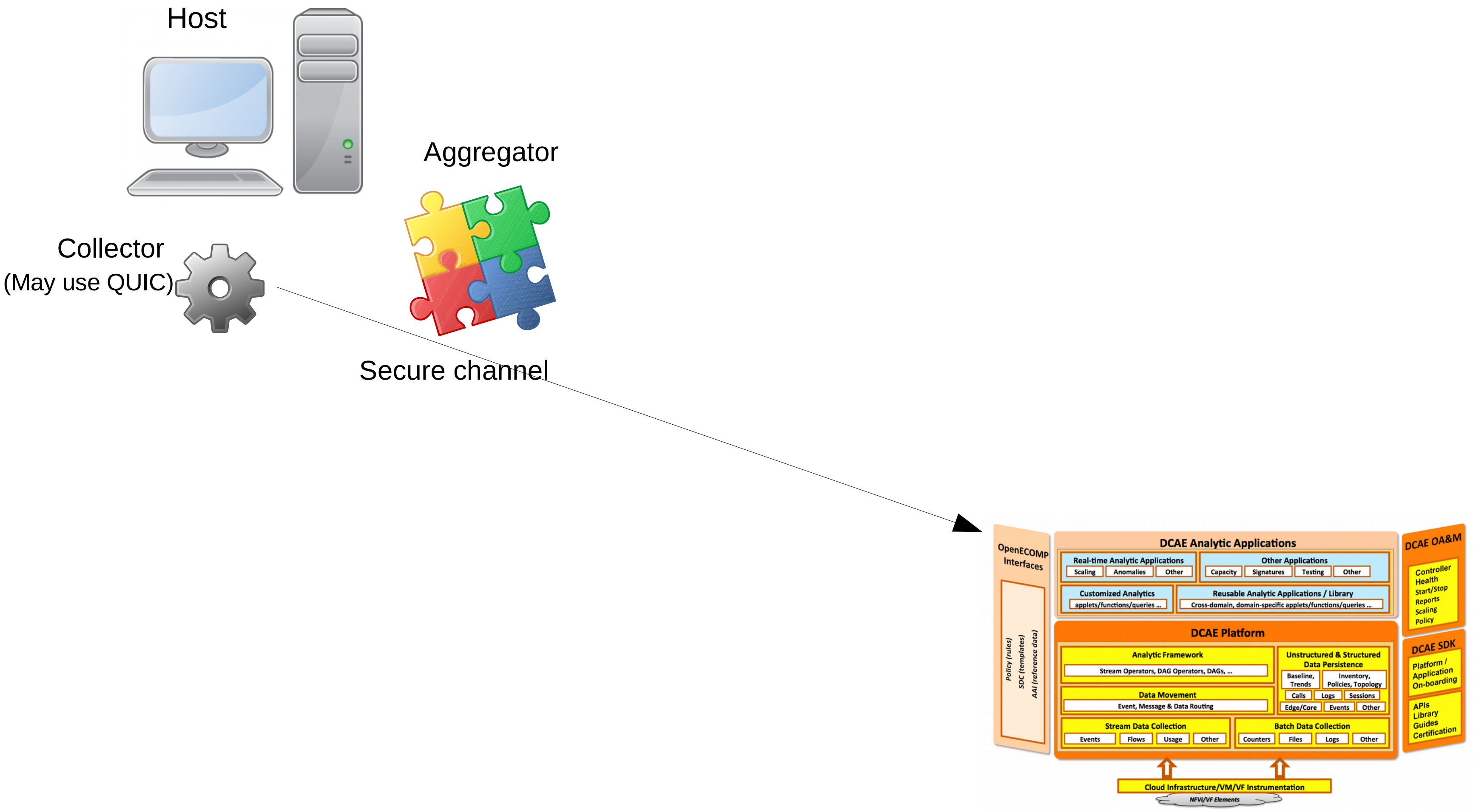
Collector
(May use QUIC)

Secure channel

We do need this!

Analytic Platform

What are the requirements?



Host

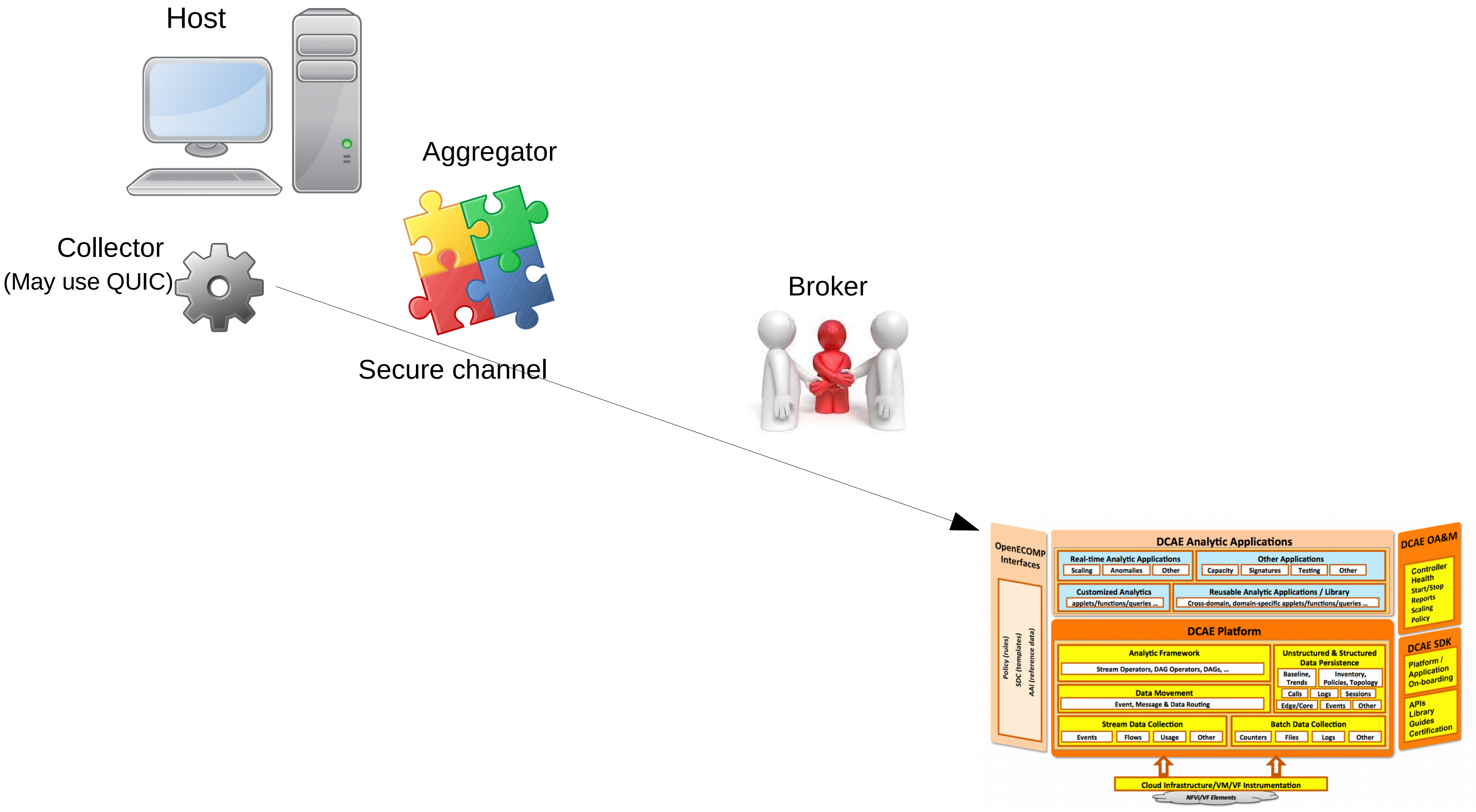
Aggregator

Collector
(May use QUIC)

Secure channel

Analytic Platform

What are the requirements?



Host

Aggregator

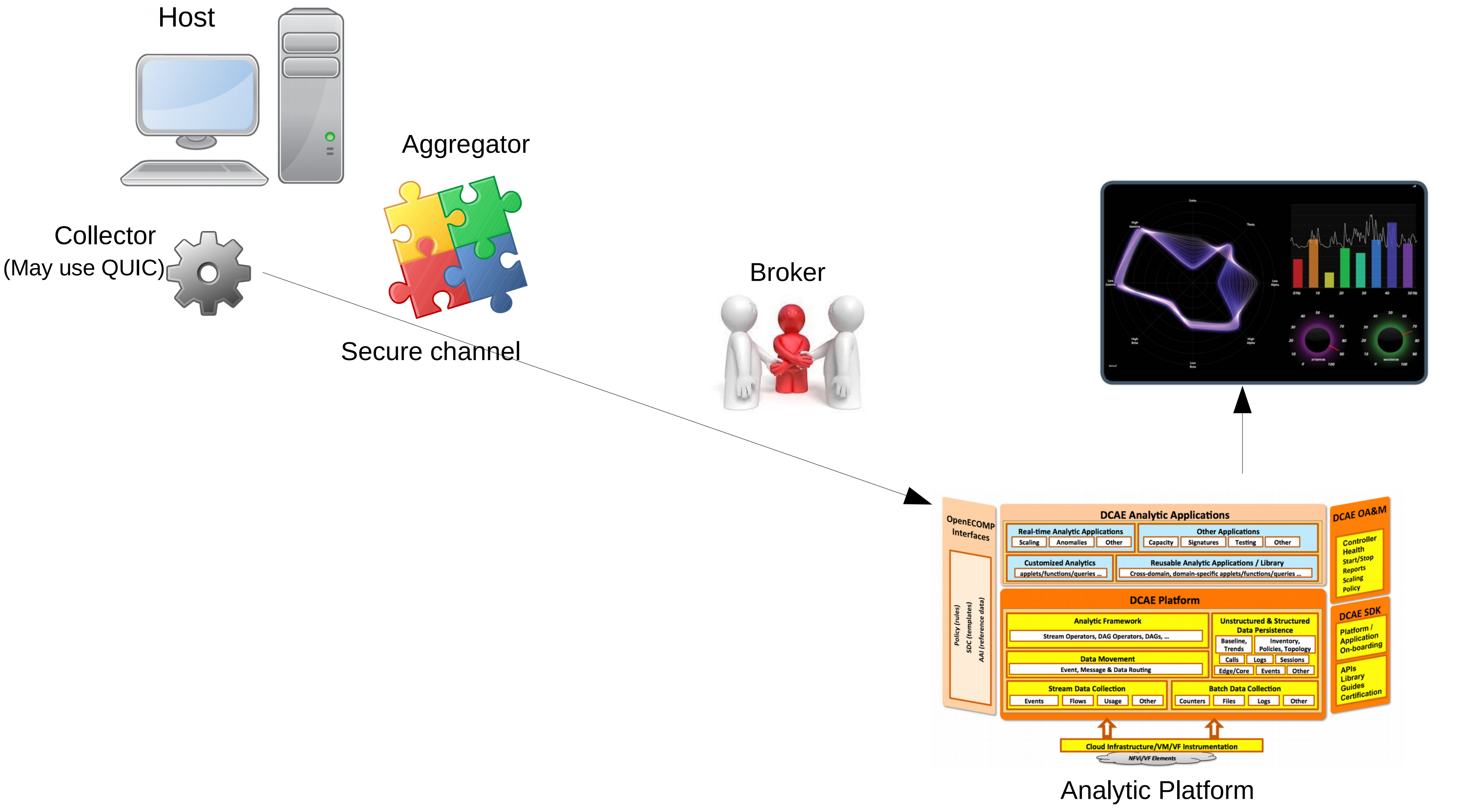
Collector
(May use QUIC)

Broker

Secure channel

Analytic Platform

What are the requirements?



Host

Aggregator

Collector
(May use QUIC)

Secure channel

Broker

Analytic Platform

How to best design for security in analytic platforms?

- Can we use:
  - Hop-by-hop decryption, analysis and re-encryption?
    - Will it scale for real-time analytics?
      - Kafka/LinkedIn mean: 2.5 million msgs/sec
      - Assume 1,500 bytes/msg and we get ~3.7e9 bytes/sec to process in real-time.
    - With data streams observing updates on millisecond frequencies, hop-by-hop decryption, analysis and re-encryption will not keep up with high-velocity streams.

How to best design for security in analytic platforms?

- Can we use:
  - Homomorphic encryption?
    - Possible to use at aggregation points to perform mathematical operations on encrypted data.
    - Disadvantages:
      - Slow.
      - Will note scale to high-velocity data streams.

How to best design for security in analytic platforms?

- If data is encrypted, can we use intraflow metadata for predictive analytics in network operations?
  - Flow metadata?
    - What is flow metadata?
      - Packet length (of encrypted packets) in a flow,
      - Packet inter-arrival time,
        - Sub-classed as: inter-packet delay variation and inter-packet generation delay;
      - Packet bursts/storms.
    - This allows:
      - Encrypted traffic identification [Gu '11]
      - Remote identification of encrypted video streams [Schuster '17]
  - Flow metadata may help in certain application analytics, but not network-level analytics.

Are we there yet?

- Takeaway:
  - There are pieces to the puzzle that help.
  - But, there is no cohesive solution that works well end to end.
  - Most analytic data today (performance measurements, KPIs, byte count) flows in cleartext from host/collector to the analytic platform.
  - We should do better by natively thinking about how to provide security in analytic platforms.

References

[Falk ‘17] Eric Falk, Vijay K. Gurbani and Radu State, “Query-able Kafka: An agile data analytics pipleline for mobile wirless networks,” PVLDB 10(12): 1646-1657 (2017)

[Gu ‘11] Chengjie Gu, Shunyi Zhang and Yanfei Sun, “Real-time Encrypted Traffic Identification using machine learning,” Journal of Software, 6(6): 1009-1016, 2011.

[Schuster ‘17] Roei Schuster and Vitaly Shmatikov, “Beauty and the Burst: Remote Identification of Encrypted Video Streams,” Proceedings of the 26th Usenix Security Symposium, 2017.