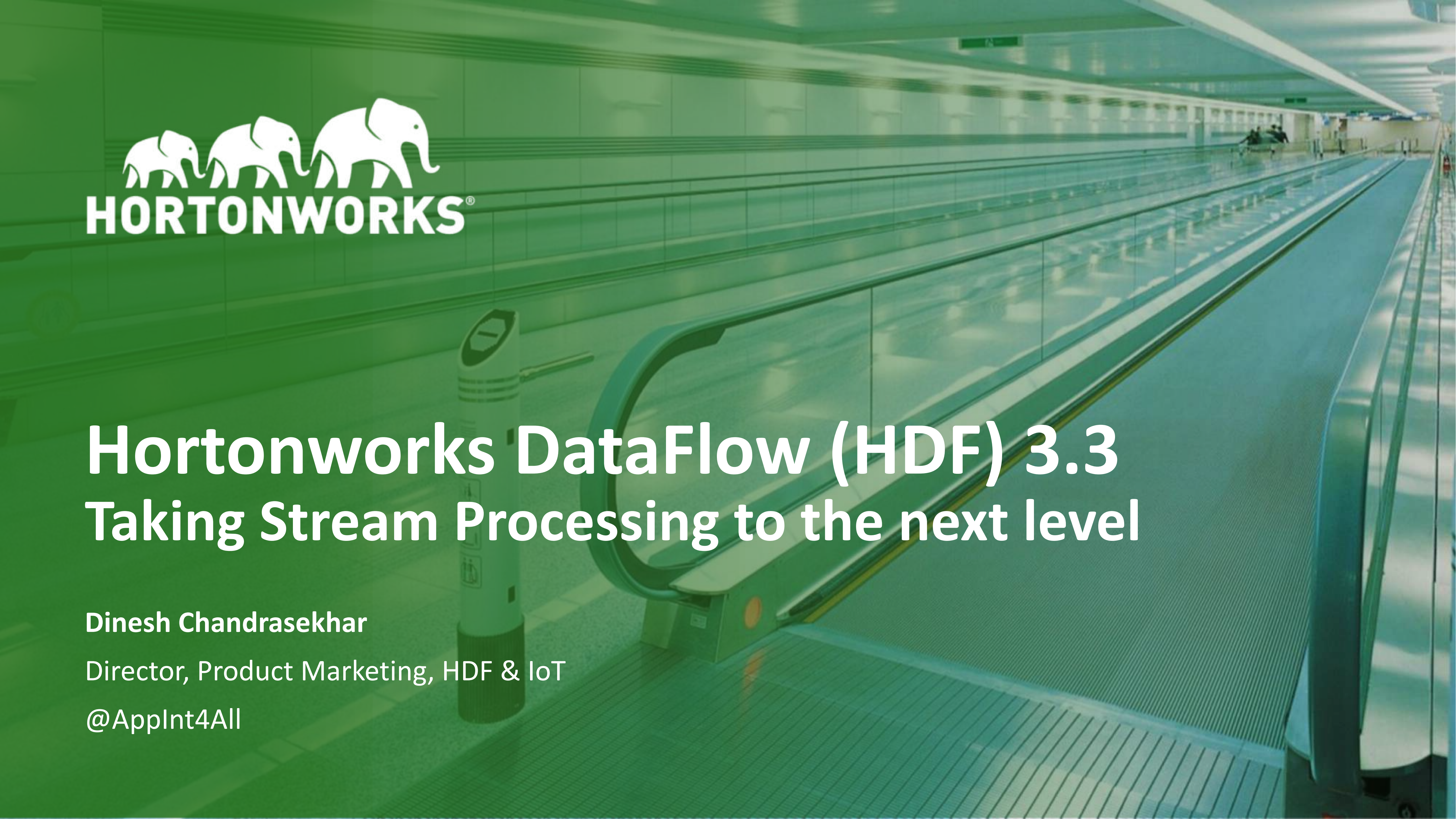


HORTONWORKS®

Hortonworks DataFlow (HDF) 3.3

Taking Stream Processing to the next level

Dinesh Chandrasekhar

Director, Product Marketing, HDF & IoT

@AppInt4All

What Is Hortonworks DataFlow (HDF)?

Hortonworks DataFlow (HDF) is a scalable, real-time streaming data platform that collects, curates, and analyzes data so customers gain key insights for immediate actionable intelligence.

HDF Data-In-Motion Platform

Flow Management

Data acquisition and delivery
Simple transformation and data routing
Simple event processing
Edge to Enterprise data lineage and provenance
Edge device connectivity and IoT data ingestion

nifi

minifi

C++ Agent

Java Agent

Stream Processing

Scalable data broker for streaming apps
Scale out streaming computation engine

kafka

STORM

Stream Analytics

Pattern Matching
Prescriptive & Predictive Stream Analytics
Complex Event Processing
Continuous Insights

STREAMING ANALYTICS MANAGER

Enterprise Services

Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance, Multi-tenancy

APACHE KNOX

SCHEMA REGISTRY

APACHE NIFI registry

Ambari

Apache Ranger

Common HDF Use Cases

Data Movement
Optimize resource utilization by moving data between data centers or between on-premises infrastructure and cloud infrastructure

Optimize Log Collection & Analysis
Optimize log analytics solutions such as Splunk by using HDF as a single platform to collect and deliver multiple data sources and using HDP for lower cost storage options

Gain key insights with Streaming Analytics
Accelerate big data ROI by analyzing streaming data for patterns, comparing with ML models and delivering actionable intelligence

Single view / 360˚ view of customer
Ingest, transform and combine customer data from multiple sources into a single data view / lake

Stream Processing
Combine multiple streams of data in real-time, enrich the data and route it to different end points based on rules

Capture IoT Data
Ingest sensor data from IoT devices and stream it for further processing and comprehensive analysis

Clearsense

Improving Healthcare with SMART data

CHALLENGE

Combine multi-format data streams, with hundreds of sources, into one platform

- Needed a platform that could combine multi-format data streaming
- Data scarcity & latency problems
- Machine learning & data science

SOLUTION

Cloud-based systems architected to deliver SMART data, using HDP and HDF

- First to deliver SMART real-time streaming data
- Clearsense's Inception™ product enables fast decisions for clinicians
- Customers have access to all data sources with HDP & HDF

RESULT

Mission-critical data and relevant insight for 2,000 rural providers

- Mission critical data is now available for doctors to make critical decisions
- Cost efficiencies led to access for 2,000 rural providers
- Real-time data helps prevent "Code Blue"

Photo by rawpixel on Unsplash

HORTONWORKS

Trimble

Positioning technology products & services empower companies worldwide

CHALLENGE

Provide accurate data for small carriers to improve business results

- 95% of small carriers (less than 50 trucks) have a deficit of data available
- Estimated data, price points and revenue base opportunity for controlling fuel cost
- Understanding of freight and lane movement

SOLUTION

Big Data in the Cloud with HDP, HDF, and Microsoft Azure

- Leveraging big data powering Blockchain, with machine learning, to revolutionize Transportation and Logistics industries
- Analyzed fuel data; can consolidate data set for small carriers to generate community data lake

RESULT

Double digit revenue increase, year over year

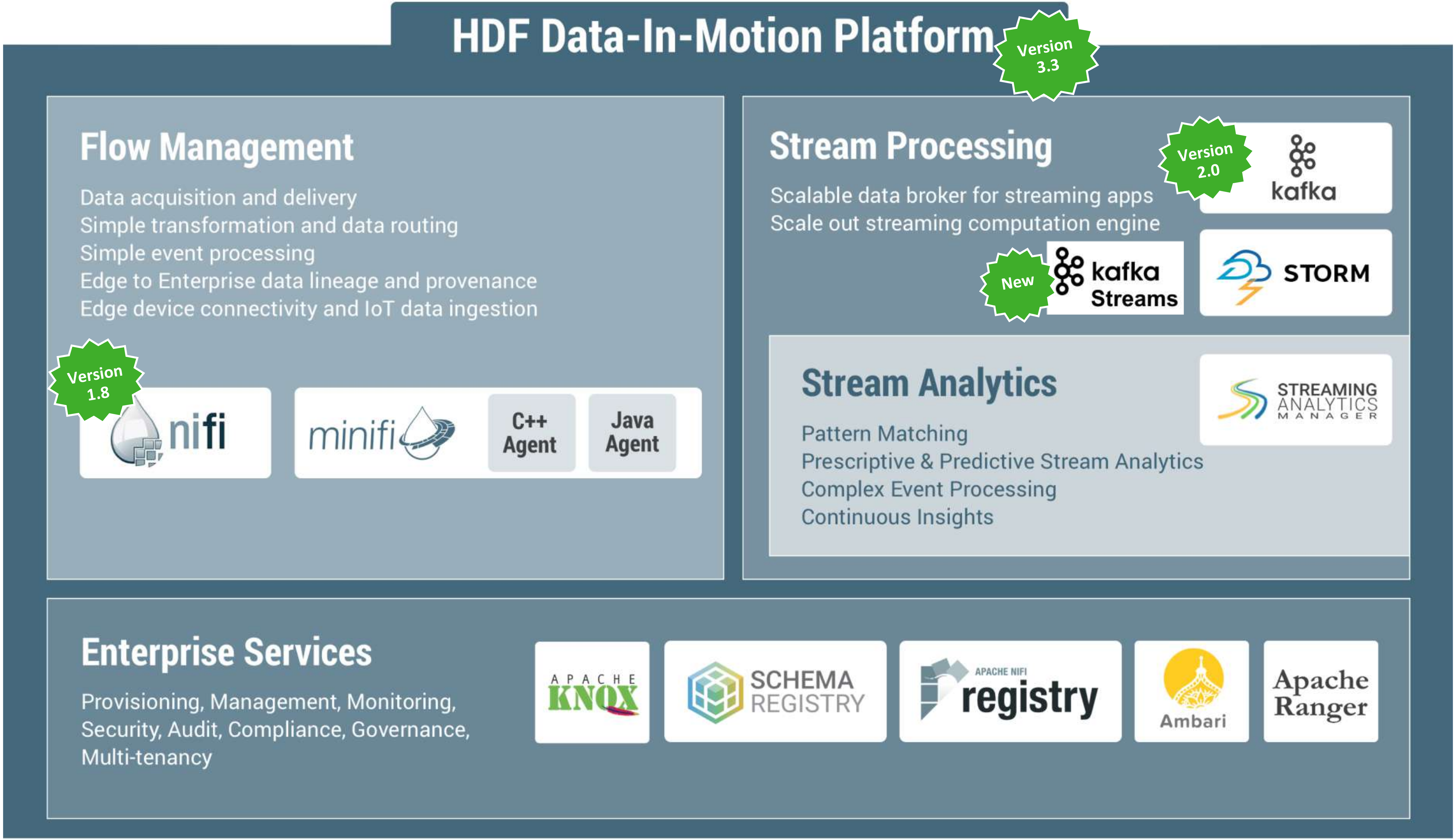
- Managing for 4 million trucks daily
- $31 billion dollars in freight movement guides customers to profitability
- Blockchain driven architecture

Photo by rawpixel.com on Unsplash

What’s new in HDF 3.3?

HDF Data-In-Motion Platform

Version 3.3



Flow Management

Data acquisition and delivery
Simple transformation and data routing
Simple event processing
Edge to Enterprise data lineage and provenance
Edge device connectivity and IoT data ingestion

Stream Processing

Scalable data broker for streaming apps
Scale out streaming computation engine

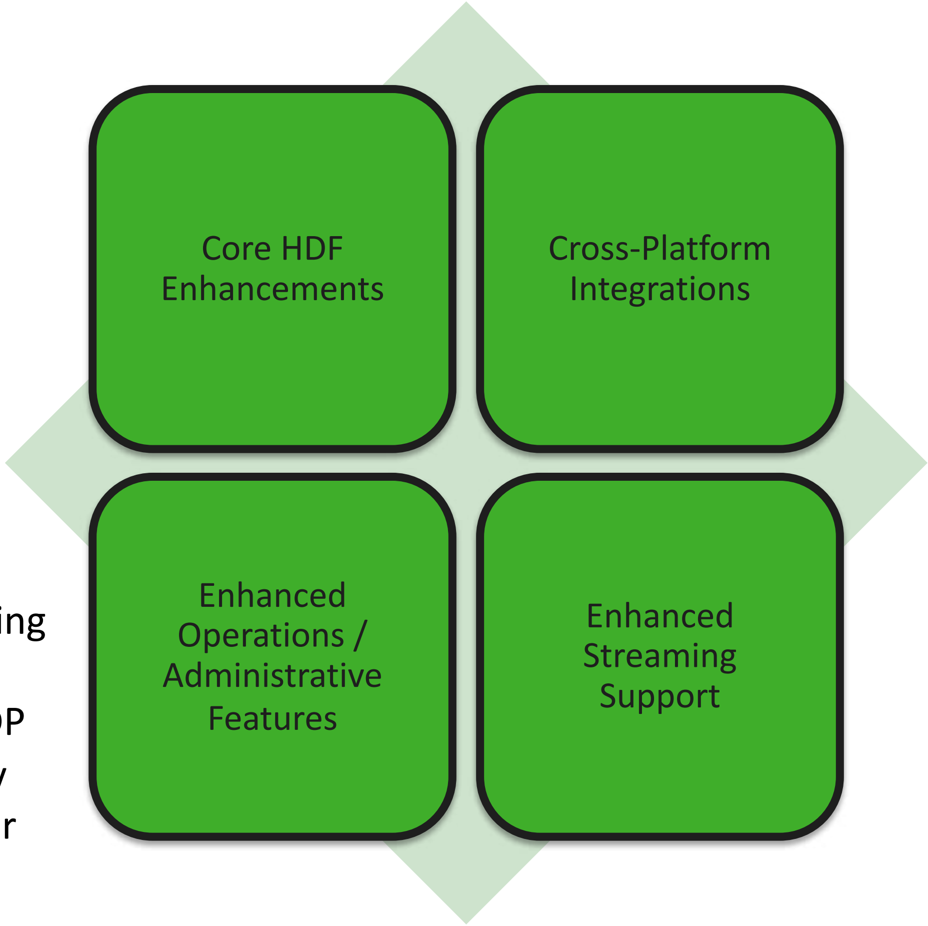
Stream Analytics

Pattern Matching
Prescriptive & Predictive Stream Analytics
Complex Event Processing
Continuous Insights

Enterprise Services

Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance, Multi-tenancy

HDF 3.3 Release Themes



Core HDF Enhancements

Cross-Platform Integrations

Enhanced Operations / Administrative Features

Enhanced Streaming Support

- Support for Kafka 2.0
- Kafka 2.0 NiFi processors
- NiFi connection load balancing
- MQTT performance improvements

- Kafka 2.0
  - Ambari and Ranger
- Kafka Streams
  - Schema Registry and Ranger
- Knox SSO Support
  - Schema Registry and SAM

- Ambari support for Express and Rolling upgrades from HDF 3.2
- Smartsense use without needing HDP
- Ranger integration with NiFi Registry
- Decommission NiFi nodes in a cluster

- Kafka Streams Support
  - Integration with SMM
- With HDP 3.1
  - New Hive Kafka Storage Handler
  - New Druid Kafka Indexing Service

Core HDF Enhancements

Core HDF Enhancements

Support for NiFi 1.8.0 and Kafka 2.0

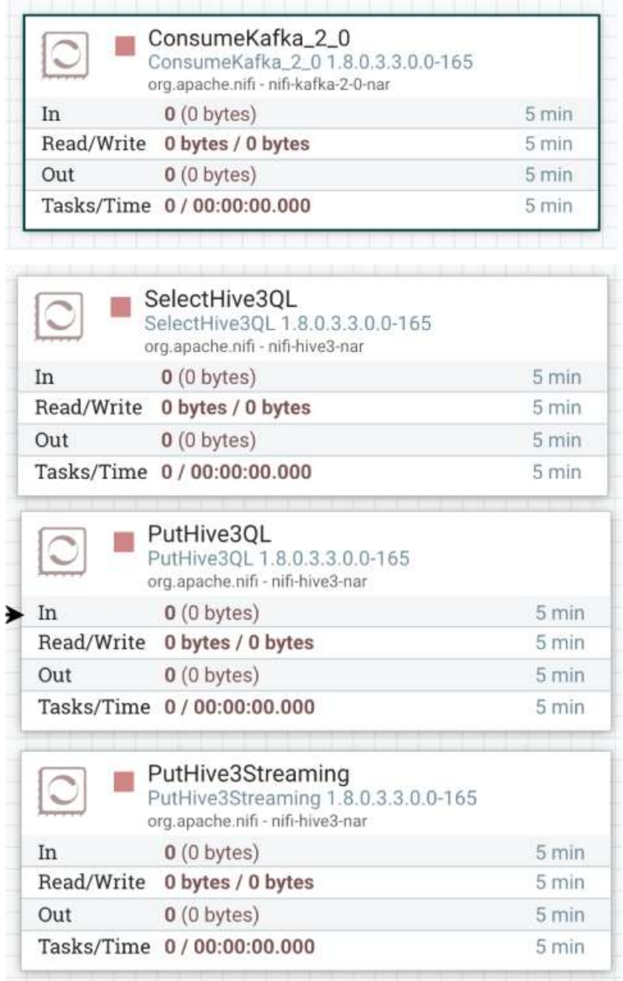
- NiFi serves as the centralized hub for managing dataflows in an enterprise environment.
- Kafka serves as the central component in every major streaming architecture.

Key New Features

- Kafka 2.0 support
- Hive 3.1.0 support
- Connection load balancing
- MQTT Performance improvements

Updated to 1.8.0

Updated to 2.0



ConsumeKafka_2_0
ConsumeKafka_2_0 1.8.0.3.3.0.0-165
org.apache.nifi - nifi-kafka-2-0-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

SelectHive3QL
SelectHive3QL 1.8.0.3.3.0.0-165
org.apache.nifi - nifi-hive3-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

PutHive3QL
PutHive3QL 1.8.0.3.3.0.0-165
org.apache.nifi - nifi-hive3-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

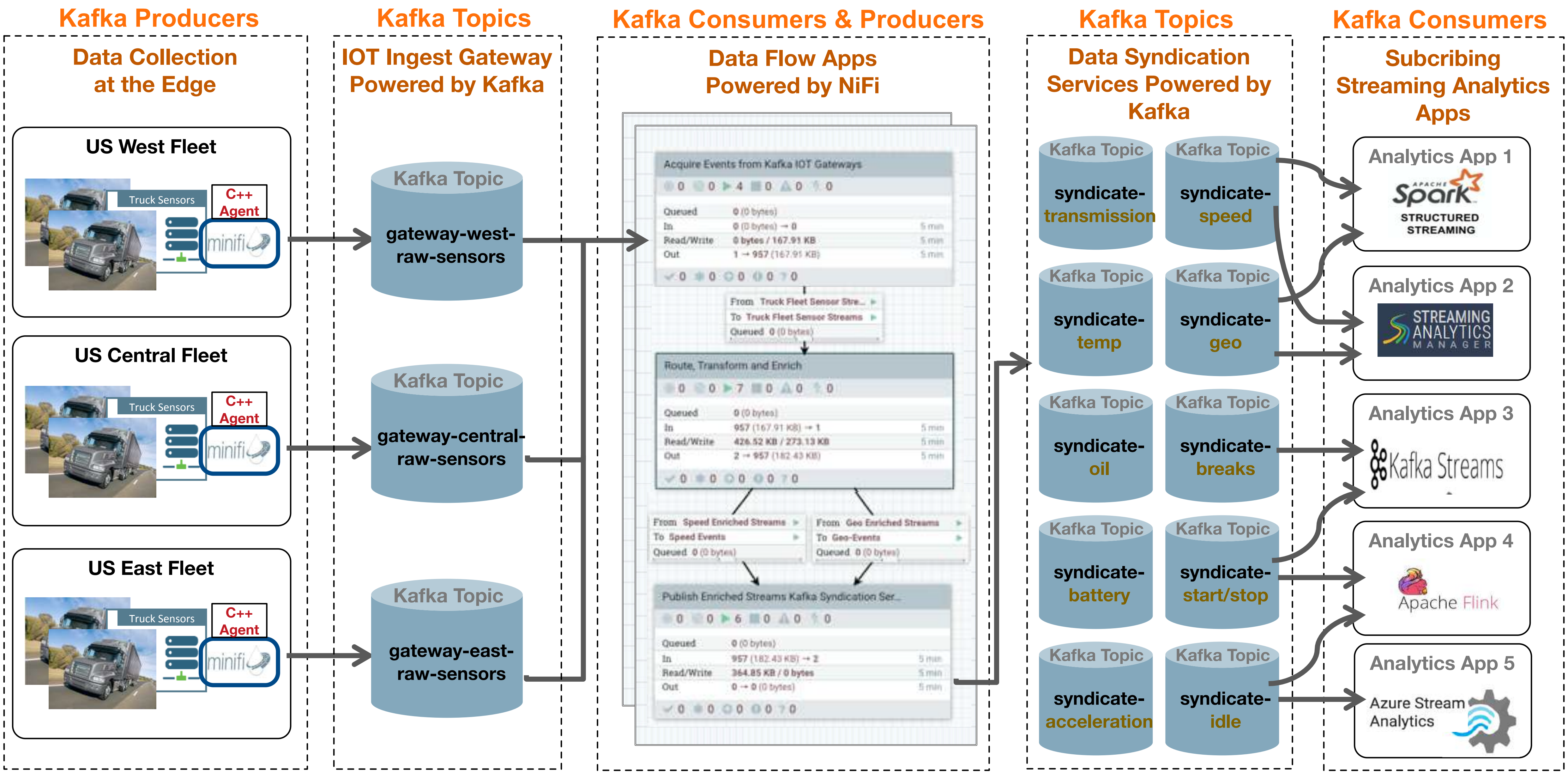
PutHive3Streaming
PutHive3Streaming 1.8.0.3.3.0.0-165
org.apache.nifi - nifi-hive3-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Enhanced Streaming Support

Streaming Analytics Reference Architecture

Kafka is Everywhere. Critical Component of Streaming Architectures



Kafka Producers

Data Collection at the Edge

US West Fleet

Truck Sensors

C++ Agent

minifi

US Central Fleet

Truck Sensors

C++ Agent

minifi

US East Fleet

Truck Sensors

C++ Agent

minifi

Kafka Topics

IOT Ingest Gateway Powered by Kafka

Kafka Topic gateway-west-raw-sensors

Kafka Topic gateway-central-raw-sensors

Kafka Topic gateway-east-raw-sensors

Kafka Consumers & Producers

Data Flow Apps Powered by NiFi

```
Acquire Events from Kafka IOT Gateways
0  0  4  0  0  0
Queued       0 (0 bytes)
In           0 (0 bytes) → 0                      5 min
Read/Write   0 bytes / 167.91 KB                  5 min
Out          1 → 957 (167.91 KB)                  5 min
0  0  0  0  0

From  Truck Fleet Sensor Stre...
To  Truck Fleet Sensor Streams
Queued  0 (0 bytes)

Route, Transform and Enrich
0  0  7  0  0  0
Queued       0 (0 bytes)
In           957 (167.91 KB) → 1                  5 min
Read/Write   426.52 KB / 273.13 KB                5 min
Out          2 → 957 (182.43 KB)                  5 min
0  0  0  0  0

From  Speed Enriched Streams        From  Geo Enriched Streams
To  Speed Events                    To  Geo-Events
Queued  0 (0 bytes)                 Queued  0 (0 bytes)

Publish Enriched Streams Kafka Syndication Ser...
0  0  6  0  0  0
Queued       0 (0 bytes)
In           957 (182.43 KB) → 2                  5 min
Read/Write   364.85 KB / 0 bytes                  5 min
Out          0 → 0 (0 bytes)                      5 min
0  0  0  0  0
```

Kafka Topics

Data Syndication Services Powered by Kafka

Kafka Topic syndicate-transmission

Kafka Topic syndicate-speed

Kafka Topic syndicate-temp

Kafka Topic syndicate-geo

Kafka Topic syndicate-oil

Kafka Topic syndicate-breaks

Kafka Topic syndicate-battery

Kafka Topic syndicate-start/stop

Kafka Topic syndicate-acceleration

Kafka Topic syndicate-idle

Kafka Consumers

Subcribing Streaming Analytics Apps

Analytics App 1

APACHE Spark STRUCTURED STREAMING

Analytics App 2

STREAMING ANALYTICS MANAGER

Analytics App 3

Kafka Streams

Analytics App 4

Apache Flink

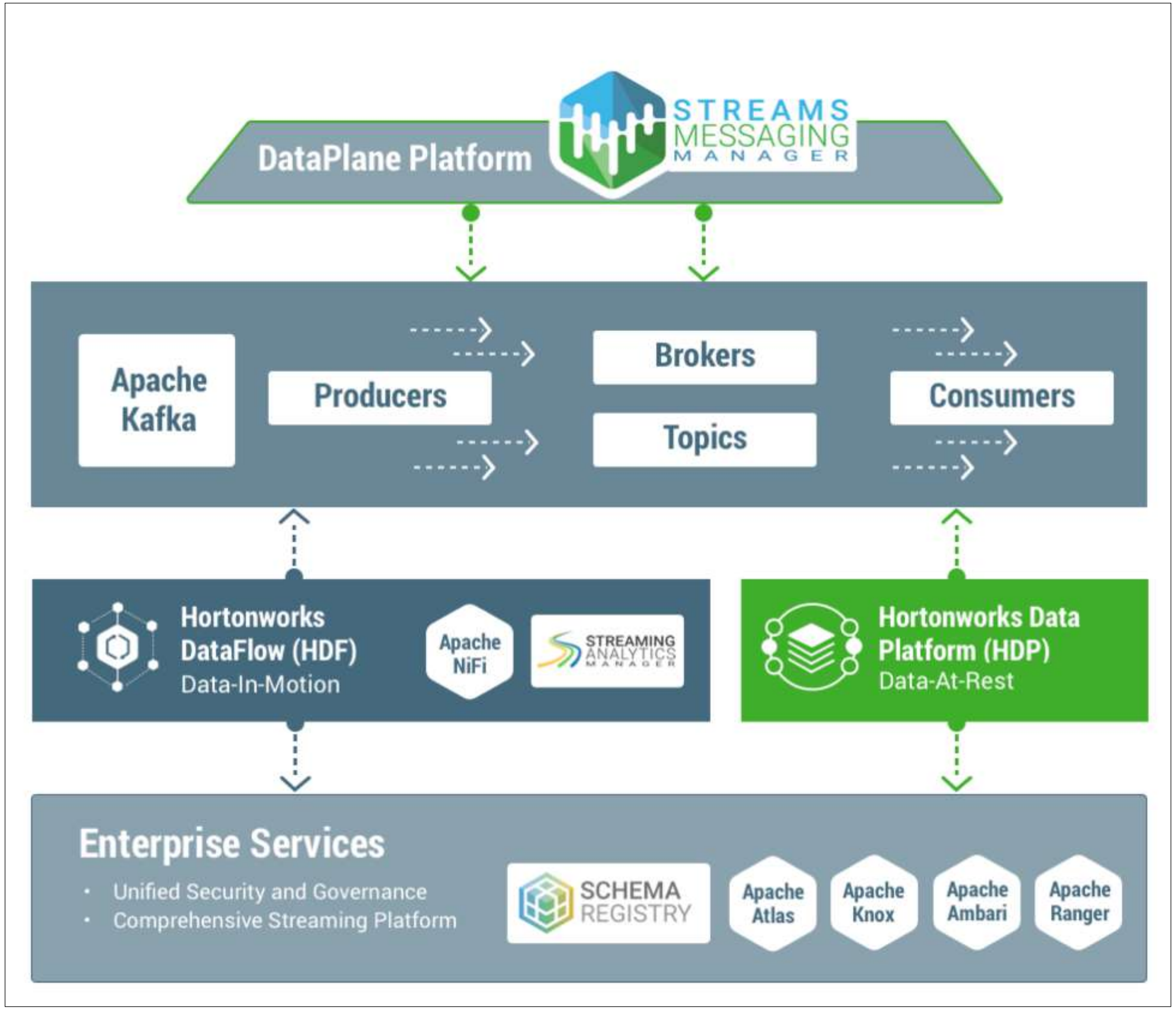
Analytics App 5

Azure Stream Analytics

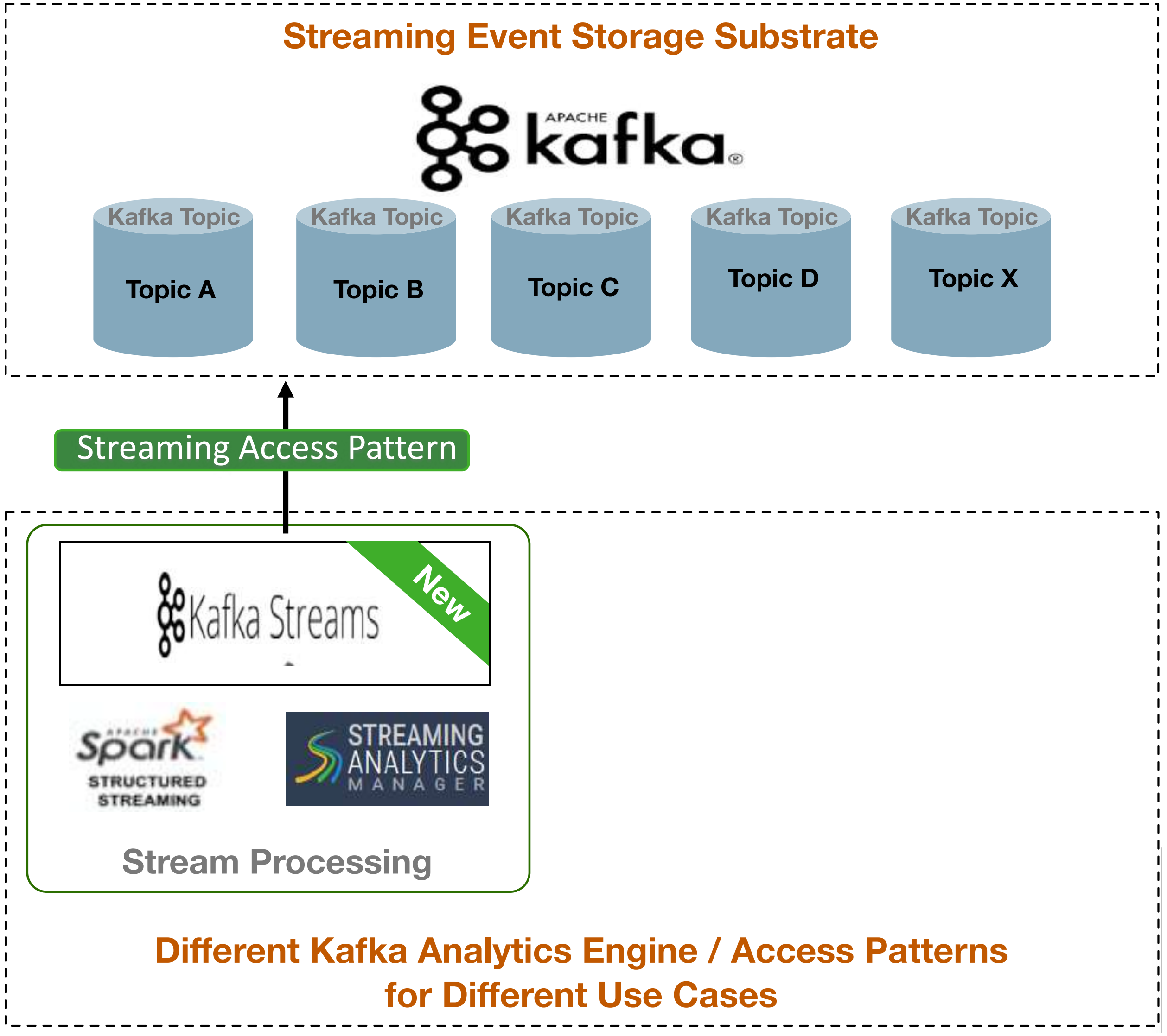
Cure is Here: Hortonworks Streams Messaging Manager (SMM)

What is SMM?

- Kafka Management and Monitoring tool
- Cure the “Kafka Blindness”
- Single Monitoring Dashboard for all your Kafka Clusters across 4 entities
  - Broker
  - Producer
  - Topic
  - Consumer
- Supports multiple HDP and/or HDF Kafka Clusters
- REST as a First Class Citizen
- Delivered as a DataPlane Service



3 New Kafka Analytics Access Patterns



Streaming Event Storage Substrate

Different Kafka Analytics Engine / Access Patterns for Different Use Cases

When is Kafka Streams an ideal choice for Stream Processing?

- Your Application consists of Kafka to Kafka Pipeline
- You don’t need/want another cluster for stream processing
- You want to perform common stream processing functions like filtering, joins, aggregations, enrichments on the stream for simpler stream processing apps
- Your target user are developers with java dev backgrounds
- Your use cases are building lightweight microservices, simple ETL and stream analytics apps

Common Microservice / Streaming Analytics Requirements

Requirement #	Requirement Description
Req. #1	Create streams consuming from the two Kafka topics.
Req. #2	Join the streams of the Geo and Speed sensors over a time based aggregation window.
Req. #3	Apply rules on the stream to filter on events of interest.
Req. #4	Calculate the average speed of driver over 3 minute window and create alert for speeding driver
Req. #5	Find all the drivers who have be speeding (> 80) over that 3 minute window
Req. #6	Send alerts for speeding drivers to downstream alert topic
Req. #7	Apply access control (ACL) to the the source kafka topics, the alert topic and intermediate topics that are created by Kafka Streams apps
Req. #8	Monitor each MicroService providing a view into producers, consumers, brokers, and key metrics like consumer group lag, etc.

Kafka Streams Microservices Architecture

```
{
    "eventTime":"2018-10-24 14:37:56.746",
    "eventTimeLong":1540391876746,
    "eventSource":"truck_geo_event",
    "truckId":885,
    "driverId":12,
    "driverName":"Joe Witt",
    "routeId":1,
    "route":"Saint Louis to Memphis",
    "eventType":"Unsafe tail distance",
    "latitude":37.47,
    "longitude":-89.71,
    "correlationId":1,
    "geoAddress":"No Address Available"
}
```

Req #1: Create streams consuming from 2 Kafka topics

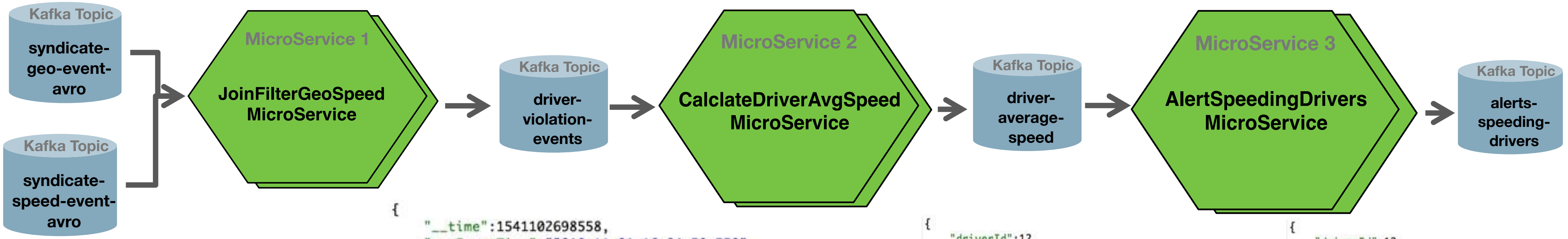
Req #2: Join the Geo & Speed sensors over a time based aggregation window.

Req #3: Apply rules on the stream to filter on events of interest.

Req #4: Calculate the average speed of driver over 3 minute window

Req #5: Find all the drivers who have be speeding (> 80) over that 3 minute window.

Req #6: Send alerts for speeding drivers to downstream alert topic



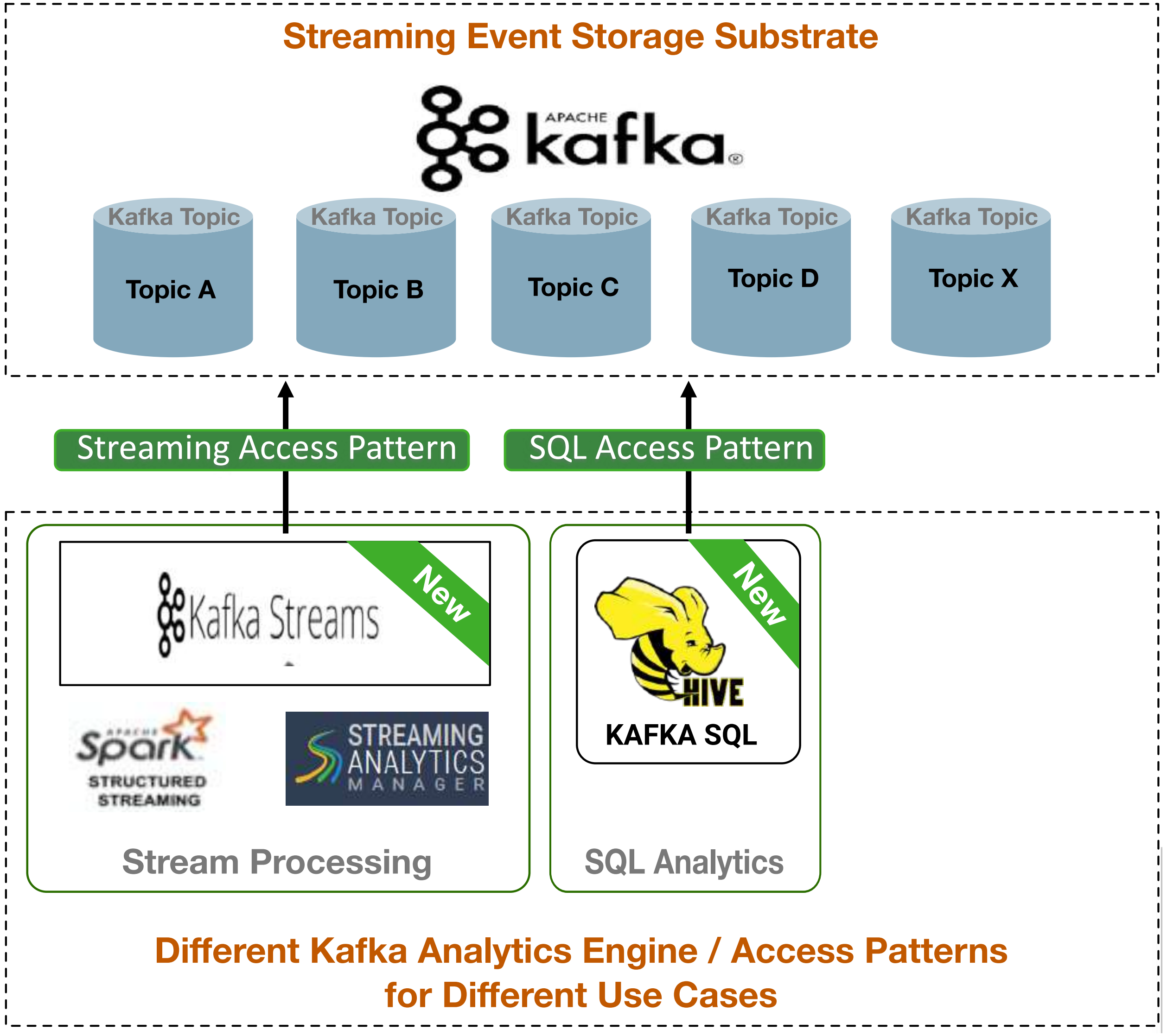
```
{
    "eventTime":"2018-10-24 14:56:24.937",
    "eventTimeLong":1540392984937,
    "eventSource":"truck_speed_event",
    "truckId":898,
    "driverId":12,
    "driverName":"Joe Witt",
    "routeId":5,
    "route":"Des Moines to Chicago",
    "speed":104
}
```

```
{
    "__time":1541102698558,
    "geoEventTime":"2018-11-01 16:04:58.558",
    "geoEventTimeLong":1541102698558,
    "eventSource":"truck_geo_event",
    "truckId":902,
    "driverId":12,
    "driverName":"Joe Witt",
    "routeId":12,
    "route":"Saint Louis to Memphis",
    "eventType":"Lane Departure",
    "latitude":38.62,
    "longitude":-90.15,
    "correlationId":1,
    "geoAddress":"No Address Available",
    "speedEventTime":"2018-11-01 16:04:58.559",
    "speedEventTimeLong":1541102698559,
    "speed":81
}
```

```
{
    "driverId":12,
    "driverName":"Joe Witt",
    "route":"Memphis to Little Rock Route 2",
    "speed_AVG":93.0,
    "processingTime":1541030916540
}
```

```
{
    "driverId":12,
    "driverName":"Joe Witt",
    "route":"Memphis to Little Rock Route 2",
    "speed_AVG":93.0,
    "processingTime":1541030916540
}
```

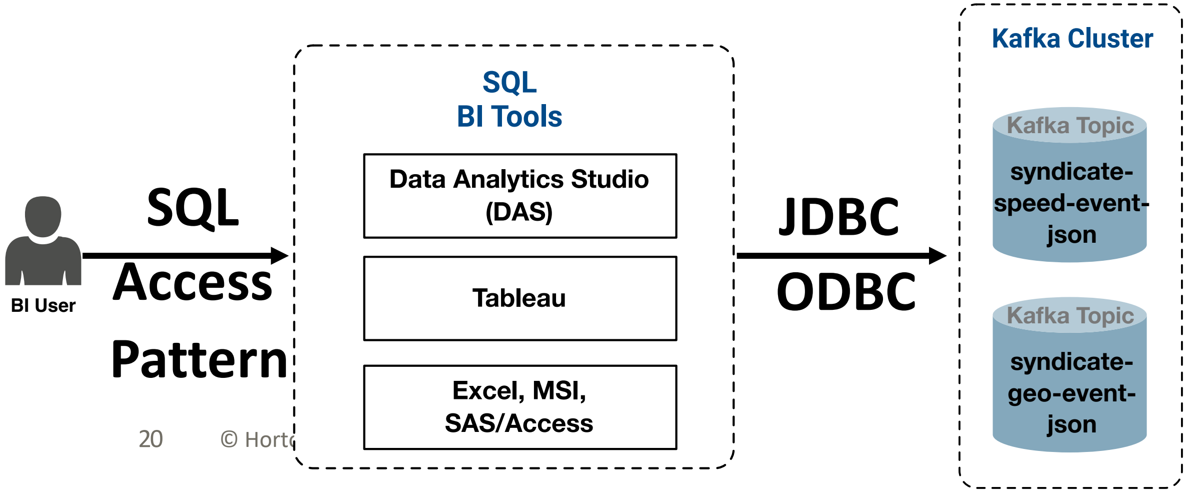
3 New Kafka Analytics Access Patterns



Kafka SQL – Interactive SQL Analytics

- Customers are starting to use **Kafka for longer term storage**. E.g.: Retention Periods for Kafka topics are getting longer

- Hence, customers want ability to query and perform interactive analytics on the data stored in Kafka topics



BI User

SQL Access Pattern

SQL BI Tools

Data Analytics Studio (DAS)

Tableau

Excel, MSI, SAS/Access

JDBC ODBC

Kafka Cluster

Kafka Topic

syndicate-speed-event-json

Kafka Topic

syndicate-geo-event-json

Key Requirements

- **SQL** is a must. Users don't want to learn a new engine/DSL to do analytics.

- **Kafka Topics becomes tables** that customers can execute SQL analytical queries

- **Joins** across multiple streaming kafka topics and traditional tables (reference data) for enrichment

- **Aggregations over windows** (group by, order by, lag, lead)

- **UDF support** for extensibility

- **JDBC/ODBC support**. Ablity to connect enterprise BI tools to do analytics on Kafka topics

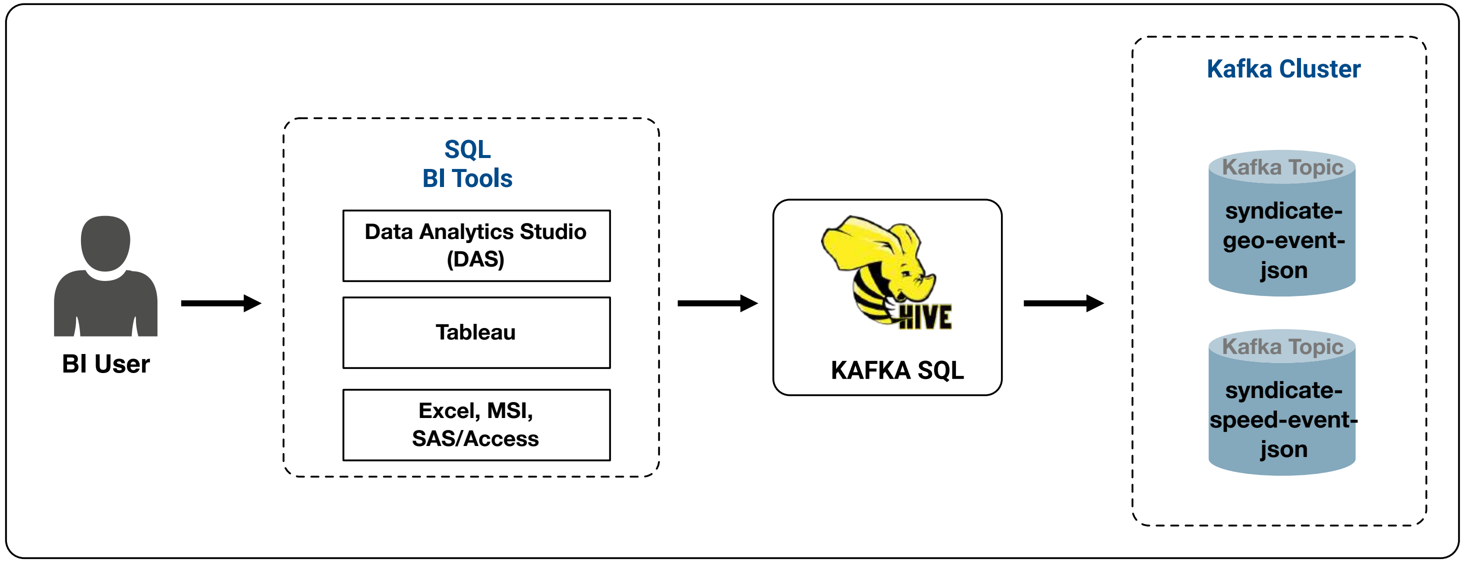
- Rich ACL support including **column level security**

- Governance with robust **Audit and Lineage**

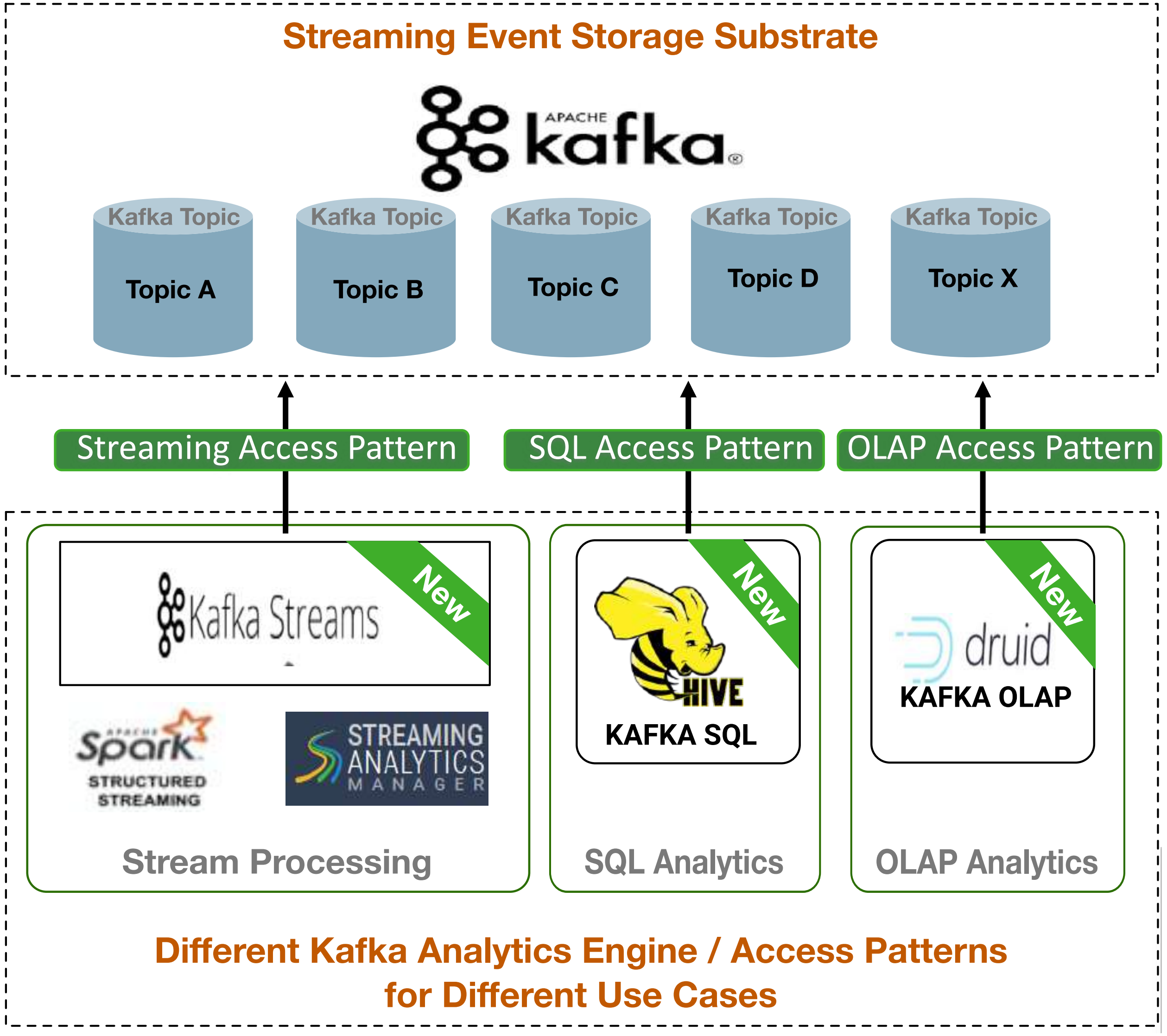
Hive Kafka SQL – “Real SQL on Real Time Stream”

What is Hive Kafka SQL?

- **New Hive Storage Handler for Kafka** that allows users to view Kafka topics as Hive tables.
- Takes full advantage of all the Hive analytical operators/capabilities supporting joins, aggregations, UDFs, push down predicate filtering, windowing etc.
- Support full Hive/Ranger integration enabling capabilities such as column level ACLs for events in Kafka topics
- Supports secure/Kerberized Hive and Kafka deployments



3 New Kafka Analytics Access Patterns



Streaming Event Storage Substrate

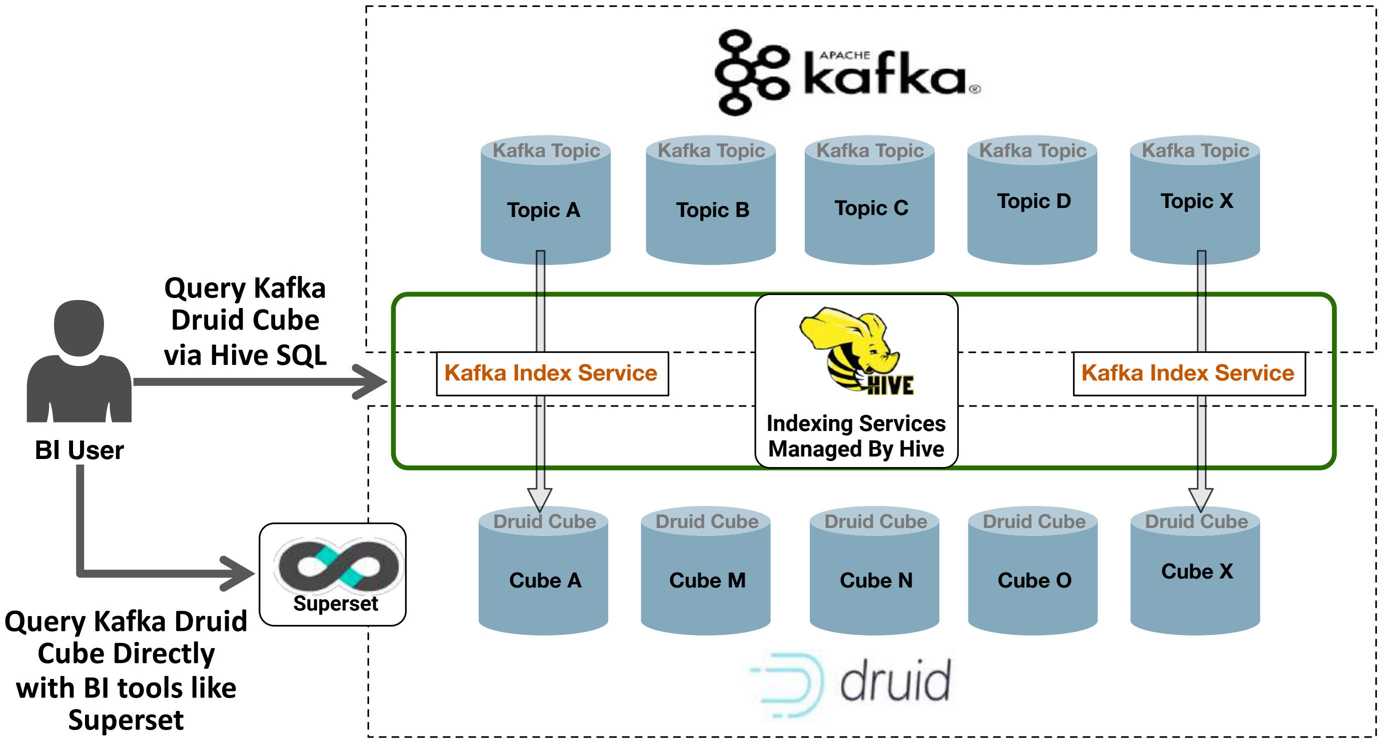
Different Kafka Analytics Engine / Access Patterns for Different Use Cases

Kafka + Druid + Hive = Powerful New Access Pattern for Streaming Data in Kafka

- Apache Druid (incubating) is a high performance analytics data store for event-driven data. Druid combines ideas from OLAP/timeseries databases, and search systems to create a unified system for operational analytics.

- The new Druid Kafka Indexing Service indexes the streaming data in a Kafka topic into a Druid cube.

- The Indexing Service can be managed by Hive as an external table providing SQL interfaces to the Druid cube.



HORTONWORKS

Layer on OLAP Analytics on top of Microservices architecture

Analytics on streaming data in driver-violation-events topic

How many violations in the last hour?

Which Drivers have the most violations in the last 30 minutes?

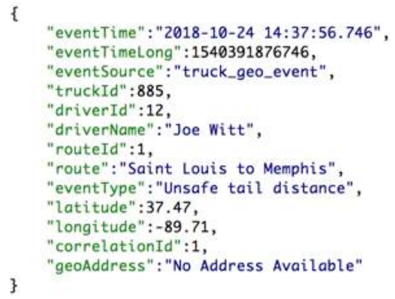
Rollup violations by driver, violation type and the route?

Are there specific routes that tend to cause more violations/incidents than others?

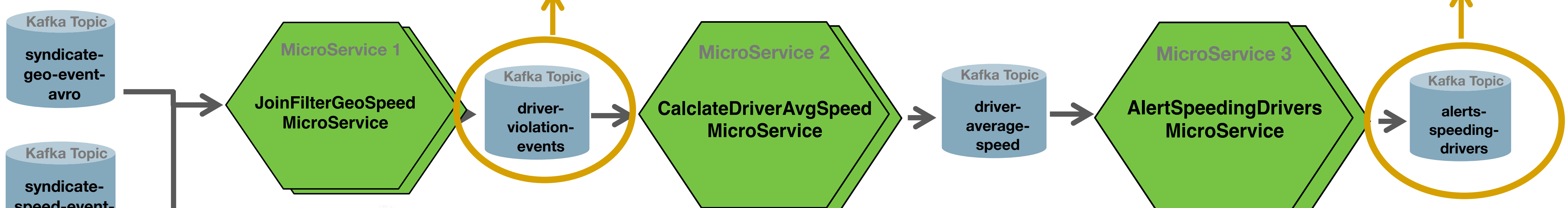
Analytics on streaming data in driver-violation-events topic

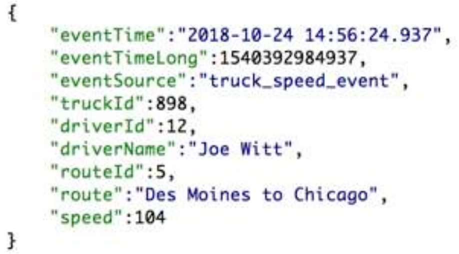
Which drivers have the most speeding alerts in last 2 hours?

How many speeding alerts in the last 5 minutes?

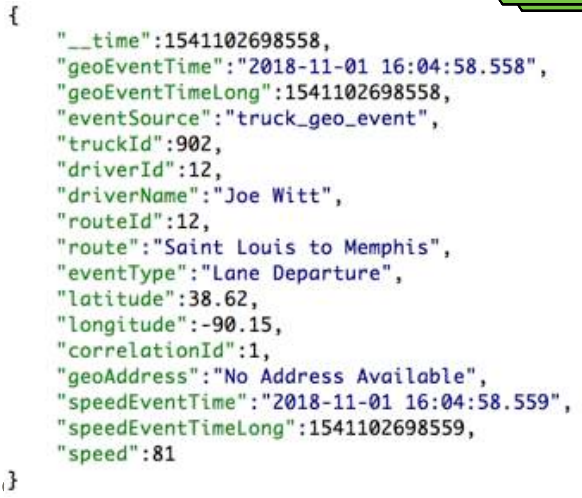


```
{
    "eventTime":"2018-10-24 14:37:56.746",
    "eventTimeLong":1540391876746,
    "eventSource":"truck_geo_event",
    "truckId":885,
    "driverId":12,
    "driverName":"Joe Witt",
    "routeId":1,
    "route":"Saint Louis to Memphis",
    "eventType":"Unsafe tail distance",
    "latitude":37.47,
    "longitude":-89.71,
    "correlationId":1,
    "geoAddress":"No Address Available"
}
```

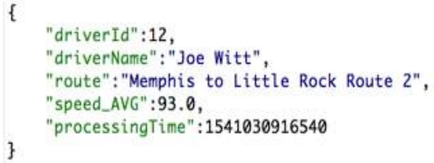




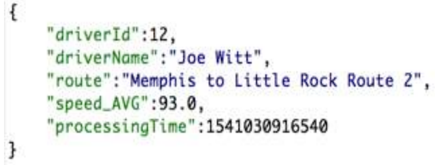
```
{
    "eventTime":"2018-10-24 14:56:24.937",
    "eventTimeLong":1540392984937,
    "eventSource":"truck_speed_event",
    "truckId":898,
    "driverId":12,
    "driverName":"Joe Witt",
    "routeId":5,
    "route":"Des Moines to Chicago",
    "speed":104
}
```



```
{
    "__time":1541102698558,
    "geoEventTime":"2018-11-01 16:04:58.558",
    "geoEventTimeLong":1541102698558,
    "eventSource":"truck_geo_event",
    "truckId":902,
    "driverId":12,
    "driverName":"Joe Witt",
    "routeId":12,
    "route":"Saint Louis to Memphis",
    "eventType":"Lane Departure",
    "latitude":38.62,
    "longitude":-90.15,
    "correlationId":1,
    "geoAddress":"No Address Available",
    "speedEventTime":"2018-11-01 16:04:58.559",
    "speedEventTimeLong":1541102698559,
    "speed":81
}
```



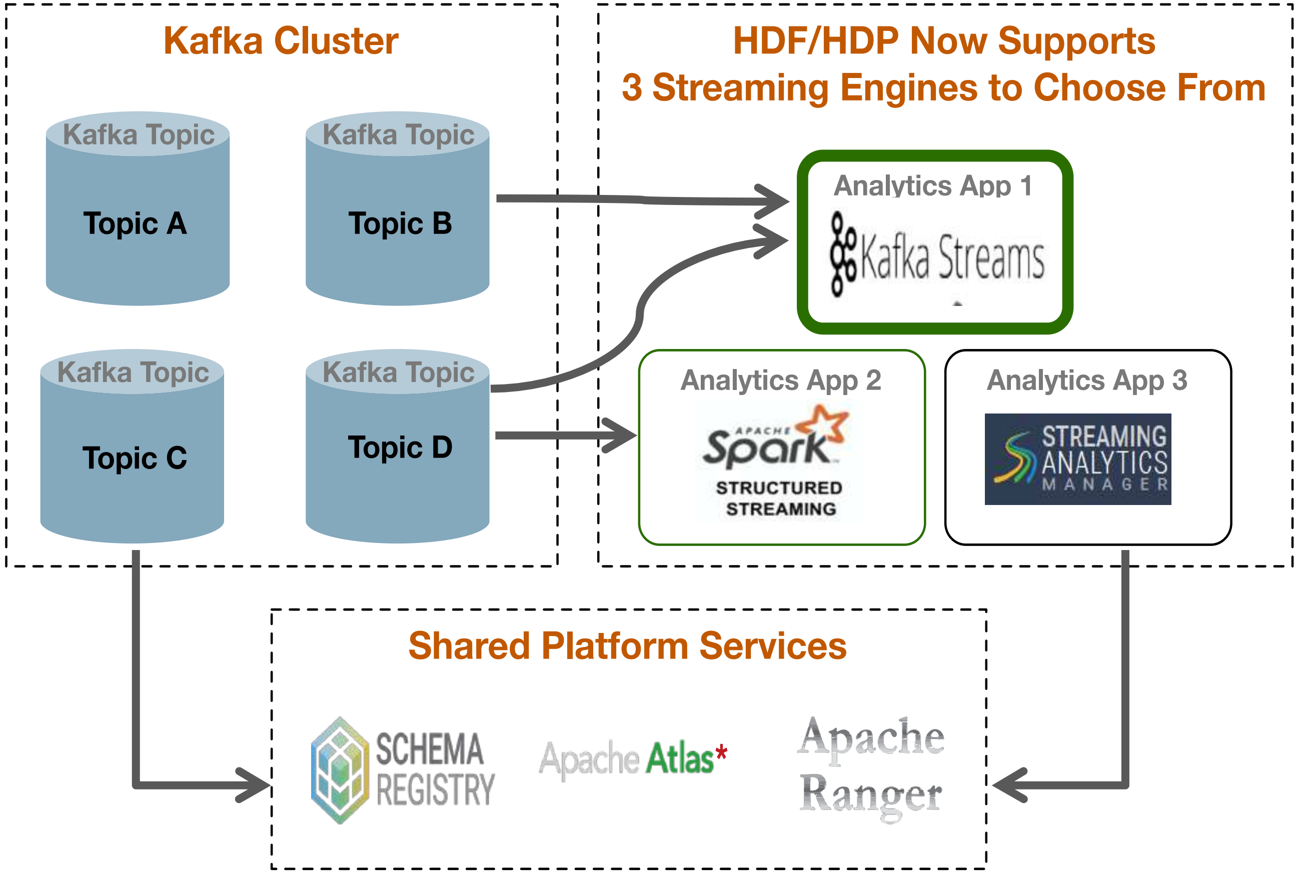
```
{
    "driverId":12,
    "driverName":"Joe Witt",
    "route":"Memphis to Little Rock Route 2",
    "speed_AVG":93.0,
    "processingTime":1541030916540
}
```



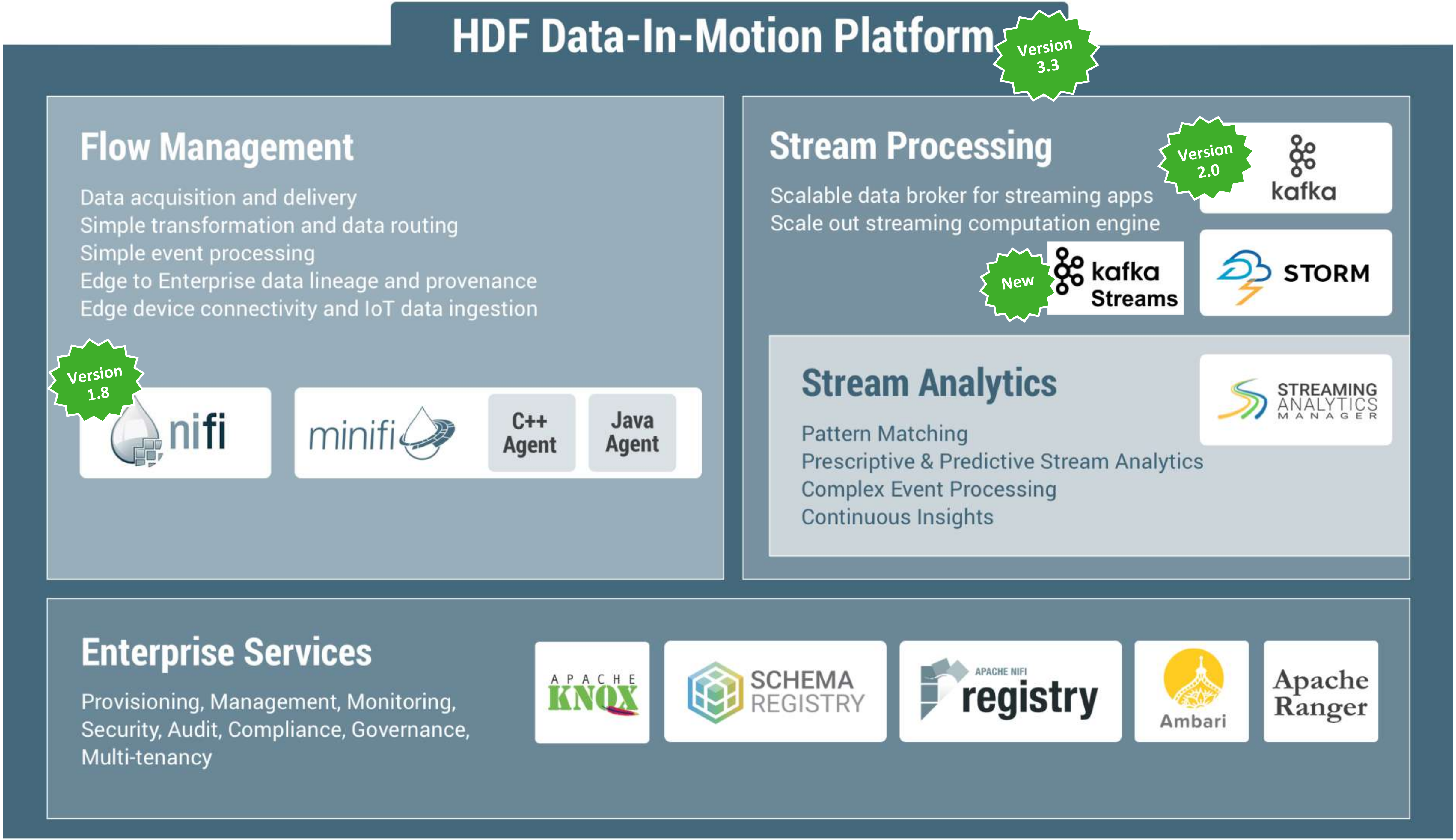
```
{
    "driverId":12,
    "driverName":"Joe Witt",
    "route":"Memphis to Little Rock Route 2",
    "speed_AVG":93.0,
    "processingTime":1541030916540
}
```

Cross-Platform Enhancements

3 Streaming Engines with the same set of shared platform services



HDF Data-In-Motion Platform



Flow Management

Data acquisition and delivery
Simple transformation and data routing
Simple event processing
Edge to Enterprise data lineage and provenance
Edge device connectivity and IoT data ingestion

Stream Processing

Scalable data broker for streaming apps
Scale out streaming computation engine

Stream Analytics

Pattern Matching
Prescriptive & Predictive Stream Analytics
Complex Event Processing
Continuous Insights

Enterprise Services

Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance, Multi-tenancy

Key Differentiators

100% open source technology – Only vendor with this strategy; prevents vendor lock-in

260+ pre-built processors – Only product to offer such comprehensive connectivity from edge to enterprise

3 Stream processing engines – Only vendor to offer a choice of three stream processing engines to customers for all their streaming architecture needs

Built-in data provenance – Only product in the market to offer out-of-the-box data provenance on data-in-motion

Comprehensive streaming platform – Only big data vendor to offer a comprehensive streaming platform from real-time data ingestion, transformation, routing to descriptive, prescriptive and predictive analytics.

Where can I find more information?

- HDF product page – www.hortonworks.com/hdf
- HDF 3.3 product documentation
- HDF 3.3 Release notes
- Blog posts –
  - What’s new in HDF 3.3?
  - Democratizing Analytics within Kafka with Three New Access Patterns
  - Kafka Streams – Is it the right Stream Processing engine for you?
  - Building Secure and Governed Microservices with Kafka Streams
- More to come
  - Monitoring Kafka Streams Microservices with Streams Messaging Manager
  - Real SQL on Real-time Streams in Kafka: Introducing the new Kafka Hive Integration

Happy Holidays!
Happy New Year!!

HORTONWORKS®