Internet of Things and Services

Service-oriented architectures

IoT and Big Data Systems

Department of Computer Science
Faculty of Electronic Engineering, University of Nis

Big Data

"Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population." - Teradata Magazine article, 2011

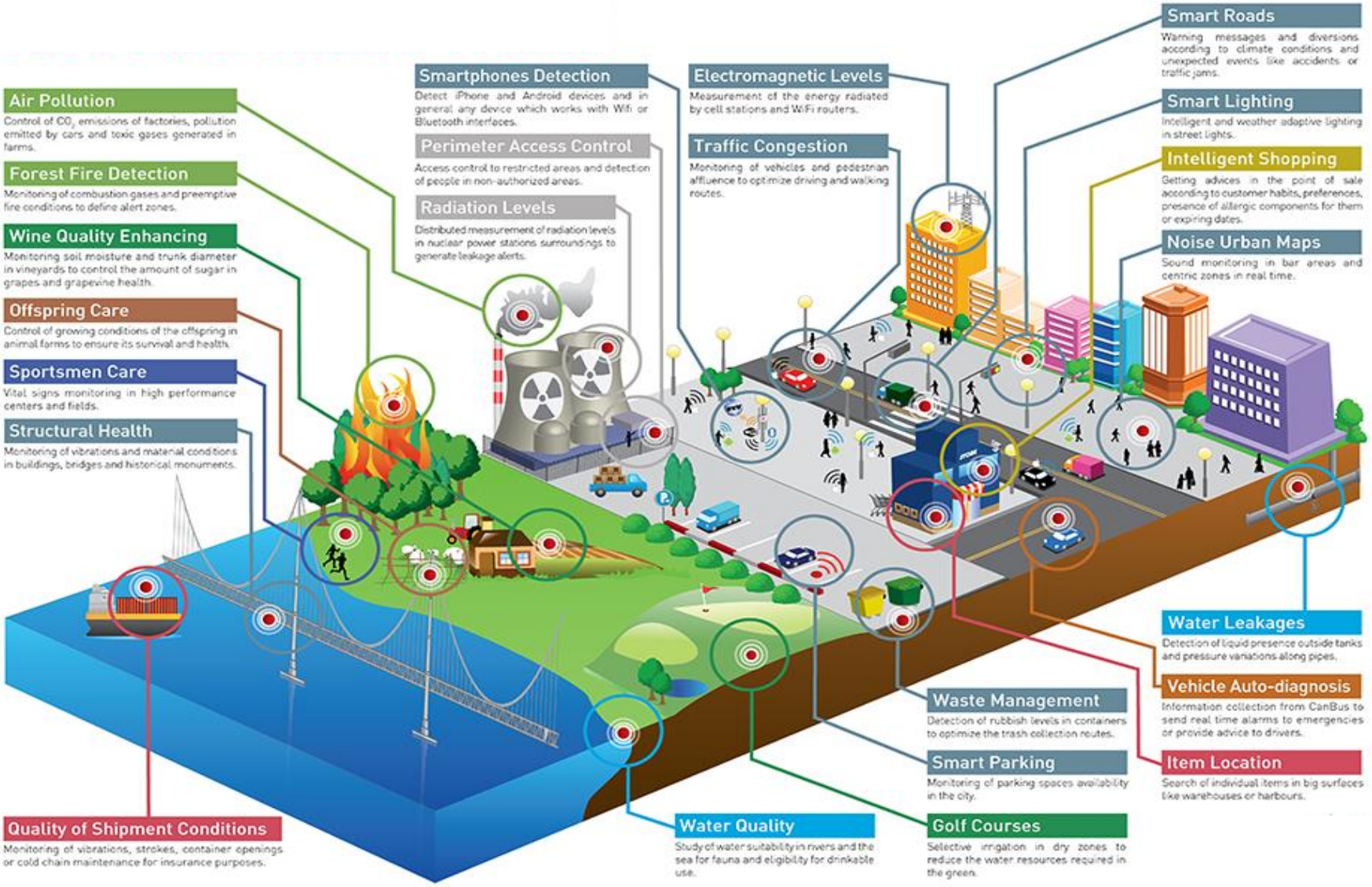
"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." - The McKinsey Global Institute, 2012

"Big data is a field that treats of ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software." - Wikipedia, 2019

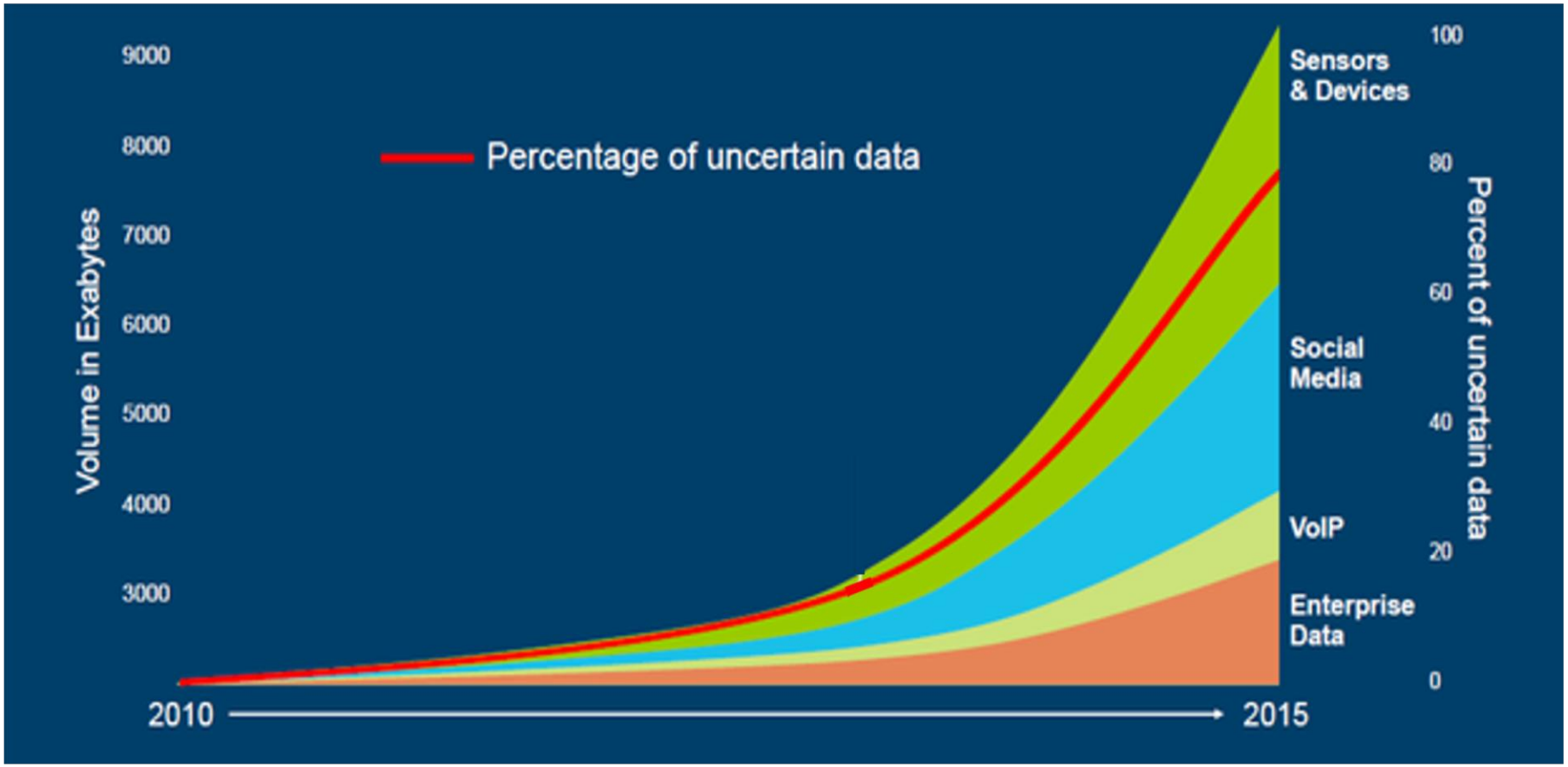
Big Data - Numbers

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006 (1EB = 10^{18}B)
  - 4.5 Zettabytes, 2013 (1ZB = 10^{21}B)
  - 44 Zettabytes by 2020
  - 163 Zettabytes by 2025
- How many data in a day?
  - 2.5 Exabytes
  - 8 TB, Twitter
  - 50 TB, Facebook
- 90% of world's data generated over last two years!

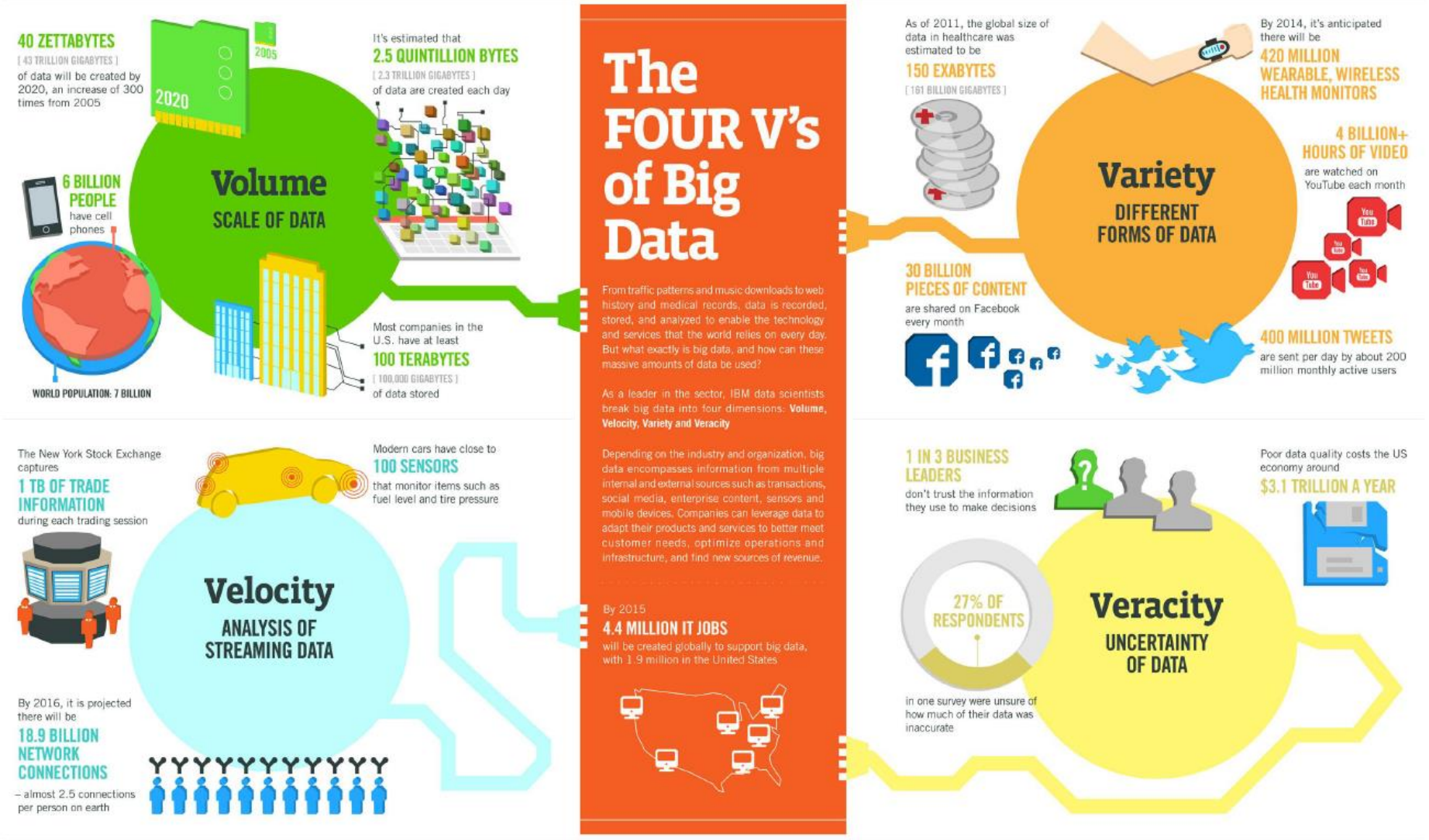
Big Data sources - Smart world



Big Data sources



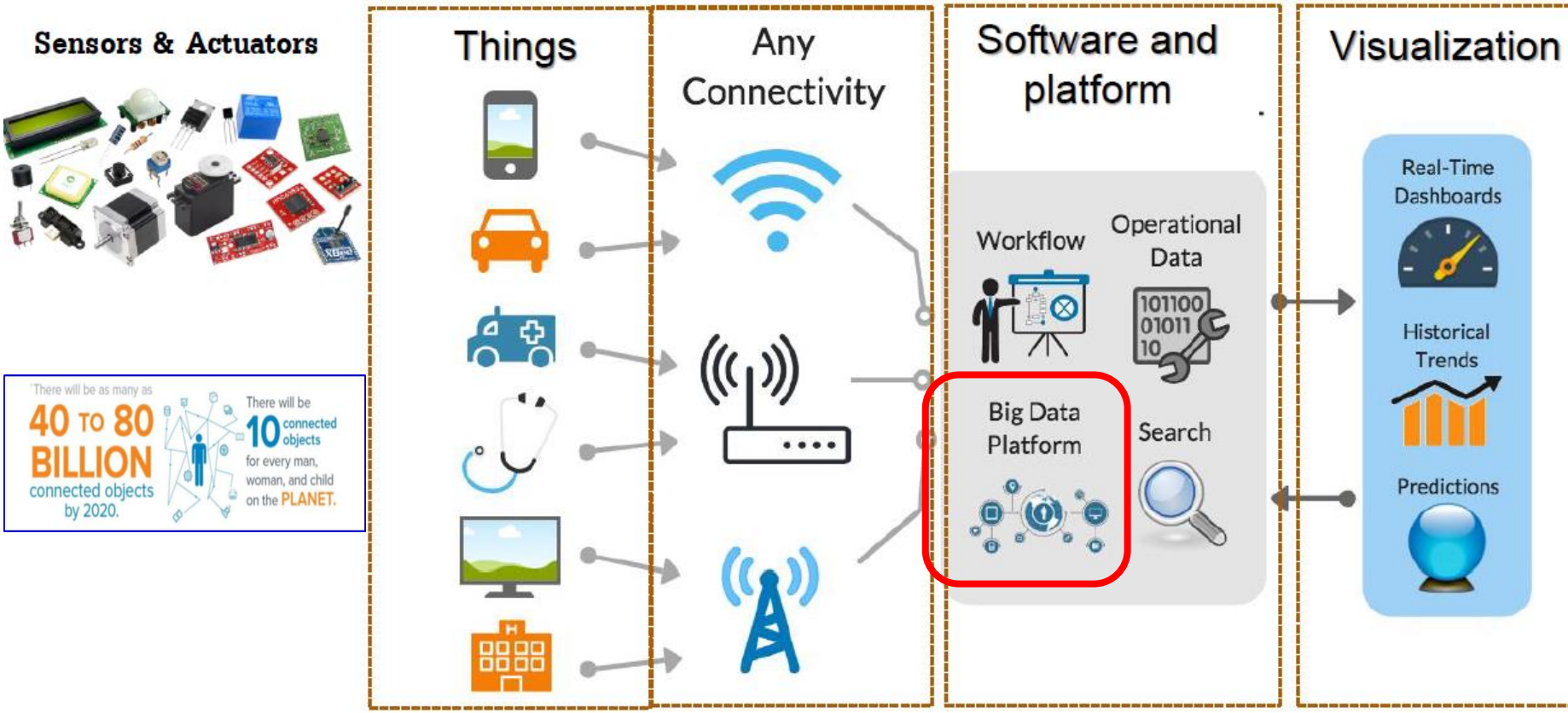
The four "V's" of Big Data



Big Data: 4V + VALUE

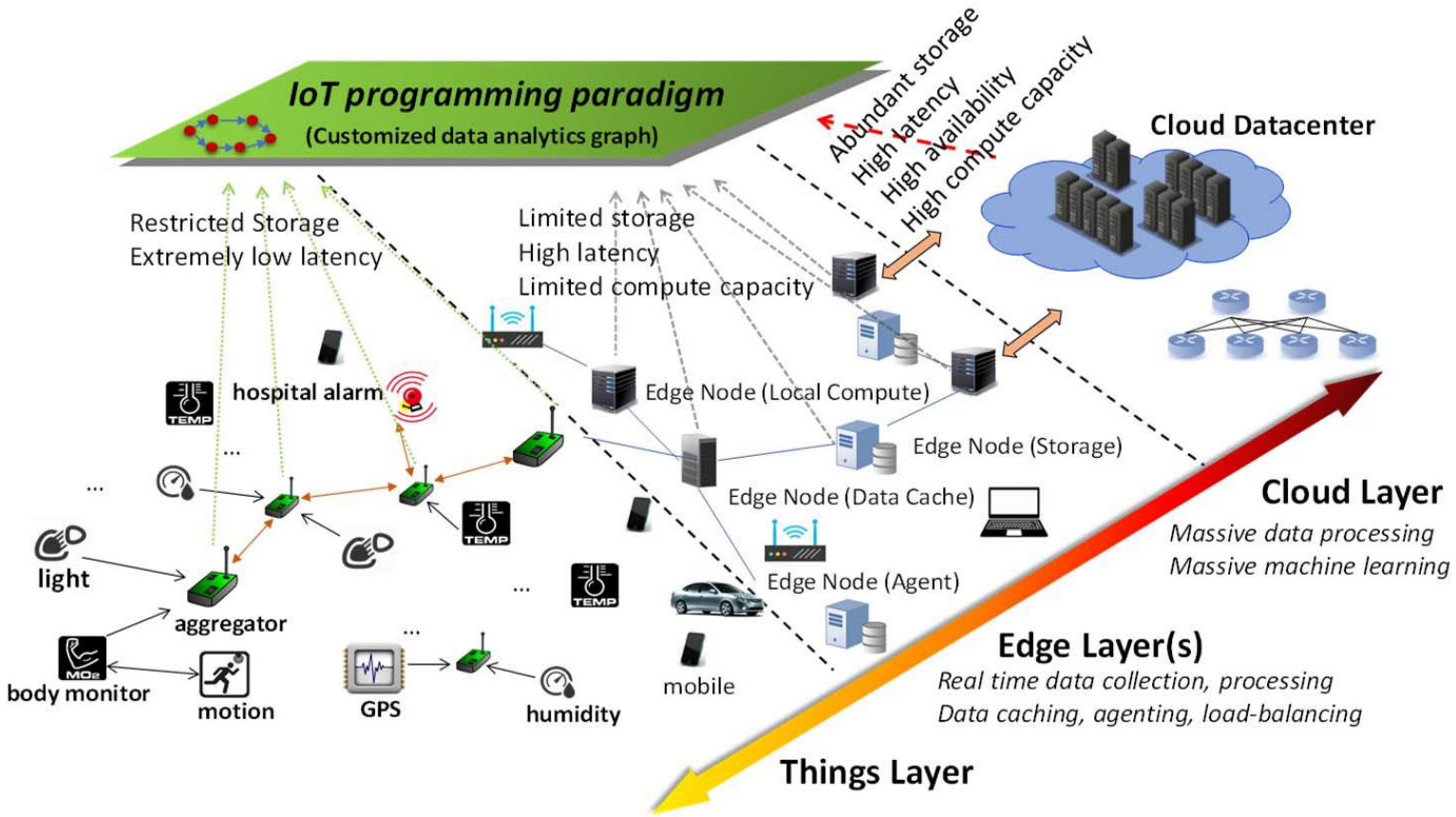
- Volume
  - Terabyte(10^{12}), Petabyte(10^{15}), Exabyte(10^{18}), Zettabyte (10^{21})
- Variety
  - Structured, semi-structured, unstructured;
  - Text, image, audio, video, record
- Velocity
  - Periodic, Near Real Time, Real Time
- Veracity
  - Quality of the data can vary greatly
- Value
  - Big Data can generate huge competitive advantages

IoT Big Data architecture



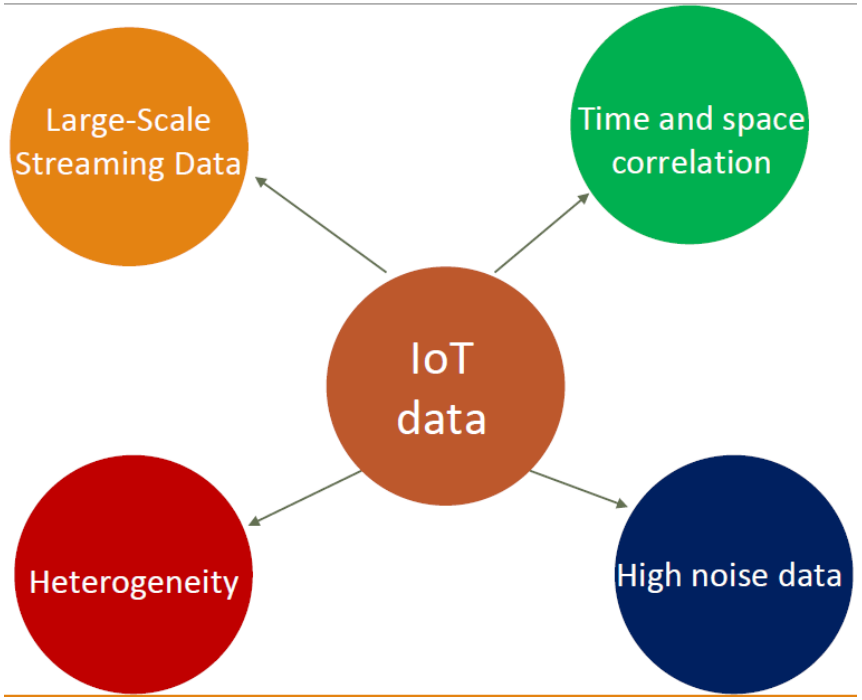
IoT & Smart health

- A typical IoT application infrastructure of a healthcare use case showing of Things, Edge, and Cloud layers.

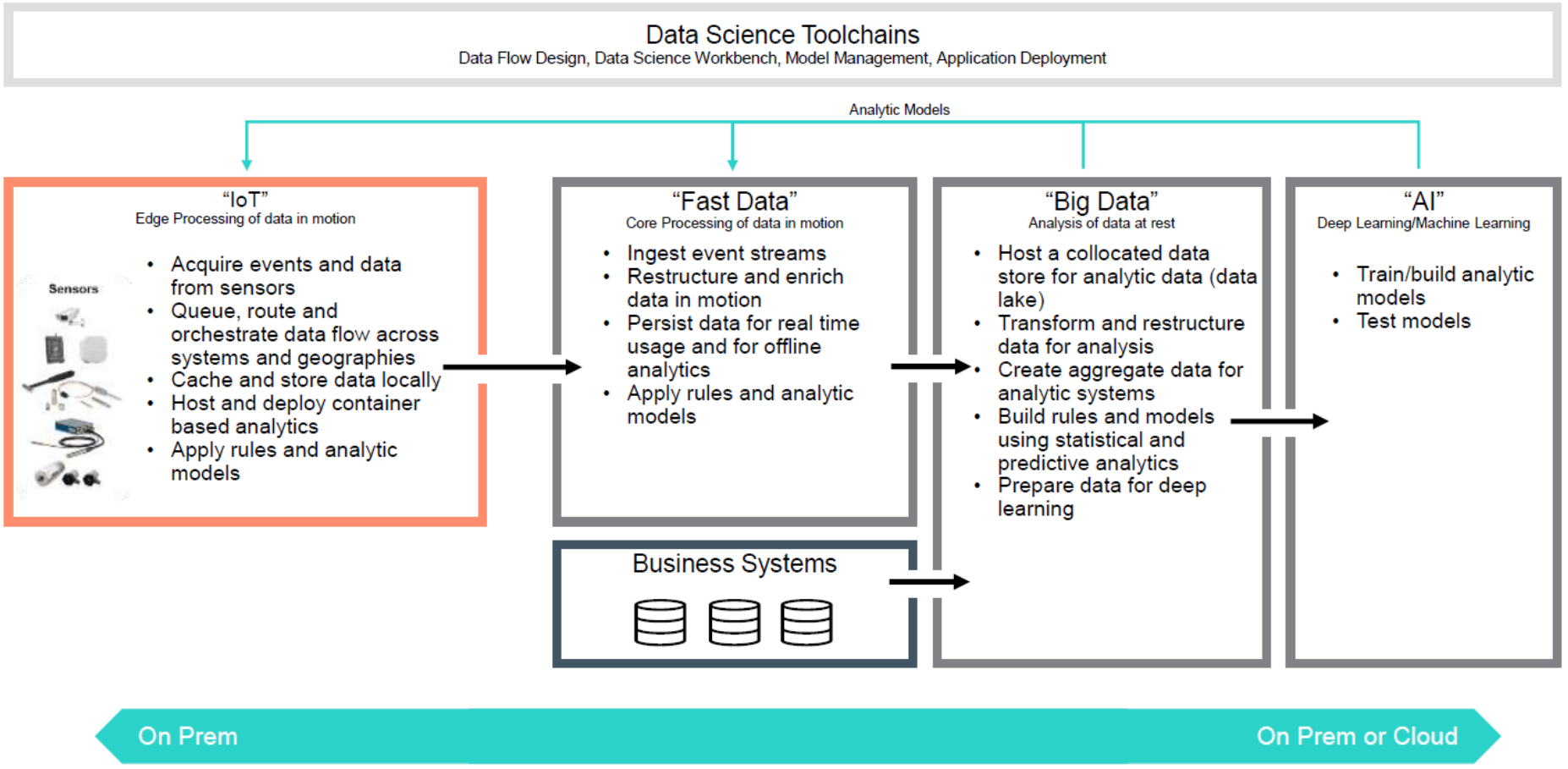


IoT Big Data characteristics

- Analytics requirement
  - Fast computing and advanced machine learning techniques require for IoT streaming data processing and IoT Big Data analytics
- IoT Applications support
  - High-speed data streams and requiring real-time or near real-time actions

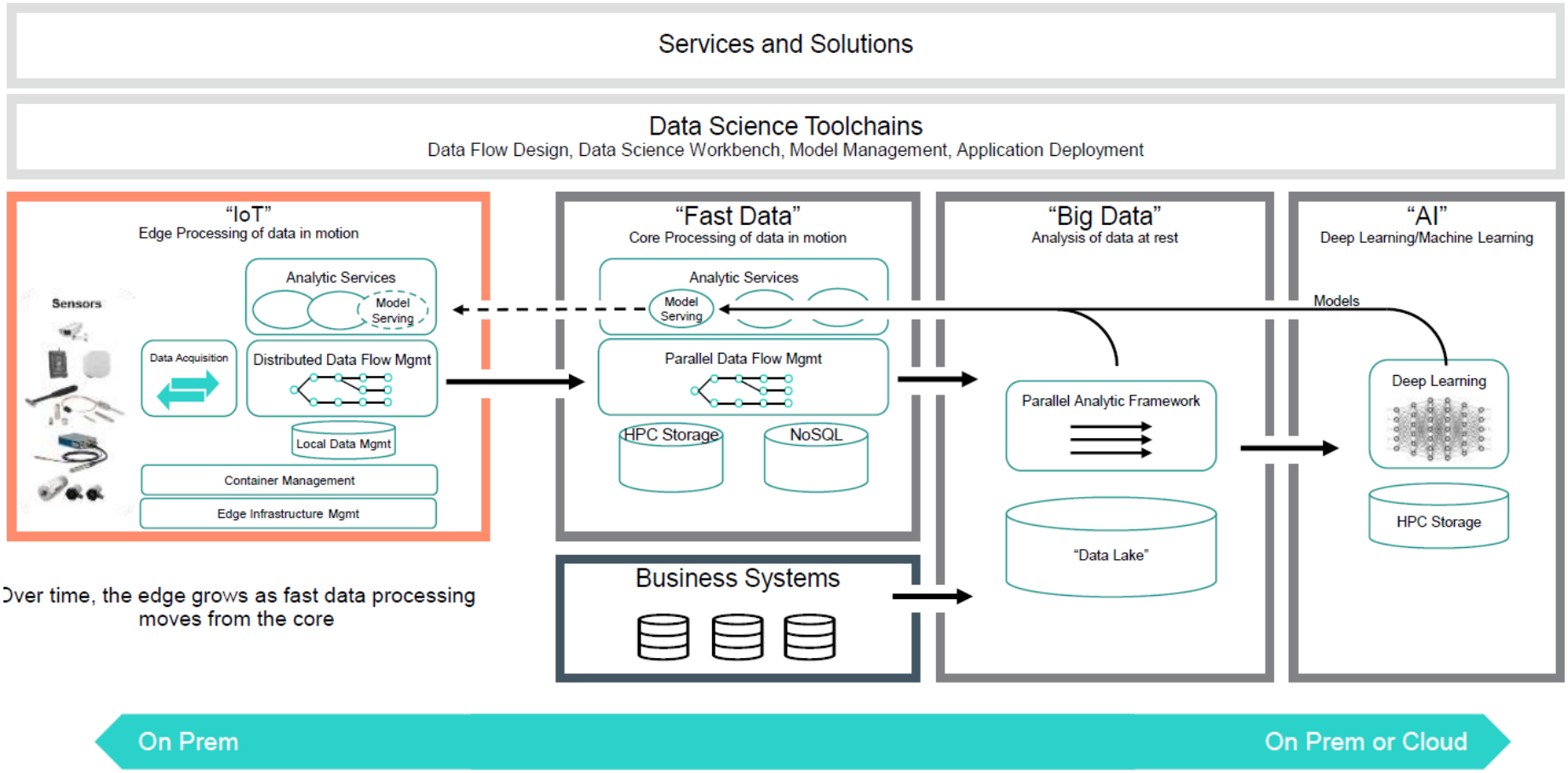


IoT Big Data flow (pipeline)

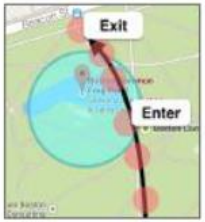


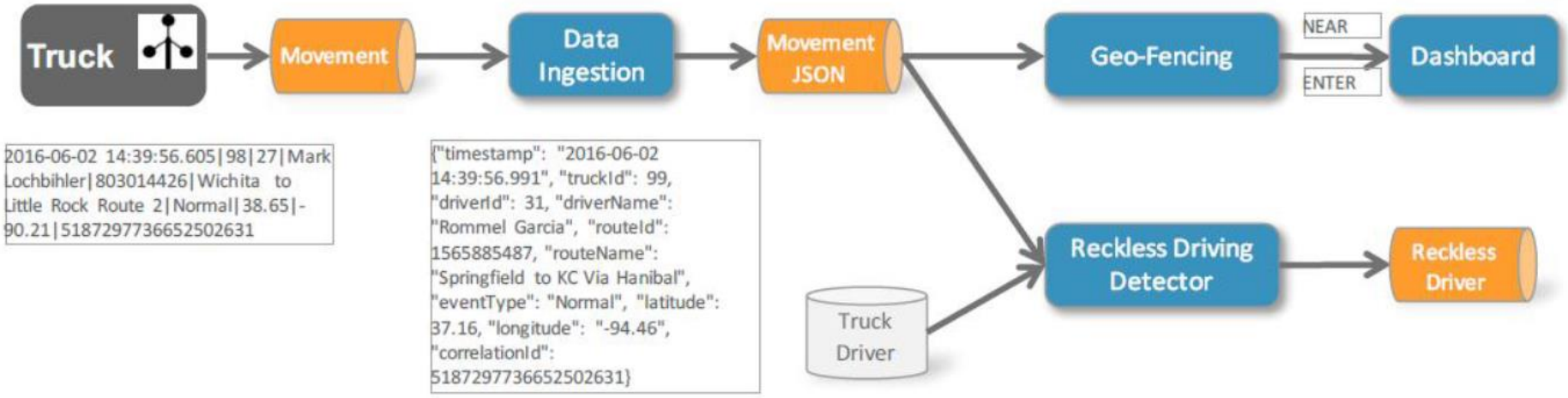
IoT Big Data flow (pipeline)

- Functional view



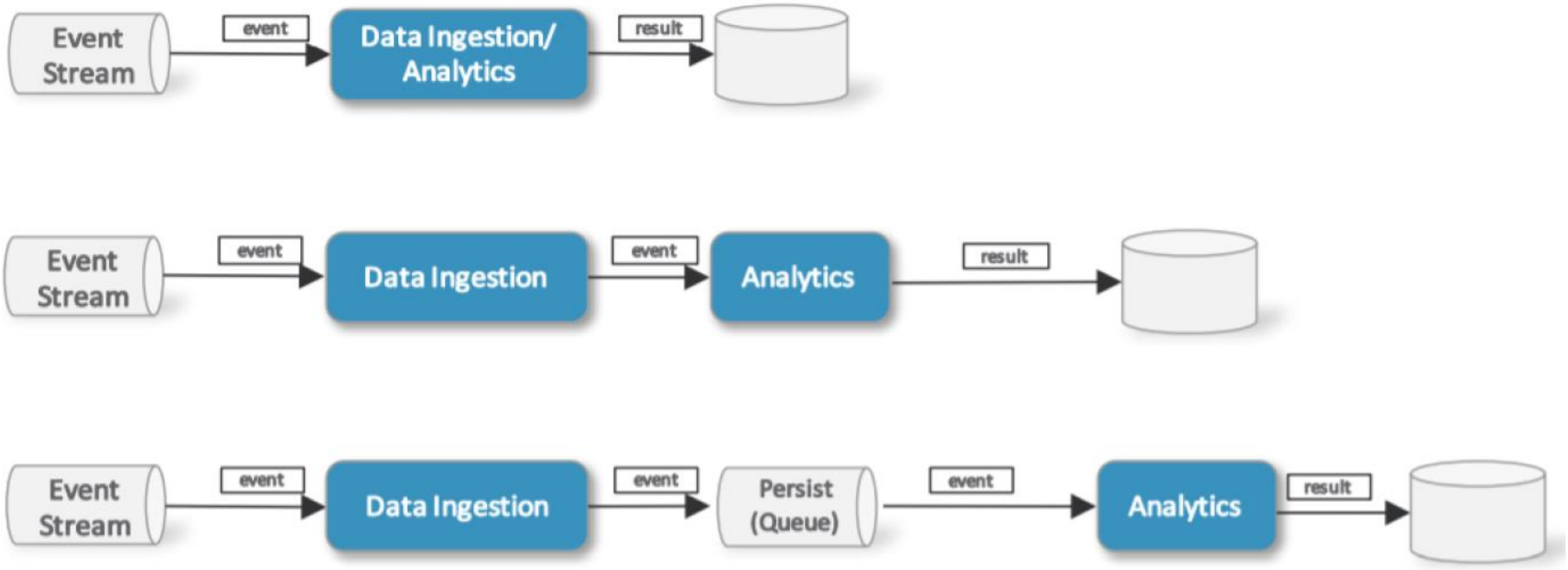
Truck monitoring - Use case





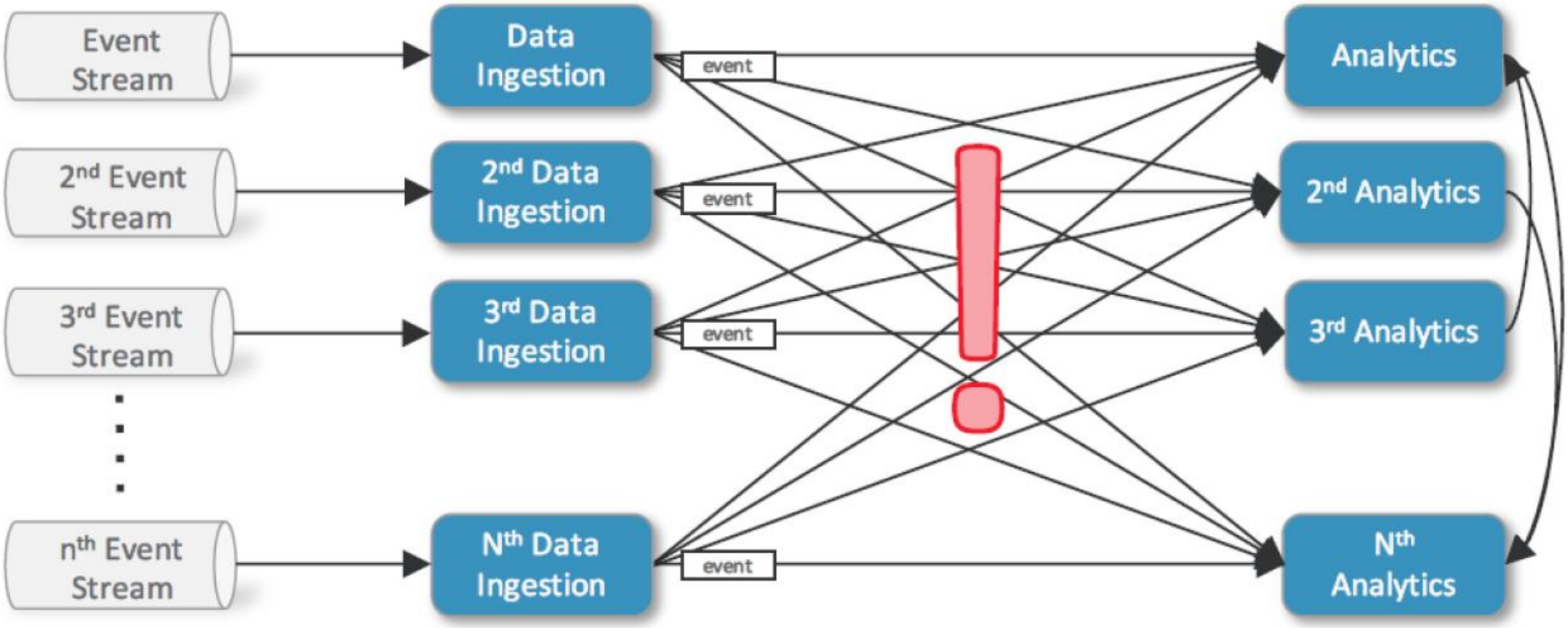
IoT streaming analytics

- How to design a Streaming Analytics Solution?



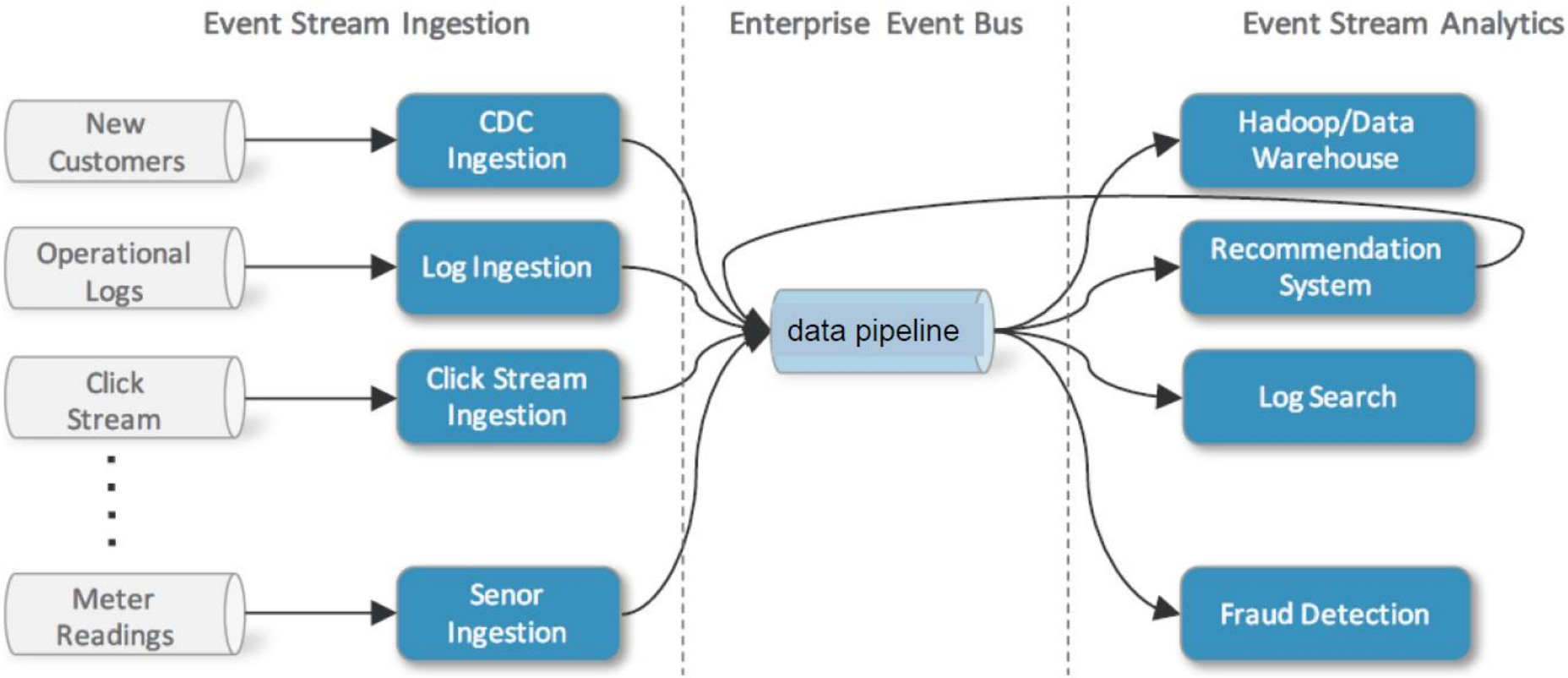
Stream sources and analytics

- New Event Stream sources are added and new Analytics are interested in the events



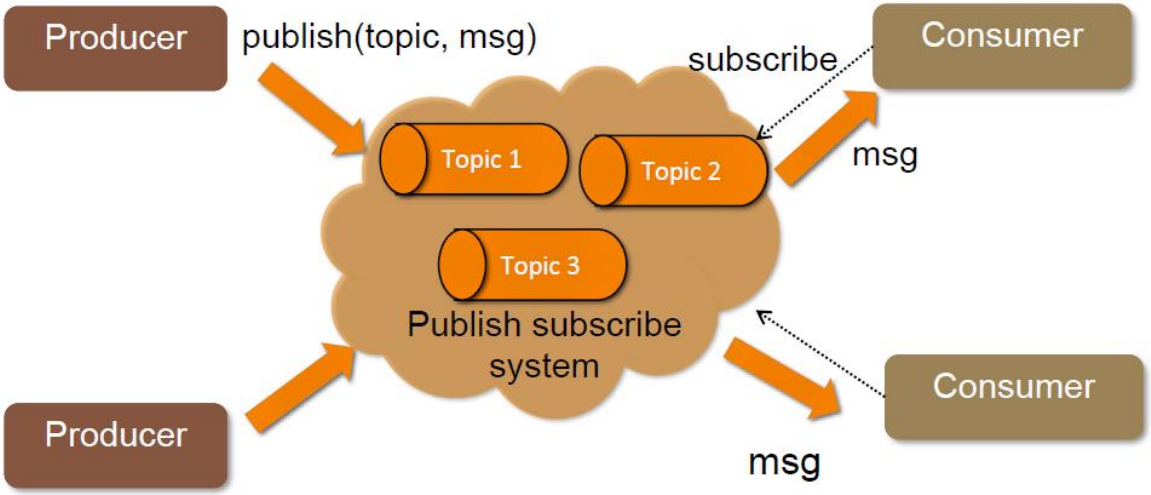
Stream sources and analytics

- Decouple event streams from consumers



Apache Kafka

- An open-source distributed streaming platform developed by LinkedIn and donated to the Apache Software Foundation.
  - https://kafka.apache.org/
- Aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds through message broker and publish/subscribe functionalities

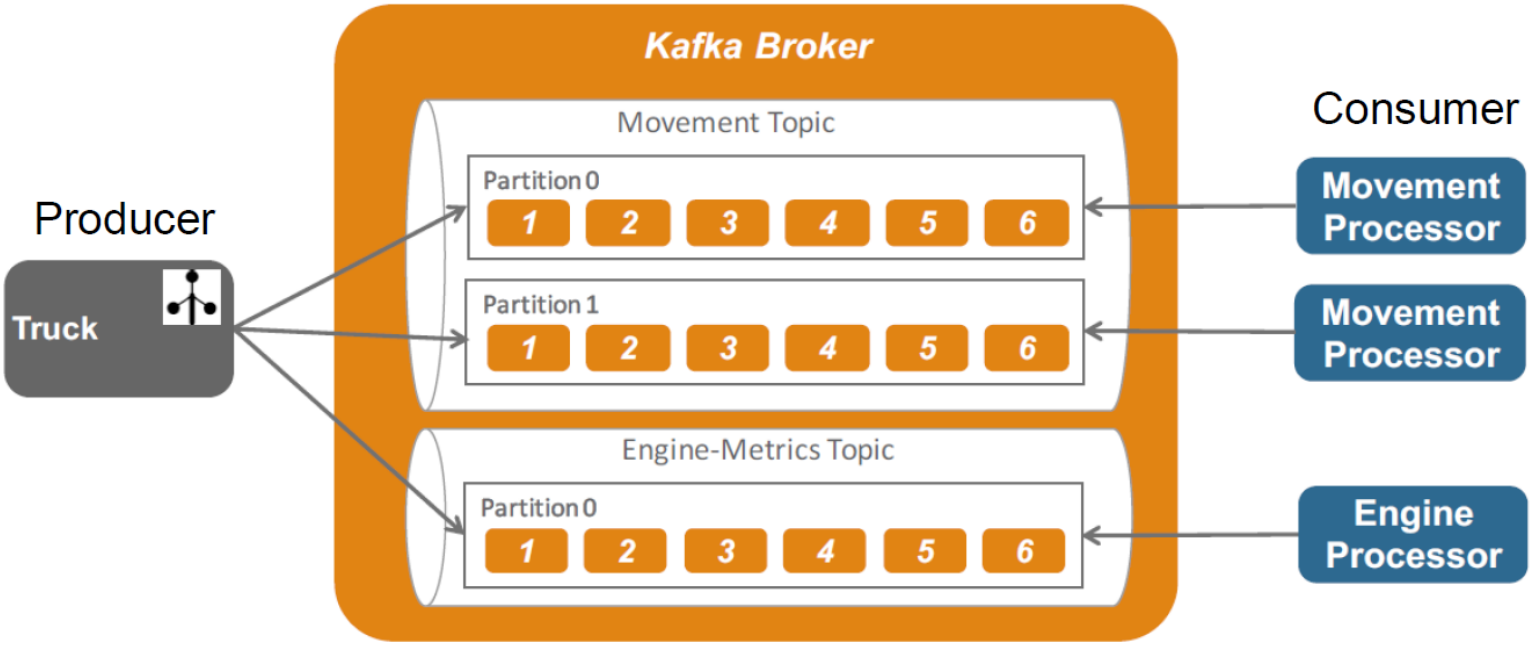


Kafka characteristics

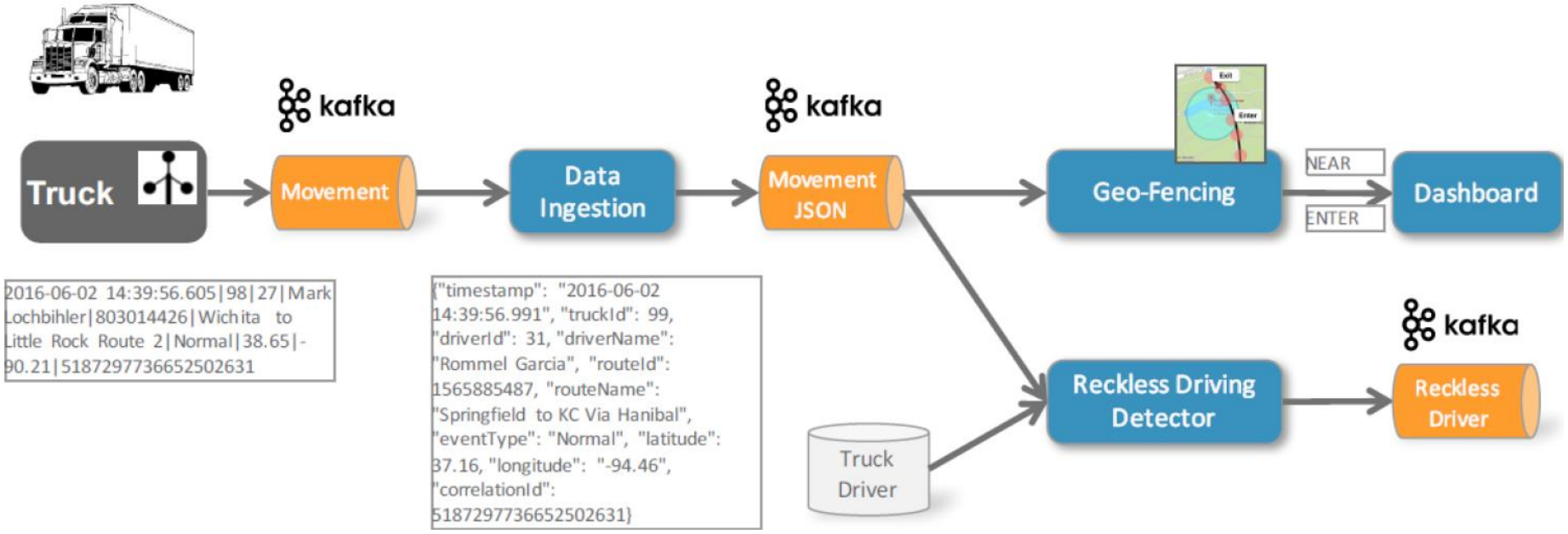
- Fast
  - Kafka can handle hundreds of megabytes of reads and writes per second from a large number of clients.
  - Designed for real time activity streaming.
- Distributed and highly scalable
  - Kafka has a cluster-centric design offers strong durability and fault-tolerance guarantees.
  - Messages partitioning spread over a cluster of machines
- Durable
  - Message persisted to disk and replicated within cluster to prevent data loss.
  - Each broker can handle terabytes of messages without performance impact

Kafka architecture

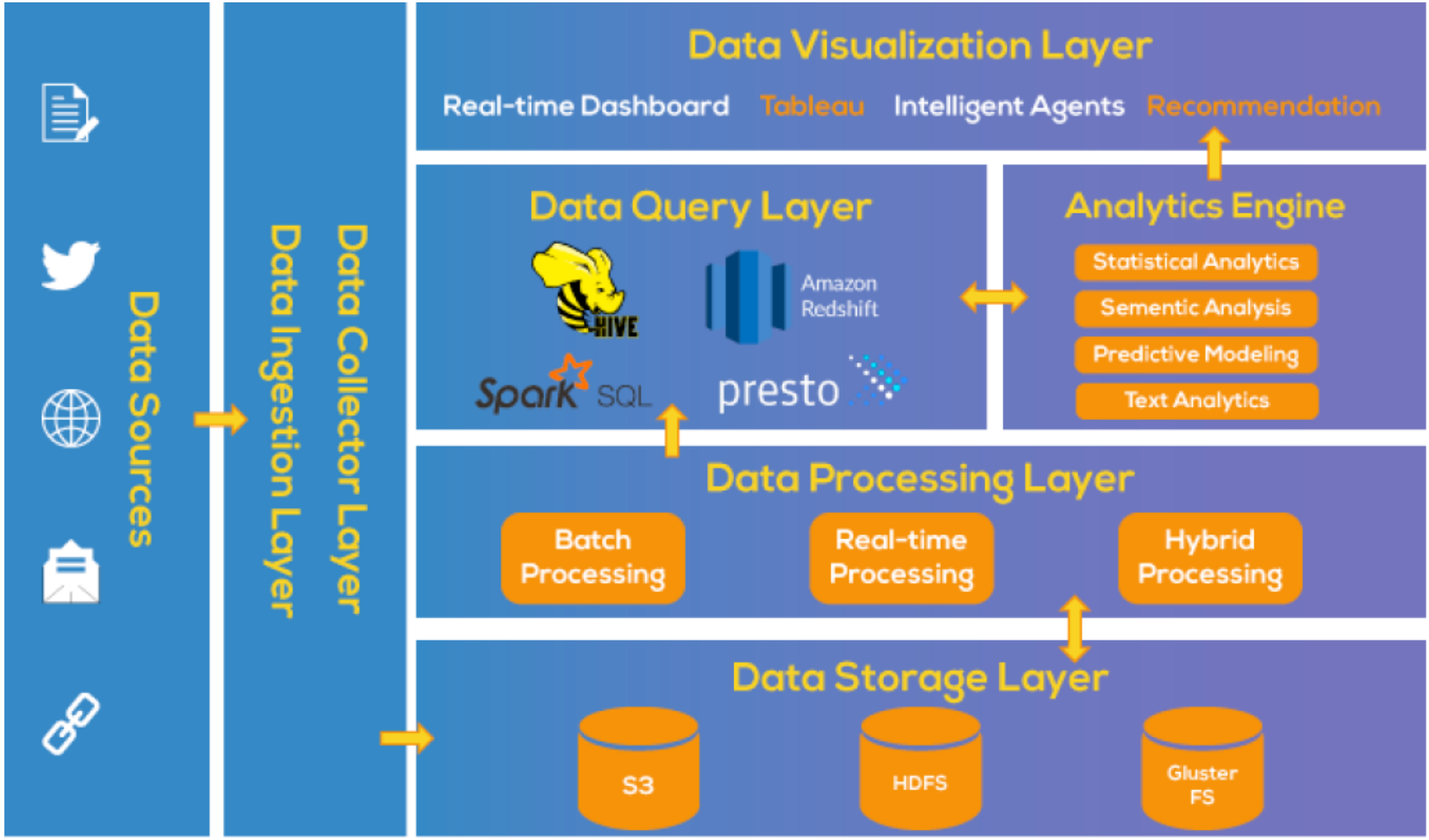
- Kafka Cluster is made up of multiple Kafka Brokers
- Producers and Consumers exchange messages via Topics
- Apache Zookeeper keeps track of status of the Kafka cluster nodes and it also keeps track of Kafka topics, partitions etc.



Truck monitoring – Use Case

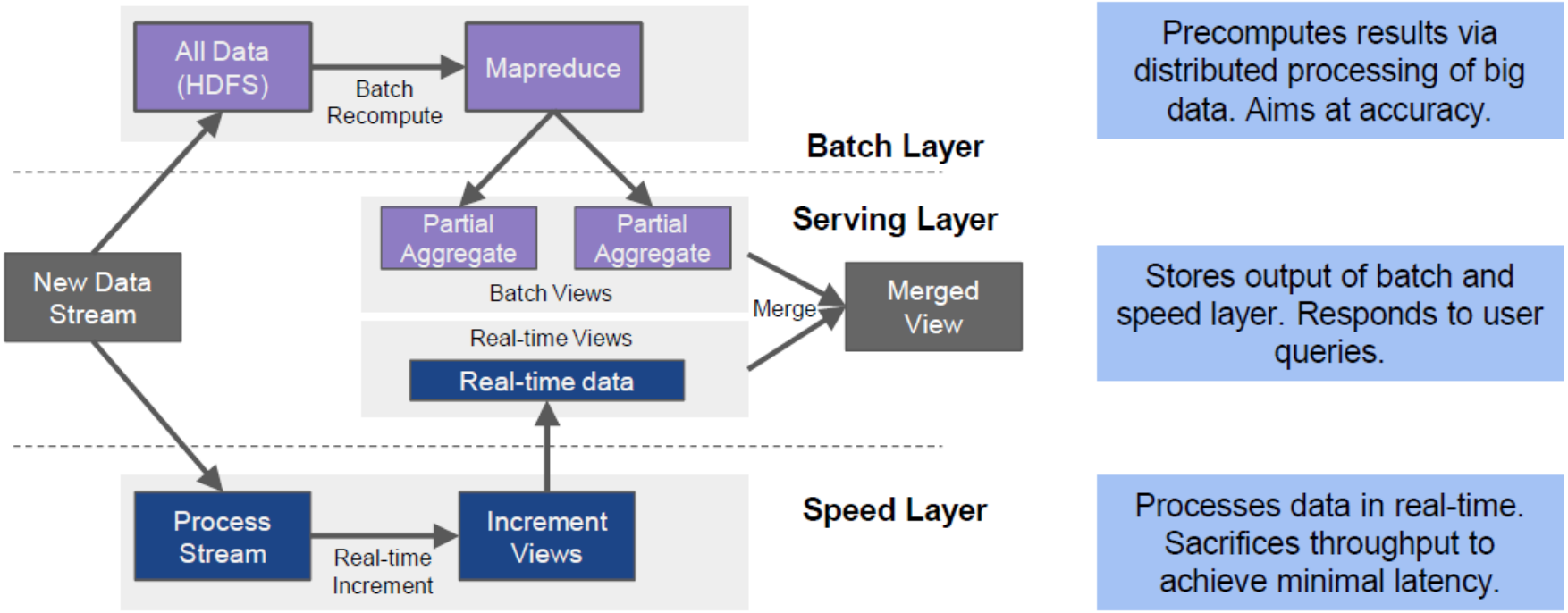


Big Data platform architecture



The Lambda architecture

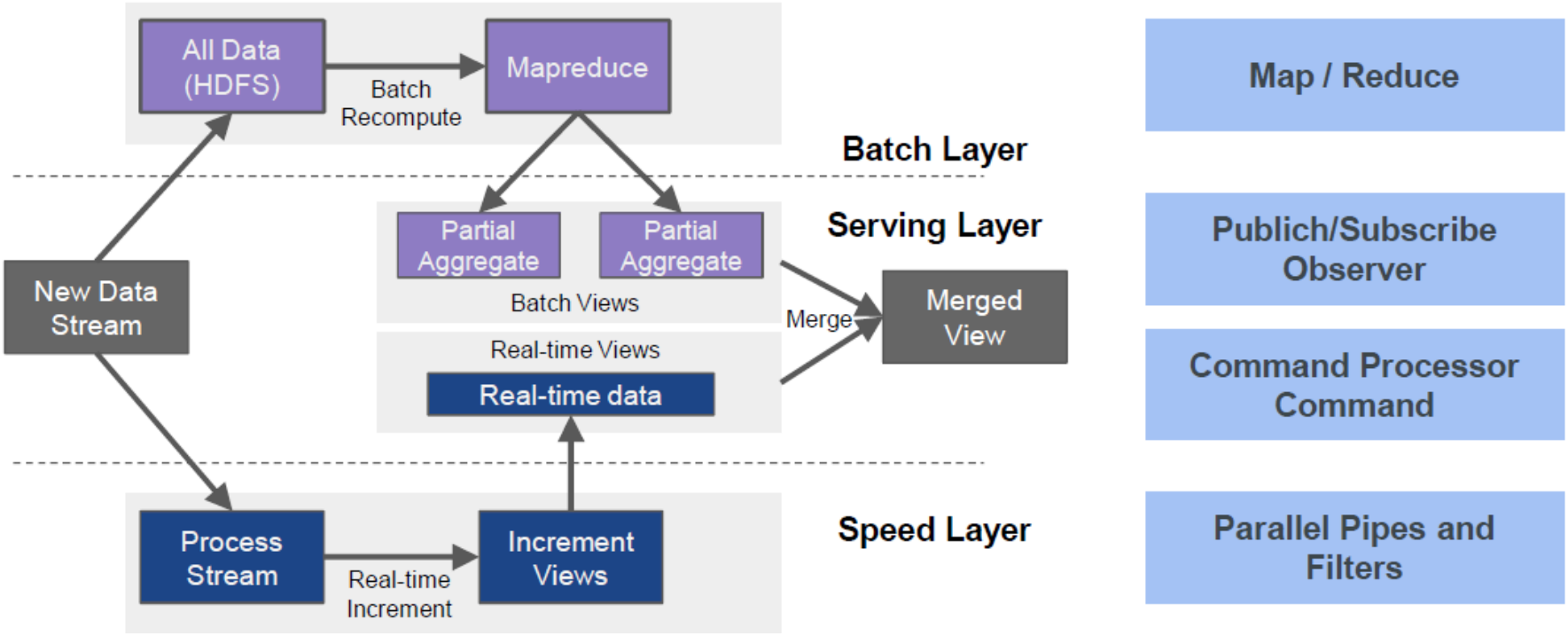
- The Lambda Architecture for Big Data processing and analytics applications



Originally designed by Nathan Marz for Twitter; now adopted by many companies

The Lambda architecture

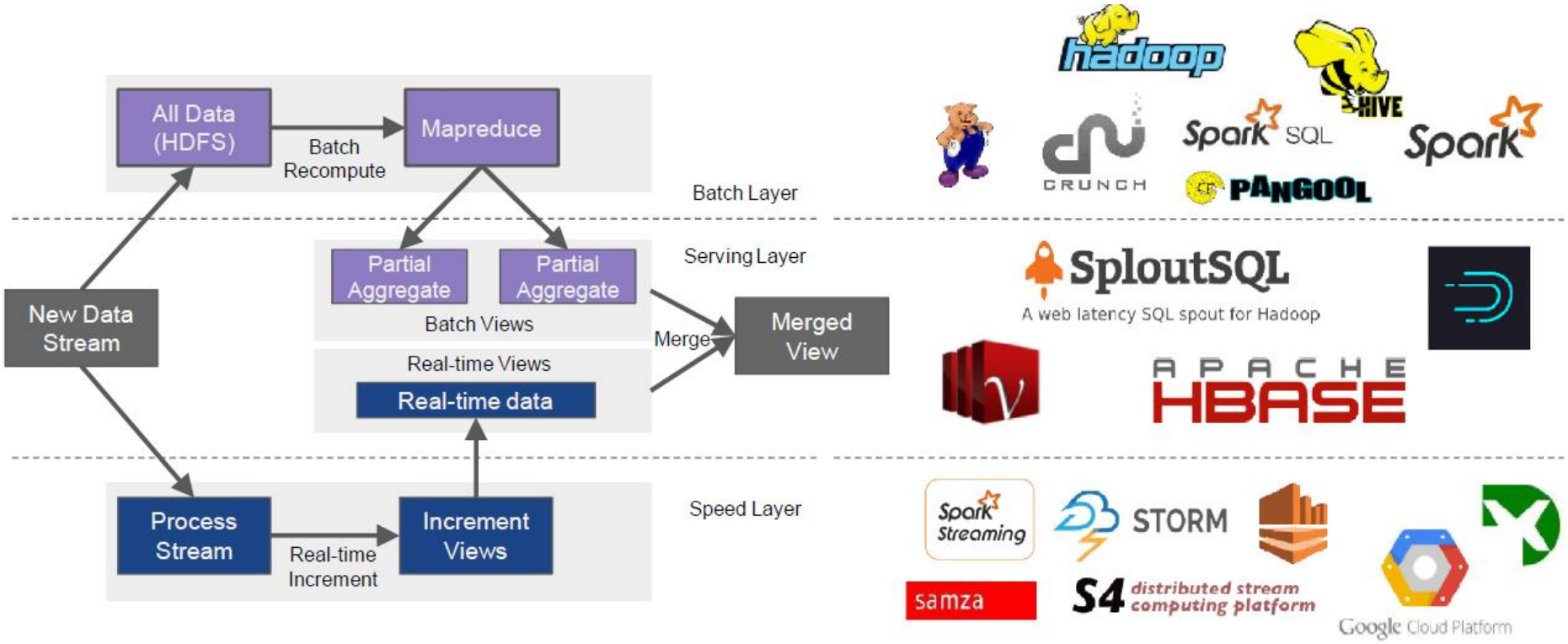
- Multiple architecture styles define the Lambda architecture



Originally designed by Nathan Marz for Twitter; now adopted by many companies

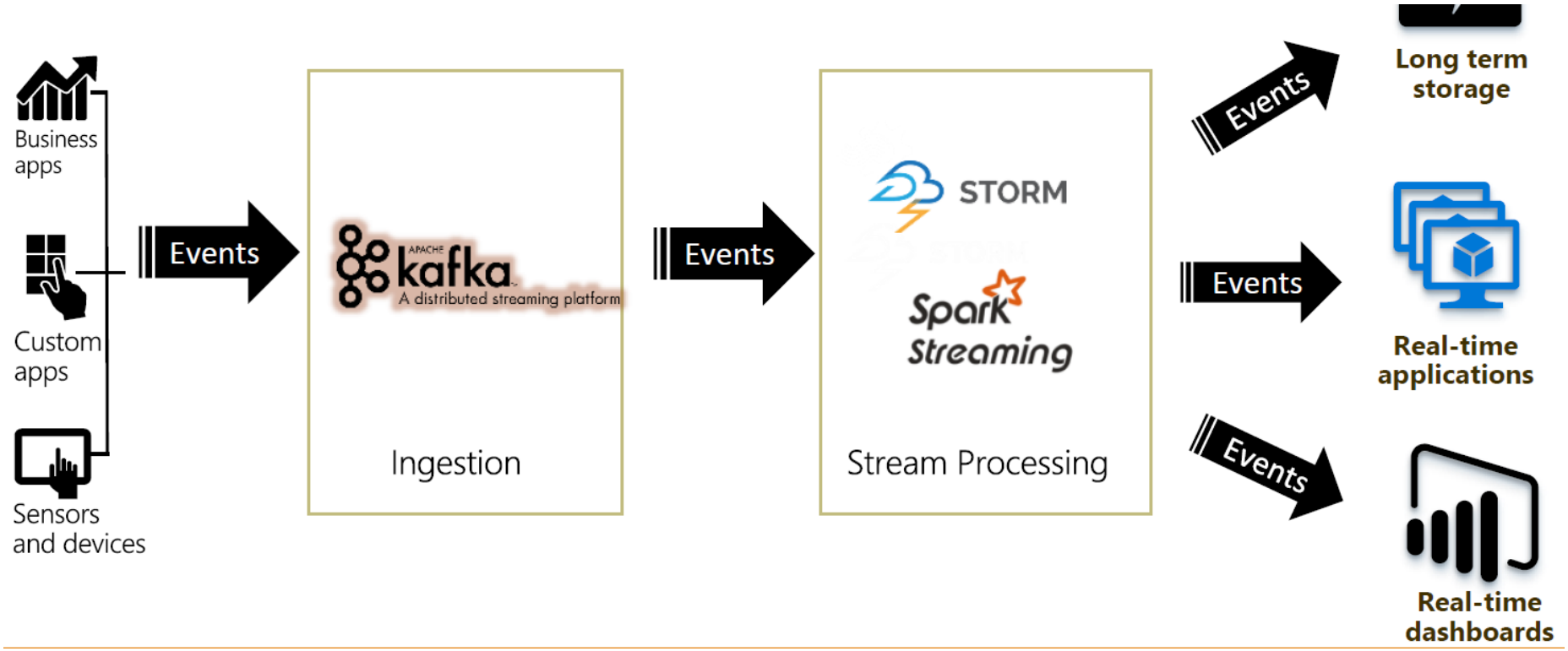
The Lambda architecture

- Many off-the-shelf and open source Big Data are technologies and frameworks available



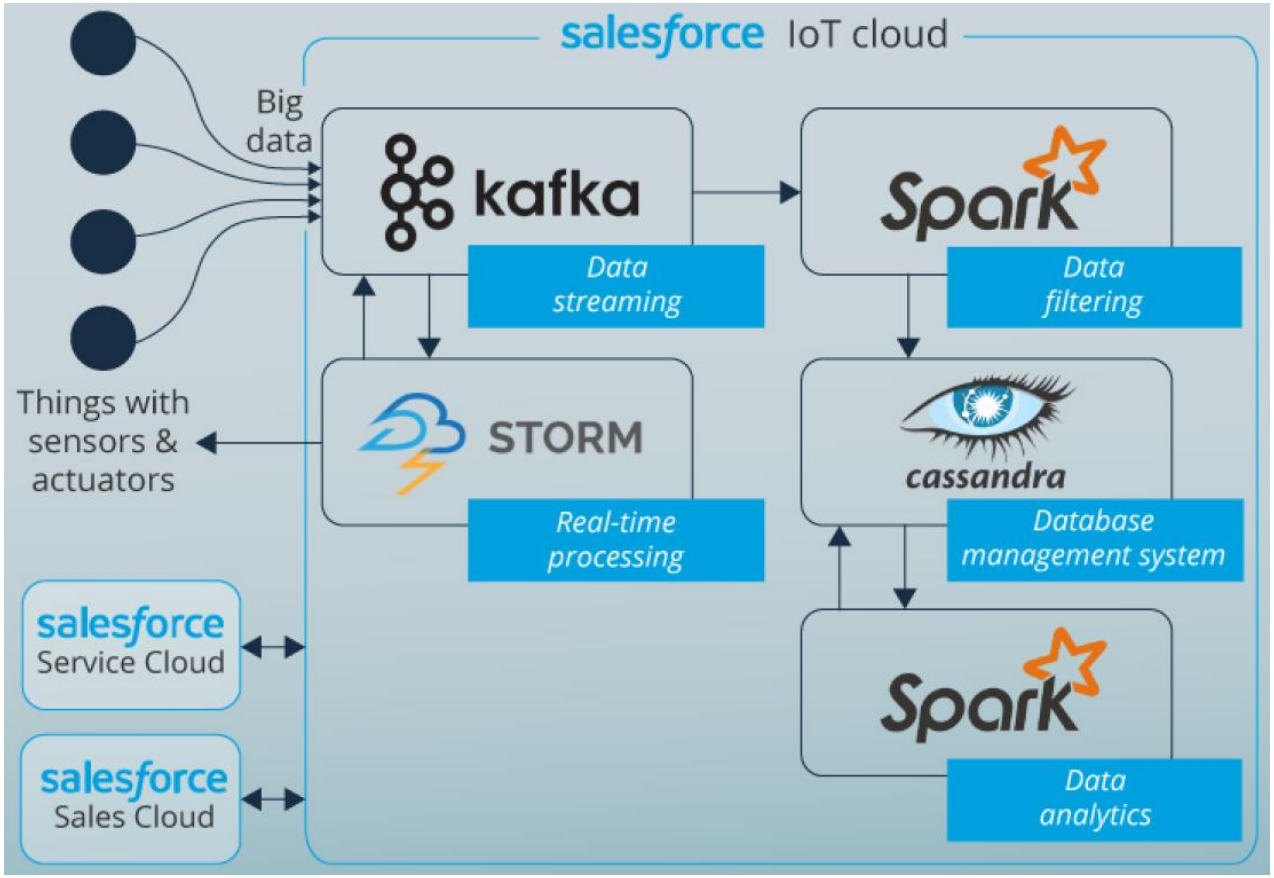
IoT Big Streaming Data

- Big Streaming Data flow and processing patterns



IoT Big Data - Example

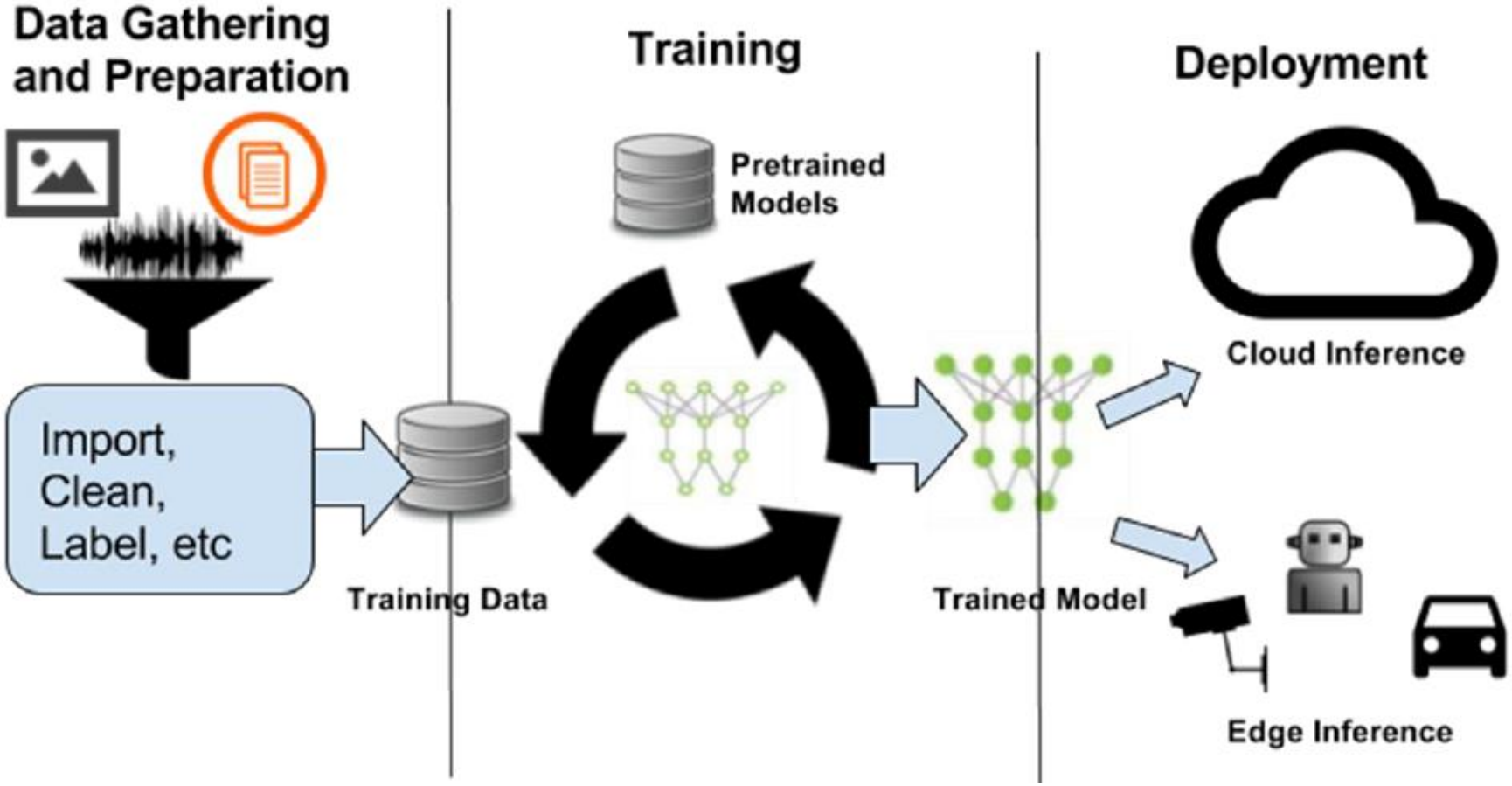
- Salesforce IoT Cloud: benefits and limitations



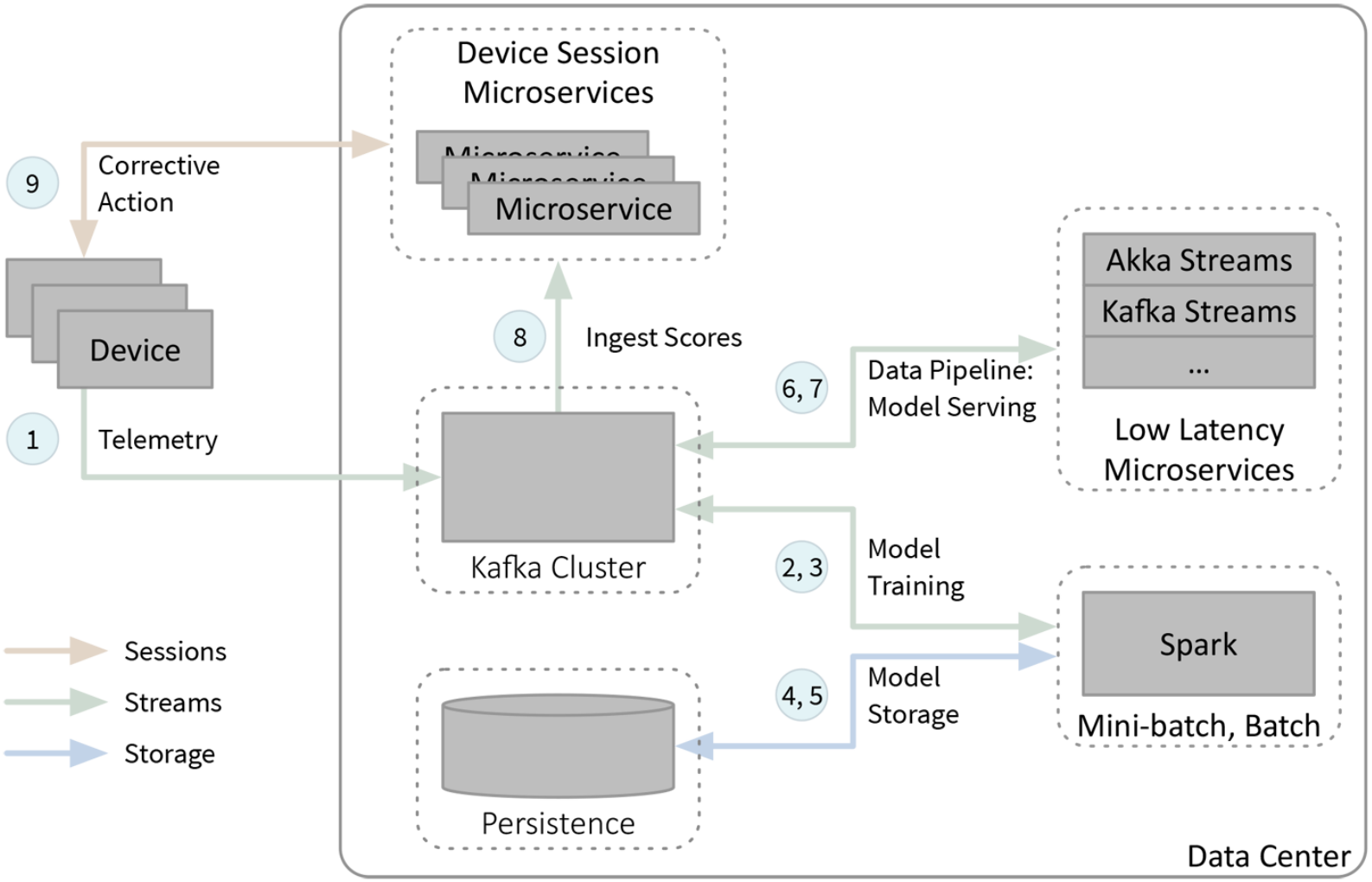
https://www.scnsoft.com/blog/salesforce-iot-cloud-benefits-and-limitations

IoT Big Data analytics

- Machine Learning in IoT Applications



IoT anomaly detection - Example



Dean Wampler, Fast Data Architectures for Streaming Applications, 2nd edition

Challenges and future directions

- Large and fast IoT datasets
  - More data is needed to achieve more accuracy
- Preprocessing
  - More complex since the system deals with data from different sources that may have various formats
- Secure and privacy preserving machine learning
  - Developing further techniques to defend and prevent the effect of this sort of attacks on models is necessary for reliable IoT applications.
- Machine learning for IoT devices
  - Consider the requirements of handling Machine learning in resource-constrained devices and Edge computer nodes.

References

- Perry Lea, ***IoT and Edge Computing for Architects,*** 2nd Edition, Packt Publishing, 2020
  - Chapter 12: Data Analytics and Machine Learning in the Cloud and in the Edge
- Dean Wampler, *Fast Data Architectures for Streaming Applications*, 2nd edition, O'Reilly Media, 2019.
- Gerard Maas, Stavros Kontopoulos, and Sean Glover, *Designing Fast Data Application Architectures*, O'Reilly Media, 2018.