Internet des Objets et Big Data
Introduction

Ons Jelassi
ons.jelassi@telecom-paritech.fr

Sommaire

- Introduction
- Technologies de l'IoT
- Complex Event Processing
- Collecte, traitement et gestion des données
- Planning du module

Definition

- **The Internet of Things (IoT)**
  - a network of physical objects that is connected to and accessed through the Internet. Connected objects contain embedded technology, such as sensors and actuator, that enable objects to sense and communicate. This ability is what will change how and where decisions are made, who makes them, and how quickly they are made.

- Traditional notions of “volume” and “velocity” will be redefined as the internet of things grows in scale and ambition.

IoT things

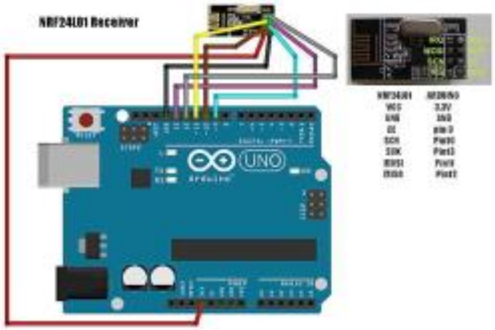
- The IoT is a complex network of things and people that are seamlessly connected through the Internet.
- Anything that can be connected will be connected and communicating using wireless sensors and RFID tags
- The IoT concept refers to identifiable things with their virtual representations in an Internet-like structure
- Once connected, these connected "things" can send data and interact with other things and people – all in real time.

IoT Technologies

- **The Internet of Things is much more than**
  - M2M communication
  - wireless sensor networks
  - 2G/3G/4G
  - RFID
  - Cloud computing capabilities
  - Transition of the Internet towards IPv6

- **These are considered as being the enabling technologies that make "Internet of Things" applications possible.**

Le rôle du hardware



- Calcul/analytics dans l’équipement (Fog Computing)
- Data Science for IoT (Synapse Chip for Deep Learning)

- Pour quel usage
- Quelle mémoire
- Quel microprocesseur
- Gestion de l’énergie
- Quel type de communications
- Temps réel
- Cloudification

Technology role

- **Technologies for the Internet of Things such as sensor networks, RFID, M2M, mobile Internet, semantic data integration, semantic search, IPv6, etc.**
  - technologies that enable “things” to acquire contextual information
  - technologies that enable “things” to process contextual information
  - technologies to improve security and privacy

IoT Technologies

- **ITU identified these technologies:**
  - Capture: Feeling Thing: Sensor technology
  - Identification: Tagging Thing: RFID technology
  - Thinking Thing: Smart technology
  - Others: Nanotechnology, Robotics, …
- **Other enabling technologies:**
  - communication technologies
  - radio communication (support mobility)
  - Powerline communication (support homenetworking)

Sensor technology

- **Electronic device**
  - detects, senses or measures physical stimuli
  - converts signals from stimuli into analogue or digital form,
- **Sensors classification according to the parameter they measure**
  - mechanical (e.g. position, force, pressure, etc.),
  - thermal (e.g. temperature, heat flow),
  - electrostatic or magnetic fields,
  - radiation intensity (e.g. electromagnetic, nuclear),
  - chemical (e.g. humidity, ion, gas concentration),
  - biological (e.g. toxicity, presence of biological organisms)
  - Military – enemy tracking or battlefield surveillance;

Operating Systems

- TinyOS
- Contiki
- RIOT
- NewAE
- MoteWare
- OpenMote
- OpenWSN.com
- OpenWSN.net Arduino-based
- OpenWSN TongJi University
- Brillo google

Smart thing

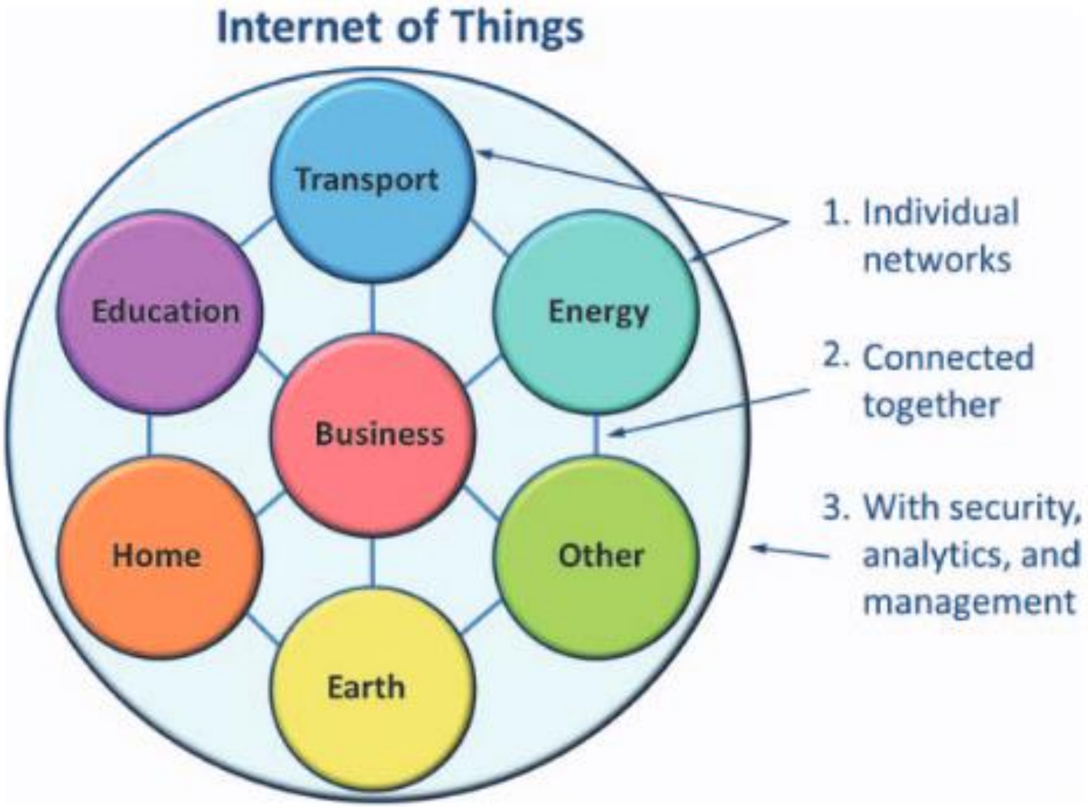
- **Pour que l'objet soit intelligent, il doit avoir:**
  - Une identité : via IPv6
  - Un mécanisme de communication (radio par exemple)
  - Des capteurs

  +
  - Un contexte physique
  - Un contexte social

  +
  - Des décisions

  => Données

Network of networks



IoT viewed as a network of networks.
(*Source*: Cisco IBSG, April 2011).

IoT Terminology

- Sensor
- Actuator
- Cyber Physical System (CPS)
- Internet of Things
- Internet of Everything
- Industrial Internet of Things (IIoT)
- Internet des objets
- Thing to Thing
- Machine to Machine, M2M
- …etc

IoT modules

- **IoT solutions comprising a number of modules**
  - Module for interaction with local IoT module, responsible for acquisition of observations and their forwarding to remote servers for analysis and permanent storage
  - Module for interaction with remote IoT devices, directly over the Internet or more likely via a proxy
  - Module for application specific data analysis and processing of observations acquired, executes appropriate data processing algorithms and generates output in terms of knowledge presented to users
  - Module for integration of IoT-generated information into the business processes of an enterprise
  - User interface (web or mobile): visual representation of measurements in a given context

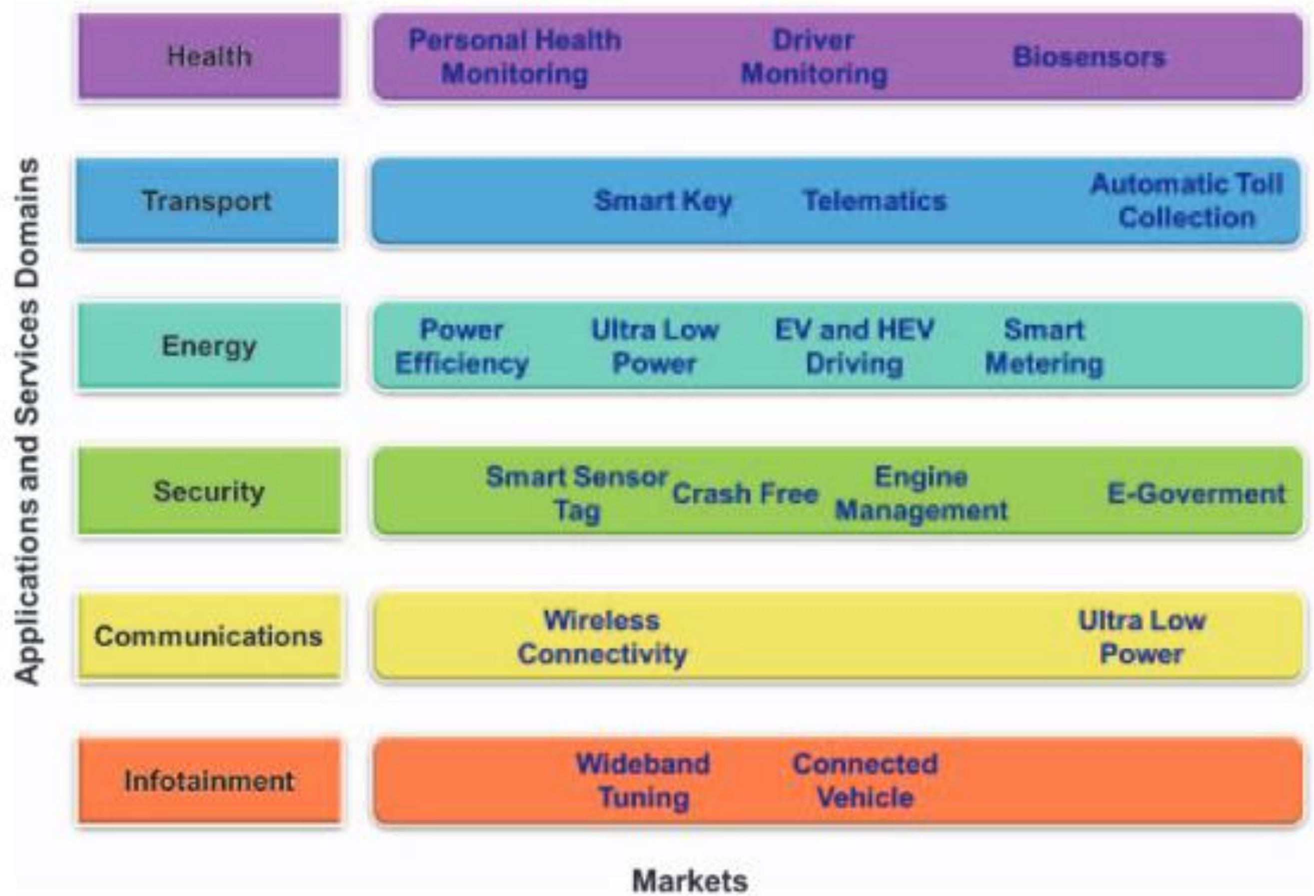
Standardization challenges

- **One of the crucial factors for the success of IoT is stepping away from vertically-oriented, closed systems towards open systems, based on open APIs and standardized protocols at various system levels.**
- **In this context innovative architecture and platforms are needed to support highly complex and inter-connected IoT applications.**

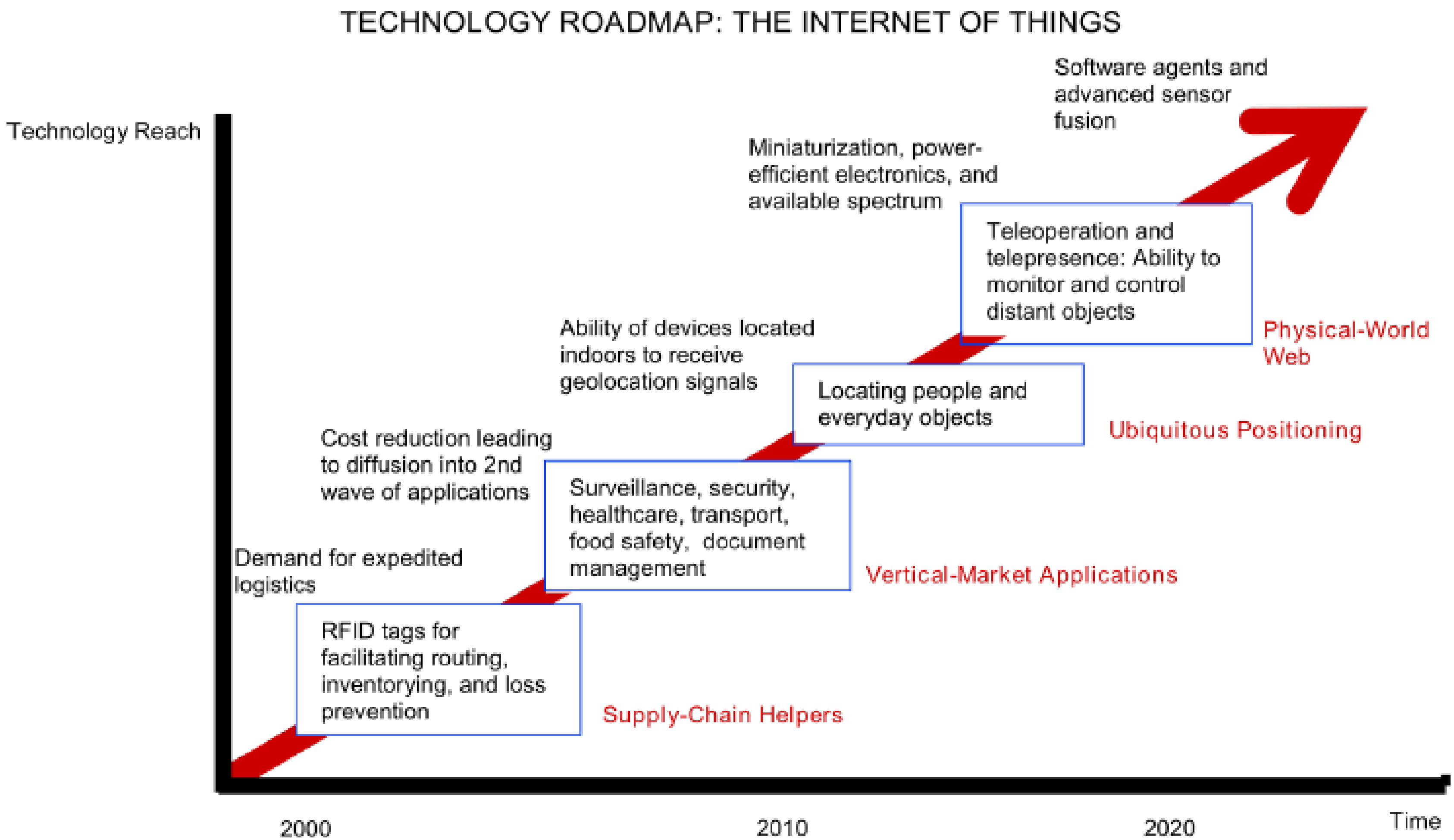
Applications

- **With the IoT, the possibilities are endless**
- **Some of the hottest areas of application, according to Goldman Sachs, are**
  - connected homes (smart thermostats and security),
  - wearables (fitness bands),
  - the industrial Internet (real-time analytics and automation),
  - connected cities (smart meters)
  - connected cars (fleet management).

IoT applications



Technology trend



Source: SRI Consulting Business Intelligence

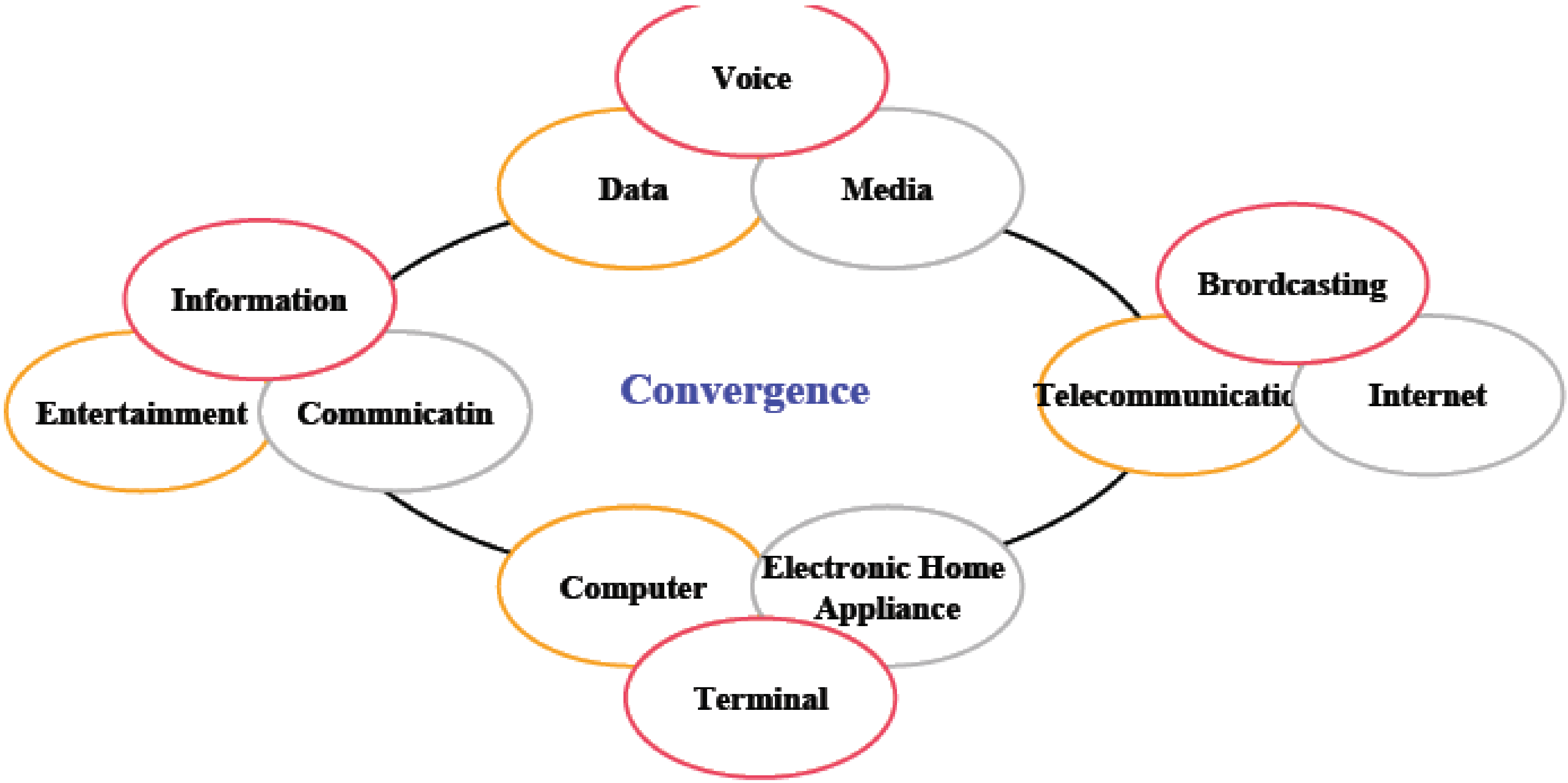
IoT Platforms

- Platforms rely on the power of network effects, as they allow more things, they become more valuable to the other things and to users that make use of the services generated.
- The success of a platform strategy for IoT can be determined by connection, attractiveness and knowledge/information/data flow.

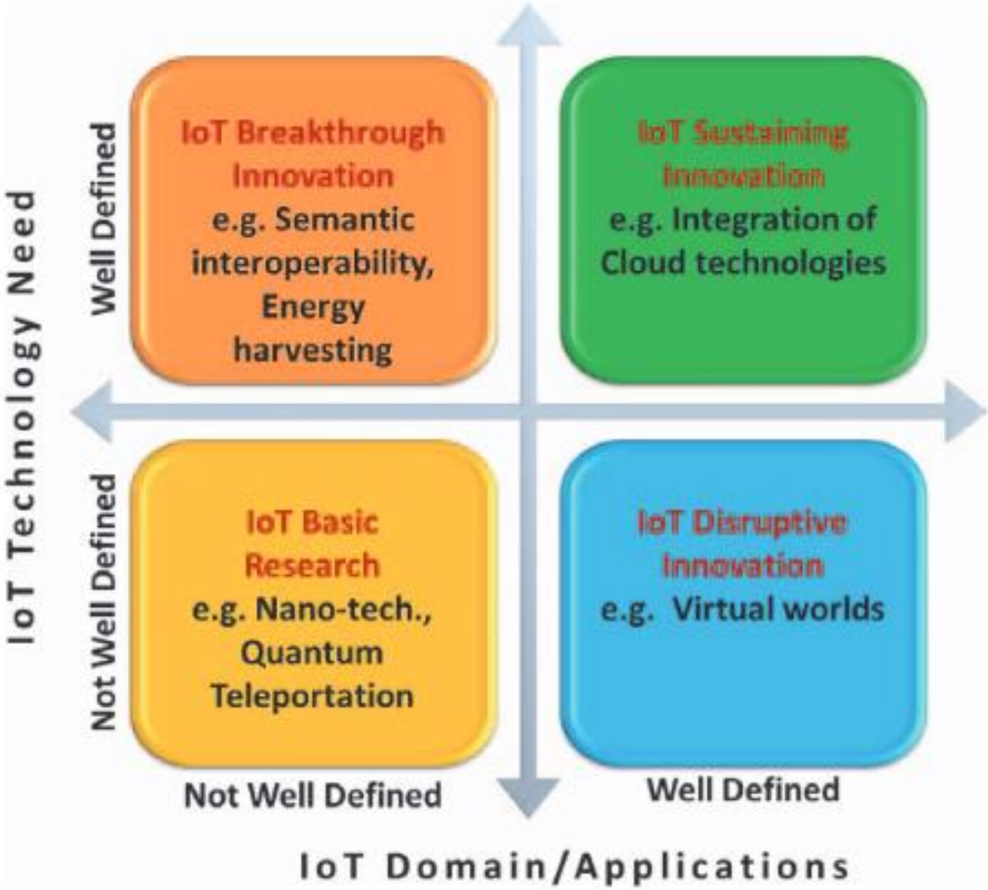
Platforms

- ORACLEIOT
- CISCO IOX
- SAP HANA, Cisco UCS
- SWARM
- PMT: (AXEDA, ThingWorks)
- OpenRemote
- Etherios
- Azure IoT Hub
- Amazon AWS IoT
- SensorLogic
- eObject

IoT convergence



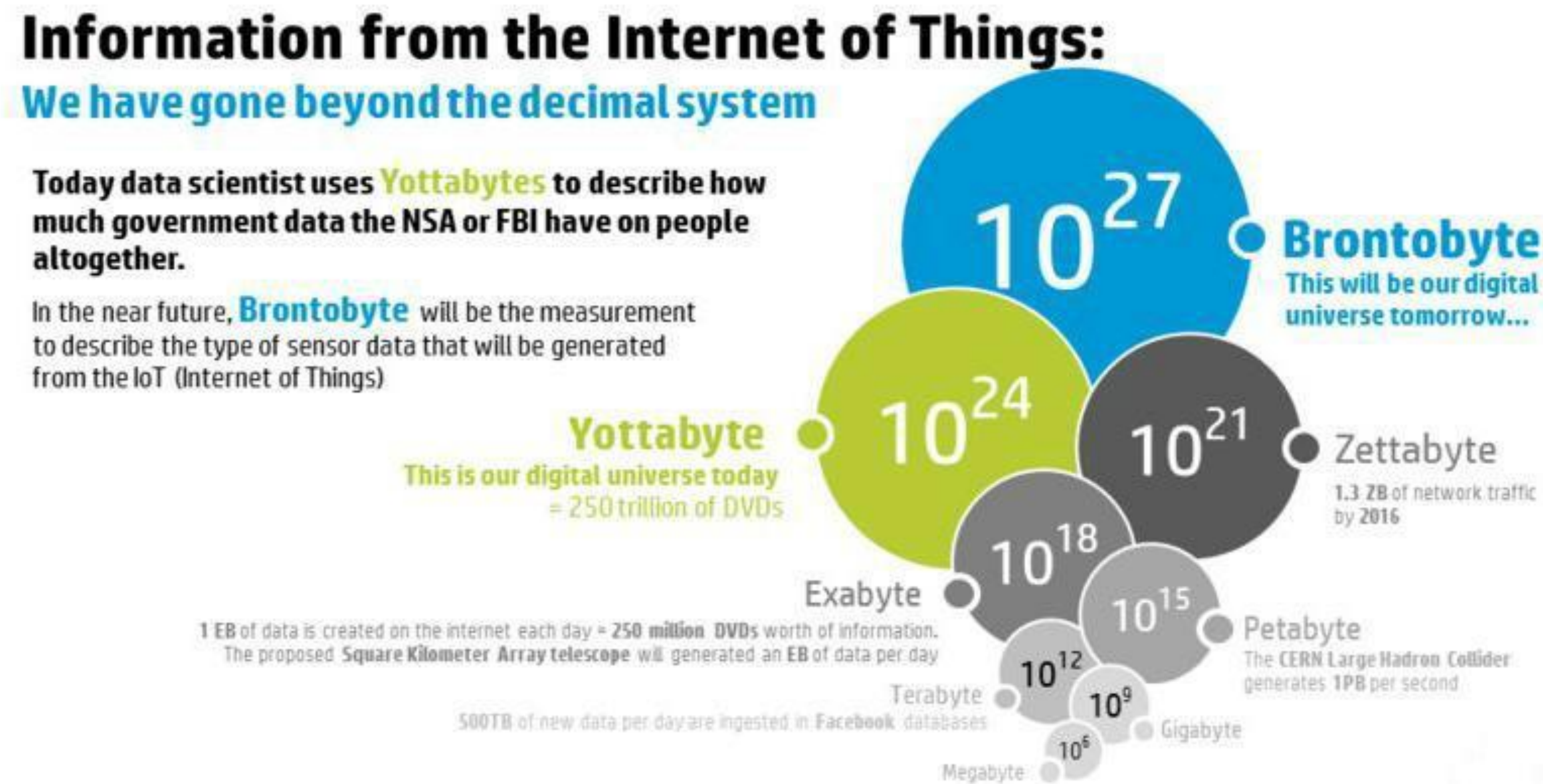
IoT technologies evolution



Les objets génèrent des données

- **Types de données**
  - Données de statut
  - Données de localisation – même statique
  - Actions – un thermostat qui réduit la température
  - Données actionnables – humain, workflow, IFTTT
  - Machine Learning non supervisé / Deep learning

Data volume

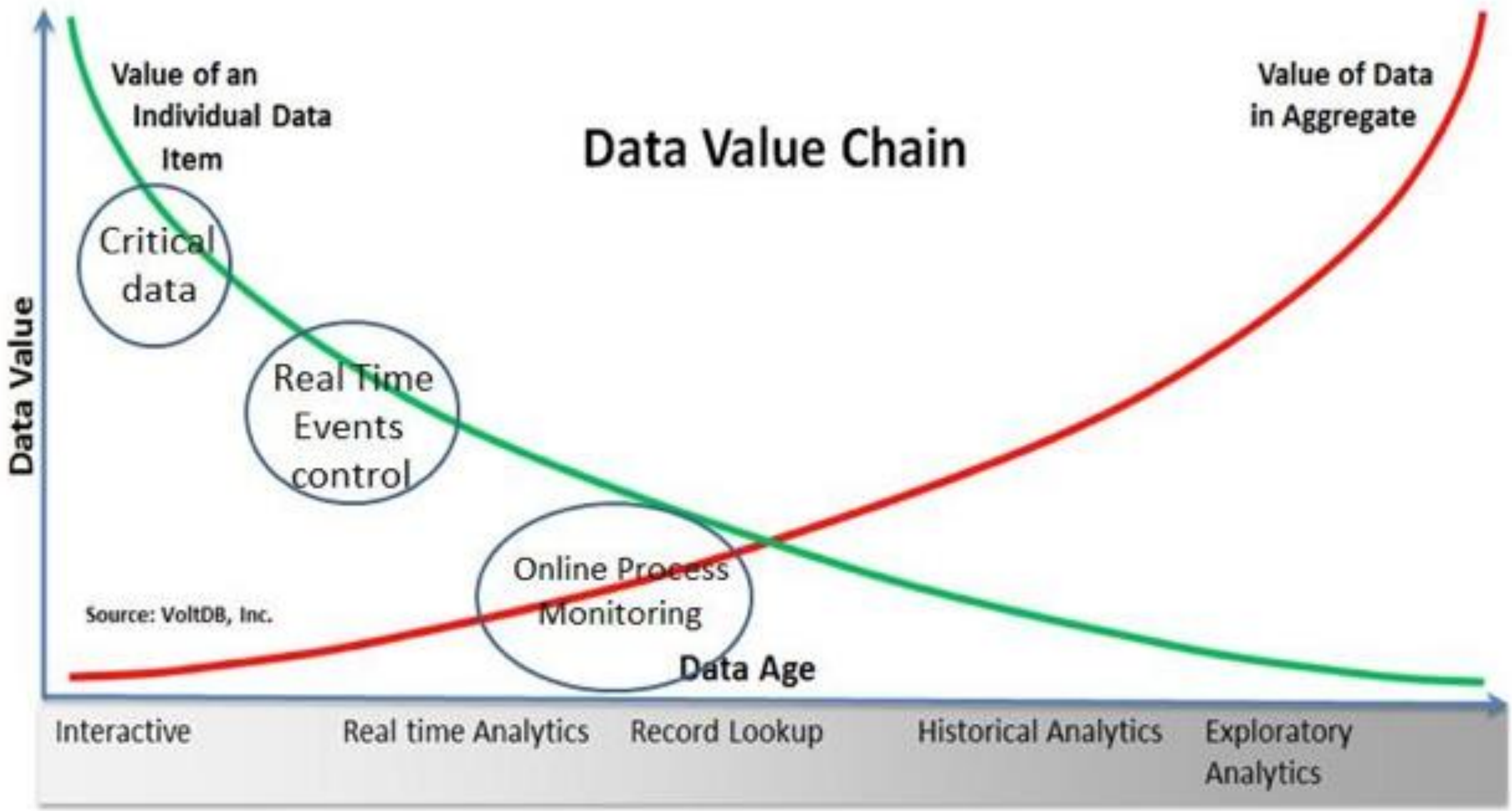


Information from the Internet of Things:
We have gone beyond the decimal system
Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.
In the near future, Brontobyte will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)
10^{27}
Brontobyte
This will be our digital universe tomorrow...
10^{24}
Yottabyte
This is our digital universe today
= 250 trillion of DVDs
10^{21}
Zettabyte
1.3 ZB of network traffic by 2016
10^{18}
Exabyte
1 EB of data is created on the internet each day = 250 million DVDs worth of information.
The proposed Square Kilometer Array telescope will generated an EB of data per day
10^{15}
Petabyte
The CERN Large Hadron Collider generates 1PB per second
10^{12}
Terabyte
500TB of new data per day are ingested in Facebook databases
10^{9}
Gigabyte
10^{6}
Megabyte

It’s All About The Data

- **Collecting, preparing and analyzing all of this streaming, fragmented data**
- **Data volumes can double every few months**
- **Data is complex – often in hundreds of different semi-structured and unstructured formats**
- **Big data analytics**
  - Combine, integrate, and analyze all of the data at once – structured, semi-structured, and unstructured – regardless of source, type, size, or format
  - Quickly and affordably scale to huge volumes of data and analyze them for insights

Value of data



Data management

- Data management is a crucial aspect in the Internet of Things. When considering a world of objects interconnected and constantly exchanging all types of information, the volume of the generated data and the processes involved in the handling of those data become critical.
- challenges and opportunities of data management are:
  - Data Collection and Analysis
  - Big Data
  - Semantic Sensor Networking
  - Virtual Sensors
  - Complex Event Processing.

Data Collection and Analysis

- Part of the core layer of any IoT platform
  - User/customer data storing: Provides storage of the customer's information collected by sensors
  - User data & operation modelling: Allows the customer to create new sensor data models to accommodate collected information and the modelling of the supported operations
  - Customer workflows: Allows the customer to create his own workflow to process the incoming events from a device
  - Multitenant structure: Provides the structure to support multiple organizations and reseller schemes.

Data collection and Analysis

- **De-centralisation**. Sensors and measurements/observations captured by them are stored in systems that are de-centralised from a single platform. It is essential that different components, geographically distributed in different locations may cooperate and exchange data.
- **Security**. Data collection platforms have a high level of data protection and security, from the transmission of messages from devices (sensors, actuators, etc.) to the data stored in the platform
- **Data mining features**. Capacities for the processing of the stored info, making it easier to extract useful data from the huge amount of contents that may be recorded.

Networking Issues: Impact on Data Collection

- **variety of technologies are used for data collection, each of which have different tradeoffs in terms of capabilities, energy efficiency, and connectivity, and may also impact both the cleanliness of the data, and how it is transmitted and managed**
- **make the data collection as energy efficient as possible**
- **appropriate data cleaning and processing methods must take issues of data quality into consideration**

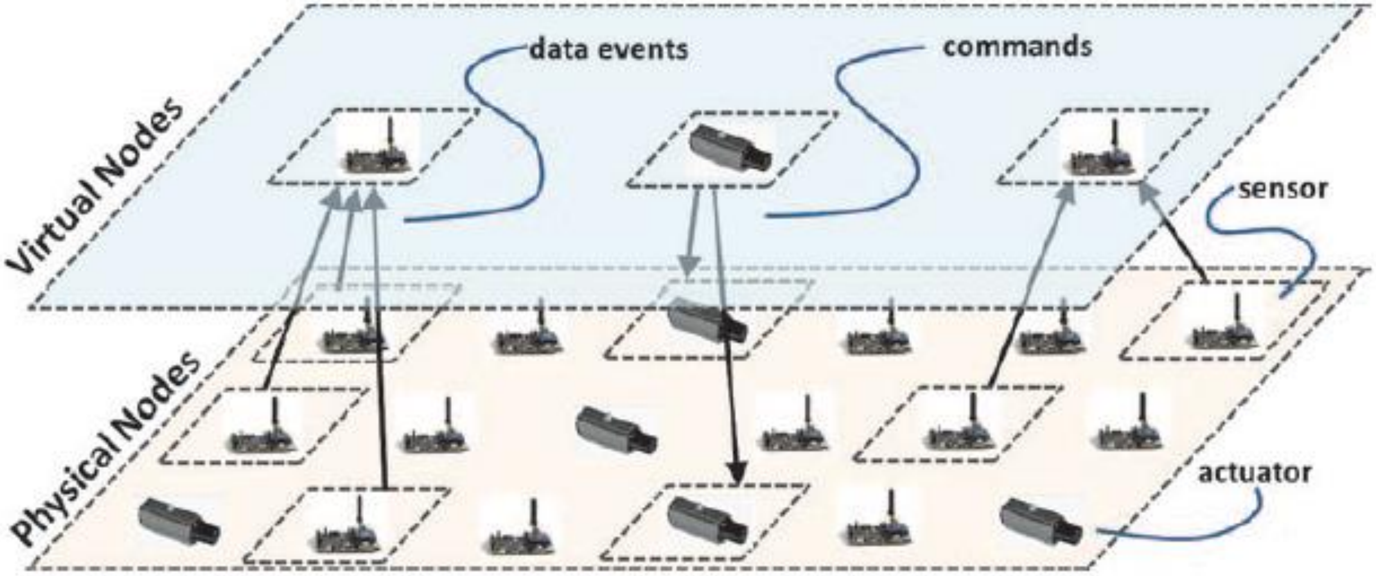
RFID and data collection

- **Limited capabilities in terms of providing more detailed sensing information, especially when passive tags are used**
- **The range of the tags is quite small, and is typically of the order of between 5 to 20 meters.**
  - significant numbers of readings are dropped
- **The data collected is massively noisy, incomplete and redundant.**
  - Challenges of data cleaning, it needs to be performed in the middleware within the sensor reader
- **Privacy challenges, especially when the tags are associated with individual**

Linked Data vision:
Semantic Sensor Web (SSW)

- Define ontologies and create frameworks to apply semantic Web technologies to sensor networks
- propose annotating sensor data with spatial, temporal, and thematic semantic metadata
- make this information widely available for different applications, front-end services and data consumers
- connect sensor descriptions to potentially endless data existing on the Web
- create business intelligence, enable smart environments, and support automated decision

Virtual Architecture



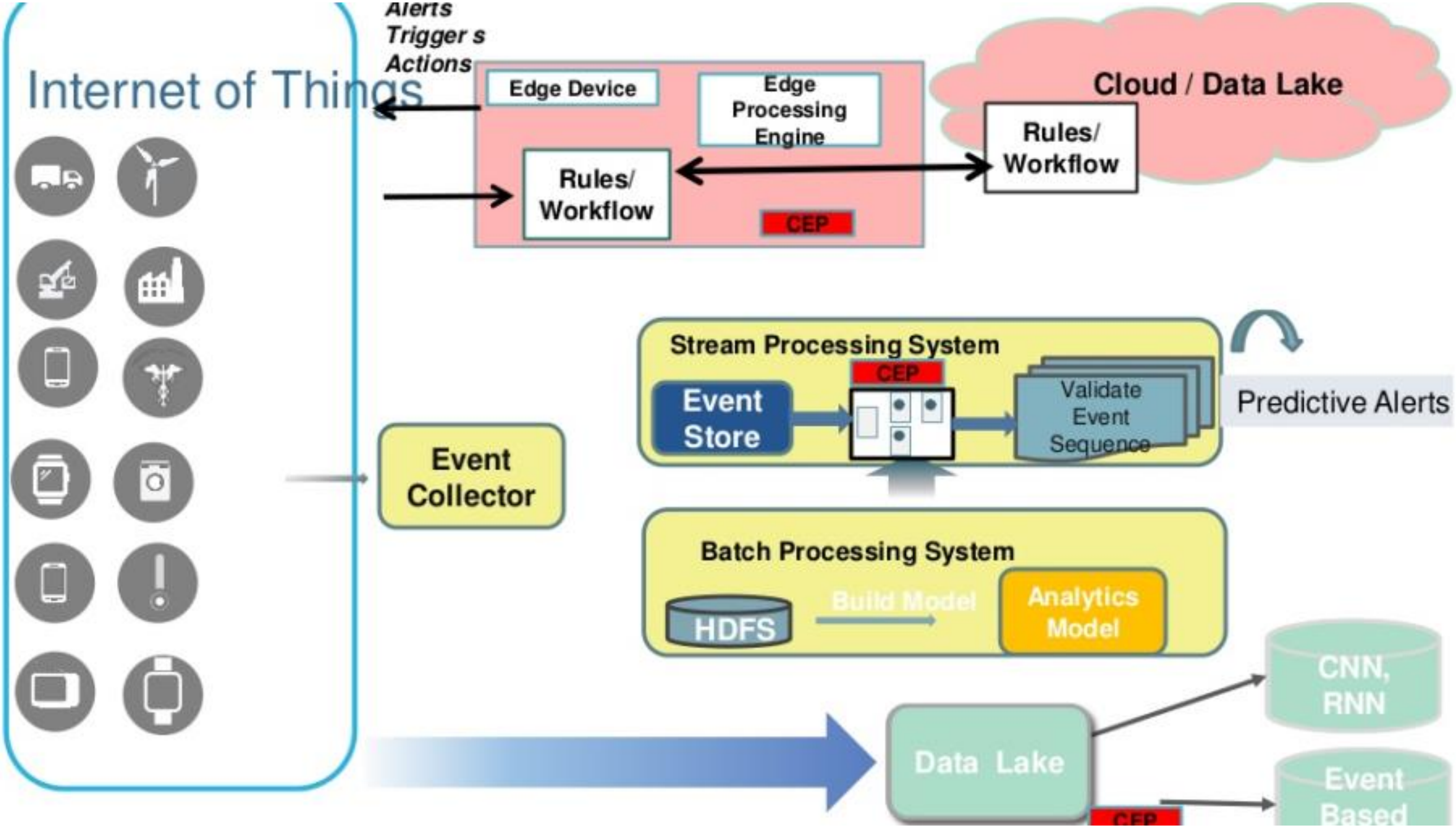
Complex event processing (CEP)

- Emerging network technology that
  - create actionable, situational knowledge from distributed message-based systems, databases and applications in real time or near real time
  - provide an organization with the capability to define, manage and predict events, situations, exceptional conditions, opportunities and threats in complex, heterogeneous networks.
  - extract higher level knowledge from situational information abstracted from processing sensory information and for low-latency filtering, correlating, aggregating, and computing on real-world event data.

CEP Challenges

- **Distributed CEP:** Since CEP core engines usually require powerful hardware and complex input data to consider, it is not easy to design and implement distributed systems capable of taking consistent decisions from non-centralised resources.
- **Improved security and privacy policies:** CEP systems often imply the handling of “private” data that are incorporated to decision taking or elaboration of more complex data. It is necessary that all processes and synthetic data can be limited by well-defined rules and security constraints (that must be measurable, traceable and verifiable).

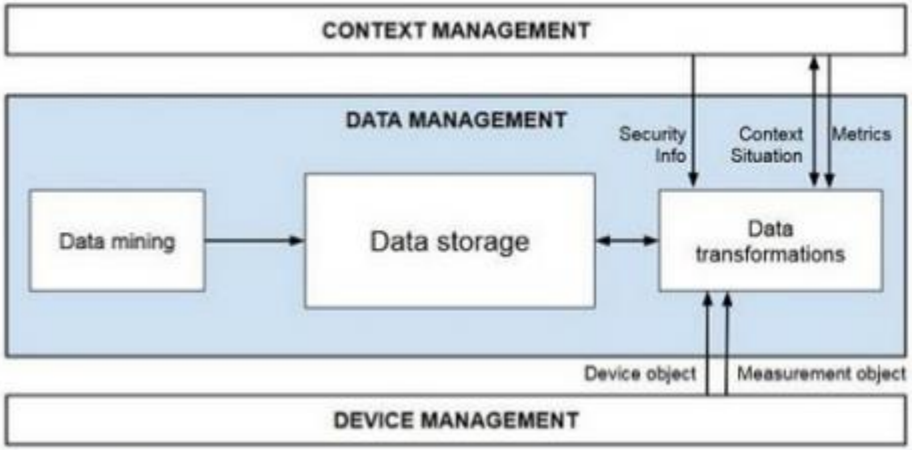
IoT architecture



Source : Futuretext

Data Management and Analytics

- the power of the internet of things paradigm is the ability to provide real time data from many different distributed sources to other machines, smart entities and people for a variety of services
- the underlying data from different resources are extremely heterogeneous, can be very noisy, and are usually very large scale and distributed
- methods are required to clean, manage, query and analyze the data in the distributed way

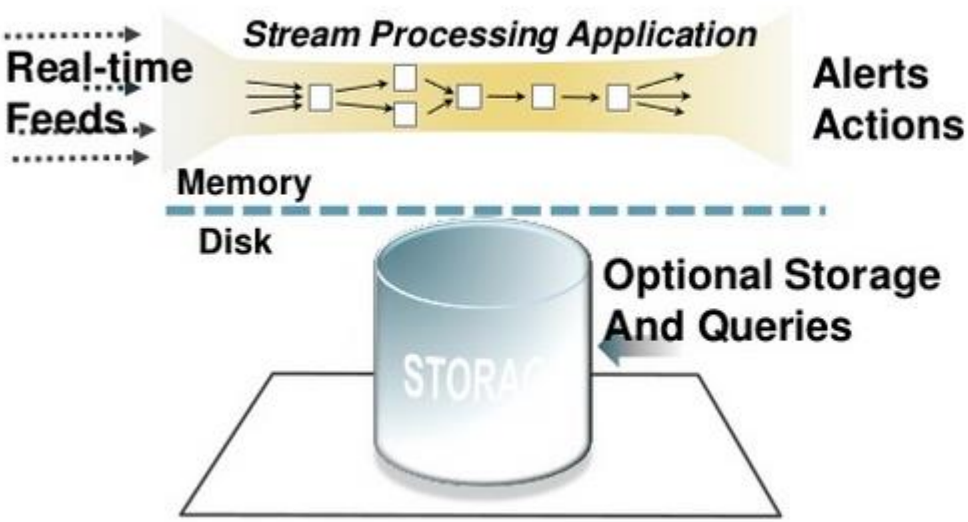


Real Time and Big Data Analytics

- **RFID and conventional sensors form the backbone of the data collection mechanisms in the IoT**
  - streaming nature of the collected data
  - smart infrastructures typically have a large number of objects simultaneously collecting data and communicating with one another
- **communications and data transfers between the objects require to enable smart analytics**
  - bandwidth and energy consumption

IoT Stream Mining

- Maintain models online
  - Incorporate data on the fly
  - Unbounded training sets
  - Detect changes and adapts
  - Dynamic models



IoT Big Data Streams

- Volume + Velocity (+ Variety)
- Too large for single commodity server main memory
- Too fast for single commodity server CPU
- A solution needs to be:
  - Distributed
  - Scalable

Big Data processing engines

- Low latency

Flink

S4

Apache Storm

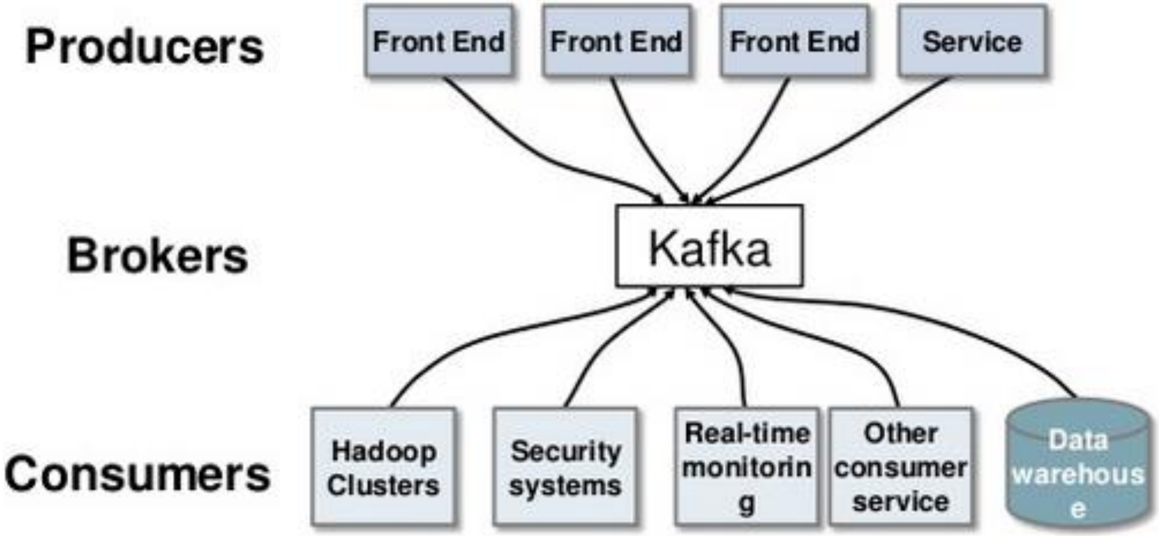
samza

- High Latency (Not real time)

Spark Streaming

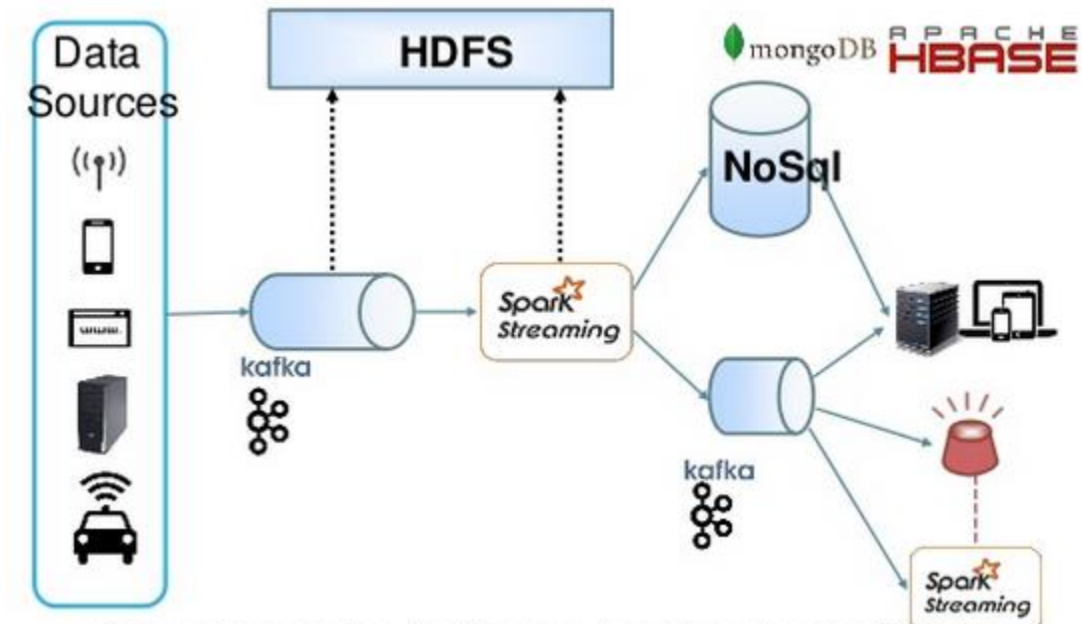
Apache Kafka

- fault-tolerant distributed streaming platform



Apache Spark Streaming

- Spark Streaming is an extension of Spark that allows processing data stream using micro-batches of data.

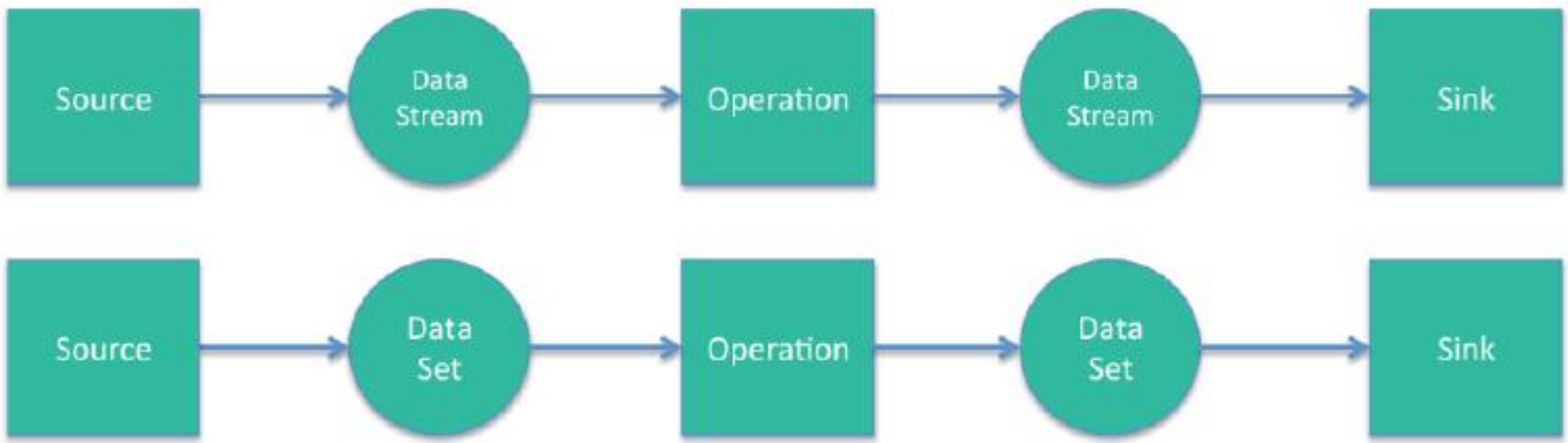


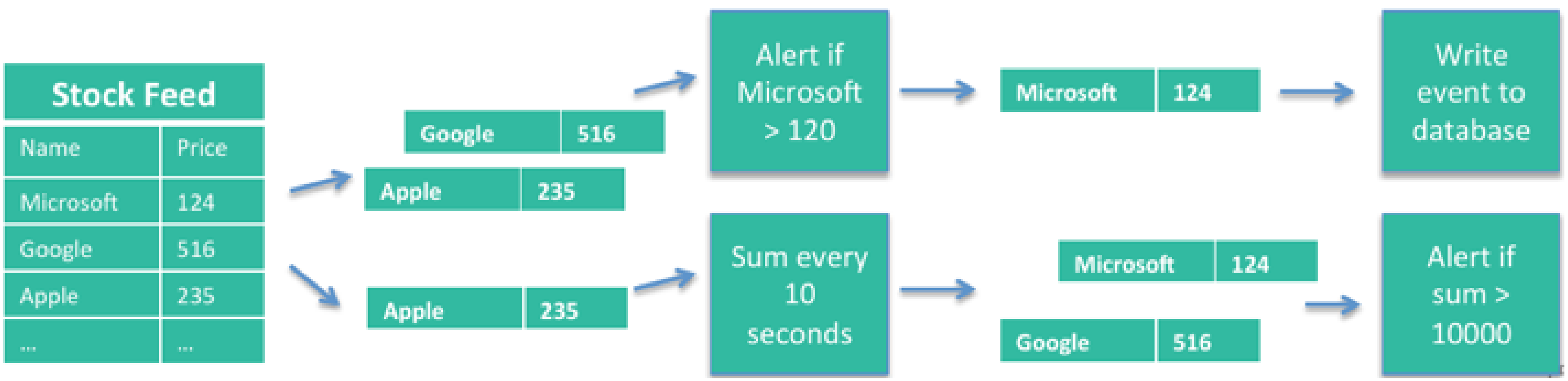
Stream Processing Architecture based on Apache Spark

Spark Streaming

- Discretized Stream or DStream represents a continuous stream of data
  - either the input data stream received from source, or
  - the processed data stream generated by transforming the input stream.
- Internally, a DStream is represented by a continuous series of RDDs

Apache Flink



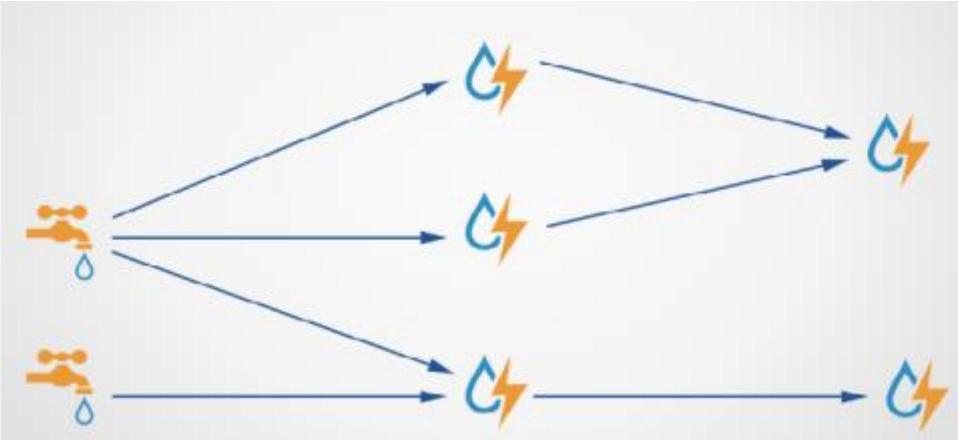


Hadoop AllReduce

- MPI AllReduce on MapReduce
- Parallel SGD + L-BFGS
- Aggregate + Redistribute
  - Each node computes partial gradient
  - Aggregate (sum) complete gradient
  - Each node gets updated model
- Hadoop for data locally (map-only job)

Apache Storm

- **Distributed real-time computation system for processing large volumes of high-velocity data**
- **Most popular streaming engine in industry**
- **Storm modes**
  - One-at-a-time processing (pure Storm)
  - Micro batch processing (Storm Trident)



Requirements

- **Algorithms to filter data before it reaches a centralized database**
- **Automated indexing at the edge ensures that important data and timely alerts are processed and delivered to the right decision makers**
  - whether they are working with streaming data or pulling historical data
  - impertinent and superfluous information is left at the edge and kept from clogging up the central system
- **Automated aggregation and classification at the edge accelerates the generation of insight from IoT data and protects databases from overwhelming data volume and velocity**
- **Properly managed IoT data is most valuable when it is connected and complemented with other data sources**

Conclusion : challenges of IoT

- **Communication protocols**
- **IP-addressability**
- **Privacy**
- **Security**
- **Data management and analytics**
  - IoT organizations need data management solutions that facilitate rapid decisions, no matter how many end points are involved
  - Data management is moving from the central data repository towards the edge of the network
    - add intelligence to the edge to streamline data processing and management
    - automated data indexing and classification
    - invest to make data available in multiple interfaces for easy access and analysis

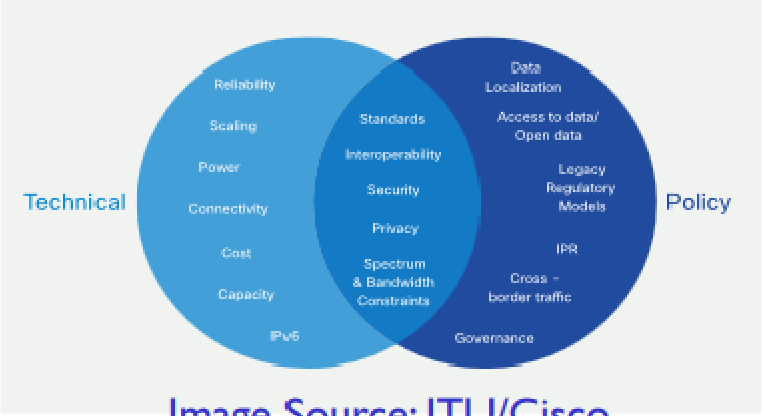


Image Source: ITU/Cisco