# HOW DATA PIPELINES SUPPORTS IOT DATA PRE-PROCESSING NEXT STEP?

## KEY CONSERNS:

**P3**

Growth and evolution of internet of things in last ten decades - A Retrospective.

**P5**

What are data pipelines and how are they functioning ?

**P9**

What can we do with data pipelines and their benefits over IoT data pre-processing.

CITYWAVE WHITE PAPER

Omal Perera
Software Engineer at 99x Technology,
CISCO Certified Network Associate,
Undergraduate Sabaragamuwa University, Sri Lanka.

Dr. K.P.N.Jayasena
Lecturer, Sabaragamuwa University, Sri Lanka
PhD in Engineering in Computer Sciences,
M.Sc in Computer Sciences & Engineering,
Wuhan University of Technology, Wuhan, PR China
B.Sc (Hons) in Information Technology, SLITT, Sri Lanka
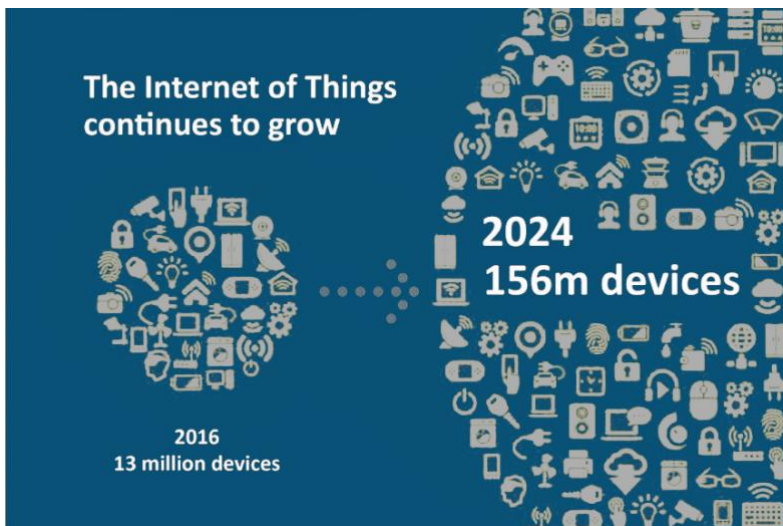
July 2018

# CONTENT

# GROWTH OF INTERNET OF THINGS IN LAST TEN DECADES. RETROSPECTIVE.

## INTERNET OF THINGS EVOLUTION: MAJOR MILESTONES

The Internet of Things (IoT) has the capability to change our world. And while we are starting to see its incredible impact, we are still very much at the beginning of the transformational journey. Soon every device you own – and nearly every element you can imagine – will be linked to the Internet. Whether it's through your phone, wearable tech or everyday household objects, the Internet of Things (IoT) will connect us in ways we can't even imagine yet.
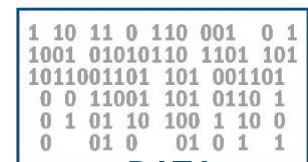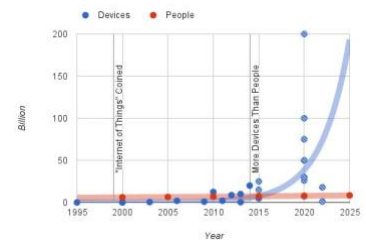
- There were more devices connected to the Internet in 2008, than there were people.
- In 2015, there were 4.9 billion (4,9000,000,000+) connected Internet of Things. Smartphone users topped at over 6.1 billion (6,100,000,000+).
- Predictions for the number of connected IoT by 2020, meets or goes over 50 billion (50,0000,000,000+). A quarter of a billion vehicles are expected to be connected to the Internet.



**Reasons for the IoT Growth..**

Mobility + connectivity made a huge motivation for Internet of things boom. Linking components via invisible information made life easier.





**DATA**

**VALUE**

The Internet of Things is not deaccelerating, and the list of connected people, systems and things keeps growing. Digital ecosystems are all around us – new business models are arising. The potential possibilities and opportunities IoT brings from businesses and clients is tremendous. Whether you are a organization or a consumer it already affects us all, every day and with everything we do in our daily lives.

# CHALLENGES THAT HINDERS IOT UTILIZATION : FEW IMPLEMENTATION CHALLENGES

## HARDWARE COMPATIBILITY ISSUES

Data capturing majorly occurs through various sensors, Programmable Logic Controllers (PLCs), etc., which are connected to IoT gateways to collect & transmit data to the cloud. Enterprises need to meticulously **identify the equipment, hardware, and existing legacy machines** based on their goals & business outcomes. When there are legacy machines that do not have the said PLCs and sensors involved, the IoT implementation challenge becomes more critical. **Adding external sensors to the legacy machines is one quick work around**, but it will not be full proof, making it a very challenging task. Therefore, identifying the physical devices and understanding the associated compatibility issues before IoT implementation is highly recommended[1].
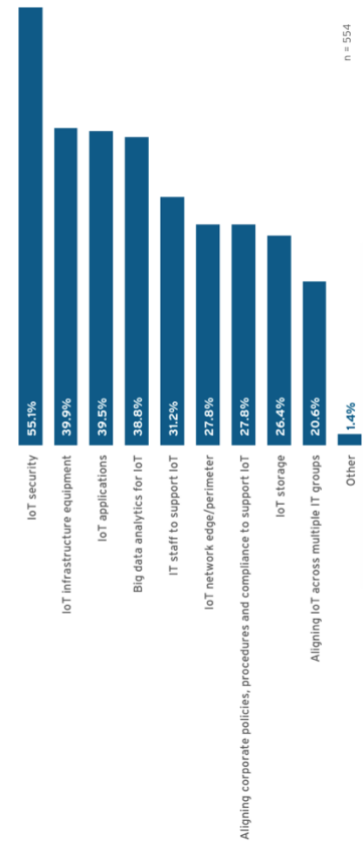
## INCORRECT DATA CAPTURE DIFFICULTIES

Let's consider that the entire setup did not face any challenges and the system is up and running, but there can still be the issue of incorrect data capture creeping up. Because of some untoward incident or the inability of the software to handle certain **anomalies in run time, incorrect data gets recorded**. This results in inaccurate analytics that may not help in taking better decisions. This IoT implementation challenge can be a major decision influencer for the enterprises as well as the customers.

## ANALYTICS CHALLENGES

The real value of an IoT solution is realized through actionable insights derived from the collected IoT data. This demands a high-performance analytics platform that is capable of handling the ginormous amount of data to be added in the solution at a later point. Data Analytics partners need to keep this in mind while devising the **IoT implementation architecture to involve data processing, cleansing and representation.** Thus, **leaving enough space for extensibility to add real-time or predictive analytics to an IoT solution can help solve this critical IoT implementation challenge.**

**Priorities for deploying IoT Initiatives [3]**

n = 554

IoT security — 55.1%
IoT infrastructure equipment — 39.9%
IoT applications — 39.5%
Big data analytics for IoT — 38.8%
IT staff to support IoT — 31.2%
IoT network edge/perimeter — 27.8%
Aligning corporate policies, procedures and compliance to support IoT — 27.8%
IoT storage — 26.4%
Aligning IoT across multiple IT groups — 20.6%
Other — 1.4%

basic Concerns when Implementing IoT : SECURITY

- Data Privacy
- Data Integrity
- Data Availability

1 6 key IoT Implementation Challenges for Enterprises to consider, 2018,
    http://www.saviantconsulting.com/blog/iot-implementation-challenges-enterprises.aspx
2 Common Concerns When Implementing IoT, 2017; https://dzone.com/articles/common-
    concerns-when-implementing-iot-and-how-to-respond
3 Survey finds security continues to be top priority in deploying IoT projects, 2018 :
    https://451research.com/blog/1934-survey-finds-security-continues-to-be-top-priority-in-
    deploying-iot-projects

## DATA SECURITY ISSUES

With occurrences of many ransomware attacks recently, Enterprises and customers are apprehensive about data security. There is also a chance of corporate espionage to gain intellectual property. Hence, IoT service providers need to ensure that their data is going to be safe. These data security issues can be taken care by using a comprehensive governance mode, which provides secure access to sensitive reports and data. This phase of planning, that defines various data related security policies is crucial for successful IoT implementation.

# WHAT ARE DATA PIPELINES ?

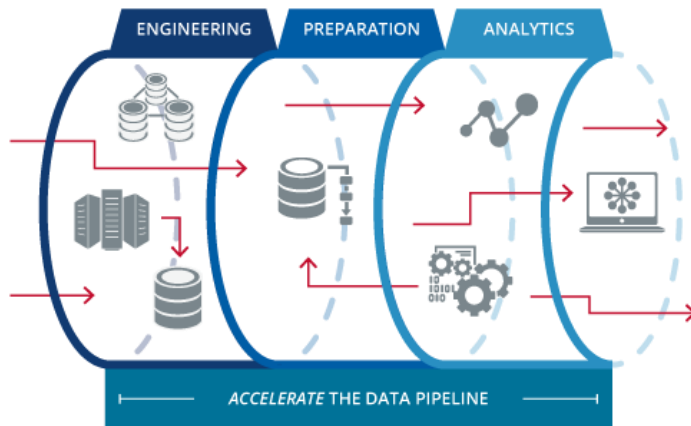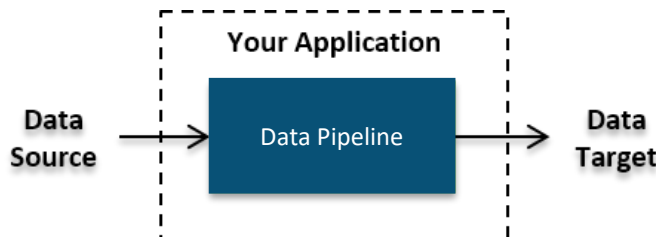A data pipeline is a **set of actions** that extract data (or directly analytics and visualization) from various sources.[4]



Figure : how a data pipeline looks like [5]

**Data pipeline design Considerations**

- Data: Origin, Type, and Timeliness
- Storage Mechanism
- Language Selection
- ETL Dependency Management
- Fault Tolerance
- Pipeline Monitoring
- Accessibility and Visualization
- System Constraints

Data Pipelines cab be embedded with  data processing engines in order to transform input data stream and bring them to target dataset.



---

4  Building a Data Pipeline from Scratch, 2016 : https://medium.com/the-data-experience/building-a-data-pipeline-from-scratch-32b712cfb1db

5 Data pipeline image (Extract, Transform, Load), 2016: https://cdn-images-1.medium.com/max/800/1*8-NNHZhRVb5EPHK5iin92Q.png
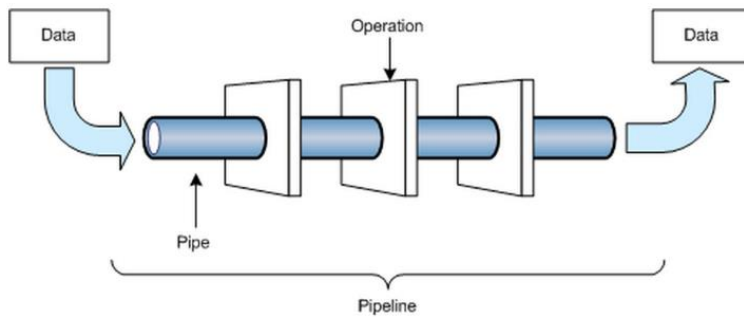
6  What is data Pipeline : https://northconcepts.com/docs/what-is-data-pipeline/

Figure : pipeline made-up of several operations [7]

As in the above figure different operations, transformation functionalities can be implemented in the pipeline which provides defined servicers to your input data set and ultimately it will bring the data set to your desired form.

If you have a well built and structured pipeline, you can go to check where is the wrong transformation, change it and that's it. If you don't have a pipeline either you go changing the coding in every analysis, transformation, merging, data whatever, or you pretend every analysis made before is to be considered void[8].

# EMPLOYING DATA PIPELINES TO MITIGATE THE IMPACT OF IOT HETEROGENITY CHALLENGES

## THERMO MONITOR SYSTEM USECASE

Let's consider a scenario with large scale of thermometers which belongs to three vendors (T1, T2, T3) that consist of heterogeneous properties between each other. Following are the data emitting structure of three types of thermometers.

**Disclaimer**: This use case for the sake of understanding only the temperature value & the timestamp is considered as the essential fields for big data processing.

**Heterogeneous industrial Thermometers**



### Properties of Thermometer T1

- Device name – Waspmotex
- Temperature Sensor id - MCP9700A
- Data structure

```
{
  SensorCities.ON();
  SensorCities.setSensorMode(SENS_ON,SENS_CITIES_TEMPERATURE);
  delay(100);
  float temperature_value;
  temperature_value=SensorCities.readValue(SENS_CITIES_TEMPERAT
    URE (Celsius));
}
```

7 Figure Data Pipeline : https://www.dataquest.io/blog/content/images/pipeline.png
8 Industrial Thermometers Images : http://sensirion.com/ | www.Lelong.my | www.myDevices.com

**Properties of thermometer T2**

- Device name – MNS-2-WF-TS-ST
- Data structure

```
{
    Date : 7-16-2017 T16:50:00,
    Signal : 36,
    Battery : 95,
    SensorReading: 80.2 F
}
```

In this system process is done respect to Celsius measuring unit. But


**Properties of thermometer T3**

- Online data set on http://iot.ee.surrey.ac.uk [9]
- Vendor – CityPulse Dataset Collection
- Data structure

```
<http://iot.ee.surrey.ac.uk/citypulse/datasets/weather/aarh
us_weather_temperature#observations_temp> a
    sao:Point ;
    sao:hasUnitOfMeasurement unit0:degree-Celsius ;
    sao:time [a tl:Instant;
        tl:at "2014-02-13T06:20:00"^^xsd:dateTime ];
    sao:value "3.0" .
```

As depicted above data samples, the claim is much valid that there is an issue in handling data streams with the heterogeneous devices.

Assuming that the system is working on common measuring unit of Celsius and uses only the fields of temperature value and timestamp, a sample pipeline can be implemented as follows.
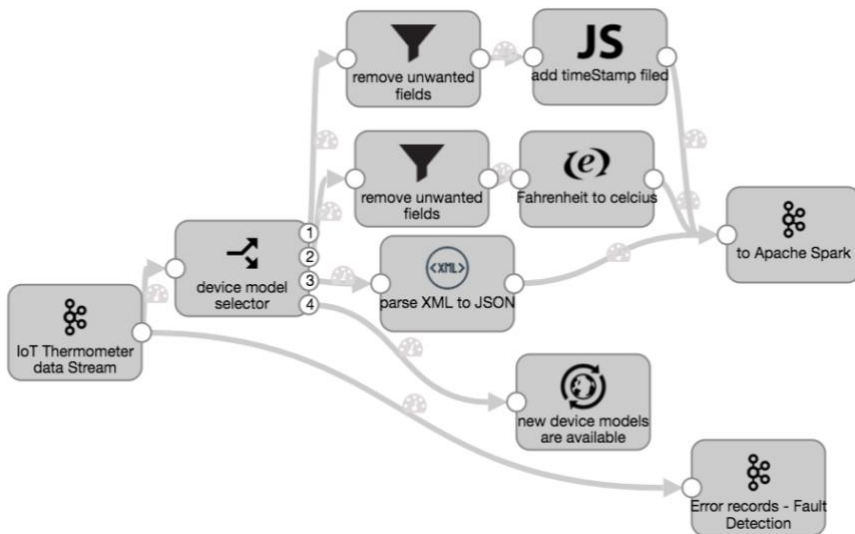


Figure : basic data pipeline designed to filter data stream for use case

Functionality of the above pipeline can be read from the diagram. As a summary it **creates the different data patterns into a unified data stream by removing unwanted key value pairs, converting XML to JSON, converting the data units as defined in the individual pipeline operations.**

With this simple transformation of data stream will creates huge impact in future computations and provides greater support for real-time data processing form multiple perspectives.

## WHAT YOU CAN DO WITH DATA PIPELINES

- Convert incoming data to a common format.
- Supports number of transformations
    - o   Validate
    - o   Filter
    - o   Lookup
    - o   Transform
    - o   Create , Update, Delete data within data records
- Prepare data for analysis and visualization.
- Share data processing logic across web apps, batch jobs, and APIs.
- Power your data ingestion and integration tools.
- Consume large XML, CSV, and fixed-width files.
- Replace batch jobs with real-time data.[10]

## BENEFITS OF PRE-PROCESSING IOT DATA USING DATA PIPELINES

**Customizability**

- Pipeline operations can be  customized as required. You can do whatever process to incoming dataset

**Metadata**

- Data Pipeline allows you to associate metadata to each individual record or field. Metadata can be used to identify the data separately & treat the data in a specific way.

---

**Flexibility**

- Data Pipeline comes with built-in readers and writers to stream data into (or out of) the pipeline. It also comes with stream operators for working with data once it's in the pipeline.
- If data stream pattern get changes, data pipeline can be adjusted to address the change. And also not the whole program, but only the relevant operation that works in particular key value pair.

**Supports to Efficient data processing**

- once data is pre-processed; noise data, irrelevant data will be removed. Therefore the volume of data is getting less. It **supports to quick network transferring data.**
- Since the data dataset is reduced it will take low **processing time to process complex processes.**
- **Low computation power will be consumed.**

**Easy to Use**

- Data Pipeline is very easy to understand, learn and use. Each part can be clearly identified.

**Monitorability**

- Data Pipeline is very easy to understand, learn and use. Each part can be clearly identified

# CityWave.github.io