

# Data mgmt. & Analytics

April, 2018



# Agenda

---

## **Introduction**

What is the problem with data?

## **Data management**

A typical pipeline

A word about disaster recovery

Data hygiene

## **Analytics**

Overview of time series analysis

Visualization demo

## **Exercises**



# Introduction

# What is the problem with data?

## Data has value

Because it can **answer questions**

Keep all data. Don't lose it.

Protect data from prying eyes.

## Data needs:

Space, bandwidth, processing power

Tools to manage and **make sense of it**

## And there are laws too (GDPR – 25 May 2018)

Give me my personal data

Forget about me

Let me know about breaches within 72 hours

Control disclosure to third parties

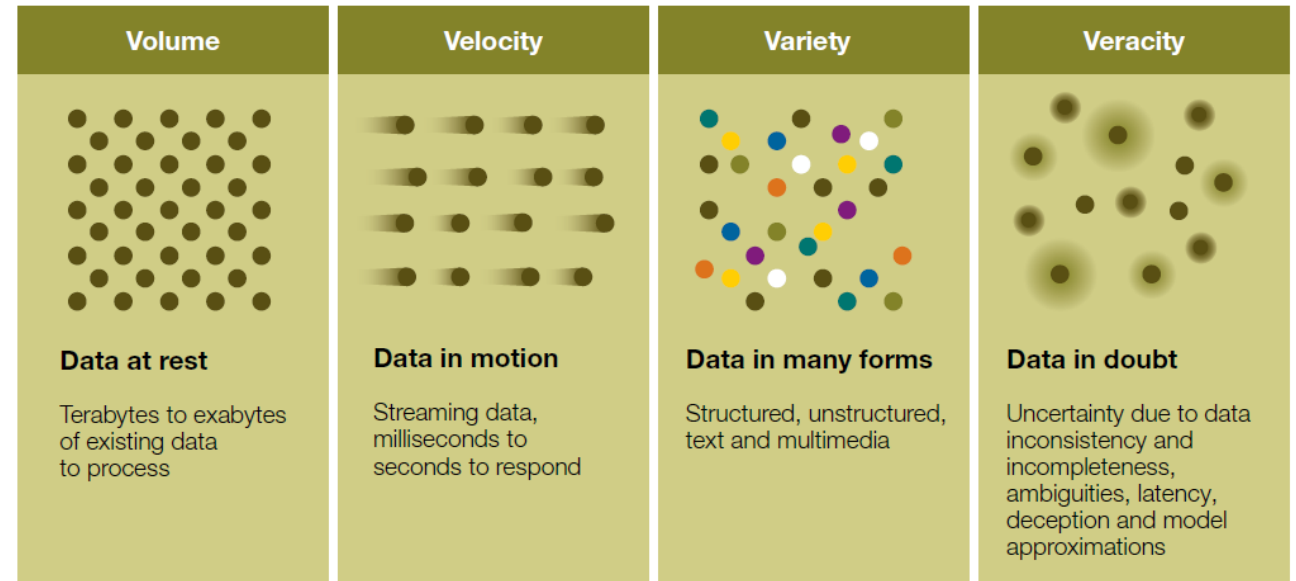


Image: [http://www.ibmbigdatahub.com/sites/default/files/public\\_images/pdf/insurance-post-2-1.png](http://www.ibmbigdatahub.com/sites/default/files/public_images/pdf/insurance-post-2-1.png)



# Data management

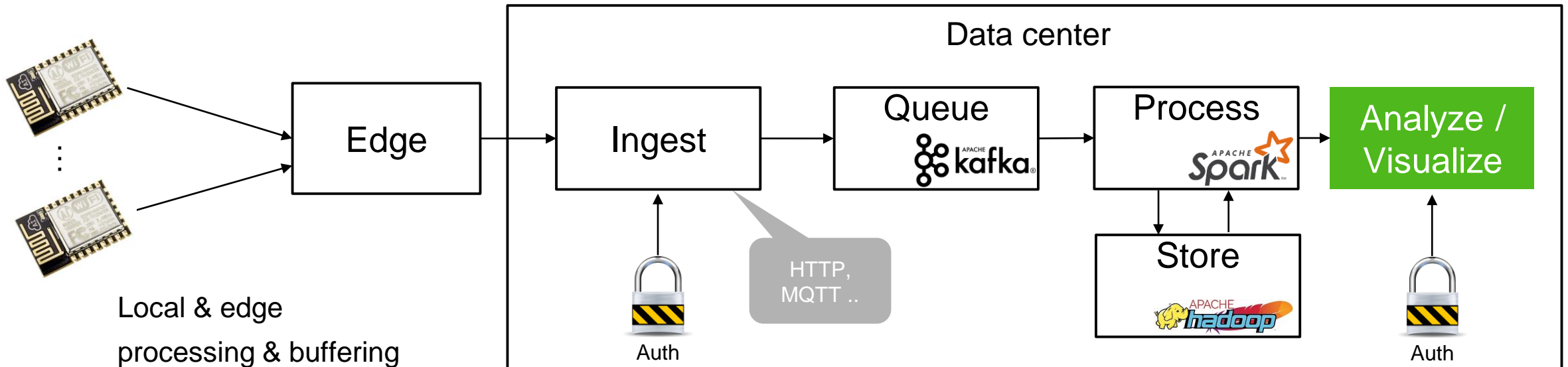
# Typical data pipeline

## Goals

Reliability & HA -> no data loss, tolerate server crashes

Scalability -> data volume & throughput can be increased

Speed: Low latency, fast and convenient analysis



# A word about disaster recovery

## What to do when the s\* happens?

How to reinstall the whole thing?

Where are my backups and how do I restore them?

How to tell the customers?

## Key metrics (SLAs)

MTTR = mean time to recover

RPO = recovery point objective

## Possible solution

Reserved/planned DR data center

Scripted installations (e.g. Bash, Ansible, Chef ...)

Staff guidelines & regular drills



Image: <https://distributedalgorithm.files.wordpress.com/2016/02/data-center-disaster-after-typhoon.jpg?w=600>

# Data hygiene

---

## Valid reasons to delete data (retention policy)

Can't cope with so much data (e.g. no space left on device)

Required by law (e.g. upon user request)

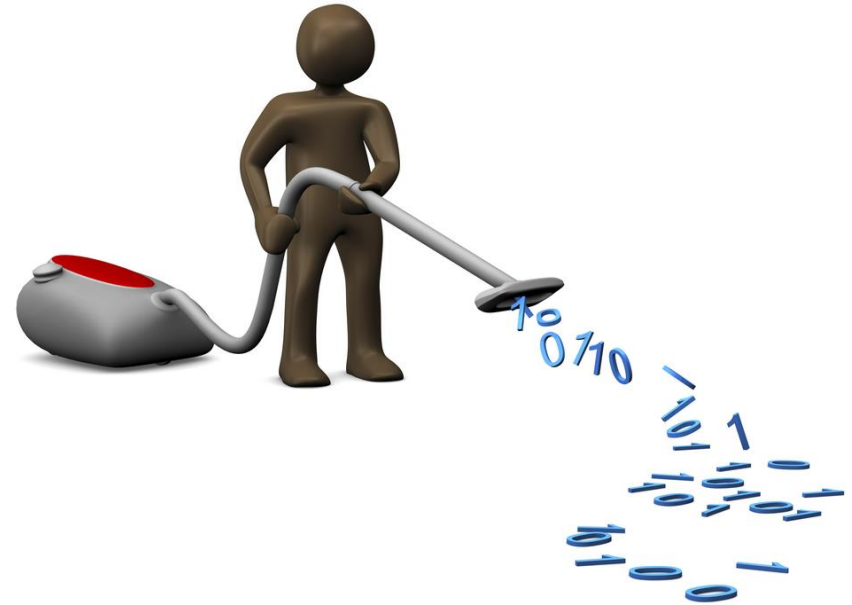
## How to do it?

Delete less valuable data

Aggregation (e.g. discard details but keep stats)

Keep data of different legal entities easily separable

Rotation (e.g. log rotation)







# **Analytics**

# Time series

## What is time series

A list of data points indexed in time

E.g. the readings of a sensor (or many sensors)

## What to do with them?

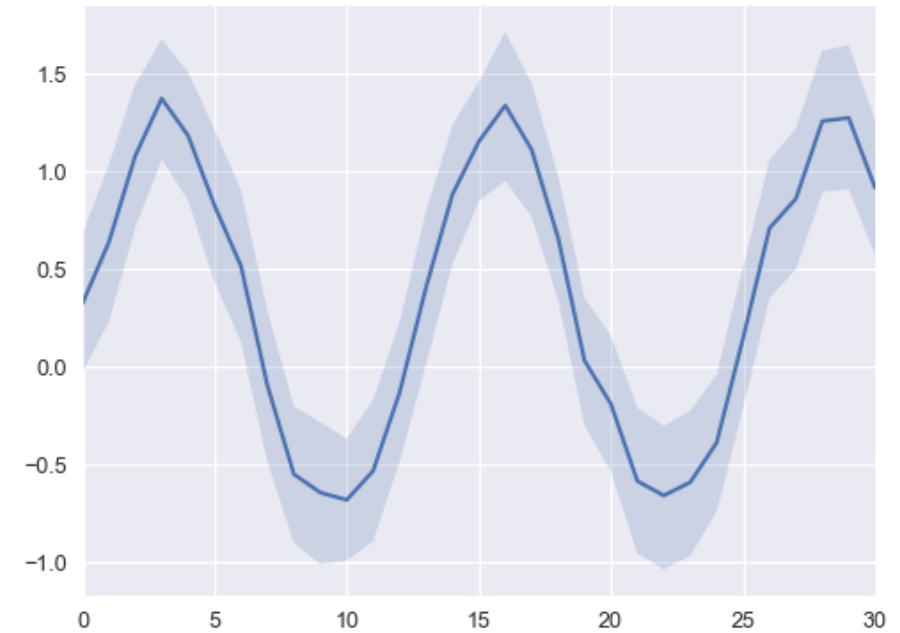
Extract high level features

- Summary: Min, Max, Med, Quantiles, Std.dev ...
- Spectrum analysis and transformations (e.g. FFT, filters)
- Prediction, Classification, Anomaly detection, Clustering

ML models: ARMA/ARIMA, RNN/LSTM

## Visualization

Demo with Jupyter & Seaborn



# Exercises