# Introduction to web crawling with StormCrawler (and Elasticsearch)

## 1 day workshop


## Overview

In this workshop, we will explore StormCrawler a collection of resources for building low-latency, large scale web crawlers on Apache Storm. After a short introduction to Apache Storm and an overview of what Storm-Crawler provides, we'll put it to use for a simple crawl before moving on to the deployed mode of Storm.

In the second part of the session, we will introduce metrics and index documents with Elasticsearch and Kibana and dive into data extraction. Finally, we'll cover recursive crawls and scalability. This course will be hands-on: attendees will run the code on their own machines.

## Duration

1 full day

## Agenda

We will cover the following topics:

- Introduction to web crawling
- Apache Storm: architecture and concepts
- StormCrawler: basic building blocks
- How to use the archetype
- Building & configuring
- URLFilters, ParseFilters
- Simple recursive crawls
- How to debug?
- Distributed mode: UI, logs, metrics
- Elasticsearch resources
  - document indexing
  - status index
  - metrics
  - Kibana
- Q&As

digitalPebble

## Audience

This course will suit Java developers with an interest in big data, stream processing, web crawling and search. It will provide a practical introduction to both Apache Storm and Elasticsearch as well of course as StormCrawler and will not require advanced programming skills.

## Prerequisites

Attendees should bring their own laptop with Apache Maven and Java 8 or above installed. The examples and instructions will be conducted on a Linux distribution and using Eclipe IDE.

Ideally, students should look at the Apache Storm and StormCrawler documentation and think about particular websites or crawl scenarios that they might be interested in.

## Trainer profile

Julien is Director of DigitalPebble Ltd and is a member of the Apache Software Foundation.

His expertise is in document engineering with a strong focus on open source tools. Julien has successfully designed and implemented solutions for Information Retrieval, Text Analysis, Information Extraction, Machine Learning, Web Crawling and Big Data for DigitalPebble's clients.

Julien contributes to several open source projects, including Apache Nutch, Tika, GATE, UIMA, CrawlerCommons and is the author of projects, such as Behemoth and StormCrawler which are used by numerous organisations worldwide.

He also speaks regularly at conferences and has reviewed several books.

**http://www.digitalpebble.com**

**https://twitter.com/digitalpebble**

**https://uk.linkedin.com/in/julien-nioche-4b7b453**