It all starts with collecting information from various sources. This is done with web crawlers or scrapers depending on the news you want to create from social media or news agencies.

DataSource Ingestion

Content Extraction

Since we’re handling large amounts of data, mostly of textual nature, it is important to store your findings from the previous steps in an easily accessible fashion.

Data

HBase is a NoSQL database that remains consistent even when network partitions occur. It strives to please random read/write access to large amounts of data.

Apache Spark is a framework for cluster computing. Spark Streaming is made for the processing of large incoming data streams and lets you easily transform packages at will.

Apache Solr is a search server based on Lucene. It has advanced full-text search capabilities and is therefore very suitable for this purpose.

Measuring engagement is a crucial aspect in planning future releases. It lets you see what content was received positively or what might have not boomed as expected.

Kafka can be a helpful tool in distributing the large data streams incoming from the searches. It can help balance the load and store data durably to disk and is known for storing streams in a very fault-tolerant way.

From all the collected data it is now time to create content. We can monitor topics or find related content, or do a keyword search.

From all the collected data it is now time to create content. There is a plethora of applications that can be built on top of the existing data.

Abstract

Inspired by the talks of Stefan Wehrmeyer and Peter Aldolfs the purpose of this poster is to visualize the modern journalistic process and how we can efficiently support it with technologies available today. I try to explain each step from ingesting data sources to publishing and engagement of the reader and a technology used to do certain tasks related to these steps.

Sarah Isabel Amon
Ringvorlesung - Data Engineering in der Praxis
WS17/18
E-Mail: sarah-isabel.amon@hpi.de