adaptTo()

APACHE SLING & FRIENDS TECH MEETUP
BERLIN, 26-28 SEPTEMBER 2016

Analyze JCR and Log Data with Apache Spark

Daniel Schley, pro!vision GmbH

Goals

- You did what?
- Why should I bother?
- How does AEM fit into a big data (big) picture?

You did what?

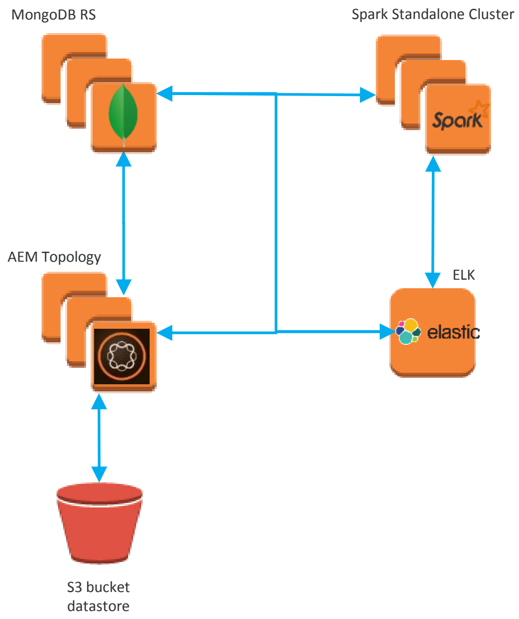
07.08.2016T12:30	some_key	4.856
07.08.2016T12:31	some_key	2.123
07.08.2016T12:32	some_key	5.234
07.08.2016T12:33	some_key	4.449
07.08.2016T12:34	some_key	4.638
…	…	…

- Queried MongoDB NodeDataStore collection (up to ~ 700 m)
- Iterative creating a time series covering 5 months, data for each minute
- Wait

You did what?

- “**Apache Spark** is a fast and general engine for large-scale data processing.”
- Get familiar with it
- Reduced the runtime from **> 1 week to ~3,5 minutes**

You did what?



- AEM topology (3 nodes)
- MongoDB RS (3 nodes)
- Elk
- Spark Standalone (4 nodes, 16 cores, ~ 50 GB)
- Custom log files & configurations

Demo
Combining JCR and Log Data

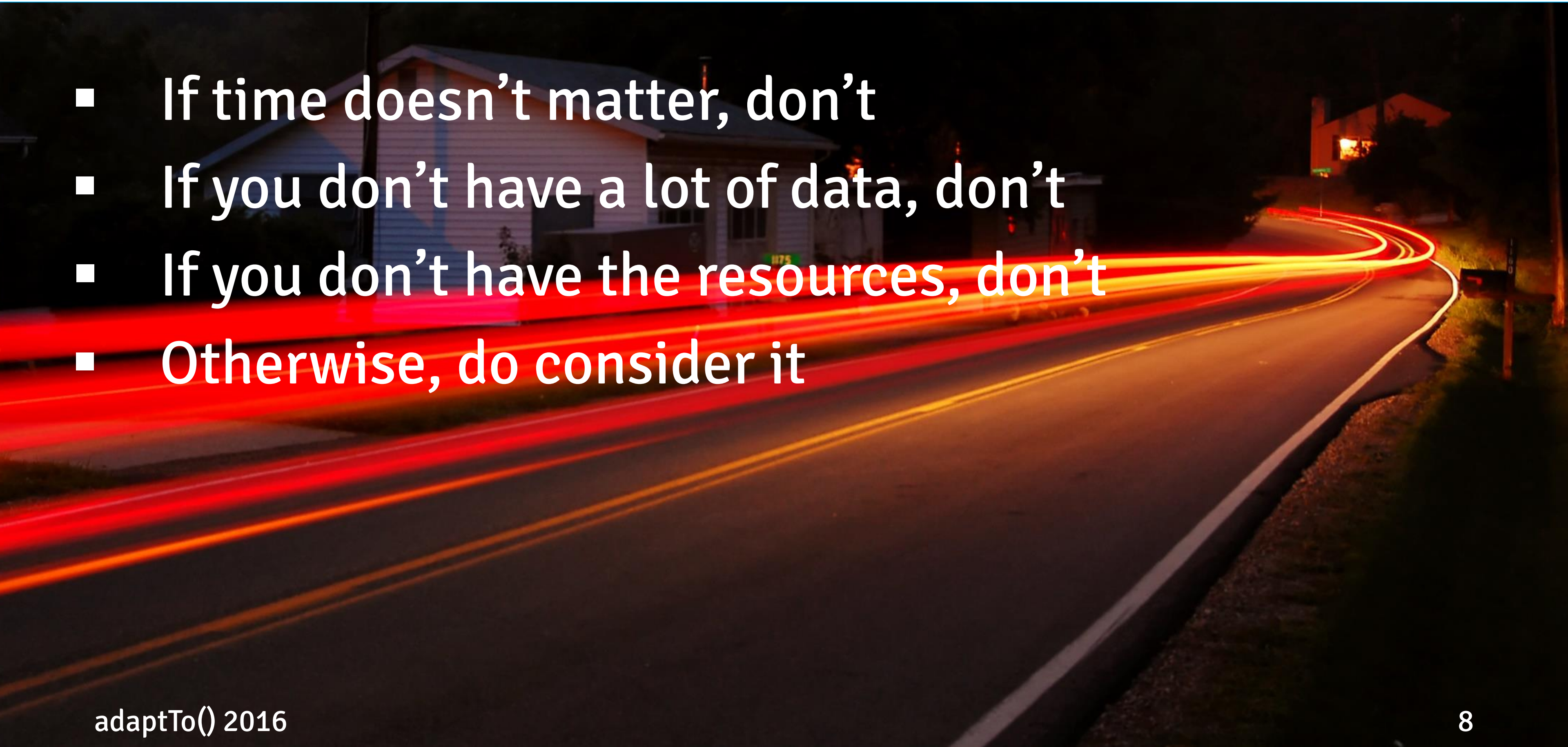
Result

- MongoDB documents

```
{
        "_id" : "/content/dam/assets/20-indd-24690101.indd",
        "modified" : "1466454695",
        "timestamp" : "2016-07-28 03:31:13.443",
        "action" : "createWfModel",
        "timeSpent" : "1"
}
```

- All actions from the (ES indexed) log associated with the node document from *aem-author* MongoDB collection

Why should I bother?



- If time doesn’t matter, don’t
- If you don’t have a lot of data, don’t
- If you don’t have the resources, don’t
- Otherwise, do consider it

adaptTo()

Questions?

Links

- MongoDB Spark Connector

  https://www.mongodb.com/blog/post/the-new-mongodb-connector-for-apache-spark-in-action-building-a-movie-recommendation-engine
- ElasticSearch for Spark

  https://www.elastic.co/guide/en/elasticsearch/hadoop/master/install.html
- MongoDB Spark Course (MongoDB University)

  https://github.com/breinero/MongoDB_Spark_Course