Data Engineering

Bringing Data Science to Scale

Hasso Plattner Institut
Digital Engineering • Universität Potsdam

Hendrik Schmidt
Bachelor

At the beginning of this lecture series, I could not really distinguish the concepts “Data Science” and “Data Engineering”. However, over the course of this semester and thanks to the speakers, I was more and more able to discern the difference between the two. As I now understand it, Data Engineering is what is needed to bring Data Science to scale in the age of Big Data. It is not so much that the core principles have changed, but rather that Data Engineering describes the methods and technologies that are employed to get the algorithms, research and questions of Data Science to production environments that need to be more stable and are way too big to run on single machines. And for Big Data, this definition fits: *“When your data sets become so large and diverse that you have to start innovating around how to collect, store, process, analyze and share them.”.* [1]

This poster now aims to get this point across. It is what I would show to my former self or people facing similar questions to get them to understand the difference between the two. It tries to clear up the many different concepts by condensing knowledge of the methods of Data Engineering. It describes and summarizes many of the technologies and steps in the pipelines the lectures spoke about. It groups similar terms and software solutions that go hand in hand and gives explanations for them. The pipeline and technologies are shown for an exemplary process of a fictitious airline. This process is not to be understood as realistic, but rather as a way to include many different softwares and steps, relying heavily on the Apache Software Foundations projects. The actual first step of course would be to find a goal: The airline for example is interested in flying more efficiently and maximising ticket sales. Then, the following could occur.

1. Data Generation

Similar: Data Capture

The raw data has to come from some kind of sources. An important question is: *“What changes faster? Data or Query?”* [2], to determine whether stream or batch processing is the better approach.

For the airline, these sources are airplane sensors, textual data from user interactions and database dumps from a flight information server.

2. Data Collection

Similar: Data Ingestion, Data Extraction

It is a good idea to collect and store as much data as possible. New questions can come up all the time, often depending on vast amounts of data from the past.

The airline deploys on Amazon Web Services to be able to easily react to changing demands. It uses Apache kafka for managing the data streams generated by the sensors, Apache Solr for storing the textual data from the users and Apache HBASE for the flight information.

3. Data Curation

Similar: Data Preparation, Data Integration

This step includes selecting the relevant data (using Data Profiling and Data Exploration), cleaning it (getting rid of null values, faulty entries) and transforming it into the correct shape.

For this task, the airline utilizes Apache Beam for the data processing. Beam runs Apache Flink for managing information from the sensor streams and Apache Spark for doing calculations on the databases, while simultaneously cleansing the inputs. Both are run distributed on local clusters and transform the data into the needed shape to write into a unified local Apache Cassandra database.

4. Data Mining

With the cleaned data now in the correct shape, one can beginn to mine it for new patterns and discoveries.

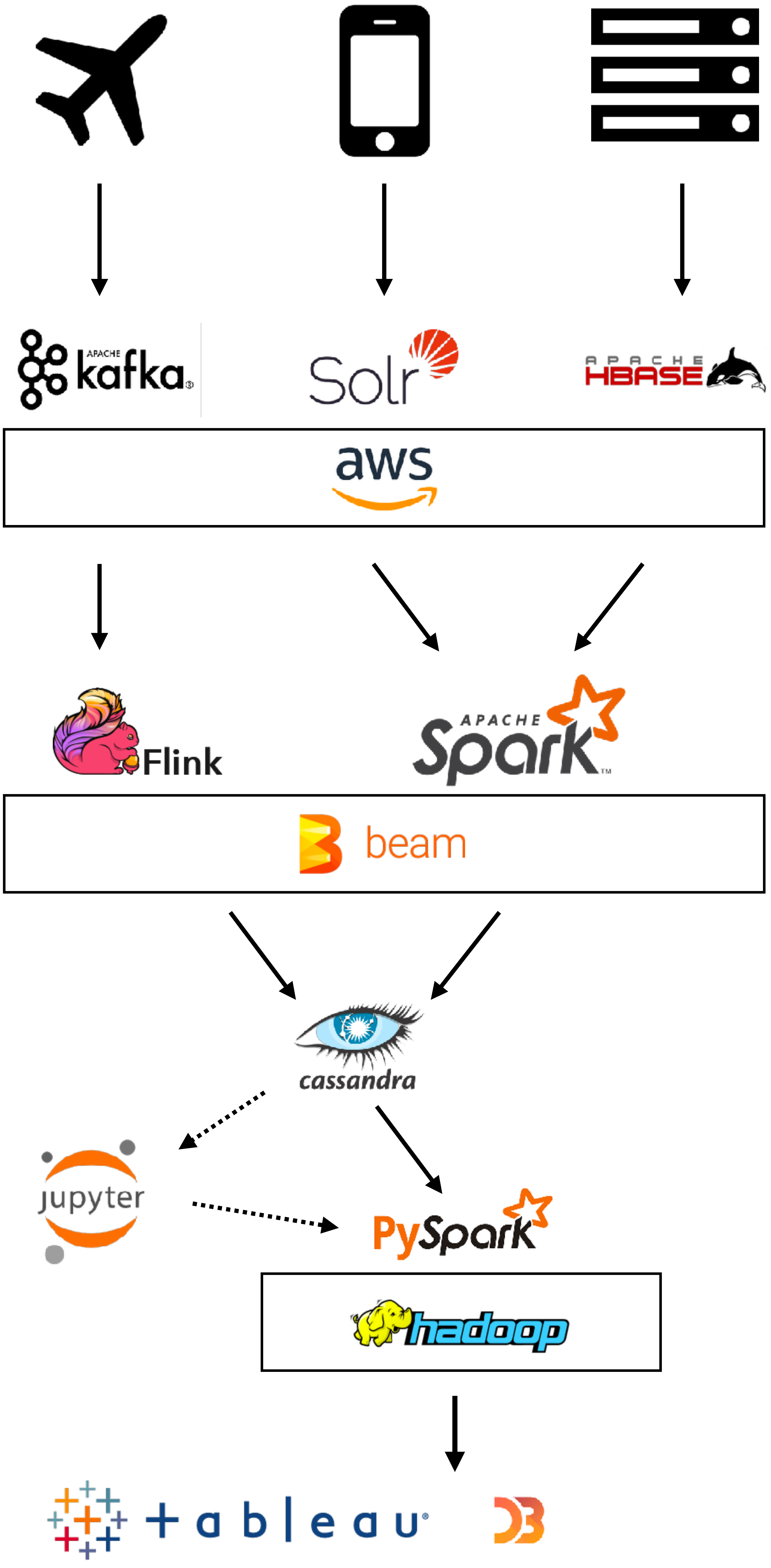
The airlines researchers first test new machine learning models and other algorithms with smaller amounts of data in Jupyter notebooks on single machines. When they are confident with their programs, the code gets ported to Apache Pyspark and is run on a local Apache Hadoop cluster, using YARN.

5. Data Interpretation

Similar: Knowledge Discovery

Finally, for any action to emerge from the gained knowledge, the data has to be interpreted by those in charge.

The airlines researchers use visualisations with tableau and D3 to present their results to the managers.



[1] Exabytes for Breakfast, Martin Grund
[2] Modern Stream Processing with Apache Flink, Fabian Hüske