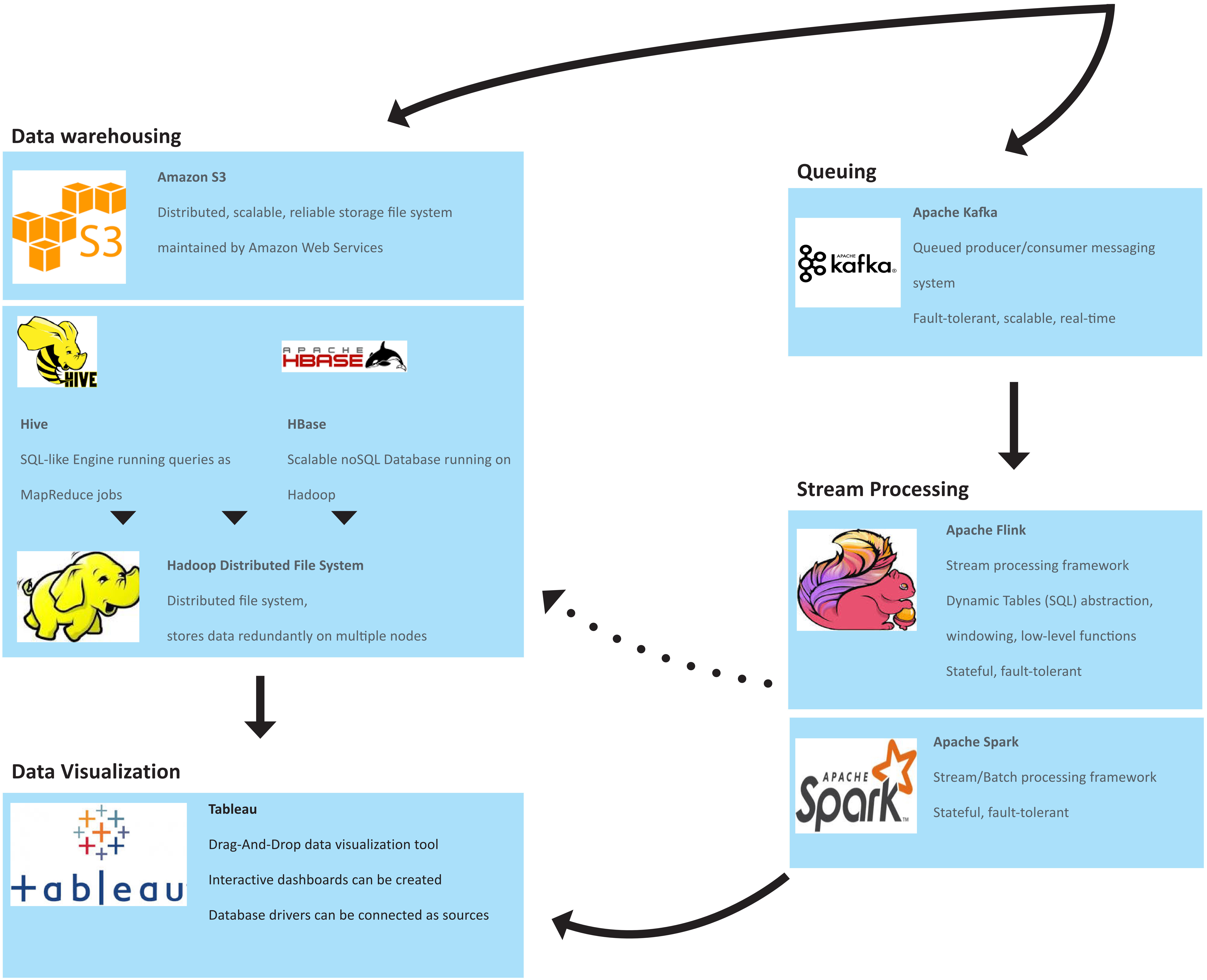
Overview Of A Data Engineering Technology Stack

Inspired by the lecture „Data Engineering in der Praxis“
Emanuel Metzenthin

Abstract: This poster aims to give an overview of a set of frameworks and products that could be used in an industrial data engineering process. Originating from various different sources the data has to be gathered in a data warehouse in order to be persisted and analyzed. Information coming in in real-time can be processed in a stream and stored afterwards. Ultimately the generated knowledge has to be visualized to be useful for the end-user. This collection does by no means provide a whole picture of the available technologies. It rather presents a choice of common tools and categorizes them into the data engineering process phases.



Data warehousing

Amazon S3

Distributed, scalable, reliable storage file system maintained by Amazon Web Services

Hive

SQL-like Engine running queries as MapReduce jobs

HBase

Scalable noSQL Database running on Hadoop

Hadoop Distributed File System

Distributed file system,

stores data redundantly on multiple nodes

Queuing

Apache Kafka

Queued producer/consumer messaging system

Fault-tolerant, scalable, real-time

Stream Processing

Apache Flink

Stream processing framework

Dynamic Tables (SQL) abstraction,

windowing, low-level functions

Stateful, fault-tolerant

Apache Spark

Stream/Batch processing framework

Stateful, fault-tolerant

Data Visualization

Tableau

Drag-And-Drop data visualization tool

Interactive dashboards can be created

Database drivers can be connected as sources

Projektpartner

Ringvorlesung „Data Engineering in der Praxis“ 2018

Dr. Krestel, Prof. Müller, Prof. Naumann,

Dr. Uflacker

Projektbeteiligte

Emanuel Metzenthin

Bachelor, IT-Systems Engineering