THE ZOO OF DATA ENGINEERING

A RANDOM WALK — PART II

TEZ

Tez is an open-source application framework that enables processing data in the schema of a complex pipeline (directed acyclic graph). It is also part of the Hadoop stack by being set atop of the YARN resource manager.

HIVE

Hive is an open-source tool for data warehousing in the Hadoop ecosystem.
It manages huge amounts of data by summarizing, querying and analyzing.
A common problem is the performance drop when it comes to large scaling. The queries need to be thoroughly optimized to reach the desired performance.

BAKDATA · XING

HDFS

The open-source Hadoop Distributed File System provides high speed access to data by saving it efficiently redundant on multiple nodes in a cluster. A common problem is that good performance becomes hard to reach with highly scaled data volumes due to sensible configuration issues.

BAKDATA · NEOFONIE · XING

FILE SYSTEM

BATCH PROCESSING

PIG

Pig is a high-level analyzing platform for applications running in the Hadoop ecosystem. It delivers a lot of built-in functionality which has – on a downside – a high percentage of not expedient methods. Maybe the community can contribute to a common-practice-set?

BAKDATA

GIRAPH

Giraph is an open-source graph processing system with a focus on high scalability. It's strength lies in generating graph models out of structured datasets at a massive scale.

UNIFIED PROCESSING

FLINK

Flink is an open-source stream processing framework, optimal for clusters running real-time applications. Certainly its deployment is not very flexible and can't be exported to other systems via e.g. Docker. So sometimes handling streams might be easier with a library like Kafka Stream.

ABSTRACT

Giraffes, elephants, pigs, hornets, and squirrels: You have probably seen all of them while looking at tools for data engineering tasks. As playful and inviting as the logos might look, what do they actually represent? What are potential pitfalls of deciding on using the one or the other?

This poster tries to give a brief introduction to the different engineering utensils that borrowed their logo from the animal kingdom. Besides giving a brief overview of their respective functions, the animals are also divided into their respective enclosures, thus clustering them by field of use. The placard also tries to raise possible issues with some of the presented technology.

STUDENT

Florian Wirtz
IT-Systems Engineering (Bachelor)

ADVISORS

Dr. Ralf Krestel,
Prof. Dr. Emmanuel Müller,
Prof. Dr. Felix Naumann,
Dr. Matthias Uflacker

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam